Monday Meeting presentation

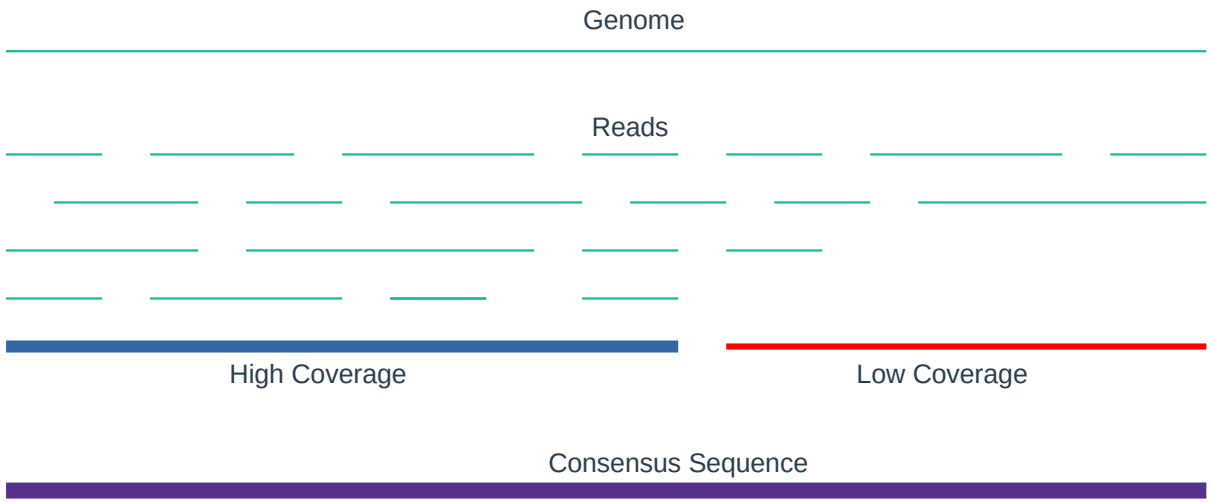
15/04/2024

Million Miles High

Nikiforos Pyrounakis
Research assistant

Gene Consensus Challenges

- Normally a consensus version of a gene is called in a simple sample.
- Most organisms being only 1-100x covered, it is difficult to distinguish – sequencing errors and real rare variants.



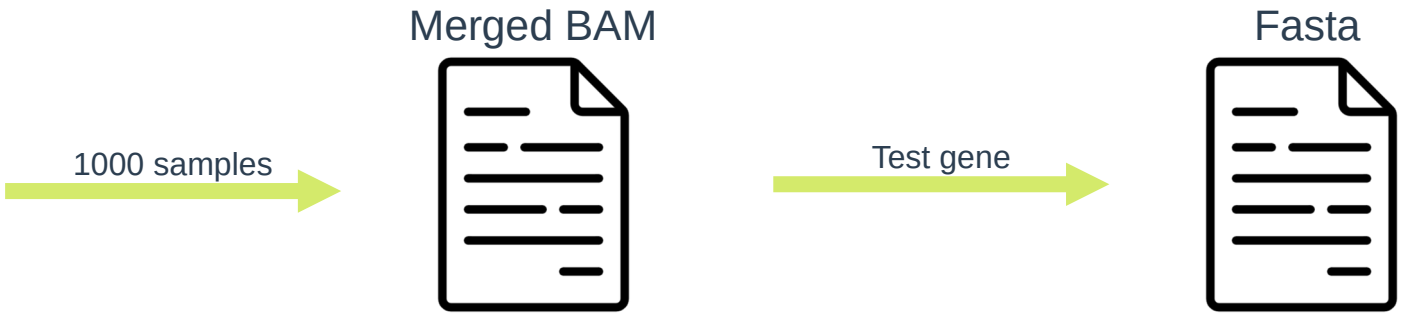
Million Miles High Alignment

- Recruit hundreds of thousand of samples
- Very deep alignments
- Subtract the background noise of sequencing errors

- Analyze how each position within the gene mutates across different bacterial populations.
  - Study mutation rate globally across all samples
  - Within specific subsets of the bacterial population

