ETH zürich

High Performance Computing for genomic applications

AWK - a quick course

Scientific IT Services

Michal Okoniewski

What is AWK?

- It is a programming/scripting language

- It is ancient – developed ca 1977 BC

- Created by **Brian Kernighan** – also known for C and Unix

Why AWK is cool and useful?

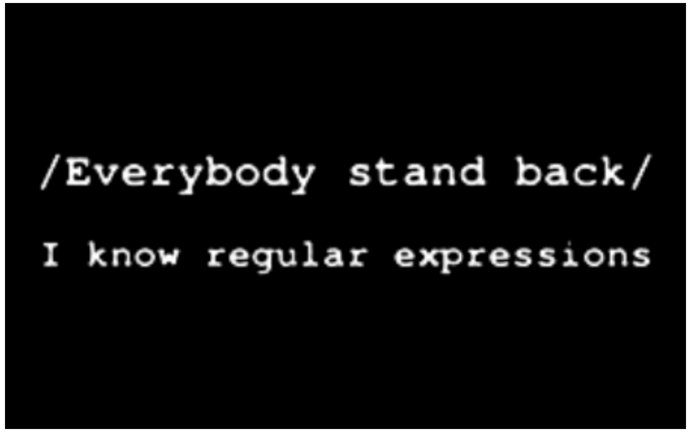
- It works

- It works quickly

- It does the job

Why AWK is cool? – technically

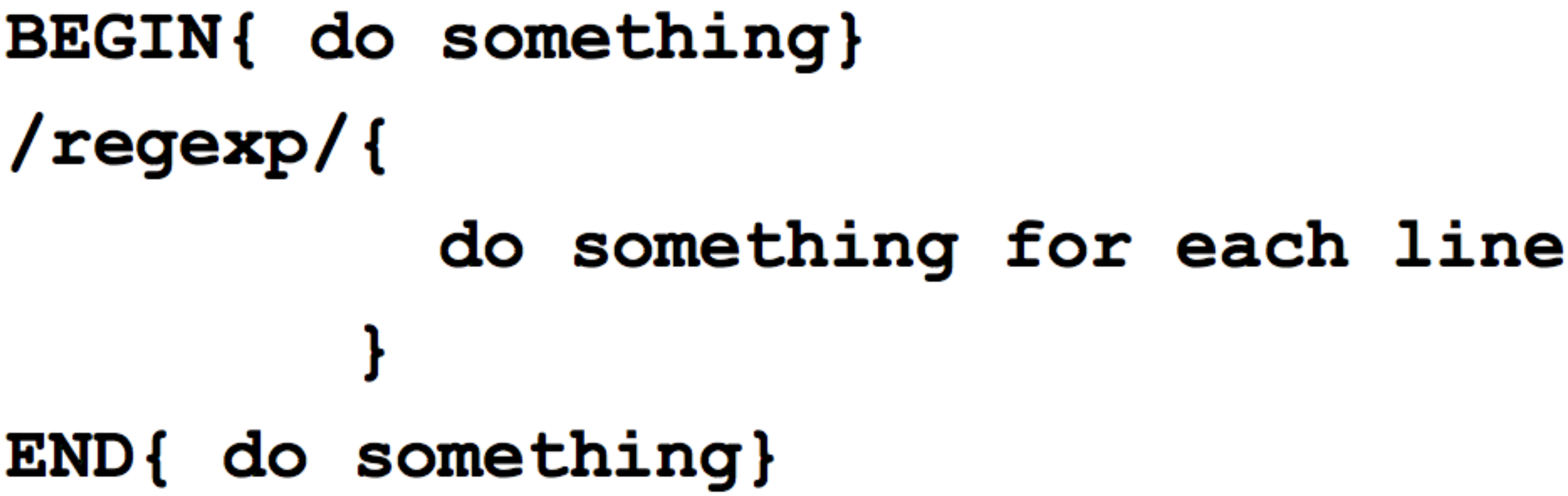
- It processes the text files, assuming that they are tables
- It processed them line-by-line, row-by-row
- It uses regular expressions
- It is so easy and reliable that I am not afraid of a live presentation ☺



/Everybody stand back/

I know regular expressions

Program structure



```
BEGIN{ do something}
/regexp/{
          do something for each line
        }
END{ do something}
```

However….

