Gene extraction & alignment for comparative genomics

Summer RCN Mentoring 2021

OUTLINE

Part 1&2: Gene Extraction

- Why do comparative genomics?
- Gene Models
- Sequence Extraction
- BLAST (Part 2 is if we get time)
- Homology
- Translation

Part 3: Multiple Alignment

- Codon aware alignment
- Alignment quality control

LOCAL VS GLOBAL ALIGNMENT

- **Local alignment:** finding matching short substrings
- BLAST is a local alignment
- **Global alignment:** match sequences along their entire length

Local Alignment

Target Sequence

```
5' ACTACTAGATTACTTACGGATCAGGTACTTTAGAGGCTTGCAACCA 3'
             |||| |||||| |||||||||||||||
Query Sequence 5' TACTCACGGATGAGGTACTTTAGAGGC 3'
```

Global Alignment

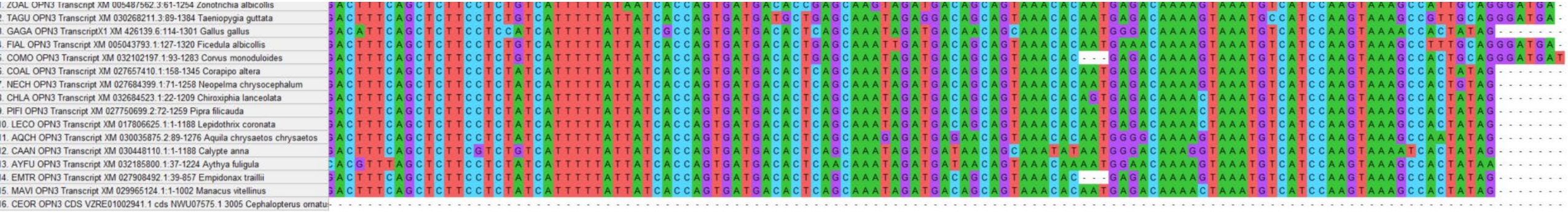
Target Sequence

```
5' ACTACTAGATTACTTACGGATCAGGTACTTTAGAGGCTTGCAACCA 3'
   |||||||||||    |||||||  |||||||||||||| |||||||
5' ACTACTAGATT----ACGGATC--GTACTTTAGAGGCTAGCAACCA 3'
```

Query Sequence

MULTIPLE SEQUENCE ALIGNMENT

- More than 2 sequences for global alignment



MULTIPLE SEQUENCE ALIGNMENT METHODS

- **Dynamic programming**
  - Slow – $O(L^N)$ computations (N sequences of L length)…
  - Programs: MSA, Multialign
- **Progressive alignment**
  - Use some search algorithm to align most similar pairs first.
  - Programs: T-Coffee, Clustal family alignment
- **Iterative alignment**
  - Progressive alignment + realignment of sequence subsets to improve initial alignments
  - Programs: MUSCLE (MAFFT has options of iterative and progressive)

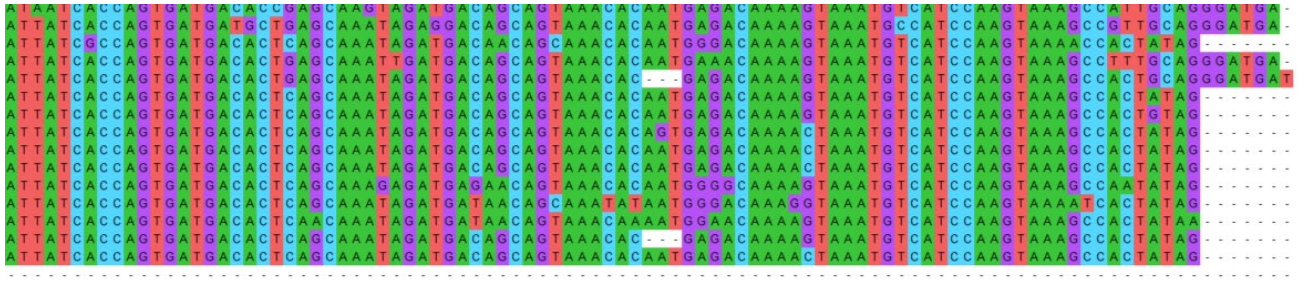
MULTIPLE SEQUENCE ALIGNMENT METHODS

- CLUSTAL: Clustal, Clustal W, Clustal X, Clustal O (latest)
  - Fast, multithreaded
  - good for sequences of similar lengths.
  - Bad for alignments of different lengths/big indels.
  - >2k input sequences
- MUSCLE
  - Good for multilength sequences
  - <1k sequences
  - Good for low homology ends
- MAFFT
  - Fast, multithreaded
  - Works well for >30k seq or long sequences
  - Good for low homology ends
- And MORE: check out https://www.ebi.ac.uk/Tools/msa/

https://help.geneious.com/hc/en-us/articles/360044627712-Which-multiple-alignment-algorithm-should-I-use-