First steps

- AvA2 samples
  - Trimmed
- KMA
  - Panres
- Samtools
  - Fixmate
  - Sort

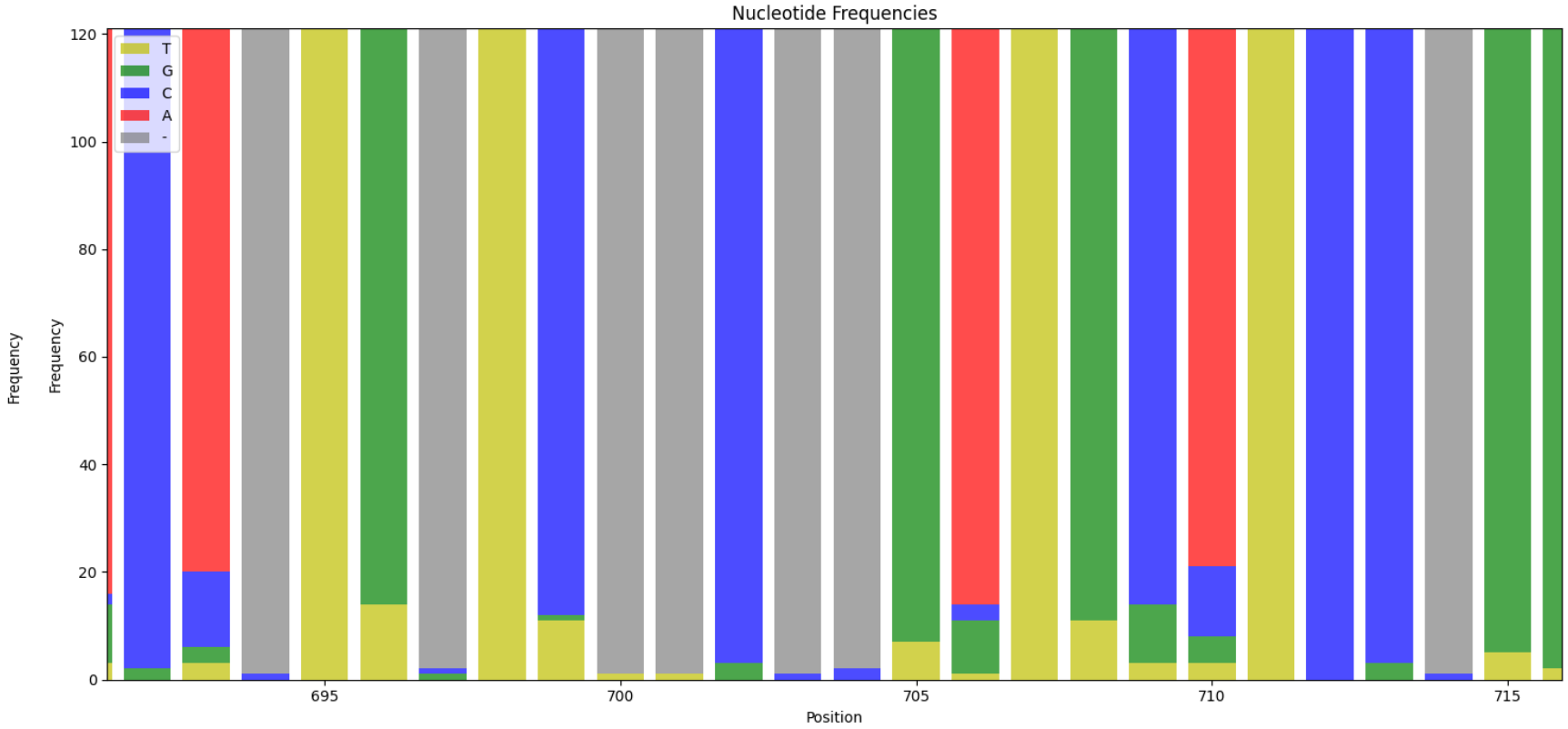


Error rates

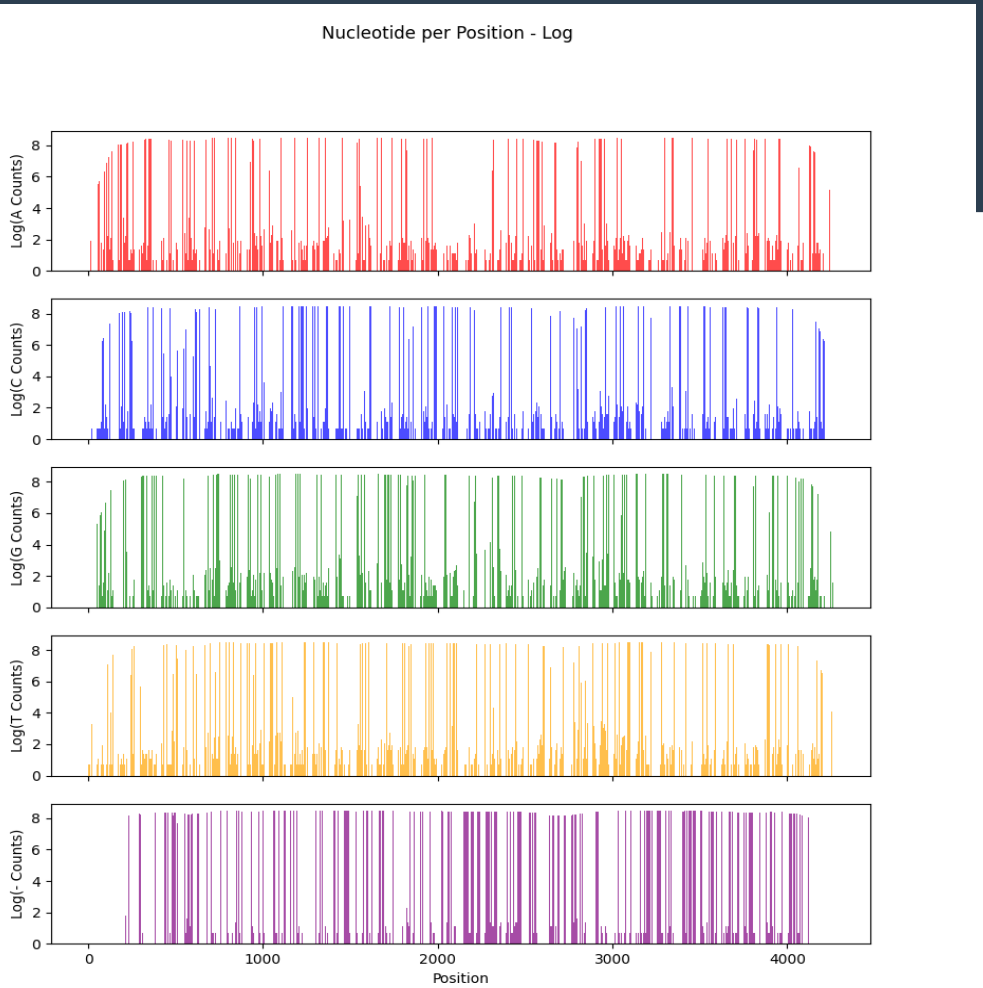
- With the aim on investigating nucleotide frequencies per position:

Position	A	C	G	T	-	A%	C%	G%	T%	-%
1	0	0	0	1	0	0.00%	0.00%	0.00%	0.00%	0.00%
2	0	0	0	1	0	0.00%	0.00%	0.00%	0.00%	0.00%
3	0	0	0	1	0	0.00%	0.00%	0.00%	0.00%	0.00%
4	0	0	0	1	0	0.00%	0.00%	0.00%	0.00%	0.00%
240	3568	6	0	1	0	8.12%	0.01%	0.00%	0.00%	0.00%
241	0	2	3578	1	0	0.00%	0.00%	8.15%	0.00%	0.00%
242	900	1	2675	18	2	2.05%	0.00%	6.09%	0.04%	0.00%
243	0	2	3609	2	0	0.00%	0.00%	8.22%	0.00%	0.00%
269	0	0	1	0	3869	0.00%	0.00%	0.00%	0.00%	8.81%
270	39	2	4	3810	44	0.09%	0.00%	0.01%	8.68%	0.10%
271	1	3811	0	103	0	0.00%	8.68%	0.00%	0.23%	0.00%
272	0	5	0	0	3910	0.00%	0.01%	0.00%	0.00%	8.90%

Error rates


Nucleotide Frequencies
T
G
C
A
-
Frequency
Frequency
120
100
80
60
40
20
0
695
700
705
710
715
Position

