Available online at www.sciencedirect.com
SCIENCE DIRECT
Gene 362 (2005) 98–108
GENE
SECTION
EVOLUTIONARY GENOMICS
www.elsevier.com/locate/gene

Genome complexity and repetitive DNA in metazoans from extreme marine environments

Kevin T. Fielman *, Adam G. Marsh

University of Delaware, College of Marine Studies, 700 Pilottown Rd, Lewes, DE 19958, USA

Received 23 November 2004; received in revised form 1 April 2005; accepted 27 June 2005
Available online 26 September 2005

Abstract

As genomics converges with ecology and evolution to identify the fundamental linkages between genome structure and function, genome and transcriptome complexity will need to be measured in organisms from more diverse habitats, most often in the absence of complete sequence data. Here, we describe the complexity of ten genomes measured by a novel, high-throughput fluorescence-based kinetic hybridization assay. We applied the Shannon information index, H, and a related, fluorescence-adjusted index, H_f, as unique metrics of the hybridization kinetics to complement the conventional rate constant, k. A strong, positive relationship was present between H_f, and the repetitive DNA content of five eukaryotic genomes previously determined by Cot kinetic analyses (*Onchorynchus keta*, *Ilyanassa obsoleta*, *Bos taurus*, *Limulus polyphemus*, *Saccharyomyces cerevisiae*). This relationship was used to characterize the complexity of previously unstudied genomic samples in five metazoan taxa from three marine environments, including deep-sea hydrothermal vents (*Alvinella pompejana*), the temperate subtidal (*Streblospio benedicti*), and Antarctic coastal bays (*Sterechinus neumayeri*, *Odontaster validus*, *Tritonia antarctica*). Contrary to the predictions of nucleotypic theory, Antarctic invertebrates consistently had the lowest quantities of repetitive DNA in conjunction with low metabolic rates and highly protracted rates of cell division and larval development. Conversely, hydrothermal vent species with rapid cell division and growth do not have significantly different genome sizes or particularly low amounts of repetitive DNA as compared to non-vent, deep-sea taxa. Furthermore, there appears to be a positive correlation between the temperature at which the most abundant repetitive sequence classes anneal and habitat thermal stability. Thus, our study reveals a potential shift in repetitive sequence representation between these extreme environments that may be related to genome function in species living at these different thermal regimes.

Keywords: Cot; Shannon index; Genomics; Hydrothermal vent; Antarctic

1. Introduction

Complexity is a fundamental structural property of the genome. Historically, genome complexity has been defined as the genome-wide total length of unique sequence measured in base pairs by Cot analysis (Bernardi, 1965; Britten and Kohne, 1968; Wetmur and Davidson, 1968; Britten and Davidson, 1971). Division of the genome into unique and repetitive sequence fractions resolves the C-value paradox underlying the discrepancy between an organism's total genome size and its perceived biological complexity (Gregory, 2001a,b). This allows more meaningful comparisons of genome structure to be made among species; however, the relationships between complexity and function have not been resolved. The field of environmental genomics is rapidly developing by identifying the

Abbreviations: ABI, Applied Bioinstruments; Cot, acrynom for a kinetic reassociation curve plotted as the amount of ssDNA (y axis) against the product of concentration (Co) and time (x axis); CpG, cytosine+guanine dinucleotide; CTAB, cetyltrimethylammonium bromide; CV, coefficient of variation; C-value, genome size of an organism defined as the total amount of DNA within its haploid chromosome set; qPCR, quantitative polymerase chain reaction; EDTA, ethylenediaminetetraacetic acid; GC, guanine+cytosine; gDNA, genomic DNA; HMW, high molecular weight; Mb, megabase; NMWL, nominal molecular weight limit; PVP, polyvinylpyrrolidone; ρ_p, Pearson's correlation coefficient; RNase, ribonuclease; ss, single stranded; TE, solution containing 10 mM Tris and 1 mM EDTA.

* Corresponding author. Current address: University of California, Santa Barbara, Department of Ecology, Evolution and Marine Biology, Santa Barbara, CA 93106-9610, USA. Tel.: +1 805 680 2248; fax: +1 805 893 4724.
E-mail addresses: fielman@lifesci.ucsb.edu (K.T. Fielman), amarsh@udel.edu (A.G. Marsh).

0378-1119/$ - see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.gene.2005.06.035

significant linkages between genome structure and function in natural populations (Jackson et al., 2002; Feder and Mitchell-Olds, 2003; Purugganan and Gibson, 2003; Thomas and Klaper, 2004). To progress further, this field will require increasingly sophisticated genome structure survey techniques to identify such functional linkages in numerous species under diverse environmental conditions and selective pressures. Because resources are typically limited, full sequencing of genomes and transcriptomes cannot meet this broad-scale need. Thus, more efficient and cost-effective methods are required.

Although the principles of Cot analysis have recently been applied to novel methods addressing fundamental problems of genome characterization (Soares et al., 1994; Peterson et al., 2002a,b), the technique itself has changed little since its inception and is not widely used today for genome characterization. Among the technical barriers to its general use are the requirements for large amounts of DNA (typically hundreds of micrograms) and the skilful execution of thermally controlled hydroxyapatite chromatography. To circumvent these limitations, we have developed a simple, high-throughput, fluorescent kinetic assay for small amounts of DNA that utilizes existing 96-well instrumentation without requiring intervening handling. Fluorometry is common in molecular biology laboratories and forms the basis of both general nucleic acid quantification and quantitative, real-time PCR (qPCR) under varying thermal regimes. By readily adapting a qPCR platform to provide the raw data from which kinetic parameters can be derived, our method minimizes sample handling and size and maximizes the number of assays that can be performed simultaneously.

Here, we use our new technique in an initial investigation of the relationship between genome complexity and the thermal stability of an organism's habitat. Specifically, we examine metazoan taxa that live in marine environments at extreme high and low temperatures of contrasting stability. Marine hydrothermal vent environments represent the upper thermal limits at which life is found. The polychaete worm, *Alvinella pompejana*, is an abundant, endemic member of the vent community and is a thermotolerant metazoan, capable of surviving highly variable temperature exposures up to 80 °C (Cary et al., 1998). In contrast, species inhabiting the stenothermal southern polar seas surrounding Antarctica have adapted to existence at a relatively constant −1.8 °C, just at a lower physiological limit for intracellular freezing. Here, the urchin, *Sterechinus neumayeri*, the starfish, *Odontaster validus*, and the nudibranch mollusc, *Tritonia antarctica*, are prominent members of the shallow water epibenthic community. Survival in both of these extreme habitats requires a specialized suite of morphological, physiological, and biochemical adaptations (Van Dover, 2000; Clark et al., 2004). Although not typically considered “extreme” environments, organisms from temperate intertidal habitats routinely face periodic, and unpredictable, large-scale changes in salinity, temperature, and oxygen tensions as a function of tidal exposure (Helmuth et al., 2002). Thus, we included a temperate, shallow water polychaete (*Streblospio benedicti*) and an intertidal mud snail (*Ilyanassa obsoleta*) in considering the potential impacts of environmental temperature on genome complexity.

2. Methods

2.1. Sample preparation

High molecular weight genomic DNA (HMW gDNA) from bacteria (*Escherichia coli*, strain DH10B), mud snail (*Ilyanassa obsoleta*) foot, Pompeii worm (*Alvinella pompejana*) body wall, and sperm from horseshoe crab (*Limulus polyphemus*) and Antarctic urchin (*Sterechinus neumayeri*) were obtained from liquid-nitrogen frozen samples that were powdered using a mortar and pestle. HMW gDNA from salmon (*Oncorynchus keta*) and calf (*Bos taurus*) were from Sigma-Aldrich Chemical Company (St. Louis, Missouri, USA). The DNAs were extracted using 4.0 M guanidine HCl, 0.5% Sarcosyl, 100 mM Tris, 10 mM EDTA, pH 8.0 buffer and standard phenol–chloroform extraction procedures for HMW gDNA at pH 8.0. A single CTAB/NaCl/chloroform extraction was added when necessary to remove polysaccharides. RNA was removed with RNAse A (Fisher Scientific, Pittsburgh, Pennsylvania, USA, 1 μg ml^{-1} at 37 °C for 1 h). DNA was extracted from whole *Streblospio benedicti* using a CTAB/PVP extraction protocol (Porebski et al., 1997) and overnight digestion. HMW gDNA from yeast (*Saccharyomyces cerevisiae*) was purchased from Novagen (San Diego, California, USA) and used without further purification.

