Bioinformatics Survival Kit

Everyday bioinformatics functions

- **Bioawk** (awk with bioinformatics super powers)
- **Seqtk** (sequence toolkit)
- **Tabix** (indexed, searchable tab files)
- **Tabtk** (tab toolkit)

### Bioawk functions

```
$ bioawk -c help

1chrom 2start 3end 4name 5score 6strand 7thickstart 8thickend 9rgb 10phase

vars:
1:chrom 2:name 3:start 4:seq 5:end 6:score 7:strand 8:quality
```

When used with a format it will populate internal variables such as:
- $name, $seq, $start, $end etc.

It also works with both gzipped or regular text files: it’s magic no less
How to characterize the effect of mutations?

What effect could it have?

Quite a few options – many with commercial connotations

- **snpEff**: Snp Effects (we’ll try this in class)
- **Annovar**: Annotating Variants - commercialized via Biobase
- **VAAST**: Variant Annotation, Analysis and Search Tool - commercialized via Omicia
- **VEP**: Variant Effect Predictor (Ensemble)
- **VAT**: Variant Annotation Tool

Task: annotate the differences between the 1999 and 2014 genomic builds

1. What is the target and what is the reference?
2. Which aligner will work? These are long sequences.
3. Which snp caller will work? We have just two sequences
4. Ensure the annotation will match
All pieces are interconnected

Note how to solve this you have to go back to the first lectures

• Identify the data: GenBank
• Get it: efetch
• Index it for bwa and samtools
• Align and create a BAM file: bwa mem
• Call variants but there is only one sequence: samtools
• Annotate variations: snpEff

Resulting visualization

Homework 24

• Find the genome that snpEff uses to annotate variants. This will be used as reference.

• Align data from one run of the 2014 outbreak against this genome.

• Call SNPs then produce a SNP effect report. Summarize two findings from the report.