D930–D934 *Nucleic Acids Research, 2012, Vol. 40, Database issue* *Published online 7 November 2011*
doi:10.1093/nar/gkr917

HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants

Lucas D. Ward[1,2,*] and Manolis Kellis[1,2,*]

[1]Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology and [2]The Broad Institute of MIT and Harvard, Cambridge, MA 02139, USA

Received August 15, 2011; Revised October 6, 2011; Accepted October 8, 2011

ABSTRACT

The resolution of genome-wide association studies (GWAS) is limited by the linkage disequilibrium (LD) structure of the population being studied. Selecting the most likely causal variants within an LD block is relatively straightforward within coding sequence, but is more difficult when all variants are intergenic. Predicting functional non-coding sequence has been recently facilitated by the availability of conservation and epigenomic information. We present HaploReg, a tool for exploring annotations of the non-coding genome among the results of published GWAS or novel sets of variants. Using LD information from the 1000 Genomes Project, linked SNPs and small indels can be visualized along with their predicted chromatin state in nine cell types, conservation across mammals and their effect on regulatory motifs. Sets of SNPs, such as those resulting from GWAS, are analyzed for an enrichment of cell type-specific enhancers. HaploReg will be useful to researchers developing mechanistic hypotheses of the impact of non-coding variants on clinical phenotypes and normal variation. The HaploReg database is available at http://compbio.mit.edu/HaploReg.

INTRODUCTION

Genome-wide association studies (GWAS) are providing a flood of data associating genetic variants with common phenotypes (1). A confounding factor in such studies is linkage disequilibrium (LD), which allows many variants at the same locus to be associated with a phenotype even if only one of them is causal. Within genes, prioritizing the likely causal variant is relatively straightforward; variants are easily annotated as synonymous, missense or nonsense, changing the consensus sequence at splice sites, or residing in introns or UTRs. Often, however, GWAS associations lie far from known genes or transcribed regions, presumably in distal tissue-specific enhancers. One of the most striking examples of such a finding is the gene desert at 8q24, within which are regions specifically and independently linked to prostate, breast, ovarian, colorectal and bladder cancer. These variants have been shown to correspond to cell-type-specific distal enhancers for the MYC oncogene (2,3). Recent systematic comparisons of expression quantitative trait loci (eQTL) and GWAS suggest that the association of intergenic variants with complex phenotyes is a result of alteration of gene expression regulatory elements (4,5).

Ernst and colleagues (6) recently developed a map of chromatin states, including enhancers, promoters, insulators and heterochromatin, in nine human cell lines based on a variety of histone modifications. Using this map, it was demonstrated that these states can be used to prioritize SNPs within LD blocks associated with disease, and in some cases reveal biologically plausible enrichments for cell type-specific enhancers. Here we present a tool, HaploReg, to systematically mine these chromatin state data, along with conservation data and regulatory motif alterations.

A wide range of resources exists to make predictions about the functional consequences of variants, as well as navigating groups of linked variants using LD information. Polyphen (7), SIFT (8) and SNPS3D (9) all make predictions of the impact of missense SNPs. Algorithms such as is-rSNP (10) and RAVEN (11) use regulatory motif changes to predict SNPs that may influence transcriptional regulation. SNPinfo (12) combines missense predictions with TRANSFAC PWM disruption predictions and conservation information across 17 vertebrates for HapMap Phase III SNPs. SNAP (13) provides LD calculations using 1000 Genomes Project pilot data with information about neighboring genes and array

*To whom correspondence should be addressed. Tel: +1 617 253 2419; Fax: +1 617 452 5034; Email: manoli@mit.edu
Correspondence may also be addressed to Lucas Ward. Tel: +1 617 715 4881; Fax: +1 617 452 5034; Email: lukeward@mit.edu

membership for proxy/tag SNP selection, but does not currently include indels. HaploReg improves on SNAP by providing LD calculation of 1000 Genomes Project pilot indels associated with query SNPs. In addition, the features of SNPinfo are improved upon by incorporating evolutionary constraint based on two alogrithms (involving the sequences of at least 29 mammals) and considering a much larger library of PWMs.

