

The PAZAR database of gene regulatory information coupled to the ORCA toolkit for the study of regulatory sequences

Elodie Portales-Casamar¹, David Arenillas¹, Jonathan Lim¹,
Magdalena I. Swanson¹, Steven Jiang¹, Anthony McCallum¹, Stefan Kirov²
and Wyeth W. Wasserman^{1,*}

¹Centre for Molecular Medicine and Therapeutics, Child and Family Research Institute, Department of Medical Genetics, University of British Columbia, Vancouver, BC, Canada and ²Applied Genomics Department, Pharmaceutical Research Institute, Bristol-Myers Squibb, NJ, USA

Received September 9, 2008; Revised and Accepted October 9, 2008

ABSTRACT

The PAZAR database unites independently created and maintained data collections of transcription factor and regulatory sequence annotation. The flexible PAZAR schema permits the representation of diverse information derived from experiments ranging from biochemical protein–DNA binding to cellular reporter gene assays. Data collections can be made available to the public, or restricted to specific system users. The data ‘boutiques’ within the shopping-mall-inspired system facilitate the analysis of genomics data and the creation of predictive models of gene regulation. Since its initial release, PAZAR has grown in terms of data, features and through the addition of an associated package of software tools called the ORCA toolkit (ORCAtk). ORCAtk allows users to rapidly develop analyses based on the information stored in the PAZAR system. PAZAR is available at <http://www.pazar.info>. ORCAtk can be accessed through convenient buttons located in the PAZAR pages or via our website at <http://www.cisreg.ca/ORCAtk>.

INTRODUCTION

A robust and growing community of life scientists studies the regulatory mechanisms governing gene expression. The cornerstones of genome sequences, high-throughput experimental technologies and computers have been laid, allowing research labs of both small and large size to generate large collections of data pertaining to the *cis*-regulatory sequences and *trans*-acting factors that

govern transcription. With the study of stem cells bringing increased focus on the programs of gene expression during developmental stages, the generation of new data is likely to continue at a dramatic rate of increase. Researchers therefore face two important data needs: a convenient means to share and manage their gene regulation data and a reliable resource of reference data. We created the PAZAR system to meet the need for an open-access repository of *cis*- and *trans*-acting gene regulation data.

The PAZAR system is predicated upon the concept of individual data providers using a shared computational infrastructure to deliver information to the community (1). Briefly, PAZAR provides a computing infrastructure for the creation, maintenance and dissemination of regulatory sequence and transcription factor (TF) annotation. Included is a web interface, an application programming interface (API) for programmatic access, and a highly flexible database schema to accommodate heterogeneous datasets. Regulatory sequences are linked to specific genomic coordinates which are updated automatically with new releases of sequence assemblies. Working with the research community, the PAZAR system includes information from focused resources such as the JASPAR database (2) and the ORegAnno collection (3). In addition, the system allows for the creation of boutique datasets by individual users. Boutique operators have the choice to limit access to a data collection, for instance to facilitate collaborative projects, or to make their data available to the public.

In this report, we describe the growth of the PAZAR system. In particular, we discuss the implementation of improved data exchange mechanisms and the expanded utility of the PAZAR system through the integration of the ORCA toolkit (ORCAtk), a new software package

*To whom correspondence should be addressed. Tel: +1 (604) 875 3812; Fax: +1 (604) 875 3819; Email: wyeth@emmt.ubc.ca

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

developed in support of diverse regulatory sequence analysis projects.

MATERIALS AND METHODS

PAZAR Data Exchange

The PAZAR XML exporter tools were developed in Perl and make extensive use of the PAZAR API. All codes are available on the Sourceforge CVS repository (<http://pazar.cvs.sourceforge.net/pazar/>).

ORCAtk

ORCAtk utilizes and builds upon existing bioinformatics software tools including BioPerl (4), BLAST (5), the TFBS modules (6), Jim Kent's UCSC software (7) and the Ensembl API (8), as well as incorporating stand-alone software code for sequence alignment and reconciliation of sequence features along an alignment. Its design facilitates code reuse through the implementation of discrete modules which communicate via standard Perl/BioPerl objects, allowing power users to integrate individual functions into their own customized solutions. For the end-user, a flexible command line script and a user-friendly web-interface provide mechanisms for performing complete end-to-end analysis.

