Genome Annotation Comparison in Ornithorhynchus anatinus

Bio 466 Final Project
Davyd Sadovskyy
11/29/2022

Introduction:

For the purpose of this project, I have chosen the Platypus (Ornithorhynchus anatinus) as the organism of interest. Platypuses live in the waterways of eastern and southern Australia, including Tasmania. Platypuses are still relatively common in the wild, but were recently reclassified as ‘vulnerable’ because of their reliance on an aquatic environment that is under stress from climate change and degradation by human activities. Water quality, erosion, destruction of habitat and food resources, and disease now threaten populations. Because the platypus has rarely bred in captivity and is the last of a long line of Ornithorhynchidae monotremes, their continued survival is of great importance.

The platypus genome, as well as the animal, is an amalgam of ancestral reptilian and derived mammalian characteristics. The platypus karyotype comprises 52 chromosomes in both sexes, with a few large and many small chromosomes, reminiscent of reptilian macro- and microchromosomes. Platypuses have multiple sex chromosomes with some homology to the bird Z chromosome. Males have five X and five Y chromosomes, which form a chain at meiosis and segregate into 5X and 5Y sperm. Sex determination and sex chromosome dosage compensation remain unclear. The most striking physical feature of the platypus is its simultaneous laying of eggs and nursing of its young with milk. Some other special features are a unique gastrointestinal system, neuroanatomy (electro-reception) and a venom delivery system.

In this study, I compare different versions of Platypus Gene Annotation data obtained from the ensembl database. Gene annotation involves the process of taking the raw DNA sequence produced by the genome-sequencing projects and adding layers of analysis and interpretation necessary to extracting biologically significant information and placing such derived details into context. Genome annotations change frequently and are becoming more and more accurate. Sometimes there can be vast differences between annotation versions. For this analysis, I measure different metrics across two annotation versions and create a website that allows a user to extract information about a gene from a database containing gene, transcript, and exon data.

Results:

	97 Release	108 Release
Total Genes	29776	26272
Total Transcripts	32009	42892
Chromosomes/Contigs	20 chromosomes 15122 contigs	27 chromosomes 62 contigs
Genes With Names	10027	12251
Non Redundant Gene Names (gene names that don’t repeat for multiple gene_ids)	7192	11791
Unique Gene Names (gene ids that are only present in one of the releases)	3959	6194
Unique Gene IDs (gene ids that are only present in one of the releases)	18879	15375

Table 1. Basic comparison information between versions. This table appears in the website.

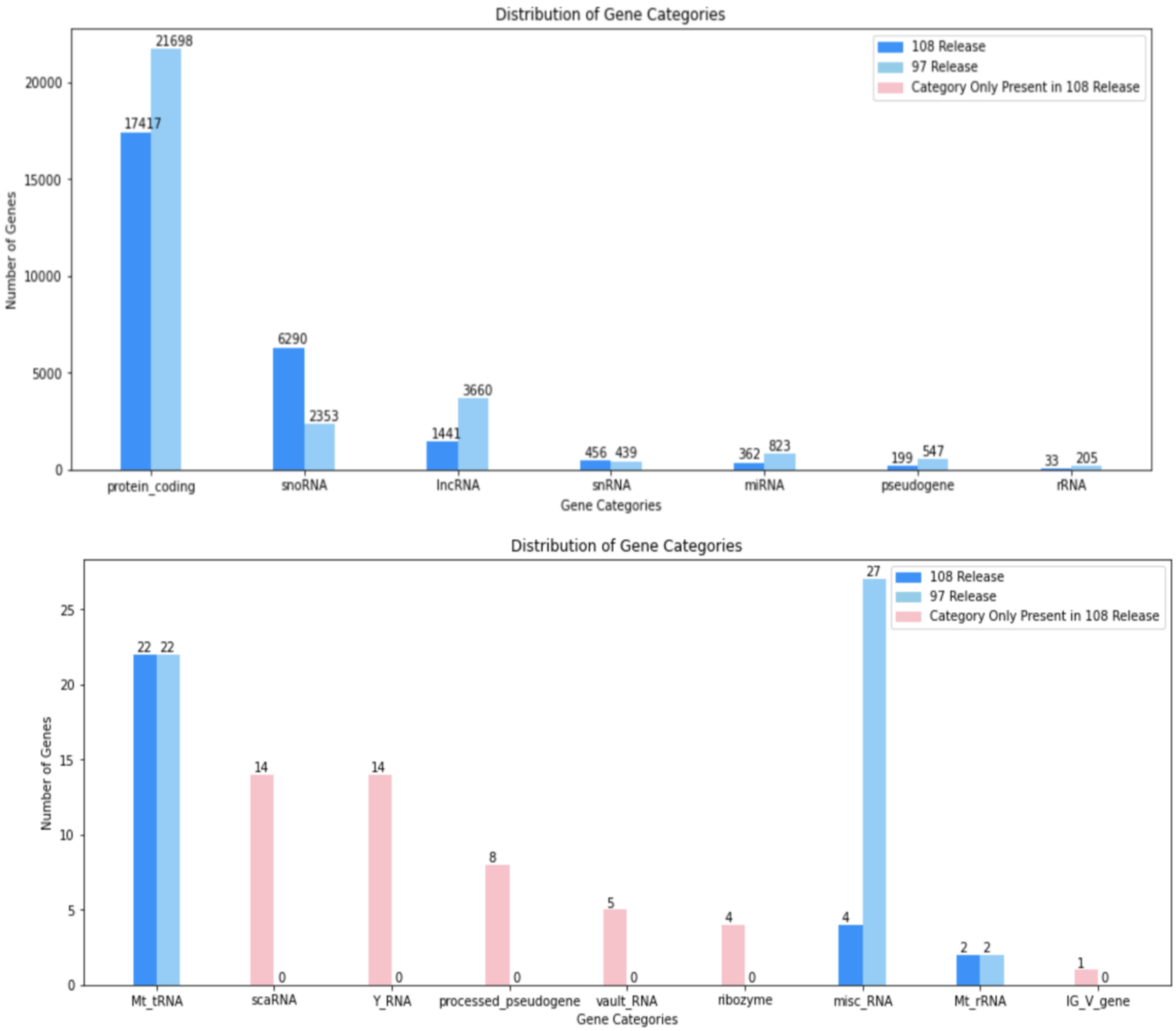

Figure 1. Looking at the Distribution of gene categories for 97 and 108 releases, 16 different gene categories were observed across the both assemblies. The Pink bars are new gene categories introduced in 108 assembly. This figure appear in the website

