ModTools: a computational toolbox for rapid detection of DNA modifications and replications using Nanopore sequencing

Qian Liu[1], Daniela Georgieva[2], Dieter Egli[2,3], Kai Wang[1,4]*

[1] *Raymond G. Perelman Center for Cellular and Molecular Therapeutics, Children's Hospital of Philadelphia, Philadelphia, PA 19104, USA*

[2] *Naomi Berrie Diabetes Center, Columbia University, New York NY 10032, USA*

[3] *Department of Pediatrics and Department of Obstetrics and Gynecology, Columbia University, New York, NY 10032, USA*

[4] *Department of Pathology and Laboratory Medicine, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104, USA*

Question? Please contact PI: Prof. Kai Wang via wangk@chop.edu (wglab.org)

1. Background:

- ***DNA base modifications***:
  - Play critical roles in various biological processes[1]
  - Base modifications have been found to contribute to disease genes in various human diseases
- ***Existing methods for detecting DNA modifications with short-read sequencing***
  - Bisulfate-conversion or antibody-based techniques
  - Have inherent limitations to detect DNA modifications[2]
- ***Feasibility of Nanopore sequencing in DNA modification detection***
  - NanoMod[3], nanoraw[4], Tombo[5], DeepMod[6], Nanopolish[1], and DeepSignal[7]
  - The performance and/or running speed needs improvement
- ***Detection of DNA replications***
  - DNA replication is a fundamental requirement for cell proliferation and DNA repair
  - No single method can identify the location/direction/speed of replication forks with high resolution
  - Now detect DNA base analogs via Nanopore sequencing (Replipore sequencing[8])
  - Computational methods are under-developed

2. Framework of ModTools:

ModTools contains three modules implemented by C++, and two of them are improved detection of DNA modifications, while the last module is to detect DNA replication origins.

- 1. NanoMod module:
  - Input: Nanopore signals between positive-control sample with modified bases and negative-control sample without any modification
  - Improvement: better base detection with rapid detection
  - Output: Genome-wide modification profile in a BED format

Advantage: without training process
Disadvantage: need unmodified samples

- 2. DeepMod module:
  - Input: Nanopore signals from a test sample
  - Improvement: better base detection with rapid detection
  - Output: Genome-wide modification profile for each position of interest

Advantage: only need modified samples for detection
Disadvantage: Train for a modification type

- 3. Detection module of DNA replications:
  - Input: Genome-wide modification profile or the output of the first 2 modules
  - Output: A list of detected DNA replication origins

3. Results:

- Modification detection via sample comparison
  - Process
    - ✓ 1. Annotate signals to basecalled sequence
    - ✓ 2. Anchor signals to reference sequence based on alignment
    - ✓ 3. Local realignment to improve signal anchoring
    - ✓ 4. Use Kolmogorov–Smirnov distance to determine signal difference from modified samples and unmodified samples
  - Datasets for evaluation
    - ✓ 3 modified samples and 1 unmodified samples
    - ✓ IdU/CIdU analogs at a know position (3072)
    - ✓ BrdU analogs for all T positions between 3046 and 3105
  - Evaluation
    - ✓ The modified positions are always ranked in top 1st as shown in Figure 1.

