DNA and Genome

Abstract

The size of a genome may change very rapidly if it fuses with another genome, or accumulates some DNA via a virus, or some other mechanism of horizontal transfer. Acquiring new DNA means acquiring new genes, but genomes do have size limits. Why? Every round of replication extracts a cost for the larger genome, and therefore genomes must balance the expense of replicating redundant DNA with the benefit of having genes that provide a selective advantage only under rare circumstances. Conversely, losing DNA and genes could be advantageous if the cell evolves to fill a new niche, such as inside another species. If genes are no longer advantageous, the DNA can be lost and the more efficient genome provides a selective advantage. From what we can tell so far, genome sizes tend to stay within a fairly narrow size range for a given group of species. For example, K- and O-islands are newly acquired DNA, but all gut bacteria tend to have genomes in the 4 to 5 Mb range. As they acquire new DNA, cells tend to return to a genome home-ostasis with an optimal size and gene count.

You might think that 200 sequenced genomes is enough and that we don't need to sequence more, but there is power in numbers for comparative genomics. Consider the analogy of living in a cave all your life and coming out one day and seeing a bluebird and a blue jay. Based on this sampling, you might conclude that all birds are blue. Later in the day, you see a red cardinal and a yellow canary, which leads you to conclude diat all birds must be primary colors. Using a small sample size leads to inaccurate conclusions. Imagine your surprise when you see a hummingbird, an ostrich, and a penguin. Just as we learn more about birds by studying their diversity, we learn more about genomes when we have a larger sample size. However, resources are limited so we must choose wisely which genomes we sequence to maximize our ability to learn from them (see Section 2.1).

DISCOVERY QUESTIONS

5. Copy and paste the *E. coli* K-12 20 kb fragment of DNA and perform a GC skew analysis on the DNA. Can you detect the origin of replication? Go to the *M. genitalium* genome picture (Figure 2.15) and use gene orientation to identify the origin of replication. Then find one gene not pointing in the "right" direction and see if you can determine if it is in the Database of Essential Genes (DEC).

7. Search *coli* BASE for the gene *nuoL* and follow the links to *E. coli* K-12 MG1655. Now go to the co/zBASE Browser and choose K-12 MG1655. Change the coloration to orthologs for all available species and mouse over the highly conserved *nuoL* gene located on the inner strand just past 6 o'clock (see the text box to confirm its location). Change the view to coloration by GC content. Is this essential and highly conserved cluster of genes above or below the average GC content? Approximately what is the GC content for *nuol2*

8. Go to the DOGS genome size web page, choose the bacterial list, and click on the upward-pointing arrow to sort by size. How small is the smallest genome? How many genes does it have? Click on the species name and see why this number seems too small. Now focus on species with "Main" listed under the "Segment" column to see entire genome sizes. Can you see a pattern between genome size and where small-genome species live? If you want to learn where one lives, read the linked abstract.

9. Go to Genomes OnLine Database (GOLD) and click the button for the published prokaryote genomes. Do a find function for *Bacillus anthmcis* Ames 0581. Click on the "MAP" link to see a complete list of genes with *DnaA* at the top of the list. Click

on the species name link at the top of this page to see a graphic depiction of the genome. Click on the circular map in the area of GBAA1887. Navigate left or right until you can see genes 1,887 and 1,888. What COG category are these genes? Click on each box to find out what the gene encodes. How does their function match your interpretation of the COG categories?

How Many Genomes Are There?

Determining the number of species in the world has been a challenging problem for many years. Biologists have sampled and counted all over the world, but we tend to count the easiest ones first. For example, halt of all described species are insects, including nearly 300,000 beetle species (which prompted naturalist J. B. S. Haldane to remark, 'The Creator, if he exists, has a special preference for beetles"). The bigger the organism, the easier it is to count, which means that prokaryotes are the least well-documented organisms. We are uncertain how many Eubacteria and Archaea live on the planet, and we still have trouble defining species-specific characteristics for most prokaryotes. As we saw with cyanobacteria, perhaps genome sequences will provide the best estimate of the number of distinct species in the world.