Sequence alignment module

The alignment module, implemented in Perl, employs a recursive global progressive alignment algorithm similar to that of LAGAN (9) and AVID (10) in which a local alignment step is used to identify short blocks of very high similarity used to anchor the alignment, followed by a global alignment performed on regions between the anchors. The ORCAtk aligner calls BLAST 2.2.11 (5) to perform the local alignment step. The high-scoring blocks are then chained together to form the maximum scoring chain which maintains colinearity. The sub-sequences anchored by these chained blocks are then aligned using an extension of the Needleman–Wunsch algorithm (11) that allows for affine gap penalties, with one penalty for the opening of a gap and another penalty for gap extensions. This is an improvement over the standard Needleman–Wunsch algorithm which utilizes linear gap penalties that unfairly penalize long gaps. The algorithm is recursive such that if a sub-sequence is too long to be aligned by Needleman–Wunsch (due to memory requirements), the local alignment step is called again on this sub-sequence, using more liberal parameters until a complete alignment is obtained. For maximum speed, the Needleman–Wunsch step has been implemented as a C program. All global and local alignment parameters are fully adjustable by the user but default values have been provided for convenience. The final alignment is returned as a BioPerl Bio::SimpleAlign object, making it easy to incorporate into other BioPerl-based applications. Conversely, the alignment module may be substituted by another alignment program.

Conservation analysis module

Once an alignment is obtained, conserved non-coding regions are identified at a specified conservation cutoff. Two conservation analysis Perl software modules are used to perform phylogenetic footprinting. The user may choose from a species pair-wise conservation analysis module or a multiple species phastCons (12) module. All conservation analysis parameters for both the pair-wise and phastCons modules are specifiable by the user, with default values automatically applied if not provided by the user.

The pair-wise conservation analysis module computes a conservation profile by scanning the alignment generated by the alignment module described above with a sliding window of user-specified fixed size. The percentage identity of the sequence is calculated at each nucleotide position along the alignment. Conserved regions are then computed by merging windows that score above a given conservation cutoff to form maximal spanning regions. The cutoff applied is controlled by two user-selected parameters—the top percentile of scoring windows and the minimum percentage identity. If only the latter is specified, then this is used as the cutoff. If only the former is specified, then the cutoff is computed dynamically by calculating a percentage identity such that the given percentiles of initial windows score above this value. If both are given, a threshold is computed dynamically as just described, but if the computed threshold falls below the user specified minimum threshold, then this minimum threshold is the one used as the cutoff, overriding the computed one. This module can also generate a conservation profile for plotting or visualization purposes, and provides useful utility functions, including conversion of sequence to alignment positions and vice versa.

The phastCons conservation analysis module utilizes the phastCons scores which are conservation scores based on a whole-genome alignment of multiple vertebrate species (12). The scores for the given sequence are extracted from the UCSC Genome Browser using the hgWiggle program (7) and form the conservation profile. The profile is then scanned to identify regions which score above a user-specified conservation cutoff. Regions which score above the cutoff are then merged to form maximal spanning conserved regions (incorporating the merging process described in the pair-wise conservation analysis module above).

Transcription factor binding site search module

The transcription factor binding site (TFBS) search modules are divided into a pair-wise version and a phastCons version. In both cases, a conservation analysis object is input to the TFBS search module, which performs a motif search via an interface to the TFBS Perl modules (6). The search is limited to the conserved regions to improve efficiency. TF binding profiles may be retrieved from the JASPAR database (2), or a user-specified file of custom position frequency matrices (PFMs). Such custom PFMs can be generated on the fly through the PAZAR interface in the TF View webpage. For the phastCons version, a binding site is reported as conserved if the site

both scores above a specified threshold and falls within or partially overlaps a region classified as conserved. The amount by which a predicted site must overlap a conserved region is specified by a user-defined parameter (with a default value of a single nucleotide). In the pair-wise version, a binding site is reported if it falls within an identified conserved region (as described for phastCons), and the corresponding sites on each of the two aligned sequences are aligned and score above a specified score threshold on both sequences. As an additional constraint, TFBS searches can be limited to a specific region within the sequence by providing sequence start and end positions.

