Supplementary Materials for 'GLANET: Genomic Loci ANnotation and Enrichment Tool'

Burçak Otlu [1], Can Firtina[2], Sündüz Keleş [3,4], & Oznur Tastan[2]
[1] Department of Computer Engineering, Middle East Technical University, Ankara, Turkey
[2] Department of Computer Engineering, Bilkent University, Ankara, Turkey
[3] Department of Statistics, University of Wisconsin, Madison, WI, U.S.A.
[4] Department of Biostatistics and Medical Informatics, University of Wisconsin, Madison, WI, U.S.A.

April 19, 2017

1 Overview of Available Tools

There are various tools that provide enrichment and/or annotation analysis on given genomic intervals. In Supplementary Table 1, we compare some of the available tools.

Tool	Version	SNPs	Genomic Intervals	Format	ENCODE	Roadmap Epigenomics	1000 Genomes	Predefined Gene-Sets	Genes	Other Data Sources	Allows user provided annotation libraries	Statistical Model or Test	Takes into account Genomic Biases	Correction for Multiple Hypothesis Testing	Provides Enrichment	DNA Elements	Predefined Gene-Set Elements	Others	User Interface
RegulomeDB	v1.1	✓	✓	dbSNP Ids, VCF, BED, GFF3	✓	✓			Gencode v7	dbSNP, GEO, published literature, eQTL, dsQTL, predicted annotations, DNase footprinting, PWMs, DNA Methylation									Web
SnpEff	v4.2	✓	✓	VCF, TXT, SAMTools Pileup Format	✓	✓		KEGG, GO, MSigDb,	Ensembl	NextProd, UCSC, Motif annotations	✓								Command Line
Ensembl SNP Effect Predictor (VEP)	Ensembl release 83	✓	✓	VCF, Pileup, HGVS notations			✓		RefSeq, Ensembl, Gencode	Ensembl transcripts, Gencode and RefSeq transcripts	✓								Web
ANNOVAR		✓	✓	VCF, GFF3	✓		✓		RefSeq, UCSC, ENSEMBL, GENCODE, AceView	dbSNP, SIFT, UCSC regions, OMIM, Exome Sequencing Project, MutationTaster, Polyphen, Complete Genomics and etc.	✓								Command Line
FunciSNP	v1.12.0	✓		dbSNP Ids	✓	✓	✓		UCSC known genes	TCGA, Faire sites, DNaseI hypersensitive sites			LD						Command Line
HaploReg	v4.1	✓	*	dbSNP Ids	✓	✓	✓		RefSeq, Gencode	dbSNP, eQTL, motif instances		Binomial Test	LD		✓	DNaseI hypersensitive sites		Enhancer	Web
ALIGATOR		✓		dbSNP Ids				GO	Entrez	dbSNP		Permutation Approach	LD	Bootstrap Approach	✓		GO		Line
Annotate-it	v0.4	✓		VCF			✓	KEGG, GO, BIOCARTA, Reactome	Ensembl	Polyphen2, SIFT, LRT, MutationTaster, Anatomical gene expression (eGenetics/SANBI dataset), HPO, EPCC and LDDB phenotype ontologies to annotate samples		Filter-based Approach, Weighted sum Approach, Gamma-based approximation for the null distribution of weighted sum statistic			✓			Annotate-it provides candidate gene lists, aggregate functionality scores, phenotype-specific gene prioritization, and statistical methods for disease-gene finding in case/control studies.	Web
Encode ChiP-Seq Significance Tool				user given gene list	✓				Ensembl, Gencode, Entrez	HAVANA, HUGO Gene Nomenclature Committee		One-tailed Hypergeometric Test		Benjamini-Hochberg FDR	✓	Transcription Factors			Web
PANOGA		✓		dbSNP Ids				KEGG		Protein-Protein Interaction Data		Two-sided test based on the hypergeometric distribution	LD	Bonferroni Correction	✓			Identified sub-networks within protein-protein interaction networks	Web
FORGE	v1.1	✓		dbSNP Ids, VCF, BED	✓	✓	✓			GEO, omni genotyping arrays, GWAS snp arrays		Background Distribution, Z score	LD, GC, minor allele frequency (maf) and distance to the nearest transcription start site (TSS)	Bonferroni Correction	✓	Dnasel hotspots		Cell Type Specific Enrichment	Web
Variant Tools	v2.7.0	✓	✓	dbSNP Ids, VCF, BED, GFF3, CSV, Plink	✓		✓	KEGG		dbSNP, Exome Sequencing Project, dbNSFP, UCSC, HapMap project, GWAS catalog	✓	Fisher's Exact Test for Single Variant Analysis, Single gene rare variant tests, Conditional rare variants analysis and etc.			✓			Use more than 20 association analysis methods to associate variants and genes with qualitative or quantitative traits.	Command Line
GEMINI	v0.18.2	✓	✓	VCF	✓		✓	KEGG		dbSNP, ClinVar, UCSC, OMIM, HPRD, Exome Sequencing Project	✓	Built-in analyses such as find de novo mutations, find compound heterozygotes and so on						You can write your own SQL queries.	Command Line
GREAT	v3.0.0		✓	BED				GO, MSigDb, Panther, BioCyc	Ensembl genes	Uses 20 ontologies including disease ontologies, phenotype ontologies, miRNA motifs, miRNA targets		Binomial Test, Hypergeometric Test		Bonferroni Correction, Benjamini-Hochberg FDR	✓	Transcription Factors	GO, MSigDb, Panther, BioCyc	Gene Expression Data, Regulatory Motifs, Gene Families	Web
INRICH	v1.1	✓	✓	dbSNP Ids				KEGG, GO, MSigDb	Entrez			Permutation Approach	LD	Bootstrap Approach	✓		KEGG, GO, MSigDb		GUI, Command Line
GAT	v1.2.2		✓	BED							✓	Sampling method	Chromosome Identity, GC and Mappability (Not tailored for each given interval)	Storey's q-value, Benjamini–Hochberg FDR	✓				Commad Line
GLANET	v1.0	✓	✓	dbSNP Ids, BED, GFF3	✓			KEGG	RefSeq		✓	Sampling Based Approach, Z score	GC, Mappability, Interval Length, Interval Chromosome	Bonferroni Correction, Benjamini-Hochberg FDR	✓	DNaseI hypersensitive sites, Histone Modifications, Transcription Factors	KEGG	User defined gene-set enrichment, user defined library enrichment	Command Line, GUI

*Only accepts a single region

Input | **Annotation Libraries** | **Statistical Tests** | **Enrichment Elements**

Supplementary Table 1. Available tools including GLANET are compared with respect to their input type, annotation and enrichment options, libraries utilized and statistical tests carried out.

2 Random Interval Sampling Procedure

To perform enrichment analysis, GLANET generates a null distribution of the test statistics by first sampling random intervals and calculating these intervals' overlap with the annotation library element intervals. The random intervals are generated such that they match properties of the each member of the input interval set as opposed to the average properties of these intervals. The algorithm for generating random intervals is outlined in Algorithm 1. Note that, we do not include the relaxation steps of the thresholds for sake of clarity. Here we provide the details of this random interval generation scheme.

The input interval set may contain overlapping intervals. In such cases, GLANET preprocesses the input by merging overlapping intervals into a single interval to avoid dependency within them. Similarly, the random intervals for an input interval set are always selected such that they do not overlap. GLANET provides four main parameters for random interval generation: `with GC (wGC)` , `with Mappability (wM)`, `with GC and Mappability (wGCM)`, and `without GC and Mappability (woGCM)`. GLANET random interval generation can also be run `with Isochore Family (wIF)` and `without Isochore Family (woIF)` (as explained in the main text). Regardless of which option is selected, for each input interval a corresponding random interval of the same length from the same chromosome is sampled. When the given interval's length is greater than 100,000 bps, GLANET does not generate random intervals by accounting for GC content and/or mappability even one of these options (`wGC`,`wM`,`wGCM`) is on. Since for very large intervals, GC content and mappabilty values are not meaningful. In case of `wGC`, `wM`, or `wGCM` options are selected in addition to the length and chromosome of given interval, GLANET also matches given interval's GC, mappability, or both GC and mappability, respectively as follows:

- **GC Option or Mappability Option Selected:** If one of the `wGC` or `wM` option is selected, GLANET tries to match the GC content or mappability value of the given interval. Same procedure applies for matching GC or mappability values. GLANET first generates a random interval and calculates its GC content or mappability depending on which option is selected. This random interval is accepted if its value is close to the corresponding value of input interval within a pre-defined threshold. Otherwise, GLANET generates a new random interval until an acceptable random interval is obtained. If after a certain number of attempts, no random interval can be found because it is not within the threshold distance to the GC or mappability of the input interval, then the threshold for the acceptable match is increased by a small increment. Again, after a certain amount of trials, if relaxing this threshold does not help, GLANET chooses the random interval with the minimum difference in GC content or mappability up to that point.
- **GC and Mappability Option Selected:** If `wGCM` option is on, GLANET selects a random interval with close GC content and mappability values to the input interval. A random interval is considered acceptable if its GC content and mappability values are within a pre-defined distance to the input interval's values. If the random interval values do not match, a new interval is sampled until an acceptable random interval is obtained. If after a certain number of attempts, no random interval can be generated because it is not within the threshold distance to the GC or mappability of the input interval, the threshold for the acceptable match is increased by a small increment. If relaxing this threshold does not help, GLANET chooses the random interval with the minimum sum of the differences in GC content and mappability up to that point.

GC and Mappability Calculation: In order to calculate the GC content and mappability of given intervals, GLANET pre-computes GC content and mappability values of genomic regions and stores them in the disk. The GC content of the genomic regions are calculated at various lengths such as 1 bp, 100 bps, 1000 bps, 10,000 bps and 100,000 bps. In runtime GLANET constructs a GC interval tree from one of these pre-computed GC content values based on the mode of the input interval lengths. Specifically, the shorter the input intervals are, the more precise the GC calculation is. If mode of given intervals' lengths is short (<= 100 bps long), GLANET calculates GC content of the given intervals at one base resolution and stores them in a byte list. Otherwise, GLANET stores GC contents of 100 bps, 1000 bps and 10000 bps long intervals in interval trees. When the mode is between ($>$ 100 and $\leq$ 1000) GLANET calculates GC content at 100 base resolution, if the mode is ($>$ 1000 and $\leq$ 10,000) GLANET calculates at 1000 base resolution. For cases between ($>$ 10,000 and $\leq$ 100,000) at 10,000 base resolution and when mode gets longer than (100,000 bps) then GLANET does not calculate GC content for intervals longer than (100,000 bps) but only for intervals shorter than (100,000 bps) at 10,000 base resolution. Mappabilities of genomic intervals are obtained from ENCODE, the source files are listed in Supplementary Table 2. A query interval can be part of a single interval or overlap with multiple intervals with different mappability values as provided in the original source. In either case, its mappability is estimated by calculating the weighted average, where the weights are the proportions of the query interval lengths that overlap with the source mappability interval.

GLANET also offers matching isochore family of the given interval: The genome is divided into five regions that are characterized by similar GC content composition. These regions are called isochores and are named as L1, L2, H1, H2, and H3 in accordance with increasing GC levels, <38%, 38–42%, 42–47%, 47–52%, >52%, respectively as defined in [1, 2]. Finally, each chromosome is divided into 100,000 bps long intervals and each such interval is tagged with its appropriate isochore family. When `wIF` option is selected, initially, input interval's isochore family is calculated and a random interval of 100,000 bps long is selected from the appropriate isochore family pool of that chromosome. Subsequently, a random interval of input interval's length is sampled from this 100,000 bps long interval.

Algorithm 1: Generating a random sample for a given set of genomic intervals.

```
Function: Generate Random Genomic Intervals (S, wGC, wM, wGCM, wIF);
  Input : S: Set of n input intervals.
          wIF: If true, isochore family pools will be used in random interval generation.
          t_M: Threshold to match mappability within this value.
          t_GC: Threshold to match GC content within this value.
          t_Value: Stands for t_M or t_GC.
          LMAX: Maximum interval length GC and mappability will be accounted for (Default is 100,000 bps).
for each chromosome chr_i do
    S_i ←subset of intervals in S that are on chr_i
    if S_i ≠∅ then
        for each sampling b in {1, ..., B} do
            S_i^(b) ← ∅
            for each given interval g in S_i do
                gLen ← length(g)
                if gLen ≤ LMAX then
                    if wGCM then
                        gGC ← calculateGC(g)
                        gM ← calculateMappability(g)
                        do
                            do
                                if wIF then
                                    gIF ← findIsochoreFamily(g)
                                    r ← getARandomInterval(chr_i, gLen, gIF)
                                else
                                    r ← getARandomInterval(chr_i, gLen)
                                end
                            while r overlaps with an already generated interval in S_i^(b);
                            rGC ← calculateGC(r)
                            rM ← calculateMappability(r)
                        while (|rGC − gGC| > t_GC) or (|rM − gM| > t_M);
                    else
                        if wGC or wM then
                            gValue ← calculateGC(g) or calculateMappability(g)
                            do
                                do
                                    if wIF then
                                        gIF ← findIsochoreFamily(g)
                                        r ← getARandomInterval(chr_i, gLen, gIF)
                                    else
                                        r ← getARandomInterval(chr_i, gLen)
                                    end
                                while r overlaps with an already generated interval in S_i^(b);
                                rValue ← calculateGC(r) or calculateMappability(r)
                            while (|rValue − gValue| > t_Value);
                        end
                    end
                end
                if gLen > LMAX or woGCM then
                    do
                        r ← getARandomInterval(chr_i, gLen)
                    while r overlaps with an already generated interval in S_i^(b);
                end
                S_i^(b) ← S_i^(b) ∪ r
            end
        end
    end
end
```

3 GLANET Data Sources

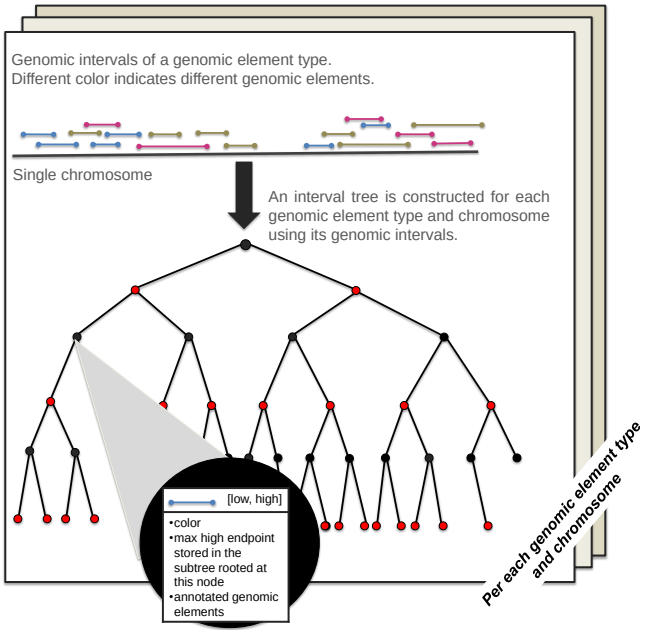
Data sources are provided in Supplementary Table 2.

Data	Source	Download Date
ENCODE DNaseI hypersensitive sites	http://ftp.ebi.ac.uk/pub/databases/ensembl/encode/supplementary/integration_data_jan2011/byDataType/openchrom/jan2011/idrPeaks/conservative/	29/03/2013
ENCODE DNaseI hypersensitive sites	http://ftp.ebi.ac.uk/pub/databases/ensembl/encode/supplementary/integration_data_jan2011/byDataType/dnase/jul2010/	29/03/2013
ENCODE Transcription factor binding sites	http://ftp.ebi.ac.uk/pub/databases/ensembl/encode/supplementary/integration_data_jan2011/byDataType/peaks/jan2011/spp/optimal/	22/03/2013
ENCODE Histone modification sites	http://ftp.ebi.ac.uk/pub/databases/ensembl/encode/supplementary/integration_data_jan2011/byDataType/peaks/jan2011/histone_macs/optimal/	29/03/2013
hg19 RefSeq genes	http://genome.ucsc.edu/	18/11/2014
hg19 chromosome sizes	http://genome.ucsc.edu/goldenPath/help/hg19.chrom.sizes	22/05/2013
KEGG pathways	http://rest.kegg.jp/list/pathway/hsa	23/09/2013
KEGG pathway to gene mapping	http://www.genome.jp/linkdb/linkdb.html	18/06/2013
GC fasta files	http://hgdownload.cse.ucsc.edu/goldenPath/hg19/chromosomes/	19/07/2013
Mappability bigWig files	ftp://hgdownload.cse.ucsc.edu/apache/htdocs/goldenPath/hg19/encodeDCC/wgEncodeMapability/	18/07/2013
JASPAR CORE pfms	http://jaspar.genereg.net/html/DOWNLOAD/JASPAR_CORE/pfm/nonredundant/pfm_all.txt	26/08/2014
ENCODE motifs	http://compbio.mit.edu/encode-motifs/	25/02/2014
NCBI REMAP API supported assemblies	Downloaded by remap_api.pl within GLANET when a Regulatory Sequence Analysis is requisted (repmap_api.pl source: ftp://ftp.ncbi.nlm.nih.gov/pub/remap).	01/04/2016
Latest ref seq assembly ids	Downloaded from ftp://ftp.ncbi.nlm.nih.gov/genomes/ASSEMBLY_REPORTS/All/ within GLANET each time Regulatory Sequence Analysis is requested.	01/04/2016
Gene ids	ftp://ftp.ncbi.nlm.nih.gov/gene/DATA/gene2refseq.gz	18/11/2014

Supplementary Table 2. GLANET data sources and their download dates.

4 GLANET Library Representation

An interval tree is a red-black tree in which each node x stores the low and high endpoints, t_1 and t_2, of an interval and an integer value max which is the maximum high endpoint stored in the subtree rooted at this node x [3]. On each node of the tree, we also store the genomic annotations associated with the interval stored on that node. For each element type in the annotation library, e.g., genomic elements representing all transcription factor binding regions across all cell lines, chromosome-specific interval trees are constructed (Supplementary Figure 1). Then, for annotation and enrichment analysis the appropriate interval trees are searched for query intervals using the interval tree search algorithm as described in [3].



Supplementary Figure 1. Genomic intervals are represented in interval trees [3]. A separate interval tree is constructed for each chromosome and genomic element type, e.g. for transcription factor binding annotations. Each node contains the low and high endpoints of the genomic interval, the color of the node (red or black), the maximum high endpoint stored in the subtree rooted at this node and the genomic elements annotated with this particular genomic interval.

