nature biotechnology

Sequencing genomes from single cells by polymerase cloning

Kun Zhang[1], Adam C Martiny[2], Nikos B Reppas[1], Kerrie W Barry[3], Joel Malek[4], Sallie W Chisholm[2] & George M Church[1]

Genome sequencing currently requires DNA from pools of numerous nearly identical cells (clones), leaving the genome sequences of many difficult-to-culture microorganisms unattainable. We report a sequencing strategy that eliminates culturing of microorganisms by using real-time isothermal amplification to form polymerase clones (plones) from the DNA of single cells. Two *Escherichia coli* plones, analyzed by Affymetrix chip hybridization, demonstrate that plonal amplification is specific and the bias is randomly distributed. Whole-genome shotgun sequencing of *Prochlorococcus* MIT9312 plones showed 62% coverage of the genome from one plone at a sequencing depth of 3.5×, and 66% coverage from a second plone at a depth of 4.7×. Genomic regions not revealed in the initial round of sequencing are recovered by sequencing PCR amplicons derived from plonal DNA. The mutation rate in single-cell amplification is $< 2 \times 10^5$, better than that of current genome sequencing standards. Polymerase cloning should provide a critical tool for systematic characterization of genome diversity in the biosphere.

Over the past two decades the exponential increase in DNA sequencing has resulted in the release of over 355 distinct genomes[1,2]. Nevertheless, most of the genetic diversity of the biosphere remains unsampled[3,4], because conventional genome sequencing is restricted to easily cultured microorganisms. Metagenomic approaches, such as environmental shotgun sequencing and large insert library sequencing, do not require large cultured clonal pools of microorganisms. Although these metagenomic techniques reveal enormous biodiversity in environmental samples[5–7], they suffer from two major drawbacks: (i) the difficulty of assembling contigs into discrete genomes[8] (such difficulty can be partially alleviated by recent computational methods[9]), and (ii) biased sampling toward abundant species[5,10,11]. The ability to sequence an entire genome from a single uncultured cell might overcome these two limitations and enable genomic analyses such as (i) the characterization of genetic heterogeneity in a population of cells, (ii) the revelation of *cis*-relationship between sequences that are more than 200 kb apart (unreachable by BAC/fosmid cloning), (iii) the study of *trans*-interactions between host and parasitic genomes (phages and viral) or cell-cell interactions (for example, predator-prey, symbionts, commensals) and (iv) the identification of rare species for genome sequencing. Here we present polymerase cloning ('ploning'), a technique for performing genome analyses at the single-cell level.

Ploning requires whole-genome amplification from a single DNA molecule to be high yield, high fidelity and without significant bias in terms of sequence coverage[12–14]. Isothermal multiple displacement amplification (MDA)[12] is superior to PCR-based methods[15–17] in all three respects, but is known to yield a dominant 'background' of undesired amplification when the template material drops below nanogram levels[18,19]. Accordingly, mixed results have been reported on such amplifications from single human cells[20–23]. Because the standard MDA protocol requires 1∼10 ng of template DNA, microorganisms with smaller genomes pose an even greater challenge as the mass of a single genome is typically at femtogram levels[12]. Genome sequencing of *Xylella fastidiosa* was possible only when genomic DNA was amplified by MDA from ∼1,000 cells[24]. Although initial success has been reported on genome amplification from single *E. coli* cells, only an estimated 30% of amplified DNA was specific to the target genome because of background amplification[25]. Reduction of reaction volume offers a way to reduce background amplification[26]. To enable single-cell genome sequencing of difficult-to-culture microorganisms, however, we must address a number of critical technical issues, such as the quantification of background, amplification bias, amplification error and the compatibility with current genome sequencing pipelines.

RESULTS

Real-time ultra-low background isothermal amplification

We hypothesized that the background amplification that currently undermines single-cell MDA arises from two sources: exogenous DNA contamination and endogenous template-independent, primer-primer interactions. To assess each source as independently as possible, we developed an ultra-sensitive, sequence-nonspecific detection system (**Supplementary Note** online, **Supplementary Figs. 1–3** online) to monitor the dynamics of isothermal amplification in real time by SYBR Green I fluorescence[27]. To suppress endogenous background amplification, we used a constrained-randomized hexanucleotide

[1]Department of Genetics, Harvard Medical School, 77 Avenue Louis Pasteur, Boston, Massachusetts 02115, USA. [2]Department of Civil and Environmental Engineering, Massachusetts Institute of Technology, 15 Vassar Street, Cambridge, Massachusetts 02139, USA. [3]United States Department of Energy Joint Genome Institute, 2800 Mitchell Drive, Walnut Creek, California 94598, USA. [4]Agencourt Bioscience, 500 Cummings Center, Suite 2450, Beverly, Massachusetts 01915, USA. Correspondence should be addressed to K.Z. (kzhang@genetics.med.harvard.edu) or G.M.C. (http://arep.med.harvard.edu/gmc/email.html).

Received 23 January; accepted 17 April; published online 28 May 2006; doi:10.1038/nbt1214

