Vol. 301: 9–22, 2005 | MARINE ECOLOGY PROGRESS SERIES
Mar Ecol Prog Ser | Published October 11

Sand DNA—a genetic library of life at the water's edge

Robert K. Naviaux[1,*], Benjamin Good[2,6], John D. McPherson[3], David L. Steffen[3], David Markusic[1,7], Barbara Ransom[4,8], Jacques Corbeil[2,5]

[1]The Mitochondrial and Metabolic Disease Center, Departments of Medicine and Pediatrics, Room C-103, Building CTF, San Diego School of Medicine, University of California, 214 Dickinson Street, San Diego, California 92103-8467, USA

[2]Genomics Core Laboratory, Center for AIDS Research and Veterans Medical Research Foundation, Room 325, Stein Clinical Research Building, University of California, San Diego, 9500 Gilman Drive, La Jolla, California 92093-0679, USA

[3]Department of Molecular and Human Genetics, and Human Genome Sequencing Center, Baylor College of Medicine, One Baylor Plaza, N1519, Houston, Texas 77030, USA

[4]The Scripps Institute of Oceanography, Geosciences Research Division, La Jolla, California, USA

[5]Laval University, Québec, Canada

[6]*Present address:* Department of Molecular Biology and Biochemistry, Room SSB 8166, Simon Fraser University, 8888 University Drive, Burnaby, British Columbia V5A 1S6, Canada

[7]*Present address:* Meibergdreef 69-71, Amsterdam 1105 BK, The Netherlands

[8]*Present address:* The National Science Foundation, Geosciences Office, 4201 Wilson Boulevard, Arlington, Virginia 20036, USA

ABSTRACT: Powdered silica has long been used for the purification of nucleic acids in the laboratory. Silicate-rich, ordinary ocean beach sand was found to concentrate dissolved DNA from seawater over 10 000-fold, providing a rich, renewable, and easily accessible genetic library that is easy to harvest and inexpensive to process. We found an average of 29 µg ml^{-1} of cell-free DNA adsorbed to silicate-rich, wave-washed sand from 14 beaches bordering 9 seas around the world. The DNA from a reference beach was shotgun cloned, 3 107 399 nucleotides of anonymous, non-redundant sequence were analyzed, and 2571 genes were found; 2562 of these genes were new. The apparent complexity of sand DNA was greater than 1.4×10^{11} nucleotides. About 90 % of the sequences identified were from prokaryotes, 10 % from eukaryotes, and 1 % were viral. Sequences from all kingdoms of life were present. Over half the sequences came from new phylotypes, reflecting the novelty of this genetic reservoir.

KEY WORDS: Sand · Genetic library · Dissolved DNA · Beach

'To see a world in a grain of sand, and heaven in a wild flower,
Hold Infinity in the palm of your hand, and eternity in an hour.'
—William Blake

INTRODUCTION

Ordinary ocean beach sand contains large amounts of quartz (SiO_2) and other silicate minerals. It is well known that nucleic acids bind to silica in high ionic strength aqueous solutions, and are released when ionic strength is decreased (Vogelstein & Gillespie 1979). Although more crystalline than the silica used in laboratories to concentrate DNA from solution, quartz and other silicates share many of silica's physicochemical properties. Conditions for binding DNA are readily found in nature, where beach sand is repeatedly soaked by seawater (a high ionic strength solution with a mean of 34.7 salinity, or about 0.5 M NaCl). Conversely, conditions favoring desorption and transport of DNA occur with exposure to freshwater and rain. At the seashore, the 4 conditions for the formation of a natural DNA library of enormous complexity converge: (1) the presence of silicate-rich sand, (2) aqueous solutions containing dissolved DNA, (3) the high ionic

*Email: naviaux@ucsd.edu

strength of seawater, and (4) the presence of low ionic strength solutions, such as rain, river, and groundwater—all of which help to mobilize, mix, transport and refresh the DNA adsorbed to sand. We use the term 'sand DNA' in this paper to refer to cell-free, unencapsidated, DNAse-sensitive DNA found adsorbed to sand on ocean beaches and other locations. The biodiversity captured by sand DNA is likely to vary according to currents, tides, local ecosystems, biomass, time of year and radial distance from the collection point. Such sand DNA libraries might be enriched in adaptive sequences from organisms that have flourished, provide the raw material for lateral gene transfer events, and be continuously refreshed in a cycle of diurnal and seasonal adsorption and desorption of DNA on beaches around the world and throughout evolutionary history.

In this paper, we report the discovery that ordinary ocean beach sand contains an adsorbed DNA library of unprecedented size, exceeding 1.4×10^{11} nucleotides in complexity. The mean, adsorbed DNA content of wet sand collected from 3 continents, 9 seas, and 14 beaches around the world was 29 µg ml^{-1}. We show that quartz-rich sand from a reference beach (Pacific Beach, PB) in San Diego, California, concentrated DNA from seawater 20000-fold. We created a library of 1.5×10^6 clones from this sand DNA, and analyzed 2571 clones picked at random. A total of 4981 reads produced 3447360 nucleotides of sequence data and 3107399 nucleotides of non-redundant sequence; 2562 previously unknown genes were identified. Gene diversity estimates suggest that the sand DNA from this reference beach contained millions of genes sampled from a large number of species in and about the sea.

MATERIALS AND METHODS

Sand collection. All samples were collected from ankle-deep water in the tidal zone of beaches around the world. The collection site used for the construction of the PB sand DNA library, Pacific Beach, was located 3 miles (~5 km) north of the confluence of the San Diego River with the Pacific Ocean in San Diego, California. The sand used as the source of DNA for library construction was collected on August 23, 1999 (for small subunit rDNA cloning and sequences see Table 2), and September 26, 2003. Sand used to determine the mean DNA yields and confidence intervals was collected from 6 San Diego beaches (Pacific Beach, PB; Mission Beach, MB; Ocean Beach, OB; Imperial Beach, IB; La Jolla shores, LJS; Torrey Pines, TP), and 8 other sites around the world (at or near Beach Haven, New Jersey; Siesta Key, Florida; Zandvoort, Netherlands; West Sussex, England; Sydney, Australia; Melbourne, Australia; Perth, Australia; Greymouth, New Zealand). These collection sites were along the margins of 9 seas (North Pacific, South Pacific, Tasman Sea, Southern Ocean, Indian Ocean, North Atlantic, Gulf of Mexico, North Sea, English Channel).

Tidal zone sand was collected from the surface 2 to 4 cm of the beach and stored in the dark at room temperature in 0.5 l wide-mouth plastic jars (Nunc #2118-0016) with a 1 to 3 cm overlayer of native seawater that had been collected with the sand. Sand stored in this way produced consistent yields and quality of DNA in repeated experiments for at least 12 mo. Care was taken not to permit sand to dry, as dehydrated (presumably A-form) DNA exhibited different physicochemical properties than well-hydrated (B-form) DNA.

Purification of sand DNA. We developed a method of purifying sand DNA based on a combination of principles used to purify high quality genomic DNA (Wang et al. 1994, Nakae et al. 1995) and for isolating DNA on powdered silica (Vogelstein & Gillespie 1979). Cellular DNA from indigenous microflora in beach sand samples was removed through a series of high-, and low-ionic strength wash steps that removed encapsidated, membrane-associated, and cell-associated nucleic acids. Briefly, 15 ml (~30 g) of wet, tidal zone sand was placed in a 50 ml, sterile, screw-top tube (Corning #25330-50), allowed to settle by gravity, and the last free drops of seawater removed by aspiration. The sand was washed 3 times in 2 volumes (30 ml) of 35 g l^{-1} NaCl or artificial seawater (35 g l^{-1}, Coralife scientific grade marine salt, Energy Savers Unlimited). The sand was then washed 3 times with 1 volume (15 ml) of NaI Wash (6 M NaI, 0.5% sodium sulfite), mixed briskly by hand for 20 s, and then allowed to settle by gravity to remove any non-adsorbed particulates. Residual NaI was aspirated between each wash. The sand was then washed again 3 times in 2 vol (30 ml) of 35 g l^{-1} NaCl or artificial seawater (35 g l^{-1}, Coralife scientific grade marine salt) to reduce the ionic strength to that of normal seawater. Next, the sand was washed 3 to 5 times in 15 ml of 70% ethanol wash (70% ethanol, 30 mM sodium acetate, pH 5.2), using the same procedure. The last drops of ethanol wash were removed by aspiration. The first fraction of DNA was eluted in 15 ml (1 vol) fractions of TE (10 mM Tris HCl pH 8.1, 1 mM Na_2EDTA), mixing briskly by hand, and incubating for 2 h at room temperature. The sample was mixed every 10 to 15 min by hand, or placed on a 20 rpm end-over-end mixer during this period. The DNA-containing TE was removed to 50 ml capped, sterile conical centrifuge tubes and placed on ice. Elutions 2 to 4 were conducted at 37°C. Insoluble debris in the 4 elution fractions was removed by centrifugation at 2500 × *g* for 15 min, and the supernatant fractions,

