BMC Bioinformatics

BioMed Central

Proceedings

SNP@Promoter: a database of human SNPs (Single Nucleotide Polymorphisms) within the putative promoter regions

Byoung-Chul Kim, Woo-Yeon Kim, Daeui Park, Won-Hyong Chung, Kwang-sik Shin and Jong Bhak*

Address: Korean BioInformation Center (KOBIC), KRIBB, Daejeon 305-806, Korea

Email: Byoung-Chul Kim - bckim@kribb.re.kr; Woo-Yeon Kim - kimplove@kribb.re.kr; Daeui Park - daeui@kribb.re.kr; Won-Hyong Chung - whchung@kribb.re.kr; Kwang-sik Shin - shinks@kribb.re.kr; Jong Bhak* - jong@kribb.re.kr

* Corresponding author

from Sixth International Conference on Bioinformatics (InCoB2007)
Hong Kong. 27–30 August 2007

Published: 13 February 2008

BMC Bioinformatics 2008, **9**(Suppl 1):S2 doi:10.1186/1471-2105-9-S1-S2

This article is available from: http://www.biomedcentral.com/1471-2105/9/S1/S2

Abstract

Background: Analysis of single nucleotide polymorphism (SNP) is becoming a key research in genomics fields. Many functional analyses of SNPs have been carried out for coding regions and splicing sites that can alter proteins and mRNA splicing. However, SNPs in non-coding regulatory regions can also influence important biological regulation. Presently, there are few databases for SNPs in non-coding regulatory regions.

Description: We identified 488,452 human SNPs in the putative promoter regions that extended from the +5000 bp to -500 bp region of the transcription start sites. Some SNPs occurring in transcription factor (TF) binding sites were also predicted (47,832 SNP; 9.8%). The result is stored in a database: SNP@promoter. Users can search the SNP@Promoter database using three entries: 1) by SNP identifier (rs number from dbSNP), 2) by gene (gene name, gene symbol, refSeq ID), and 3) by disease term. The SNP@Promoter database provides extensive genetic information and graphical views of queried terms.

Conclusion: We present the SNP@Promoter database. It was created in order to predict functional SNPs in putative promoter regions and predicted transcription factor binding sites. SNP@Promoter will help researchers to identify functional SNPs in non-coding regions.

Background

After finishing the Human Genome Project, biologists' interest has shifted to non-repetitive sequence variants in genome, by far the most common of which are single nucleotide polymorphisms (SNPs). For a variation to be considered an SNP, it must occur in at least 1% of the population. SNPs, which make up about 90% of all human genetic variation, occur every 100 to 300 bases along the 3-billion-base human genome [1,2]. It is generally believed that the complete human sequence will reveal at least a million SNPs of coding regions, including introns and promoters. As a general rule, many SNPs have no effect on cell function, but some SNPs are reported to be highly related to diseases or to influence cells' response to a drug. Although more than 99% of human DNA sequences are the same across all populations, some SNPs can have a major impact on how humans respond to diseases; environmental insults such as bacteria, viruses, toxins, and chemicals; and drugs and other therapies. This makes SNPs of great value for biomedical research and for developing pharmaceutical products and for medical diagnostics.

New bioinformatics tools and public SNP resources for SNP studies, specifically for linkage disequilibrium and disease association studies, will form part of the new scientific landscape [3-9]. These public SNP resources are possible through the large-scale and high-throughput systems to screen SNPs on many individuals. The challenge is to accomplish this while reducing the cost per genotype and required completion time. The public SNP resources are producing information about SNPs which are related to diseases or that modify biological function. Many functional studies of SNPs were focused on SNPs located in coding regions that can influence phenotype by altering the encoded proteins [9,10]. They can also influence premature termination that can cause nonsense-mediated mRNA decay (NMD) [11]. Another function of SNPs is that they affect splice sites which results in alternative splicing [12].