Purified DNAs were dissolved in TE (pH 8.0) and then diluted 10-fold in sterile glycerol. Glycerol increases the mixture's viscosity to facilitate shearing. The DNA in glycerol was sheared with a nebulizer (Invitrogen, Carlsbad, California, USA) at 1.41 kg cm^{-1} (20 psi) in an ice water bath to a mean length of 300–500 bp among species as determined by agarose gel electrophoresis. Glycerol and metal ions were removed by centrifugal dialysis (Centricon Plus-20, Ultrafree-0.5, NMWL = 10 kDa or 3 kDa; Millipore, Billerica, Massachusetts, USA) with a minimum of 5 sample volumes of TE (pH 8.0). DNA in the dialyzed solution was quantified by spectrophotometry and ensured to have an $A_{260/280}$ ratio of 1.8 or greater.

2.2. Annealing curves

Samples were prepared as a master mix that was transferred to a 96-well optical reaction plate using a multi-channel pipetter. Typically, 8 to 24 replicate wells within a plate were used for each sample. These were considered technical replicates. Independent plate preparations on separate days were considered experimental replicates and were most often $n = 3$. Each 20 μl sample in 0.12 M sodium phosphate buffer contained 400 ng sheared genomic DNA and 0.8 μM PicoGreen (Molecular Probes, Inc., Oregon, USA). PicoGreen is a highly sensitive, intercalating cyanine dye that exhibits >1000-fold increase in fluorescence when bound to dsDNA in solution. The dye binds rapidly (within 10 s) and quantitatively over four orders of magnitude in a sequence-independent fashion with

little signal interference from ssDNA (Singer et al., 1997). DNA annealing data were collected on an ABI 7700 qPCR instrument at an excitation wavelength of 488 nm for 25 ms (sensitivity) and an emission wavelength of 520 nm. Following an initial denaturation at 99.9 °C for 2 min the instrument was programmed for stepwise 1 °C decreases from 95 °C to 67 °C each held for 45 min. Preliminary studies indicated that no notable DNA degradation was evident as a large shift in electrophoretic mobility on an agarose gel at the end of the assay.

2.3. Data analysis

Raw data were collected and pre-processed using the ABI Sequence Detection Software v1.7 set at default parameters. A C++ program with a graphical user interface was written to process the fluorescence data automatically for each well (96) and each temperature plateau (30) giving ~2880 kinetic curves per run. Annealing curves for each well at each temperature plateau were iteratively fit to the second-order equation:

$$y = B_0 + B_1(1 - e^{-B_2 t}) \quad (1)$$

in which B_0 is the intercept, B_1 is the average maximum fluorescence value, t is time in minutes, and B_2 defines the rate constant "k", where $k = -1/B_2$ and represents time t where the second derivative of Eq. (1) is equal to zero (i.e., the point at which the rate of reannealing is maximal). Final regression model solutions having an $r^2 < 0.6$ were considered not to be representative of a second-order rate process and were not included in subsequent complexity calculations. The discarded values typically represented less than 12% of all curves and were predominantly found at the highest temperatures where deviations from second-order behaviour might be expected. Parameter estimates that lay beyond the upper and lower inner fences of the data's distribution were considered to be outliers and discarded. Discarded values were not included in the summary statistics calculated by the program. The program provided second-order, iterative solutions that were an excellent fit to the experimental data with at least 85% of the r^2 values greater than 0.9 at the optimal data limit, providing an automatic criterion for assessing data quality across the 2880 kinetic curves.

The C++ program also calculated the Shannon information index (H) (Shannon, 1948) for each well at each temperature for the first derivative of the observed fluorescence vs. time data. H was calculated as:

$$H(X) = -K \sum_{i=1}^{n} p_i \log_2 p_i \quad (2)$$

where p_i is the proportion of the change in fluorescence observed at time interval i relative to the total change in fluorescence, and K is a scalar value (equal to 1 in the Shannon–Weaver calculation, Eq. (2)). Thus, the shape of this function is directly related to the slope and curvature of the kinetic trajectory. To incorporate the overall abundance of double-stranded DNA within a temperature plateau the H statistic was scaled to the maximum fluorescence observed within the well at that temperature ($F_{max} = K$). This scaled index was designated H_f. H and H_f were not calculated for samples that did not fit the second-order regression model (Eq. (1)).

2.4. Data analysis optimization

We chose four criteria to optimize the number of data points (data limit within the C++ program) used to fit the second-order curve model and to calculate the complexity statistics at each temperature plateau. These criteria were: (1) maximize the number of replicate wells retained in the analysis, (2) maximize the kinetic curve fit to the second-order model (r^2), (3) maximize the differences among genomes, and, (4) minimize variability within and among wells. Variability was expressed as the coefficient of variation (CV = standard deviation/mean) to enable comparisons among means of markedly different magnitude (over an order of magnitude in this study). The extent of heteroscedasticity (the presence of unequal variability among means) was measured by the coefficient of variation calculated from the CVs associated with each mean. All criteria were evaluated by visual examination of data plots from a subset of the genomes analyzed.

2.5. Complexity measurements in uncharacterized genomes

The total repetitive DNA content (%) of previously uncharacterized genomes was estimated from the relationship between H_f and the repetitive DNA content of genomes previously studied by kinetic hybridization assays. To account for stochasticity in our measurements and to maximize the function's range, the H_f value used in the calculations was an average across the five highest, contiguous H_f values. Because peak values were not observed for *S. cerevisiae*, an average was calculated across all temperatures. Published C-values were used to estimate chemical complexity from our kinetic complexity measurements by multiplying genome size (Mb) by the percentage of unique sequence in each genome (100% minus repetitive DNA%).

3. Results and discussion

3.1. Annealing curves

The kinetic complexity of ten genomes was measured with a high-throughput fluorescence-based method. By using the highly sensitive, intercalator PicoGreen to monitor the reannealing process of sheared, total genomic DNA, measurements were made in real time on a single sample, rather than at predetermined, discrete time points requiring multiple samples. In a further refinement of the original Cot methodology, reannealing kinetics were determined at 1 °C intervals by holding the samples at a temperature plateau for 45 min while monitoring reannealing rates, and then cooling the sample by 1 °C to the next temperature plateau. Although temperature affects which fragments can form duplex DNA depending on their GC (guanine + cytosine) content, there is little dependence of

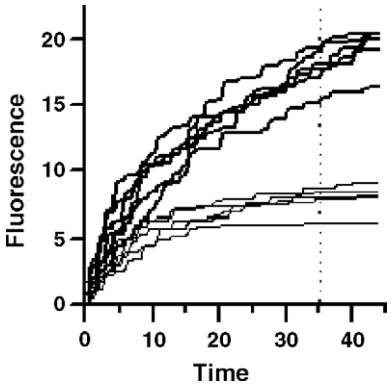

Fig. 1. Formation of double-stranded DNA measured as a change in PicoGreen fluorescence over time at 75 °C for two genomes of markedly different complexity. Salmon genomic DNA (upper, heavy lines) exhibits rapid annealing, indicating low complexity. By contrast, cow genomic DNA (lower, light lines) anneals more slowly, indicating higher sequence complexity. The fluorescence data, collected as simultaneous measurements in replicate wells, have been smoothed and scaled to the baseline fluorescence value at the start of the annealing period. The dotted line indicates the time point selected as the optimal data limit for our complexity calculations.