containing dissolved DNA, were transferred to clean Oak Ridge tubes. DNA was precipitated by adding a half volume of 7.5 M ammonium acetate, and an equal volume of isopropanol. The tubes were placed at –20°C for at least 2 h to overnight, then DNA was concentrated by centrifugation at 12 000 × *g* for 30 min. The supernatants were decanted and the tubes inverted at room temperature for 5 min to drain. The tan-to brown-colored, DNA-containing pellets were either resuspended in 100 µl of TE if the properties of crude sand DNA were to be tested, or solubilized in 500 µl of Proteinase K lysis buffer (50 mM Tris-HCl, pH 8.1, 100 mM NaCl, 5 mM EDTA, 0.5% SDS) and transferred to labeled 1.5 ml, sterile, Eppendorf tubes. Proteinase K was added to a final concentration of 500 µg ml^{-1} and the samples were incubated at 50°C overnight (>16 h). The samples were extracted twice with an equal volume of phenol and once with equal volumes of phenol and chloroform, precipitated in 0.3 M sodium acetate and 2 volumes of 100% ice cold ethanol, concentrated by centrifugation at 12 000 × *g* for 10 min at 4°C, washed once in 70% ethanol, then resuspended in 100 µl TE. DNA yields were quantified by UV-spectroscopy. A260:A280 ratios of DNA at this stage of purification were typically 1.5 to 1.8. When highly purified sand DNA was required, the 4 eluted fractions were pooled and further purified by ultracentrifugation and double banding in isopycnic CsCl-ethidium bromide gradients using standard methods (Sambrook et al. 1989).

Scanning electron microscopy. After washing and elution of DNA, sand samples were dried and processed for scanning electron microscopy. Sand was mounted on carbon tape and sputter coated with a gold-palladium film at a current of 10 mA for 5 min (20 nm). A Cambridge Instrument Model 360 scanning electron microscope was used to acquire images at several magnifications. The images shown were taken at 90× magnification of representative fields. The proportions of minerals present in sand were estimated by examination under the microscope, and the elemental composition was confirmed by X-ray energy dispersion spectrometry.

Cloning. The DNA isolated from sand was of sufficient quality and size to be used in standard Human Genome Sequence Center protocols for the preparation of whole genome shotgun libraries (Andersson et al. 1996) without further purification. Briefly, 50 µg of sand DNA was randomly fragmented using a GeneMachines HydroShear instrument with a targeted fragment size of 2.5 to 4.5 kb. To remove small fragments that may form chimeras and to further narrow the size distribution, the sheared DNA was separated using agarose gel electrophoresis and the 2.5 to 4.5 kb fragment sizes recovered in a gel slice. DNA was recovered from the gel slice using Qiagen QIAquick protocols. The random overhangs left by the shearing process were polished using T4 DNA polymerase and the repaired DNA was recovered using Qiagen PCR clean up columns. Adapters blunt at one end and with a 12-base overhang at the other end were ligated to the blunt-end genomic fragments. These fragments were then annealed to pUC18 vector prepared with complementary 12-base overhang adapters. Bacteria (XL-10 Gold, Stratagene) were transformed by electroporation without a ligation step; 5000 colonies were plated for picking. Approximately 1.5×10^6 colonies were available for plating. Automated colony picking robots were used to select and array colonies into 96 deep-well plates. The colonies were cultured overnight.

DNA sequencing. Plasmid DNA was isolated at fully automated robotic workstations. An aliquot of the culture was used to prepare a glycerol stock archive plate. After centrifugation to pellet cells, plasmid DNA was isolated using the Concert 96 plasmid purification system (LTI-Invitrogen). An aliquot of the resuspended DNA was used as a template for Sanger sequencing using fluorescent dye labeled terminators (ABI BigDye Terminator Version 3.1) using standard thermal cycling conditions. After cycling, excess dye was removed by ethanol precipitation. The reactions were sequenced on an Applied Biosystems 3730 DNA analyzer. Results are reported for a sample of 2889 randomly selected clones and 5778 associated forward and reverse sequence reads, with a success rate of 89%. This yielded information on 2571 clones isolated from the sand DNA library.

Contig assembly. Sequence reads were clustered using blastclust from the NCBI Blast software suite using a 10% overlap and 97% identity cutoff. Read pair information was then used to combine reads for determining the most significant Blast result for each representative clone.

Functional analysis. Conceptual translations of open reading frames were scanned for functional cues by automated comparison with the Cluster of Orthologous Groups (COGs) database at the National Center for Bioinformatics (NCBI) (http://www.ncbi.nlm.nih.gov/COG). For extension of functional searches to genes unique to eukaryotes and viruses, conceptual translations of sand DNA clones were used to search the Gene Ontology (GO) Annotation database of the European Bioinformatics Institute (EBI) (http://www.ebi.ac.uk/GOA) (Camon et al. 2004).

Small subunit (ssu) ribosomal DNA cloning. The PCR primers (see Table 2) were previously designed to amplify eukaryal 18S (Palumbi 1996), bacterial 16S, and archaebacterial 16S ribosomal DNA loci (Hinrichs et al. 1999). PCR reactions contained 1 µg of CsCl-purified sand DNA, 0.5 µg each of forward and reverse primers, 2.5 mM $MgCl_2$, 200 µM each of dGTP, dATP,

dTTP, and dCTP, 50 mM KCl, and 10 mM Tris-HCl (pH 8.3) at room temperature, and 1 to 2.5 U AmpliTaq DNA polymerase (Perkin-Elmer), in 50 µl reaction volume. Amplifications were initiated by 'hot start' as follows: all reaction components, minus *Taq* polymerase, were combined in a volume of 25 µl and overlayed with 50 µl of mineral oil in a 0.3 ml thin-walled PCR tube. Reactions were denatured at 95°C for 5 min, then 25 µl of a 2× dilution of AmpliTaq in water (containing 1 to 2.5 U of polymerase) was added by pipetting through the oil vapor barrier to produce the final reaction volume of 50 µl. The thermocyle program for the eukaryal 18S primers was as follows: (94°C × 30 s, 55°C × 1 min, 72°C × 1 min) × 5, (94°C × 30 s, 65°C × 1 min, 72°C × 1 min) × 25, followed by a 10 min extension at 72°C. The thermocycle program for the bacterial and archaebacterial 16S primers was: (95°C × 20 s, 50°C × 1.5 min, 72°C × 3 min) × 35, followed by a 10 min extension at 72°C. The amplified fragments were gel purified by adsorption and elution from powdered glass (Vogelstein & Gillespie 1979), ligated in the TA vector pCRII (Invitrogen) and cloned in NM522 or Inv-F′ *Escherichia coli*. Recombinant, miniprep plasmid DNAs were purified on QIAquick silica matrix columns (Qiagen #28104) and sequenced by automated ABI 373 DNA sequencer with an XL upgrade using Big Dye terminator chemistry. Cloned rDNA sequences were analyzed for matching sequences in the National Center for Biotechnology Information (NCBI) database using a basic BLASTX and BLASTN searches available at Website www.ncbi.nlm.nih.gov/BLAST.

Taxonomic analysis. To give a high-level assessment of the taxonomic composition of the sand DNA library, the sequences were compared with the Genbank protein database as of April 17, 2004 on the basis of amino acid (BLASTX), and the Genbank DNA database as of April 25, 2004 on the basis of nucleotide (BLASTN) matches. The 6-frame translation of each sequence was compared to the database using the BlastX program from the NCBI toolkit (Altschul & Lipman 1990, Altschul et al. 1990). Default parameters were used, with the following exceptions: Blastx –F 'm S' –v 300 –b 300; Megablast -D 2 –W 12 –G 2 –E 2 –q -1 –r 1 –F 'm D' –v 300 –b 300. Scripts were developed to parse out the taxonomic information from each of the matching Genbank records and to tally the counts for each taxonomic group. Only GenBank sequence matches with e-values of 1×10^{-5} or less were considered significant for purposes of gene and species assignments in this report. An e-value of $\leq 10^{-5}$ means that the probability of a sequence match by chance is less than 1 in 100 000. Because marine biota remain largely uncharacterized, our taxonomic assignments must be considered best guesses or inferences, made with the limitations of the world's current genetic databases, as of April 17, 2004. The degrees of certainty captured by our analysis were measured by e-scores of 0 (a near-perfect match) to 1×10^{-5}, with corresponding bit-scores of 1306 (a near-perfect match) to 57. The best match for each clone was used to make the taxonomic assignment. In our 1st-pass analysis, 639 clones had a best match to 'unclassified' environmental sequences. As these matches provided no taxonomic information, we removed 994 617 'unclassified environmental' peptides from the April 17, 2004 version of the NCBI protein nr (non-restricted) database. This resulted in a revised database with 1 773 385 amino acid sequences for which taxonomic information was available, and we repeated the analysis. We report the results of this 2nd-pass analysis. For DNA sequence alignments, we used the April 25, 2004 version of the NCBI database. This contained 2 200 567 DNA sequences and 10 470 868 853 nucleotides. We have adopted Cavalier-Smith's treatment of archaebacteria as a division of the kingdom Prokaryotes (Cavalier-Smith 1998).

