

Range queries

Genomic arithmetic with BedTools - <https://bedtools.readthedocs.io/en/latest/>

Often want to ask questions about genomic ranges. For example.

- What are genes that are found on Chromosome 1?
- What are sequence reads that are aligned to Gene X
- How many RNA sequence reads are aligned to Protein coding genes?

Some starter example code:

https://github.com/biodataprogram/GEN220_2019_examples/tree/master/Bioinformatics_1/Ranges

```
#!/bin/bash -l
```

```
module load bedtools
```

```
set -euo pipefail
```

```
bedtools intersect -a rice_chr6.fixed_Chrom.gff -b rice_chr6_3kSNPs_filt.bed -wo > snp_gene_intersect.tab
```

```
# how many features have SNPs?
```

```
cut -f3 snp_gene_intersect.tab | sort | uniq -c
```

```
# how many SNPs does each gene have?
```

```
grep -P "\tgene\t" snp_gene_intersect.tab > snp_gene_intersect.genes_only.tab
```

```
# this outputs gene SNP counts ordered by gene name which is actually chromosome
```

```
# position nicely
```

```
cut -f9 snp_gene_intersect.genes_only.tab | sed 's/^ID=//; s/;Name=.*//' | sort | uniq -c > gene_snp_count.txt
```

```
# which genes have the most snps?
```

```
sort -nr gene_snp_count.txt > gene_snps_count.by_number.txt
```