Intraspecies sequence comparisons for annotating genomes

Dario Boffelli,1,2 Claire V. Weer,1,2 Li Weng,1,2 Keith D. Lewis,1,2 Malak I. Shoukry,1,2 Lior Pachter,2,3 David N. Keys,1,2 and Edward M. Rubin1,2,4

1US Dept. of Energy Joint Genome Institute, Walnut Creek, California 94598, USA; 2Genomics Division, Lawrence Berkeley National Laboratory, Berkeley, California 94720, USA; 3Department of Mathematics, University of California, Berkeley, California 94720, USA

Analysis of sequence variation among members of a single species offers a potential approach to identify functional DNA elements responsible for biological features unique to that species. Due to its high rate of allelic polymorphism and ease of genetic manipulability, we chose the sea squirt, Ciona intestinalis, to explore intraspecies sequence comparisons for genome annotation. A large number of C. intestinalis specimens were collected from four continents, and a set of genomic intervals were amplified, resequenced, and analyzed to determine the mutation rates at each nucleotide in the sequence. We found that regions with low mutation rates efficiently demarcated functionally constrained sequences: these include a set of noncoding elements, which we showed in C. intestinalis transgenic assays to act as tissue-specific enhancers, as well as the location of coding sequences. This illustrates that comparisons of multiple members of a species can be used for genome annotation, suggesting a path for the annotation of the sequenced genomes of organisms occupying uncharacterized phylogenetic branches of the animal kingdom. It also raises the possibility that the resequencing of a large number of Homo sapiens individuals might be used to annotate the human genome and identify sequences defining traits unique to our species.

[Supplemental material is available online at www.genome.org. The sequence data from this study were submitted to GenBank under accession nos. AY667278–AY667407. The following individuals kindly provided reagents, samples, or unpublished information as indicated in the paper: S. Fujiwara, A. Gittenberger, K. Heasman, H. Huelvan, D. Jiang, S. Kano, A. Phillippi, A. Sexton, and S. Shimeld.]

Sequence comparisons between the genomes of organisms separated by a varying degree of evolutionary distances currently serve as an essential means to identify genes as well as gene regulatory elements (Ansari-Lari et al. 1998; Nobrega et al. 2003; Thomas et al. 2003). These comparisons are based on the well established molecular evolution principle that negative selection reduces the accumulation of sequence differences in functional sequences of related species (Hartl and Clark 1997; Hardison 2003). An important limitation of interspecies comparisons is that they can only be used to identify sequences underlying biological traits shared by the species examined. Although recently described approaches, leveraging the sequences of many closely related species to increase the total evolutionary branch length of the comparisons, have begun to address this issue (Boffelli et al. 2003), they are nevertheless ill-suited to uncover species-specific features. Intraspecies comparisons of the genomes of numerous members of the same species offer a theoretical strategy to tackle this problem. Although this approach is clearly expected to require the sequences of a very large number of individuals of the same species, the progressive lowering of the barrier to large-scale resequencing made possible by advances in sequencing technology now provides the opportunity to determine whether the theoretical advantages of intraspecies comparisons for genome annotation can be supported by experimental data.

The sequencing of the ascidian Ciona intestinalis (sea squirt) genome recently revealed this organism to be a particularly attractive candidate target for testing the feasibility of resequencing for annotating genomes. An important attribute is its very high allelic polymorphism, with an average 1.2% of the nucleotides differing between chromosome pairs of a single individual (Dehal et al. 2002). This high degree of allelic variation, more than 15-fold that noted in humans, is probably a consequence of the large effective population size of C. intestinalis. In addition, the genetic manipulability of the sea squirts offers a rigorous in vivo experimental system to test the functional activity of identified candidate regulatory elements (Satoh 2003).

