Predicting G-quadruplex Formation

Jacob Calvert[1], Alex Kreig[1], Saurabh Sinha[2] and Sua Myong[1]

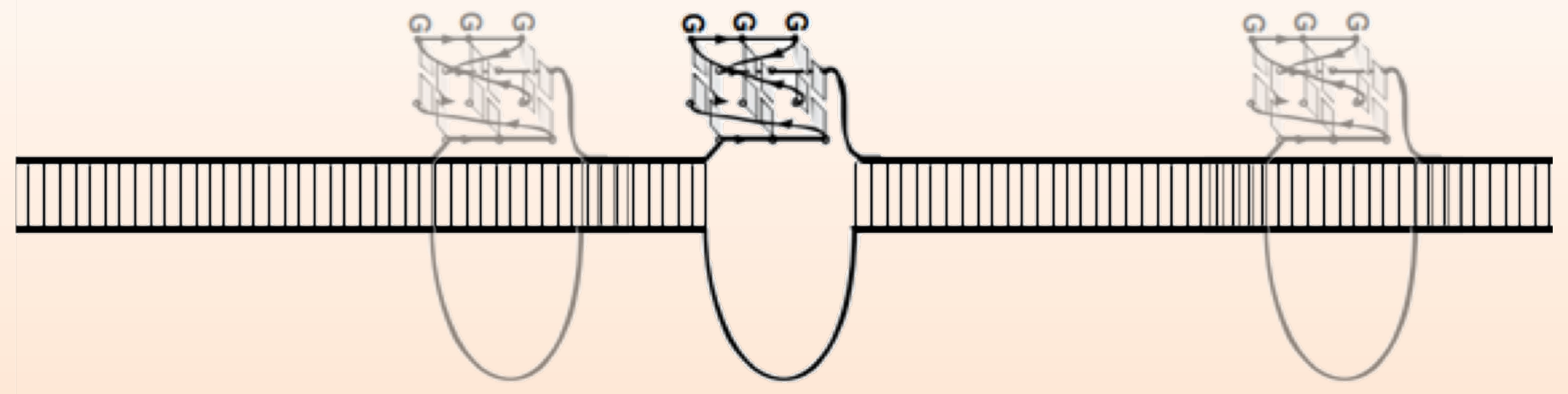
[1]Department of Bioengineering and [2]Department of Computer Science, College of Engineering, University of Illinois at Urbana-Champaign

Aims

- Model G-quadruplex folding
- Predict the folding of new sequences

Introduction

Q: What is a G-quadruplex (GQ)?
A: *A region of guanine-rich DNA that **can** fold into a hill-like structure (Fig. 1a,c)*



Q: Where are GQs found?
A: *In regulatory regions like gene promoters (Fig. 1d) and telomeres (Fig. 1a,b)*

Q: Why model GQs?
A: *To better our understanding of gene regulation and motivate new disease therapies*