Input a multi-fasta file (multiple sequences in a single text file (.fasta, .fa, .fna, .fas))

Alignment program:
ClustalO
MUSCLE
MAFFT
PRANK… etc



Output multi-fasta file (file extensions may be .fasta or .mfa, .msa) or Clustal, Phylip, Nexus

CODON-AWARE ALIGNMENT

- What is it?
- Why do we need this for estimation of positive selection?

CODON-AWARE ALIGNMENT

- TranslatorX.co.uk

					16							34						52						70						88						106	
Hummingbird	ATG	AAG	CAG	AAA	TTT	ATA	CCA	TCA	GTT	CAA	ACT	ATT	TTG	CTG	CTT	GCT	CTG	ACT	GCA	ATG	GGC	CTG	ACA	GGT	CAA	TCA	CAC	CCT	GGA	AAG	CCT	ATG	ATA	ACG	AGA	TGT
HoodedCrow	ATG	AAA	CAG	AGA	TTC	ATT	TCA	TCA	GTT	CAA	ATT	ATT	TTG	CAG	CTT	GCT	CTG	ACT	ACA	GTA	GGT	CTG	ACA	GGA	CAA	ACG	TAT	CCT	GGA	AAA	CCT	CAG	ATA	ATA	AGA	TGT
Flycatcher	ATG	AAA	CAG	AGA	TTC	ATT	TCA	TCA	GCT	CAA	ATA	ATT	TTG	CAG	CTG	GCT	CTG	ACT	ACA	GTA	GGT	CTG	ACT	GGT	CAA	ACA	TAC	CCT	GGA	AAA	CCT	AAG	ATA	ATA	AGA	TGT
GroundTit	ATG	AAA	CAG	AGA	TTC	ATT	TCA	TCA	GTT	CAA	ATA	ATT	TTG	CAG	CTT	GCT	CTG	ACT	ACA	GTA	GGT	CTG	ACT	GGT	CAG	TCG	TAC	CCT	GGA	AAA	CCT	CAG	ATA	ATA	AGA	TGT
Parus	ATG	AAA	AAG	AGA	TTC	ATT	TCA	TCA	GTT	CAA	ATC	ATT	TTG	CAG	CTT	GCT	CTG	ACT	ACA	GTA	GGT	CTG	ACT	GGT	CAG	TCG	TAC	CCT	GGA	AAA	CCT	CAG	ATA	ATA	AGA	TGT
Indigobird	ATG	AAA	GAG	AGA	TTC	ATT	TCA	TCA	GTT	CAA	ATT	ATG	TTG	CAG	CTT	GCG	CTG	ACT	ACA	GTA	GGT	CTG	ACT	GGT	CAA	TCA	TAC	CCT	GGA	AAA	CCT	CAG	ATA	GTA	AGA	TGT
