

Doing more in UNIX: Command-line tools

HORT 530

Lecture 3

Instructor: Kranthi Varala

UNIX features

- Command-line based.
- Supports thousands of small programs running simultaneously.
- Easy to create pipelines from individual programs.
- Each command has 3 Input/Output streams: STDIN, STDOUT, STDERR.
- Be aware of your location in the file system and permissions.

Control over process

- Foreground: Default mode for running commands. The shell waits on the process to finish.
 - Process retains control of the command line.
 - Key input is directed to the active process.
- Background: Process is initiated and pushed to the background.
 - Control of command-line is returned to the user.
 - Key input and other interactions are no longer passed to the process.
 - Processes can be pushed to background at initiation using &
 - `cat North_of_Boston.txt &`

Stopping a process

- Active processes can be stopped or terminated (killed) using the SIGSTOP and SIGKILL signals.
- SIGSTOP in UNIX shell is issued by `ctr ̣+z`
- Once a process receives SIGSTOP it is suspended and the job number `[j]` is shown.
- Suspended processes can be pushed to background using the command `bg %j`.

```
kvarala@scholar-fe06: $ wc SRR6473489.fastq
^Z
[1]+  Stopped                  wc SRR6473489.fastq
kvarala@scholar-fe06: $ bg %1
[1]+ wc SRR6473489.fastq &
kvarala@scholar-fe06: $ ls
GSE49418_series_matrix.tsv  GSE49418_series_matrix.txt  North_of_Boston.txt
```

```
kvarala@scholar-fe06: $      4751112   9502224 268304258 SRR6473489.fastq
[1]+  Done                    wc SRR6473489.fastq
```

Killing a process

- Active processes can be stopped or terminated (killed) using the SIGSTOP and SIGKILL signals.
- SIGKILL in UNIX shell is given by `ctrl+c`
- SIGKILL kills the process immediately and returns control to the user.
- Stopped processes can be killed using `kill` command.

```
kvarala@scholar-fe06: $ wc SRR6473489.fastq
^Z
[1]+  Stopped                  wc SRR6473489.fastq
kvarala@scholar-fe06: $ bg %1
[1]+  wc SRR6473489.fastq &
kvarala@scholar-fe06: $ kill %1
kvarala@scholar-fe06: $
[1]+  Terminated              wc SRR6473489.fastq
```

Monitoring processes

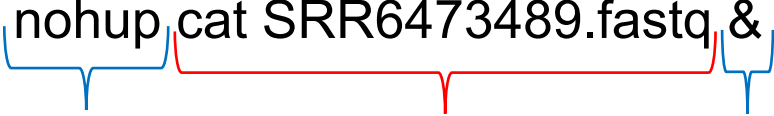
- Processes in the current shell can be listed using the `jobs` command.
- `SIGKILL` can then be issued for any job using the `kill %j` command where `j` is the job number.
- To list all processes on the current machine use the `ps` command.
- E.g., `ps -ae` gives a snapshot of all processes on the current machine.
- A more dynamic view is given by the command `top`.

Tasks: 2402 total, 6 running, 2394 sleeping, 0 stopped, 2 zombie
 %Cpu(s): 6.4 us, 1.3 sy, 0.0 ni, 92.2 id, 0.1 wa, 0.0 hi, 0.0 si, 0.0 st
 KiB Mem : 79122854+total, 69281785+free, 81134160 used, 17276488 buff/cache
 KiB Swap: 23436284 total, 23436284 free, 0 used. 69297376+avail Mem