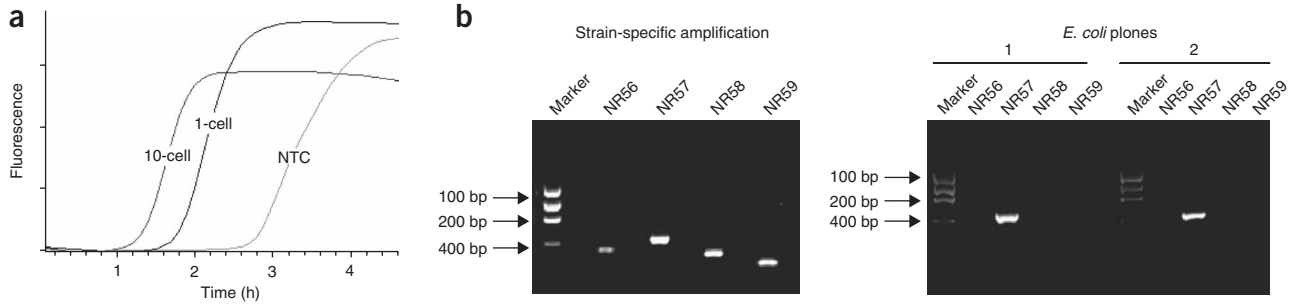

Figure 1 Ploning on *E. coli* single cells. (**a**) The genome of single *E. coli* cells was successfully amplified when the background was suppressed to a sub-femtogram level. The amplification curves from one cell and from nontemplate control (NTC) are well-separated, indicating background amplification did not interfere with template-specific amplification. The 10-cell amplification was conducted on 1 μl of *E. coli* cells diluted at the ten cell/μl level. The number of cells in the one-cell dilution was confirmed as described in the **Supplementary Note** online. The *y*-axis represents arbitrary fluorescent units. (**b**) Strain-specific amplification showed that both plones were amplified from single cells of the NR57 strain.

primer, R6 (R = A/G) that cannot cross-hybridize. This primer permitted us to estimate exogenous DNA contamination present in reagents and labware subjected to different preparative procedures. Using this constrained-randomized primer, we arrived at a strict sample-handling protocol that reliably reduced background amplification below $\sim 10^{-4}$ femtogram/reaction—10,000× lower than a single copy of the *E. coli* genome. Next, we asked whether the level of endogenous background amplification using the totally degenerate primer N6—the most appropriate primer for non-biased MDA, but most susceptible to primer-primer interactions—was below the femtogram level. Because the effective background of nontemplate amplifications was consistently in the ~0.03 femtogram range (**Fig. 1a**), ploning of single genomes became possible.

Ploning single *E. coli* genomes

Having optimized a protocol to achieve ultra-low exogenous background, we next sought to develop a procedure to plone single cells. Ploning requires a method to assess whether an amplicon is truly from a single cell. We found that single cells prepared by standard flow-sorting are not suitable for amplification, because we could not prevent introduction of contamination during sorting. Therefore, we established a system to verify the clonality of our amplicon by using a mixture of four *E. coli* strains (NR56, NR57, NR58, NR59, all derivatives of MG1655) that can be distinguished genotypically (**Supplementary Note** online). When only one strain-specific marker is identified in an appropriate dilution of the mixed-cell population (**Fig. 1b**), the Poisson-based probability that this amplicon is from a single cell is equal to 88%, a percentage that is similar to the success rate of flow-sorted single cells[28] (see **Supplementary Note** online for detailed calculation).

We monitored ploning reactions in real time to ensure a clear kinetic separation of the amplification curves of our target sample from those of the nontemplate control (**Fig. 1a**). The two *E. coli* plones we identified in our dilution were both derived from NR57 (**Fig. 1b**). After a second round of amplification to ensure a sufficient quantity of DNA, we characterized the specificity, amplification bias and genome coverage by hybridizing the ploned DNA to Affymetrix *E. coli* Antisense Genome chips. Using *E. coli* MG1655 genomic DNA isolated from cell culture as an unamplified control, we calculated the amplified/unamplified ratio of hybridization intensities of 2,231 nonoverlapping 2-kb windows (covering 96.2% of the *E. coli* genome) for each plone. This 'ratio profile' represents genome-wide relative locus enrichment after amplification (**Fig. 2a**). By detecting the dips in the ratio-profile characteristic of the two engineered deletions present in NR57 (**Fig. 2a**), we demonstrated that ploning has high specificity to the target genome. However, a 6.5-fold of increase of variability (in $\log_2$ space) in the two plones compared with unamplified reference genomic DNA (**Fig. 2b**) suggested that ploning results in various degrees of local over- and underrepresentation. Because we observed a low correlation between ratio profiles of the two independent plones (Pearson $R^2 = 0.014$), most of the observed ploning bias appears sequence-nonspecific. To further explore this bias, we also characterized two *E. coli* plones amplified with the constrained-randomized primer D6 (D = A/G/T), a primer with endogenous background orders of magnitude lower than that of N6 (**Supplementary Table 1** online). D6, however, leads to much higher amplification bias (**Supplementary Fig. 4** online).

To investigate whether the dips in ratio profiles represent sequences that are completely missing after amplification, we performed real-time quantitative PCR targeting the three regions with the lowest ratios (**Supplementary Fig. 5** online). All three regions were present in the amplicon, albeit in lower copy numbers. Therefore, it might be possible to recover such underrepresented regions by sequencing at a greater depth, or by targeted PCR amplification before sequencing.

Whole-genome shotgun sequencing of *Prochlorococcus* plones

Prochlorococcus[29,30] is one of the most abundant bacterial lineages in the ocean. Although several ecotypes adapted to low and high light have been identified and sequenced[31,32], *Prochlorococcus* is highly heterogeneous in the open ocean[33]. The high level of 'microheterogeneity' leads to great difficulty in genome assemblies using environmental shotgun sequencing data[6,7]. Thus, this organism is another model system suitable for conducting proof-of-concept ploning experiments. As described above for *E. coli*, we mixed cells of three *Prochlorococcus* strains (MIT9312, MIT9313, MED4) in a 1:1:1 ratio. We stored the cells in 7.5% DMSO at –80 °C to mimic a typical environmental sampling procedure, and carried out ploning by amplifying genomic DNA at a dilution level of 0.5 cell/reaction. To ensure the presence of single cells in each well, we screened each plone with eight PCR primer sets specific for each strain. A second round of amplification was performed to generate enough DNA for shotgun library construction and sequencing.

