

High-Throughput Sequencing and Metagenomics

William J. Jones

Received: 16 January 2009 / Revised: 18 May 2009 / Accepted: 20 May 2009
© Coastal and Estuarine Research Federation 2009

Abstract The fields of biology and environmental health are undergoing an unprecedented transformation in the way communities are characterized and studied. One of the catalysts for this change has been the advent of high-throughput sequencing (aka next-generation sequencing). Though high-throughput sequencing is a relatively new concept, the principles behind it are fairly straightforward (PCR and pyrosequencing) and have been in the published literature since the early 1990s. Herein, I will highlight some of the more commonly encountered issues involving experimental design as well as important considerations for metagenomic sample collection (both quantity and quality) and data handling for high-throughput sequencing using the Roche/454 Life Sciences genome sequencing platform.

Keywords Pyrosequencing · Emulsion PCR · Metagenomic · High-throughput sequencing

Introduction

Now is an exciting time to be a marine scientist involved with estuarine-coastal systems. Whether the focus be viruses in a water sample, diversity of methanotrophs in a sediment sample, or characterizing a de novo genome of an organism, high-throughput sequencing (HTS) has fundamentally changed the way questions in estuarine-coastal science are approached. One of the most common application of HTS is metagenomics. Metagenomics is “the application of modern

genomic techniques to the study of communities of microbial organisms directly in their natural environments, bypassing the need for isolation and lab cultivation of individual species” (Chen and Pachter 2005). Metagenomics has existed since the late 1990s (Handelsman et al. 1998) and has exploded as a field since the availability of long read, high-throughput sequencing offered by the Roche/454 Life Sciences Genome Sequencer (Margulies et al. 2005). Beginning in 2006, a variety of papers surveying *16S rRNA* microbial diversity (Sogin et al. 2006), marine viromes (Angly et al. 2006), and environmental genetics of microbes (Edwards et al. 2006) began to appear. With the advent of a true random sampling of genomes (i.e., “shotgun-approach” to sequencing) and new protocols (Frias-Lopez et al. 2008; Pernthaler et al. 2008), estuarine science and the study of coastal environments is experiencing a renaissance via genomics. By definition, metagenomics is the study of the genetic material from an environmental sample. While much of metagenomics focuses on bacteria (especially the *16S rRNA* gene), the field is expanding rapidly to encompass the entire spectrum of organisms in an environmental sample, including bacteria, archaea, viruses, small eukaryotes, plasmids, and short RNAs.

HTS provides researchers and scientists with a revolutionary tool for quickly and affordably obtaining DNA sequence. The cost per unit of DNA sequence (i.e., base pair) obtained with HTS compared to traditional Sanger sequencing is dramatically more cost and time effective. The initial human genome was sequenced by an army of labs filled with ABI3730 Sanger sequencing instruments. This approach took 13 years and \$3 billion US to go from sample collection to publication of the completed human genome (Venter et al. 2001). Today, a human genome can be sequenced de novo in 2 months for \$1 million US (Wheeler et al. 2008). To put this in broader terms, a

W. J. Jones (✉)
Environmental Genomics Core Facility,
University of South Carolina,
Columbia, SC 29208, USA
e-mail: engencor@mailbox.sc.edu

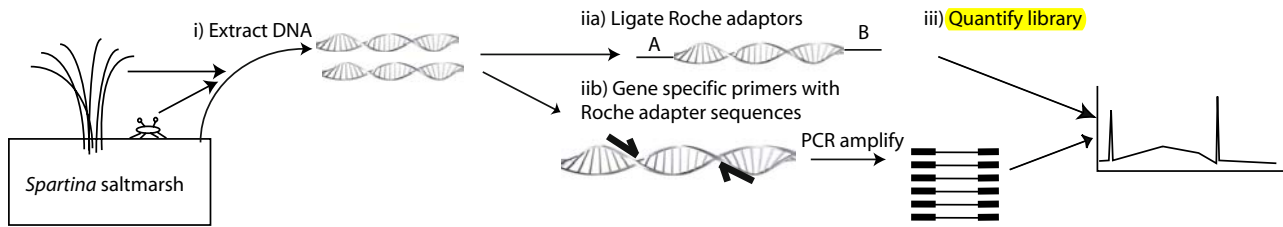
Roche/454 Life Science genome sequencer can produce the same sequence data at 10% of the cost and in 0.9% of the time that a single ABI3730 48 capillary can.

As of December 2008, there are three commercially available HTS platforms: the Applied Biosystems SOLiD, Illumina/Solexa 1G, and the Roche/454 Life Sciences Genome Sequencer. While the SOLiD and 1G systems currently use short reads (50 and 75 bp, respectively), the Roche/454 Life Sciences platform with the FLX chemistry produces average reads of 250 bp. As metagenomics is a random sampling of genetic information in a discrete sample (usually DNA extracted from a water, soil, or other environmental sample), short reads (i.e., less than 100 bp) make it practically impossible to be able to glean enough information to characterize DNA segments to either functional groups (i.e., Clusters of Orthologous Groups of proteins) or species (Krause et al. 2008; though see Wommack et al. 2008). A recent review by Mardis (2008)

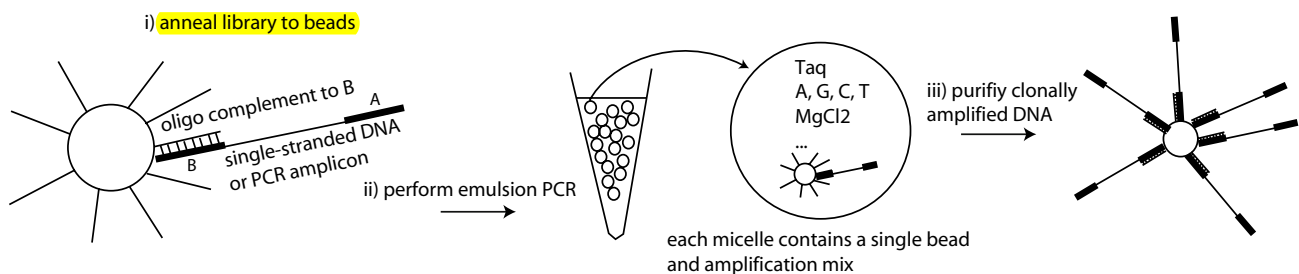
provides a point-by-point comparison of the applications as well as the pros/cons for each of these three platforms. In this paper, I will focus on the Roche/454 Life Science Genome Sequencing platform, as it is the most widely used HTS platform for metagenomics due to the relatively long read lengths.

