

The Complete Nucleic Acid Sequence of the Adenovirus Type 5 Reference Material (ARM) Genome

BY BARRY J. SUGARMAN,
BETH M. HUTCHINS, DIANE
L. MCALLISTER, FEI LU, and
KENNETH B. THOMAS

The Adenovirus Reference Material Working Group (ARMWG) oversaw development of an adenovirus reference material (ARM) with the intent to provide a way to standardize assay measurements from different laboratories.^{1,2} The ARM, which was manufactured in stages by various organizations including Canji (San Diego, CA) and Introgen Therapeutics (Houston, TX), is available from American Type Culture Collection (Manassas, VA).³ Upon completion of its manufacture, the characterization phase primarily defined viral particle concentration as well as infectious titer for this product.⁴ However, many other concurrent characterization studies were conducted including an assessment of vector purity (e.g., host cell DNA, host cell protein, reversed-phase HPLC), a short-term field use and shipping stability study, and a long-term stability study.^{5,6} Also included in these studies was a coordinated effort to determine the complete DNA sequence of the ARM vector genome.

FDA has recommended that the ARM be used in conjunction with assay

development, assay validation, and product characterization.⁷ FDA's Center for Biologics Evaluation and Research (CBER) revised their requirements for pre-Phase I sequence analysis of viral vectors used for gene transfer trials such that it is now necessary to sequence and analyze the entire genome of viral vectors ≤ 40 kilobases (kb) in length (e.g., adenovirus, adeno-associated virus, and retro/lentivirus), but for vectors > 40 kb sequence only those regions of special interest (for example, inserted cDNA and transcriptional domains regulating expression of exogenous cDNAs).⁸ FDA has asked sponsors to compare the sequence of exogenous genes as well as the vector backbone to equivalent sources in public databases. As part of the characterization process for the ARM, DNA was isolated from purified ARM samples and the viral genome was sequenced, analyzed, and the information deposited in GenBank.

This article summarizes efforts related to the determination of the entire ARM genomic sequence. A column-purified wild-type adenovirus type 5 preparation was used as the source for viral DNA. Its purity and identity was confirmed prior to DNA sequence analysis. Sequencing results were compared to information in the National Center for Biotechnology Information (NCBI, a division of the National Institutes of Health) public database GenBank. A total of forty differences

were noted between the corresponding nucleic acid sequence in GenBank (accession number M73260) and the sequence derived for the ARM genome. Because this information will be available through public databases, the expectation is that it will be used by sponsors to compare their vector-specific DNA sequence data with the ARM sequence as part of the IND submission process.

Material and Methods

Participants

The following organizations participated in these studies: Canji Inc. (San Diego, CA) and SeqWright DNA Technology Services (Houston, TX).

Starting Material

Dr. Schrock (Introgen Therapeutics, Houston, TX) provided a 45-ml aliquot of the Adenovirus Reference Material final product to Canji for these studies. Unlike material currently available from the American Type Culture Collection (Manassas, VA), the starting material used herein was from a concentrated bulk sample collected before final dilution (part number 09-00159, C/N 001471) and fill of the four sub-lots. Subsequent analysis using anion-exchange chromatography indicated that the viral particle concentration was 3.5×10^{12} particles/ml.⁹

Barry J. Sugarman, Ph.D. (barry.sugarman@canji.com) is associate director of process sciences, Beth M. Hutchins, Ph.D. is director of process sciences, and Diane L. McAllister is a scientist II in the department of process sciences, Canji Inc., San Diego, CA; Fei Lu, M.D. is chief executive officer and Kenneth B. Thomas, Ph.D. is director of bioinformatics, SeqWright DNA Technology Services, Houston, TX.

DNA Purification

DNA was purified using the QIAamp Maxi Kit protocol (QIAGEN, Valencia, CA). Briefly, the capsid was removed using protease digestion at 62° C for 1 hour. Intact viral DNA was loaded onto

a spin column and bound to the resin in high salt, washed, and eluted in 10 mM Tris, 1 mM EDTA, pH 8.0. The concentration of the resulting DNA preparation was determined by measuring absorbance at 260 nm.

