CSCI 652-BioInformatics - I David Blerc

Vamsi Chunduri

Ashton Dsouza

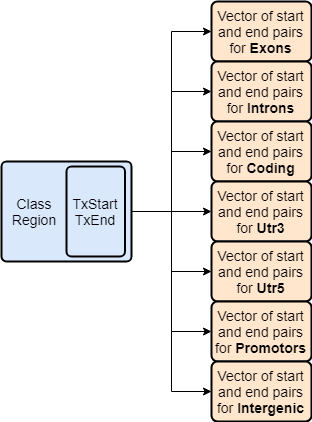
**Study of Divergence of Mammalian Genomes**

**Purpose:**

In this project, we study the divergence among human vs dog, mouse and chimpanzee for different regions i.e. intergenic regions, intron regions, exons, coding regions, 5’ UTRs, 3’ UTRs, and promoters.

**Framework:**

Our program is written in C++ and compiled on the Northern Illinois University hopper server. The program is based on labeling region Start and End areas mentioned above and shown in figure 1. We first build a look up structure of regions from the KnownGene.txt. Next after properly sorting the array of regions by TxStart. We read in a pair from .maf file and ran the calculations for the proper overlapping sections. Finally, the results were printed and graphs were made.



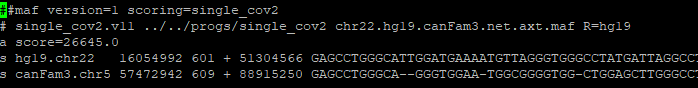
**Figure 1**

**Problems Encountered:**

1. The Length of .maf sequence. How does it properly align to the known gene text file?

Human sequence = 601 (only characters in sequence)

Sequence Length in program = 622 (expanded with gaps to align properly)

****

Our solution, cut all gaps from the human sequence while also cutting the other sequence in the same position.

2) Do inversions or mismatches of sequence segments effect calculation?

Both files .maf and Knowngene.txt had strand info with all cases (++,+-,-+,--)

Our solution, we just did the calculation for each position completely ignoring strand +/-.

3) How should separated exons be calculated when separated?

Example for an Exon case:

Region: Exon 1: Intron: Exon 2:

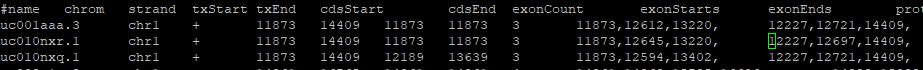
Species 1: ACTG- AGCT -GTCA

Species 2: GTCAC GGGG CCAAA

Would you calculate Exon1 and Exon2 together or separate?

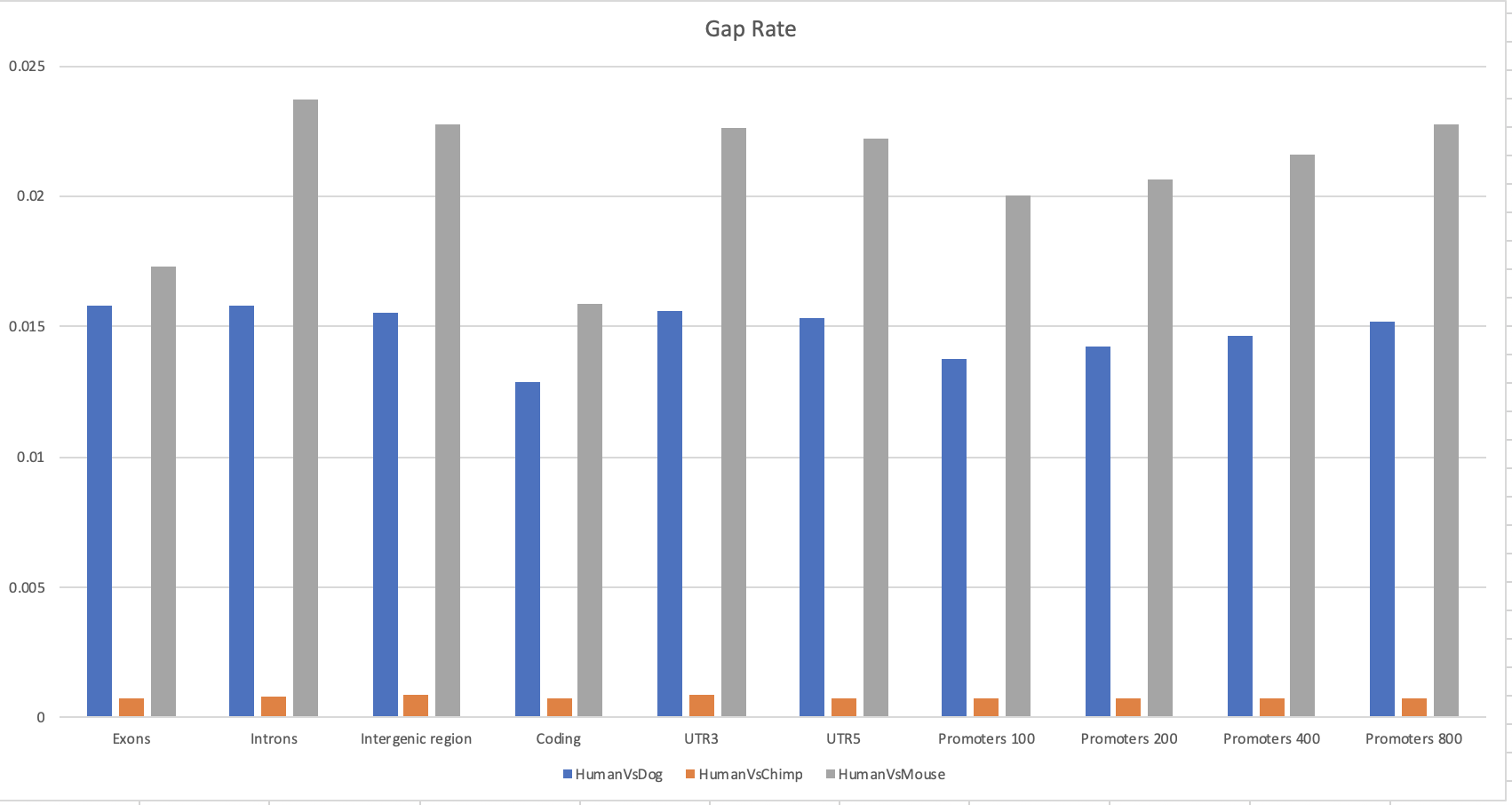
* Seems to only affect the gap lengths and counts.
* Our solution, just run calculations all separate.

