JOURNAL OF BACTERIOLOGY, Dec. 2002, p. 6403–6405 Vol. 184, No. 23
0021-9193/02/$04.00+0 DOI: 10.1128/JB.184.23.6403–6405.2002

DIALOG

The Value of Complete Microbial Genome Sequencing (You Get What You Pay For)

Claire M. Fraser,* Jonathan A. Eisen, Karen E. Nelson, Ian T. Paulsen, and Steven L. Salzberg

The Institute for Genomic Research, Rockville, Maryland 20850

Since the publication of the complete *Haemophilus influenzae* genome sequence in July 1995 (4), the field of microbiology has been one of the largest beneficiaries of the breakthroughs in genomics and computational biology that made this accomplishment possible. When the 1.8-Mbp *H. influenzae* project began in 1994, it was not certain that the whole-genome shotgun sequencing strategy would succeed because it had never been attempted on any piece of DNA larger than an average lambda clone (~40 kbp) (9).

During the past 7 years, progress in DNA sequencing technology, the design of new vectors for library construction for use in shotgun sequencing projects, significant improvements in closure and finishing strategies, and more sophisticated and robust methods for gene finding and annotation have dramatically reduced the time required for each stage of a genome project and the cost per base pair while at the same time producing a finished product of higher quality than was possible just a few years ago. Today, the random sequencing phase of a genome project, representing more than 99% of the genome sequence, can easily be completed in just a few days at a cost of approximately 3 to 4¢ per bp. The early release of such draft data is of benefit to the scientific community, and there are numerous examples of how access to incomplete data has had a significant impact in many areas of microbial research.

Comparable breakthroughs have also been achieved in closure strategies in centers such as The Institute for Genomic Research (TIGR) and the Pathogen Sequencing Unit at the Sanger Centre, which routinely produce complete microbial genome sequencing data, and closure and annotation can usually be accomplished in a matter of a few months. The cost for generating a closed microbial genome sequence with the shotgun approach has been reduced by an order of magnitude from what it was in 1995 to approximately 8 to 9¢/bp today in large centers that routinely handle large numbers of projects.

Given the advances in sequencing technologies, we were dismayed when the Department of Energy changed the strategy for its microbial genomic sequencing program in 1998 to one in which only high-coverage draft sequences for organisms of interest would be generated by the Joint Genome Institute. The rationale for such a change was that this would allow more organisms to be sampled because of the cost savings that would come from not taking each project to completion. While this strategy does achieve a cost savings, today it is only approximately 50%, and this comes at a cost in terms of the quality and utility of the finished product.

A complete genome sequence represents a finished product in which the order and accuracy of every base pair have been verified. In contrast, a draft sequence, even one of high coverage, represents a collection of contigs of various sizes, with unknown order and orientation, that contain sequencing errors and possible misassemblies. As stated by Selkov et al. in a 2000 paper on a draft sequence of *Thiobacillus ferrooxidans*, "It is clear that such sequencing. . .produces more errors than complete genome sequencing. . . . The current error rate is estimated to be 1 per 1,000 to 2,000 base pairs vs. 1 in 10,000 base pairs for complete sequencing" (10). In fact, the difference is much greater; recent studies show the error rate for completed microbial genomes to be closer to 1 in 100,000 (3). Another problem associated with draft sequence data is library contamination with DNA from foreign sources that can represent 5 to 10% of the total number of sequence reads for libraries prepared from DNA isolated from endosymbiotic and parasitic microbes that must be grown in animal cells. Until a genome project has been closed, it is often difficult to identify contaminating sequences, and these can confound subsequent comparative and functional genomics studies.

A retrospective analysis of 17 microbial genome sequences completed at TIGR during the past few years also revealed that when these genome projects entered closure, the extent of genome completion and the accuracy of assembly varied significantly (I. Paulsen, unpublished data). For example, at eightfold sequence coverage, the *Thermotoga maritima* genome was represented by 98 contigs (>1 kb in size) and was missing only 26 genes (~1.5% of the total) in the final annotation (7). This contrasts with the *Streptococcus pneumoniae* genome of similar size, whose initial assembly contained 265 contigs and was missing 115 genes (~6% of the total) (11). This difference likely reflects the fact that the genomes of some microbes (gram-positive organisms, for example) are not well represented in random DNA libraries. Some of the most interesting biology may be encoded in the missing genes of each organism. Due to the larger percentage of repetitive DNA in the *S. pneumoniae* genome, many of the initial contigs contained misassemblies that were revealed only during genome closure. Currently there is no method for assigning quality values to

* Corresponding author. Mailing address: The Institute for Genomic Research, 9712 Medical Center Dr., Rockville, MD 20850. Phone: (301) 838-3504. Fax: (301) 838-0209. E-mail: cmfraser@tigr.org.

genome assemblies as there is for DNA sequence reads, and this makes it difficult for anybody wishing to make use of preliminary data to know how reliable they really are. As a result, considerable additional work may be required to make full use of draft sequence data. Any such ad hoc attempts to improve draft sequences by additional sequencing or to close gaps in a draft project are inefficient and expensive and rapidly negate any initial cost savings.

