Lab Meeting Slides

10-18-17

Annotator

- **What:** Annotates
- **Why:** People doing analysis like to know where their features lie
- **How:**
  - `module load annotator`
- **Code:** https://github.com/byee4/annotator
- **Input:**
  - **BED6 File** (chrom, start, stop, name, score, strand)
  - **GTFDB File**
  - **Species** (default: hg19)
- **Output:**
  - BED6 + 5
    - **GeneID** (ENSG00000142949.12)
    - **GeneName** (PTPRF)

Annotator: Method

- For each line, get every overlapping feature
- If 2+ transcripts overlap:
  - Group every transcript into respective gene
  - Prioritize transcripts
- If 2+ genes overlap:
  - Prioritize transcripts first (so every gene has ONE transcript)
  - Prioritize genes

Annotator: Required Arguments

```
annotator \
--input BED6_FILE \              # Will truncate the 7+ column!!
--output OUTPUT_FILE \           # Output
--gtfdb GTF_DB_FILE \            # gffutils.create_db()*
--genome SPECIES                 # either: hg19 (default), mm10,
                                 or ce11
```

***Current locations (may change):**

(hg19):
/projects/ps-yeolab/genomes/hg19/gencode_v19/gencode.v19.annotation.gtf.db
(mm10):
/projects/ps-yeolab/genomes/mm10/gencode/gencode.vM10.annotation.gtf.db

Annotator: Optional Arguments

```
annotator \

...                                     ...

--gene-priority-file TXT \      # sets the gene priority

--transcript-priority-file TXT # sets the transcript priority
```

Example priority format:

```
1  protein_coding,CDS
2  protein_coding,start_codon
3  protein_coding,stop_codon
4  protein_coding,5utr
5  protein_coding,3utr
6  protein_coding,intron
7  protein_coding,Selenocysteine
8  non_coding,exon
9  non_coding,intron
```

- Unlisted regions will be lowest priority
- Needs an interactive node