SHARCNET™

Bioinformatics in the terminal

Tips and tricks to make your life easier

Common sequence formats

FASTA FORMAT

Description	>Sequence1
Sequence	AAATGACATCAGCAACATACCAAGTTTC
Description	>Sequence2
Sequence	CTATCCTTACGTTAGAATCGCATCGTGG

- Plain text of nucleotide sequences or peptides
- Sequences represented with single-letter codes.
- Single-line description, marked by “>”
- Description is followed by lines of sequence data
- Often sequence lines are wrapped to 80 characters in length.

FASTQ FORMAT

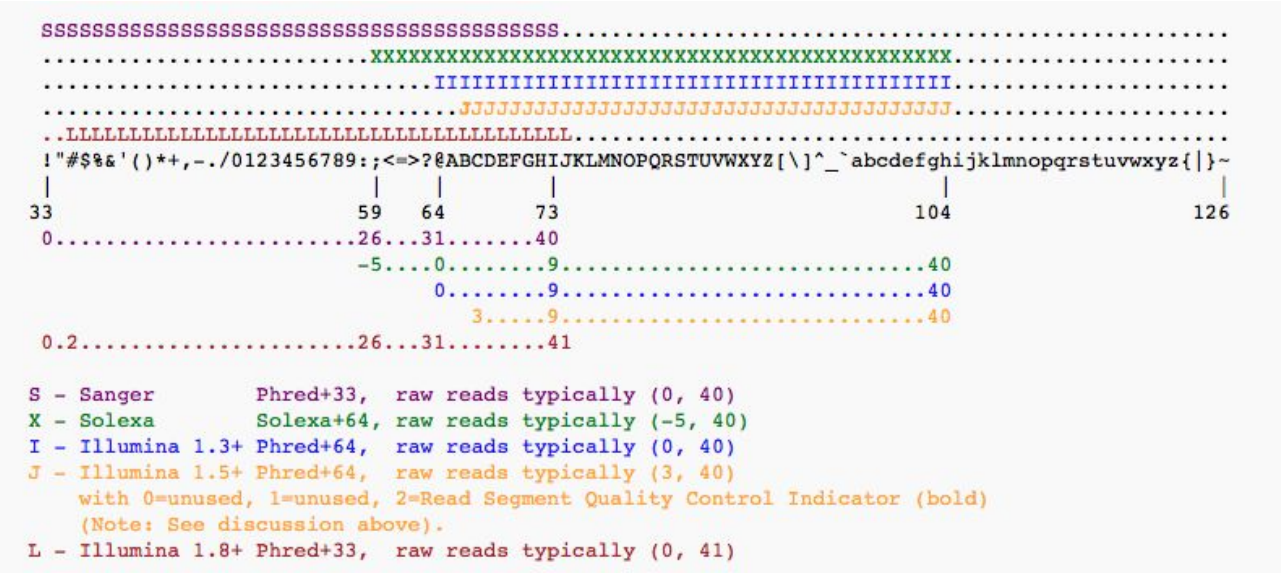
Description	@Sequence1
Sequence	AAAGGACAGCAGCAACATACCAAGTTTC
Info	+
Sequence	G=CGG=GG=GCCCGGCC=GGGCGCGJ$

- Plain text of nucleotide sequences
- Sequences represented with single-letter codes.
- Single-line description, marked by “@”
- Description is followed by lines of sequence data, then info line and finally quality (encoded) line.
- Not wrapped

Quality and its Encoding in FASTQ format

$$Q = -10\log_{10}(e)$$

Phred Q	P(incorrect)	Base call accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.9%
40	1 in 10000	99.99%
50	1 in 100000	99.999%
60	1 in 1000000	99.9999%



ASCII control characters

DEC	HEX	Simbolo ASCII	
00	00h	NULL	(carácter nulo)
01	01h	SOH	(inicio encabezado)
02	02h	STX	(inicio texto)
03	03h	ETX	(fin de texto)
04	04h	EOT	(fin transmisión)
05	05h	ENQ	(enquiry)
06	06h	ACK	(acknowledgement)
07	07h	BEL	(timbre)
08	08h	BS	(retroceso)
09	09h	HT	(tab horizontal)
10	0Ah	LF	(salto de linea)
11	0Bh	VT	(tab vertical)
12	0Ch	FF	(form feed)
13	0Dh	CR	(retorno de carro)
14	0Eh	SO	(shift Out)
15	0Fh	SI	(shift In)
16	10h	DLE	(data link escape)
17	11h	DC1	(device control 1)
18	12h	DC2	(device control 2)
19	13h	DC3	(device control 3)
20	14h	DC4	(device control 4)
21	15h	NAK	(negative acknowle.)
22	16h	SYN	(synchronous idle)
23	17h	ETB	(end of trans. block)
24	18h	CAN	(cancel)
25	19h	EM	(end of medium)
26	1Ah	SUB	(substitute)
27	1Bh	ESC	(escape)
28	1Ch	FS	(file separator)
29	1Dh	GS	(group separator)
30	1Eh	RS	(record separator)
31	1Fh	US	(unit separator)
127	20h	DEL	(delete)