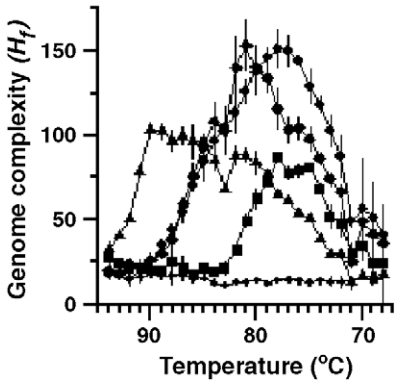

Fig. 3. Relationship between the fluorescence-adjusted complexity index, H_f, and genomic complexity partitioned by reannealing temperature among five genomes previously studied by Cot kinetic analyses. Note that as the magnitude of H_f increases, genome complexity decreases (greater proportion of repetitive sequence). Data are from $n=7$ to 24 technical replicates and $n=3$ experimental replicates per species (total replication of 21 to 72 wells). Error bars are 1 S.D. In their order of decreasing amounts of repetitive sequence: (●) salmon, (♦) mud snail, (▲) cow, (■) horseshoe crab, and (+) yeast.

the reannealing rate at a constant temperature on GC content (Wetmur and Davidson, 1968). Partitioning the sample by GC content in 1 °C annealing intervals reveals far more complexity within a sequence population than has been previously possible to quantify. Another advantage of this method over traditional Cot kinetic analyses is the elimination of hydroxyapatite chromatography, greatly reducing sample handling. These combined factors decrease the amount of DNA required for analysis to as little as 100 ng per replicate and allow for rapid sample measurements in a 96-well microtiter plate format with existing instrumentation. Even with small amounts of DNA, the kinetic annealing profiles we have measured with PicoGreen readily discriminate between genomes of different complexity within an annealing temperature (Fig. 1). Although the sequences that underlie these patterns may not be known, it is still possible to discriminate among genomes and quantify their relative differences in complexity. One disadvantage of our approach is that it may make direct comparison of the Cot values from this method and those from HAP-based techniques more challenging. In this initial study we have not undertaken the comprehensive calibration required to account for the interactive effects of temperature, GC content, and fluorescence output (quantity of reannealed, dsDNA) that might allow for a direct comparison with hydroxyapatite-based Cot approaches. Rather, we have used the technique to generate a unique complexity profile that can be used to ascertain relative differences in complexity among genomes.

We applied the Shannon information index, H (Shannon, 1948), and a modified, fluorescence-adjusted index, H_f, as unique, robust metrics of reannealing kinetics to augment conventional comparisons of the annealing rate constant, k. The H index and the kinetic rate constant for a sample were independently derived from the fluorescence response at each temperature plateau. These unrelated metrics were significantly correlated with each other ($\rho_p=0.757, p<0.001, N=155$; Fig. 2). Thus, an increase in the amount of repetitive sequence results in an increase in k, H, and H_f metrics that each quantifies the decrease in chemical complexity. H_f is scaled by the maximum fluorescence value observed within each temperature plateau and correctly characterizes the kinetic profiles of five genomes

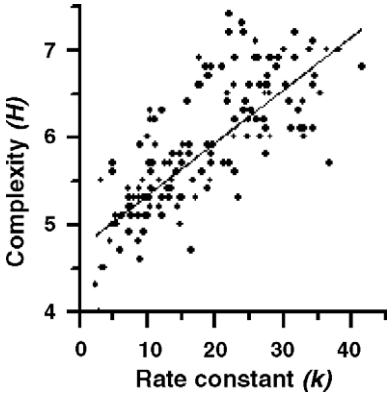

Fig. 2. Correlation between the kinetic rate constant, k, and the Shannon index, H.

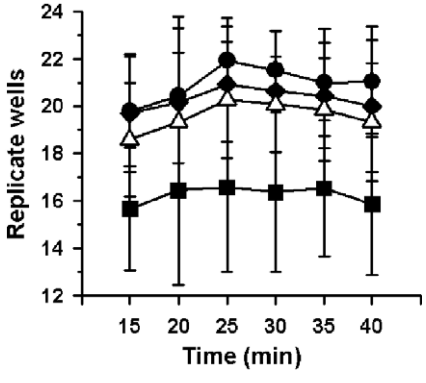

Fig. 4. Average number of replicate wells retained for analysis as a function of the data limit expressed as time. Error bars are 1 S.D. (●) salmon, (♦) mud snail, (▲) cow, (■) horseshoe crab, and (+) yeast.

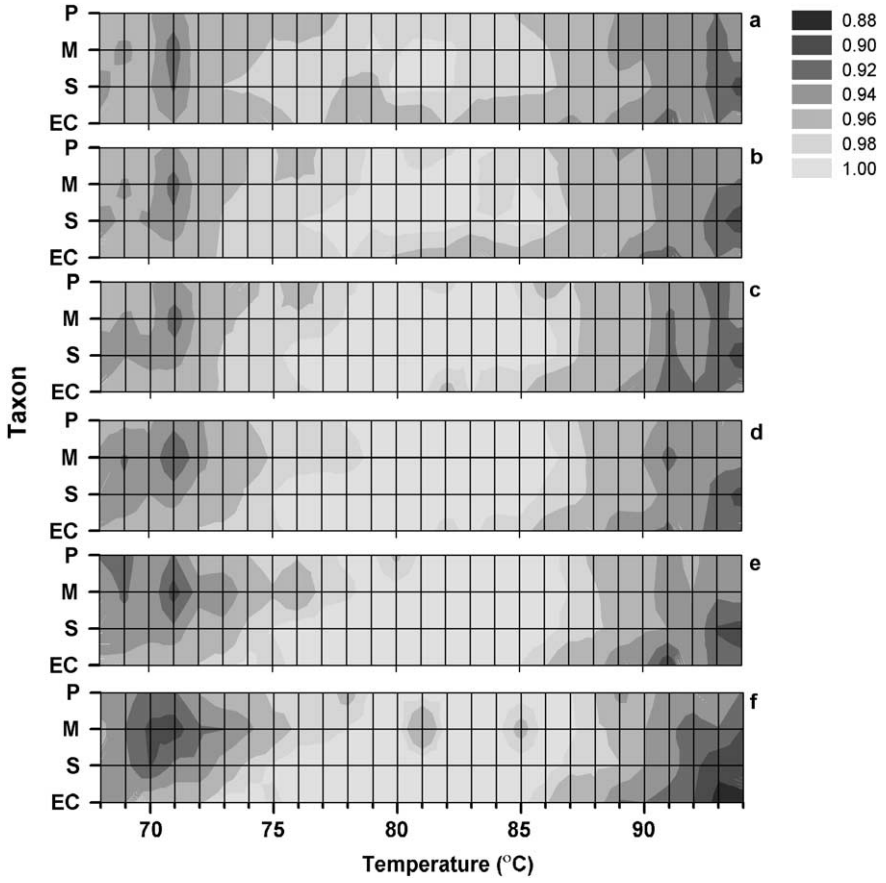

Fig. 5. Surface contour plots of the average r^2 among replicate wells for the second-order curve solutions that describe genomic DNA reannealing at each temperature plateau. Panels (a–f) are results with data limits as time from 15 to 40 min at 5-min intervals. P—Pompeii worm, M—mud snail, S—chum salmon, EC—*E. coli.*

of known complexity in their order of decreasing amounts of repetitive sequence, from chum salmon (*Oncorynchus keta*), mud snail (*Ilyanassa obsoleta*), calf (*Bos taurus*), horseshoe crab (*Limulus polyphemus*), and yeast (*Saccharyomyces cerevisiae*) (Fig. 3).

