BMB
reports

CONVIRT: A web-based tool for transcriptional regulatory site identification using a conserved virtual chromosome

Taewoo Ryu[1], *Sejoon Lee*[2], *Cheol-Goo Hur*[1,*] *& Doheon Lee*[2,*]

[1]Bioinformatics Research Center, KRIBB, Daejeon 305-806, [2]Department of Bio and Brain Engineering, KAIST, Daejeon 305-701, Korea

Techniques for analyzing protein-DNA interactions on a genome-wide scale have recently established regulatory roles for distal enhancers. However, the large sizes of higher eukaryotic genomes have made identification of these elements difficult. Information regarding sequence conservation, exon annotation and repetitive regions can be used to reduce the size of the search region. However, previously developed resources are inadequate for consolidating such information. CONVIRT is a web resource for the identification of transcription factor binding sites and also features comparative genomics. Genomic information on ortholog-independent conserved regions, exons, repeats and sequences is integrated into the virtual chromosome, and statistically over-represented single or combinations of transcription factor binding sites are sought. CONVIRT provides regulatory network analysis for several organisms with long promoter regions and permits inter-species genome alignments. CONVIRT is freely available at http://biosoft.kaist.ac.kr/convirt. [BMB reports 2009; 42(12): 823-828]

INTRODUCTION

Gene expression is mediated by transcription machinery consisting of RNA polymerase and transcription factors (TFs). TFs interact with short DNA sequences, called *cis*-elements, which are concentrated proximal to promoters in simple organisms or dispersed over broad regions in complex organisms (1-3). Identification of TF binding sites (TFBSs) is vital for understanding regulatory mechanisms within the nucleus, and much effort has been devoted to this task. However, the size of the search region presents obstacles for the identification of TFBSs of higher eukaryotes. Phylogenetic footprinting methods considering only regions that are conserved among closely related organisms have emerged as efficient tools for eliminating less informative regions (4-6).

However, these tools have several practical disadvantages: (i) Most resources only analyze promoter-proximal regions, even though many TFBSs are located often several hundred kilobases from the transcription start site (TSS) (2, 7); (ii) Existing resources focus mainly on the identification of human or mouse regulatory sites, and hence cannot be utilized for analyzing other organisms; (iii) Input genes cannot be analyzed if ortholog genes are unknown; (iv) It is very useful to analyze *cis*-elements for a specific set of TFs using several search parameters, but many tools do not provide this option; (v) Most existing tools can provide the statistical significance for the occurrence of a TFBS for a single TF only, and do not address the increasing demand to identify combinatorial TF regulation (8-10).

To overcome these problems, we have developed a Web resource, CONVIRT, which performs regulatory network analysis in higher eukaryotes. Whole genome alignments, gene annotation, repeats and sequence information are integrated into a virtual chromosome, which allows any region at any distance from a chosen gene to be easily retrieved. For improved TF-target network analysis, TFBS search parameters are user-specified or optimized automatically by the server, and combinatorial TFBS enrichment is then analyzed.

RESULTS

Rationale

The first step in phylogenetic footprinting is the identification of homologous genes from a set of organisms, followed by comparison of promoter sequences near the TSSs. However, the queried gene cannot be analyzed if no counterpart is found in another organism. This is often the case as many genes unfortunately lack orthologs in other organisms. For instance, only 16,751 human genes, or about 66% of known genes, have mouse counterparts according to NCBI Homologene (11). To overcome this problem, we used sequence conservation information from whole genome alignment (12). Identification of homologous genes was thus rendered unnecessary, and genes without orthologs were analyzed.

