

# DNA Sequencing and Data Analysis

Final Project

Tomer, Elad, Shir, Omer, Rotem, Gil

# The Team



Gil



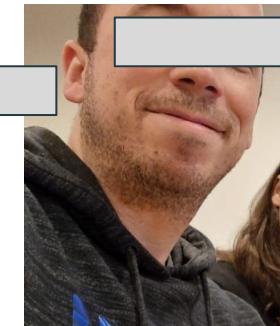
Rotem



Omer



Tomer



Elad

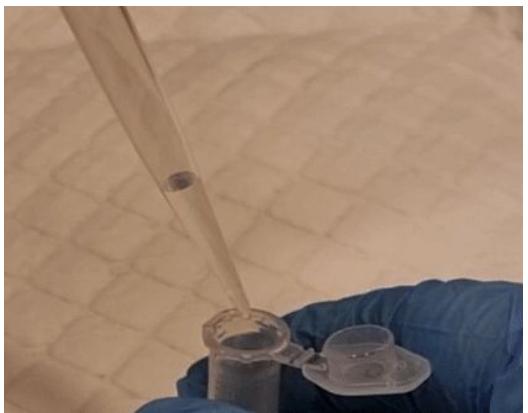


Shir

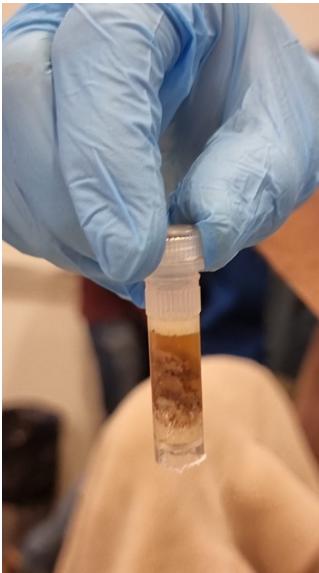
# Agenda

- From poo to DNA
- From DNA to Reads
- From Reads to a DNA Sequences
- From Sequences to -
  - Animal detection
  - Gene Analysis
  - Microbiome analysis