Web interface

The web interface allows the user to enter a sequence for analysis by uploading a FastA-formatted DNA sequence file, by entering genomic coordinates, or by pasting a sequence directly into the browser. The user specifies pair-wise or multi-species (phastCons) analysis. In the pair-wise case, additional pages are provided for input of an orthologous sequence in the same manner as the initial sequence. The web browser also allows for input of a gene identifier, to which matches are identified via an Ensembl API-based search (8). The user is then required to select from amongst matching genes. For pair-wise analysis, the user also provides the comparison species, and the software automatically identifies the orthologous gene(s) and transcript(s) (again via the Ensembl API). Once the relevant transcript is chosen, the amount of upstream and downstream sequence to be included in the analysis can be specified. The user is then provided with a page to specify phylogenetic footprinting parameters. Finally, TFBS profiles may be selected from the JASPAR database (2), uploaded from a file (custom profiles can be generated through the PAZAR interface) or pasted directly into the browser. The results of the analysis are displayed in the browser as a plot file showing the conservation profile, conserved regions, sequence features such as exon positions and CpG islands, and the

TFBS positions (if any were selected). Text files of the alignment, conserved regions, conserved sequences and conserved TFBS positions are also generated and can be viewed in the browser or downloaded. For additional convenience, these files are also compressed into a single downloadable file. In addition, a link is generated to automatically display the results on the UCSC browser (7) as a custom annotation track.

RESULTS

Expansion of PAZAR Database Content

Since the initial publication describing the PAZAR system (1), more scientists have created 'boutiques' to share their data and the number of annotated regulatory sequences has grown (Table 1). New boutique collections include a set of HNF4 binding sites (provided by Dr Frances Sladek; 70 annotated publications and 107 regulatory sequences), a collection of regulatory target sites for the Olf/Ebf proteins and the transcription factor encyclopedia (TFe) project that includes 49 publications and 68 regulatory sequences associated with 25 TFs. In addition, in a continuing effort to synergize our endeavors with the ORegAnno resource (3), one additional data collection, the Stanford ENCODE promoters (13), has been imported into PAZAR using our customized GFF exchange format. The total content of curated regulatory data now includes 1433 annotated publications, 1284 regulated genes, 6869 regulatory sequences and 708 TFs.

Expanded Functionality

Improved XML data exchange. The initial release of PAZAR included a limited XML (extensible markup language) data exchange format for input of large data sets into PAZAR. The XML format has since been improved and updates were made to the document type declaration (DTD) file and the associated documentation. New software tools efficiently export data from the public projects in PAZAR. The most direct access to the data in

Table 1. PAZAR database content on August 29, 2008*

Project	Regulated genes	Regulatory sequence (genomic)	Regulatory sequence (artificial)	Transcription factors	Transcription factor profiles	Annotated publications
ABS	205	611	–	152	–	110
AREs	20	20	–	1	–	20
HNF4	79	107	–	1	–	70
JASPAR core	–	–	3229	84	138	106
Liver set	14	62	–	–	–	–
MUC5AC	2	23	–	13	–	9
Muscle set	15	49	–	–	–	–
Olf_Ebf_TFBS	19	25	–	1	–	12
ORegAnno	256	690	–	114	–	305
ORegAnno ENCODEprom	302	457	–	–	–	1
ORegAnno Erythroid	8	33	–	1	–	1
ORegAnno STAT1 lit	28	37	–	1	–	29
Pleiades genes	285	1302	135	313	–	714
TFe	47	68	6	25	–	49
TOTAL	1284	3499	3370	708	138	1433

*This table includes only the experimentally validated annotations available in PAZAR and therefore excludes the Kellis predictions.

PAZAR is provided by an automated procedure which stores the entire contents of each public data collection every week. These resulting XML files are available in the new 'downloads' section on the website.

ORCAtk software. To extend the software services offered by the PAZAR database, we now provide the integrated ORCAtk software for analysis of regulatory sequences. Based on the observation that functional sequences shared by species will be more highly conserved than those that are not, cross-species comparison, or phylogenetic footprinting, has been identified as a useful means to predict regulatory regions (14–16; see also supplementary material for discussion). ORCAtk builds on this observation to predict *cis*-regulatory elements in genomic sequences by identifying conserved non-coding regions and searching for TF binding sites (TFBSs) within them.

This functionality overlaps the functions provided by a variety of Internet sites (17–19), but ORCAtk serves an important role in making such analyses available to software developers using the PAZAR system and in providing mechanisms for scientists to perform analyses incorporating information such as TF binding profiles from the data collections they create.

