BEDTOOLS COMPARISON

AKA an abuse of bedtools intersect for my own devilish purposes :)

Combined.gtf has 14623 transcripts

- Includes mRNA, cuts, nuts, xuts, etc

Merge_downloaded_gtf has ~5500 transcripts

- Flattened contained isoforms, results in ~1000 transcripts being thrown out

First - lets do a comparison between merged_final and combined. Let's says we pull out anything that is

No overlap with any known transcript

```
bedtools intersect -v -a downloadable_collapsed.gtf -b combined.gtf >
totally_unique.gtf
```

- 322 transcripts

No overlap with any known transcript -s forces strandedness

```
bedtools intersect -v -s -a downloadable_collapsed.gtf -b combined.gtf >
totally_unique_stranded.gtf
```

- 1700 transcripts

```
/home/agreenla/gffcompare-0.12.2.Linux_x86_64/gffcompare -C -o completely.unique
totally_unique_stranded.gtf
```

Let's compare between mRNA and annotations

```
bedtools intersect -F 1 -s -wa -a downloadable_collapsed.gtf -b
mRNAonly.gtf > mRNA_overlap_completely.gtf
```

- Must cover 100% of gene
- 714 extended transcripts - but some genes double cause spliced

How many antisense transcripts

```
bedtools intersect -F .3 -S -wa -a downloadable_collapsed.gtf -b
mRNAonly.gtf > Antisense_tomRNA_overlap_30p.gtf
```

- 1804 antisense transcripts
    - Probably an undercount

```
bedtools intersect -F .2 -S -wa -a downloadable_collapsed.gtf -b
mRNAonly.gtf > Antisense_tomRNA_overlap_20p.gtf
/home/agreenla/gffcompare-0.12.2.Linux_x86_64/gffcompare -C -o antisense_20
Antisense_tomRNA_overlap_20p.gtf
```

```
bedtools intersect -F .4 -S -wa -a downloadable_collapsed.gtf -b
mRNAonly.gtf > Antisense_tomRNA_overlap_40p.gtf
```

```
/home/agreenla/gffcompare-0.12.2.Linux_x86_64/gffcompare -C -o antisense_40
Antisense_tomRNA_overlap_40p.gtf
```

```
/home/agreenla/gffcompare-0.12.2.Linux_x86_64/gffcompare -C -o antisense_30
Antisense_tomRNA_overlap_30p.gtf
```

How many transcripts are both antisense and extended?

```
bedtools intersect -F 1 -s -wa -a  mRNA_overlap_completely.gtf -b
Antisense_tomRNA_overlap_30p.gtf > antisense_extended2.gtf
```

- 486

- 70% different from any known transcript

How many transcripts are partial genes? - short

```
bedtools intersect -F .2 -s -wa -a downloadable_collapsed.gtf -b
mRNAonly.gtf > mRNA_overlap_20.gtf
```

```
Then subtract all extended transcripts - if this works it should be small
amount
```

```
bedtools intersect -v -s -a mRNA_overlap_20.gtf -b mRNA_overlap_1.gtf >
mRNA_frag1.gtf
```

- This also pulls transcripts which are extended but with weird start site - not total overlap with gene
- ~50 or fewer fragment genes

How many transcripts are completely intergenic? - no mRNA overlap

```
bedtools intersect -v -wa -a downloadable_collapsed.gtf -b mRNAonly.gtf >
totally_intergenic_2.gtf
```

- 1275

```
/home/agreenla/gffcompare-0.12.2.Linux_x86_64/gffcompare -C -o intergenic
totally_intergenic_2.gtf
```