PID	USER	PR	NI	VIRT	RES	SHR	S	%CPU	%MEM	TIME+	COMMAND
226045	chakra37	20	0	1664588	47268	1212	R	99.7	0.0	1126:00	Prob_B_1
119128	xie310	20	0	3275692	1.494g	66496	R	11.1	0.2	2453:30	Web Content
181229	gupta513	20	0	588172	21816	13668	S	9.5	0.0	402:50.56	Thunar
1	root	20	0	192620	5700	2524	S	5.9	0.0	620:03.67	systemd
227563	root	20	0	231300	12064	2108	R	5.2	0.0	7:38.31	python-thinlinc
1297	root	20	0	497056	12788	4296	S	4.6	0.0	539:05.12	cgroups_py
458487	swisherr	20	0	979584	121808	19800	S	4.3	0.0	771:36.35	rsession
245517	gupta513	20	0	2047500	295792	74912	S	3.3	0.0	168:43.79	firefox
179395	gupta513	20	0	99456	69680	19084	S	3.0	0.0	147:41.84	Xvnc
118082	ningb	20	0	2058556	267884	64192	S	1.6	0.0	23:39.74	Web Content
118924	xie310	20	0	2094740	357332	76868	S	1.6	0.0	438:33.60	firefox
121643	kvarala	20	0	162184	4680	1556	R	1.6	0.0	0:00.10	top
351998	oakleyc	20	0	2504216	175964	51456	S	1.6	0.0	68:58.91	rstudio.exe
60231	chou71	20	0	2591324	727184	60752	R	1.3	0.1	152:37.61	Web Content
245741	gupta513	20	0	2328344	513048	61704	S	1.3	0.1	55:18.25	Web Content
322072	ssuman	20	0	162908	5424	1580	S	1.3	0.0	133:50.04	top
74180	oakleyc	20	0	2724724	191436	61288	S	1.0	0.0	237:56.65	rstudio.exe
387457	mohan18	20	0	2041332	234244	81408	S	1.0	0.0	111:04.32	firefox
17567	tshera	20	0	472424	61764	9124	S	0.7	0.0	1:03.27	jupyterhub-sing
24720	asivasa	20	0	2024008	321180	72756	S	0.7	0.0	25:01.83	firefox
41851	gbonnett	20	0	3741376	258328	69992	S	0.7	0.0	32:30.91	spyder
181223	gupta513	20	0	440204	17912	11968	S	0.7	0.0	22:40.66	xfwm4
198289	jwisecav	20	0	2049692	290872	100780	S	0.7	0.0	13:28.55	firefox
224246	huan1368	20	0	2048980	361028	67360	S	0.7	0.0	39:06.68	Web Content
431731	rnandyma	20	0	88836	56096	9372	S	0.7	0.0	26:30.69	Xvnc
10	root	20	0	0	0	0	S	0.3	0.0	49:51.85	rcu_sched

Running long processes

- On a cluster we use the job/queue management systems to run long jobs. Eg., PBS system on Scholar.
- On a remote non-cluster server, you can initiate a process using the nohup command.
- nohup stands for no hangup, which means keep the process running even after the current shell closes.
- Remember to start nohup commands in background by using & at the end of the command.

- 

```
nohup cat SRR6473489.fastq &
```

No hangup Actual command Background

Command line tools

- Common tasks users perform are greatly helped by standard command-line tools in UNIX.
- Two most common user tasks are:
 - File manipulation
 - Text manipulation
- We learnt some file commands already:
 - `ls`, `cd`, `chmod`, `mkdir`, `cp`, `mv` etc.
- Other common tasks with files and folders are compression, archiving and linking.

Compression

- Files are compressed to reduce their size on the disk.
- Typically most efficient with compressing text files.
- gzip command is most commonly used to compress and expand files.
- Replaces original file with compressed file.

```
kvarala@scholar-fe06: $ ll SRR6473489.fastq
-rw-r--r-- 1 kvarala student 10838938338 Jan 22 13:17 SRR6473489.fastq
kvarala@scholar-fe06: $ gzip SRR6473489.fastq
kvarala@scholar-fe06: $ ll SRR6473489.fastq
ls: cannot access SRR6473489.fastq: No such file or directory
kvarala@scholar-fe06: $ ll SRR6473489.fastq.gz
-rw-r--r-- 1 kvarala student 1075635123 Jan 22 13:17 SRR6473489.fastq.gz
```

- bzip2 is an alternative compression algorithm that may provide more compression but takes more time to compress and expand.