# Extracting DNA - O



# Extracting DNA - O

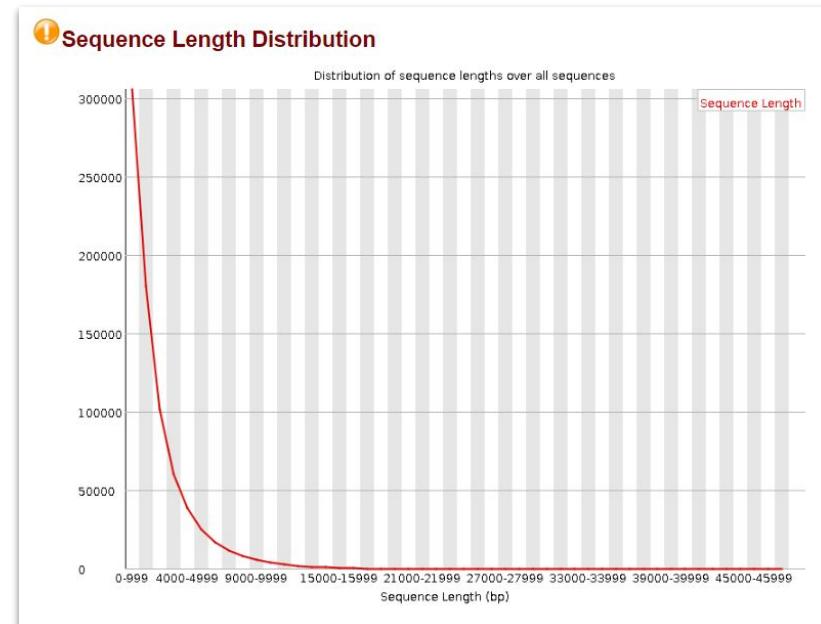


# Sequencing - A



# Analyzing the Reads File - M2

- 773,977 reads
- Length - 50-47,639

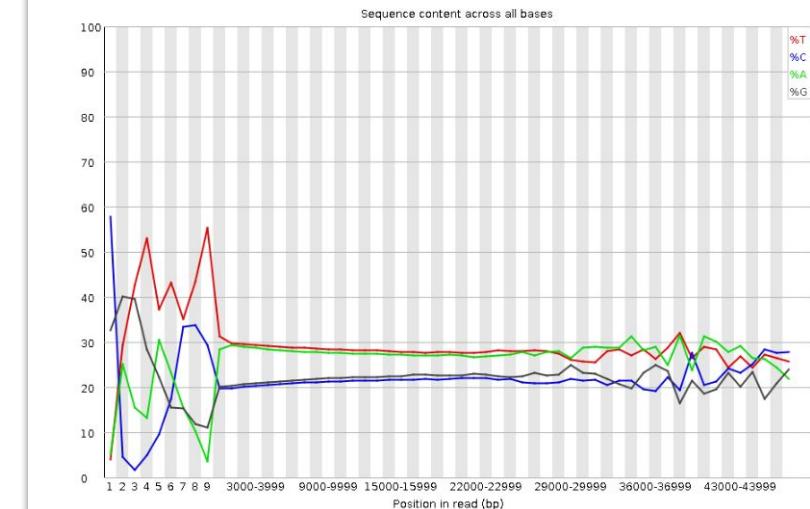


# Analyzing the Reads File

## ✖ Per base sequence quality



## ✖ Per base sequence content



# Processing

- Trimming and filtering bad reads - NanoFilt
- Alignment - Minimap2
- We aligned the FASTA vs all reference genomes

```
fastas=("GCF_015227675.2_mRatBN7.2_genomic.fna" \  
"GCF_000001635.27_GRCh39_genomic.fna" \  
"GCA_019924945.1_FelChav1.0_genomic.fa" \  
"GCF_003160815.1_VulVul2.2_genomic.fna" \  
"GCF_014441545.1_ROS_Cfam_1.0_genomic.fna")
```

```
function align() {  
    for fasta in ${fastas[@]}; do
```

```
        echo "---- Finished ---"  
        done  
    }
```

```
function count_alignments() {
```

```
    done  
}
```

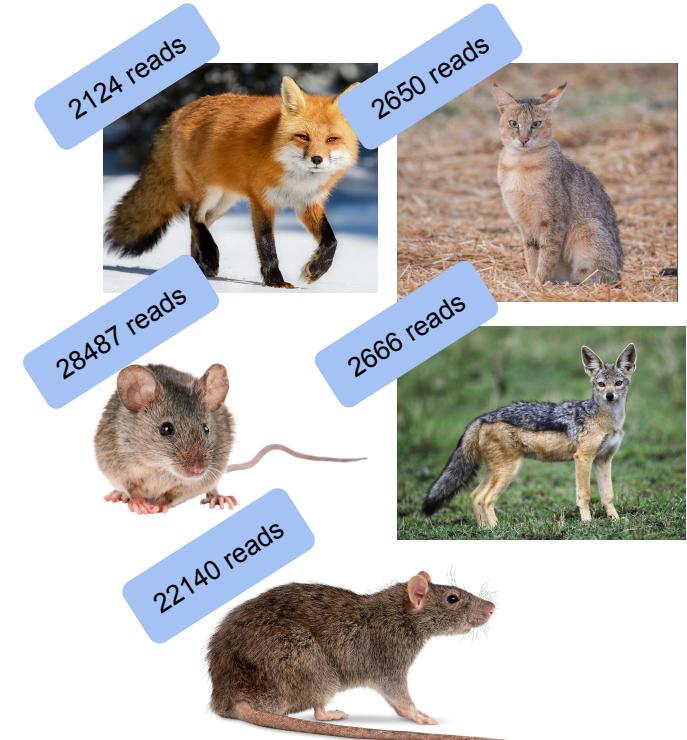
# Determining the Animal - Strategy

- We compared the number of aligned reads in each SAM file
- The maximal number of aligned reads - this will be the animal

# First Try

- We filtered out reads with less than 500 bases
- Got very similar numbers between rat and mouse

```
gunzip -c ./M2.raw.fastq.gz | \  
  --logfile $fasta.nanofilt.log| \  
  \
```