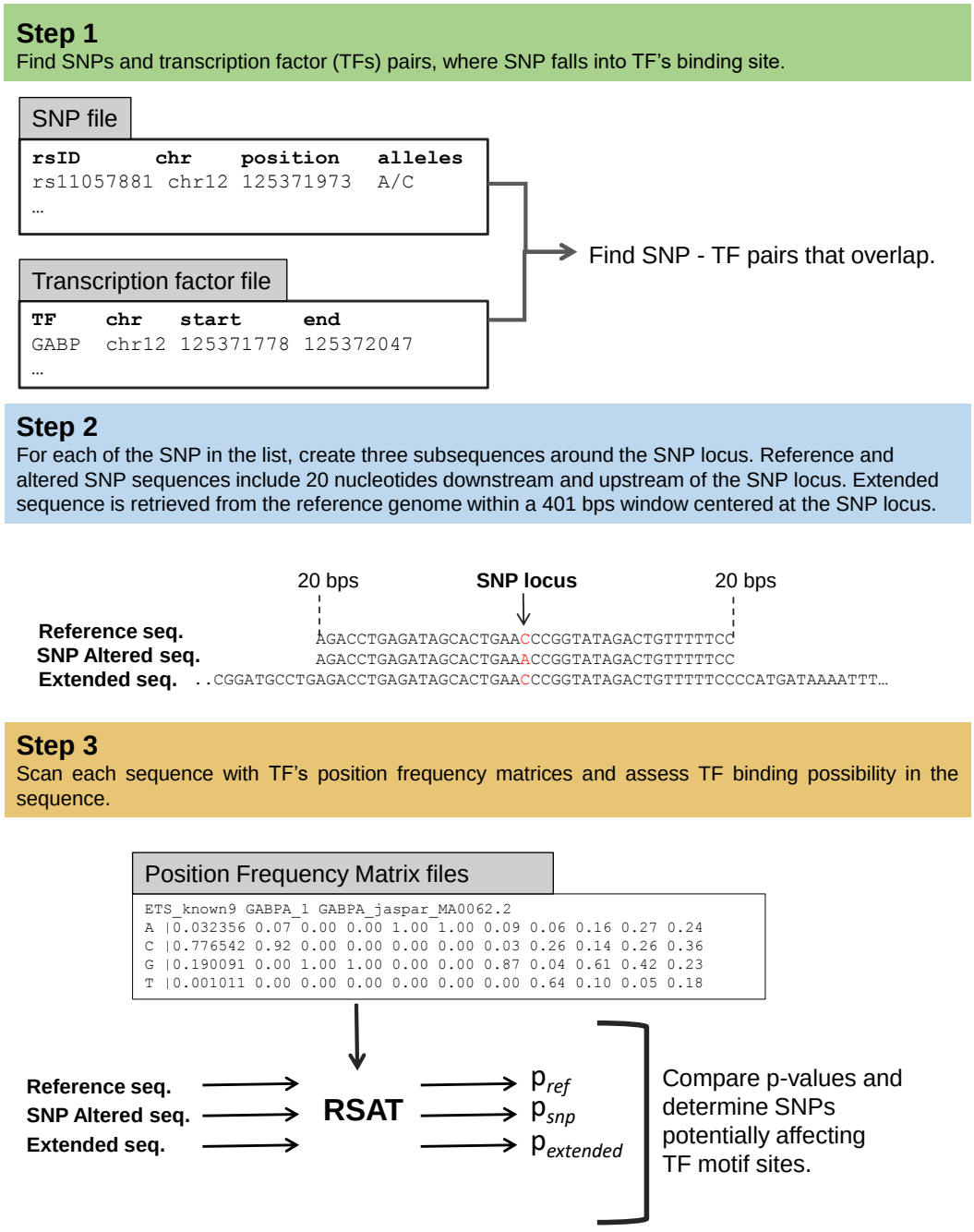
5 GLANET Regulatory Sequence Analysis

Specific to SNP inputs, GLANET offers regulatory sequence analysis (RSA). Regulatory sequence analysis takes advantage of the available ENCODE transcription factor binding regions in the default GLANET annotation library. This analysis can be conducted after an annotation analysis that involves transcription factor elements. GLANET performs regulatory sequence analysis in three main steps (depicted in Supplementary Figure 2):

In the first step, SNP and TF pairs for which SNP resides in the binding region of the TF are found. This is accomplished by overlapping the positions of the SNPs with transcription factor binding sites provided in the annotation library.

In the second step, GLANET generates three subsequences around the SNP site: reference, SNP and extended sequences. These sequences are used to statistically assess whether the SNP can alter the transcription factor binding. Reference and SNP sequences are 41 bps long and they are created by taking ±20 bps upstream and downstream sequences around SNP locus. Extended sequence is obtained by taking ±200 bps upstream and downstream sequences around SNP locus and is used to check if the SNP site is actually the most likely binding site in the vicinity of the SNP.

In the third step, GLANET scans the subsequences for a matching motif site in each of the sequences (Reference, SNP, Extended) and evaluate the statistical significance of the match using RSAT [4]. For this, the position frequency matrices (PFMs) for the annotated TFs are obtained from Jaspar Core and Encode motifs [5, 6]. This step results with three p-values: p_{ref}, p_{snp} and p_{extended}.



Supplementary Figure 2. Three main steps of regulatory sequence analysis in GLANET.

6 Data-driven Computational Experiments Results

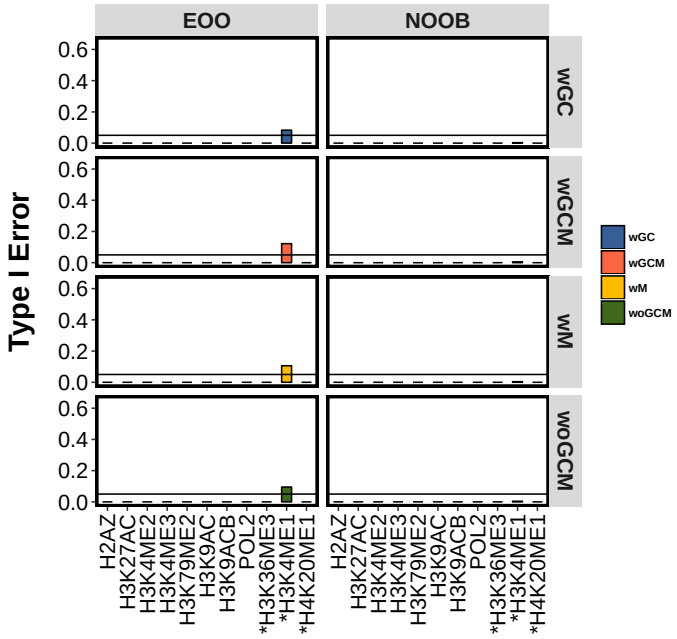
In this section, we provide the results of data-driven computational experiments for all GLANET parameter settings under different experiment settings. Results for activator elements are displayed in Supplementary Figures 3-10 and results for repressor elements are listed in Supplementary Tables 3-6.

- **Data-driven Computational Experiment Settings:** We conduct experiments in two cell lines, GM12878 and K562, with expressed genes under `Top5` and `Top20` and non-expressed genes under `CompletelyDiscard` and `TakeTheLongest` settings. Type-I error and power are reported for each activator and repressor element for all GLANET parameter settings at significance levels of $\alpha = 0.05$ and $\alpha = 0.001$.
- **GLANET Parameter Settings:** For each data-driven computational experiment setting, GLANET is run with all possible parameter combinations. We present data-driven computational experiment results for two different association measures: `Existence of Overlap (EOO)` and `Number of Overlapping Bases (NOOB)`, for four different null distribution generation modes: `with GC (wGC)`, `with Mappability (wM)`, `with GC and Mappability (wGCM)` and `without GC and Mappability (woGCM)`, and finally for two different isochore family options: `with Isochore Family (wIF)` and `without Isochore Family (woIF)`.

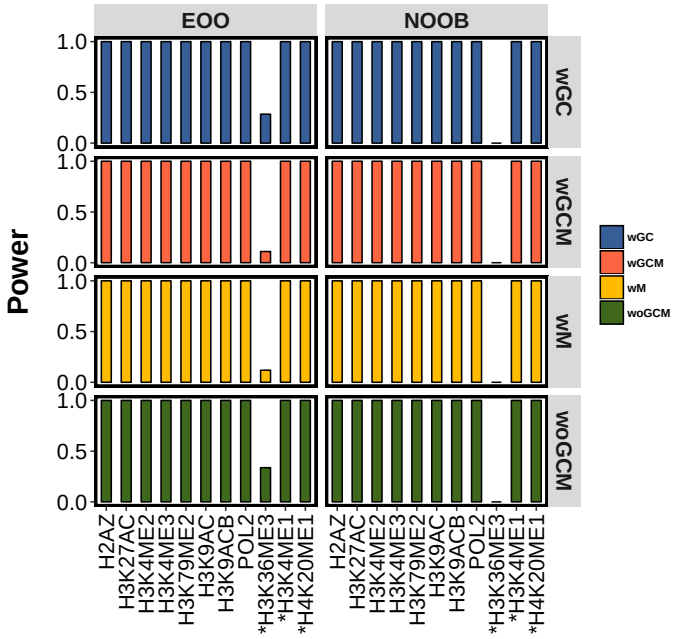
6.1 Data-driven Computational Experiments Results for Activator Elements

6.1.1 K562 NonExpressed (CompletelyDiscard), Expressed (Top5), $\alpha = 0.05$

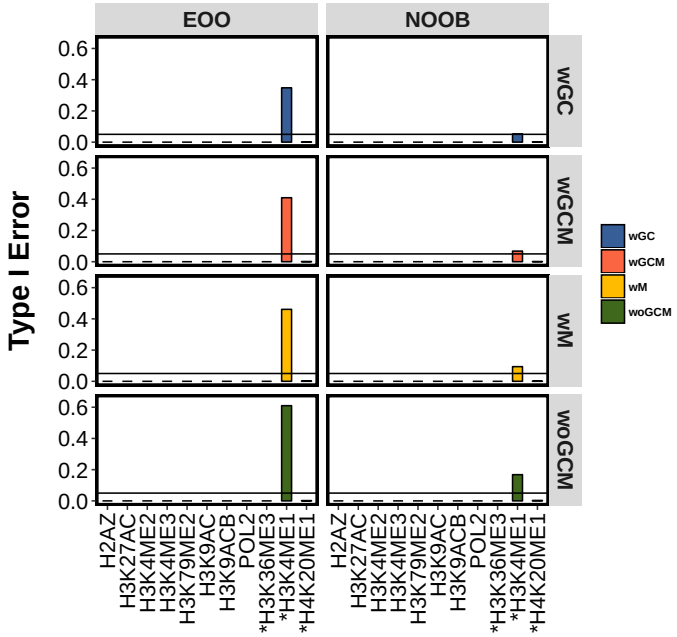
(a) K562, Non-expressed genes, $\alpha = 0.05$, wIF



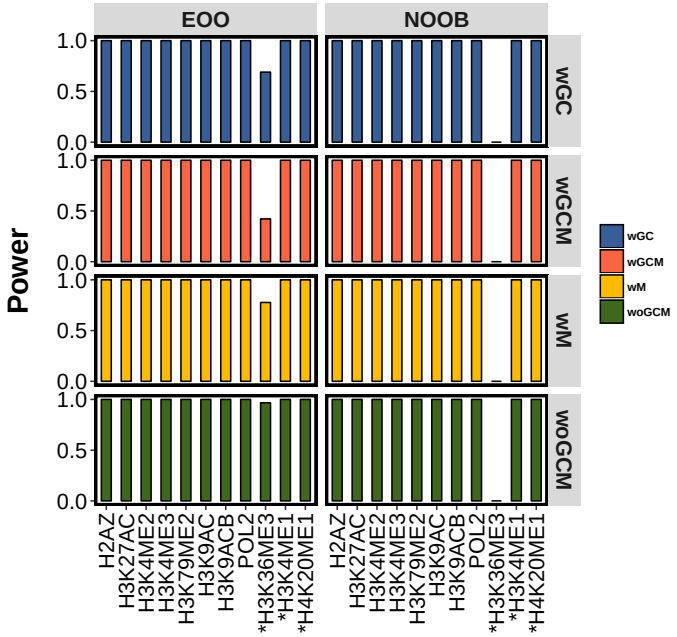
(b) K562, Expressed genes, $\alpha = 0.05$, wIF



(c) K562, Non-expressed genes, $\alpha = 0.05$, woIF



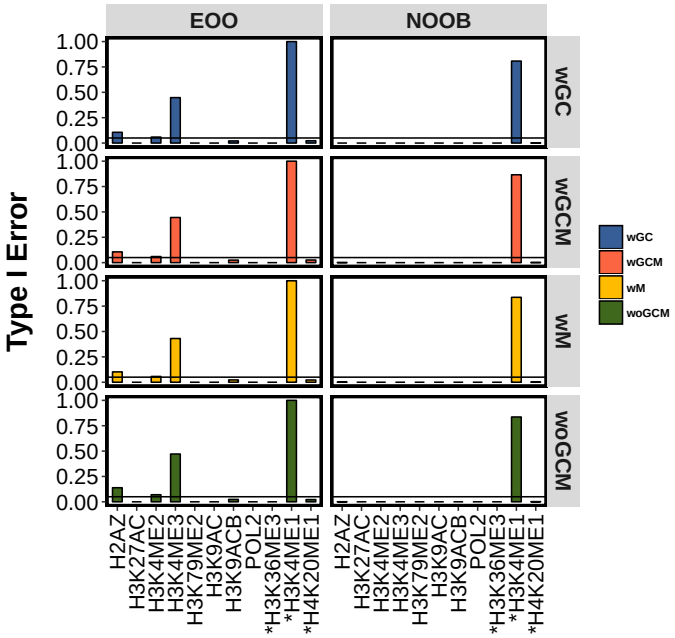
(d) K562, Expressed genes, $\alpha = 0.05$, woIF



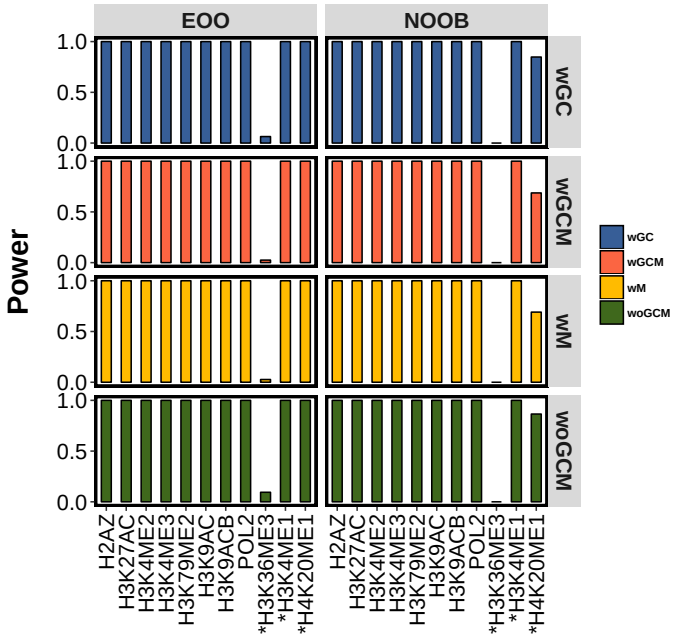
Supplementary Figure 3. Assessment of GLANET Type-I error and power with data-driven computational experiments. Results for the two association statistics - existence of overlap (EOO) and the number of overlapping bases (NOOB) - together with - with GC (wGC), with Mappability (wM), with GC and Mappability (wGCM), and without GC and mappability (woGCM) null distribution generation modes are displayed. Histone marks with ambiguous activator roles are marked with $*$. (a, b) Type-I error and power estimated with Isochore Family (wIF) heuristic using K562, (Non-expressed Genes,CompletelyDiscard) and (Expressed Genes,Top5) results, for significance level of 0.05. (c, d) Type-I error and power estimated without Isochore Family (woIF) heuristic using K562, (Non-expressed Genes,CompletelyDiscard) and (Expressed Genes,Top5) results, for significance level of 0.05.

6.1.2 K562 NonExpressed (TakeTheLongest), Expressed (Top20), $\alpha = 0.05$

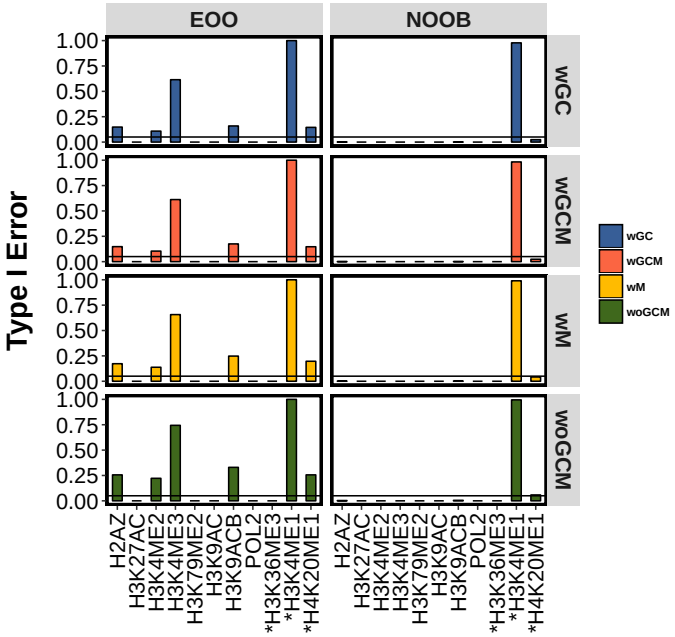
(a) K562, Non-expressed genes, $\alpha = 0.05$, wIF



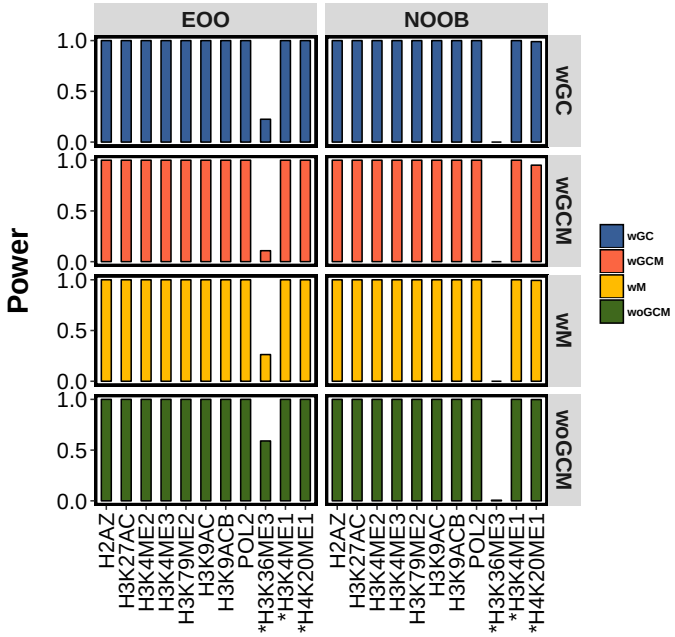
(b) K562, Expressed genes, $\alpha = 0.05$, wIF



(c) K562, Non-expressed genes, $\alpha = 0.05$, woIF



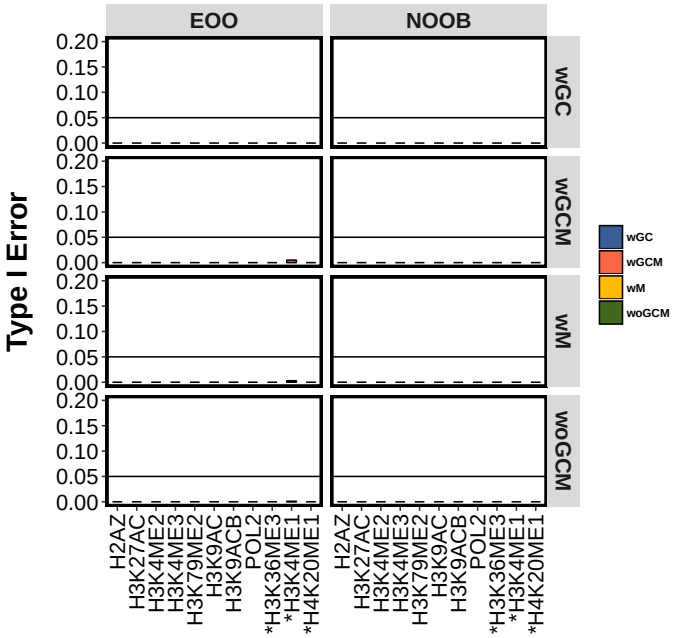
(d) K562, Expressed genes, $\alpha = 0.05$, woIF



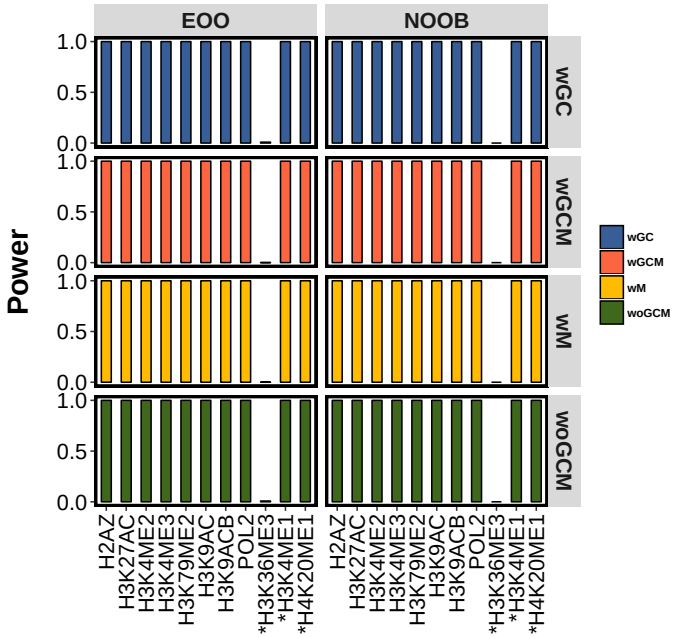
Supplementary Figure 4. Assessment of GLANET Type-I error and power with data-driven computational experiments. Results for the two association statistics - existence of overlap (EOO) and the number of overlapping bases (NOOB) - together with - with GC (wGC), with Mappability (wM), with GC and Mappability (wGCM), and without GC and mappability (woGCM) null distribution generation modes are displayed. Histone marks with ambiguous activator roles are marked with $*$. (a, b) Type-I error and power estimated with Isochore Family (wIF) heuristic using K562, (Non-expressed Genes,TakeTheLongest) and (Expressed Genes,Top20) results, for significance level of 0.05. (c, d) Type-I error and power estimated without Isochore Family (woIF) heuristic using K562, (Non-expressed Genes,TakeTheLongest) and (Expressed Genes,Top20) results, for significance level of 0.05.

