Bioinformatics
doi.10.1093/bioinformatics/bty1010
Advance Access Publication Date: 8 December 2018
Applications Note

OXFORD

Genome Analysis

Supplement Material for "atSNP Search: a web resource for statistically evaluating influence of human genetic variation on transcription factor binding"

Sunyoung Shin, Rebecca Hudson, Christopher Harrison, Mark Craven, and Sündüz Keleş

1 Comparison of the atSNP Search with existing resources

Several motif-based web resources are currently available to quantify the regulatory impacts of human SNPs, among which the atSNP Search is one of the most comprehensive and up-to-date tools (Supplementary Table 1). In generation of the atSNP Search contents, we matched all 132,946,852 SNPs in dbSNP 144 on hg 38 to 2,270 motifs in total. The initial SNP set of SNP2TFBS is a SNP catalogue from 1000 Genomes project, which contains approximately 64% of the initial SNP set of the atSNP Search. The final database of SNP2TFBS itself contains a much smaller subset of variant-motif pairs that survive at p-value cutoff of 3×10^{-6}. While the atSNP Search, SNP2TFBS, and Raven harbor pre-computed results on web servers and return immediate search results, OncoCis implements motif searches on the fly using the Possum tool (Haverty *et al*., 2004).

Table 1. Comparison of motif-based regulatory SNP discovery tools

Tools	JASPAR	ENCODE	hg version	# initial SNPs	Pre-computed data
atSNP Search	✓	✓	hg38	133M	✓
SNP2TFBS	✓		hg19	85M	✓
Raven	✓		hg17	30K	✓
OncoCis	✓		hg19	NA	

Tools	Statistical significance	User-defined thresholds	Genome-wide search given a motif	Graphics
atSNP Search	✓	✓	✓	✓
SNP2TFBS	✓			✓
Raven				✓
OncoCis				

	Annotation
atSNP Search	UCSC Genome Browser hyperlink
SNP2TFBS	RefSeq gene
Raven	phastCons score
OncoCis	Gene expression, phastCons score, Histon ChIP/DNase-seq peak UCSC Genome Browser/DGIdb hyperlink FANTOM5 enhancer/promoter TSS prediction

SNP2TFBS (Kumar *et al*., 2016), Raven (Andersen *et al*., 2008), OncoCis (Perera *et al*., 2014)

2 Additional database contents

The atSNP (Zuo *et al*., 2015) testing framework estimates a background distribution to use as the null distribution for evaluating motif matches. We paid special attention to the GC content in these evaluations because GC content has been found to diversify mutation rates, as evidenced by their explanatory power in human genome variability (Hellmann *et al*., 2005). Specifically, we computed the GC content for all the 201-base-long windows centered at the SNP positions on reference alleles and classified each variant location into one of the two GC classes depicted with a mixture of two normal distributions (Supplementary Figure 1). Then the first order Markov models were fitted separately to the two sub-populations in order to impose adjacent base dependencies. Next, for every SNP-motif pair, we identified the best motif matches in the 61-base DNA sequence, centered at the variant location with both the reference and SNP alleles and quantified both the significance of the motif matches and the change in the motif matches using a likelihood-based approach.

atSNP Search utilizes the p-value of the log rank statistic evaluated at the best motif matches with both reference and SNP alleles, which is named *p-value SNP impact*, as the key sequence-based measure of detecting regulatory variants. Statistical evidence for the regulatory roles of variants is further assessed with three additional statistical hypothesis

tests on the match alteration: (1) log likelihood ratio evaluated at the best matches of both alleles, (2) log likelihood ratio evaluated at the matches of both alleles at the best match position of the reference allele, (3) log likelihood ratio evaluated at the matches of both alleles at the best match position of the SNP allele. The p-values from these calculations are named *p-value Difference*, *p-value Condition Ref*, and *p-value Condition SNP*, respectively. The scores and p-values are reported in the detail page. Users can quickly retrieve each detail page using the intuitive URL, which is a combination of the motif ID, RSID, and variant nucleotide, e.g., http://atsnp.biostat.wisc.edu/detail/motifID_RSID_ N. This feature enables programmatic access to atSNP Search results. For studying human variants other than SNPs or non-human genetic variations, we suggest the `R` package atSNP, which is publicly available at https://github.com/keleslab/atSNP.

