BMC Bioinformatics

Genome-wide prediction, display and refinement of binding sites with information theory-based models

BMC Bioinformatics 2003, 4:38

Sashidhar Gadiraju (sashidharg@yahoo.com)
Carrie A Vyhlidal (cvyhlidal@cmh.edu)
J. Steven Leeder (sleeder@cmh.edu)
Peter K. Rogan (progan@cmh.edu)

ISSN 1471-2105

Article type Software

Submission date 07 May 2003

Acceptance date 08 Sep 2003

Publication date 08 Sep 2003

Article URL http://www.biomedcentral.com/1471-2105/4/38

Genome-wide prediction, display and refinement of binding sites with information theory-based models

Sashidhar Gadiraju (sgadiraju@cmh.edu)[1,3],
Carrie A. Vyhlidal(cvyhlidal@cmh.edu)[2],
J. Steven Leeder (sleeder@cmh.edu)[2]
and Peter K. Rogan[1,3]

[1]Laboratory of Human Molecular Genetics and
[2]Section of Developmental and Experimental Pharmacology and Therapeutics,
Children's Mercy Hospital and Clinics,
School of Medicine and
[3]School of Interdisciplinary Computer Science and Engineering
University of Missouri-Kansas City
Kansas City MO 64108 USA

Correspondence to:
Peter K. Rogan, Ph.D.
Laboratory of Human Molecular Genetics
Children's Mercy Hospital and Clinics
2401 Gillham Rd.
Kansas City MO 64108
816-960-4820
progan@cmh.edu

Abstract

Background

We present *Delila genome,* a software system for identification, visualization and analysis of protein binding sites in complete genome sequences. Binding sites are predicted by scanning genomic sequences with information theory-based (or user-defined) weight matrices. Matrices are refined by adding experimentally-defined binding sites to published binding sites. *Delila Genome* was used to examine the accuracy of individual information contents of binding sites detected with refined matrices as a measure of the strengths of the corresponding protein-nucleic acid interactions. The software can then be used to predict novel sites by rescanning the genome with the refined matrices.

Results

Parameters for genome scans are entered using a Java-based GUI interface and backend scripts in Perl. Multi-processor CPU load-sharing minimized the average response time for scans of different chromosomes. Scans of human genome assemblies required 4-6 hours for transcription factor binding sites and 10-19 hours for splice sites, respectively, on 24- and 3-node Mosix and Beowulf clusters. Individual binding sites are displayed either as high-resolution sequence walkers or in low-resolution custom tracks in the UCSC genome browser. For large datasets, we applied a data reduction strategy that limited displays of binding sites exceeding a threshold information content to specific chromosomal regions within or adjacent to genes. An HTML document is produced listing binding sites ranked by binding site strength or chromosomal location hyperlinked

to the UCSC custom track, other annotation databases and binding site sequences. Post-genome scan tools parse binding site annotations of selected chromosome intervals and compare the results of genome scans using different weight matrices. Comparisons of multiple genome scans can display binding sites that are unique to each scan and identify sites with significantly altered binding strengths.

Conclusions

Delila-Genome was used to scan the human genome sequence with information weight matrices of transcription factor binding sites, including PXR/RXRα, AHR and NF-κB p50/p65, and matrices for RNA binding sites including splice donor, acceptor, and SC35 recognition sites. Comparisons of genome scans with the original and refined PXR/RXRα information weight matrices indicate that the refined model more accurately predicts the strengths of known binding sites and is more sensitive for detection of novel binding sites.

Background

We describe a system to identify and display significant non-coding genomic sequences that are important for transcriptional regulation and post-transcriptional mRNA processing. Our system builds on *Delila* [1], a series of programs designed to scan sets of sequence fragments (or small genomes, ie. bacterial) for potential binding sites. The regulatory sequences that are bound by proteins are detected by the tools provided with the *Delila* system, which defines binding sites according to Shannon information theory [2].

Information content is the number of choices needed to describe a sequence pattern and has units of bits [3]. In the analysis of nucleic acid binding sites, functional site sequences are aligned and the frequencies of nucleotides at each position are used to calculate the individual information weight matrix, $R_i(b,l)$ [4] of each base *b* at position *l*. Computation of binding site $R_i(b,l)$ information weight matrices based upon published and laboratory-derived sites is a prerequisite to detecting and visualizing predicted binding sites with *Delila-Genome*. The procedures and software used to derive these matrices have been previously described [1, 4, 5] for different types of protein binding sites [6-13]. This matrix is used to scan the genome and evaluate the individual information content (R_i , in bits) of potential binding sites. Functional binding sites have values > 0 bits and the consensus sequence has the maximum R_i value. A single bit difference in R_i value corresponds to at least a two-fold difference in binding site strength. Changes in information content resulting from mutations correspond to observed phenotypes both in vitro and in vivo [6-8, 11]; by contrast, non-deleterious polymorphisms result in nominal

changes in R_i. value. Therefore, scans with information weight matrices can be used to measure the relative strengths of potential binding sites throughout the genome.

Scans of eukaryotic genomes [14, 15] often require longer execution times and generate considerably larger outputs than prokaryotic genome scans [10] due to increased genome sizes and the quantities of sites detected. The development of *Delila-Genome* was motivated by the need to streamline the detection and display of sets of the most relevant binding sites in eukaryotic genomic or heteronuclear RNA sequences. Visual juxtaposition of these results with other genomic annotation facilitates the prediction and interpretation of binding sites. In order to limit the presentation of weak binding sites (with lower than average information content, ie. $<<R_{sequence}$) which can be densely distributed in both expressed and non-expressed genomic intervals, we developed visualization tools in *Delila Genome* to mine relevant binding sites in gene-rich regions and to display clusters of sites with their respective information contents. Details of individual binding sites can also be presented at high resolution as sequence walkers [5], which depict contributions of each nucleotide to the overall information content of the site.

The number and R_i values of the sites that define the information weight matrix, $R_i(b,l)$, dictate which binding sites are predicted and the corresponding strengths of these sites found in genome scans. Models based on small numbers of proven binding sites may fail to detect valid binding sites and can tend to predict R_i inaccurately. Iterative selection of functional binding sites has been used to optimize [7, 8, 16] and to introduce bias [17]

into the frequencies of each nucleotide in computing the information theory-based weight matrices of binding sites. Significant differences between information weight matrices have been determined from their respective evolutionary distance metrics (for example, see [10]). *Delila Genome* monitors the effects of model alterations by comparing the genome scan results for pairs of information weight matrices. Although the primary application is to compare sets of binding sites with successive versions of the same weight matrix, other potential applications include determining the locations of overlapping binding sites recognized by different proteins and comparisons of binding sites detected with information models of orthologous proteins from different species.

Delila-Genome has been optimized to compute the locations of prospective transcription factor and splicing recognition sites by information theory-based analyses of recent human genome draft and finished sequences. We describe this software system, measure its performance, and illustrate the results of genome scans using visualization and post-genomic analytic tools which monitor the effects of matrix refinement on genome-wide identification of binding sites.

Implementation

The *Delila-Genome* system has a client-server architecture which is comprised of three functional modules: (A) the *Delila-Genome* Front End, (B) the *Delila Genome* Server and (C) Post-genomic scan analysis tools (Figure 1). The front end is a graphical interface that takes user input to set parameters for scanning the genome sequence and processing the results. It interacts with the system tools, and while it currently does not

have a WWW interface like the UCSC genome browser, it is available as an installable module. The server is the actual engine of the system where all the tools are hosted and all the computations are performed. For multiprocessor servers, a load balancing feature has been written for the Scyld operating system (for Beowulf clusters) using the 'mpprun' utility. This feature is not supported in operating systems like Mosix, where load balancing is done automatically based on CPU utilization. We now describe each of these modules and their respective interactions and dependencies.

