Intra-species 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. Rubin[1,2,*]

[1] DOE Joint Genome Institute

Walnut Creek, CA 94598

[2] Genomics Division

Lawrence Berkeley National Laboratory

Berkeley, CA 94720

[3] Department of Mathematics

University of California, Berkeley

Berkeley, CA 94720

[*]to whom correspondence should be addressed at: emrubin@lbl.gov

ABSTRACT

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 intra-species sequence comparisons for genome annotation. A large number of *C. intestinalis* specimens were collected from four continents and a set of genomic intervals 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 and 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.

The sequence data from this study has been submitted to GenBank under accession nos. AY667278-AY667407.

INTRODUCTION

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 (Hardison 2003; Hartl and Clark 1997). An important limitation of inter-species comparisons is that they can only be used to identify sequences underlying biological traits shared by the species examined. While 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. Intra-species comparisons of the genomes of numerous members of the same species offer a theoretical strategy to tackle this problem, and would also provide a convenient approach to the annotation of the genomes of unusual organisms who are not likely to have the genomes of related species sequenced for comparative analysis. While 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 intra-species comparisons for genome annotation can be supported by experimental data.

The sequencing of the ascidian *C. intestinalis* genome has 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 experimentally characterized their functional roles. These studies illustrate that slowly evolving regions correspond to protein-coding or enhancer regions, indicating the feasibility of using intra-species 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. While 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 approximately 4kb 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 (Di Gregorio et al. 2001; Erives et al. 1998).

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. We were able to obtain amplified genomic targets from approximately 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 approximately every 250bp 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 Fig. 1. Ciona 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 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.

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* 5' regulatory region revealed five distinct minima representing the sequences likely under the strongest selective constraints (Fig 2A, regions 1, 2, 3, 4 and 5). 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 3 in the notochord, endoderm and neural tube, region 4 in the notochord and region 5 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 N1 and N2, 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.

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* 5' region through sequential deletions, had shown that removal of a segment containing region 2 abolished neural tube expression in the *C. intestinalis* tadpole(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 deletion of regions 1 and 3 resulted in a loss of neural tube, endoderm and notochord expression consistent with our results.

We applied a similar analysis to the genomic interval containing *snail*'s 5' region. One *5'* 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 5' 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 *collagen* and *patched*

To further test our ability to predict the location of exonic sequences through intra-species sequence comparisons, we investigated two genomic intervals containing

several exons of the *col5al* 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 *col5al* 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 variation plot were not functionally investigated in this study but could represent additional *col5a1* regulatory elements. Consistent with the results for the *col5al* 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 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 intra-species sequence polymorphism can be effectively used to identify gene regulatory elements and exons through a combination of phylogenetic analysis and polymorphism diversity. Intra-species 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 show the usefulness of intraspecies sequence comparisons for identifying functional regions in a genome.

Intra-species 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*, approximately 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, intra-species comparisons suggests 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 then their surrounding regions. Applying the same intra-species 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. (Yu et al. 2002) suggests that sequencing of a few thousands individuals would yield 0.3 snp/site, a number comparable to the total tree length of the sequences used in the analysis of *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.

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, Massachussetts; Kobe, Japan; Kochi, Japan; Marlborough Sounds, New Zealand; Reading, England; Roscoff, France; Grevelingen, Netherlands; Den Hosse, Netherlands; Viaggio Coppola, Italy; Fusaro, Italy.

Direct resequencing of target regions. Genomic DNA was isolated from *C. intestinalis* muscle tissue using the Puregene kit (Gentra Systems). Target regions were amplified by PCR using the Elongase systems (Invitrogen) following the manufacturer's recommendations, with primers described in the Supplementary 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 250bp on each strand. Fluorescence automated DNA sequencing was carried out using BygDye 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 Phil 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 phylogenetic trees for the consensus

sequences were reconstructed using FastDNAml. In order to identify conserved sequences, a likelihood ratio (conserved vs. non-conserved) 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 phylogenetic 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 phylogenetic 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 phylogenetic relationships between groups as well as polymorphic differences among same-population individuals. The rates for the slow (conserved) and fast (non-conserved) 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 18C in re-circulating artificial seawater.

Construction of Plasmids and Electroporation. All constructs were made with the pCES (plasmid Ciona enhancer screen) vector previously reported (Harafuji 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 Supplementary 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. (Corbo et al. 1997b). Aliquots containing 100μg 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 from approximately 70 to 200.