The UCSC Genome Browser (14) and ENSEMBL Genome Browser (15,16) both allow genomic regions to be annotated with the results of cutting-edge genomic data, including chromatin state segmentations, ENCODE data, 1000 Genomes variants, evolutionary constraint, LD calculations and NHGRI catalog variants. However, the output of these browsers can be overwhelming, especially when one is interested only in a limited subset of loci (such as the variants linked to a GWAS hit.) To this end, HaploReg combines the focus on haplotype blocks provided by tools such as SNAP and SNPinfo with the breadth of genomic annotation provided by the full-featured genome browsers.

METHODS

HaploReg consists of a PHP interface to a MySQL database. The initial database table was populated using genomic coordinates and sequences for 16 151 841 biallelic SNPs and small indels from the pilot release of the 1000 Genomes Project (17). In some cases, such as novel indels, the variant call format (VCF) file from the pilot release did not have a RefSNP identifier (rsid); for the purpose of creating a unique identifier for this database, these variants were assigned the label of 'chromosome:position' in hg18 coordinates. To provide backward compatibility with obsolete rsids, dbSNP release 132 was checked for variants at the same position as 1000 Genomes pilot variants with multiple rsids (18). In addition, annotations of functional consequences were extracted from dbSNP.

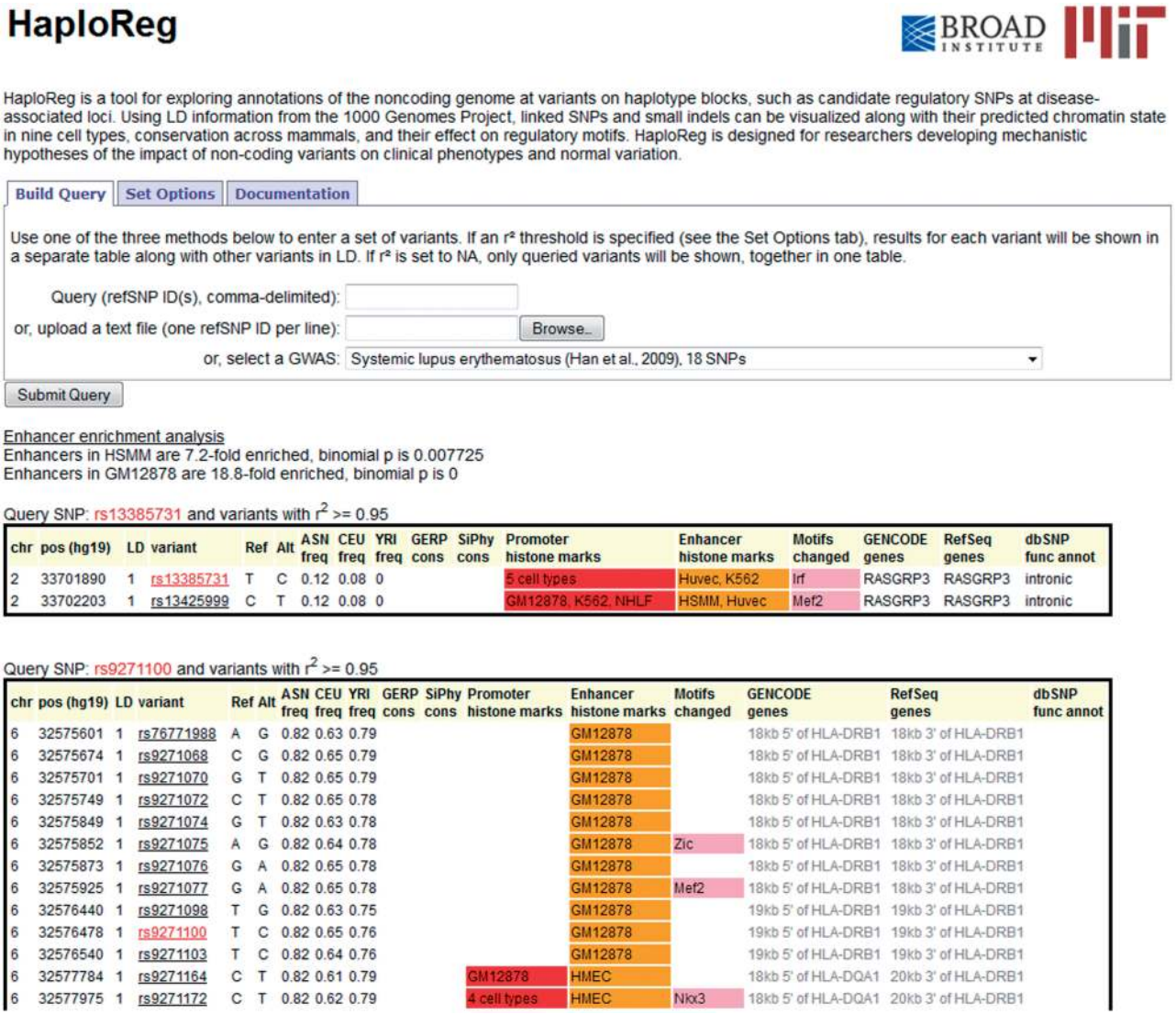

Figure 1. HaploReg view of the SNPs from the lupus GWAS by Han *et al.*

A variety of functional annotations were then intersected with the set of variants using the BEDTools package (19), including the chromatin state segmentation of Ernst *et al.* (6), and conserved regions by GERP (20) and SiPhy (21,22). To obtain gene annotations, RefSeq genes (23) were downloaded from the UCSC Genome Browser and GENCODE version 7 (24) was downloaded from the project website. BEDTools was then used to calculate the proximity of each variant to a gene by either annotation, as well as the orientation (3′ or 5′) relative to the nearest end of the gene, based on the strand of the gene.

