GRAFIMO: variant and haplotype aware motif scanning on pangenome graphs

Manuel Tognon[1], Vincenzo Bonnici[1], Erik Garrison[2], Rosalba Giugno[1,*], Luca Pinello[3,4,5,*]

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

* rosalba.giugno@univr.it or lpinello@mgh.harvard.edu

Abstract

Transcription factors (TFs) are proteins that promote or reduce the expression of genes by binding short genomic DNA sequences known as transcription factor binding sites (TFBS). While several tools have been developed to scan for potential occurrences of TFBS in linear DNA sequences or reference genomes, no tool exists to find them in pangenome variation graphs (VGs). VGs are sequence-labelled graphs that can efficiently encode collections of genomes and their variants in a single, compact data structure. Because VGs can losslessly compress large pangenomes, 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 considering variations and alternative haplotypes encoded in a VG. Using GRAFIMO on a VG based on individuals from the 1000 Genomes project we recover several potential binding sites that are enhanced, weakened or missed when scanning only the reference genome, and which could constitute individual-specific binding events. GRAFIMO is available as an open-source tool, under the MIT license, at https://github.com/pinellolab/GRAFIMO and https://github.com/InfOmics/GRAFIMO.

Author summary

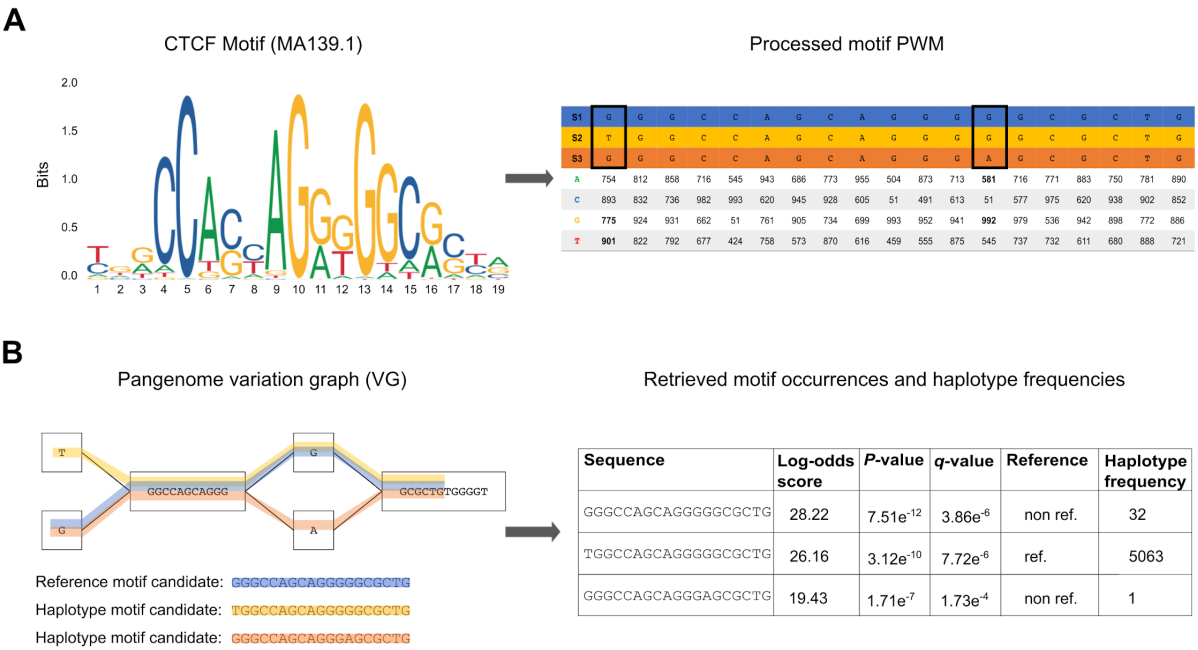
Transcription factors (TFs) are key regulatory proteins and mutations occurring in their binding sites can alter the normal transcriptional landscape of a cell and lead to disease states. Pangenome variation graphs (VGs) efficiently encode genomes from a population of individuals and their genetic variations. GRAFIMO is an open-source tool that extends the traditional PWM scanning procedure to VGs. By scanning for potential TBFS in VGs, GRAFIMO can simultaneously search thousands of genomes while accounting for SNPs, indels, and structural variants. GRAFIMO reports motif occurrences, their statistical significance, frequency, and location within the reference or alternative haplotypes in a given VG. GRAFIMO makes it possible to study how genetic variation affects the binding landscape of known TFs within a population of individuals.

Introduction

Transcription factors (TFs) are fundamental proteins that regulate transcriptional processes. They bind short (7-20bp) genomic DNA sequences called transcription factor binding sites (TFBS) [1]. Often, the binding sites of a given TF show recurring sequence patterns, which are referred to as motifs. Motifs can be represented and summarized using Position Weight Matrices (PWMs) [2], which encode the probability of observing a given nucleotide in a given position of a binding site. In recent years, several tools have been proposed for scanning regulatory DNA regions, such as enhancers or promoters, with the goal of predicting which TF may bind these genomic locations. Importantly, it has been shown that regulatory motifs are under purifying selection [3, 4], and mutations occurring in these regions can lead to deleterious consequences on the transcriptional states of a cell [5]. In fact, mutations can weaken, disrupt or create new TFBS and therefore alter expression of nearby genes. Mutations altering TFBS can occur in haplotypes that are conserved within a population or private to even a single individual, and can correspond to different phenotypic behaviour [6, 7]. For these reasons, population-level analysis of variability in TFBSs is of crucial importance to understand the effect of common or rare variants to gene regulation. Recently, a new class of methods and data structures based on genome graphs have enabled us to succinctly record and efficiently query thousands of genomes [8]. Genome graphs optimally encode shared and individual haplotypes based on a population of individuals. An efficient and scalable implementation of this approach called variation graphs (VGs) has been recently proposed [9]. Briefly, a VG is a graph where nodes correspond to DNA sequences and edges describe allowed links between successive sequences. Paths through the graph, which may be labelled (such as in the case of a reference genome), correspond to haplotypes belonging to different genomes [10]. Variants like SNPs and indels form bubbles in the graph, where diverging paths through the graph are anchored by a common start and end sequence on the reference [11]. VGs offer new opportunities to extend classic genome analyses originally designed for a single reference sequence to a panel of individuals. Moreover, by encoding individual haplotypes, VGs have been shown to be an effective framework to capture the potential effects of personal genetic variants on functional genomic regions profiled by ChIP-seq of histone marks [12]. During the last decade, several methods have been developed to search TFBS on linear reference genomes, such as FIMO [13] and MOODS [14] or to account for SNPs and short indels such as is-rSNP, TRAP and atSNP [15, 16, 17], however these tools do not account for individual haplotypes nor provide summary on the frequency of these events in a population. To solve these challenges, we have developed GRAFIMO, a tool that offers a variation- and haplotype-aware identification of TFBS in VGs. Here, we show the utility of GRAFIMO by searching TFBS on a VG encoding the haplotypes from all the individuals sequenced by the 1000 Genomes Project (1000GP) [18, 19].

Design and implementation

GRAFIMO is a command-line tool, which enables a variant- and haplotype- aware search of TFBS, within a population of individuals encoded in a VG. GRAFIMO offers two main functionalities: the construction of custom VGs, from user data, and the search of one or more TF motifs, in precomputed VGs. Briefly, given a TF model (PWM) and a set of genomic regions, GRAFIMO leverages the VG to efficiently scan and report all the TFBS candidates and their frequency in the different haplotypes in a



Sequence	Log-odds score	*P*-value	*q*-value	Reference	Haplotype frequency
GGGCCAGCAGGGGGCGCTG	28.22	$7.51e^{-12}$	$3.86e^{-6}$	non ref.	32
TGGCCAGCAGGGGGCGCTG	26.16	$3.12e^{-10}$	$7.72e^{-6}$	ref.	5063
GGGCCAGCAGGGAGCGCTG	19.43	$1.71e^{-7}$	$1.73e^{-4}$	non ref.	1

Fig 1. GRAFIMO TF motif search workflow. (A) The motif PWM (in MEME or JASPAR format) is processed and its values are scaled in the range [0, 1000]. The resulting score matrix is used to assign a score and a corresponding *P*-value to each motif occurrence candidate. In the final report GRAFIMO returns the corresponding log-odds scores, which are retrieved from the scaled values. (B) GRAFIMO slides a window of length *k*, where *k* is the motif width, along the haplotypes (paths in the graph) of the genomes used to build the VG. The resulting sequences are scored using the motif scoring matrix and are statistically tested assigning them the corresponding *P*-value and *q*-value. Moreover, for each entry is assigned a flag value stating if it belongs to the reference genome sequence ("ref") or contains genomic variants ("non.ref") and is computed the number of haplotypes in which the sequence appears.

single pass together with the predicted changes in binding affinity mediated by genetic variations. GRAFIMO is written in Python3 and Cython and it has been designed to easily interface with the *vg* software suite [9]. For details on how to install and run GRAFIMO see **S1 Text section 7**.

Genome variation graph construction

GRAFIMO provides a simple command-line interface to build custom genome variation graphs if necessary. Given a reference genome (FASTA format) and a set of genomic variants with respect to the reference (VCF format), GRAFIMO interfaces with the VG software suite to build the main VG data structure, the XG graph index [9] and the GBWT index [10, 20] used to track the haplotypes within the VG. To minimize the footprint of these files and speedup the computation, GRAFIMO constructs the genome variation graph by building a VG for each chromosome. This also speeds-up the search operation since we can scan different chromosomes in parallel. Alternatively, the search can be performed one chromosome at the time for machines with limited RAM.

