Bioinformatics, 32(4), 2016, 490–496
doi: 10.1093/bioinformatics/btv565
Advance Access Publication Date: 17 October 2015
Original Paper

OXFORD

Genome analysis

GERV: a statistical method for generative evaluation of regulatory variants for transcription factor binding

Haoyang Zeng[1], Tatsunori Hashimoto[1], Daniel D. Kang[1] and David K. Gifford[1,2,*]

[1]Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA 02142, USA and [2]Department of Stem Cell and Regenerative Biology, Harvard University and Harvard Medical School, Cambridge, MA 02138, USA

*To whom correspondence should be addressed.

Associate Editor: John Hancock

Received on July 5, 2015; revised on September 11, 2015; accepted on September 22, 2015

Abstract

Motivation: The majority of disease-associated variants identified in genome-wide association studies reside in noncoding regions of the genome with regulatory roles. Thus being able to interpret the functional consequence of a variant is essential for identifying causal variants in the analysis of genome-wide association studies.

Results: We present GERV (generative evaluation of regulatory variants), a novel computational method for predicting regulatory variants that affect transcription factor binding. GERV learns a k-mer-based generative model of transcription factor binding from ChIP-seq and DNase-seq data, and scores variants by computing the change of predicted ChIP-seq reads between the reference and alternate allele. The k-mers learned by GERV capture more sequence determinants of transcription factor binding than a motif-based approach alone, including both a transcription factor's canonical motif and associated co-factor motifs. We show that GERV outperforms existing methods in predicting single-nucleotide polymorphisms associated with allele-specific binding. GERV correctly predicts a validated causal variant among linked single-nucleotide polymorphisms and prioritizes the variants previously reported to modulate the binding of FOXA1 in breast cancer cell lines. Thus, GERV provides a powerful approach for functionally annotating and prioritizing causal variants for experimental follow-up analysis.

Availability and implementation: The implementation of GERV and related data are available at http://gerv.csail.mit.edu/.

Contact: gifford@mit.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Genome-wide association studies (GWAS) have revealed genetic polymorphisms that are strongly associated with complex traits and diseases (Hindorff *et al.*, 2009; Manolio, 2010; McCarthy *et al.*, 2008; Stranger *et al.*, 2011). Missense and nonsense variants that occur in protein coding sequences are simple to characterize. However, many GWAS-detected variants reside in non-coding regions with regulatory function (Frazer *et al.*, 2009; Hindorff *et al.*, 2009). The influence of non-coding variation on gene expression and other cellular functions is not well understood. Previous work has observed that non-coding DNA changes in the recognition sequences of transcription factors can affect gene expression and cellular phenotypes (Ward and Kellis,

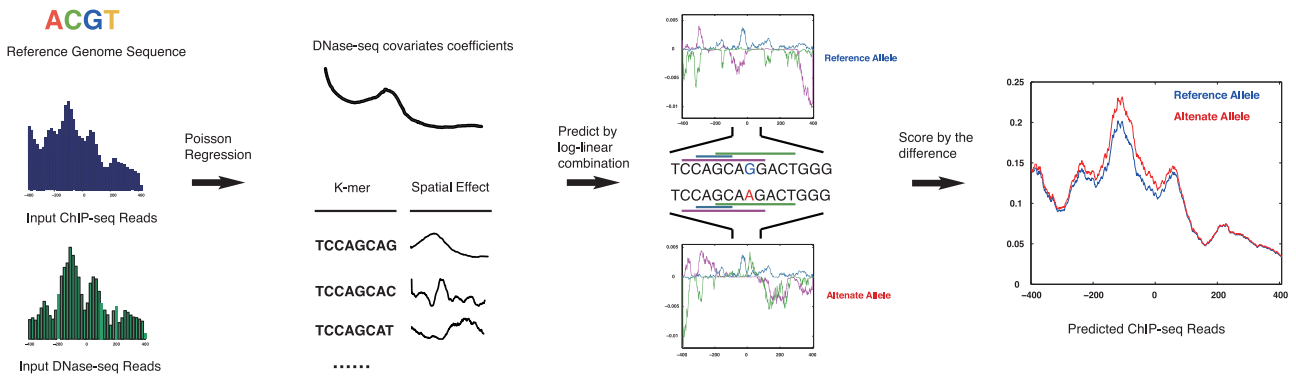

Fig. 1. The schematic of GERV. The spatial effects of all the k-mers and the DNase-seq covariates are learned from the reference genome sequence and ChIP-seq, DNase-seq datasets. Then the spatial effects (purple, cyan andgreen) of the k-mers underlying the reference (blue) and alternate (red) allele for a variant are aggregated with DNase-seq covariates by log-linear combination to yield a spatial prediction of localChIP-seq reads for the two alleles. GERV scores the variant by the l^2-norm of the predicted change of reads

2012b). Thus, predicting the effect of genomic variants on transcription factor (TF) binding is an essential part of interpreting the role of non-coding variants in pathogenesis. Most of existing computational approaches to predict the effect of single-nucleotide polymorphism (SNPs) on TF binding such as sTRAP and HaplogReg are based on quantifying the difference between the presented reference and alternate alleles in the context of canonical TF binding motifs (Andersen *et al.*, 2008; Macintyre *et al.*, 2010; Manke *et al.*, 2010; Molineris *et al.*, 2013; Riva, 2012; Teng *et al.*, 2012; Ward and Kellis, 2012a; Zuo *et al.*, 2015). Recent work (Lee *et al.*, 2015) uses k-mer weights learned from a gapped-kmer SVM (Ghandi *et al.*, 2014) to score the effect of variants, taking into account the frequency of k-mer occurrences but not the spatial effect of k-mers.