Transcript distributions (number of transcripts per gene category) were almost identical to the gene distributions shown in figure 1. Only the protein and lncRNA categories had a higher transcript number than gene number.

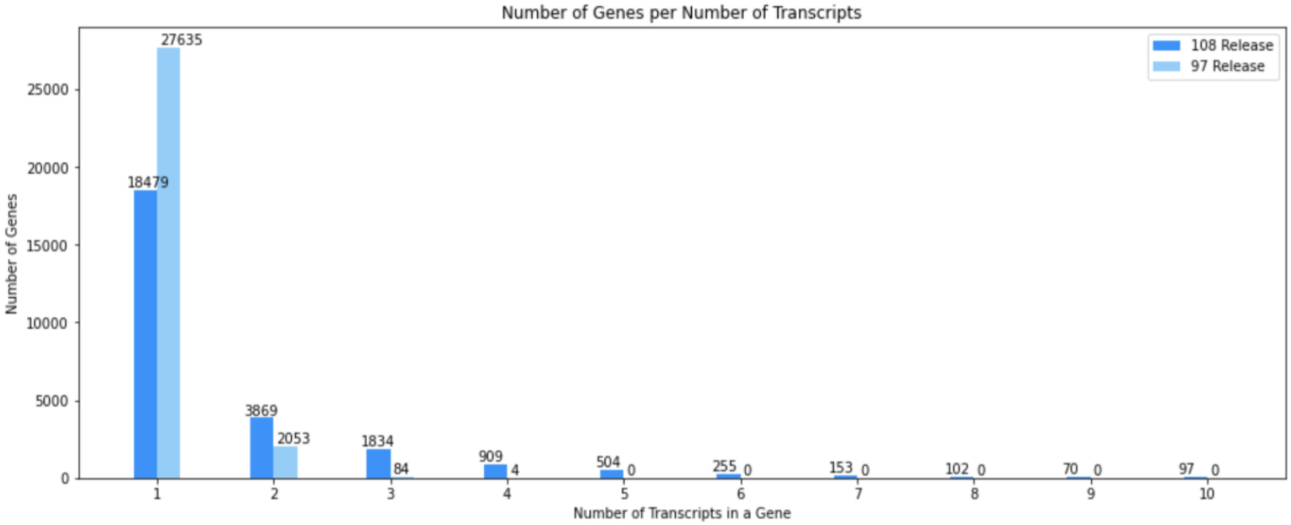

Figure 2. Most genes are **1 to 1** i.e, they have 1 transcript per gene. The 108 Assembly has more genes with multiple transcript numbers In 97 version, the highest # of transcripts for a gene is 4. In the 108 version, it's 10. In fact, 97 genes in the 108 version have 10 transcripts.

Discussion:

Performing genome annotations is as much of a biological and computational problem as it is philosophical. Every organism contains a certain defined set of genes in specific locations in its genome - every gene and transcript already exists, and does not change, except with long term evolution. Even the latest genome annotations are merely a representation or model of what the actual genome is. The annotations produced throughout the years vary widely in their gene and transcript information, but each successive version carries us closer to the base truth.

A central problem of comparing different annotations is finding which genes map to each other between the versions. Intuitively, it would seem that the gene id is enough of a discriminant between genes across annotation versions. In such a model, gene ids that are unique to a newer annotation would be newly discovered genes, or past genes that were redefined into new ones. This is not entirely the case, however. My analysis shows that some genes only present in release 108, ones that are expected to be completely new genes, actually have counterparts in release 97 - they have a gene name that is identical to that of a gene in release 97 with a completely different gene id. There are 1109 such genes in release 108. The converse is also true. There are 1114 genes with unique gene ids in release 97, each mapping to a gene in release 108 with a different gene id but the same name. Subtracting 1109 from 15375, the number of unique gene ids in release 108, would show that 14266 genes are apparently completely new genes discovered in the new annotation, something that is obviously not possible. These examples show that it is challenging to discover which genes have been altered or updated between annotations and which genes are completely new additions. Tracking progress between versions is a difficult task that requires encapsulating a gene with its critical features, more than just its gene id. Such attributes could be start and end positions, gene version, chromosome, or even other attributes that aren't indicated in a GTF file. Given my current skill set and bioinformatics knowledge, it is possible that these problems could very well have trivial solutions. Future steps will involve researching the topic of gene annotations and finding what the current criteria are by which to compare assembly versions.

For future assembly comparison steps, I could further examine genes that have clear one-to-one mapping between annotations (the ones with the same gene id). For these genes I could do things like find the set that gained or lost transcripts between versions.

Possible future steps for my user interface would be allowing a user to input queries about the type of gene to extract from a release. These queries could already be hardcoded into sql in the backend and various inputs could be made available to the user that represent such queries. For example, a button labeled “snoRNA” could have a dropdown menu of how many of all the snoRNA genes to display on a page. Also, another search option could be added that allows a user to input a specific gene name instead of id.