Craig Venter, the genomicist who pioneered whole genome shotgun (WGS) sequencing, has taken a sabbatical of sorts to sail around the world, collecting DNA samples from all the oceans. He is confident that he can measure species diversity around the world because in 2004 he and 22 coauthors published the first environmental WGS sequencing results from the Sargasso Sea. The Sargasso Sea, located just south of Bermuda in the Atlantic Ocean, is probably the most-studied oceanic region in the world. Venter sampled about 1,500 liters of surface waters 7 times in 4 locations and sorted the species by size to collect organisms bigger than 0.1 urn but smaller than 3.0 um (i.e., primarily prokaryotes). The cells were lysed and the resulting 7 mixtures of DNA were cut into small pieces and ligated into plasmids. An average of 818 bases from both ends of 990,000 plasmid inserts were sequenced to produce a total of 1.62 billion bp of data. Using a combination of assembly software and hand-curation, the investigators assembled 64,398 scaffolds (a collection of contigs lumped together) of 826 bp to 2.1 Mb. The two main questions they wanted to answer were:

1. How many species are there?
2. What is the relative abundance of each species?

Let's look at the number of species first. The investigators identified 1,412 different small subunit rRNA genes or fragments, with 148 of these being new to die database. This indicates that 10% of these ribosomal genes were from species never before sequenced, and illustrates the limitations of previous sampling methods. Most of these species cannot be grown in the lab and therefore we cannot use standard microbiology methods to characterize them. Some investigators have used "universal PCR primers" to amplify all ribosomal genes, but PCR amplification is not uniform and some genes may not bind to these universal primers. Therefore, WGS sequencing identified 148 previously unknown rRNA genes. However, ribosomal genes were sampled randomly with the WGS method, and other genes may give different species counts.

The investigators used 6 additional genes to estimate a range of 341—569 sampled species, or phylotypes (Table 3.1). The term phylotypes is the newest effort to clarify the term "species" when classifying prokaryotes. Intended as a functionally equivalent term to "species," phylotypes recognizes that arbitrary distinctions are used to classify a species, since the mating criterion cannot be used widi prokaryotes.

Table 3.1 Diversity of species defined by six different proteins. Ortholog cutoff refers to the **E-value** used to determine if a sequence was a true ortholog when the *E. coli* gene was queried with BLASTx against the collection of Sargasso DNA.

Protein Name	Sequence ID	Ortholog Cutoff	Observed Phylotypes
AtpD	NTL01EC03653	$1 \times 10^{-[illegible]}$	456
GyrB	NTL01EC03620	$1 \times 10^{-[illegible]}$	569
Hsp70	NT01EC0015	$1 \times 10^{-[illegible]}$	515
RecA	NTL01EC02539	$1 \times 10^{-[illegible]}$	341
RpoB	NTL01EC03885	$1 \times 10^{-[illegible]}$	428
TufA	NTL01EC03262	$1 \times 10^{-[illegible]}$	397

The exact number of identifiable phylotypes varies depending on the gene chosen, but the advantage of using ribosomal genes is a huge database of orthologs; rRNA changes very little over time, and every species has ribosomes. The caveat about using ribosomal genes is diat two phylotypes with highly conserved ribosomal genes might be collapsed into a single phylotype by the assembly software, causing the loss of one species in our counting. Another bonus of the WGS sampling method was that sequences included dsDNA viruses in the water. Using the virus database as a standard, the investigators identified 71 scaffolds at least 10 kb long, with 50 different viral genes in the scaffolds, and another 150 viral genes in sequence reads that did not assemble into scaffolds.

The authors acknowledged their sampling method was not comprehensive and that they probably failed to sequence DNA from less abundant species. Other species probably were missed due to the random nature of the DNA cloning process and the inability to find an identifiable gene. To address the sampling omissions, Venter's group performed three different calculations to account for missed species. The most conservative estimate was about 1,800 different phylotypes in the water collected. To sequence 95% of all species in their samples, they would need 12 times more sequence data which permits a cost/benefit calculation of expense versus the value of an estimated number of species from a small sample of the world's ocean.