In this study, we determined the extent of sequence polymorphism in several C. intestinalis subpopulations collected at multiple locations worldwide. We exploited sequence variation within C. intestinalis to computationally identify regions subjected to fast and slow rates of evolution, and we experimentally characterized their functional roles. These studies illustrate that slowly evolving regions correspond to protein-coding or enhancer regions, indicating the feasibility of using intraspecies polymorphism to annotate a species’ genome.

Results

Amplification of genomic targets and phylogenetic analysis

C. intestinalis specimens were collected from several coastal locations in North America, Europe, Eastern Asia, and Oceania.
Samples were defined as *C. intestinalis* based on characteristic morphological features supported by sequence analysis. Although the formal proof that all the samples analyzed are members of the same species would require successful sexual mating, the sequence difference among the samples examined indicated that they are several-fold more similar to each other than to their closest intra-genus relative, *C. savignyi*. Any two pairs of *C. intestinalis* analyzed in this study were at least 85% identical, a value within the range of allelic polymorphism reported for a single *C. intestinalis* individual (Dehal et al. 2002). Conversely, BLAST comparisons between *C. intestinalis* and *C. savignyi*, carried out at the relaxed threshold of $e = -1$, revealed sequence alignments for less than 10% of the sequences. Four genomic intervals, each ~4 kb long, were chosen for analysis in this study based on previous knowledge of genes and gene regulatory elements located within these intervals. Target regions included exons 18–25 of *patched* homolog, exons 1–4 of *col5a1*, and the 5′ sequences of *forkhead* and *snail*. The targeted coding regions all had strong gene structure predictions supported by EST sequences, while several 5′ tissue-specific enhancers for *forkhead* and the promoter of *snail* had been previously defined and characterized (Erives et al. 1998; Di Gregorio et al. 2001).

Using the consensus sequence of an individual collected from a West Coast location for PCR primer design, amplification was attempted for each of the four genomic target regions from 140 animals. Successful PCR reactions yielded unique bands in most cases. When two bands were obtained, they were both sequenced independently and in all instances resulted in heterozygous allelic variants due to insertion/deletion events. Even though all four target regions were unique in the current *C. intestinalis* assembly, its draft status raises the possibility that amplimers from different individuals represent paralogous, rather than allelic, copies. This is unlikely, since all the sequences for the same target locus had stretches of perfect identity longer than 100 base pairs, which are not normally observed between paralogous loci.

We were able to obtain amplified genomic targets from ~50% of the individuals collected from West Coast locations, 20% of the individuals from New Zealand and Japan, and less than 20% of the individuals collected from coastal locations on the Atlantic Ocean or the Mediterranean Sea (Table 1). The difficulty in amplifying *C. intestinalis* genomic samples, explained by the high levels of polymorphism of this species, suggests that only a reduced subset of the polymorphism present in *C. intestinalis* was likely captured by this study.

Successfully amplified target regions were fully sequenced using custom primers designed –every 250 bp on each strand, for a total of 40 sequencing reads for each amplified target region. This level of coverage ensured that each base was read at least four times. Analysis of the sequences revealed that specimens from the same collection location clustered nearly exclusively close to each other as estimated by their degree of sequence similarity, supporting the conjecture that they are members of largely isolated subpopulations (Fig. 1). Surprisingly, individuals from Mediterranean locations appeared more related to individuals collected from the Pacific rather than Atlantic Ocean. In addition, subpopulations collected from locations on the Atlantic Ocean showed much higher heterozygosity than subpopulations from locations on the Pacific Ocean, as reflected by the size of the circles in Figure 1. *C. intestinalis* is an invasive species reported to be spread in ship bilges along shipping routes. This is supported by the clustering of samples from the Pacific Ocean, reflecting the greater shipping activity between California, Japan, and New Zealand ports. The lower heterozygosity among those samples also suggests that the Pacific Ocean was colonized after the Atlantic Ocean.

