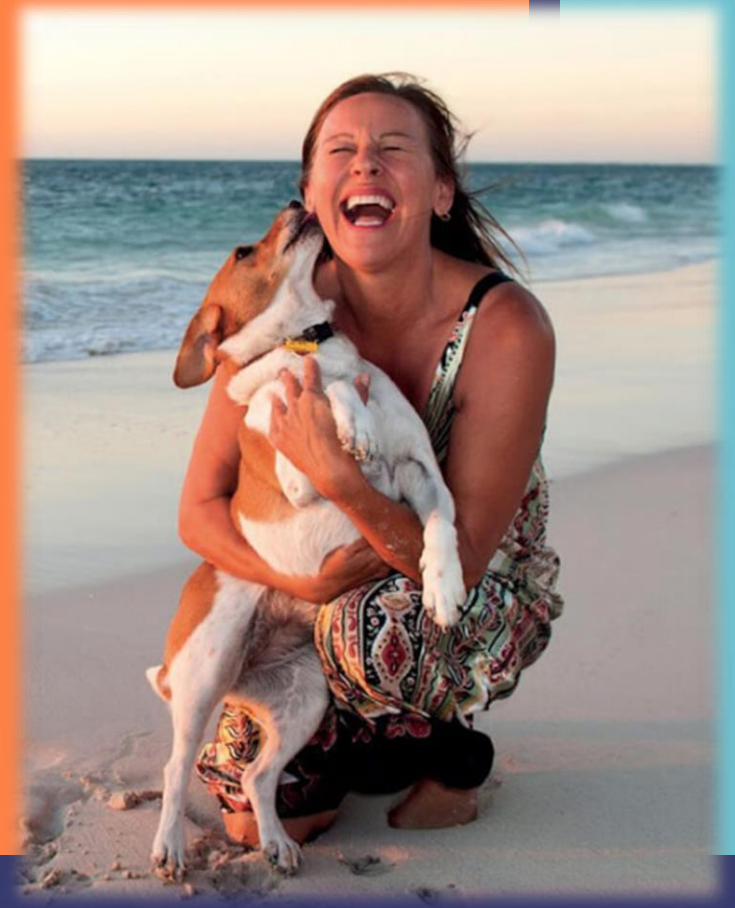


Comparison between cross-species gene and Human gene with William Syndrome for potential similarity