Our initial efforts on sequencing plonal *Prochlorococcus* DNA failed owing to problems with library construction, including low cloning efficiency, abnormal insert size distribution and a high percentage of vector sequence among inserts. We reasoned that these problems

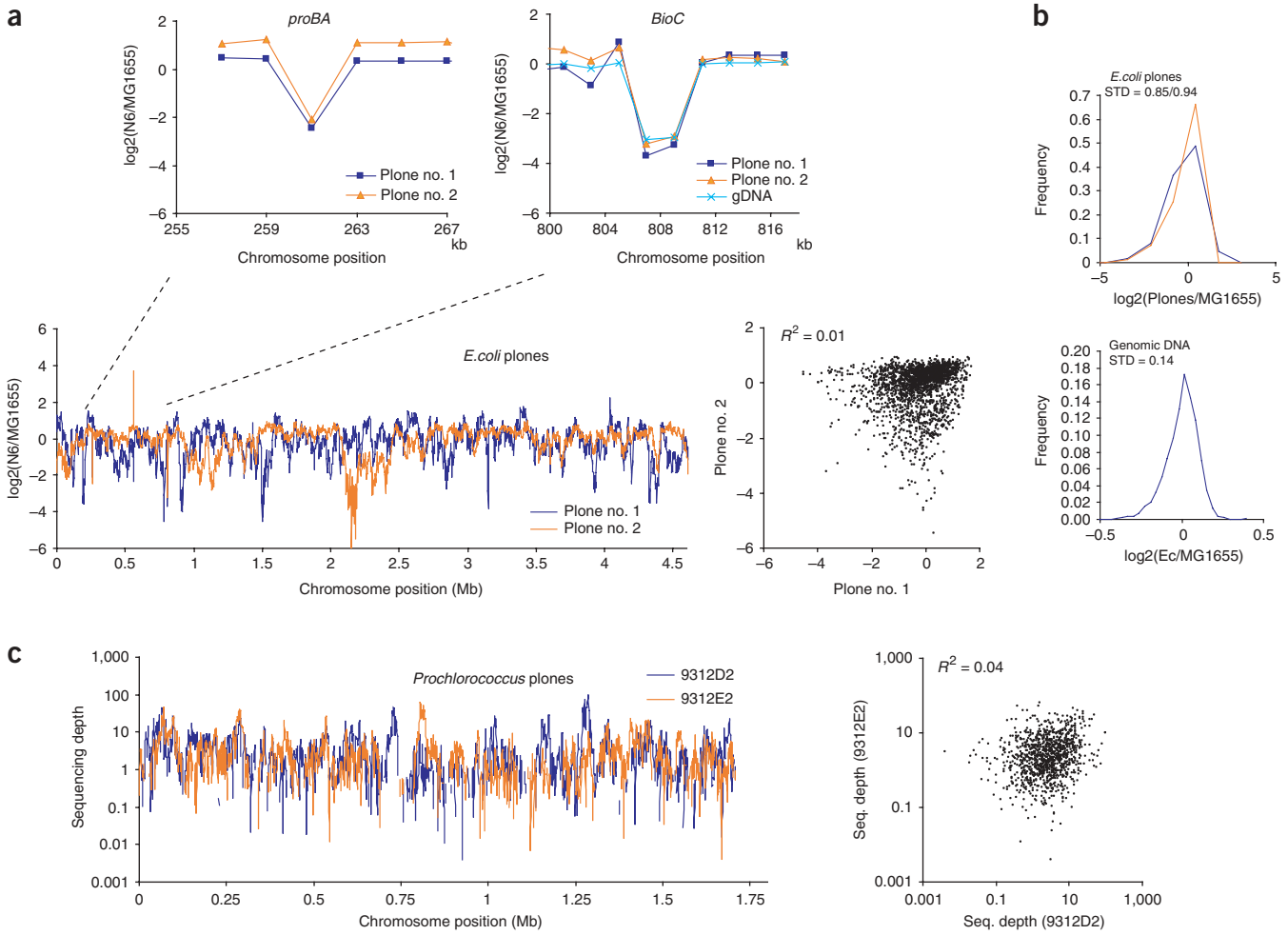

Figure 2 Characterization of plones. (**a**) Hybridization of two *E. coli* plones to Affymetrix *E. coli* genomic chips showed that the engineered deletions at the *bio* and *proBA* (NR57) loci were accurately preserved during amplification. Amplification was not even across the genome. The over- and underrepresented regions in the two plones do not overlap. In addition, there is little correlation between the ratio profiles of the two plones as shown in the right panel. (**b**) The Affymetrix chip hybridization ratios of the *E. coli* plones (upper panel) have a wider distribution compared with nonamplified genomic DNA control (lower panel), suggesting that the amplification is biased. (**c**) Distribution of sequencing depth of plones from two *Prochlorococcus* MIT9312 single cells across their genomes. The sequencing depth is calculated as the total length (in base pairs) of raw sequencing reads that mapped to a 1-kb window divided by the window size (1,000 bp).

arose because of hyperbranched structures generated during strand-displacement amplification. During library construction, such branched DNA could be ligated into the vector cloning sites, and the branches are somehow resolved by *E. coli* to form chimeras. To remove hyperbranched structures, we used S1 nuclease to cut the junctions of branched DNA molecules, and constructed a 3-kb sequencing library from an MIT9312 *Prochlorococcus* plone (9312E2) by using a one-step ligation protocol at the DOE Joint Genome Institute. Subsequently, we performed shotgun sequencing at a depth of 3.5×. We sampled 62.2% (including 63.5% of coding sequences and 44.6% intact genes) of the MIT9,312 genome at least once by 7,484 sequencing reads from the 9312E2 library (**Fig. 2c**). These raw sequences were assembled into 477 contigs, including 174 contigs >2 kb. In comparison, in previous efforts of sequencing the MIT9312 strain from a genomic DNA library, the same amount of sequencing reads were assembled into 311 contigs with 211 contigs >2kb.

Although the 9312E2 library represented an improvement over our initial library, it contained an unusually high percentage of chimeric sequences (19.3%; see **Table 1**) and therefore limited the quality of genome assemblies. An improved assembly with longer contigs was obtained by computationally splitting these chimeric sequences at their junction points based on the MIT9312 reference genome. In an effort to improve assembly, we implemented an iterative assembly strategy by taking advantage of the fact that 85.1% of chimeric junctions could be mapped to genomic regions covered by at least two nonchimeric reads. We generated an assembly of higher quality by successfully identifying 698/1,481 chimeric reads (47.1%) without the reference genome. The longest contig was improved from 35.4 kb to 58.3 kb, and the percentage of misassembled contigs dropped from 20% to 13% (**Supplementary Fig. 6** online).