As described in Figure 1, ORCAtk consists of three main software modules which perform the following tasks: (i) alignment of orthologous sequences (required for pair-wise analysis only); (ii) computation of a conservation profile based on the pair-wise alignment or extraction of a multi-species phastCons score profile from the UCSC database and identification of conserved (non-coding) regions from this score profile; and (iii) identification of putative TFBSs within those conserved regions. Each of these modules can be used independently in

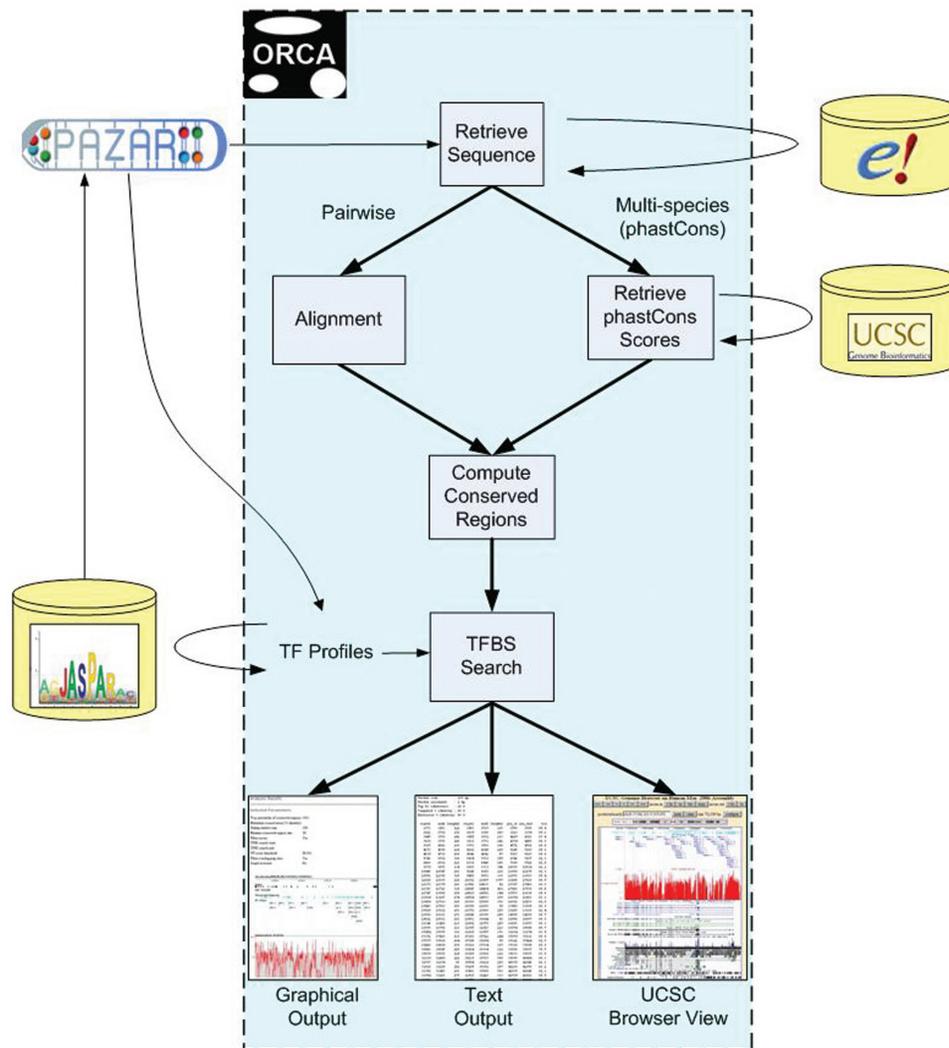


Figure 1. ORCAtk analysis pipeline. ORCAtk can be launched either directly or through PAZAR. To initiate analysis, ORCAtk first retrieves the user-specified sequence from the Ensembl database ('e!'). Second, based on user selection, ORCAtk either performs a pairwise alignment or retrieves a multi-species alignment-based phastCons score profile from the UCSC database. Third, conserved regions are identified. If the user has chosen TFBS analysis, ORCAtk then performs a search for TFBSs in the conserved regions. The TF binding profiles for this optional analysis step are provided by the user, either by selection from the JASPAR database or by upload from a file previously downloaded from PAZAR or other sources. As output, ORCAtk provides both a graphical display and text files of the results. In addition, ORCAtk results can be viewed as tracks in the UCSC genome browser.

customized applications or together to provide a complete analysis pipeline, and can be used through a command-line script as well as a web interface.

The core algorithms employed by ORCAtk are not novel. We have implemented proven bioinformatics methods in an easy-to-use and efficient manner and built upon and extended existing bioinformatics tools and Perl modules. The ORCAtk software features improvements in the Needleman–Wunsch algorithm to facilitate global progressive alignments of regulatory sequences, as well as an efficient procedure for comparison of sequence features along aligned DNA sequences.