6.1.3 K562 NonExpressed (CompletelyDiscard), Expressed (Top5), $\alpha = 0.001$

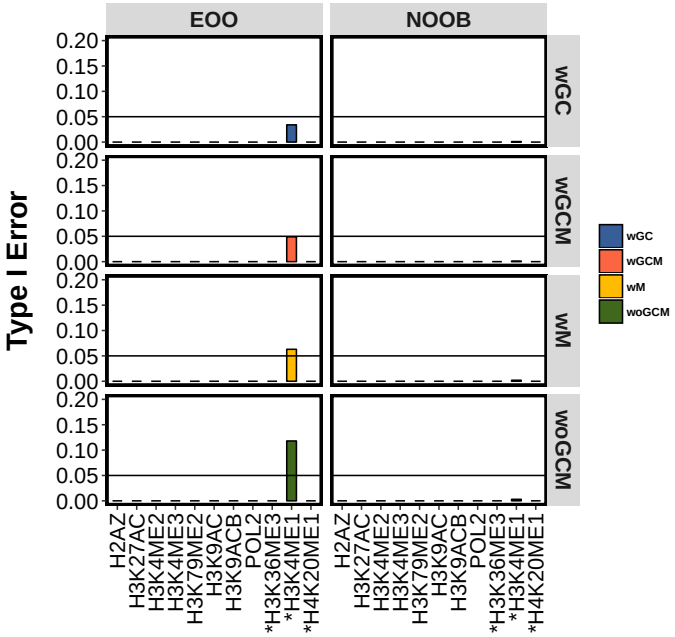
(a) K562, Non-expressed genes, $\alpha = 0.001$, wIF



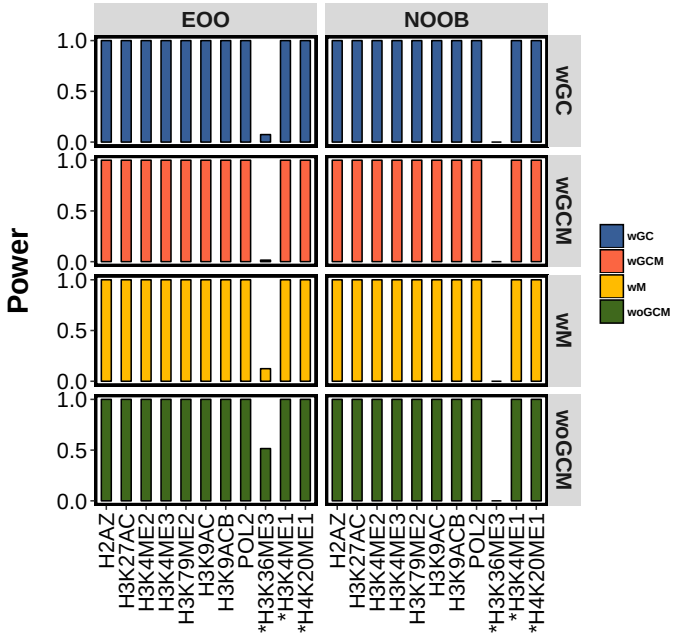
(b) K562, Expressed genes, $\alpha = 0.001$, wIF



(c) K562, Non-expressed genes, $\alpha = 0.001$, woIF



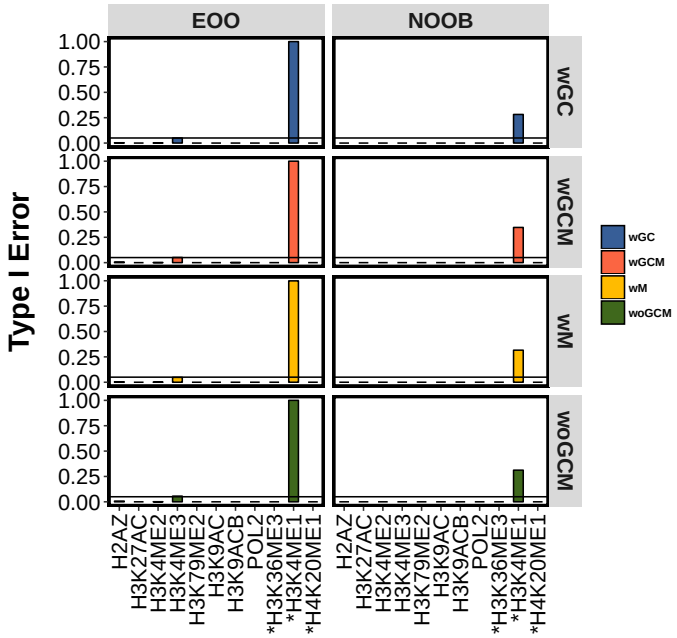
(d) K562, Expressed genes, $\alpha = 0.001$, woIF



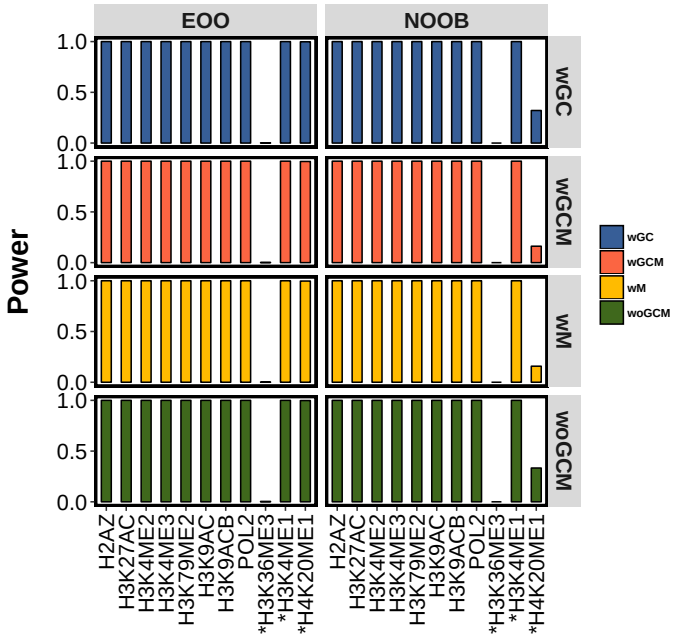
Supplementary Figure 5. Assessment of GLANET Type-I error and power with data-driven computational experiments. Results for the two association statistics - existence of overlap (EOO) and the number of overlapping bases (NOOB) - together with - with GC (wGC), with Mappability (wM), with GC and Mappability (wGCM), and without GC and mappability (woGCM) null distribution generation modes are displayed. Histone marks with ambiguous activator roles are marked with $*$. (a, b) Type-I error and power estimated with Isochore Family (wIF) heuristic using K562, (Non-expressed Genes,CompletelyDiscard) and (Expressed Genes,Top5) results, for significance level of 0.001. (c, d) Type-I error and power estimated without Isochore Family (woIF) heuristic using K562, (Non-expressed Genes,CompletelyDiscard) and (Expressed Genes,Top5) results, for significance level of 0.001.

6.1.4 K562 NonExpressed (TakeTheLongest), Expressed (Top20), $\alpha = 0.001$

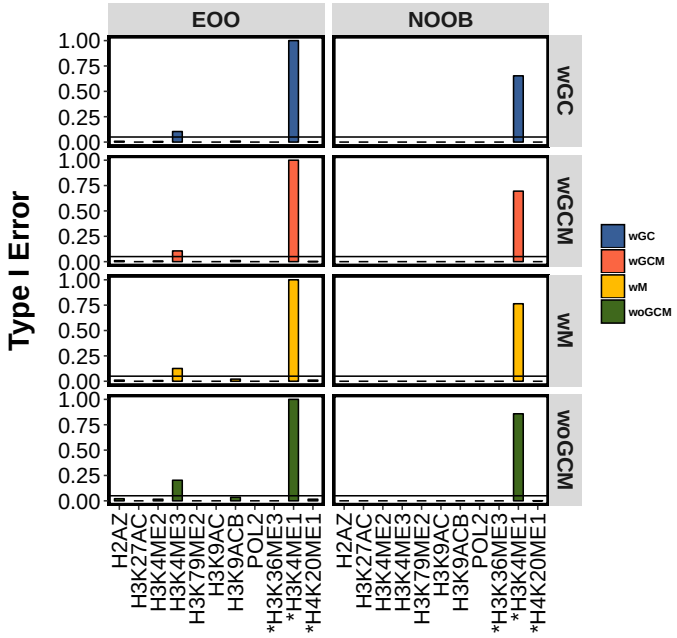
(a) K562, Non-expressed genes, $\alpha = 0.001$, wIF



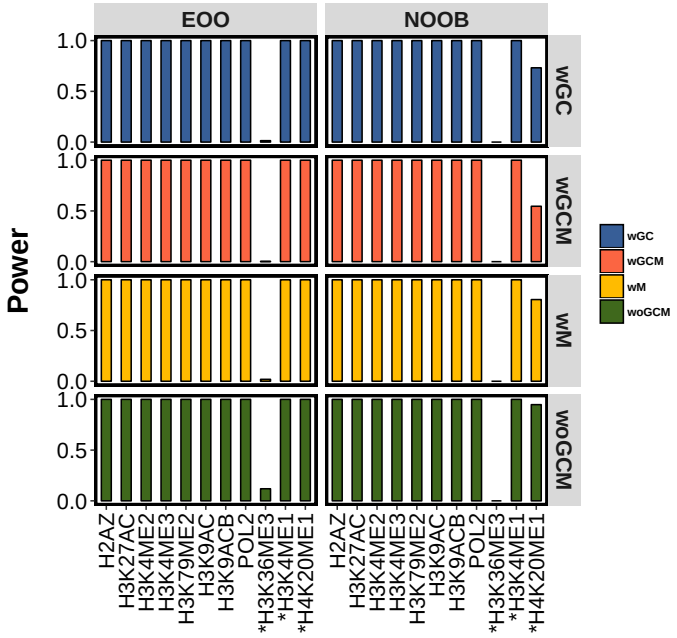
(b) K562, Expressed genes, $\alpha = 0.001$, wIF



(c) K562, Non-expressed genes, $\alpha = 0.001$, woIF



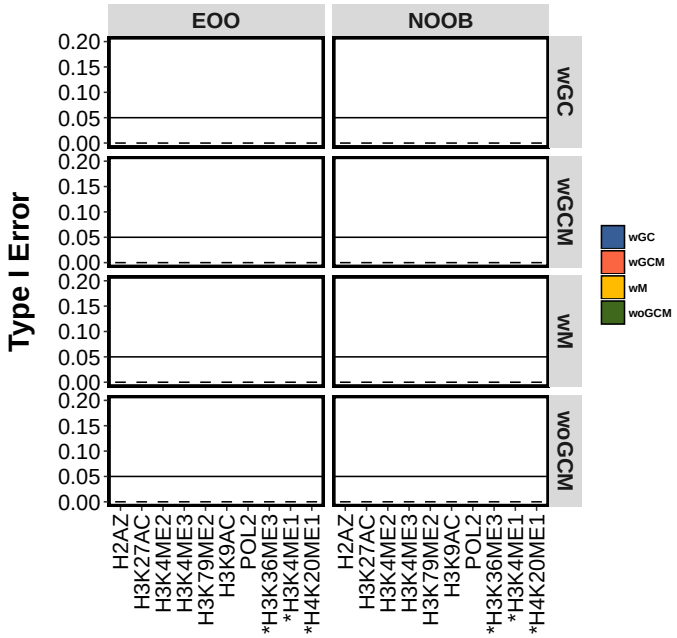
(d) K562, Expressed genes, $\alpha = 0.001$, woIF



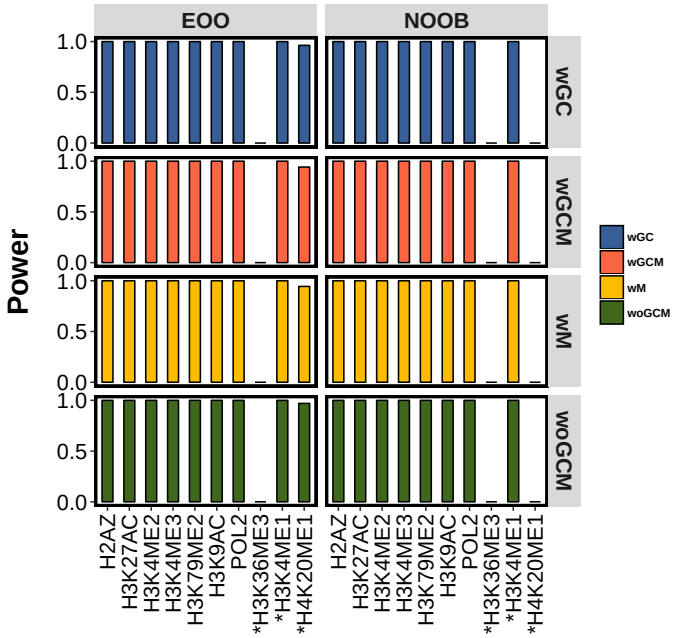
Supplementary Figure 6. Assessment of GLANET Type-I error and power with data-driven computational experiments. Results for the two association statistics - existence of overlap (EOO) and the number of overlapping bases (NOOB) - together with - with GC (wGC), with Mappability (wM), with GC and Mappability (wGCM), and without GC and mappability (woGCM) null distribution generation modes are displayed. Histone marks with ambiguous activator roles are marked with $*$. (a, b) Type-I error and power estimated with Isochore Family (wIF) heuristic using K562, (Non-expressed Genes,TakeTheLongest) and (Expressed Genes,Top20) results, for significance level of 0.001. (c, d) Type-I error and power estimated without Isochore Family (woIF) heuristic using K562, (Non-expressed Genes,TakeTheLongest) and (Expressed Genes,Top20) results, for significance level of 0.001.

6.1.5 GM12878 NonExpressed (CompletelyDiscard), Expressed (Top5), $\alpha = 0.05$

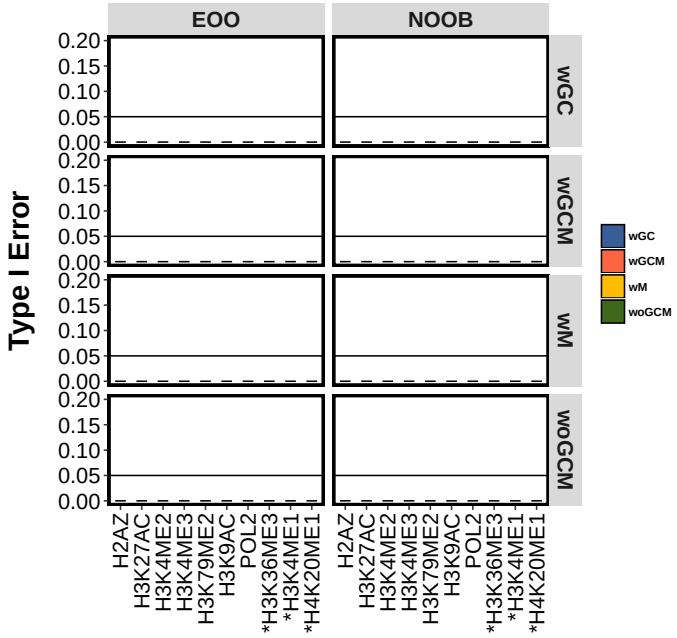
(a) GM12878, Non-expressed genes, $\alpha = 0.05$, wIF



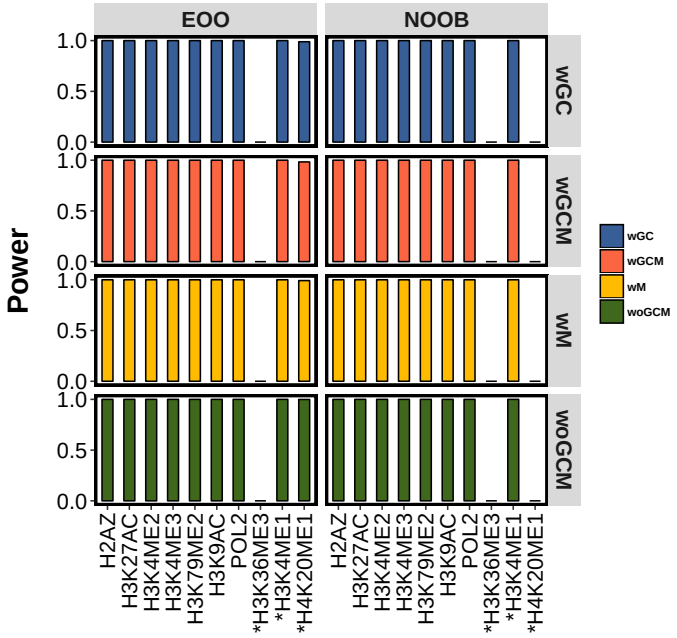
(b) GM12878, Expressed genes, $\alpha = 0.05$, wIF



(c) GM12878, Non-expressed genes, $\alpha = 0.05$, woIF



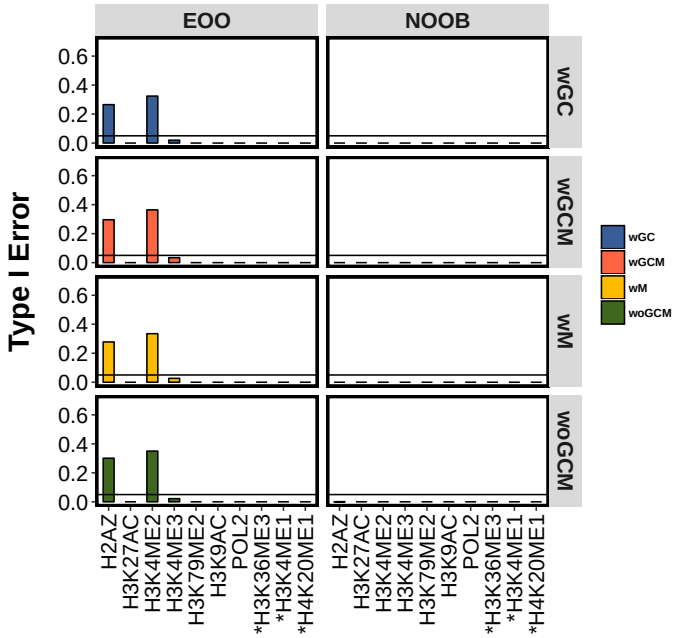
(d) GM12878, Expressed genes, $\alpha = 0.05$, woIF



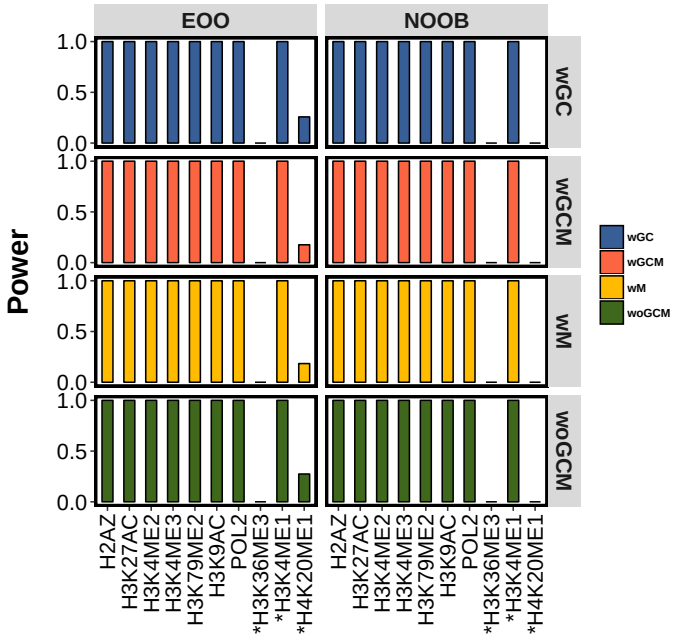
Supplementary Figure 7. Assessment of GLANET Type-I error rate and statistical power with data-driven computational experiments. Results for the two association statistics - existence of overlap (EOO) and the number of overlapping bases (NOOB) - together with GC (wGC), with Mappability (wM), with GC and Mappability (wGCM), and without GC and mappability (woGCM) null distribution generation modes are displayed. Histone marks with ambiguous activator roles are marked with $*$. (a, b) Type-I error and power estimated with Isochore Family (wIF) heuristic using GM12878, (Non-expressed Genes,CompletelyDiscard) and (Expressed Genes,Top5) results, for significance level of 0.05. (c, d) Type-I error and power estimated without Isochore Family (woIF) heuristic using GM12878, (Non-expressed Genes,CompletelyDiscard) and (Expressed Genes,Top5) results, for significance level of 0.05.