Almost half of the chimeric artifacts can be computationally removed by our iterative assembling procedure. However, a high chimeric rate can compromise the accuracy of pair-end information and therefore undesirably limit the ploning method to simple genomes. As we did not detect chimeras in the plonal DNA by PCR, we hypothesized that they were introduced after ploning

Table 1 Chimeric rates of sequencing libraries from *Prochlorococcus* MIT9312 plones constructed with different post-amplification treatments

Library	phi29 debranching	S1 nuclease	Mung bean nuclease	T4 endo-nuclease VII	DNA pol I	Chimeric rate
A	–	+	–	–	–	19.28%
B	–	+	–	–	–	18.68%
C	–	+	–	–	–	17.00%
D	–	+	–	–	+	15.63%
E	+	–	–	–	+	12.50%
F	+	–	+	–	+	51.28%
G	–	–	–	+	+	23.40%
H	–	–	–	–	+	31.82%
I	+	+	–	–	+	6.25%
J	+	+	–	–	+	8.33%

Libraries A–C were constructed from the same S1-digested plonal DNA (9312E2) but three different methods were used. A was made by JGI with a one-step ligation protocol; B was made by Agencourt with a two-step ligation protocol; C was made in-house with the Invitrogen TOPO cloning system. The Invitrogen TOPO blunt-end cloning protocol was used to make Libraries D–I from the same plonal DNA (9312D2) with different treatments. I and J are two independent libraries generated using the same protocol.

(**Supplementary Fig. 7** online). Reasoning that the chimeric artifacts were derived from our library construction procedure, we tested another library construction strategy using two-step oligo-based ligation developed in Agencourt Bioscience. This technique also resulted in low cloning efficiency (~20-fold lower than regular genomic DNA) and a high chimeric rate (18.7%). Furthermore, out of a total of 5,314 pair-end reads obtained from 2,657 clones, only 6.5% sequencing reads (465 kb in total length) could be mapped to the reference sequence, as the majority of reads represented vector sequence. The chimeric rate of an in-house library made by the TOPO cloning method was not significantly better (17%; see **Table 1**).

We hypothesized that even after S1 nuclease treatment, plonal DNA retained enough noncanonical structure to interfere with cloning. We sought to address this by testing other postamplification enzymatic treatments (**Table 1**). We observed different chimeric rates with different treatments on the same plonal DNA, confirming that chimeras were generated during library construction. We achieved a chimeric rate as low as 6.25% with the combination of three treatments: phi29 polymerase debranching, S1 nuclease digestion and DNA polymerase I nick translation. Skipping any of these treatments resulted in higher chimeric rates, suggesting a three-step model of linearizing hyperbranched DNA (**Fig. 3**).

A sequencing library for a second plone (9312D2) was constructed using our three-step, enzymatic treatment, library-construction protocol, and was sequenced to a depth of 4.7×. Approximately 66.0% of the genome was recovered with a total of 7.2 Mb of high-quality reads; the largest gap was 17 kb. Because of the biases introduced by amplification, ploning requires more total sequencing to achieve the same level of coverage as that obtained from sequencing unamplified genomic DNA (**Supplementary Fig. 8** online). At a comparable level of sequencing depth, ~96.4% of the genome can be recovered from unamplified MIT9312 genomic DNA. Owing to the biases inherent to single molecule amplification, some genomic regions are repeatedly sampled whereas others are barely covered once (**Fig. 2c** and **Supplementary Fig. 9** online). By fitting the coverage curves in **Supplementary Fig. 8** online, we estimate that it would take ~26 Mbp (~15×) of sequencing reads to sample 90% of the *Prochlorococcus* genome. Alternatively, the unsampled regions can be amplified by PCR on the plones and sequenced. To illustrate the feasibility of this approach, we designed primers to target the center regions of the largest gaps in the two *Prochlorococcus* plones. All the target regions were successfully recovered (**Supplementary Fig. 10** online), consistent with the real-time PCR results from the *E. coli* plones.

We observed ten sequencing reads (0.09%) in the 9312D2 library and 12 (0.16%) in the 9312E2 library that do not share homology with any known sequences in the NCBI nonredundant database. These were probably artificial sequences generated by N6 primer-primer interaction (endogenous background amplification). There were also two sequencing reads (0.02%) in the 9312D2 library and 74 reads (1%) in the 9312E2 library that mapped to the MED4 genome. These sequences were probably amplified from DNA from lysed MED4 cells in the initial mixed cell population. No other known sequence (except for the target genome, MIT9312) was detected in the libraries, confirming that our ploning method has extremely low background.

To identify potential mutations that could have arisen during single-cell amplification, the largest contig of the 9312E2 plone (59,652 bp) was compared with the reference genome; 81 mismatches were identified. We visually inspected the assembly at each of the mismatched positions and found that all of the mismatches were caused by assembly errors or discrepancies among raw reads, especially in homopolymeric regions (that is, some reads have five Ts in a row and others have four). We did not find a single mismatch that was free from sequencing errors. Therefore, the estimated amplification error is $< 1.7 \times 10^{-5}$, well below the Bermuda standard for genome sequencing (10^{-4}).