Presented By

Rocco DiGiorgio V

& Nabila Shahnaz Khan

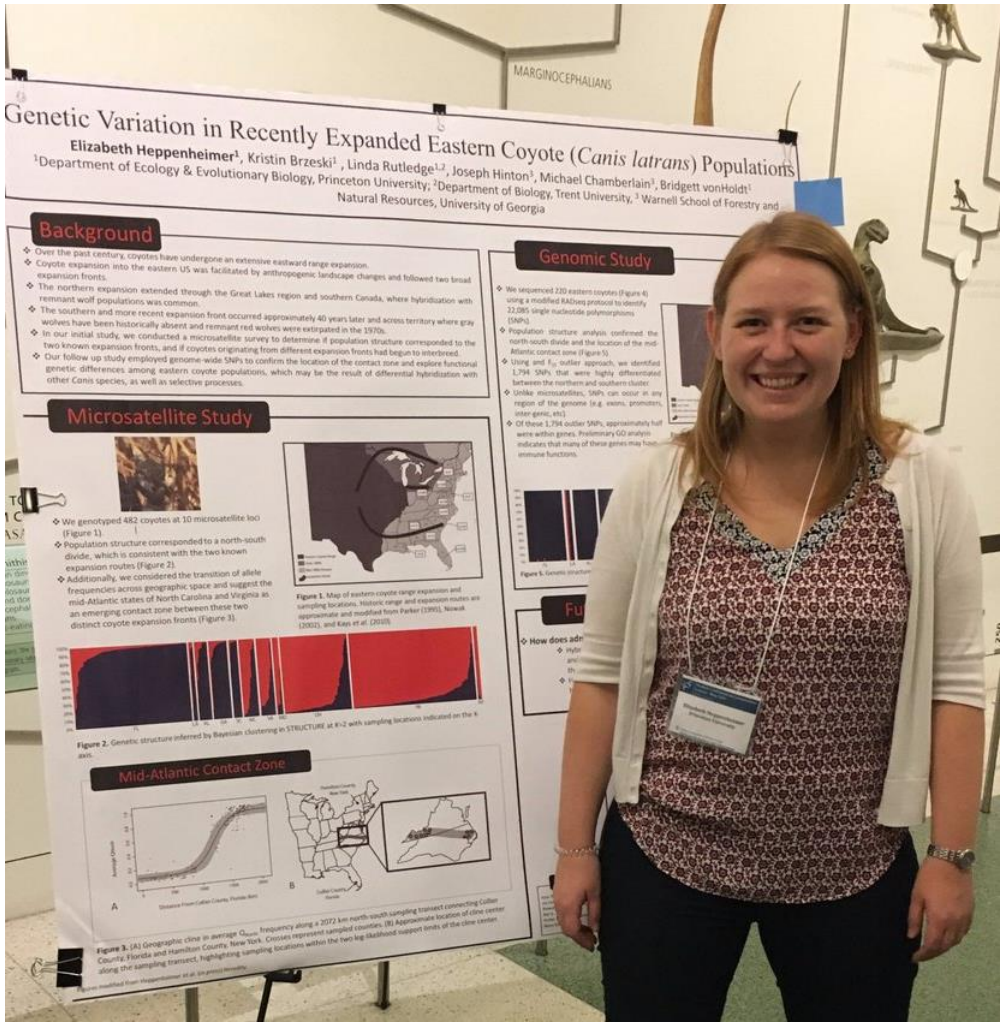
Williams-Beuren Syndrome

- Williams syndrome occurs when people are **missing some genes (a chunk of DNA) on chromosome 7** containing about 27 genes
- It results in traits such as: bubbly, extroverted personalities, heart defects, intellectual disability etc.
- The first hint of a link between dogs and Williams syndrome came in **2010**, when biologist VonHoldt et al. found **WBSCR17** and other genes near it were important in dog evolution and domestication.