Taxonomic assignment biases were inescapable in this analysis of anonymous environmental sequences from marine ecosystems. Many matches were to proteins found in a variety of different species, and in the cases of highly conserved proteins, solid amino acid matches were observed among organisms from different kingdoms of life. Therefore, the species matches indicated should not be interpreted as the only possible origin for a given DNA sequence. They represent merely the best match at the time of this analysis. Assignment biases had a number of identifiable sources. Among these were the disproportionate number of sequences from terrestrial bacteria, archaebacteria, viruses, mammals, plants, protozoan pathogens, and other terrestrial sequences in the world's genetic databases. Biases were most obvious for the detailed, species-level assignments. It is likely that many more marine species were captured in the sand DNA library but could not be accurately assigned because of the terrestrial bias of the existing databases.

Comparison with sequences from the Sargasso Sea. BLASTN analysis of the PB sand DNA sequences was used to determine the number of similar sequences recently identified by drag filter collection of pelagic bacteria from the Sargasso Sea (Venter et al. 2004). Sequence alignments with e-values $<10^{-5}$ and bit scores of more than 57 were considered significant for this analysis.

Pulsed field gel electrophoresis. We poured 1% LE agarose (FMC) gels containing 1 × TAE buffer (40 mM Tris, 20 mM acetic acid, 1 mM Na_2EDTA) in a 13 × 14 cm casting stand of a Bio-Rad CHEF-DR III pulsed field electrophoresis system. Run parameters were 6 V cm^{-1} field strength, 120° field angle, with switch times ramped from 1 to 5 s over the 20 h run in 1 × TAE that

was maintained at 14°C and recirculated by a cooling pump. The gel was stained for 20 min in 0.5 µg ml^{-1} ethidium bromide and photographed under shortwave UV illumination.

Quantitative Southern blot analysis. A 0.7% agarose gel was loaded with 5 µg of total sand DNA cut with *Bam*HI, and an internal calibration curve containing 5, 2, 1, 0.5 and 0.2 pg of a 1.4 kb BamHI fragment of sand DNA Clone A4 (GenBank Accession No. AF298086, also see Appendix 1). The conceptual translation of Clone A4 had strong amino acid sequence similarity to the ATP-dependent chaperon CDC48 from *Deinococcus radiodurans* (NCBI protein Accession No. D75493). A high specific activity probe was synthesized from the purified 1.4 kb BamHI fragment of Clone A4, and used for hybridization. The CDC48-like, A4 clone was selected for use in quantitative Southern blot analysis in an effort to obtain a minimum estimate of the complexity of the sand DNA 'genome'. After electrophoresis and transfer, the blot was hybridized and washed under stringent conditions. After a 3 wk exposure at –80°C with an intensifying screen, the 0.2 pg internal control band was easily visible. A single hybridizing band was identified in the BamHI-cut sand DNA lane. The size of this genomic band was 1.4 kb, corresponding precisely to the expected 1.4 kb BamHI genomic fragment originally cloned from an M13 library. The genomic band had an intensity estimated to correspond to 0.05 pg (5×10^{-14} g) of DNA. The lower limit of the effective single-copy gene complexity of the cell-free sand DNA was calculated according to the formula: (5×10^{-6} g total sand DNA digested) ÷ (5×10^{-14} g of the genomic 1.4 kb BamHI fragment observed) × (1400 bp in the genomic fragment) = 1.4×10^{11} bp. This was considered 1 genome-equivalent. Its molecular weight was 9.1×10^{13} g mol^{-1} (1.4×10^{11} bp × 650 g mol^{-1} bp^{-1}). The mass of 1 genome-equivalent was 150 pg (9.1×10^{13} g mol^{-1} ÷ 6.0×10^{23} copies mol^{-1}). On an average beach, containing 29 µg of DNA ml^{-1} wet sand, 1 genome-equivalent (150 pg) was present in each 5.2 nl (~10 µg) of sand.

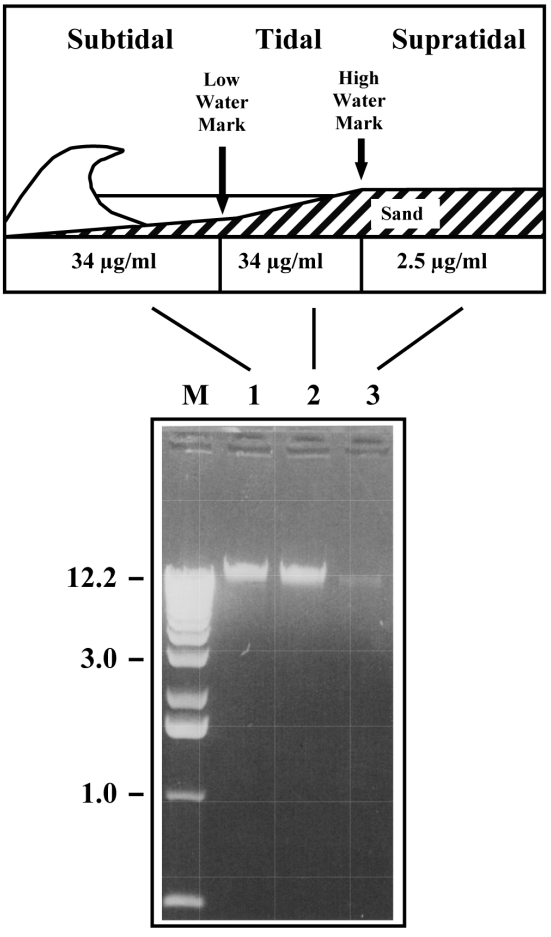


Fig. 1. Cell-free DNA content of sand collected from different tidal zones on an ocean beach. Lanes 1 to 3 represent $^1/_{500}$ of DNA isolated from 20 ml (40 g) samples of wet sand collected from each of zones indicated on Pacific Beach, a southern California beach. Purified sand DNA was resolved in a 1% agarose gel and visualized by ethidium bromide staining under UV illumination. Lane M: molecular weight markers

RESULTS

Location of DNA on beach

DNA was released from ocean beach sand collected from subtidal, tidal, and supratidal zones. The DNA sequences described here have been deposited in Genbank, Accession Nos. AF298077–AF298115 and, NCBI Genome Project #13729, Trace Archives 711564901 to 711569881. Fig. 1 illustrates the location of these zones and the sand DNA found within them. The yields of DNA from the subtidal zone and the tidal zone were equivalent, and equal to 34 µg ml^{-1} of wet sand. This was equivalent to 17 µg g^{-1}, since the density of the wet sand samples tested was 2 g ml^{-1}. The yield of DNA from the supratidal zone (which remained dry, beyond the reach of waves during high tide, throughout most of the year) was only 2 to 20% of that observed from the 2 areas that were lapped more regularly by the waves. For convenience, sand for all subsequent experiments was collected from ankle-deep water in the tidal zone.

Cell-free DNA from seawater concentrated on sand

The concentration of dissolved, cell-free DNA in seawater collected at the same time as sand from the beach (PB) in San Diego, California was 1.7 ng ml^{-1} (1.7 µg l^{-1}).