4) How should overlapping exons be counted?

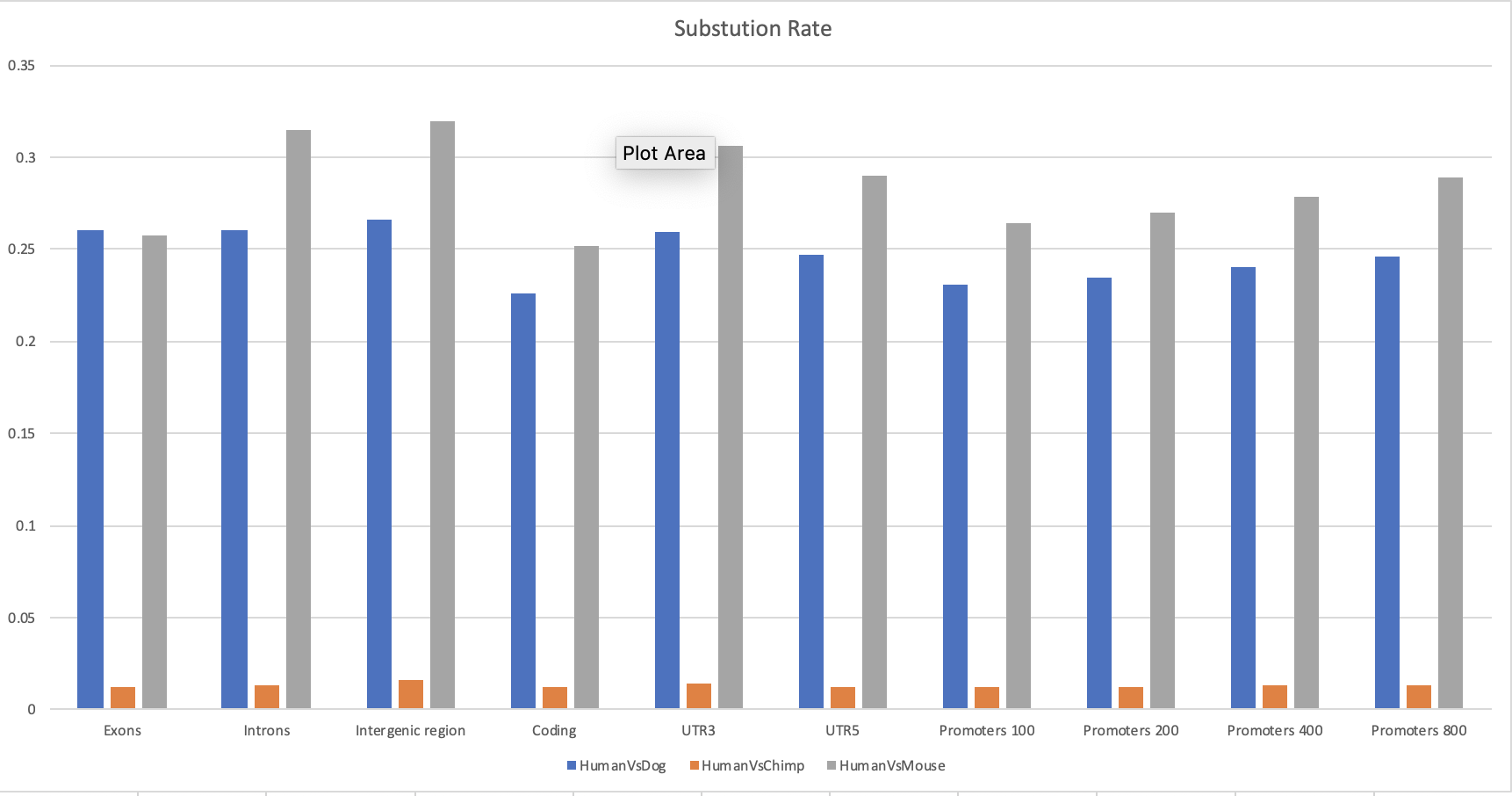


Our solution, just count everything that was given.