ZebraFinch	ATG	AAA	CAG	AGA	TAC	ATT	TCA	Tca	gtt	caa	ata	att	ttt	cag	ctt	gCA	CTG	ACT	ACA	GTA	GGT	CTG	ACT	GGT	CAA	TCG	TAC	CCT	GGA	AAA	CCT	CAG	ATA	ATA	AGA	TGT
Cowbird	ATG	AAA	CAG	AGA	TTC	ATT	TCA	TCA	GTC	CCA	ATT	ATT	TTG	CAG	CTT	GCT	CTG	ACT	ACA	GTA	GGT	CTG	ACT	GGT	CAA	TCA	TAC	CCT	GGA	AAA	CCT	CAG	ATA	ATA	AGA	TGT
Redwing	ATG	AAA	CAG	AGA	TTC	ATT	TCA	TCA	GTC	CCA	ATT	ATT	TTG	CAG	CTT	GCT	CTG	ACT	ATA	GTA	GGT	CTG	ACT	GGT	CAA	TCA	TAC	CCT	GGA	AAA	CCT	CAG	ATA	ATA	AGA	TGT
Canary	ATG	AAA	CAG	AGA	TTC	ATT	TCA	TCA	Gtt	caa	att	att	ttg	cag	cTT	GCT	CTG	ACT	ACA	GTA	GGT	CTG	GCT	GGT	CAA	TCA	TAC	CCT	GGA	AAA	CCT	CAG	ATA	ATA	AGA	TGT
Sparrow	ATG	AAA	CAG	AGA	TTC	ATT	TCA	TCA	GTC	CCA	ATT	ATT	TTG	CAG	CTT	GCT	CTG	ACT	ACA	CTA	GGT	CTA	CCT	GGT	CAA	TCA	TAC	CCT	GGA	AAA	CCT	CAG	ATA	ATA	AGA	TGT
Chicken	ATG	AAA	CAG	GAT	TTG	ATA	TCG	TCT	GTT	CAA	ATC	ATA	TTG	TTC	CTT	CCT	CTG	ACC	ACA	GTG	GGT	CTG	GCA	GGT	CAA	TCA	TTC	CCT	GGA	AAA	CCT	AAG	ATA	ATA	AGA	TGT
Cuckoo	ATG	AAG	CAG	AAA	TTG	ATA	TCA	TCT	GTT	CAA	ATT	ATT	TTG	CTA	CAT	GCT	GTG	GCT	GTC	ATG	GGT	CTG	ACT	GGT	CAA	TCA	TTC	CCT	GGA	AAA	CCT	AAG	ATA	ATA	AGA	TGT
Falcon	ATG	AAG	CAG	AAA	TTG	ACG	TTA	TCA	GTT	CAA	ATT	ACT	TTG	CTG	CTC	ACT	GTG	GCT	GCA	GTG	GGT	CTG	ACT	GGT	CAA	TCA	CAC	CCT	GGA	AAA	CCT	AAG	ATA	ATA	AGG	TGT
Pigeon	ATG	AAG	CAG	AAA	TTG	AGA	TCA	TCA	GTT	CAA	ATT	ATT	TTG	CTA	TTT	GCT	CTG	ACT	GCA	GTG	GGT	CTG	ACT	GGT	CAA	TCA	TAC	CCT	GGA	AAA	CCT	AAG	ATA	ATA	AGA	TGT