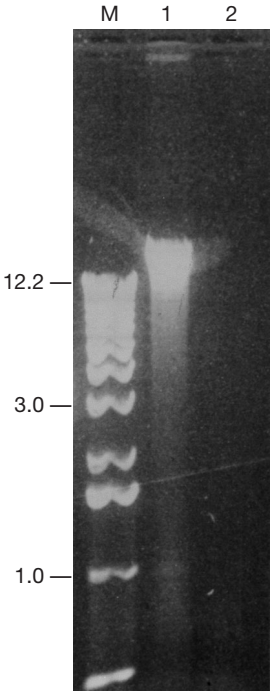


Fig. 2. DNAse sensitivity of DNA released from ocean beach sand. Crude sand DNA was tested after elution in low salt buffer, before further purification; 5 µg of crude sand DNA was incubated in 50 mM NaCl, 10 mM Tris pH 8, 10 mM $MgCl_2$, and 10 mM dithiothreitol, either alone (Lane 1), or with 20 U of RQ1 DNAse (Promega) (Lane 2) in 20 µl reactions for 30 min at 37°C, resolved in a 0.7% agarose TAE gel, and visualized by ethidium bromide staining under UV illumination. Lane M: high molecular weight marker

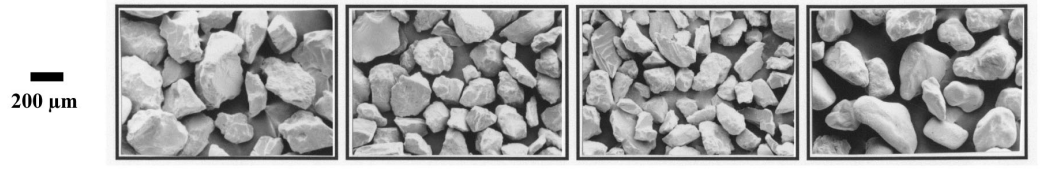
Since the yield of cell-free DNA from this beach was 34 µg ml^{-1} (34 000 µg l^{-1}), we calculated that dissolved DNA was concentrated 20 000-fold (34 000 µg l^{-1} ÷ 1.7 µg l^{-1}) on sand from this beach. Crude nucleic acid released from sand was shown to be cell-free and unencapsidated DNA by its sensitivity to DNAse (Fig. 2).

Table 1. Minerals present in beach sands studied

Mineral	Formula
Quartz	SiO_2
Feldspar (Plagioclase, K-feldspar)	$NaAlSi_3O_8$, $KAlSi_3O_8$
Amphibole (mostly hornblende)	$(Ca,Na)(Mg,Fe)_4$ $(Al,Fe,Ti)_3Si_6O_{22}(OH)_2$
Biotite	$K(Mg,Fe)AlSi_3O_{10}(OH)_2$
Chlorite	$(Mg,Fe,Al)_6(Al,Si)_4O_{10}(OH)_8$
Calcite (shell fragments)	$CaCO_3$
Magnetite	$FeFe_2O_4$
Gypsum	$CaSO_4 \cdot 2(H_2O)$

Physical properties of beach sand

The geologic and physical characteristics and chemical composition of the beach sand used in this study are summarized in Table 1 and Fig. 3. The size of grains varied from 0.1 to 1 mm, as seen in scanning electron micrographs (Fig. 2). X-ray energy dispersion (XRD) spectrometry was used to confirm the elemental composition of the sands. Silicon was the dominant element in the bulk samples, followed in order of abundance by aluminum, potassium, iron, calcium, titanium and magnesium. Chloride and iodine peaks were present in dried samples of sand before washing and DNA elution.



	PB	LJS	IB	ZV
Grain Size	0.15–1 mm	0.15–0.3 mm	0.1–1 mm	0.1–0.8 mm
Quartz	94%	94%	82%	94%
Feldspars	3%	1%	4%	trace
Biotite	trace	trace	8%	2%
Chlorite	<1%	nd	2%	2%
Amphiboles	2%	5%	2%	1%
Magnetite	<1%	nd	nd	trace
Gypsum	trace	nd	nd	nd
Calcite	trace	trace	2%	1%
DNA Yield (µg/ml)	34	19	15	91

Fig. 3. Scanning electron microscopy, mineralogic characteristics, and DNA yields of ocean beach sand. Sand collection sites were from 3 San Diego, California beaches (Pacific Beach: PB; La Jolla Shores: LJS; Imperial Beach: IB); and 1 from the Netherlands (Zandvoort: ZV). DNA yields were calculated as mean of at least 3 isolations from 15 ml of sand from each beach. nd: none detected

Yields and size distribution of sand DNA

The yield of DNA isolated from the sand of the 4 beaches shown in Fig. 3 ranged from 15 to 91 µg ml^{-1} of wet sand. The mean yield of DNA released and purified from sand collected from 3 continents (North America, Europe, Australia) and New Zealand, from the margins of 9 seas and 14 beaches, was 29.0 µg ml^{-1} (±22). The range was 2.2 to 91 µg ml^{-1} of wet sand. The 95% confidence interval was 16 to 42 µg ml^{-1}.

Pulsed field gel electrophoresis was used to determine the physical length of DNA isolated from beach sand; 6 other beaches sampled yielded similar size distributions, with DNA ranging from 5 to 300 kb in length, and a modal size of 30 kb (Fig. 4).

Characteristics of sand DNA 'genome'

We analyzed 2571 clones randomly selected from a plasmid library of 1.5×10^6 independent clones. The average insert size was 2.5 to 3 kb. The average reading length per sequence was 692 nucleotides. A total of 3447360 nucleotides were sequenced and analyzed.

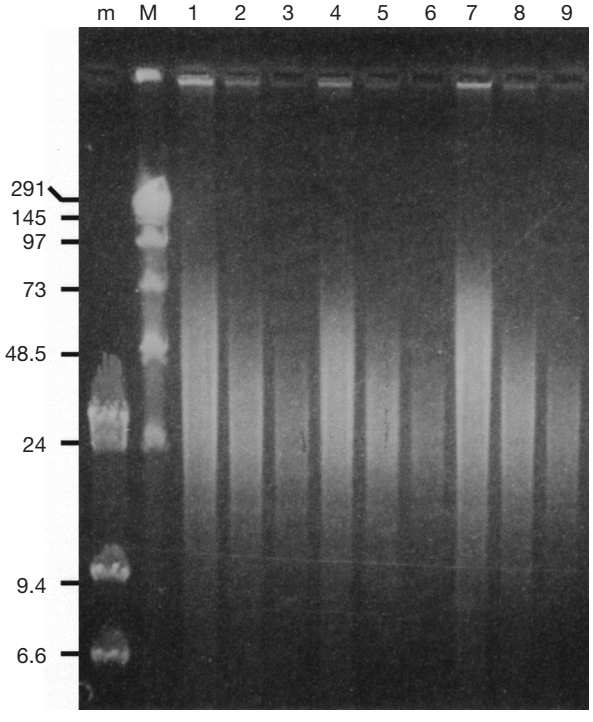


Fig. 4. Physical length of sand DNA determined by pulse field gel electrophoresis. Lane m: low molecular weight marker, 6.6 to 24 kb; Lane M: high molecular weight marker, 24 to 291 kb; Lanes 1 to 9: sand DNA from southern California beaches Pacific Beach (Lanes 1 to 3), La Jolla Shores (Lanes 4 to 6), and Imperial Beach (Lanes 7 to 9); Lanes 1, 4, 7: first low salt elution; Lanes 2, 5, 8: second low salt elution; Lanes 3, 6, 9: third low salt elution

This yielded 3107399 nucleotides of non-redundant sequence. The modal GC content of the cloned sequences was 61%, with a range of 21% to 84%. Open reading frames were found in all sequences, and ranged in size from 21 to 100% of the sequenced length. The average open reading frame was 594 nucleotides (198 amino acids) long, when the standard translational code was used; 2562 (99.6%) of the 2571 sequences from the southern California sand DNA library we sampled (PB) were new to the world's genetic databases. The 9 previously seen sequences had DNA sequence matches with e-values of 0, bit scores of 643 to 1306, and identities ranging from 76 to 99%. The strongest amino acid match was to the sequence of the beta subunit of an RNA polymerase (Accession No. Q7URW6) from the bacterium *Pirellula* sp., which was 90.8% identical, had an e-value of 1×10^{-125} and a bit score of 455. Based on DNA sequence alignments, 1584 (62%) of the sand DNA sequences were from new phylotypes, not yet represented in the world's public genetic databases. A total of 987 (38%) of the sand DNA sequences matched sequences in the current public databases. No significant contiguous blocks of overlapping DNA sequences (contigs) could be assembled among our sample of 2571 clones. This was expected, since we sampled only about 1/45000th of the DNA complexity contained in sand DNA from our reference beach.

Self-similarity of sequences in PB sand DNA library

A BLASTN (Altschul et al. 1990) analysis was used to search for DNA sequence matches among the 2571 clones within the PB library. To maximize the probability that our self-similarity analysis identified genes that were physically sampled twice, rather than merely showing phylogenetic similarity, we searched for DNA alignments of ≥99%, e-values equal to 0, and bit scores >600. This analysis revealed that 2567 (99.84%) of the sequences were unique within the library. We found 4 genes exactly twice. Of these, 1 was a proline-rich protein from the nitrogen-fixing symbiotic bacterium *Bradyrhizobium japonicum* (blr0521). The other 3 genes were new, and had no matches to genes in GenBank.