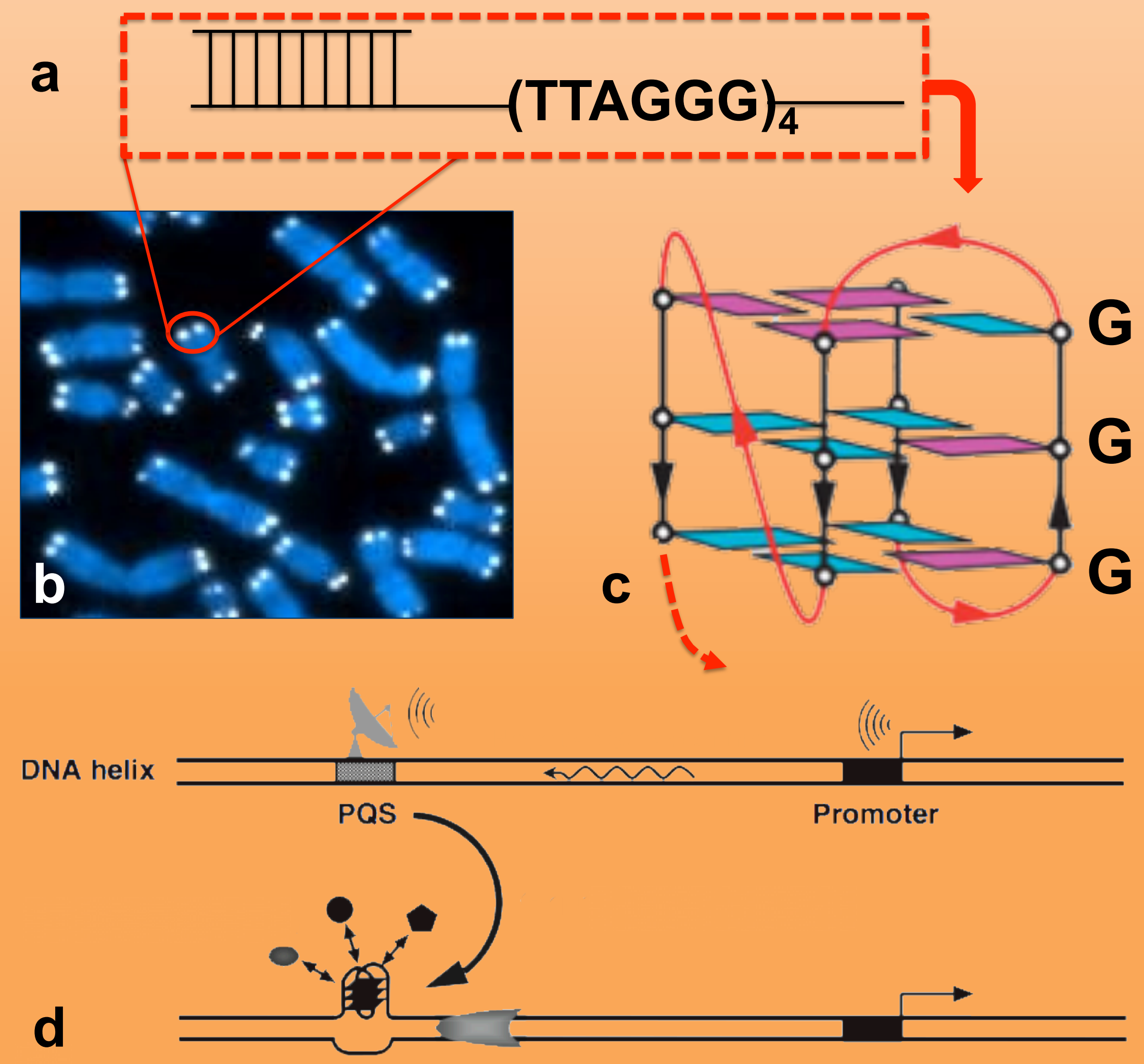


Figure 1. *(a) Telomere scheme. Adapted from Yildiz Lab. (b) Telomeres (white) cap the ends of chromosomes. (c) Parallel, planar interactions stabilize the GQ. Adapted from Phan AT et al. (2007). (d) Transcription-activated GQ formation in gene promoter. Adapted from Zhang C et al. (2013).*

Data

Q: Where did the data come from?
A: *"Pull-down" experiments detected folded GQs in human cells. We compared folded GQ sequences to GQ sequence motifs that did not fold (Fig. 2a).*