Additionally, there are many SNPs in non-coding regulatory regions. The exact functions of the non-coding regulatory region SNPs are not clear yet. However, some SNPs are predicted to be related to genes by influencing the binding affinity of transcription factors. For example, the G/C polymorphism in the promoter region of the FCGR2B promoter regulates gene expression [13]. -783A/G and -1438A/G polymorphisms in the promoter of HTR2A gene regulate gene expression. -783 G allele and -1438 G allele are known to reduce the binding activity of transcription factors [14]. However, there are no public resources that provide promoter information of SNPs influencing the non-coding regulatory regions in the human genome. The rSNP_Guide system is the only one that has reported SNPs that are related to potential transcription factor candidates among 41 types of known transcription factor binding sites. [15,16]. ORegAnno is focused not on SNP information of the regulatory regions in the human genome but on the registration and validation of SNPs from promoters, transcription factor binding sites, and regulatory variation [17].

SNP@Promoter is a large database that contains various types of information on the location and function for putative promoter regions in the human genome for gene regulation study. In particular, SNP@Promoter provides a platform for biologists including disease associated genes, transcription factor binding sites, and a graphic viewer.

Methods and results

We developed an integrated computational system for identifying SNPs in non-coding regulation regions (Fig 1). In this system, we: 1) predicted TF binding sites in putative promoter regions, 2) identified SNPs in the putative promoter regions and selected SNPs within predicted TF binding sites, 3) examined evolutionary conservation of predicted TF binding sites, and 4) integrated a variety of gene annotation information.

Prediction of TF binding sites in putative promoter region

We identified TF binding sites in the putative promoter regions in the human genome. The promoter region is defined as the sequence of 5 kb upstream to 500 downstream bases of a transcription start site. The annotation information of genes, which is mapped to the genome,

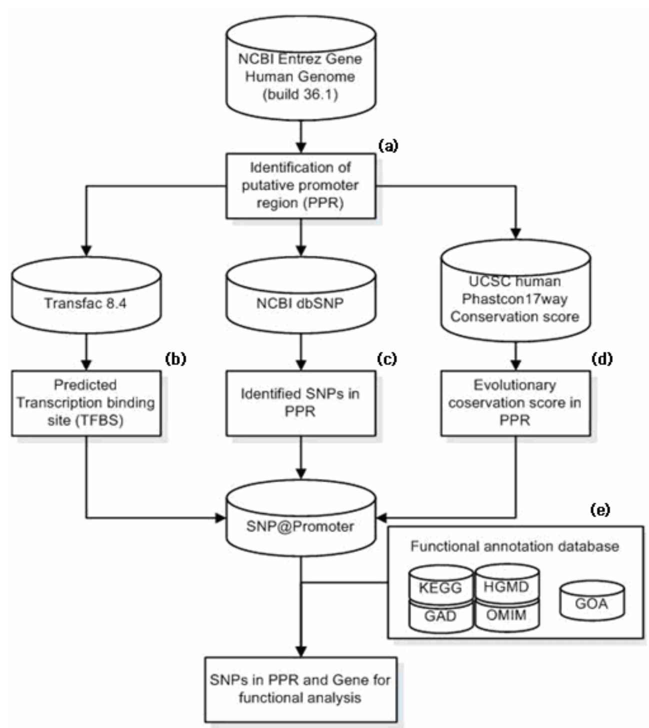

Figure 1

Flow chart for identifying SNPs in putative promoter regions. Cylinders represent databases. Rectangles are computational applications. (a) Putative promoter regions are identified in the human genome sequence. (b) Transcription binding sites are predicted in the putative promoter regions by using TransFac database. (c) SNPs are mapped. (d) Evolution conservation scores are calculated within transcription factor binding sites. (e) The disease association and functional annotation of target genes carried out by using an in-house functional annotation database.