While we agree that a draft sequence can be of tremendous benefit to interested investigators and acknowledge that every completed project generates draft sequence data as part of the process, there are several reasons why we believe that complete genome sequence should be the standard whenever possible, particularly with microbial genome projects.

FUNCTIONAL GENOMICS STUDIES DEMAND AN ERROR-FREE GENOME SEQUENCE AS A STARTING POINT

Regardless of how one defines functional genomics, all downstream work beginning with genome annotation is greatly facilitated by a complete, high-quality DNA sequence. This is true whether we are talking about defining gene coordinates in a genome, identifying paralogous gene families, or designing PCR primers for microarray analysis. Robust annotation of any genome sequence will ultimately require experimental work that will proceed more quickly and economically with a complete genome sequence as a starting point.

AVAILABILITY OF DATA ON GENOME ORGANIZATION PROVIDES BIOLOGICAL INSIGHTS

The results from a number of completed genome projects have demonstrated that information on overall genome organization can provide biological insights. As an example, assembly and closure of the *Borrelia burgdorferi* genome revealed the presence of several novel linear and circular plasmids that could not be distinguished in pulsed-field gel electrophoresis (1, 5). These plasmids contained novel genes encoding a number of lipoproteins that may be involved in virulence and infectivity. Another example is the segregation of genes with related functions among the four elements of the *Deinococcus radiodurans* genome (12). While the relevance of this observation is not yet entirely clear, it has suggested that one or more of the megaplasmids in *D. radiodurans* may be differentially regulated in response to stress. A third example is the presence of a second chromosome in *Vibrio cholerae* that appears to have been acquired as a separate genome element in the history of this species (6). With a draft sequence alone, these observations and the follow-up experiments that they suggest would not have been possible.

COMPARATIVE GENOMICS IS MEANINGFUL ONLY IN TERMS OF COMPLETE GENOME SEQUENCES

Work from many laboratories over the past several years has confirmed that there is much to be gleaned from genome sequence data beyond the identification of predicted coding sequences, in particular, the study of the forces that have shaped microbial genome evolution. When working with draft sequences alone it is impossible to know what regions of a given genome are not represented in any given data set. The absence of a gene from draft sequence data cannot be taken as evidence of its absence. Information on the presence or absence of genes is necessary to infer certain events in genome evolution, such as gene duplication, gene loss, and lateral gene transfer. If we had had only a draft sequence of the *T. maritima* genome, it is quite possible that the novel insights on lateral gene transfer between archaea and *Thermotoga* would have been missed because we would not have been able to identify the large regions in the *T. maritima* genome that differed in nucleotide composition from the remainder of the genome sequence and were presumably acquired by lateral gene transfer (7). Comparison of the completed genome sequences of *V. cholerae*, *S. pneumoniae*, and *Mycobacterium tuberculosis* to those of the closely related species *Escherichia coli*, *Streptococcus pyogenes*, and *Mycobacterium leprae*, respectively, have identified an unusual and previously unobserved feature of bacterial genome structure, termed X alignments, that reflects symmetric inversions around the replication origin and terminus (2). The finding of these X alignments between many pairs of species suggests that chromosomal inversions around the origin are a common feature of bacterial genome evolution. The detection of this relationship within the genomes of closely related organisms would not have been possible without complete genome data for all species examined.

MICROBIAL FORENSICS REQUIRES AT LEAST ONE COMPLETE REFERENCE GENOME SEQUENCE

Whole-genome sequencing represents the most powerful approach to identification of genomic diversity among closely related strains or isolates. Scanning whole genomes to detect genetic differences has the advantage that there is no inherent bias, in contrast to sampling methods such as multilocus sequence typing. A recent comparative study at TIGR of two Ames isolates of *Bacillus anthracis* revealed that polymorphic loci that distinguish between strains can be extracted from whole-genome sequence data (8). Such intergenome comparisons are greatly facilitated if at least one of the genomes is completely finished to a high degree of accuracy, rather than in the multiple unordered assemblies typical of a draft project. Prior to the anthrax letter attacks of the fall of 2001, the need for a robust program in microbial forensics was not appreciated. However, we now realize that we must work towards developing a comprehensive genotyping database of important pathogens that will allow investigators to quickly pinpoint the isolate that is most closely related to a natural or deliberately released outbreak strain, which will greatly accelerate investigations and may also be a deterrent to future attacks.