- This region of the genome is similar in dogs and humans, and the human version of WBSCR17 is located near the sequence that is deleted in people with Williams syndrome.

Previous Findings

- **Researcher VonHoldt** previously had found that chromosome 6 on dogs (similar to regions of chromosome 7 in Humans) have important role in evolution.

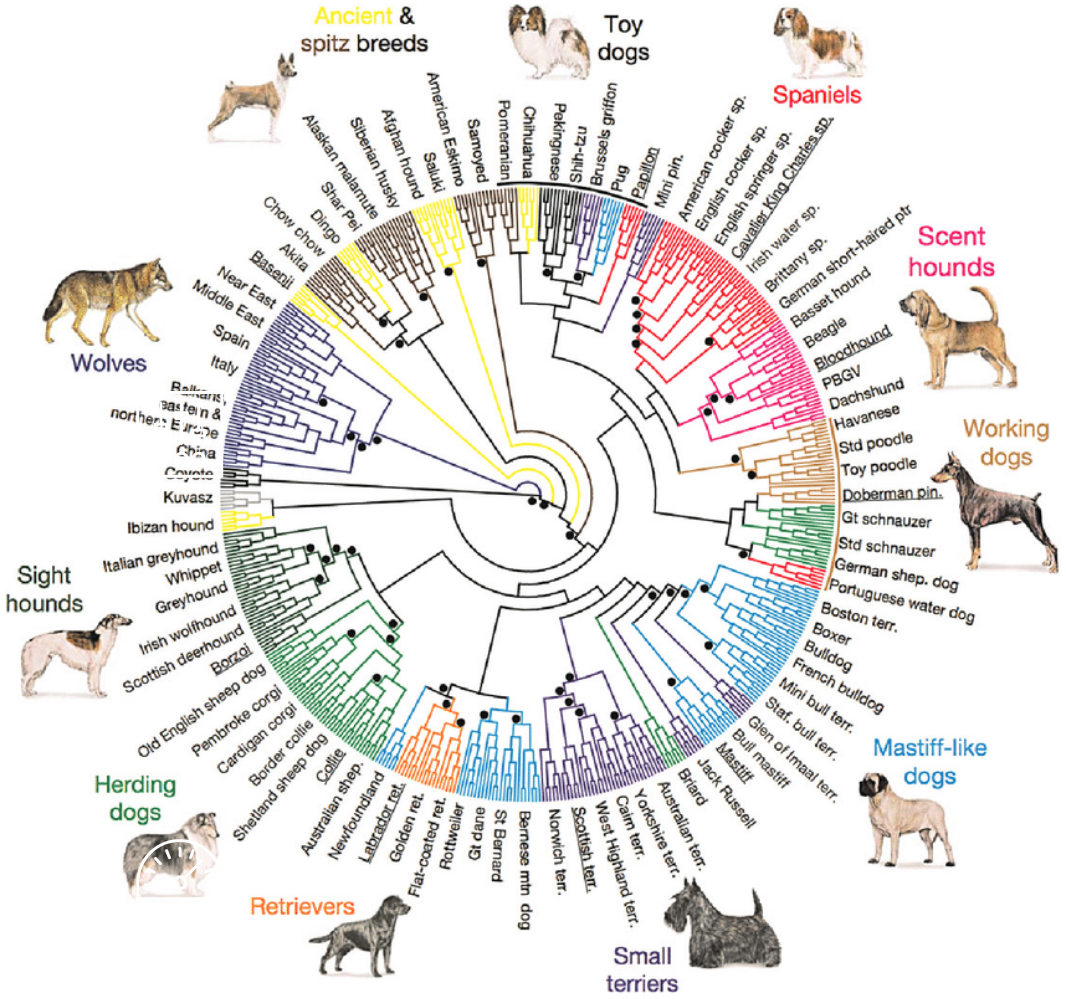
- A relative lack of changes in that gene seems to lead to aloof, wolf like behavior in dogs and wolves.