OR) in TAE for 40 minutes at ambient temperature.

DNA Sequencing

After the fragment array was observed to be consistent with the predicted array, approximately 100 µg of DNA was shipped to SeqWright (Houston, TX). Prior to sequencing, the DNA concentration was confirmed by electrophoresis using a pGEM DNA standard.

A shotgun library was constructed using sheared template DNA. Library construction was essentially performed using the double-adaptor method.¹⁰ DNA was sheared using a nebulizer, precipitated, and treated with the enzymes T4 and Klenow to generate blunt ends prior to adaptor ligation. Fragments were size-selected for the 2–4 kb range before insertion into the pUC18 vector.

Dye-terminator sequencing chemistry was used for sequencing random clones from the library using a kit purchased from Applied Biosystems, Inc. (Mountain View, CA). M13 forward and reverse sequencing primers were used to generate sequence from a total of 288 clones. Custom primers, 25 in total, were used to generate reads to fill in gaps in the random data.

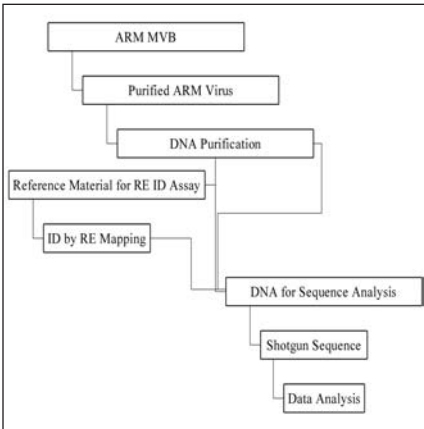


Figure 1. Outline of the manufacturing and testing scheme used to generate the ARM DNA sequence. The ARM Master Virus Bank was manufactured and characterized by Canji, Inc. The purified ARM material was manufactured and initially characterized by Introgen Therapeutics, Incorporated. A concentrated purified bulk drug substance ARM sample was used as the source for DNA purification. It was this material that was subsequently used for restriction enzyme (RE) analysis and DNA sequencing.

Identity Confirmation by Restriction Enzyme Analysis

Identity was checked to determine whether or not it was consistent with the fragment array predicted for human adenovirus type 5 DNA before it was shipped for sequencing. Restriction enzyme analysis was performed by digesting 600–700 ng ARM DNA per reaction with the following enzymes: *Acc65 I*, *Apa I*, *Bgl II*, *Hind III*, *Nco I*, *Not I*, *Nsi I*, *Sma I*, *Ssp I*, and *Xho I* (New England Biolabs, Beverly, MA). Viral DNA was digested with these endonucleases for 2 hours at 37° C. Restriction fragments were resolved by electrophoresis in a 15 cm x 25 cm, 1% (w/v) SeaKem ME agarose gel (BioWhittaker Molecular Biology Applications, Rockland, ME) containing 40 mM Tris-acetate, 1 mM EDTA (TAE) (Sigma Chemicals, St. Louis, MO) at 60 volts for 18 hours. Bands were visualized by staining with SYBR Green I (Molecular Probes, Eugene,