A COMPLETE GENOME SEQUENCE IS A PERMANENT, VALUABLE, SCIENTIFIC RESOURCE

If an organism is sufficiently important to study in the first place, then a complete, closed genome sequence of at least one strain provides the basis for decades, perhaps centuries, of future investigations. The complete and correct sequence represents a permanent snapshot of one moment in evolutionary

history, one that will always remain accurate even though the organism will continue to evolve.

Despite all of the arguments in favor of complete microbial genome sequences, there certainly are situations where draft sequence data can be useful, for example, as a means of surveying species and metabolic diversity in communities of microbes that cannot be grown in culture. Draft sequence data can also be useful in comparative studies when a complete genome sequence for a closely related strain or species exists and can be used as a scaffold to order and orient contigs. While it is true that the generation of a draft sequence for an organism of interest does not preclude its completion at some point in the future when closure costs may be lower, there is a considerable overhead that comes from taking such an approach—the loss of efficiency that comes with doing a project over time, the false leads that can come from trying to work with imperfect data, and the limitations of working with incomplete data sets. The only way to continue to drive the costs of genome closure down even further is to continue to fund projects to take genome sequences to completion. At the end of the day, you get what you pay for in terms of microbial genome sequencing projects. When one considers how much there still is to learn about the diversity of microbial life on our planet, the investment in complete microbial genome sequencing is some of the best money that will ever be spent.

