Week 1, Lecture 3

István Albert

Biochemistry and Molecular Biology
and Bioinformatics Consulting Center

Penn State
Some Programming Required

- Existing software tools can rarely do all steps

- Source data in different files may be formatted differently

- We need to bridge the differences with simple transformations

- Some programming ability is necessary for every project
Possibly the first scripting language: \texttt{awk}

- appeared in 1977, strong UNIX (1972) roots

- it is the precursor of languages such as \texttt{Perl} (1987) and \texttt{Python} (1989)

- Has fallen into disuse for a while; the large text datasets have lead to a resurgence of the language
The Art of Unix Programming: TAOUP

• Seminal open source book by Eric Raymond

For a few years after the release of Perl in 1987, `awk` remained competitive simply because it had a smaller, faster implementation. But as the cost of compute cycles and memory dropped, the economic reasons for favoring a special-purpose language that was relatively thrifty with both lost their force. Programmers increasingly chose to do awk-like things with Perl or (later) Python, rather than keep two different scripting languages in their heads.\[90\] By the year 2000 `awk` had become little more than a memory for most old-school Unix hackers, and not a particularly nostalgic one.

Not quite right: **The more things change the more they stay the same.**
The structure of an **awk** program

```awk
' pattern { action } '
```

**Line oriented**

1. try to match the pattern to the line
2. If there is a match process the line and produce a new line

**Default pattern** = match everything
Simple Awk Example – no pattern specified

```
ialbert@porthos ~/work
$ ls
features.gff sc.gff

ialbert@porthos ~/work
$ awk '{ print $3 }' features.gff | head -3
chromosome
telomeric_repeat
telomere

ialbert@porthos ~/work
$ # equivalent to

ialbert@porthos ~/work
$ cut -f 3 features.gff | head -3
chromosome
telomeric_repeat
telomere

ialbert@porthos ~/work
$ 
```
Advice
write pipelines with clear inputs and output

• Keep the flow of information from left to right
• It is a lot easier to modify/alter later on

$ cat features.gff | awk '{ print $3 }' | head -3
Awk **automatically splits** the input by whitespace (spaces and tabs) and assigns names to them:

- $0$ the entire line
- $1$ first field
- $2$ second field
- ...
- **NF** the number of fields
- **NR** the number of the current line
The whitespace curse: spaces and tabs

- Many tools will autosplit by whitespace → this was thought to be convenient but is also the source of extremely subtle errors → leads to a column shift in a tab file if a field contains spaces

- Always specify the character to be split by!

- This refers to programming languages as well!

  Do not use the `split()` methods with their default behavior (Python, Perl etc) unless you perfectly understand what they do)
Customize awk to use tabs as both the input and output field separator

```
alias awk="awk -F '\t' -v OFS='\t'"
```

Tip: you can add this to the `.profile` or `.bash_profile` file in your root folder so that it is activate all the time.

**Note:** file names that start with a dot . are only listed if you do a `ls -a`
Recall the GFF format from lecture 3

Search for GFF3 → http://www.sequenceontology.org/gff3.shtml

Tab separated with 9 columns. Missing attributes may be replaced with a dot → .

<table>
<thead>
<tr>
<th></th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td><strong>Seqid</strong> (usually chromosome)</td>
</tr>
<tr>
<td>2.</td>
<td><strong>Source</strong> (where is the data coming from)</td>
</tr>
<tr>
<td>3.</td>
<td><strong>Type</strong> (usually a term from the sequence ontology)</td>
</tr>
<tr>
<td>4.</td>
<td><strong>Start</strong> (interval start relative to the seqid)</td>
</tr>
<tr>
<td>5.</td>
<td><strong>End</strong> (interval end relative to the seqid)</td>
</tr>
<tr>
<td>6.</td>
<td><strong>Score</strong> (the score of the feature, a floating point number)</td>
</tr>
<tr>
<td>7.</td>
<td><strong>Strand</strong> (+/-/.).</td>
</tr>
<tr>
<td>8.</td>
<td><strong>Phase</strong> (used to indicate reading frame for coding sequences)</td>
</tr>
<tr>
<td>9.</td>
<td><strong>Attributes</strong> (semicolon separated attributes → <strong>Name=ABC;ID=1</strong>)</td>
</tr>
</tbody>
</table>