The Roche/454 Life Sciences HTS process begins with 3–5 μg of any double-stranded starting material whether it be from a soil, water, plant, animal, or environmental sample (Fig. 1a, i). While high-molecular weight DNA is preferred (1.8 OD, 1.5 kb pieces or larger), fragmented DNA from museum collections or mummified samples is acceptable and has been used to sequence ancient genomes (Miller et al. 2008). Following an accurate quantification of the DNA concentration by a fluorescence method (i.e., PicoGreen), the DNA is fractionated into a size-range conducive for emulsion PCR (emPCR) and subsequent pyrosequencing.

A) DNA extraction and library preparation



B) Emulsion PCR (emPCR)



C) Performing the run

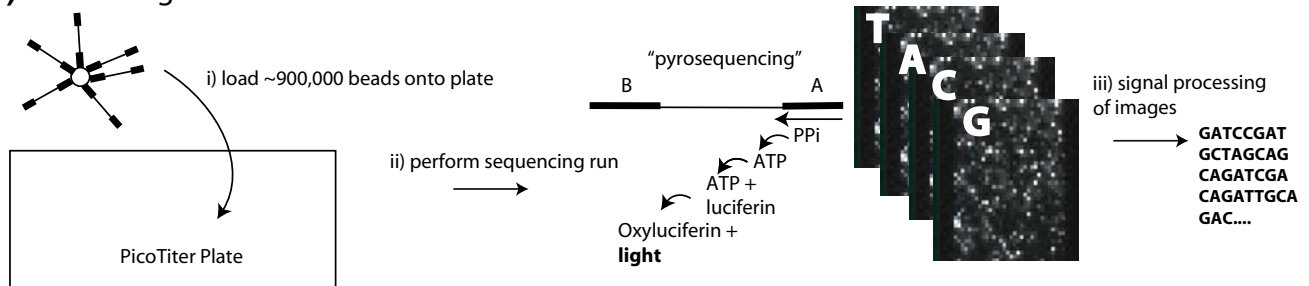


Fig. 1 Schematic portrayal of the Roche/454 Life Science work flow. DNA is extracted from environment/organism of interest and converted into a library (a), which is then used as a template for the emulsion PCR (b), which clonally amplifies the DNA. Clonally

amplified DNA containing beads are then loaded on a PicoTiter plate (c), which is processed for the sequencing run. Raw data images from the pyrosequencing reactions are finally converted via on-instrument software to nucleotide sequence data

The power behind pyrosequencing lies in the combination of emulsion PCR and pyrosequencing. The ability to perform a very large number of independent, clonally amplified (i.e., PCR reactions) beads in a single tube using a mixture of oil and water to create emulsions provides a way of physically separating PCR reactions on a very small spatial scale. Emulsion PCR (emPCR) is a modification of standard PCR where micelles are created wherein each micelle contains a single oligo-coated bead, template DNA, and PCR reagents/enzymes. For the emPCR amplification to be effective, DNA fragments must be no larger than 500–600 bp (Roche protocol and Jones per obs). After preparation of the sample following established protocols (below), clonally amplified beads are sequenced using pyrosequencing, which involves the production of light via a luciferase-catalyzed reaction when a complementary nucleotide is incorporated during the sequencing by synthesis. Though fundamentally different from traditional dye-labeled sequencing approaches, the data and handling are the same for pyrosequencing.

Typically, DNA for Roche/454 pyrosequencing is sheared (=nebulized) using compressed nitrogen gas that forces the sample through a small opening in a nebulizer and randomly shears the DNA. Next, the nebulized DNA is checked on a Bioanalyzer DNAChip to ensure the size-fractionated DNA is suitable (ideally between 300 and 800 bp with a median of 500 bp). If the nebulized sample is acceptable, the ends of the DNA fragments are blunt-end-polished with T4 PNK and Roche library adaptors A and B are added via standard blunt-end ligation (Fig. 1a, iia). The complement of the Roche A adaptor is used as the sequencing primer (see Fig. 2), while the complement of the Roche B adaptor is on the Sephadrose bead used in the emPCR reactions (Fig. 1b, i). One option at this point is to ligate adaptors that have barcodes (aka molecular IDs or MIDs) in their sequence (also see Fig. 2). The barcodes allow multiple discrete samples to be pooled together following library preparation and sequenced together in a pool with other like-treated samples for pyrosequencing. After the pyrosequencing run is complete and the DNA sequence data are collected, the software is then told to detect the barcodes in order to bin the reads and data accordingly.

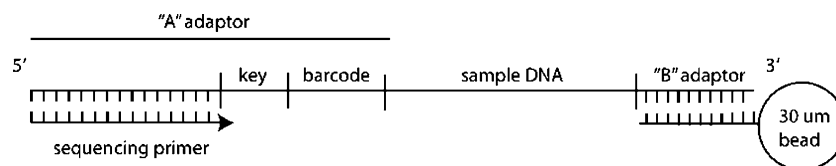


Fig. 2 Schematic of a Roche/454 single-stranded library. Adaptors “A” and “B” are blunt-end-ligated to the sample DNA fragment. The key (TCAG for sample DNA or ATCG for control beads) and the barcode are part of the larger Roche/454 A adaptor that is added

At this point in the process, the genomic sample is now fully converted into a single-stranded DNA library (sstDNA) that is compatible with the emPCR and Roche/454 pyrosequencing chemistry. After quantifying the amount of single-stranded DNA (a typical sstDNA library is 10–20 ng/μl) and estimating the size of the library using a Bioanalyzer RNA chip (Fig. 1b, iii), the molarity (number of DNA molecules per microliters) of the sample can be estimated. As each emPCR micelle is an independent clonal amplification of the template DNA inside (Fig. 1b), hundreds of thousands of discrete PCR products can be generated in a single 200 μl PCR tube.