In order to annotate variants by their effect on regulatory motifs, a library of position weight matrices (PWMs) was constructed from literature sources and was scored on genomic sequences as described previously (6). Briefly, a set of PWMs was collected from TRANSFAC (25), JASPAR (26), and protein-binding microarray (PBM) experiments (27–29). The reference and alternate alleles for each of the 1000 Genomes pilot SNPs and indels were concatenated with 29 bp of genomic context on each side, using the hg18 sequence obtained from the UCSC Genome Browser (30). PWMs were then scored for instances that passed either of two thresholds, a stringent threshold of $P < 4^{-8}$ and a less-stringent threshold of $P < 4^{-7}$ (31). Only instances where a motif in the sequence (i) passed the stringent threshold of a PWM in either the reference or the alternate genomic sequence, and (ii) overlapped the variable nucleotide(s) (thus changing the PWM score) were considered. Then, the change in log-odds (LOD) score was calculated. In cases where the weaker match was did not pass the less-stringent threshold, an approximate minimum change of LOD score was reported, corresponding to the difference between the score of the stronger match to the score required to pass the less-stringent threshold. In cases where both allelic variants surpassed the less-stringent threshold, the exact difference in score was reported.

GWAS results were obtained from the table curated by NHGRI (32) (accessed June 29, 2011.) In cases where multiple studies were annotated as pertaining to the same phenotype, unique independent SNPs were consolidated into a single list.

LD was calculated using the phased genotype information accompanying the 1000 Genomes Project pilot release (17). VCFTools (33) was used to perform the calculation, using an LD threshold of $r^2 = 0.80$, and a maximum distance between variants of 200 kb. Results from VCFTools were then consolidated such that for every variant in our database, a list of linked variants is accessible for each of the three populations, along with an r^2 value.

To perform enhancer enrichment analysis on sets of variants, tables of common array designs were obtained from the UCSC Table Browser (34) and lists were constructed of 1000 Genomes SNPs segregating in each of the three pilot populations, as well as all SNPs in the database. Then, a background frequency of coverage was calculated for variants annotated as overlapping a strong enhancer state in each cell type. When a user submits a query list of variants, the coverage of strong enhancers in each cell type is calculated. If the coverage exceeds that of the background set selected by the user, a binomial test is performed, and enrichment is reported if it passes an uncorrected significance threshold of 0.05.