# First Try



Rat (mRatBN7.2)			Mouse (GRCm39)		
Count	Sequence	Chromosome	Count	Sequence	Chromosome
2134	NC_051336.1	chr1	1160	NC_000067.7	chr1
1881	NC_051337.1	chr2	2789	NC_000068.8	chr2
1245	NC_051338.1	chr3	1129	NC_000069.7	chr3
1633	NC_051339.1	chr4	1006	NC_000070.7	chr4
1270	NC_051340.1	chr5	1165	NC_000071.7	chr5
1377	NC_051341.1	chr6	1245	NC_000072.7	chr6
2204	NC_051342.1	chr7	986	NC_000073.7	chr7
979	NC_051343.1	chr8	802	NC_000074.7	chr8
1006	NC_051344.1	chr9	5997	NC_000075.7	chr9
666	NC_051345.1	chr10	817	NC_000076.7	chr10
662	NC_051346.1	chr11	637	NC_000077.7	chr11
301	NC_051347.1	chr12	930	NC_000078.7	chr12
749	NC_051348.1	chr13	776	NC_000079.7	chr13
852	NC_051349.1	chr14	914	NC_000080.7	chr14
812	NC_051350.1	chr15	563	NC_000081.7	chr15
610	NC_051351.1	chr16	585	NC_000082.7	chr16
861	NC_051352.1	chr17	616	NC_000083.7	chr17
682	NC_051353.1	chr18	508	NC_000084.7	chr18
359	NC_051354.1	chr19	335	NC_000085.7	chr19
416	NC_051355.1	chr20			
1321	NC_051356.1	X	1565	NC_000086.8	X
30	NC_051357.1	Y	367	NC_000087.8	Y



# First Try

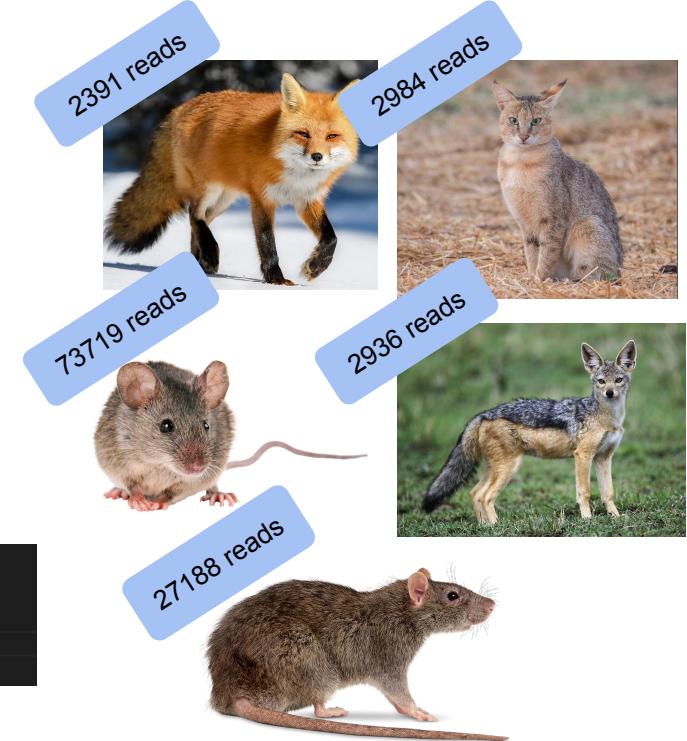


Rat (mRatBN7.2)			Mouse (GRCm39)		
Count	Sequence	Chromosome	Count	Sequence	Chromosome
2134	NC_051336.1	chr1	1160	NC_000067.7	chr1
1881	NC_051337.1	chr2	2789	NC_000068.8	chr2
1245	NC_051338.1	chr3	1129	NC_000069.7	chr3
1633	NC_051339.1	chr4	1006	NC_000070.7	chr4
1270	NC_051340.1	chr5	1165	NC_000071.7	chr5
1377	NC_051341.1	chr6	1245	NC_000072.7	chr6
2204	NC_051342.1	chr7	986	NC_000073.7	chr7
979	NC_051343.1	chr8	802	NC_000074.7	chr8
1006	NC_051344.1	chr9	5997	NC_000075.7	chr9
666	NC_051345.1	chr10	817	NC_000076.7	chr10
662	NC_051346.1	chr11	637	NC_000077.7	chr11
301	NC_051347.1	chr12	930	NC_000078.7	chr12
749	NC_051348.1	chr13	776	NC_000079.7	chr13
852	NC_051349.1	chr14	914	NC_000080.7	chr14
812	NC_051350.1	chr15	563	NC_000081.7	chr15
610	NC_051351.1	chr16	585	NC_000082.7	chr16
861	NC_051352.1	chr17	616	NC_000083.7	chr17
682	NC_051353.1	chr18	508	NC_000084.7	chr18
359	NC_051354.1	chr19	335	NC_000085.7	chr19
416	NC_051355.1	chr20			
1321	NC_051356.1	X	1565	NC_000086.8	X
30	NC_051357.1	Y	367	NC_000087.8	Y