Transcription Factor motif search

The motif search operation takes as input a set of genomes encoded in a VG (.xg format), a database of known TF motifs (PWM in JASPAR [21] or MEME format [22]) and a set of genomic regions (BED format), and reports in output all the TFBS motifs occurrences in those regions and their estimated significance (**Fig 1**). To search for potential TFBS, GRAFIMO slides a window of length *k* (where *k* is

the width of the query motif) along the paths of the VG corresponding to the genomic sequences encoded in it (**Fig 1 (B)**). This is accomplished by an extension to the *vg find* function, which uses the GBWT index of the graph to explore the *k*-mer space of the graph while accounting for the haplotypes embedded in it [10]. By default, GRAFIMO considers only paths that correspond to observed haplotypes, however it is possible also to consider all possible recombinants even if they are not present in any individual. The significance (log-likelihood) of each potential binding site is calculated by considering the nucleotide preferences encoded in the PWM as in FIMO [13]. More precisely, the PWM is processed to a Position Specific Scoring Matrix (PSSM) (**Fig 1 (A)**) and the resulting log-likelihood values are then scaled in the range [0, 1000] to efficiently calculate a statistical significance i.e. a *P*-value by dynamic programming [23] as in FIMO [13]. P-values are then converted to q-values by using the Benjamini-Hochberg procedure to account for multiple hypothesis testing. For this procedure, we consider all the P-values corresponding to all the k-mer-paths extracted within the scanned regions on the VG. GRAFIMO computes also the number of haplotypes in which a significant motif is observed and if it is present in the reference genome and/or in alternative genomes. (**Fig 1 (B)**).

Report generation

We have designed the interface of GRAFIMO based on FIMO, 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 compatible with the UCSC Genome Browser [24]. The TSV report (**S1 Fig**) contains for each candidate TFBS its score, genomic location (start, stop and strand), *P*-value, *q*-value, the number of haplotypes in which it is observed and a flag value to assess if it belongs to the reference or to the other genomes in VG. The HTML version of the TSV report (**S2 Fig**) can be viewed with any web browser. The GFF3 file (**S3 Fig**) can be loaded on the UCSC genome browser as a custom track, to visualize and explore the recovered TFBS with other annotations such as nearby genes, enhancers, promoters or pathogenic variants from the ClinVar database [25].

Results

As discussed above, GRAFIMO can be used to study how genetic variants may affect the binding affinity of potential TFBS within a set of individuals and may recover additional sites that are missed when considering only linear reference genomes without information about variants. To showcase its utility, we first constructed a VG based on 2548 individuals from the 1000GP phase 3 (hg38 human genome assembly) encoding their genomic variants and phased haplotypes (see **S1 Text section 1** for details). We then searched this VG for putative TFBS for three TF motifs with different lengths (from 11 to 19 bp), evolutionary conservation, and information content from the JASPAR database [21]: CTCF (JASPAR ID MA0139.1), ATF3 (JASPAR ID MA0605.2) and GATA1 (JASPAR ID MA0035.4) (**S4 Fig**) (see **S1 Text section 3-4-5**). To study regions with likely true binding events, for each factor we selected regions corresponding to peaks (top 3000 sorted by *q*-value) obtained by ChIP-

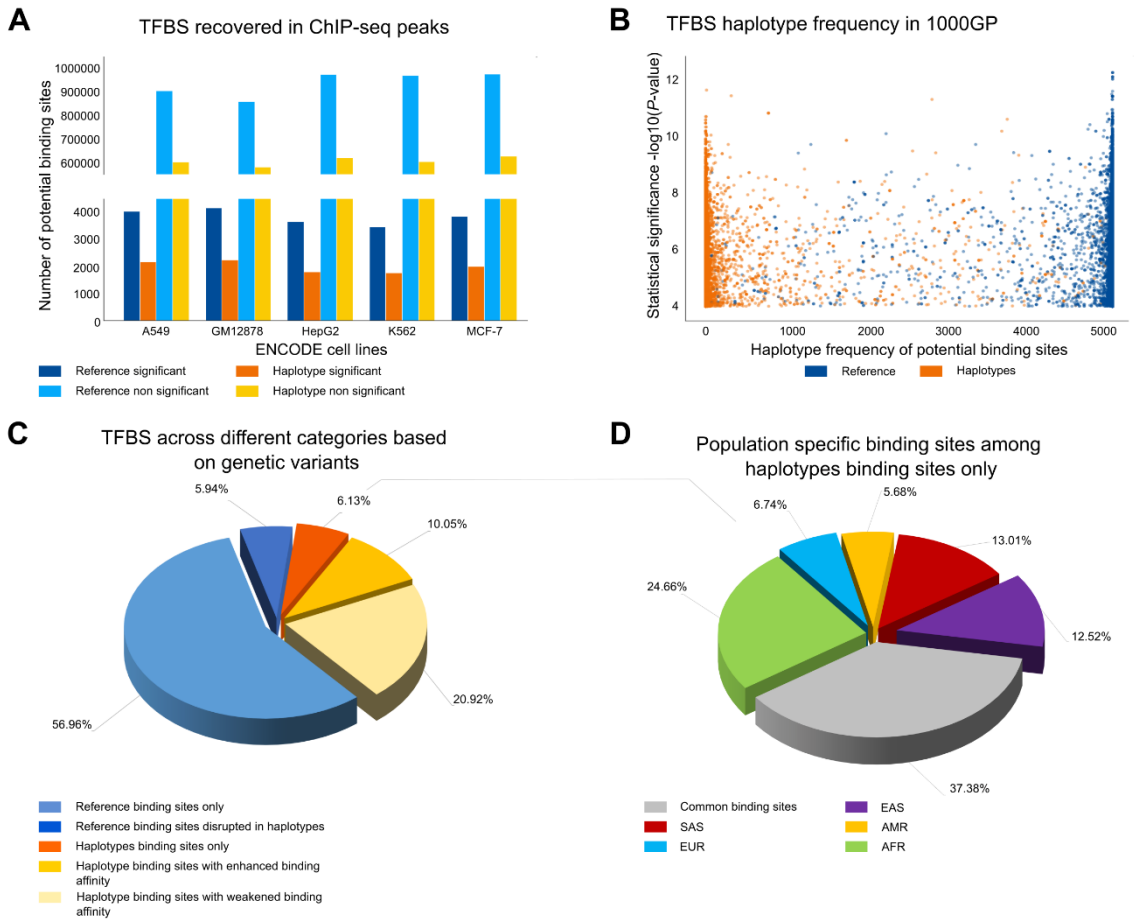
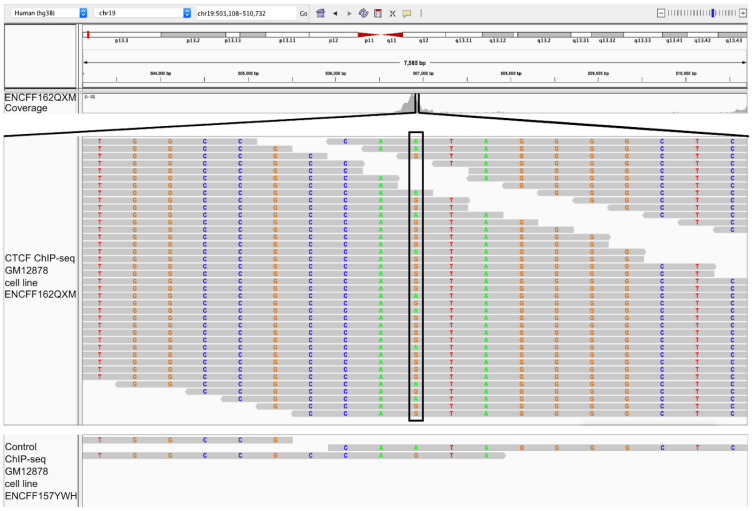
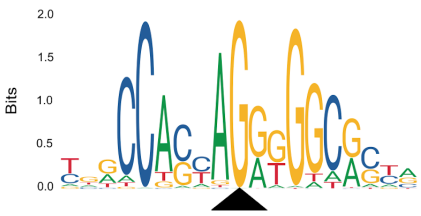


Fig 2. Searching CTCF motif on VG with GRAFIMO provides an insight on how genetic variation affects putative binding sites. (A) Potential CTCF occurrences statistically significant (P-value < 1e-4) and non-significant found in the reference and in the haplotype sequences found with GRAFIMO oh hg38 1000GP VG. (B) Statistical significance of retrieved potential CTCF motif occurrences and frequency of the corresponding haplotypes embedded in the VG. (C) Percentage of statistically significant CTCF potential binding sites found only in the reference genome or alternative haplotypes and with modulated binding scores based on 1000GP genetic variants (D) Percentage of population specific and common (shared by two or more populations) potential CTCF binding sites present on individual haplotypes.

seq experiments in 6 different cell types (A549, GM12878, H1, HepG2, K562, MCF-7) from the ENCODE project [26, 27] (see **S1 Text section 2**). We used GRAFIMO to scan these regions and selected for our downstream analyses only sites with a *P*-value < $1e^{-4}$ and considered them as potential TFBS for these factors. Based on the recovered sites, we consistently observed across the 3 studied TFs that genetic variants can significantly affect estimated binding affinity. In fact, we found that thousands of CTCF motif occurrences are found only in non-reference haplotypes, suggesting that a considerable number of TFBS candidates are lost when scanning for TFBS the genome without accounting for genetic variants (**Fig 2 (A)**). Similar results were obtained searching for ATF3 (**S5 Fig (A)**) and GATA1 (**S6 Fig (A)**). We also found several highly significant CTCF motif occurrences in rare haplotypes that may potentially modulate gene expression in these individuals (**Fig 2 (B)**). Similar behaviours were observed for ATF3 (**S5 Fig (B)**) and GATA1 (**S6 Fig (B)**).



Reads	#A in REF	#G in ALT
ChIP-seq	10 (29.41%)	24 (70.59%)
Control	1 (50%)	1 (50%)



Sequence	Genomic location	Strand	*P*-value	Reference	Haplotype frequency
TGGCCGCCAGTAGGGGCTC	chr19:506,910-506,529	+	$7.64e^{-5}$	non ref.	3529
TGGCCGCCAATAGGGGCTC	chr19:506,910-506,529	+	$1.31e^{-3}$	ref.	1567

Fig 3. Considering genomic variation, GRAFIMO captures more potential binding events. GRAFIMO reports a potential CTCF binding site at chr19:506,910-506,929 found only in haplotype sequences, searching the motif in ChIP-seq peaks called on cell line GM12878 (experiment code ENCSR000DZN). The reads used to call for ChIP-seq peaks (ENCFF162QXM) show an allelic imbalance at position 10 of the motif sequence towards the alternative allele G, instead of the reference allele A. The imbalance is captured by GRAFIMO which reports the sequence presenting G at position 10 (found in the haplotypes), while the potential TFBS on the reference carrying an A is not reported as statistically significant (*P*-value > $1e^{-4}$). CTCF motif logo shows that the G is the dominant nucleotide in position 10.

