# Genome Summary Statistics

After all the post-bakta processing, it is useful to generate some genome summary information. This includes; genome size, number of contigs, number of tRNAs, number of coding sequences, coding density, hypotheticals, total length of rage regions, total length of interRAGE regions, percentage of the genome in RAGE regions, the number of RAGE regions, the number of complete RAGEs, the number of spoT-synthase, bifunctional spoT and truncated spoT/RelA, and the number of both ankyrins and tetratricopeptide containing proteins.

As this information comes from 5 different sources for each genome, it was decided not to turn this into a single package/chunk of code. But rather to provide the code for each individual bit of information.

## Sources of information used:

1. The genome.fasta file (post-polishing genome file)
2. The bakta genome.txt output (in the bakta output directory)
3. RAGE derived bed files (found in processing outputs after running the main.sh script)
4. Complete RAGE bed files (found in processing outputs after running the main.sh script)
5. Fully annotated gbff file (found in processing outputs after running the main.sh script)

## Command line instructions to obtain this information:

|  |  |
| --- | --- |
| Statistic | Method (gbff files are the ones post processing) |
| genome\_length\_(bases) | seqkit stats genome.fasta (seqkit needs to be installed) |
| number\_of\_contigs | seqkit stats genome.fasta |
| number\_of\_tRNAs | less genome.txt (bakta genome.txt output) |
| number\_of\_coding\_sequences | less genome.txt (bakta genome.txt output) |
| coding\_density | less genome.txt (bakta genome.txt output) |
| hypotheticals | less genome.txt (bakta genome.txt output)(this may be lower with ankyrin and tpr annotation... I’ve decided this probably doesn’t matter as these proteins are still largely hypothetical) |
| size\_of\_RAGE\_regions | for file in \*.bed; do awk '{sum += $3 - $2} END {print FILENAME, sum; sum=0}' "$file"; done (files in rage\_derived) |
| size\_of\_interRAGE\_regions | =genome\_size-size\_of\_RAGE\_regions (done in excel spreadsheet) |
| percentage\_genome\_RAGE | =size\_of\_RAGE\_regions/genome\_size\*100 (done in excel spreadsheet) |
| number\_of\_RAGE\_regions | wc -l processing\_outputs/rage\_derived/\* |
| number\_of\_complete\_RAGEs | wc -l processing\_outputs/complete\_rage/\* |
| number\_spoT-synthase | grep -c "Orientia SpoT-synthetase" \*.gbff |
| number\_bifunctional\_spoT | grep -c "full length bifunctional spoT" \*.gbff |
| number\_of\_truncated\_RelA/SpoT\_homolog\_proteins | grep -c "truncated RelA/SpoT homolog (RSH) protein" \*.gbff |
| number\_of\_ankyrins | grep -c -i "ankyrin" \*.gbff |
| number\_of\_tetratricopeptides | grep -c -iE "tpr |tetratricopeptide" \*.gbff |