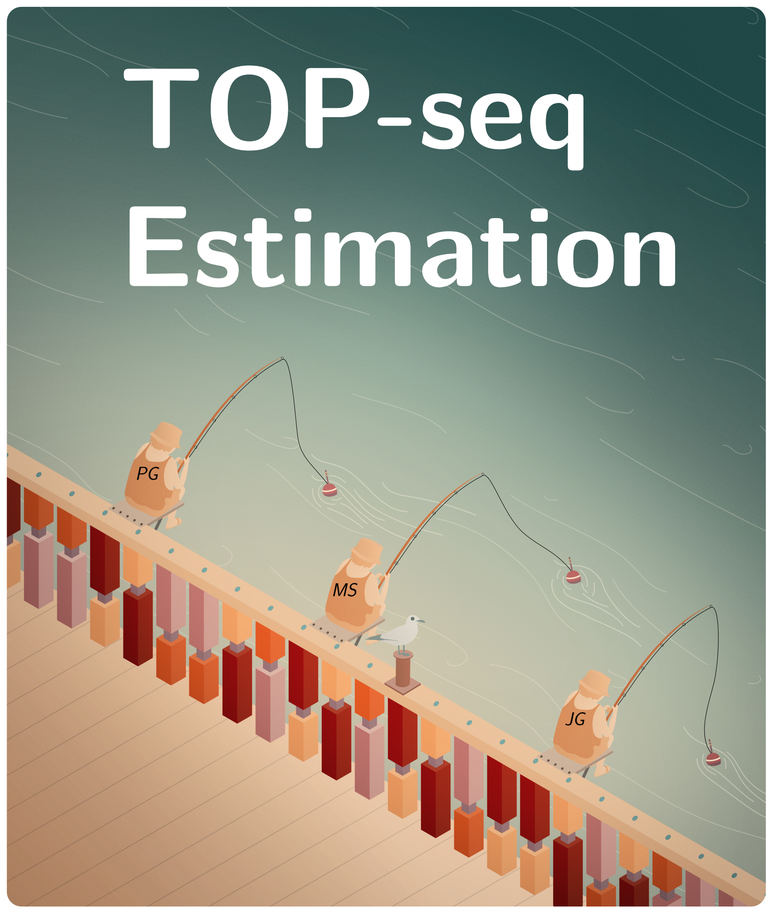
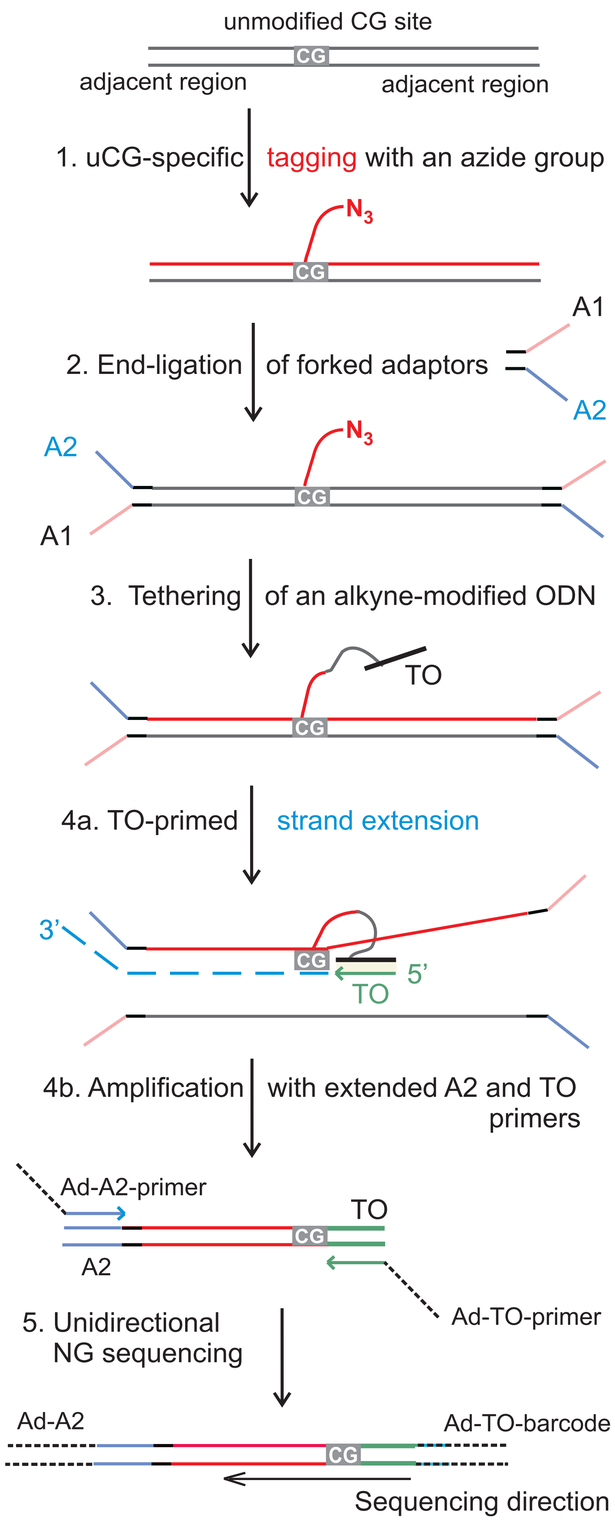