In theory, 1 copy of DNA per bead is the preferred ratio. Because of the Poisson distribution of DNA at low concentration encountering a Sephadrose bead in solution, a typical estimate of the DNA to bead ratio is done by **taking the molarity and performing what’s referred to as a “titration”**. The concentration of DNA to emPCR beads that will maximize data output is **estimated by performing a series of test emPCRs where different amounts of DNA are added to emPCR beads** (i.e., “titration”). In a typical sample library titration, **four conditions are chosen** (see below) to bracket the idealized concentration of 1 copy of DNA/bead.

As mentioned earlier, *any* double-stranded DNA can be sequenced with the Roche/454 Life Sciences genome sequencer. In addition to being double-stranded, PCR products also have two useful characteristics that we can use to our advantage with HTS. The first is that one can order PCR primers of interest with the Roche A and B adaptors as part of the sequence (Fig. 1a, iib). This principle is analogous to the M13-tailed primers for dye-terminator sequencing. During the PCR reaction, the Roche adaptors are being created as part of the region of interest being amplified and thus create the equivalent of the genomic sstDNA library. The second benefit of PCR products is that they are often easy to quantify for concentration and size by standard gel electrophoresis or Bioanalyzer DNAChip, thus making the calculation of molarity relatively simple. As such, the need for a titration step is eliminated, saving time and money.

While one DNA copy per bead for genomic material is the theoretical recommendation, numbers can range from 0.125 up to 32 copies in our experience. This variation

during the blunt-end ligation. The sequencing primer anneals to the portion of the A adaptor immediately prior to the key, thus making the first base of the key the first nucleotide sequenced. The B adaptor is complementary to the oligo on the Sephadrose bead

seems dependent on the accurate estimation of the size and concentration of the sample library. For example, the sstDNA size may range from 300 to 1,200 bp with a median of 500 bp. However, since fragments larger than 600 bp do not amplify well in emPCR, the Bioanalyzer concentration of the sstDNA can dramatically overestimate the DNA to bead ratio.

Each individual library to be sequenced must be titrated to accurately calculate the DNA concentration for the emPCR. If too much DNA is added, then too many of the reads will fail during the run because of the presence of “mixed” signals (analogous to having two separate plasmid clones in one Sanger sequencing reaction). If too little DNA is added, not enough data will be collected, and the advantage of pyrosequencing is lost.

To determine the ideal amount of library to add to the beads, we set up small tests with two emulsion reactions per treatment (usually 0.5, 2, 4, and 16 copies/bead). Following the emulsion step, beads are enriched for those containing DNA, and the DNA negative beads are separated and counted on a coulter counter. The percentage of positive beads is then calculated by dividing the number of DNA positive beads by the sum of the DNA positive and the DNA negative beads. At our facility, we find that if an emulsion produces 12–15% DNA positive beads, this particular condition will produce the optimum amount of material for a Roche/454 pyrosequencing run while minimizing beads with no DNA and beads with more than one copy of DNA. If the beads are too positive (i.e., over 20%), the condition contains too much DNA per bead and will result in a high number of mixed signal sequences, which will be discarded by the software during the sequencing run.

Once the appropriate amount of beads are obtained from the emPCR for performing a run, the DNA positive beads are enriched and loaded onto a picotiter plate (PTP, Fig. 1b, iii) along with enzyme beads (containing the necessary enzymes for catalyzing the light production) and packing beads that help to hold the DNA bead in place (Fig. 1c, i). Since each PTP well is 40 μm in diameter and each DNA bead is 33 μm in diameter, each well will contain at most one DNA bead. At this point, the instrument run begins (see Fig. 1c, ii). Reagents are “flowed” over the PTP plate in the order of T, A, C, and G (=one flow). Each of the four nucleotides in a flow is recorded by a high-resolution CCD camera that is flush with the PTP. The onboard instrument software records each well for location (relative X, Y), intensity (two As are roughly twice as bright as one A), and key (TCAG for sample DNA and ATCG for control DNA; Fig. 1c, iii). A FLX LR70 run consists of 100 flows with each flow producing roughly 2.5 nucleotides for an average of 250 nucleotides for each of the 400,000 reads. Further details

on the principles and chemistry of pyrosequencing can be found in Ronaghi et al. (1998).

Comparison of HTS with Sanger Sequencing

The cost/time savings benefit of HTS versus Sanger sequencing has already been illustrated, but how do the two chemistries compare in terms of coverage, accuracy, and error rate? Because there is no cloning step where the sample DNA is ligated into *Escherichia coli* plasmids, there is no selective step where pieces of DNA may be “missed” because of sequence composition, GC content, or secondary structure. As such, the shotgun sequencing approach with pyrosequencing is more inclusive of typically hard-to-sequence regions such as AT-rich regions.

Accuracy of HTS with the Roche/454 Life Sciences chemistry is touted to be 0.5%/bp, while general consensus with Sanger sequencing is that the error rate is \sim 0.7%/bp. One of the key innovations of Roche/454 Life Sciences HTS is that for each run, a set of control beads are loaded along with the sample beads. With the current FLX chemistry, there are six control bead types that vary in composition (AT), homopolymers (a continuous stretch of a single nucleotide type), and length (150–250 bp). As the sequence of each of the six bead types are known and are sorted by the software as control beads (due to the ATGC rather than TCAG key at the beginning of the read), the accuracy of the individual run error rate can be easily estimated. On a standard LR70 with 420,000 sample reads, there are typically 15,000–30,000 control reads present. On a typical LR70 run, we see \sim 0.3–0.8%/bp error rates.

Considerations

With the obvious benefits of HTS for cost and time, what are some drawbacks or limitations that one should be aware of prior to designing a HTS experiment for a metagenomics project?