Archiving

- Creates a single archive that contains multiple files and/or directories.
- tar is the most common archiving tool used in UNIX
- Supports compression via compression programs such as gzip and bzip2.
- `$ tar -cvzf TextFiles.tar.gz *txt`
- Creates an archive called TextFiles.tar.gz from all txt files in the current folder. Does NOT replace.
- `$ tar -xvzf TextFiles.tar.gz`
- Extract the files from TextFiles.tar.gz to current folder.
- Preserves the original directory structure

Links

- A link points to a file/directory on the file system.
- `$ ln -s SRR6473489.fastq Example.fastq`
 - Creates a link called Example.fastq

```
kvarala@scholar-fe06: $ ln -s SRR6473489.fastq Example.fastq
kvarala@scholar-fe06: $ ll Example.fastq
lrwxrwxrwx 1 kvarala student 16 Jan 22 14:05 Example.fastq -> SRR6473489.fastq
```

- Similar to the concept of shortcuts on Windows/OS X.
- Removing a link does not remove the original file.
- Removing the original file does not remove the link, only makes it non-functional.

Common text formats

- Simple text files contain blocks of text with no imposed structure beyond the line breaks.
 - Eg., `North_of_Boston.txt`
- Text files can also store tables with data arranged in rows and columns.
 - Defined column separators eg., `<TAB>`, Comma etc.
 - Each row is one data collection.
- Data may be arranged in blocks that span multiple lines.
 - Eg., FASTA and FASTQ formats in Biology.

Example tabular data

- Each row represents one gene.
- Each column represents expression of gene in that sample. Column separator <TAB>
- First row and First column contain respective labels.

```
kvarala@scholar-fe00:/scratch/scholar/k/kvarala/Week3/Files $ head -n20 GSE49418.top250.tsv
Gene      WT-CK1  WT-CK2  WT-CK3  MT-CK1  MT-CK2  MT-CK3
245041_at 7.6044173 8.447944 7.777406 8.246566 8.562556 8.739321
245119_at 7.8671436 8.00699 7.9079247 7.501124 8.054494 7.5438313
245247_at 8.140869 8.010652 8.181079 7.9125094 8.316227 8.123076
245250_at 7.742004 8.042652 7.3609304 7.384283 7.645018 7.692195
245329_at 6.0456758 6.0853715 6.7108784 7.7846346 8.612887 8.28583
245627_at 0.6136984 2.5572119 1.7915204 1.6312857 0.5355964 2.1927164
245765_at 9.176895 9.558297 10.050962 9.510457 9.776581 9.219471
245777_at 7.4506702 8.334939 6.9342628 7.6104965 7.508419 7.4884295
246270_at 10.288048 9.246562 9.370121 9.395946 9.554836 9.3737755
246289_at 10.853704 11.123942 11.493572 10.820894 11.0875845 10.795305
246490_at 6.854518 7.1305866 7.319211 8.059608 7.6570344 7.3294134
246495_at 8.385275 7.754309 7.818665 7.3373976 7.6959305 7.1948853
246777_at 7.6047335 7.203148 6.480317 7.068931 7.288151 7.2908893
246821_at 6.0464168 5.100028 6.0297427 6.4661326 6.465997 6.3456426
246858_at 7.594398 8.540037 7.764797 8.517765 8.694667 8.651038
246870_at 9.778225 9.399788 9.515823 9.403256 9.387745 9.425886
246889_at 6.125349 5.0277743 4.941466 7.471291 7.8759556 8.14821
246943_at 7.3501644 7.628934 7.2901397 8.436425 8.710248 8.083563
247047_at 8.316461 8.423967 8.446023 7.858939 7.9224544 8.031709
```