Delila-Genome Front End

Submission of the genome scan A front end was developed for submission of the genome-wide or chromosomal scans and for tailoring the output to filter and view the most relevant results. A Java-based GUI tool (developed with Java Swing technology) enables submission of scans to the server. Besides the *Delila* books containing chromosomal sequences, the only required input file is the $R_i(b,l)$ information weight matrix (ribl) of the protein binding site. This file is output by the *ri* program, and the procedure for generating this file has been described [4]. In order to assess the degree to which the computed information depends on these weights, an option is provided to modify this matrix by uploading a file containing these weights or entering them as integers on a Java form. Parameters are requested for the *Delila scan* program [18] which performs the genome scan, and *promotsite* (see below: **Delila Genome Server**), a program that produces files for displaying binding sites within or adjacent to genes. The user selects the program to execute on the server and then either fills in the parameters required by the selected program or the front-end can pull the default parameters from the

server. The front end also displays all of the genome assembly versions installed on the server (at our institution: human genome versions April, 2003, November, 2002, and October, 2000). The front end validates the parameters before submission. Java socket programming is used to connect to the server.

Visualization To present the most relevant results from the scans, Delila-Genome uses Javascript to produce an HTML page listing binding sites within or adjacent to expressed loci in the human genome sequence. The user can view these binding sites at *low resolution* (relative to genes and other sites) or at *high resolution* (at the nucleotide level). Figure 2 shows an HTML page with corresponding high and low resolution links associated with each binding site. Binding sites are selected based on their proximity to the 5' termini of transcripts mapped onto the human genome draft at the UCSC Genome Browser database (http://genome.ucsc.edu). The coordinates of mapped transcripts are read from each chromosome-specific, mRNA annotation table (downloaded from the UCSC Genome Browser annotation database (files: chrXX_mrna.txt) into the chromosome-specific directories containing the corresponding genomic sequences). Currently, the genome contains numerous expressed sequences that have not been definitively established as genes in public databases. By defining binding sites in the context of such mRNAs mapped onto the genome sequence, it may be possible to annotate regulatory or other features in otherwise poorly-characterized, expressed coding sequences.

Low resolution tools The server generates a list of predicted binding sites as a BED-

formatted file (http://genome.ucsc.edu/goldenPath/help/customTrack.html) which is uploaded to the appropriate human genome draft browser at the inception of the session. The name assigned to a site is a concatenation of the GenBank accession number associated with the site (described below: Delila-Genome server, *promotsite*), the name of $R_i(b,l)$ matrix, ie. type of site, and the strength of the site in bits. Sites are represented as a color-shaded block in the custom track of the UCSC browser. The *score* field of the BED file controls the degree of shading of the site, with the strongest sites being the most opaque and the weakest being the most transparent. The score used in *Delila Genome* BED files is a linear scaling of the R_i value. The start and end coordinates of a site correspond to the thick- and thin-ends of the BED features, respectively, so that its orientation can be visualized at high magnification. The *Scandiff* program generates BED files for different categories of output, each of which has a unique color coding. The *genvis* Perl tool selects genes with sites either within user-defined chromosomal intervals or sorted by information content from input BED files and generates HTML pages hyperlinked to the UCSC genome browser custom track. The user can either retrieve the BED files from the server and upload them to the genome browser locally, or connect to the server using X terminal software and upload them from the server to the genome browser.

By navigating the other hyperlinks on the HTML page, one can view (i) the DNA sequence of a binding site (Fig. 2C), (ii) detailed characteristics of the binding site on the UCSC genome browser custom track (Fig. 2B), (iii) GenBank (Fig. 2E) and Stanford SOURCE (Fig. 2F) relational data describing the mRNA associated with this site, and

(iv) all binding sites adjacent to the accession number on the UCSC browser within a user-defined window size (Fig. 2G).

High resolution tools The contributions of each nucleotide (in bits) to the overall individual information content of a single binding site can be at viewed at high resolution using sequence walkers ([5]; shown in Figure 2D). A walker graphically represents the weight of each nucleotide at each position in a single possible binding site, with the height of the nucleotide indicating how well the bases match the individual information weight matrix.

To display a sequence walker, the DNA sequence containing the binding site (through a hyperlink on the HTML page) should be stored in the user's autolister directory on the server (or a Linux/Unix client running *Delila*). The *Delila atchange* script is configured to display the sequence walker by running the *Delila-Genome autolist* script which scans the downloaded sequence for binding sites, executes the *lister* program to generate a postscript image of the sequence walker, and pops up the image in a new X-window with *ghostview*. Longer sequences may also be retrieved, permitting walkers from multiple, adjacent binding sites and the genomic context of the binding site to be visualized.

The *Delila-Genome* Server

The first step in building *Delila-Genome* was to port the *Delila* individual information programs to the Linux platform. The *Delila* software library is distributed by the National Cancer Institute as binaries for the Sun Sparc system. Source code written in Pascal was

translated to C using *p2c* and debugged.

The main components of the server are the *scan* (from *Delila*), *promotsite*, *scandiff* and *genvis* programs. The server module generally runs directly on top of the *Delila* system however it can be run using a reduced set of *Delila* binaries. Besides the *scan* program, the only *Delila* programs required by *Delila Genome* are *lister*, *mkdb*, and *dbbk* (for displaying sequence walkers). The Delila-Genome server programs are described below.

Scan evaluates the strength (in bits) of each binding site and reports those sites whose strength (R_i) lies within a user defined range [5]. The parameters for scan are defined in the front-end Java program. The minimum threshold R_i value ($R_{i,minimum}$) is set at or above zero bits. Genome scans with an $R_i(b,l)$ matrix derived from a limited number of binding sites, $n \leq 50$ can significantly contribute to Type 1 errors (false positive detection of weak binding sites). To decrease the source of this error, $R_{i,minimum}$ is generally set to the R_i value of the weakest binding site used to compute the weight matrix. Alternatively, sites whose Z scores or probabilities of the binding strengths fall within a user-defined range may be selected. The user also specifies which portion of the individual information matrix is scanned and which strand to evaluate (positive, negative or both). *Scan* can output *data* (locations and strengths of sites), *scanfeatures* (features for display with *lister*) and *scaninst* (instructions for extracting sites as *Delila* book files) files for each chromosome, however, only the *data* file is required as input to the *promotsite* program. Each record in the *data* file contains the R_i values of all predicted binding sites in the genome, their respective coordinates, the Z scores of these R_i values and their

corresponding probabilities. *Scan* has numerous other features, the details of which are presented in [18]. The Z score for user-defined matrices is based upon the mean of the distribution of scores derived from these matrices. The mean is determined first by simulating a set of binding sites based upon this weight matrix (with the *ridi* program [18]) and then computing $R_{sequence}$ from a book of sequences containing these sites with the *encode*, *dalvec* and *rseq* programs (eg. [19]).

Promotsite was developed to filter the output produced by *scan*, since these results may potentially contain large numbers of potential binding sites ($>>10^6$), many of which are distant from expressed sequences. *Promotsite* prunes the *data* file produced by *scan* and reports only relevant sites which are within or adjacent to expressed genomic templates. The user defines a search window either upstream or downstream (or both) relative to the beginning genomic coordinate (often the transcriptional initiation site) of each gene. The upstream and downstream window lengths may be specified independently. *Promotsite* modifies the data file format produced by scan so that the associated GenBank accession number is appended to the record containing the binding site (psdataop file). Typical analyses of splice sites within human coding regions selected sites up to 1 Mb downstream of the transcription initiation site in order to ensure that even the longest genes would be encompassed by these searches. We have limited the analyses of promoters to a 10 kb interval upstream (in some cases, downstream) of the transcription initiation site. However, these parameters should be set (and subsequently optimized) based upon previous experimental or published binding site studies for specific factors. For example, to comprehensively detect insulator elements bound by the protein CTCF,

this window has been specified bi-directionally and increased in length (to 50 kb; not shown).