# Second Try

- We decided to align again with different parameters
- This time we didn't filter short reads
- Now the mouse was significantly better



```
gunzip -c ./M2.raw.fastq.gz | \
```

# Second Try



Rat (mRatBN7.2)			Mouse (GRCm39)		
Count	Sequence	Chromosome	Count	Sequence	Chromosome
2697	NC_051336.1	chr1	2742	NC_000067.7	chr1
2353	NC_051337.1	chr2	8237	NC_000068.8	chr2
1585	NC_051338.1	chr3	2581	NC_000069.7	chr3
1985	NC_051339.1	chr4	2456	NC_000070.7	chr4
1546	NC_051340.1	chr5	2529	NC_000071.7	chr5
1642	NC_051341.1	chr6	2758	NC_000072.7	chr6
2472	NC_051342.1	chr7	2405	NC_000073.7	chr7
1228	NC_051343.1	chr8	1859	NC_000074.7	chr8
1236	NC_051344.1	chr9	17155	NC_000075.7	chr9
873	NC_051345.1	chr10	1904	NC_000076.7	chr10
836	NC_051346.1	chr11	1533	NC_000077.7	chr11
341	NC_051347.1	chr12	2206	NC_000078.7	chr12
926	NC_051348.1	chr13	1728	NC_000079.7	chr13
1067	NC_051349.1	chr14	2811	NC_000080.7	chr14
1006	NC_051350.1	chr15	1385	NC_000081.7	chr15
758	NC_051351.1	chr16	1362	NC_000082.7	chr16
1037	NC_051352.1	chr17	1357	NC_000083.7	chr17
859	NC_051353.1	chr18	1221	NC_000084.7	chr18
446	NC_051354.1	chr19	809	NC_000085.7	chr19
513	NC_051355.1	chr20			
1640	NC_051356.1	X	3815	NC_000086.8	X
37	NC_051357.1	Y	603	NC_000087.8	Y
14	NC_001665.2	?	34	NC_005089.1	?
<b>27097</b>			<b>63490</b>		



# A Male Mouse



# Let's just verify with the MAPQ values...



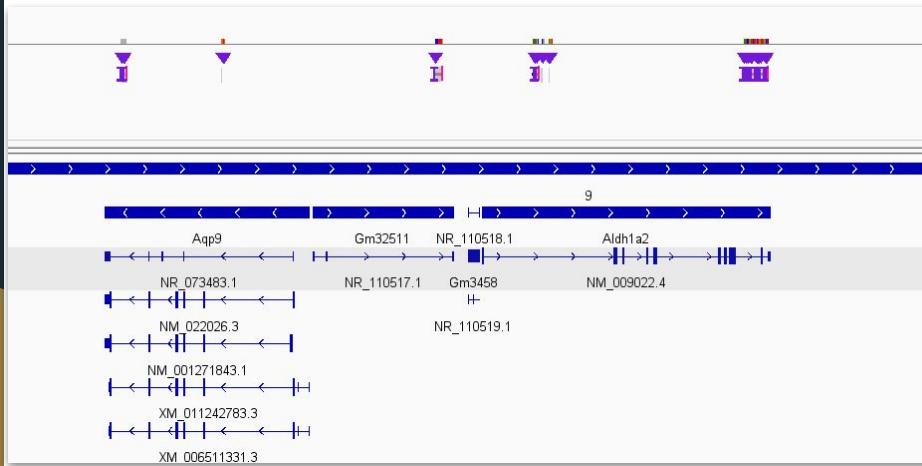
Rat (mRatBN7.2)				Mouse (GRCm39)			
Average MAPQ	Count	Sequence	Chromosome	Average MAPQ	Count	Sequence	Chromosome
15.20	2697	NC_051336.1	chr1	50.06	2742	NC_000067.7	chr1
16.33	2353	NC_051337.1	chr2	22.45	8237	NC_000068.8	chr2
18.78	1585	NC_051338.1	chr3	45.91	2581	NC_000069.7	chr3
15.51	1985	NC_051339.1	chr4	40.82	2456	NC_000070.7	chr4
16.34	1546	NC_051340.1	chr5	40.72	2529	NC_000071.7	chr5
14.42	1642	NC_051341.1	chr6	40.53	2758	NC_000072.7	chr6
8.36	2472	NC_051342.1	chr7	36.54	2405	NC_000073.7	chr7
17.49	1228	NC_051343.1	chr8	45.45	1859	NC_000074.7	chr8
15.58	1236	NC_051344.1	chr9	5.51	17155	NC_000075.7	chr9
19.83	873	NC_051345.1	chr10	50.79	1904	NC_000076.7	chr10
18.14	836	NC_051346.1	chr11	51.36	1533	NC_000077.7	chr11
16.54	341	NC_051347.1	chr12	36.34	2206	NC_000078.7	chr12
16.37	926	NC_051348.1	chr13	46.51	1728	NC_000079.7	chr13
17.29	1067	NC_051349.1	chr14	29.64	2811	NC_000080.7	chr14
16.06	1006	NC_051350.1	chr15	50.15	1385	NC_000081.7	chr15
18.90	758	NC_051351.1	chr16	51.96	1362	NC_000082.7	chr16
13.31	1037	NC_051352.1	chr17	46.24	1357	NC_000083.7	chr17
16.26	859	NC_051353.1	chr18	50.93	1221	NC_000084.7	chr18
18.60	446	NC_051354.1	chr19	48.96	809	NC_000085.7	chr19
14.93	513	NC_051355.1	chr20				
15.70	1640	NC_051356.1	X	29.82	3815	NC_000086.8	X
0.35	37	NC_051357.1	Y	1.36	603	NC_000087.8	Y
40.67	14	NC_001665.2	?	39.06	34	NC_005089.1	?
	<b>27097</b>					<b>63490</b>	