Another advantage of utilizing whole genome alignment is the ability to identify functional regions that evolved independently from associated genes. The sequence of a genome is continuously modified during evolution by genomic rearrangement processes. The most common mechanism is replication and random insertion of genome sequences by mo-

*Corresponding author. Doheon Lee, Tel: 82-42-350-4316; Fax: 82-42-350-8680; E-mail: dhlee@kaist.ac.kr, Cheol-Goo Hur, Tel: 82-42-879-8560; Fax: 82-42-879-8569; E-mail: hurlee@kribb. re.kr

Received 22 April 2009, Accepted 18 August 2009

Keywords: *Cis*-element, Comparative genomics, Phylogenetic footprinting, TFBS, Transcription factor

bile elements. Interestingly, TFBSs within these elements are also transplanted to new loci. For example, the genome-wide distribution of p53 binding sites within ERV retroelements was recently reported, and their transcriptional activities were experimentally validated (13). Alu repeat sequences were also suggested to harbor many functional TFBSs involved in various gene regulation mechanisms (14). RE1, the binding site for REST (the neuronal fate determinant), was found in many transposable elements, and the positions of many RE1 sites were not related to those of orthologs (15). Taken together, these data suggest that many functional regions are not related to orthologs. Our analysis also supports this idea, as shown in Supplemental Table 1. For all genome alignments tested, the majority of conserved regions were not related to orthologs.

To construct the system, we selected organisms whose gene annotations and genome alignments were available in the UCSC database (16) (Supplemental Table 2). Genome alignment results by BLASTZ are provided in either chain or net format. The chain format contains both paralogous and orthologous alignments, whereas the net format yields only orthologous alignments (17). Netted alignments are thus used in our method since phylogenetic footprinting focuses only on orthologous regions. A large amount of information may be obtained from genome alignments and sequence files, which increases the complexity and time of retrieval. We therefore established a virtual chromosome in order to integrate various resources and retrieve information quickly (Fig. 1). Using this framework, one can retrieve any specific region using various options, i.e., inclusion or exclusion of conserved regions, non- conserved regions, exons or repeats (see Materials and Methods for details).

To identify candidate TFBSs in retrieved sequences, our system allows users to specify a position frequency matrix (PFM) and a matrix similarity score (MSS) threshold because new TFs and binding profiles are continuously discovered. The decision on the similarity threshold for each PFM is important because dramatic changes in the number of TFBS candidates can occur,

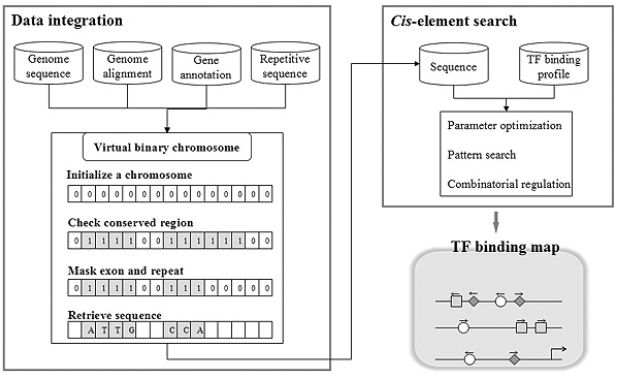

Fig. 1. Schematic diagram of the CONVIRT system. Information on conserved regions, exons, repeats and the sequence are integrated into the virtual chromosome. See text for details.

and many false-positives or -negatives may appear, depending on the chosen threshold, even though a uniform threshold (e.g., 80%,) for all TFs has been used in many studies (4, 6, 18). Our system therefore allows two choices for the MSS threshold. If the threshold for each PFM is user-specified, this value is used as the threshold. If not, the threshold is automatically calculated using a previously described optimization method (7). This approach defines the optimal threshold for each PFM so that maximum *n* TFBSs are allowed per limited length of random sequence. This limits the density of TFBSs and thus balances the occurrence frequencies between highly specific and less specific TFs.

Combinatorial regulation by multiple TFs is of interest since it allows a limited number of TFs to regulate many genes in response to diverse environmental signals (8-10). Moreover, owing to the nature of TFBS prediction, the false-positive rate for an individual TFBS is extremely high, and identification of significant TFBS combinations is useful in reducing such unwanted predictions (10, 18, 19). If two or more TF-binding profiles are uploaded by the user, CONVIRT automatically generates all possible TF combinations as well as the statistical significance of each combination in a given gene set.

We tested our system using three examples: (i) muscle-specific TF and target genes, (ii) estrogen receptor and target genes, and (iii) REST and target genes. The first example above is well-known to exhibit cooperative TF binding to the proximal regions of muscle gene promoters; the second example was used to demonstrate TFBS identification in regions distal to TSSs; and the final example explored TF binding to repetitive regions. The TFs and targets used in our case studies are listed in Supplemental Table 3 and our website.