Since a site may, in some instances, fall within the search window of multiple mRNAs, the mRNA whose start position is closest to the binding site coordinate is assigned to be the associated mRNA for that site. The list of reported binding sites may also be pruned based on a range of chromosomal coordinates and by specifying particular chromosomes. *Promotsite* also defines a parameter known as the *paralog distance*. Since the same mRNA sequence may be mapped based upon its similarity to multiple genomic locations, paralogous genes on the same chromosome designated with the same mRNA accession number were distinguished from large genes containing multiple widely-dispersed exons by defining a parameter for the minimum distance between paralogous loci. Binding sites separated by less than the *paralog distance* are labeled with the same GenBank accession number and are considered part of the same gene, whereas sites exceeding this distance were assumed to be derived from different genes that were similar to the same GenBank accession. Typically, we set the paralog distance to 10^5 or 10^6 bp, depending upon the lengths and density of genes or gene families thought to contain relevant binding sites. Using the associated mRNA for each site, *promotsite* creates a BED-formatted file that can be uploaded as a custom track on the UCSC human genome browser (http://genome.ucsc.edu).

The execution time of *scan* depends on the length of the chromosome and the nucleotide length, l, of the $R_i(b,l)$ weight matrix that defines the binding site. For hardware platforms

with multiple computational nodes, the server can distribute *scan* and *promotsite* runs for each chromosome between these nodes so that the execution time over the whole genome is minimized. As *l* is constant over the whole genome, this load-balancing is based upon the length of each chromosome. Since execution times are generally several hours, the server informs the user of job completion by email.

Relevant binding sites identified with *promotsite* or *scandiff* (see below) can be viewed with the *genvis* program. Like these programs, *genvis* also uses Javascript to generate HTML pages that display the binding site list extracted from the BED files. Since, in some instances, too many sites may be produced by *promotsite* and *scandiff* for browser uploading, *genvis* offers several options to select subsets of binding sites from a chromosome or genome scan. Groups of sites may be extracted by writing subsets of the BED files specified either by genomic strand, the chromosomal coordinates, or a list of accession numbers corresponding to mRNAs mapped onto the genome sequence.

Post-genome scan analysis

Inaccuracies in the genome draft coordinates of splice junction recognition sites motivated the development of an automated strategy to select correctly localized splice sites. Information weight matrices were iteratively recomputed from the set of sites with positive R_i values [6]. More recently, we have built models of transcription factor binding sites by cyclical refinement of weight matrices based on published data from established regulated gene targets, supplemented with binding sites in these genes predicted by information theory and experimentally validated [7, 8]. With *Delila-Genome*, potential novel binding sites identified can be verified in the laboratory and

included in subsequent refinements of the weight matrix.

Previous approaches for comparing information weight matrices have involved determining the Euclidean or positional distances between related $R_i(b,l)$ matrices [10, 20]. Comparisons of the results of successive genome scans offer an alternative approach for monitoring the progress of weight matrix refinement. The *scandiff* program computes model-to-model changes in information at experimentally-proven and predicted binding sites by scanning the same genome sequence with two different information weight matrices. This enables the user to monitor genome-wide sensitivity and specificity of binding site prediction. The psdataop output file generated by *promotsite* is the input to the *scandiff* program. The output files generated by *scandiff* categorize binding sites based upon their identification of unique sets of sites by each of the matrices (models A and B; columns A-B and B-A; Table 1), and sites detected with both weight matrices that show differences in information content (columns A $\cap$ B; in Table 1). *Scandiff* can display differences in binding strength at the same coordinate based upon either exceeding thresholds of absolute changes in R_i (ΔR_i), changes in their respective Z scores (ΔZ) or distinct confidence intervals computed from each of the $R_i(b,l)$ matrices [11].

The criteria of measuring changes in binding site strength is dictated by the stage of model refinement (see below). Absolute comparisons of R_i values are not as meaningful at early stages of refinement, since addition of experimentally-defined binding sites to an information model can substantially alter the distribution of R_i values of the binding sites that underlie these weight matrices. At early stages of refinement, the information

models are based on fewer binding sites, resulting in larger confidence intervals for individual R_i values. Comparisons of R_i values based upon the sizes of confidence intervals are therefore not as reliable measure of significant change in information as changes in their respective Z scores.

Upon model convergence, the proportion of sites in successive models with significant differences in information content should be quite small (*S/[S+I]* (S=significant, I=insignificant) for confidence intervals of ≤ 3 S.D. The proportion of sites common to both models relative to discordant sites found in only one model ([*S+I]* /[*A-B*] + [*B-A*]), should stabilize as successive versions of the information weight matrix are refined.

Scandiff generates BED-formatted files and data files similar in format to that produced by *promotsite* from the identified and categorized binding sites. We used the following color shading convention for the different types of binding sites. The sites with significant changes in R_i are shaded gray; sites identified only by scanning the first matrix are shaded brown; and sites found only with the second matrix are shaded blue. An example of this output is shown in Figure 3, which indicates the results for PXR/RXRα models 1 and 2 in the vicinity of the *CYP3A4* gene.

Results and Discussion

We tested the *Delila Genome* system by scanning the human genome draft sequence (November, 2002) with information weight matrices developed from human transcription factor binding sites (PXR/RXRα [pregnane-X receptor], NF-kB [p50/p65 heterodimer],

and AHR [aryl hydrocarbon receptor]) and with models of sites required for post-transcriptional processing of heteronuclear RNA (donor and acceptor splice sites, and the SR protein, SC35). All binding site sequences were derived from published studies, and in some instances (PXR/RXRα, NF-kB), supplemented by binding sites validated in our laboratory [7, 8]. The information weight matrices were derived with the *Delila* system using previously established procedures [19].

Performance metrics

Table 2 indicates the execution times of complete genome scans for various types of binding sites on two different Linux hardware platforms: a Beowulf cluster of three dual 1.1 Ghz CPU nodes running the Scyld operating system and a Mosix cluster of 24 single processor 500 Mhz nodes. Due to limitations in disk storage, Scyld Beowulf cluster was used to genome scans with PXR/RXRα matrix only. The execution times given in Table 2 represent combined results of running of both *scan* on the genome sequence and *promotsite* on the results of the *scan* program. The execution time for both programs depends upon the length of the binding site, $R_{sequence}$ of the weight matrix, and $R_{i,minimum}$ (specified by the user). The length of the site contributes to the CPU time, and the last two factors contribute to the I/O access time. From the table, we can see that for successive models of PXR/RXRα, $R_{sequence}$ decreases, and consequentially, the number of sites predicted, increases. Additional novel sites that are predicted by information analysis and validated by laboratory testing are introduced with each successive model. The additional sites in the model account for the decrease in $R_{sequence}$, and the increase in

the number of predicted sites in the genome. $R_{sequence}$ decreases from 17 bits to 14.9 bits from models 2 to 3, and there is a steep rise (more than a 2 fold increase) in the number of sites. With the addition of (somewhat weaker) binding sites to model 3, this resultant matrix is less biased towards the consensus sequence, resulting in a large genome-wide increase in predicted sites. The median execution times in the Mosix cluster were approximately 6.5 hrs and 3.5 hrs for the Scyld cluster for all PXR/RXRα models, despite an increase of 3.5 fold in the number of sites from models 1 to 4. The effect of increased I/O access time on the total execution time is evident in the case of the SR protein SC35 site (which has a low $R_{sequence}$ value of 3.64 bits), where the run time is 19 hours due to 76-fold increase in the quantity of sites predicted compared with the scan of PXR/RXRα Model 4.

Analysis of the splice acceptor and donor runs required a modification of the published genome sequence. In the original genome drafts, a very large number of binding sites ($>>10^8$) were initially found. Many of these sites were composed of long runs of undefined polynucleotides (ie. $\geq N_{(10)}$) in heterochromatin and in gaps in the draft sequence. The *Delila* program defaults to adenine in these cases, and in the case of splice acceptor sites, these substitutions generated sites comprised of polyadenine, which itself has an R_i value exceeding the user-defined threshold (2.4 bits; $R_{i,minimum}$). These runs exceeded our available disk storage, and to reduce the quantity of false positive sites, we generated and substituted random nucleotides for every sequence of undefined polynucleotides ≥10 bp in length. Our previous studies have shown that sequence randomization produces fewer than 2% of binding sites with R_i values above zero bits [6],

and none above the minimum R_i threshold value [11]. The genome scans of the substituted genome sequences with splice donor and acceptor $R_i(b,l)$ weight matrices were completed in 10.5 and 14.5 hours, respectively.