3 Heterogeneity of GC content around genomic locations of variants

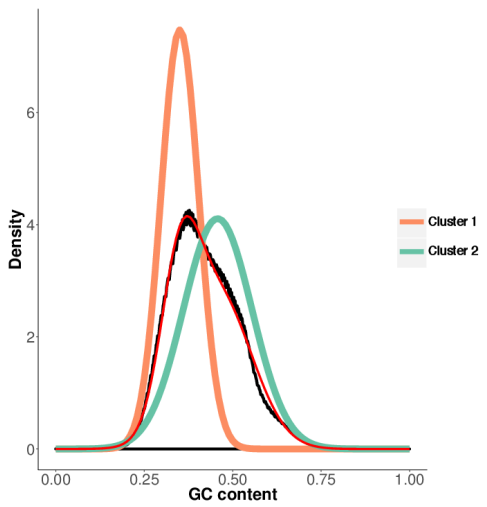

Fig. 1. Grouping of SNPs based on their local neighborhood GC content with a mixture modeling framework. The black curve denotes the observed GC content, whereas the red curve is the fitted probability density function of the mixture of two normal distributions. The curves labeled as Cluster 1 and 2 denote the two identified components.

Table 2. Estimated stationary distributions and transition matrices of the two SNP groups based on their GC contents.

Cluster 1					Cluster 2				
	A	C	G	T		A	C	G	T
	0.34	0.16	0.16	0.34		0.26	0.24	0.24	0.26
	A	C	G	T		A	C	G	T
A	0.37	0.15	0.18	0.30		0.28	0.20	0.30	0.22
C	0.41	0.19	0.03	0.37		0.32	0.29	0.08	0.31
G	0.34	0.17	0.19	0.30		0.26	0.24	0.29	0.21
T	0.27	0.16	0.20	0.37		0.17	0.24	0.30	0.28

Supplementary Figure 1 displays the distribution of GC content in the local neighbourhood of SNPs, i.e., a 201-base-long windows centered at the SNPs, and the two mixture components that are identified. The two groups have significantly different transition patterns, and in the stationary state, the second cluster has higher GC content than the first cluster (Supplementary Table 2).

4 The atSNP Search infrastructure

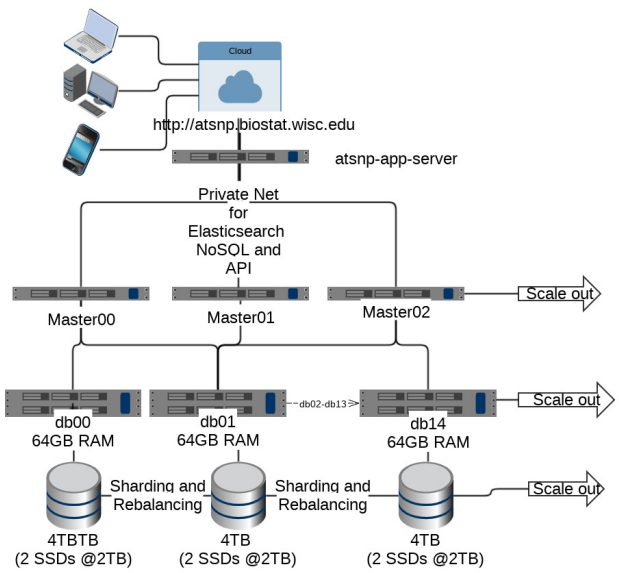

Fig. 2. The atSNP Search design.