Accurate Quantification of Starting Material

One of the biggest drawbacks with current HTS technology is the amount of starting material required. According to the Roche/454 Life Science library protocol, 3–5 μg of starting material is required. The reasons for this relatively large amount of material (which is often not feasible for environmental samples) is that the purification steps for the Roche/454 library preparation uses Qiagen columns (where up to 50% of material can be lost at each of the three Qiagen steps in the library preparation protocol) and the fact that the final check of the single-stranded library

with the ligated adaptors must be quantified (typically by PicoGreen or an Agilent Bioanalyzer 2100) in order to estimate the ratio of DNA copies to beads for emulsion PCR (see above).

One of the most critical aspects of Roche/454 Life Sciences pyrosequencing protocol is accurate quantification of the sstDNA. Meyer et al. (2008b) have developed a protocol that takes advantage of the sensitivity of quantitative PCR (qPCR; aka real-time PCR) and the known sequences of the adaptors that have been ligated to the unknown sample. In this assay, a standard of known size and concentration is run alongside with experimental samples at specified dilutions to estimate a regression line. The number of molecules per microliter of experimental sample can be calculated at levels down to picograms levels rather than micrograms. As Meyer et al. (2008b) note, only 50 pg of initial DNA is needed for 12,000 reads on a single 16th region of a Roche/454 picotitre plate. For a full LR70 (420,000) reads, 1.75 ng of starting material would be sufficient for approximately 100 Mb of data.

Working with Low Amounts of Starting Material

The other drawback of HTS is actually having enough material to begin the library preparation. Roche/454 does not recommend library construction on samples where there is less than 1 μg of starting material. What does one do when their environmental sample produces only 0.5 μg of material? One option that has been formalized by Blow et al. (2008) is similar in concept to the qPCR protocol of Meyer et al. (2008b) in that it takes advantage of the adaptor sequences attached to the sample DNA. The twist is that the amplified product is used as the template for the PCR. In effect, a whole-genome amplification (WGA) takes place on the sample only after the library prep, rather than before. The benefit of the Blow et al. (2008) protocol is that if you have marginal levels of starting DNA (1–3 μg), you can still have the library prepared, and if there is too little as determined by a Bioanalyzer (range of detection is 5–50 ng/ μl), then the sstDNA can be used as a template in a standard PCR with the complement of the Roche/454 adaptors as the primers. We have successfully used this protocol for BACs, bacterial complementary DNA (cDNA), eukaryotic cDNA, and bacterial genomes with no discernable bias in coverage as also noted by Blow et al. (2008). The one obvious drawback of this approach is an additional amplification step, which can be negated, to some degree by using a high quality *Taq* with 5' and 3' proofreading capability. One alternative to the Blow et al. (2008) protocol is the use of a commercially available WGA kit such as the GE Healthcare Phi29 DNA polymerase amplification or Qiagen REPLI-g, which are capable of increasing sample amount by an order of magnitude. While

WGA is used in metagenomic projects (Abulencia et al. 2006; Lasken 2007), one should be aware of the introduction of biases attributed by an additional amplification step (Pinard et al. 2006).

Partial Plate Runs

A smaller, though important drawback to keep in mind is the vast amount of data obtained by HTS. If one is only interested in a small piece of DNA (i.e., BAC) or only need a small amount of data for a grant or proof-of-concept paper, 100 Mb of data is overkill. One solution is to pool together samples with other researchers and colleagues to fill out a plate. A second alternative is to run a partial plate. This latter option is often not available for people outsourcing their HTS work because many core facilities do not offer this or have limits on partial plates. Many in-house genome centers process partial plates for their researchers but it may take on the order of months for a plate to be sufficiently full to justify the cost of a run.

Homopolymers and Special Considerations for cDNA

HTS sequencing of cDNA poses a unique set of drawbacks, but is ultimately related to the issues of pyrosequencing homopolymers. Eukaryotic cDNA is transcribed from messenger RNA (mRNA) typically using commercially available kits that use a Poly-T primer. Because eukaryotic mRNA is poly-adenylated on the 3' end, the poly-T primer is used to target this region, as T is the complement of A. The one drawback of pyrosequencing cDNA libraries is the presence of these long regions of homopolymers. In our experience, the accuracy of Roche/454 Life Sciences base calling on homopolymers diminishes markedly after the fifth or sixth base in a homopolymer stretch. Additionally, the most nucleotides that can be incorporated in a single flow for a given nucleotide is ~ 10 . If three consecutive flows fail to get through a homopolymer (i.e., more than 30 As), the software discards that read as a "dot" which means that the software has interpreted the bead as containing no DNA and that particular read is discarded. This problem of data loss by the A homopolymers is exasperated by the fact that the intensity of the bead with the 30+ As also obscures neighboring beads (compare Fig. 3b vs. a), thus causing additional loss of data. Finally, the faulty reads obscure the control reads such that even the data collected cannot be used with much confidence (Table 1). Fortunately, a recent, simple modification solves the polyA dilemma with Roche/454 Life Sciences pyrosequencing.

Since the confounding issue is the presence of the polyA tail on the 3' end of the cDNA, Novaes et al. (2008) have