GenBank Accession Numbers. Forkhead region: AY667314-AY667347. Snail region: AY667371-AY667407. Col5a1region: AY667278-AY667313. Patched region: AY667348-AY667370.

Acknowledgements We thank Shigeki Fujiwara, Arjan Gittenberger, Kevin Heasman, Helene Huelvan, Di Jiang, Shungo Kano, Aimee Phillippi, 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 a grant from the Programs for Genomic Application, NHLBI (E.M.R.); a grant from NIH (L.P.); and performed under Department of Energy Contract DE-AC0378SF00098, University of California.

Figure legends

Fig. 1. Phylogenetic relationships of *C. intestinalis* subpopulations. Consensus sequences for the *col5a1* interval, obtained for each of the six subpopulation 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.

Fig. 2. Panel A. Mutation rate analysis of the genomic interval containing the 5' region of the *forkhead* gene. The x-axis represents the position in the multiple alignment consensus sequence, the y-axis the log likelihood ratio for a fast- over a slow-mutation regime at that position. The plot is smoothed using a 20%-trimmed mean over the 24-base window centered at each aligned site. A lower ratio indicates a low mutation rate. The sequence of 33 individuals (total tree length = 0.28) was used to generate this plot. The blue bar labeled P indicates the position of the *forkhead* promoter, the red and purple bars indicate the position of low- and high-mutation rate intervals, respectively, that were functionally analyzed in this study. Panel B. Transgenic analysis of intervals identified by mutation rate analysis of the 5' region of the *forkhead* gene. *C. intestinalis* larvae were electroporated with a reporter construct containing the genomic fragments 1, 2, 3, 4, and 5, respectively and the expression was visualized by histochemical staining with X-gal. Red arrows indicate expression in the neural tube, yellow arrows in the notochord and green arrows in the endoderm. Constructs for region 2 failed to yield tissue-specific expression.

Fig. 3. Mutation rate analysis of the genomic interval containing the 5' region and the first exon of the *snail* gene. The plot was drawn as described in the Fig 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.

Fig. 4. Mutation rate analysis of the genomic interval containing the 5' region of the *col5a1* (Panel A) and *patched* (Panel B) genes. The plot was drawn as described in the Fig 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*.