Visualization of binding sites in subgenomic intervals

We have found that uploads of large BED files of binding sites to the remote UCSC genome browser can be time-consuming and sometimes fail. The BED file for all binding sites found with PXR/RXRα Model 4, for example, is ~30 MB and required 5-10 minutes to upload. Furthermore, the large numbers of sites found with some information weight matrices (eg. splice donor and acceptor sites; 254 Mb for acceptor sites on chromosome 1 alone) produce BED file sizes exceeding browser/server limits. We therefore created and viewed subsets of binding sites for genomic regions of specific interest with the *genvis* tool.

Figure 2 depicts the HTML page generated by *genvis*, containing a partial list of binding sites on chromosome 1 for the PXR/RXRα model 4 $R_i(b,l)$ weight matrix. The websites linked to this page are also shown (but have been resized or truncated) to reflect only the important details of each. When the HTML page is initially loaded, a window for the UCSC browser pops up. The BED file is uploaded using a button in this window upon selecting the appropriate version of the genome draft at the UCSC website. When the genome browser target links (entries in the R_i, Seq and UCSC Browser columns) are activated, the genome browser displays the information based on this uploaded file.

The second row of the HTML table in Figure 2 corresponds to the binding site associated with the GenBank Accession L13278. This is a strong binding site (R_i value of ~20.1 bits) which is hyper-linked to the custom track detail in the genome browser. This track detail page indicates the size of the site and the orientation of the recognition sequence on the draft genome sequence. The user can obtain the DNA sequence of the site either from from the *Seq* cell in the HTML table or from the corresponding custom track detail. The pop up sequence walker indicates the relative contributions of each nucleotide in the site [5].

The linked GenBank and SOURCE database entries indicate that accession L13278 encodes the zeta-crystallin/quinone reductase gene. We selected this example to illustrate that *Delila-Genome* can be used to potentially discover novel transcriptional regulatory targets, since this gene has not been previously demonstrated to be regulated by PXR/RXRα. The SOURCE entry is based on a dynamic collection and compilation of gene data from many scientific databases associated with the GenBank accession, whereas the GenBank entry, in some instances, is not curated and guaranteed only to contain the corresponding sequence. The SOURCE entry also indicates other information such as the aliases for the gene name, the locus link designation, expression profile, etc.

The UCSC genome browser entry displays the binding site custom track and sequences in the proximity of the associated GenBank accession. The coordinates delineate a display window concordant with the search window defined in *promotsite* for generating the list of binding sites given in the HTML page. In Figure 2, the predicted site is 1112 bp

upstream of L13278 and ~7.2 kb upstream of an as yet uncharacterized gene corresponding to both AK098237 and BC009514. Although we cannot exclude the possibility that this site regulates the gene encoded by AK098237/BC009514, its closer proximity to the zeta-crystallin gene and the common orientation of both the site and gene on the antisense strand suggests that this site may function as a potential transcriptional enhancer element. There are no other predicted binding sites in the vicinity of this gene.

Comparison of genome scans produced from successive transcription factor information weight matrices

The results of genome scans with successive refinements of PXR/RXRα information weight matrices were compared using *scandiff*. The refinement procedure was validated by detecting binding sites in well-established PXR/RXRα target genes. Initial models based on published sites were used to scan target genes that were known to be induced by PXR/RXRα binding, but where additional sites had not been previously identified. Sites detected in these scans were assayed for binding to PXR/RXRα and those found to bind were incorporated in subsequent rounds of refinement.

The *genvis* program was used to display *scandiff* results for *CYP3A4*,which is a single gene known to be regulated by PXR/RXRα (Figure 3). BED custom tracks of this gene for scans of the initial and second PXR/RXRα models (1 and 2) are indicated. Both information models recognize experimentally-verified binding sites [21, 22]: a strong,

potential proximal enhancer binding site (custom track M18907_pxr_R17) 204 bp upstream of the transcription initiation site and a cluster of distal enhancer elements 7.2-7.8 kb upstream. Model 1 identified a 7 bit site (AF182273_pxr_R7) in the first intron, which is absent in the scan of model 2. However, model 2 also identifies an additional site (M18907_pxr_R7) within the distal enhancer cluster, which is consistent with the possibility that Model 2 more specifically recognizes promoter binding sites. Similar results were obtained confirming detection of experimentally-defined binding sites in the promoters of other PXR/RXRα regulated genes (*CYP3A7, CYP2B6*; results not shown) induced by this transcription factor.

Scandiff also produces a summary statistics file which can be used to monitor the progress of information theory-based model refinement. The following example indicates how the results of complete genome scans with four successive PXR/RXRα $R_i(b,l)$ matrices can be interpreted from these summaries (each successive model is based on increasing numbers of experimentally validated binding sites; Table 1). The tables indicate the differences in the number of predicted binding sites in each category of these models. By selecting high thresholds for either ΔR_i values, ΔZ scores or confidence intervals, it is possible to identify binding sites with the most significant model-to-model changes. The following analysis is based on changes in information content of at least 3 bits (ΔR_i), Z score differences of $\geq$1, and confidence intervals $\geq$ 3 standard deviations, ie. 95%.

Newly identified sites (B-A) predicted with model 2 are 3.8 fold more abundant than

those found only with model 1. Scanning the genome with model 3 (vs. model 2) resulted in an even greater disproportionate distribution of unique sites (9.2 fold). This trend continues in model 4, but the fraction of novel binding sites is decreased (2.4 fold). The findings indicate that increasing the diversity of the sequences underlying the matrix affects which binding sites are found in the genome scan. It is apparent that the PXR/RXRα weight matrix has not converged, since large numbers of novel sites continue to be found with successive information models.

Only a modest fraction of sites (*S/[S+I];* S=significant, I=insignificant) exhibit the largest significant changes in binding site strength ($\Delta R_i \geq 3$ bits; ranging from 3-11%), regardless of which pair of scans are analyzed. Most changes in information content are ≤2 bits. As ΔR_i values give no indication of the strengths of the sites that have changed (only the magnitude of those changes), we also cataloged significant changes by comparing the Z scores of the same binding sites found by successive models. The most stringent test ($\Delta Z \geq 1$) revealed that the transition from model 2 to model 3 produced the largest proportion of significant changes (48% of sites; n = 48,657), in comparison with more modest changes in Z score from models 1 to 2 (0.8%) and models 3 to 4 (2.5%). We interpret these results to indicate that model 3 may have altered the strengths of binding sites at outlying R_i values to a greater extent than the transitions either from models 1 to 2 or from models 3 to 4.

Binding sites that are added to the models in subsequent rounds of experimental refinement have increasingly diverse sequences, resulting in lower measures of $R_{sequence}$

and therefore detect additional predicted sites. Shorter binding sites, such as those recognized by AHR, with lower $R_{sequence}$ values, are predicted to be even more abundant. The vast majority of the newly detected binding sites are considered "weak" ($R_i << R_{sequence}$; Table 2). The lower threshold R_i value of binding sites reported by *scan* is typically set to the strength of the weakest binding site used to define the information weight matrix. The confidence intervals on binding sites with low R_i values are still fairly large [see Appendix to reference 11], and some of these sites may turn out to have $R_i < 0$ bits. In any case, the affinities for sites with low R_i values, especially those $\sim R_{i,minimum}$ are likely to be negligible and may not be detectable experimentally [6]. Nevertheless, the increased sequence diversity introduced by these refinement procedures augments the dynamic range of site binding strengths found with later versions of refined models. The increased sequence diversity affects the frequencies of the nucleotides underlying the weight matrix and can significantly alter the information contents of predicted "strong" sites [9].