6.1.6 GM12878 NonExpressed (TakeTheLongest), Expressed (Top20), $\alpha = 0.05$

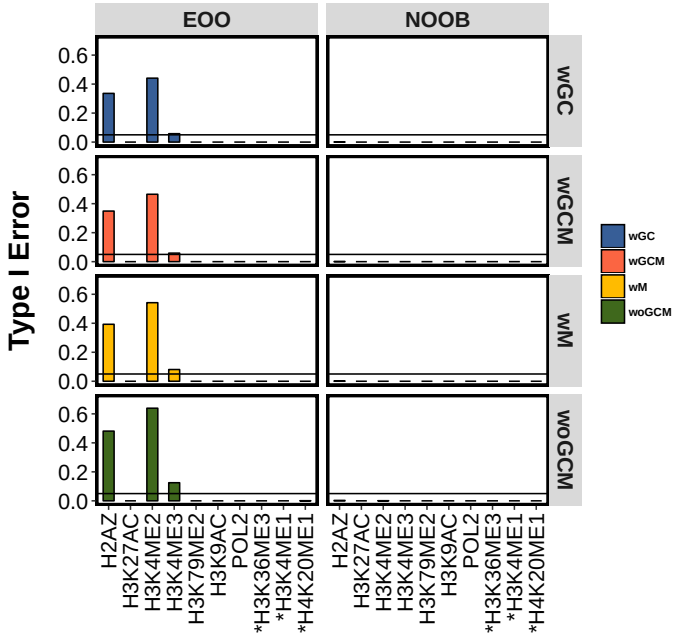
(a) GM12878, Non-expressed genes, $\alpha = 0.05$, wIF



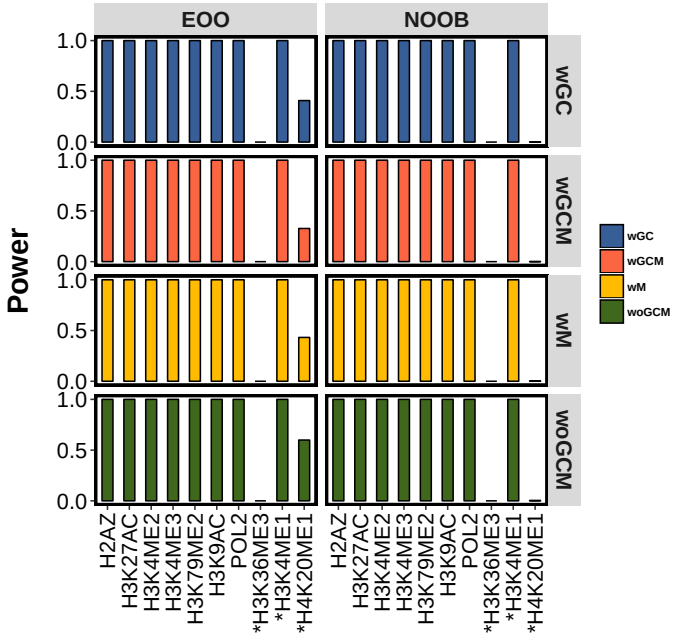
(b) GM12878, Expressed genes, $\alpha = 0.05$, wIF



(c) GM12878, Non-expressed genes, $\alpha = 0.05$, woIF



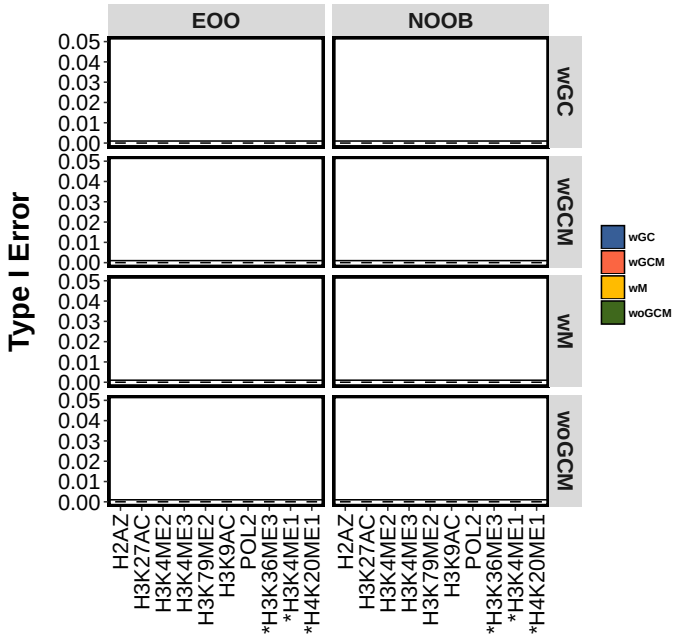
(d) GM12878, Expressed genes, $\alpha = 0.05$, woIF



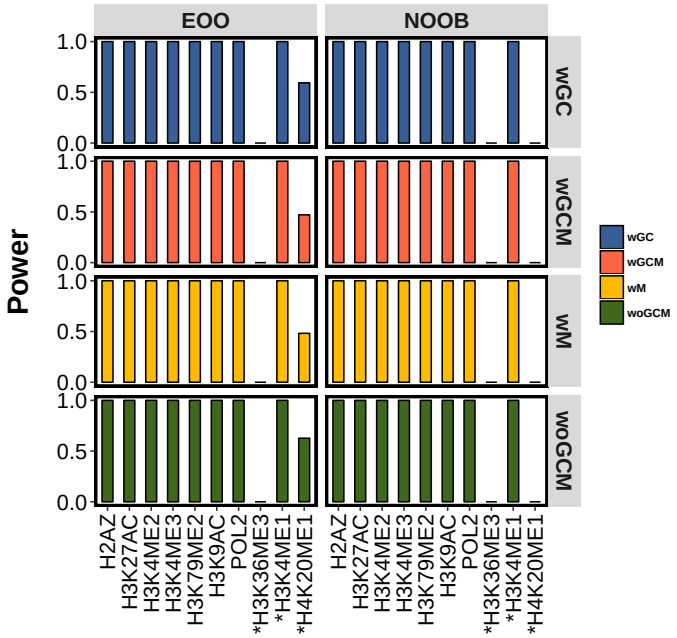
Supplementary Figure 8. Assessment of GLANET Type-I error and power with data-driven computational experiments. Results for the two association statistics - existence of overlap (EOO) and the number of overlapping bases (NOOB) - together with - with GC (wGC), with Mappability (wM), with GC and Mappability (wGCM), and without GC and mappability (woGCM) null distribution generation modes are displayed. Histone marks with ambiguous activator roles are marked with $*$. (a, b) Type-I error and power estimated with Isochore Family (wIF) heuristic using GM12878, (Non-expressed Genes,TakeTheLongest) and (Expressed Genes,Top20) results, for significance level of 0.05. (c, d) Type-I error and power estimated without Isochore Family (woIF) heuristic using GM12878, (Non-expressed Genes,TakeTheLongest) and (Expressed Genes,Top20) results, for significance level of 0.05.

6.1.7 GM12878 NonExpressed (CompletelyDiscard), Expressed (Top5), $\alpha = 0.001$

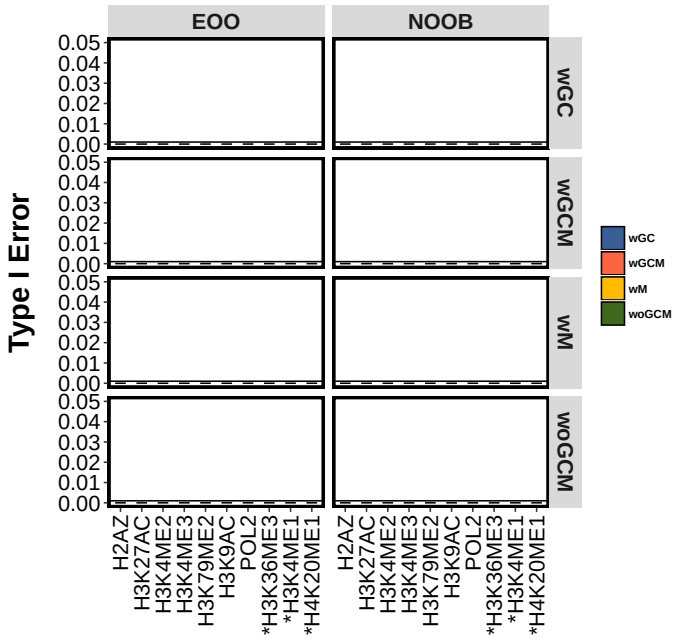
(a) GM12878, Non-expressed genes, $\alpha = 0.001$, wIF



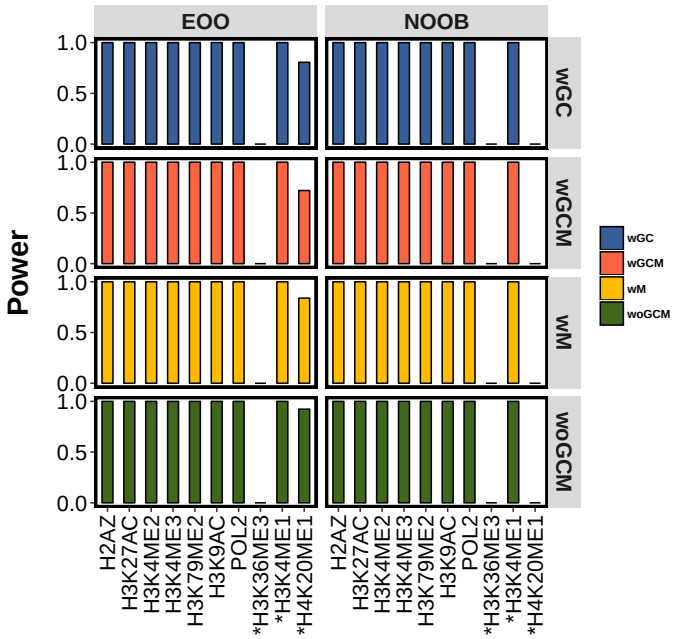
(b) GM12878, Expressed genes, $\alpha = 0.001$, wIF



(c) GM12878, Non-expressed genes, $\alpha = 0.001$, woIF



(d) GM12878, Expressed genes, $\alpha = 0.001$, woIF

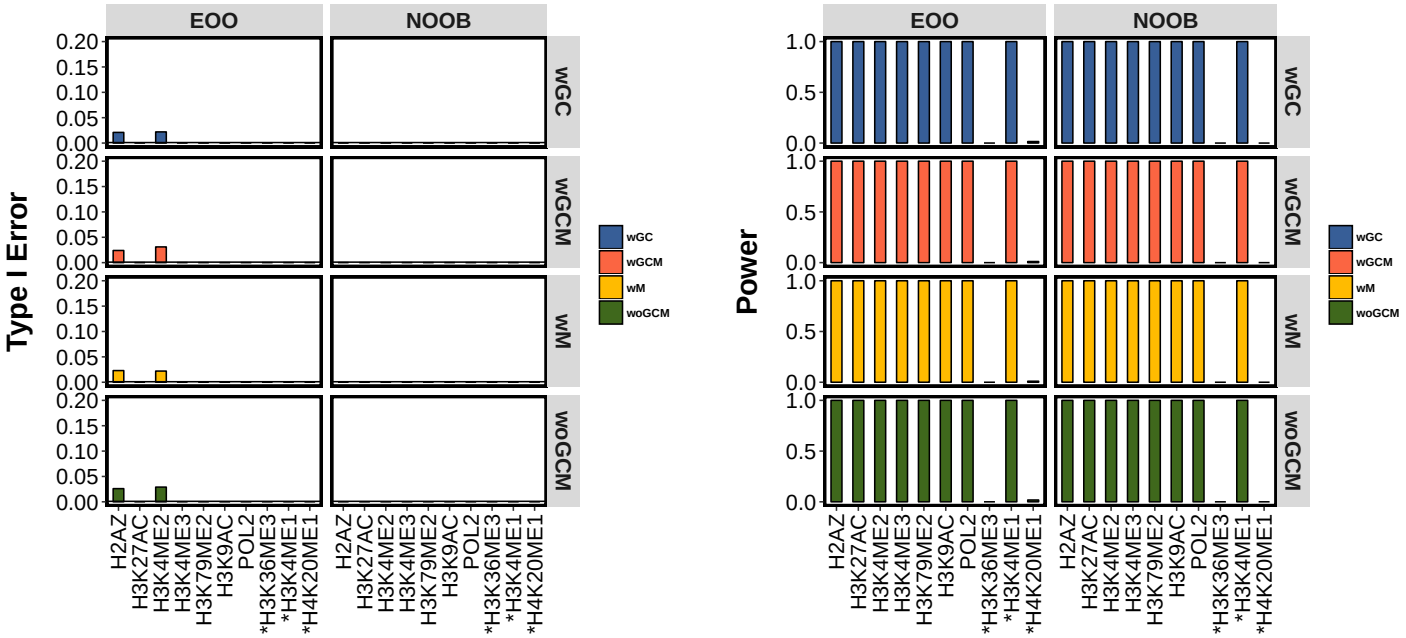


Supplementary Figure 9. Assessment of GLANET Type-I error and power with data-driven computational experiments. Results for the two association statistics - existence of overlap (EOO) and the number of overlapping bases (NOOB) - together with - with GC (wGC), with Mappability (wM), with GC and Mappability (wGCM), and without GC and mappability (woGCM) null distribution generation modes are displayed. Histone marks with ambiguous activator roles are marked with $*$. (a, b) Type-I error and power estimated with Isochore Family (wIF) heuristic using GM12878, (Non-expressed Genes,CompletelyDiscard) and (Expressed Genes,Top5) results, for significance level of 0.001. (c, d) Type-I error and power estimated without Isochore Family (woIF) heuristic using GM12878, (Non-expressed Genes,CompletelyDiscard) and (Expressed Genes,Top5) results, for significance level of 0.001.

6.1.8 GM12878 NonExpressed (TakeTheLongest), Expressed (Top20), $\alpha = 0.001$

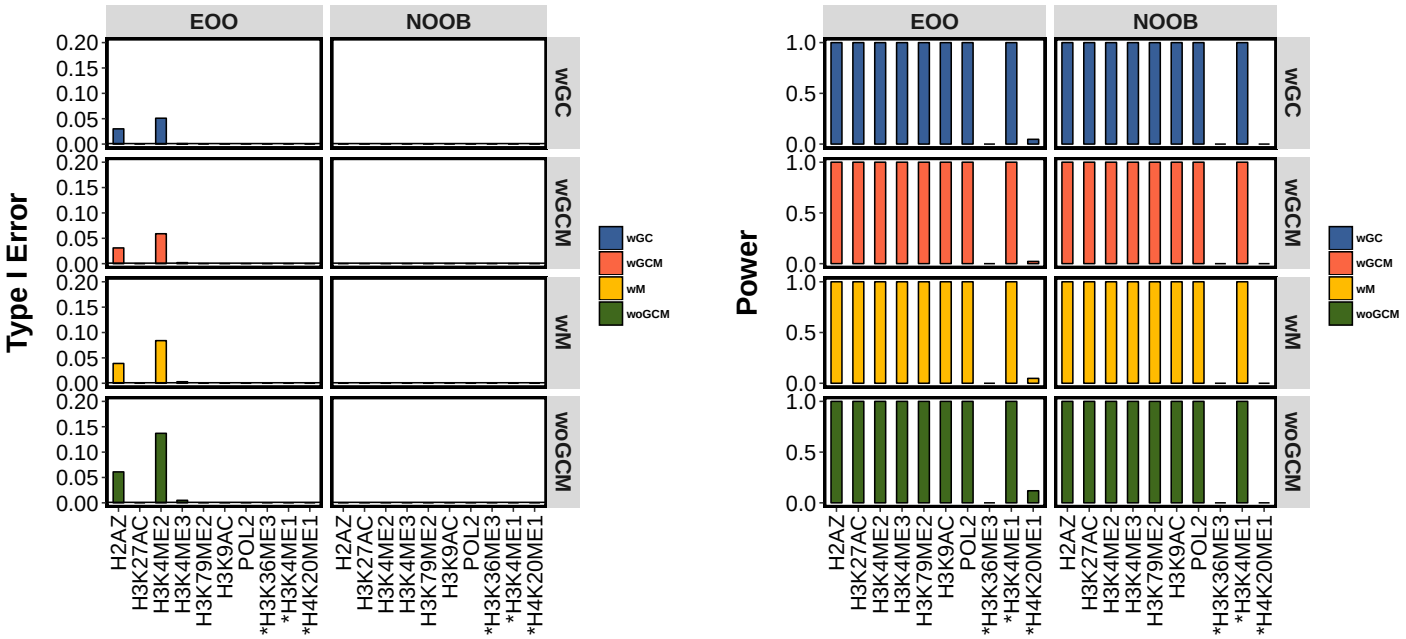
(a) GM12878, Non-expressed genes, $\alpha = 0.001$, wIF

(b) GM12878, Expressed genes, $\alpha = 0.001$, wIF



(c) GM12878, Non-expressed genes, $\alpha = 0.001$, woIF

(d) GM12878, Expressed genes, $\alpha = 0.001$, woIF



Supplementary Figure 10. Assessment of GLANET Type-I error and power with data-driven computational experiments. Results for the two association statistics - existence of overlap (EOO) and the number of overlapping bases (NOOB) - together with - with GC (wGC), with Mappability (wM), with GC and Mappability (wGCM), and without GC and mappability (woGCM) null distribution generation modes are displayed. Histone marks with ambiguous activator roles are marked with $*$. (a, b) Type-I error and power estimated with Isochore Family (wIF) heuristic using GM12878, (Non-expressed Genes,TakeTheLongest) and (Expressed Genes,Top20) results, for significance level of 0.001. (c, d) Type-I error and power estimated without Isochore Family (woIF) heuristic using GM12878, (Non-expressed Genes,TakeTheLongest) and (Expressed Genes,Top20) results, for significance level of 0.001.