Fig. 1

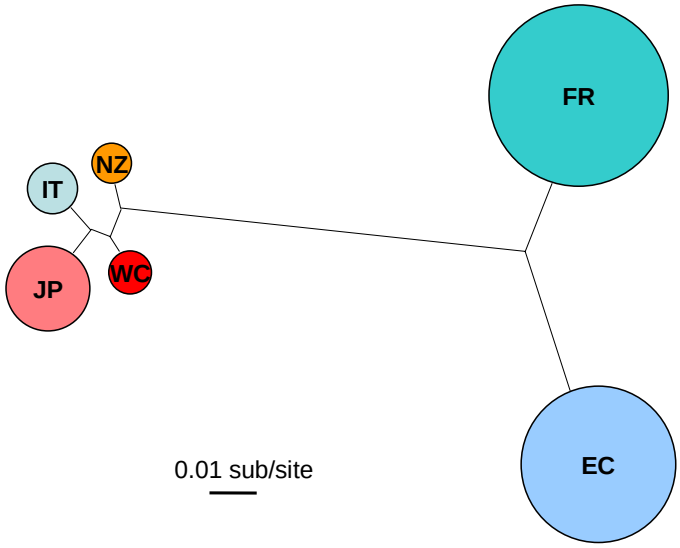


Fig. 2A

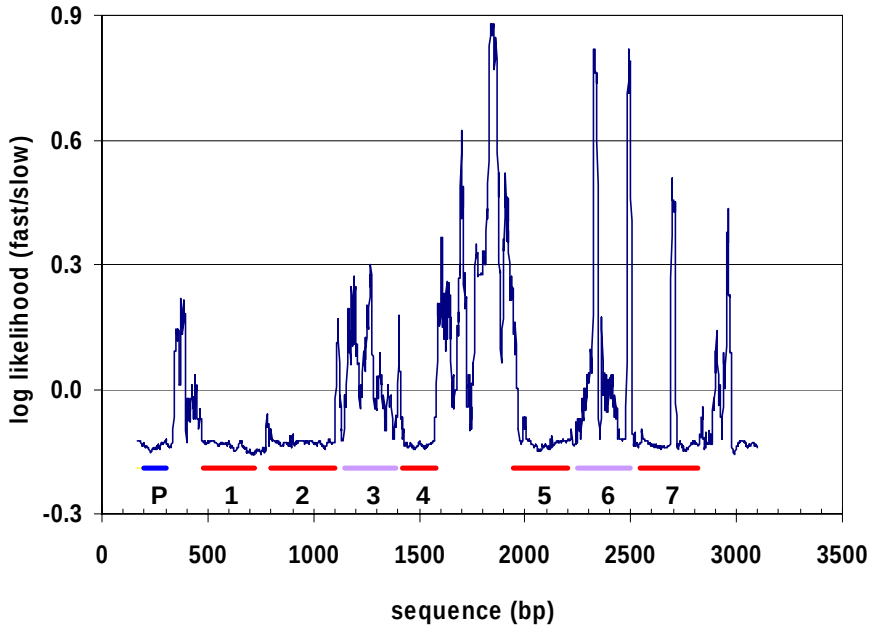

0.9
0.6
0.3
0.0
-0.3
log likelihood (fast/slow)
P
1
2
3
4
5
6
7
0
500
1000
1500
2000
2500
3000
3500
sequence (bp)

Fig. 2B

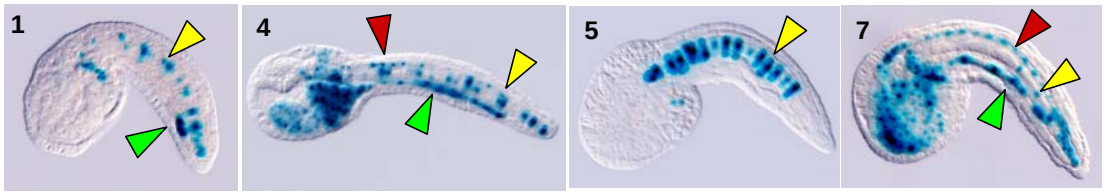

1
4
5
7

Fig. 3

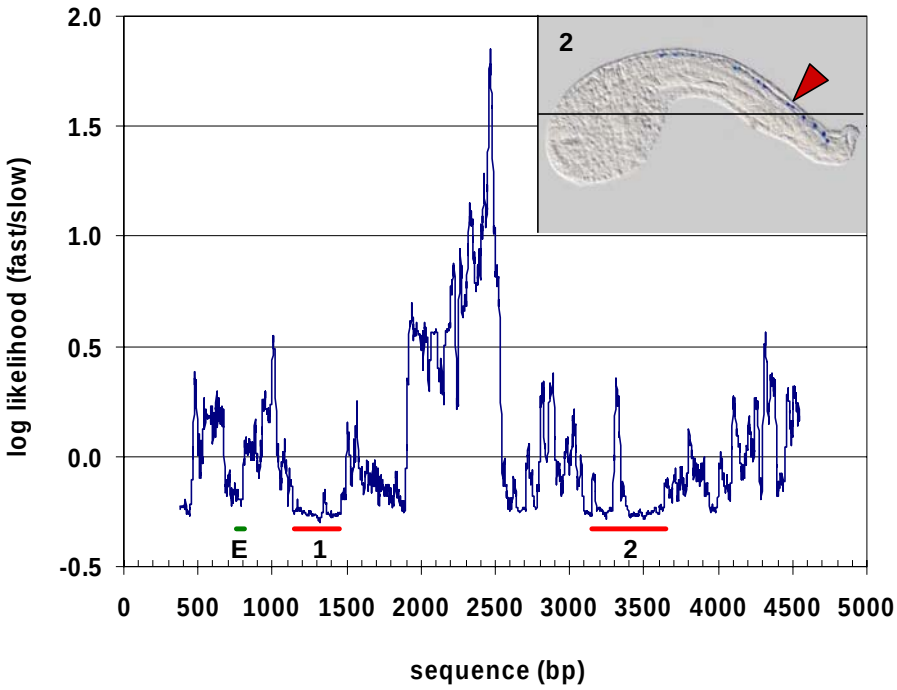

2.0
1.5
1.0
0.5
0.0
-0.5
log likelihood (fast/slow)
0
500
1000
1500
2000
2500
3000
3500
4000
4500
5000
sequence (bp)
E
1
2
2