Here, we present GERV (generative evaluation of regulatory variants), a novel computational model that learns the spatial effect of k-mers on TF binding *de novo* from whole-genome ChIP-seq and DNase-seq data, and scores variants by the change in predicted ChIP-seq read counts between the reference and alternate alleles. GERV improves on existing models in three ways. First, GERV does not assume the existence of a canonical TF binding motif. Instead it models transcription factor binding by learning the effects of specific k-mers on observed binding. This allows GERV to capture more subtle sequence features underlying transcription factor binding including non-canonical motifs. Second, GERV accounts for the spatial effect of k-mers and learns the effect of cis-regulatory regions outside of the canonical TF motif. This enables us to model the role of important auxiliary sequences in transcription factor binding, such as cofactors. Third, GERV incorporates chromatin openness information as a covariate in the model which boosts the accuracy of the predicted functional consequence of a variant.

We first demonstrate the power of GERV on the ChIP-seq data for transcription factor NF-κB. We show that GERV learns a vocabulary of k-mers that accurately predicts held-out NF-κB ChIP-seq data and captures the canonical NF-κB motifs and associated sequences such as known co-factors. Applying GERV to six transcription factors on which allele-specific binding (ASB) analysis is available, we show GERV outperforms existing approaches in prioritizing SNPs associated with ASB. We demonstrate the application of GERV in post-GWAS analysis by scoring risk-associated SNPs and their linked SNPs for breast cancer and show that GERV trained on FOXA1 ChIP-seq data achieves superior performance in prioritizing SNPs previously reported to modulate FOXA1 binding in breast cancer cell lines.

2 Methods

2.1 GERV model overview

GERV is a fully generative model of ChIP-seq reads. We assume that the genome is a long regulatory sequence containing k-mer 'code words' that induce invariant spatial effects on proximal transcription factor binding. We use the level of chromatin openness in a region as a functional prior to predict the magnitude of a sequence-induced binding signal. Following this assumption, we model the read counts produced by transcription factor ChIP-seq at a given base as the log-linear combination of the DNase-seq signal on nearby bases and the spatial effect of a set of learned k-mers whose effect range covers that base.

The GERV procedure of variant scoring consists of the following three steps (Fig. 1):

1. GERV first learns the spatial effects of all the k-mers ($k = 1$–8) and the DNase-seq covariates over a spatial window of ± 200 bp *de novo* from ChIP-seq data using regularized Poisson regression
2. GERV then computes the predicted ChIP-seq read counts for the reference and alternate allele of a variant from the log-linear combination of the local DNase-seq signal and spatial effect of the learned k-mers.
3. GERV predicts the effect of a genomic variant on transcription factor binding by the l^2-norm of the change of predicted reads between two alleles

2.2 Learning the spatial effect of k-mers

The effect profile of a k-mer is defined as a real-valued vector of length $2M$ that corresponds to a spatial window of $[-M, M-1]$ relative to the start position of the k-mer. Specifically, the jth entry of the profile for a k-mer is the expected log-change in read counts at the jth base relative to the start of the k-mer. Here, we consider k-mers with k from 1 to 8 ($k_{\max} = 8$) as this is the maximum that would fit in memory in an Amazon EC2 c3.8 xlarge instance. Larger k-mers tested on a larger memory machine did not perform substantially better than 8-mers. As ChIP-seq signals are relatively sparse and spikey, we chose an effect range of ± 200 bp for each k-mer ($M = 200$).

For notational convenience, we use i for genomic coordinate, k for k-mer length and j for coordinate offset from the start of a k-mer. We assume that the genome consists of one large chromosome with coordinates 0–N. In practice, we will construct this by concatenating chromosomes with the telomeres acting as a spacer.

We represent the effect vector of all k-mers of length k as a parameter matrix θ^k of size $4^k \times 2M$. For any particular k-mer of length k starting at base i on the reference genome, we define g_i^k as its row index in θ^k. So $\theta^k_{(g_i^k, j)}$ would denote the effect of this kmer at offset $j \in [-M, M-1]$. Additionally, a special parameter θ_0 is used to set the average read rate of the genome globally.

The DNase-seq covariate κ is defined as a binary vector of length N that denotes whether each base of the genome has any DNase-seq read, and we assume that ChIP-seq reads can be predicted with this covariate and the contributions from surrounding k-mers. We define the spatial effect of the covariate as β, a vector of length $2L$ which can be thought of as analogous to the k-mer effect θ but occurring everywhere and scaled by the binary covariate κ. In all the experiments in this analysis, we chose an $L = 200$ to balance between computational complexity and prediction power.

Given these definitions, we define a generative model for ChIP-seq reads on the genome. Observed counts at position i on the genome are generated from a Poisson distribution with rate parameter λ_i, which is defined as:

$$\lambda_i = \exp\left(\left(\sum_{k\in[1,k_{\max}]}\sum_{j\in[-M,M-1]}\theta^k_{(g^k_{i+j},-j)}\right) + \left(\sum_{l\in[-L,L-1]}\beta_{-l}\times\kappa_{i+l}\right) - \theta_0\right) \quad (1)$$

The problem we solve is a regularized Poisson regression. Particularly, we would like to maximize the following:

$$\max_{\theta,\beta}\{(\sum_i c_i \log(\lambda_i) - \lambda_i) - \eta\sum|\theta^k|_1\} \quad (2)$$

To efficiently optimize this objective function, we performed an accelerated gradient descend method. The detail of implementation can be found in the Supplementary Data (Supplementary Text S1).

