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

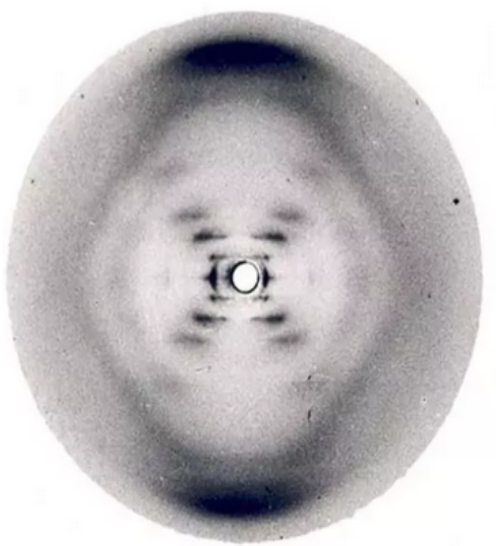
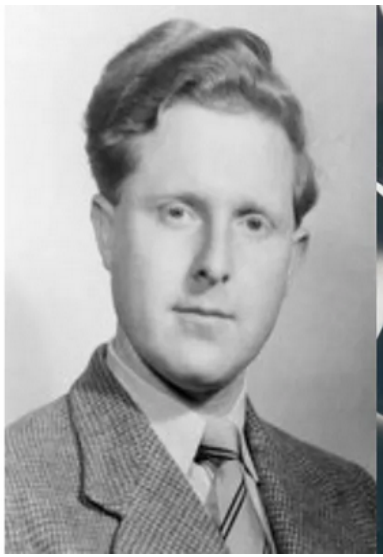
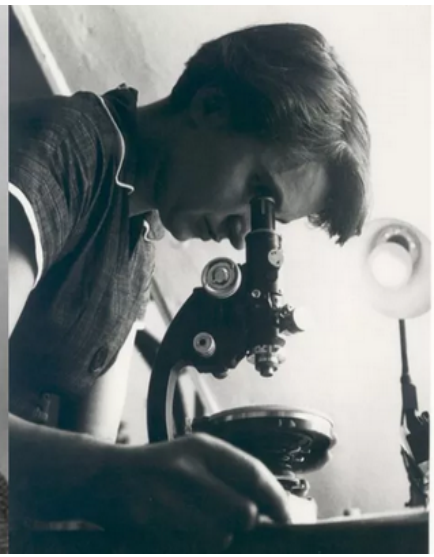
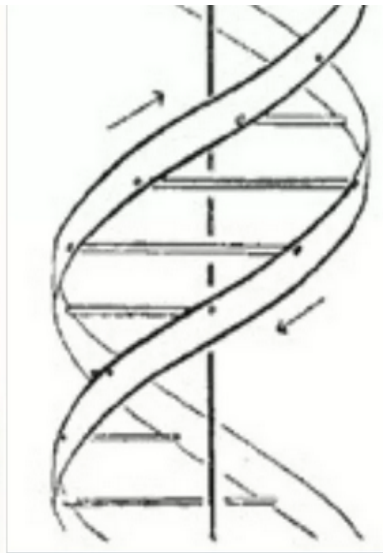
Alexander Turco

