UCRL-TR-228361

LAWRENCE
LIVERMORE
NATIONAL
LABORATORY

04-ERD-052-Final Report

G. G. Loots, I. Ovcharenko, N. Collette, P. Babu, J. Chang, L. Stubbs, X. Lu, C. Pennachio, R. M. Harland

February 26, 2007

Disclaimer

This document was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor the University of California nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or the University of California. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or the University of California, and shall not be used for advertising or product endorsement purposes.

This work was performed under the auspices of the U.S. Department of Energy by University of California, Lawrence Livermore National Laboratory under Contract W-7405-Eng-48.

FY06 LDRD Final Report

A coupled computational and experimental approach to determine functions of deeply conserved "anonymous" human genes

LDRD Project Tracking Code: *04-ERD-052*
Gabriela G. Loots, Principal Investigator

Abstract

Generating the sequence of the human genome represents a colossal achievement for science and mankind. The technical use for the human genome project information holds great promise to cure disease, prevent bioterror threats, as well as to learn about human origins. Yet converting the sequence data into biological meaningful information has not been immediately obvious, and we are still in the preliminary stages of understanding how the genome is organized, what are the functional building blocks and how do these sequences mediate complex biological processes. The overarching goal of this program was to develop novel methods and high throughput strategies for determining the functions of 'anonymous' human genes that are evolutionarily deeply conserved in other vertebrates. We coupled analytical tool development and computational predictions regarding gene function with novel high throughput experimental strategies and tested biological predictions in the laboratory. The tools required for comparative genomic data-mining are fundamentally the same whether they are applied to scientific studies of related microbes or the search for functions of novel human genes. For this reason the tools, conceptual framework and the coupled informatics-experimental biology paradigm we developed in this LDRD has many potential scientific applications relevant to LLNL multidisciplinary research in bio-defense, bioengineering, bio-nanosciences and microbial and environmental genomics.

Introduction/Background

In the past few years, genomic DNA sequences have emerged for numerous microbes, plants, and animals from yeast to humans, presenting us with the unique opportunity to answer fundamental questions regarding genome architecture, evolution and the genetic causes of human disease. Genomic sequence analysis holds the key to new pioneering discoveries in biology. The availability of several completely sequenced animal genomes also facilitates the creation of new tools and strategies for identifying the interconnecting building blocks necessary and sufficient to build a developmentally complex vertebrate life form. Although vertebrate genomic sequences have been available to the public for some time now, we have only begun to mine the information embedded within the raw sequence data of the human genome. One significant reason for this slow progress is the paucity of computational tools and high-throughput experimental strategies available to efficiently interpret the information encoded within the DNA of complex living organisms. Whereas gene-finding tools are improving, a large percentage of the vertebrate genes that have been identified are 'anonymous', and of completely unknown function. This proposal has focused on developing new high-throughput tools to determine the *in vivo* function of highly conserved, anonymous human genes.

The overarching goal of this program has been to couple analytical tool development and computational predictions regarding gene function with novel high throughput experimental strategies to test the biological predictions in the laboratory. Similar couplings of computations and experimentation have been the mainstay of LLNL research in other disciplines for years, however this coupling is a new concept in biological sciences and one that represents the "next wave" for future biomedical research. The tools required for comparative genomic data-mining are fundamentally the same whether they are applied to scientific studies of related microbes or the search for functions of novel human genes. For this reason the tools, conceptual framework and the coupled informatics-experimental biology paradigm we propose to establish in this LDRD has many potential scientific applications relevant to LLNL multidisciplinary biological research.

Research Activities

This study has focused on three major aims:

1. Development of new computational algorithms and analytical tools to predict and statistically analyze features that are key to the function of conserved human genes and the associated sequence elements. Using whole-genome comparisons and other tools, we were able to define the "core" vertebrate genome: a set of highly conserved human genes that have been previously uncharacterized, but are predicted to have important functions. These genes were further prioritized for further experimental study.

2. Identification of candidate genes with interesting biological functions by examining expression patterns of unknown transcripts in frog and mouse at comparable stages of development. Genes with significant/interesting embryonic expression patterns were prioritized as targets for *in vivo* experimental analysis.