Case study 1: muscle-specific TF and target genes

The DNA binding profiles of five muscle-specific TFs were obtained from the work of Wasserman and Fickett (20), and relevant several target genes were retrieved from TRANSFAC (21). The proposed method was applied to this test set with default options, which found the regulatory module targeted by three TFs (SRF, SP1, and MYF) to be significant (P value: 1.098E-4). Among the target genes in this module, CHRND, a acetylcholine receptor involved in muscle contraction, was chosen as an example with its regulatory elements depicted in Fig. 2A. Even though the biological roles of acetylcholine receptors are well known, the regulatory mechanism governing receptor transcription has been poorly studied thus far. In this example, four TFs are suggested as candidate regulators of the CHRND gene, and all binding sites are located within the human-mouse conserved region. This should be validated experimentally since this particular TF combination has already been suggested (20) and the functions of the TFs and target genes are consistent.

Case study 2: estrogen receptor and target genes

The estrogen receptor (ER), which is essential for sexual devel-

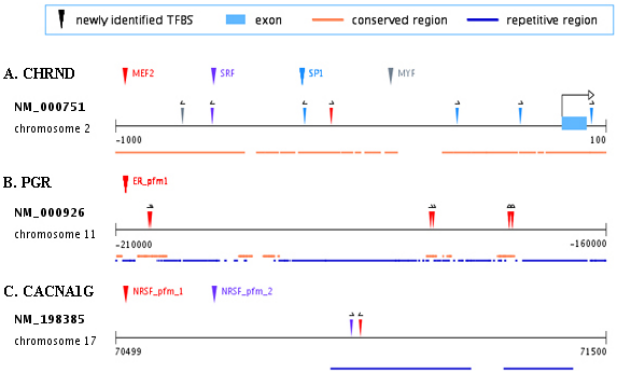

Fig. 2. Examples of TF-target interactions identified by the CONVIRT system. (A) Multiple regulation of the nicotinic cholinergic receptor gene by muscle-specific TFs. (B) Estrogen receptor binding to the distal enhancer of the progesterone receptor gene. (C) REST, the neuronal fate regulator, binding to the downstream repetitive region of the CACNA1G gene. The results are depicted using the graphic module of CONVIRT. All arrowheads show newly identified TFBSs and colors discriminate between different TFs. The term 'Newly identified TFBS' indicates a TFBS identified using our system. Even though no known TFBSs are shown in this figure, users can easily identify previously known TFBSs from the TRANSFAC database using individual queries. The arrow above each TFBS indicates the direction of the TFBS relative to the TSS. Some TFBSs are shown as overlapped due to their close proximity. The conservation information is from the human-mouse netted alignment. Other search options were set at default values, except that non-conserved regions and repetitive sequences were not masked in the third example.

opment, has been shown experimentally to bind an enhancer region several hundreds of kilobases away from the TSS (2). We used an ER binding profile compiled by TRANSFAC and the progesterone receptor (PGR) gene as a target. Eight ER binding sites were identified by inspecting the region from -210 kb to -160 kb of the PGR gene (Fig. 2B) while two sites were identified in a ChIP-chip experiment (2). The difference in the number of described TFBSs is probably because the ChIP-chip experiment provides a snapshot of TF-DNA interactions under specific conditions, whereas all possible binding sites are identified in the computational analysis.

Case study 3: REST and target genes

Even though repetitive DNA sequences are often regarded as irrelevant, recent studies have verified the transcriptional regulatory activity of transposable elements (13, 14). Chromosomal rearrangement by such mobile elements may move regulatory sequences to new genomic regions during evolution. This might explain why most conserved regions are not related to orthologs (Supplemental Table 1). Therefore, comparison of only non-repetitive sequences that flank ortholog genes may miss numerous potential regulatory regions. For example, the CACNA1G gene, which encodes a voltage-gated calcium channel protein, contains the target site for REST, which represses the expression of neuronal genes in non-neuronal tissues, in its downstream region (15). We analyzed the downstream region of the gene and found two TFBSs in the repeat region near +71 kb that show high similarity to the REST binding profile (Fig. 2C). The positions of these sites are the same as those previously identified (15).