Additional gene promoters are found with successive PXR/RXRα models (Table 1). In each pairwise comparison of information models, novel binding sites detected by the later model substantially outnumbered unique sites found only by the earlier model (by 4 to 11.2 fold). Nevertheless, it is encouraging that the increased number of genes containing these binding sites does not proportionally increase with the numbers of binding sites, which suggests that the subsequent models are predicting additional sites in the same genes. This is not surprising, since multiple PXR/RXRα enhancer binding elements with "moderate-to-strong" R_i values have been documented in known targets of this

transcription factor, including several *CYP3A* gene family members. We examined the distributions of such sites in genome scans of promoters with the different PXR/RXRα weight matrices.

The "moderate-to-strong" binding sites in the genome-wide promoter scans ($R_i > R_{sequence}$; Table 2) are a small percentage of all sites detected (0.06 % in Model 1, increasing to 0.5 % in Model 4). The refinement procedure may improve the sensitivity of detecting such sites. PXR/RXRα models 1 and 2 actually detect *fewer* of these sites in gene promoters (and genes) than the numbers of genes thatexhi bit changes in expression by microarray studies [21, 22], suggesting that these models predict fewer binding sites, and consequently fewer target genes than expected. In subsequent models, increasingly higher frequencies of multiplex sites are found in the same promoters (8% in Model 1 versus 16% in Model 4). This degree of redundancy (in Model 4) substantially exceeds the expected frequency of promoters with multiple binding sites, and the information required to find these sites in the genome ($R_{frequency}$ ~4 bits). We also find that multiplex binding sites within promoters recognized by transcription factors with smaller footprints are considerably more frequent (NF-κB p50/p65 and AHR), as expected from their lower $R_{sequence}$ values.

Conclusions

Delila-Genome can be used to scan eukaryotic genomes with information theory-based models for transcription factor and post-transcriptional protein binding sites and displays the most relevant sites. Complete scans of human genome draft sequences with

information-weight matrices of transcription factor binding sites (PXR/RXRα, AHR and NF-κB p50/p65) and sequences required for mRNA splicing (donor, acceptor, and SC35 splicing enhancer protein binding sites) were completed within several hours on small Linux clusters. Binding sites can be visualized at either high or low sequence resolution juxtaposed with other genome annotation. The software can also be used to compare the distributions of predicted sites in multiple or successive binding site models. Refinement of successive binding site models should enable more accurate and specific predictions of site strength, which in turn, may facilitate discovery of novel regulatory gene targets and assist in the prediction of mRNA splicing patterns.

Availability and Requirements

- **Project Name:** Delila-Genome
- **Project Home Page:** http://www.sice.umkc.edu/~roganp/Information/delgen.html
- **Operating System(s):**

  Server - Linux; can be ported to Unix/Solaris with little or no modification.

  Client [Front end] – Any system with JRE (Java Runtime Environment) 1.4 or higher installed
- **Programming Language:**

  Server - Perl, Pascal, C/C++, Bash shell scripts, Javascript

  Client [Front end] – Java
- **Other requirements:**.Individual information program package (for details, see http://www.lecb.ncifcrf.gov/~toms/walker/iipp.html)

- **License**: Delila-Genome is deposited at www.bioinformatics.org under GNU GPL. The Individual Information programs are available from the National Cancer Institute via transfer agreements **(**see http://www.lecb.ncifcrf.gov/~toms/contacts.html). Linux binaries and the source code of the *Delila* programs are available to NCI-authorized users from the authors.
- **Any restrictions to use by non-academics:** None

Authors' contributions

PKR developed and implemented the model refinement procedures and designed the *Delila-Genome* system. SG implemented the *Delila-Genome* architecture and wrote the code. SG and PKR have tested the system. PKR and JSL refined the AHR, NF-kB, and PXR/RXRα information models (PXR/RXRα with CAV); PKR developed and refined the splice donor, acceptor and SC35 models. CAV and JSL validated the predicted PXR/RXRα binding sites in the laboratory. All authors have approved the manuscript.

Acknowledgements

This work was sponsored by grant ES 10855 from the National Institute of Environmental Health. We are grateful to Tom Schneider and Joan Knoll for their valuable comments on the manuscript. We thank Information Services at the University of Missouri-Kansas City for access to the Mosix cluster.

References

1. Schneider TD, Stormo GD, Haemer JS, Gold L: **A design for computer nucleic-acid-sequence storage, retrieval, and manipulation**. *Nucleic Acids Res* 1982, **10**:3013-24.
2. Shannon CE: **A mathematical theory of communication**. *Bell System Technical Journal* 1948, **27**:379-423 and 623-656.
3. Schneider TD: **Sequence logos, machine/ channel capacity, Maxwell's demon, and molecular computers: a review of the theory of molecular machines**. *Nanotechnology* 1994, **5**:1-18.
4. Schneider TD: **Information content of individual genetic sequences**. *J Theor Biol* 1997, **189**:427-41.
5. Schneider TD: **Sequence walkers: a graphical method to display how binding proteins interact with DNA or RNA sequences**. *Nucleic Acids Res* 1997, **25**:4408-15.
6. Rogan PK, Faux BM, Schneider TD: **Information analysis of human splice site mutations**. *Hum Mutat* 1998, **12**:153-71.
7. Hurwitz I, Svojanovsky S, Leeder JS, Rogan PK: **Modeling differential binding of NF-kB p50 to a CYP2D6 promotor variant by information theory [abstract]**. *American Journal of Human Genetics* 2001, **69**:s476.
8. Rogan PK, Svojanovsky S, Hurwitz I, Schneider TD, Leeder JS: **Modeling splice site and transcription factor binding site variation by information theory [abstract]**. *American Journal of Human Genetics* 2002, **71**:s333.
9. Vyhlidal CA, Rogan PK, Leeder JS: **Modeling PXR/RXR Binding Using**

Information Theory [abstract]. *7th Annual Meeting of the International Society for Study of Xenobiotics* 2002.

10. Shultzaberger RK, Bucheimer RE, Rudd KE, Schneider TD: **Anatomy of Escherichia coli ribosome binding sites**. *J Mol Biol* 2001, **313**:215-28.

11. Rogan PK, Svojanovsky S, Leeder JS: **Information theory-based analysis of CYP2C19, CYP2D6 and CYP3A5 splicing mutations**. *Pharmacogenetics* 2003, **13**:207-18.

12. Hengen PN, Bartram SL, Stewart LE, Schneider TD: **Information analysis of Fis binding sites**. *Nucleic Acids Res* 1997, **25**:4994-5002.

13. Zheng M, Doan B, Schneider TD, Storz G: **OxyR and SoxRS regulation of fur**. *J Bacteriol* 1999, **181**:4639-43.

14. Berman BP, Nibu Y, Pfeiffer BD, Tomancak P, Celniker SE, Levine M, Rubin GM, Eisen MB: **Exploiting transcription factor binding site clustering to identify cis-regulatory modules involved in pattern formation in the Drosophila genome**. *Proc Natl Acad Sci U S A* 2002, **99**:757-62.

15. Rebeiz M, Reeves NL, Posakony JW: **SCORE: a computational approach to the identification of cis-regulatory modules and target genes in whole-genome sequence data. Site clustering over random expectation**. *Proc Natl Acad Sci U S A* 2002, **99**:9888-93.

16. Lund M, Tange TO, Dyhr-Mikkelsen H, Hansen J, Kjems J: **Characterization of human RNA splice signals by iterative functional selection of splice sites**. *RNA* 2000, **6**:528-44.

17. Shultzaberger RK, Schneider TD: **Using sequence logos and information**

analysis of Lrp DNA binding sites to investigate discrepanciesbetween natural selection and SELEX. *Nucleic Acids Res* 1999, **27**:882-7.

18. Schneider TD: **Delila programs documentation.** *http://www.lecb.ncifcrf.gov/~toms/delila/delilaprograms.html* 2003.