Gene functional analysis

Based on amino acid sequence matches to the COG database, 1377 (53%) clones were assigned to 1180 functions. Of the clones with assignable functions, 53% were dedicated to intermediary metabolism, 10% to translation, 10% to DNA replication, recombination and repair, 9% to cell wall or membrane synthesis, 7%

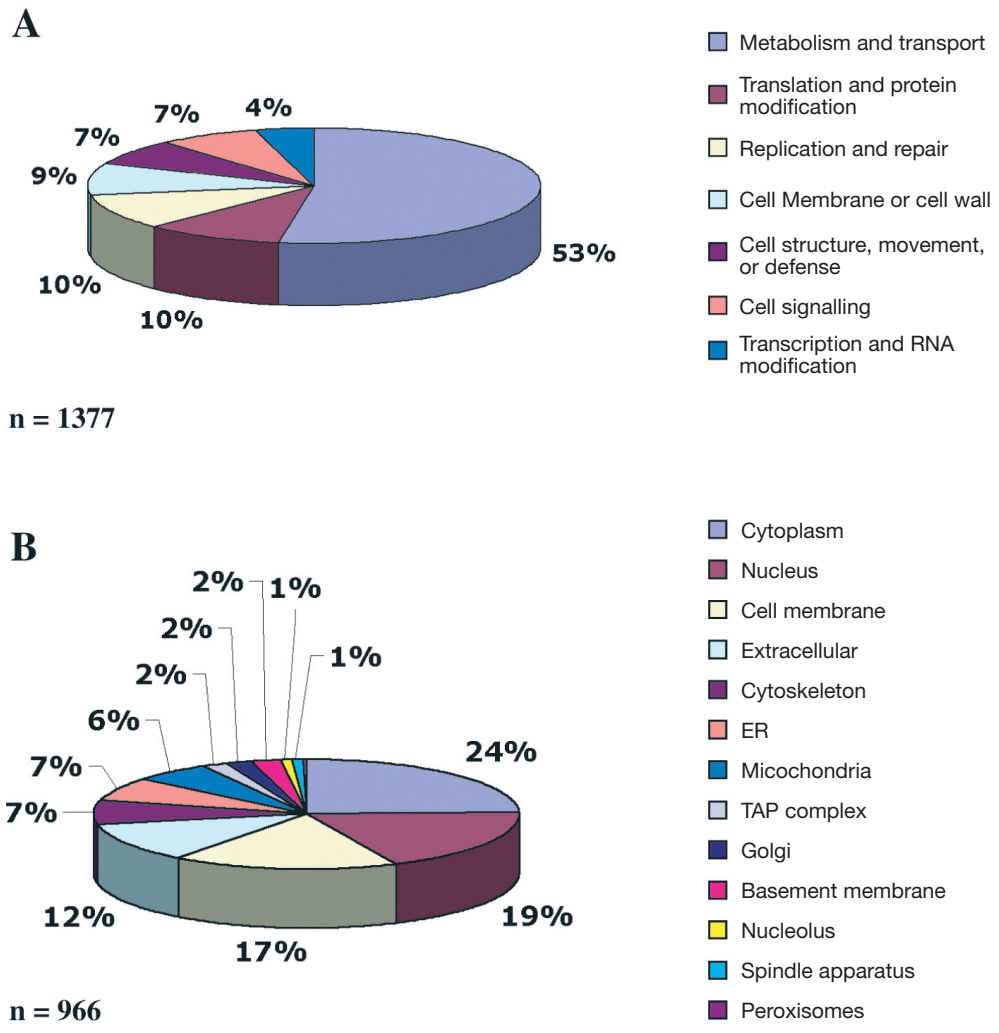


Fig. 5. Functional distribution of genes from PB sand DNA library. (A) Cluster of orthologous group (COG) gene functions; (B) gene ontology (GO)-predicted subcellular locations. ER: endoplasmic reticulum; TAP: peptide transport complex

to cell signalling, 7% to cell structure, movement or defense, and 4% to RNA transcription (Fig. 5A).

A weakness of the COG analysis is that it does not include genes that are not conserved down through bacteria. We augmented the COG analysis with a BLASTX search of the seqdblite protein database created by the GO consortium of the EBI. The GO analysis permitted us to identify 1544 genes (60%) that encode 165 cellular functions unique to eukaryotes and viruses, as well as conserved down through, or unique to bacteria. The distribution of sand DNA among the top 13 GO subcellular locations is illustrated in Fig. 5B. Of the clones with assignable GO functions, 24% encoded genes from the cytoplasm, 19% from the nucleus, 17% cell membrane, 7% cytoskeleton, and 6% from mitochondria (Fig. 5B). The majority of eukaryotic genes in the sand DNA library appeared as cDNAs, without introns.

Multi-kingdom ribosomal DNA

In addition to library construction, we used the polymerase chain reaction (PCR) to amplify kingdom-selective ssu rDNA from sand DNA isolated from the same beach (PB) in San Diego. Primers were designed to amplify eukaryotic 18S (Palumbi 1996), bacterial 16S (Hinrichs et al. 1999) and archaebacterial 16S rDNA (Hinrichs et al. 1999). Of the 6 rDNAs, 2 (Clones E18S5, E16S17) were similar to sequences previously known from samples obtained from deep-sea sediments or pelagic ecosystems (Table 2). Another 2 of the 6 rDNAs (Clones E18S9 and E16S16) were similar to sequences from established endosymbionts. The remaining rDNA (Clone A16S14) showed closest similarity to mesohalophilic archaebacteria (Table 2).

Table 2. Details of 6 random small subunit rDNA sequences cloned from sand DNA. NCBI: National Center for Biotechnology Information: Dozens to hundreds of matches were identified for each clone; only top 1 or 2 matches are shown

Primers PCR	Clone No.	Best match in NCBI database	Accession No.	Sequenced DNA (nt)	% identity	Gaps	Probability of match by chance
Eukarya	E18S5	*Dimorpha*-like Cercozoan protist	AF174374	376	87	0	2×10^{-87}
185-F-5′CTGGTTGATCCTGCCAGT 18S-R-5′AACCCTGATTCCCCGTCACC Typical amplicon size = 419 bp	E18S9	Photosynthetic eukaryotic endo-symbiont of *Peridinium balticum*	ES28SPF	358	94	1	1×10^{-124}
Bacteria	E16S16	Chloroplast DNA of diatom *Bacillaria paxillifer*	BPA536452	662	96	0	$<1 \times 10^{-180}$
E16S-F-5′-AGAGTTTGATCCTGGCTCAG E16S-R-5′-GFGTTACCTTGTTACGACTT Typical amplicon size = 1465 bp	E16S17	Unidentified gamma proteobacterium from deep-sea sediments	AB015583	633	97	0	$<1 \times 10^{-180}$
Archaea	A16S13	Uncultured Archeon from a saltmarsh	AF015981	657	97	0	$<1 \times 10^{-180}$
A16S-F-5′-TTCCGGTTGATCCYGCCRG A16S-R-5′-YCCGGCGTTGAMTCCAATT Typical amplicon size = 938 bp	A16S14	Halophilic Archeon 14AHC	AY292398	662	96	0	$<1 \times 10^{-180}$

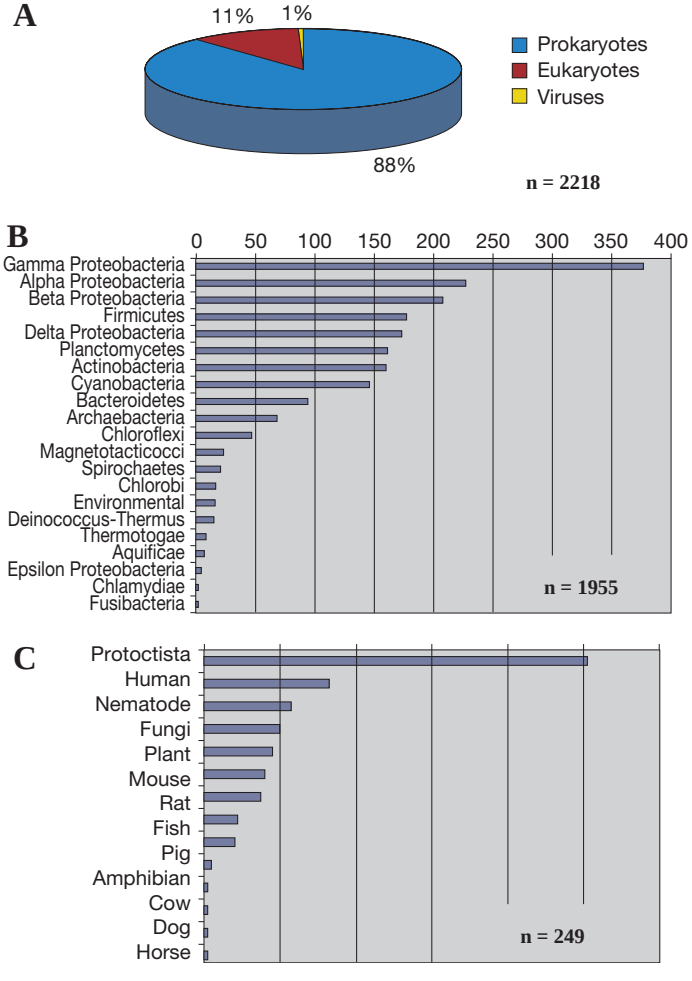


Fig. 6. Taxonomy of PB sand DNA sequences. (A) Prokaryotes, eukaryotes and viruses; (B) prokaryotes; (C) eukaryotes