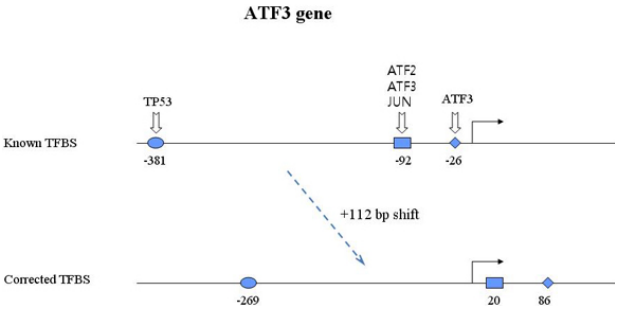

Fig. 3. Inconsistency in the positions of TFBSs. Positions of previously known TFBSs were searched for in the NCBI reference genome, and many were shifted or mismatched. In this example, three known TFBSs upstream of the human ATF3 gene are located 112 bp downstream of the reference genome sequence.

Inconsistency of TFBSs in previously known position and genomic sequences

When the results of the TFBS prediction were compared to previously identified sites, we found many inconsistencies. Surprisingly, we found that TFBSs were shifted either up- or down-stream in many cases (Fig. 3). Experimentally identified TFBSs upstream of the human ATF3 gene were searched for in the reference genome and were found 112 bp downstream from previously reported positions. We further analyzed the positions of known TFBSs listed in the TRANSFAC database (21). TFBSs with known sequences and positions relative to TSSs were extracted and mapped to the promoter sequences of RefSeq genes. Among 444 human TFBSs, 65 sites were not found in the promoter sequences, whereas the positions of 227 sites were shifted (Supplemental Table 4).

This inconsistency could be due to one of the following: (i) NCBI RefSeq and genome sequences are continuously updated. Thus, TSS annotations might have changed from the original annotations, which would shift the positions of TFBSs relative to TSSs; (ii) Since the reference genome used in whole genome sequencing is different from the cell lines used in particular research projects, the positions and sequences of TFBSs do not exactly match. The human genome project utilized blood cells of anonymous donors (22), whereas laboratory human cell lines originate from humans of different race, various tissues or patients with various diseases (e.g., the HeLa cell line came from a cervical cancer patient). Tissue-specific deletions and mutations during development, aging, disease progression (23-25), as well as the development of genomic variants such as single/multi-nucleotide polymorphisms and insertion/deletion events (26), might explain the observed incon-

sistencies. These possibilities are further supported by the data of Smith and colleagues (27), who found that the positions of several TFBSs did not agree with previous annotations. Therefore, previously reported TFBS positions should be thoroughly checked in the reference genome. The corrected positions of individual sites are displayed in the web search result.

DISCUSSION

Regulatory element analysis using comparative genomics (phylogenetic footprinting) is a powerful tool for the study of higher eukaryotes in which TF binding sites are dispersed over several hundreds of kilobases. The basic assumption of phylogenetic footprinting is that functional, non-coding elements are under high selection pressure and therefore are conserved through evolution. A simple method for identifying such sequence elements is to align the promoter sequences of orthologs from closely related species. Many methods have been developed based on this approach. However, Supplemental Table 1 shows that conserved elements near orthologs form only a small proportion of all conserved elements. Moreover, previously published tools are biased to human-mouse alignment, restrict TFBS searches to promoter-proximal regions and do not allow TFBS search parameters to be adjusted. CONVIRT was developed to overcome these problems and to facilitate regulatory network analysis in higher eukaryotes. Even though only seven organisms are covered in the current version, other organism data can be easily added to the system when genome sequences and alignments become more available.