a G_3 N_{1-9} G_3 N_{1-9} G_3 N_{1-9} G_3

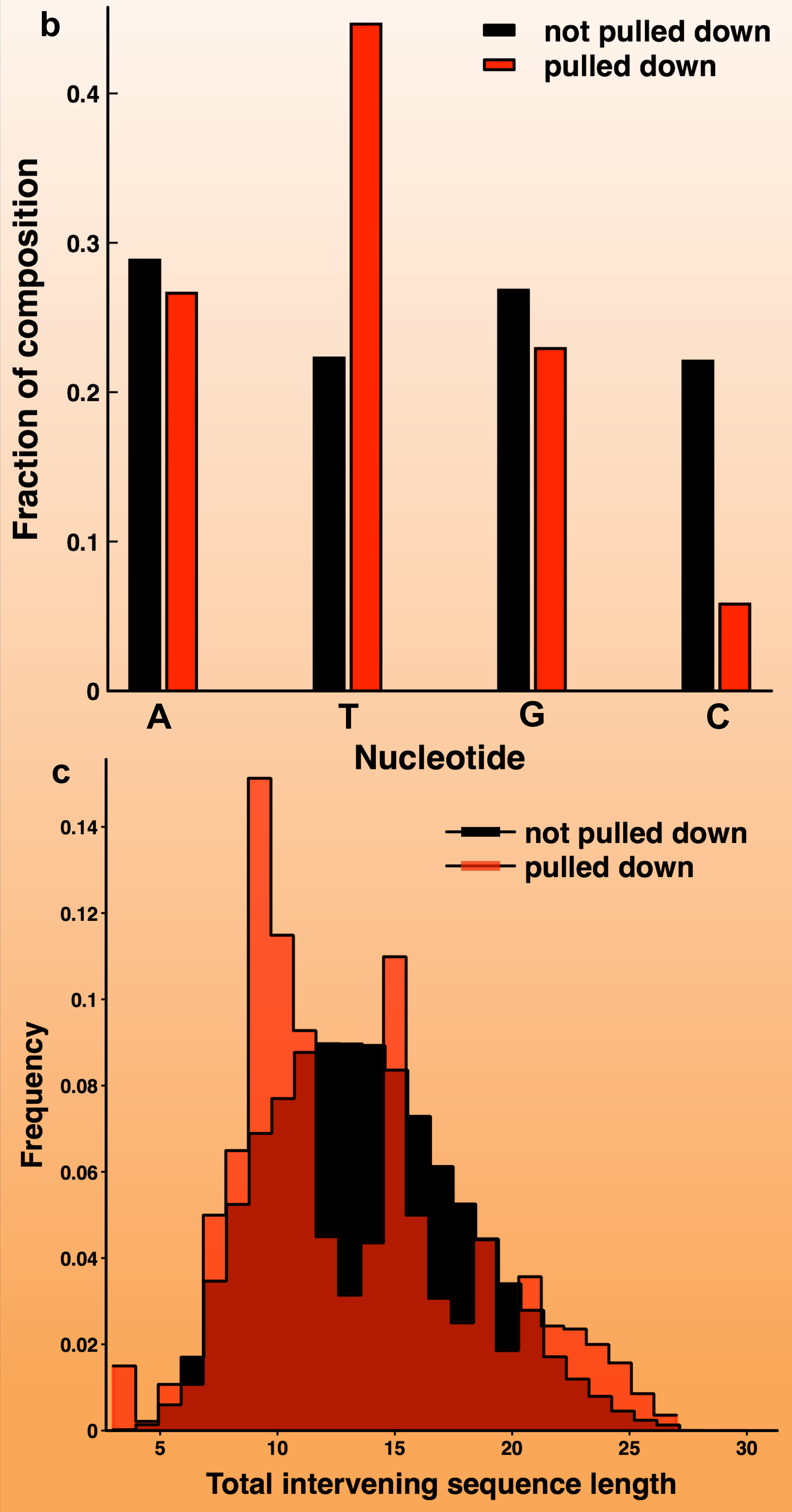


Figure 2. *(a) GQ sequence motif.* **N** *represents any base,* $\mathbf{N_{1-9}}$ *are called intervening sequences. (b) Comparison of base composition and (c) the total intervening length (sum of three* $\mathbf{N_{1-9}}$ *loops).*

Results

Q: How are the pull-down data used?
A: *We use a probabilistic model to detect the unique features of the pulled-down sequences. We translate the probability that a sequence folds into a score, which we call the "QPD Score."*