Taxonomic complexity of sand DNA

Taxonomic assignment of environmental DNA sequences is difficult and requires a number of assumptions. In most cases, a single new DNA sequence will show similarities to cognate genes from several different species, simply because of the evolutionary relationships between all life on earth. We adopted the convention of assigning a taxonomic identifier to a sand DNA sequence according to the best match (lowest e-value) to a known sequence as of the date of this analysis (April 2004). Using this criterion, amino acid sequence alignments permitted the best-fit assignment of taxonomic origins of 2218 (86.3%) of the clones from sand DNA. Of these clones, 14% (353) had no significant matches in public databases at the time of writing, and therefore constituted DNA from novel genes for which no taxonomic data could be adduced. Of the assignable clones, 88% (1955) were prokaryotic, 11% (249) eukaryotic, and 1% (14) viral (Fig. 6A). Of the prokaryote sequences, 1887 were bacterial, and 68 were archaebacterial. The gamma subdivision of Proteobacteria was the largest taxonomic group among the bacteria, comprising 377 sequences, and including over 35 different species (Fig. 6B). Of the eukaryote sequences, 114 were animals, 101 protoctista (single-celled eukaryotes, including zooplankton, certain phytoplankton and algae), 18 fungi, and 16 plant. In our samples, 33 human sequences, 23 arthropod, 20 nematode, 15 mouse, 9 rat, 8 fish, 2 pig, and single frog, horse, cow, and dog sequences were present (Fig. 6C).

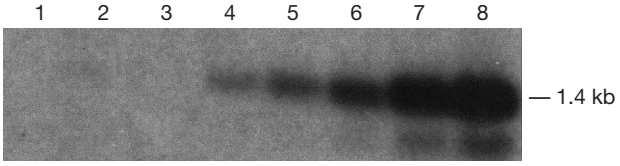


Fig. 7. Estimaton of single gene complexity of sand DNA by quantitative Southern blot analysis. Lane 1: uncut PB sand DNA; Lane 2: *Bam*HI cut PB (Pacific Beach) sand DNA. Lanes 3 to 8 were loaded with internal calibration curve of Clone A4 containing 5 pg M13mp18 control (Lane 3), 0.2 pg (Lane 4), 0.5 pg (Lane 5), 1 pg (Lane 6), 2 pg (Lane 7) and 5 pg (Lane 8) of gel-purified, 1.4 kb *Bam*HI fragment of sand DNA Clone A4

Comparison with microbial sequences from the Sargasso Sea

DNA sequence alignments identified just 54 clones from the sand DNA library that matched microbial sequences isolated from the Sargasso Sea (Venter et al. 2004). Of these, 40% were Proteobacteria, 11% Planctomycetes, 9% Firmicutes (including Clostridia, Bacilliaceae, Streptococci, Listeriacieae and other groups), 8% Cyanobacteria, 8% Actinobacteriacea, 3% Bacteroidetes, and the remaining 6% were distributed among 6 other large bacterial groups.

Estimation of genomic complexity of sand DNA

Quantitative Southern blot analysis was used to obtain an estimate of the effective genome size of sand DNA. This is illustrated in Fig. 7, which shows the results of a quantitative Southern blot transfer of known amounts of the 1.4 kb fragment of a CDC48-like clone purified from the PB sand DNA reference library. The hybridization intensity of the corresponding 1.4 kb band in 5 µg of total sand DNA was <0.05 pg (Fig. 7: Lane 2). Based on these results, the calculated, effective genome size of sand DNA from the PB reference beach was 1.4×10^{11} bp. This genome equivalent was contained in 150 pg of DNA present on 5.2 nl (~10 µg) of wet ocean beach sand (see 'Materials and methods').

In these experiments, a 1.4 kb clone from sand DNA was used as probe that was strongly related to CDC48 from *Deinococcus radiodurans* (NCBI Protein Accession No. D75493). The selection of the CDC48-like clone and quantitative Southern blot analysis was based on the assumption that highly conserved sequences might be over-represented in the sand DNA library. The existence of a multigene family or cross-hybridizing genes that fortuitously produced 1.4 kb *Bam*HI restriction fragments that comigrated with the marker gene would intensify the band on the Southern blot and produce a falsely low estimate of complexity.

Table 3. Species distribution of sand DNA determined by amino acid alignments: 2218 of 2571 (86.3%) clones sequenced could be assigned to species by amino acid alignments

Taxonomic group	Sequences (n)	Species (n)
Bacteria	1887	144
Protoctista	101	11
Vertebrates	71	11
Archaebacteria	68	15
Invertebrates	43	4
Fungi	18	10
Plants	16	3
Viruses	14	9
Total	2218	207

On the other hand, the presence of satellite DNA, or other genes that might be present in high copy number within the sand DNA genome would lower the effective single-copy gene complexity. In our sample of 2571 clones, we found that fewer than 1.4% of the nucleotides consisted of short repeat sequences; however, the fractions of single-copy and repetitive DNA will be characterized in future studies by classical DNA renaturation kinetics and $C_o t$ analysis (Peterson et al. 2002).

Gene and species diversity

Throughout this paper we use the term diversity to mean richness, or the number of identifiable genes or species. Amino acid sequence alignments identified a total of 207 different species (Table 3) that contributed 2218 (86%) of the 2571 sequences and 3.11 million nucleotides analyzed from the PB sand DNA library. This produced a species:non-redundant nucleotide ratio of 7.74×10^{-5} $\left(207 \text{ species} \div (0.86 \times 3.11 \times 10^6 \text{ bp})\right)$. The gene:nucleotide ratio for both amino acid- and DNA-

Table 4. Species distribution of sand DNA determined by DNA sequence alignments: 987 of 2571 (38.4%) clones sequenced could be assigned to species by DNA alignments

Taxonomic group	Sequences (n)	Species (n)
Bacteria	894	94
Vertebrates	36	8
Archaebacteria	18	11
Plants	11	4
Invertebrates	10	3
Protoctista	10	7
Fungi	6	4
Viruses	2	2
Total	987	133

sequence alignments was 8.27×10^{-4} (2571 genes ÷ 3.11×10^6 bp). Using our estimate of 1.4×10^{11} bp for the minimum size of the sand DNA genome and assuming linear scaling based on the gene and species abundance obtained from amino acid sequence alignments, we calculated that sand DNA from PB contained more than 116 million genes (8.27×10^{-4} gene $bp^{-1} \times 1.4 \times 10^{11}$ bp) from up to 11 million species (7.74×10^{-5} species $bp^{-1} \times 1.4 \times 10^{11}$ bp).

DNA sequence alignments identified a total of 133 different species (Table 4) from 987 (38.4%) of the 2571 sequences and 3.11 million nucleotides analyzed from the PB sand DNA library. This produced a species:non-redundant nucleotide ratio of 1.11×10^{-4} $\left(133 \div (0.384 \times 3.11 \times 10^6 \text{ bp})\right)$, and a gene:nucleotide ratio of 8.27×10^{-4} ($2571 \div 3.11 \times 10^6$ bp). Using our estimate of 1.4×10^{11} bp for the minimum size of the sand DNA genome and assuming linear scaling based on gene and species abundance, we calculated that sand DNA from PB contains more than 116 million genes (8.27×10^{-4} gene $bp^{-1} \times 1.4 \times 10^{11}$ bp) from 15.6 million species (1.11×10^{-4} species $bp^{-1} \times 1.4 \times 10^{11}$ bp). Combining the estimates from protein and DNA sequence alignments, the sand DNA library contained a sample of genes from up to 11 to 16 million species. We consider these numbers to be soft estimates because of the inherent limitations of species assignments by GenBank analysis.

