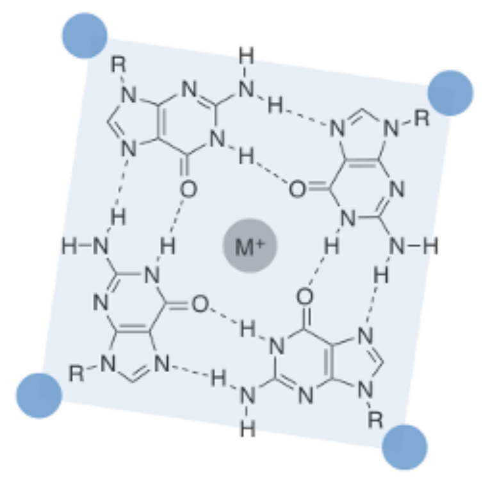
High-throughput sequencing of DNA G-quadruplex structures in the human genome

Vicki Chambers*, Giovanni Marsico*, Jonathan Boutell, Marco Di Antonio, Geoffrey Smith, Shankar Balasubramanian

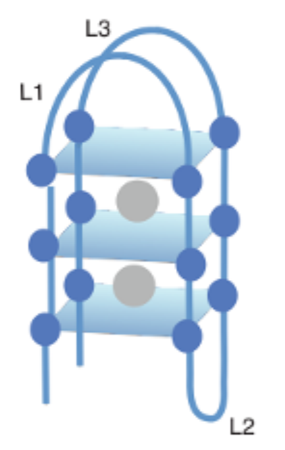
Department of Chemistry, University of Cambridge, Lensfield Road, Cambridge, CB2 1EW, UK
Illumina Cambridge Ltd., Chesterford Research Park, Little Chesterford, Essex, CB10 1XL, UK
vsc29@cam.ac.uk or sb10031@cam.ac.uk

Introduction

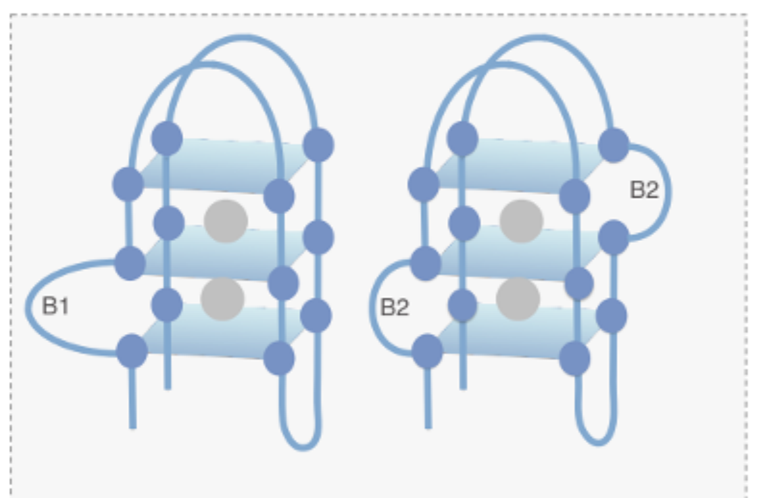
G-quadruplexes (G4s) are nucleic acid secondary structures that form within guanine-rich DNA or RNA sequences. Their formation can influence key biological processes, such as replication, translation and splicing. G4s have been associated with genomic instability, genetic diseases and cancer progression and so experimental evidence for their prevalence and formation in the human genome is essential. Therefore, it is important to develop a method to map these structures.