The Shannon information index (H) provides three major advantages over the annealing rate constant, k, for comparative studies. First, it is computationally simple and is commonly applied to complexity measurements in biological systems (Spellerberg and Fedor, 2003). Second, it is methodologically

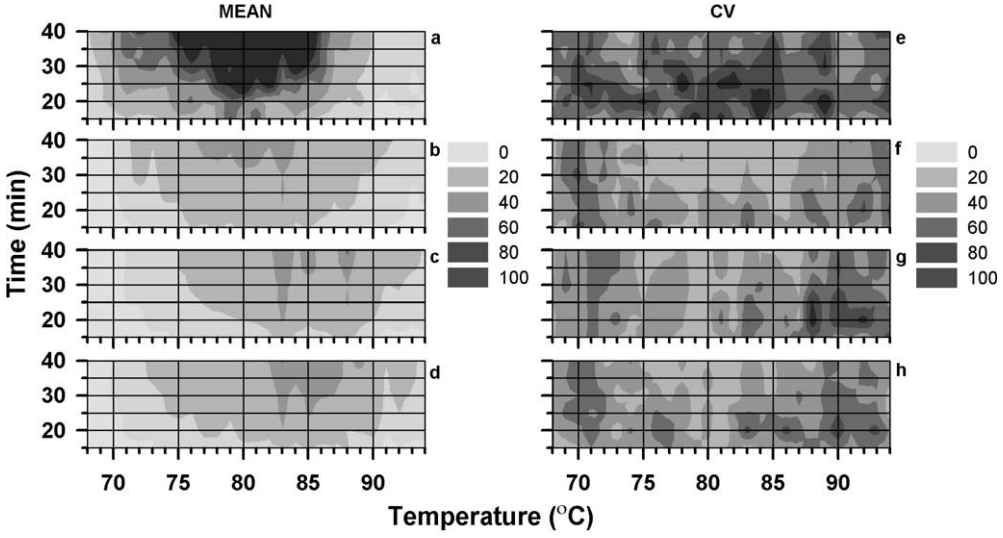

Fig. 6. Mean and coefficient of variation (CV%) surface contour plots of the kinetic rate constant, k, for replicate wells at each annealing temperature plateau with the data limits expressed as time. The data limit describes the number of observations within each plateau that were used in the regression analysis to calculate k. *E. coli* (a, e), chum salmon (b, f), mud snail (c, g), Pompeii worm (d, h).

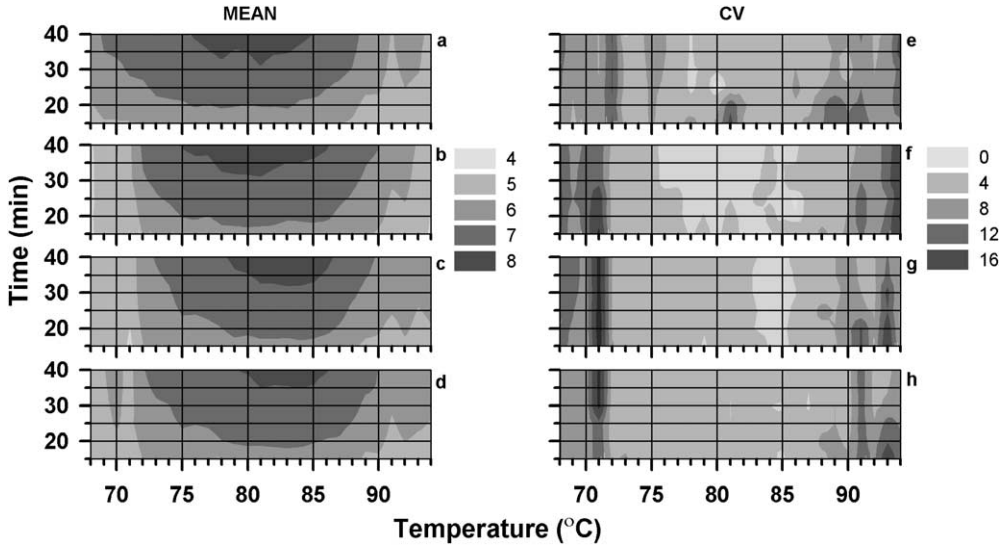

Fig. 7. Mean and coefficient of variation (CV%) surface contour plots of the Shannon complexity index, *H*, for replicate wells at each annealing temperature plateau with data limits expressed as time. The data limit describes the number of observations within each plateau that were used to calculate *H*. *E. coli* (a, e), chum salmon (b, f), mud snail (c, g), Pompeii worm (d, h).

robust and yields lower between-replicate differences in the complexity estimates, decreasing the average CV among these values (see Section 3.2). This greater precision helps reduce the number of replicates required. Third, by scaling the index calculation to sample fluorescence, the resolution of differences in repetitive sequence content is increased among different genome samples (see Section 3.2). We believe that these advantages, and the fact that *H* is strongly correlated with *k* (Fig. 2), warrant use of the index in this context.

3.2. Data analysis optimization

After smoothing, the average number of data points available for analysis within each well at each temperature plateau was 442. The average number of replicate wells (*N*) retained for analysis among temperatures peaked at a data limit of 250, although within a sample type (i.e., species) the average *N* varied by no more than two replicates among limit values (Fig. 4). The maximum average r^2 values were observed at a data limit of 300 (Fig. 5d), except for the mud snail data that exhibited a decreasing r^2 with limit increases. The distribution of r^2 values among temperatures and taxa across data limits indicated that these decreases were concentrated at the highest and lowest temperature plateaus where few differences are observed among the kinetic profiles. These outer regions also had less change in magnitude and greater variability (higher CVs) in the mean value of each complexity metric when compared to the central temperature plateaus (Figs. 6–8).

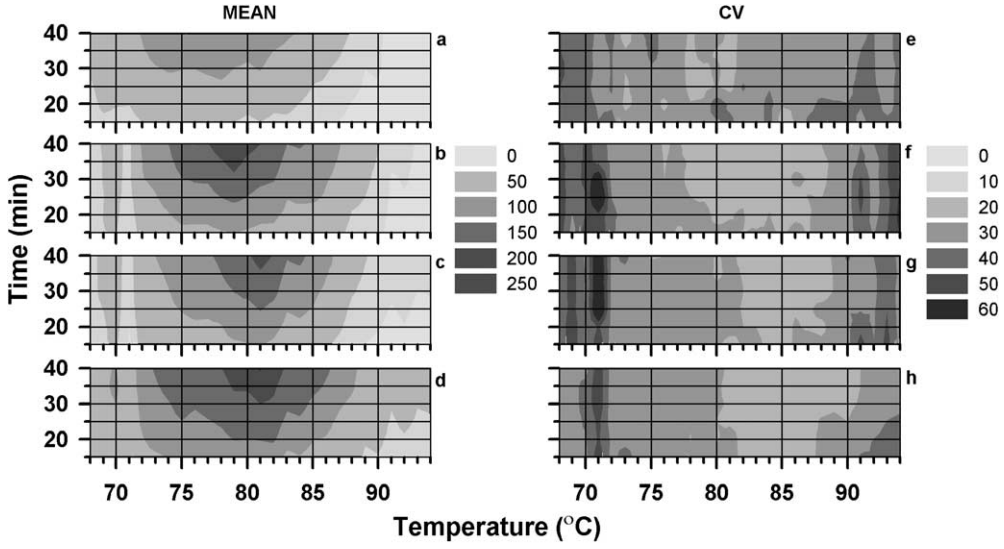

Fig. 8. Mean and coefficient of variation (CV%) surface contour plots of the fluorescence-scaled Shannon complexity index, H_f, for replicate wells at each annealing temperature plateau with data limits expressed as time. The data limit describes the number of observations within each plateau that were used in the regression analysis to calculate H_f. *E. coli* (a, e), chum salmon (b, f), mud snail (c, g), Pompeii worm (d, h).