Fig. 4A

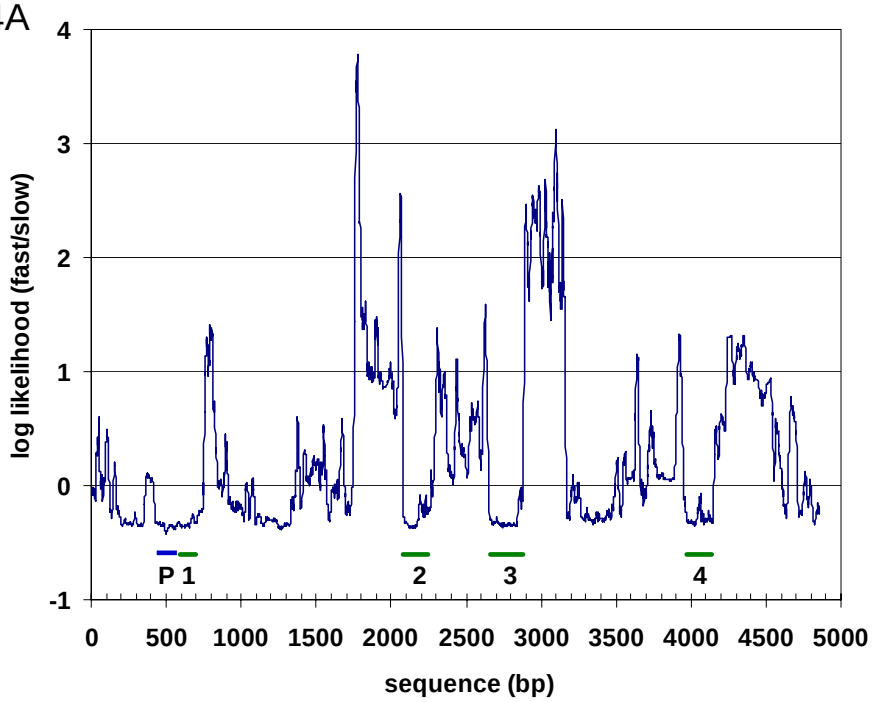
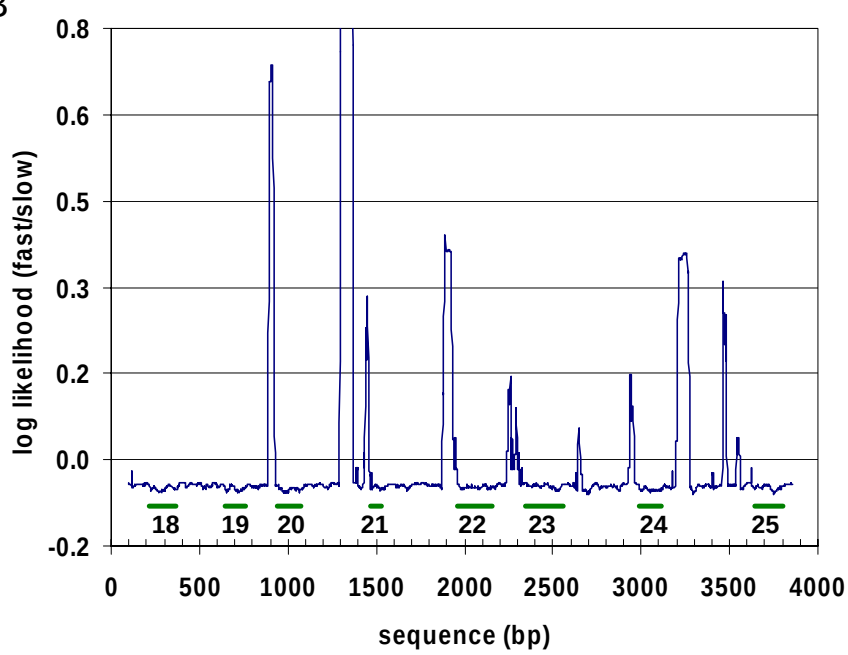

log likelihood (fast/slow)
4
3
2
1
0
-1
P 1
2
3
4
0
500
1000
1500
2000
2500
3000
3500
4000
4500
5000
sequence (bp)

Fig. 4B


log likelihood (fast/slow)
0.8
0.6
0.5
0.3
0.2
0.0
-0.2
18
19
20
21
22
23
24
25
0
500
1000
1500
2000
2500
3000
3500
4000
sequence (bp)