Error rates

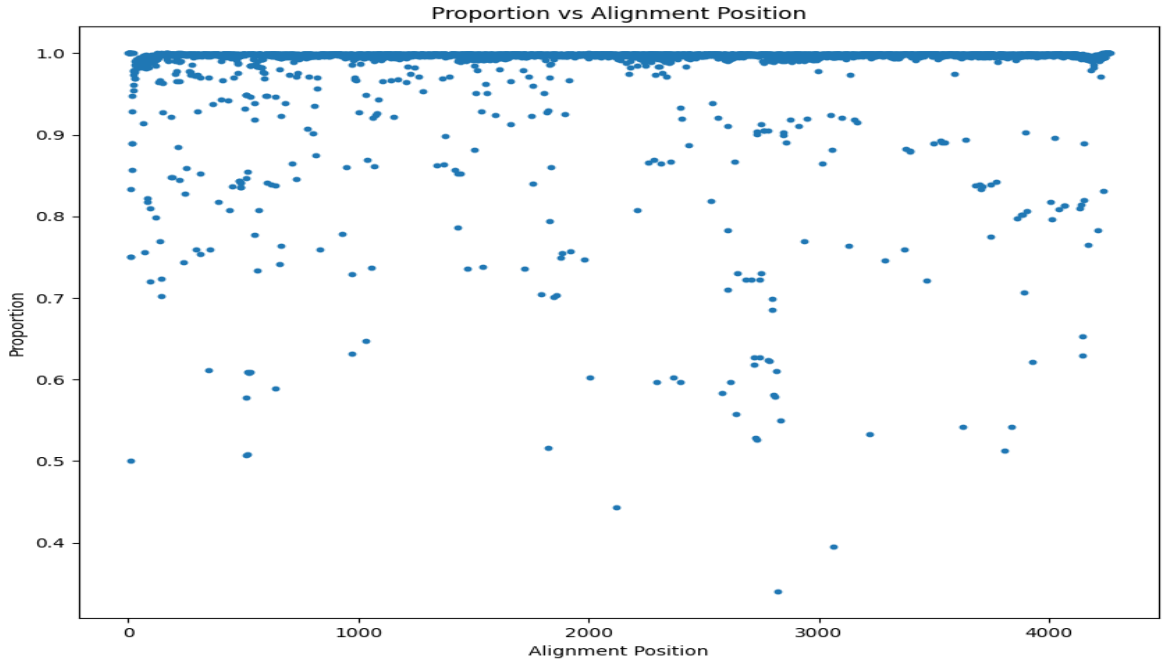

Nucleotide per Position - Log
Log(A Counts)
Log(C Counts)
Log(G Counts)
Log(T Counts)
Log(- Counts)
0
2
4
6
8
0
1000
2000
3000
4000
Position

Error rates

- For each position we wanted to know:
  - Which nucl was #1, #2, #3, #4, #5
  - What is the ratio between #1 and the sum of the other

1	0	0	0	1	0	0.00%	0.00%	0.00%	0.00%	0.00%	1	1	1
2	0	0	0	1	0	0.00%	0.00%	0.00%	0.00%	0.00%	1	1	1
3	0	0	0	1	0	0.00%	0.00%	0.00%	0.00%	0.00%	1	1	1
4	0	0	0	1	0	0.00%	0.00%	0.00%	0.00%	0.00%	1	1	1
240	3568	6	0	1	0	8.12%	0.01%	0.00%	0.00%	0.00%	3568	3575	0.998041958041958
241	0	2	3578	1	0	0.00%	0.00%	8.15%	0.00%	0.00%	3578	3581	0.99916224518291
242	900	1	2675	18	2	2.05%	0.00%	6.09%	0.04%	0.00%	2675	3596	0.743882091212458
243	0	2	3609	2	0	0.00%	0.00%	8.22%	0.00%	0.00%	3609	3613	0.998892886797675
269	0	0	1	0	3869	0.00%	0.00%	0.00%	0.00%	8.81%	3869	3870	0.999741602067183
270	39	2	4	3810	44	0.09%	0.00%	0.01%	8.68%	0.10%	3810	3899	0.977173634265196
271	1	3811	0	103	0	0.00%	8.68%	0.00%	0.23%	0.00%	3811	3915	0.973435504469987
272	0	5	0	0	3910	0.00%	0.01%	0.00%	0.00%	8.90%	3910	3915	0.998722860791826

Error rates


Proportion vs Alignment Position
Proportion
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0
1000
2000
3000
4000
Alignment Position

Already existing tools

- We decided to try already existing tools that focus on variant calling:
  - Bcftools
  - Vcftools
  - freebayes
  - GATK
  - InStrain
  - metaSNV

Struggles with KMA

- KMA removes some valuable variant calling information:
  - Bcftools – Runs with KMA
  - Vcftools – Runs with KMA
  - freebayes – Runs with KMA
  - GATK – Complains for NM tag
  - InStrain – Complains for NM tag
  - MetaSNV – Complains for NM tag

NM tag: Edit distance, which is the number of mismatches and gap openings in the alignment

freebayes

- Developed in 2012

- Bayesian genetic variant detector designed to find small polymorphisms, specifically SNPs , indels, MNPs, and complex events (composite insertion and substitution events) smaller than the length of a short-read sequencing alignment.

- Important flags:
  - **min-alternate-fraction**: Sets the minimum fraction of observations that must support the alternate allele. Here, a fraction of 0 means that even a single observation of the alternate allele is sufficient.
  - **pooled-continuous**: Estimate the frequency across the entire pool of samples
  - **report-monomorphic**: Report positions where all samples in the input data have the same allele
  - **haplotype-length**: Consider every possible combination of variants

freebayes

- **Run 1**
  - Min-alternate-fraction: 0.1
    - 143 variants
    - 14 min
- **Run 2**
  - Min-alternate-fraction: 0.01
    - 238 variants
    - 14 min
- **Run 3**
  - Min-alternate-fraction: 0.001
    - 1362 variants
    - 12 min
- **Run 4**
  - Min-alternate-fraction: 0.0001
    - 1256 variants
    - 10 min
- **Run 5**
  - Min-alternate-fraction: 0.5
    - 64 variants
    - 14 min

Moving forward

- Continue testing freebayes
  - Filter variant results
- Try aligning again the samples with BWA
  - Designed for short-reads
  - Keep useful tags

Thank you!