2.3 ChIP-seq signal prediction for reference and alternate allele

In step 2, given the effect profiles of all the k-mers and the DNase-seq covariates trained from step 1, we first predict the ChIP-seq count λ at each position across the reference genome by combining the effect of proximal k-mers and DNase-seq level into the log-linear model using Equation (1). Then in similar manner, we predict the read counts λ' of the alternate allele after replacing the k-mers that are affected by the variant. If we assume an SNP, at most $\frac{4}{3}\times(4^{k_{\max}}-1)$ k-mers will change.

2.4 Variant scoring

In step 3, we score an SNP at locus on the genome by the square root of the sum of squared per-base change (l^2-norm of the change) of binding signal at all bases within the effect range of any k-mers affected by the variant:

$$s_i = \sqrt{\sum_{j\in[-M-k_{\max}+1,M-1]}(\lambda_{i+j}{}' - \lambda_{i+j})^2} \quad (3)$$

2.5 Collapsing GERV k-mers into a position weight matrix

We interpret the active k-mers captured by GERV with a post-processing framework that aggregates similar k-mers into position weight matrixes (PWMs):

1. We filter k-mers based on the sum of spatial effect to eliminate inactive k-mers.
2. We calculate the Levenshtein distance (number of single character edits) between the remaining k-mers.
3. We perform UPGMA hierarchical clustering over the candidate k-mers until the minimal distance among clusters is larger than 2.
4. For each cluster, we define its key k-mer as the one with the largest sum of spatial effect. We obtain the PWM for this cluster by aligning all k-mers in the cluster against the key k-mer.
5. All the clusters are ranked by the average sum of spatial effect of all the k-mers in the cluster.

2.6 ChIP-seq peak prediction comparison

Gapped-kmer SVM was downloaded from http://www.beerlab.org/gkmsvm/index.html. To match with the training data for GERV, the positive training set for gapped-kmer SVM consists of the all the NF-κB ChIP-seq peaks on chr1-13 of GM12878 from ENCODE, and the negative training set consists of the same number of randomly sampled regions of similar size on chr1-13. The default parameter set ('-d 3') was used. Both GERV and gapped-kmer SVM were evaluated on the same test set. The positive test set consists of all the NF-κB ChIP-seq peaks on chr14-22 of GM12878 from ENCODE, and the negative test set consists of the same number of randomly sampled regions of similar size on chr14-22.

2.7 Benchmark the performance in prioritizing SNPs with ASB

2.7.1 deltaSVM

deltaSVM source code was downloaded from http://www.beerlab.org/deltasvm/. For each transcription factor included in the benchmarking, a gapped-kmer SVM model was trained using ChIP-seq peaks of that factor on chr1-13 of GM12878 from ENCODE as positive sets and the same number of randomly sampled region of similar size on chr1-13 as negative sets. The default parameter set ('-d 3') was used. As instructed by the software, the gapped-kmer SVM model was then used to score all the possible 10-mers, the result of which was input as the kmer-weight parameter to deltaSVM.

2.7.2 sTRAP

We used the R version of sTRAP downloaded from the website (http://trap.molgen.mpg.de/download/TRAP_R_package/) for scalability. The built-in JASPAR and TRANSFAC motif data included in the package were used. Specifically, MA0105.1, MA0105.2, MA0105.3, MA0107.1, MA0061.1, V$NFKAPPAB_01, V$NFKB_Q6, V$NFKAPPAB65_01, V$NFKAPPAB50_01, V$P50_Q6, V$NFKB_C and V$RELA_Q6 were used for NF-κB. MA0139.1, MA0531.1, V$CTCF_01, V$CTCF_02 were used for CTCF. MA0099.1, MA0099.2, MA0476.1 and V$CFOS_Q6 were used for FOS. MA0059.1, MA0058.1, MA0058.2, PB0043.1, PB00147.1, VMAX_01, VMAX_04, VMAX_Q6, VMYCMAX_01, V$MYCMAX_02, V$MYCMAX_03 and V$MYCMAX_B were used for MAX. MA0059.1, MA0147.1,MA0147.2, V$CMYC_01, V$CMYC_02, VMYC_01, VMYCMAX_01, V$MYCMAX_02, V$MYCMAX_03 and V$MYCMAX_B were used for MYC. None of the JUND motifs were included in the built-in motif database of sTRAP. For each variant, the scores from different matrices of the same factor were combined by taking the highest one.

3 Materials

3.1 ChIP-seq data

ChIP-seq data for all the factors used in this analysis were downloaded from ENCODE. The full list of GEO accession numbers can be found in Supplementary Table S1.

3.2 DNase-seq data

DNase-seq data of GM12878 were downloaded from ENCODE (GEO accession GSM816665)

3.3 Allele-specific binding SNPs

As a gold standard for SNPs that affect TF binding, we used the list of SNPs that are reported to induce ASB of NF-κB, CTCF, FOS, JUND, MAX and MYC in GM12878. The NF-κB ASB SNPs are collected from Rozowsky *et al.* (2011) and Karczewski *et al.* (2013). The ASB SNPs data for all other transcription factors are collected from Rozowsky *et al.* (2011).