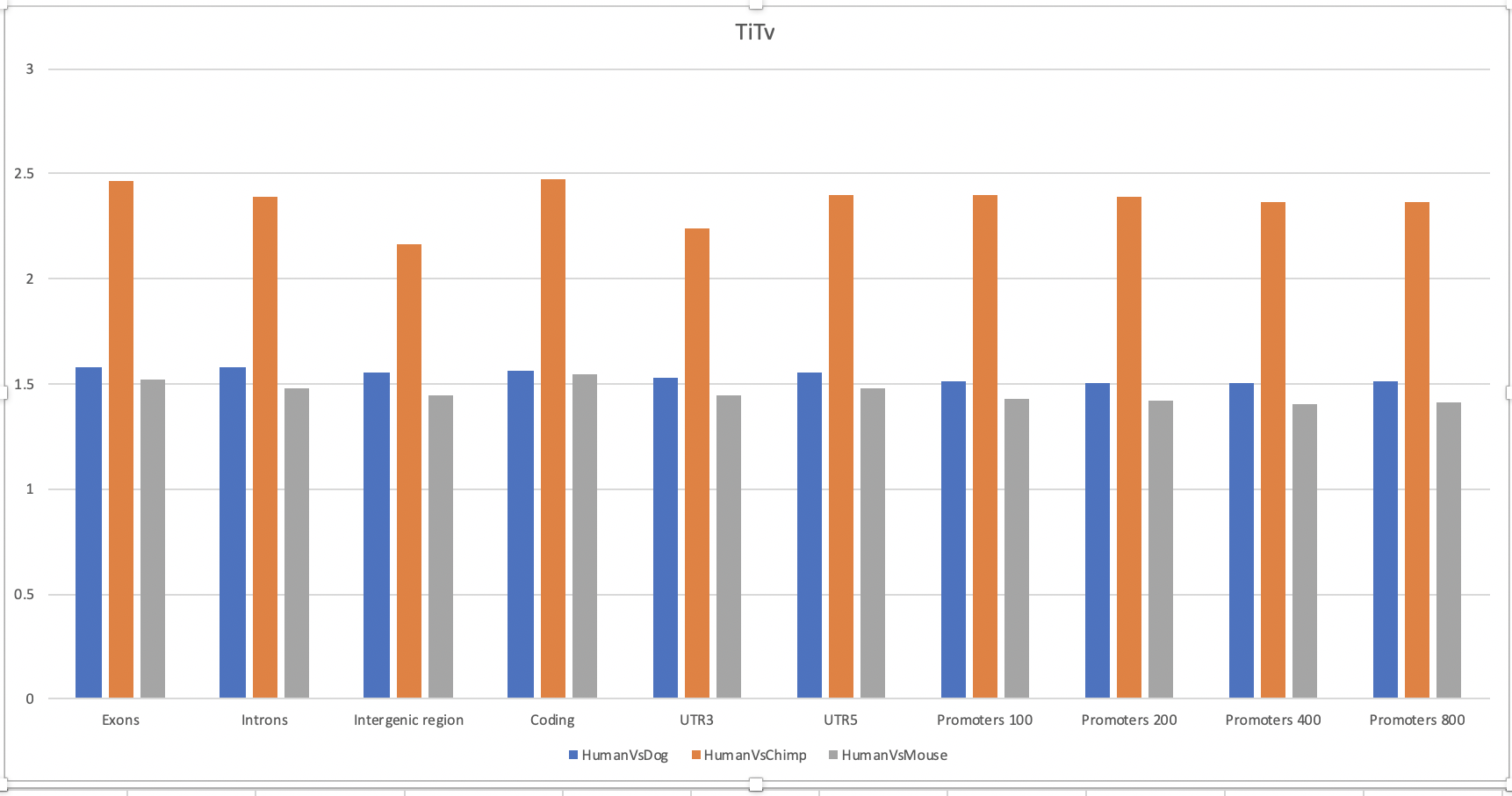
**Results:**

Gap Rate:****

Substitution Rate:

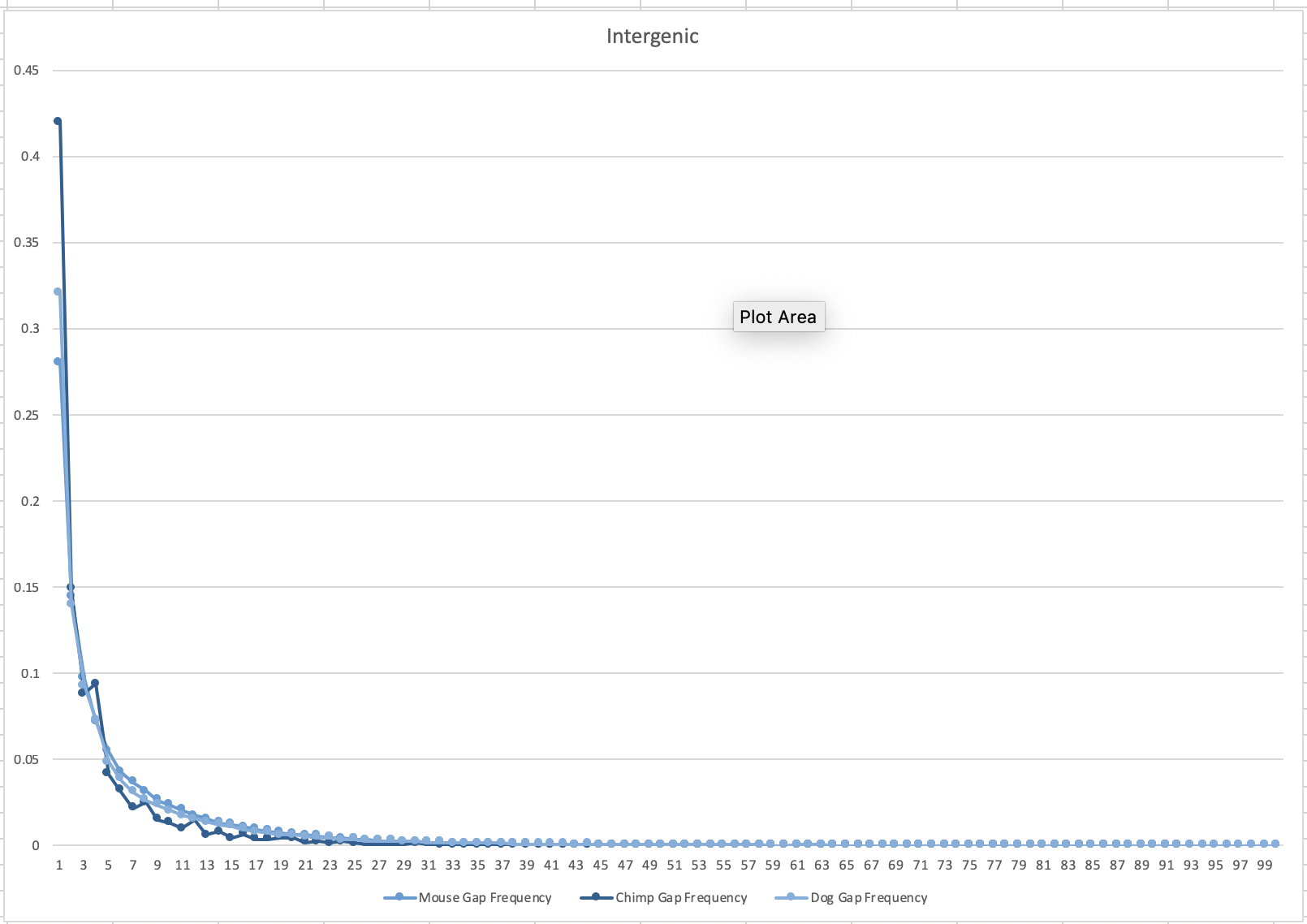
****

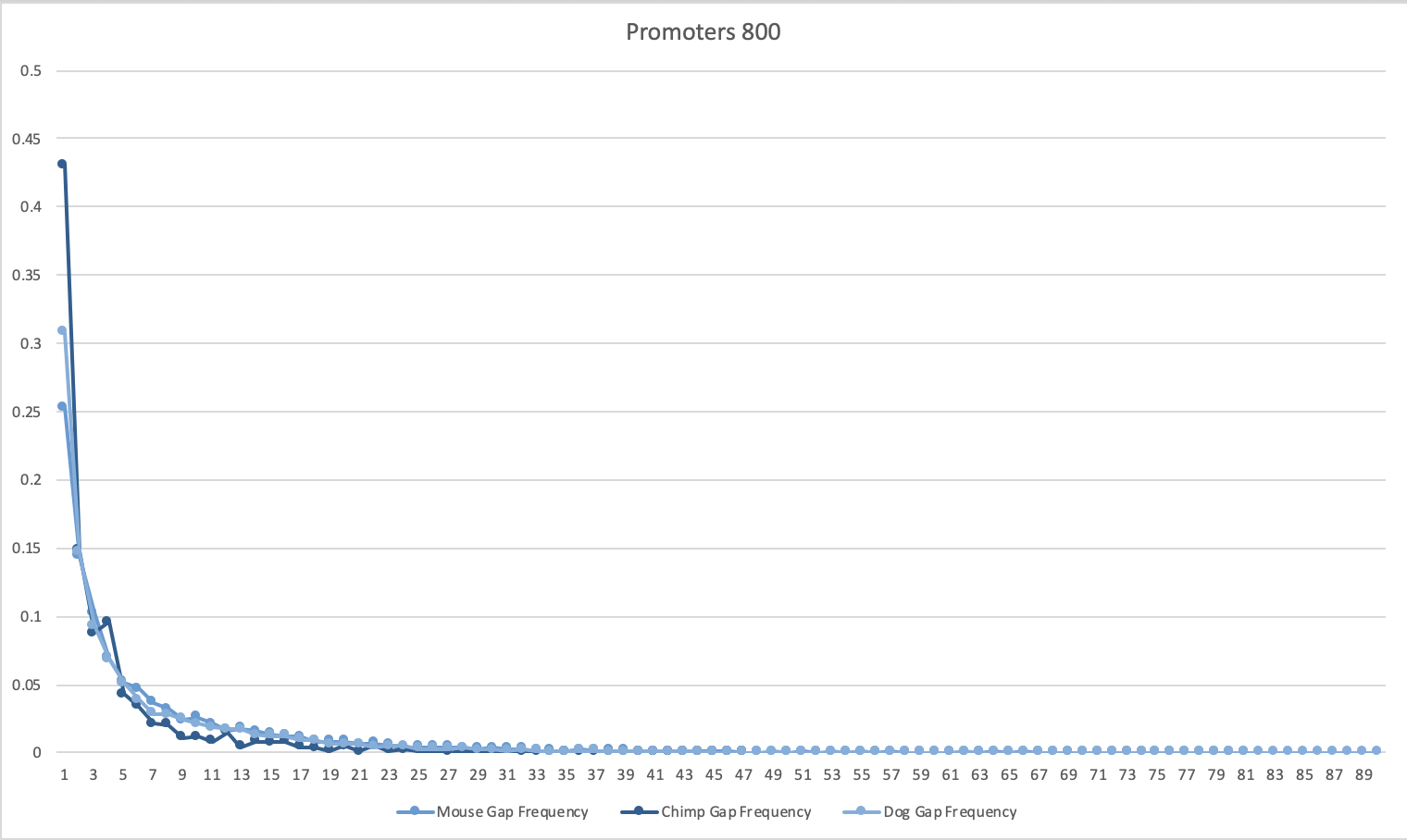
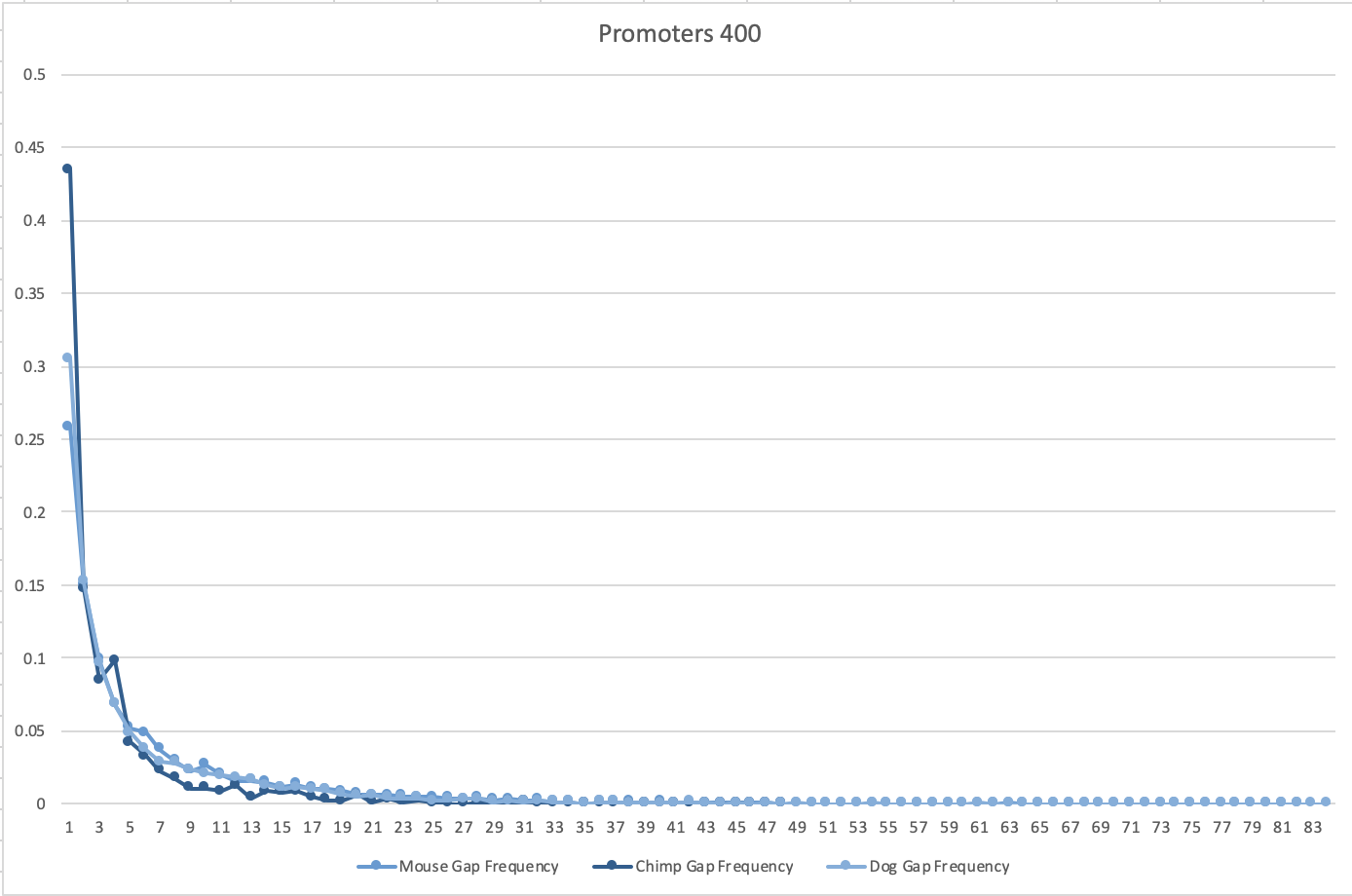
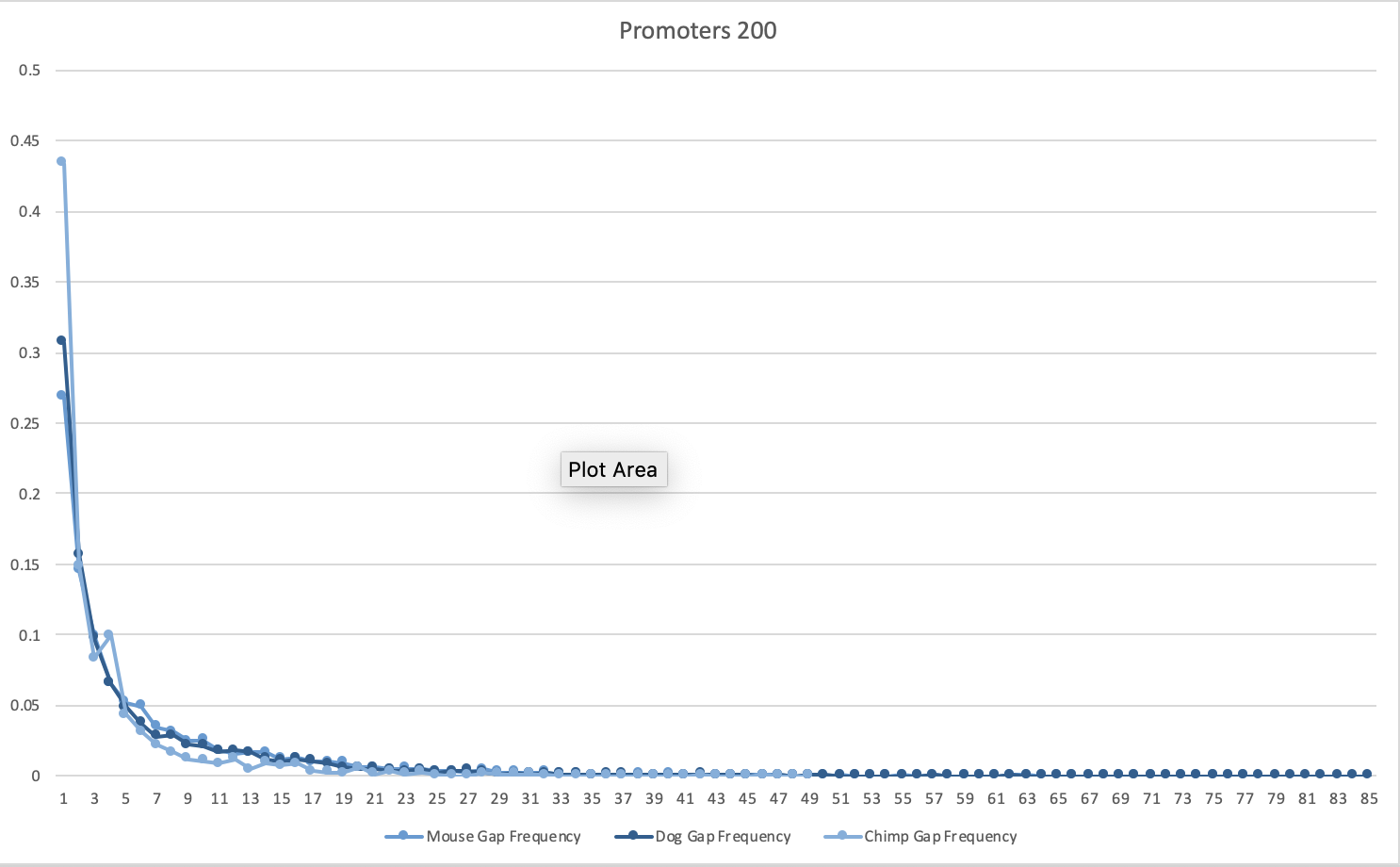
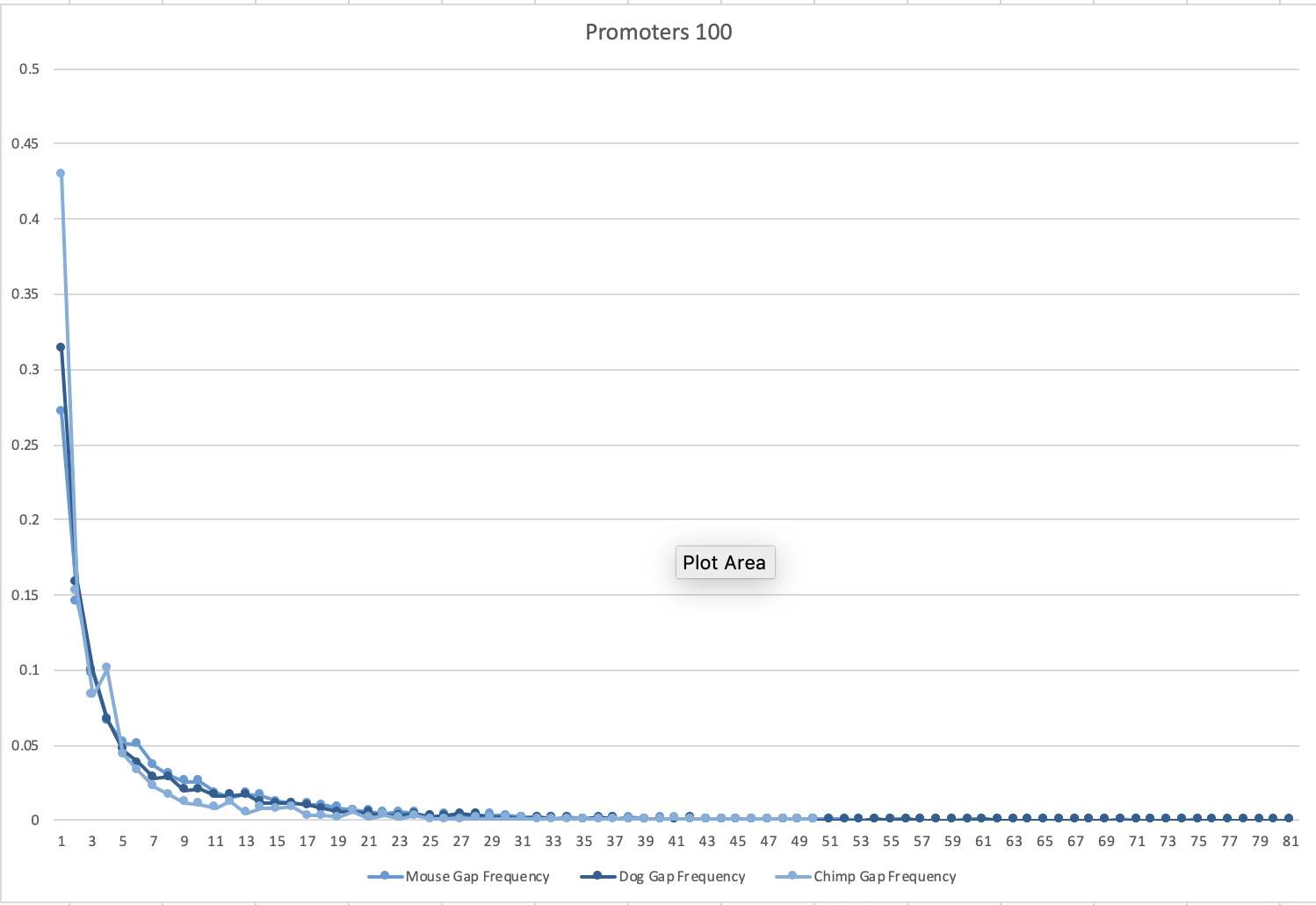
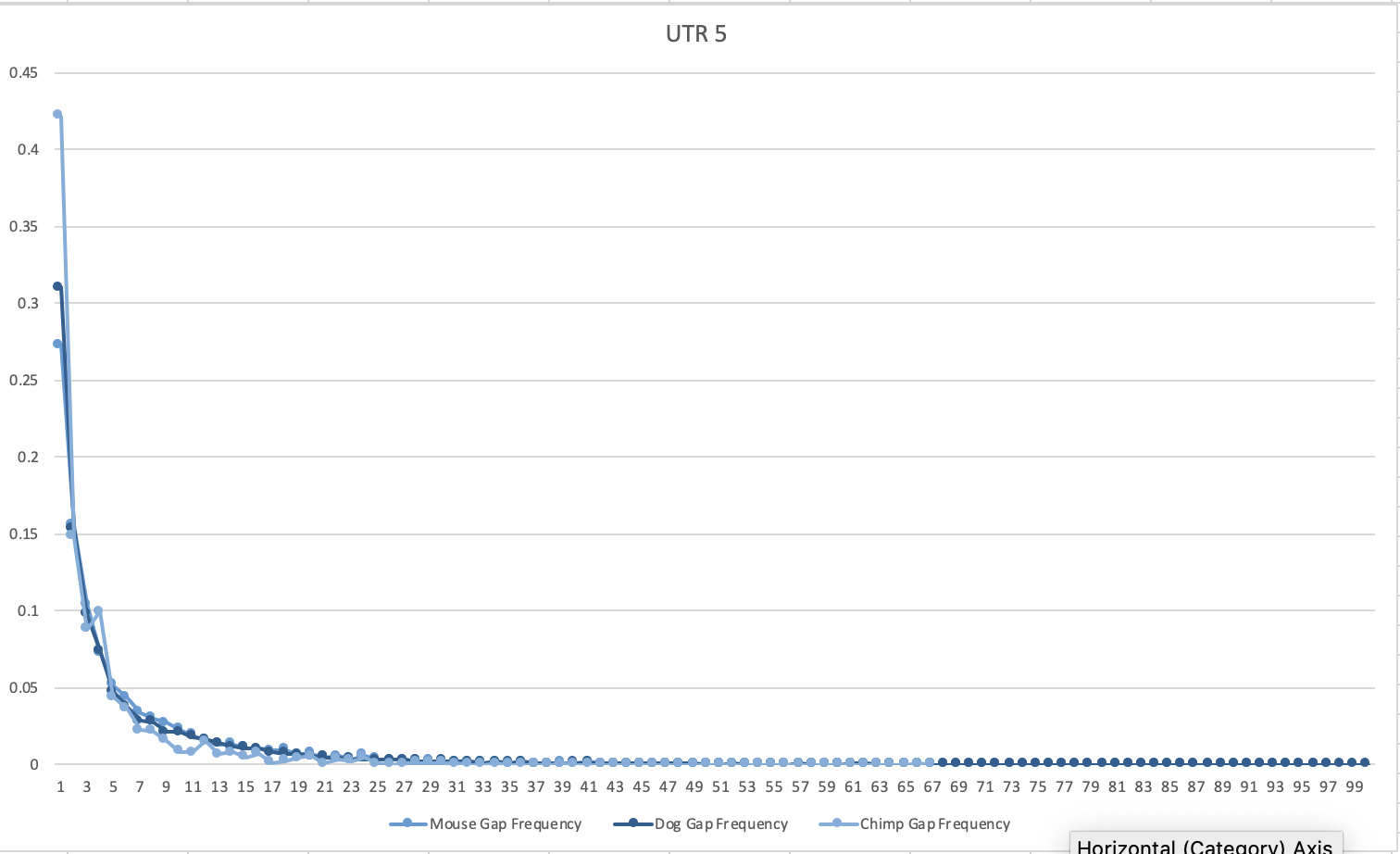