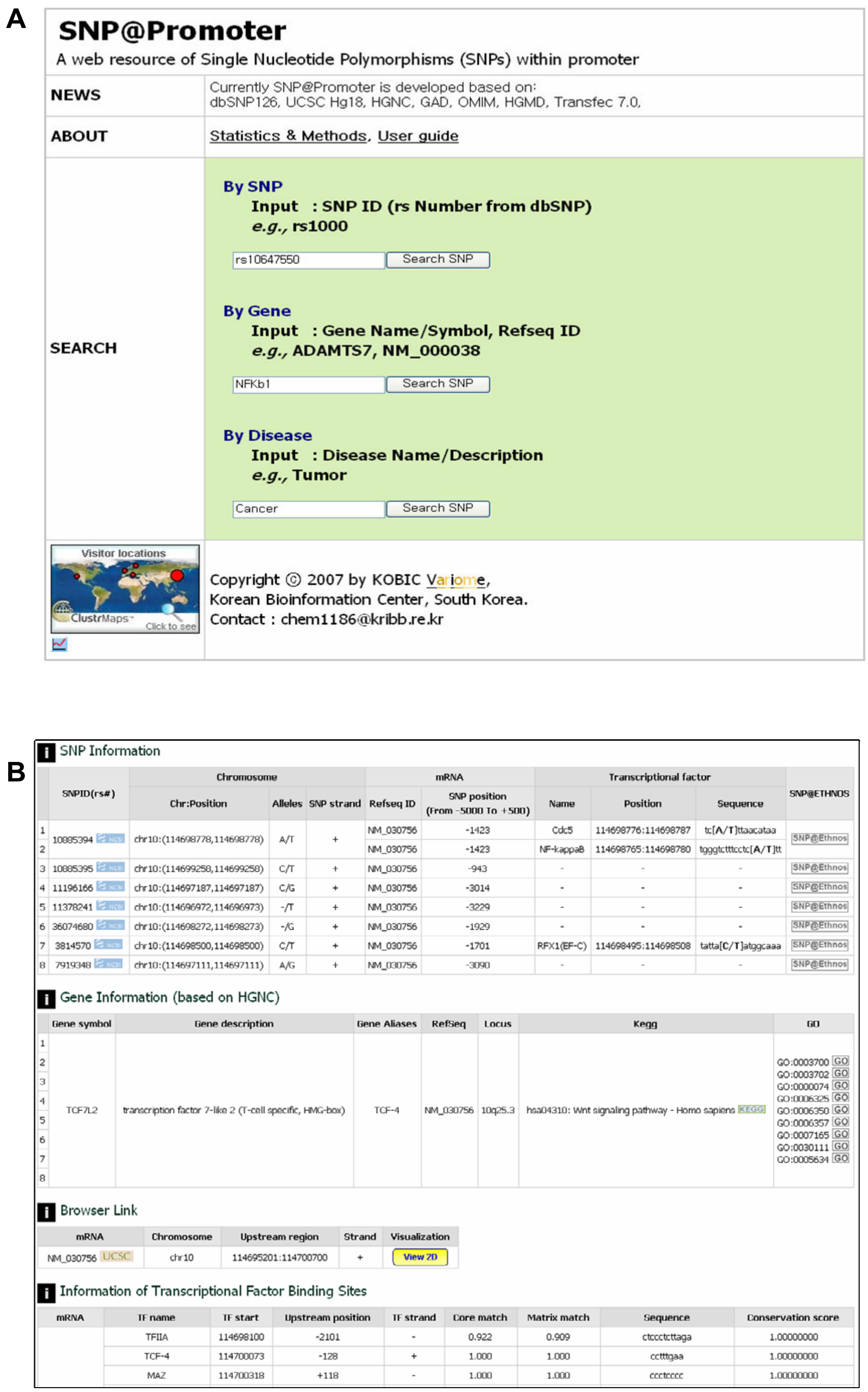

Figure 2

SNP@Promoter user interface. SNP@Promoter main page. (A) Users can search using three entries: 1) an SNP identifier (rs number from dbSNP), 2) a gene (Gene name, gene symbol, refSeq ID), or 3) a disease term. (B) SNP@Promoter gene retrieval page. The SNP Information table shows identified SNPs within putative promoter region and TF biding sites. The Gene Information table shows various gene annotations including pathways (KEGG), gene ontology (GOA). The Information of Transcription Factor Binding Sites table shows a variety off TF information such as TF start position, upstream position, TF strand, match score, TF binding sequences, conservations score.