	Forkhead	Snail	Col5a1	Patched
WC	16 (.48)	16 (.48)	17 (.52)	19 (.58)
JP	5 (.25)	6 (.30)	4 (.20)	4 (.20)
NZ	3 (.21)	1 (.7)	5 (.36)	0 (.0)
EC	8 (.17)	8 (.17)	5 (.11)	0 (.0)
FR	1 (.7)	4 (.27)	3 (.20)	0 (.0)
IT	0 (.0)	2 (.17)	2 (.17)	0 (.0)

Table 1. Summary of PCR amplification of *forkhead, snail, col5a1* and *patched*. Number of individuals amplified from each subpopulation for each of the four target regions analyzed in this study. The numbers within parentheses indicate the fraction of individuals that yielded successful amplification out of all the amplifications attempted from that subpopulation. WC: samples from 3 collection locations on the pacific coast of the United States. JP: samples from 2 collection location in western Japan. NZ: samples from 1 collection location in New Zealand. FR: samples from 1 collection location on the Atlantic coast of France. EC: samples from 3 collection locations on the northern Atlantic coast of the United States. IT: samples from 1 collection location in Italy.

References

Ansari-Lari, M.A., J.C. Oeltjen, S. Schwartz, Z. Zhang, D.M. Muzny, J. Lu, J.H. Gorrell, A.C. Chinault, J.W. Belmont, W. Miller et al. 1998. Comparative sequence analysis of a gene-rich cluster at human chromosome 12p13 and its syntenic region in mouse chromosome 6. *Genome Res* **8:** 29-40.

Boffelli, D., J. McAuliffe, D. Ovcharenko, K.D. Lewis, I. Ovcharenko, L. Pachter, and E.M. Rubin. 2003. Phylogenetic shadowing of primate sequences to find functional regions of the human genome. *Science* **299:** 1391-1394.

Bray, N. and L. Pachter. 2003. MAVID multiple alignment server. *Nucleic Acids Res* **31:** 3525-3526.

Collins, F.S., E.D. Green, A.E. Guttmacher, and M.S. Guyer. 2003. A vision for the future of genomics research. *Nature* **422:** 835-847.

Corbo, J.C., A. Erives, A. Di Gregorio, A. Chang, and M. Levine. 1997a. Dorsoventral patterning of the vertebrate neural tube is conserved in a protochordate. *Development* **124:** 2335-2344.

Corbo, J.C., M. Levine, and R.W. Zeller. 1997b. Characterization of a notochord-specific enhancer from the Brachyury promoter region of the ascidian, Ciona intestinalis. *Development* **124:** 589-602.

Dehal, P., Y. Satou, R.K. Campbell, J. Chapman, B. Degnan, A. De Tomaso, B. Davidson, A. Di Gregorio, M. Gelpke, D.M. Goodstein et al. 2002. The draft genome of Ciona intestinalis: insights into chordate and vertebrate origins. *Science* **298:** 2157-2167.

Di Gregorio, A., J.C. Corbo, and M. Levine. 2001. The regulation of forkhead/HNF-3beta expression in the Ciona embryo. *Dev Biol* **229:** 31-43.

Erives, A., J.C. Corbo, and M. Levine. 1998. Lineage-specific regulation of the Ciona snail gene in the embryonic mesoderm and neuroectoderm. *Dev Biol* **194:** 213-225.

Harafuji, N., D.N. Keys, and M. Levine. 2002. Genome-wide identification of tissue-specific enhancers in the Ciona tadpole. *Proc Natl Acad Sci U S A* **99:** 6802-6805.

Hardison, R.C. 2003. Comparative Genomics. *PLoS Biol* **1:** E58.

Hartl, D.L. and A.G. Clark. 1997. *Principles of Population Genetics*. Sinauer Associates.

McAuliffe, J.D., L. Pachter, and M.I. Jordan. 2004. Multiple-sequence functional annotation and the generalized hidden Markov phylogeny. *Bioinformatics***:** bth153.

Nobrega, M.A., I. Ovcharenko, V. Afzal, and E.M. Rubin. 2003. Scanning human gene deserts for long-range enhancers. *Science* **302:** 413.

Pamilo, P., M. Nei, and W.H. Li. 1987. Accumulation of mutations in sexual and asexual populations. *Genet Res* **49:** 135-146.