The phylogenetic tree of the individuals sequenced in each genomic interval was obtained in two steps. First, phylogenetic relationships between subpopulations were calculated from the consensus sequences for each subpopulation, defined by their collection locations (see legend, Table 1). Phylogenetic relationships for individuals from the same subpopulation were then estimated from the average distance of all members of that subpopulation, since the degree of sequence similarity among individuals from the same subpopulation did not allow the computation of statistically significant trees within subpopulations. The resulting composite trees were used to calculate the likelihood that each nucleotide site in the multiple sequence alignment is mutating at a high or at a low rate (Boffelli et al. 2003). Variation profiles of the genomic intervals analyzed were displayed through likelihood ratio curves to identify regions undergoing the slowest mutation rates relative to the rate of their surrounding regions.

Table 1. Summary of PCR amplification of *forkhead*, *snail*, *col5a1* and *patched*

<table>
<thead>
<tr>
<th></th>
<th>forkhead</th>
<th>snail</th>
<th>col5a1</th>
<th>patched</th>
</tr>
</thead>
<tbody>
<tr>
<td>WC</td>
<td>16 (.48)</td>
<td>16 (.48)</td>
<td>17 (.52)</td>
<td>19 (.58)</td>
</tr>
<tr>
<td>JP</td>
<td>5 (.25)</td>
<td>6 (.30)</td>
<td>4 (.20)</td>
<td>4 (.20)</td>
</tr>
<tr>
<td>NZ</td>
<td>3 (.21)</td>
<td>1 (.7)</td>
<td>5 (.36)</td>
<td>3 (.20)</td>
</tr>
<tr>
<td>EC</td>
<td>8 (.17)</td>
<td>8 (.17)</td>
<td>5 (.31)</td>
<td>0 (.0)</td>
</tr>
<tr>
<td>FR</td>
<td>1 (.7)</td>
<td>4 (.27)</td>
<td>3 (.20)</td>
<td>0 (.0)</td>
</tr>
<tr>
<td>IT</td>
<td>0 (.0)</td>
<td>2 (.17)</td>
<td>2 (.17)</td>
<td>0 (.0)</td>
</tr>
</tbody>
</table>

Number of individuals amplified from each subpopulation for each of the four target regions analyzed. The numbers in parentheses indicate the fraction of individuals that yielded successful amplification out of all the amplifications attempted from that subpopulation. WC: samples from three collection locations on the Pacific coast of the United States. JP: samples from two collection locations in western Japan. NZ: samples from one collection location in New Zealand. FR: samples from one collection location on the Atlantic coast of France. EC: samples from three collection locations on the northern Atlantic coast of the United States. IT: samples from two collection locations in Italy.

![Figure 1. Phylogenetic relationships of *C. intestinalis* subpopulations. Consensus sequences for the *col5a1* interval, obtained for each of the six subpopulations analyzed in this study, were used to calculate the population tree. Subpopulations are defined by their collection locations, as in Table 1. The size of the circle surrounding each subpopulation is proportional to the heterozygosity of that subpopulation.](image-url)
Identification of regulatory elements of forkhead and snail

The identification of gene regulatory elements in the proximity of two early-development genes, forkhead and snail, was sought using C. intestinalis sequence comparisons. Both genes are expressed in the larval stage of C. intestinalis development and are therefore amenable to in vivo assessments in transgenic C. intestinalis tadpoles. The mutation rate ratio plot for the forkhead S′ regulatory region revealed five distinct minima representing the sequences likely under the strongest selective constraints (Fig. 2A, regions 1, 2, 4, 5, and 7). We explored the ability of these five regions to function as enhancers in vivo using reporter constructs assayed in transgenic C. intestinalis tadpoles. Constructs which reproducibly drove expression in a tissue-specific manner included that for region 1 in the notochord and endoderm, region 4 in the notochord, endoderm, and neural tube, region 5 in the notochord, and region 7 in the neural tube, endoderm, and notochord (Fig. 2B). These patterns are consistent with the endogenous forkhead expression characteristics (Corbo et al. 1997a). As putative negative controls, we examined two regions with high mutation rates (regions 3 and 6, Fig. 2A). Consistent with the expectation that fast-evolving regions likely lack gene regulatory activity, these two regions failed to drive gene expression in this assay (data not shown).