- G-tetrad



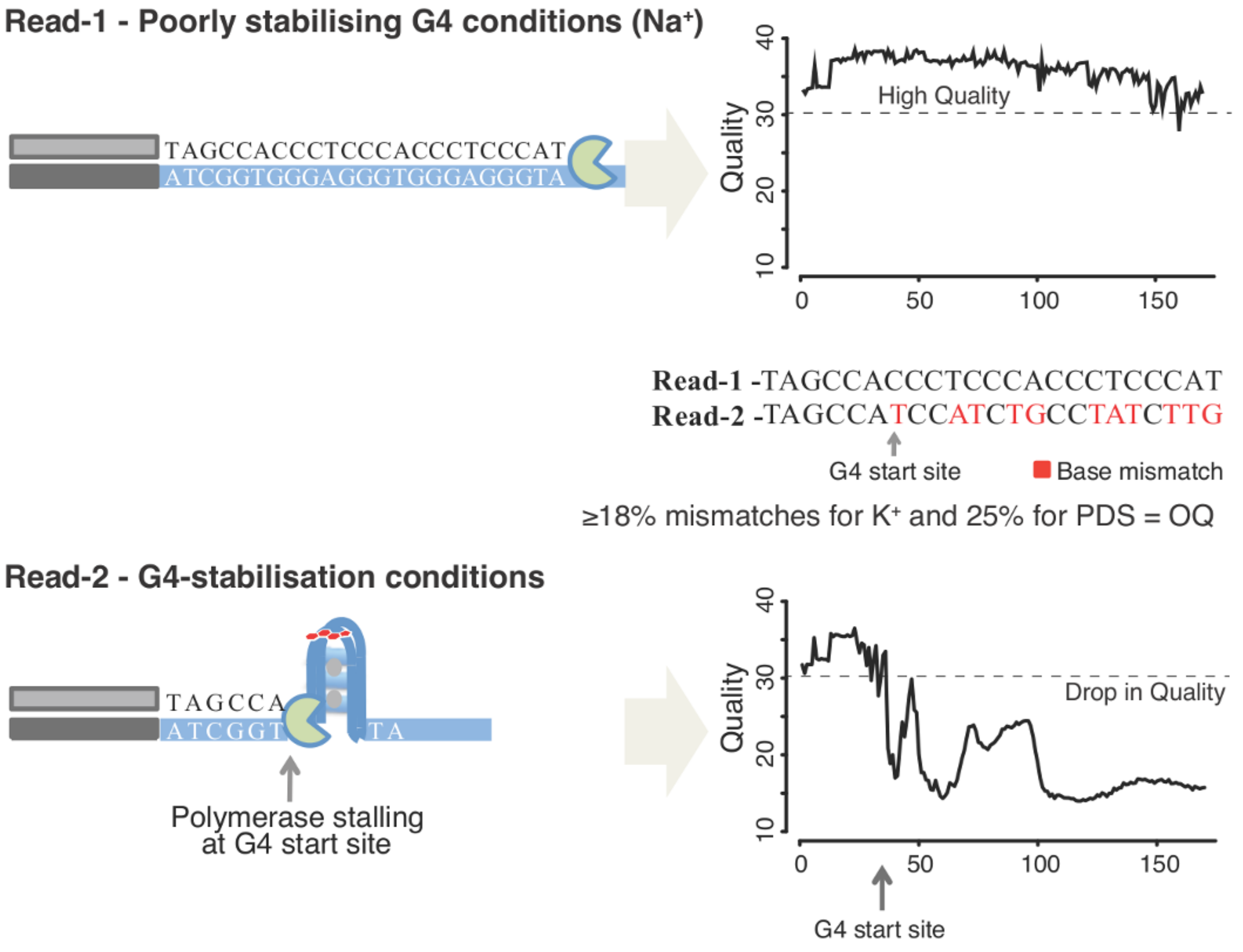
- Canonical G4



- Discovery of new G4 structures?

G4-Seq

We present G4-Seq, a method to detect G4 structures across the human genome.



Read-1 - Poorly stabilising G4 conditions (Na⁺)