ASCII printable characters

DEC	HEX	Simbolo	DEC	HEX	Simbolo	DEC	HEX	Simbolo
32	20h	espacio	64	40h	@	96	60h	`
33	21h	!	65	41h	A	97	61h	a
34	22h	"	66	42h	B	98	62h	b
35	23h	#	67	43h	C	99	63h	c
36	24h	$	68	44h	D	100	64h	d
37	25h	%	69	45h	E	101	65h	e
38	26h	&	70	46h	F	102	66h	f
39	27h	'	71	47h	G	103	67h	g
40	28h	(	72	48h	H	104	68h	h
41	29h	)	73	49h	I	105	69h	i
42	2Ah	*	74	4Ah	J	106	6Ah	j
43	2Bh	+	75	4Bh	K	107	6Bh	k
44	2Ch	,	76	4Ch	L	108	6Ch	l
45	2Dh	-	77	4Dh	M	109	6Dh	m
46	2Eh	.	78	4Eh	N	110	6Eh	n
47	2Fh	/	79	4Fh	O	111	6Fh	o
48	30h	0	80	50h	P	112	70h	p
49	31h	1	81	51h	Q	113	71h	q
50	32h	2	82	52h	R	114	72h	r
51	33h	3	83	53h	S	115	73h	s
52	34h	4	84	54h	T	116	74h	t
53	35h	5	85	55h	U	117	75h	u
54	36h	6	86	56h	V	118	76h	v
55	37h	7	87	57h	W	119	77h	w
56	38h	8	88	58h	X	120	78h	x
57	39h	9	89	59h	Y	121	79h	y
58	3Ah	:	90	5Ah	Z	122	7Ah	z
59	3Bh	;	91	5Bh	[	123	7Bh	{
60	3Ch	<	92	5Ch	\	124	7Ch	|
61	3Dh	=	93	5Dh	]	125	7Dh	}
62	3Eh	>	94	5Eh	^	126	7Eh	~
63	3Fh	?	95	5Fh	_			

Extended ASCII characters

DEC	HEX	Simbolo	DEC	HEX	Simbolo	DEC	HEX	Simbolo	DEC	HEX	Simbolo
128	80h	Ç	160	A0h	á	192	C0h	└	224	E0h	Ó
129	81h	ü	161	A1h	í	193	C1h	┴	225	E1h	ß
130	82h	é	162	A2h	ó	194	C2h	┬	226	E2h	Ô
131	83h	â	163	A3h	ú	195	C3h	├	227	E3h	Ò
132	84h	ä	164	A4h	ñ	196	C4h	─	228	E4h	õ
133	85h	à	165	A5h	Ñ	197	C5h	┼	229	E5h	Õ
134	86h	å	166	A6h	ª	198	C6h	ã	230	E6h	µ
135	87h	ç	167	A7h	º	199	C7h	Ã	231	E7h	þ
136	88h	ê	168	A8h	¿	200	C8h	╚	232	E8h	Þ
137	89h	ë	169	A9h	®	201	C9h	╔	233	E9h	Ú
138	8Ah	è	170	AAh	¬	202	CAh	╩	234	EAh	Û
139	8Bh	ï	171	ABh	½	203	CBh	╦	235	EBh	Ù
140	8Ch	î	172	ACh	¼	204	CCh	╠	236	ECh	ý
141	8Dh	ì	173	ADh	¡	205	CDh	═	237	EDh	Ý
142	8Eh	Ä	174	AEh	«	206	CEh	╬	238	EEh	¯
143	8Fh	Å	175	AFh	»	207	CFh	¤	239	EFh	´
144	90h	É	176	B0h	░	208	D0h	ð	240	F0h	
145	91h	æ	177	B1h	▒	209	D1h	Ð	241	F1h	±
146	92h	Æ	178	B2h	▓	210	D2h	Ê	242	F2h	‗
147	93h	ô	179	B3h	│	211	D3h	Ë	243	F3h	¾
148	94h	ò	180	B4h	┤	212	D4h	È	244	F4h	¶
149	95h	ò	181	B5h	Á	213	D5h	ı	245	F5h	§
150	96h	û	182	B6h	Â	214	D6h	Í	246	F6h	÷
151	97h	ù	183	B7h	À	215	D7h	Î	247	F7h	¸
152	98h	ÿ	184	B8h	©	216	D8h	Ï	248	F8h	°
153	99h	Ö	185	B9h	╣	217	D9h	┘	249	F9h	¨
154	9Ah	Ü	186	BAh	║	218	DAh	┌	250	FAh	·
155	9Bh	ø	187	BBh	╗	219	DBh	█	251	FBh	¹
156	9Ch	£	188	BCh	╝	220	DCh	▄	252	FCh	³
157	9Dh	Ø	189	BDh	¢	221	DDh	¦	253	FDh	²
158	9Eh	×	190	BEh	¥	222	DEh	Ì	254	FEh	■
159	9Fh	ƒ	191	BFh	┐	223	DFh	▀	255	FFh	