PAZAR-ORCAtk integration. ORCAtk analysis can be launched from either the PAZAR Gene View pages or the PAZAR Sequence View pages by the click of a button. The ORCAtk web interface receives the gene or sequence information via CGI query string parameters encoded in the URL passed to it from PAZAR. ORCAtk checks these parameters, performs the necessary processing and returns the appropriate webpage in the analysis pipeline. In the case of gene information, it retrieves the gene which matches the passed gene symbol from a local Ensembl database and launches the ‘Gene Verification’ page. In the case of sequence information, ORCAtk launches into the ‘Sequence Input’ page that displays the selected sequence. Once the correct gene or sequence is specified, the user can choose either the multi-species (phastCons) or the pair-wise sequence analysis option.

In PAZAR’s TF View pages, TF binding profiles are dynamically generated from sets of binding sites for a given TF, using the MEME pattern discovery algorithm (20). A custom set of TFBS sequences can also be selected, and the profile is generated on demand via an AJAX-based process. Those profiles are then downloadable as a text file representing the position frequency matrix (PFM) calculated from the current set of selected sequences. Optionally, the user may specify a name for the downloaded file in the text field next to the download button. The downloaded text file can then be used as input for ORCAtk. ORCAtk has the capacity to read one or more custom TFBS profiles encoded in a flat file as PFMs. Once the user saves a TFBS profile from PAZAR, he or she can simply click the appropriate button in the ‘Select TFBS Profiles’ page of ORCAtk to upload the profile from the saved file.

Figure 2 shows a working example of the use of the PAZAR-ORCAtk connection. The user is interested in the human nestin gene which has five regulatory sequences annotated in PAZAR, one of them (RS0000993) is a verified binding site for various nuclear receptors. Simply using the link provided on the PAZAR interface, the user can perform a regulatory analysis of the gene and then compare the results to the *bona fide* regulatory sequences from PAZAR in the UCSC genome browser. In this specific example, ORCAtk predicts a ‘RORA_1’ binding site confirming the known nuclear receptor element. The additional predictions, included in a known regulatory region (annotated in PAZAR as RS0000992,

RS0000994, RS0001004 and RS0001005), provide new hypotheses for the user to test in experiments.

DISCUSSION

The development and expansion of the PAZAR information mall continues to progress. In addition to a growth in the number and size of data collections, we have described the new integrated ORCAtk to facilitate analyses based on the data in PAZAR. Enhancements to the XML format and introduction of download and upload tools makes it easier for researchers to interact with the system. While much work remains to achieve the goal of making PAZAR the primary open-access repository for transcriptional regulatory sequence annotation, the described updates move the system closer to the mark.

PAZAR is located at URL <http://www.pazar.info> and is distributed under the GNU Lesser General Public License (<http://www.gnu.org/copyleft/lesser.html>) to emphasize the open-source and open-access character of the system. The code is freely available on the SourceForge server (<http://www.sourceforge.net>) and we encourage participation to extend the system to any specific requirements. The ORCAtk software is open-source and freely available to all users. It can be downloaded from our website at <http://www.cisreg.ca> and is distributed under the GNU General Public License (<http://www.gnu.org/licenses/gpl.html>).

The ORCAtk has proven to be widely useful for applied analysis of regulatory sequence data. The underlying software has been used in a variety of applied projects, such as the oPOSSUM system for motif over-representation analysis (21), gene-centric studies [e.g. PIK3CA (22)] and the Pleiades Promoter Project (<http://www.pleiades.org>). By integrating the toolkit into the PAZAR system, researchers are enabled to build a binding profile for a TF and immediately apply the model to promoter sequence analysis. Moreover, users can now easily integrate and compare experimentally verified and predicted regulatory elements.

ORCAtk will continue to grow as new research approaches become mature. Possible extensions include more focus on TFBS modules rather than individual binding sites, detection of clusters of binding sites involved in cooperative binding and integration of tools for the analysis of variably spaced half-sites such as nuclear receptor binding sites.

