Genome Matching

Team Number: 20
Team name: The Dominant Trait

Members:
Aditya Gupta Armaan Khetarpaul Sahil Chaudhary
Sirjan Hansda Suhas Vundavilli

Introduction

The human genome is highly complex and exhibits natural variations across individuals. One of the key challenges in genomics is to identify these variations and understand their biological significance. In this project, we focused on the analysis of the X chromosome, specifically aiming to reconstruct an individual's X chromosome and identify exon regions and single nucleotide variants.

Exons are coding regions that are crucial for protein synthesis and identifying them helps us understand gene function and expression. Additionally, we aimed to pinpoint locations where the individual's genome differs from the reference sequence focusing on mismatches and regions with maximum alignment.

This study has implications for genetic diversity, identifying disease-associated mutations, and contributing to personalized medicine efforts. Using efficient algorithms for sequence alignment, coverage analysis, we performed a small investigation of the X chromosome to highlight exon regions and look into their biological significance.

Methodology

Succinct Rank Data Structure

This data structure helps us compute the rank of an index (number of 1s before the index) in a binary array. It's implemented using a 2-level approach. For this, divide the array into larger blocks of size 32 ($O(logn)$ where n is the length of the array). Then precompute the rank of each block and stored it in a an array.Next, divide each block into smaller blocks of size 4 and precomputed the rank of each of these blocks and stored it in a 2D array.
To perform a query for the prefix sum at index k:

1. Divide k by 32 to get the index of the block containing k. Write down the prefix sum at the start of that block.

2. Divide k mod 32 by 4 to get the index of the miniblock containing k. Write down the prefix sum at the start of the miniblock.
3. Count the number of 1s in the miniblock from the starrting of the block to $k - 1$.
4. Add these values together.

This approach has a space complexity of roughly $O(n \log \log n)$ and a query time complexity of $O(1)$.

Searching and Aligning Reads

Define a valid read to be one, which can be matched into the reference with two mismatches, or two insertions + two deletions. In such a case, at least one of the 3 parts must match exactly. Also, define the edit distance between two strings to be the minimum number of deletions + insertions required to convert one string to the other.

For every read do the following:

- Divide the read into 3 equal parts.
- We find a matching part.
- For every position that the matching part can be potentially aligned to, do the following:
  - Find the corresponding sub string in reference.
  - Use a linear matching to count the number of mismatches, and if the number of mismatches is less than or equal to two and less than the current best number of matches, update the current best
  - Otherwise, find the edit distance between the read and sub string of reference. If number of insertions is less than or equal to two, and less than min number of errors, update the current best.
- Find minimum errors over all positions where the matching part could be put.
- Repeat the exercise for read reverse complement.
- Consider the configuration which led to minimum number of errors (i.e. minimum number of mismatches or minimum number of insertions/deletions) from read or read reverse complement.

Coverage Based Calculations

Coverage: The coverage array, is initially set to zeros and with a length matching the reference genome. It tracks the depth of read alignment at each position. For every read that aligns to the reference, the array gets incremented by 1 at each covered index, reflecting the total coverage across the genome.

Next, based on this coverage array, we set a threshold(T). This threshold is set according to this equation:

$$T = \pi_{coverage} + k \times \sigma_{coverage}$$

where,

1. $\pi_{coverage}$ is the average coverage of all the the base pair in the reference genome.
2. $\sigma_{coverage}$ is the standard deviation of the coverage of the base pair
3. k is a hyper-parameter that we choose. Currently set to 7

Next we iterate through the entire coverage array and follow the following steps:

1. If the coverage of the current base pair is greater than the threshold, then, **if the flag is not already set to true**, set the `exon` flag to `True` and mark the current positions as a potential starting position of an exon.
2. If the coverage of the current base pair falls below threshold , then **if the flag is not set to false**, mark the `exon` flag to `false` and set the end position as the current base pair location.
3. Add the start and end positions as a tuple into the potential exon regions.