Read-1 -TAGCCACCCTCCCACCCTCCCAT
Read-2 -TAGCCATCCATCTGCCTATCTTG

G4 start site 🟥 Base mismatch

≥18% mismatches for K⁺ and 25% for PDS = OQ

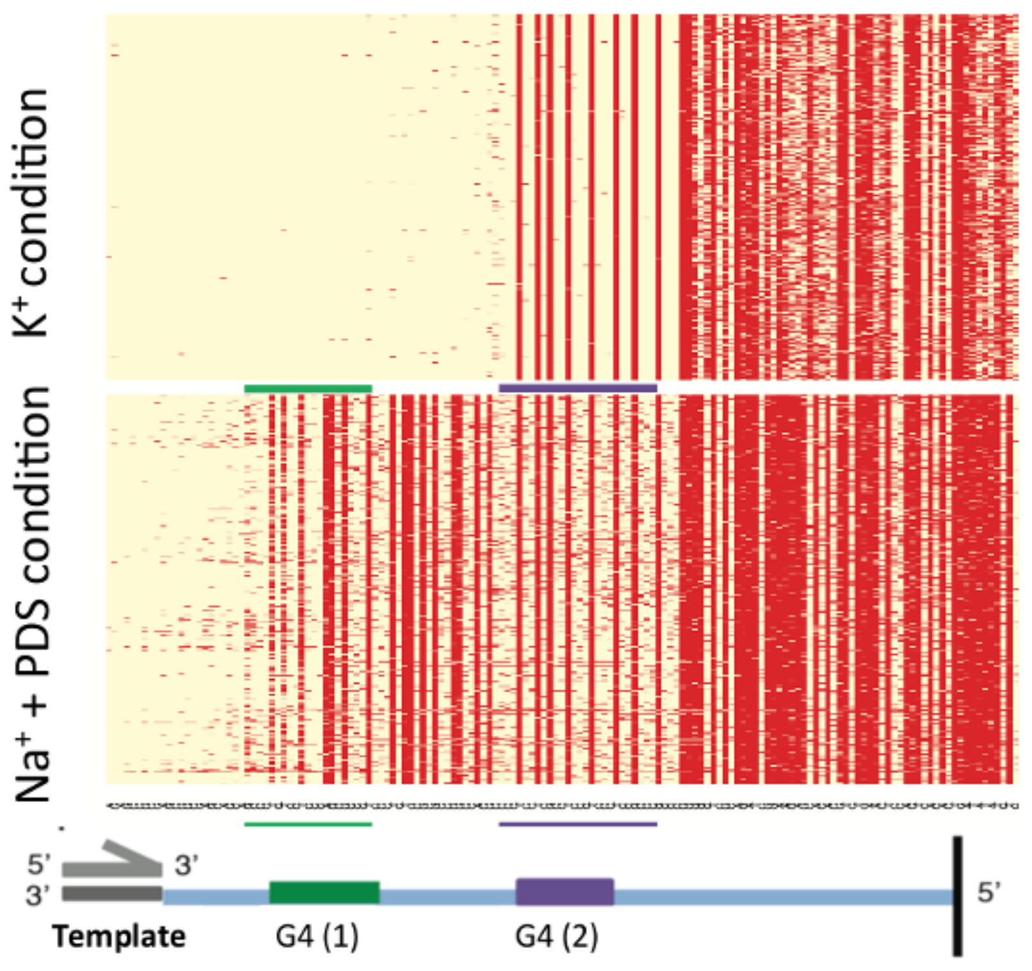
Read-2 - G4-stabilisation conditions

Polymerase stalling at G4 start site

G4 start site

Differences in sequencing quality and base mismatches between Read-1 and Read-2 are analysed to provide a map of G4 structures.

Condition dependent G4 mapping



- Inspection of a gene containing two G4 motifs, shows that base mismatches (shown in red) only accumulate after the G4 start site.
- Also, as G4 stability is increased by the use of the G4 ligand Pyridostatin (PDS), even more base mismatches are observed.

G4s in the human genome

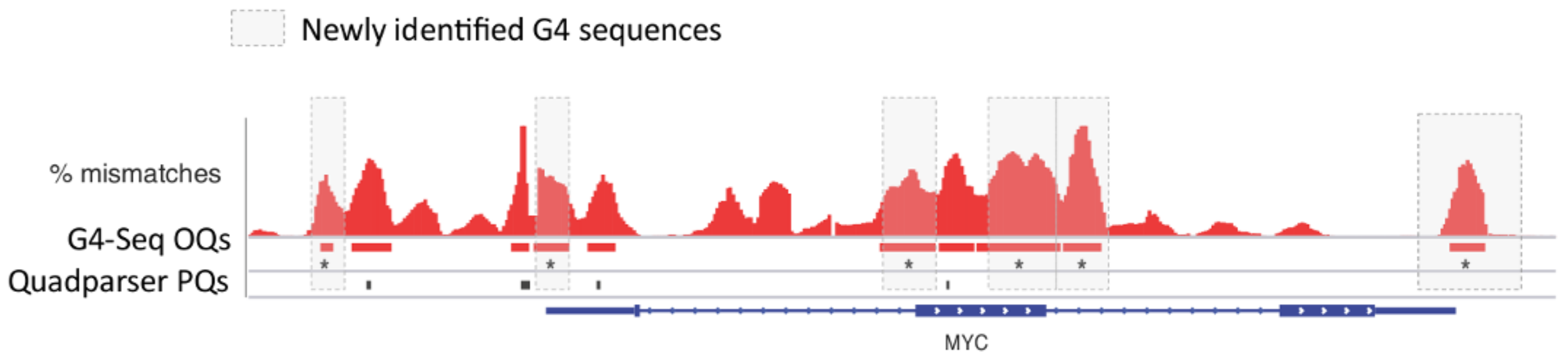
- *Computational predictions*
  2005: Quadparser predicted 361,424 sequences (PQs).
  2005: G4-calculator was developed
  2006: QGRS-Mapper was developed

