DiNAMO: Exact method for degenerate IUPAC motifs discovery, characterization of sequence-specific errors

Chadi Saad, Laurent Noé, Hugues Richard, Julie Leclerc, Marie-Pierre Buisine, Helene Touzet, Martin Figeac

DiNAMO: Exact method for degenerate IUPAC motifs discovery, characterization of sequence-specific errors

Chadi SAAD[1,2], Laurent NOÉ[2], Hugues RICHARD[3], Julie LECLERC[1], Marie-Pierre BUISINE[1], Hélène TOUZET[2] and Martin FIGEAC[4]

[1] JPARC (UMR1172 Inserm, Lille 2 University and Lille University Hospital), 59000, Lille, France

[2] CRIStAL (UMR CNRS 9189 Lille 1 University and Inria Lille), Team BONSAI, 59000, Lille, France

[3] LCQB (UMR 7238 CNRS Pierre Marie Curie University), 75006, Paris, France

[4] Functional and Structural Genomic Platform, Lille 2 University, 59000, Lille, France

Corresponding author: Chadi.Saad@univ-lille1.fr

Next generation sequencing technologies are still associated with relatively high error rates, about 1%, which correspond to thousands of errors in the scale of a complete genome. Each region needs therefore to be sequenced several times and variants are usually filtered based on depth criteria. The significant number of artifacts, in spite of those filters, shows the limit of conventional approaches and indicates that some sequencing artifacts are recurrent. This recurrence underlines that sequencing errors can depend on the upstream nucleotide sequence context. Our goal is to search for overrepresented motifs that tend to induce sequencing errors.

Previous studies showed that some motifs, such as GGT [1,2], induce sequencing errors in the Illumina technologies. However, these studies were dedicated to exact motifs, and did not take into account approximate motifs, limiting the statistical power of such approaches. On the other hand, some tools, such as FIRE [3], DREME [4] and Discrover [5], were developed to search for degenerate motifs over the 15-letter IUPAC alphabet in the context of chip-seq studies. However, these tools use greedy algorithms, implying a lack of sensitivity. So we developed an exact algorithm to search for degenerate motifs by enumerating all possible IUPAC motifs. This algorithm is based on mutual information and uses hashtables with graphs data structure to store the motifs. It is independent from the sequencing technology.

Experimental results on real data show that there are many overrepresented motifs upstream of sequencing artifacts. These latter are identified through the strand bias between forward and reverse reads. The homopolymer of length 3 CCC seems to be sufficient to induce errors on IonTorrent. On Illumina, motifs are mainly composed of GGC followed by GGT (like: TGGCNGGT) or homopolymers. We have also noticed a base quality fall after the detected motifs. Our exact algorithm requires less than one minute (Intel® Core™ i5-4570 CPU, 3.20GHz), and less than 2GB of RAM to search for full degenerate motifs of length 6 on a dataset of approximately 24000 sequences, extracted from 11 exomes sequenced on IonTorrent Proton.

Availability: https://github.com/bonsai-team/DiNAMO

Acknowledgements

This work is supported by *Lille University Hospital* and *Hauts-de-France region*

References

[1] Manuel Allhoff, Alexander Schönhuth, Marcel Martin, Ivan G Costa, Sven Rahmann, and Tobias Marschall. Discovering motifs that induce sequencing errors. In *BMC bioinformatics*, volume 14, page S1. BioMed Central Ltd, 2013.

[2] Frazer Meacham, Dario Boffelli, Joseph Dhahbi, David IK Martin, Meromit Singer, and Lior Pachter. Identification and correction of systematic error in high-throughput sequence data. *BMC bioinformatics*, 12(1):451, 2011.

[3] Olivier Elemento, Noam Slonim, and Saeed Tavazoie. A universal framework for regulatory element discovery across all genomes and data types. *Molecular cell*, 28(2):337–350, 2007.

[4] Timothy L Bailey. Dreme: motif discovery in transcription factor chip-seq data. *Bioinformatics*, 27(12):1653–1659, 2011.

[5] Jonas Maaskola and Nikolaus Rajewsky. Binding site discovery from nucleic acid sequences by discriminative learning of hidden markov models. *Nucleic acids research*, 42(21):12995–13011, 2014.