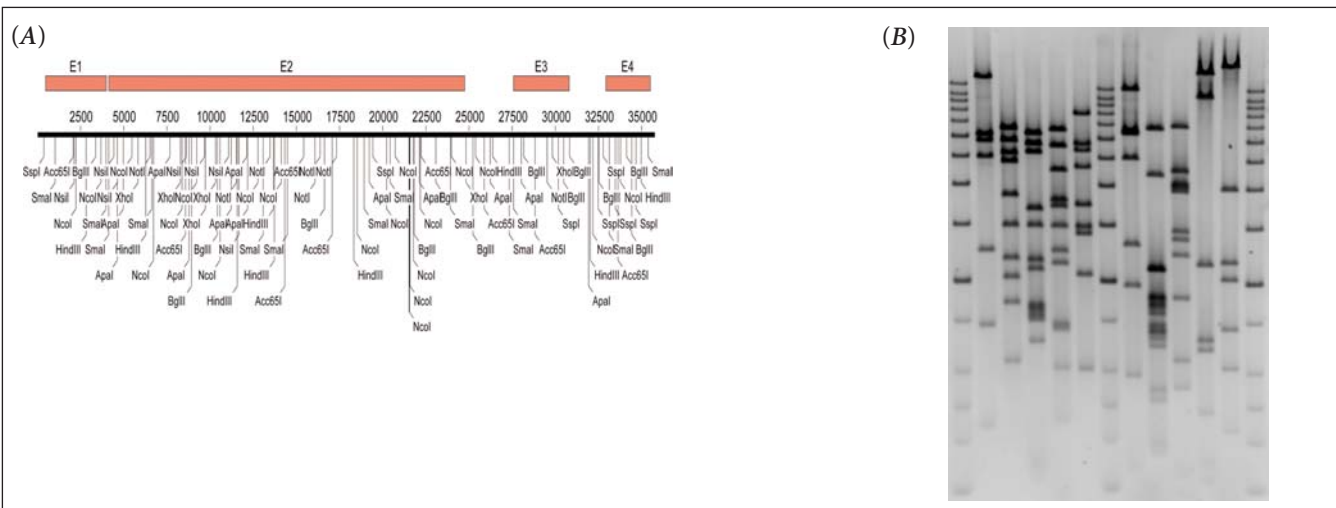


Figure 2. Confirmation that the ARM DNA was wild-type Adenovirus type 5. (A) Predicted locations of different restriction enzyme recognition sites within the wild-type Ad5 genome based upon GenBank accession number M73260. The restriction enzyme sites in the illustration include *Apa I*, *Acc65 I*, *Bgl II*, *Hind III*, *Nco I*, *Not I*, *Nsi I*, *Sma I*, *Ssp I*, and *Xho I*. Numbers denote their location within the vector genome. (B) Electrophoretic profile of DNA fragments from the ARM DNA digested with various restriction enzymes. Resulting fragments were resolved in a 1% (w/v) agarose gel and visualized using SYBR Green I. The contents of each lane are as follows (i.e., lane 1 is on the left): lanes 1, 7 and 13, DNA molecular weight standards (kb ladder, Stratagene); lane 2, ARM (*Xho I*); lane 3, ARM (*Acc65 I*); lane 4, ARM (*Bgl II*); lane 5, ARM (*Sma I*); lane 6, ARM (*Hind III*); lane 8, ARM (*Not I*); lane 9, ARM (*Nco I*); lane 10, ARM (*Apa I*); lane 11, ARM (*Ssp I*); lane 12, ARM (*Nsi I*). The size of each DNA molecular weight standard (from top to bottom) was as follows: 12000, 10000, 9000, 8000, 7000, 6000, 5000, 4000, 3000, 2000, 1500, 1000, 750, 500, and 250 base pairs.

Table 1. Summary of the restriction enzyme analyses of ARM DNA. Below are listed both the expected and observed fragment sizes. Size was determined using DNA molecular weight standards (250 to 12,000 base pairs) and AlphaEase analysis software (Alpha Innotech, San Leandro, CA).