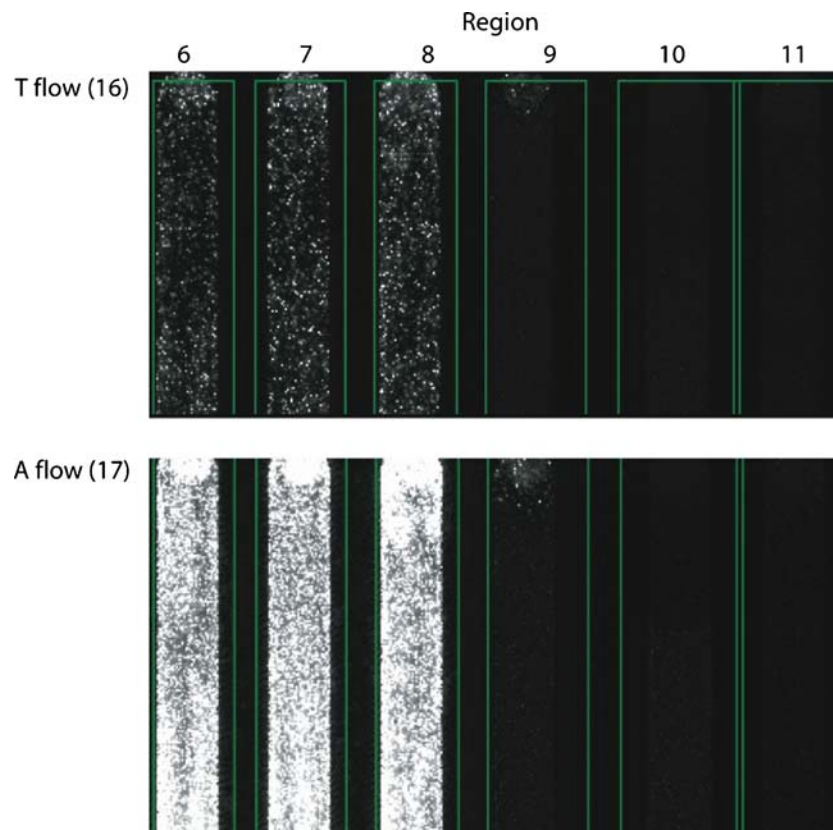


Fig. 3 Framegrab of plate images from a Roche/454 run of a bacterial cDNA sample with polyA tails present (regions 6–8) and absent (regions 9–11). Wells in regions 9–11 are not visible due to the

extreme white balance (>5,000) required to show the intensity of the A flow in the sequencing reaction for the polyA present sample

incorporated a modified 3' primer that has a rare restriction site (*Sfi*I) built in the 3' cDNA adaptor primer. Following relatively standard cDNA protocols, the additional step of digesting the cDNA with *Sfi*I and removal of the polyA tail via column or Ampure purification greatly increases the percent passing reads and data yield. A similar solution has been proposed by Frias-Lopez et al. (2008) for bacterial cDNA. Bacterial cDNA (which lack adenylation on their mRNA) are poly-adenylated via ligation, thus making them effectively eukaryotic cDNA for downstream applications. A rare restriction enzyme (*Bpm*I) recognition sequence is inserted between the poly-T and the T7 promoter sequence, thus allowing removal of the confounding polyA region prior to pyrosequencing library construction.

Coastal Ecosystems and Applications of HTS

As of December 2008, a total of 677 papers have referenced the Margulies et al. (2005) paper describing the proof-of-concept for the Roche/454 Life Sciences Genome Sequencer platform. Most applications of Roche/454 Life Sciences HTS have been for de novo characterization such as metagenomics, amplicons, and whole-genome sequencing of new organisms.

For metagenomics, there are two main approaches. The first is a systems-based approach where the entire sample of DNA is processed and analyzed. Depending on the sample origin, DNA from bacteria, archaea, eukaryotes, viruses, and plasmids will be detected by HTS. A widely used

Table 1 Summary statistics for regions 6–11 in Fig. 3a and b

	Region 6	Region 7	Region 8	Region 9	Region 10	Region 11
Raw wells	34,852	33,441	28,898	15,961	14,006	12,998
Keypass reads	16,660	18,372	15,020	14,700	12,908	12,090
Dot reads	10,794	12,393	10,174	852	866	594
Pass filter reads	55	67	101	11,556	10,232	10,039
Control reads	0	0	0	657	385	327
95% match	0	0	0	94.7	91	97

program to assist in the daunting task of characterizing a HTS metagenomic pyrosequencing run is the Metagenomic Rapid Annotation Subsystem Technology (MG-RAST, <http://metagenomics.nmpdr.org>; Meyer et al. 2008a). Through the use of a subsystem-based annotation approach, the MG-RAST web server provides annotation, phylogenetic as well as functional classification, and the ability to compare metagenomic samples to see both shared and unique genes/subsystems. Another widely used approach for processing and exploring metagenomic data is the MEGAN program (Huson et al. 2007).

The whole-system approach has been used to simultaneously characterize a variety of taxa: viruses in the ocean (Angly et al. 2006), multiple comparisons of biomes (Dinsdale et al. 2008), and bacterial community composition (Edwards et al. 2006; Sogin et al. 2006; Fierer et al. 2007; Huber et al. 2007; Pernthaler et al. 2008). One promising development has been the advancement of metagenomics beyond simple characterization of what species are present. Now that protocols exist for the construction of cDNA from bacteria, temporal and spatial studies of changes in bacterial communities are possible (Frias-Lopez et al. 2008). Additionally, the combination of metagenomics with the analysis of microbial community metabolism provides a powerful approach to understanding the function of microbial communities in a suite of environments (Turnbaugh and Gordon 2008). With the advent of autonomous times-series transcriptomics for coastal bacteria populations (Preston et al. 2009), our understanding of coastal communities and changes in both space and time promise to be an exciting area of research and development. Additionally, new protocols are being developed to pull a species of interest from the metagenomic haystack (Pernthaler et al. 2008; Zehr et al. 2008). As HTS becomes a more accessible tool in the biologist's toolbox, more creative and exciting modifications are bound to change the way we view science in the ocean and beyond. The whole-system approach has the benefit that one is characterizing a community by the transcriptomic potential (i.e., are ammonia oxidizers present in a particular sample/environment?).

The second approach to metagenomics is designed to identify which species are present in a sample. Unlike the first approach, this approach has the distinct likelihood of potentially missing certain taxa through the use of PCR primers for amplifying specific regions (usually variable regions of the *16S rRNA*). Even with the known bias of PCR amplification and missing certain taxa due to primer mismatches (Fodor et al. 2009), the microbial diversity estimates being discovered using HTS are truly astounding (Fierer et al. 2007; Huber et al. 2007). The one drawback of this approach is that the transcriptomic potential of a community is being inferred

by species' assignment based on a single molecule (typically *16S rRNA*).