We also investigated the potential effects of the different length and type of mutations i.e. SNPs and indels on the CTCF, ATF3 and GATA1 binding sites. However, we did not observe a clear and general trend (**S7 Fig).** 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 the potential CTCF binding sites can be found only on non-reference haplotype sequences, 5.94% are disrupted by variants in non-reference haplotypes and ~30% are still significant in non-reference haplotypes but with different binding scores (**Fig 2 (C)**). Similar results were observed for ATF3 (**S5 Fig (C)**) and GATA1 (**S6 Fig (C)**). Interestingly, we observed that a large fraction of putative binding sites recovered only on individual haplotypes are population specific. For CTCF we found that 24.66%, 6.74%, 5.68%, 13.01%, 12.52% of potential CTCF TFBS retrieved on individual haplotype sequences only are specific for AFR, EUR, AMR, SAS and EAS populations, respectively (**Fig 2 (D)**). Similar results were observed for ATF3 and GATA1 (**S1 Text sections 4-5**).

Among the unique CTCF motif occurrences found only on non-reference haplotypes in CTCF ChIP-seq peaks we uncovered one TFBS (chr19:506,910-506,929) that clearly illustrates the danger of only using reference genomes for motif scanning. Within this region we recovered a heterozygous SNP that overlaps (position 10 of the CTCF matrix) and significantly modulates the binding affinity of this TFBS. In fact, by inspecting the ChIP-seq reads (experiment ENCSR000DZN, GM12878 cell line), we observed a clear allelic imbalance towards the alternative allele G (70.59% of reads) with respect to the reference allele A (29.41% of reads). This allelic imbalance is not observed in the reads used as control (experiment code ENCSR000EYX) (**Fig 3**).

Taken together these results highlight the importance of considering non-reference genomes when searching for potential TFBS or to characterize their potential activity in a population of individuals.

We also compared the performance of GRAFIMO against FIMO [13] (**S1 Text Section 6** and **S8 Fig**). FIMO is faster and requires less memory, when scanning a single linear genome. However, when considering the 2548 individual genomes and their genetic variation, GRAFIMO proves to be generally faster than FIMO. Moreover, we benchmarked how GRAFIMO running time and memory usage change using an increasing number of threads (**S9 Fig**). By increasing the number of threads, we observed a dramatical drop in running time, while memory usage remained similar.

Conclusion

By leveraging VGs, GRAFIMO provides an efficient method to study how genetic variation affects the binding landscape of a TF within a population of individuals. Moreover, we show that several potential and private TFBS are found in individual haplotype sequences and that genomic variants significantly also affect the binding affinity of several motif occurrence candidates found in the reference genome sequence. Our tool therefore can help in prioritizing potential regions that may mediate individual specific changes in gene expression, which may be missed by using only reference genomes.

Availability and future directions

GRAFIMO can be downloaded and installed via PyPI, source code or Bioconda. Its Python3 source code is available on Github at https://github.com/pinellolab/GRAFIMO and at https://github.com/InfOmics/GRAFIMO under MIT license. Since GRAFIMO is based on VG data structure, has the potential to be applied to future pangenomic reference systems that are currently under development (https://news.ucsc.edu/2019/09/pangenome-project.html). The genome variation graphs enriched with 1000GP on GRCh38 phase 3 used to obtain the results presented in this manuscript can be downloaded at http://ncrnadb.scienze.univr.it/vgs.

Supporting information

S1 Text. Additional information about experiments design, GATA1 and ATF3 search on genome variation graph, and how to install and run GRAFIMO.

S1 Fig. Example of TSV summary report. The tab-delimited report (TSV report) shows the first 25 potential CTCF occurrences retrieved by GRAFIMO, searching the motif in ChIP-seq peak regions defined in ENCODE experiment ENCFF816XLT (cell line A549).

S2 Fig. Example of HTML summary report. The HTML report shows the first 25 potential CTCF occurrences retrieved by GRAFIMO, searching the motif in ChIP-seq peak regions defined in ENCODE experiment ENCFF816XLY (cell line A549).

S3 Fig. Example of GFF3 track produced by GRAFIMO, loaded on the UCSC genome browser. GRAFIMO returns also a GFF3 report which can be loaded on the UCSC genome browser; the loaded custom track shows three potential CTCF occurrences (region chr8:142,782,661-142,782,680) retrieved by GRAFIMO overlapping a dbSNP annotated variant (rs892844).

S4 Fig. Structure of transcription factor motifs used to test GRAFIMO. Transcription factor binding site motifs of (A) CTCF, (B) ATF3 and (C) GATA1.

S5 Fig. Searching ATF3 motif on VG with GRAFIMO provides an insight on how genetic variation affects the binding site sequence. (A) Potential ATF3 occurrences statistically significant (P-value < $1e^{-4}$) and non-significant found in the reference and in the haplotype sequences found with GRAFIMO oh hg38 1000GP VG. (B) Statistical significance of retrieved potential ATF3 motif occurrences and their frequency in the haplotypes embedded in the VG. (C) Percentage of statistically significant ATF3 potential binding sites found only in genome reference sequence, percentage of potential TFBS found in the reference for which genetic variants cause the sequence to be no more significant, percentage of binding sites found only in the haplotypes, percentage of potential TFBS found in the reference with increased statistical significance by the action of genomic variants and percentage of those with a decreased significance by the action of variants (with P-value still significant). (D) Fraction of population specific potential ATF3 binding sites recovered on individual haplotype sequences.

S6 Fig. Searching GATA1 motif on VG with GRAFIMO provides an insight on how genetic variation affects the binding site sequence. (A) Potential GATA1 occurrences statistically significant (P-value < $1e^{-4}$) and non-significant found in the reference and in the haplotype sequences found with GRAFIMO oh hg38 1000GP VG. (B) Statistical significance of retrieved potential GATA1 motif occurrences and their frequency in the haplotypes embedded in the VG. (C) Percentage of statistically significant GATA1 potential binding sites found only in genome reference sequence, percentage of potential TFBS found in the reference for which genetic variants cause the sequence to be no more significant, percentage of binding sites found only in the haplotypes, percentage of potential TFBS found in the reference with increased statistical significance by the action of genomic variants and percentage of those with a decreased significance by the action of variants (with *P*-value still significant). (D) Percentage of population specific potential GATA1 binding sites, among those TFBS retrieved uniquely on individual genome sequences.

S7 Fig. Influence of the different length and type of mutations on binding affinity score. (A) CTCF, (B) ATF3, (C) GATA1.

S8 Fig. Comparing GRAFIMO and FIMO performance. (A) Searching CTCF motif (JASPAR ID MA0139.1) on human chr22 regions (total width ranging from 1 to 9 millions of bp) without accounting for genetic variants FIMO is faster than GRAFIMO (using a single thread). (B) FIMO uses less memory resources than GRAFIMO, however they work on different frameworks. (C) When considering the genetic variation present in large panels of individuals as 1000GP on GRCh38 phase 3 (2548 samples), GRAFIMO proves to be faster than FIMO in searching potential CTCF occurrences. It is faster when run with a single execution thread, and significantly faster when run with 16.

S9 Fig. GRAFIMO running time efficiently scales with the number of threads used. By running GRAFIMO with multiple threads (A) the running time significantly decreases, while (B) memory usage remains similar.

Acknowledgements

L.P. was supported by National Human Genome Research Institute R00HG008399 and Genomic Innovator Award R35HG010717; R.G was supported from the European Union's Horizon 2020 research and innovation programme under grant agreement 814978 and JPcofuND2 Personalised Medicine for Neurodegenerative Diseases project JPND2019-466-037. We would like to thank Centro Piattaforme Tecnologiche (CPT) located in the University of Verona that provided us with all the hardware necessary to perform all the tests.

References

1. Stewart AJ, Hannehanhalli S, Plotkin JB. Why transcription factor binding sites are ten nucleotides long. Genetics. 2012;192(3): 973–985.
2. Stormo GD. Modeling the specificity of protein—dna interactions. Quantitative Biology. 2013; 1(2): 115–130.
3. Li S, Ovcharenko I. Human enhancers are fragile and prone to deactivating mutations. Mol Bio Evol. 2015;32(18): 2161–2180.
4. Vorontsov IE, Khimulya G, Lukianova EN, Nikolaeva DD, Eliseeva IA, Kulakovskiy IV, et al. Negative selection maintains transcription factors binding motifs in human cancer. BMC genomics.2016;17(2): 395.
5. Guo YA, Chang MM, Huang W, Ooi WF, Xing M, Tan P, et al. Mutation hotspots at CTCF binding sites coupled to chromosomal instability in gastrointestinal cancers. Nature communications. 2018;9(1): 1–14.
6. Albert FW, Kruglyak L. The role of regulatory variation complex traits and diseases. Nature Reviews Genetics. 2015;16(4): 197—212.
7. Kasowski M, Grubert F, Heffelfinger C, Hariharan M, Asabere A, Waszak SM, et al. Variation in transcription factor binding among humans. Science. 2010;328(5975):232—235.
8. Paten B, Novak AM, Eizenga JM, Garrison E. Genome graphs and the evolution of genome inference. Genome research. 2017;27(5): 665—676.
9. Garrison E, Sirén J, Novak AM, Hickey G, Eizenga JM, Dawson ET, et al. Variation graph toolkit improves read mapping by representing genetic variation in the reference. Nature biotechnology. 2018;36(9): 875–879.
10. Sirén J, Garrison E, Novak AM, Paten B, Durbin R. Haplotype-aware graph indexes. Bioinformatics. 2020;36(2): 400-407.
11. Paten B, Eizenga JM, Rosen YM, Novak AM, Garrison E, Hickey G. Superbubbles, ultrabubbles and cacti. Journal of Computational Biology. 2018;25(7): 649–663.
12. Groza C, Kwan T, Soranzo N, Pastinen T, Bourque G. Personalized and graph genomes reveal missing signal in epigenomic data. Genome biology. 2020;21: 1—22.
13. Grant CE, Bailey TL, Noble WS. Fimo: scanning for occurrences of a given motif. Bioinformatics. 2011;27(7): 1017–1018.
14. Kohronen J, Martinmäki P, Pizzi C, Rastas P, Ukkonen E. Moods: fast search for position weight matrix matches in dna sequences. Bioinformatics. 2009;25(23):3181–3182.