While the slow-evolving region 2 failed to show tissue-specific expression in this assay, a previous study, analyzing a reporter construct containing the forkhead S′ region through sequential deletions, had shown that removal of a segment containing region 2 abolished neural tube expression in the C. intestinalis tadpole assay (Di Gregorio et al. 2001). This suggests that this functional element might require interactions with nearby enhancers for driving gene expression in the neural tube. The same previous study also demonstrated that deletions corresponding to regions 1 and 3 resulted in a loss of neural tube, endoderm, and notochord expression, in a manner consistent with our results.

We applied a similar analysis to the genomic interval containing snail’s S′ region. One S′ noncoding region characterized by a mutation rate similar to that of forkhead enhancers was also investigated using the transgenic C. intestinalis tadpole assay (region 2, Fig. 3). The reporter construct for this region drove expression in the neuronal lineage (inset, Fig. 3). The other S′ noncoding region showing a similarly low variation rate (region 1, Fig. 3) corresponded exactly to the previously described minimal promoter/mesodermal enhancer of snail (Erives et al. 1998). The mutation rate ratio plot for this interval also correctly identified the first exon of snail (region E, Fig. 3).

Identification of coding exons of col5a1 and patched

To further test our ability to predict the location of exonic sequences through interspecies sequence comparisons, we investigated two genomic intervals containing several exons of the col5a1 and patched genes. Both intervals revealed a similarly uneven distribution of mutation rates, with several regions characterized by a low rate interspersed among high mutation rate regions (Fig. 4). Of the six regions with low mutation in the col5a1 interval, four corresponded to the exon-coding regions annotated by gene prediction programs and EST sequences available for this interval (1–4, Fig. 4A). The two additional regions of very low mutation rate revealed by the variation plot were not functionally investigated in this study but could represent additional col5a1 regulatory elements. Consistent with the results for the col5a1 interval, the position of the eight patched exons were all coincident with regions of low mutation rate (Fig. 4B). Overall sequence diversity was very low in the patched interval (total tree length = 0.10 substitutions per site), likely due to the successful amplification of DNA from only 20 individuals exclusively from California or Japanese locations, but was nevertheless sufficient to identify regions encompassing the exon locations.
Discussion

In contrast to all previous comparative genomic studies, which used evolutionarily fixed sequence differences between two or more species to estimate mutation rates, we are here able to show that intraspecies sequence polymorphism can be effectively used to identify gene regulatory elements and exons through a combination of phylogenetic analysis and polymorphism diversity. Intraspecies sequence polymorphism has been extensively used by population geneticists to study the action of selection on protein coding sequences and by biologists to detect linkage disequilibrium and associate deleterious allelic variants with disease. Our results extend the interest in using polymorphism beyond those applications and suggest that intraspecies sequence comparisons may be useful for identifying functional regions in a genome.

Intraspecies comparisons fill several important niches in the comparative genomics analysis of sequenced genomes. First, they have the potential to identify sequences underlying biological traits unique to a single species. They are also useful when a species under investigation is too far away from its nearest evolutionary neighbor for useful pair-wise comparisons. For example, *C. intestinalis* shared a last common ancestor with its sister species, *C. savignyi*, ∼100 million years ago. Consequently, these two species show limited sequence similarity that, while often sufficient to identify many highly constrained sequences, might lead to failure in identifying many other functional sequences, which can otherwise be detected by the approach described here. Finally, intraspecies comparisons suggest a path for the annotation of organisms occupying previously uncharacterized phylogenetic branches of the animal kingdom, including the green alga *C. reinhardtii*, the diatom *T. pseudonana*, and human malaria parasite *P. falciparum*, whose sequences are now becoming available.