- Just the main body of the program is necessary
- Most of the programs are ad-hoc one-liners

- Can be run from command line ad-hoc way:
  awk '{print}' test.txt

- Or run from command line using a script file:
  awk -f my.awk.script test.input.txt > output.txt

Basics 1

- Variables:

$0 the whole row being processed right now $1, $2, $3 ... the fields in the line

NR – number of current row

FS – field separator

OFS – output field separator

Basics 2

- Commands that you may need

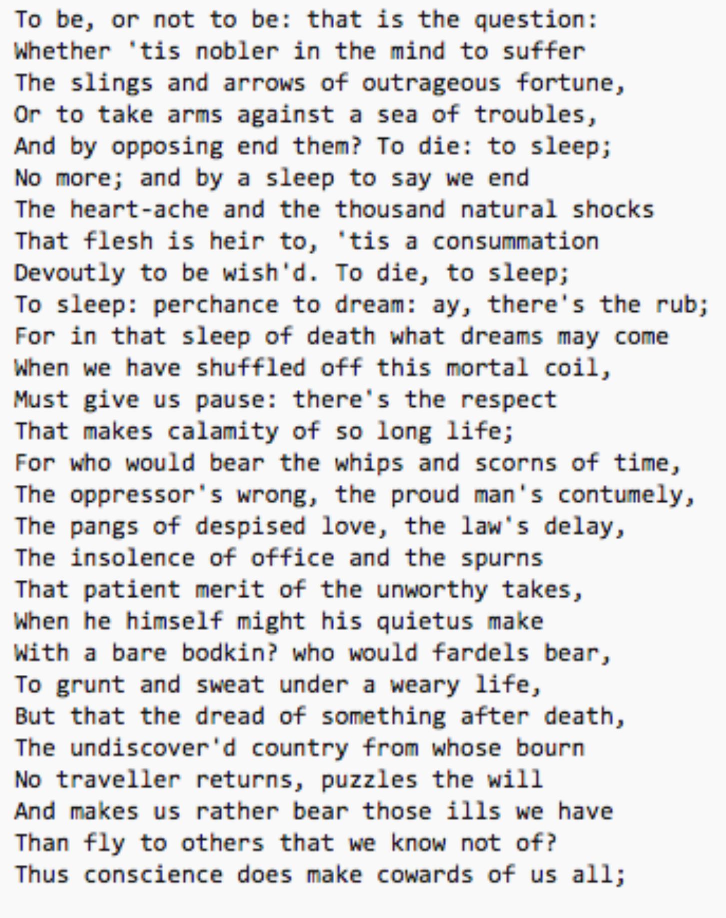
```
print

if (condition) command

if (var~/regexp/) command
```

Variable assignment =

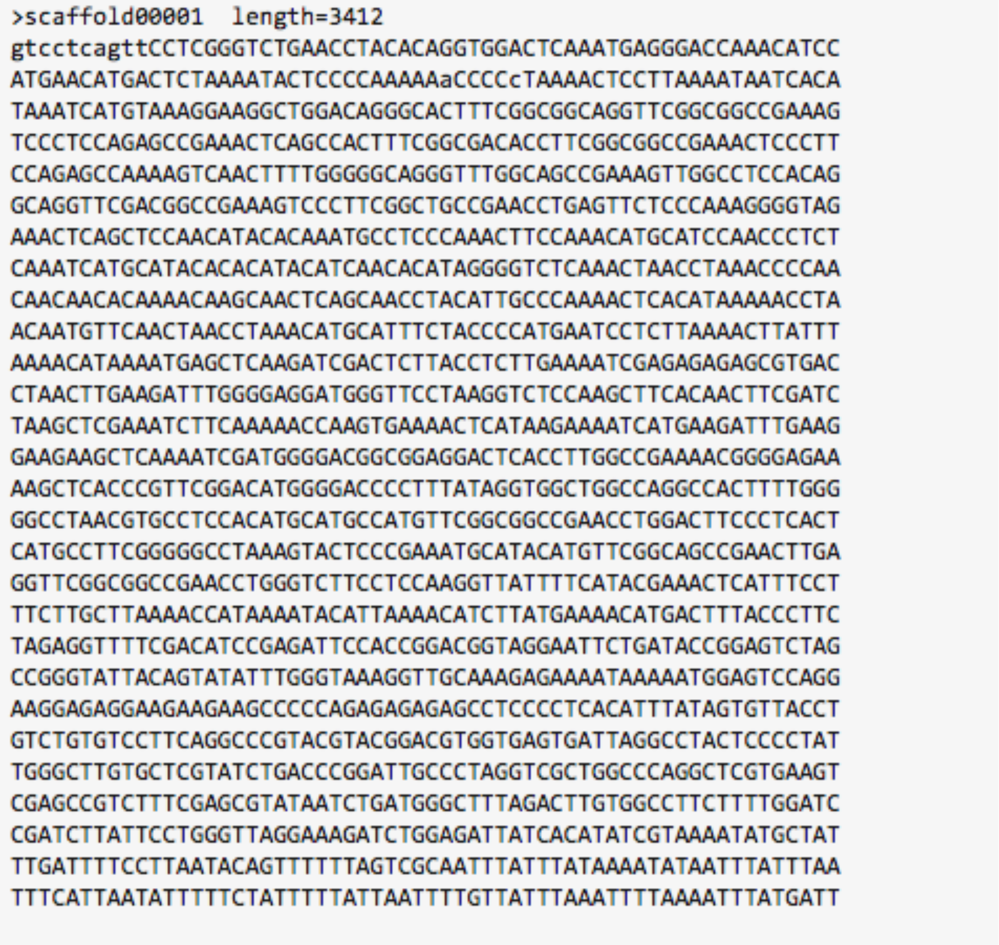
Concatenation of strings – just put them next to each other:

```
    “Manuel is ”$3
```