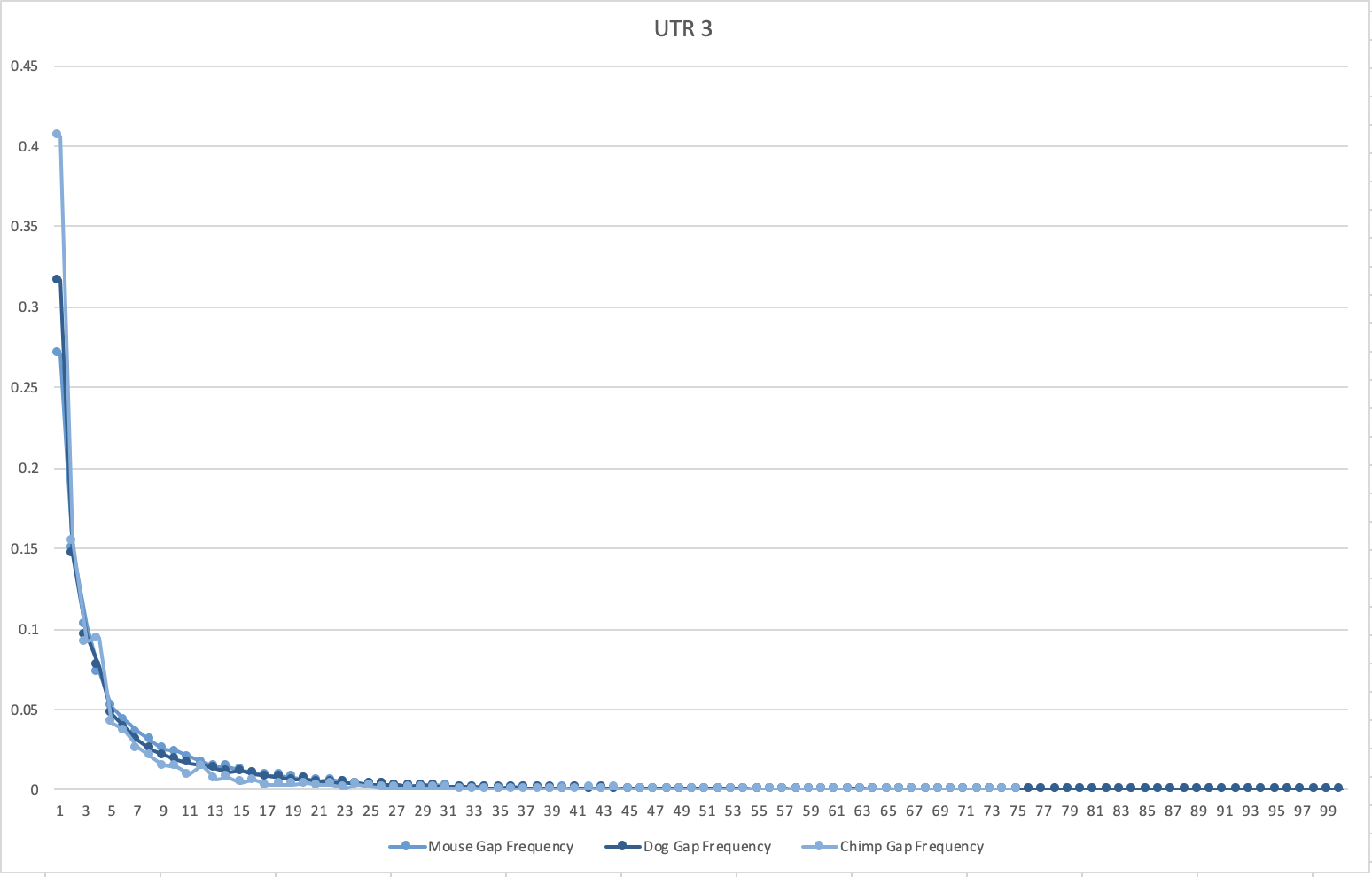
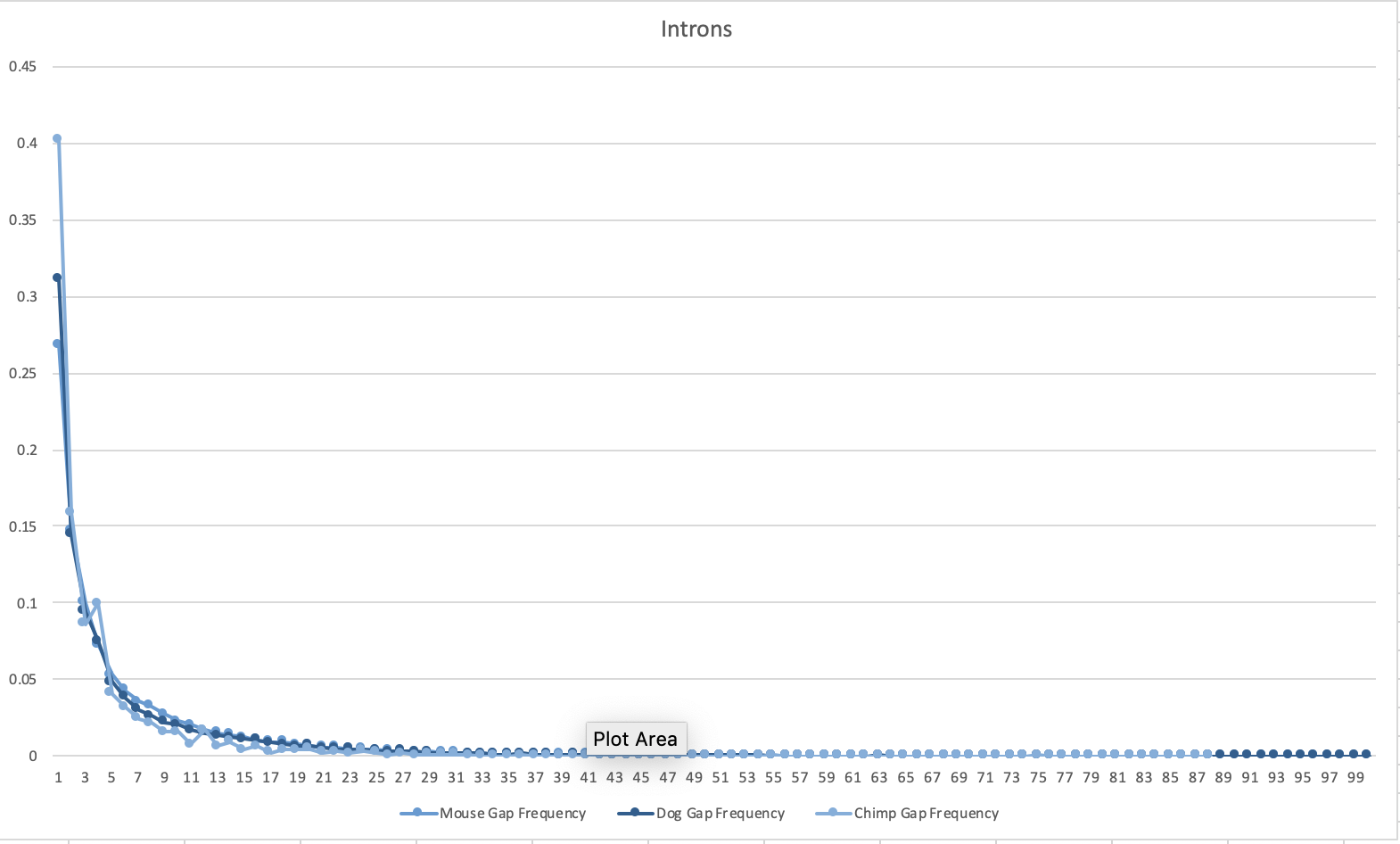
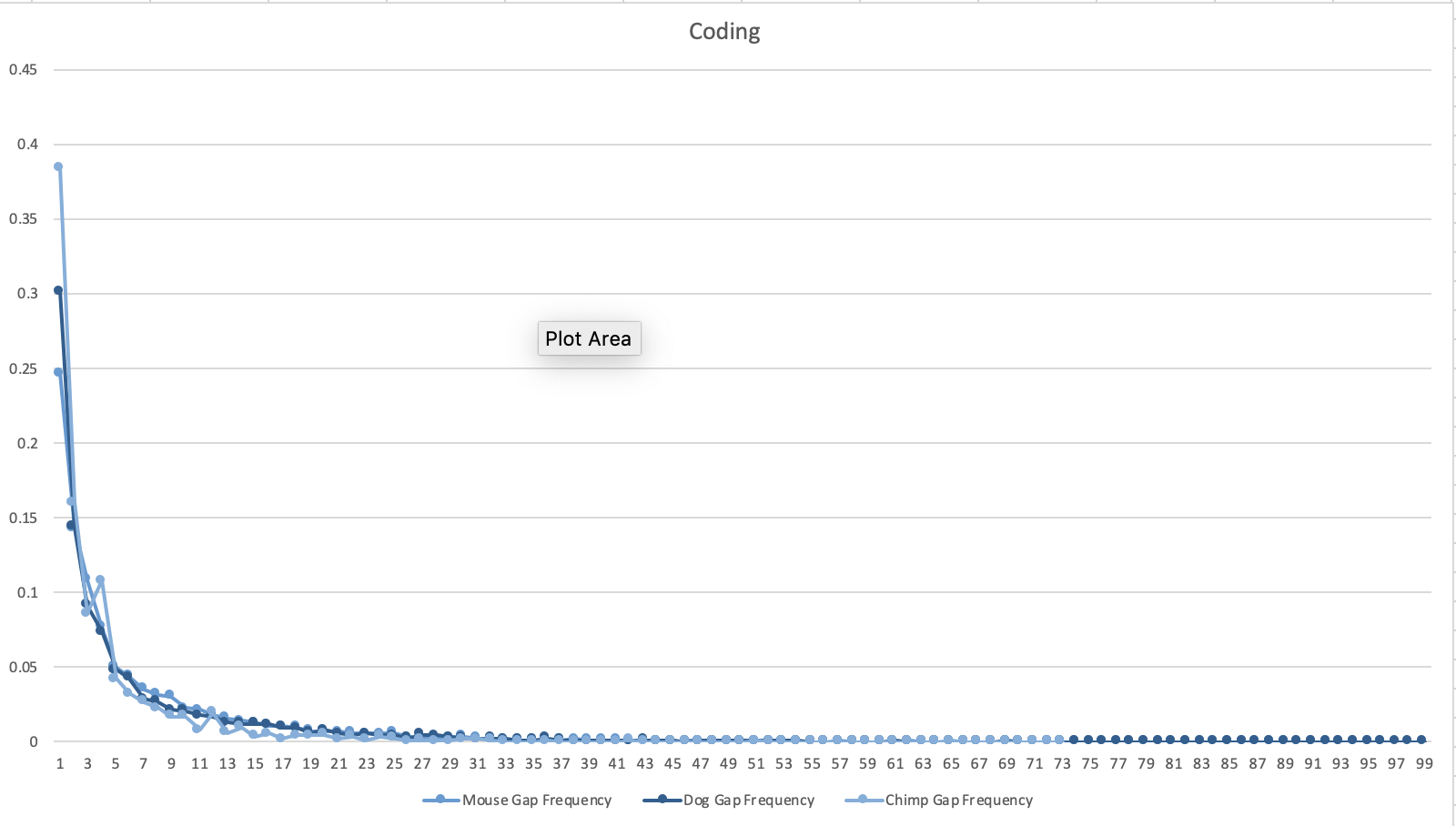
TiTv:

****

Gap Frequencies vs the Gap Length:

Y-axis x-axis

****

****

**Conclusion:**

Based on the above data humans and chimps are closer in gene structure when compared against a dog or mouse with a overall gene match of 99%.The mouse and dog are a good model since they represent a very close to the human genome. This is further reinforced by the graphs shown above. The substitution rates and gap rates for the Human-Chimp genomes are much lower than other species comparisons, which means than there are less differences in ACGT amongst these genomes. Fewer differences lead to similarities in genome which is evident in the data. Similarly the the values for Ti/Tv ratio for humans and chimps lie between 2 and 2.5 compared to values ranging from 1.3 to 1.7 for the other species.