4 Results

4.1 GERV learns a vocabulary of k-mers that regulate factor binding

We first tested if GERV could predict held-out ChIP-seq data. We trained a GERV model on ENCODE NF-κB ChIP-seq data and DNase-seq data from chromosomes 1 to 13 of GM12878 and compared the predicted ChIP-seq signal from GERV to actual ChIP-seq reads on the held-out chromosomes 14–22. The predicted ChIP-seq signals are very similar to actual ChIP-seq reads (Fig. 2A and B), with a chromosome-wide Pearson's correlation of 0.76. We measured correlation after smoothing predicted and actual reads over 400 bp windows since actual reads are insufficiently sampled to produce base-pair resolution measurements. To further examine the ability of GERV to model ChIP-seq peaks, we used the GERV model trained above to score a positive set of regions defined as all the ENCODE GM12878 NF-κB ChIP-seq peaks on chr14-22, and a negative set of regions defined as same number of randomly sampled region of similar length on chr14-22. Each region was scored by the sum of predicted signal in the region. We compared GERV with a previously published kmer-based model for TF peak prediction by training a gapped-kmer SVM (Ghandi *et al.*, 2014) on ENCODE NF-κB peaks and same number of randomly sampled region of similar length on chr1-13 of GM12878 and then performing the same scoring task on the same positive and negative set. We quantified the performance of these two models in prioritizing positive regions over negative regions by calculating the area under receiver operating characteristic (ROC) curve (Fig. 2C). Our model achieved a better area under ROC curve of 0.972 than that of 0.949 for gapped-kmer SVM. Thus, GERV learns a vocabulary of k-mers that can accurately predict the ChIP-seq data.

Although GERV fits a model with a potentially large parameter space ($\pm$200 bp window for 87 380 k-mers when $k_{\max} = 8$), it uses sparsifying regularization to avoid overfitting and to limit the number of active k-mers (Equation 2). For example, in the NF-κB GERV model, most of the l^1-norm of the parameter matrix is contained in the top 1% of the 87 380 k-mers (Supplementary Fig. S1). GERV is also robust to the choice of the window size for a k-mer's spatial effect and DNase-seq covariates (Supplementary Table S2).

4.2 GERV captures the binding sequence of a TF and its co-factors

We then examined if GERV learned the sequence features important for transcription factor binding. We trained a GERV model on

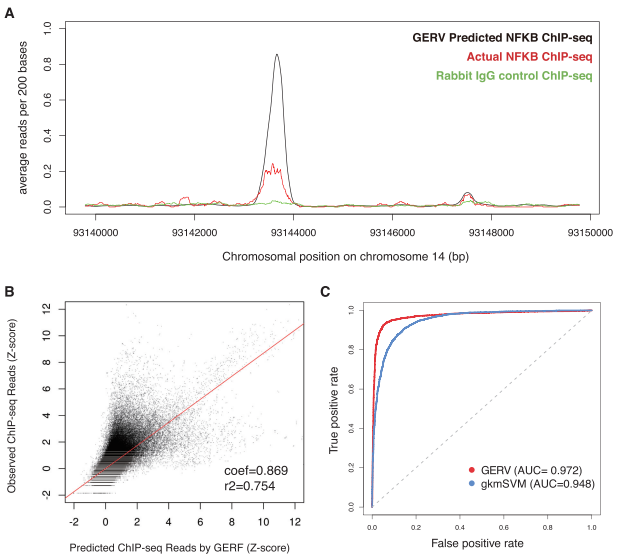

Fig. 2. (**A**) Example held-out genomic region on chromosome 14 showing GERV-predicted NF-κB reads (black), actual NF-κB ChIP-seq reads (red) and rabbit IgG control ChIP-seq reads (green). (**B**) Comparison of GERV-predicted (*x*-axis) and observed (*y*-axis) NF-κB ChIP-seq reads in binned regions of held-out chromosomes 14–22. The coefficient and r^2 of a linear regression on predicted and actual *z*-score is plotted. (**C**) ROC curve for discriminating NF-κB peaks from negative control sets using GERV and gapped-kmer SVM (gkmSVM)

DNase-seq data and NF-κB ChIP-seq data combined from 10 lymphoblastoid B cell lines (LCL) individuals. PWMs were generated for visualization purposes by hierarchical clustering of the active k-mers in GERV (Section 2.5) and matched to known TF motifs in JASPAR and TRANSFAC using STAMP (Mahony and Benos, 2007). With a threshold of significant matching at 1e-7, many clusters of the active k-mers correspond to known motifs (Table 1). The top two k-mer clusters for NF-κB were matched to motifs from NF-κB family (Supplementary Fig. S2A), indicating that GERV correctly learned the strongest expected sequence features for the binding. Moreover, many of the other k-mer clusters learned by GERV correspond to transcription factors, which have been associated with NF-κB regulation (Supplementary Fig. S2B), including ETS1, AP1, IRF1 and SP1 (Bartels *et al.*, 2007; Fujioka *et al.*, 2004; Sgarbanti *et al.*, 2008; Thomas *et al.*, 1997). To validate the role of these transcription factors in NF-κB binding, we performed co-factor analysis on the same NF-κB data using GEM (Guo *et al.*, 2012) to search for transcription factors that have spatially binding constraint with NF-κB. This analysis identified AP-1 and IRF1 as the strongest co-factors of NF-κB binding. Interestingly, some of the active-kmer clusters in GERV were matched to transcription factors such as ELF1, ERF2, CTCF and SUT1, which have not been associated with NF-κB binding in previous studies.

