monthly thereafter for detailed clinical history, complete physical examination (including skin examination for the presence of GvHD), and general laboratory evaluation [including complete blood count and differential, urinalysis, blood urea nitrogen creatinine, bilirubin, aspartate transaminase (AST), alanine transferase (ALT), alkaline phosphatase, Na, K, Cl, albumin, total protein, glucose, and radiographic evaluation] as indicated. Disease status was assessed through examination of marrow aspirates and biopsy, cytogenetic analysis, and molecular analysis.

21. C. Bonini and S. Rossini, data not shown.
22. M. Ponzoni and L. Ruggieri, data not shown.
23. If grade II GvHD or higher occurred, patients who had previously received transduced donor lymphocytes were treated with ganciclovir (10 mg/kg per day) for 7 days or less if all the GvHD signs and symptoms regressed. Patients who had previously received infusions of both transduced and untransduced donor lymphocytes were initially treated with ganciclovir at 10 mg/kg per day, to down-regulate GvHD.
24. H. J. Kolb, personal communication.
25. C. Traversari *et al.*, in preparation.
26. G. B. Vogelsang and A. D. Hess, *Blood* **84**, 2061 (1994); L. M. Faber, S. A. van Luxemburg Heijs, W. F. Veenhof, R. Willemze, J. H. Falkenburg, *ibid.* **86**, 2821 (1995).
27. S. A. Giralt and R. E. Champlin, *ibid.* **84**, 3603 (1994).
28. H. Waldman, S. Cobbold, G. Hale, *Curr. Opin. Immunol.* **6**, 777 (1994).
29. T. Friedmann, *Science* **244**, 1275 (1989); K. W. Culver, W. F. Anderson, R. M. Blaese, *Hum. Gen. Ther.* **2**, 107 (1991); G. Ferrari *et al.*, *Science* **251**, 1363 (1991); W. F. Anderson, *ibid.* **256**, 808 (1992); R. C. Mulligan, *ibid.* **260**, 926 (1993); C. Bordignon *et al.*, *ibid.* **270**, 470 (1995); E. Marshall, *ibid.*, p. 1751.
30. Y. C. Cheng *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **80**, 2767 (1983).
31. P. d. C. d. M. Comitato Nazionale per la Bioetica, Dipartimento per l'informazione e l'editoria (Società e Istituzioni publisher, 1991), pp. 1–48.

32. We thank N. Nobili, D. Maggioni, L. Parma, and G. Torriani for technical assistance; F. Candotti, H. J. Kolb, and P. Panina for help and support; and the nurses and clinical staff of the Gene Therapy and Bone Marrow Transplantation Unit. Supported by grants from Telethon, the Italian Association for Cancer Research (AIRC), the European Community (Bio 4-CT 95-0284), and Boehringer Mannheim (Penzberg, Germany).

16 December 1996; accepted 30 April 1997

Dealing with Database Explosion: A Cautionary Note

Carol J. Bult *et al.* (*1*) report the first entire archea genome sequence of *Methanococcus jannaschii* (Mja). Because the initial gene assignments were conservative (*1*, *2*), we anticipated that much interesting biological information would be missing. We searched the database for additional open reading frames (ORFs), and found 15 ORFs: four within intergenic regions (M1 through M4, Table 1); five overlapping with previously identified ORFs (*1*, *2*) but that read off in a different frame (M5 through M9, Table 1); and six that are extended or truncated as a result of potential frameshifts (M10 through M15, Table 2).

Table 1. New ORFs in *M. jannaschii* (Mja) identified on the basis of similarity. ORFs were identified after purging out protein coding regions reported for the organism (*1*) and searched using BLASTX against the combined SwissProt+PIR+Genbank translations database through the NCBI Network BLAST server using a score cutoff of 60, as described previously (*6*). Corresponding matching protein, matching species–*Methanococcus vannielii* (Mva), *Bacillus subtilis* (Bsu), *Haemophilus influenzae* (Hin)—5′ start position, + or − strand, length of the ORF in amino acids (AA), 5′ to 3′ flanking ORFs, and the Poisson probability estimates are provided for each ORF. Other details available at www.golgi.harvard.edu/bhatia/neworfs/mja/table1.html