Ploning of 'wild' *Prochlorococcus* cells

Having established the ploning method using lab strains of *E. coli* K-12 and *Prochlorococcus* MIT9312, we next applied this method to a Pacific Ocean sample collected at a depth of 85 m in October 2003 at the Station ALOHA (2′45′N, 158′00′W) under the Hawaii Ocean Time-series project. We have successfully ploned two single *Prochlorococcus* cells. Sequencing of the 16S rDNA and the *Prochlorococcus* specific internal transcribed spacers of the ribosomal operon indicates that both cells are closely related to the MIT9312 strain. We have

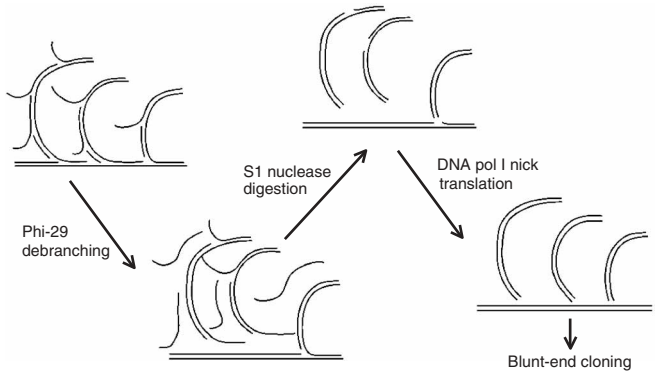

Figure 3 Resolving hyperbranched DNA structure for sequencing library construction. In the first step, hyperbranched DNA is incubated with phi29 DNA polymerase and dNTPs, but without any primer. Because of the strand-replacement activity of the phi29 polymerase, the density of branching junctions is reduced. This step also gives rise to some 3′ single-stranded overhangs. Junctions are broken by the S1 nuclease digestion at the second step and 3′ single-stranded overhangs are also removed. The resulting DNA molecules are double-stranded with some nicks. After shearing and gel-size selection, these nicks are removed by nick translation using DNA polymerase I, which has not only polymerase activity, but also 5′- and 3′-exonuclease activity.

performed trial sequencing of two in-house shotgun libraries, and found a similar level of bias in genome coverage as seen in the 9312D2 and 9312E2 plones (**Supplementary Fig. 11** online). The average sequence identity to the MIT9312 genome is 91.7% for both plones, and these two plones also differ between themselves. Full genome shotgun sequencing of these two plones is currently being undertaken and will elucidate the level of genome diversity between lab strains and cells in the wild.

DISCUSSION

In summary, we have developed a method for genome sequencing from single cells using a real-time isothermal amplification system to guarantee target-specific amplification from one single cell and a three-step enzymatic treatment method to resolve hyperbranched DNA structures before shotgun library construction. Detailed characterization of *E. coli* and *Prochlorococcus* plones with Affymetrix chip hybridization and whole-genome shotgun sequencing, respectively, clearly demonstrate that genome sequencing from single cells is feasible.

Our current method suffers from two clear limitations. First, the resulting genome sequence coverage is noticeably biased. Two non-mutually exclusive mechanisms probably explain the observed bias in plonal amplification. One is the initial stochastic priming on a single template molecule. This is supported by the observation that constrained-randomized primer D6 led to a much higher bias than the completely degenerate primer N6, which can perfectly anneal to $\sim 4^6/2^6 = 64$ times as many locations as D6. Another plausible explanation is chromosomal breakage resulting in underrepresentation at breakpoint ends[34]. Although the degree of chromosomal breakage is hard to characterize at the single-molecule level, our experimental protocol was optimized (for example, releasing DNA from the cells with lysozyme, denaturing DNA with alkaline solution instead of heat, gentle pipetting) to reduce the chance of breaking DNA. Owing to the bias in plonal amplification, deeper sequencing is required for the ploning method relative to traditional sequencing. Given the recent breakthroughs of DNA sequencing technologies that are drastically reducing sequencing costs, we feel that this necessary extra sequencing is a minor limitation of our method[35,36]. Because of the random distribution in amplification bias (**Fig. 2a,c**), combining the sequencing reads of the two *Prochlorococcus* plones improves coverage to 84%, which is more efficient than sequencing just one cell. Yet, although the chance of catching two identical cells might exist in some simple ecosystem, such as the acidic Iron Mountain mine drainage[5], pooling identical plones for genome assembly could require impractically extensive prescreening of millions of nearby cells in environmental samples (such as soil, seawater or microbes inhabiting plants and animals) because of the great diversity of genotypes[37,38]. It is worth noting that a single cell can contain more than one chromosomal copy or a partially duplicated genome during cell division. In such cases, the amplification bias will be lower in regions of more than one copy. However, it is unclear how feasible it is to capture dividing cells, especially for organisms with long doubling times. The most efficient strategy is probably to perform $\sim 15\times$ shotgun sequencing on plonal DNA to obtain $\sim 90\%$ of the genome, then use PCR sequencing to close the gaps. PCR requires sequencing information for primer design, which can usually be obtained from the huge amount of data generated by metagenomic sequencing efforts, or ploning of closely related cells from the same sample.

Second, although we have successfully reduced the chimeric rate to 6.25%, such a rate is still higher than that typically seen in sequencing libraries ($<1\%$). Assembling sequencing reads from a biased library with $\sim 6\%$ of chimeras merits further development of genome assembly algorithms. Although chimerism can be addressed moderately well using an iterative genome assembly algorithm, we are actively investigating enzymatic processing methods to minimize it further. As S1 nuclease can digest double-stranded DNA at the 'bubbles' of the double helix, which transiently form because of thermodynamic fluctuation, we did not perform complete digestion to avoid loss of DNA. This is probably why we have not been able to completely eliminate chimeras, and where further improvement could be made.

Despite these two limitations, ploning opens a window to genomic information not evident with current metagenomic or population-based methods. It represents an important step in charting the largely unmapped genomic biosphere. Furthermore, ploning has generality and impact beyond microbial genome sampling, for example, in sequencing isolated human chromosomes and microdissected chromosomal fragments.