REFERENCES

1. **Casjens, S., W. Huang, N. Palmer, R. van Vugt, B. Stevenson, P. Rosa, R. Lathigra, G. Sutton, J. Peterson, R. Dodson, E. Hickey, M. Gwinn, O. White, and C. Fraser.** 2000. A genome in flux: the twelve linear and nine circular extrachromosomal DNAs of an infectious isolate of the Lyme disease spirochete *Borrelia burgdorferi*. Mol. Microbiol. **35:**490–516.
2. **Eisen, J. A., J. F. Heidelberg, O. White, and S. L. Salzberg.** 2000. Evidence for symmetric chromosomal inversions around the replication origin in bacteria. Genome Biol. **1:**RESEARCH0011.1–RESEARCH0011.9. [Online.]
3. **Fleischmann, R. D.** 2001. Single nucleotide polymorphisms in *Mycobacterium tuberculosis* structural genes. Emerg. Infect. Dis. **7:**487–488.
4. **Fleischmann, R. D., M. D. Adams, O. White, R. A. Clayton, E. F. Kirkness, A. R. Kerlavage, C. J. Bult, J.-F. Tomb, B. A. Dougherty, J. M. Merrick, K. McKenney, G. Sutton, W. FitzHugh, C. Fields, J. D. Gocayne, J. Scott, R. Shirley, L.-I. Liu, A. Glodek, J. M. Kelley, J. F. Weidman, C. A. Phillips, T. Spriggs, E. Hedblom, M. D. Cotton, T. R. Utterback, M. C. Hanna, D. T. Nguyen, D. M. Saudek, R. C. Brandon, L. D. Fine, J. L. Fritchman, J. L. Fuhrmann, N. S. M. Geoghagen, C. L. Gnehm, L. A. McDonald, K. V. Small, C. M. Fraser, H. O. Smith, and J. C. Venter.** 1995. Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. Science **269:**496–512.
5. **Fraser, C. M., S. Casjens, W. M. Huang,. G. G. Sutton, R. Clayton, R. Lathigra, O. White, K. A. Ketchum, R. Dodson, E. K. Hickey, M. Gwinn, B. Dougherty, J. F. Tomb, R. D. Fleischmann, D. Richardson, J. Peterson, A. R. Kerlavage, J. Quackenbush, S. Salzberg, M. Hanson, R. van Vugt, N. Palmer, M. D. Adams, J. Gocayne, J. Weidman, T. Utterback, L. Watthey, L. McDonald, P. Artiach, C. Bowman, S. Garland, C. Fujii, M. D. Cotton, K. Horst, K. Roberts, B. Hatch, H. O. Smith, and J. C. Venter.** 1997. Genomic sequence of a Lyme disease spirochete, *Borrelia burgdorferi*. Nature **390:**580–586.
6. **Heidelberg, J. F., J. A. Eisen, W. C. Nelson, R. A. Clayton, M. L. Gwinn, R. J. Dodson, D. H. Haft, E. K. Hickey, J. D. Peterson, L. Umayam, S. R. Gill, K. E. Nelson, T. D. Read, H. Tettelin, D. Richardson, M. D. Ermolaeva, J. Vamathevan, S. Bass, H. Qin, I. Dragoi, P. Sellars, H. O. Smith, R. R. Colwell, J. J. Mekalanos, J. C. Venter, and C. M. Fraser.** 2000. DNA sequence of both chromosomes of the cholera pathogen *Vibrio cholerae*. Nature **406:**477–483.
7. **Nelson, K. E., R. A. Clayton, S. R. Gill, M. L. Gwinn, R. J. Dodson, D. H. Haft, E. K. Hickey, J. D. Peterson, W. C. Nelson, K. A. Ketchum, L. McDonald, T. R. Utterback, J. A. Malek, K. D. Linher, M. M. Garrett, A. M. Stewart, M. D. Cotton, M. S. Pratt, C. A. Phillips, D. Richardson, J. Heidelberg, G. G. Sutton, R. D. Fleischmann, J. A. Eisen, O. White, S. L. Salzberg, H. O. Smith, J. C. Venter, and C. M. Fraser.** 1999. Evidence for lateral gene transfer between Archaea and Bacteria from genome sequence of *Thermotoga maritima*. Nature **399:**323–329.
8. **Read, T. D., S. L. Salzberg, M. Pop, M. Shumway, L. Umayam, L. Jiang, E. Holtzapple, J. D. Busch, K. L. Smith, J. M. Schupp, D. Solomon, P. Keim, and C. M. Fraser.** 2002. Comparative genome sequencing for discovery of novel polymorphisms in *Bacillus anthracis*. Science **296:**2028–2033.
9. **Sanger, F., A. R. Coulson, G. F. Hong, D. F. Hill, and G. B. Peterson.** 1982. Nucleotide sequence of bacteriophage lambda DNA. J. Mol. Biol. **162:**729–773.
10. **Selkov, E., R. Overbeek, Y. Kogan, L. Chu, V. Vonstein, D. Holmes, S. Silver, R. Haselkorn, and M. Fonstein.** 2000. Functional analysis of gapped microbial genomes: amino acid metabolism of *Thiobacillus ferrooxidans*. Proc. Natl. Acad. Sci. USA **97:**3509–3514.
11. **Tettelin, H., K. E. Nelson, I. T. Paulsen, J. A. Eisen, T. D. Read, S. Peterson, J. Heidelbert, R. T. DeBoy, D. H. Haft, R. J. Dodson, A. S. Durkin, M. Gwinn, J. F. Kolonay, W. C. Nelson, J. D. Peterson, L. A. Umayam, O. White, S. L. Salzberg, M. R. Lewis, D. Radune, E. Holtzapple, H. Khouri, A. M. Wolf, T. R. Utterback, C. L. Hansen, L. A. McDonald, R. V. Feldblyum, S. Angiuoli, T. Dickinson, E. K. Hickey, I. E. Holt, B. J. Loftus, F. Yang, H. O. Smith, J. C. Venter, B. A. Dougherty, D. A. Morrison, S. K. Hollingshead, and C. M. Fraser.** 2001. Complete genome sequence of a virulent isolate of *Streptococcus pneumoniae*. Science **293:**498–506.
12. **White, O., J. A. Eisen, J. F. Heidelberg, E. K. Hickey, J. D. Peterson, R. J. Dodson, D. H. Haft M. L. Gwinn, W. C. Nelson, D. L. Richardson, K. S. Moffat, H. Qin, L. Jiang, W. Pamphile, M. Crosby, M. Shen J. J. Vamathevan, P. Lam, L. McDonald, T. Utterback, C. Zalewski, K. S. Makarova, L. Aravind, M. J. Daly, K. W. Minton, R. D. Fleischmann, K. A. Ketchum, K. E. Nelson, S. Salzberg, H. O. Smith, J. C. Venter, and C. M. Fraser.** 1999. Genome sequence of the radioresistant bacterium *Deinococcus radiodurans* R1. Science **286:**1571–1577.

Dialog

*In the article above, we present arguments in favor of complete genome sequencing as an important research tool. In the following article by Branscomb and Predki (J. Bacteriol. **184:**6406–6409, 2002) about draft microbial genome sequence data, the authors raise the notion that "delayed, third-party, or targeted finishing can be made to work very efficiently as a second step to draft sequencing." We believe that it is too early to make such conclusions based on what has been accomplished to date. In fact, this contradicts another statement made by Branscomb and Predki that "it may be substantially more expensive on average to finish draft sequence data later, . . .than to do so at the start and in the same laboratory." Given that cost savings is the primary reason for carrying out draft sequencing, we still strongly believe that except under unusual circumstances, such cost savings will be minimal. However, we agree with Branscomb and Predki on many other points, in particular the hidden costs that derive from (i) errors and imperfections in the data that may be misleading, (ii) the additional costs that come from uncoupling the finishing of draft sequence from its generation, and (iii) the risk that many of these projects will never be finished because of shifts in research priorities. We also agree that there is value in generating draft sequence data in some circumstances, but we maintain that complete genome data should be the target whenever possible.*