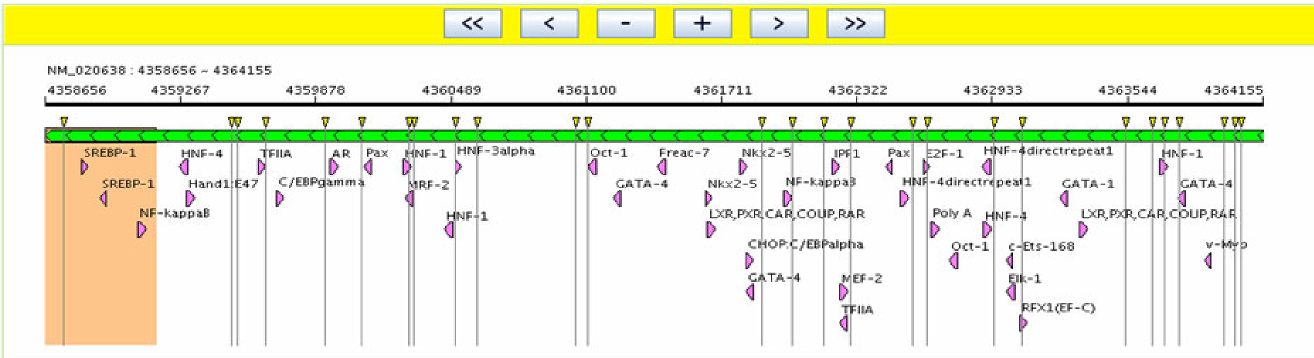

Figure 3

A graphic viewer of transcription regulatory region. The green bar represents a putative promoter region (5500 bp). The arrows in the green bar show a strand of transcription, orange box is transcription start region, yellow inverted triangles are SNP positions, and purple triangles are predicted transcription binding sites.

was obtained from the NCBI Gene database. To find TF binding sites in the putative promoter regions, we used the MATCH (Matrix Search For Transcription Factor Binding Site) program from the Transfac database (ver. 8.4) [18,19]. As a result, we predicted 1,497,317 TF binding sites from 28,644 human genes.

Identification of SNPs on predicted TF binding sites

The SNP annotation information was derived from a public SNP database (dbSNP ver. 126). We identified SNPs in putative promoter regions and selected SNPs that are predicted to be within TF binding sites. As a result, we mapped 488,452 SNPs and filtered out 47,832 SNPs within the putative TF binding sites.

Applying a conservation score

Using computational methods for predicting TFBS (TF binding sites) is not optimal due to a high false positive rate. However, recent algorithms have been improved in their reliability in TFBS prediction. Popular algorithms examine well-conserved regulatory sequences by comparing upstream sequences of orthologous genes across species [20-28]. Therefore, as an index of reliability for such an approach, we calculated an evolutionary conservation score for all the predicted TF binding sites. Users can see how reliable their predicted TF binding sites are. We used the phastcons16way file derived from UCSC human genome data. This file contains a conservation score from multiple genome alignment data calculated by the phastCons program [29].

Integration with functional annotation

The SNP@Promoter database adopted various gene annotations including pathways (KEGG), gene ontology (GOA), and disease information such as GAD, HGMD, and OMIM. The raw data files were integrated into the SNP@Promoter database based on a gene synonym table from HGNC (HUGO). These annotations provide insight into the effects of SNPs within TF binding sites and help users to characterize target genes regulated by SNPs.

User interface

As shown in Fig. 2(A), a user can search the SNP@Promoter database using three kinds of entries: 1) an SNP identifier (rs number from dbSNP), 2) a gene (Gene name, gene symbol, refSeq ID), or (3) a disease term. When the user submits a gene or a disease term, SNP@Promoter returns a gene list related to queries. In the case of accessing details of the query gene, it shows SNP information, gene information, and transcription factor binding site information of target genes as shown Fig. 2(B). SNP@Promoter provides graphical views of the queried SNPs and genes. Fig. 3 shows a putative promoter region browser.

Conclusion

SNP@Promoter is a database for functional SNPs within putative promoter regions and predicted TF binding sites. The database provides genetic information and graphical views of queried terms. SNP@Promoter will help researchers to identify functional SNPs in non-coding regions. Users can access the SNP@Promoter at http://variome.net or directly at http://variome.kobic.re.kr/SNPatPromoter.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

BK constructed the database. WYK developed the website and assisted to construction of database. WH and KS helped to develop the website. BK initiated this project and wrote the manuscript. DP assisted the manuscript writing. JB directed the study and helped to draft the manuscript.