METHODS

Ultra-low background real-time isothermal amplification. We developed a strict sample handling and experimental procedure, which we found was essential to achieve sub-femtogram levels of background. All experiments were conducted in an AirClean 1000 PCR hood (AirClean System) with a dedicated set of pipettes. Unopened pipette tips were used for every experiment. Tubes, tube caps and all reagents, except for the primers, dNTPs, SYBR Green I and polymerases, were treated with UV for 5 ~ 10 min in a Stratalinker (Stratagene, model no. 1800). Primers and SYBR Green I were diluted with UV-treated RT-PCR grade water (Ambion). Isothermal amplifications that contain various amounts of templates, 8 U/μl RepliPHI phi29 DNA polymerase (Epicentre), 1 mM dNTP, 1 mM N6 primer with two 3′ phosphothioate bonds[39], 0.1× SYBR Green I (Molecular Probes) and 1× RepliPHI reaction buffer were performed in volumes of 20 μl or 50 μl in a real-time PCR thermocycler (Opticon 2, MJ Research) at 30 °C for 10 h. Fluorescent intensities were collected via the SYBR Green I channel every 6 or 15 min. Random primers were purchased from IDT. When necessary, we UV-treated the phi29 DNA polymerase in inverted strip-tube caps placed on top of a chilled 96-well PCR cooler (Eppendorf) filled with water to avoid sample heating. Real-time isothermal amplification data were exported by the Opticon2 program, and analyzed using a Perl script.

Polymerase cloning on *E. coli* and *Prochlorococcus*. Six *E. coli* strains, MG1655, NR1, NR56, NR57, NR58 and NR59 were cultured in LB medium in a 30 °C shaker overnight. NR1 is MG1655 with a defective λ-phage in place of the *bioA* and *bioF* genes and was used to construct the four strains by recombineering[40]. Each of these strains has a *cat* marker replacing a particular gene operon, that is, NR56 is NR1 Δ*glyA::cat*, NR57 is NR1 Δ*proBA::cat*, NR58 is NR1 Δ*thrBC::cat* and NR59 is NR1 Δ*trpLEDBCA::cat*. Genomic DNA from the MG1655 and EcNR1 strains were extracted with the Genomic-tip 20/G (Qiagen). To prepare single-cell dilutions, cells of strains NR56, NR57, NR58 and NR59 were washed twice in UV-treated PBS. Cell densities were determined by direct counting using a hemocytometer. Cells were then mixed in equal ratio, and diluted to the single-cell level. Cell density was reconfirmed by performing 36 4-plex single-cell PCR reactions on single-cell dilutions using the strain-specific primers, and checking PCR products by electrophoresis. Single-cell dilutions were treated with 10 units of lysozyme at room temperature (25 °C) for 10 min before amplification, and denatured by alkaline solution as described[12]. After real-time isothermal amplification at 30 °C for 10 h, we performed single-plex PCR with the same primer set on the amplicons to identify those that were amplified from single cells. We used the HotStar PCR MasterMix (Qiagen) for PCR amplification with 1 μl of 1:100 diluted amplicons (or 1 μl of a single-cell dilution) and a final primer concentration of 0.2 μM. The thermocycling protocol is: 95 °C, 15 min, followed by 35 cycles of 94 °C, 30 s; 64 °C, 30 s; and 72 °C, 30 s, and a final step of 3 min at 72 °C. PCR products were checked by gel electrophoresis. To prepare sufficient DNA from single-cell amplicons for both Affymetrix chip hybridizations and library construction, we performed a second round of amplification on the amplicons using the standard multiple displacement amplification protocol[12]. Real-time

quantitative PCR assays confirmed that additional locus-specific biases introduced in this step is negligible compared with the first-round amplification (data not shown).

To prepare *Prochlorococcus* plones, cells of three lab strains (MIT9312, MIT9313, MED4) were mixed in 1:1:1 ratio and stored in 7.5% DMSO at –80 °C. Cell density was determined with flow cytometry. Amplification was performed at the dilution level of 0.5 cell/aliquot. Amplicons were 1:100 diluted and screened for positive *Prochlorococcus* plones by performing PCR with primers targeting the internal transcribed spacers of the ribosomal operon (ITS, 2F: GAAGTCGTTACTCCAACCC; 3R: TCATCGCCTCTGTGTGCC).

Affymetrix *E. coli* chip hybridization and analysis. We purified single-cell amplicons with Microcon YM30 columns (Millipore), and performed fragmentation with DNase I (Amersham), and labeling with the BioArray terminal labeling kit (Affymetrix). Unamplified genomic DNA from EcNR1 and MG1655 were hybridized in triplicate and very low interexperiment variation was observed. Therefore, we did one hybridization experiment for each of the four amplicons. Hybridization and scanning were performed with ~2 μg of labeled DNA by the Biopolymer Core Facility (Harvard Medical School). Data analyses were primarily conducted with the Bioconductor affy package with a customized probe set package, in which probes were grouped into nonoverlapping 2-kb bins along the chromosome. To reduce potential cross-hybridization signals, we performed BLAST searches of all probes on the Affymetrix *E. coli* Antisense Genome chip against the *E. coli* K12 genome sequence (GenBank accession number NC_000913), and excluded those having more than one match of >75% identity. As a result, our analyses were based on a total of 133,203 pairs of perfect matched-mismatched probes. The oligonucleotide probes on the Affymetrix chip are not evenly spaced across the genome, so that the 2-kb bins do not have equal numbers of probes. Because too few probes may lead to probe-specific bias, we excluded bins having less than ten probes. Additional probe sets (each contains ten pairs of probes) representing the four strain-specific deletion regions were included, because a bin size of 2 kb is too large compared with the size of these deletions. The average normalized intensity of the three MG1655 replicates at each 2-kb bin was used (as denominators) to calculate the ratios for the other experiment to cancel the hybridization biases at the probe set level. The Bioconductor Affy package provides several different methods for background correction, normalization and probe set summary. We compared the performance of all methods based on the results at the *bio* locus, and found that the MAS5 method was most appropriate for this study. Therefore, all analyses were based on the results generated by MAS5.