Restriction Enzyme	Expected Size (bp)	Observed Size (bp)	Restriction Enzyme	Expected Size (bp)	Observed Size (bp)
Acc65 I	6485	6779	Hind III	8010	7934
	5753	6000		5665	5851
	5120	5291		5324	5358
	4811	4902		4597	4665
	3646	3750		3437	3501
	2949	3026		2937	2987
	2337	2366		2804	2836
	2052	2061		2081	2097
	1696	1717		1005	1029
	1086	1091		75	Not Observed
Apa I	7010	7149	Nco I	6775	7000
	4623	4735		4386	4484
	4099	4247		2206	2262
	4026	4022		2204	2162
	3925	3946		2192	1935
	2910	2973		1803	1811
	2731	2789		1785	Doublet
	2447	2455		1715	1744
	1803	1794		1670	1648
	1110	1115		1604	1587
	873	890		1463	1500
	306	325		1398	1414
	72	Not Observed		1383	1359
	Bgl II	6182		6357	Ssp I
5586		5851	10797	10830	
5178		5358	2298	2282	
3328		3383	1301	1319	
2943		3000	1221	1207	
2334		2346	682	723	
2151		2190	341	388	
1672		1717	271	293	
1625		1641	14502	>12000	
1548		1555	6144	6276	
1497		1500	5788	6000	
1268		1285	4995	5032	
351		365	2466	2514	
272		289	1445	1457	
Not I	12200	>12000	Xho I	595	608
	6503	6887		6480	6522
	6425	6560		5563	5705
	5001	5000		4456	4619
	2589	2689		3540	3718
	1931	2000		3386	3472
	960	995		2933	3000
	326	343		2463	2514
Nsi I	24308	>12000	Sma I	2262	2267
	3984	4000		1455	1500
	2344	2324		1398	1416
	2075	2028		1009	1039
	1431	1428		580	678
	1042	1040		230	Not Observed
	573	574		180	Not Observed
	178	Not Observed			

An Applied Biosystems automated fluorescent DNA sequencer (ABI Prism 3700) was used to produce the raw data with average gel reads between 400 and 700 bases. Two samples were also run on an ABI Prism 377 DNA sequencer with an ABI dye terminator mix designed to resolve problems caused by

high template GC content. Sequences were edited and aligned using the Sequencher 3.0 assembly and editing software (Gene Codes Corporation, Ann Arbor, MI). Contig alignment, editing of sequencing data, and production of the consensus sequence with double-stranded coverage and > 99.9%

accuracy were generated by SeqWright (Houston, TX). SeqWright stored all sequencing data and reports under project number S16549.CNJ.

Results and Summary

An overview of the manufacturing and characterization steps preceding sequencing of the ARM genomic DNA is depicted in Figure 1. Prior to analyzing the purified ARM DNA by DNA sequencing, the identity was confirmed by restriction enzyme analysis using ten different enzymes. The expected distribution of these restriction enzyme recognition sites, which are located throughout the wild-type adenovirus 5 (wtAd5) genome based on the nucleic acid sequence in GenBank (accession number M73260), is depicted in Figure 2A. Because the observed restriction enzyme fragment array (Fig. 2B) was consistent with the hypothetical pattern (Table 1), we presumed that this DNA was in fact wtAd5. The only anomaly was that an Acc65 I recognition site predicted to reside at base pairs 11282–11287 did not exist. This was later confirmed by the DNA sequencing.

The consensus ARM DNA sequence was determined to be 35,934 base pairs in length (GenBank accession number AY339865). A detailed base-by-base analysis showed that the sequence of the ARM genome is similar but not identical to the Ad5 sequence in GenBank. These differences are enumerated in Table 2. A total of forty items — including substitutions, deletions, and insertions — were observed to differ between these sequences. Within this group, only one results in significant change, *i.e.* a premature termination of the E3 10.4K (RID α) protein. Changes were confirmed by reviewing individual electropherograms within each domain as well as by using multiple sequence overlaps. Two possible reasons for why so many differences were observed could be significant improvements in sequencing technologies subsequent to the publication of M73260 and/or that sequences from different Ad5 variants were combined to produce GenBank sequence file M73260.¹¹

Table 2. Comparison of the ARM Genomic DNA Sequence with Information in Public Databases

