BITS :: Call for Abstracts 2021 - Oral communication

Type	Oral communication
Session	Algorithms for Bioinformatics
Title	GRAFIMO: variant and haplotype aware motif scanning on pangenome graphs
All Authors	Tognon M(1), Bonnici V(1), Garrison E(2), Giugno R(1), Pinello L(3,4,5).

Affiliation

(1) Computer Science Department, University of Verona, Italy
(2) University of Tennessee Health Science Center, Memphis, TN, USA
(3) Molecular Pathology Unit, Center for Computational and Integrative Biology and Center for Cancer Research, Massachusetts General Hospital Charlestown, MA, USA
(4) Department of Pathology, Harvard Medical School, Boston, MA, USA
(5) Broad Institute of MIT and Harvard, Cambridge, MA, USA

Motivation

Transcription factors are key regulatory proteins that promotes or reduce the expression of genes by binding short (7-20 bp) DNA sequences known as transcription factors binding sites (TFBS) [1]. TFBS can be summarized using Position Weighted Matrices (PWMs) [2], which encode the probability of observing a given nucleotide in a given position of a binding site. Recently, several studies showed that mutations occurring in regulatory motifs can enhance, weaken or even create new binding sites [3—6]. Moreover, mutations altering TFBS can occur in haplotypes conserved within a population or even private to a single individual. While several tools have been developed to scan for potential motif occurrences on reference genome sequences [7, 8], no tool exists to find them in pangenome variation graphs (VGs) [9]. VGs are sequence-labelled graphs that can efficiently encode collections of genomes and their genetic variants in a single efficient data structure. Because VGs can losslessly compress large genomes from large panels of individuals, TFBS scanning in VGs can efficiently capture how genomic variation affects the potential binding landscape of TFs in a population of individuals. Here we present GRAFIMO (GRAph-based Finding of Individual Motif Occurrences), a command-line tool for the scanning of known TF DNA motifs represented as Position Weight Matrices (PWMs) in VGs. GRAFIMO extends the standard PWM scanning procedure by offering a variant- and haplotype-aware search for TFBS in a VG.

Methods

GRAFIMO is a command line tools which enables a variant- and haplotype-aware search of TFBS motifs within a population of individuals encoded in VGs. GRAFIMO offers two main functionalities: construction of VGs from user data and the search of one or more motifs in precomputed VGs. Briefly, given a TF binding site model (PWM) and a set of genomic regions, GRAFIMO leverages the VG to efficiently scan and report all potential motif occurrences with their frequencies in the haplotypes encoded in the VG, in a single pass. Moreover, GRAFIMO provides the predicted changes in binding affinity mediated by genetic variants encoded in the scanned VG. We have designed the interface of GRAFIMO based on FIMO [7], so it can be used as in-drop replacement for tools built on top of FIMO. As in FIMO, the results are available in three files: a tab-delimited file (TSV), a HTML report and a GFF3 file.

Results

We tested GRAFIMO by searching CTCF, ATF3 and GATA1 potential motif occurrences on a VG based encoding 2548 individuals from 1000 Genomes Project phase 3 [10, 11]. Each TFBS was searched in ChIP-seq peak regions retrieved from the ENCODE Project data portal [12]. For our downstream analysis we considered as potential motif occurrences those candidates with P-value < 1e-4. Based on the recovered sites, we consistently observed across the 3 studied TFs that genetic variants can significantly affect estimated binding affinity. We observed that thousands of potential CTCF motif occurrences are found only in non-reference haplotypes, suggesting that a considerable number of TFBS candidates are lost when scanning genomes without accounting for variants. Interestingly, we observed that several potential CTCF motif occurrences found in rare haplotypes have a high statistical significance. By considering the genomic locations of the significant motif occurrences we next investigated how often individual TFBS may be disrupted, created or modulated. We observed that 6.13% of potential CTCF TFBS can be found only in non-reference haplotypes, 5.94% are disrupted genetic variation in non-reference haplotypes, while ~30% are still significant in non-reference haplotypes but with different binding affinity scores. Similar results were observed for ATF3 and GATA1.

Info

References

[1] Stewart AJ et al. Genetics 2012;192(3): 973–985.

[2] Stormo GD. Quantitative Biology. 2013; 1(2): 115–130

[3] Wienert B et al. Nature communications. 2015;6(1): 1-8.

[4] Weinhold N et al. Nature genetics. 2014;46(11): 1160-1165.

[5] De Gobbi M et al. Science. 2006;312(5777): 1215-1217.

[6] Guo YA et al. Nature communications. 2018;9(1): 1–14.

[7] Grant CE et al. Bioinformatics. 2011;27(7): 1017–1018.

[8] Kohronen J et al. Bioinformatics. 2009;25(23):3181–3182.

[9] Garrison E et al. Nature biotechnology. 2018;36(9): 875–879.

[10] 1000 Genomes Project Consortium. Nature. 2015;526(7571): 68–74.

[11] Zheng-Bradley X et al. GigaScience. 2017;6(7): 1-8.

[12] ENCODE Project Consortium et al. Nature. 2012;489(7414): 57—74.

Figure	
Availability	https://www.biorxiv.org/content/10.1101/2021.02.04.429752v1

Corresponding Author

Name, Surname	Manuel, Tognon
Email	manuel.tognon@univr.it

Submitted on	27.04.2021