ORF	Matching protein	Matching species	Start 5′	Length (AA)	Flanking ORFs		*p*
					5′	3′	
M1	30S Ribosomal protein S14	Mva	415652+	55	469	470	10^{-19}
M2	Yqgp protein	Bsu	540515+	190	610	611	10^{-9}
M3	Amido phosphoribosyl transferase	Mja	1301085−	362	1352	1351	10^{-5}
M4	Unknown	Mja	1230530−	255	1283	282	10^{-18}
M5	Asparagine synthetase	Bsu	994621+	318	1055	1056	10^{-26}
M6	Modification methylase	Mja	1153501−	81	1208	1206	10^{-12}
M7	Modification methylase HINCII	Hin	1277783+	286	1327	1328	10^{-35}
M8	Helicase	Bsu	1548365−	182	1573	1572	10^{-6}
M9	Unknown	Mja	1329000+	58	1380	1381	10^{-12}

Table 2. Identification of potential frameshift(s) by similarity. Highly significant BLAST matches, of similar genes in alternative coding frames, were classified as frameshifts, manually assembled, and confirmed. Effect of the frameshift (extension or truncation) and length of the ORF as a result of the frameshift are also provided. M10 through M14 have suffered a single frameshift event, while M15 has apparently undergone a second frameshift. Other details available at www.golgi.harvard.edu/bhatia/neworfs/mja/table2.html

ORF	Matching protein	Matching species	Start 5′	Length (AA)	Frameshift		*p*
					Effect	Length (AA)	
M10	Restriction modification enzyme subunit M1	Mja	128577−	359	extension	583	10^{-93}
M11	Transposase	Mja	276289+	91	truncation	38	10^{-43}
M12	Polyferredoxin	Mja	457630−	410	extension	567	10^{-40}
M13	Unknown	Mja	14344−	72	truncation	16	10^{-35}
M14	Unknown	Mja	202169+	177	truncation	50	10^{-8}
M15	Unknown	Mja	809431+	32	extension	131	10^{-20}

Although the potential frameshifts we describe might be bona fide, it cannot be ruled out that they represent actual sequencing artifacts. Erroneous sequences in public databases are a substantial problem and have been estimated to be in the range of 0.37 to 2.9 errors per 1000 nucleotides (*3*), making data interpretation sometimes difficult. This is especially true, for example, in studies that utilize protein and DNA sequence information to estimate evolutionary distances (*4*). It is not known how the error rate in this study (*1*) compares with error rates in the database, but a previous study suggests that error rates generally vary between 1 in 5000 to 1 in 10,000 nucleotides (*5*).

The issue of sequencing artifacts is important and is expected to be a continuing problem in the future, considering the heightened surge of genome sequencing projects from model organisms, as well as from the human genome sequencing initiative.

Umesh Bhatia
Department of Molecular and Cellular Biology,
Harvard University,
Cambridge, MA 02138, USA
E-mail: bhatia@nucleus.harvard.edu
Keith Robison
Millennium Pharmaceuticals, Inc.,
640 Memorial Drive,
Cambridge, MA 02139, USA
Walter Gilbert
Department of Molecular and Cellular Biology,
Harvard University

REFERENCES

1. C. J. Bult *et al.*, *Science* **273**, 1058 (1996).
2. N. C. Kyrpides *et al.*, *Microb. & Comp. Genomics* **1**, 329 (1996).
3. S. A. Krawetz, *Nucleic Acids Res.* **17**, 3951 (1989); J. Claverie, *J. Mol. Biol.* **234**, 1140 (1993).
4. W. H. Li *et al.*, *Genetics* **129**, 513 (1991).
5. R. D. Fleischmann *et al.*, *Science* **269**, 496 (1995).
6. K. Robison *et al.*, *Nature Genet.* **7**, 205 (1994); K. Robison *et al.*, *Science* **271**, 1302 (1996).