Conventional approaches to TFBS identification have yielded good results in many studies. However, as pointed out in several reports, many predicted sites are not functional despite successful TF binding *in vitro* (4,19). This is partly due to incompleteness of the current position frequency matrix model as well as an inadequate search algorithm that relies on the base frequency at each position. Further elucidation of the binding mechanisms of TFs to DNA is crucial for reducing false-positives. For example, Michal and co-workers studied the effect of *cis*-element variation on the pattern of target gene expression (28). Relationships between TFBSs and flanking sequences also require further study (29). As TF binding models and TFBS search mechanisms are improved, functional binding sites will be correctly identified even in large genomes.

MATERIALS AND METHODS

Integration of genomics information using CONVIRT

Pairs of organisms whose pairwise genomic alignments and gene annotations were available were chosen from the UCSC database (Supplemental Table 2). As genomic alignment results included both orthologs and paralogs, only top-level chains in the net alignment file were used to exclude paralogous alignments. A virtual chromosome was prepared for efficient integration and retrieval of information. The system first identifies the chromosome for each query upon user specification of organisms, alignments and genes (or specific regions), followed by initialization of a binary array equal in length to each chromosome with 0 at all positions. Indices corresponding to conserved positions are then converted to 1 by reading of complex net alignment files. A researcher may want to inspect non-conserved regions rather than conserved regions for some purposes. CONVIRT provides options to perform this task, and only non-conserved regions are converted to 1 in this case. The masking of exons or repetitive regions is a useful option for reducing search space in many studies (15, 30, 31). Repeat-masked chromosome sequences were downloaded from the UCSC database. If the masking option is selected, exon or repeat positions on the chromosome are set to 0. Finally, the base at each position is retrieved if the index at the virtual chromosome is 1. For users wishing to employ retrieved information in further analysis, we provide such information on the results page.

TFBS search and statistical significance

The matrix-based pattern search algorithm defined in MATCH (30) is employed to seek candidate TFBSs in retrieved sequences. PFMs uploaded by users in the appropriate format (see our website) are normalized by the system. Two choices are permitted to establish the MSS threshold. If a user specifies the threshold for each PFM, this value is used as the MSS cut-off. If not, the system automatically defines a balanced MSS threshold using a previously described method (7), which predicts *n* binding sites every 10 kb. We generated a random background sequence for each organism. A TFBS is searched for in this random sequence employing a user-specified PFM, and the occurrence number of each TFBS is restricted to less than the user-inputted number (default: 3) per 10 kb.

The significance of predicted TFBSs can be evaluated statistically. In our system, P values are calculated by Fisher's exact test to indicate the enrichment of each TFBS for each input gene compared to random background. A P value is given by:

$$P = \sum_{i=t}^{\min(T,g)} \frac{[\frac{T}{i}][\frac{G-T}{g-i}]}{\frac{G}{g}}$$

where G is the total number of genes in the dataset, g is the number of genes in the foreground set, T is the number of genes targeted by the TF in the dataset and t is the number of genes bound by the TF in the foreground set.

To identify cooperative regulation by multiple TFs, CONVIRT automatically calculates the statistical significance of all possible TF combinations once two or more TF-binding profiles are uploaded. Fisher's exact test is again adopted to measure the significance of each TF combination; here, T and t are the numbers of genes targeted by all TFs in each subset (32, 33).

Hence, the resulting P value is the probability of TFBS co-occurrence within the selected genes, or within a greater number of genes by chance. The TF or TF set with a Bonferroni-corrected P value smaller than 0.05 is shown on the results page.

Acknowledgements
This work was supported by a NRL Grant (2005-01450) from the Ministry of Education, Science, and Technology (MEST), Korea. TR and CH were additionally supported by grants from the Plant Diversity Resource Center (PDRC) of the 21st Century Frontier Research Programs of MEST.

