The Dominant Trait Gene Matching

Aditya Gupta
Armaan Khetarpaul
Sahil Chaudhary
Sirjan Hansda
Suhas Vundavilli

Problem

Identify genomic variants between Prof. Ramesh's X chromosome and the reference genome.

Why do genome matching?

- Provide useful insights into where a person is different from the reference human.
- Investigate the regions of mismatches closely.
- Study the consequences of these mismatches on their overall being.

Goals

With this project, we aim to

- Analyze Prof. Ramesh's DNA reads and compare them with the reference Genome.
- Use various statistical analysis methods to find out the regions of interest, such as regions of high matching and higher mismatches.
- Analyze how the mismatches in Prof. Ramesh's DNA with the reference genome might affect him.

Method

Succint Rank data structure

- Traditional Rank: Divide into blocks, compute and store block sums, do a linear pass to get rank. O(nlogn) space, O(1) time
- Problem? Linear pass is slightly heavy and takes space
- Idea: Treat the linear scan as another rank query
- Divide each block into mini blocks
- Compute mini block sums
- O(nloglogn) space, O(1) time [O(32) to O(4)]

Method

Searching and Aligning

- Valid Read : At most two mismatches, or two insertions or deletions.
- For every read :
  a) Divide into 3 parts
  b) If it is a valid read, at least one of the three parts must match
  c) For all potentially valid positions of the part :
     i. Find the corresponding position in the reference
     ii. Count mismatches (O(n)), if <= 2 and < current best, then update
     iii. Otherwise, find edit distance ($O(n^2)$), if <= 2 and < current best, then update
  d) Repeat for reverse complement of the read
  e) Place the read at position of least error

Method

Heuristic and coverage based calculations

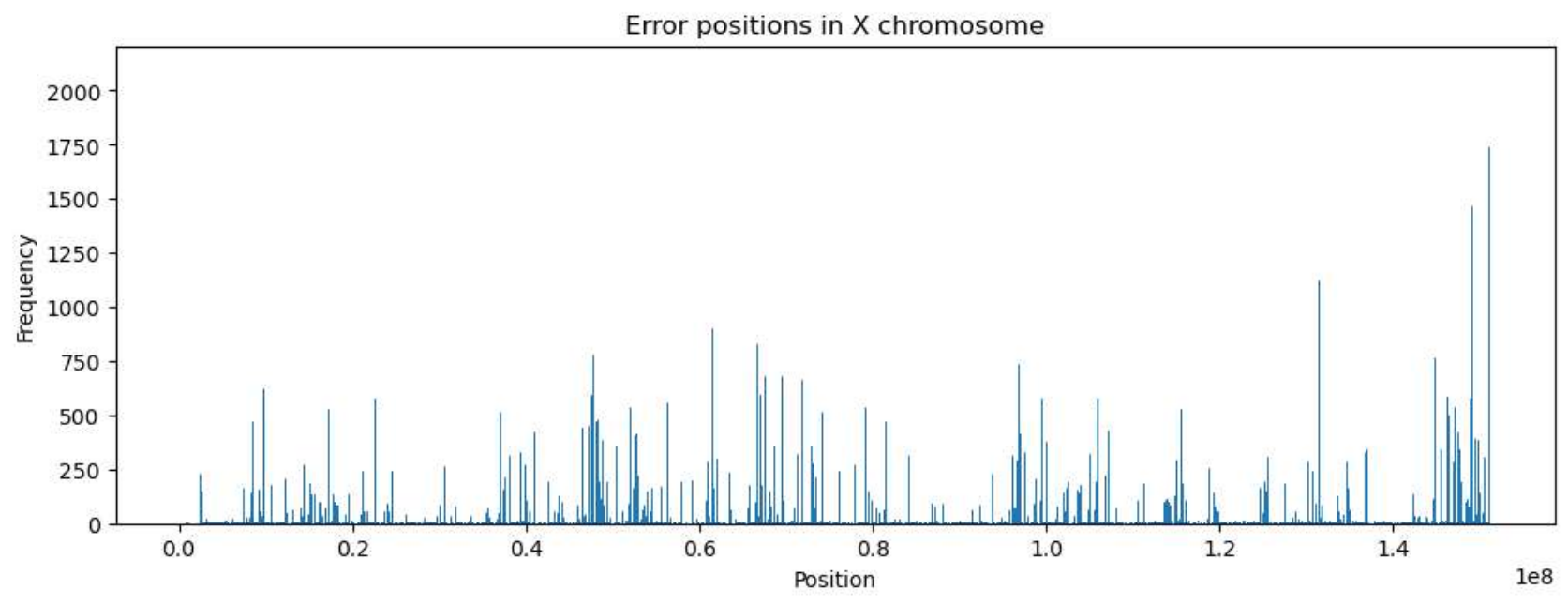
- Coverage Array : Frequency array with same size as reference. Counts the number of matches at every index
- μ —> Avg(Coverage), σ —> Stdev(Coverage), k —> Hyper param (7 here)
- $t = \mu + k \times \sigma$
- Mark portions in the coverage array where coverage is > t
- These positions will serve as regions of high exon concentration