atSNP Search is written with Django, a high-level Python Web framework that encourages rapid development and pragmatic design (Forcier *et al*., 2008). atSNP Search contents were first generated in RData format using UW Madison HTCondor, an open-source high-throughput computing software framework for coarse-grained distributed parallelization of computationally intensive tasks (Thain *et al*., 2005). "Years of compute hours" on the entire task are roughly 13 years (113,500 HTC hours) on a single core CPU machine with at least 7GB of disk space and 10GB of memory. Records on variant-motif pairs with marginal significance in motif matches or alteration were provided as input to the atSNP Search server in JSON format. Custom Python scripts for ETL (Extract, Transform and Load) were utilized for data loading (Harrison *et al*., 2018). Elasticsearch, a NoSQL database, runs on the atSNP Search server, utilizing a distributed scale-out system architecture for large workloads (Gormley and Tong, 2015). It accomplishes the task of search and retrieval by distributing requests for searches among the scaled computing resources. As requirements for storage and performance increase with user demand, we can scale out by adding more machines. A restAPI (Masse, 2011) handles communication between the search page and the Elasticsearch data store. The complete atSNP Search infrastructure is illustrated in Supplementary Figure 2. Composite sequence logos are generated on the fly using D3.js, which is a JavaScript library for dynamic and interactive data visualizations on web (Bostock *et al*, 2011).

5 atSNP query response time

Table 3. Response time for SNPid List and SNPid Window Searches (in seconds).

SNPid List			SNPid Window		
# of	p-value		Window	p-value	
SNP IDs	0.05	0.01	size	0.05	0.01
15	13.7-19.6	14.7-32.1	100	2.7-9.6	2.7-3.3
50	9.2-40.1	3.5-27.0	1K	2.7-9.6	2.7-3.3
100	29.7-32.5	13.6-14.4	10K	40.5-53.4	12.3-24.1
500	API timeout	API timeout	100K	API timeout	API timeout

Table 4. Response time for Genomic Location and Gene Searches (in seconds).

Genomic Location			Gene		
Loscation	p-value		Window	p-value	
size	0.05	0.01	size	0.05	0.01
1K	3.9-12.5	3.1-5.2	100	4.9-11.5	2.2-2.7
10K	22.8-51.1	7.9-32.3	1K	9.5-29.3	2.6-7.3
50K	39.4-55.9	36.1-37.4	5K	46.7-61.2	24.5-33.2
100K	API timeout	API timeout	10K	API timeout	API timeout

Table 5. Response time for Transcription Factor Search (in seconds).

Library	Transcription	p-value	
	Factor	0.05	0.0005
JASPAR	ZNF263	2.3-4.7	2.5-2.6
	CTCF	3.8-5.1	2.6-3.9
ENCODE	AFP	2.7-3.3	1.7-5.4
	GATA	40.3-47.7	27.9-43.6

Supplementary Tables 3-5 report response time for the five search types under various combinations of query parameters and significance levels. We performed two experimental runs under one combination of each search type at a time and recorded both response times. Our empirical studies suggest that, overall, both query type and size of query results determine the response time albeit some exceptions exist. The four types in Supplementary Tables 3-4 search through all variant-pairs which meet the user-defined criteria within a collection of SNPs or a specified genomic range. Transcription factor search, which needs no access to genomic coordinates, returns thousands to hundreds of millions of variant-motif pairs within a minute.

6 Enrichment analysis of acute myeloid leukemia SNPs

UK Biobank genotyped 820,967 SNPs using the Affymetrix Axiom arrays, a subset of which are annotated with disease genes in Online Mendelian Inheritance in Man (OMIM) database (Hamosh *et al*., 2005). We used atSNP Search to assess whether the 1,475 acute myeloid leukemia (AML) SNPs in the UK Biobank are enriched for impact on a set of transcription factors. To assess enrichment, we utilized the 21,529 non-AML cancer SNPs in the biobank as the background set of SNPs. Binding enhancement or disruption of a transcription factor by a SNP are assumed to occur when