Example tabular data

- Each row represents one gene.
- Each column represents expression of gene in that sample. Column separator ,
- First row and First column contain respective labels.

```
kvarala@scholar-fe00:/scratch/scholar/k/kvarala/Week3/Files $ head -n20 GSE49418.top250.csv
```

```
Gene,WT-CK1,WT-CK2,WT-CK3,MT-CK1,MT-CK2,MT-CK3  
245041_at,7.6044173,8.447944,7.777406,8.246566,8.562556,8.739321  
245119_at,7.8671436,8.00699,7.9079247,7.501124,8.054494,7.5438313  
245247_at,8.140869,8.010652,8.181079,7.9125094,8.316227,8.123076  
245250_at,7.742004,8.042652,7.3609304,7.384283,7.645018,7.692195  
245329_at,6.0456758,6.0853715,6.7108784,7.7846346,8.612887,8.28583  
245627_at,0.6136984,2.5572119,1.7915204,1.6312857,0.5355964,2.1927164  
245765_at,9.176895,9.558297,10.050962,9.510457,9.776581,9.219471  
245777_at,7.4506702,8.334939,6.9342628,7.6104965,7.508419,7.4884295  
246270_at,10.288048,9.246562,9.370121,9.395946,9.554836,9.3737755  
246289_at,10.853704,11.123942,11.493572,10.820894,11.0875845,10.795305  
246490_at,6.854518,7.1305866,7.319211,8.059608,7.6570344,7.3294134  
246495_at,8.385275,7.754309,7.818665,7.3373976,7.6959305,7.1948853  
246777_at,7.6047335,7.203148,6.480317,7.068931,7.288151,7.2908893  
246821_at,6.0464168,5.100028,6.0297427,6.4661326,6.465997,6.3456426  
246858_at,7.594398,8.540037,7.764797,8.517765,8.694667,8.651038  
246870_at,9.778225,9.399788,9.515823,9.403256,9.387745,9.425886  
246889_at,6.125349,5.0277743,4.941466,7.471291,7.8759556,8.14821  
246943_at,7.3501644,7.628934,7.2901397,8.436425,8.710248,8.083563  
247047_at,8.316461,8.423967,8.446023,7.858939,7.9224544,8.031709
```

Example block data

FASTQ file contains 4 lines per block:

1. Sequence Header
2. Sequence
3. Quality Header
4. Quality

```
kvarala@scholar-fe00:/scratch/scholar/k/kvarala/Week3/Files $ head -n 20 SRR6473489.fastq
1 { @SRR6473489.1 1 length=76
  { CGATTTCAATGGTTTCCGGGTAAAGAGCTTCGCCGTCGATCTCTATCGCTCTCTGTAATCTGTATTTCTCCGATTA
  { +SRR6473489.1 1 length=76
  { AAAAAEEEEEEA/AEEAA/AEEEEEEEEEEEE/EEAEAAEEEE//EEE<EAEEE/EEEEEAEEAE/AEAE<EAAE
2 { @SRR6473489.2 2 length=76
  { CCGATTTCAATGGTTTCCGGGTAAAGAGCTTCGCCGTCGATCTCTCTCGCTCTCTGTAATCTGTATTTCTCCGATT
  { +SRR6473489.2 2 length=76
  { AAAAAEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEE
3 { @SRR6473489.3 3 length=76
  { GTCCGATTTCAATGGTTTCCGGGTAAAGAGCTTCGCCGTCGATCTCTCTCGCTCTCTGTAATCTGTATTTCTCCGA
  { +SRR6473489.3 3 length=76
  { /AAAAEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEE/
4 { @SRR6473489.4 4 length=76
  { GTCCGATTTCAATGGTTTCCGGGTAAAGAGCTTCGCCGTCGATCTCTCTCGCTCTCTGTAATCTGTATTTCTCCGA
  { +SRR6473489.4 4 length=76
  { 6AAAAEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEAEAE
  { @SRR6473489.5 5 length=76
  { GGAGAAATACAGATTACAGAGAGCGAGAGAGATCGACGGCGAAGCTCTTTACCCGGAACCATTGAAATCGGACGG
  { +SRR6473489.5 5 length=76
  { AAAA6EEEEEEEEEEEEEEAEE6EEEEEEEEEE6EEEEEEEEEEEEEEEEEEEE/E/EEEEEEEE//E/EEEEEE//EE
```