Other Statistics :

- Distinct positions where mismatches have occurred
- Distinct positions where matches have occurred
- Ratio of mismatches/reads over 10,000 bins

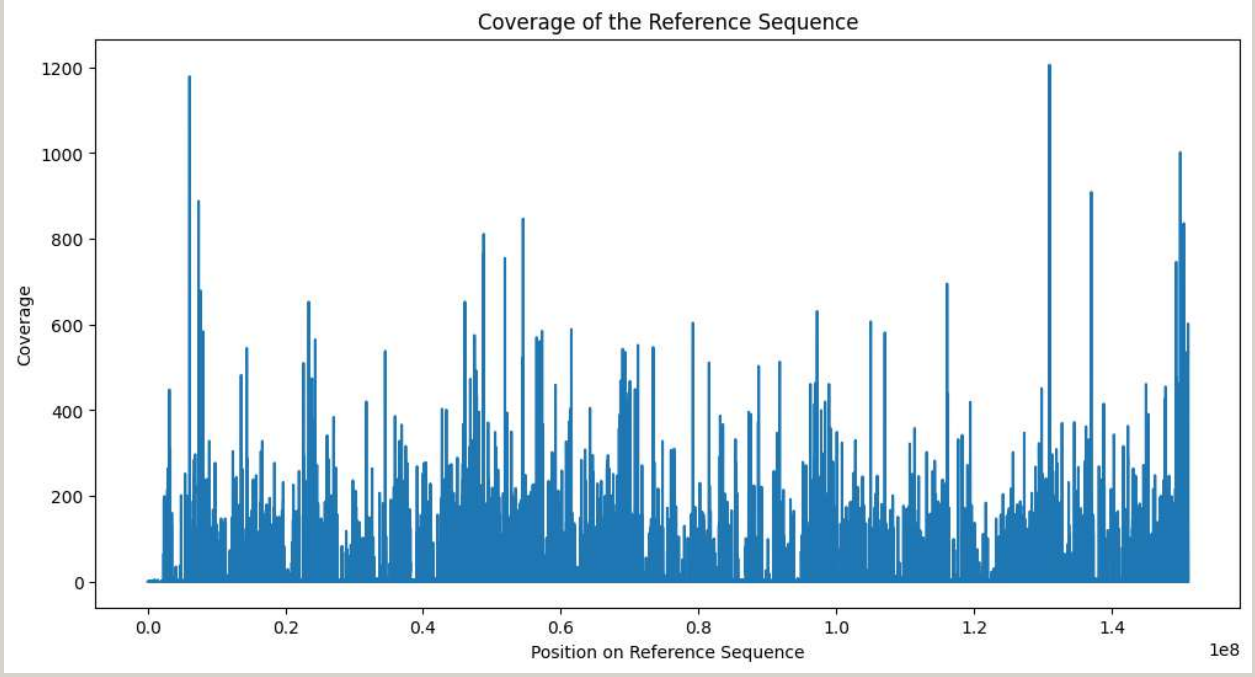
Results

We were able to fit about 11% of the total X chromosome with the reads and found a 0.4% total mismatch