January 27, 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] Nijman, 2011

DNA Can Adopt Alternative Structures

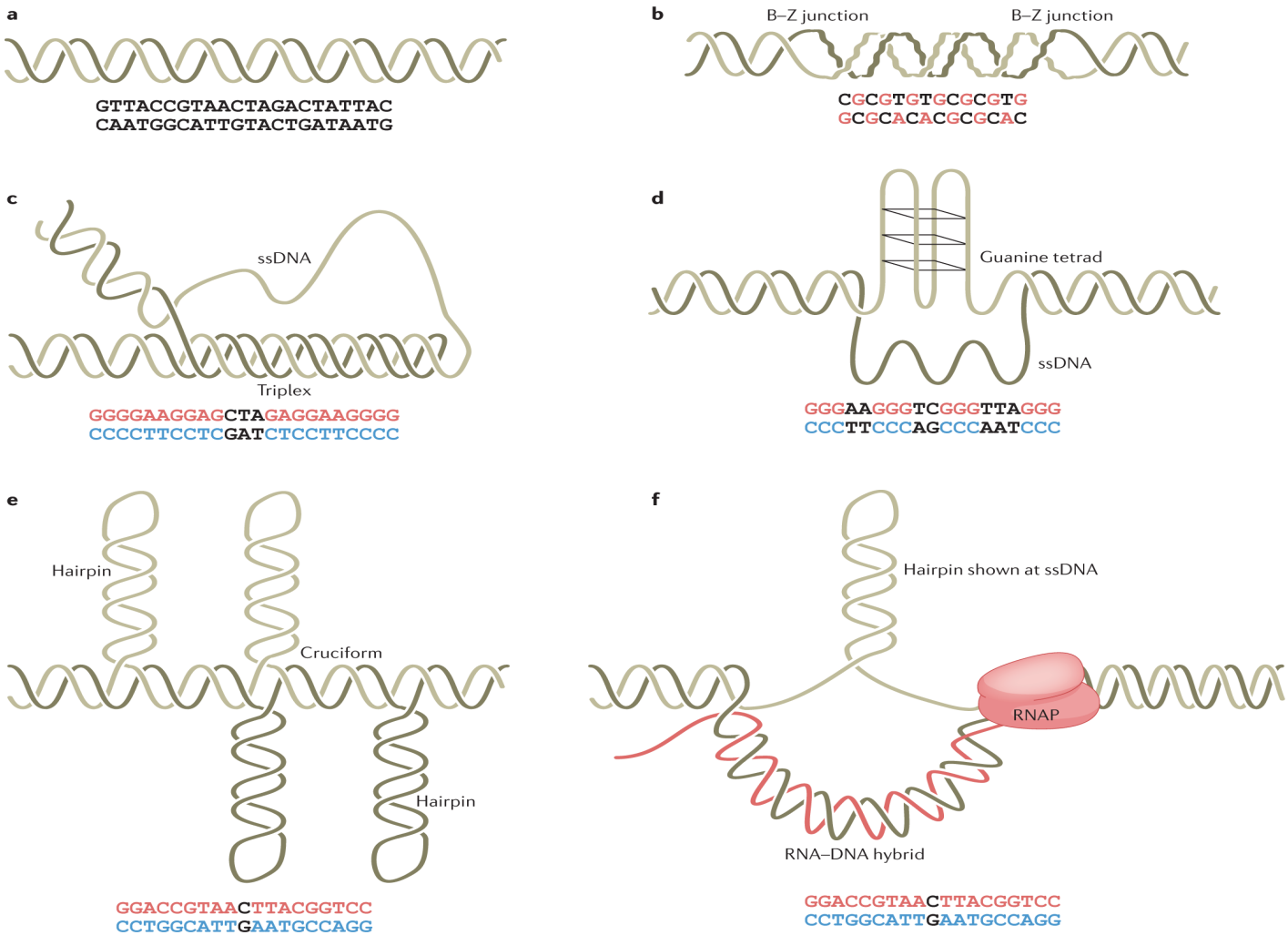
- Now, more than 15 types of DNA structure that differ from the canonical B-form have been reported (non-canonical or non-B form DNA)