According to Sarkar et. al [4], "80% of the exons on each chromosome are < **200** bp in length" which aligns with our result, where we found the average exon length to be: **123.3265**.

Results

After matching all the reads to the reference, we were able to fit about 11% of the total reference. We also obtained a total of 595762 mismatches, which for a chromosome of length 150 million is about 0.4% mismatches. This is in agreement with the standard value of the number of mismatches.

Now we can fit the reads like a jigsaw and find the number of mismatches per region. Upon plotting a histogram of bin size 10000 for this, we obtain the following graph:

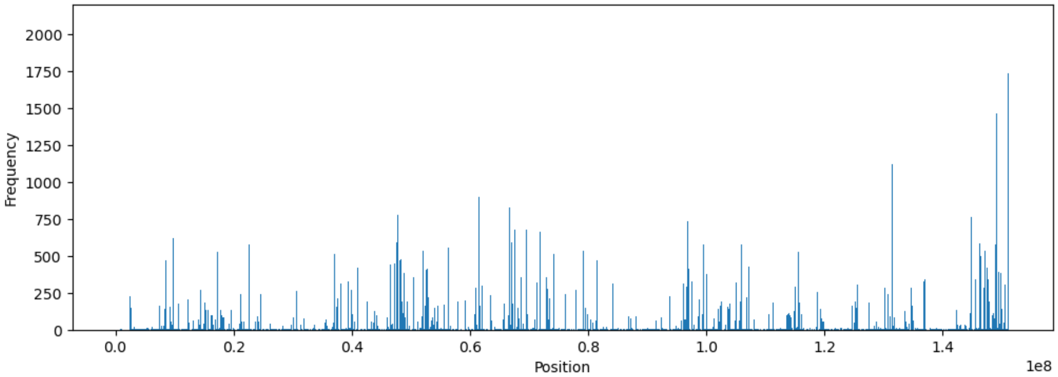

Figure 1: Total no of mismatches per location

We can see that there are a high number of mismatches at the end, which correspond to the red and green exon regions. This is expected since the reads belong to Prof. Ramesh, who is diagnosed to be red-green colorblind. We also obtain a large number of mismatches in the middle regions, which requires further investigation. These regions of mismatch correspond to some known genes of the X chromosome. A high number of mismatches in these regions could affect their functioning and cause severe issues.

Next we plotted the number of reads that match to each location. This will help us locate the exon regions, since we expect to see a high number of reads from the exon regions. The plot is as follows:

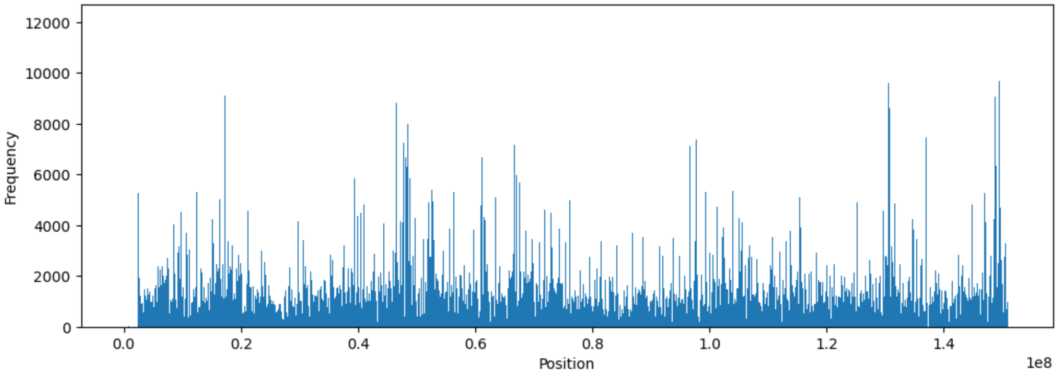

Figure 2: No of reads aligning with each location

From the graph we can see a lot of regions with a high number of reads. All of these are possible exon regions. By taking an average of the total matches and labelling the regions with a number of reads significantly higher than the average, we were able to identify 3323 exon regions, which is close to the actual number of 4373 exons [3].