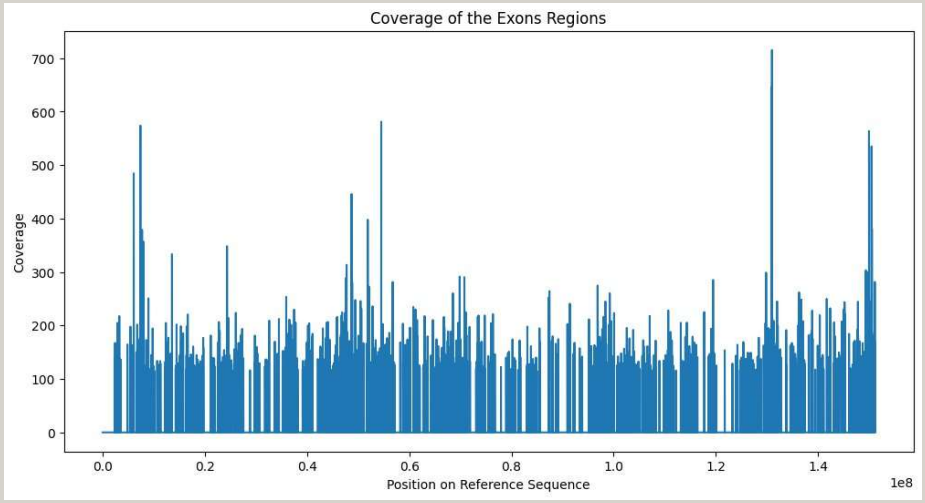
Exon Prediction

Locations with high number of reads correspond to exon regions

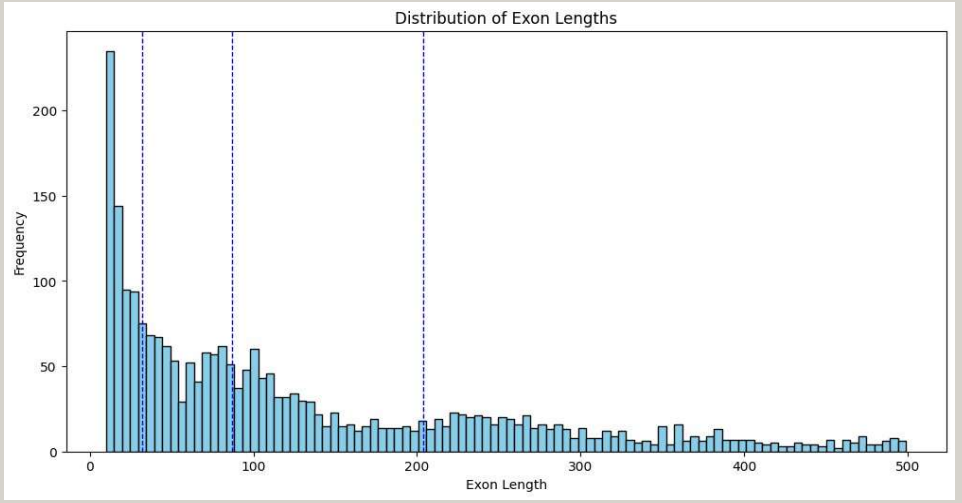


Exon Prediction

We used heuristic analysis to find potential coverage regions



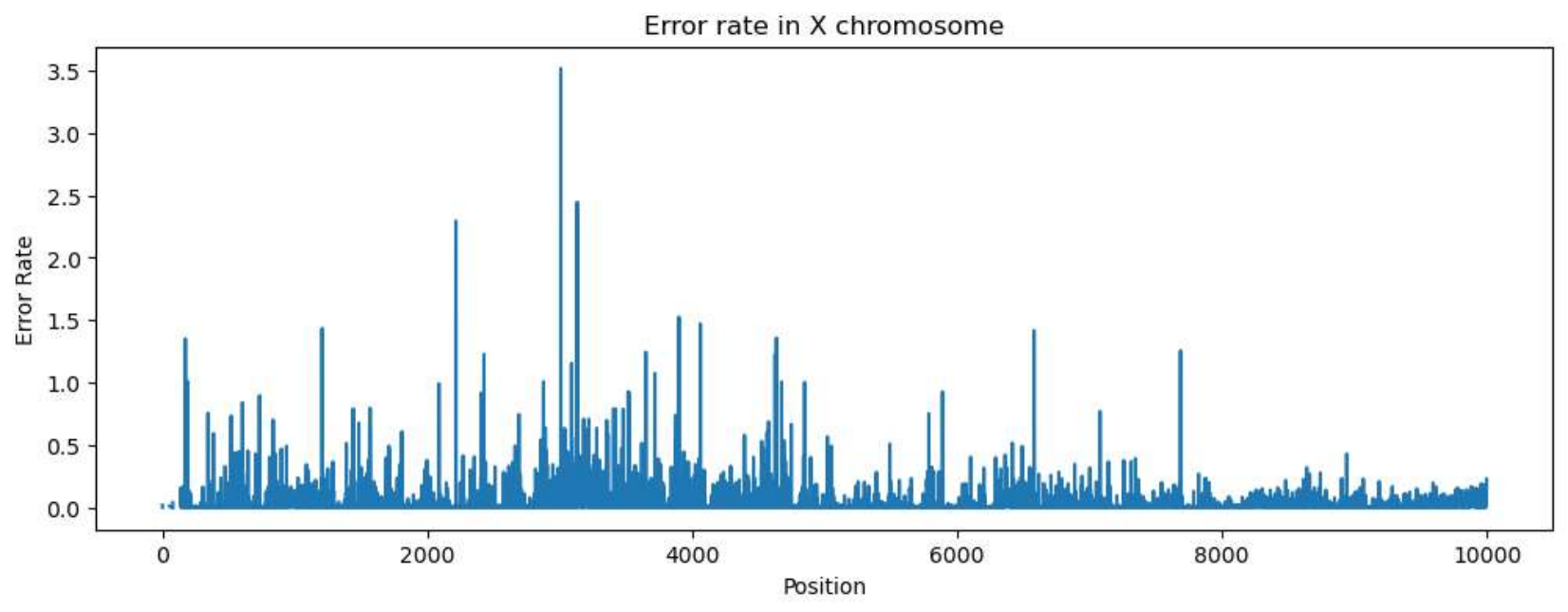
About 3323 exons region found



About 75% of exons are < 200 bp in length

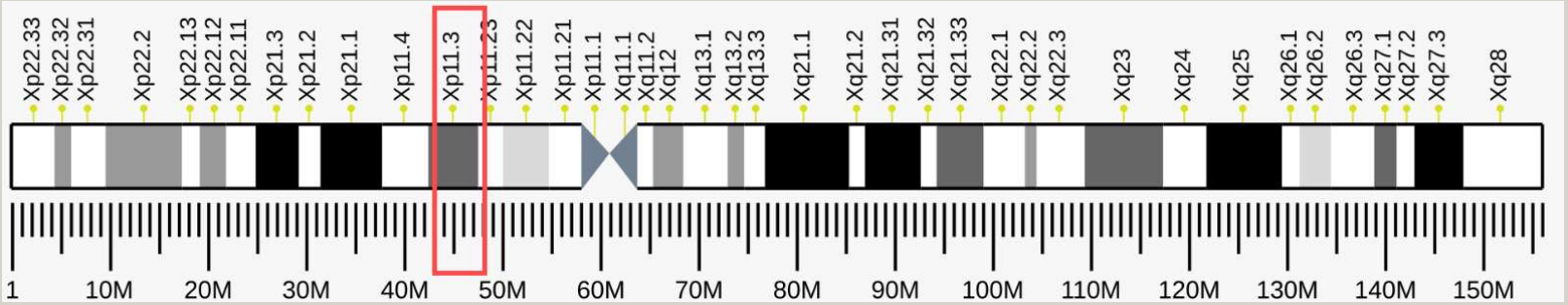
Gene Anomalies

The average mismatches per read gives us a clear indication of exons with anomalies



Gene Anomalies

- On analyzing the histogram, we find that the highest average mismatches per read occurred in bins labelled 3008 and 3132. These roughly correspond to the locations between 46 million and 48 million base pairs in the X chromosome



Source : Wikipedia

- As is evident from the diagram, this location lies on the Xp11.3 band.

Gene Anomalies

- Some important genes on the Xp11.3 band are :
  a) MAOA
  b) MAOB
  c) EFHC2
  d) NDP
  e) CDK16
- Of these, the most interesting ones are NDP and CDK16.
- The NDP gene is responsible for encoding the Norrin protein, which is responsible for retinal development. Mismatches here might signal towards potentially abnormal development of the retina, which may even lead to blindness.
- The CDK16 gene is responsible for encoding the enzyme Serine/threonine-protein kinase PCTAIRE-1, which helps in autophagy, particularly decreasing cancer cell proliferation. Abnormalities in this gene are shown to be linked to lung, skin, breast, prostate and many other such cancers.

Conclusions

- Read Alignment: Successfully matched 11% of the reference genome with a mismatch rate of 0.4%, aligning closely with expected values.
- Exon Identification: Identified 3323 exon regions, closely approximating the literature-reported 4373.
- High-Mismatch Regions: Found significant mismatch regions in the Xp 11.3 band of the X chromosome, containing critical genes such as **NDP** and **CDK16.**
- Biological Implications: Highlighted the potential consequences of mismatches in functional gene regions, underlining the importance of accurate genome reconstruction.

Drawbacks

- Methodological Limitations:
  a) Only 11% of the reference genome was analyzed, limiting the scope of findings.
  b) Assumed uniformity in mismatch patterns; further studies are needed to validate biological relevance.
- Data Constraints:
  a) Limited dataset size may have excluded rare variants or mutations.
  b) Coverage thresholds could potentially overlook minor exons.

Further Scope

- Broader genome coverage: Expand the study to include the entire genome for a more comprehensive view of genetic variations
- Enhanced data collection: Integrate larger, more diverse datasets to capture rare variants and improve accuracy in identifying significant mutations
- Refinement of Exon identification: Implement adaptive thresholds or multi-parameter analysis to improve the detection of smaller or less common exon regions

Thank you