Item Number	Description of Change	Location in ARM Sequence	Location in GenBank DNA Sequence ¹	Additional Evidence / Comments	Modifications to Coding Domain
1	SUBSTITUTION. Residue is a "C" instead of a "G"	4952	4952	Residue is a "C" in Ad2WT (GenBank NC_001405, bp 4942)	Changes a HIS to a GLN residue at amino acid 162 of the E2B IVa2 protein CDS (exon 2)
2	SUBSTITUTION. Residue is an "A" instead of a "G"	8783	8783	Residue is an "A" in Ad2WT (NC_001405, bp 8773)	Changes a LEU to a PHE residue at amino acid 588 of the E2B pTP
3	SUBSTITUTION. Residue is a "C" instead of a "T"	11284	11284	Residue is a "C" in GenBank M73260 (11284)	Changes a TYR to a HIS residue at amino acid 79 of the L1 52,55K protein
4	INSERTION. Extra "A" residue in poly "A" stretch (i.e., 13 "A" versus 12 "A" in the GenBank M73260)	14086	Does not exist; would reside after bp 14085 and before bp 14086	Not present in either GenBank M73260; poly "A" stretch is only 12 residues in length in this sequence file	Sequence in question is not within a known coding domain or other known regulatory element
5	SUBSTITUTION. Residue is a "C" instead of a "G"	17388	17387	—	Changes a GLY to a ARG residue at amino acid 282 of the L2 pV protein
6	DELETION. Missing an "A" residue present in GenBank M73260	Does not exist; would reside after bp 18756 and before bp 18757	18756	Although an "A" residue is present in GenBank M73260, no "A" residue is present in ARM DNA sequence file	Sequence in question is not within a known coding domain; 3' to the terminal codon for L3 pVI (hexon-associated precursor)
7	SUBSTITUTION. Residue is an "A" instead of a "T"	19483	19483	—	No changes in the L3 pII (hexon) amino acid sequence
8	SUBSTITUTION. Residue is an "A" instead of a "T"	19513	19513	Is an "A" residue ARM DNA Sequence (19513) and a "T" in GenBank M73260 (19513)	No changes in the L3 pII (hexon) amino acid sequence
9	SUBSTITUTION. Residue is an "A" instead of a "G"	19657	19657	—	No changes in the L3 pII (hexon) amino acid sequence
10	SUBSTITUTION. Residue is a "G" instead of an "A"	19658	19658	—	Changes a THR to a ALA residue at amino acid 273 of the L3 pII (hexon) amino acid sequence
11	SUBSTITUTION. Residue is a "C" instead of a "T"	20378	20378	"T" in GenBank X02997 (1761)	No changes in the L3 pII (hexon) amino acid sequence
12	SUBSTITUTION. Residue is a "T" instead of a "C"	21163	21163	"C" in GenBank X02997 (2546)	No changes in the L3 pII (hexon) amino acid sequence
13	SUBSTITUTION. Residue is an "A" instead of a "G"	21630	21630	"G" in GenBank X02997 (3013)	Changes a ARG to a GLN residue at amino acid 930 of the L3 pII (hexon) amino acid sequence
14	SUBSTITUTION. Residue is a "T" instead of an "A"	25995	25995	—	No changes in the L4 100K protein (hexon assembly) amino acid sequence
15	DELETION. Missing a "G" residue present in GenBank M73260	Does not exist; would reside after bp 26740 and before bp 26741	26741	Although a "G" residue is present in GenBank M73260, no "G" residue is present in the ARM DNA sequence file	Sequence in question is not within a known coding domain or other known element
16	DELETION. Missing a "G" residue present in GenBank M73260	Does not exist; would reside after bp 26740 and before bp 26741	26742	Although a "G" residue is present in GenBank M73260, no "G" residue is present in the ARM DNA sequence file	—

¹GenBank Accession Number M73260

Table 2. Comparison of the ARM Genomic DNA Sequence with Information in Public Databases