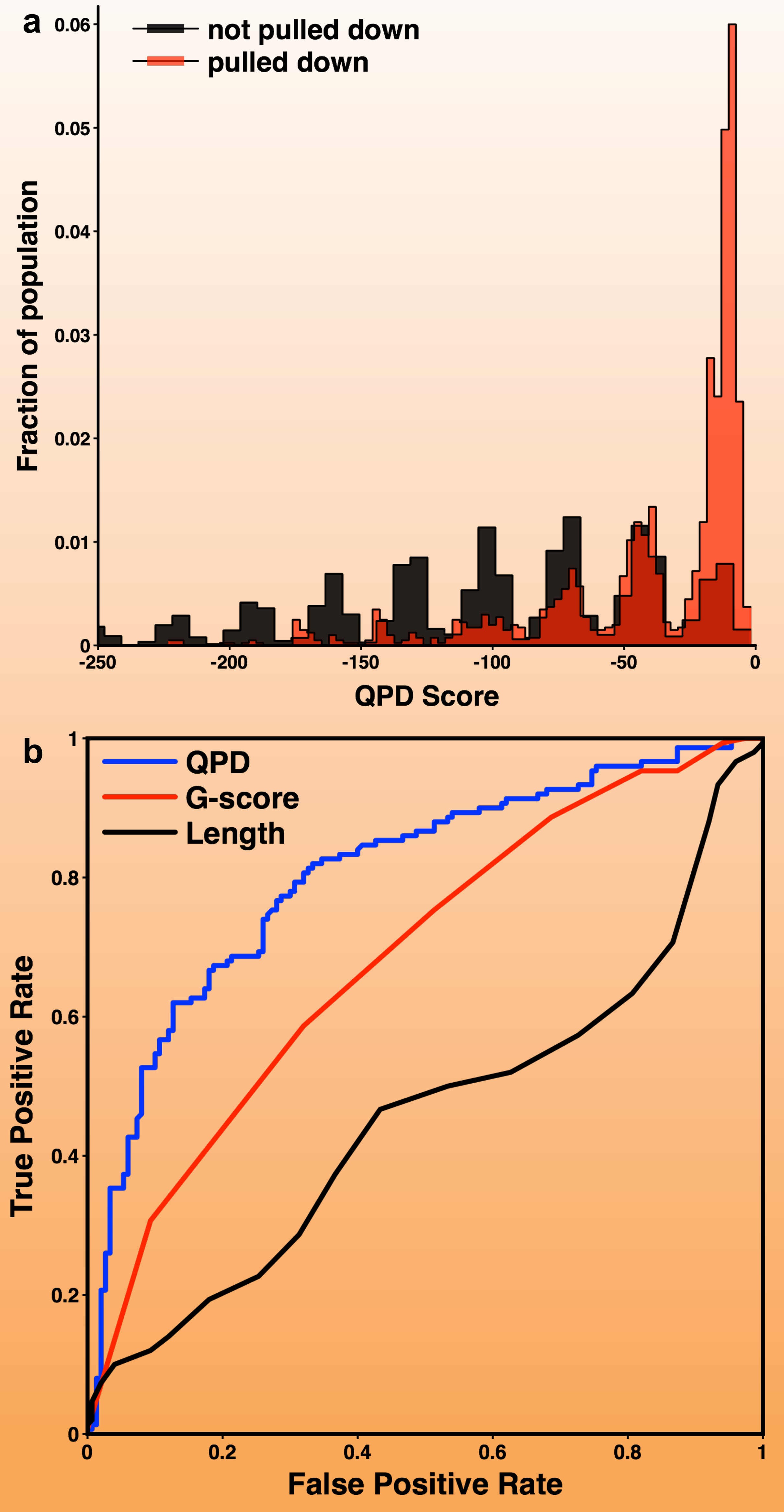


Figure 3. *(a) Distribution of QPD scores for all genomic GQs. (b) ROC curve comparison of QPD score against two existing methods for predicting GQ folding. Plot was generated with sequences not included in the training set.*