6.2 Data-driven Computational Experiments Results for Repressor Elements

			Type-I Error, $\alpha = 0.05$							
			wIF				**woIF**			
	Expressed Genes		wGC	wM	wGCM	woGCM	wGC	wM	wGCM	woGCM
H3K27me3 GM12878	Top5	EOO	0	0	0	0	0	0	0	0
	Top20	EOO	0.001	0	0	0.001	0.002	0.001	0	0.006
	Top5	NOOB	0	0	0	0	0	0	0	0
	Top20	NOOB	0	0	0	0.001	0.001	0.001	0	0.008
H3K27me3 K562	Top5	EOO	0	0	0	0	0	0	0	0
	Top20	EOO	0	0	0	0	0	0	0	0
	Top5	NOOB	0	0	0	0	0	0	0	0
	Top20	NOOB	0	0	0	0	0	0	0	0
H3K9me3 GM12878	Top5	EOO	0	0	0	0	0	0	0	0
	Top20	EOO	0	0	0	0	0	0	0	0
	Top5	NOOB	0	0	0	0	0	0	0	0
	Top20	NOOB	0	0	0	0	0	0	0	0
H3K9me3 K562	Top5	EOO	0	0	0	0	0	0	0	0
	Top20	EOO	0.079	0.051	0.052	0.083	0.103	0.081	0.066	0.126
	Top5	NOOB	0	0	0	0	0	0	0	0
	Top20	NOOB	0.042	0.023	0.025	0.041	0.06	0.039	0.035	0.085

Supplementary Table 3. Type-I error rates calculated in data-driven experiments conducted with repressor elements, H3K27me3 and H3K9me3, in GM12878 and K562 cell lines for $\alpha = 0.05$.

			Type-I Error, $\alpha = 0.001$							
			wIF				**woIF**			
	Expressed Genes		wGC	wM	wGCM	woGCM	wGC	wM	wGCM	woGCM
H3K27me3 GM12878	Top5	EOO	0	0	0	0	0	0	0	0
	Top20	EOO	0	0	0	0	0	0	0	0
	Top5	NOOB	0	0	0	0	0	0	0	0
	Top20	NOOB	0	0	0	0	0	0	0	0
H3K27me3 K562	Top5	EOO	0	0	0	0	0	0	0	0
	Top20	EOO	0	0	0	0	0	0	0	0
	Top5	NOOB	0	0	0	0	0	0	0	0
	Top20	NOOB	0	0	0	0	0	0	0	0
H3K9me3 GM12878	Top5	EOO	0	0	0	0	0	0	0	0
	Top20	EOO	0	0	0	0	0	0	0	0
	Top5	NOOB	0	0	0	0	0	0	0	0
	Top20	NOOB	0	0	0	0	0	0	0	0
H3K9me3 K562	Top5	EOO	0	0	0	0	0	0	0	0
	Top20	EOO	0.002	0.001	0.001	0.002	0.003	0.001	0.001	0.005
	Top5	NOOB	0	0	0	0	0	0	0	0
	Top20	NOOB	0.001	0	0	0.001	0.001	0	0	0.001

Supplementary Table 4. Type-I error rates calculated in data-driven experiments conducted with repressor elements, H3K27me3 and H3K9me3, in GM12878 and K562 cell lines for $\alpha = 0.001$.

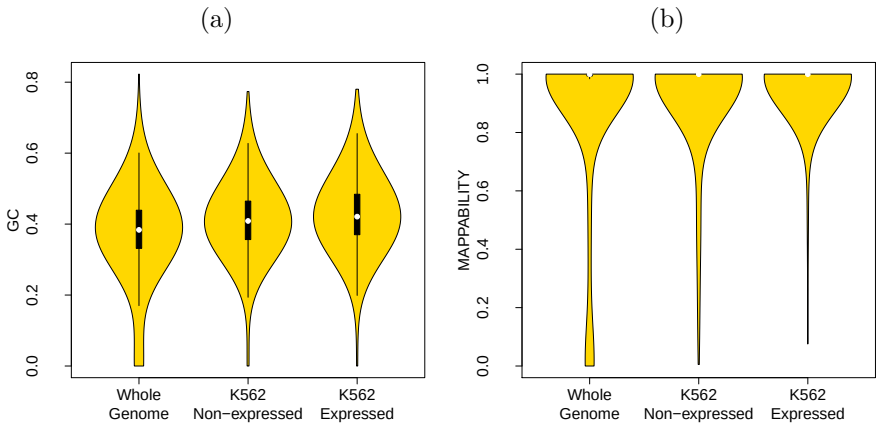
			Power, $\alpha = 0.05$							
			wIF				woIF			
	Non-expressed Genes		wGC	wM	wGCM	woGCM	wGC	wM	wGCM	woGCM
H3K27me3 GM12878	CompletelyDiscard	EOO	1	1	1	1	1	1	1	1
	TakeTheLongest	EOO	1	1	1	1	1	1	1	1
	CompletelyDiscard	NOOB	1	1	1	1	1	1	1	1
	TakeTheLongest	NOOB	1	1	1	1	1	1	1	1
H3K27me3 K562	CompletelyDiscard	EOO	1	1	1	1	1	1	1	1
	TakeTheLongest	EOO	1	1	1	1	1	1	1	1
	CompletelyDiscard	NOOB	1	1	1	1	1	1	1	1
	TakeTheLongest	NOOB	1	1	1	1	1	1	1	1
H3K9me3 GM12878	CompletelyDiscard	EOO	0.134	0.151	0.163	0.154	0.161	0.182	0.177	0.214
	TakeTheLongest	EOO	0.186	0.199	0.209	0.211	0.221	0.244	0.234	0.299
	CompletelyDiscard	NOOB	0.076	0.098	0.103	0.095	0.094	0.113	0.106	0.133
	TakeTheLongest	NOOB	0.096	0.113	0.124	0.119	0.12	0.134	0.129	0.168
H3K9me3 K562	CompletelyDiscard	EOO	0.003	0.004	0.003	0.004	0.004	0.005	0.005	0.007
	TakeTheLongest	EOO	0.002	0.003	0.002	0.003	0.002	0.005	0.004	0.006
	CompletelyDiscard	NOOB	0.003	0.004	0.004	0.004	0.004	0.005	0.005	0.006
	TakeTheLongest	NOOB	0.005	0.004	0.005	0.005	0.005	0.005	0.005	0.005

Supplementary Table 5. Power calculated in data-driven experiments conducted with repressor elements, H3K27me3 and H3K9me3, in GM12878 and K562 cell lines for $\alpha = 0.05$.

			Power, $\alpha = 0.001$							
			wIF				woIF			
	Non-expressed Genes		wGC	wM	wGCM	woGCM	wGC	wM	wGCM	woGCM
H3K27me3 GM12878	CompletelyDiscard	EOO	1	1	1	1	1	1	1	1
	TakeTheLongest	EOO	1	1	1	1	1	1	1	1
	CompletelyDiscard	NOOB	1	1	1	1	1	1	1	1
	TakeTheLongest	NOOB	1	1	1	1	1	1	1	1
H3K27me3 K562	CompletelyDiscard	EOO	1	1	1	1	1	1	1	1
	TakeTheLongest	EOO	1	1	1	1	1	1	1	1
	CompletelyDiscard	NOOB	1	1	1	1	1	1	1	1
	TakeTheLongest	NOOB	1	1	1	1	1	1	1	1
H3K9me3 GM12878	CompletelyDiscard	EOO	0.003	0.005	0.004	0.005	0.006	0.006	0.008	0.017
	TakeTheLongest	EOO	0.008	0.009	0.009	0.011	0.012	0.013	0.013	0.023
	CompletelyDiscard	NOOB	0	0.002	0.001	0.001	0	0.003	0.001	0.004
	TakeTheLongest	NOOB	0.005	0.005	0.004	0.005	0.005	0.005	0.004	0.007
H3K9me3 K562	CompletelyDiscard	EOO	0	0	0	0	0	0	0	0
	TakeTheLongest	EOO	0	0	0	0	0	0	0	0
	CompletelyDiscard	NOOB	0	0	0	0	0	0	0	0
	TakeTheLongest	NOOB	0	0	0	0	0	0	0	0

Supplementary Table 6. Power calculated in data-driven experiments conducted with repressor elements, H3K27me3 and H3K9me3, in GM12878 and K562 cell lines for $\alpha = 0.001$.

7 Assessing best GLANET parameter settings through Data-driven Computational Experiments Results



Supplementary Figure 11. Violin plots for (a) GC of randomly sampled intervals from human genome, GC of intervals of K562 non-expressed genes and expressed genes. (b) Mappability of randomly sampled intervals from human genome, mappability of intervals from non-expressed and expressed gene-sets of K562.

Kolmogorov-Smirnov Test Results		
Property	**Interval Set**	**Maximum Distance**
GC	Non-expressed (GM12878)	0.1454
GC	Expressed (GM12878)	0.1462
GC	Non-expressed (K562)	0.1241
GC	Expressed (K562)	0.1897
Mappability	Non-expressed (GM12878)	**0.0794**
Mappability	Expressed (GM12878)	0.1693
Mappability	Non-expressed (K562)	**0.0898**
Mappability	Expressed (K562)	0.1585

Supplementary Table 7. Kolmogorov-Smirnov test results. Null hypothesis states that the distribution of GC content or mappability values calculated for $50,000$ randomly sampled intervals from human genome and the corresponding interval set are not different. Each row corresponds to Kolmogorov-Smirnov testing of this null hypothesis. In all tests, the null hypothesis is rejected ($p\text{-}value < 2.2e\text{-}16$). The first column lists the property of the genome in question, the second column lists the distribution that is compared with the genome, finally the last column lists the maximum distance between the two distributions.

8 Wilcoxon Signed Rank Test Results

To assess the different parameter settings, we carried out one-sided paired Wilcoxon signed rank tests and we compared the distribution of Type-I errors generated under these different parameters. The null states there is no difference in the mean of the ranks of the two distributions whereas alternative hypothesis is that the first distribution has lower mean of ranks than the second one.

The distribution generated under the GLANET parameter setting specified in the row is tested against the Type I error distribution generated under the parameter setting specified in the column. A p-value less than or equal to the value in the cell indicates that setting in the corresponding row has a lower mean of ranks in Type-I error distribution than the setting in the corresponding column.

We pooled the Type-I errors of parameter setting results to compare **wIF** and **woIF**. Results indicate **wIF** achieves lower Type-I errors than **woIF** (Supplementary Table 8). Similarly, we pooled simulation results to compare **NOOB** with **EOO**, we observed that **NOOB** provides lower Type-I errors than **EOO** (Supplementary Table 9).

	Wilcoxon signed rank test p-values	
	woIF	wIF
woIF		
wIF	**2.2e-16**	

Supplementary Table 8. Wilcoxon Signed Rank test results that compares Type I error distribution of all simulations generated under **(woIF)** setting against the Type-I errors generated under **(wIF)** setting. A p-value less than or equal to the value in the cell indicates that setting in the corresponding row has a lower mean of ranks in Type-I error distribution than the setting in the corresponding column.

	Wilcoxon signed rank test p-values	
	EOO	NOOB
EOO		
NOOB	**2.2e-16**	

Supplementary Table 9. Wilcoxon Signed Rank test results that compares Type I error distribution of all simulations generated under **(EOO)** setting against the Type-I errors generated under **(NOOB)** setting. A p-value less than or equal to the value in the cell indicates that setting in the corresponding row has a lower mean of ranks in Type-I error distribution than the setting in the corresponding column.

9 ROC Curves for Assessing GLANET's Different Parameter Settings and GAT's Performance

To compare quantitatively how GLANET parameters affect the enrichment performance, we also analyzed ROC curves. To compare GLANET's performance with GAT, we also include GAT results in the ROC curves. To plot a single ROC curve per an element in cell line, simulation results that are conducted under the same parameter setting for expressed and non-expressed genes are combined. While drawing element-based ROC curves, we label each activator element as "enriched" in expressed gene scenario and "not enriched" in non-expressed genes scenarios. Similarly, the true label for each repressor element as "not-enriched" and "enriched" under expressed and non-expressed genes simulations, respectively.

We compared the difference in AUC of two ROC curves with each other using pROC R package [7]. We utilized "delong" method and count the number of wins, ties and losses. A win is registered whenever the first ROC curve is found to be higher than the second tested ROC curve at 0.05 significance level. A lose registers the reverse scenario; the first curve is found to be below the second one, and a tie indicates that there is no statistically significant difference between the two compared curves. We accumulate the number of wins, ties and losses across different histone modification elements and POL2 and cell line to summarize the results. The results are summarized in Supplementary Tables 10-17.

(EOO,woIF)	GAT (woGCM)	GLANET (woGCM)	GLANET (wGC)	GLANET (wM)	GLANET (wGCM)	Number of Wins	Number of Ties	Number of Losses
GAT(woGCM)		1/44/5	3/37/10	3/38/9	3/37/10	10	156	34
GLANET(woGCM)	5/44/1		3/38/9	3/38/9	3/38/9	14	158	28
GLANET(wGC)	10/37/3	9/38/3		5/41/4	3/43/4	27	159	14
GLANET(wM)	9/38/3	9/38/3	4/41/5		3/42/5	25	159	16
GLANET(wGCM)	10/37/3	9/38/3	4/43/3	5/42/3		28	160	12

Supplementary Table 10. ROC curves of different parameter settings where `(EOO,woIF)` setting is on are compared. A Win indicates a case where the ROC curve obtained with settings specified in the row is statistically significantly above the ROC curve obtained with the settings specified in the column at significance level 0.05. A lose indicates the opposite, while a tie indicates that there is no statistically significant difference between the two compared curves. The counts indicate the number of times win/tie/lose cases occur when the results for different elements, cell lines and other experimental conditions are compared.

(EOO,wIF)	GAT (woGCM)	GLANET (woGCM)	GLANET (wGC)	GLANET (wM)	GLANET (wGCM)	Number of Wins	Number of Ties	Number of Losses
GAT(woGCM)		5/38/7	3/40/7	3/39/8	5/39/6	16	156	28
GLANET(woGCM)	7/38/5		2/45/3	3/42/5	3/42/5	15	167	18
GLANET(wGC)	7/40/3	3/45/2		3/43/4	3/43/4	16	171	13
GLANET(wM)	8/39/3	5/42/3	4/43/3		3/47/0	20	171	9
GLANET(wGCM)	6/39/5	5/42/3	4/43/3	0/47/3		15	171	14

Supplementary Table 11. ROC curves of different parameter settings where `(EOO,wIF)` setting is on are compared. A Win indicates a case where the ROC curve obtained with settings specified in the row is statistically significantly above the ROC curve obtained with the settings specified in the column at significance level 0.05. A lose indicates the opposite, while a tie indicates that there is no statistically significant difference between the two compared curves. The counts indicate the number of times win/tie/lose cases occur when the results for different elements, cell lines and other experimental conditions are compared.

(NOOB,woIF)	GAT (woGCM)	GLANET (woGCM)	GLANET (wGC)	GLANET (wM)	GLANET (wGCM)	Number of Wins	Number of Ties	Number of Losses
GAT(woGCM)		2/44/4	6/39/5	7/38/5	6/39/5	21	160	19
GLANET(woGCM)	4/44/2		6/39/5	6/39/5	6/39/5	22	161	17
GLANET(wGC)	5/39/6	5/39/6		4/42/4	6/40/4	20	160	20
GLANET(wM)	5/38/7	5/39/6	4/42/4		5/40/5	19	159	22
GLANET(wGCM)	5/39/6	5/39/6	4/40/6	5/40/5		19	158	23

Supplementary Table 12. ROC curves of different parameter settings where `(NOOB,woIF)` setting is on are compared. A Win indicates a case where the ROC curve obtained with settings specified in the row is statistically significantly above the ROC curve obtained with the settings specified in the column at significance level 0.05. A lose indicates the opposite, while a tie indicates that there is no statistically significant difference between the two compared curves. The counts indicate the number of times win/tie/lose cases occur when the results for different elements, cell lines and other experimental conditions are compared.

(NOOB,wIF)	GAT (woGCM)	GLANET (woGCM)	GLANET (wGC)	GLANET (wM)	GLANET (wGCM)	Number of Wins	Number of Ties	Number of Losses
GAT(woGCM)		1/41/8	1/41/8	0/40/10	1/39/10	3	161	36
GLANET(woGCM)	8/41/1		6/42/2	5/41/4	5/41/4	24	165	11
GLANET(wGC)	8/41/1	2/42/6		5/41/4	7/39/4	22	163	15
GLANET(wM)	10/40/0	4/41/5	4/41/5		6/44/0	24	166	10
GLANET(wGCM)	10/39/1	4/41/5	4/39/7	0/44/6		18	163	19

Supplementary Table 13. ROC curves of different parameter settings where `(NOOB,wIF)` setting is on are compared. A Win indicates a case where the ROC curve obtained with settings specified in the row is statistically significantly above the ROC curve obtained with the settings specified in the column at significance level 0.05. A lose indicates the opposite, while a tie indicates that there is no statistically significant difference between the two compared curves. The counts indicate the number of times win/tie/lose cases occur when the results for different elements, cell lines and other experimental conditions are compared.

	GLANET (EOO,woIF,wGCM)	GLANET (EOO,wIF,wM)	GLANET (NOOB,woIF,woGCM)	GLANET (NOOB,wIF,wM)	Number of Wins	Number of Ties	Number of Losses
GLANET(EOO,woIF,wGCM)		4/40/6	7/39/4	7/36/7	18	115	17
GLANET(EOO,wIF,wM)	6/40/4		9/38/3	7/38/5	22	116	12
GLANET(NOOB,woIF,woGCM)	4/39/7	3/38/9		6/38/6	13	115	22
GLANET(NOOB,wIF,wM)	7/36/7	5/38/7	6/38/6		18	112	20

Supplementary Table 14. Winner settings from Supplementary Tables 10-13 are compared with each other. A Win indicates a case where the ROC curve obtained with settings specified in the row is statistically significantly above the ROC curve obtained with the settings specified in the column at significance level 0.05. A lose indicates the opposite, while a tie indicates that there is no statistically significant difference between the two compared curves. The counts indicate the number of times win/tie/lose cases occur when the results for different elements, cell lines and other experimental conditions are compared.

(woIF)	GAT (woGCM)	GLANET (woGCM)	GLANET (wGC)	GLANET (wM)	GLANET (wGCM)	Number of Wins	Number of Ties	Number of Losses
GAT(woGCM)		3/88/9	9/76/15	10/76/14	9/76/15	31	316	53
GLANET(woGCM)	9/88/3		9/77/14	9/77/14	9/77/14	36	319	45
GLANET(wGC)	15/76/9	14/77/9		9/83/8	9/83/8	47	319	34
GLANET(wM)	14/76/10	14/77/9	8/83/9		8/82/10	44	318	38
GLANET(wGCM)	15/76/9	14/77/9	8/83/9	10/82/8		47	318	35

Supplementary Table 15. ROC curves of different parameter settings where `(woIF)` setting is on are compared. A Win indicates a case where the ROC curve obtained with settings specified in the row is statistically significantly above the ROC curve obtained with the settings specified in the column at significance level 0.05. A lose indicates the opposite, while a tie indicates that there is no statistically significant difference between the two compared curves. The counts indicate the number of times win/tie/lose cases occur when the results for different elements, cell lines and other experimental conditions are compared.