Demo1



```
To be, or not to be: that is the question:
Whether 'tis nobler in the mind to suffer
The slings and arrows of outrageous fortune,
Or to take arms against a sea of troubles,
And by opposing end them? To die: to sleep;
No more; and by a sleep to say we end
The heart-ache and the thousand natural shocks
That flesh is heir to, 'tis a consummation
Devoutly to be wish'd. To die, to sleep;
To sleep: perchance to dream: ay, there's the rub;
For in that sleep of death what dreams may come
When we have shuffled off this mortal coil,
Must give us pause: there's the respect
That makes calamity of so long life;
For who would bear the whips and scorns of time,
The oppressor's wrong, the proud man's contumely,
The pangs of despised love, the law's delay,
The insolence of office and the spurns
That patient merit of the unworthy takes,
When he himself might his quietus make
With a bare bodkin? who would fardels bear,
To grunt and sweat under a weary life,
But that the dread of something after death,
The undiscover'd country from whose bourn
No traveller returns, puzzles the will
And makes us rather bear those ills we have
Than fly to others that we know not of?
Thus conscience does make cowards of us all;
```



```
>scaffold00001  length=3412
gtcctcagttCCTCGGGTCTGAACCTACACAGGTGGACTCAAATGAGGGACCAAACATCC
ATGAACATGACTCTAAAATACTCCCCAAAAAaCCCCcTAAAACTCCTTAAAATAATCACA
TAAATCATGTAAAGGAAGGCTGGACAGGGCACTTTCGGCGGCAGGTTCGGCGGCCGAAAG
TCCCTCCAGAGCCGAAACTCAGCCACTTTCGGCGACACCTTCGGCGGCCGAAACTCCCTT
CCAGAGCCAAAAGTCAACTTTTGGGGGCAGGGTTTGGCAGCCGAAAGTTGGCCTCCACAG
GCAGGTTCGACGGCCGAAAGTCCCTTCGGCTGCCGAACCTGAGTTCTCCCAAAGGGGTAG
AAACTCAGCTCCAACATACACAAATGCCTCCCAAACTTCCAAACATGCATCCAACCCTCT
CAAATCATGCATACACACATACATCAACACATAGGGGTCTCAAACTAACCTAAACCCCAA
CAACAACACAAAACAAGCAACTCAGCAACCTACATTGCCCAAAACTCACATAAAAACCTA
ACAATGTTCAACTAACCTAAACATGCATTTCTACCCCATGAATCCTCTTAAAACTTATTT
AAAACATAAAATGAGCTCAAGATCGACTCTTACCTCTTGAAAATCGAGAGAGAGCGTGAC
CTAACTTGAAGATTTGGGGAGGATGGGTTCCTAAGGTCTCCAAGCTTCACAACTTCGATC
TAAGCTCGAAATCTTCAAAAACCAAGTGAAAACTCATAAGAAAATCATGAAGATTTGAAG
GAAGAAGCTCAAAATCGATGGGGACGGCGGAGGACTCACCTTGGCCGAAAACGGGGAGAA
AAGCTCACCCGTTCGGACATGGGGACCCCTTTATAGGTGGCTGGCCAGGCCACTTTTGGG
GGCCTAACGTGCCTCCACATGCATGCCATGTTCGGCGGCCGAACCTGGACTTCCCTCACT
CATGCCTTCGGGGGCCTAAAGTACTCCCGAAATGCATACATGTTCGGCAGCCGAACTTGA
GGTTCGGCGGCCGAACCTGGGTCTTCCTCCAAGGTTATTTTCATACGAAACTCATTTCCT
TTCTTGCTTAAAACCATAAAATACATTAAAACATCTTATGAAAACATGACTTTACCCTTC
TAGAGGTTTTCGACATCCGAGATTCCACCGGACGGTAGGAATTCTGATACCGGAGTCTAG
CCGGGTATTACAGTATATTTGGGTAAAGGTTGCAAAGAGAAAATAAAAATGGAGTCCAGG
AAGGAGAGGAAGAAGAAGCCCCCAGAGAGAGAGCCTCCCCTCACATTTATAGTGTTACCT
GTCTGTGTCCTTCAGGCCCGTACGTACGGACGTGGTGAGTGATTAGGCCTACTCCCCTAT
TGGGCTTGTGCTCGTATCTGACCCGGATTGCCCTAGGTCGCTGGCCCAGGCTCGTGAAGT
CGAGCCGTCTTTCGAGCGTATAATCTGATGGGCTTTAGACTTGTGGCCTTCTTTTGGATC
CGATCTTATTCCTGGGTTAGGAAAGATCTGGAGATTATCACATATCGTAAAATATGCTAT
TTGATTTTCCTTAATACAGTTTTTTAGTCGCAATTTATTTATAAAATATAATTTATTTAA
TTTCATTAATATTTTTCTATTTTTATTAATTTTGTTATTTAAATTTTAAAATTTATGATT
```