Because the cloned DNA fragments were sequenced randomly, you would expect that more abundant species would have their DNA sequenced more often; thus, the coverage of a particular gene in an abundant species would be greater and a greater number of different genes would be sequenced per species. For example, 53% of all sequenced DNA in water sample #1 were from two genera, *Shewanella* and *Burkholderia*—but this confounds previous knowledge of these species. *Shewanella* is usually found in nutrient-rich water, not nutrient-poor water like the Sargasso Sea. *Burkholderia* is considered a terrestrial species. However, open ocean water contains marine snow, tiny bits of decaying organic matter (including animal feces). If the sampling of water happened in an area with decaying feces, the microbial composition of the sample would be altered.

Unfortunately, Venter's group did not view the samples microscopically, so we do not know if their samples contained marine snow. In other water samples, they found abundant cyanobacteria DNA, especially *Prochlorococcus* and *Synechococcus*, whose genomes we studied in Section 2.2. Ninety percent of the cyanobacteria DNA was from *Prochlorococcus,* but again, the sampling methods may have created this bias since *Prochlorococcus* is smaller and thus may have fit into the niters better than the larger Synechococcus. Taking advantage of the completed *Prochlorococcus* MED4 genome, the researchers assembled four distinguishable *Prochlorococcus* genomes, indicating the diversity of this phylotype is greater than we had known (Figure 3.2). Find an area where 4 rings of genomes overlap in the circle (e.g., at -10 o'clock) and you can see how they arrived at their minimal estimate of *ProMorococcus* diversity. Notice the large gap (at 8 o'clock) in the 4 inner circles;

these genes encode surface polysaccharide synthesizing enzymes and may be either unique to MED4 or highly divergent alleles and thus not recognized as orthologs in the Sargasso DNA. Only by comparing multiple *ProMorococcus* genomes were we able to identify a diverse cluster of genes in the reference genome sequence.

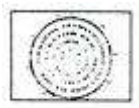

Figure 3.2 Gene conservation among closely related *Prochlorococcus*.
Go to www.GeneticsPlace.com to view this figure.

As you can imagine, sequencing random environmental DNA is bound to uncover some new genes, but you may be surprised by the number of new genes. A total of 1,214,207 genes were identified and added to the databases, including 69,901 novel genes. One class of genes that was very abundant was a recently discovered gene family in marine bacteria called bacteriorhodopsin. Bacteriorhodopsin permits cells to harvest solar energy in the absence of chlorophyll. Previous sampling by PCR had uncovered 67 bacteriorhodopsin homologs, but the Sargasso DNA contained 782 new bacteriorhodopsin homologs—more than 10 times the previous total! The investigators clustered all the bacteriorhodopsin genes and found 13 families from a wider range of phylotypes than we had known before (Figure 3.3).

The purpose of this figure is to impress upon you the degree of our ignorance of the oceans, which influence our global climate (e.g., CO_2 balance; see Section 4.1) and nutrient cycles (nitrogen, phosphorus, and the food chain). Keep in mind, this phylogenetic tree was taken from only 1,500 liters of water in one area, compared to the ocean's estimated volume of 1.37 X 10" liters. Clearly, we have only begun to sample the oceans' diversity, not to mention the different terrestrial prokaryotes in diverse environments. The survey raises new questions, such as how nutrient-poor environments can sustain so much genomic diversity (see Section 4.1).

DISCOVERY QUESTIONS

10. Search ProbeBase, modify the pull-down menu • under the heading "List probes by category" to list the "organisms of medical or hygienical relevance," and click "enter." Do a find function for the probe "Hpl6S-2" and view the probe. Copy and paste the sequence into a BLASTn search. Do you get only one species with 100% matching?

11. Based on work published in 1990, oceanogra-phers estimated that 1% to 4% of planktonic bacteria are infected with phage. Based on the number of species in the Sargasso environmental sampling, calculate the number of viruses that should be in that sample. Given their sampling methods, would you expect the number of viruses identified to be higher or lower than the viruses actually present in the sample? Explain your answer.

12. Read a segment from a National Academy of Sciences report on biodiversity. Given the bias in sampling, what effect do you think environmental genomics can have on our awareness of the unseen diversity? In 2005, Venter announced an "Air Genome Project" in which his group will sequence DNA sampled from New York City air. Who knows what we're breathing?

