Current Affairs: Utilizing Oxford Nanopore Sequencing Data to Detect Non-Canonical DNA Structures

Alexander Turco

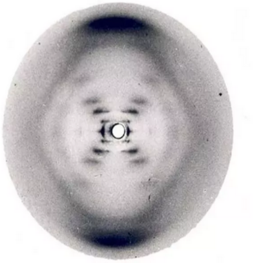
January 30, 2024

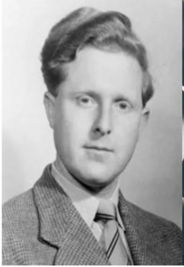
UHN Princess Margaret Cancer Centre

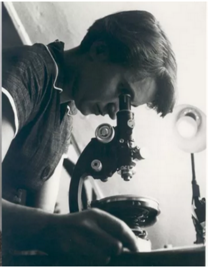
The Structure of DNA: A Brief History Lesson

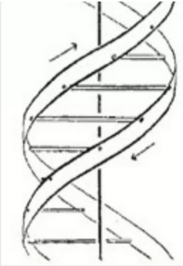
- In 1953, Watson, Crick, Wilkins, Franklin, and Gosling were the first to describe the structure of DNA

- They discovered the right-handed double helix (canonical B-form DNA), the most common form found in cells







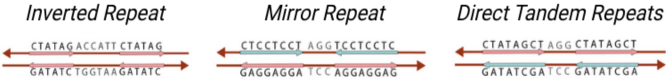


[1] Watson and Crick, 1953

DNA Can Adopt Alternative Structures

- Now, more than 15 types of DNA structure that differ from the canonical B-form have been reported (non-canonical or non-B form DNA)

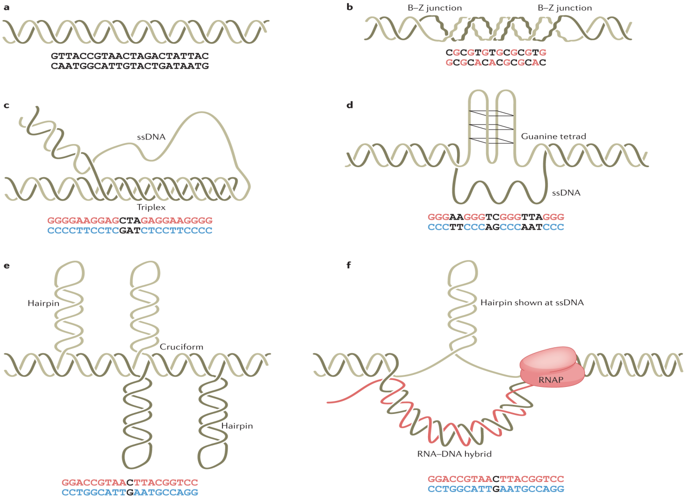
- Through sequencing of the human genome, we now know over half the genome is composed of repetitive elements - these were initially thought to be ‘junk DNA’



- A crucial feature of some repetitive sequences is their ability to fold into non-canonical DNA structures (non-B DNA)

[2] Wang and Vasquez, 2023

Types of Non-canonical DNA structures



[2] Figure from Wang and Vasquez, 2023

Non-Canonical DNA Structures are Involved in Biological Processes

Non-B DNA structures have been shown to co-localize with functional genomic loci (promoters, enhancers, etc) and genetic instability hotspots

This suggests a role for non-B DNA in vital cellular events such as;

- Regulation of transcription
- Regulation of DNA replication and recombination
- Regulating genome integrity

Diseases Associated with Non-Canonical DNA structures

Repeat Expansion Diseases: Expansions of non-B DNA structure-forming repeats have been implicated in many neurodegenerative and neuromuscular diseases.

Genetic Instability Diseases: Non-canonical DNA structures are associated with increased mutability (point mutations, deletions, insertions and chromosomal translocations)

- Enriched at chromosomal breakpoints in translocation-related cancers such as lymphomas and leukaemias.
- Can be recognized by DNA repair proteins, triggering error-generating repair processes
- G-quadruplexes are present within most human oncogenic promoters and at telomeres - a current theraputic target to downregulate transcription or block telomere elongation in cancer cells.

How are Non-B Structures Detected in the Genome?