# A Female Mouse!



# Exploring the results on IGV



# Genes Analysis

- We found interesting alignments in the IGV and looked for their source in the DB
- We used the famous website for Mouse Genome



# Genes Analysis



# Genes Analysis - ADAMTS7

Coronary Artery Disease -  
מחלה לב כילילית

**ADAMTS7**

From Wikipedia, the free encyclopedia

A disintegrin and metalloproteinase with thrombospondin motifs 7 (ADAMTS7) is an enzyme that in humans is encoded by the *ADAMTS7* gene on chromosome 15.<sup>[1]</sup> It is ubiquitously expressed in many tissues and cell types.<sup>[2]</sup> This enzyme catalyzes the degradation of cartilage oligomeric matrix protein (COMP) degradation.<sup>[3]</sup> ADAMTS7 has been associated with cancer and arthritis in multiple tissue types.<sup>[4][5]</sup> The *ADAMTS7* gene also contains one of 27 SNPs associated with increased risk of coronary artery disease.<sup>[6]</sup>

**Structure** [edit]

**Gene** [edit]

The *ADAMTS7* gene resides on chromosome 15 at the band 15q24.2 and contains 25 exons.<sup>[1]</sup>

Enzymes are proteins that act as biological catalysts by accelerating chemical reactions

דלקת מפרקים = Arthritis

# Genes Analysis - ADAMTS7

They turned off the gene in mice (probably using CRISPR)

## Knockout of Adamts7, a novel coronary artery disease locus in humans, reduces atherosclerosis in mice

Robert C Bauer <sup>1 2</sup>, Junichiro Tohyama <sup>1 2</sup>, Jian Cui <sup>1 3</sup>, Lan Cheng <sup>1 3</sup>, Jifu Yang <sup>1 3</sup>, Xuan Zhang <sup>1 3</sup>, Kristy Ou <sup>1 3</sup>, Georgios K Paschos <sup>1 4</sup>, X Long Zheng <sup>5</sup>, Michael S Parmacek <sup>1 3</sup>, Daniel J Rader <sup># 1 2 3 4</sup>, Muredach P Reilly <sup># 1 3</sup>

Affiliations + expand

PMID: 25712206 PMCID: PMC4382454 DOI: 10.1161/CIRCULATIONAHA.114.012669

[Free PMC article](#)

### Abstract

**Background:** Genome-wide association studies have established ADAMTS7 as a locus for coronary artery disease in humans. However, these studies fail to provide directionality for the association between ADAMTS7 and coronary artery disease. Previous reports have implicated ADAMTS7 in the regulation of vascular smooth muscle cell migration, but a role for and the direction of impact of this gene in atherogenesis have not been shown in relevant model systems.