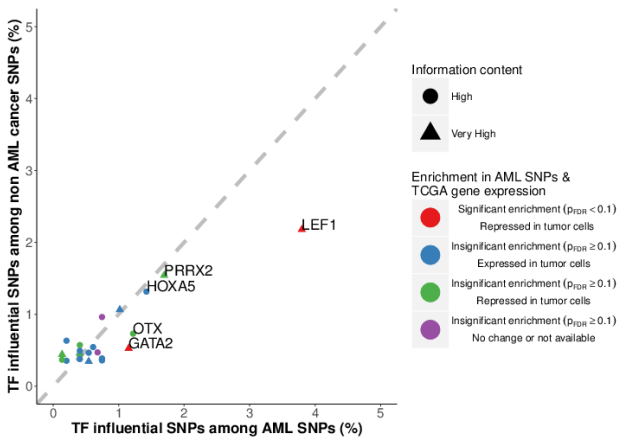

Fig. 3. Proportions of SNPs impacting binding of the 23 transcription factors among non-AML and AML SNP sets. For each transcription factor, the proportion of SNPs with significant *P-value SNP Impact* (Bonferroni correction at level 0.05) for the AML SNPs was compared to that for the background set of SNPs from non-AML cancers.

the SNP significantly impacts matches of at least one motif corresponding to the transcription factor at the significance level of 0.05 after Bonferroni multiple testing correction. Using atSNP Search queries to conduct this analysis results in 13,578 non-AML cancer SNPs and 906 AML SNPs as impacting at least one of the 102 transcription factors that have motifs with high information content (median IC ≥ 1.1). For each transcription factor, we evaluated whether the proportion of SNPs with significant impact differed between the two SNP sets after constructing a contingency table. Supplementary Figure 3 summarizes the results on 23 transcription factors, the contingency tables of which have all expected cell frequencies larger than or equal to 5. We found the proportion of SNPs impacting binding of LEF1 and GATA2 significantly differ between the two groups at a false discovery rate of 0.1.

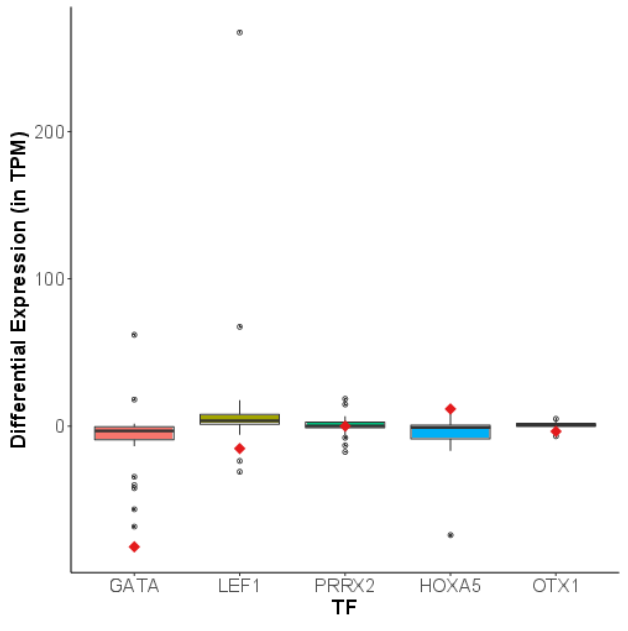

Fig. 4. Differential expression level distribution of the 31 TCGA cancer types for the five transcription factors from GEPIA. Red rhombi indicate differential expression levels in AML tumor samples.

We next asked whether expression of these transcription factors across cancer types are supportive of this finding. Specifically, we computed differences in their median expression levels in TPM (Transcripts Per Million) between AML tumor samples and matched normal tissues with the GEPIA web server (Tang *et al*., 2017). GATA2 is repressed in AML compared to matched controls by 82.04 TPM, and LEF1 is repressed by 15.23 TPM (Supplementary Figure 4). Both differential repression levels in AML are identified as outliers with respect to their distributions in all 31 TCGA cancer types, thus both transcription factors are considered having AML-specific expressional differences. Furthermore, recent research on GATA2 (Hsu *et al*., 2013; Johnson *et al*., 2012) showed that mutations of a GATA2 intronic binding site cause a primary immunodeficiency (MonoMAC) associated with myelodysplastic syndrome that progresses to AML. PRRX2 and HOXA5, which are affected by a larger proportion of AML SNPs compared to GATA2, exhibit less specificity to AML compared to the rest of cancer types. OTX1 is more repressed in AML tumor; however, overall differential OTX1 expression levels are marginal, thus its AML-specificity may not be appreciable.