3. Establish gene-manipulation technologies such as transgenic over-expression and morpholinos (modified antisense oligonucleotides that block protein translation) to "knock-up or -down" the expression of unknown genes in the frog and study the effects of these manipulations during development. This work was carried out in close collaboration with Dr. Harland's group at UC Berkeley. We aimed to determine the function of 20-30 conserved genes in this pilot study.

Computational Algorithms for Comparative Studies

Comparisons between deeply diverged vertebrate species, and in particular, multispecies comparisons, provide the best method for identifying and classifying functionally relevant DNA sequences (Loots et al. 2000; Pennacchio and Rubin 2001). Although several excellent methods have been developed, finding the best and most informative strategy for aligning whole-genomes remains a challenging computational task. Before initiating this program, we have successfully developed novel technologies for aligning whole vertebrate genomes (http://ecrbrowser.dcode.org/), and we used this platform to continue our tool and algorithm developments. Initially we focused on generating a tool that would increase the flexibility of comparative alignment tools to better define the extent and structure of evolutionarily conserved regions (ECRs) in multi-genome comparisons. We also focused on improvements of data display and retrieval tools that would make comparative data more accessible to experimental biologists. We spent significant effort classifying ECRs according to their specific noncoding or coding function to enrich for gene targets likely to be functional. We prioritized highly conserved genes with similar expression patterns in frogs since we believe that these genes are most likely to have critical roles during early vertebrate development, and to recapitulate the human function. Our main goal was to dramatically improve the reliability of computational predictions such that detected ECRs are likely to have predictable biological function. All data and tool generated have been made publicly available at http://dcode.org.

Expression Pattern Analysis

We focused our functional analysis on genes for which both coding sequence and noncoding elements have been highly conserved. By adapting a revolutionary technique developed by Dr. Stubbs' team of scientists, we aimed to characterize the expression pattern of genes with unknown functions in whole mount and sectioned frog embryos of different developmental stages. We also aimed to carry out a pilot study for high throughput upscaling capabilities that would eventually we applied to the entire human genome gene analysis.

Genetic Manipulation of Gene Expression

Sequences that generated interesting expression patterns were used in functional studies *in vivo*, in frogs. For each gene, we made mRNA or designed morpholinos and inject embryos. We assessed the effects of each mRNA based on the predicted function, for example, a gene predicted to function during kidney development was tested for its role during kidney patterning.

Results/Technical Accomplishments

Computational Algorithms for Comparative Studies

In the course of this program we have created a suite of bioinformatic tools for biological discovery, all of which are publicly were described in publications and are available at http://www.dcode.org/. Briefly, we have created a comparative genome browser, the ECR browser (http://ecrbrowser.dcode.org/) which currently includes all available vertebrate genomes. We have also created an alignment visualization engine (http://zpicture.dcode.org/), a transcription factor binding site analyzer (http://rvista.dcode.org/), a multiple alignment program for analyzing closely related sequences (http://eshadow.dcode.org/), and a *cis*-regulatory module analyzer (http://crème.dcode.org/). Our tools currently receive hundreds of requests per day, and we have thousands of users on all continents and with both academic and commercial affiliations. For one of our tools, zPicture, we have obtained licensing rights, and is now available for distribution at the following website: http://www.llnl.gov/IPandC/technology/software/softwaretitles/zpicture.php.

Expression Analysis

We have aligned the 10X frog genome assembly (http://ecrbrowser.dcode.org/) to the human genome and found ~65% of all human genes to be highly conserved to frog (conservation criteria of >100bp/>70%ID; nucleotide level). This represents over 12,000 transcripts, a data comparable to the defined human/chicken conservation. Using various computational filters we have identified ~2,000 transcripts categorized as 'novel anonymous conserved genes'. We have used this dataset to perfect our high throughput *in situ* capabilities, using the newly purchased *in situ* robot (Holle&Huttner AG BioLane HTI), optimized for mRNA probes from IMAGE clones (http://image.llnl.gov/), in 40- and 80- well format. Up to date we have generated *in situ* patterns for over 300 genes, exceeding our projected 30 genes goal/year, and have created an *in situ* database that is available at http://insitu.dcode.org

Other Accomplishments

We have established a state of the art frog transgenic and genetic facility. We have three re-circulating water tanks that house two species of frogs, X.laevis and X.tropicalis. One of the systems in particular has small tanks that can house individual families of genetically modified X.tropicalis frogs. We have also in this program acquired the necessary equipment and technical expertise to carry out transgenic experiments and high throughput in situ patterning determination. This project has created a sustainable foundation for future xenopus research in our group as well as a resource to all other investigators at the lab who can use gene manipulation in frogs as part of their research.