A HIGHLY CONSERVED PORTION OF A GENE (PURIFYING SELECTION)

			844						862						880						898						916						934
GAA	TGT	CCT	GAT	TAC	AGA	AGT	GGG	GGC	CCC	AAT	TCA	TGC	TAC	TTT	GAT	AAA	AAG	CAC	ACC	TCT	TTC	TGG	ACC	GTA	TAC	AAT	ATT	ACT	GTG	AGG	GCA	ACC	AAT
GAA	TGT	CCA	GAT	TAC	AAA	ACT	GCA	GGC	CCC	AAT	TCA	TGC	TAC	TTC	GAT	AAA	AAG	CAC	ACC	TCT	TTC	TGG	ACC	ATA	TAC	AAT	ATT	ACT	GTG	AGG	GCA	ACT	AAT
GAA	TGT	CCA	GAT	TAC	AAA	ACT	GCA	GGC	CCC	AAT	TCA	TGC	TAC	TTT	GAT	AAA	AAG	CAC	ACC	TCT	TTC	TGG	ACC	ATA	TAC	AAT	ATA	ACT	GTG	AGG	GCA	ACT	AAT
GAA	TGT	CCA	GAT	TAC	AAA	ACT	GCA	GGC	CCC	AAT	TCG	TGC	TAC	TTC	GAT	AAA	AAG	CAC	ACC	TCT	TTC	TGG	ACC	ATA	TAC	AAT	ATT	ACT	GTG	AGG	GCA	ACT	AAT
gAA	TGT	CCA	GAT	TAC	AAA	ACT	GCA	GGC	CCC	AAT	TCG	TGC	TAC	TTC	GAT	AAA	AAG	CAC	ACC	TCT	TTC	TGG	ACC	ATA	TAC	AAT	ATT	ACT	GTG	AGG	GCA	ACT	AAT
GAA	TGT	CCA	GAT	TAC	AAA	ACT	GCA	GGC	CCC	AAT	TCA	TGC	TAC	TTC	GAT	AAA	AAG	CAC	ACC	TCT	TTC	TGG	ACC	ATA	TAC	AAT	ATT	ACT	GTG	AGG	GCA	ACT	AAT
GAA	TGT	CCA	GAT	TAC	AAA	ACT	GCA	GGC	CCC	AAT	TCA	TGC	TAC	TTC	AAT	AAA	AAG	CAC	ACC	TCT	TTC	TGG	ACC	ATA	TAC	AAT	ATT	ACT	GTG	AGG	GCA	ACT	AAT
GAA	TGT	CCA	GAT	TAC	AAA	ACT	GCA	GGC	CCC	AAT	TCG	TGC	TAC	TTC	GAT	AAA	AAG	CAC	ACC	TCT	TTC	TGG	ACC	ATA	TAC	AAT	ATT	ACT	GTG	AGG	GCA	ACT	AAT
GAA	TGT	CCA	GAT	TAC	AAA	ACT	GCA	GGC	CCC	AAT	TCG	TGC	TAC	TTC	GAT	AAA	AAG	CAC	ACC	TCT	TTC	TGG	ACC	ATA	TAC	AAT	ATT	ACT	GTG	AGG	GCA	ACT	AAT
GAA	TGT	CCA	GAT	TAC	AAA	ACT	GCA	GGC	CCC	AAT	TCG	TGC	TAC	TTC	GAT	AAA	AAG	CAC	ACC	TCT	TTC	TGG	ACC	ATA	TAC	AAT	ATT	ACT	GTG	AGG	GCT	ACT	AAT
GAA	TGT	CCA	GAT	TAC	AAA	ACT	GCA	GGC	CCC	AAT	TCG	TGC	TAC	TTC	GAT	AAA	AAG	CAC	ACC	TCT	TTC	TGG	ACC	ATA	TAC	AAT	ATT	ACT	GTG	AGG	GCA	ACT	AAT
GAA	TGT	CCA	GAT	TAC	AGA	ACT	GCA	GGC	CCC	AAT	TCC	TGC	TAC	TTT	GAT	AAA	AAA	CAC	ACT	TCT	TTC	TGG	ACC	ATA	TAC	AAC	ATT	ACT	GTC	AGG	GCA	ACT	AAT
GAA	TGT	CCA	GAT	TAC	AGA	ACT	GCA	GGC	CCC	AAT	TCG	TGC	TAC	TTT	GAT	AAA	AAA	CAT	ACC	TCT	TTC	TGG	ACC	ATA	TAC	AAT	ATT	ACT	GTG	AAG	GCA	ACT	AAT
GAA	TGT	CCA	GAT	TAC	AGA	ACT	GCA	GGC	CCC	AAT	TCA	TGC	TAC	TTC	AAT	AAA	AAG	TAT	ACA	TCT	TTC	TGG	ACC	ATA	TAC	AAT	ATT	ACT	GTG	AAG	GCA	ACT	AAT
GAG	TGT	CCA	GAT	TAC	AAA	ACT	GCA	GGC	CCC	AAT	TCA	TGC	TAC	TTC	GAT	AAA	AAG	CAC	ACC	TCT	TTC	TGG	ACT	ATA	TAC	AAT	ATA	ACT	GTG	AAG	GCA	ACT	AAT

