

# The Misidentification of Bacterial Genes as Human cDNAs: Was the Human D-1 Tumor Antigen Gene Acquired from Bacteria?

Eric P. Skaar and H. Steven Seifert\*

Department of Microbiology-Immunology, Northwestern University, Feinberg School of Medicine, 303 East Chicago Avenue, Chicago, Illinois 60611, USA

\*To whom correspondence and reprint requests should be addressed. Fax: (312) 503-1339. E-mail: [h-seifert@northwestern.edu](mailto:h-seifert@northwestern.edu).

---

**The initial analysis of the draft copy of the human genome sequence revealed the presence of several genes that were proposed to have been directly transferred from bacteria. We investigated the human D-1 antigen as a potential lateral transfer event. We report that although the human D-1 antigen seems to be an excellent candidate for lateral transfer, it is a contaminating bacterial sequence present in a human cDNA library that was included in the human genome analysis. Furthermore, several other genes present in the publicly available databases that were included in the analysis of the human genome are also likely contaminating bacterial sequences present in cDNA libraries.**

---

In the report by the International Human Genome Sequencing Consortium describing the first draft of the human genome, it was proposed that 113 genes were potentially horizontally transferred from bacteria into vertebrates [1,2]. Since this initial report, most of the cases have been discounted due to more rigorous analyses [3-5]. Of the approximately 50 genes that have not been discounted, several potential candidates remain. We investigated whether one of these genes, encoding the human D-1 antigen, is an example of a gene transferred directly from bacteria to humans.

The human D-1 antigen was identified from a cDNA expression library derived from the A375 melanoma cell line as a clone reactive to pooled sera from patients with melanoma [6]. The D-1 transcript was identified in multiple transformed cell lines, but was absent from primary cells [6]. The D-1 cDNA sequence is highly similar to a 5'-portion of the *recN* genes from many bacteria. Extensive BLAST analyses of all publicly available sequence databases showed that there are no known eukaryotic homologs to the human D-1 cDNA sequence. Thus D-1 is an excellent candidate for bacterial to vertebrate transfer.

Neighbor-joining analysis revealed that of the 28 *RecN* sequences scoring with the highest probability in a BLASTP search of the NCBI database, the closest homologs to the D-1 antigen are the *RecN* proteins from the genera *Ralstonia*

and *Neisseria* (Fig. 1). *Ralstonia solanacearum* is a plant pathogen and would not be expected to have the opportunity to transfer DNA into the human germ line. *Neisseria gonorrhoeae* is an obligate human pathogen, a characteristic consistent with the exchange of genetic information between bacteria and humans. The *Neisseria recN* genes are almost identical and share 51% nucleotide sequence identity with the D-1 antigen over their shared sequence (Fig. 2). *N. gonorrhoeae* lives predominately in the human genital tract, can adhere to human sperm, is extremely autolytic, and is capable of secreting its DNA into the surrounding milieu [7-9]. Thus, *N. gonorrhoeae* has the means and opportunity to transfer genes into the human germ line.

If the D-1 cDNA was produced from a gene that had originated from *N. gonorrhoeae*, there would be three features that would strongly support this phylogeny: first, the flanking sequences might be similar to the sequences flanking the *recN* gene in *N. gonorrhoeae*; second, the gene would not carry introns; and third, the codon usage would be closer to that of *Neisseria* than humans. Analysis of the codon usage in the D-1 cDNA showed it did not match *Neisseria* or human codon preferences (data not shown). The D-1 genomic sequence was not found in any publicly available human sequence database, and none of the people contacted who have previously published results using the D-1 cDNA were able to supply us with a D-1 cDNA clone. Therefore, we attempted to amplify the D-1 gene from human genomic DNA isolated from neuroblastoma, erythroleukemia, and B-lymphoid cell lines reported previously to express D-1 mRNA, using PCR with multiple primer pairs spanning different portions of the cDNA sequence. We were unable to amplify a D-1 gene product from any genomic DNA samples even though the  $\beta$ -actin gene was amplified from all samples (data not shown). To circumvent the possibility that all D-1 primer pairs used flanked an intron, we carried out PCR using cDNA as template. We were unable to amplify the D-1 gene from multiple cDNA sources, including the A375 cell line from which the original D-1 cDNA was isolated, even though  $\beta$ -actin and *gap-dh* specific primers produced the predicted products (data not shown). Finally, examination of the human genome revealed that the D-1 cDNA sequence was