Publications generated from this project:

1. Ovcharenko I, Loots GG, Hardison RC, Miller W and Stubbs L. (2004) zPicture: Dynamic alignment and visualization tool for analyzing conservation profiles. Genome Research 14:472-7.
2. Ovcharenko I and Loots GG. Comparative Genomics. (2004) Tools for Exploring the Human Genome. CSHL Symposia on Quantitative Biology; The genome of *Homo Sapiens*, volume LXVIII.
3. Ovcharenko I, Boffelli D and Loots GG. (2004) eShadow: A tool for comparing closely related species. Genome Research 14: 1191-1198.
4. Ovcharenko I, Nobrega MA, Loots GG and Stubbs L. (2004) ECR Browser: A Tool For Visualizing And Accessing Data From Comparisons Of Multiple Vertebrate Genomes. Nucleic Acids Research 32(13).
5. Loots GG and Ovcharenko I. (2004) rVISTA 2.0: Evolutionary Analysis of Transcription Factor Binding Sites. Nucleic Acid Research 32(13).
6. Sharan R, Ben-Hur A, Loots GG and Ovcharenko I. (2004) CREME: Cis-Regulatory Module Explorer for the Human Genome. Nucleic Acid Research 32(13).
7. Ovcharenko I, Stubbs L and Loots GG. (2004) Interpreting Mammalian Evolution using Fugu Genome Comparisons. Genomics 84(5):890-5.
8. Ovcharenko I, Loots GG, Giardine BM, Hou M, Ma J, Hardison RC, Stubbs L and Miller W. (2005). Mulan: multiple-sequence local alignment and visualization for stuying function and evolution. Genome Research 15(1):184-94.
9. Ovcharenko I, Loots GG, Nobrega MA, Hardison RC, Miller W and Stubbs L. (2005). Evolution and functional classification of vertebrate gene deserts. Genome Research 15(1):137-45. (featured on the cover)
10. Loots GG and Ovcharenko I. (2005). Dcode.org anthology of comparative genomic tools. Nucleic Acid Research; Web Server Issue, July (33):W1-W9.
11. Loots GG, Kneissel M, Keller H, Baptist M, Chang J, Collette NM, Ovcharenko D, Plajzer-Frick I and Rubin EM. (2005). Genomic deletion of a long-range bone enhancer misregulated sclerostin in Van Buchem Disease. Genome Research. 15(7):928-35. (featured on the cover)
12. Khokha M and Loots GG. (2005) Strategies for characterizing cis-regulatory elements in *Xenopus tropicalis*. *Briefings in Functional Genomics & Proteomics* 4(1):58-68.

13. Loots GG, Chain PS, Mabery S, Rasley A, Garcia E and Ovcharenko I. (2006). Array2Bio: from Microarray expression data to functional annotation of co-regulated genes. *BMC Bioinformatics* 16(7):307.

14. Loots GG and Ovcharenko I. (2006). ECRBase: Database of Evolutionary Conserved Regions, Promoters, and Transcription Factor Binding Sites in Vertebrate Genomes. *Bioinformatics* 22(23):1-3.

Book Chapters:

15. Gabriela G. Loots. Modifying Yeast Artificial Chromosomes to Generate Cre/LoxP and FLP/FRT Site-Specific Deletions and Inversions. In Methods in Molecular Biology: YAC Protocols. (Alasdair MacKenzie, ed.), 2th edition, Chapter 8, Humana Press Inc, USA (2006).

16. Gabriela G. Loots and Ivan Ovcharenko. Mulan Multiple-Sequence Alignment to Predict Functional Elements in Genomic Sequences. In Methods in Molecular Biology: Comparative Genomics. (Nicholas Bergman, ed.), Humana Press Inc, USA (2007, in press)

Exit Plan

Success in this pilot study has positioned us very favorably for obtaining funding from NIH (NHGRI, NICHD), which is presently focused intently on developing methods for comparative genomic analysis and high throughput methods for determining the function of human genes. With its focus on genes that control the most basic mechanisms of vertebrate life, this work is also in step with DOE/OBER's current plan for future research in genomics. The program fits well with JGI sequencing goals and may provide a new direction for JGI high throughput functional genomic studies as methods are established and streamlined in future years. The methods developed here may also translate directly into enhanced competition for other DOE programs including low-dose radiation studies (by providing a high throughput vertebrate model system for the study of radiation response, DNA repair and cellular-stress pathways). *Grants*--We have written and submitted several RO1 grant proposals to NIH during the duration of this project, all of which have received positive feedback from reviewers, praising the scientific value of this work. One proposal in particular is under review at has received a fundable score. If funded we anticipate this grant to replace the LDRD as early as June 2007.