Shotgun-sequencing library construction. To prepare a sufficient amount of DNA from the 9312D2 plone for library construction, a second round of amplification was performed on >1 μg of plone DNA with the regular MDA protocol. The amplicon was purified using a Microcon YM-100 column, then incubated with 8 U/μl RepliPHI phi29 DNA polymerase, 1 mM dNTP and 1× RepliPHI reaction buffer in 50 μl at 30 °C for 2 h, 65 °C for 3 min, then digested with 1 U/μl S1 nuclease (USB) in 200 μl 1× buffer (30 mM sodium acetate, pH 4.5, 50 mM NaCl, 1 mM $ZnCl_2$) at 37 °C for 30 min. Debranched DNA was extracted with phenol/chloroform, and sheared with a homemade shearing device (equivalent to Genomic Solutions' HydroShear) at speed code 13. Sheared DNA was concentrated with Microcon YM-100 column (Millipore), size-selected with agarose gel electrophoresis and purified with Qiaquick gel extraction kit (Qiagen). DNA (0.1–1 μg) was polished with 3 U of T4 DNA polymerase (New England Biolabs (NEB)) and 10 U of DNA polymerase I (Invitrogen) in 50 μl of 1× NEB buffer 2 and 0.5 mM dNTP at room temperature (25 °C) for 1 h, inactivated at 75 °C for 10 min and dephosphorylated by adding 50 U of calf intestinal phosphatase (NEB), 10 μl of NEB buffer 3 (10×) and 35 μl dH_2O and incubating at 37 °C for 1 h. After extraction with phenol/chloroform and purification with ethanol precipitation, 4 μl of DNA (30~100 ng) was incubated with 1 μl pCR4Blunt TOPO vector (Invitrogen) and 1 μl salt solution (Invitrogen) at room temperature (25 °C) for 15 min. The ligation product was purified with ethanol precipitation, resuspended in 3 μl dH_2O, and transfected to 50 μl of TOP10 ElectroComp cells (Invitrogen) by electroporation at 20 kV. The transformation was incubated in 500 μl SOC medium in a 37 °C shaker at 250 r.p.m. for 1 h, and stored at –80 °C with 20% glycerol before plating. Whole-genome shotgun sequencing of the 9312D2 plone was conducted at Agencourt Biosciences. For the 9312E2 plone, DNA was only digested with S1 nuclease after amplification, and the library was constructed and sequenced at the US Department of Energy Joint Genome Institute (JGI) with JGI's standard protocol. Small-scale sequencing was conducted at the Harvard Medical School Biopolymers Facility or Genaissance Pharmaceuticals. Sequence analyses/genome assembling was performed in-house with phred/phrap/consed or at the DOE JGI.

Iterative genome assembling. To improve genome assembly in the presence of chimeric sequences, we performed multiple rounds of genome assembling and chimeric sequence detection: (i) assembled raw reads into contigs with phrap; (ii) assuming all contigs were nonchimeric, compared all raw reads with the contigs, detected chimeric sequences and broke them at each chimeric junction; (iii) fed the resulting sequences to Phrap for the next round of assembly. This iterative assembling procedure was repeated until the chimeric rate stopped improving. This algorithm was implemented with Perl (IterativeAssembler, K. Zhang, Harvard Medical School).

URLs. Bioconductor web site: http://www.bioconductor.org/. The IterativeAssembler is available at http://arep.med.harvard.edu/kzhang/Ploning/IterativeAssembler.zip. The DOE-JGI sequencing library construction protocol is available at http://www.jgi.doe.gov/sequencing/protocols/prots_production.html. The raw DNA sequences are available at http://arep.med.harvard.edu/kzhang/Ploning/Plone_raw_seqs.tar.gz.

Note: Supplementary information is available on the Nature Biotechnology website.

ACKNOWLEDGMENTS

We thank J. Shendure, M. Umbarger and M. Wright for critical comments on the manuscript; C. Detter (DOE-JGI) for technical assistance on sequencing library construction. We would also like to thank the US Department of Energy for Genomes-to-Life Center support (G.M.C. and S.W.C.), and the National Science Foundation and the Moore Foundation for additional support (S.W.C.).

AUTHOR CONTRIBUTIONS

A.C.M. contributed to *Prochlorococcus* single-cell amplification, data analyses and manuscript writing, N.B.R. contributed to *E. coli* single-cell amplification, Affymetrix gene-chip analyses and writing, K.W.B. contributed to genome assembly, J.M. contributed to the development of sequencing library construction protocol, S.W.C. contributed to the design of the project and writing. K.Z. contributed to the development of the real-time amplification method, polymerase cloning, characterization of plones by Affymetrix gene-chip analyses and genome sequencing, development of the library construction protocol, data analyses and manuscript writing. S.W.C. and G.M.C. contributed to the planning and design of the project, and manuscript writing.

COMPETING INTERESTS STATEMENT

The authors declare that they have no competing financial interests.

Published online at http://www.nature.com/naturebiotechnology/
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions/

1. Shendure, J., Mitra, R.D., Varma, C. & Church, G.M. Advanced sequencing technologies: methods and goals. *Nat. Rev. Genet.* **5**, 335–344 (2004).
2. Kyrpides, N.C. Genomes OnLine Database (GOLD 1.0): a monitor of complete and ongoing genome projects world-wide. *Bioinformatics* **15**, 773–774 (1999).
3. Moreira, D. & Lopez-Garcia, P. The molecular ecology of microbial eukaryotes unveils a hidden world. *Trends Microbiol.* **10**, 31–38 (2002).
4. Falkowski, P.G. & de Vargas, C. Genomics and evolution. Shotgun sequencing in the sea: a blast from the past? *Science* **304**, 58–60 (2004).
5. Tyson, G.W. *et al.* Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature* **428**, 37–43 (2004).
6. Venter, J.C. *et al.* Environmental genome shotgun sequencing of the Sargasso Sea. *Science* **304**, 66–74 (2004).
7. DeLong, E.F. Microbial community genomics in the ocean. *Nat. Rev. Microbiol.* **3**, 459–469 (2005).
8. Beja, O. *et al.* Comparative genomic analysis of archaeal genotypic variants in a single population and in two different oceanic provinces. *Appl. Environ. Microbiol.* **68**, 335–345 (2002).
9. Tringe, S.G. *et al.* Comparative metagenomics of microbial communities. *Science* **308**, 554–557 (2005).