**Methods and results:** We bred an Adamts7 whole-body knockout mouse onto both the Ldlr and Apoe knockout hyperlipidemic mouse models. Adamts7(-/-)/Ldlr(-/-) and Adamts7(-/-)/Apoe(-/-) mice displayed significant reductions in lesion formation in aortas and aortic roots compared with controls. Adamts7 knockout mice also showed reduced neointimal formation after femoral wire injury. Adamts7 expression was induced in response to injury and hyperlipidemia but was absent at later time points, and primary Adamts7 knockout vascular smooth muscle cells showed reduced migration in the setting of tumor necrosis factor- $\alpha$  stimulation. ADAMTS7 localized to cells positive for smooth muscle cell markers in human coronary artery disease lesions, and subcellular localization studies in cultured vascular smooth muscle cells placed ADAMTS7 at the cytoplasm and cell membrane, where it colocalized with markers of podosomes.

**Conclusions:** These data represent the first *in vivo* experimental validation of the association of Adamts7 with atherogenesis, likely through modulation of vascular cell migration and matrix in atherosclerotic lesions. These results demonstrate that Adamts7 is proatherogenic, lending directionality to the original genetic association and supporting the concept that pharmacological inhibition of ADAMTS7 should be atheroprotective in humans, making it an attractive target for novel therapeutic interventions.