13. Perform an Entrez search for the accession number AACYO1000000. What submission is this? Click on the link and then click on the organism link to see how it is classified. Click on the link "1,986,782" that appears in the bottom right corner of the table. Change the display from "FASTA" to "Trace," click on the color box, then hit the "Show" button. Examine these first few sequences and determine if all the reads were of equal quality (turn on the "Confidence" option). ; How many bases would you trust from each of the first three reads?

14. Go to NCBI's main page, choose "conserved domain," search for "bacteriorhodopsin," and follow the "pfam" link. Change the view to show as many as are available in the menu; set the color to ; "identity" and die "Type Selection" to the most diverse members, then hit the "Show Alignment" button. Copy the first stretch of uninterrupted amino acids in the consensus line and perform a BLASTp search. How many good hits did you get? Now try a BLASTp using a region with high conservation as revealed in your modified display. Did you get more hits the second time?

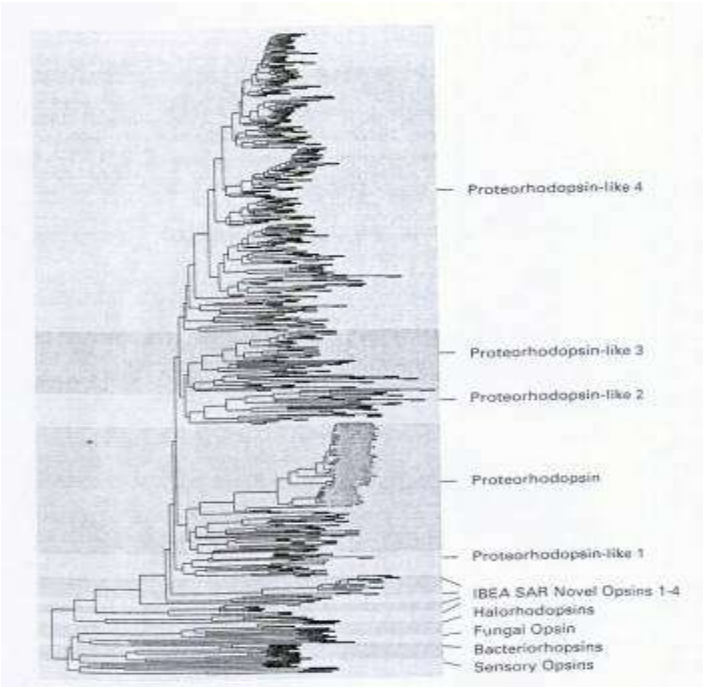

Figure 3.3 Phylogenetic tree of rhodopsinlike genes in the Sargasso Sea data and GenBank.
Sargasso sequences are colored purple, cultured species are black, and other environmental samples are gray Sequences from uncultured species in the Sargasso bea. Subfamilies of rhodopsins are indicated on the right. Sequences greater than 75 ammo acids long were aligned to each other using CLUSTALw, and a neighbor-]Oinmg Phylogenetic tree was inferred using Phylip.

What Can We Learn by Comparing Many Whole Genomes?

When we examine a lot of data, what we see depends on what we want to know. If you examine a forest from a distance of 1 meter, you notice the bark. From 20 meters, you notice the shapes of the trees and the foliage; 200 meters reveals species distribution and patchiness; from an airplane you see large patterns of trees combined with other geographical features. Similarly, when we examine many genomes, our perspective determines what we will notice. Zooming in to the amino acid level of the proteome, a group of investigators measured the frequency of amino acid usage in fifteen different species from all three domains of life using three-way comparisons. They identified amino acids in orthologs that were identical in two distantly related species but different in one of two closely related species. By identifying conserved amino acids (in distantly related species) that had drifted during a short evolutionary period (closely related species), they compiled a large number of amino acid changes and then charted die frequency for every amino acid, regardless of its position in a protein. It may surprise you to learn that they found a pattern in all species, including humans: cysteine, methionine, histidine, serine, and phenylalanine all showed increased frequency in these.