Summary

Understanding the functional blocks in the human genome represents the greatest challenge in genomics, but has the potential to unlock information that would cure disease, combat bioterror threats, as well as teach us about human origins, and human interaction with the environment. Despite tremendous technical advances both in computational and experimental biology, we still lack the ability to predict the function of DNA *de novo*. The LDRD funding received for our project has allowed us to establish a program in developmental and computational genomics that can expedite the identification and characterization of functional coding and noncoding elements in the human genome. We have developed computational and high-throughput experimental tools that we can now harmoniously apply to specific biological questions, and facilitate the discovery of novel genes, regulatory elements and signaling pathways that are essential for organogenesis and physiological processes. This project has resulted in 14 publications in internationally recognized journals, and two of these manuscripts have been featured on the cover of the journal Genome Research, one of the most esteemed journals in Genomics. We have also written two book chapters and presented out work at several International Meetings. During the course of this program, we have developed a conglomerate of computational tools (ecrbrowser, eshadow, zpicture, mulan, ecrbase, array2bio) and methods that are easy to use and are aimed to aid computational analysis in wet-bench biology labs. We have also developed new high throughput experimental methods in a new developmental system, the frog X.tropicalis. This frog has been sequenced by the Joint Genome Institute, and has great potential for evolving into a high throughput genomic experimental model, for quick validation of human sequences. As part of this project we have built the infrastructure which allows us now to carry out all aspects of frog experimentation. We have a state of the art aquatic facility that houses up to 1000 adult frogs. We have recirculating tank systems with small individual tanks that allow us to house families of genetically modified strains of frogs. We also have all the instruments required for microinjection and in situ hybridization, as well as the technical expertise to carry out the experiments. The tools required for comparative genomic data-mining are fundamentally the same whether they are applied to scientific studies of related microbes or the search for functions of novel human genes. For this reason the tools, conceptual framework and the coupled informatics-experimental biology paradigm we have developed as part of this LDRD has many potential scientific ramifications relevant to LLNL multidisciplinary research in bio-defense, bioengineering, bio-nanosciences and microbial and environmental genomics.

Acknowledgements: We are grateful to Dr. Harland's generosity and mentorship, for allowing us to use his facilities and resources while we were setting up our laboratory, this work would not have been possible without his kindness. We would also like to thank Cate Oldenburg, Mustafa Khokha and Tim Grammer who has always been available to train us, set up our equipment and provide us with invaluable advise on frog techniques and husbandry.

References

Beck, C.W. and J.M. Slack. 2001. An amphibian with ambition: a new role for Xenopus in the 21st century. *Genome Biol* **2:** REVIEWS1029.

Grammer, T.C., K.J. Liu, F.V. Mariani, and R.M. Harland. 2000. Use of large-scale expression cloning screens in the Xenopus laevis tadpole to identify gene function. *Dev Biol* **228:** 197-210.

Knecht, A.K. and R.M. Harland. 1997. Mechanisms of dorsal-ventral patterning in noggin-induced neural tissue. *Development* **124:** 2477-2488.

Lander, E.S. et al. 2001. Initial sequencing and analysis of the human genome. *Nature* **409:** 860-921.

Loots, G.G., R.M. Locksley, C.M. Blankespoor, Z.E. Wang, W. Miller, E.M. Rubin, and K.A. Frazer. 2000. Identification of a coordinate regulator of interleukins 4, 13, and 5 by cross-species sequence comparisons. *Science* **288:** 136-140.

Pennacchio, L.A. and E.M. Rubin. 2001. Genomic strategies to identify mammalian regulatory sequences. *Nat Rev Genet* **2:** 100-109.

Venter, J.C. et al. 2001. The sequence of the human genome. *Science* **291:** 1304-1351.

Vogel, G. 1999. Frog is a prince of a new model organism. *Science* **285:** 25.