And reduced atherosclerosis (ערקם)  
עורקים

# Microbiome Analysis

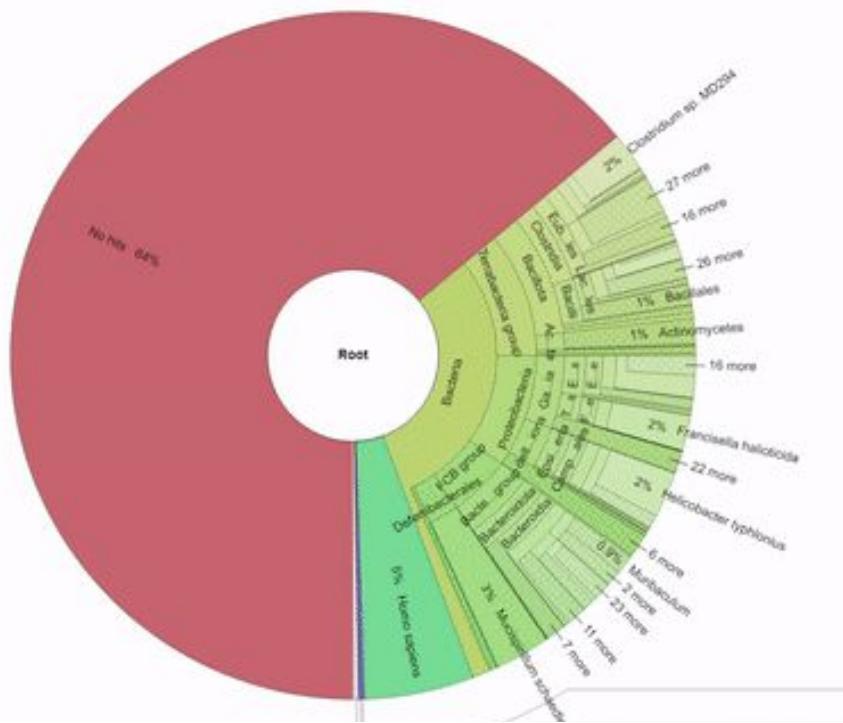
We mapped the unaligned reads in the sequenced DNA using Kraken2, and got a mapping of the other DNA sources in the sample.

```
# Download standard 16Gb DB
mkdir db-16 && cd db-16
kraken-build --db杜鹃花_20221209.tar.gz

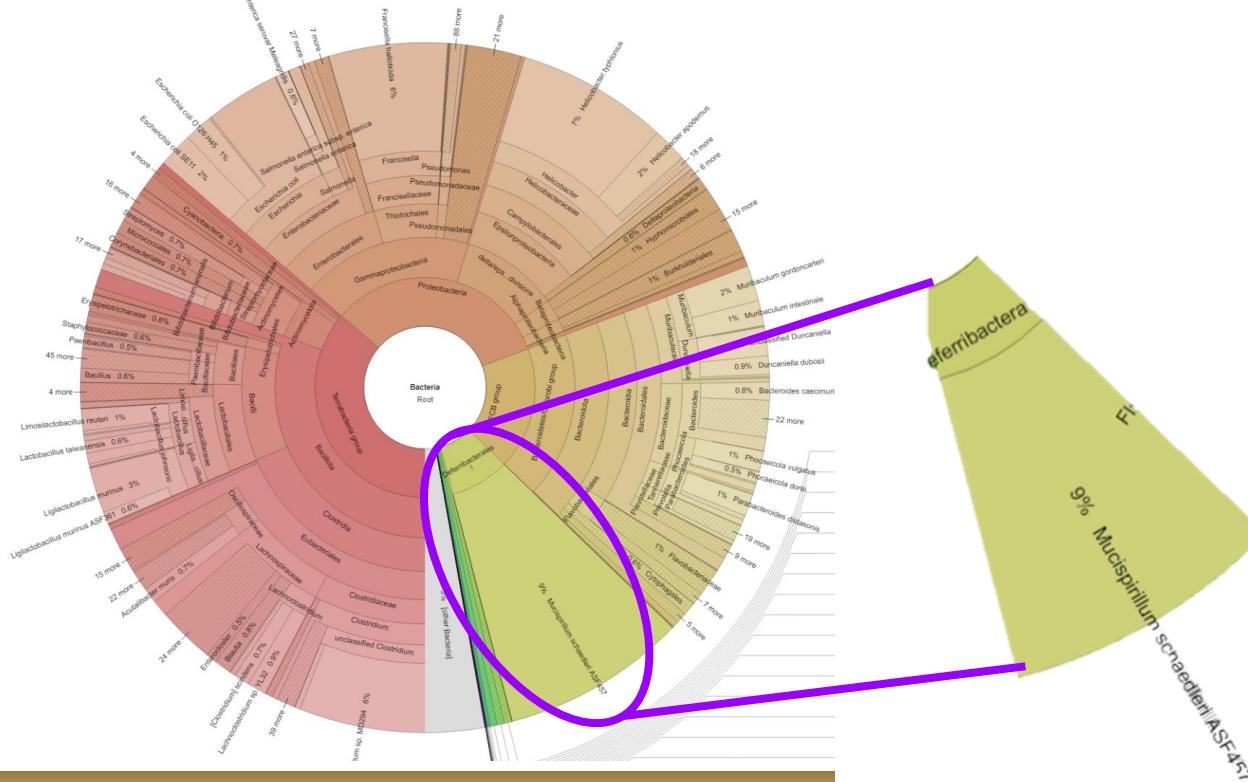
# Invoke Kraken
kraken-train
kraken-align -t 5 -m 3 sample.fq.gz sample.ed.bam > grc.fasta
kraken-report sample.ed.bam --output report.txt --output out.txt

# Invoke Krona
ktImportTaxonomy -t 5 -m 3 report.txt
```