(wIF)	GAT (woGCM)	GLANET (woGCM)	GLANET (wGC)	GLANET (wM)	GLANET (wGCM)	Number of Wins	Number of Ties	Number of Losses
GAT(woGCM)		6/79/15	4/81/15	3/79/18	6/78/16	19	317	64
GLANET(woGCM)	15/79/6		8/87/5	8/83/9	8/83/9	39	332	29
GLANET(wGC)	15/81/4	5/87/8		8/84/8	10/82/8	38	334	28
GLANET(wM)	18/79/3	9/83/8	8/84/8		9/91/0	44	337	19
GLANET(wGCM)	16/78/6	9/83/8	8/82/10	0/91/9		33	334	33

Supplementary Table 16. ROC curves of different parameter settings where `(wIF)` setting is on are compared. A Win indicates a case where the ROC curve obtained with settings specified in the row is statistically significantly above the ROC curve obtained with the settings specified in the column at significance level 0.05. A lose indicates the opposite, while a tie indicates that there is no statistically significant difference between the two compared curves. The counts indicate the number of times win/tie/lose cases occur when the results for different elements, cell lines and other experimental conditions are compared.

All Pooles	GAT (woGCM)	GLANET (woGCM)	GLANET (wGC)	GLANET (wM)	GLANET (wGCM)	Number of Wins	Number of Ties	Number of Losses
GAT(woGCM)	0/0/0	9/167/24	13/157/30	13/155/32	15/154/31	50	633	117
GLANET(woGCM)	24/167/9	0/0/0	17/164/19	17/160/23	17/160/23	75	651	74
GLANET(wGC)	30/157/13	19/164/17	0/0/0	17/167/16	19/165/16	85	653	62
GLANET(wM)	32/155/13	23/160/17	16/167/17	0/0/0	17/173/10	88	655	57
GLANET(wGCM)	31/154/15	23/160/17	16/165/19	10/173/17	0/0/0	80	652	68

Supplementary Table 17. ROC curves of different "Generate Random Data Options" are compared. A Win indicates a case where the ROC curve obtained with settings specified in the row is statistically significantly above the ROC curve obtained with the settings specified in the column at significance level 0.05. A lose indicates the opposite, while a tie indicates that there is no statistically significant difference between the two compared curves. The counts indicate the number of times win/tie/lose cases occur when the results for different elements, cell lines and other experimental conditions are compared.

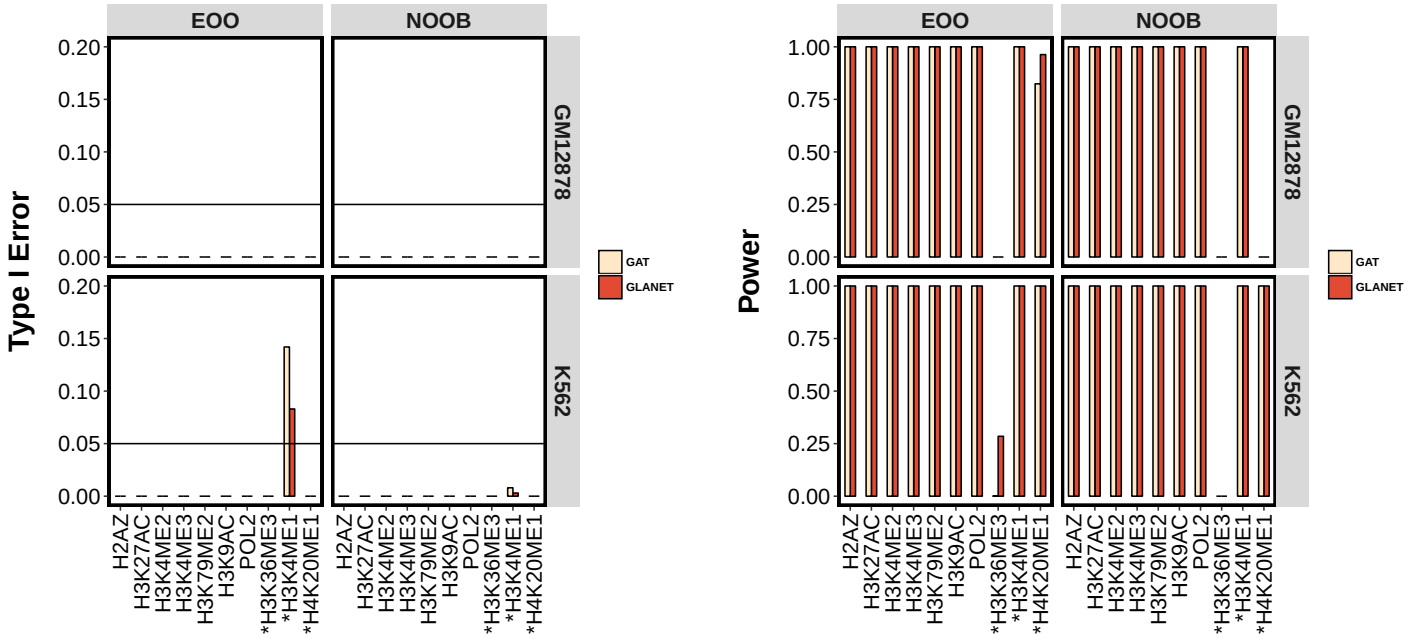
10 GLANET GAT Comparison

10.1 GLANET GAT Comparison Results for Activator and Repressor Elements through Data-driven Computational Experiments

10.1.1 NonExpressed (CompletelyDiscard), Expressed (Top5), $\alpha = 0.05$

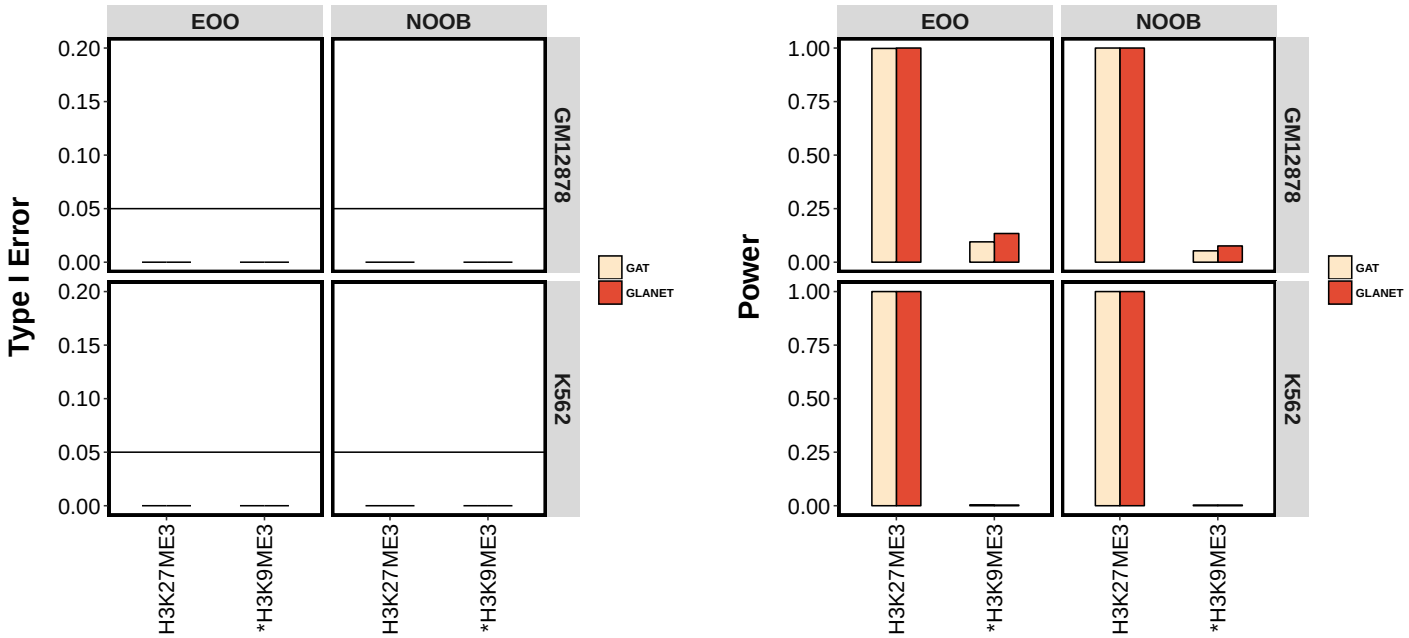
(a) Non-expressed Genes, CompletelyDiscard, $\alpha = 0.05$

(b) Expressed Genes, Top5, $\alpha = 0.05$



(c) Expressed Genes, Top5, $\alpha = 0.05$

(d) Non-expressed Genes, CompletelyDiscard, $\alpha = 0.05$

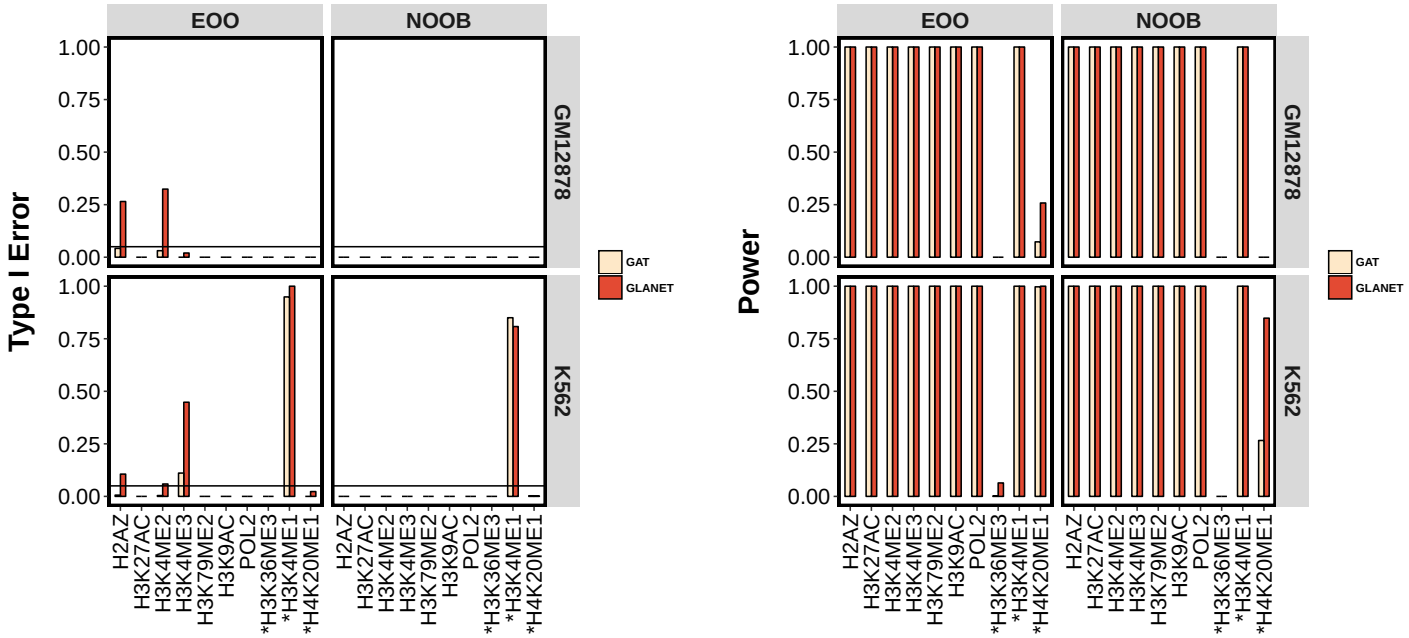


Supplementary Figure 12. Comparison of GLANET and GAT with respect to data-driven computational experiments in terms of Type-I Error and Power for significance level of 0.05. `GLANET(wIF,wGC)` and `GAT(wIF)` parameter settings results are used. Results for the two association statistics - existence of overlap (`EOO`) and the number of overlapping bases (`NOOB`) are displayed. (a, b) Type-I error and power of activator elements in (`Non-expressed Genes`,`CompletelyDiscard`) and (`Expressed Genes`,`Top5`) experiment settings, respectively. GLANET attains lower Type-I error for H3K4me1 and higher power for H3K36me3 and H4K20me1 elements than GAT. (c, d) Type-I error and power of repressor elements in (`Expressed Genes`,`Top5`) and (`Non-expressed Genes`,`CompletelyDiscard`) experiment settings, respectively. GLANET achieves higher power for H3K9me3 than GAT.

10.1.2 NonExpressed (TakeTheLongest), Expressed (Top20), $\alpha = 0.05$

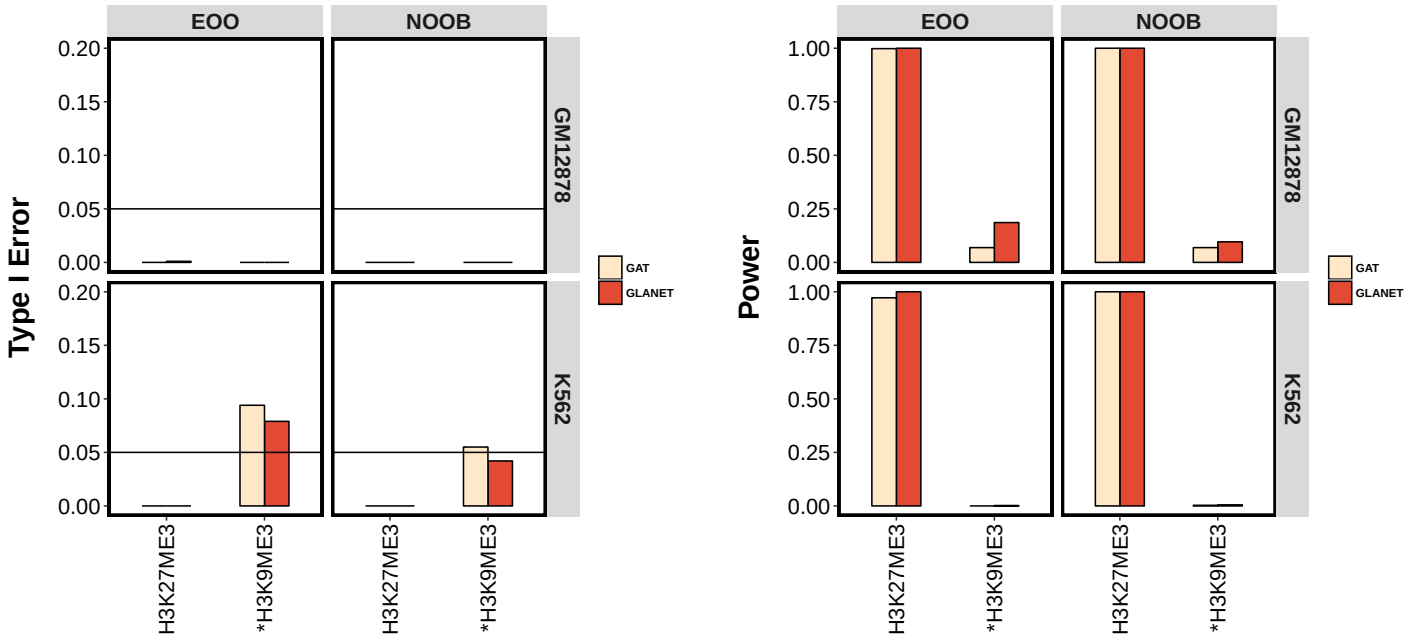
(a) Non-expressed Genes, TakeTheLongest, $\alpha = 0.05$

(b) Expressed Genes, Top20, $\alpha = 0.05$



(c) Expressed Genes, Top20, $\alpha = 0.05$

(d) Non-expressed Genes, TakeTheLongest, $\alpha = 0.05$

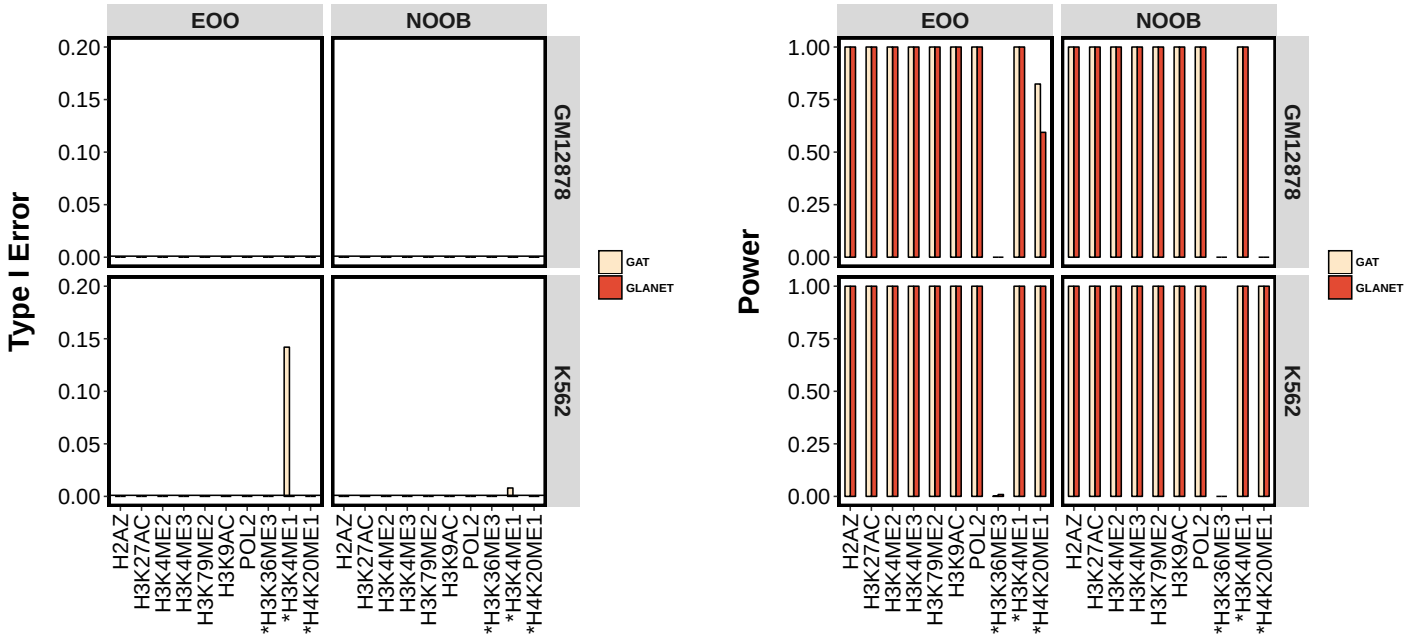


Supplementary Figure 13. Comparison of GLANET and GAT with respect to data-driven computational experiments in terms of Type-I Error and Power for significance level of 0.05. `GLANET(wIF,wGC)` and `GAT(wIF)` parameter settings results are used. Results for the two association statistics - existence of overlap (`EOO`) and the number of overlapping bases (`NOOB`) are displayed. (a, b) Type-I error and power of activator elements in (`Non-expressed Genes`,`TakeTheLongest`) and (`Expressed Genes`,`Top20`) experiment settings, respectively. (c, d) Type-I error and power of repressor elements in (`Expressed Genes`,`Top20`) and (`Non-expressed Genes`,`TakeTheLongest`) experiment settings, respectively.

