Unix - Advanced I

(plain text file modification - basics)

Libor Mořkovský, Václav Janoušek

https://ngs-course.readthedocs.io/en/praha-february-2019/

Pattern Search: `grep`

- pattern specification & matching
- use:

```
grep pattern file # Match lines having a pattern
```

```
grep -v pattern file # Match lines not having a pattern
```

`grep`: Regular expressions

- matching string patterns according to certain rules
- -E = extended grep

```
^A
A$
[0-9]
[A-Z]
[ATGC]
.
A*
A{2}
A{1,} or A+
A{1,3}
AATT|TTAA
\s
```

Exercise (nightingale VCF)

Work with nightingale variant call file (VCF):

1. Count the number variants in the file

```
< /data-shared/vcf_examples/
luscinia_vars_flags.vcf.gz zcat |
grep -v '^#' | wc -l
```

Exercise (nightingale VCF)

2. Count the number of variants passing/failing the quality threshold

```
< /data-shared/vcf_examples/
luscinia_vars_flags.vcf.gz zcat |
grep -v '^#' | grep 'PASS' | wc -l


< /data-shared/vcf_examples/
luscinia_vars_flags.vcf.gz zcat |
grep -v '^#' | grep 'FAIL' | wc -l
```

Exercise (nightingale VCF)

3. Count the number of variants on the chromosome Z passing the quality threshold

```
< /data-shared/vcf_examples/
luscinia_vars_flags.vcf.gz zcat |
grep -v '^#' |
grep 'PASS' |
grep '^chrZ\s' | wc -l
```

Coffee break…

Cutting out, sorting and unique records

```
cut -f
sort -rn -k1,1 -k2,2
uniq -c
```

Try these commands using VCF file:

```
cut -f
sort -rn -k1,1 -k2,2
uniq -c
```

String replacing/deleting (`tr` vs. `sed`)

```
# Removal line endings
tr -d "\n"

# Replacement all ; to TAB separators
tr ";" "\t"
```

String replacing/deleting (`tr` vs. `sed`)

```
sed 's/pattern/replacement/'

# Replace one or more A or C or G or T by N
sed 's/^[AGCT]\{1,\}/N/'

# The same thing using extended regular expressions:
sed -r 's/^[AGCT]+/N/'
```

Exercise (nightingale VCF)

1. Which chromosome has the highest and the least number of variants?

```
< /data-shared/vcf_examples/
luscinia_vars_flags.vcf.gz zcat |
grep -v '^#' |
cut -f 1 |
sort |
uniq -c |
sed -r 's/^ +//' |
sort -k1,1nr
```

Exercise (nightingale VCF)

2. What is the number of samples in the VCF file?

```
< data-shared/luscinia_vars_flags.vcf
grep -v '^##' |
head -n1 |
cut --complement -f 1-9 |
tr "\t" "\n" |
wc -l
```

Exercise

3. Count the number of bases sequenced in nightingale FASTQ files (data/fastq/.fastq)?*

...you have help at the website

Lunch, Lunch!!