19. Stephens RM, Schneider TD: **Features of spliceosome evolution and function inferred from an analysis of the information at human splice sites**. *J Mol Biol* 1992, **228**:1124-36.

20. Schneider TD: **Measuring molecular information**. *J Theor Biol* 1999, **201**:87-92.

21. Gerhold D, Lu M, Xu J, Austin C, Caskey CT, Rushmore T: **Monitoring expression of genes involved in drug metabolism and toxicology using DNA microarrays**. *Physiol Genomics* 2001, **5**:161-70.

22. Rae JM, Johnson MD, Lippman ME, Flockhart DA: **Rifampin is a selective, pleiotropic inducer of drug metabolism genes in human hepatocytes: studies with cDNA and oligonucleotide expression arrays**. *J Pharmacol Exp Ther* 2001, **299**:849-57.

Figure Legends

Figure 1. Architecture of the *Delila Genome* system. Server programs are shown on the right side of the schema and client programs shown on the left side. A Java-based GUI application (*Delgenfront*) is run on a desktop client that prompts entry of a series of

parameters (server, results directory, genome draft, email address) and the location of ribl file or entry of a weight matrix. These data are sent to a Linux server which runs the *scan* and *promotsite* programs to display predicted binding sites. The *scan* and *promotsite* jobs may be submitted individually or sequentially. Since *scan* operates on *Delila*books, scripts have been provided to automate the downloading and build *Delila* books of the genome drafts from UCSC (documented in the package: Readme.txt). The *genvis* program uses the results of previous chromosome or genome analyses with *scan* and *promotsite* to generate BED and HTML files of predicted binding sites within a user-defined genomic interval. Upon opening the HTML page, the user uploads the BED file to the corresponding version of the UCSC genome browser, which then displays the custom binding site track of the interval containing the site juxtaposed with other genome annotations. The HTML page is also hyperlinked to the binding site sequence (which can be used to generate a sequence walker using the *autolist* script), details of the binding site location, and the GenBank and SOURCE entries of the transcript associated with the site. Results obtained with different information matrices can be compared with the *scandiff* program, which generates BED files for binding sites found with each of the matrices and summary output indicating these differences. While *promotsite* takes input parameters in a file, all other *Delila-Genome* programs have command line options to specify the required and optional parameters and most support an '-h' switch that displays these options.

Figure 2. Screen shot of results generated by *Delila-Genome* visualization tools. This example shows predicted PXR/RXRα binding sites at the zeta crystalline locus.

Genome-wide HTML and BED files have been generated by the *promotsite* program. Sites are in the HTML ordered by information content. Hyperlinked pages (arrows from *Delila-Genome* HTML page) reveal details about binding sites and annotations of the gene associated with the binding site. Panels indicate: (A) *Delila Genome* HTML page for viewing sorted binding sites with associated genes; (B) UCSC browser custom track detail for specific binding site; (C) Sequence of binding site; (D) Sequence walker of the binding site (computed on the server and displayed on client running X-windows); (E) GenBank entry for mRNA accession number associated with binding site (F) Stanford SOURCE database entry providing current information about gene template of GenBank mRNA accession (G) UCSC browser for viewing sites in the gene associated with the GenBank accession.

Figure 3. Screen shot of UCSC Genome Browser indicating binding sites found in genome scans using different information weight matrices. Binding sites in the promoter of the *CYP3A4* gene found with PXR/RXRα weight matrices are indicated by color-coded custom tracks. Sites uniquely identified with the weight matrices from Models 1 and 2 are respectively indicated with brown and blue tracks. The grey track shows binding sites with significantly different binding strengths that were identified by scanning with both of the matrices. The Custom tracks were generated by the *scandiff* program and uploaded to the Genome Browser.

Table 1: Total binding site counts based on genome scans of promoters with PXR/RXRα information weight matrices

Models Compared		Numbers of sites in each category										
		Unique sites		Z scores			R_i			Confidence intervals[+]		
A	**B**	**A-B***	**B-A^**	**Threshold (ΔZ)**	**(A ∩ B), $S^{\sim}$**	**(A ∩ B), $I^{@}$**	**Threshold (ΔR_i, bits)**	**(A ∩ B), *S***	**(A ∩ B), *I***	**Threshold (±S.D.)**	**(A ∩ B), *S***	**(A ∩ B), *I***
1	2	11758	45219	0.5	27945	44302	1	29378	42869	1	30982	41265
				0.75	7492	64755	2	9080	63167	2	26931	45316
				1.0	589	71658	3	2293	69954	3	23625	48622
2	3	17065	157922	0.5	90459	9942	1	54426	45975	1	55431	44970
				0.75	73309	27092	2	26038	74363	2	45069	55332
				1.0	48657	51744	3	11044	89357	3	37822	62579
3	4	61906	148894	0.5	54586	141831	1	93585	102832	1	104397	92020
				0.75	17891	178526	2	33843	162574	2	80088	116329
				1.0	5044	191373	3	11069	185348	3	68846	127571

[+] Standard error computation for individual R_i values is based on derivation given in reference 11; *Sites found with model A but not with model B; ^sites found with model B, but not with model A; $^{\sim}$ Number of sites with differences in R_i values exceeding threshold Z scores; $^{@}$Number of sites with differences in R_i values less than threshold.

Table 2: Performance metrics for genome scans

Site	Length	Weight matrix version	Num. sites in Model	$R_{i,min}$	R_{seq}	Execution time (hrs)*		Number of sites found ^			
						Mosix	Scyld	$R_i \geq R_{i,min}$	$R_i \geq R_{Seq}$	Unique Promoters with $R_i \geq R_{seq}$	Promoters with multiple sites (%) $R_i \geq R_{seq}$
PXR	23	1	15	7.1	17.1	6.5	4.3	3.48e5	218	200	8.3
PXR	23	2	19	7.1	17.0	6	3.5	4.97e5	391	365	6.6
PXR	23	3	32	7.1	14.9	7.1	4	1.10e6	3393	3036	10.5
PXR	23	4	48	7.1	14.4	6.8	3.8	1.44e6	7694	6439	16.3
NF-κB	10	3	75	2.6	10.9	5.8	-	1.16e7	74050	33340	54.9
AHR	17	1	30	2.8	9.4	6.3	-	1.20e7	42487	24764	41.7
Acc	28	12	1.08e5	2.4	7.4	14.5	-	4.87e7	-	-	-
Don	7	5	1.11e5	2.4	6.7	10.5	-	4.85e7	-	-	-
SC35	8	1	30	0.4	3.6	19	-	1.07e8	-	-	-

Abbreviations. Site: Binding site information matrix; PXR: PXR/RXRα; NF-κB: NF-κB p50/p65 subunits; Acc: Splice Acceptor; Don: Splice Donor; Length: Length of the site in nucleotides; $R_{i,min}$: $R_{i,minimum}$ (in bits); R_{Seq} : $R_{sequence}$ (in bits)

*****total runtime for both *scan* and *promotsite* programs

^Results of information analysis with the PXR/RXRα, NF-kB and AHR matrices of promoter regions (10 kb upstream of transcription initiation site) for all transcripts mapped in reference genome sequence. Complete gene sequences (from the transcription initiation site to the terminal sequence of the 3' UTR) were analyzed with the Acc, Don and SC35 matrices.

Descriptions of additional data files

A package of *Delila-Genome* software and documentation and *Delila* books of the human genome sequence assembly (April 2003) are available at http://www.sice.umkc.edu/~roganp/Information/delgen.html. Examples of HTML pages produced by *Delila Genome* with corresponding BED custom tracks can also be downloaded from this website.