To further interpret the role of the transcription factors whose motifs were matched to an active-kmer clusters in the NF-κB GERV model, we performed motif analysis on the SNPs known to alter transcription factor binding. ASB studies have identified SNPs associated with significantly imbalanced binding events on heterozygous sites (Rozowsky *et al.*, 2011; Karczewski *et al.*, 2013). Therefore, we collected a list of 56 ASB SNPs for NF-κB and use HaploReg (Ward and Kellis, 2012b) to query for the motifs that these ASB SNPs altered (Supplementary Table S3). Among the 56 ASB SNPs tested, only 16 (29%) were found to alter the canonical motif of

Table 1. TF motifs matched to active-kmer clusters in NF-κB GERV model using STAMP with *E*-value cutoff of 1e-07

Cluster	Matched motif	Motif database	Matched TF	*E*
PWM1	M00053	TRANSFAC	REL	5.1842e-08
	MA0101.1	JASPAR	REL	1.2145e-09
PWM2	M00053	TRANSFAC	REL	7.0388e-14
	MA0101.1	JASPAR	REL	1.1385e-12
PWM3	M00495	TRANSFAC	Bach1	8.0813e-13
	MA0099.2	JASPAR	AP1	9.4186e-10
PWM4	M01111	TRANSFAC	RBP-Jkappa	6.0791e-08
PWM5	M00339	TRANSFAC	ETS1	1.3508e-11
	MA0080.2	JASPAR	SPI1	3.6387e-10
PWM7	M01057	TRANSFAC	ERF2	2.0724e-08
	MA0123.1	JASPAR	abi4	1.9650e-08
PWM12	MA0139.1	JASPAR	CTCF	1.1289e-08
PWM15	M00062	TRANSFAC	IRF1	2.7655e-10
	MA0050.1	JASPAR	IRF1	3.1444e-09
PWM18	MA0399.1	JASPAR	SUT1	2.7097e-08
PWM20	M00722	TRANSFAC	core-binding	3.6831e-09
PWM22	MA0242.1	JASPAR	run_Bgb	3.1286e-11
PWM23	M01066	TRANSFAC	BLIMP1	1.4886e-09
PWM27	MA0453.1	JASPAR	nub	4.2762e-09
PWM32	M00345	TRANSFAC	GAMYB	8.8633e-08
PWM33	MA0344.1	JASPAR	NHP10	2.3002e-09
PWM38	MA0403.1	JASPAR	TBF1	5.6499e-08
PWM43	M00181	TRANSFAC	E2	2.9261e-08
PWM49	MA0152.1	JASPAR	NFATC2	5.3905e-11

For each cluster, only the strongest match in each motif database (TRANSFAC and JASPAR) is shown. PWMs are ordered by the average sum of spatial effect of all the k-mers in the corresponding cluster.

NF-κB, while another 11 (20%) were found to alter the TF motif matched to other active-kmer clusters in the GERV model. Thus, GERV captures the sequence context of factor binding, which provides additional descriptive power and biological insight for auxiliary elements in TF binding.

4.3 GERV outperforms existing approaches in prioritizing ASB SNPs

To demonstrate the power of GERV in detecting regulatory variants, we compared GERV's performance against existing approaches in discriminating ASB SNPs from negative control variants. We collected ASB SNPs with known differential binding for NF-κB, CTCF, JUND, MAX, MYC and FOS from previous studies (Karczewski *et al.*, 2013; Rozowsky *et al.*, 2011) as positive sets, resulting in a total of 56 SNPs for NF-κB, 1225 SNPs for CTCF, 26 SNPs for FOX, 233 SNPs for JUND, 71 SNPs for MAX and 69 SNPs for MYC (Section 3.3). Note that these ASB SNPs were completely held-out in the training process of any model compared in this analysis and were only used as the test set.

For each of the six transcription factors, we constructed two types of negative SNP sets that we assume do not exhibit differential factor binding. Both kinds of negative sets are subsets of 1000 Genome Project (1KG) SNPs. In the first case, we randomly sampled 100 negative samples for each positive sample, to get a reasonable sample of the background while making analyses computationally tractable. The second set is a fine-mapping task which is an important topic in post-GWAS analysis where a list of lead SNPs and their linked SNPs are under interrogation for regulatory consequence. To simulate such tasks, this second set was constructed as random selection of 1KG SNPs within 10 kb from any ASB SNP. To reflect the number of SNPs typically in a single LD block, we calculated LD information from phased genotype data in the 1KG pilot release using PLINK (Purcell *et al.*, 2007). With a r^2 cutoff of 0.8, the median number of linked SNPs for a variant is 10 (Supplementary Fig. S3). Thus, in this set, we sampled 10 negative samples for each positive sample. For both types of negative sets, we sampled 10 sets with replacement, so that we could obtain the mean and confidence intervals. For each of the 10 negative sets, we constructed a paired positive set, same size as the corresponding ASB SNP set, by sampling with replacement from the ASB SNPs.

For each transcription factor, we evaluated the performance of GERV and two published regulatory variant scoring methods sTRAP (Manke *et al.*, 2010) (motif-based) and deltaSVM (Lee *et al.*, 2015) (kmer-based) in discriminating the positive set from each of the two negative sets. The other motif-based methods are not included due to either the inability to produce numerical scores for the queried variants or the low throughput that cannot scale up to thousands of SNPs. For each factor, a GERV model was trained on ENCODE ChIP-seq data from chr1-13 of GM12878 and a deltaSVM model was trained on ENCODE ChIP-seq peaks and same number of random regions of similar length on chr1-13 of GM12878. The built-in JASPAR and TRANSFAC motif dataset was used for sTRAP, which includes the motif for all the factors but JUND (Section 2.7).