REFERENCES

1. Bussemaker, H. J., Li, H. and Siggia, E. D. (2001) Regulatory element detection using correlation with expression. *Nat. Genet.* **27**, 167-171.
2. Carroll, J. S., Meyer, C. A., Song, J., Li, W., Geistlinger, T. R., Eeckhoute, J., Brodsky, A. S., Keeton, E. K., Fertuck, K. C., Hall, G. F., Wang, Q., Bekiranov, S., Sementchenko, V., Fox, E. A., Silver, P. A., Gingeras, T. R., Liu, X. S. and Brown, M. (2006) Genome-wide analysis of estrogen receptor binding sites. *Nat. Genet.* **38**, 1289-1297.
3. Du, R., An, X. R., Chen, Y. F. and Qin, J. (2007) Some motifs were important for myostatin transcriptional regulation in sheep (Ovis aries). *J. Biochem. Mol. Biol.* **40**, 547-553.
4. Karanam, S. and Moreno, C. S. (2004) CONFAC: automated application of comparative genomic promoter analysis to DNA microarray datasets. *Nucleic. Acids. Res.* **32**, W475-484.
5. Loots, G. G. and Ovcharenko, I. (2004) rVISTA 2.0: evolutionary analysis of transcription factor binding sites. *Nucleic. Acids. Res.* **32**, W217-221.
6. Berezikov, E., Guryev, V. and Cuppen, E. (2005) CONREAL web server: identification and visualization of conserved transcription factor binding sites. *Nucleic. Acids. Res.* **33**, W447-450.
7. Pennacchio, L. A., Loots, G. G., Nobrega, M. A. and Ovcharenko, I. (2007) Predicting tissue-specific enhancers in the human genome. *Genome. Res.* **17**, 201-211.
8. Beer, M. A. and Tavazoie, S. (2004) Predicting gene expression from sequence. *Cell.* **117**, 185-198.
9. Kato, M., Hata, N., Banerjee, N., Futcher, B. and Zhang, M. Q. (2004) Identifying combinatorial regulation of transcription factors and binding motifs. *Genome Biol.* **5**, R56.
10. Ryu, T., Kim, Y., Kim, D. W. and Lee, D. (2007) Computational identification of combinatorial regulation and transcription factor binding sites. *Biotechnol. Bioeng.* **97**, 1594-1602.
11. NCBI Homologene. [http://www.ncbi.nlm.nih.gov/sites/ entrez?db=homologene].
12. Schwartz, S., Kent, W. J., Smit, A., Zhang, Z., Baertsch, R., Hardison, R. C., Haussler, D. and Miller, W. (2003) Human-mouse alignments with BLASTZ. *Genome Res.* **13**, 103-107.
13. Wang, T., Zeng, J., Lowe, C. B., Sellers, R. G., Salama, S. R., Yang, M., Burgess, S. M., Brachmann, R. K. and Haussler, D. (2007) Species-specific endogenous retroviruses shape the transcriptional network of the human tumor suppressor protein p53. *Proc. Natl. Acad. Sci. U.S.A.* **104**, 18613-18618.
14. Polak, P. and Domany, E. (2006) Alu elements contain many binding sites for transcription factors and may play a role in regulation of developmental processes. *BMC Genomics.* **7**, 133.
15. Johnson, R., Gamblin, R. J., Ooi, L., Bruce, A. W., Donaldson, I. J., Westhead, D. R., Wood, I. C., Jackson, R. M. and Buckley, N. J. (2006) Identification of the REST regulon reveals extensive transposable element-mediated binding site duplication. *Nucleic. Acids. Res.* **34**, 3862-3877.
16. UCSC Genome Browser. [http://hgdownload.cse.ucsc.edu/downloads.html].
17. Kent, W. J., Baertsch, R., Hinrichs, A., Miller, W. and Haussler, D. (2003) Evolution's cauldron: duplication, deletion, and rearrangement in the mouse and human genomes. *Proc. Natl. Acad. Sci. U.S.A.* **100**, 11484-11489.
18. Lenhard, B., Sandelin, A., Mendoza, L., Engstrom, P., Jareborg, N. and Wasserman, W. W. (2003) Identification of conserved regulatory elements by comparative genome analysis. *J. Biol.* **2**, 13.
19. Wasserman, W. W. and Sandelin, A. (2004) Applied bioinformatics for the identification of regulatory elements. *Nat. Rev. Genet.* **5**, 276-287.
20. Wasserman, W. W. and Fickett, J. W. (1998) Identification of regulatory regions which confer muscle-specific gene expression. *J. Mol. Biol.* **278**, 167-181.
21. Matys, V., Fricke, E., Geffers, R., Gossling, E., Haubrock, M., Hehl, R., Hornischer, K., Karas, D., Kel, A. E., Kel-Margoulis, O. V., Kloos, D. U., Land, S., Lewicki-Potapov, B., Michael, H., Munch, R., Reuter, I., Rotert, S., Saxel, H., Scheer, M., Thiele, S. and Wingender, E. (2003) TRANSFAC: transcriptional regulation, from patterns to profiles. *Nucleic. Acids. Res.* **31**, 374-378.
22. Osoegawa, K., Mammoser, A. G., Wu, C., Frengen, E., Zeng, C., Catanese, J. J. and de Jong, P. J. (2001) A bacterial artificial chromosome library for sequencing the complete human genome. *Genome Res.* **11**, 483-496.
23. Ponomareva, O. N., Rose, J. A., Lasarev, M., Rasey, J. and Turker, M. S. (2002) Tissue-specific deletion and discontinuous loss of heterozygosity are signatures for the mutagenic effects of ionizing radiation in solid tissues. *Cancer Res.* **62**, 1518-1523.
24. Semina, E. V., Murray, J. C., Reiter, R., Hrstka, R. F. and Graw, J. (2000) Deletion in the promoter region and altered expression of Pitx3 homeobox gene in aphakia mice. *Hum. Mol. Genet.* **9**, 1575-1585.
25. Takada, H., Imoto, I., Tsuda, H., Nakanishi, Y., Sakakura, C., Mitsufuji, S., Hirohashi, S. and Inazawa, J. (2006) Genomic loss and epigenetic silencing of very-low-density lipoprotein receptor involved in gastric carcinogenesis. *Oncogene.* **25**, 6554-6562.
26. Levy, S., Sutton, G., Ng, P. C., Feuk, L., Halpern, A. L., Walenz, B. P., Axelrod, N., Huang, J., Kirkness, E. F., Denisov, G., Lin, Y., MacDonald, J. R., Pang, A. W., Shago, M., Stockwell, T. B., Tsiamouri, A., Bafna, V., Bansal, V., Kravitz, S. A., Busam, D. A., Beeson, K. Y., McIntosh, T. C.,