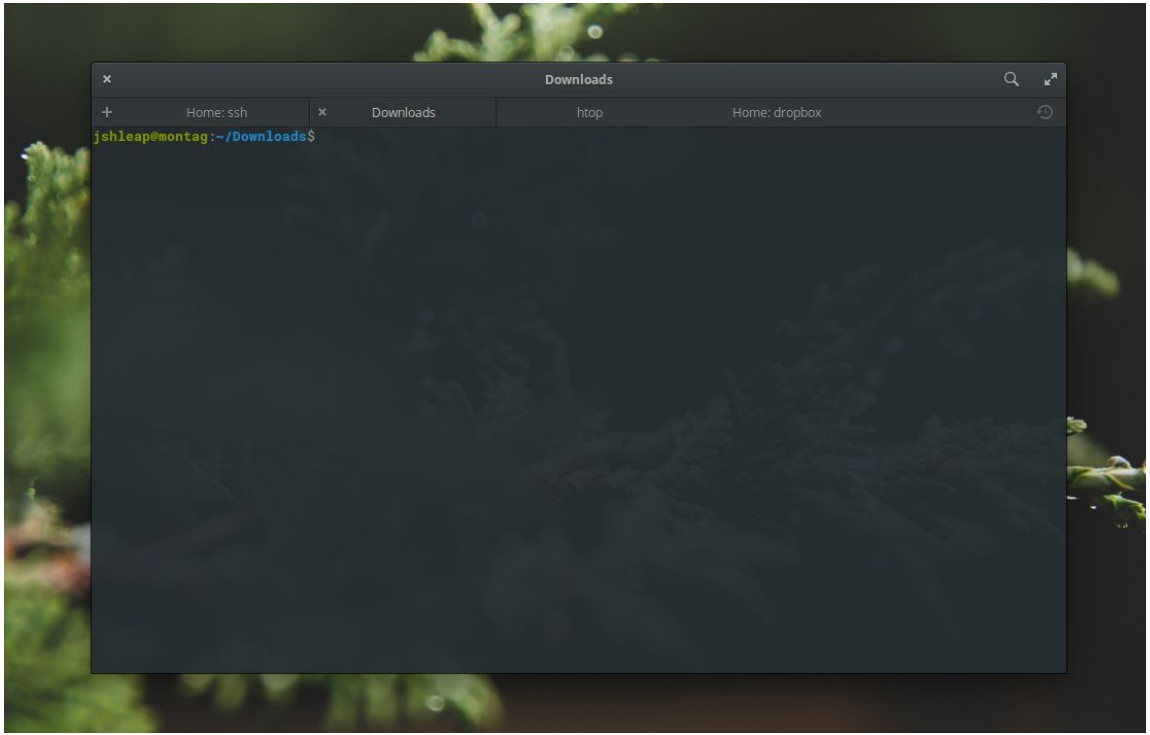
Tabular format

Eg. blastn -db adb -query afasta.fa -out ablast.txt -outfmt "6 qaccver saccver pident qcovs length evalue"