Sachidanandam, R., D. Weissman, S.C. Schmidt, J.M. Kakol, L.D. Stein, G. Marth, S. Sherry, J.C. Mullikin, B.J. Mortimore, D.L. Willey et al. 2001. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* **409:** 928-933.

Satoh, N. 2003. The ascidian tadpole larva: comparative molecular development and genomics. *Nat Rev Genet* **4:** 285-295.

Shendure, J., R. Mitra, C. Varma, and G. Church. 2004. Advanced Sequencing Technologies: Methods and Goals. *Nat Rev Genet* **5:** 335-343.

Thomas, J.W., J.W. Touchman, R.W. Blakesley, G.G. Bouffard, S.M. Beckstrom-Sternberg, E.H. Margulies, M. Blanchette, A.C. Siepel, P.J. Thomas, J.C. McDowell et al. 2003. Comparative analyses of multi-species sequences from targeted genomic regions. *Nature* **424:** 788-793.

Yu, N., F.C. Chen, S. Ota, L.B. Jorde, P. Pamilo, L. Patthy, M. Ramsay, T. Jenkins, S.K. Shyue, and W.H. Li. 2002. Larger genetic differences within africans than between Africans and Eurasians. *Genetics* **161:** 269-274.

Supplementary dats

Primers sequences

1. Amplification of genomic target regions

```
Forkhead:
FORKHEAD.EXT.f: TCATGCAAGGAATGCGTGTT
FORKHEAD.EXT.r: CACAGGTGCACGCGAGATAG

FORKHEAD.INT.f: GGGTGTTGGATGGCGAACTA
FORKHEAD.INT.r: TGACCACTCTCCACCGCTTT

Snail:
SNAIL.EXT.f: CTGTCCTGAGGGCCGTAAAT
SNAIL.EXT.r: CCGGTAATACACCTCCTGCAT

SNAIL.INT.f: AAGCAGCAGTTGTGCCGAGT
SNAIL.INT.r: CCAGCTCCCATAGCAGCATT

Patched:
PATCHED.EXT.f: TTTGGGGTCCCTGCTTTTTA
PATCHED.EXT.r: TCCTTCGCCTTGGACCTGT

PATCHED.INT.f: AGGAAATAATATGATGCTAATGGT
PATCHED.INT.r: TGACCATAATTAACGCTTGTTC

Col5A1:
COL5A1.EXT.f: CTTGGCCTAGGCTTATGACG
COL5A1.EXT.r: TTCTGGTTCTTCCCTGATGG

COL5A1.INT.f: CGAGAAGGCAAGCCAGAATA
COL5A1.INT.r: GGATGCTGGGAAACACATCT
```

2. Amplification of candidate functional regions for subcloning into reporter construct

```
Forkhead Region 1
Fkh_1f: CCCTCGAGTGTTTGTTCACTAAGAAATGATAA
Fkh_1f: TGGATATCTTATTCTACTGAATATTTGGCGAC

Forkhead Region 2
Fkh_2f: CCCTCGAGAAGGAAATTGGTCTGTTAAGGTGC
Fkh_2f: TGGATATCTCATTATCATTTCTTAGTGAACAA

Forkhead Region 3
Fkh_3f: CCCTCGAGGACTTAAATTTTGAAAGGAAATTGGT
Fkh_3f: TGGATATCATAAAATGGAGCATCATAGTTTGGC

Forkhead Region 4
Fkh_4f: CCCTCGAGTCTAATGCACGTTTCTTATCAATG
Fkh_4f: TGGATATCACATTTGCAACTCGTCTTGTCTTT

Forkhead Region 5
Fkh_5f: CCCTCGAGAAGCACAACAGGCGAATTAAAGTC
Fkh_5f: TGGATATCATACGGTTATAAATTATTCAAAGT
```

Forkhead Region N1
Fkh_N1f: CCCTCGAGCTAATTCATGCTGGTTATAAA
Fkh_N1f: TGGATATCACGCATACATTTAAACATAAACTA

Forkhead Region N2
Fkh_N2f: CCCTCGAGTATGTGTACTTTATTTATTT
Fkh_N2f: TGGATATCCATTTAGAACTTGTTTTATTTAAC

Snail Region 2
Snail_2f: CCCTCGAGCGAGAGCAAACTAATAACTGGACT
Snail_2r: TGGATATCATGTACGATATAGAGAATAAACGG