References

Andersen, M. C., Engström, P. G., Lithwick, S., Arenillas, D., Eriksson, P., Lenhard, B., Wasserman, W. W., & Odeberg, J. (2008). In silico detection of sequence variations modifying transcriptional regulation. *PLoS computational biology*, **4(1)**, e5.

Bostock, M., Ogievetsky, V., & Heer, J. (2011). D3 data-driven documents. *IEEE transactions on visualization and computer graphics*, **17(12)**, 2301-2309.

Forcier, J., Bissex, P., & Chun, W. J. (2008). Python web development with Django. *Addison-Wesley Professional*.

Gormley, C., & Tong, Z. (2015). Elasticsearch: The Definitive Guide: A Distributed Real-Time Search and Analytics Engine. *O'Reilly Media, Inc*.

Hamosh, A., Scott, A. F., Amberger, J. S., Bocchini, C. A., & McKusick, V. A. (2005). Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic acids research*, **33** (suppl_1), D514-D517.

Harrison, C., Keleş, S., Hudson, R., Shin, S., & Dutra, I. (2018). atSNPInfrastructure, a case study for searching billions of records while providing significant cost savings over cloud providers. In *Proceedings of the 32nd IEEE International Parallel and Distributed Processing Symposium Workshops*, pp. 497-506.

Haverty, P. M., Hansen, U., & Weng, Z. (2004) Computational inference of transcriptional regulatory networks from expression profiling and transcription factor binding site identification. *Nucleic acids research*, **32**, 179-188.

Hellmann, I., Prüfer, K., Ji, H., Zody, M. C., Pääbo, S., & Ptak, S. E. (2005). Why do human diversity levels vary at a megabase scale?. *Genome research*, **15** (9), 1222-1231.

Hsu, A.P., Johnson, K.D., Falcone, E.L., Sanalkumar, R., Sanchez, L., Hickstein, D.D., Cuellar-Rodriguez, J., Lemieux, J.E., Zerbe, C.S., Bresnick, E.H., & Holland, S.M. (2013). GATA2 haploinsufficiency caused by mutations in a conserved intronic element leads to MonoMAC syndrome. *Blood*, **121** (19), 3830-3837, S1-S7.

Johnson, K.D., Hsu, A.P., Ryu, M.J., Wang, J., Gao, X., Boyer, M.E., Liu, Y., Lee, Y., Calvo, K.R., Keles, S., Zhang, J., Holland, S. M., & Bresnick E. H. (2012). Cis-element mutated in GATA2- dependent immunodeficiency governs hematopoiesis and vascular integrity. *The Journal of clinical investigation*, **122** (10), 3692-3704.

Kumar, S., Ambrosini, G., & Bucher, P. (2016). SNP2TFBS-a database of regulatory SNPs affecting predicted transcription factor binding site affinity. *Nucleic acids research*, **45(D1)**, D139-D144.

Masse, M. (2011). REST API Design Rulebook: Designing Consistent RESTful Web Service Interfaces. *O'Reilly Media, Inc*.

Perera, D., Chacon, D., Thoms, J. A., Poulos, R. C., Shlien, A., Beck, D., Campbell, P. J., Pimanda, J. E. & Wong, J. W. (2014). OncoCis: annotation of cis-regulatory mutations in cancer. *Genome biology*, **15(10)**, 485.

Tang, Z., Li, C., Kang, B., Gao, G., Li, C., & Zhang, Z. (2017). GEPIA: a web server for cancer and normal gene expression profiling and interactive analyses. *Nucleic acids research*, **45** (W1), W98-W102.

Thain, D., Tannenbaum, T., & Livny, M. (2005). Distributed computing in practice: the Condor experience. *Concurrency and computation: practice and experience*, **17(2-4)**, 323-356.

Zuo, C., Shin, S., & Keleş, S. (2015). atSNP: transcription factor binding affinity testing for regulatory SNP detection. *Bioinformatics*, **31(20)**, 3353-3355.