A growing community of researchers needs a convenient system for sharing regulatory sequence data. As increasing numbers of researchers develop data collections, demand for a shared repository naturally grows. This has been evident for DNA sequences [GENBANK (23)], genome sequences [UCSC (7) and ENSEMBL (8)], microarrays [ArrayExpress (24) and GEO (25)] and polymorphisms [dbSNP (26)]. Unfortunately, experimental results of *cis*-regulatory sequence and TF analyses are often contained in disparate sites on the Internet and are not usually kept up-to-date, a fact that limits utility. In addition, there are many regulatory sequence databases, but none currently allows data producers to

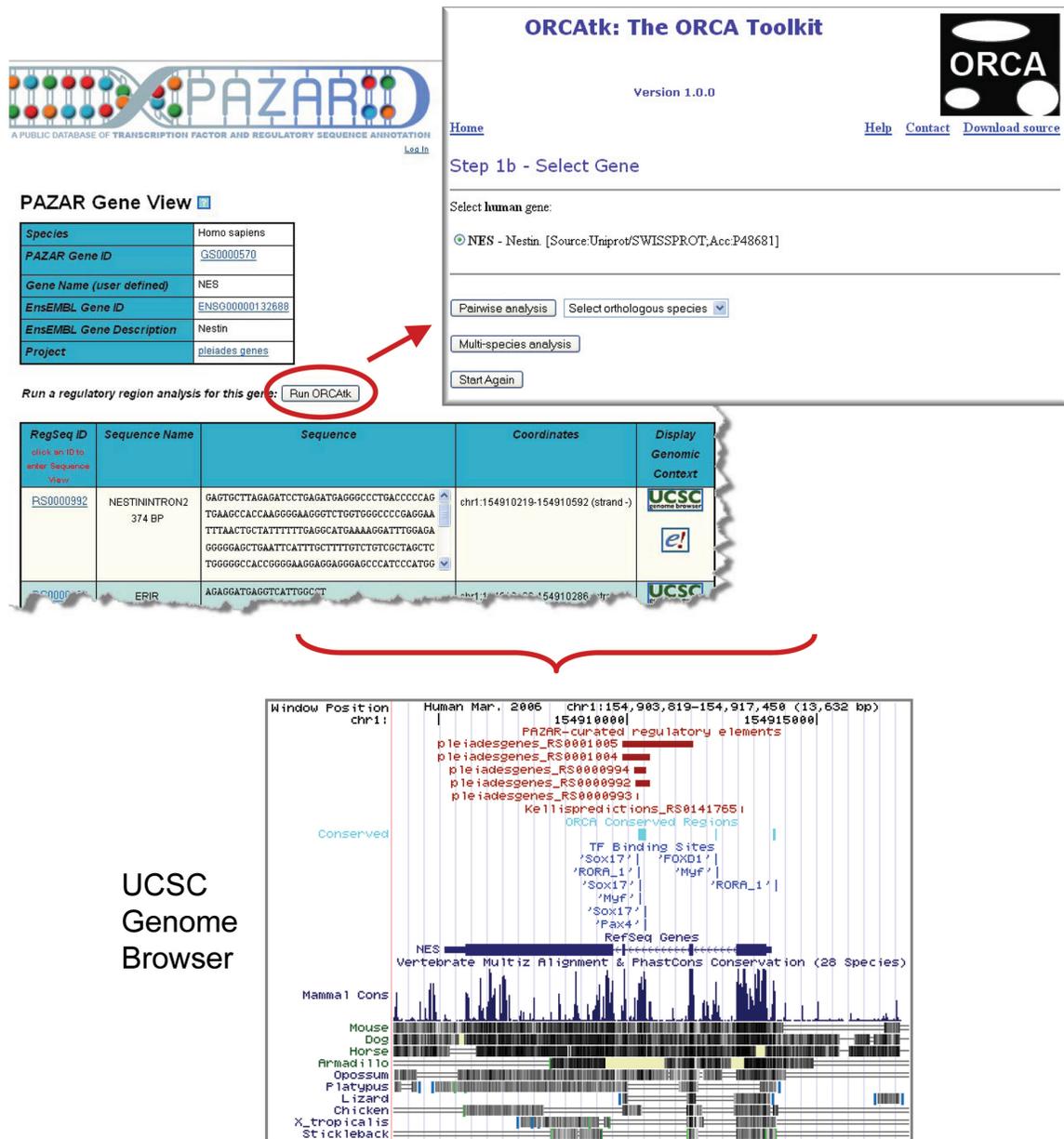


Figure 2. The PAZAR-ORCA link provides a portal to gene regulation studies. On the PAZAR Gene View page, a button is available for the user to launch an analysis using the ORCAtk. Clicking the button opens a new window displaying the ORCAtk interface with the appropriate gene already selected. The user can now proceed to the analysis including multiple steps where custom parameters can be defined or the default parameters used. The results of the analysis, as well as the annotations from PAZAR, can then be displayed on the UCSC genome browser which provides a user-friendly platform to compare experimentally-verified and predicted binding sites.

