Dealing with large files in UNIX

SNP and indel calling using samtools

- samtools view -b -o ler.bam -S ler.sam
- samtools sort ler.bam ler.sorted
- samtools index ler.sorted.bam
- samtools faidx mt.fa
- samtools tview ler.sorted.bam mt.fa
- samtools mpileup -uf mt.fa ler.sorted.bam | bcftools view -vcg - > ler_snps_indels.vcf
- less -S ler_snps_indels.vcf

Now we have some large files

- cat
  - Print a file or files line by line
- less
  - Display a file so you can scroll through it
- head -n X
  - Print X lines from the beginning of the file
- tail -n X
  - Prints X lines from the end of the file

Finding what you want in a large file

- grep
  - Search for a string of characters
  - e.g.

```
grep 'word' filename
```

Finding what you want in a large file

- grep
  - Search for a string of characters

```
grep '#' ler_snps_indels.vcf
```

Finding what you want in a large file

```
grep -c '#' ler_snps_indels.vcf
```

Finding what you want in a large file

```
grep -v '##' ler_snps_indels.vcf > ler_snps_indels.txt
```

- Useful for filtering out lines that you want / don't want in a file, as well as counting, etc.

Finding what you want in a large file

- grep

–Search for a string of characters

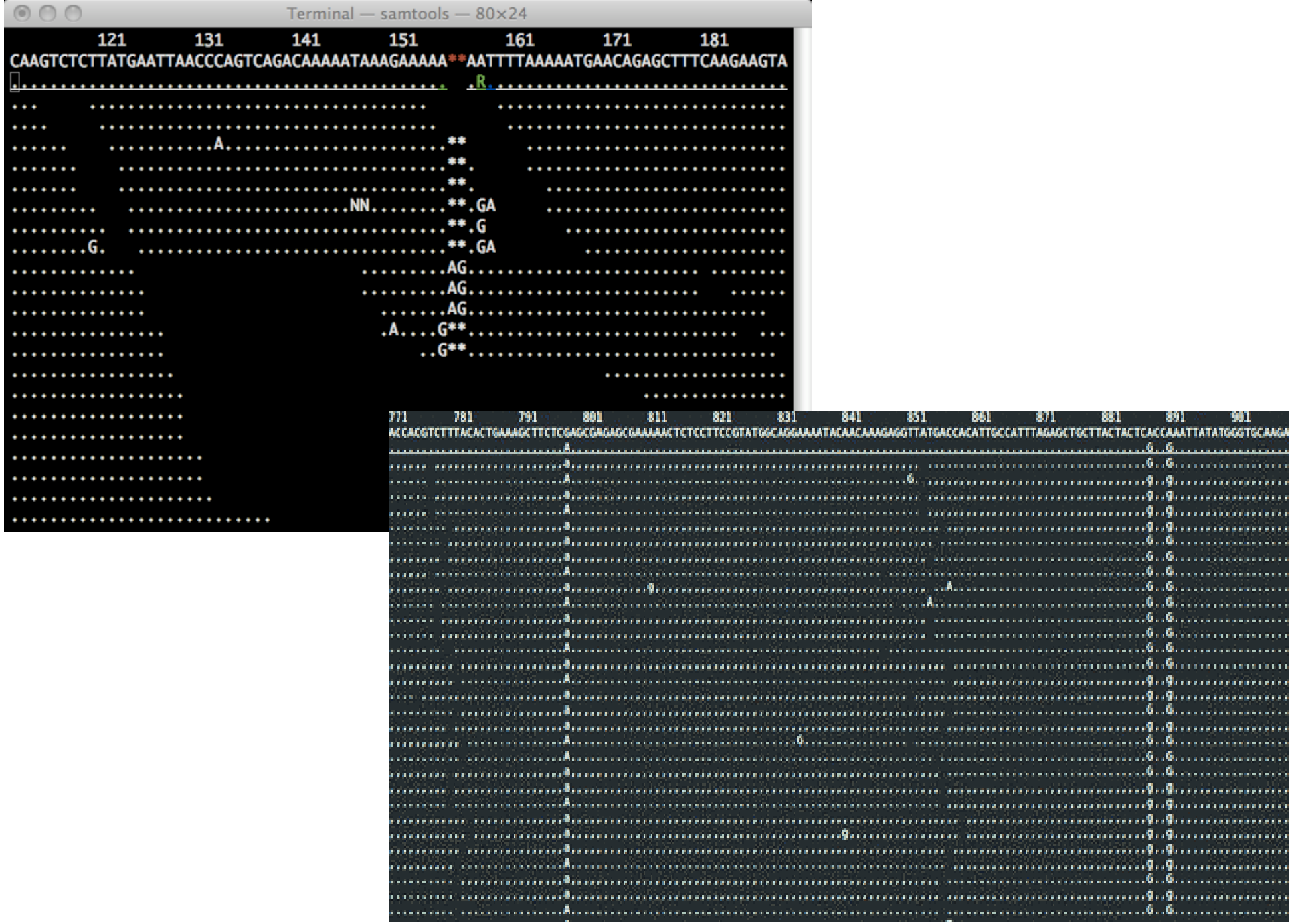
```
grep '#' ler_snps_indels.vcf

grep -c '#' ler_snps_indels.vcf

grep -v '##' ler_snps_indels.vcf > ler_snps_indels.txt
```