Remington, K. A., Abril, J. F., Gill, J., Borman, J., Rogers, Y. H., Frazier, M. E., Scherer, S. W., Strausberg, R. L. and Venter, J. C. (2007) The diploid genome sequence of an individual human. *PLoS. Biol*. **5**, e254.
27. Smith, A. D., Sumazin, P. and Zhang, M. Q. (2007) Tissue-specific regulatory elements in mammalian promoters. *Mol. Syst. Biol*. **3**, 73.
28. Michal, L., Mizrahi-Man, O. and Pilpel, Y. (2008) Functional characterization of variations on regulatory motifs. *PLoS. Genet*. **4**, e1000018.
29. Vega, V. B., Lin, C. Y., Lai, K. S., Kong, S. L., Xie, M., Su, X., Teh, H. F., Thomsen, J. S., Yeo, A. L., Sung, W. K., Bourque, G. and Liu, E. T. (2006) Multiplatform genome-wide identification and modeling of functional human estrogen receptor binding sites. *Genome. Biol*. **7**, R82.
30. Kel, A. E., Gossling, E., Reuter, I., Cheremushkin, E., Kel-Margoulis, O. V. and Wingender, E. (2003) MATCH: A tool for searching transcription factor binding sites in DNA sequences. *Nucleic. Acids. Res*. **31**, 3576-3579.
31. Bedell, J. A., Korf, I. and Gish, W. (2000) MaskerAid: a performance enhancement to RepeatMasker. *Bioinformatics*. **16**, 1040-1041.
32. Ho Sui, S. J., Fulton, D. L., Arenillas, D. J., Kwon, A. T. and Wasserman, W. W. (2007) oPOSSUM: integrated tools for analysis of regulatory motif over-representation. *Nucleic. Acids. Res*. **35**, W245-252.
33. Zhou, Y., Ferguson, J., Chang, J. T. and Kluger, Y. (2007) Inter- and intra-combinatorial regulation by transcription factors and microRNAs. *BMC Genomics*. **8**, 396.