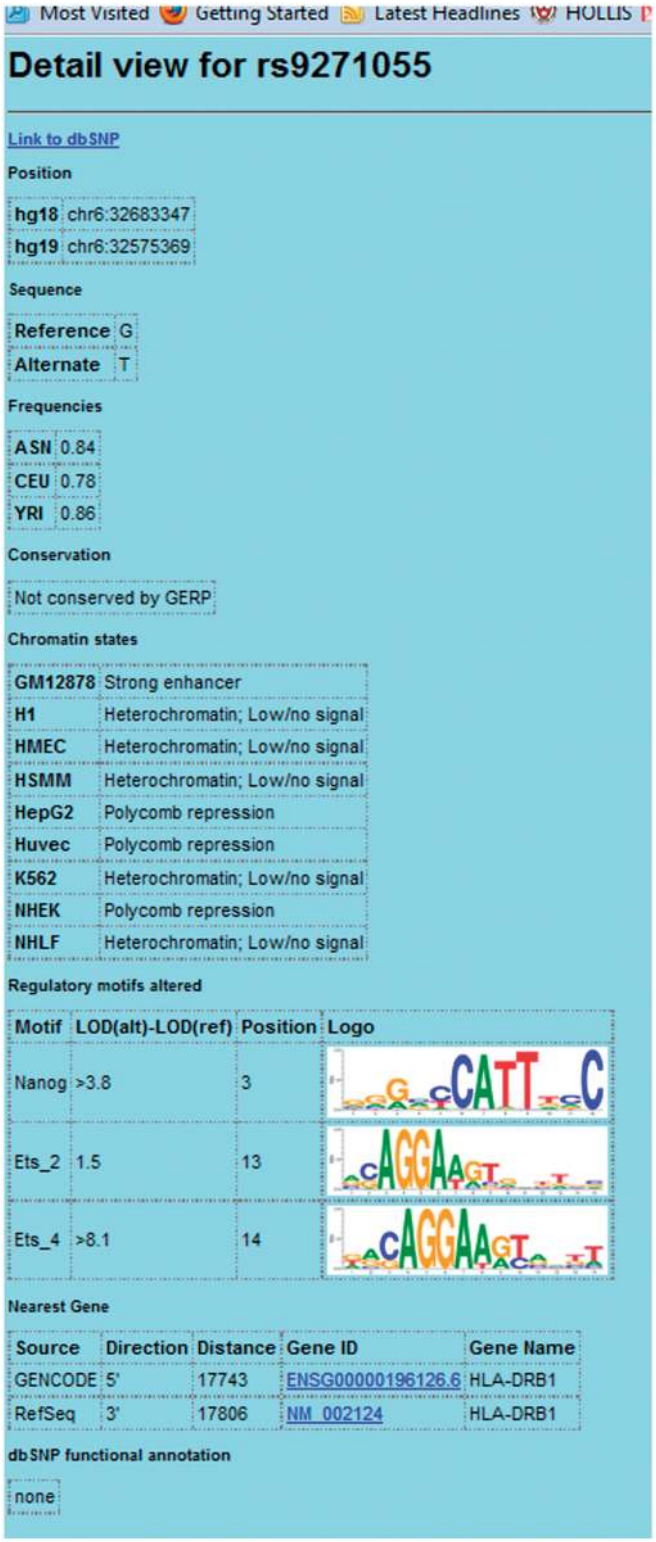

Figure 2. HaploReg detail view for the SNP rs9271055.

USAGE

A user may submit queries in two formats: a comma-delimited list of rsids, or a one of the GWAS or traits from the NHGRI catalog. To illustrate (Figure 1), we select the lupus study by Han *et al.* (35). Since the study was conducted in Han Chinese, we select ASN (CHB + JPT) as the population for LD calculation, and we select all SNPs in the ASN population as the background for enhancer enrichment analysis. As was reported by Ernst *et al.* (6), there is a strong enrichment for GM12878 (lymphoblastoid) enhancers. To demonstrate LD blocks, we select an LD threshold of $r^2 = 0.95$. In the LD block with lead SNP rs9271100, there is a SNP rs9271055 which affects an Ets-family binding site. Clicking on rs9271055 leads to a detail view (Figure 2) in which the complete chromatin state data are available. The positions in two literature motifs for Ets-family proteins can be seen, where the alternate T allele strengthens the predicted affinity relative to the reference G allele. In addition, links to NCBI RefSeq and ENSEMBL pages detailing the neighboring HLA-DRB1 gene are provided.

ACKNOWLEDGEMENTS

We thank Pouya Kheradpour for valuable assistance with PWM curation and scoring, and other members of the Kellis lab for helpful discussions.

FUNDING

National Institutes of Health (R01-HG004037, RC1-HG005334); National Science Foundation (0644282). Funding for open access charge: National Institutes of Health.

Conflict of interest statement. None declared.