10. Riesenfeld, C.S., Schloss, P.D. & Handelsman, J. Metagenomics: genomic analysis of microbial communities. *Annu. Rev. Genet.* **38**, 525–552 (2004).
11. Rodriguez-Valera, F. Environmental genomics, the big picture? *FEMS Microbiol. Lett.* **231**, 153–158 (2004).
12. Dean, F.B. *et al.* Comprehensive human genome amplification using multiple displacement amplification. *Proc. Natl. Acad. Sci. USA* **99**, 5261–5266 (2002).
13. Hosono, S. *et al.* Unbiased whole-genome amplification directly from clinical samples. *Genome Res.* **13**, 954–964 (2003).
14. Paez, J.G. *et al.* Genome coverage and sequence fidelity of phi29 polymerase-based multiple strand displacement whole genome amplification. *Nucleic Acids Res.* **32**, e71 (2004).
15. Telenius, H. *et al.* Degenerate oligonucleotide-primed PCR: general amplification of target DNA by a single degenerate primer. *Genomics* **13**, 718–725 (1992).
16. Zhang, L. *et al.* Whole genome amplification from a single cell: implications for genetic analysis. *Proc. Natl. Acad. Sci. USA* **89**, 5847–5851 (1992).
17. Dietmaier, W. *et al.* Multiple mutation analyses in single tumor cells with improved whole genome amplification. *Am. J. Pathol.* **154**, 83–95 (1999).
18. Nelson, J.R. *et al.* TempliPhi, phi29 DNA polymerase based rolling circle amplification of templates for DNA sequencing. *Biotechniques* Suppl., 44–47 (2002).
19. Lage, J.M. *et al.* Whole genome analysis of genetic alterations in small DNA samples using hyperbranched strand displacement amplification and array-CGH. *Genome Res.* **13**, 294–307 (2003).
20. Handyside, A.H. *et al.* Isothermal whole genome amplification from single and small numbers of cells: a new era for preimplantation genetic diagnosis of inherited disease. *Mol. Hum. Reprod.* **10**, 767–772 (2004).
21. Hellani, A. *et al.* Multiple displacement amplification on single cell and possible PGD applications. *Mol. Hum. Reprod.* **10**, 847–852 (2004).
22. Sorensen, K.J., Turteltaub, K., Vrankovich, G., Williams, J. & Christian, A.T. Whole-genome amplification of DNA from residual cells left by incidental contact. *Anal. Biochem.* **324**, 312–314 (2004).
23. Jiang, Z., Zhang, X., Deka, R. & Jin, L. Genome amplification of single sperm using multiple displacement amplification. *Nucleic Acids Res.* **33**, e91 (2005).
24. Detter, J.C. *et al.* Isothermal strand-displacement amplification applications for high-throughput genomics. *Genomics* **80**, 691–698 (2002).
25. Raghunathan, A. *et al.* Genomic DNA amplification from a single bacterium. *Appl. Environ. Microbiol.* **71**, 3342–3347 (2005).
26. Hutchison, C.A., III, Smith, H.O., Pfannkoch, C. & Venter, J.C. Cell-free cloning using {phi}29 DNA polymerase. *Proc. Natl. Acad. Sci. USA* **102**, 17332–17336 (2005).
27. Hafner, G.J., Yang, I.C., Wolter, L.C., Stafford, M.R. & Giffard, P.M. Isothermal amplification and multimerization of DNA by Bst DNA polymerase. *Biotechniques* **30**, 852–856, 858, 860 passim (2001).
28. Gray, J.W. *et al.* High-speed chromosome sorting. *Science* **238**, 323–329 (1987).
29. Chisholm, S.W., Olson, R.J., Zettler, E.R. & Goericke, R. A novel free-living prochlorophyte abundant in the oceanic euphotic zone. *Nature* **334**, 340–343 (1988).
30. Partensky, F., Hess, W.R. & Vaulot, D. Prochlorococcus, a marine photosynthetic prokaryote of global significance. *Microbiol. Mol. Biol. Rev.* **63**, 106–127 (1999).
31. Rocap, G. *et al.* Genome divergence in two Prochlorococcus ecotypes reflects oceanic niche differentiation. *Nature* **424**, 1042–1047 (2003).
32. Dufresne, A. *et al.* Genome sequence of the cyanobacterium Prochlorococcus marinus SS120, a nearly minimal oxyphototrophic genome. *Proc. Natl. Acad. Sci. USA* **100**, 10020–10025 (2003).
33. Urbach, E. & Chisholm, S.W. Genetic diversity in Prochlorococcus Populations flow-cytometrically sorted from the Sargasso Sea and Gulf Stream. *Limnol. Oceanogr.* **43**, 1615–1630 (1998).
34. Panelli, S., Damiani, G., Espen, L. & Sgaramella, V. Ligation overcomes terminal underrepresentation in multiple displacement amplification of linear DNA. *Biotechniques* **39** 174, 176, 178 passim (2005).
35. Shendure, J. *et al.* Accurate multiplex polony sequencing of an evolved bacterial genome. *Science* **309**, 1728–1732 (2005).
36. Margulies, M. *et al.* Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**, 376–380 (2005).
37. Thompson, J.R. *et al.* Genotypic diversity within a natural coastal bacterioplankton population. *Science* **307**, 1311–1313 (2005).
38. Acinas, S.G. *et al.* Fine-scale phylogenetic architecture of a complex bacterial community. *Nature* **430**, 551–554 (2004).
39. Dean, F.B., Nelson, J.R., Giesler, T.L. & Lasken, R.S. Rapid amplification of plasmid and phage DNA using Phi 29 DNA polymerase and multiply-primed rolling circle amplification. *Genome Res.* **11**, 1095–1099 (2001).
40. Yu, D. *et al.* An efficient recombination system for chromosome engineering in *Escherichia coli*. *Proc. Natl. Acad. Sci. USA* **97**, 5978–5983 (2000).