We show the averaged ROC curves and precision recall curves (PRC) (Supplementary Fig. S4 for the first control set, Fig. 3 for the second control set) of all the methods for different transcription factors and negative sets. We evaluated two aspects of the curves. The first metric is the area under curve (AUC) (Supplementary Table S4), which summarizes the overall performance in prioritizing the positive set over negative set. The second metric is the true-positive rate at low false-positive rate (for ROC) or the recall at high precision (for PRC), which reflects the practical need for low false discovery rate in post-GWAS analysis where thousands of lead and linked SNPs are tested for regulatory consequence.

The ROC curves for GERV consistently dominated the competing methods for all factors and control scenarios, with much better AUC and higher true-positive rate at low false-positive rates. In PRCs, because of the small size of the positive set, the confidence intervals of precision when the recall is low tend to be large, making the left-most part of the curves less informative for comparison. For transcription factor FOS, MAX and MYC, GERV achieved a PRC clearly superior to the others, without overlapping in the confidence interval. For factor JUND, NF-κB and CTCF, GERV had a similarly precision for low recall but outperformed the other methods with consistently high precision for larger recall. Given the fact that CTCF has a motif (19 bp) more than twice as long as the maximum length of k-mer (8 bp) learnable for GERV (Section 2.2), the competitive performance on CTCF demonstrates the strong descriptive power of GERV in modeling TF binding. We can also see that even without DNase-seq covariates, the GERV model still achieved a performance superior to the competing methods, demonstrating the power of the model in capturing sequence determinants of the TF binding. We also found that in our second control scenario, choosing 50 instead of 10 negative SNPs for each positive SNP did not change the relative performance of the methods compared.

To mimic the original ASB analysis, we constructed an additional type of negative set by sampling 10 negative samples for each positive sample from heterozygous SNPs in GM12878 with the distribution of SNP's distance to the closest ChIP-seq peak matched to that of the positive sets. This is a more difficult and partially

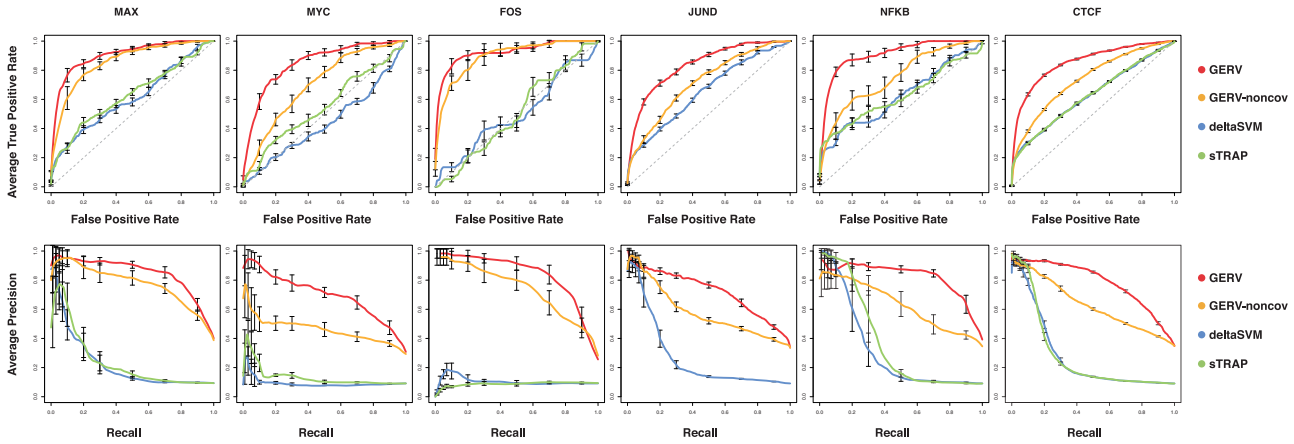

Fig. 3. ROC curve (first row) and PRC (second row) for discriminating ASB SNPs from the second type of negative variant set (10 times of the size of positive set) using GERV (red), GERV without covariates (yellow), deltaSVM (blue) and sTRAP (green). Gray-dashed line in ROC curves indicates random chance. In each figure, 95% confidence intervals of the true-positive rate (for ROC) or precision (for PRC) are plotted. The performance of sTRAP on JUND is not measurable as JUND motif is not included in its built-in motif database