REFERENCES

1. McCarthy,M.I., Abecasis,G.R., Cardon,L.R., Goldstein,D.B., Little,J., Ioannidis,J.P. and Hirschhorn,J.N. (2008) Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nat. Rev. Genet.*, **9**, 356–369.
2. Ghoussaini,M., Song,H., Koessler,T., Al Olama,A.A., Kote-Jarai,Z., Driver,K.E., Pooley,K.A., Ramus,S.J., Kjaer,S.K., Hogdall,E. *et al.* (2008) Multiple loci with different cancer specificities within the 8q24 gene desert. *J. Natl Cancer Inst.*, **100**, 962–966.
3. Wasserman,N.F., Aneas,I. and Nobrega,M.A. (2010) An 8q24 gene desert variant associated with prostate cancer risk confers differential in vivo activity to a MYC enhancer. *Genome Res.*, **20**, 1191–1197.
4. Gamazon,E.R., Huang,R.S., Cox,N.J. and Dolan,M.E. (2010) Chemotherapeutic drug susceptibility associated SNPs are enriched in expression quantitative trait loci. *Proc. Natl Acad. Sci. USA*, **107**, 9287–9292.
5. Nicolae,D.L., Gamazon,E., Zhang,W., Duan,S., Dolan,M.E. and Cox,N.J. (2010) Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. *PLoS Genet.*, **6**, e1000888.
6. Ernst,J., Kheradpour,P., Mikkelsen,T.S., Shoresh,N., Ward,L.D., Epstein,C.B., Zhang,X., Wang,L., Issner,R., Coyne,M. *et al.* (2011) Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature*, **473**, 43–49.
7. Adzhubei,I.A., Schmidt,S., Peshkin,L., Ramensky,V.E., Gerasimova,A., Bork,P., Kondrashov,A.S. and Sunyaev,S.R. (2010) A method and server for predicting damaging missense mutations. *Nat. Methods*, **7**, 248–249.
8. Ng,P.C. and Henikoff,S. (2001) Predicting deleterious amino acid substitutions. *Genome Res.*, **11**, 863–874.
9. Yue,P., Melamud,E. and Moult,J. (2006) SNPs3D: candidate gene and SNP selection for association studies. *BMC Bioinformatics*, **7**, 166.
10. Macintyre,G., Bailey,J., Haviv,I. and Kowalczyk,A. (2010) is-rSNP: a novel technique for in silico regulatory SNP detection. *Bioinformatics*, **26**, i524–i530.
11. Andersen,M.C., Engstrom,P.G., Lithwick,S., Arenillas,D., Eriksson,P., Lenhard,B., Wasserman,W.W. and Odeberg,J. (2008) In silico detection of sequence variations modifying transcriptional regulation. *PLoS Comput. Biol.*, **4**, e5.
12. Xu,Z. and Taylor,J.A. (2009) SNPinfo: integrating GWAS and candidate gene information into functional SNP selection for genetic association studies. *Nucleic Acids Res.*, **37**, W600–W605.
13. Johnson,A.D., Handsaker,R.E., Pulit,S.L., Nizzari,M.M., O'Donnell,C.J. and de Bakker,P.I. (2008) SNAP: a web-based tool for identification and annotation of proxy SNPs using HapMap. *Bioinformatics*, **24**, 2938–2939.
14. Kent,W.J., Sugnet,C.W., Furey,T.S., Roskin,K.M., Pringle,T.H., Zahler,A.M. and Haussler,D. (2002) The human genome browser at UCSC. *Genome Res.*, **12**, 996–1006.
15. Flicek,P., Amode,M.R., Barrell,D., Beal,K., Brent,S., Chen,Y., Clapham,P., Coates,G., Fairley,S., Fitzgerald,S. *et al.* (2011) Ensembl 2011. *Nucleic Acids Res.*, **39**, D800–D806.
16. Chen,Y., Cunningham,F., Rios,D., McLaren,W.M., Smith,J., Pritchard,B., Spudich,G.M., Brent,S., Kulesha,E., Marin-Garcia,P. *et al.* (2010) Ensembl variation resources. *BMC Genomics*, **11**, 293.
17. The 1000 Genomes Project Consortium. (2010) A map of human genome variation from population-scale sequencing. *Nature*, **467**, 1061–1073.
18. Sherry,S.T., Ward,M.H., Kholodov,M., Baker,J., Phan,L., Smigielski,E.M. and Sirotkin,K. (2001) dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.*, **29**, 308–311.
19. Quinlan,A.R. and Hall,I.M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, **26**, 841–842.
20. Davydov,E.V., Goode,D.L., Sirota,M., Cooper,G.M., Sidow,A. and Batzoglou,S. (2010) Identifying a high fraction of the human genome to be under selective constraint using GERP++. *PLoS Comput. Biol.*, **6**, e1001025.
21. Garber,M., Guttman,M., Clamp,M., Zody,M.C., Friedman,N. and Xie,X. (2009) Identifying novel constrained elements by exploiting biased substitution patterns. *Bioinformatics*, **25**, i54–i62.
22. Lindblad-Toh,K., Garber,M., Zuk,O., Lin,M.F., Parker,B.J., Washietl,S., Kheradpour,P., Ernst,J., Jordan,G., Mauceli,E. *et al.* (2011) A high-resolution map of human evolutionary constraint using 29 mammals. *Nature,* (epub ahead of print).
23. Pruitt,K.D., Tatusova,T. and Maglott,D.R. (2007) NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res*, **35**, D61–D65.
24. Harrow,J., Denoeud,F., Frankish,A., Reymond,A., Chen,C.K., Chrast,J., Lagarde,J., Gilbert,J.G., Storey,R., Swarbreck,D. *et al.* (2006) GENCODE: producing a reference annotation for ENCODE. *Genome Biol.*, **7(Suppl. 1)**, S4 1–9.
25. Matys,V., Fricke,E., Geffers,R., Gossling,E., Haubrock,M., Hehl,R., Hornischer,K., Karas,D., Kel,A.E., Kel-Margoulis,O.V. *et al.* (2003) TRANSFAC: transcriptional regulation, from patterns to profiles. *Nucleic Acids Res.*, **31**, 374–378.
26. Portales-Casamar,E., Thongjuea,S., Kwon,A.T., Arenillas,D., Zhao,X., Valen,E., Yusuf,D., Lenhard,B., Wasserman,W.W. and Sandelin,A. (2010) JASPAR 2010: the greatly expanded open-access database of transcription factor binding profiles. *Nucleic Acids Res.*, **38**, D105–D110.