Because of the high rate of allelic polymorphism of *C. intestinalis*, the sequences from as few as 30 individuals were required to achieve sufficient total sequence variation to identify intervals evolving significantly more slowly than their surrounding regions. Applying the same intraspecies approach to human is made more difficult by its low rate of polymorphism (Sachidanandam et al. 2001). The complexity of human population dynamics and haplotype structure complicates estimating the number of individuals required for such a study. Nonetheless, making several simplifications and assumptions, extrapolation of the data from Yu et al. (2002) suggests that sequencing of a few thousand individuals would yield 0.3 single nucleotide polymorphism (SNP)/site, a number comparable to the total tree length of the sequences used in the analysis of the forkhead 5′ region in this study. Irregardless of the exact number of humans required for such a study, with the increase in human genome resequencing predicted for the near future (Collins et al. 2003; Shendure et al. 2004), intra- *Homo sapiens* comparisons exploiting the approaches described here for *C. intestinalis* may prove fruitful for the identification of functional sequences shared with other species as well as human-specific ones.

Figure 3. Mutation rate analysis of the genomic interval containing the 5′ region of the *col5a1* (A) and *patched* (B) genes. The plot was drawn as described in the Figure 2 legend. The sequence of 36 and 22 individuals was used to generate the *col5a1* and *patched* plots (total tree lengths were 0.69 and 0.10), respectively. The blue bar labeled “P” indicates the position of *col5a1*’s promoter; the numbered green bars indicate the position of exons 1–4 of *col5a1* and exons 18–25 of *patched*.

Figure 4. Mutation rate analysis of the genomic interval containing the 5′ region of the *col5a1* (A) and *patched* (B) genes. The plot was drawn as described in the Figure 2 legend. The sequence of 37 individuals (total tree length = 0.52) was used to generate this plot. The position of the first exon is indicated by the green bar labeled “E”; region 1 is *snail*’s promoter, and region 2 is a constrained interval upstream of *snail*. The inset shows the transgenic analysis of region 2. *C. intestinalis* larvae were electroporated with a reporter construct containing region 2, and the expression was visualized by histochemical staining with X-gal. The red arrow indicates expression in the neural tube.
Methods

Collection of *C. intestinalis* specimens

Specimens were collected from the following coastal locations: South San Francisco, California; Half-Moon Bay, California; Santa Barbara, California; Cobscook Bay, Maine; Darling Marine Center, Maine; Woods Hole, Massachusetts; Kobe, Japan; Kochi, Japan; Marlborough Sounds, New Zealand; Reading, England; Roscoff, France; Grevelingen, Netherlands; Den Hoser, Netherlands; Viaggio Coppola, Italy; Fusaro, Italy.

Direct resequencing of target regions

Genomic DNA was isolated from *C. intestinalis* muscle tissue using the Puregene kit (Genta Systems). Target regions were amplified by PCR using the Elongase systems (Invitrogen) following the manufacturer’s recommendations, with primers described in the Supplemental information. Single-band PCR products were gel-purified (SNAP UV-free Gel purification Kit, Invitrogen) and subjected to direct microsequencing using custom primers designed every 250 bp on each strand. Fluorescence automated DNA sequencing was carried out using BytDye chemistry in an ABI3700 Sequencer (Applied Biosystems). Both the (+) and (−) strands were sequenced at least twice. Base calling, quality assessment, and assembly were carried out using the phred, Phrap, Consed software suite developed by Paul Green (www.phrap.org). All the sequences generated in this study have been submitted to GenBank.

Data analysis

Sequences were aligned using MAVID (Bray and Pachter 2003) (baboon.math.berkeley.edu/mavid) or ClustalW (www.ebi.ac.uk/clustalw/index.html) and the alignments were manually verified. Consensus sequences for individuals collected from California, Japan, New Zealand, East Coast, and Europe were derived from the multiple alignment. Maximum likelihood phylogeographic trees for the consensus sequences were reconstructed using FastDNAML. In order to identify conserved sequences, a likelihood ratio (conserved vs. nonconserved) was calculated for each position in the multiple alignment of the individuals (McAuliffe et al. 2004) (http://bonaire.lbl.gov/newshadower/).