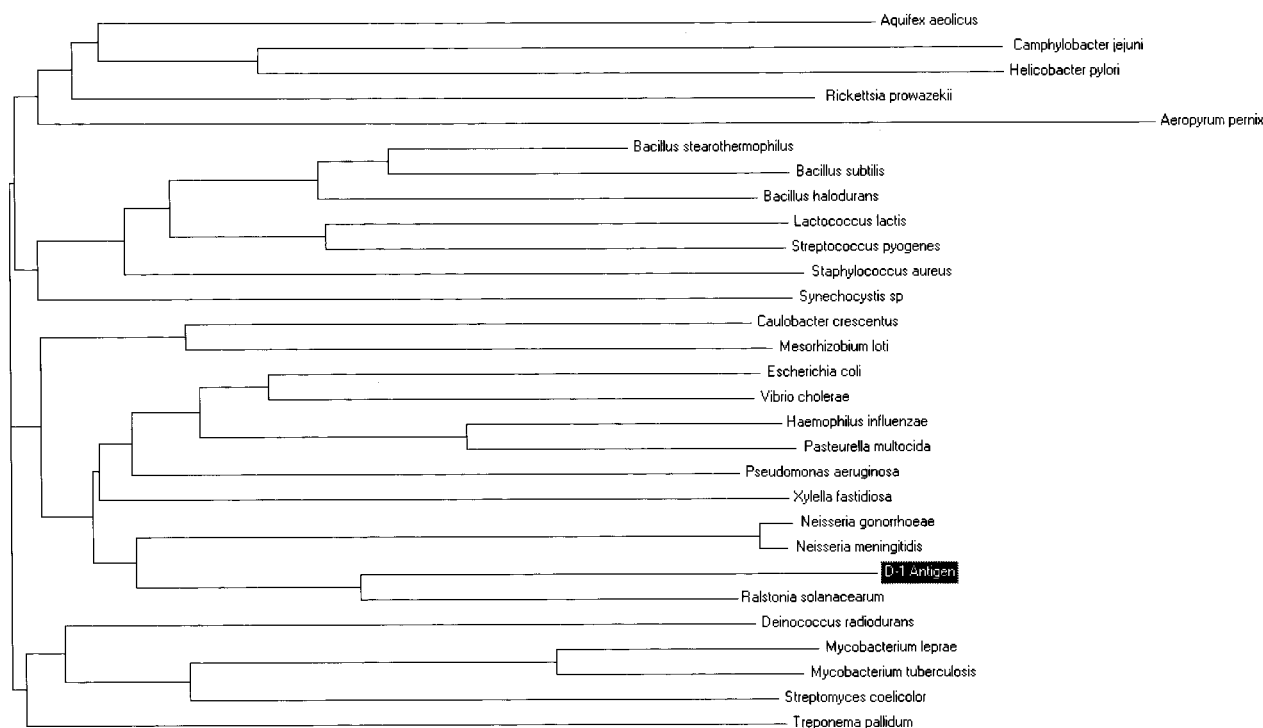


FIG. 1. Neighbor-joining phylogenetic analysis of the predicted D-1 protein sequence and RecN homologs. All organisms listed represent the top hits in a BLASTP search of the NCBI database. The location of the D-1 antigen in the tree is highlighted in black.

not present in any available sequence data. Therefore, our inability to analyze the genomic sequence within and surrounding D-1 have made it impossible to determine if the D-1 gene carries introns, or if the sequences flanking D-1 are similar to the sequences flanking the *recN* gene in *N. gonorrhoeae*. In combination, these negative results suggest that the D-1 cDNA is not a product of a human gene, but was present in the cDNA library as a contaminating bacterial sequence that reacted with the pooled melanoma sera. The close similarity of the D-1 sequence with bacterial *recN* genes supports this conclusion, and suggests that the source organism was an unsequenced bacterial species.

The misidentification of the D-1 cDNA as the product of a human gene may not be unique. Of the 52 genes with no obvious eukaryotic homologs suggested to be potential cases of lateral transfer [1], approximately 26 have not yet been eliminated by other groups [2-5]. Of these, six (including the D-1 cDNA) cannot be found in the publicly available human sequence databases. We have not tried to isolate these genes from human DNA, but five of these six genes were identified as cloned cDNAs from human cell line expression libraries. Additionally, BLAST analysis revealed that genes scoring with the highest E-values to the six are found in bacterial genomes. These observations

support the hypothesis that several of the genes reported to be demonstrations of lateral transfer of DNA from bacteria to humans are actually contaminating bacterial sequence present in cDNA expression libraries.