To find the regions of high genomic mismatch, we divide the number of mismatches by the number of reads at each location. This will give us the average number of mismatches per read for that location. We then plot this data to get the following graph:

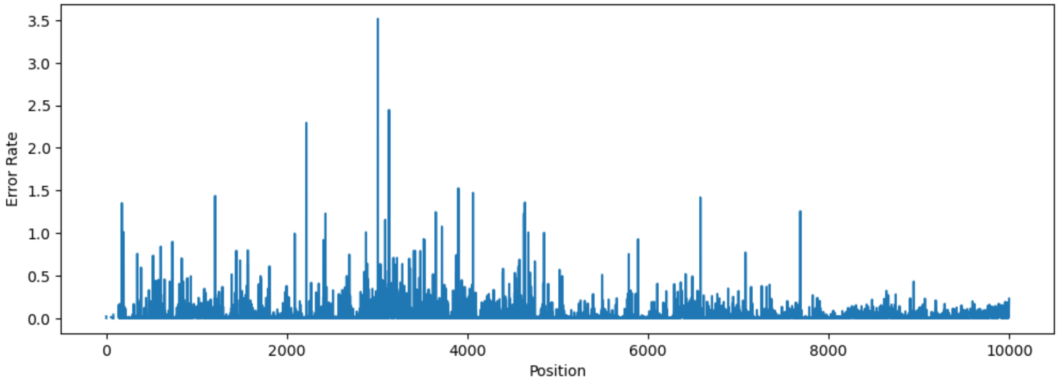

Figure 3: Average number of mismatches per read

We notice that this graph peaks somewhere near the middle. On taking a maximum, we identify the regions with the highest average mismatches per read, which in this case are the 3008 and 3132 bins (out of 10000). Since we know the total length of the X chromosome, we can calculate the actual locations of these regions.

The above regions lie in the Xp 11.3 band of the X chromosome. This band contains some important genes like MAOA, MAOB, EFHC2, NDP and CDK16. Out of these, the NDP and CDK16 genes are the ones closest to the identified regions. The NDP gene is associated with Norrie disease, which is a rare genetic disorder which causes abnormal development of the retina of the eye [1]. The CDK16 gene is associated with producing enzymes that are involved in autophagy, mainly decreasing cancer cell proliferation [2].

Conclusion

To conclude, we have matched the reads to the reference and obtained graphs depicting the locations of gene mismatches and read matches.

We were able to fit 11% of the total reference with the given reads, and 0.4% of mismatches of the entire chromosome length. We were able to correctly identify 3323 exon regions, which is close to the actual number of exon regions, specifically 4373, as seen in literature. By doing so, we were able to obtain the exon regions and also exons with a high number of mismatches per read.

This led us to the Xp 11.3 band in the X chromosome as the region of highest average mismatches. Since this band corresponds to genes like NDP, which affects retinal development, and CDK16 which affects cancer cell proliferation, a higher number of mismatches here could result in severe complications.

References

[1] ZY Chen et al. "Isolation and characterization of a candidate gene for Norrie disease". In: *Nature Genetics* 1.3 (June 1992), pp. 204–208. DOI: 10.1038/ng0692-204.

[2] Javad Karimbayli et al. "Insights into the structural and functional activities of forgotten Kinases: PCTAIREs CDKs". In: *Molecular Cancer* 23.1 (2024), p. 135. ISSN: 1476-4598. DOI: 10.1186/s12943-024-02043-6. URL: https://doi.org/10.1186/s12943-024-02043-6.

[3] M. Ross, D. Grafham, A. Coffey, et al. "The DNA sequence of the human X chromosome". In: *Nature* 434.7031 (2005), pp. 325–337. DOI: 10.1038/nature03440.

[4] Meena Sakharkar, Vincent Chow, and Pandjassarame Kangueane. "Distributions of exons and introns in the human genome". In: *In silico biology* 4 (Jan. 2004), pp. 387–93.