Word count

- `wc` command returns the word count in file.
- Default is to return counts of words, lines and characters.

```
kvarala@scholar-fe00:/scratch/scholar/k/kvarala/Week3/Files $ wc North_of_Boston.txt
 2607  19985 118808 North_of_Boston.txt
kvarala@scholar-fe00:/scratch/scholar/k/kvarala/Week3/Files $ wc -l North_of_Boston.txt
2607 North_of_Boston.txt
kvarala@scholar-fe00:/scratch/scholar/k/kvarala/Week3/Files $ wc -w North_of_Boston.txt
19985 North_of_Boston.txt
kvarala@scholar-fe00:/scratch/scholar/k/kvarala/Week3/Files $ wc -c North_of_Boston.txt
118808 North_of_Boston.txt
```

Sort file contents

- `sort` command sorts the file by the line content.
- Can be applied to tabular data to sort by specific columns.
- Default sort is by ASCII code (as modified by locale).

```
kvarala@scholar-fe00:/scratch/scholar/k/kvarala/Week3/Files $ sort -k3,3 GSE49418.top250.tsv |head -n 5
257919_at      0.3471166      0.20855379     0.828144       1.9871551     4.676103      3.4019423
264661_at      2.0964775     0.79441476     0.5991168     2.1336834     0.8073569     3.5307076
267139_s_at    1.5610524     0.961639      1.5409981     5.0253215     4.9877644     5.4757285
254784_at     10.315967     10.020184     10.047566     9.9493685     10.172818     10.167413
263478_at      9.870353      10.088003     10.271939     10.87601      10.983693     11.1485405
kvarala@scholar-fe00:/scratch/scholar/k/kvarala/Week3/Files $ sort -n -k3,3 GSE49418.top250.tsv |head -n 5
Gene   WT-CK1  WT-CK2  WT-CK3  MT-CK1  MT-CK2  MT-CK3
257919_at      0.3471166      0.20855379     0.828144       1.9871551     4.676103      3.4019423
264661_at      2.0964775     0.79441476     0.5991168     2.1336834     0.8073569     3.5307076
267139_s_at    1.5610524     0.961639      1.5409981     5.0253215     4.9877644     5.4757285
265709_at      1.7578125     1.4652753     1.90664 2.3643675      1.8901684     2.1367643
```

ASCII code

The ASCII code

American Standard Code for Information Interchange

ASCII control characters				ASCII printable characters								
DEC	HEX	Simbolo ASCII		DEC	HEX	Simbolo	DEC	HEX	Simbolo	DEC	HEX	Simbolo
00	00h	NULL	(carácter nulo)	32	20h	espacio	64	40h	@	96	60h	`
01	01h	SOH	(inicio encabezado)	33	21h	!	65	41h	A	97	61h	a
02	02h	STX	(inicio texto)	34	22h	"	66	42h	B	98	62h	b
03	03h	ETX	(fin de texto)	35	23h	#	67	43h	C	99	63h	c
04	04h	EOT	(fin transmisión)	36	24h	\$	68	44h	D	100	64h	d
05	05h	ENQ	(enquiry)	37	25h	%	69	45h	E	101	65h	e
06	06h	ACK	(acknowledgement)	38	26h	&	70	46h	F	102	66h	f
07	07h	BEL	(timbre)	39	27h	'	71	47h	G	103	67h	g
08	08h	BS	(retroceso)	40	28h	(72	48h	H	104	68h	h
09	09h	HT	(tab horizontal)	41	29h)	73	49h	I	105	69h	i
10	0Ah	LF	(salto de línea)	42	2Ah	*	74	4Ah	J	106	6Ah	j
11	0Bh	VT	(tab vertical)	43	2Bh	+	75	4Bh	K	107	6Bh	k
12	0Ch	FF	(form feed)	44	2Ch	,	76	4Ch	L	108	6Ch	l
13	0Dh	CR	(retorno de carro)	45	2Dh	-	77	4Dh	M	109	6Dh	m
14	0Eh	SO	(shift Out)	46	2Eh	.	78	4Eh	N	110	6Eh	n
15	0Fh	SI	(shift In)	47	2Fh	/	79	4Fh	O	111	6Fh	o
16	10h	DLE	(data link escape)	48	30h	0	80	50h	P	112	70h	p
17	11h	DC1	(device control 1)	49	31h	1	81	51h	Q	113	71h	q
18	12h	DC2	(device control 2)	50	32h	2	82	52h	R	114	72h	r
19	13h	DC3	(device control 3)	51	33h	3	83	53h	S	115	73h	s
20	14h	DC4	(device control 4)	52	34h	4	84	54h	T	116	74h	t
21	15h	NAK	(negative acknowle.)	53	35h	5	85	55h	U	117	75h	u
22	16h	SYN	(synchronous idle)	54	36h	6	86	56h	V	118	76h	v
23	17h	ETB	(end of trans. block)	55	37h	7	87	57h	W	119	77h	w
24	18h	CAN	(cancel)	56	38h	8	88	58h	X	120	78h	x
25	19h	EM	(end of medium)	57	39h	9	89	59h	Y	121	79h	y
26	1Ah	SUB	(substitute)	58	3Ah	:	90	5Ah	Z	122	7Ah	z
27	1Bh	ESC	(escape)	59	3Bh	;	91	5Bh	[123	7Bh	{
28	1Ch	FS	(file separator)	60	3Ch	<	92	5Ch	\	124	7Ch	
29	1Dh	GS	(group separator)	61	3Dh	=	93	5Dh]	125	7Dh	}
30	1Eh	RS	(record separator)	62	3Eh	>	94	5Eh	^	126	7Eh	~
31	1Fh	US	(unit separator)	63	3Fh	?	95	5Fh	-			
127	20h	DEL	(delete)									theASCIIcode.com.ar