Computational Approaches

- Sequence based computer algorithms
- Deep learning approaches
- Molecular dynamics simulations

Wet-lab Approaches

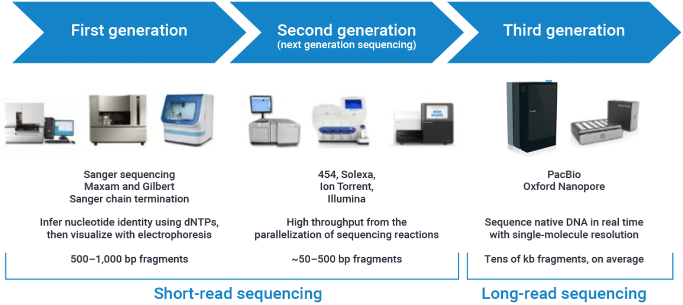
- circular dichroism spectra analysis
- Polymerase stop assays
- Immunoflourescence studies

These approaches are based primarily on DNA sequence motifs, which are necessary, but insufficient for formation and are not available for all non-B DNA structures

Third Generation Sequencing: A Promising New Approach

Single Molecule, Real Time Sequencing (SMRT): Pacbio's third generation sequencing machine

- Emits a fluorescent pulse when nucleotide is detected - the time interval between two pulses is called the interpulse duration (IPD)
- Guiblet et al (2018), showed that there is a significant divergence between IPDs in non-B DNA motif regions compared to B-DNA regions



First generation

Second generation
(next generation sequencing)

Third generation

Sanger sequencing
Maxam and Gilbert
Sanger chain termination

Infer nucleotide identity using dNTPs, then visualize with electrophoresis

500–1,000 bp fragments

454, Solexa,
Ion Torrent,
Illumina

High throughput from the parallelization of sequencing reactions

~50–500 bp fragments

PacBio
Oxford Nanopore

Sequence native DNA in real time with single-molecule resolution

Tens of kb fragments, on average

Short-read sequencing

Long-read sequencing

Oxford Nanopore Sequencing Technology

ONT Sequencer

placeholder box

Inside the Nanopore

Predicting Non-B Structures From Nanopore Sequencing

A recently published paper utilized translocation times from ONT sequencing to predict non-B DNA structures (citation)

- Developed the first computational pipeline and a novel unsupervised deep statistical model for predicting non-B DNA structures

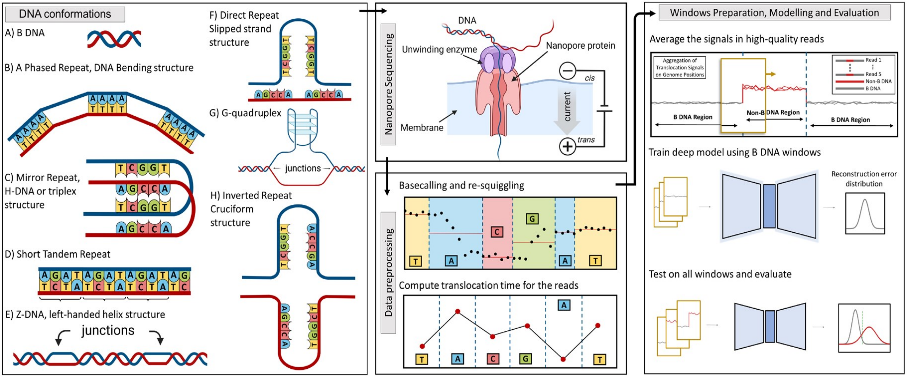
Benefits of unsupervised approach;

1. non-B database labels are noisy (just because motif is present does not mean structure is)
2. Even if high quality labeling for non-B DNA were available, substantially more B-DNA samples are available
3. Unknown non-B structures or non-B DNA without sequence motifs cannot be modelled by a supervised approach

[2] Hosseini et al., 2023

GoFAE-DND: Deep Statistical modelling of non-B DNA

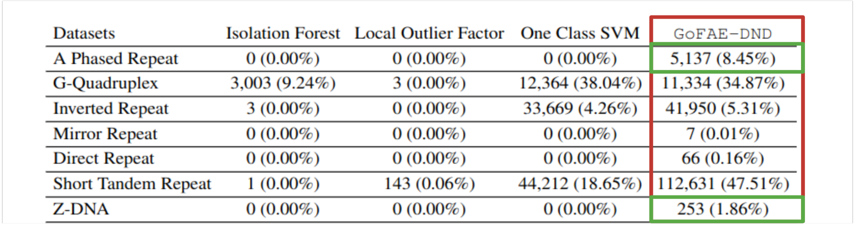
Anomaly Detection Problem: Identifying patterns within data that deviate significantly from the norm or expected behaviour of the majority of the data


DNA conformations
A) B DNA
B) A Phased Repeat, DNA Bending structure
C) Mirror Repeat, H-DNA or triplex structure
D) Short Tandem Repeat
E) Z-DNA, left-handed helix structure
junctions
F) Direct Repeat Slipped strand structure
G) G-quadruplex
junctions
H) Inverted Repeat Cruciform structure
Nanopore Sequencing
DNA
Unwinding enzyme
Nanopore protein
Membrane
current
cis
trans
Data preprocessing
Basecalling and re-squiggling
Compute translocation time for the reads
T
A
C
G
A
T
Windows Preparation, Modelling and Evaluation
Average the signals in high-quality reads
Aggregation of Translocation Signals on Genome Positions
Read 1
Read 5
Non-B DNA
B DNA
B DNA Region
Non-B DNA Region
B DNA Region
Train deep model using B DNA windows
Reconstruction error distribution
Test on all windows and evaluate