15. Macintyre G, Bailey J, Haviv I, Kowalczyk A. is-rSNP: a novel technique for in silico regulatory SNP detection. Bioinformatics. 2010;26(18): i524—i530.
16. Thomas-Chollier M, Hufton A, Heining M, O'Keefe S, El Masri N, Roider H, et al. Transcription factor binding prediction using TRAP for the analysis of ChIP-seq data and regulatory SNPs. Nature protocols. 2011;6(12): 1860.
17. Zuo C, Shin S, Keles S. atsnp: transcription factor binding affinity testing for regulatory snp detection. Bioinformatics. 2015;31(20): 3353–3355.
18. 1000 Genomes Project Consortium. A global reference for human genetic variation. Nature. 2015;526(7571): 68–74.
19. Zheng-Bradley X, Streeter I, Fairley S, Richardson D, Clarke L, Flicek P, et al. Alignment of 1000 Genomes Project reads to reference assembly GRCh38. GigaScience. 2017;6(7): 1-8.
20. Novak, AM, Garrison E, Paten B. A graph extension of the positional Burrows-Wheeler transform and its applications. Algorithms for Molecular Biology. 2017;12(1): 18.
21. Fornes O, Castro-Mondragon JA, Khan A, van der Lee R, Zhang X, Richmond PA, et al. JASPAR 2020: update of the open- access database of transcription factor binding profiles. Nucleic Acid Research. 2019;48(D1): D87—D92.
22. Bailey TL, Boden M, Buske FA, Frith M, Grant CE, Clementi L, et al. Meme suite: tools for motif discovery and searching. Nucleic Acid Research. 2009;37(suppl): W202—W208.
23. Staden R. Searching for motifs in nucleic acid sequences. Methods in molecular biology. 1994;25: 93–102.
24. Lee CM, Barber GP, Casper J, Clawson H, Diekhans M, Navarro Gonzalez J, et al. UCSC Genome Browser enters 20th year. Nucleic Acid Research. 2020;48(D1): D756–D761.
25. Landrum MJ, Chitipiralla S, Brown GR, Chen C, Gu B, Hart J, et al. ClinVar: improvements to accessing data. Nucleic Acid Research. 2020;48(D1): D835—D844
26. ENCODE Project Consortium. An Integrated encyclopedia of DNA elements in the human genome. Nature. 2012;489(7414): 57—74.
27. Davis CA, Hitz BC, Sloan CA, Chan ET, Davidson JM, Gabdank I, et al. The encyclopedia of DNA elements (ENCODE): data portal update. Nucleic Acid Research. 2018;46(D1): D794—D801.

Supplementary materials of "GRAFIMO: variant and haplotype aware motif scanning on pangenome graphs"

Manuel Tognon[1], Vincenzo Bonnici[1], Erik Garrison[2], Rosalba Giugno[1,*], Luca Pinello[3,4,5,*]

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

* rosalba.giugno@univr.it or lpinello@mgh.harvard.edu

1. 1000 Genomes genomic variants

To validate GRAFIMO, we tested it on a pangenome variation graph (VG) [1] based on 2548 individuals from the 1000 Genomes Project (1000GP) phase 3 on GRCh38 cohort [2, 3], encoding all their genomic variants and their haplotypes. We constructed the 1000GP VG using ~78 millions of genomic variants (SNPs and indels) (see **Table A in S1 Text**), belonging to the 2548 considered subjects and constituted a total of 5096 haplotypes.

2. GRAFIMO on ENCODE Project's ChIP-seq data

To test GRAFIMO we selected three transcription factor (TF) motifs of different length, evolutionary conservation, information content, and tolerance to point mutations: CTCF (**S4 Fig (A)**), ATF3 (**S4 Fig (B)**) and GATA1 (**S4 Fig (C)**). For each TF we obtained ChIP-seq optimal IDR thresholded peak regions (in ENCODE bigBED format) from the ENCODE Project database [4, 5]. The ChIP-seq peaks were mapped on the hg38 human genome assembly and were retrieved from different cell types (see **Table B in S1 Text**). In order to obtain a suitable input for GRAFIMO, the bigBED files were converted to the corresponding BED files with UCSC bigBedToBed tool [6]. The resulting BED files were filtered in order to contain only features related to the canonical chromosomes and were sorted by *q*-values to select the most informative regions (top 3000 for each experiment).

3. Searching for CTCF occurrences

CTCF is a zinc-finger transcription factor involved in transcriptional regulation, which plays a fundamental role in epigenetic regulation [7] and acts as tumor suppressor [8]. During our tests, we searched CTCF motif (**S4 Fig (A)**) (JASPAR ID MA0139.1) on the 1000GP VG (hg38 genome assembly) (see **S1 Text section 1**). In order to have binding events likely to happen, CTCF motif has been searched on regions corresponding to the top 3000 significant ChIP-seq peaks regions (sorted and filtered by *q*-value, see **S1 Text section 2**) for CTCF on the A549, HepG2, GM12878, K562, and MCF-7 cell lines (see **Table B in S1 Text**).

4. Searching for ATF3 occurrences

Activating Transcription Factor 3 (ATF3) is a transcription factor belonging to the family of cAMP responsive element-binding. Its activity is induced by physiological stress in a wide variety of tissues [9] and has been shown to have significant roles in both immunity and cancer [10]. ATF3 binds short, conserved DNA sequences. We searched the ATF3 motif (**S4 Fig (B)**) (JASPAR ID MA0605.2) on a hg38 VG enriched with SNPs and indels from 2548 individuals of 1000GP phase 3. As done for CTCF, to have likely to happen binding events we searched ATF3 motif on regions corresponding to the top 3000 significant ChIP-seq optimal IDR thresholded peaks (sorted and filtered by *q*-value, see **S1 Text section 2**) for ATF3 on the H1, HepG2 and K562 cell lines (see **Table B in S1 Text**).

For our downstream analysis we selected only the ATF3 motif occurrence whose *P*-value was < $1e^{-4}$ and we considered them as potential binding sites. We found several potential motif occurrences which would be lost scanning only the reference genome sequence (**S5 Fig (A)**). Moreover, we observed that many ATF3 motif candidates with highly statistically significant *P*-values are found in rare haplotypes (**S5 Fig (B)**). We also found that 7.03% of the potential ATF3 binding sites can be found only in non-reference haplotype sequences, 11.28% are disrupted by genomic variants in non-reference haplotypes and ~13% of the potential ATF3 TFBS are still significant in non-reference haplotypes but showing different binding scores (**S5 Fig (C)**). We also observed that a considerable fraction of putative ATF3 binding sites are population specific. In fact, 19.81%, 3.77%, 7.55%, 19.81%, 19.81% of potential TFBS retrieved on individual haplotypes are specific for AFR, EUR, AMR, SAS and EAS populations, respectively (**S5 Fig (D)**).

5. Searching for GATA1 occurrences

GATA1 is a zinc-finger transcription factor having a fundamental role during the development of hematopoietic cell lineages [11]. GATA1 binds short (11 bp) highly conserved DNA sequences. GATA1 motif (**S4 Fig (C)**) (JASPAR ID MA0035.4) has been searched with GRAFIMO on a hg38 VG enriched with SNPs and indels from 2548 individuals of 1000GP phase 3. In order to have likely to happen binding events, we searched GATA1 motif occurrences in regions corresponding to the top 3000 significant ChIP-seq optimal IDR thresholded peaks (sorted and filtered by *q*-value, see **S1 Text section 2**) for GATA1 on the K562 cell line (see **Table B in S1 Text**).

In our downstream analysis we considered only the motif occurrences, whose *P*-value was < $1e^{-4}$ and we considered them as potential binding sites. We observed that many potential GATA1 occurrences are lost when searching the motif only in the reference genome (**S6 Fig (A)**). We also found that several potential motif occurrences with a highly statistically significant *P*-value are detected in rare haplotypes (**S6 Fig (B)**). Moreover, we found that 9.78% of the potential GATA1 TFBS can be found only in non-reference haplotype sequences, 12.58% are disrupted by genomic variants and ~4% are still significant in non-reference haplotypes but with different binding scores (**S6 Fig (C)**). We also found that several potential GATA1 binding sites among those retrieved only in individual genome sequences are population specific. We observed that 25.97% of putative binding sites are specific for AFR, 3.90% for EUR, 9.09% for AMR, 19.48% for SAS and 11.69% for EAS populations (**S6 Fig (D)**).

6. Comparing FIMO and GRAFIMO

To assess GRAFIMO correctness we compared the obtained results with those retrieved running FIMO [12] on the same ChIP-seq regions used to test our tool. We run both GRAFIMO and FIMO on a Linux-based machine (OS Ubuntu 18.04), with an Intel(R) Core (TM) i7- 5960X 3.00GHz CPU (16 cores) and 64GB of RAM. FIMO requires in input a set of sequences given via a FASTA file. For each tested TF motif, we obtained the reference genome sequences corresponding to the ChIP-seq optimal IDR thresholded peaks used to test GRAFIMO using BEDTools [13] *getfasta* functionality. We observed

that, for each TF motif, GRAFIMO does not lose any potential motif occurrence with respect to those retrieved running FIMO. Thus, GRAFIMO detects more motif occurrence candidates, located on the alternative haplotypes embedded in the VG, without losing any potential binding sites in the reference genome sequence.