- 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] Lee et al., 2018

Types of Non-canonical DNA structures



[3] Nijman, 2011; Shen and Ideker, 2018

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

[4] Figure From O'Connor, 2015

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

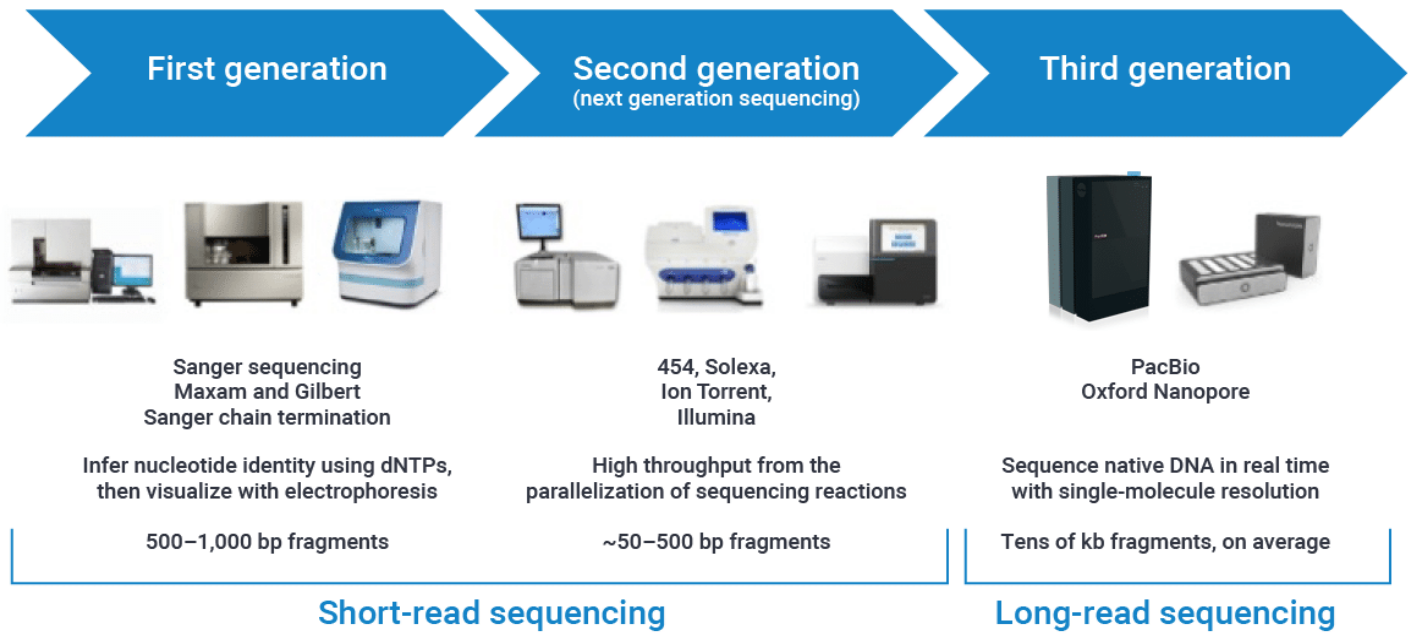
- circular dichroism spectra analysis
- Polymerase stop assays
- Immunoflourescence studies

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



Oxford Nanopore Sequencing Technology

ONT Sequencer

placeholder box

Inside the Nanopore

[5] Cheng et al., 2021

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

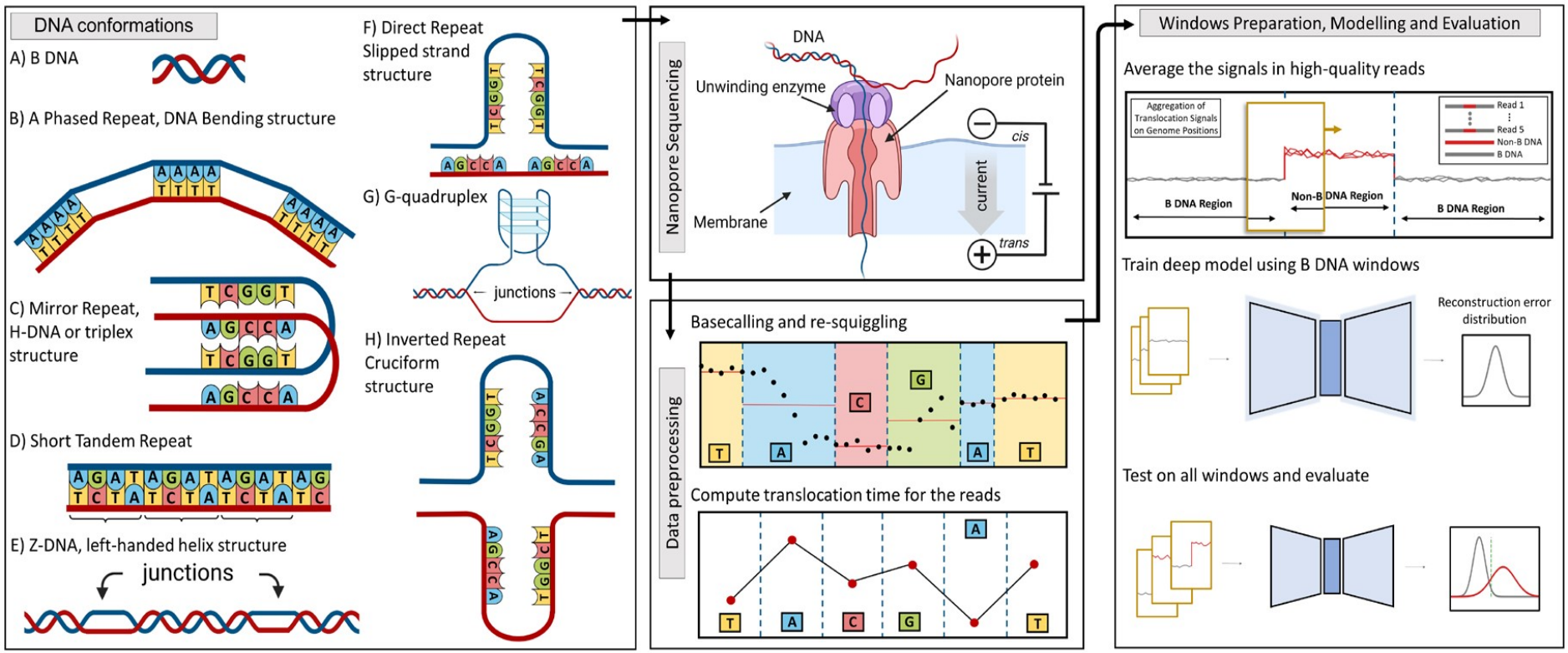
Benefits of unsupervised approach;