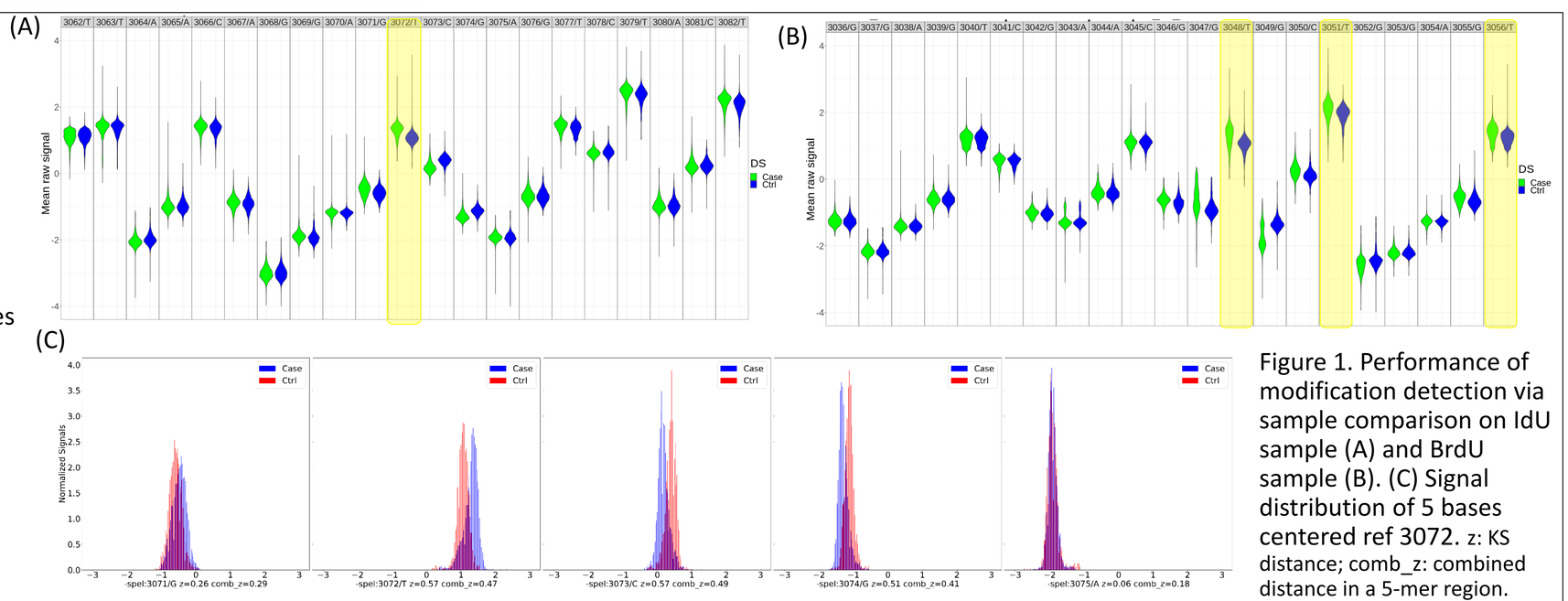

Figure 1. Performance of modification detection via sample comparison on IdU sample (A) and BrdU sample (B). (C) Signal distribution of 5 bases centered ref 3072. z: KS distance; comb_z: combined distance in a 5-mer region.

- Modification detection via deep learning
  - Process
    - ✓ Consider all thymidine in negative control to be unmodified and thymidine in positive to be modified
    - ✓ Train deep learning framework on negative control and 91% positive control in Table 1
    - ✓ Retrain the model after considering 5% long reads with lowest prediction modifications in 91% positive control as unmodified
  - Datasets for evaluation
    - ✓ Tested on 13 datasets in Table 1 with different BrdU percentages.
  - Evaluation:
    - ✓ ModTools is able to more accurately estimate modifications

			#Long reads	#Bases	Average Length	Map-read (%)	Map-base (%)	Predicted modification %	
								ModTools	DNAScent2
R9.5	Neg	*FAH14273*	45,856	458,323,198	9,994	90.60	99.67	0.30	1.30
	Pos	**FAH14319**	64,354	684,630,838	10,638	89.90	99.13	27.80	22.80
R9.4	Neg	*FAH58492*	75,461	945,361,460	12,527	99.71	99.90	2.21	0.50
	Pos: 91%	**FAH58548**	272,188	950,341,389	3,491	91.16	97.22	81.00	66.30
	Pos: 69%	FAH58543	182,295	2,761,235,612	15,147	99.42	99.93	86.50	69.60
	Pos: 38%	FAK06596	90,079	1,715,525,634	19,044	99.44	99.88	34.40	25.80
	Pos: 9%	FAK06634	113,550	2,273,975,351	20,026	99.61	99.95	9.30	6.10
DNAScent data	0%	*barcode08*	155,081	751,119,344	4,843	99.50	96.76	1.70	
	15%	*barcode09*	134,022	642,487,360	4,793	99.30	96.64	16.80	
	26%	*barcode10*	73,690	371,459,982	5,040	98.93	96.71	33.60	
	49%	*barcode11*	31,549	138,867,127	4,401	96.30	95.81	53.90	
	78%	*barcode12*	9,768	43,798,448	4,483	93.45	95.71	28.10	
	10%	*HU*	230,418	4,722,433,353	20,495	99.33	99.80	14.20	8.10
	18%	*cell-cycle*	735,060	10,929,872,403	14,869	99.28	99.09	20.35	11.20

Table 1. Modification detection on 7 datasets generated by Fork-seq[9] and 6 datasets produced by DNAScent[10], together with the statistics of each datasets.

R9.4/5 is for Nanopore flowcell versions.
Neg: a thymidine-only control without thymidine modification.
Percentage (%): the total BrdU contents determined by mass spectrometry.

- Replication detection
  - Given a genome-wide modification profile
    - ✓ 1. Get 75% quantile modification percentage for a fixed region size (100bp)
    - ✓ 2. Calculate Pearson correlation in 8kb region
    - ✓ 3. Call peaks based on Pearson correlation as shown in Figure 2.
  - Evaluated on 4 datasets as shown in Table 2.