We also benchmarked GRAFIMO comparing its running time and memory usage to FIMO. To this end we searched for potential occurrences of the CTCF motif (19 bp wide, JASPAR ID MA0139.1) on forward and reverse strand in 1000 genomic regions of the human chr22 of increasing length (1 to 9 million). To run FIMO, for each set of genomic regions we created FASTA file as required by this tool.

We assessed GRAFIMO performance scanning the chr22 VG without any encoded genetic variant for CTCF motif occurrences, on the previously computed sets of genomic regions. Moreover, since FIMO does not provide a parallel implementation, we run GRAFIMO using a single thread. In this scenario FIMO is faster and requires less memory than GRAFIMO (**S8 Fig (A)** and **(B)**).

We also benchmarked GRAFIMO running time with a single thread when performing a motif search on the VG encoded chr22 with 1000GP genetic variants (2548 individuals). In this test setting, excluding the preprocessing required to retrieve the FASTA genomic sequences for each of the 2548 1000GP subjects from the running time calculations, GRAFIMO proves to be faster, although FIMO speed is similar scanning regions up to 1 million of bp (**S8 Fig (C)**).

These results are expected since FIMO is a highly optimized tool to scan linear genomic sequences, while GRAFIMO has been designed to efficiently scan panels of individuals simultaneously and to run in parallel using multiple threads.

We also assessed GRAFIMO running time and memory usage when using 1, 4, 8 and 16 threads, respectively. This analysis was performed using the chr22 VG with 1000GP genetic variants (**S9 Fig**). As expected, running the tool with multiple threads resulted in a dramatical reduction of the running time, however the memory usage remains similar increasing the number of threads used. Not surprisingly, running GRAFIMO in parallel on the chr22 VG enriched with variants produces the major improvements in running time.

To obtain the results presented in the manuscript on average each scan required ~15 minutes and used ~24 GB of memory.

7. Installing and running GRAFIMO

In this section will be presented how to install and run GRAFIMO. For further details on installation and run refer to GRAFIMO's README and Wiki at https://github.com/pinellolab/GRAFIMO and https://github.com/InfOmics/GRAFIMO.

Before installing GRAFIMO the user must have installed:

- VG, v1.27.1 or later (https://github.com/vgteam/vg)
- Tabix (https://github.com/samtools/htslib)
- Graphviz (https://graphviz.org)

GRAFIMO has been written in Python3 and Cython. To build GRAFIMO, Cython is required to be installed. To build GRAFIMO are also used *Setuptools* and *Wheel* Python packages. Moreover, GRAFIMO depends on the following Python packages:

- *Pandas*

- *NumPy*
- *Statsmodels*
- *Sphinx*
- *Numba*
- *Colorama*

Once all the dependencies have been satisfied, GRAFIMO can be installed via pip typing on terminal:

```
pip3 install grafimo
```

GRAFIMO can also be built from source code, by cloning GRAFIMO repository on Github. To clone the repository, type:

```
git clone https://github.com/pinellolab/GRAFIMO.git.
```

To build GRAFIMO, type:

```
cd GRAFIMO; python3 setup.py install –user
```

To install GRAFIMO via Bioconda (for Linux users only), type:

```
conda install grafimo
```

To scan a pangenome variation graph with GRAFIMO are required the path to a directory containing the VGs of all the chromosome (XG and GBWT indexes) or the path to a whole pangenome variation graph (note that the XG and GBWT index of the VG must be stored in the same location), a motif PWM given in JASPAR or MEME format and a BED file containing the genomic regions where the motif will be searched.

Let us assume that we built the pangenome variation graph by constructing a VG for each chromosome. To find potential motif occurrences in the VG, type

```
grafimo -d /path/to/directory/storing/my/graphs/ -b /path/to/my/bedfile -m path/to/my/motif
```

To scan a whole pangenome variation graph for the occurrence of the given motif, type

```
grafimo -g /path/to/my/whole/genome/vg -b /path/to/my/bedfile -m /path/to/my/motif
```

With GRAFIMO it is also possible to build a pangenome variation graph from user data. To construct a VG are required a FASTA file containing the reference genome and a VCF file containing the phased genomic variants to enrich the reference sequence.

To build the pangenome variation graph with GRAFIMO, type

```
grafimo buildvg -l /path/to/reference/genome -v /path/to/vcf/file
```

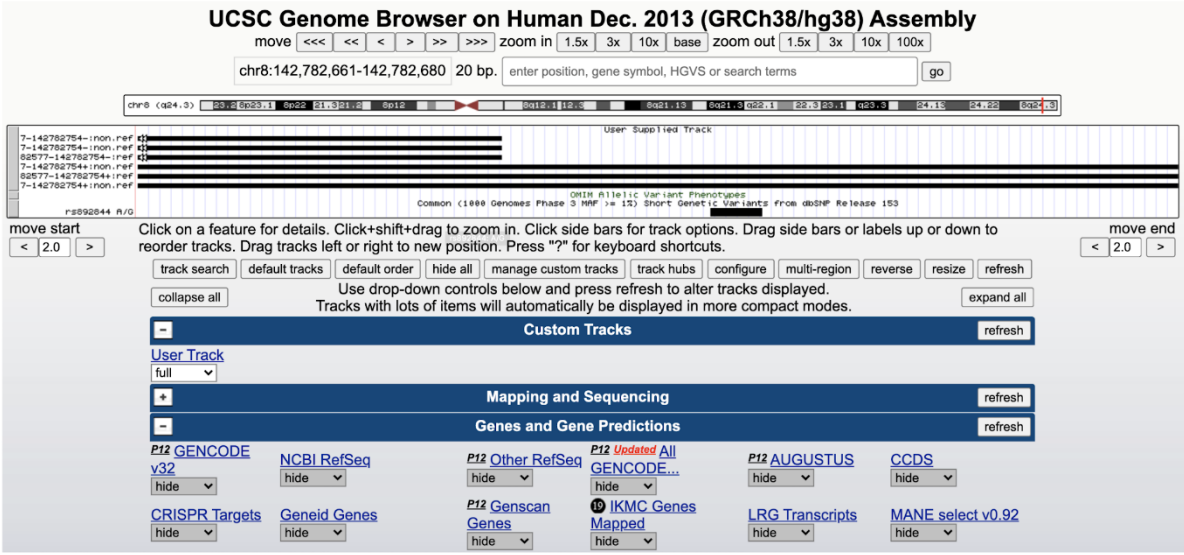
Hands-on tutorials on how to run GRAFIMO are available at https://github.com/pinellolab/GRAFIMO and https://github.com/InfOmics/GRAFIMO.

	motif_id	motif_alt_id	sequence_name	start	stop	strand	score	p-value	q-value	matched_sequence	haplotype_frequency	reference
1	MA0139.1	CTCF	chr2:231612642-231612803	231612720	231612739	+	29.114754098360663	5.988799754224379e-13	6.040843464356013e-07	TGGCCACCAGGGGGCGCCG	5096	ref
2	MA0139.1	CTCF	chr3:98023327-98023504	98023430	98023411	-	29.049180327868884	8.440969121103595e-13	6.040843464356013e-07	CGGCCACCAGGGGGCGCCA	5096	ref
3	MA0139.1	CTCF	chr7:98021604-98021781	98021693	98021712	+	28.983606557377072	1.0141213123241087e-12	6.040843464356013e-07	CGGCCACCAGGGGGCGCCG	5096	ref
4	MA0139.1	CTCF	chr12:121499983-121500138	121500062	121500081	+	28.42622950819674	4.0897345623855974e-12	1.218071546367531e-06	CGGCCACCAGGGGGCGCCC	5096	ref
5	MA0139.1	CTCF	chr16:67993308-67993508	67993402	67993421	+	28.42622950819674	4.0897345623855974e-12	1.218071546367531e-06	CGGCCACCAGGGGGCGCCC	5094	ref
6	MA0139.1	CTCF	chr12:108618200-108618398	108618301	108618282	-	28.42622950819674	4.0897345623855974e-12	1.218071546367531e-06	CGGCCACCAGGGGGCGCCC	5096	ref
7	MA0139.1	CTCF	chr8:142782577-142782754	142782661	142782680	+	28.327868852459005	5.273880274923456e-12	1.3463598544475949e-06	TGGCCACCAGGGGGCGCTC	2835	non.ref
8	MA0139.1	CTCF	chr1:156090811-156090991	156090877	156090896	+	28.032786885245912	1.1988060111075724e-11	2.3803199115082577e-06	TGGCCACCAGGTGGCGCCG	5096	ref
9	MA0139.1	CTCF	chr16:66608629-66608805	66608709	66608728	+	28.032786885245912	1.1988060111075724e-11	2.3803199115082577e-06	TGGCCACCAGGTGGCGCCG	5088	ref
10	MA0139.1	CTCF	chr19:35067213-35067372	35067283	35067302	+	27.918032786885525	1.5942099323583996e-11	2.559578506660857e-06	TGGCCACCAGGGGGCACTG	5096	ref
11	MA0139.1	CTCF	chr15:34210116-34210292	34210206	34210187	-	27.918032786885525	1.5942099323583996e-11	2.559578506660857e-06	TGGCCACCAGGGGGCACTG	787	non.ref
12	MA0139.1	CTCF	chr9:130461251-130461422	130461337	130461356	+	27.885245901639337	1.7187819018057012e-11	2.559578506660857e-06	TGGCCACCAGAGGGCGCCA	5096	ref
13	MA0139.1	CTCF	chr1:7647328-7647487	7647424	7647405	-	27.803278688524586	2.1113539694844316e-11	2.695019677028664e-06	TGGCCACCAGGTGGCGCTG	5096	ref
14	MA0139.1	CTCF	chr16:66608629-66608805	66608709	66608728	+	27.803278688524586	2.1113539694844316e-11	2.695019677028664e-06	TGGCCACCAGGTGGCGCTG	8	non.ref
15	MA0139.1	CTCF	chr3:120558991-120559170	120559086	120559067	-	27.672131147540995	2.6237024511211056e-11	3.0606872796384374e-06	TGGCCAGCAGGGGGCGCTC	5052	ref
16	MA0139.1	CTCF	chr19:4809674-4809833	4809756	4809775	+	27.590163934426243	2.945541904678063e-11	3.0606872796384374e-06	TGGCCACCAGAGGGCGCTG	5096	ref
17	MA0139.1	CTCF	chr20:44685269-44685461	44685354	44685373	+	27.57377049180326	3.0829219981831114e-11	3.0606872796384374e-06	TGGCCAGCAGGGGGCGGTG	5091	ref
18	MA0139.1	CTCF	chr19:47863504-47863679	47863606	47863587	-	27.57377049180326	3.0829219981831114e-11	3.0606872796384374e-06	TGGCCACTAGGGGGCGGCA	5095	ref
19	MA0139.1	CTCF	chr10:123994704-123994883	123994783	123994802	+	27.491803278688508	3.5727124816933935e-11	3.3144635195783872e-06	TGGCCACCAGGGGGCAGCG	2	non.ref
20	MA0139.1	CTCF	chr5:177501044-177501224	177501155	177501136	-	27.459016393442596	3.7094909167992564e-11	3.3144635195783872e-06	CGGCCACCAGAGGGCGCTG	5096	ref
21	MA0139.1	CTCF	chr10:123994704-123994883	123994783	123994802	+	27.360655737704917	4.4407968997956725e-11	3.7789447576156337e-06	CGGCCACCAGGGGGCAGCG	5086	ref
22	MA0139.1	CTCF	chr20:49979066-49979252	49979155	49979174	+	27.229508196721326	5.882201242820514e-11	4.509085382907359e-06	TGGCCAGCAGAGGGCGCCA	4789	ref
23	MA0139.1	CTCF	chr12:111397009-111397212	111397090	111397109	+	27.213114754098342	6.217717166529091e-11	4.509085382907359e-06	CAGCCACCAGGGGGCGCCA	5096	ref
24	MA0139.1	CTCF	chr20:63973975-63974156	63974071	63974052	-	27.13114754098359	7.240291664795245e-11	4.509085382907359e-06	CGGCCACCAGGGGGCAGTG	5094	ref
25	MA0139.1	CTCF	chr10:123994704-123994883	123994783	123994802	+	27.13114754098359	7.240291664795245e-11	4.509085382907359e-06	CGGCCACCAGGGGGCAGTG	2	non.ref