Acknowledgements

We thank our colleagues at KOBIC, especially Areum Han. This project was supported by a grant from the KRIBB Research Initiative Program of Korea, by the Korea Science and Engineering Foundation (KOSEF) grant funded by the Korea government (MOST) (No. M10508040002-07N0804-00210), and by the MIC (Ministry of Information and Communication), Korea, under the KADO (Korea Agency Digital Opportunity and Promotion) support program (07-83).

This article has been published as part of *BMC Bioinformatics* Volume 9 Supplement 1, 2008: Asia Pacific Bioinformatics Network (APBioNet) Sixth International Conference on Bioinformatics (InCoB2007). The full contents of the supplement are available online at http://www.biomedcentral.com/1471-2105/9?issue=S1.

References

1. Collins FS, Brooks LD, Chakravarti A: **A DNA polymorphism discovery resource for research on human genetic variation.** *Genome Res* 1998, **8:**1229-1231.
2. Brookes AJ: **The essence of SNPs.** *Gene* 1999, **234:**177-186.
3. Kang HJ, Choi KO, Kim BD, Kim S, Kim YJ: **FESD: a Functional Element SNPs Database in human.** *Nucleic Acids Res* 2005, **33:**D518-522.
4. Chang H, Fujita T: **PicSNP: a browsable catalog of nonsynonymous single nucleotide polymorphisms in the human genome.** *Biochem Biophys Res Commun* 2001, **287:**288-291.
5. Riva A, Kohane IS: **SNPper: retrieval and analysis of human SNPs.** *Bioinformatics* 2002, **18:**1681-1685.
6. Yue P, Melamud E, Moult J: **SNPs3D: candidate gene and SNP selection for association studies.** *BMC Bioinformatics* 2006, **7:**166.
7. Reumers J, Schymkowitz J, Ferkinghoff-Borg J, Stricher F, Serrano L, Rousseau F: **SNPeffect: a database mapping molecular phenotypic effects of human non-synonymous coding SNPs.** *Nucleic Acids Res* 2005, **33:**D527-532.
8. Dantzer J, Moad C, Heiland R, Mooney S: **MutDB services: interactive structural analysis of mutation data.** *Nucleic Acids Res* 2005, **33:**W311-314.
9. Han A, Kang HJ, Cho Y, Lee S, Kim YJ, Gong S: **SNP@Domain: a web resource of single nucleotide polymorphisms (SNPs) within protein domain structures and sequences.** *Nucleic Acids Res* 2006, **1(34):**W642-W644.
10. Ng PC, Henikoff S: **SIFT: predicting amino acid changes that affect protein function.** *Nucleic Acids Res* 2003, **31:**3812-3814.
11. Han A, Kim WY, Park SM: **SNP2NMD: a database of human single nucleotide polymorphisms causing nonsense-mediated mRNA decay.** *Bioinformatics* 2007, **23:**397-399.
12. ElSharawy A, Manaster C, Teuber M, Rosenstiel P, Kwiatkowski R, Huse K, Platzer M, Becker A, Nurnberg P, Schreiber S, Hampe J: **SNPSplicer: systematic analysis of SNP-dependent splicing in genotyped cDNAs.** *Hum Mutat* 2006, **27:**1129-1134.
13. Blank MC, Stefanescu RN, Masuda E, Marti F, King PD, Redecha PB, Wurzburger RJ, Peterson MG, Tanaka S, Pricop L: **Decreased transcription of the human FCGR2B gene mediated by the -343 G/C promoter polymorphism and association with systemic lupus erythematosus.** *Hum Genet* 2005, **117:**220-227.
14. Myers RL, Airey DC, Manier DH, Shelton RC, Sanders-Bush E: **Polymorphisms in the regulatory region of the human serotonin 5-HT2A receptor gene (HTR2A) influence gene expression.** *Biol Psychiatry* 2007, **61:**167-173.
15. Ponomarenko JV, Orlova GV, Merkulova TI, Gorshkova EV, Fokin ON, Vasiliev GV, Frolov AS, Ponomarenko MP: **rSNP_Guide: An integrated database-tools system for studying SNPs and site-directed mutations in transcription factor binding sites.** *Hum Mutat* 2002, **20:**239-248.