sq1;size=59022;	sq1;size=59022	100.000	100	313	1.33e-168
sq2;size=30981;	sq2;size=30981	100.000	100	313	1.33e-168
sq2;size=30981;	sq28;size=15	88.782	99	312	1.12e-109
sq2;size=30981;	sq22;size=31	84.345	100	313	6.92e-87
sq3;size=13885;	sq3;size=13885	100.000	100	313	1.33e-168
sq3;size=13885;	sq22;size=31	91.667	99	312	1.09e-124
sq3;size=13885;	sq28;size=15	87.859	100	313	3.13e-105
sq4;size=13553;	sq4;size=13553	100.000	100	313	1.33e-168
sq4;size=13553;	sq19;size=34	81.553	99	309	2.54e-71
sq4;size=13553;	sq5;size=1760	80.192	100	313	3.31e-65
sq5;size=1760;	sq5;size=1760	100.000	100	313	1.33e-168
sq5;size=1760;	sq20;size=33	81.529	100	314	7.07e-72
sq5;size=1760;	sq4;size=13553	80.192	100	313	3.31e-65
sq6;size=307;	sq6;size=307	100.000	100	313	1.33e-168
sq7;size=267;	sq7;size=267	100.000	100	313	1.33e-168
sq8;size=258;	sq8;size=258	100.000	100	313	1.33e-168
sq9;size=258;	sq9;size=258	100.000	100	313	1.33e-168
sq10;size=231;	sq10;size=231	100.000	100	313	1.33e-168
sq10;size=231;	sq12;size=127	91.054	100	313	6.59e-122
sq10;size=231;	sq27;size=21	89.744	99	312	1.11e-114

The terminal

- Bash
- First contact with cluster
- Command line
- Let’s check it out (if you haven’t)



Common basic commands

- **ls**: list files/directories
- **cd**: change directory
- **rm**: Remove file/directory
- **wc**: word count, line count
- **pwd:** print working directory
- **mkdir**: Make a directory
- **nano**: Open file editor
- **wget**: web get
- **cat**: Print/concatenate files
- **head**: Print first n lines
- **tail**: Print last n lines
- **more/less**: Interactively read file by page

Dealing with tabular data

Say you have a blast result from the command:

```
blastn -db adb -query afasta.fa -out ablast.txt -outfmt "6 qaccver saccver pident qcovs length evalue"
```

Can we subset the columns?

Commands to use:

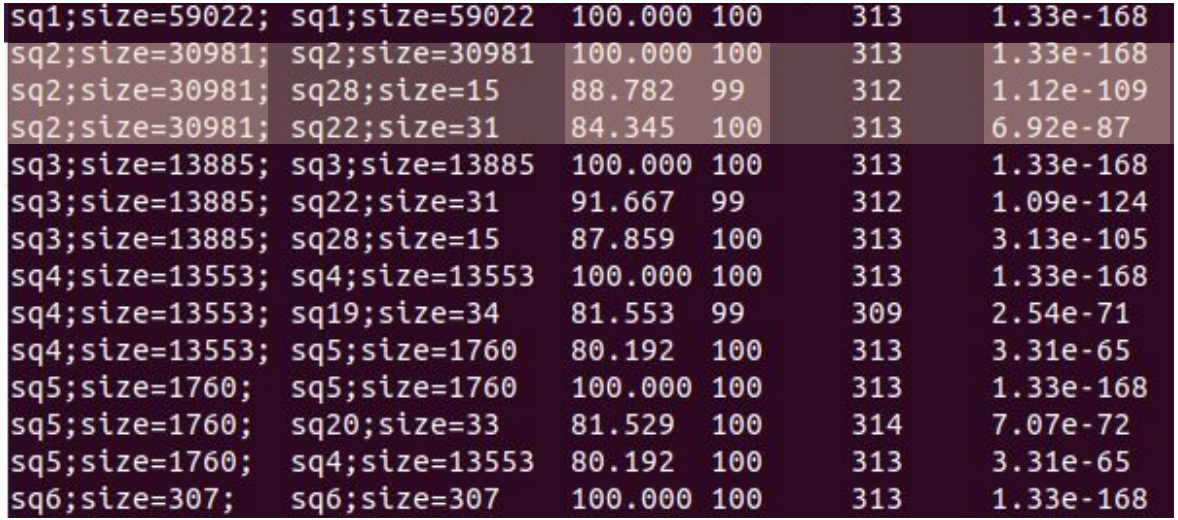
- cut, or awk

Dealing with tabular data

What about subsetting both the rows and columns?