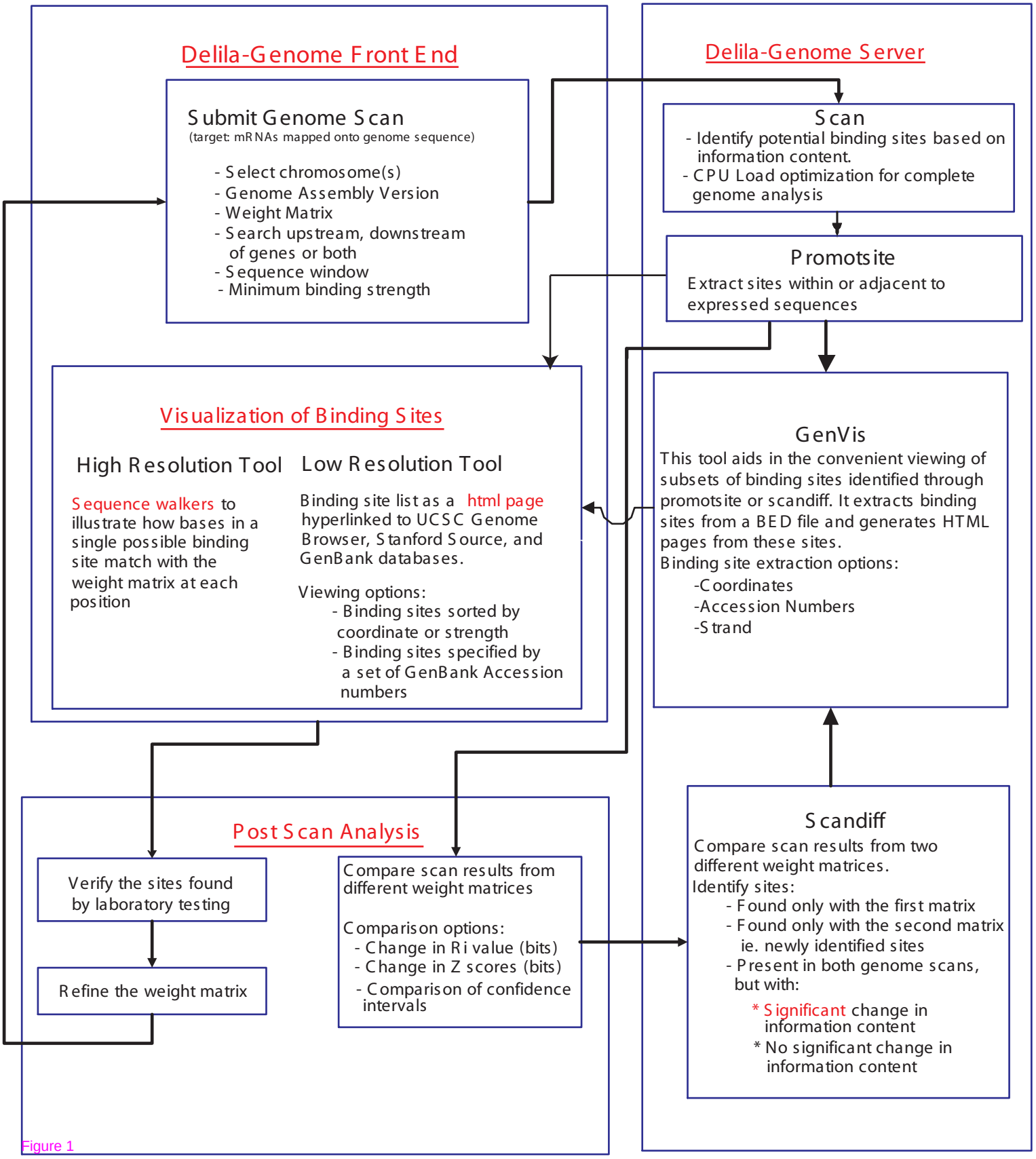
Delila-Genome Front End
Submit Genome Scan
(target: mRNAs mapped onto genome sequence)
- Select chromosome(s)
- Genome Assembly Version
- Weight Matrix
- Search upstream, downstream of genes or both
- Sequence window
- Minimum binding strength
Delila-Genome Server
Scan
- Identify potential binding sites based on information content.
- CPU Load optimization for complete genome analysis
Promotsite
Extract sites within or adjacent to expressed sequences
Visualization of Binding Sites
High Resolution Tool
Sequence walkers to illustrate how bases in a single possible binding site match with the weight matrix at each position
Low Resolution Tool
Binding site list as a html page hyperlinked to UCSC Genome Browser, Stanford Source, and GenBank databases.
Viewing options:
- Binding sites sorted by coordinate or strength
- Binding sites specified by a set of GenBank Accession numbers
GenVis
This tool aids in the convenient viewing of subsets of binding sites identified through promotsite or scandiff. It extracts binding sites from a BED file and generates HTML pages from these sites.
Binding site extraction options:
-Coordinates
-Accession Numbers
-Strand
Post Scan Analysis
Verify the sites found by laboratory testing
Refine the weight matrix
Compare scan results from different weight matrices
Comparison options:
- Change in Ri value (bits)
- Change in Z scores (bits)
- Comparison of confidence intervals
Scandiff
Compare scan results from two different weight matrices.
Identify sites:
- Found only with the first matrix
- Found only with the second matrix ie. newly identified sites
- Present in both genome scans, but with:
* Significant change in information content
* No significant change in information content