10.1.3 NonExpressed (CompletelyDiscard), Expressed (Top5), $\alpha = 0.001$

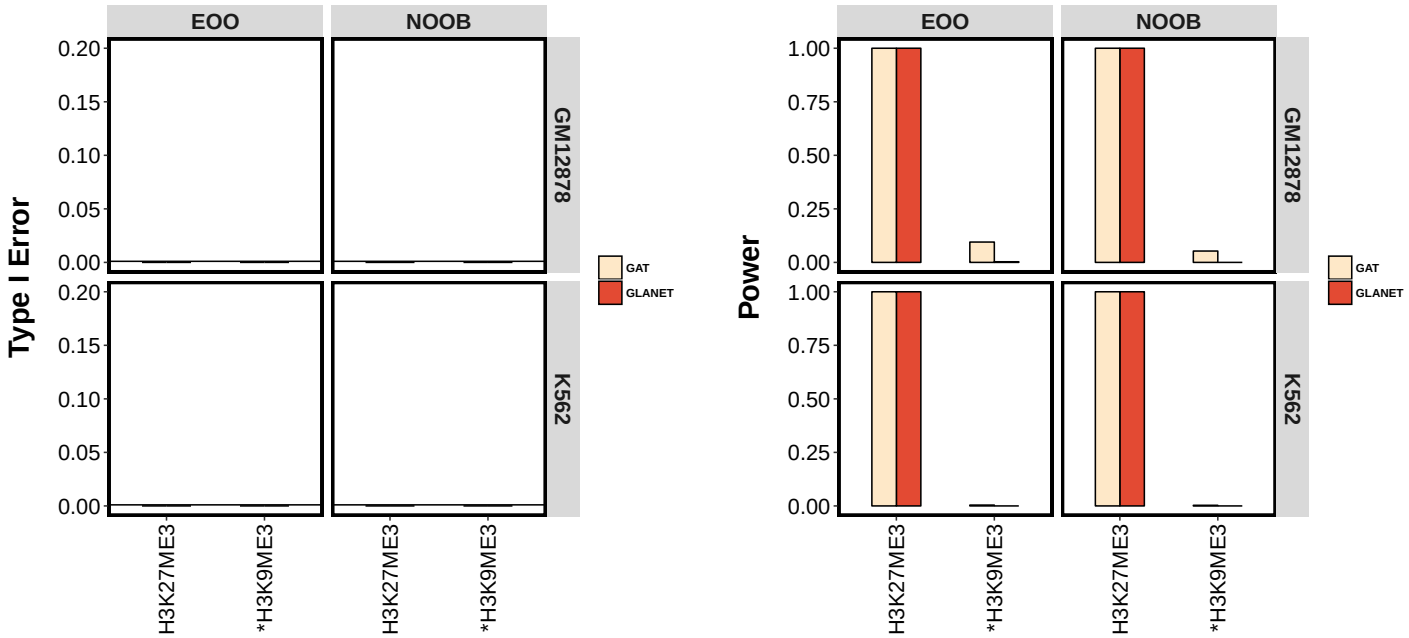
(a) Non-expressed Genes, CompletelyDiscard, $\alpha = 0.001$

(b) Expressed Genes, Top5, $\alpha = 0.001$



(c) Expressed Genes, Top5, $\alpha = 0.001$

(d) Non-expressed Genes, CompletelyDiscard, $\alpha = 0.001$

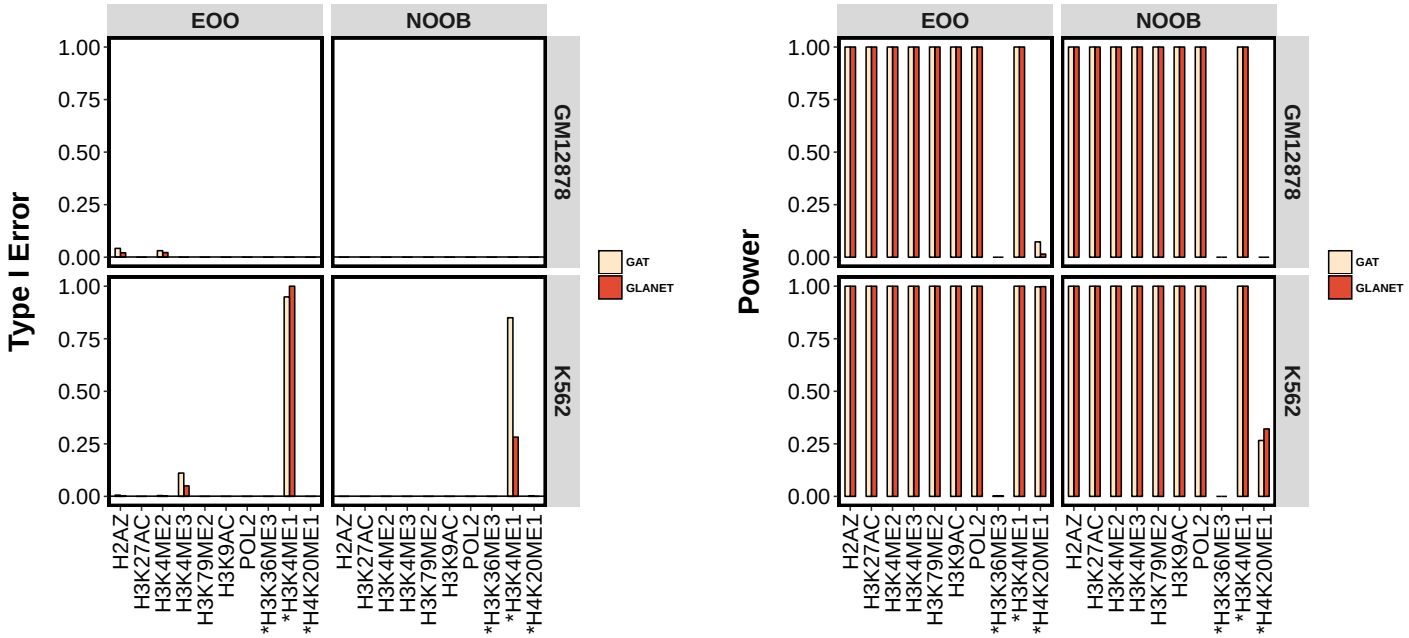


Supplementary Figure 14. Comparison of GLANET and GAT with respect to data-driven computational experiments in terms of Type-I Error and Power for significance level of 0.001. `GLANET(wIF,wGC)` and `GAT(wIF)` parameter settings results are used. Results for the two association statistics - existence of overlap (`EOO`) and the number of overlapping bases (`NOOB`) are displayed. (a, b) Type-I error and power of activator elements in (`Non-expressed Genes`,`CompletelyDiscard`) and (`Expressed Genes`,`Top5`) experiment settings, respectively. (c, d) Type-I error and power of repressor elements in (`Expressed Genes`,`Top5`) and (`Non-expressed Genes`,`CompletelyDiscard`) experiment settings, respectively.

10.1.4 NonExpressed (TakeTheLongest), Expressed (Top20), $\alpha = 0.001$

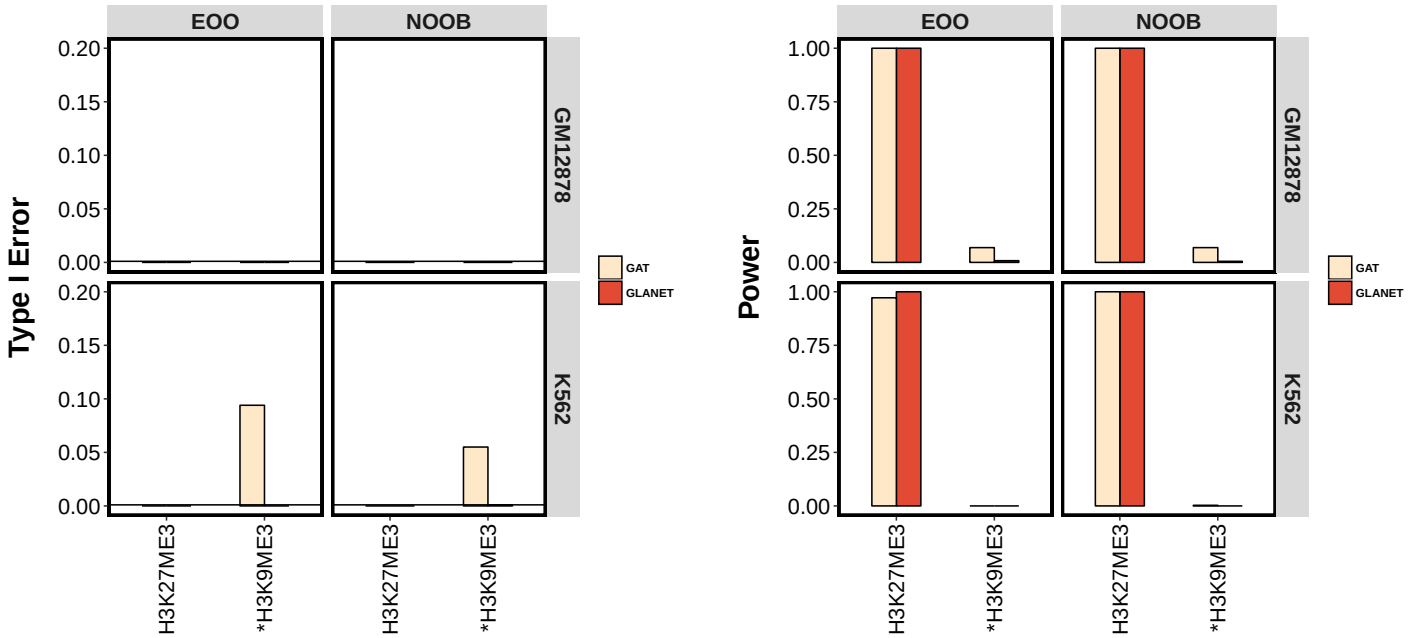
(a) Non-expressed Genes, TakeTheLongest, $\alpha = 0.001$

(b) Expressed Genes, Top20, $\alpha = 0.001$



(c) Expressed Genes, Top20, $\alpha = 0.001$

(d) Non-expressed Genes, TakeTheLongest, $\alpha = 0.001$



Supplementary Figure 15. Comparison of GLANET and GAT with respect to data-driven computational experiments in terms of Type-I Error and Power for significance level of 0.001. `GLANET(wIF,wGC)` and `GAT(wIF)` parameter settings results are used. Results for the two association statistics - existence of overlap (`EOO`) and the number of overlapping bases (`NOOB`) are displayed. (a, b) Type-I error and power of activator elements in (`Non-expressed Genes`,`TakeTheLongest`) and (`Expressed Genes`,`Top20`) experiment settings, respectively. (c, d) Type-I error and power of repressor elements in (`Expressed Genes`,`Top20`) and (`Non-expressed Genes`,`TakeTheLongest`) experiment settings, respectively.

10.2 Results for Additional GAT Experiments

We repeated the experiments provided in the GAT supplementary website [8] with GLANET. The detailed results for these additional experiments are provided in Supplementary Tables 18-21. Results for GAT runs are obtained from the GAT tutorial (`http://gat.readthedocs.org/en/latest/tutorialIntervalOverlap.html`). For each experiment, GLANET results are computed in sixteen different parameter settings. GLANET is run with different modes of random data generation (`wGC,wM,wGCM,woGCM`), isochore family (`woIF,wIF`) and association measure (`EOO,NOOB`). In each of the Supplementary Tables 18-21, *Observed* column shows the association measure value calculated between the given sets, set1 and set2. *Expected* and *StdDev* columns show the mean and standard deviation of association measure values of samplings, respectively. *Fold change* is one plus *Observed* divided by one plus *Expected*. Enrichment result is provided by the *p-value* column.

Experiment1 Set1: Srf(Jurkat) Set2: DNaseI(Jurkat)						
Tool	**Parameter Settings**	**Observed**	**Expected**	**StdDev**	**FoldChange**	**pValue**
GAT	(NOOB,woGCM,woIF)	20183	246.5650	105.5933	81.5301	1.0e-03
GLANET	(EOO,wGC,woIF)	450	15.7577	3.8662	26.9130	0
GLANET	(EOO,wM,woIF)	450	7.6723	2.7149	52.0046	0
GLANET	(EOO,wGCM,woIF)	450	17.3464	4.0456	24.5824	0
GLANET	(EOO,woGCM,woIF)	450	6.6257	2.5610	59.1421	0
GLANET	(EOO,wGC,wIF)	450	15.5799	3.8328	27.2016	0
GLANET	(EOO,wM,wIF)	450	11.9761	3.4328	34.7562	0
GLANET	(EOO,wGCM,wIF)	450	17.3041	4.0071	24.6392	0
GLANET	(EOO,woGCM,wIF)	450	10.9239	3.2333	37.8231	0
GLANET	(NOOB,wGC,woIF)	20183	599.3644	158.8155	33.6195	0
GLANET	(NOOB,wM,woIF)	20183	288.3931	112.5672	69.7459	0
GLANET	(NOOB,wGCM,woIF)	20183	668.5556	169.8404	30.1453	0
GLANET	(NOOB,woGCM,woIF)	20183	247.9067	105.5192	81.0906	0
GLANET	(NOOB,wGC,wIF)	20183	595.9552	160.3645	33.8115	0
GLANET	(NOOB,wM,wIF)	20183	453.3407	140.4382	44.4248	0
GLANET	(NOOB,wGCM,wIF)	20183	657.1246	168.5808	30.6689	0
GLANET	(NOOB,woGCM,wIF)	20183	413.4114	136.8533	48.7052	0

Supplementary Table 18. Experiment1: Intervals of transcriptor factor Srf in Jurkat cell line are overlapped with DNaseI hypersensitive sites in Jurkat cell line. Both GAT and GLANET find enrichment of DNaseI(Jurkat) for Srf(Jurkat).

Experiment2 Set1: Srf(Jurkat) Set2: DNaseI(HepG2)						
Tool	**Parameter Settings**	**Observed**	**Expected**	**StdDev**	**FoldChange**	**pValue**
GAT	(NOOB, woGCM, woIF)	18965	597.1380	166.9945	31.7084	1.0e-03
GLANET	(EOO,wGC,woIF)	381	49.4944	6.1386	7.5651	0
GLANET	(EOO,wM,woIF)	381	15.8633	3.9072	22.6527	0
GLANET	(EOO,wGCM,woIF)	381	55.9002	6.3335	6.7135	0
GLANET	(EOO,woGCM,woIF)	381	13.5410	3.6388	26.2705	0
GLANET	(EOO,wGC,wIF)	381	55.2896	6.5083	6.7863	0
GLANET	(EOO,wM,wIF)	381	34.2100	5.5440	10.8491	0
GLANET	(EOO,wGCM,wIF)	381	62.4521	6.6809	6.0202	0
GLANET	(EOO,woGCM,wIF)	381	30.8020	5.3329	12.0118	0
GLANET	(NOOB,wGC,woIF)	18965	2298.8933	295.0334	8.2464	0
GLANET	(NOOB,wM,woIF)	18965	699.2644	177.4524	27.0840	0
GLANET	(NOOB,wGCM,woIF)	18965	2592.5174	305.0763	7.3128	0
GLANET	(NOOB,woGCM,woIF)	18965	595.3543	165.2816	31.8032	0
GLANET	(NOOB,wGC,wIF)	18965	2532.3832	310.1418	7.4864	0
GLANET	(NOOB,wM,wIF)	18965	1531.4211	257.4727	12.3764	0
GLANET	(NOOB,wGCM,wIF)	18965	2874.7601	316.5953	6.5951	0
GLANET	(NOOB,woGCM,wIF)	18965	1375.0903	246.4372	13.7825	0

Supplementary Table 19. Experiment2: Intervals of transcriptor factor Srf in Jurkat cell line are overlapped with DNaseI hypersensitive sites in HepG2 cell line. Both GAT and GLANET find enrichment of DNaseI(HepG2) for Srf(Jurkat).

Experiment3 Set1: DNaseI(HepG2) Set2: DNaseI(Jurkat)						
Tool	**Parameter Settings**	**Observed**	**Expected**	**StdDev**	**FoldChange**	**pValue**
GAT	(NOOB,woGCM,woIF)	6163503	456928.2770	8119.7800	13.4890	1.0e-03
GLANET	(EOO,wGC,woIF)	37863	4486.2310	63.3604	8.4381	0
GLANET	(EOO,wM,woIF)	37863	4729.1280	62.8720	8.0048	0
GLANET	(EOO,wGCM,woIF)	37863	4980.2900	63.7331	7.6012	0
GLANET	(EOO,woGCM,woIF)	37863	4021.9370	61.3296	9.4120	0
GLANET	(EOO,wGC,wIF)	37863	4779.9930	62.7600	7.9196	0
GLANET	(EOO,wM,wIF)	37863	5330.1410	66.3065	7.1024	0
GLANET	(EOO,wGCM,wIF)	37863	5304.6820	67.4277	7.1365	0
GLANET	(EOO,woGCM,wIF)	37863	4679.8590	62.3539	8.0891	0
GLANET	(NOOB,wGC,woIF)	6163503	514669.0700	8361.7736	11.9756	0
GLANET	(NOOB,wM,woIF)	6163503	542634.9810	8866.3420	11.3584	0
GLANET	(NOOB,wGCM,woIF)	6163503	577794.1580	9186.0057	10.6672	0
GLANET	(NOOB,woGCM,woIF)	6163503	457457.8130	7800.8096	13.4733	0
GLANET	(NOOB,wGC,wIF)	6163503	548311.7080	8391.4861	11.2408	0
GLANET	(NOOB,wM,wIF)	6163503	616187.0040	9160.8373	10.0026	0
GLANET	(NOOB,wGCM,wIF)	6163503	614923.7840	8718.6997	10.0231	0
GLANET	(NOOB,woGCM,wIF)	6163503	536616.5930	8472.0299	11.4858	0

Supplementary Table 20. Experiment3: DNaseI hypersensitive sites in HepG2 cell line are overlapped with DNaseI hypersensitive sites in Jurkat cell line. Both GAT and GLANET find enrichment of DNaseI(Jurkat) for DNaseI(HepG2).

Experiment4 Set1: Srf(Jurkat) Set2: DNaseI(HepG2-Unique)						
Tool	**Parameter Settings**	**Observed**	**Expected**	**StdDev**	**FoldChange**	**pValue**
GAT	(NOOB,woGCM,woIF)	425	324.6790	117.8233	1.3080	1.85e-01
GLANET	(EOO,wGC,woIF)	9	21.5893	4.4931	0.4426	9.995e-01
GLANET	(EOO,wM,woIF)	9	8.9383	2.9387	1.0062	5.403e-01
GLANET	(EOO,wGCM,woIF)	9	24.3285	4.6873	0.3948	9.998e-01
GLANET	(EOO,woGCM,woIF)	9	7.5673	2.7146	1.1672	3.486e-01
GLANET	(EOO,wGC,wIF)	9	27.1950	4.9426	0.3546	1e+00
GLANET	(EOO,wM,wIF)	9	18.8889	4.2593	0.5027	9.956e-01
GLANET	(EOO,wGCM,wIF)	9	29.8631	5.1835	0.3240	1e+00
GLANET	(EOO,woGCM,wIF)	9	17.0837	4.0756	0.5529	9.878e-01
GLANET	(NOOB,wGC,woIF)	425	951.4744	206.6389	0.4472	9.973e-01
GLANET	(NOOB,wM,woIF)	425	379.5031	131.6084	1.1195	3.46e-01
GLANET	(NOOB,wGCM,woIF)	425	1066.5066	216.2852	0.3990	9.998e-01
GLANET	(NOOB,woGCM,woIF)	425	324.0335	122.5103	1.3106	2.053e-01
GLANET	(NOOB,wGC,wIF)	425	1186.2319	224.0918	0.3588	9.998e-01
GLANET	(NOOB,wM,wIF)	425	816.2769	189.3228	0.5212	9.867e-01
GLANET	(NOOB,wGCM,wIF)	425	1309.7033	235.8895	0.3250	1e+00
GLANET	(NOOB,woGCM,wIF)	425	731.7741	182.2007	0.5813	9.603e-01

Supplementary Table 21. Experiment4: Intervals of transcriptor factor Srf in Jurkat cell line are overlapped with DNaseI hypersensitive sites in HepG2-Unique cell line. Both GAT and GLANET find no enrichment of DNaseI(HepG2-Unique) for Srf(Jurkat).