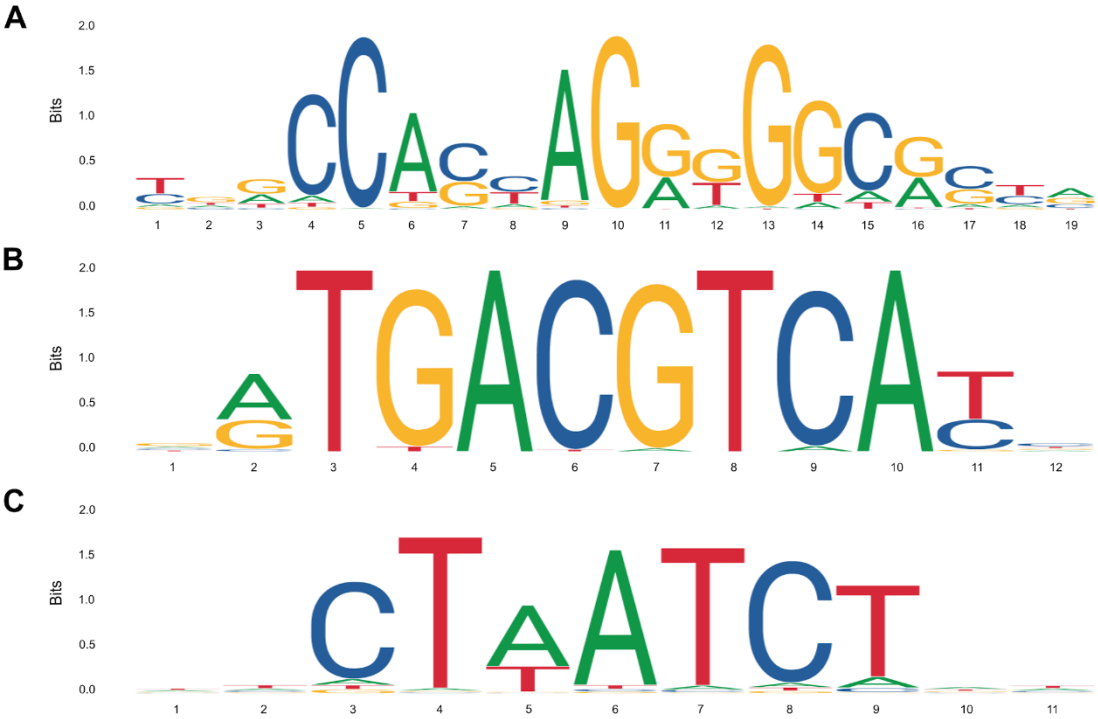
S1 Fig. Example of TSV summary report. The tab-delimited report (TSV report) shows the first 25 potential CTCF occurrences retrieved by GRAFIMO, searching the motif in ChIP-seq peak regions defined in ENCODE experiment ENCFF816XLT (cell line A549).

	motif_id	motif_alt_id	sequence_name	start	stop	strand	score	p-value	q-value	matched_sequence	haplotype_frequency	reference
1	MA0139.1	CTCF	chr2:231612642-231612803	231612720	231612739	+	29.114754	5.988800e-13	6.040843e-07	TGGCCACCAGGGGGCGCCG	5096	ref
2	MA0139.1	CTCF	chr3:98023327-98023504	98023430	98023411	-	29.049180	8.440969e-13	6.040843e-07	CGGCCACCAGGGGGCGCCA	5096	ref
3	MA0139.1	CTCF	chr7:98021604-98021781	98021693	98021712	+	28.983607	1.014121e-12	6.040843e-07	CGGCCACCAGGGGGCGCCG	5096	ref
4	MA0139.1	CTCF	chr12:121499983-121500138	121500062	121500081	+	28.426230	4.089735e-12	1.218072e-06	CGGCCACCAGGGGGCGCCC	5096	ref
5	MA0139.1	CTCF	chr16:67993308-67993508	67993402	67993421	+	28.426230	4.089735e-12	1.218072e-06	CGGCCACCAGGGGGCGCCC	5094	ref
6	MA0139.1	CTCF	chr12:108618200-108618398	108618301	108618282	-	28.426230	4.089735e-12	1.218072e-06	CGGCCACCAGGGGGCGCCC	5096	ref
7	MA0139.1	CTCF	chr8:142782577-142782754	142782661	142782680	+	28.327869	5.273880e-12	1.346360e-06	TGGCCACCAGGGGGCGCTC	2835	non.ref
8	MA0139.1	CTCF	chr1:156090811-156090991	156090877	156090896	+	28.032787	1.198806e-11	2.380320e-06	TGGCCACCAGGTGGCGCCG	5096	ref
9	MA0139.1	CTCF	chr16:66608629-66608805	66608709	66608728	+	28.032787	1.198806e-11	2.380320e-06	TGGCCACCAGGTGGCGCCG	5088	ref
10	MA0139.1	CTCF	chr19:35067213-35067372	35067283	35067302	+	27.918033	1.594210e-11	2.559579e-06	TGGCCACCAGGGGGCACTG	5096	ref
11	MA0139.1	CTCF	chr15:34210116-34210292	34210206	34210187	-	27.918033	1.594210e-11	2.559579e-06	TGGCCACCAGGGGGCACTG	787	non.ref
12	MA0139.1	CTCF	chr9:130461251-130461422	130461337	130461356	+	27.885246	1.718782e-11	2.559579e-06	TGGCCACCAGAGGGCGCCA	5096	ref
13	MA0139.1	CTCF	chr1:7647328-7647487	7647424	7647405	-	27.803279	2.111354e-11	2.695020e-06	TGGCCACCAGGTGGCGCTG	5096	ref
14	MA0139.1	CTCF	chr16:66608629-66608805	66608709	66608728	+	27.803279	2.111354e-11	2.695020e-06	TGGCCACCAGGTGGCGCTG	8	non.ref
15	MA0139.1	CTCF	chr3:120558991-120559170	120559086	120559067	-	27.672131	2.623702e-11	3.060687e-06	TGGCCAGCAGGGGGCGCTC	5052	ref
16	MA0139.1	CTCF	chr19:4809674-4809833	4809756	4809775	+	27.590164	2.945542e-11	3.060687e-06	TGGCCACCAGAGGGCGCTG	5096	ref
17	MA0139.1	CTCF	chr20:44685269-44685461	44685354	44685373	+	27.573770	3.082922e-11	3.060687e-06	TGGCCAGCAGGGGGCGGTG	5091	ref
18	MA0139.1	CTCF	chr19:47863504-47863679	47863606	47863587	-	27.573770	3.082922e-11	3.060687e-06	TGGCCACTAGGGGGCGGCA	5095	ref
19	MA0139.1	CTCF	chr10:123994704-123994883	123994783	123994802	+	27.491803	3.572712e-11	3.314464e-06	TGGCCACCAGGGGGCAGCG	2	non.ref
20	MA0139.1	CTCF	chr5:177501044-177501224	177501155	177501136	-	27.459016	3.709491e-11	3.314464e-06	CGGCCACCAGAGGGCGCTG	5096	ref
21	MA0139.1	CTCF	chr10:123994704-123994883	123994783	123994802	+	27.360656	4.440797e-11	3.778945e-06	CGGCCACCAGGGGGCAGCG	5086	ref
22	MA0139.1	CTCF	chr20:49979066-49979252	49979155	49979174	+	27.229508	5.882201e-11	4.509085e-06	TGGCCAGCAGAGGGCGCCA	4789	ref
23	MA0139.1	CTCF	chr12:111397009-111397212	111397090	111397109	+	27.213115	6.217717e-11	4.509085e-06	CAGCCACCAGGGGGCGCCA	5096	ref
24	MA0139.1	CTCF	chr20:63973975-63974156	63974071	63974052	-	27.131148	7.240292e-11	4.509085e-06	CGGCCACCAGGGGGCAGTG	5094	ref
25	MA0139.1	CTCF	chr10:123994704-123994883	123994783	123994802	+	27.131148	7.240292e-11	4.509085e-06	CGGCCACCAGGGGGCAGTG	2	non.ref