Figure 1

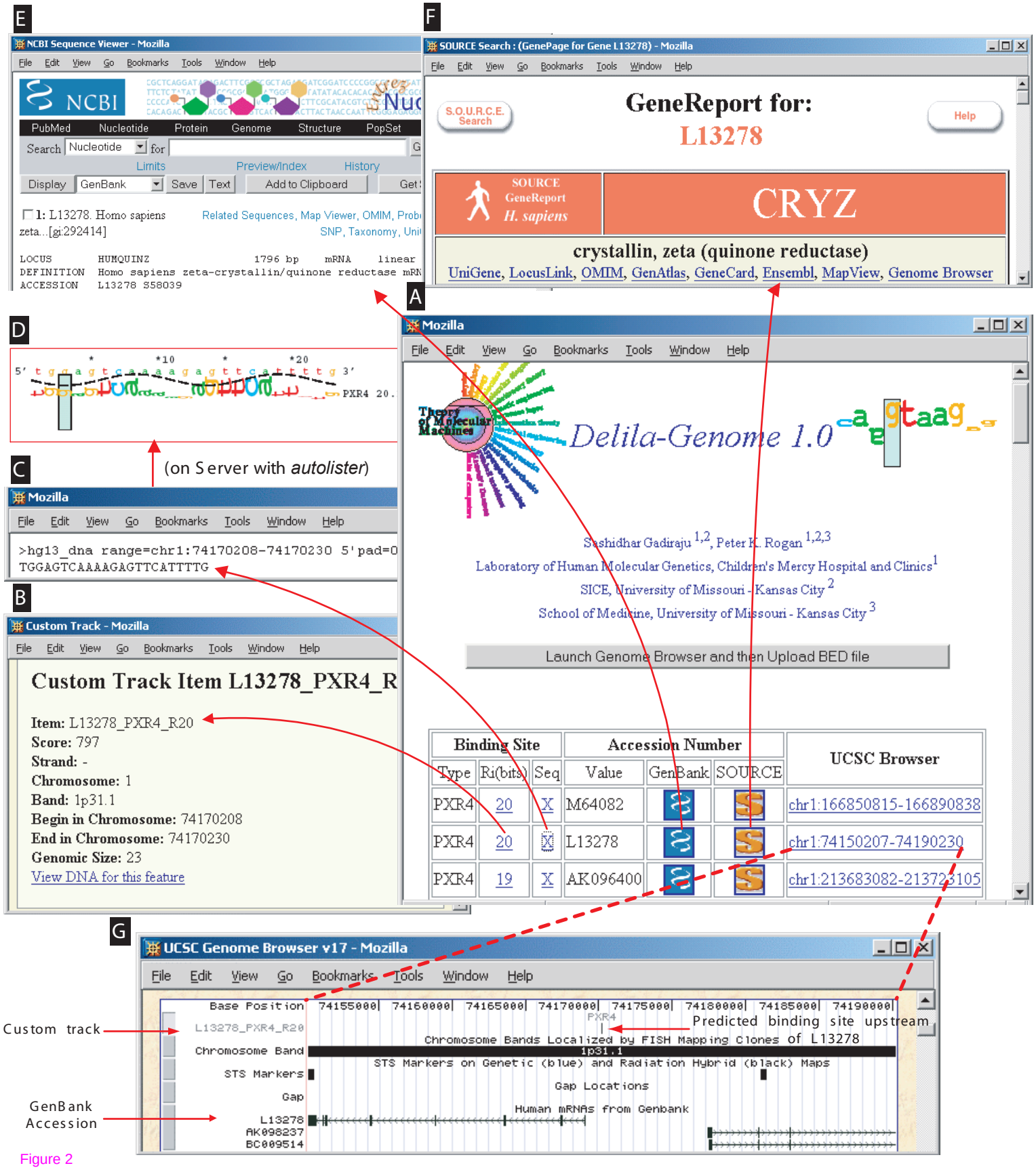
E
NCBI Sequence Viewer - Mozilla
File Edit View Go Bookmarks Tools Window Help
NCBI
PubMed Nucleotide Protein Genome Structure PopSet
Search Nucleotide for
Limits Preview/Index History
Display GenBank Save Text Add to Clipboard
1: L13278. Homo sapiens zeta...[gi:292414]
Related Sequences, Map Viewer, OMIM, Probe
SNP, Taxonomy, Uni
LOCUS HUMQUINZ 1796 bp mRNA linear
DEFINITION Homo sapiens zeta-crystallin/quinone reductase mRN
ACCESSION L13278 S58039
F
SOURCE Search : (GenePage for Gene L13278) - Mozilla
File Edit View Go Bookmarks Tools Window Help
S.O.U.R.C.E. Search
GeneReport for:
L13278
Help
SOURCE GeneReport H. sapiens
CRYZ
crystallin, zeta (quinone reductase)
UniGene, LocusLink, OMIM, GenAtlas, GeneCard, Ensembl, MapView, Genome Browser
A
Mozilla
File Edit View Go Bookmarks Tools Window Help
Theory of Molecular Machines
Delila-Genome 1.0
Sashidhar Gadiraju [1,2], Peter K. Rogan [1,2,3]
Laboratory of Human Molecular Genetics, Children's Mercy Hospital and Clinics [1]
SICE, University of Missouri - Kansas City [2]
School of Medicine, University of Missouri - Kansas City [3]
Launch Genome Browser and then Upload BED file
Binding Site
Accession Number
UCSC Browser
Type Ri(bits) Seq Value GenBank SOURCE
PXR4 20 X M64082 chr1:166850815-166890838
PXR4 20 X L13278 chr1:74150207-74190230
PXR4 19 X AK096400 chr1:213683082-213723105
D
5' t g g a g t c a a a a g a g t t c a t t t g 3'
PXR4 20.
(on Server with autolister)
C
Mozilla
File Edit View Go Bookmarks Tools Window Help
>hg13_dna range=chr1:74170208-74170230 5'pad=0
TGGAGTCAAAAGAGTTCATTTTG
B
Custom Track - Mozilla
File Edit View Go Bookmarks Tools Window Help
Custom Track Item L13278_PXR4_R
Item: L13278_PXR4_R20
Score: 797
Strand: -
Chromosome: 1
Band: 1p31.1
Begin in Chromosome: 74170208
End in Chromosome: 74170230
Genomic Size: 23
View DNA for this feature
G
UCSC Genome Browser v17 - Mozilla
File Edit View Go Bookmarks Tools Window Help
Base Position 74155000 74160000 74165000 74170000 74175000 74180000 74185000 74190000
L13278_PXR4_R20
PXR4
Predicted binding site upstream
Chromosome Band
Chromosome Bands Localized by FISH Mapping Clones of L13278
1p31.1
STS Markers
STS Markers on Genetic (blue) and Radiation Hybrid (black) Maps
Gap
Gap Locations
Human mRNAs from Genbank
L13278
AK098237
BC009514
Custom track
GenBank Accession

Figure 2

Significantly different Ri values, sites present in both models

Model 1 only

Model 2 only

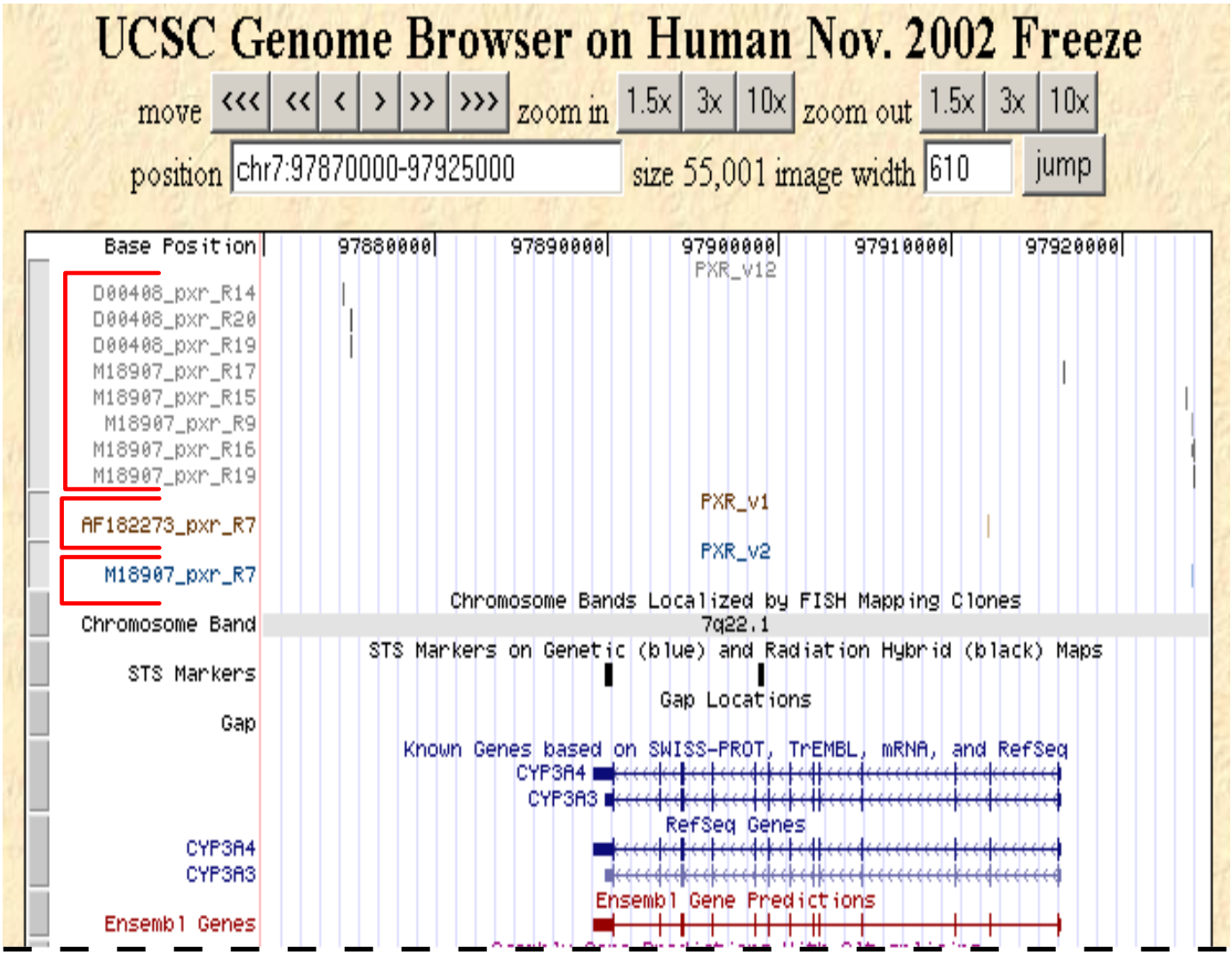

Figure 3