16. Ponomarenko JV, Merkulova TI, Orlova GV, Fokin ON, Gorshkova EV, Frolov AS, Valuev VP, Ponomarenko MP: **rSNP_Guide, a database system for analysis of transcription factor binding to DNA with variations: application to genome annotation.** *Nucleic Acids Res* 2003, **31:**118-121.
17. Montgomery SB, Griffith OL, Sleumer MC, Bergman CM, Bilenky M, Pleasance ED, Prychyna Y, Zhang X, Jones SJ: **ORegAnno: an open access database and curation system for literature-derived promoters, transcription factor binding sites and regulatory variation.** *Bioinformatics* 2006, **22:**637-640.
18. Matys V, Fricke E, Geffers R, Goßling E, Haubrock M, Hehl R, Hornischer K, Karas D, Kel AE, Kel-Margoulis OV, Kloos DU, Land S, Lewicki-Potapov B, Michael H, Munch R, Reuter I, Rotert S, Saxel H, Scheer M, Thiele S, Wingender E: **TRANSFAC: transcriptional regulation, from patterns to profiles.** *Nucleic Acids Res* 2003, **31:**374-378.
19. Kel AE, Gossling E, Reuter I, Cheremushkin E, Kel-Margoulis OV, Wingender E: **MATCH: A tool for searching transcription factor binding sites in DNA sequences.** *Nucleic Acids Res* 2003, **31:**3576-3579.
20. Gumucio DL, Shelton DA, Zhu W, Millinoff D, Gray T, Bock JH, Slightom JL, Goodman M: **Evolutionary strategies for the elucidation of cis and trans factors that regulate the developmental switching programs of the beta-like globin genes.** *Mol Phylogenet Evol* 1996, **5:**18-32.
21. Hardison RC, Oeltjen J, Miller W: **Long human-mouse sequence alignments reveal novel regulatory elements: A reason to sequence the mouse genome.** *Genome Res* 1997, **7:**959-966.
22. Hardison RC: **Conserved noncoding sequences are reliable guides to regulatory elements.** *Trends Genet* 2000, **16:**369-372.
23. Levy S, Hannenhalli S, Workman C: **Enrichment of regulatory signals in conserved non-coding genomic sequence.** *Bioinformatics* 2001, **17:**871-877.
24. Loots GG, Ovcharenko I, Pachter L, Dubchak I, Rubin EM: **rVista for Comparative Sequence-Based Discovery of Functional Transcription Factor Binding Sites.** *Genome Res* 2002, **12:**832-839.
25. Steffens NO, Galuschka C, Schindler M, Bulow L, Hehl R: **AthaMap web tools for database-assisted identification of combinatorial cis-regulatory elements and the display of highly conserved transcription factor binding sites in Arabidopsis thaliana.** *Nucleic Acids Res* 2005, **33:**W397-W402.
26. Sinha S, Schroeder MD, Unnerstall U, Gaul U, Siggia ED: **Cross-species comparison significantly improves genome-wide prediction of cis-regulatory modules in Drosophila.** *BMC Bioinformatics* 2004, **5:**129.
27. Elnitski L, King D, Hardison RC: **Computational Prediction of cis-Regulatory Modules from Multispecies Alignments Using Galaxy, Table Browser, and GALA.** *Methods Mol Biol* 2006, **338:**91-103.
28. Pierstorff N, Bergman CM, Wiehe T: **Identifying cis-regulatory modules by combining comparative and compositional analysis of DNA.** *Bioinformatics* 2006, **22:**2858-64.
29. Siepel A, Bejerano G, Pedersen JS, Hinrichs A, Hou M, Rosenbloom K, Clawson H, Spieth J, Hillier LW, Richards S, Weinstock GM, Wilson RK, Gibbs RA, Kent WJ, Miller W, Haussler D: **Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes.** *Genome Res* 2005, **15:**1034-1050.