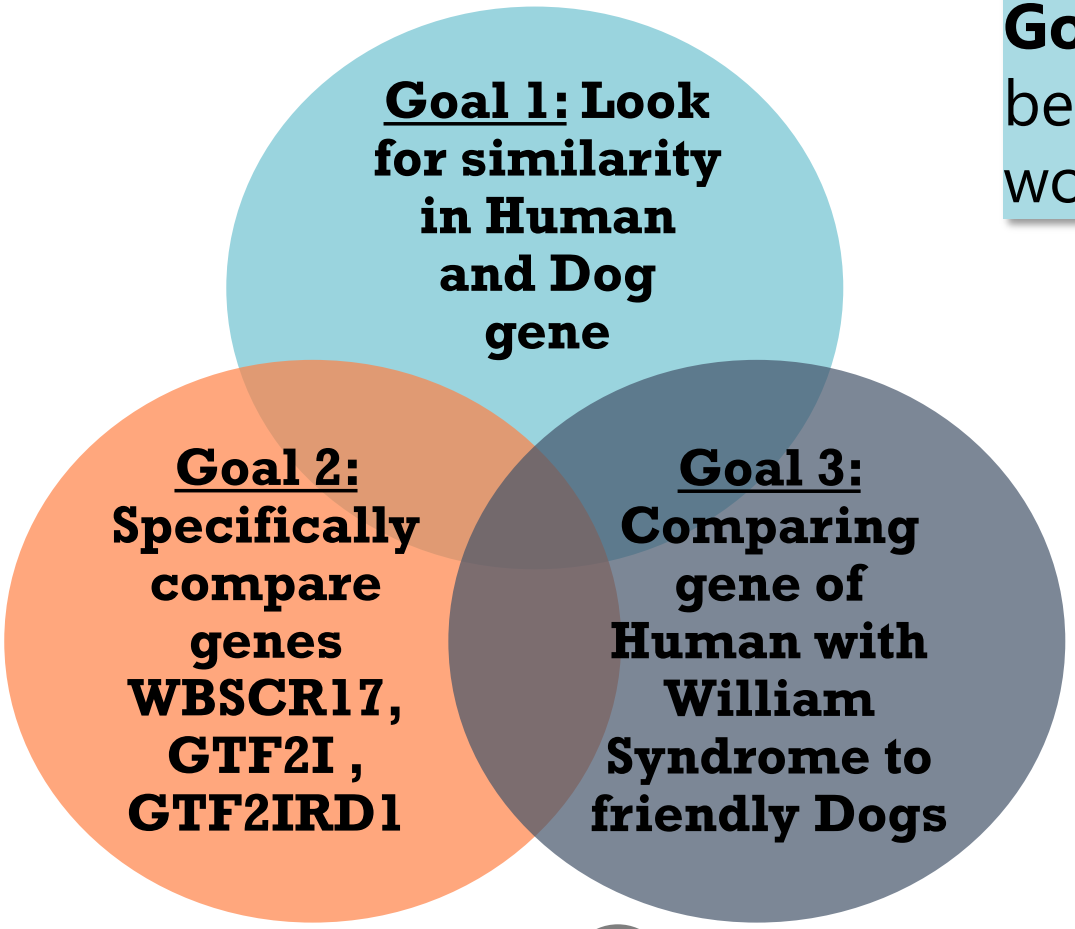
Biologist Bridgett VonHoldt

Previous Findings

- Researchers found that threes genes (WBSCR17, GTF2I , GTF2IRD1) were correlated with the behavioral traits of dogs as compared to wolves
- This region of genome had similarity with human genome which were located near the sequence that exist in the DNA of people with Williams syndrome
- This information hints towards the possibility that there might be some link between the gene sequence of Human with William Syndromes to dogs



Our Goal



Goal 1: Figuring out if there is more similarity between the genes of Human to dogs compared to wolves as dogs show more social traits

Goal 2: According to the study, these genes play most important role in the evolution of dogs from wolves

Goal 3: Comparing chromosome 7 of Human with and without William Syndrome to chromosome 6 of dogs to see if dog's genes have more similarity to people with William Syndrome

Approach Followed

Appraoch 1
Cross-species gene comparison

- Compare Human Gene to Dog's gene (WBSCR17, GTF2I , GTF2IRD1)
- Compare Human Gene to Wolves' gene (WBSCR17, GTF2I , GTF2IRD1)
- Analyze the results to see if they have any significance

Approach 2
similarity testing of Human with William syndrome and Dogs

- Compare Human Chromosome 7 to Dog Chromosome 6
- Compare Chromosome 7 of Human with William syndrome to Dog Chromosome 6
- Analyze the results to see if there's any significant difference

Challenges Faced and Solutions

Challenge 1
Finding the dataset of specific genes for dogs and wolves

Emailed author VonHoldt

Challenge 2
Finding Data of Human with William Syndrome

Constructed artificial data using information from NCBI and NIH

Challenge 3
Huge Data size (on average more than 400K!!!)

Dividing assembled sequences to reads

Challenge 4
Huge time and space complexity

1. Implementing Linear-space version of algorithms

2. Parallel Processing

Data Collection

Dogs and Wolves

1. Data from VonHoldt's study was obtained after a short email exchange

Typical Humans

1. The human reference chromosome 7 was obtained from NCBI

William's Syndrome Humans

1. Multiple authors were emailed, however, none responded
2. We found that William's Syndrome is caused by deletion of an arbitrary subset of adjacent genes on 7q11.23
3. We decided to create artificial data