- ***The first experimental map***
  2015: G4-Seq established

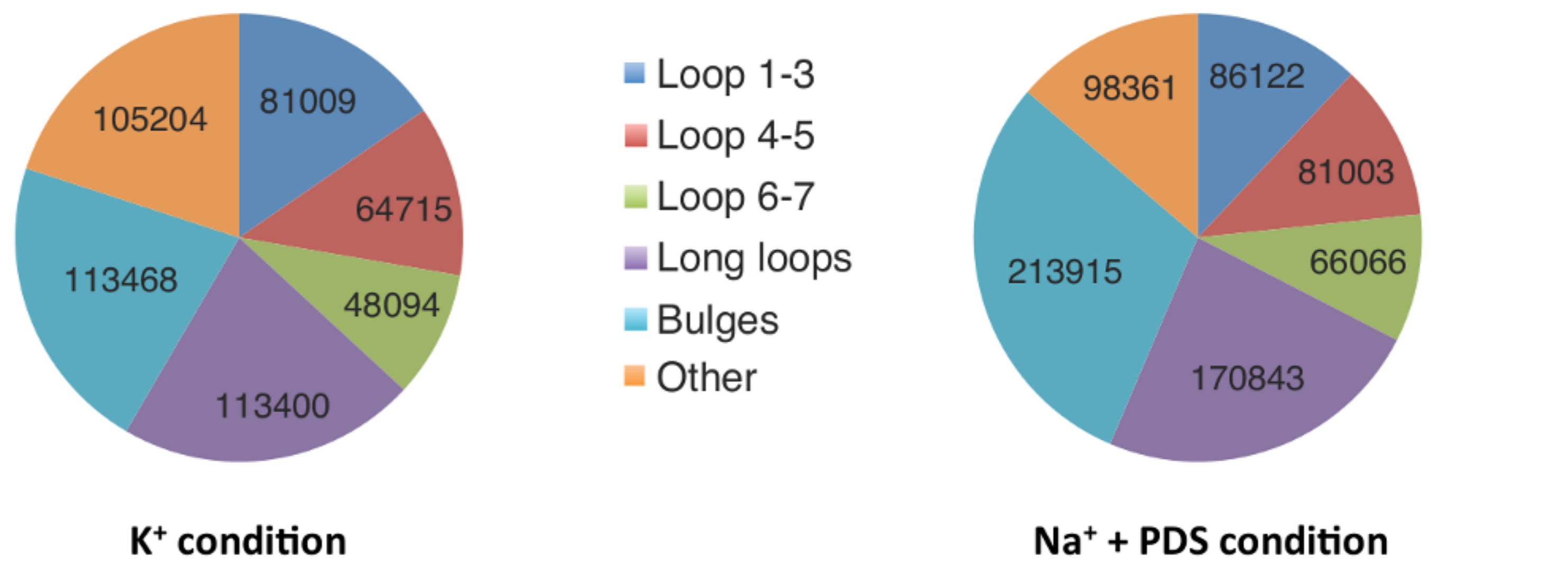
Observed G-quadruplex sequences (OQs)

Sequencing Condition	Number of OQs
K⁺	525,890
Na⁺ + PDS	716,310

An example of the MYC oncogene:

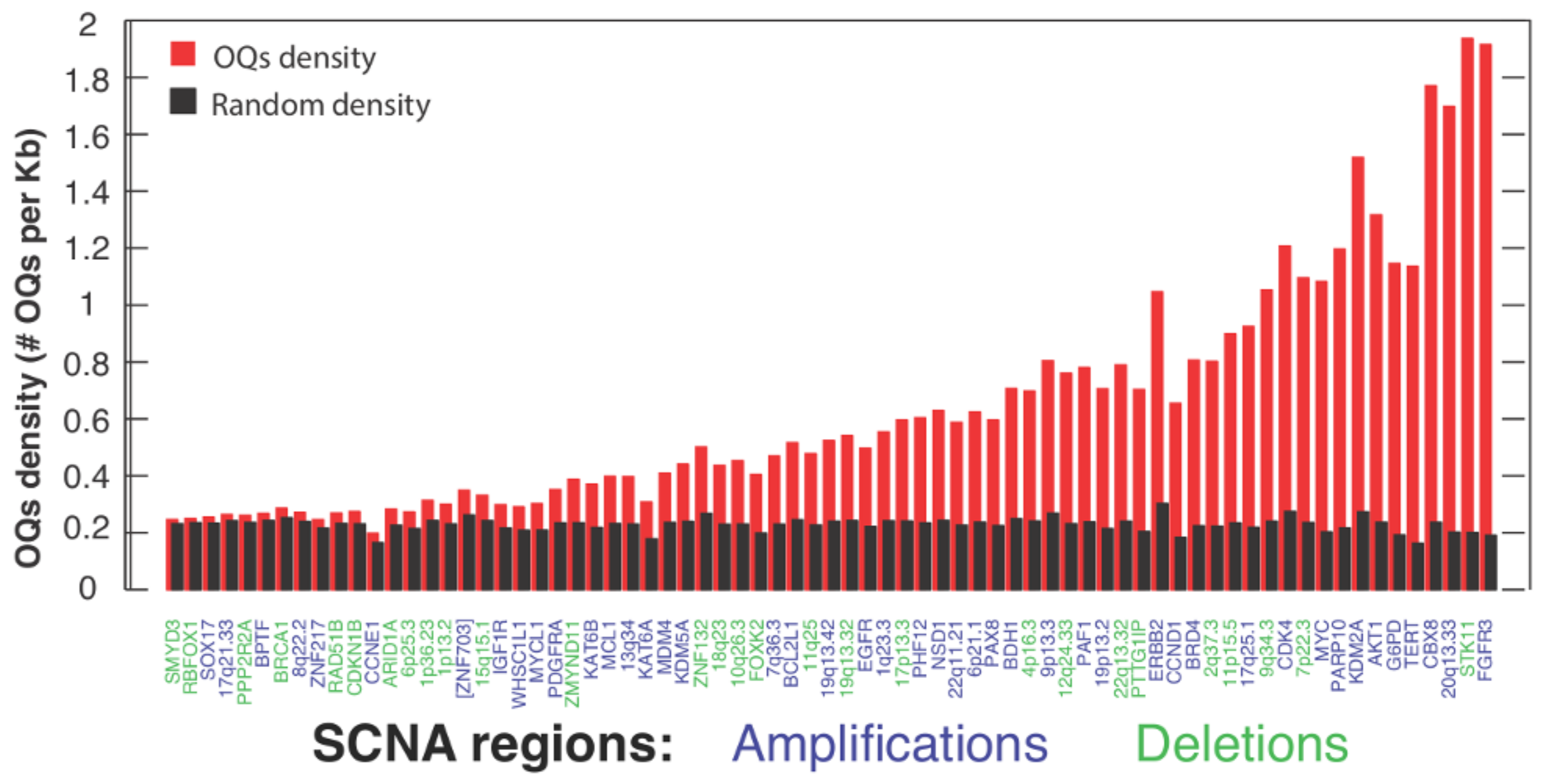


Nature and distribution of OQs



- Analysis of distinct structural features of the OQs, unraveled a dataset of stable G4s that were not easily identified *a priori* by computational approaches.
- G4s were found predominately in regulatory regions, especially in 5'UTRs and splicing sites.

Correlation of OQs with SCNAs



G4-Seq reveals a high OQs density in somatic copy number alterations (SCNAs) in cancer-related genes associated with amplifications (blue).

Summary

G4-Seq enables the genome-wide profiling of DNA G4 structures with high-resolution and provides insights into the nature of G4s, including non-canonical features such as longer loops and bulges that were previously not fully characterised. This method provides a resource of genomic targets for further biological and mechanistic studies. Our data suggests that G4s are strongly associated with SCNAs in cancer related genes, highlighting the potential of G4-targeting for therapeutic intervention. This universal method is applicable to the study of any genome and to the screening of other DNA-small-molecule interactions.

Acknowledgements:

This study is supported by BBSRC and Illumina. The S.B. Lab is funded by CRUK and ERC. We thank P. McCauley for preparation of sequencing buffers and Dr. Lowe and Dr. Tannahill for helpful discussion.
The data is available at the NCBI's GEO repository, accession number GSE63874 (http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE63874).
doi:10.1038/nbt.3295