create and maintain their own collections. PAZAR was created to bring these disparate ‘boutique’ datasets together under one roof and maintain their relevancy through reference to current genomic coordinates. The association of data collections with individual research labs provides an important recognition of the work, promotes longer-term maintenance of the information and allows users to access the collections they deem to be reliable. In comparison to commercial systems, PAZAR is more likely to attract participation from researchers wishing to share their data with the research community. With the increasing production of high-throughput

TFBS data, the depositing of data into PAZAR is likely to increase.

The advances described in this report largely address the curated annotation of regulatory sequences based on individual gene studies in the scientific literature. In order to make the system suitable for sequencing or microarray derived binding data, the data depositing process will be expanded. Bulk uploading of target sequence coordinates coupled to a common TF (or TF complex) would allow for high-throughput data to be rapidly collected. The identification of the TF remains cumbersome in the current implementation. In the future, we will incorporate the

transcription factor catalog (TFCat) (Fulton D.L. *et al.*, under revision) which provides an organized structural classification of human and mouse DNA binding proteins. Future work will also provide access to the data in the system via web services.

The regulation of gene transcription is a fundamental process in health and disease. PAZAR serves as a data resource of growing importance to researchers committed to understanding how and when genes will be transcribed.

SUPPLEMENTARY DATA

Supplementary data are available at NAR Online.

ACKNOWLEDGEMENTS

We acknowledge Dr Jay Snoddy for contribution to the early PAZAR development, Amy Ticoll and Stuart Lithwick for addition of curated data to the system, Dimas Yusuf for the drawing of the PAZAR mall map, Jerome Bacconnier for the PAZAR logo and Web interface design. Dr Luis Mendoza developed a global progressive alignment procedure which helped define the challenges for the ORCA_tk to overcome. Dr David Martin provided the algorithm for building the maximum scoring chain of BLAST hits used in the ORCA aligner. Dr Pär Engström provided some testing and bug fixes to the ConservationAnalysis modules.

FUNDING

GenomeCanada (via the Pleiades Promoter Project to PAZAR and ORCA_tk); the Canadian Institute of Health Research; Canada Foundation for Innovation, the Canadian Genetic Diseases Network; Merck and IBM. W.W.W. is a CIHR New Investigator and a Scholar of the Michael Smith Foundation for Health Research. Funding for open access publication charge has been waived by Oxford University Press.

Conflict of interest statement. None declared.