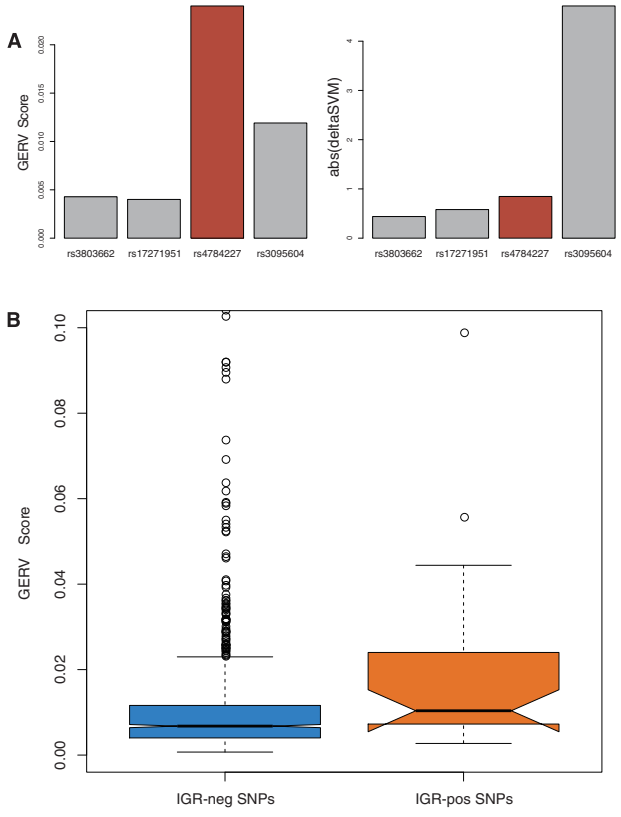

Fig. 4. (**A**) GERV correctly predicted the effect of validated causal SNP rs4784227 on FOXA1 binding, while deltaSVM failed. (**B**) The 29 variants previously reported to modulate FOXA1 binding had significantly higher (Mann–Whitney *U* test $P = 0.0011$) GERV scores than the rest of the AVS

confounded task with potentially a much higher ratio of false-negative ASB SNPs included in the negative set. GERV outperformed the other methods using ROC analysis for four out of six factors in this task, with precision-recall analysis showing improved performance for one of six factors (Supplementary Fig. S6 and Table S4). The presence of false-negative ASB SNPs could explain the precision-recall performance and the close-to-random performance on FOS for all methods.

4.4 GERV prioritizes linked-SNPs that modulate FOXA1 binding in breast cancer

To demonstrate the application of GERV in post-GWAS analysis, we applied GERV to a breast-cancer-associated variant set (AVS) collected by a previous study (Cowper-Sal lari *et al.*, 2012). It is composed of 44 risk-associated SNPs discovered from GWAS studies and 1053 'linked' SNPs that were not discovered in GWAS but are in strong linkage disequilibrium ($r^2 > 0.8$) with any risk-associated SNP. It has been shown that breast-cancer-associated SNPs are enriched for the binding sites of FOXA1, a pioneer transcription factor essential for chromatin opening and nucleosome positioning favorable to transcription factor recruitment (Carroll *et al.*, 2005, 2006; Eeckhoute *et al.*, 2006; He *et al.*, 2010; Lupien *et al.*, 2008).

The rs4784227 breast-cancer-associated SNP has been shown to disrupt the binding of FOXA1 with several lines of evidence (Cowper-Sal lari *et al.*, 2012; Long *et al.*, 2010). We trained a GERV model and a deltaSVM model on ENCODE FOXA1 ChIP-seq data from a breast cancer cell line T47D. Using these two models, we scored rs4784227 and all of its linked SNPs collected in the AVS (rs3803662, rs17271951 and rs3095604). GERV correctly predicted the effect of rs4784227 on FOXA1 binding, while deltaSVM failed (Fig. 4A).

Having probed a single risk-associated SNP, we then applied GERV to all the SNPs in the breast cancer AVS. The 29 variants previously reported to modulate FOXA1 binding (Cowper-Sal lari *et al.*, 2012) had significantly higher GERV scores than the rest of the AVS (Fig. 4B, Mann–Whitney *U* test $P = 0.0011$, AUC = 0.68, Supplementary Fig. S7). In contrast, deltaSVM could not distinguish the positive set from the rest of the AVS (Mann–Whitney *U* test $P = 0.19$, AUC = 0.57, Supplementary Fig. S7)

5 Discussion

Despite the recent substantial advances in characterizing the genome-wide transcription factor binding sites with ChIP-seq experiments, it remains a challenge to interpret variation in the noncoding region of the genome and to determine variants that cause transcription factor binding changes in post-GWAS analysis. Our work improves the prediction of causal non-coding variants when compared with other contemporary methods.

As the first generative model that directly predicts the ChIP-seq signal, GERV achieved greater accuracy than other methods in predicting ChIP-seq peaks. GERV models the spatial effect of all the k-mers and thus captures the effect of the primary motif and auxiliary sequences on TF binding. We have shown that many of these auxiliary sequences correspond to known binding cofactors, while others were matched to transcription factors whose roles in the binding regulation have not been previously characterized. Since GERV is trained on cell-type-specific ChIP-seq and DNase-seq data each GERV model is cell-type specific. The effect size of kmers across cell types is generally stable, with differences that reflect cell-type-specific effects (Supplementary Table S5).

The generative nature of the GERV model scores each variant as the predicted change to a proximal ChIP-seq signal. The analysis on six transcription factors NF-κB, CTCF, FOS, JUND, MAX and MYC demonstrated that GERV outperforms existing methods in discriminating variants known to alter TF binding from negative control sets. In a few cases (Fig. 3F, Supplementary Fig. S4F), the discriminative nature of the competing methods equipped them with higher precision for recalling a small fraction of positives. However, their inability to model auxiliary sequences led to the dramatic precision decrease afterward, while GERV achieved constantly high precision for larger recall.

Applied to an AVS of breast cancer, GERV correctly predicted the effect of previous validated causal SNP rs4784227 and highly prioritized variants reported to affect FOXA1 binding in breast cancer cell line. With the superior performance exemplified in this task, we expect GERV to play an important role in functionally annotating and prioritizing putative causal variants for downstream experimental analysis.

Acknowledgements

We thank Yuchun Guo for technical support in co-factor analysis using GEM. We also thank Matthew Edwards for many helpful comments and discussions.