Demo1

```
awk '{print}' test.txt

awk '{print $1}' test.txt

awk '{print $10}' test.txt

awk '/the/{print NR}' test.txt

awk '/the|The/{print NR}' test.txt

awk '/the/{c=c+1}END{print c}' test.txt

awk '/GATCGATC/{c=c+1}END{print c}' sequence.fa

awk '/>/{c=c+1}END{print c}' sequence.fa

awk '/>/{getline; if ($0~/AAAA/) c=c+1}END{print c}' sequence.fa
```

Demo 2

- Using Homo_sapiens.GRCh38.86.chr.gtf

Demo 2

- Print all lines that include a specific identifier

```
awk '/MITF/{print}' Homo_sapiens.GRCh38.86.chr.gtf
```

- Print all the genes in a given chromosome and strand, as tab-delimited

```
awk '{OFS="\t"; if ($1== "22" && $7=="+" && $3=="gene") print}'
Homo_sapiens.GRCh38.86.chr.gtf
```

- Print their genome coordinates

```
awk '{OFS="\t"; if ($1== "22" && $7=="+" && $3=="gene") print$1, $4, $5, $7, $9)
}' Homo_sapiens.GRCh38.86.chr.gtf
```

- Count known exons for MITF gene

```
awk '/MITF/{OFS="\t"; if ($3=="exon") c=c+1}END{print c}'
Homo_sapiens.GRCh38.86.chr.gtf
```

Extras

- Other useful functions:

- split

```
awk '/scaffold12946/{split($9,a,";"); print a[2]}' Mesculenta_147_gene.gff3
```

- getline

```
awk '/>/{getline; if ($0~/AAAA/) c=c+1}END{print c}' Mesculenta_147.fa
```

- many others

Extras

- One-liners:

  http://www.pement.org/awk/awk1line.txt
  http://nixshell.wordpress.com/2009/04/01/awk-one-liners/

- Manuals

  http://www.gnu.org/software/gawk/manual/gawk.html

  http://www.grymoire.com/Unix/Awk.html

  http://pubs.opengroup.org/onlinepubs/7908799/xcu/awk.html

ETH zürich

Enjoy

AWK