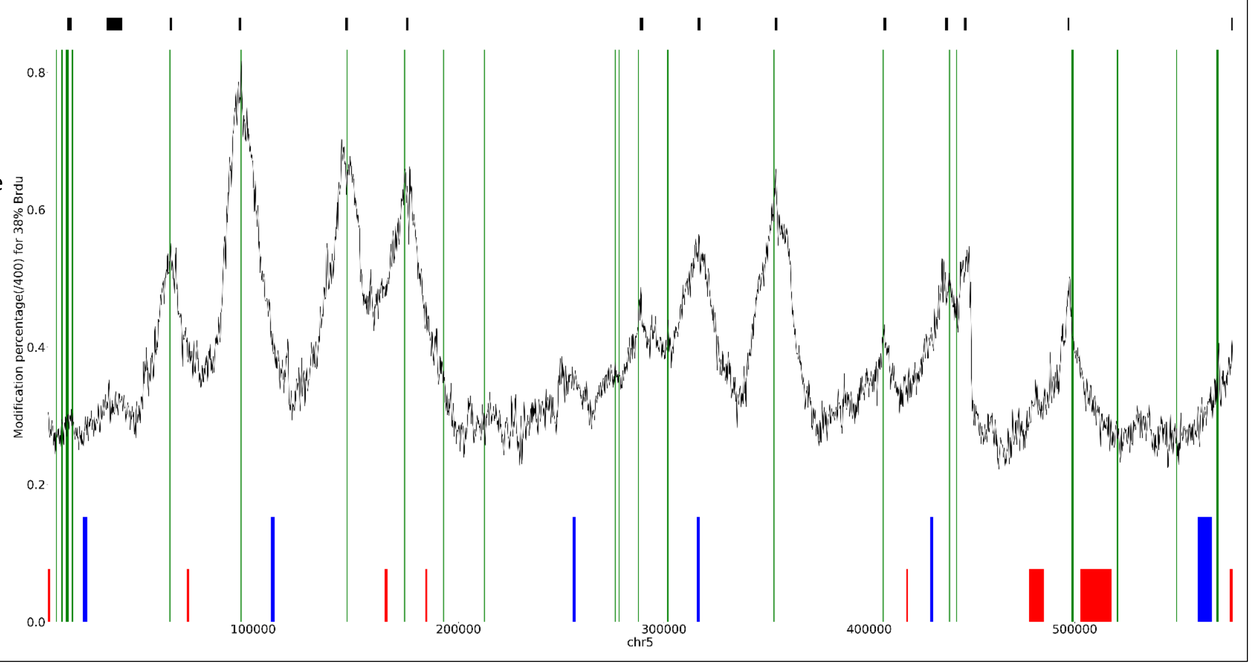

Figure 2. Performance of the detection of DNA replication origins for Chr5 of yeast.

X: refence position
Y: predicted modification percentage
Green rectangle: confirmed origins
Blue rectangle: Likely origins
Red rectangle: Dubious origins
Squares in black: predicted origins

ModTools	Dataset	Peak regions	Precision			
			Confirmed	Likely	Dubious	All
Fork-seq data	9% BrdU	179	0.76	0.145	0.112	0.894
	38% BrdU	206	0.641	0.155	0.117	0.816
DNAScent Data	Cell cycle	245	0.816	0.151	0.098	0.943
	HU	223	0.83	0.17	0.099	0.946

Confirmed/Likely/Dubious: Confirmed/Likely/Dubious Origins in OriDB
OriDB[11] origins: 410 confirmed, 216 likely and 203 dubious replication origins.

Table 2. Performance of the detection of DNA replication origins compared against OriDB.

4. Conclusion:

Our evaluation suggests that

- ModTools can detect both DNA modifications and replications.
- ModTools is faster and more accurate than existing tools.
- Has great potential to speed up genome-scale analysis of DNA modification and replications.

ModTools: https://github.com/WGLab/ModTools soon

Contact: liuq1@chop.edu; wangk@chop.edu

6. Acknowledgement:

NIH R21 grant (1R21HG010165-01A1) to Egli, Dieter Meinrad and Wang, Kai

Reference:

1. Simpson et al. *Nat. Methods* 14, 407-410 (2017)
2. Miura et al. *Nuceic Acids Res.* 40,e136 (2012)
3. Liu et al. *BMC Genomics* 20, 78 (2019)
4. Stoiber et al. *BioRxiv* (2017)
5. Stoiber et al. *Nanopore github*
6. Liu et al. *Nat. Commun* 10, 2449 (2019)
7. Ni et al. *Bioinformatics* 35(22):4586-4595 (2019)
8. Georgieva et al. *Nucleic Acids Res* 48(15):e88 (2020)
9. Hennion et al. *Genome Biology* 21:125 (2020
10. Muller et al. *Nat Methods* 16(5):429-436 (2019)
11. Nieduszynski et al. *Nucleic Acids Res* 35: D40-D46 (2006)