A MORE COMPLICATED REGION

			1744						1762						1780						1798						1816		
Hummingbir	CTT	---	CAA	AGA	CAA	GAT	GGA	AGA	GAT	GCT	GAA	GAA	AAT	AAA	GAA	GGA	AAA	AGG	AGC	TGG	GAA	GCT	CAG	GGT	ATA	GCC	---	---	TCA
HoodedCrow	CTT	---	CAA	ACA	CAG	GTT	GTA	AGA	---	GCC	AAA	CAA	AGG	AAA	GGA	GGA	AAG	GGG	AGC	TGG	GAA	GCT	CAG	TCT	ATG	GCC	---	---	TCA
Flycatcher	CTT	---	CAA	ACA	CAA	GAT	GTA	AGa	---	gct	aaa	gaa	aag	aaa	gga	gga	aaa	ggg	agc	tGG	GAA	ACT	CAG	TGT	TCG	GCC	---	---	TCA
GroundTit	CTT	---	CAA	ACA	CAA	GAT	GTA	AGA	---	GCT	AAA	GAA	AAA	AAA	GGA	GGA	AAA	GGG	AGC	TGG	GAA	ACT	CAG	TGT	ACC	GCC	---	---	TCA
Parus	CTT	---	CAA	ACA	CAA	GAT	GTa	aga	---	gct	aaa	gaa	aaa	aaa	gga	gga	aaa	ggg	agc	tgg	gaa	aCT	CAG	TGT	ACG	GCC	---	---	TCA
Indigobird	GAT	GGA	AGA	GCT	AAA	GAT	GTA	AGA	---	GCT	AAA	GAA	AAG	AAA	GGG	GGG	AAA	GCG	AGC	TGG	GAC	ACT	CGG	TGT	ACA	GCC	---	---	TCA
ZebraFinch	CTT		CAA	ACA	CAA	GAT	GTA	AGa		gct	aaa	gaa	aag	aaa	ggg	ggg	AAA	GGG	AGC	TGG	GAC	ACT	CGG	TGT	ATG	GCC	---	---	TCA
Cowbird	CTT	---	CAA	ACA	CAA	GAT	GTA	AGA	---	GCT	AAA	GAA	AAG	AAA	GGG	GGG	AAG	GGG	AGC	TGG	GAC	ACT	CGG	TGT	GCG	GCC	TCG	ACC	TCA
Redwing	CTT	---	CAA	ACA	CAA	GAT	GTA	AGA	---	GCT	AAA	GAA	AAG	AAA	GGG	GGG	AAG	GGG	AGC	TGG	GAC	ACT	CGG	TGT	GCG	GCC	TCG	ACC	TCA
Canary	CTT	---	CAA	ACA	CAA	GAT	GTA	AGa	---	gct	aaa	gaa	aag	aaa	ggg	ggg	aaa	ggg	agc	TGG	GAC	ACT	cgc	tgt	gca	gcc	cca	gcc	tca
Sparrow	CTT	---	CAA	ACA	CAA	GAT	GTA	CGa	---	gct	aaa	gaa	aag	aaa	ggg	ggg	aaa	ggg	agc	TGG	GAC	ACT	CGG	TGT	GCG	GCC	TCG	ACC	TCA
Chicken	CTT	---	CAA	ACA	CAA	GAA	GTA	AGA	GAT	GTT	CAA	GAA	AAG	AAA	GCG	GCG	AAA	AGG	AGC	TGG	GAA	ACT	CAG	TAT	GTA	GCC	---	---	TCA
Cuckoo	CTT	---	CAA	ACA	CAA	GAA	GTA	GGA	GAT	GCT	CAA	GAA	AAT	AAC	GAA	GGA	AAA	ATG	AGC	TGG	GAA	ACT	CAG	TGC	ATA	GTC	---	---	TCA
Falcon	CTT	---	CAA	AGA	CAA	GAC	CTA	AGA	GAC	ATC	CAA	GAA	AAT	AAA	AGA	GGA	AAA	AGG	AGC	TGG	GAA	ACT	CAA	TGT	ATA	GCC	---	---	TCA
Pigeon	CTT	---	CAA	ACC	CAA	GAC	ATA	AGA	---	GAT	GTT	CAA	GAA	AAT	AAT	GGA	AGA	AGG	CAT	TGG	GAA	ACT	CAG	TGT	ATA	GCC	---	---	TCA