Ideally, one would use a multi-tiered approach to characterizing a metagenomic community. Just as one should not rely on a single gene for a phylogeny or a particular dataset, one should consider all accessible avenues of data in order to strengthen ones conclusions. In order to accurately characterize a community, one would look at the genes expressed in a particular sample (i.e., cDNA) in addition to all genes present (i.e., metagenomic DNA). With advances in methodology such as Frias-Lopez et al. (2008) and the ability to barcode samples, HTS is a tool with great potential for revolutionizing the way we design studies for coastal ecosystems.

Data Handling and Storage

Many open-source programs and scripts (especially Perl and Python) exist and are freely available in the HTS/metagenomics community. The main issue that many people first entering into the HTS world encounter is that the raw data (the Ts, As, etc.) are rarely viewed by a human eye due to the enormous amount of data. The accuracy of Roche/454 pyrosequencing is over 99% for a single read and close to 100% with coverage (usually 10–20× depending on the project). Additionally, the individual nucleotide quality scores (Q , where $Q = -10 \log p$ and p is the estimated error probability of the base) are invariably over Q35 (usually closer to Q40), which is considerably better than the Q20 score (i.e., odds of miscalling an individual nucleotide is 1%) used as a cutoff with traditional Sanger sequencing. It is also not possible to manually go through all nucleotides in a HTS dataset to verify the accuracy of nucleotides as one would do with Sanger reads. Today's biologist must be adept and knowledgeable with Unix/Linux commands and how to apply pipelines for data processing. The final note of warning for those considering HTS as part of their experimental toolbox is data storage. The current Roche/454 Life Sciences FLX run produces on average 100 Mb of data in a folder of approximately 12–18 Gb of data. The new chemistry upgrade ("Titanium") that has become recently available produces ~400 Mb of data in a folder of ~35 Gb.

Conclusions

Our understanding of the environment has been revolutionized by findings one could not have possibly imagined before the application of high-throughput sequencing to metagenomics. With new protocols and technologies

continually being developed, the metagenomics field will continue to evolve and revolutionize our understanding of complex coastal communities and how they function and change over space and time.