27. Berger,M.F., Badis,G., Gehrke,A.R., Talukder,S., Philippakis,A.A., Pena-Castillo,L., Alleyne,T.M., Mnaimneh,S., Botvinnik,O.B., Chan,E.T. *et al.* (2008) Variation in homeodomain DNA binding revealed by high-resolution analysis of sequence preferences. *Cell*, **133**, 1266–1276.
28. Badis,G., Berger,M.F., Philippakis,A.A., Talukder,S., Gehrke,A.R., Jaeger,S.A., Chan,E.T., Metzler,G., Vedenko,A., Chen,X. *et al.* (2009) Diversity and complexity in DNA recognition by transcription factors. *Science*, **324**, 1720–1723.
29. Berger,M.F., Philippakis,A.A., Qureshi,A.M., He,F.S., Estep,P.W. III and Bulyk,M.L. (2006) Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. *Nat. Biotechnol.*, **24**, 1429–1435.
30. Lander,E.S., Linton,L.M., Birren,B., Nusbaum,C., Zody,M.C., Baldwin,J., Devon,K., Dewar,K., Doyle,M., FitzHugh,W. *et al.* (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.
31. Touzet,H. and Varre,J.S. (2007) Efficient and accurate P-value computation for position weight matrices. *Algorithms Mol. Biol.*, **2**, 15.
32. Hindorff,L.A., Sethupathy,P., Junkins,H.A., Ramos,E.M., Mehta,J.P., Collins,F.S. and Manolio,T.A. (2009) Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl Acad. Sci. USA*, **106**, 9362–9367.
33. Danecek,P., Auton,A., Abecasis,G., Albers,C.A., Banks,E., Depristo,M.A., Handsaker,R.E., Lunter,G., Marth,G.T., Sherry,S.T. *et al.* (2011) The variant call format and VCFtools. *Bioinformatics*, **27**, 2156–2158.
34. Karolchik,D., Hinrichs,A.S., Furey,T.S., Roskin,K.M., Sugnet,C.W., Haussler,D. and Kent,W.J. (2004) The UCSC Table Browser data retrieval tool. *Nucleic Acids Res.*, **32**, D493–D496.
35. Han,J.W., Zheng,H.F., Cui,Y., Sun,L.D., Ye,D.Q., Hu,Z., Xu,J.H., Cai,Z.M., Huang,W., Zhao,G.P. *et al.* (2009) Genome-wide association study in a Chinese Han population identifies nine new susceptibility loci for systemic lupus erythematosus. *Nat. Genet.*, **41**, 1234–1237.