Reasons for mistakenly ascribing a bacterial origin to many of these genes include the failure to compare candidate genes with all known non-vertebrate sequences [4,5] and the failure to use appropriate phylogenetic analyses [3]. We have identified another possible explanation: the gene in question may not actually be of human origin. Our findings emphasize that the public databases may contain many gene sequences with misleading phylogenies and we suggest that efforts be made to verify the source organism of deposited genes.

#### ACKNOWLEDGEMENTS

We thank David Eide (University of Missouri-Columbia), David Tidd (University of Liverpool), Sue Cohn (Northwestern University), Ashok Aiyar (Northwestern University), Nancy L. Krett (Northwestern University), and Boris Pashche (Northwestern University) for providing reagents. This work was supported by PHS grant R01AI33493. E.P.S. was partially supported by PHS training grant T32GM08061.

RECEIVED FOR PUBLICATION DECEMBER 27, 2001;  
ACCEPTED MARCH 7, 2002.

**FIG. 2.** Nucleotide alignment of the entire human D-1 antigen and a portion of the *N. gonorrhoeae* *recN* gene over their shared sequence. Identical nucleotides are boxed in black.

```

D-1      1 .....
recN     1 GCGATTGGTCTGCTGTTGGCGATAAGGCCGATTACAGCCAAGTCCGCGCGGTGCAAAA

D-1      1 .....GAATTCACCCGCGACCCGGTCGCGCGCTGGCTC
recN     61 GAAGCGCAGCTGTCCGCATTGTTCGATAATTTC.CCATTTCCTCGCTTTAAAAAGCAAAATT

D-1      37 FAC.GAACACCGCTTC.....GACGCGAAGACACCGTCATGCTCCGCGCGCTGATCGA
recN     120 CCGTGAACAAGGCTGTTGGATGAGCCGGGGAAGAACTCAGTATCCGCCGATTATCGA

D-1      90 CGCGAACCGCGCTCGCGGCCTTCATCAACGGCACCAGCGCGACCTCCGGCAACTGCG
recN     180 TGCCAAAGGCAAAAGCCCGAGTTTATCAACAATCAGGCCGCTACCTTGGCGCAACTCAA

D-1      150 GAACTCCGGCGAGATGCTCGTCGACATTCATGGGCAACCCGACCAAGTTGCTGATGCG
recN     240 AGCCCTCCGGCGGCAGCTTTCGACATCCAAGGCAAAACGACATCATTCGCTTATCA

D-1      210 TCCCACCGCCGAGCGCGAGCTGTTGACACGCACCGGGGCTGCTCGCCGAGGCCCGAA
recN     300 GGAACTGCCAGCGCGAATGTTGGACGATTTTGGCGGAGGGTGCAGCGCGAAAC

D-1      270 CGTCGCGCGCGCGTGGCGTGTCTGCCGCGACCGGACGCGAGGCGATCGACCGCGCGAAGGC
recN     360 CGTCAGCGAGCTTTATCAAAACTGGGCAAAAGCGAATAAGCCCTCCAAAGCGCCAGGA

D-1      330 GCACCAACGCGAACTGCAGCTCGAGCCGAAAGCTCCGCTGGCAGCTCCGCGAGCTCGA
recN     420 ACACGCCGATCCCTCATTATCGAGCCGCGCGCTTGGAAATGGCAGTTTAAAGCACTGAA

D-1      390 CAAGCTCCGCGCCGACCGCGGCGAATGGGAAGAAGTACGCGCCGAGCACAAAGCCCTGTC
recN     480 TCAGTTGGACATTAAACAAGCGCAATGGGAAGCTTTACGCAAAAGCCACGACAGCCTGCG

D-1      450 GCATTCCGGGAATCTCATTCGAGGG...CGTCGCGCGCGC...GCTCGATCGCTTCTCCGA
recN     540 CCATTCTGCCGAATTGTTGCAGGCTGCCAAGAACTCCGAAGCAAGATGAGCGCGCAAA

D-1      505 TCGGA...CGAGGCATGCTCACGAGCTCGCGCGCATCCCTGTCGAAGCTCGCAGGCTCG
recN     600 CGGCAATCCAAACCGCATATCTATCAATCTCAAAAAC...TATTCGCAATTCGCAAAAGATCG

D-1      563 CCGACTACGATGCCGCGCTCGGCGACGCGCTCGCGTCTCGAGCC.....GGCGAAA
recN     659 AGCCGCGCTTGGCGAGAGCTGAATATGTTGGCAAGCATCGAGGCCGAATTCCGCGAAA

D-1      617 TCCAGCTGCAGGAAGCGGCTCTATTCCCTGTGCGCATACCGCGAGCCTCTGGATCTCGATC
recN     719 TC.AGTGCCAATAAGCG.....CGATGTGCCAGGACAC...AGTGACA....TCAACC

D-1      677 CCGACCGCTCCGCGCAGGTCGAAATGGCCTCGATTCGACCGCGCGCAAAAT
recN     764 CCAATGAATGGCGCGCAGGAAACAGCGTATGGGCGAACTGATGGGGATGGCGCGCAAAAT

D-1      737 TCCGCTTCCGCGCCGACACGCTGCACGAGGACACCGCGCGCGCGCGCGAGCTTCCCG
recN     824 ACCGCAATGAAACCGAAGAGTTGCCCTGCCAAGTTGGCAAAATCGAAACAC.GCCTGCCAA

D-1      797 CG.CTCGACCGCGCGCGGATCTCGCGCGCTCGAAGCCGCGCAGGCGAAGCGTGGCAC
recN     883 AGTCTTCAAGCTCCCGCTGATTTCGACCGCGCTCGAGCATAAATGTTGCCACAAATTTGCC

D-1      856 CCGTATCTGCCACCGCGAAGCCCTCTCGAAGGCGCGCGCAGGCGCGAAGCGCTC
recN     943 GAATATCAGGAAGCTGCCACATCCTTTCTGCCATCGCGCATCAGGCGCG.AGGCGTTTT

D-1      916 C.GCACCGCGGTGACGGCCGGCATGCAAGGACTGTCATGGCGCGCGCAGCTTCGAAGT
recN     1002 GAGTAGCGACACGACCGAGCATATGCAACAACCTTCCCATGAAAAGCGCGCTTTCGACAT

D-1      975 CGCGCTCGTACCGCTCCGCGACCGCGCCCGC.ACGGCTCGAACAGATCGAATTC.....
recN     1062 CGTCTCTTCCCT.TCGTCGCCACCGCAC...ACGGTTTGGACAGGTTCAATTTCAAGT