Item Number	Description of Change	Location in ARM Sequence	Location in GenBank DNA Sequence ¹	Additional Evidence / Comments	Modifications to Coding Domain
17	DELETION. Missing a "C" residue present in GenBank M73260	Does not exist; would reside after bp 26740 and before bp 26741	26743	Although a "C" residue is present in GenBank M73260, no "C" residue is present in the ARM DNA sequence file	—
18	DELETION. Missing an "A" residue present in GenBank M73260	Does not exist; would reside after bp 26740 and before bp 26741	26744	Although an "A" residue is present in GenBank M73260, no "A" residue is present in the ARM DNA sequence file	—
19	DELETION. Missing a "G" residue present in GenBank M73260	Does not exist; would reside after bp 26740 and before bp 26741	26745	Although a "G" residue is present in GenBank M73260, no "G" residue is present in the ARM DNA sequence file	—
20	DELETION. Missing a "C" residue present in GenBank M73260	Does not exist; would reside after bp 26740 and before bp 26741	26746	Although a "C" residue is present in GenBank M73260, no "C" residue is present in the ARM DNA sequence file	—
21	SUBSTITUTION. Residue is a "T" instead of a "C"	27155	27161	—	5' to L4 pVIII protein (hexon associated precursor) CDS
22	SUBSTITUTION. Residue is an "A" instead of a "C"	27308	27314	—	No changes in the L4 pVIII (hexon associated precursor) amino acid sequence
23	SUBSTITUTION. Residue is a "C" instead of a "T"	27333	27339	"T" in GenBank X03002 (Ad5 E3 region, base 8)	No changes in the L4 pVIII (hexon associated precursor) amino acid sequence
24	SUBSTITUTION. Residue is a "C" instead of a "T"	27644	27650	"T" in GenBank X03002 (bp 319, Ad5 E3)	No changes in the L4 pVIII (hexon associated precursor) amino acid sequence
25	SUBSTITUTION. Residue is a "T" instead of a "C"	27645	27651	"C" in GenBank X03002 (bp 320, Ad5 E3)	Changes a PRO to a SER residue at amino acid 160 of the L4 pVIII (hexon associated precursor) amino acid sequence
26	SUBSTITUTION. Residue is a "C" instead of a "T"	28114	28120	"T" in GenBank X03002 (789)	Changes a LEU to a PRO residue at amino acid 88 of the E3 12.5K protein
27	INSERTION. Extra "T" residue in poly "T" stretch in ARM sequence (i.e., 7 "T" versus 6 "T" in GenBank M73260)	29824	Does not exist; would reside after bp 29829 and before bp 29830	Insertion of additional "T" residue differs from all other GenBank entries	Causes pre-mature termination of the E3 10.4K (i.e., RIDa) protein after 63 instead of 91 amino acids; first 15 amino acids of both protein sequences are identical
28	SUBSTITUTION. Residue is a "C" instead of an "A"	30296	30301	Matches the corresponding sequence in GenBank K02559	Changes a LYS to a ASN residue at amino acid 80 in the E3 14.5K protein
29	SUBSTITUTION. Residue is a "G" instead of a "C"	30297	30302	Matches the corresponding sequence in GenBank K02559	Changes a ARG to a ALA residue at amino acid 81 in the E3 14.5K protein
30	SUBSTITUTION. Residue is a "C" instead of a "G"	30298	30303	Matches the corresponding sequence in GenBank K02559	—
31	SUBSTITUTION. Residue is an "A" instead of a "C"	30398	30403	—	No changes in the E3 14.5K protein

Table 2. (cont.) Comparison of the ARM Genomic DNA Sequence with Information in Public Databases