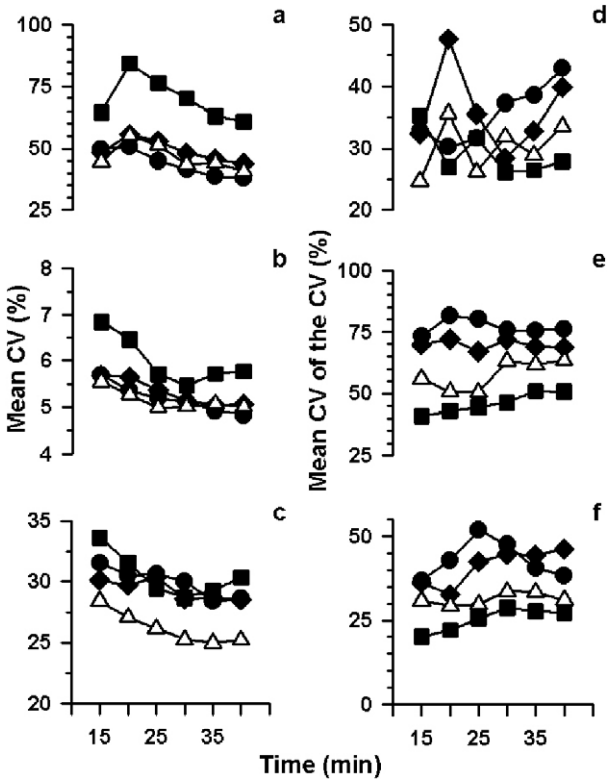

Fig. 9. The mean coefficient of variation (CV) and its corresponding CV among temperature plateaus for the complexity metrics, k (a, d), H (b, e), and H_f, (c, f) at data limits expressed as time. *E. coli* (■), chum salmon (●), mud snail (♦), and Pompeii worm (△). Means are presented without error bars for visual clarity.

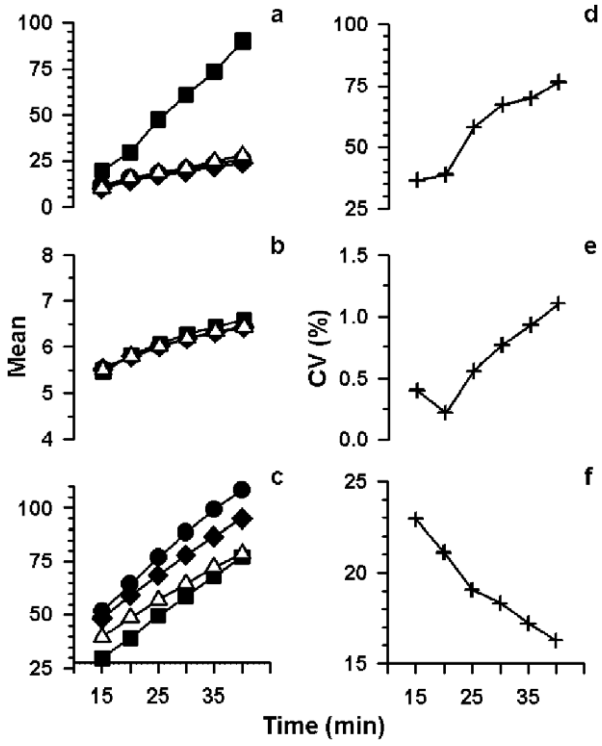

Fig. 10. Change in the mean and the coefficient of variation (CV) among taxa for the complexity metrics, k (a, d), H (b, e), and H_f (c, f) with increasing data limits expressed as time. *E. coli* (■), chum salmon (●), mud snail (♦), and Pompeii worm (△). Means are presented without error bars for visual clarity.

The magnitude and variability of the CVs among temperatures were evaluated from the mean and CV of the CVs (Fig. 9). The average CV exhibited a net decrease with an increasing data limit among all complexity indices (Fig. 9a, b, c). A minimum CV was typically observed at a data limit of 300 or 350. No consistent pattern of variation among the CVs (i.e., systematic heteroscedasticity) was evident, although the majority of the curves reached a relatively stable value at a data limit of 300 or 350 (Fig. 9d, e, f).

The ability of k, H, and H_f to differentiate among genomes was equated with the average coefficient of variation (CV) among the measured complexity profiles; the higher the CV, the greater their differences. Both the magnitude and spread of k and H increased with an increased data limit; however, the CV of k was, on average, 100-fold greater than that observed for H. This discrepancy was largely due to the rapid increase of the average k in the *E. coli* sample with an increasing data limit as compared to the other taxa. Although the magnitude of H_f exhibited the same general pattern of increase observed for k and H, its CV decreased with an increasing data limit, but only by 7% (Fig. 10).

Based on the results of these analyses, we chose an optimum data limit value of 350, roughly corresponding to 35 min within each temperature plateau.

3.3. Complexity measurements in uncharacterized genomes

One of the biggest challenges facing environmental genomics is the limited number of phylogenetically informed contrasts that can be made between related metazoan taxa in diverse environments. For example, no animal genera span both hydrothermal vent and polar sea habitats. Consequently, we have begun to assemble samples from a diverse suite of organisms within extreme marine environments to uncover potentially convergent genomic characters among species by habitat. These comparisons were made possible by a strong, positive linear relationship found between H_f and the repetitive DNA content of the eukaryotic genomes that have been characterized previously by Cot kinetic analyses ($y = 1.95 + 14.64$, $r^2 = 0.99$; Fig. 11). From this relationship, we estimated the repetitive DNA content

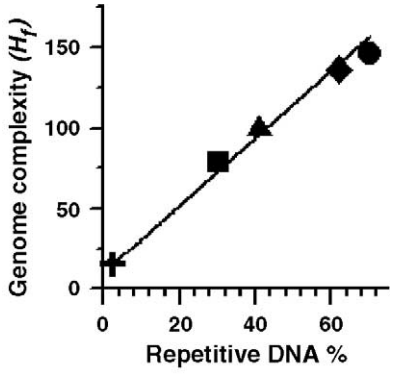

Fig. 11. Relationship between H_f and repetitive DNA content. In their order of decreasing amounts of repetitive sequence: (●) salmon, (♦) mud snail, (▲) cow, (■) horseshoe crab, and (+) yeast.

and kinetic complexity of the uncharacterized genomes from five marine invertebrates across a wide range of environments (Table 1). Notably, we have observed a substantial shift among environments in both the quantity and composition of the kinetic reannealing classes that represent repetitive DNA (see below).

In general, a negative correlation between repetitive DNA content (i.e., genome size) and cell division/metabolic rate is thought to exist, forming the basis of nucleotypic theory (Kozlowski et al., 2003). Thus, increases in both latitude and altitude have been positively correlated with the amount of repetitive DNA, which is thought to be a function of reduced metabolism in organisms from cold environments (Gregory, 2001a,b). Contrary to this expectation, however, Antarctic invertebrates across all taxonomic groupings consistently had the lowest quantities of repetitive DNA within their taxonomic group (Table 1) in conjunction with their low metabolic rates and highly protracted rates of cell division and larval development (Marsh et al., 2001), lasting for many months before metamorphosis and settlement. Conversely, hydrothermal vent species have rapid cell division and growth, but do not have significantly different genome sizes or particularly low amounts of repetitive DNA (Table 1) as compared to non-vent, deep-sea taxa (Dixon et al., 2001). Thus, our data support the view that in addition to nucleotypic effects, the complexity of both the process and products of development (Gregory, 2003), as well as other organismal and environmental factors (Gregory, 2002), must be considered when evaluating the relationships among repetitive DNA, genome size, development, physiology, and habitat.