REFERENCES

- Portales-Casamar,E., Kirov,S., Lim,J., Lithwick,S., Swanson,M.I., Ticoll,A., Snoddy,J. and Wasserman,W.W. (2007) PAZAR: a framework for collection and dissemination of cis-regulatory sequence annotation. *Genome Biol.*, **8**, R207.
- Vlieghe,D., Sandelin,A., De Bleser,P.J., Vlemingck,K., Wasserman,W.W., van Roy,F. and Lenhard,B. (2006) A new generation of JASPAR, the open-access repository for transcription factor binding site profiles. *Nucleic Acids Res.*, **34**, D95–D97.
- Montgomery,S.B., Griffith,O.L., Sleumer,M.C., Bergman,C.M., Bilenky,M., Pleasance,E.D., Prychyna,Y., Zhang,X. and Jones,S.J. (2006) ORegAnno: an open access database and curation system for literature-derived promoters, transcription factor binding sites and regulatory variation. *Bioinformatics*, **22**, 637–640.
- Stajich,J.E., Block,D., Boulez,K., Brenner,S.E., Chervitz,S.A., Dagdigan,C., Fuellen,G., Gilbert,J.G., Korf,I., Lapp,H. *et al.* (2002) The Bioperl toolkit: Perl modules for the life sciences. *Genome Res.*, **12**, 1611–1618.
- Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
- Lenhard,B. and Wasserman,W.W. (2002) TFBS: computational framework for transcription factor binding site analysis. *Bioinformatics*, **18**, 1135–1136.
- Karolchik,D., Kuhn,R.M., Baertsch,R., Barber,G.P., Clawson,H., Diekhans,M., Giardine,B., Harte,R.A., Hinrichs,A.S., Hsu,F. *et al.* (2008) The UCSC Genome Browser Database: 2008 update. *Nucleic Acids Res.*, **36**, D773–D779.
- Flicek,P., Aken,B.L., Beal,K., Ballester,B., Caccamo,M., Chen,Y., Clarke,L., Coates,G., Cunningham,F., Cutts,T. *et al.* (2008) Ensembl 2008. *Nucleic Acids Res.*, **36**, D707–D714.
- Brudno,M., Do,C.B., Cooper,G.M., Kim,M.F., Davydov,E., Green,E.D., Sidow,A. and Batzoglou,S. (2003) LAGAN and Multi-LAGAN: efficient tools for large-scale multiple alignment of genomic DNA. *Genome Res.*, **13**, 721–731.
- Bray,N., Dubchak,I. and Pachter,L. (2003) AVID: a global alignment program. *Genome Res.*, **13**, 97–102.
- Needleman,S.B. and Wunsch,C.D. (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.*, **48**, 443–453.
- Siepel,A., Bejerano,G., Pedersen,J.S., Hinrichs,A.S., Hou,M., Rosenbloom,K., Clawson,H., Spieth,J., Hillier,L.W., Richards,S. *et al.* (2005) Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.*, **15**, 1034–1050.
- Trinklein,N.D., Aldred,S.J., Saldanha,A.J. and Myers,R.M. (2003) Identification and functional analysis of human transcriptional promoters. *Genome Res.*, **13**, 308–312.
- Tagle,D.A., Koop,B.F., Goodman,M., Slightom,J.L., Hess,D.L. and Jones,R.T. (1988) Embryonic epsilon and gamma globin genes of a prosimian primate (*Galago crassicaudatus*). Nucleotide and amino acid sequences, developmental regulation and phylogenetic footprints. *J. Mol. Biol.*, **203**, 439–455.
- Levy,S., Hannehalli,S. and Workman,C. (2001) Enrichment of regulatory signals in conserved non-coding genomic sequence. *Bioinformatics*, **17**, 871–877.
- Fickett,J.W. and Wasserman,W.W. (2000) Discovery and modeling of transcriptional regulatory regions. *Curr. Opin. Biotechnol.*, **11**, 19–24.
- Sandelin,A., Wasserman,W.W. and Lenhard,B. (2004) ConSite: web-based prediction of regulatory elements using cross-species comparison. *Nucleic Acids Res.*, **32**, W249–W252.
- Corcoran,D.L., Feingold,E. and Benos,P.V. (2005) FOOTER: a web tool for finding mammalian DNA regulatory regions using phylogenetic footprinting. *Nucleic Acids Res.*, **33**, W442–W446.
- Aerts,S., Van Loo,P., Thijs,G., Mayer,H., de Martin,R., Moreau,Y. and De Moor,B. (2005) TOUCAN 2: the all-inclusive open source workbench for regulatory sequence analysis. *Nucleic Acids Res.*, **33**, W393–W396.
- Bailey,T.L. and Elkan,C. (1994) Fitting a Mixture Model By Expectation Maximization To Discover Motifs In Biopolymers. *Proceedings of the International Conference on Intelligent Systems for Molecular Biology*, Vol. 2, pp. 28–36.
- 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–W252.
- Astanehe,A., Arenillas,D., Wasserman,W.W., Leung,P.C., Dunn,S.E., Davies,B.R., Mills,G.B. and Auersperg,N. (2008) Mechanisms underlying p53 regulation of PIK3CA transcription in ovarian surface epithelium and in ovarian cancer. *J. Cell Sci.*, **121**, 664–674.
- Benson,D.A., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J. and Wheeler,D.L. (2008) GenBank. *Nucleic Acids Res.*, **36**, D25–D30.
- Parkinson,H., Kapushesky,M., Shojatalab,M., Abeygunawardena,N., Coulson,R., Farne,A., Holloway,E., Kolesnykov,N., Lilja,P., Lukk,M. *et al.* (2007) ArrayExpress—a public database of microarray experiments and gene expression profiles. *Nucleic Acids Res.*, **35**, D747–D750.
- Barrett,T., Troup,D.B., Wilhite,S.E., Ledoux,P., Rudnev,D., Evangelista,C., Kim,I.F., Soboleva,A., Tomashevsky,M. and Edgar,R. (2007) NCBI GEO: mining tens of millions of expression profiles—database and tools update. *Nucleic Acids Res.*, **35**, D760–D765.
- Sherry,S.T., Ward,M.H., Kholodov,M., Baker,J., Phan,L., Smigielski,E.M. and Sirotkin,K. (2001) dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.*, **29**, 308–311.