# Microbiome Analysis



# Microbiome Analysis - Bacteria

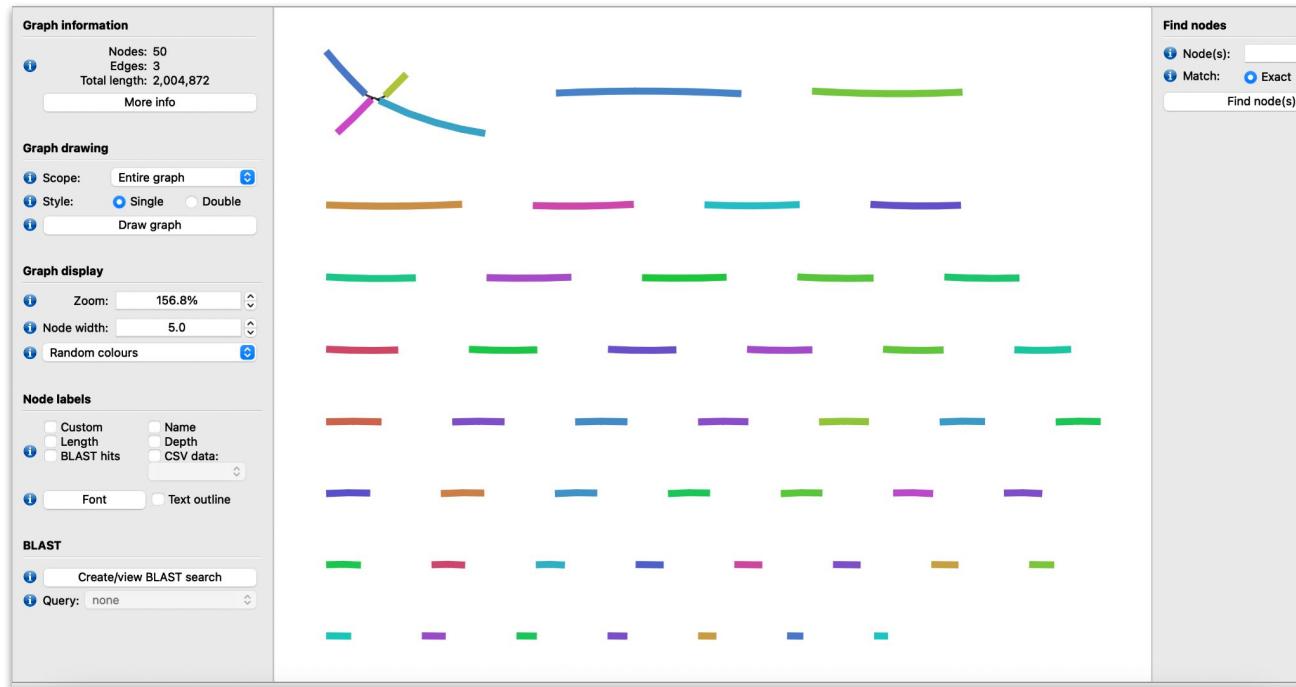


# Mucispirillum Schaedleri ASF457

The screenshot shows the NCBI Taxonomy Browser interface. At the top, there is a navigation bar with links for Entrez, PubMed, Nucleotide, and Protein. Below the navigation bar is a search bar with the query "Mucispirillum schaedleri ASF457". The search results page displays the following information:

- Taxonomy ID: 1379858** (for references in articles please use NCBI:txid1379858)
- current name**: **Mucispirillum schaedleri ASF457**
- NCBI BLAST name: bacteria**
- Rank: strain**
- Genetic code: [Translation table 11 \(Bacterial, Archaeal and Plant Plastid\)](#)**
- Lineage (full)**: [cellular organisms](#); [Bacteria](#); [Deferribacteres](#); [Deferribacteres](#); [Deferribacterales](#); [Mu](#)

# Mucispirillum Schaedleri ASF457 - Assembly



# Assembly Assessment

 **Graph information**

**File name:** /Users/omerwachman/omer/DNA/FinalAssignment/asf.gfa

**Graph size**

Node count:	50
Edge count:	3
Edge overlaps:	7,865 to 9,501 bp
Total length:	2,004,872 bp
Total length (no overlaps):	1,970,041 bp

**Node sizes**

N50:	48,555 bp
Shortest node:	9,949 bp
Lower quartile node:	21,177 bp
Median node:	33,421 bp
Upper quartile node:	50,542 bp
Longest node:	131,538 bp

**Graph connectivity**

Dead ends:	96
Percentage dead ends:	96.00%
Connected components:	47
Largest component:	177,529 bp (8.85%)
Total length orphaned nodes:	1,827,343 bp (91.15%)

**Depth**

Median depth:	1.00x
Estimated sequence length:	1,977,906 bp

**Close**

	[REF]	[QRY]
[Sequences]		
TotalSeqs	1	50
AlignedSeqs	1(100.00%)	50(100.00%)
UnalignedSeqs	0(0.00%)	0(0.00%)
[Bases]		
TotalBases	2347083	2004872
AlignedBases	1975841(84.18%)	1901458(94.84%)
UnalignedBases	371242(15.82%)	103414(5.16%)

N50	48,555 bp
L50	14 contigs

# Mucispirillum schaedleri ASF457

Article

*Mucispirillum schaedleri*  
Antagonizes *Salmonella*  
Virulence to Protect Mice  
against Colitis

Simone Herp<sup>1, 11</sup> ... Bärbel Stecher<sup>1, 2, 14</sup>  

Show more 

 Outline



Share



Cite

WALT DISNEY'S  
**MICKEY MOUSE**

TRADE MARK

REGISTERED

THE  
END