Povilas Gibas, Mantas Šarauskas and Juozas Gordevičius
Institute of Biotechnology, Life Sciences Centre,
Vilnius university, Lithuania

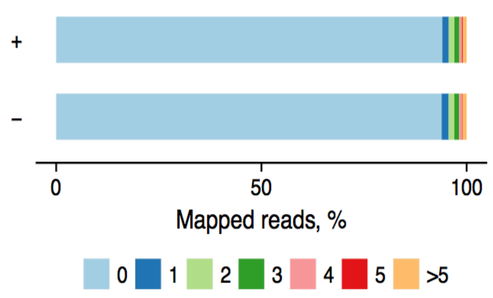
Estimation of DNA modification using artificial neural networks from TOP-seq data augmented with genomic context information

- DNA modifications have been implicated in aetiology of complex disease
- Epigenome-wide association studies are limited either by resolution or by the cost of existing technologies
- We leverage novel TOP-seq[1] unmethylome profiling technology that yields 3.9x coverage of 7.3M genomic CG sites with 57M reads
- Our computational strategy leads to r=0.77 genome-wide correlation of TOP-seq[1] with whole genome bisulfite seq (WGBS) and r=0.9 in gene regions

TOP-seq[1]: direct readout of individual unmethylated CG sites



98% of reads start within 2bp of CG



Method

- IMR90 cell line sequenced using TOP-seq[1] used for prediction
- IMR90 WGBS data (Lister et al., Ziller et al.) for training and evaluation
- A multi-layer perceptron regressor with 2 hidden layers (44 and 22 nodes respectively) was trained using stochastic gradient descent on chromosome 20 (2.5% CG sites in the genome)

Input features

- TOP-[1] coverage and u-density for each CG in the genome
- Genome element information (gene, CpG island, repeat etc. status)
- Sequence context information around each CG in 50bp window (CG%, GC%, C ratio, mappability, GERP score)
- Nucleotide sequence around CG in 3bp window

Results

- Correlation between WGBS and predicted values in protein coding genes (n = 18.5k):
  with WGBS 1 (used for training) Pearson r = 0.9,
  with WGBS 2 (used for evaluation) Pearson r = 0.8
- Applying regressor model on CpG islands with simulated differential coverage (observed average coverage was scaled by a given factor) yields a corresponding change in predicted values.
- Correlation between WGBS 1 and predicted values in different chromosomes. For a subset of random CG (20K) scatter plots between the two measurements are shown on the left.

[1] Staševskij et al. Tethered Oligonucleotide-Primed Sequencing, TOP-Seq: A High-Resolution Economical Approach for DNA Epigenome Profiling. *Mol Cell*, 65:554–564, 2017.

[2] Lister et al. Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature*, 462:313–322, 2009.

[3] Ziller et al. Charting a dynamic DNA methylation landscape of the human genome. *Nature*, 500:477–481, 2013.