Item Number	Description of Change	Location in ARM Sequence	Location in GenBank DNA Sequence ¹	Additional Evidence / Comments	Modifications to Coding Domain
32	SUBSTITUTION. Residue is a "C" instead of an "A"	30399	30404	—	Changes a SER to a PRO residue at amino acid 115 in the E3 14.5K protein
33	INSERTION. Extra "T" residue in poly "T" stretch within ARM sequence (i.e., 12 "T" versus 11 "T" in GenBank M73260)	34350	Does not exist; would reside after bp 34354 and before bp 34355	Not present in either GenBank_001406; poly "T" stretch is only 11 residues in length in this sequence file	At the the 3' end of E4 orf3; no effect on a coding domain
34	SUBSTITUTION. Residue is an "A" instead of a "T"	34856	34860	"T" in GenBank J01969 (505, Ad5 E4); "A" in GenBank D12587, Ad5 E4)	A MET residue at amino acid 84 of E4 orf2; no equivalent for M73260 E4 orf2 because of the insertion at bp 34936, it terminates prematurely after 67 versus 136 amino acids in the ARM sequence
35	DELETION. Missing a "T" residue present in GenBank M73260	Does not exist; would reside after 34931 and before 34932	34936	Although a "T" residue is present in GenBank M73260, no "T" residue is present in ARM, or GenBank D12587 (i.e., after 145 and 146) sequence files	The predicted sequence for the protein coded for by E4 orf 2 is identical through amino acid number 59 and begins to differ at amino acid 60
36	INSERTION. Extra "A" residue not present in the GenBank M73260	35317	Does not exist; would reside after bp 35321 and before bp 35322	Does not exist in GenBank J01969 (between bp 967 and 968); present in GenBank D12587 (532-533)	A VAL residue at amino acid 69 of E4 orf1; no equivalent for M73260 E4 orf1 because of the various insertion [item numbers 36-39]
37	INSERTION. Extra "C" not present in the GenBank M73260	35318	Does not exist; would reside after bp 35321 and before bp 35322	—	—
38	INSERTION. Extra "C" residue not present in the GenBank M73260	35508	Does not exist; would reside after bp 35510 and before bp 35511	Does not exist in GenBank J01969 (between bp 1155 and 1156); present in GenBank D12587 (723)	A VAL residue at amino acid 5 of E4 orf1; no equivalent for M73260 E4 orf1 because of the various insertion [item numbers 36-39]
39	INSERTION. Extra "A" residue not present in the GenBank M73260	35521	Does not exist; would reside after bp 35522 and before bp 35523	Does not exist in GenBank J01969 (between bp 1167 and 1168); present in GenBank D12587 (736)	A MET residue at amino acid 1 of E4 orf1; no equivalent for M73260 E4 orf1 because of various insertions [item numbers 36-39]
40	SUBSTITUTION. Residue is a "C" instead of an "A"	35772	35773	"A" in GenBank J01969 (1418)	—

Acknowledgements

We thank Dr. Stephen Chang for his support, advice, and critical reading of this manuscript; Susan Miller for determining the particle concentration of the initial purified vector preparation; and Introgen Therapeutics for providing the source material and documentation.

References

- Hutchins B, Sajjadi N, Seaver S, Shepherd A, Bauer SR, Simek S, Carson K, Aguilar-Cordova E. Working toward an adenoviral standard. *Molecular Therapy* 2002;2(6):532–534.
- Hutchins B. Development of a reference material for characterizing adenovirus vectors. *BioProcessing J* 2002;1(1):25–28.
- American Type Culture Collection (Manassas, VA), catalog no. VR-1516.
- Callahan JD. A statistical analysis of adenovirus material assay results. *BioProcessing J* 2002;1(3):43–47.
- Vellekamp G. A contaminant in the Adenovirus Reference Material. *BioProcessing J* 2002;1(3):57–61.
- Adadevoh K, Croyle M, Malarme D, Bonfils E, Bowe MA. A short-term field use and shipping stability study of a wild type Ad5 adenoviral reference material. *BioProcessing J* 2002;1(2):62–69.
- Simek S, Byrnes A, Bauer S. FDA perspectives on the use of the Adenovirus Reference Material. *BioProcessing J* 2002;1(3):40–42.
- McIntyre MC. Development of viral vectors for gene transfer trials. *Presentation at the 8th Annual Williamsburg BioProcessing Foundation Conference on Viral Vectors and Vaccines*; 2001 November 12–15; Lake Tahoe, NV.
- Shabram PW, Giroux DD, Goudreau AM, Gregory RJ, Horn MT, Huyghe BG, Lui X, Nunnally MH, Sugarman BJ, Sutjipto S. Analytical anion-exchange HPLC of recombinant Type-5 adenovirus particles. *Human Gene Therapy* 1997;8:453–465.
- Andersson B, Wentland MA, Ricafrente JY, Liu W, Gibbs RA. A "double adaptor" method for improved shotgun library construction. *Analytical Biochemistry* 1996;236:107–113.
- Chroboczek J, Bieber F, Jacrot B. The sequence of the genome of adenovirus Type 5 and its comparison with the genome of adenovirus Type 2. *Virology* 1992;186:280–285.