The estimates of gene and species diversity in sand DNA required several assumptions: (1) The assignment of each gene and species was based on a match to a DNA sequence in GenBank. This made our results strongly dependent on the coverage of relevant marine and terrestrial species in GenBank at the time of this analysis in 2004. (2) When a sand DNA sequence matched a gene or genes from several different species, only the best match was used to make the species assignment. The 'true' species may have been one that is not yet covered by a sequence deposited in GenBank. We could not use a GenBank-independent, sequence difference measurement like the operational taxonomic unit (OTU) used in ribosomal DNA studies of species diversity (Hughes et al. 2001, Schloss & Handelsman 2004), because we studied thousands of genes, not just 1. Since different genes evolve at different rates, different gene-specific OTU thresholds apply. An OTU-based study could be done by expanding the ssu rDNA PCR experiments in Table 2. However, that was not the focus of this report. (3) We used linear scaling to estimate the diversity of genes and species in our reference library. The gene and species diversity contained in sand DNA are likely to be smaller than that predicted by linear scaling. However, the degree of departure from linear scaling is difficult to estimate in the face of such incomplete sampling and the extreme genetic diversity found in sand DNA. We sampled just 1/45 000th of the estimated sequence content of sand DNA (3.11×10^6 bp sequenced ÷ 1.4×10^{11} bp/sand DNA genome = 1/45 016). The application of classical sampling theories in ecology such as the nonparametric Chao1 estimator (Chao 1984) is attractive, but not appropriate for our data. The Chao1 estimate of diversity does not become accurate until the sample size exceeds a value equal to the square root of twice the true diversity (Colwell & Coddington 1994). If, for example, the true diversity of sand DNA is close to our estimate of 116 million genes (1.16×10^8), then Chao1 will begin to estimate the true diversity only after about 15 200 clones—$(2 \times 1.16 \times 10^8)^{1/2}$ = 15 232—have been sequenced.

Accurate estimates of species diversity are even more difficult than those of gene diversity. We could not use Chao1 to estimate the species diversity in the sand DNA assemblage because our species counts were not based on a single gene known to be present in every species. The species counts we report are dependent on the species content, coverage, and a statistical match to sequences deposited in GenBank. At the time of our analysis, 159 prokaryotic, 4 eukaryotic, and 1400 viral genomes had been completely sequenced, at least 1 gene from 185 000 species was archived (about 150 000 of these were from eukaryotic species, and 35 000 were from prokaryotes and viruses), and a total of 38 989 342 565 nucleotides of DNA sequence were available for alignment searches.

DISCUSSION

Current estimates of global prokaryotic biodiversity range from 6×10^6 (Curtis et al. 2002) to 1×10^9 (Dykhuizen 1998) species, with a standing population size of about 5×10^{30} individuals and an associated biomass of about 5.2×10^{17} g of carbon (Whitman et al. 1998). The biomass of prokaryotes is nearly equal to that of all the plants and rainforests on earth (5.6×10^{17} g of carbon; Schlesinger 1991). Estimates of global eukaryotic biodiversity range from 1 to 5×10^7 (May 1988). Taxonomists have given names to about 1.5×10^6 eukaryotic species (Stork 1997). Most of these belong to the more charismatic species, visible to the human eye, that have attracted both popular and scientific attention. On the other hand, prokaryotes and microbial eukaryotes have received significantly less attention. There are names for just 6200 of the prokaryotic species (Oren 2004).

DNA-based methods have revolutionized the taxonomy of microbial species (Lane et al. 1985, Stackebrandt et al. 1993), and are contributing to a revolution in the taxonomy of animal species through the use of

genetic 'barcodes' derived from the DNA sequences from selected mitochondrial genes (Hebert et al. 2003). The genetic assessment of biodiversity by methods of environmental population genomics, sometimes called 'metagenomics' (Handelsman 2004), has revealed a richness of genetic variation that has exceeded all expectations. In a recent large-scale sequencing effort, Venter et al. (2004) sequenced 1.045×10^9 nucleotides of DNA isolated from microbes in the Sargasso Sea and discovered 1.2×10^6 previously unknown genes. The ratio of DNA sequenced to genes discovered was 871 nucleotides to 1 new gene. In our study of sand DNA, we sequenced 3.1×10^6 nucleotides and discovered 2562 new genes. The ratio of DNA sequenced to genes discovered in our sample was 1210 nucleotides to 1 new gene, in keeping with the expected larger size of eukaryotic genes mixed with shorter prokaryotic sequences in sand DNA. While the sequencing effort from the Sargasso sea exceeded ours by some 340-fold, it suggests that the genetic diversity in the sea is so vast that the use of linear scaling to estimate the gene diversity of sand DNA may be accurate up to at least the 1.045×10^9 nucleotides sequenced by Venter et al. (2004). For example, if 1×10^9 nucleotides of sand DNA were sequenced, linear scaling would predict that 8.3×10^5 new genes (1×10^9 bp ÷ 1210 bp gene^{-1}) would be identified. Environmental DNA sequencing cannot replace the need for studying the natural history of single species and the trophic interactions among species, but it can rapidly fill in our picture of the true diversity of life on Earth, and help single out interesting species for more detailed study.

The concentration of dissolved DNA in seawater is low (1 to 12 µg l^{-1}) (Jiang & Paul 1995), making direct isolation of dissolved DNA from seawater difficult and expensive. Current DNA-based methods for evaluating the biodiversity of the oceans typically employ drag filters and other equipment deployed from seagoing research vessels, submersibles, piers, or used by divers. These methods permit the direct capture of cells (Venter et al. 2004), viral particles (Breitbart et al. 2004, Venter et al. 2004), or pelagic microcosms of organic material known as sea snow (Azam & Long 2001). However, they were not designed or intended to assess the presence or relative abundance of cells from multicellular organisms such as seaweeds, coral reef and pelagic invertebrates, and marine vertebrates such as fishes, birds and mammals. On the other hand, sand DNA is not limited by these restrictions. Silicate-rich, wave-washed sand concentrates the DNA in seawater over 10 000-fold, contains the DNA of many multicellular animals and, in theory, may contain a sample of DNA from any virus or life form that has released its DNA to the hydrosphere.

Sand DNA is ubiquitous along continental and island coastlines, easy to collect, and easy to process. In our study of 14 beaches bordering 9 seas, the mean yield of sand DNA was 29 µg ml^{-1} wet sand. Using a conservative value of 20 µg DNA ml^{-1} beach sand, a swath of ocean beach measuring 1 km long × 10 m wide potentially contains 5 kg of cell-free DNA in the top 2.5 cm of wave-washed sand. Using our estimate of 1.4×10^{11} bp for the minimum size of the sand DNA genome, we calculated that the DNA from a single handful of ocean beach sand contains millions of genes, sampled from a large number of species in and about the sea. Over 99% of the genes (2562 of 2571) in sand DNA were new, not previously available in the world's public genetic databases. The availability of a concentrated, easily accessible, and protected natural deposit of diverse DNAs on beaches around the world opens several doors for gene discovery and the molecular analysis of coastal ecosystems, and creates a powerful new resource to help explore, monitor and protect the living diversity of the oceans of the world.

Acknowledgements. R.K.N. was supported by grants from the Lennox Foundation, the UCSD Christini Fund, the USCD Department of Medicine, and a generous gift from Madett and Dorothy Engs. He thanks Miriam Kastner, and Charles Graham for helpful discussions, electron microscopy, and assistance in analyzing the elemental composition of sand; Richard Haas and William Nyhan for helpful discussions and support; Pierre Sonigo, Jean-Jacques Kupiec, and Marc Sitbon for expert advice, and Anthony Mazeroll and Nicolas Manel for many helpful suggestions. R.K.N. also thanks the following people for collecting sand: Peter Pearson and the Buckingham Friends School, Beach Haven, New Jersey; Mike Pelloth, Siesta Key, Florida; Robert Jansen, Sydney, Australia; Diane Holland, Greymouth, New Zealand; Elizabeth Symes, West Sussex, UK; Anthony Linnane, Melbourne, Australia; and David Nolan, Perth, Australia. J.C. is holder of a Senior Canadian Research Chair in medical genomics and gratefully acknowledges the support of the UCSD Center for AIDS Research (CFAR) genomics core laboratory.