Here, we provide command line arguments for one of the GLANET runs of Experiment1 in Supplementary Table 18. The input is the genomic intervals of Srf in Jurkat cell line and annotation/enrichment is conducted with respect to DNaseI hypersensitive sites of Jurkat cell line using GLANET's user-defined library feature. The enrichment analysis is conducted with $10,000$ samplings. In this run, GLANET parameter setting is as follows, (`NOOB,wGCM,wIF`). This run took 5 minutes on Intel(R) Core i7-3630QM CPU, 2.40 GHz with 16GB RAM.

```
java -Xms8G -Xmx8G -jar "path/to/GLANET.jar" -c -g "path/to/GLANET Folder/" -i
"path/to/GLANET Folder/Data/demo_input_data/GAT_Comparison_Data/srf.hg19.bed"
-fbed -udl -udldf0exc -udlinput
"path/to/GLANET Folder/Data/demo_input_data/GAT_Comparison_Data/GAT_UDL_InputFile_jurkat.txt"
-e -noob -wgcm -wif -s 10000 -se 10000 -l -j "GLANET_SRF_JURKAT_NOOB_wGCM_wIF"
```

We supply another command line arguments for one of the GLANET runs of Experiment3 in Supplementary Table 20. The input is the genomic intervals of DNaseI hypersensitive sites of HepG2 cell line and annotation/enrichment is

conducted with respect to DNaseI hypersensitive sites of Jurkat cell line using GLANET's user-defined library feature. The enrichment analysis is conducted with 1000 samplings. In this run, GLANET parameter setting is as follows, (`EOO,wM,woIF`). This run took 7 minutes on Intel(R) Core i7-3630QM CPU, 2.40 GHz with 16GB RAM.

```
java -Xms16G -Xmx16G -jar "path/to/GLANET.jar" -c -g "path/to/GLANET Folder/" -i
"path/to/GLANET Folder/Data/demo_input_data/GAT_Comparison_Data/hepg2.hg19.dhs.bed"
-fbed -udl -udldf0exc -udlinput
"path/to/GLANET Folder/Data/demo_input_data/GAT_Comparison_Data/GAT_UDL_InputFile_jurkat.txt"
-e -eoo -wm -woif -s 1000 -se 1000 -l -j "GLANET_HEPG2_JURKAT_EOO_wM_woIF"
```

11 GLANET Runtime Comparison

We compare GLANET against GAT and GREAT with respect to run-time. GLANET and GAT are compared based on genomic interval enrichment, as GAT does not offer gene set enrichment. GREAT comparisons were on the basis of gene set enrichment, as GREAT only offers enrichment based on annotations of nearby genes. All GAT and GLANET run were run on the following system configuration: CPU: Intel(R) Xeon(R) CPU E7-4850 v3 @ 2.20GHz CPU. Memory: 1TB. Operating system: Ubuntu 16.04.2 LTS.

11.1 Comparison with GAT

We compare GAT and GLANET in two different experimental settings. For the first comparison setting, we used the genomic intervals randomly selected from promoter regions of non-expressing genes in GM12878 cell line as input. We conducted two different experiments with this input. In the first one the enrichment analysis of two different genomic element sets, in the first one the entire ENCODE transcription factor and histone modification elements in various cell lines were included (568 files). In the second analysis only the subset of ENCODE elements are checked, these included the 12 histone modification elements and POL2 in cell line GM12878 that are used in the data driven experiment as described in manuscript Section 2.7 (13 files). It's worth noticing that increasing the library size did not increase the runtime that much. We varied the number of intervals and the number of samplings. The resulting runtimes are provided in Supplementary Table 22.

Input Query	**Number of Input Intervals**	**Number of Samplings**	**Running times of tools (in secs)**		
			GLANET - all ENCODE	**GLANET - subset ENCODE**	**GAT - subset ENCODE**
Promoter regions of Non-expressing genes in GM12878	500	1,000	826	690	**145**
	500	10,000	1,169	**856**	1,463
	500	100,000	4,447	**2,140**	14,353
	1000	1,000	1,395	1,283	**147**
	1000	10,000	1,650	**1,165**	1,538
	1000	100,000	9,137	**3,866**	14,341
	2000	1,000	1,396	1,179	**155**
	2000	10,000	2,429	**1,270**	1,583
	2000	100,000	14,724	**6,257**	16,039

Supplementary Table 22. Elapsed CPU times (in seconds) for GLANET and GAT runs for a given input query are provided. Input intervals are randomly selected sets from the promoter regions of non-expressing genes in GM12878 cell line from (Non-Expressing,CompletelyDiscard), where each interval is 601 bps long. All ENCODE checks the enrichment of all ENCODE elements in the GLANET library which encompass histone modifications, transcription factor sites, and DNase I hypersensitive sites for all cell lines. ENCODE subset only include 12 histone modifications and POL2 as described in Section 2.6 of the main manuscript. Both GLANET and GAT are run under the parameter setting `(NOOB,wIF,woGCM)`. Results for 1,000 and 10,000 samplings are averaged over 10 runs. For 100,000 samplings, each run time in the table denotes the average run-time from 5 individual runs.

Input Query	**User Defined Library**	**Number of Samplings**	**Running times of tools (in secs)**					
			GLANET					GAT
			woIF	wIF				wIF
			wGC	wM	wGCM	wGC	woGCM	woGCM
Srf(Jurkat)	DNaseI(Jurkat)	1,000	505	498	741	492	473	**86**
Srf(Jurkat)	DNaseI(Jurkat)	10,000	923	589	1,056	712	**582**	792
Srf(Jurkat)	DNaseI(Jurkat)	100,000	4,158	1,812	3,856	2,710	**1,777**	7,383
DNaseI(HepG2)	DNaseI(Jurkat)	1,000	16,843	7,942	24,602	13,386	2,428	**1,125**
DNaseI(HepG2)	DNaseI(Jurkat)	10,000	167,079	69,248	250,360	127,134	16,693	**12,476**
DNaseI(HepG2)	DNaseI(Jurkat)	100,000	2,066,470	766,951	2,700,420	1,447,620	262,553	**97,659**
Srf(Jurkat)	DNaseI(HepG2)	1,000	518	499	741	509	495	**82**
Srf(Jurkat)	DNaseI(HepG2)	10,000	951	585	1,056	715	**551**	792
Srf(Jurkat)	DNaseI(HepG2)	100,000	4,312	1,779	4,002	2,712	**1,746**	7,296
Srf(Jurkat)	DNaseI(HepG2Unique)	1,000	519	499	752	492	485	**76**
Srf(Jurkat)	DNaseI(HepG2Unique)	10,000	945	596	1,049	692	**565**	701
Srf(Jurkat)	DNaseI(HepG2Unique)	100,000	4,042	1,745	3,987	2,728	**1,734**	6,924

Supplementary Table 23. CPU time in seconds spent for GLANET and GAT runs given the input query specified. For 1,000 and 10,000 samplings, run time is the average of 10 runs. For 100,000 samplings, each run time shows the average run-time from 5 individual runs.

In the second comparison setting, we used data provided in GAT readthedocs web page as described in the Section 3.3 of the manuscript. Srf(Jurkat) is the transcription binding sites of 556 intervals each 51 bps long from Jurkat cell line.

DNaseI(Jurkat) and DNaseI(HepG2) comprised of DNaseI hypersensitive sites in Jurkat and HepG2 cell line, respectively. DNaseI(HepG2Unique) consists of DNaseI hypersensitive sites in HepG2 but not in Jurkat cell line. For your information, DNaseI(Jurkat) have 159,613 intervals each 151 bps long. DNaseI(HepG2) have 144,171 intervals of average 360 bps long. DNaseI(HepG2Unique) have 106,308 intervals of average 275 bps long. The results are listed in Supplementary Table 23.

All the running time results for GLANET and GAT are shown in terms of CPU time in seconds. This is the actual time that one CPU would need to complete its process. Thus, these running times are the sum of the times taken in each thread for a run if multithreading is available (time command in Unix). Since GLANET is a multi-threaded java application, as the same runs are conducted with a larger number of threads, the elapsed time for an individual run would decrease accordingly. During these runs, GLANET and GAT 16GB of memory is reserved, except for DNaseI(Hepg2)-DNaseI(Jurkat) runs of 100,000 samplings, GLANET required 64GB of memory.

11.2 Comparison with GREAT

Input Query	Enrichment	Number of Samplings	Running times of tools (in secs)			
			GLANET			
			Association Measure	Random Interval Generation	Isochore Family	Runtime
GATA2 Binding sites in K562	BP GO Terms	1,000	NOOB	woGCM	woIF	522,75
	BP, MF and CC GO Terms	1,000	NOOB	woGCM	woIF	658,34
	BP GO Terms	10,000	NOOB	woGCM	woIF	3069,14
	BP, MF and CC GO Terms	10,000	NOOB	woGCM	woIF	5808,08
	BP GO Terms	10,000	NOOB	wGC	wIF	6838,32
	BP, MF and CC GO Terms	10,000	NOOB	wGC	wIF	9718,03

Supplementary Table 24. CPU time in seconds spent for GLANET runs given the input query specified. For 1,000 and 10,000 samplings, each run time is the average of 10 individual runs.

We compared GREAT on the basis of GO term gene set enrichment. The input was GATA2 transcription bindings sites in K562 cell line and their enrichment is checked against gene sets derived from GO terms as described in Section 3.4.3 in the main text. The input included 7407 intervals of average 256 bps long. Supplementary Table 24 includes the results. GREAT is not available as a stand-alone command line application, thus, the results are obtained from the online web service. Nevertheless, when run on from the server, GREAT was very fast, it completed one analysis in less than 1 minute, as its enrichment procedure does not do include sampling but instead assumes a parametric distribution. Since we do not know how each GREAT run is parallelized in their server, we do not know the actual CPU hour. Therefore, it is not possible for us to compare the running times in terms of CPU hours and we do not have information on the actual memory used for the GREAT analysis.

12 Enrichment Analysis of OCD GWAS SNPs

The GLANET command line arguments to replicate GLANET run for OCD GWAS SNPs is provided below. Input query is the set of SNPs found to be associated with OCD. These SNPs are overlapped with the intervals in default annotation library of GLANET. The enrichment analysis included $10,000$ samplings and were conducted with respect to the GLANET default annotation library that included all ENCODE data (DNaseI hypersensitive sites, transcription factor binding sites and histone modification sites in all cell lines), RefSeq Genes, KEGG pathways, joint TF KEGG pathways (cell line pooled). The complete analysis took 19 minutes on Intel(R) Core i7-3630QM CPU, 2.40 GHz with 16GB RAM.

```
java -Xms16G -Xmx16G jar "path/to/GLANET.jar" -c -g "path/to/GLANET Folder/"
-i "path/to/GLANET Folder/Data/demo_input_data/OCD_GWAS_chrNumber_1Based_GRCh37_p13_Coordinates.txt"
-f1 -dnase -histone -tf -gene -kegg -tfkegg
-e -noob -wgcm -wif -s 10000 -se 10000 -l -j "GLANET_OCD_GWAS_SNPs_NOOB_wGCM_wIF"
```

13 Regulatory Sequence Analysis of OCD GWAS SNPs

For a transcription factor, there can be more than one position frequency matrices (PFMs) available. We scan the sequences with all of the PFMs of that TF. RSAT returns the subsequence that achieves the smallest p-value along with the PFM that leads to that p-value.

RSAT's best matching motif might not contain the SNP locus. If that is the case, GLANET finds the subsequence containing the SNP with the smallest p-value among the remaining results. Therefore, as it can be seen in the provided RSA results of rs1891215 (STAT1) and rs10946279 (MAX), we have two results line for each sequence. RSA results of STAT1 and MAX are shown in Supplementary Table 25 and 26.

rs1891215(chr1,7667794) STAT1								
Sequence	PFM Name	Database	Direction	Start	End	Sequence	pValue	log
Reference sequence	MA0137.3 STAT1	JASPAR CORE	R	18	28	CTTCTGGAAAA	1.10e-03	
Reference sequence containing SNP locus	MA0137.3 STAT1	JASPAR CORE	R	18	28	CTTCTGGAAAA	1.10e-03	
SNP sequence	MA0137.3 STAT1	JASPAR CORE	R	18	28	CTTCTGGGAAA	6.10e-05	
SNP sequence containing SNP locus	MA0137.3 STAT1	JASPAR CORE	R	18	28	CTTCTGGGAAA	6.10e-05	1.26e+00
TFExtendedPeak sequence	MA0137.3 STAT1	JASPAR CORE	R	34	44	CTTCTGGAAAA	1.10e-03	
TFExtendedPeak sequence containing SNP locus	MA0137.3 STAT1	JASPAR CORE	R	34	44	CTTCTGGAAAA	1.10e-03	

Supplementary Table 25. Regulatory sequence analysis result for rs1891215 and transcription factor, STAT1. For each of the subsequences, the best matching motif site is provided together with the position frequency matrix that leads to that result and the source of the matrix. The direction column indicates whether the match is within the forward (D) or the reverse (R) strand. rs1891215 is an enhancer SNP for transcription factor STAT1 because nucleotide change at SNP locus, from A to G increases STAT1's binding affinity. p_{snp} is 6.10e-05 while p_{ref} is 1.10e-03. Last column lists $\log(p_{ref}/p_{snp})$ value where p_{ref} is the p-value of reference sequence containing SNP position and p_{snp} is the p-value of the match for the altered SNP sequence containing SNP position.

rs10946279 (chr6,170553248) MAX								
Sequence	PFM Name	Database	Direction	Start	End	Sequence	pValue	log
Reference sequence	MYC known8 MAX 3	JASPAR CORE	R	14	23	GCCGTGCGAT	6.10e-05	
Reference sequence containing SNP locus	MYC known8 MAX 3	JASPAR CORE	R	14	23	GCCGTGCGAT	6.10e-05	
SNP sequence	MYC disc9 MAX HUVEC encode Snyder	ENCODE MOTIFS	D	26	35	CGCCTGCGGA	5.50e-04	
SNP sequence containing SNP locus	MYC known19 MAX 3	JASPAR CORE	R	14	29	GGCGCAGCTG TGCGAT	1.50e-03	-1.39e+00
TFExtendedPeak sequence	MYC known8 MAX 3	JASPAR CORE	R	65	74	GCCGTGCGAT	6.10e-05	
TFExtendedPeak sequence containing SNP locus	MYC known8 MAX 3	JASPAR CORE	R	65	74	GCCGTGCGAT	6.10e-05	

Supplementary Table 26. Regulatory sequence analysis result for rs10946279 and transcription factor, MAX. For each of the subsequences, the best matching motif site is provided together with the position frequency matrix that lead to that result and the source of the matrix. The direction column indicates whether the match is within the forward (D) or reverse (R) strand. rs10946279 acts as a repressor SNP for transcription factor MAX, since nucleotide change at SNP locus, from C to T decreases the TF's binding affinity. p_{ref} is 6.10e-05 where as p_{snp} is 1.50e-03. Last column lists $\log(p_{ref}/p_{snp})$ value where p_{ref} is the p-value of reference sequence containing SNP position and p_{snp} is the p-value of the match for the altered SNP sequence containing SNP position..

GLANET's command line arguments for Regulatory Sequence Analysis for OCD GWAS SNPs:

```
java -Xms4G -Xmx4G -jar "path/to/GLANET.jar" -c -g "path/to/GLANET Folder/"
-i "path/to/GLANET Folder/Data/demo_input_data/OCD_GWAS_chrNumber_1Based_GRCh37_p13_Coordinates.txt"
-f1 -tf -rsa -l -j "GLANET_OCD_GWAS_SNPs_RSA"
```

14 Analysis with User-Defined Gene Sets: Gene Ontology Enrichment Analysis for GATA2 Binding Regions

We ran GLANET with genomic intervals of GATA2 in K562 cell line as input query and loaded the GO terms in the library using the user-defined gene set feature of GLANET. Annotation and enrichment of the input with respect to the new library is conducted. The enrichment analysis is conducted with $10,000$ samplings. The analysis can be repeated with the with the command line arguments below:

```
java -Xms16G -Xmx16G -jar "path/to/GLANET.jar" -c -g "path/to/GLANET Folder/" -i
"path/to/GLANET Folder/Data/demo_input_data/RDA2/
spp.optimal.wgEncodeSydhTfbsK562bGata2UcdAlnRep0_VS_wgEncodeSydhTfbsK562bInputUcdAlnRep1.narrowPeak"
-fbed -udgs -udgsinput
"path/to/GLANET Folder/Data/demo_input_data/UserDefinedGeneSet/GO/GO_gene_associations_human_ref.txt"
-genesym -udgsname "GO" -udgsdfile
"path/to/GLANET Folder/Data/demo_input_data/UserDefinedGeneSet/GO/GO_ids2terms.txt"
-e -noob -wgcm -wif -s 10000 -se 1000 -l -j "GLANET_SydhGATA2K562_GoTerms_NOOB_wGCM_wIF"
```

15 Additional Supplementary Tables

Supplementary Table 27: OCD GWAS SNPs that overlap with genes of glutamatergic synapse pathway (hsa04724) in the regulation based analysis. The table lists the locations of SNPs and genes, which region of the gene the SNP resides in. Excel file: SupplementaryTable27.xlsx

Supplementary Table 28: OCD GWAS SNP list enrichment analysis with GLANET default library. Results include enriched DNaseI hypersensitive sites, transcription factor binding sites, modification regions for multiple histones across all cell lines available in ENCODE, RefSeq genes, KEGG pathways, TF-KEGG pathway pairs. KEGG pathway enrichment analyses are conducted in three modes: exon, regulation and all-based. Excel file: SupplementaryTable28.xlsx

Supplementary Table 29: List of interesting SNP and transcription factor pairs where SNP allele potentially affects the transcription factor binding event. Excel file: SupplementaryTable29.xlsx

Supplementary Table 30: Output of GLANET enrichment analysis for the user-defined gene set option. Gene ontology (GO) enrichment analysis for GATA2 binding regions in K562 cell line conducted in three modes: exon-based, regulation-based and all-based. Excel file: SupplementaryTable30.xlsx

References

[1] Costantini, M., Clay, O., Auletta, F., and Bernardi, G. (April, 2006) An isochore map of human chromosomes. *Genome research,* **16**(4), 536–541.

[2] Bernardi, G. (October, 2001) Misunderstandings about isochores. Part 1. *Gene,* **276**(1-2), 3–13.

[3] Cormen, T. H. (2009) Introduction to algorithms, MIT Press, Cambridge, Mass. 3rd edition.

[4] Thomas-Chollier, M., Sand, O., Turatsinze, J.-V., Janky, R., Defrance, M., Vervisch, E., Brohée, S., and van Helden, J. (July, 2008) RSAT: regulatory sequence analysis tools. *Nucleic Acids Research,* **36**(suppl 2), W119–W127.

[5] Mathelier, A., Zhao, X., Zhang, A. W., Parcy, F., Worsley-Hunt, R., Arenillas, D. J., Buchman, S., Chen, C.-y. Y., Chou, A., Ienasescu, H., Lim, J., Shyr, C., Tan, G., Zhou, M., Lenhard, B., Sandelin, A., and Wasserman, W. W. (January, 2014) JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles.. *Nucleic acids research,* **42**(Database issue), D142–D147.

[6] Kheradpour, P. and Kellis, M. (December, 2013) Systematic discovery and characterization of regulatory motifs in ENCODE TF binding experiments. *Nucleic Acids Research,* **42**(5), gkt1249–2987.

[7] Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., and Mller, M. (2011) pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics,* **12**, 77.

[8] GAT Tutorial. `https://gat.readthedocs.org` (2013) Last Accessed: 2016-05-13.