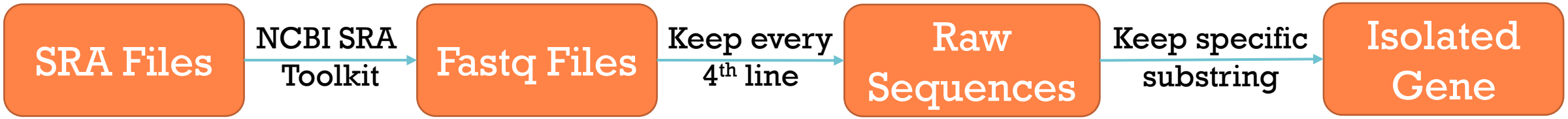
NCBI

NIH

PubMed.gov

Data Processing

WBSCR17

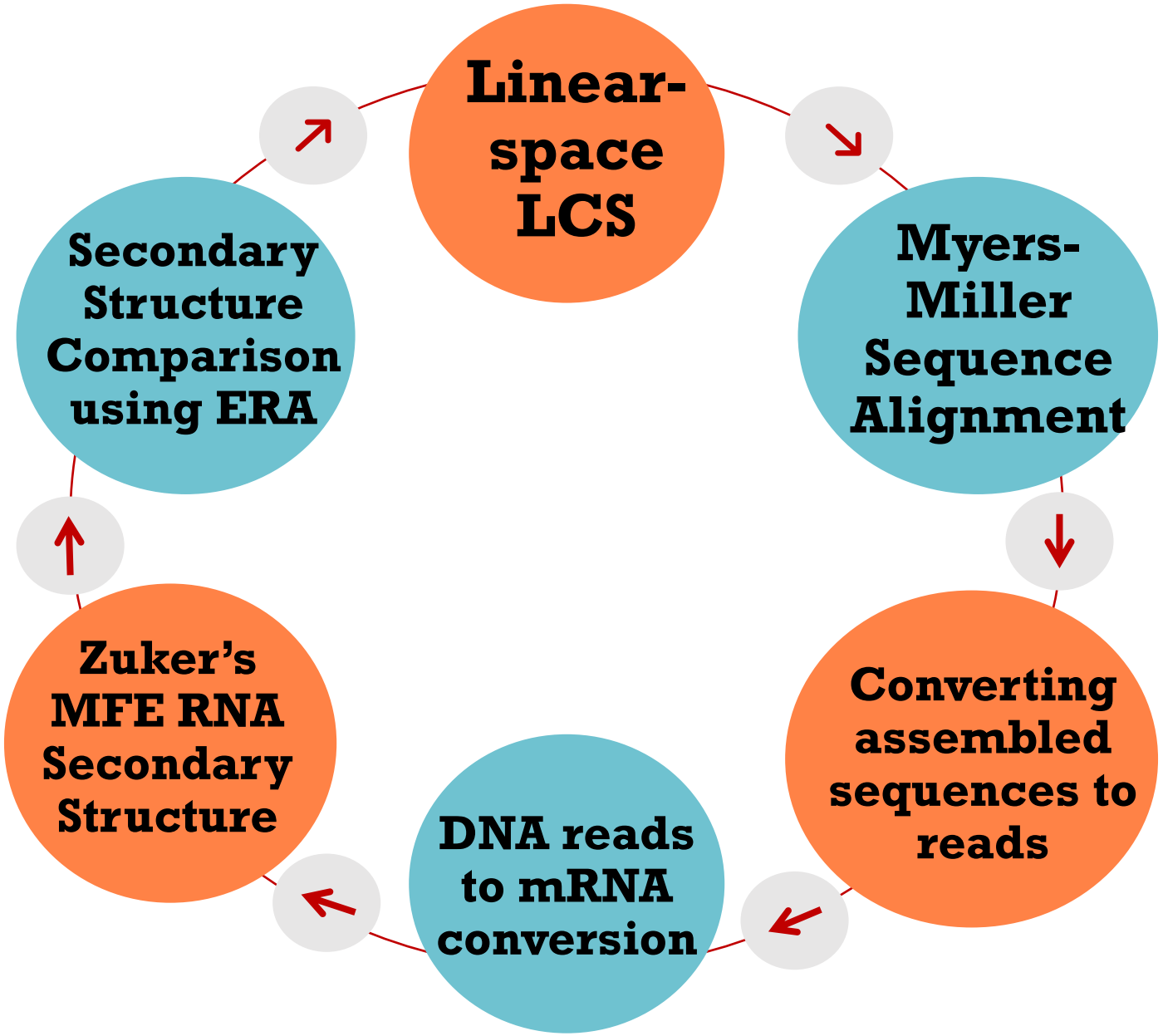


Whole 7q11.23

Using the human reference chromosome 7:

1. 7q11.23 was isolated
2. A subset of the genes associated with William's syndrome were deleted
3. Multiple artificial William's Syndrome 7q11.23 sequences generated
4. Assembled input sequences were sliced into reads of 400kb

Computational Methods Used



Linear-Space LCS Algorithm

- Normally LCS requires **two (n×n) arrays**, one for calculating the length and another for backtracking
- But sequences of length near about **400,000**! 8GB/16GB/32GB RAM can't handle this
- LCS can be implemented easily using **one (2×n) array** but that can **only calculate length**
- To get both length and longest common sequence, we implemented a code that **calculates length using one (2×n) array** and **for backtracking**, it **stores** the (n×n) matrix **in** one **file** and later stores it in reverse order in another file.
- This code can be run in a computer with 16GB RAM