References

- Abulencia, C.B., D.L. Wyborski, J.A. Garcia, M. Podar, W.Q. Chen, S.H. Chang, H.W. Chang, D. Watson, E.L. Brodie, T.C. Hazen, and M. Keller. 2006. Environmental whole-genome amplification to access microbial populations in contaminated sediments. *Applied and Environmental Microbiology* 72: 3291–3301. doi:10.1128/AEM.72.5.3291-3301.2006.
- Angly, F.E., B. Felts, M. Breitbart, P. Salamon, R.A. Edwards, C. Carlson, A.M. Chan, M. Haynes, S. Kelley, H. Liu, J.M. Mahaffy, J.E. Mueller, J. Nulton, R. Olson, R. Parsons, S. Rayhawk, C.A. Suttle, and F. Rohwer. 2006. The marine viromes of four oceanic regions. *Plos Biology* 4: 2121–2131. doi:10.1371/journal.pbio.0040368.
- Blow, M.J., T. Zhang, T. Woyke, C.F. Speller, A. Krivoshapkin, D.Y. Yang, A. Derevianko, and E.M. Rubin. 2008. Identification of ancient remains through genomic sequencing. *Genome Research* 18: 1347–1353. doi:10.1101/gr.076091.108.
- Chen, K. and L. Pachter. 2005. Bioinformatics for whole-genome shotgun sequencing of microbial communities. *PLoS Computational Biology* 1: e24. doi:10.1371/journal.pcbi.0010024.
- Dinsdale, E.A., R.A. Edwards, D. Hall, F. Angly, M. Breitbart, J.M. Brulc, M. Furlan, C. Desnues, M. Haynes, L.L. Li, L. McDaniel, M.A. Moran, K.E. Nelson, C. Nilsson, R. Olson, J. Paul, B.R. Brito, Y.J. Ruan, B.K. Swan, R. Stevens, D.L. Valentine, R.V. Thurber, L. Wegley, B.A. White, and F. Rohwer. 2008. Functional metagenomic profiling of nine biomes. *Nature* 452: 629–U628. doi:10.1038/nature06810.
- Edwards, R.A., B. Rodriguez-Brito, L. Wegley, M. Haynes, M. Breitbart, D.M. Peterson, M.O. Saar, S. Alexander, E.C. Alexander, and F. Rohwer. 2006. Using pyrosequencing to shed light on deep mine microbial ecology. *BMC Genomics* 7: 57.
- Fierer, N., M. Breitbart, J. Nulton, P. Salamon, C. Lozupone, R. Jones, M. Robeson, R.A. Edwards, B. Felts, S. Rayhawk, R. Knight, F. Rohwer, and R.B. Jackson. 2007. Metagenomic and small-subunit rRNA analyses reveal the genetic diversity of bacteria, archaea, fungi, and viruses in soil. *Applied and Environmental Microbiology* 73: 7059–7066. doi:10.1128/AEM.00358-07.
- Fodor, A., W. Jones, and T. Hamp. 2009. The effects of experimental choices and analysis noise on the rare biosphere. *Applied Environmental Microbiology*. doi:10.1128/AEM.01931-08.
- Frias-Lopez, J., Y. Shi, G.W. Tyson, M.L. Coleman, S.C. Schuster, S. W. Chisholm, and E.F. DeLong. 2008. Microbial community gene expression in ocean surface waters. *Proceedings of the National Academy of Sciences of the United States of America* 105: 3805–3810. doi:10.1073/pnas.0708897105.
- Handelsman, J., M.R. Rondon, S.F. Brady, J. Clardy, and R.M. Goodman. 1998. Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. *Chemistry & Biology* 5: R245–R249. doi:10.1016/S1074-5521(98)90108-9.
- Huber, J.A., D. Mark Welch, H.G. Morrison, S.M. Huse, P.R. Neal, D. A. Butterfield, and M.L. Sogin. 2007. Microbial population structures in the deep marine biosphere. *Science* 318: 97–100. doi:10.1126/science.1146689.
- Huson, D.H., A.F. Auch, J. Qi, and S.C. Schuster. 2007. MEGAN analysis of metagenomic data. *Genome Research* 17: 377–386. doi:10.1101/gr.5969107.
- Krause, L., N.N. Diaz, A. Goesmann, S. Kelley, T.W. Nattkemper, F. Rohwer, R.A. Edwards, and J. Stoye. 2008. Phylogenetic classification of short environmental DNA fragments. *Nucleic Acids Research* 36: 2230–2239. doi:10.1093/nar/gkn038.
- Lasken, R.S. 2007. Single-cell genomic sequencing using multiple displacement amplification. *Current Opinion in Microbiology* 10: 510–516. doi:10.1016/j.mib.2007.08.005.
- Mardis, E.R. 2008. Next-generation DNA sequencing methods. *Annual Review of Genomics and Human Genetics* 9: 387–402. doi:10.1146/annurev.genom.9.081307.164359.
- Margulies, M., M. Egholm, W.E. Altman, S. Attiya, J.S. Bader, L.A. Bembien, J. Berka, M.S. Braverman, Y.J. Chen, Z.T. Chen, S.B. Dewell, L. Du, J.M. Fierro, X.V. Gomes, B.C. Godwin, W. He, S. Helgesen, C.H. Ho, G.P. Irzyk, S.C. Jando, M.L.I. Alenquer, T.P. Jarvie, K.B. Jirage, J.B. Kim, J.R. Knight, J.R. Lanza, J.H. Leamon, S.M. Lefkowitz, M. Lei, J. Li, K.L. Lohman, H. Lu, V. B. Makhijani, K.E. McDade, M.P. McKenna, E.W. Myers, E. Nickerson, J.R. Nobile, R. Plant, B.P. Puc, M.T. Ronan, G.T. Roth, G.J. Sarkis, J.F. Simons, J.W. Simpson, M. Srinivasan, K. R. Tartaro, A. Tomasz, K.A. Vogt, G.A. Volkmer, S.H. Wang, Y. Wang, M.P. Weiner, P.G. Yu, R.F. Begley, and J.M. Rothberg. 2005. Genome sequencing in microfabricated high-density picolitre reactors. *Nature* 437: 376–380.
- Meyer, F., D. Paarmann, M. D'Souza, R. Olson, E.M. Glass, M. Kubal, T. Paczian, A. Rodriguez, R. Stevens, A. Wilke, J. Wilkening, and R.A. Edwards. 2008a. The metagenomics RAST server—a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics* 9: 386.
- Meyer, M., A.W. Briggs, T. Maricic, B. Hober, B.H. Hoffner, J. Krause, A. Wehmann, S. Paabo, and M. Hofreiter. 2008b. From micrograms to picograms: quantitative PCR reduces the material demands of high-throughput sequencing. *Nucleic Acids Research* 36: e5.
- Miller, W., D.I. Drautz, A. Ratan, B. Pusey, J. Qi, A.M. Lesk, L.P. Tomsho, M.D. Packard, F.Q. Zhao, A. Sher, A. Tikhonov, B. Raney, N. Patterson, K. Lindblad-Toh, E.S. Lander, J.R. Knight, G.P. Irzyk, K.M. Fredrikson, T.T. Harkins, S. Sheridan, T. Pringle, and S.C. Schuster. 2008. Sequencing the nuclear genome of the extinct woolly mammoth. *Nature* 456: 387–U351. doi:10.1038/nature07446.
- Novaes, E., D.R. Drost, W.G. Farmerie, G.J. Pappas, D. Grattapaglia, R.R. Sederoff, and M. Kirst. 2008. High-throughput gene and SNP discovery in *Eucalyptus grandis*, an uncharacterized genome. *BMC Genomics* 9: 312.
- Pernthaler, A., A.E. Dekas, C.T. Brown, S.K. Goffredi, T. Embaye, and V.J. Orphan. 2008. Diverse syntrophic partnerships from deep-sea methane vents revealed by direct cell capture and metagenomics. *Proceedings of the National Academy of Sciences of the United States of America* 105: 7052–7057. doi:10.1073/pnas.0711303105.
- Pinard, R., A. de Winter, G.J. Sarkis, M.B. Gerstein, K.R. Tartaro, R. N. Plant, M. Egholm, J.M. Rothberg, and J.H. Leamon. 2006. Assessment of whole genome amplification-induced bias through high-throughput, massively parallel whole genome sequencing. *BMC Genomics* 7: 216.
- Preston, C.M., R. Marin 3rd, S.D. Jensen, J. Feldman, J.M. Birch, E.I. Massion, E.F. DeLong, M. Suzuki, K. Wheeler, and C.A. Scholin. 2009. Near real-time autonomous detection of marine bacterioplankton on a coastal mooring in Monterey Bay, California, using rRNA-targeted DNA probes. *Environmental Microbiology* 11: 1168–1180. doi:10.1111/j.1462-2920.2009.01848.x.
- Ronaghi, M., M. Uhlen, and P. Nyren. 1998. A sequencing method based on real-time pyrophosphate. *Science* 281: 363. doi:10.1126/science.281.5375.363.
- Sogin, M.L., H.G. Morrison, J.A. Huber, D. Mark Welch, S.M. Huse, P. R. Neal, J.M. Arrieta, and G.J. Herndl. 2006. Microbial diversity in the deep sea and the underexplored “rare biosphere”. *Proceedings of*