Example attribute specification: **name=REB1;id=YP33546**
Will attempt to extract the name from 9th column -> the attribute

Semicolon separated attributes one of them is **Name=YAL069W**;

1. Split by semicolon and keep the second element, then split this again by the = sign and keep the second element
You may also put the program into a file

dollar3 == "gene" { print $9 }

$ cat features.gff | awk -f name.awk | head -1
Write a single step at a time and test it every time

```awk
$3=="gene" {

    # split column 9 and place it into x
    split($9, x, ",");

    # print second element of x
    print x[2]
}
```

```bash
ialbert@porthos ~/work
$ cat features.gff | awk -f name.awk | head -1
Name=YAL069W

ialbert@porthos ~/work
$```
```awk
$3=="gene" {
    # split column 9 and place it into x
    split($9, x, ";");

    # rename for easier handling
    pair = x[2]

    split(pair, y, "=")
    # print second element of x

    print y[2]
}
```

```
ialbert@porthos ~/work
$ cat features.gff | awk -f name.awk | head -1
YAL069W

ialbert@porthos ~/work
$ 
```
Create a gff file that contains the gene name in the type column

```awk
split($9, x, ";");

# rename for easier handling
pair = x[2]

split(pair, y, "=");
# print second element of x

name = y[2]

print $1, $2, name, $4, $5, $6, $7, $8, "."
```
Make that new simpler gff: short.gff contains only genes and has name in the second column

```
ialbert@porthos ~/work
$ cat features.gff | awk -f name.awk > short.gff

ialbert@porthos ~/work
$ ls
features.gff name.awk sc.gff short.gff

ialbert@porthos ~/work
$ head short.gff
chrI SGD YAL069W 335 649 . + . .
chrI SGD YAL068W-A 538 792 . + . .
chrI SGD YAL068C 1807 2169 . - . .
chrI SGD YAL067W-A 2480 2707 . + . .
chrI SGD YAL067C 7235 9016 . - . .
chrI SGD YAL066W 10091 10399 . + . .
```
Interval lengths for genes

```bash
ialbert@porthos ~/work
$ cat short.gff | awk ' { size = $5 - $4 + 1; print $3, size } ' | head -1
YAL069W 315

ialbert@porthos ~/work
$
```

```bash
ialbert@porthos ~/work
$ # match only on chromosome chrV

ialbert@porthos ~/work
$ cat short.gff | awk ' $1=='chrV'' { size = $5 - $4 + 1; print $3, size } ' | head -1
YEL077C 3834

ialbert@porthos ~/work
$
```
Operators

- `+ - * /` for numerical context
- `%` modulo division (remainder of division)
- `<space>` string concatenation
- `==, !=` equal, not equal
- `~, !~` match, no match (regular expressions)
Special patterns

- BEGIN → before the stream starts
- END → after the stream ends

```awk
# a running tally of the total length of genomic intervals in the data
BEGIN { num = 0 }
{ size= $5 - $4 + 1; num += size; }
END { print num }
```

```bash
ialbert@porthos ~/work
$ cat short.gff | awk -f counter.awk | more
8886525
```

```bash
ialbert@porthos ~/work
$ ```
Multiple patterns \( \rightarrow \) awk will try them each in turn

```awk
# all patterns will be matched unless the next command is seen
# the command next causes the next line to be loaded

$3=="chromosome" { print "found chrom at line", NR; }

{ print "found data at line ", NR}

$3=="telomeric_repeat" { print "found telo at line", NR }
```

```
$ cat features.gff | awk -f multi.awk | head -5
found chrom at line 1
found data at line 1
found data at line 2
found data at line 2
found telo at line 2
found data at line 3
```
Advanced awk – usually not needed

• conditionals: if
• loops: for, while
• break, continue
• associative data structures (hash, dictionary)

You can do all that though at that point it is probably better to learn Python

But you can do a lot with just basic awk!

Awk’s power comes from its simplicity – one liners!
A few awk resources

Large number of resources, the quirky name makes it very searchable ➔ How to do X with awk?

- How to use awk:
  http://sparky.rice.edu/awk.html
1. Write the awk code that finds the length of the genome (sum of all chromosome lengths) from the features.gff file.

2. Write the awk code that finds the longest and shortest gene in the features.gff file