To explicitly account for both the phylogenetic relationships between the consensus sequences, as well as the polymorphic diversity between individuals from the same subpopulation, a phylogeographic likelihood computation was coupled with a nucleotide diversity calculation for both fast and slow rates (Pamilo et al. 1987). Although nucleotide diversity is typically computed with a Jukes-Cantor correction, it can be viewed as a phylogeographic likelihood computation on a star tree and can be based on any evolutionary model. We matched the model to that used for the phylogenetic computations on the consensus sequences (Boffelli et al. 2003), and by conditioning on the consensus sequences we obtain a probabilistic model for modeling phylogeographic relationships between groups as well as polymorphic differences among same-population individuals. The rates for the slow (conserved) and fast (nonconserved) computations were based on previous estimates. Regions characterized by the slowest rates were defined relative to the rates of their surroundings, implicitly accounting for the local rate of mutation.

Maintenance of *C. intestinalis* colony

Adult *C. intestinalis* were collected from several locations in Northern California, purchased from Woods Hole, Massachusetts, and Long Beach, Southern California. The animals were kept at 18°C in recirculating artificial seawater.

Construction of plasmids and electroporation

All constructs were made with the pCES (plasmid Ciona enhancer screen) vector previously reported (Harafujii et al. 2002). Different fragments from the Ci-fkh 5′ and Ci-snail 5′ flanking regions were amplified by PCR using the Ciona genomic DNA isolated from Northern California animals, with primers described in the Supplemental information. The DNA fragments were then ligated into the BamHI site of the pCES vector. Electroporations, fixation, and staining reactions were done as described by Corbo et al. (1997b). Aliquots containing 100 μg of purified plasmid were used in each electroporation. Each transgene was tested at least twice. The number of positive embryos scored in each experiment ranged –between 70 and 200.

GenBank accession numbers


Acknowledgments

We thank Shigeki Fujiwara, Arjan Gittenberger, Kevin Heisman, Helene Huelvan, Di Jiang, Shungo Kano, Aimee Philipp, Andy Sexton, and Seb Shimeld for providing *C. intestinalis* samples. Research was conducted at the E.O. Lawrence Berkeley National Laboratory and at the Joint Genome Institute, with support by grants from the Programs for Genomic Application, NHLBI (E.M.R.) and NIH (L.P.), and performed under Dept. of Energy Contract DE-AC0378SF00098, Univ. of California.

References


Web site references


www.phrap.org; consed suite.

www.ebi.ac.uk/clustalw/index.html; multiple sequence alignment.

baboon.math.berkeley.edu/mavid; multiple sequence alignment.

Received May 18, 2004; accepted in revised form October 5, 2004.
Intraspecies sequence comparisons for annotating genomes

Dario Boffelli, Claire V. Weer, Li Weng, et al.

Genome Res. 2004 14: 2406-2411
Access the most recent version at doi:10.1101/gr.3199704

Supplemental Material

http://genome.cshlp.org/content/suppl/2004/11/15/gr.3199704.DC1

References

This article cites 18 articles, 9 of which can be accessed free at:
http://genome.cshlp.org/content/14/12/2406.full.html#ref-list-1

Creative Commons License

This article is distributed exclusively by Cold Spring Harbor Laboratory Press for the first six months after the full-issue publication date (see http://genome.cshlp.org/site/misc/terms.xhtml). After six months, it is available under a Creative Commons License (Attribution-NonCommercial 3.0 Unported License), as described at http://creativecommons.org/licenses/by-nc/3.0/.

Email Alerting Service

Receive free email alerts when new articles cite this article - sign up in the box at the top right corner of the article or click here.

To subscribe to Genome Research go to:
http://genome.cshlp.org/subscriptions

Cold Spring Harbor Laboratory Press