Make a gtf to determine if extension i 3’ or 5’

```
bedtools slop -i mRNAonly.gtf -g saccer3.genome -l 100 -r 0 -s >
100_3_mRNA.gtf
```

```
bedtools slop -i mRNAonly.gtf -g saccer3.genome -l 200 -r 0 -s >
200_3_mRNA.gtf
```

```
bedtools slop -i mRNAonly.gtf -g saccer3.genome -l 0 -r 100 -s >
100_5_mRNA.gtf
```

```
bedtools slop -i mRNAonly.gtf -g saccer3.genome -l 0 -r 200 -s >
200_5_mRNA.gtf
```

```
/home/agreenla/gffcompare-0.12.2.Linux_x86_64/gffcompare -C -o extended_total_number
mRNA_overlap_completely.gtf
714 extended transcripts or 10% of genes
```

```
bedtools intersect -F 1 -s -wa -a mRNA_overlap_completely.gtf -b
100_3_mRNA.gtf > 3_extended_100_annos.gtf
   -  489 of 789
```

```
/home/agreenla/gffcompare-0.12.2.Linux_x86_64/gffcompare -C -o 3_extended_100_annos
3_extended_100_annos.gtf
```

```
bedtools intersect -F 1 -s -wa -a mRNA_overlap_completely.gtf -b
200_3_mRNA.gtf > 3_extended_200_annos.gtf
```

```
/home/agreenla/gffcompare-0.12.2.Linux_x86_64/gffcompare -C -o 3_extended_200_annos
3_extended_200_annos.gtf
   -  489 of 714
```

```
bedtools intersect -F 1 -s -wa -a mRNA_overlap_completely.gtf -b
100_5_mRNA.gtf > 5_extended_100_annos.gtf
```

```
bedtools intersect -F 1 -s -wa -a mRNA_overlap_completely.gtf -b
200_5_mRNA.gtf > 5_extended_200_annos.gtf
```

```
/home/agreenla/gffcompare-0.12.2.Linux_x86_64/gffcompare -C -o 5_extended_200_annos
5_extended_200_annos.gtf
```

```
490 of 714
```

```
bedtools intersect -F 1 -s -wa -a  5_extended_200_annos.combined.gtf -b
3_extended_200_annos.combined.gtf > extended_in_both_directions.gtf
```

```
/home/agreenla/gffcompare-0.12.2.Linux_x86_64/gffcompare -C -o 5_3_extended_200_annos
extended_in_both_directions.gtf
276
```

```
bedtools intersect -F 1 -s -v -wa -a  downloadable_collapsed.gtf -b
3_extended_200_annos.combined.gtf > unclassed.gtf
```

OKAY what’s left???
Bedtools subtract
Bleh

```
bedtools subtract -r -f 1 -s -a downloadable_collapsed.gtf -b
Antisense_tomRNA_overlap_20p.gtf > no_antisense.gtf
```

```
bedtools subtract -r -f 1 -s -a no_antisense.gtf -b
totally_intergenic_2.gtf > no_antisense_no_intergenic.gtf
```

```
bedtools subtract -r -f 1 -s -a no_antisense_no_intergenic.gtf -b
mRNA_overlap_1.gtf > no_antisense_no_intergenic_no_extended.gtf
```

```
/home/agreenla/gffcompare-0.12.2.Linux_x86_64/gffcompare -C -o unclassed
no_antisense_no_intergenic_no_extended.gtf
```

GO TERMS

```
Pull Upstream extended
bedtools intersect -f 1 -s -wa -a mRNAonly.gtf -b 3_extended_200_annos.gtf
> mRNAs_with_extended_200.gtf

bedtools intersect -f .30 -S -wa -a mRNAonly.gtf -b
Antisense_tomRNA_overlap_30p.gtf > mRNAs_with_AS.gtf
```

All not extended antisense transcripts:

```
bedtools subtract -s -F 1 -A –a Antisense_tomRNA_overlap_30p.gtf –b
antisense_extended2.gtf > AS_not_Extended.gtf
```

Antisense transcripts near large intergenics:

```
bedtools intersect -S -F .3 -wa –a AS_not_Extended.gtf –b
mRNA_600_intergenics.gtf > as_class_3_try1.gtf
```

```
/home/agreenla/gffcompare-0.12.2.Linux_x86_64/gffcompare -C -o AS_class3_try1gff
as_class_3_try1.gtf
```

```
awk '{if($4 == transcript) print}' AS_class3_gff2bed.bed >
AS_class3_gff2bed_transcriptsonly.bed
```

```
awk '($4=="transcript")' AS_class3_gff2bed.bed >
AS_class3_gff2bed_transcriptsonly.bed
```

Get those class 2 genes

Get those intergenics

```
bedtools subtract -a chroms.bed -b mRNAonly.gtf > allintergenics.bed
```

```
bedtools intersect -wa -s -f 1 -a allintergenics.bed  -b
antisense_extended2.gtf > class2_intergenics.bed
```

```
bedtools sort -i class2_intergenics.bed >
class2_intergenics_Sorted.bed
```

```
bedtools merge -s -i class2_intergenics_Sorted.bed >
merged_class2_intergenics.bed
```