Week 12, Lecture 24

István Albert

Biochemistry and Molecular Biology and Bioinformatics Consulting Center

Penn State
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

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
Bioawk examples

```bash
$ cat r1.fq | bioawk -c fastx '{ print $name }' | head -1
NC_002549_8570_9028_8:0:0:6:0:0:0/1
$
$ cat r1.fq | bioawk -c fastx '{ print $seq }' | head -1
AGGACCACCCCTGATCGAGCAGCTGATCATCCATAGATAATTATCGAGGTGCCTAGCGTCAATCAAGGAG
$
$ cat mutations.gff | bioawk -c gff '{ print $start }' | head -1
244
$ ```
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
SnpEff

Genetic variant annotation and effect prediction toolbox.

Download SnpEff

Latest version 4.0 E (2014-09-13)
Requires Java 1.7

SnpEff
Genetic variant annotation and effect prediction toolbox. It annotates and predicts the effects of variants on genes (such as amino acid changes). Features:
- Supports over 20,000 genomes.
- Cancer variants analysis
- GATK compatible (-o gatk)
- HGVS notation
- Sequence Ontology standardized terms

Version 4.0
Major improvements and support for standards:
- HGVS notations
- Sequence Ontology terms
- Easier to use
- SnpEff downloads databases automatically
- Automatic third party databases downloads
- Support for GRCh38
- Support for Ebola Zaire Virus (2014 West Africa outbreak)

SnpSift
SnpSift helps filtering and manipulating genomic annotated files (VCF). Once you annotated your files using SnpEff, you can use SnpSift to help you filter large genomic datasets in order to find the most significant variants

View details »
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
snpEff report

Number of effects by type and region

<table>
<thead>
<tr>
<th>Type (alphabetical order)</th>
<th>Count</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>downstream_gene_variant</td>
<td>1,287</td>
<td>42.9%</td>
</tr>
<tr>
<td>intergenic_region</td>
<td>107</td>
<td>3.567%</td>
</tr>
<tr>
<td>intragenic_variant</td>
<td>68</td>
<td>2.267%</td>
</tr>
<tr>
<td>missense_variant</td>
<td>68</td>
<td>2.267%</td>
</tr>
<tr>
<td>stop_retained_variant</td>
<td>1</td>
<td>0.033%</td>
</tr>
<tr>
<td>synonymous_variant</td>
<td>305</td>
<td>10.167%</td>
</tr>
<tr>
<td>upstream_gene_variant</td>
<td>1,164</td>
<td>38.8%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Region</th>
<th>Count</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>DOWNSTREAM</td>
<td>1,287</td>
<td>42.9%</td>
</tr>
<tr>
<td>EXON</td>
<td>374</td>
<td>12.467%</td>
</tr>
<tr>
<td>INTERGENIC</td>
<td>107</td>
<td>3.567%</td>
</tr>
<tr>
<td>NONE</td>
<td>68</td>
<td>2.267%</td>
</tr>
<tr>
<td>UPSTREAM</td>
<td>1,164</td>
<td>38.8%</td>
</tr>
</tbody>
</table>

Variations

- Upstream
- 5'UTR
- Exon
- Intron
- Splice Donor
- Splice Acceptor
- 3'UTR
- Downstream
- Intergenic
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.