Compositional heterogeneity in DNA results from local changes in GC content that arise via potentially diverse mechanisms, forming isochores that occur in all metazoan genomes (Nekrutenko and Li, 2000; Paces et al., 2004). We observed a distinct compositional heterogeneity in the kinetic

Table 1
Kinetic hybridization estimates of repetitive DNA content and the calculated complexity of unique DNA fractions

Species	Taxon [a]	Common name	Genome size (Mb) [b]	% Repeats	Unique complexity (Mb)	Cot reference
Dermasterias imbricata	Order: Spinulosida	Starfish	528	55	238	(Smith et al., 1980)
Evasterias retifera	Order: Forcipulatida	Starfish	–	45	–	(Brykov et al., 1979)
Asterias amurensis	Order: Forcipulatida	Starfish	–	42	–	(Brykov et al., 1979)
Distolasterias nipon	Order: Forcipulatida	Starfish	–	40	–	(Brykov et al., 1979)
Pisaster ochraceus	Order: Forcipulatida	Starfish	636	40	382	(Smith and Boal, 1978)
Odontaster validus [c]	**Order: Valvatida**	**Antarctic starfish**	**719** [d]	**19**	**585**	
Strongylocentrotus intermedius	Family: Strongylocentrotidae	Urchin	–	40	–	(Brykov et al., 1979)
Strongylocentrotus nudus	Family: Strongylocentrotidae	Urchin	–	40	–	(Brykov et al., 1979)
Strongylocentrotus drobachiensis	Family: Strongylocentrotidae	Urchin	880	30	616	(Vorobev and Kosjuk, 1974)
Strongylocentrotus purpuratus	Family: Strongylocentrotidae	Purple urchin	870	30	609	(Graham et al., 1974; Angerer et al., 1976)
Strongylocentrotus franciscanus	Family: Strongylocentrotidae	Urchin	812	29	576	(Angerer et al., 1976)
Sterechinus neumayeri	**Family: Echinidae**	**Antarctic urchin**	**670** [e]	**24**	**509**	
Ilyanassa obsoleta	**Class: Gastropoda**	**Eastern mud snail**	**2840**	**62**	**1100**	(Davidson et al., 1971)
Crassostrea virginica	Class: Bivalvia	Eastern oyster	675	40	380	(Kamalay et al., 1976)
Aplysia californica	Class: Gastropoda	California sea hare	1760	40	840	(Angerer et al., 1975)
Crassostrea gigas	Class: Bivalvia	Pacific oyster	890	30	623	(McLean and Whiteley, 1974)
Spisula solidissima	Class: Bivalvia	Atlantic surf clam	1174	25	820	(Goldberg et al., 1975)
Loligo sp.	Class: Cephalopoda	Squid	2738	25	2054	(Galau et al., 1978)
Tritonia antarctica	**Class: Gastropoda**	**Antarctic nudibranch**	**1138** [d]	**8**	**1053**	
Siboglinum fiordicum	Family: Siboglinidae	Polychaete worm	5477	70	1643	(Petrov et al., 1980)
Alvinella pompejana	**Family: Alvinellidae**	**Vent polychaete**	**782**	**41**	**462**	
Streblospio benedicti	**Family: Spionidae**	**Polychaete worm**	**1128** [d]	**21**	**890**	

[a] United States Department of Agriculture. 2004 Integrated Taxonomic Information System http://www.itis.usda.gov/index.html.
[b] Genome size estimates from Marsh et al. (1999); Gregory (2001a,b) unless specified otherwise.
[c] Bold type indicates species and data from this study.
[d] Mean of Asteroidia, Nudibranchia, and Polychaeta, respectively.
[e] Genome size estimate from (Marsh et al., 1999).

reannealing classes that represent repetitive DNA between the vertebrate genomes we examined (Fig. 12a). This separation is reminiscent of the GC content patterns obtained by cesium chloride gradient ultracentrifugation and parallels the compositional differences known from warm and cold-blooded vertebrate genomes (Bernardi, 2000). Because isochores are both large and abundant in vertebrates, their relationship to genome function has been most thoroughly investigated in this animal group. Despite much controversy, overall, there are most certainly associations between compositional heterogeneity and gene density, the extent of CpG methylation, and the complexity of gene expression (magnitude and specificity) that may well correlate with general biological complexity (Tweedie et al., 1997; Fazzari and Greally, 2004). What we find most intriguing about the interpretation of the vertebrate data is that compositional heterogeneity is positively correlated with homeostasis/environmental stability (Bernardi, 1990) and regulatory complexity. Because the relationship between temperature and compositional heterogeneity of the vertebrate genome remains under debate, inclusion of invertebrate taxa from thermal extremes will provide much needed new data and help address questions surrounding the relationship between environment and the mechanistic complexity of adaptive gene regulation, about which we know remarkably little.

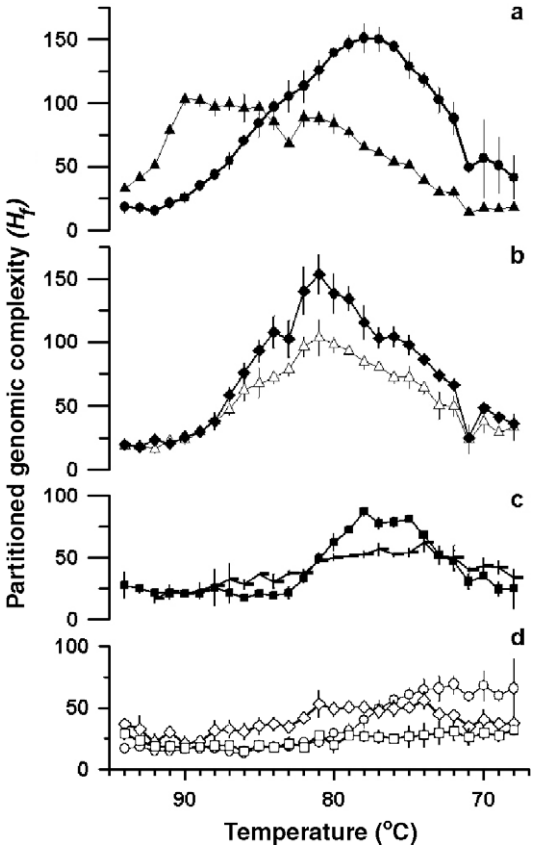

Fig. 12. Genome complexity partitioned by reannealing temperature for nine genomes illustrating an apparent relationship between heterogeneity in the GC composition of invertebrate repetitive DNA and the thermal stability of their environment: (a) the vertebrates, (●) salmon and (▲) cow; (b) invertebrates from high-temperature hydrothermal vent and intertidal habitats, (△) Pompeii worm and (♦) mud snail; (c) subtidal invertebrates, (■) horseshoe crab, (—) polychaete (*S. benedicti)*; (d) Antarctic invertebrates, (○) urchin, (◇) starfish and (□) nudibranch. Replication and error bars as in Fig. 3.

In this study, the invertebrates from thermally variable habitats that include high-temperature exposures (40 °C or greater, *I. obsoleta* and *A. pompejana*) have abundant repetitive sequence classes that reanneal at 80 °C and higher that is evident as a peak in their H_f complexity profile (Fig. 12b). These kinetic classes are less prevalent in the other invertebrates (Fig. 12c–d). The invertebrate taxa from subtidal habitats (*L. polyphemus, S. benedicti*) where temperature change decreases in magnitude with depth and occurs predominately on slow seasonal time scales have more repetitive DNA at a lower GC content (Fig. 12c). Even more striking is the abundance of repetitive DNA of low GC content in the Antarctic invertebrates (*S. neumayeri*, *O. validus*, *T. antarctica*) that experience a stenothermal, low temperature (<2 °C) habitat (Fig. 12d). Thus, if compositional heterogeneity parallels regulatory complexity, as proposed for vertebrates, then stability of the physical environment (temperature in this context) and the complexity of gene regulation relative to homeostasis in invertebrate poikilotherms appear to correlate.