–Useful for filtering out lines that you want / don't want in a file, as well as counting, etc.

- Use grep to remove INDELS from ler_snps_indels.txt
- Save it as allsnps.txt

SNP and indel calling using samtools

- samtools view -b -o ler.bam -S ler.sam
- samtools sort ler.bam ler.sorted
- samtools index ler.sorted.bam
- samtools faidx mt.fa
- samtools tview ler.sorted.bam mt.fa
- samtools mpileup -uf mt.fa ler.sorted.bam | bcftools view -vcg - > ler_snps_indels.vcf
- less -S ler_snps_indels.vcf
- grep -v '##' ler_snps_indels.vcf > ler_snps_indels.txt


Terminal — samtools — 80×24
121 131 141 151 161 171 181
CAAGTCTCTTATGAATTAACCCAGTCAGACAAAAATAAAGAAAAA**AATTTTAAAAATGAACAGAGCTTTCAAGAAGTA
771 781 791 801 811 821 831 841 851 861 871 881 891 901

AWK

AWK

- Created in 1977
- Named after creators Aho, Weinberger, Kernighan
- Inspired perl, a more common modern language.
- Very powerful for 1-line commands
- See wikipedia for more background on AWK
- http://en.wikipedia.org/wiki/AWK

AWK

- awk 'IF {what you want to do }'
- http://www.catonmat.net/blog/awk-one-liners-explained-part-one/

Printing in AWK

- An AWK program is a series of pattern action pairs:
- condition { action }
- The default action is to print the current line ($0 in awk), so if no action is specified, that is what is performed, if the condition is true.
- awk '1 {print $0}' mt.vcf
- Is the same as
- cat mt.vcf
- You can also print out particular columns (fields)
- $1 is column 1, $2 is column 2, etc.
- Please use awk to print out the quality scores

‘if’ statements in awk

- The other half of awk commands is the if statement, outside of the brackets
- awk '$4 == “A” {print $0}' mt.vcf
- Which is also the same as:
- awk '$4 == “A”' mt.vcf
- What does this do?
- awk '$2 < 40000' mt.vcf

AWK

- Built for parsing large text files and tables
- awk '{ print FNR "\t" $0 }' allsnps.txt > numbered.txt

- # Looks only as snps – no indels!
- awk '$4 ~ /^[AGCT]$/ && $5 ~ /^[AGCT]$/' ler_snps_indels.txt
- Also:
- grep -v 'INDEL' ler_snps_indels.txt

- http://www.catonmat.net/blog/awk-one-liners-explained-part-one/

Homework to turn in

- Email me with:
  - A count of how many total SNPs and INDELS are in the vcf at first
  - How many good SNPs vs total SNPs
  - How many good INDELS vs total INDELs
  - How you got those numbers
  - Describe one example of a good / not good SNP that you looked at in tview, and why you think it is good / not good

- Also in that email:
- Tell me what you think of the ‘awk one-liners’ website. List at least one command you could use on the files we have worked with, and what you would use it for