# Part I: Project Review

## What is the general scientific question?

How to improve Zhao et al. (2015) 's pipeline and analyse gene loss in birds?

## What is my specific scientific question?

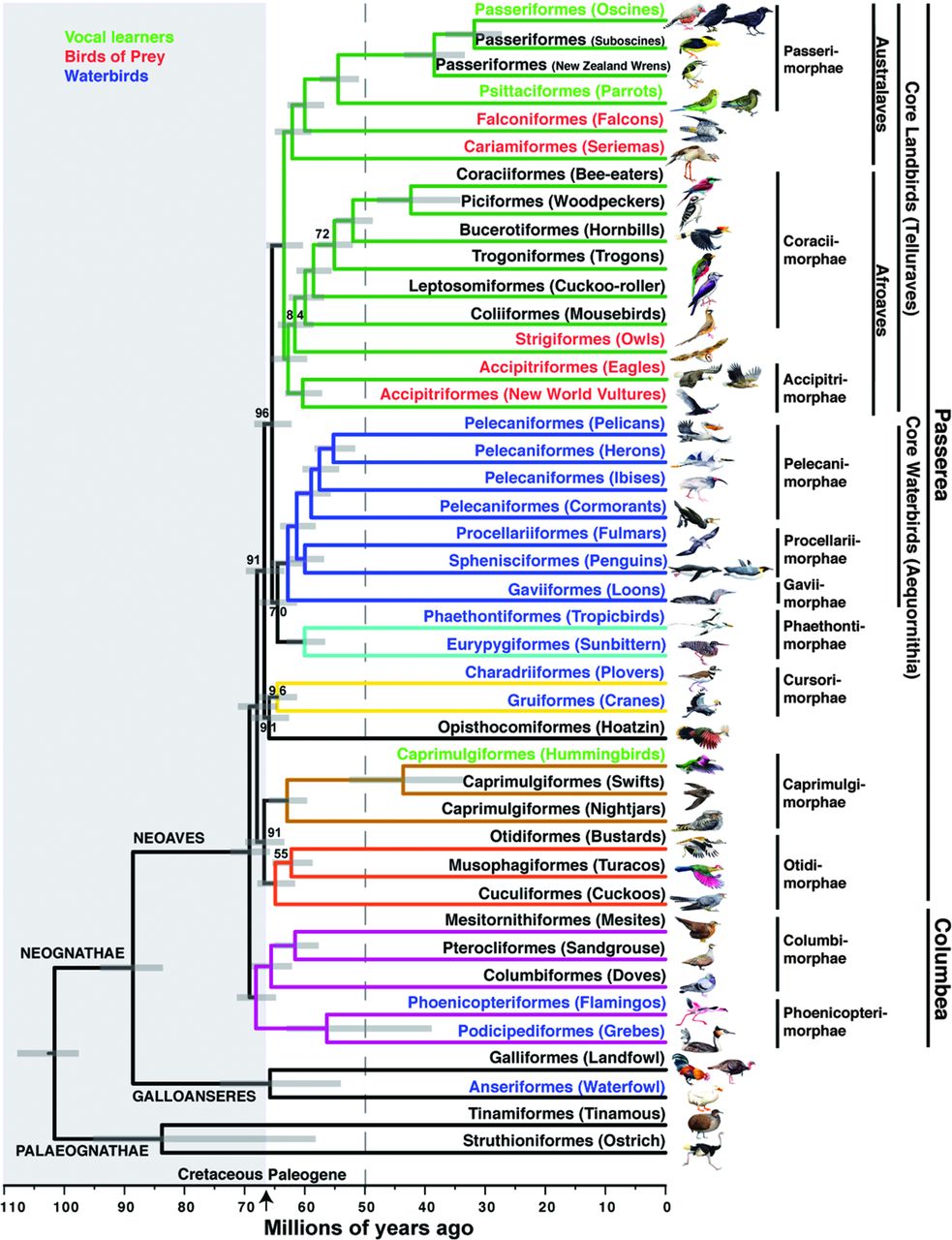
1. How to further filter the result of PseudoPipe?
2. How to detect unitary gene loss in birds? (How to choose *long-established* protein-coding genes? How to choose appropriate divergent time? How will the pipeline perform to a large number of species?)
3. How does gene loss correlate with genome size/gain/loss?

## What is my hypothesis?

1. We can further filter the result of PseudoPipe by perfrom ortholog prediction again using [HaMStR](https://bmcevolbiol.biomedcentral.com/articles/10.1186/1471-2148-9-157).
2. We can use Palaeognathae （古颚总目) as outgroup to investige unitary gene loss in Galloanseres (鸡雁小纲) and neoaves (新鸟小纲), which diverged about **90** Myrs ago. Or use Palaeognathae and Galloanseres as outgroup to investige lineages within neoaves (新鸟小纲), which diverged time about **67** Myrs ago ( close to the WGD

## What is my experiment design plan to test the hypothesis?

1. 收集鸟类的基因组数据和系统发育关系。



Jarvis et al., 2014, figure 1

## What is the new technology that I need to develop?

1. 在超级计算机上执行任务

* 已经熟练掌握 Linux 基本操作和 R、C++语言。
* 需要学习如何在超算上运行程序。

1. 处理 PseudoPipe 的输出

* 目前解决程序的问题，可以顺利运行。但只是把其输出直接传给 pipeline， 还需要掌握如何自己分析输出结果。

1. 学习 Kapusta et al. (2017) 的分析方法。

## What is the timeline and milestones?

Week01~04 收集数据，运行完 pipeline，学习如何自己分析 PseudoPipe 的输出。

Week05~08 使用 HaMStR 处理 "genic hits" 和 "intergenic hits" 中ORF完好的情况，得出 gene loss

Week09~12 利用 Kapusta et al. (2017) 的方法，分析本项研究的类群的 genome gain/loss。

Week13~16 分析 gene loss number 与 genome size/gain/loss 之间的关系，撰写毕业论文。

# Part II: Status Update

## What I have already finished?

基本完成 pipeline 的运行。

1. 进一步验证 PseudoPipe 需要的 chr\*\_exLocs

* 为了进一步验证 mysql/exon.txt 和 gtf 是一致的，我将 gene name，start coordinate，end coordinate 三者 paste 成一个字符串，结果表明两个文件的信息完全一样。
* 也就是说，在外显子的坐标上，两个文件是完全一致的。不过之前已经发现了在某些 mRNA 的剪接上二者有分歧，按目前的需求下，这件事可以画上一个句号了[^严格来说，以蜀黎为例，gtf 中含有很多没有确定染色体的 exon（比如 gl002604.1）。也许 Ensembl ftp 的 mysql 文件夹中的其他文件或其它某个文件夹中储存了染色体信息。至少我检查了 ftp 的每个文件夹，以及查阅了 https://www.ensembl.org/info/docs/api/core/core\_schema.html 中关于mysql 文件夹中的每个文件的每列意义的文档，还是没找到染色体信息。]。

1. synteny relationships

* 关于之前提到的 Schnable et al. 提供的 EVS009\_Supplemental\_dataset\_S1\_pluslinks.xls 中的基因名与 Ensembl 15 中不一致的问题，我又核实了一遍。我从 gtf 文件第 9 列 和 pep.fa 文件的序列民中提取出全部基因名称，然后用 EVS009\_Supplemental\_dataset\_S1\_pluslinks.xls 中的基因名来检索。结果还真有找不到的。

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| species | brachypodium | rice | maize | sorghum | brachypodium & rice | maize & sorghum |
| gene loss | 132 | 102 | 245 | 88 | 27 | 35 |