As an alternative to metabolic constraint-based explanations, we hypothesize that shifts in the quantity and composition of repetitive, non-coding DNA reflect differences in total genomic regulatory complexity relative to the stability of an organism's physical environment. In part, we draw this conclusion from recent analyses of genomic data indicating a significant contribution of non-coding DNA to genome structure, function, and evolution (Wray et al., 2003; Fazzari and Greally, 2004; Kazazian, 2004; Mattick, 2004; Nelson et al., 2004). Non-coding, repetitive DNA consists of simple repeats (e.g., micro- and minisatellites), inverted repeats, and sequences from transposable elements. Simple and inverted repeats are well known as regulatory elements that bind proteins and serve as enhancer elements in the eukaryotic genome; variation in repeat number and composition is associated with phenotypic variation at biochemical, physiological, and developmental levels (Kashi and Soller, 1999). For transposons, genome sequencing projects have brought into clear focus the significant structural contribution of these elements and an acceptance of their important role in controlling eukaryotic gene expression directly by cis regulatory modification or via epigenetic mechanisms (Fazzari and Greally, 2004; Kazazian, 2004; Lippman et al., 2004). Notably, an increasing number of studies have found a positive relationship between repetitive DNA (especially transposons) and a species' environmental conditions, particularly temperature (Kalendar et al., 2000; Ceccarelli et al., 2002; Vieira and Biemont, 2004). Broadly, the data from sequenced organisms suggest that the number of conserved, functional non-coding nucleotides is equivalent to their protein-coding counterpart and that roughly half of all phenotypically relevant molecular evolution involves non-coding sequence (Wray et al., 2003). On a local scale, a positive correlation exists between a gene's regulatory complexity and the surrounding amount of non-coding DNA (Nelson et al., 2004). Reduced regulatory complexity for poikilotherms in cold, stenothermal habitats seems likely and has been documented from disparate

Antarctic eukaryotes for myoglobin (Small et al., 2003) and *hsp70* (Hofmann et al., 2000; La Terza et al., 2004). On a genomic scale, reduction in regulatory complexity is thought to be associated with elimination of repetitive DNA by chromatin diminution during development in multiple groups (Kloc and Zagrodzinska, 2001) and could well involve loss of abundant, non-protein coding regulatory RNAs (Mattick, 2004). Thus, the quantitative, bulk decreases in repetitive DNA we have observed in Antarctic species would be consistent with a global reduction in regulatory complexity that is not evident in species from more thermally variable habitats (i.e., hydrothermal vent and intertidal mudflat).

The relationships described above suggest a shift in the repetitive sequence representation between metazoans living in deep-sea hydrothermal vents and Antarctic coastal margins that may ultimately reflect differences in genome function. Naturally, our conclusions at this time are tempered by the phylogenetically limited sample in this first study; pinpointing either the exact selective force or neutral molecular and population genetic processes (Schlotterer, 2000; Eyre-Walker and Hurst, 2001; Petrov, 2001; Lynch and Conery, 2003) that may underlie this pattern will require a far more comprehensive investigation. However, the methodology we have developed is novel in terms of the ease and efficiency with which new species from diverse habitats can now be screened for patterns of genome and transcriptome sequence complexity. Its broad applicability can rapidly add to our understanding of the evolutionary linkages between genome structure and function in an environmental context.

4. Conclusions

In summary, we have developed a new fluorescence method for the rapid and easy determination of genome complexity. The method is based on the fundamentals of Cot analysis, in which the reassociation rate of sheared, total genomic DNA is dependent on the abundance of complementary sequence. A key feature of the new method is the novel application of a fluorescence-scaled Shannon complexity index, H_f, to describe the hybridization kinetics. H_f is well-correlated with the traditional kinetic parameter, k, and has the advantages of being computationally simple and less variable than k. A second key feature is that our technique sorts the sample by reannealing temperature (%GC), providing greater resolution and revealing more complexity than previously possible. After calibrating the technique to genomic samples of known complexity, we characterized the genomes of five marine invertebrates from diverse, extreme habitats. Contrary to the predictions of nucleotypic theory, Antarctic invertebrates consistently had the lowest quantities of repetitive DNA in conjunction with low metabolic rates and highly protracted rates of cell division and larval development. Conversely, hydrothermal vent species with rapid cell division and growth do not have significantly different genome sizes or particularly low amounts of repetitive DNA as compared to non-vent, deep-sea taxa. Furthermore, there appears to be a positive correlation between the temperature at which the most abundant repetitive sequence classes anneal and habitat thermal stability. Thus, our study reveals a potential shift in repetitive sequence representation between these extreme environments that may be related to genome function in species living at these different thermal regimes.

Acknowledgments

The authors would like to thank Lisa Waidner, Dr. Frances Weaver and Dr. Craig Cary for providing access to bacterial, horseshoe crab, and Pompeii worm DNA, respectively. This work was supported by grants from the National Science Foundation programs in Biological Oceanography and Antarctic Biology and Medicine and the University of Delaware Research Foundation.

References

Angerer, R.C., Davidson, E.H., Britten, R.J., 1975. DNA sequence organization in mollusk *Aplysia californica*. Cell 6, 29–40.

Angerer, R.C., Davidson, E.H., Britten, R.J., 1976. Single copy DNA and structural gene sequence relationships among 4 sea-urchin species. Chromosoma 56, 213–226.

Bernardi, G., 1965. Chromatography of nucleic acids on hydroxyapatite. Nature 206, 779–783.

Bernardi, G., 1990. Compositional transitions in the nuclear genomes of cold-blooded vertebrates. J. Mol. Evol. 31, 282–293.

Bernardi, G., 2000. Isochores and the evolutionary genomics of vertebrates. Gene 241, 3–17.

Britten, R.J., Davidson, E.H., 1971. Repetitive and non-repetitive DNA sequences and a speculation on origins of evolutionary novelty. Q. Rev. Biol. 46, 111–138.

Britten, R.J., Kohne, D.E., 1968. Repeated sequences in DNA. Science 161, 529–540.

Brykov, V.A., Volfson, V.G., Vorobev, V.I., 1979. Genome structure and divergence of nucleotide-sequences in echinodermata. Chromosoma 74, 105–124.

Cary, S.C., Shank, T., Stein, J., 1998. Worms bask in extreme temperatures. Nature 391, 545–546.

Ceccarelli, M., et al., 2002. Genome plasticity in *Festuca arundinacea*: direct response to temperature changes by redundancy modulation of interspersed DNA repeats. Theor. Appl. Genet. 104, 901–907.

Clark, M.S., et al., 2004. Antarctic genomics. Compar. Funct. Genom. 5, 230–238.

Davidson, E.H., Hough, B.R., Chamberlin, M.E., Britten, R.J., 1971. Sequence repetition in DNA of *Nassaria* (*Ilyanassa*) *obsoleta*. Dev. Biol. 25, 445–463.

Dixon, D.R., Dixon, L.R.J., Pascoe, P.L., Wilson, J.T., 2001. Chromosomal and nuclear characteristics of deep-sea and hydrothermal-vent organisms: correlates of increased growth rates. Mar. Biol. 139, 251–255.

Eyre-Walker, A., Hurst, L.D., 2001. The evolution of isochores. Nat. Rev., Genet. 2, 549–555.

Fazzari, M.J., Greally, J.M., 2004. Epigenomics: beyond CpG islands. Nat. Rev., Genet. 5, 446–455.

Feder, M.E., Mitchell-Olds, T., 2003. Evolutionary and ecological functional genomics. Nat. Rev., Genet. 4, 651–657.

Galau, G.A., Chamberlain, M.E., Hough, B.R., Britten, R.J., Davidson, E.H., 1978. Evolution of repetitive and nonrepetitive DNA. Molecular Evolution. Sinauer Associates, Inc., Massachusetts, pp. 200–262.

Goldberg, R.B., et al., 1975. Sequence organization in genomes of 5 marine invertebrates. Chromosoma 51, 225–251.

Graham, D.E., Neufeld, B.R., Davidson, E.H., Britten, R.J., 1974. Interspersion of repetitive and non-repetitive DNA sequences in sea-urchin genome. Cell 1, 127–137.

Gregory, T.R., 2001a. Animal Genome Size Database. http://www.genomesize.com.

Gregory, T.R., 2001b. Coincidence, coevolution, or causation? DNA content, cell size, and the C-value enigma. Biol. Rev. 76, 65–101.