24 February 1997; accepted 23 April 1997

Response: Bhatia *et al.* express concerns about erroneous sequences in public databases, which make the interpretation of sequence data sometimes difficult. We share these concerns because faulty entries in public databases, especially sequence annotations, often complicate our research efforts. Therefore, we dedicate considerable resources to maintain a curated in-house database and to carefully check the sequences and annotations provided by us to the public. The challenge is to find a suitable compromise between the quick release of newly sequenced genomes and responsible sequence quality and annotation. We estimate our error rate at the time of release to be 1 base in 5000 to 10,000 (*1*), which is about the quality requested for the Human Genome Project. For the 1.7-Mbp M. *jannaschii* genome (*1*), this would account for about 250 putative errors, which would mainly result in frameshifts in ORFs that as yet have no recognizable homologs in any database. Bhatia *et al.* specify 15 regions in this genome where they suspect ORFs or frameshift problems resulting from sequencing artifacts. We encourage the input of the scientific community in ongoing efforts to further elucidate the wealth of biological information still hidden in this genome; however, without access to the original electropherograms that were used to generate the final genome sequence data, it is not always possible to definitively determine whether a presumed frameshift reflects an error in the DNA sequence or not (Table 1). For example, ORF M11 in table 2 of the comment suggests that we truncated a transposase gene by a frameshift, but this "ORF" is a vestigial gene that is missing a significant portion of the central part of its homologues. The nucleotide necessary for a correction of the frameshift, A-276,294, is absent in all 12 sequences covering this area of the genome.

Table 1. Comments and additional biological information for some of the ORFs criticized by Bhatia *et al.* M1 through M4, ORFs in intergenic regions missed in the initial publication; M5 through M9, "additional" ORFs overlapping with previously identified ORFs—all are frameshifts or vestigial genes. All the frameshifts mentioned in this table have been confirmed by extensive reediting or resequencing of the problem areas with alternative sequencing chemistry. A complete list of the corrected ORFs will shortly become available on our *M. jannaschii* genome web page at www.tigr.org/tdb/mdb/mjdb/mjdb.html. ORF, open reading frame; aa, amino acid.

ORF	Information updates on some of the criticized ORFs
M1 = MJ0469.1	53 aa long ribosomal protein S14P; available in SWISS-PROT since October 1996.
M2 = MJ0610.1	Member of uncharacterized family of membrane proteins of unknown function.
M3 = MJ1351.1	Functionally uncharacterized third member of type-2 glutamine amidotransferase family.
M5 = MJ1056	Frameshift at position C995,616 extends the asparagine synthetase gene by 342 codons.
M8 = MJ1574	Multiple frameshifts extend the eIF-4A family member by 197 aa at the COOH-terminus.
M11 = MJ0293.1	Vestigial transposase of ISAMJ1-type; none of the 12 sequences available for this area supports the presence of an additional A at 276,294 to correct the criticized frameshift.
M12 = MJ0514.2	MJ0514.2, MJ0514.1, and MJ0514 are a set of polyferredoxin genes; the MJ0514.2/MJ0514.1 junction area is covered by 16 sequences, none of which supports the supposed frameshift; MJ0514.1 is preceded by a typical Shine-Darlgarno sequence.

No automated computer system will discover some of the treasures (and some of the errors) still hidden in the genome of M. *jannaschii*. We are therefore grateful to colleagues who, after the release of the M. *jannaschii* genome sequence, contacted us to provide their biological, biochemical, and genetic experience and expertise, which has resulted in quick updates and corrections of our freely accessible database at www.tigr.org.

Hans-Peter Klenk
E-mail: hpklenk@tigr.org
Owen White
E-mail: owhite@tigr.org
J. Craig Venter
The Institute for Genomic Research, TIGR,
9712 Medical Center Drive,
Rockville, MD 20850, USA

REFERENCES

1. R. D. Fleischmann *et al.*, *Science* **269**, 496 (1995); C. M. Fraser *et al.*, *ibid.* **270**, 397 (1995); C. J. Bult *et al.*, *ibid.* **273**, 1058 (1996).

22 April 1997; accepted 3 April 1997

Discover a new sequence.

Visit the SCIENCE Online Web site and you just may find the key piece of information you need for your research. The fully searchable database of research abstracts and news summaries allows you to look through current and back issues of SCIENCE on the World Wide Web. Tap into the sequence below and see SCIENCE Online for yourself.

www.sciencemag.org

SCIENCE