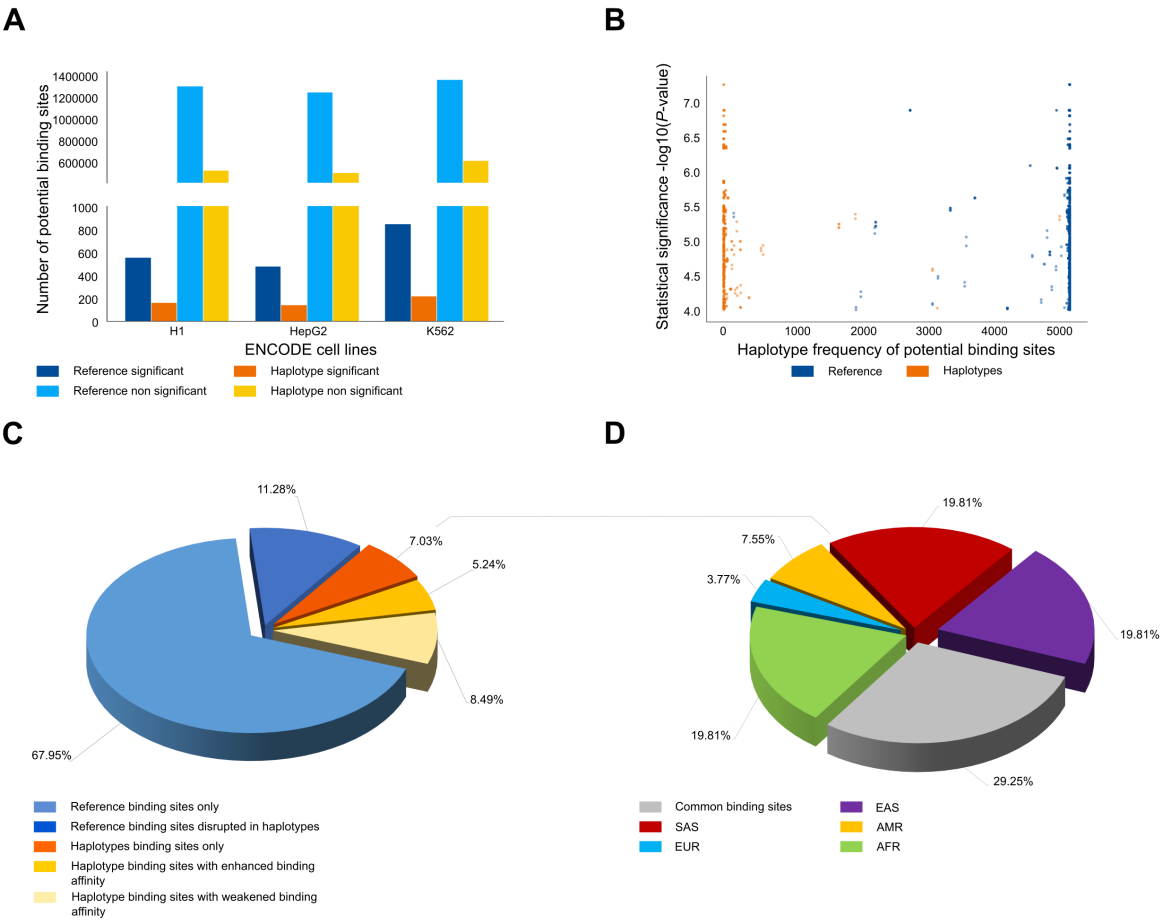
S2 Fig. Example of HTML summary report. The HTML report shows the first 25 potential CTCF occurrences retrieved by GRAFIMO, searching the motif in ChIP-seq peak regions defined in ENCODE experiment ENCFF816XLY (cell line A549).



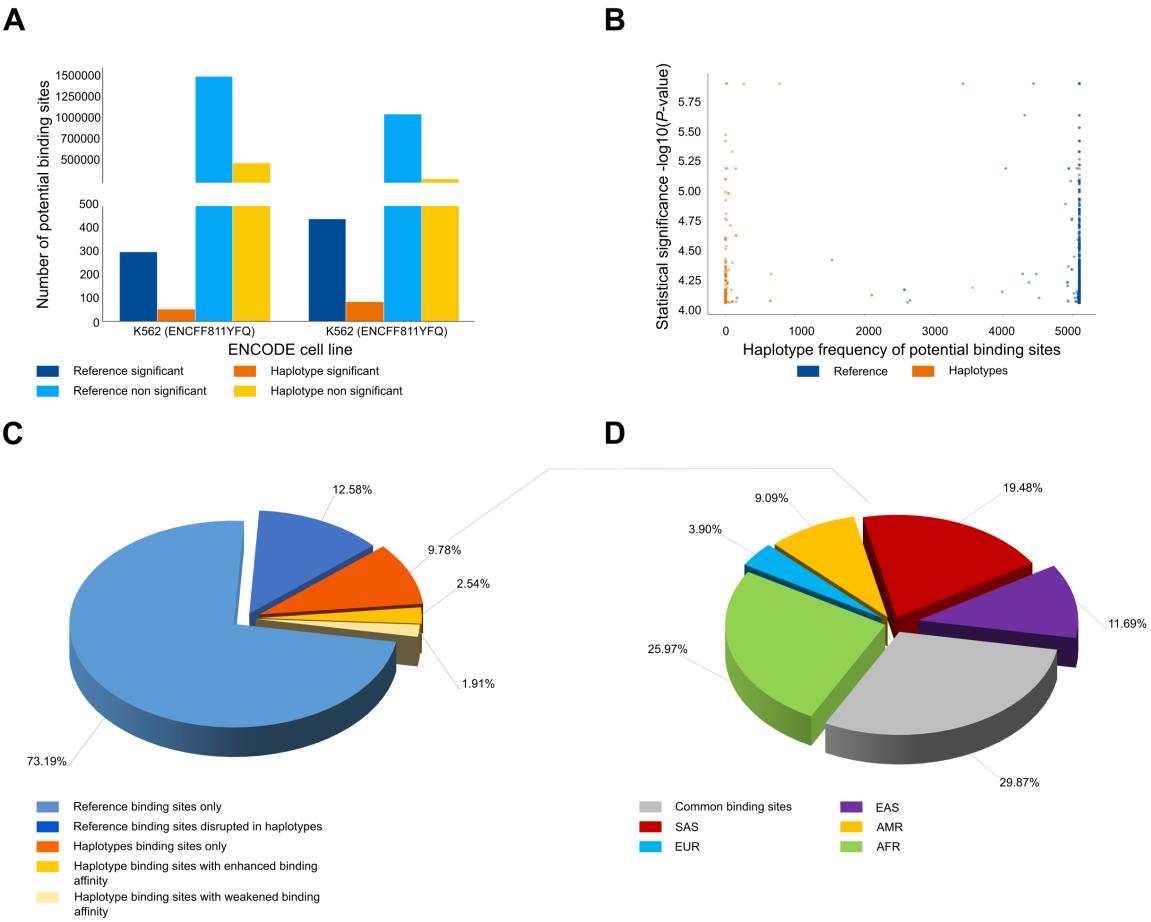
S3 Fig. Example of GFF3 track produced by GRAFIMO, loaded on the UCSC genome browser. GRAFIMO returns also a GFF3 report which can be loaded on the UCSC genome browser; the loaded custom track shows three potential CTCF occurrences (region chr8:142,782,661-142,782,680) retrieved by GRAFIMO overlapping a dbSNP annotated variant (rs892844).



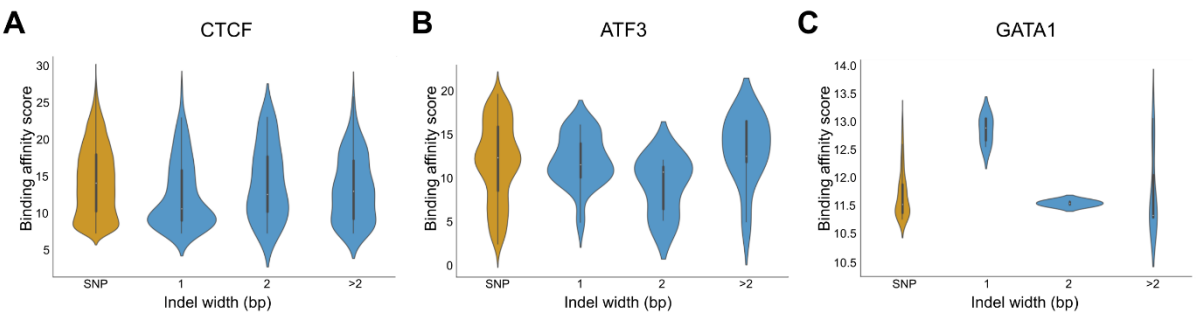
S4 Fig. Structure of transcription factor motifs used to test GRAFIMO. Transcription factor binding site motifs of (A) CTCF, (B) ATF3 and (C) GATA1.