Gregory, T.R., 2002. Genome size and developmental parameters in the homeothermic vertebrates. Genome 45, 833–838.

Gregory, T.R., 2003. Variation across amphibian species in the size of the nuclear genome supports a pluralistic, hierarchical approch to the C-value enigma. Biol. J. Linn. Soc. 79, 329–339.

Helmuth, B., Harley, C.D.G., Halpin, P.M., O'Donnell, M., Hofmann, G.E., Blanchette, C.A., 2002. Climate change and latitudinal patterns of intertidal thermal stress. Science 298, 1015–1017.

Hofmann, G.E., Buckley, B.A., Airaksinen, S., Keen, J.E., Somero, G.N., 2000. Heat-shock protein expression is absent in the Antarctic fish *Trematomus bernacchii* (family Nototheniidae). J. Exp. Biol. 203, 2331–2339.

Jackson, R.B., Linder, C.R., Lynch, M., Purugganan, M., Somerville, S., Thayer, S.S., 2002. Linking molecular insight and ecological research. Trends Ecol. Evol. 17, 409–414.

Kalendar, R., Tanskanen, J., Immonen, S., Nevo, E., Schulman, A.H., 2000. Genome evolution of wild barley (*Hordeum spontaneum*) by BARE-1 retrotransposon dynamics in response to sharp microclimatic divergence. Proc. Natl. Acad. Sci. U. S. A. 97, 6603–6607.

Kamalay, J.C., Ruderman, J.V., Goldberg, R.B., 1976. DNA sequence repetition in the genome of the American oyster. Biochim. Biophys. Acta, Nucleic Acids Protein Synth. 432, 121–128.

Kashi, Y., Soller, M., 1999. Functional roles of microsatellites and minisatellites. In: Goldstein, D.B., Schlotterer, C. (Eds.), Microsatellites. Evolution and Applications. Oxford University Press, New York, pp. 10–23.

Kazazian, H.H., 2004. Mobile elements: drivers of genome evolution. Science 303, 1626–1632.

Kloc, M., Zagrodzinska, B., 2001. Chromatin elimination—an oddity or a common mechanism in differentiation and development? Differentiation 68, 84–91.

Kozlowski, J., Konarzewski, M., Gawelczyk, A.T., 2003. Cell size as a link between noncoding DNA and metabolic rate scaling. Proc. Natl. Acad. Sci. U. S. A. 100, 14080–14085.

La Terza, A., Miceli, C., Luporini, P., 2004. The gene for the heat-shock protein 70 of *Euplotes focardii*, an Antarctic psychrophilic ciliate. Antarct. Sci. 16, 23–28.

Lippman, Z., et al., 2004. Role of transposable elements in heterochromatin and epigenetic control. Nature 430, 471–476.

Lynch, M., Conery, J.S., 2003. The origins of genome complexity. Science 302, 1401–1404.

Marsh, A.G., Leong, P.K.K., Manahan, D.T., 1999. Energy metabolism during embryonic development and larval growth of an Antarctic sea urchin. J. Exp. Biol. 202, 2041–2050.

Marsh, A.G., Maxson, R.E., Manahan, D.T., 2001. High macromolecular synthesis with low metabolic cost in Antarctic sea urchin embryos. Science 291, 1950–1952.

Mattick, J.S., 2004. RNA regulation: a new genetics? Nat. Rev., Genet. 5, 316–323.

McLean, K.W., Whiteley, A.H., 1974. Characteristics of DNA from oyster, *Crassostrea-gigas*. Biochim. Biophys. Acta 335, 35–41.

Nekrutenko, A., Li, W.H., 2000. Assessment of compositional heterogeneity within and between eukaryotic genomes. Genome Res. 10, 1986–1995.

Nelson, C.E., Hersh, B.M., Carroll, S.B., 2004. The regulatory content of intergenic DNA shapes genome architecture. Genome Biol. 5, R25.1–R25.15.

Paces, J., Zika, R., Paces, V., Pavlicek, A., Clay, O., Bernardi, G., 2004. Representing GC variation along eukaryotic chromosomes. Gene 333, 135–141.

Peterson, D.G., et al., 2002a. Integration of Cot analysis, DNA cloning, and high-throughput sequencing facilitates genome characterization and gene discovery. Genome Res. 12, 795–807.

Peterson, D.G., Wessler, S.R., Paterson, A.H., 2002b. Efficient capture of unique sequences from eukaryotic genomes. Trends Genet. 18, 547–550.

Petrov, D.A., 2001. Evolution of genome size: new approaches to an old problem. Trends Genet. 17, 23–28.

Petrov, N.B., Poltaraus, A.B., Antonov, A.S., 1980. The genome of the pogonophoran *Siboglinum-fiordicum*—characteristics of its organization and divergence from the genomes of several representative invertebrate animals. Mol. Biol. 14, 340–348.

Porebski, S., Bailey, L.G., Baum, B.R., 1997. Modification of a CTAB DNA extraction protocol for plants containing high polysaccharide and polyphenol components. Plant. Mol. Biol. Report. 15, 8–15.

Purugganan, M., Gibson, G., 2003. Merging ecology, molecular evolution, and functional genetics. Mol. Ecol. 12, 1109–1112.

Schlotterer, C., 2000. Evolutionary dynamics of microsatellite DNA. Chromosoma 109, 365–371.

Shannon, C.E., 1948. A mathematical theory of communication. Bell Syst. Tech. J. 27, 623–656.

Singer, V.L., Jones, L.J., Yue, S.T., Haugland, R.P., 1997. Characterization of PicoGreen reagent and development of a fluorescence-based solution assay for double-stranded DNA quantitation. Anal. Biochem. 249, 228–238.

Small, D.J., Moylan, T., Vayda, M.E., Sidell, B.D., 2003. The myoglobin gene of the Antarctic icefish, *Chaenocephalus aceratus*, contains a duplicated TATAAAA sequence that interferes with transcription. J. Exp. Biol. 206, 131–139.

Smith, M.L., Boal, R., 1978. DNA-sequence organization in common pacific starfish *Pisaster-ochraceous*. Can. J. Biochem. 56, 1048–1054.

Smith, M.J., Lui, A., Gibson, K.K., Etzkorn, J.K., 1980. DNA-sequence organization in the starfish *Dermasterias-imbricata*. Can. J. Biochem. 58, 352–360.

Soares, M.B., Bonaldo, M.D., Jelene, P., Su, L., Lawton, L., Efstratiadis, A., 1994. Construction and characterization of a normalized cDNA library. Proc. Natl. Acad. Sci. U. S. A. 91, 9228–9232.

Spellerberg, I.F., Fedor, P.J., 2003. A tribute to Claude Shannon (1916–2001) and a plea for more rigorous use of species richness, species diversity and the 'Shannon–Wiener' Index. Glob. Ecol. Biogeogr. 12, 177–179.

Thomas, M.A., Klaper, R., 2004. Genomics for the ecological toolbox. Trends Ecol. Evol. 19, 439–445.

Tweedie, S., Charlton, J., Clark, V., Bird, A., 1997. Methylation of genomes and genes at the invertebrate–vertebrate boundary. Mol. Cell. Biol. 17, 1469–1475.

Van Dover, C.L., 2000. The Ecology of Deep-sea Hydrothermal Vents. Princeton University Press, Princeton.

Vieira, C., Biemont, C., 2004. Transposable element dynamics in two sibling species: *Drosophila melanogaster* and *Drosophila simulans*. Genetica 120, 115–123.

Vorobev, V.I., Kosjuk, G.N., 1974. Distribution of repetitive and non-repetitive nucleotide-sequences in DNA of sea-urchin. FEBS Lett. 47, 43–46.

Wetmur, J.G., Davidson, N., 1968. Kinetics of renaturation of DNA. J. Mol. Biol. 31, 349–370.

Wray, G.A., et al., 2003. The evolution of transcriptional regulation in eukaryotes. Mol. Biol. Evol. 20, 1377–1419.