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] Lee et al., 2018

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



[2] Lee et al., 2018

Results of the Model

Pan-Cancer Analysis: Analysis of multiple types of cancer simultaneously, aims to identify commonalities shared across different types
Cancer-Specific Analysis: Analysis of individual cancer types to better understand the variation within a specific cancer type

- Pan-cancer analyses may overlook cancer-specific differences
- Availability of datasets for all cancer types varies, this may limit statistical power of pan-cancer analyses

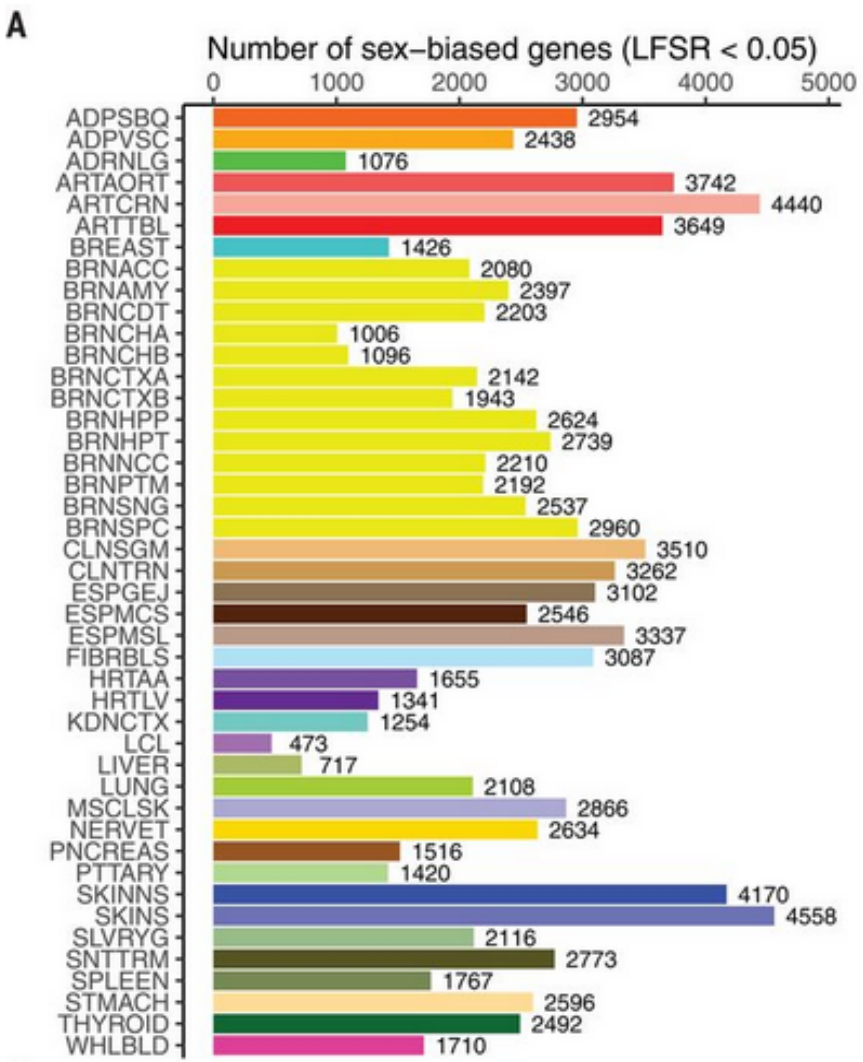
The Objective

Human sex differences are mainly caused by;

1. Gonadal hormone secretions
2. Genes located on the sex chromosomes (X and Y)

This leads to differences in the frequency of certain cancer types and the efficacy of treatments in males and females

Preprocessing Workflow



[6] Oliva et al., 2020