Compared to Human WBSCR17

LCS for Dog's WBSCR17		LCS for Wolfs' WBSCR17	
Dog1	316974	Wolf1	316996
Dog2	317326	Wolf2	314335
Dog3	316634	Wolf3	314646
Dog4	318036	Wolf4	316799
Dog5	317990	Wolf5	316548

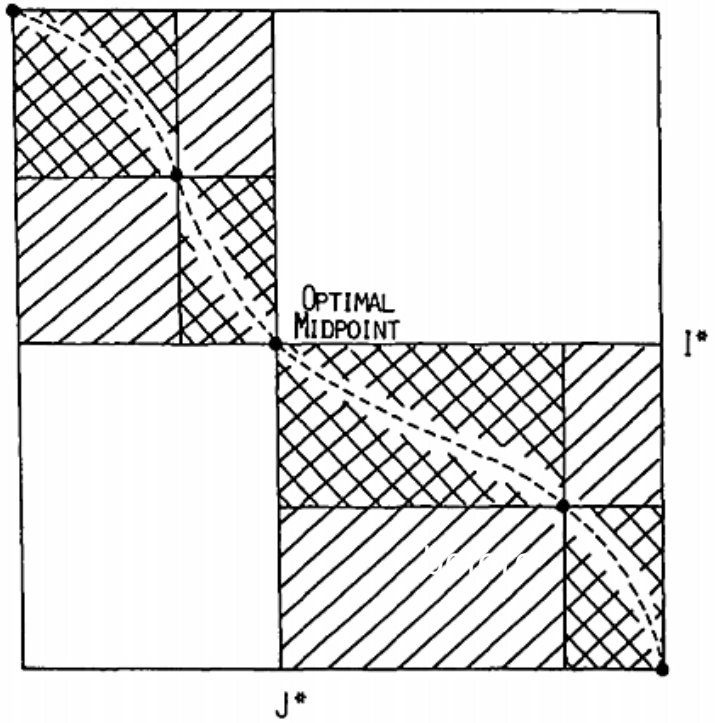
Myers-Miller Sequence Alignment Algorithm

- **Gotoh's algorithm** (1982) for sequence alignment (minimizing the cost) required **O(MN) space**
- Implemented in **O(N) space** if **only cost required**
- **Myers et al**. (1988) developed a **linear-space version of Gotoh's algorithm** based on Hirschberg's idea
- Our code outputs minimum **cost**, **aligned sequence**, number of **insertions**, **deletions**, **mutations**, **matches** and based on that calculated **alignment identity**

Uses Divide-and-Conquer Along with Dynamic Programming

```
A C C G G T A T C C T A G G A C
! ! ! - - ! ! ! ! | ! ! ! ! ! !
A C C     T A T C T T A G G A C


Insert: 0
Delete: 2
Match: 13
Replace: 1
Minimum cost 4
Percent Sequence identity 81.25
```



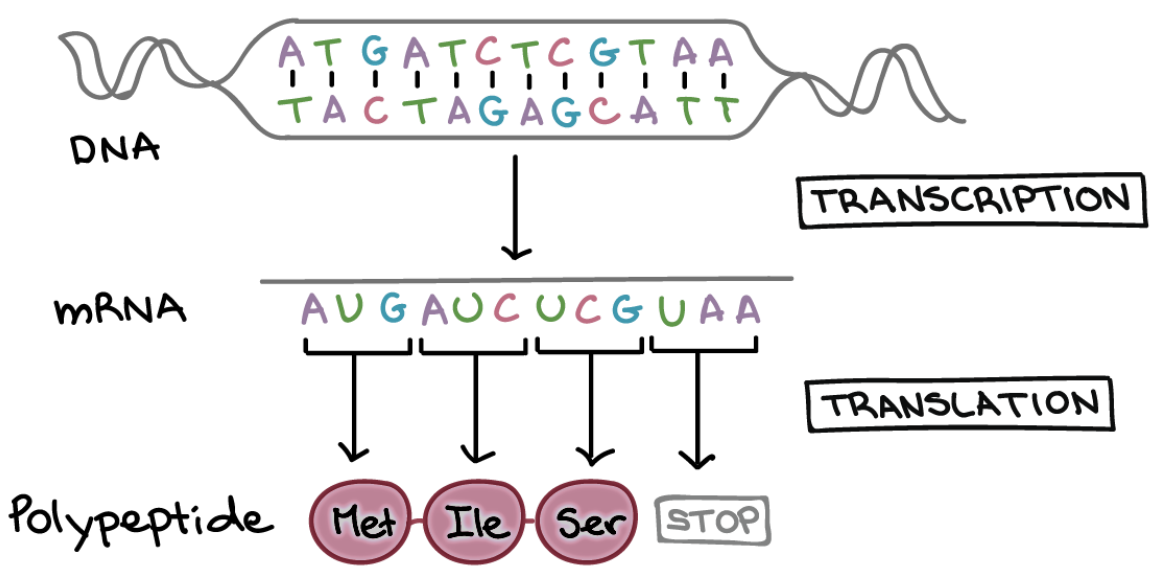
$$\text{BLAST Identity Score} = \frac{\#Matches}{\#Matches + \#Replacements + \#Insertions + \#Deletions}$$