A MORE COMPLICATED REGION

			1744						1762						1780						1798					1816		
Hummingbir	CTT	---	CAA	AGA	CAA	GAT	GGA	AGA	GAT	GCT	GAA	GAA	AAT	AAA	GAA	GGA	AAA	AGG	AGC	TGG	GAA	GCT	CAG	GGT	ATA	GCC	---	---	TCA
HoodedCrow	CTT	---	CAA	ACA	CAG	GTT	GTA	AGA	---	GCC	AAA	CAA	AGG	AAA	GGA	GGA	AAG	GGG	AGC	TGG	GAA	GCT	CAG	TCT	ATG	GCC	---	---	TCA
Flycatcher	CTT	---	CAA	ACA	CAA	GAT	GTA	AGa	---	gct	aaa	gaa	aag	aaa	gga	gga	aaa	ggg	agc	tGG	GAA	ACT	CAG	TGT	TCG	GCC	---	---	TCA
GroundTit	CTT	---	CAA	ACA	CAA	GAT	GTA	AGA	---	GCT	AAA	GAA	AAA	AAA	GGA	GGA	AAA	GGG	AGC	TGG	GAA	ACT	CAG	TGT	ACC	GCC	---	---	TCA
Parus	CTT	---	CAA	ACA	CAA	GAT	GTa	aga	---	gct	aaa	gaa	aaa	aaa	gga	gga	aaa	ggg	agc	tgg	gaa	aCT	CAG	TGT	ACG	GCC	---	---	TCA
Indigobird	GAT	GGA	AGA	GCT	AAA	GAT	GTA	AGA	---	GCT	AAA	GAA	AAG	AAA	GGG	GGG	AAA	GCG	AGC	TGG	GAC	ACT	CGG	TGT	ACA	GCC	---	---	TCA
ZebraFinch	CTT	---	CAA	ACA	CAA	GAT	GTA	AGa	---	gct	aaa	gaa	aag	aaa	ggg	ggg	AAA	GGG	AGC	TGG	GAC	ACT	CGG	TGT	ATG	GCC	---	---	TCA
Cowbird	CTT	---	CAA	ACA	CAA	GAT	GTA	AGA	---	GCT	AAA	GAA	AAG	AAA	GGG	GGG	AAG	GGG	AGC	TGG	GAC	ACT	CGG	TGT	GCG	GCC	TCG	ACC	TCA
Redwing	CTT	---	CAA	ACA	CAA	GAT	GTA	AGA	---	GCT	AAA	GAA	AAG	AAA	GGG	GGG	AAG	GGG	AGC	TGG	GAC	ACT	CGG	TGT	GCG	GCC	TCG	ACC	TCA
Canary	CTT	---	CAA	ACA	CAA	GAT	GTA	AGa	---	gct	aaa	gaa	aag	aaa	ggg	ggg	aaa	ggg	agc	TGG	GAC	ACT	cgc	tgt	gca	gcc	cca	gcc	tca
Sparrow	CTT	---	CAA	ACA	CAA	GAT	GTA	CGa	---	gct	aaa	gaa	aag	aaa	ggg	ggg	aaa	ggg	agc	TGG	GAC	ACT	CGG	TGT	GCG	GCC	TCG	ACC	TCA
Chicken	CTT	---	CAA	ACA	CAA	GAA	GTA	AGA	GAT	GTT	CAA	GAA	AAG	AAA	GCG	GCG	AAA	AGG	AGC	TGG	GAA	ACT	CAG	TAT	GTA	GCC	---	---	TCA
Cuckoo	CTT	---	CAA	ACA	CAA	GAA	GTA	GGA	GAT	GCT	CAA	GAA	AAT	AAC	GAA	GGA	AAA	ATG	AGC	TGG	GAA	ACT	CAG	TGC	ATA	GTC	---	---	TCA
Falcon	CTT	---	CAA	AGA	CAA	GAC	CTA	AGA	GAC	ATC	CAA	GAA	AAT	AAA	AGA	GGA	AAA	AGG	AGC	TGG	GAA	ACT	CAA	TGT	ATA	GCC	---	---	TCA
Pigeon	CTT	---	CAA	ACC	CAA	GAC	ATA	AGA	---	GAT	GTT	CAA	GAA	AAT	AAT	GGA	AGA	AGG	CAT	TGG	GAA	ACT	CAG	TGT	ATA	GCC	---	---	TCA

QC

- Don't want multiple nonsynonymous changes in succession, particularly within interest lineages
- Avoid indels where suspicious
- Clean 5' and 3' end to nearest codon
- May need to create rules about cutting poorly aligned individuals or partial sequences from alignment (e.g. Beichman et al 2019 used min 8/13 individuals) – want to avoid interest group removal if possible!
- May need to create rules about cutting whole genes (minimum trimmed alignment length)

FURTHER READING & REFERENCES

- Pais et al (2014) Assessing the efficiency of multiple sequence alignment programs. *Algorithms for molecular biology* 9: 4
- Oregon State Applied Bioinformatics. Chapter 4 Multiple Sequence Alignments, Molecular Evolution, and Phylogenetics. https://open.oregonstate.education/appliedbioinformatics/chapter/chapter-4/

TO MEGA!