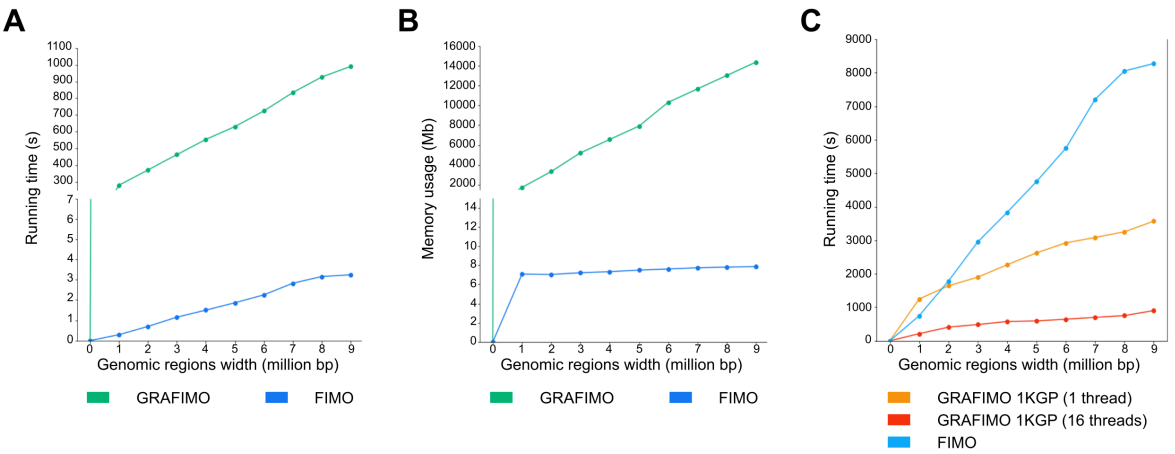
S5 Fig. Searching ATF3 motif on VG with GRAFIMO provides an insight on how genetic variation affects putative binding sites. (A) Potential ATF3 occurrences statistically significant (P-value < 1e-4) and non-significant found in the reference and in the haplotype sequences found with GRAFIMO oh hg38 1000GP VG. (B) Statistical significance of retrieved potential ATF3 motif occurrences and their frequency in the haplotypes embedded in the VG. (C) Percentage of statistically significant ATF3 potential binding sites found only in the reference genome or alternative haplotypes and with modulated binding scores based on 1000GP genetic variants. (D) Percentage of population specific and common (shared by two or more populations) potential ATF3 binding sites present on individual haplotypes.



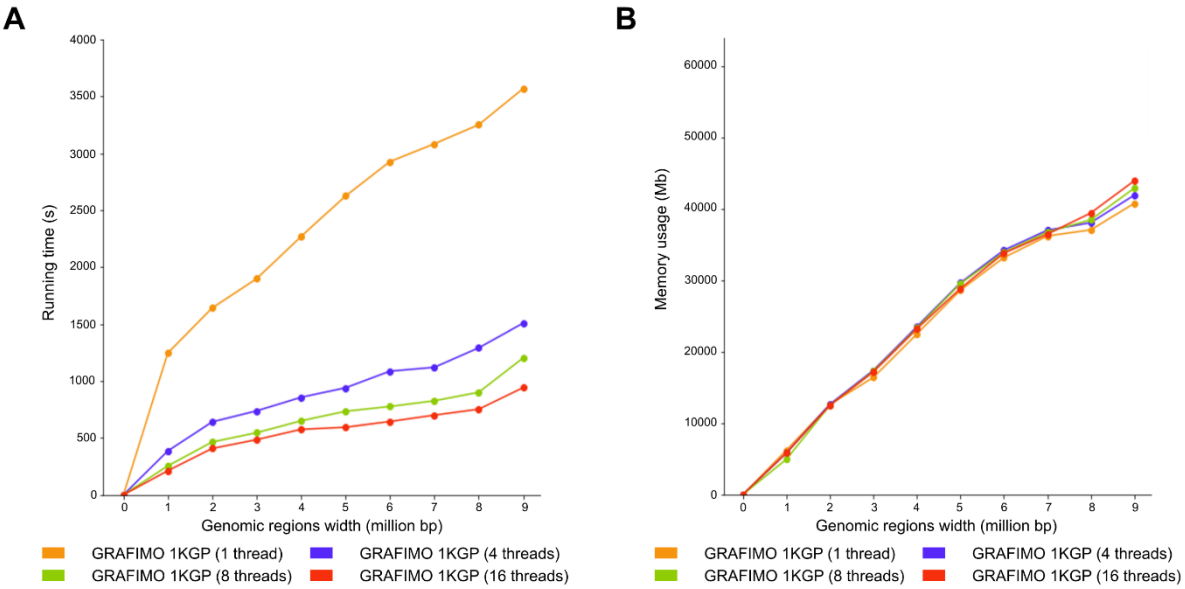
S6 Fig. Searching GATA1 motif on VG with GRAFIMO provides an insight on how genetic variation affects putative binding sites. (A) Potential GATA1 occurrences statistically significant (P-value < 1e-4) and non-significant found in the reference and in the haplotype sequences found with GRAFIMO oh hg38 1000GP VG. (B) Statistical significance of retrieved potential GATA1 motif occurrences and their frequency in the haplotypes embedded in the VG. (C). Percentage of statistically significant GATA1 potential binding sites found only in the reference genome or alternative haplotypes and with modulated binding scores based on 1000GP genetic variants. (D) Percentage of population specific and common (shared by two or more populations) potential GATA1 binding sites present on individual haplotypes.



S7 Fig. Influence of different type of mutations and length on binding affinity score. (A) CTCF, (B) ATF3, (C) GATA1.



S8 Fig. Comparing GRAFIMO and FIMO performance. (A) FIMO is faster than GRAFIMO (using a single thread) when searching CTCF motif (JASPAR ID MA0139.1) on human chr22 regions (total width ranging from 1 to 9 millions of bp) and without accounting for genetic variants. (B) FIMO uses less memory resources than GRAFIMO, however it can only scan linear sequences. (C) GRAFIMO is generally faster than FIMO in searching potential CTCF occurrences after 1Mbp when considering genetic variation present in large panels of individuals as 1000GP on GRCh38 phase 3 (2548 samples), even with a single thread.



S9 Fig. GRAFIMO running time efficiently scales with the number of threads used. By running GRAFIMO with multiple threads (A) the running time significantly decreases, while (B) memory usage remains similar.

Table A. Number of genomic variants used to test GRAFIMO. Number of genomic variants used to test GRAFIMO, divided by chromosome. The variants were obtained from 1000 Genomes Project on GRCh38 phase 3 and belongs to 2548 individuals, from 26 populations. The number of variants refers to SNPs and indels together. In total were considered ~78 million variants.

Chromosome	Number of SNPs and indels
Chr1	6,191,833
Chr2	6,790,551
Chr3	5,641,493
Chr4	5,477,810
Chr5	5,115,036
Chr6	4,863,337
Chr7	4,511,408
Chr8	4,425,449
Chr9	3,384,360
Chr10	3,874,259
Chr11	3,881,791
Chr12	3,745,465
Chr13	2,760,845
Chr14	2,548,903
Chr15	2,301,453
Chr16	2,548,920
Chr17	2,209,149
Chr18	2,189,529
Chr19	1,738,824
Chr20	1,817,492
Chr21	1,045,269
Chr22	1,059,079
ChrX	106,963

Table B. ENCODE ChIP-seq experiment codes. To test our software, we searched the potential occurrences of three transcription factor motifs (CTCF, ATF3 and GATA1) in a hg38 pangenome variation graph enriched with genomic variants and haplotypes of 2548 individuals from 1000 Genomes project phase 3. In order to have likely to happen binding events, the motifs have been searched in ChIP-seq peak regions, obtained from the ENCODE project data portal.

Motif	Cell line A549	Cell line GM12878	Cell line H1	Cell line HepG2	Cell line K562	Cell line MCF-7
CTCF	ENCFF816XLT	ENCFF267NYF		ENCFF015OJG	ENCFF895HAG	ENCFF088JWU
ATF3			ENCFF207AVV	ENCFF753WNT	ENCFF787GVU	
GATA1					ENCFF811YFQ	
					ENCFF939ODZ	

References

1. Garrison E, Sirén J, Novak AM, Hickey G, Eizenga JM, Dawson ET, et al. Variation graph toolkit improves read mapping by representing genetic variation in the reference. Nature biotechnology. 2018;36(9): 875—879.
2. Zheng-Bradley X, Streeter I, Fairley S, Richardson D, Clarke L, Flicek P, et al. Alignment of 1000 Genomes Project reads to reference assembly GRCh38. GigaScience. 2017;6(7): 1—8.
3. Lowy-Gallego E, Fairley S, Zheng-Bradley X, Ruffier M, Clarke L, Flicek P. Variant Calling on the GRCh38 assembly with the data from phase three of the 1000 Genomes Project. Wellcome Open Research, 2019;4.
4. ENCODE Project Consortium. An Integrated encyclopedia of DNA elements in the human genome. Nature. 2012;489(7414): 57—74.
5. Davis CA, Hitz BC, Sloan CA, Chan ET, Davidson JM, Gabdank I, et al. The encyclopedia of DNA elements (ENCODE): data portal update. Nucleic Acid Research. 2018;46(D1): D794—D801.
6. Kent WJ, Zweig AS, Barber G, Hinrichs AS, Karolchik D. BigWig and BigBed: enabling browsing of large distributed datasets. Bioinformatics. 2010;26(17): 2204—2207.
7. Ishihara K, Oshimura M, Nakao M. CTCF-dependent chromatin insulator is linked to epigenetic remodeling. Molecular Cell. 2006;23(5): 733—742.
8. Fiorentino FP, Giordano A. The tumor suppressor role of CTCF. Journal of cellular physiology. 2012;227(2): 479—492.
9. Chen BP, Wolfgang CD, Hai T. Analysis of ATF3, a transcription factor induced by physiological stresses and modulated by gadd153/Chop10. Molecular and cellular biology. 1996;16(3): 1157—1168.
10. Thompson MR, Xu D, Williams BRG. ATF3 Transcription factor and its emerging roles in immunity and cancer. Journal of molecular medicine. 2009;87(11): 1053.
11. Calligaris R, Bottardi S, Cogoi S, Apezteguia I, Santoro C. Alternative translation initiation site usage results in two functionally distinct forms of the gata-1 transcription factor. Proceedings of the National Academy of Sciences. 1995;92(25): 11598—11602.
12. Grant CE, Bailey TL, Noble WS. Fimo: scanning for occurrences of a given motif. Bioinformatics. 2011;27(7): 1017—1018.
13. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26(6): 841—842.