Converting Assembled Sequences to Reads

- Each gene were sliced into **overlapping reads of 40kb** as the whole gene doesn't get converted to mRNA
- It was done using a custom **python script**

DNA reads to mRNA conversion

- mRNA carries the instructions of portions of DNA to cytoplasm for protein synthesis
- So DNA reads were converted to mRNA (**transcription**), Later secondary structure of mRNA will be predicted



Zuker's RNA Secondary Structure Prediction Algorithm

- **Zuker's algorithm** predicts the most stable secondary structure for a RNA sequence by computing its **minimal free energy** (MFE).
- Our code takes input in **FASTA** format
- It give's output in a **dotted-bracket** format
- We **merged Vienna-RNA package tool** with our code to **visualize the secondary structure** (on Linux operating system)

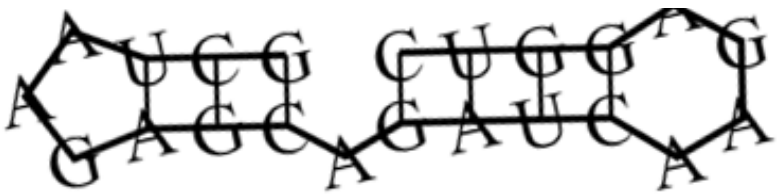
$$W(i) = \min\{W(i-1), \min_{0 \le k < i}\{W(k) + V(k+1, i)\}\}.$$

$$V(i,j) = \min\{eH(i,j), eS(i,j,i+1,j-1) + V(i+1,j-1), \min_{i<i'<j'<j \text{ and } i'-i+j-j'>2}\{eL(i,j,i',j') + V(i',j')\}, \min_{i+1<k<j}\{WM(i+1,k-1) + WM(k,j-1) + a\}\},$$

$$WM(i,j) = \min\{V(i,j) + b, WM(i,j-1) + c, WM(i+1,j) + c, \min_{i<k\le j}\{WM(i,k-1) + WM(k,j)\}\},$$

GCUAAGAGCAGAUCAAGAGGUC

(((...)))(((....)))



Secondary Structure Comparison using ERA

- For comparing the secondary structures, we plan to use **Efficient alignment of RNA** (ERA) tool
- It was implemented by **Zhang et al.** using Sparse Dynamic Programming
- It takes input dotted bracket format
- Outputs an alignment score
- We have download and installed the tool

```
>AB013372
GCGCCCGUAGCUCAAUUGGAUAGAGCGUUUGACUACGGAUCAAAAGGUUAGGGGUUCGACUCCUCUCGGGCGCG
(((((((..((((.........)))).((((((....).)))))).....((((.....))))).)))))))).
>AB013373
GCGGAAGUAGUUCAGUGGUAGAACACCACCUUGCCAAGGUGGGGGUCGCGGGUUCGAAUCCCGUCUUCCGCU
(((((((..((((.......)))).((((((.....))))))).....((((.......)))))))))))).
```

Input

```
base deletion (w_d) = 2
base mismatch (w_m) = 1
arc  removing (w_r) = 2
arc  breaking (w_b) = 1.5
arc  mismatch (w_am)= 1.8
```

Output

Further Work To be Done

Finish compiling the algorithms on our input dataset

A

Analyzing the output result (both for approach1 and approach2)

B

Generate structure variant table (if still have time)

C

D

THANK YOU