Commands to use:

- ★ awk



```
sq1;size=59022;  sq1;size=59022   100.000 100     313     1.33e-168
sq2;size=30981;  sq2;size=30981   100.000 100     313     1.33e-168
sq2;size=30981;  sq28;size=15     88.782  99      312     1.12e-109
sq2;size=30981;  sq22;size=31     84.345  100     313     6.92e-87
sq3;size=13885;  sq3;size=13885   100.000 100     313     1.33e-168
sq3;size=13885;  sq22;size=31     91.667  99      312     1.09e-124
sq3;size=13885;  sq28;size=15     87.859  100     313     3.13e-105
sq4;size=13553;  sq4;size=13553   100.000 100     313     1.33e-168
sq4;size=13553;  sq19;size=34     81.553  99      309     2.54e-71
sq4;size=13553;  sq5;size=1760    80.192  100     313     3.31e-65
sq5;size=1760;   sq5;size=1760    100.000 100     313     1.33e-168
sq5;size=1760;   sq20;size=33     81.529  100     314     7.07e-72
sq5;size=1760;   sq4;size=13553   80.192  100     313     3.31e-65
sq6;size=307;    sq6;size=307     100.000 100     313     1.33e-168
```

Dealing with tabular data

Can you filter by column value?

```
awk '{if ($3>=90 && $4 > 99 && $6<1E-70 ) {print $0}}'  file
```

Commands to use:

- ★ awk

```
sq1;size=59022; sq1;size=59022  100.000 100     313     1.33e-168
sq2;size=30981; sq2;size=30981  100.000 100     313     1.33e-168
sq2;size=30981; sq28;size=15    88.782  99      312     1.12e-109
sq2;size=30981; sq22;size=31    84.345  100     313     6.92e-87
sq3;size=13885; sq3;size=13885  100.000 100     313     1.33e-168
sq3;size=13885; sq22;size=31    91.667  99      312     1.09e-124
sq3;size=13885; sq28;size=15    87.859  100     313     3.13e-105
sq4;size=13553; sq4;size=13553  100.000 100     313     1.33e-168
sq4;size=13553; sq19;size=34    81.553  99      309     2.54e-71
sq4;size=13553; sq5;size=1760   80.192  100     313     3.31e-65
sq5;size=1760;  sq5;size=1760   100.000 100     313     1.33e-168
sq5;size=1760;  sq20;size=33    81.529  100     314     7.07e-72
sq5;size=1760;  sq4;size=13553  80.192  100     313     3.31e-65
sq6;size=307;   sq6;size=307    100.000 100     313     1.33e-168
```

SHARCNET™

Dealing with tabular data

Don’t like tab-delimited format? No problem:

Note:

In unix tab is represented by \t

Commands to use:

- ★ sed

How many sequences do I have?

It is all about figuring out the pattern:

- Fasta files' headers always start with '>'
- FastQ files headers always start with '@'
  - OK?

Commands to use:

- grep (you might do the same with awk)

Are they unique?

Often you might have composite files, how to get the unique headers and their counts?

Commands to use:

- grep
- sort

Make a fasta sequence single lined?

This is a bit more complex, so we will use this awk code:

```
awk '/^>/{print s?
s"\n"$0:$0;s="";next}{s=s
$0}END{if(s)print s}' infasta > outfasta
```

Let's break it down:

Find lines that start with ">" → `/^>/ {`

Print s + EOL + line if s is set else print line → `print s ? s "\n" $0 : $0;`

Reset s to empty string → `s = "";`

Seek for next entry that matches pattern → `next;}`

`{`

Append lines that do not start with ">" into s → `s = s $0;`

`}`

Before the program ends, this block print the last s stored → `END {`

`if (s) print s;}`

```
>Sequence1
AAATGACATCAGCAA
CATACCAAGTTTCTAC
GTTAGAATC
>Sequence2
CTATCCTTACGTTAGA
ATCGCATCGTGG
```

Convert a fastQ into a fastA

Often you would like to convert a fastq into a fasta.
Can you do that in pure bash?

YES!! Sed is your friend:

```
sed -n '1~4s/^@/>/p;2~4p'
```

Commands to use:

- ★ sed

There is a program for that!

Actually many, but let's look at seqkit:

https://bioinf.shenwei.me/seqkit/

Useful resources

https://github.com/stephenturner/oneliners

https://www.grymoire.com/Unix/index.html

https://wiki.bash-hackers.org/scripting/tutoriallist