Extract specific columns

- cut allows extraction of 'fields' (columns) from the file.

```
kvarala@scholar-fe00:/scratch/scholar/k/kvarala/Week3/Files $ cut -f1 GSE49418.top250.tsv
Gene
245041_at
245119_at
245247_at
245250_at
```

- Default delimiter is <TAB> but can be substituted using the -d argument.

```
kvarala@scholar-fe00:/scratch/scholar/k/kvarala/Week3/Files $ cut -d, -f1 GSE49418.top250.csv
Gene
245041_at
245119_at
245247_at
245250_at
```

Extract specific columns

- Multiple columns can be specified by giving their column index in `-f` argument.
 - E.g., `-f1,5,7` would extract columns 1,5 and 7
- Range of columns may also be specified.
 - E.g., `-f1-4` would extract columns 1,2,3,4

```
kvarala@scholar-fe00:/scratch/scholar/k/kvarala/Week3/Files $ cut -d, -f1-4 GSE49418.top250.csv
Gene,WT-CK1,WT-CK2,WT-CK3
245041_at,7.6044173,8.447944,7.777406
245119_at,7.8671436,8.00699,7.9079247
245247_at,8.140869,8.010652,8.181079
245250_at,7.742004,8.042652,7.3609304
```

Merge column data

- paste command allows combining files at a column level.

```
kvarala@scholar-fe03:/scratch/scholar/k/kvarala/Week3/Files $ head -n 5 GSE49418.top250.ids
Gene
245041_at
245119_at
245247_at
245250_at
kvarala@scholar-fe03:/scratch/scholar/k/kvarala/Week3/Files $ head -n 5 GSE49418.top250.WT-CK1.vals
WT-CK1
7.6044173
7.8671436
8.140869
7.742004
kvarala@scholar-fe03:/scratch/scholar/k/kvarala/Week3/Files $ head -n 5 GSE49418.top250.MT-CK1.vals
MT-CK1
8.246566
7.501124
7.9125094
7.384283
kvarala@scholar-fe03:/scratch/scholar/k/kvarala/Week3/Files $ paste GSE49418.top250.ids GSE49418.top250.WT-CK1.vals GSE49418.top250.MT-CK1.vals
Gene    WT-CK1  MT-CK1
245041_at    7.6044173    8.246566
245119_at    7.8671436    7.501124
245247_at    8.140869     7.9125094
245250_at    7.742004     7.384283
```

Matching text via patterns

- Pattern matching via *regular expressions* is a powerful tool to match text within files.
- It forms the basis for text searches and manipulation in multiple UNIX tools such as: grep, sed, awk etc.
- We will cover regular expressions and these commands in the next lecture.