- the National Academy of Sciences of the United States of America* 103: 12115–12120. doi:10.1073/pnas.0605127103.
- Turnbaugh, P.J. and J.I. Gordon. 2008. An invitation to the marriage of metagenomics and metabolomics. *Cell* 134: 708–713. doi:10.1016/j.cell.2008.08.025.
- Venter, J.C., M.D. Adams, E.W. Myers, P.W. Li, R.J. Mural, G.G. Sutton, H.O. Smith, M. Yandell, C.A. Evans, R.A. Holt, J.D. Gocayne, P. Amanatides, R.M. Ballew, D.H. Huson, J.R. Wortman, Q. Zhang, C.D. Kodira, X.Q.H. Zheng, L. Chen, M. Skupski, G. Subramanian, P.D. Thomas, J.H. Zhang, G.L.G. Miklos, C. Nelson, S. Broder, A.G. Clark, C. Nadeau, V.A. McKusick, N. Zinder, A.J. Levine, R.J. Roberts, M. Simon, C. Slayman, M. Hunkapiller, R. Bolanos, A. Delcher, I. Dew, D. Fasulo, M. Flanigan, L. Florea, A. Halpern, S. Hannenhalli, S. Kravitz, S. Levy, C. Mobarry, K. Reinert, K. Remington, J. Abu-Threideh, E. Beasley, K. Biddick, V. Bonazzi, R. Brandon, M. Cargill, I. Chandramouliswaran, R. Charlab, K. Chaturvedi, Z.M. Deng, V. Di Francesco, P. Dunn, K. Eilbeck, C. Evangelista, A.E. Gabrielian, W. Gan, W.M. Ge, F.C. Gong, Z.P. Gu, P. Guan, T.J. Heiman, M.E. Higgins, R.R. Ji, Z.X. Ke, K.A. Ketchum, Z.W. Lai, Y.D. Lei, Z.Y. Li, J.Y. Li, Y. Liang, X.Y. Lin, F. Lu, G.V. Merkulov, N. Milshina, H.M. Moore, A.K. Naik, V.A. Narayan, B. Neelam, D. Nusskern, D.B. Rusch, S. Salzberg, W. Shao, B.X. Shue, J.T. Sun, Z.Y. Wang, A.H. Wang, X. Wang, J. Wang, M.H. Wei, R. Wides, C.L. Xiao, C.H. Yan, A. Yao, J. Ye, M. Zhan, W. Q. Zhang, H.Y. Zhang, Q. Zhao, L.S. Zheng, F. Zhong, W.Y. Zhong, S.P.C. Zhu, S.Y. Zhao, D. Gilbert, S. Baumhueter, G. Spier, C. Carter, A. Cravchik, T. Woodage, F. Ali, H.J. An, A. Awe, D. Baldwin, H. Baden, M. Barnstead, I. Barrow, K. Beeson, D. Busam, A. Carver, A. Center, M.L. Cheng, L. Curry, S. Danaher, L. Davenport, R. Desilets, S. Dietz, K. Dodson, L. Doup, S. Ferriera, N. Garg, A. Gluecksmann, B. Hart, J. Haynes, C. Haynes, C. Heiner, S. Hladun, D. Hostin, J. Houck, T. Howland, C. Ibegwam, J. Johnson, F. Kalush, L. Kline, S. Koduru, A. Love, F. Mann, D. May, S. McCawley, T. McIntosh, I. McMullen, M. Moy, L. Moy, B. Murphy, K. Nelson, C. Pfannkoch, E. Pratts, V. Puri, H. Qureshi, M. Reardon, R. Rodriguez, Y.H. Rogers, D. Romblad, B. Ruhfel, R. Scott, C. Sitter, M. Smallwood, E. Stewart, R. Strong, E. Suh, R. Thomas, N.N. Tint, S. Tse, C. Vech, G. Wang, J. Wetter, S. Williams, M. Williams, S. Windsor, E. Winn-Deen, K. Wolfe, J. Zaveri, K. Zaveri, J.F. Abril, R. Guigo, M.J. Campbell, K.V. Sjolander, B. Karlak, A. Kejariwal, H.Y. Mi, B. Lazareva, T. Hatton, A. Narechania, K. Diemer, A. Muruganujan, N. Guo, S. Sato, V. Bafna, S. Istrail, R. Lippert, R. Schwartz, B. Walenz, S. Yoosheph, D. Allen, A. Basu, J. Baxendale, L. Blick, M. Caminha, J. Carnes-Stine, P. Caulk, Y.H. Chiang, M. Coyne, C. Dahlke, A.D. Mays, M. Dombroski, M. Donnelly, D. Ely, S. Esparham, C. Fosler, H. Gire, S. Glanowski, K. Glasser, A. Glodek, M. Gorokhov, K. Graham, B. Gropman, M. Harris, J. Heil, S. Henderson, J. Hoover, D. Jennings, C. Jordan, J. Jordan, J. Kagan, L. Kagan, C. Kraft, A. Levitsky, M. Lewis, X.J. Liu, J. Lopez, D. Ma, W. Majoros, J. McDaniel, S. Murphy, M. Newman, T. Nguyen, N. Nguyen, M. Nodell, S. Pan, J. Peck, M. Peterson, W. Rowe, R. Sanders, J. Scott, M. Simpson, T. Smith, A. Sprague, T. Stockwell, R. Turner, E. Venter, M. Wang, M.Y. Wen, D. Wu, M. Wu, A. Xia, A. Zandieh, and X.H. Zhu. 2001. The sequence of the human genome. *Science* 291: 1304. doi:10.1126/science.1058040.
- Wheeler, D.A., M. Srinivasan, M. Egholm, Y. Shen, L. Chen, A. McGuire, W. He, Y.J. Chen, V. Makhijani, G.T. Roth, X. Gomes, K. Tartaro, F. Niazi, C.L. Turcotte, G.P. Irzyk, J.R. Lupski, C. Chinault, X.Z. Song, Y. Liu, Y. Yuan, L. Nazareth, X. Qin, D.M. Muzny, M. Margulies, G.M. Weinstock, R.A. Gibbs, and J.M. Rothberg. 2008. The complete genome of an individual by massively parallel DNA sequencing. *Nature* 452: 872–U875. doi:10.1038/nature06884.
- Wommack, K.E., J. Bhavsar, and J. Ravel. 2008. Metagenomics: read length matters. *Applied and Environmental Microbiology* 74: 1453–1463. doi:10.1128/AEM.02181-07.
- Zehr, J.P., S.R. Bench, B.J. Carter, I. Hewson, F. Niazi, T. Shi, H.J. Tripp, and J.P. Affourtit. 2008. Globally distributed uncultivated oceanic N-2-fixing cyanobacteria lack oxygenic photosystem II. *Science* 322: 1110–1112. doi:10.1126/science.1165340.