LITERATURE CITED

Altschul SF, Lipman DJ (1990) Protein database searches for multiple alignments. Proc Natl Acad Sci USA 87: 5509–5513

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215:403–410

Andersson B, Wentland MA, Ricafrente JY, Liu W, Gibbs RA (1996) A 'double adaptor' method for improved shotgun library construction. Anal Biochem 236:107–113

Azam F, Long RA (2001) Oceanography—sea snow microcosms. Nature 414:495–497

Breitbart M, Felts B, Kelley S, Mahaffy JM, Nulton J, Salamon P, Rohwer F (2004) Diversity and population structure of a near-shore marine-sediment viral community. Proc R Soc Lond Ser B 271:565–574

Camon E, Magrane M, Barrell D, Lee V and 6 others (2004)

The gene ontology annotation (GOA) database: sharing knowledge in uniprot with gene ontology. Nucleic Acids Res 32:D262–266

Cavalier-Smith T (1998) A revised six-kingdom system of life. Biol Rev Camb Philos Soc 73:203–266

Chao A (1984) Nonparametric-estimation of the number of classes in a population. Scand J Stat 11:265–270

Colwell RK, Coddington JA (1994) Estimating terrestrial biodiversity through extrapolation. Philos Trans R Soc Lond B 345:101–118

Curtis TP, Sloan WT, Scannell JW (2002) Estimating prokaryotic diversity and its limits. Proc Natl Acad Sci USA 99:10494–10499

Dykhuizen DE (1998) Santa Rosalia revisited: why are there so many species of bacteria? Antonie Leeuwenhoek 73:25–33

Handelsman J (2004) Metagenomics: application of genomics to uncultured microorganisms. Microbiol Mol Biol Rev 68: 669–685

Hebert PD, Cywinska A, Ball SL, deWaard JR (2003) Biological identifications through DNA barcodes. Proc R Soc Lond Ser B 270:313–321

Hinrichs KU, Hayes JM, Sylva SP, Brewer PG, DeLong EF (1999) Methane-consuming archaebacteria in marine sediments [see comments]. Nature 398:802–805

Hughes JB, Hellmann JJ, Ricketts TH, Bohannan BJ (2001) Counting the uncountable: statistical approaches to estimating microbial diversity. Appl Environ Microbiol 67: 4399–4406

Jiang SC, Paul JH (1995) Viral contribution to dissolved DNA in the marine environment as determined by differential centrifugation and kingdom probing. Appl Environ Microbiol 61:317–325

Lane DJ, Pace B, Olsen GJ, Stahl DA, Sogin ML, Pace NR (1985) Rapid determination of 16S ribosomal RNA sequences for phylogenetic analyses. Proc Natl Acad Sci USA 82:6955–6959

Margulis L (1992) Symbiosis theory: cells as microbial communities. In: Margulis L, Olendzenski L (eds) Environmental evolution. MIT Press, Cambridge, MA, p 149–172

Margulis L, Schwartz KV (1998) Five kingdoms—an illustrated guide to the phyla of life on Earth. WH Freeman & Co, New York

May RM (1988) How many species are there on Earth? Science 241:1441–1449

Muller WE, Krasko A, Le Pennec G, Schroder HC (2003) Biochemistry and cell biology of silica formation in sponges. Microsc Res Tech 62:368–377

Nakae D, Mizumoto Y, Kobayashi E, Noguchi O, Konishi Y (1995) Improved genomic/nuclear DNA extraction for 8-hydroxydeoxyguanosine analysis of small amounts of rat liver tissue. Cancer Lett 97:233–239

Oren A (2004) Prokaryote diversity and taxonomy: current status and future challenges. Philos Trans R Soc Lond B 359:623–638

Palumbi SR (1996) Nucleic acids. II. The polymerase chain reaction. In: Hillis DM, Moritz C, Mable BK (eds) Molecular systematics. Sinauer, Sunderland, MA, p 233–234

Peterson DG, Schulze SR, Sciara EB, Lee SA and 6 others (2002) Integration of Cot analysis, DNA cloning, and high-throughput sequencing facilitates genome characterization and gene discovery. Genome Res 12:795–807

Sambrook J, Fritsch EF, Maniatis T (1989) Molecular cloning, a laboratory manual; 2nd edn. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY

Schlesinger WH (1991) Biogeochemistry: an analysis of global change. Academic Press, San Diego, CA

Schloss PD, Handelsman J (2004) Status of the microbial census. Microbiol Mol Biol Rev 68:686–691

Stackebrandt E, Liesack W, Goebel BM (1993) Bacterial diversity in a soil sample from a subtropical Australian environment as determined by 16S rDNA analysis. FASEB J 7:232–236

Stork NE (1997) Measuring global biodiversity and its decline. In: Reaka-Kulda ML, Wilson DE, Wilson EO (eds) Biodiversity. II. Joseph Henry Press, Washington, DC, p 41–68

Venter JC, Remington K, Heidelberg JF, Halpern AL and 19 others (2004) Environmental genome shotgun sequencing of the Sargasso Sea. Science 304:66–74

Vogelstein B, Gillespie D (1979) Preparative and analytical purification of DNA from agarose. Proc Natl Acad Sci USA 76:615–619

Wang L, Hirayasu K, Ishizawa M, Kobayashi Y (1994) Purification of genomic DNA from human whole blood by isopropanol-fractionation with concentrated NaI and SDS. Nucleic Acids Res 22:1774–1775

Whitman WB, Coleman DC, Wiebe WJ (1998) Prokaryotes: the unseen majority. Proc Natl Acad Sci USA 95:6578–6583

Appendix 1. Thirty-three random clones from a sand DNA library. The ORF underlined is the one that produced the strongest match to a protein or proteins in the GenBank database as of 10 August 2000. Clone A4 was used in the quantitative Southern hybridization analyses. Best protein match in NCBI database: many matches shown were to proteins found in a variety of different species, and in the cases of highly conserved proteins, solid amino acid matches were observed among organisms from different kingdoms of life. Therefore, the species matches indicated should not be interpreted as the only possible origin for a given clone

No.	Clone	Insert size (kb)	Sequenced DNA (nt)	Longest ORF (AA)	Best of DNA match in NCBI database (standard translational code)	Best protein match in NCBI database	Accession No.	% identity	% similarity	Gaps (%)	e-value
1	A3	1.5	537	164/90	None	None					
2	A4	1.4	616	173/63	None	Deinococcus CDC48	D75493	48	72	4	1×10^{-27}
3	A5	5.3	578	100/62	None	None					
4	A6	3.1	705	232/205	None	Cyanobacterial permease	S75996	31	53	8	2×10^{-10}
5	A7	1.6	676	172/135	None	None					
6	A11	0.7	574	186/185	Streptomyces/AL161755	Streptomyces kinase	AL161755	41	54	9	6×10^{-12}
7	A13	4.1	585	192/100	None	None					
8	A14	1.2	652	138/132	None	Methanoococcus glycosyltransferase	Q58619	31	50	3	4×10^{-4}
9	A20	1.4	640	180/133	None	None					
10	A21	1.1	697	149/102	None	None					
11	A22	1.1	441	136/129	None	Xenopus skin secretory protein	P17437	33	37	7	3×10^{-3}
12	B2	2.0	665	106/105	None	None					
13	B5	0.4	345	94/78	None	Eubacterial formyl CoA transferase	C65011	79	85	1	2×10^{-21}
14	B8	2.8	705	175/135	None	None					
15	B10	0.4	445	132/75	None	None					
16	B12	6.1	699	137/53	None	None					
17	B13	0.7	656	172/167	None	Myxococcus protein kinase	AAD42851	43	66	3	2×10^{-32}
18	B14	1.1	605	180/159	None	Neisseria spermidine synthase	AAF41280	32	52	2	7×10^{-15}
19	B15	1.4	581	181/170	None	Sea urchin hyalin	AAC72757	30	43	8	2×10^{-6}
20	B17	1.5	688	133/120	None	Epstein-Barr virus protein	P03181	40	40	16	3×10^{-5}
21	B18	3.0	638	204/78	None	Thermophilic, eubacterial anaerobe	AAB06263	24	42	1	8×10^{-7}
22	B20	4.5	588	95/83	None	None					
23	B23	6.0	725	182/113	None	Neisseria DNA repair polymerase	CAB72023	48	67	1	2×10^{-42}
24	B24	1.1	566	148/96	None	Abalone lustrin A	T08852	28	39	4	3×10^{-3}
25	B25	1.2	699	139/110	None	None					
26	B26	2.0	619	139/101	None	None					
27	B27	1.3	647	112/109	Rho/L27275	Pseudomonas rho terminator	P52155	71	81	0	7×10^{-34}
28	B30	3.0	566	136/99	None	None					
29	B31	2.2	598	194/97	None	Mycobacterial RNA/DNA helicase	B70841	34	51	12	2×10^{-8}
30	B32	2.0	649	193/90	None	Eubacterial subtilisin-like protease	A69587	49	66	5	5×10^{-31}
31	B33	1.9	668	193/153	None	Tunicate myosin	BAA08111	29	51	12	4×10^{-5}
32	B34	2.9	568	151/69	None	None					
33	B35	2.3	530	165/70	None	Unknown Drosophila protein	AAF56337	34	46	8	7×10^{-3}
Mean		2.2	611	157/111	2/33 (6%)	18/33 (55%) with protein matches 12/18 (67%) bacterial proteins 5/18 (28%) eukaryote proteins 1/18 (6%) viral proteins					
SD		1.5	84	34/37							

Editorial responsibility: Otto Kinne (Editor-in-Chief), Oldendorf/Luhe, Germany

Submitted: October 19, 2004; Accepted: April 28, 2005
Proofs received from author(s): September 20, 2005