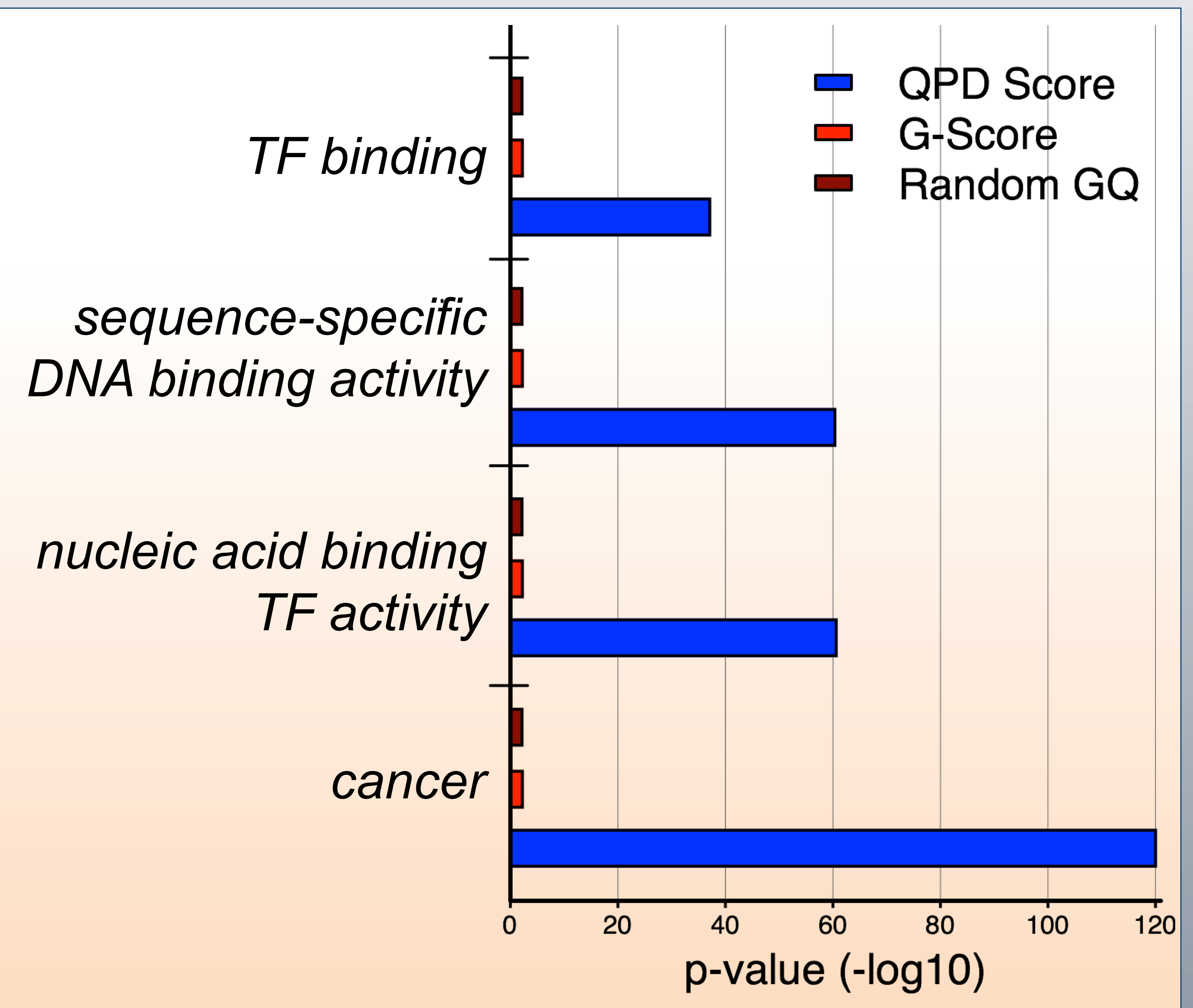


Figure 4. *Selected ontologies. High QPD-scoring GQs tend to localize near genes of importance to transcription factor (TF) activity, regulation, and cancer. Sequences chosen in other ways do not.*

Table 1. *A selection of genes whose promoters contain GQs predicted to fold by QPD. Sequences are listed with* $\mathbf{G_3}$ *omitted.*

Name	Sequence	G-Score Percentile	QPD Percentile
ABL2	--AAGGA--A--A--	54	98
RAB31	--T--A--GTAGA--	69	99
MSLN	--T--TGAA--GT--	69	99
PBX1	--AATA--GT--AGT--	82	97
BCL3 & ERG	--A--A--A--	95	99
WNT10A	--T--T--G--	95	99

Conclusions

- Our model outperforms existing methods of GQ folding prediction.
- Highly-scoring sequences localize near genes important in regulation and disease

Acknowledgments

We'd like to acknowledge our funding sources: the **NSF Center for the Physics of Living Cells** and the **NIH New Innovator Award**. We'd also like to thank Dr. Ramreddy Tipanna for helpful discussions.