Funding

This work was supported by the National Institutes of Health [1U01HG007037 to D.K.G.]

Conflict of Interest: none declared.

References

Andersen,M.C. *et al*. (2008) In silico detection of sequence variations modifying transcriptional regulation. *PLoS Comput. Biol.*, **4**, e5.

Bartels,M. *et al*. (2007) Peptide-mediated disruption of NFkappaB/NRF interaction inhibits IL-8 gene activation by IL-1 or *Helicobacter pylori*. *J. Immunol.*, **179**, 7605–7613.

Carroll,J.S. *et al*. (2005) Chromosome-wide mapping of estrogen receptor binding reveals long-range regulation requiring the forkhead protein FoxA1. *Cell*, **122**, 33–43.

Carroll,J.S. *et al*. (2006) Genome-wide analysis of estrogen receptor binding sites. *Nat. Genet.*, **38**, 1289–1297.

Cowper-Sal Lari,R. *et al*. (2012) Breast cancer risk-associated SNPs modulate the affinity of chromatin for FOXA1 and alter gene expression. *Nat. Genet.*, **44**, 1191–1198.

Eeckhoute,J. *et al*. (2006) A cell-type-specific transcriptional network required for estrogen regulation of cyclin D1 and cell cycle progression in breast cancer. *Genes Dev.*, **20**, 2513–2526.

Frazer,K.A. *et al*. (2009) Human genetic variation and its contribution to complex traits. *Nat. Rev. Genet.*, **10**, 241–251.

Fujioka,S. *et al*. (2004) NF-κB and AP-1 connection: mechanism of NF- κ B-dependent regulation of AP-1 activity. *Society*, **24**, 7806–7819.

Ghandi,M. *et al*. (2014) Enhanced regulatory sequence prediction using gapped k-mer features. *PLoS Comput. Biol.*, **10**, e1003711.

Guo,Y. *et al*. (2012) High resolution genome wide binding event finding and motif discovery reveals transcription factor spatial binding constraints. *PLoS Comput. Biol.*, **8**, e1002638.

He,H.H. *et al*. (2010) Nucleosome dynamics define transcriptional enhancers. *Nat. Genet.*, **42**, 343–347.

Hindorff,L.A. *et al*. (2009) Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl. Acad. Sci. USA*, **106**, 9362–9367.

Karczewski,K.J. *et al*. (2013) Systematic functional regulatory assessment of disease-associated variants. *Proc. Natl. Acad. Sci. USA*, **110**, 9607–9612.

Lee,D. *et al*. (2015) A method to predict the impact of regulatory variants from DNA sequence. *Nat. Genet.*, **47**, 955–961.

Long,J. *et al*. (2010) Identification of a functional genetic variant at 16q12.1 for breast cancer risk: results from the Asia Breast Cancer Consortium. *PLoS Genet.*, **6**, e1001002.

Lupien,M. *et al*. (2008) FoxA1 translates epigenetic signatures into enhancer-driven lineage-specific transcription. *Cell*, **132**, 958–970.

Macintyre,G. *et al*. (2010) is-rSNP: a novel technique for in silico regulatory SNP detection. *Bioinformatics*, **26**, i524–i530.

Mahony,S. and Benos,P.V. (2007) STAMP: a web tool for exploring DNA-binding motif similarities. *Nucleic Acids Res.*, **35**, 253–258.

Manke,T. *et al*. (2010) Quantifying the effect of sequence variation on regulatory interactions. *Hum. Mutat.*, **31**, 477–483.

Manolio,T.A. (2010) Genomewide association studies and assessment of the risk of disease. *N. Engl. J. Med.*, **363**, 166–176.

McCarthy,M.I. *et al*. (2008) Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nat. Rev. Genet.*, **9**, 356–369.

Molineris,I. *et al*. (2013) Identification of functional cis-regulatory polymorphisms in the human genome. *Hum. Mut.*, **34**, 735–742.

Purcell,S. *et al*. (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.*, **81**, 559–575.

Riva,A. (2012) Large-scale computational identification of regulatory SNPs with rSNP-MAPPER. *BMC Genomics*, **13**(Suppl 4), S7.

Rozowsky,J. *et al*. (2011) AlleleSeq: analysis of allele-specific expression and binding in a network framework. *Mol. Syst. Biol.*, **7**, 522.

Sgarbanti,M. *et al*. (2008) IRF-1 is required for full NF-kappaB transcriptional activity at the human immunodeficiency virus type 1 long terminal repeat enhancer. *J. Virol.*, **82**, 3632–3641.

Stranger,B.E. *et al*. (2011) Progress and promise of genome-wide association studies for human complex trait genetics. *Genetics*, **187**, 367–383.

Teng,M. *et al*. (2012) Regsnps: a strategy for prioritizing regulatory single nucleotide substitutions. *Bioinformatics*, **28**, 1879–1886.

Thomas,R.S. *et al*. (1997) ETS1, NFkappaB and AP1 synergistically transactivate the human GM-CSF promoter. *Oncogene*, **14**, 2845–2855.

Ward,L.D. and Kellis,M. (2012a) HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. *Nucleic Acids Res.*, **40**(Database issue), D930–D934.

Ward,L.D. and Kellis,M. (2012b) Interpreting noncoding genetic variation in complex traits and human disease. *Nat. Biotechnol.*, **30**, 1095–1106.

Zuo,C. *et al*. (2015) atSNP: transcription factor binding affinity testing for regulatory SNP detection. *Bioinformatics*, **31**, 3353–3355.