[2] Hosseini et al., 2023

Model Performance

At an FDR control level $\alpha = 0.2$, SVM and GoFAE-DND generated the most novelties, with GoFAE-DND yielding the most predictions for all non-B types besides G4



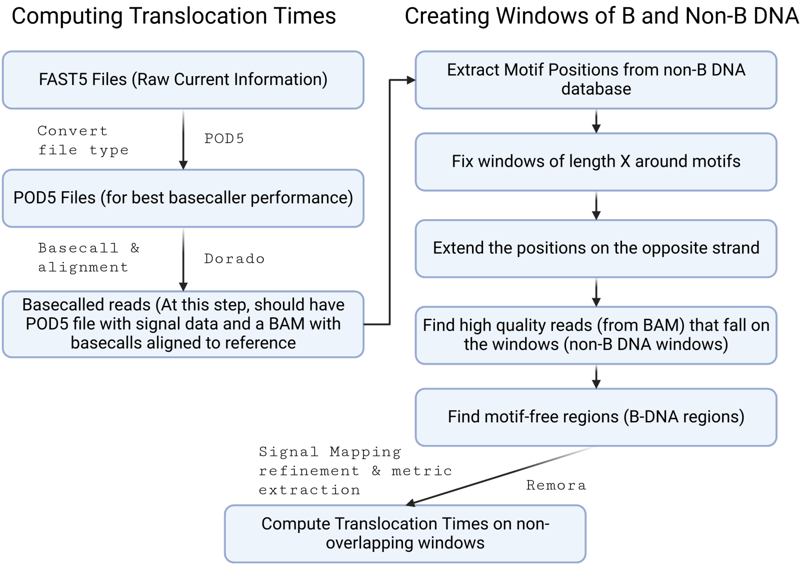
Datasets	Isolation Forest	Local Outlier Factor	One Class SVM	GoFAE-DND
A Phased Repeat	0 (0.00%)	0 (0.00%)	0 (0.00%)	5,137 (8.45%)
G-Quadruplex	3,003 (9.24%)	3 (0.00%)	12,364 (38.04%)	11,334 (34.87%)
Inverted Repeat	3 (0.00%)	0 (0.00%)	33,669 (4.26%)	41,950 (5.31%)
Mirror Repeat	0 (0.00%)	0 (0.00%)	0 (0.00%)	7 (0.01%)
Direct Repeat	0 (0.00%)	0 (0.00%)	0 (0.00%)	66 (0.16%)
Short Tandem Repeat	1 (0.00%)	143 (0.06%)	44,212 (18.65%)	112,631 (47.51%)
Z-DNA	0 (0.00%)	0 (0.00%)	0 (0.00%)	253 (1.86%)

The Objective

Given the dramatic increase in genome-scale data produced using ONT platforms, and the relevance of non-B structures in human cancers;

1. Utilize the model to analyze nanopore samples from the human pangenome reference consortium (HPRC)
2. Eventually improve the model, with an emphasis on the detection of G4 quadruplexes - Linking methylation, gene expression profiles which are available for HPRC samples

Preprocessing Workflow



Computing Translocation Times

FAST5 Files (Raw Current Information)

POD5 Files (for best basecaller performance)

Basecalled reads (At this step, should have POD5 file with signal data and a BAM with basecalls aligned to reference

Creating Windows of B and Non-B DNA

Extract Motif Positions from non-B DNA database

Fix windows of length X around motifs

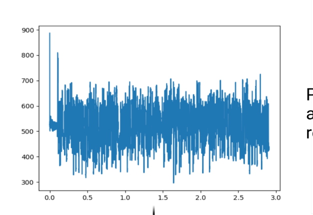
Extend the positions on the opposite strand

Find high quality reads (from BAM) that fall on the windows (non-B DNA windows)

Find motif-free regions (B-DNA regions)

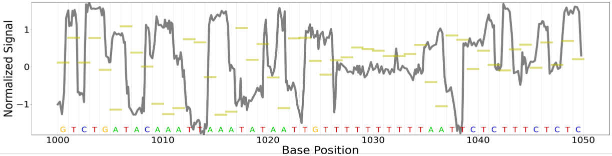
Compute Translocation Times on non-overlapping windows

Preprocessing Visualization

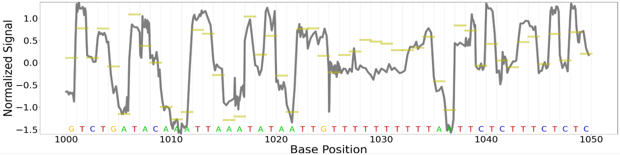


Plotting raw signal data against time for a single read

After basecalling and alignment



After signal mapping refinement


Normalized Signal
G T C T G A T A C A A A T T A A A T A T A A T T G T T T T T T T T T T A A T T C T C T T T C T C T C
Base Position