```

## REFERENCES

1. International Human Genome Sequencing Consortium. (2001). Initial sequencing and analysis of the human genome. *Nature* **409**: 860-921.
2. International Human Genome Sequencing Consortium. (2001). Correction: Initial sequencing and analysis of the human genome. *Nature* **412**: 565-566.
3. Stanhope, S. J., et al. (2001). Phylogenetic analyses do not support horizontal gene transfers from bacteria to vertebrates. *Nature* **411**: 940-944.
4. Salzberg, S. L., White, O., Peterson, J., and Eisen, J. A. (2001). Microbial genes in the human genome: lateral transfer or gene loss? *Science* **292**: 1903-1906.
5. Roelofs, J., and Van Haastert, J. M. V. (2001). Genes lost during evolution. *Nature* **411**: 1013-1014.
6. Hayashibe, K., Mishima, Y., and Ferrone, S. (1991). Cloning and in vitro expression of a melanoma-associated antigen immunogenic in patients with melanoma. *J. Immunol.* **147**: 1098-1104.
7. Elmros, T., Burnam, L. G., and Bloom, G. D. (1976). Autolysis of *Neisseria gonorrhoeae*. *J. Bacteriol.* **126**: 969-976.
8. Harvey, H. A., et al. (2000). Gonococcal lipooligosaccharide is a ligand for the asialoglycoprotein receptor on human sperm. *Mol. Microbiol.* **36**: 1059-1070.
9. Dillard, J. P., and Seifert, H. S. (2001). A variable genetic island specific for *Neisseria gonorrhoeae* is involved in providing DNA for natural transformation and is found more often in disseminated infection isolates. *Mol. Microbiol.* **41**: 263-277.