

Identification of foreign gene sequences by transcript filtering against the human genome

Published online: 14 January 2002, DOI: 10.1038/ng818

We have developed a computational subtraction approach to detect microbial causes for putative infectious diseases by filtering a set of human tissue-derived sequences against the human genome. We demonstrate the potential of this method by identifying sequences from known pathogens in established expressed-sequence tag libraries.

Many human diseases are likely to be caused by unknown pathogens¹. A variety of molecular methods have been used successfully to discover new microbes. Among these, amplification of conserved sequences using PCR identified new bacteria that cause bacillary angiomatosis² and Whipple syndrome³, and representational difference analysis⁴ revealed the herpesvirus causing Kaposi sarcoma⁵. Each of these methods has limitations: PCR amplification requires homology to known organisms, and representational difference analysis requires that the infectious agent is absent in control tissue from the same individual.

We have developed an *in silico* approach that uses the draft sequence of the human genome^{6,7} to identify nonhuman DNA sequence in expressed-sequence tag (EST) libraries of human origin. In this approach, called computational subtraction, sequences with significant similarity to the human genome are subtracted from the libraries. The remaining set is enriched for sequences of nonhuman origin. The potential power of sequence-based computational subtraction lies in its ability to identify new nonhuman sequences in a comprehensive and unbiased manner.

We tested computational subtraction on more than 3.2 million sequences from the GenBank human EST database^{8,9}. We used the MEGABLAST program^{10,11} to screen transcript sequences for similarity to a panel of human and mouse sequence databases, followed by a crude sequence quality filter. Sequences with BLAST alignment scores of 60 bits or higher, corresponding to an expectation value of 10^{-7} or less, were subtracted (Web Note A). Subtraction with seven database filters left a set of 65,839 ESTs, roughly 2.0% of the 3,287,578 sequences analyzed (Fig. 1).

We compared subtracted ESTs with the GenBank nucleotide database using MEGABLAST. Our analysis showed that the subtracted ESTs include: (i) sequences from known pathogenic and commensal organisms (Web Tables A–C and Web Notes B–E) (ii) microbial sequences likely to represent experimental contamination from sources such as *Escherichia coli* genomic DNA, PCR amplification and non-sterile reagents (Web Note F and Web Table D) (iii) transcripts from unsequenced regions of the human genome (Web Note G) and (iv) poor-quality sequences (Web Note H). In addition, more than 51,000 subtracted ESTs

have no detectable homology to database sequences; these are likely to include sequences in the latter three categories above, as well as new microbial sequences without homologs in the current databases.

Subtracted ESTs match several pathogenic virus genomes, including hepatitis B and C viruses, human papillomavirus types 16 and 18 (HPV-16 and -18), cytomegalovirus, Kaposi sarcoma herpesvirus, and Epstein-Barr virus (Table 1). We found 18 ESTs matching hepatitis B virus in a library of 16,743 sequences derived from non-cancerous liver of an individual with hepatocellular carcinoma. Hepatitis B virus is a major cause of both hepatitis and liver carcinoma¹². We also found sequence alignments between subtracted ESTs and bacterial, fungal, and protozoan genomes (Web Notes C–E).

As a proof-of-principle exercise for the feasibility of computational subtraction, we analyzed an EST library from HeLa cervical carcinoma cells, which harbor HPV-18, a viral agent of cervical cancer¹³. Using the filters described above (Fig. 1), we reduced 7,073 sequences from a HeLa EST library (UniGene #271) to 44, and further filtered these to 22 by eliminating matches to

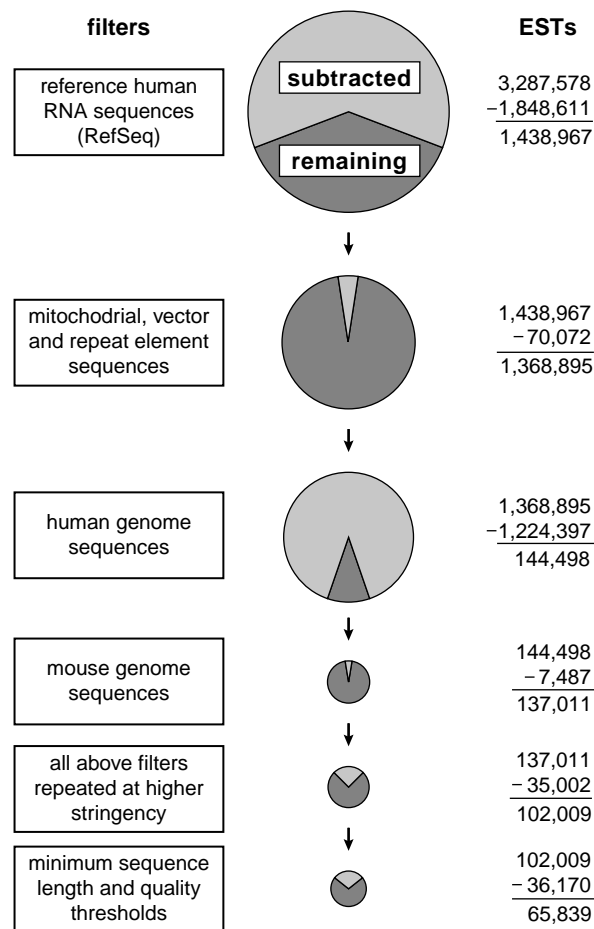


Fig. 1 Computational subtraction of the human transcriptome. Sequential subtractions were carried out on 3,287,578 ESTs in the NCBI human EST database using the MEGABLAST algorithm. Each circle represents the relative input number of ESTs at a given filtering step. The ESTs subtracted at each step are shown in light gray; the ESTs remaining after each step are shown in dark gray.



Table 1 • Sequences in human EST libraries matching known viral genomes

Species	Total NT hits	Unique NT hits	Unique libraries	Length	Best NT hit by score		Score
					Alignment	E value	
Hepatitis B virus	18	7	1	699	97.42%	< 10 ⁻¹⁸⁰	1243.4
Human cytomegalovirus	9	4	3	248	97.98%	10 ⁻¹²²	444.5
Human adenovirus type 2	7	1	6	531	99.06%	< 10 ⁻¹⁸⁰	1017.4
Human spumaretrovirus	7	1	1	299	99.00%	50 ⁻¹⁵⁹	565.5
Squirrel monkey retrovirus	6	1	5	587	99.15%	< 10 ⁻¹⁸⁰	1108.6
Human papillomavirus type 18	4	4	3	621	98.55%	< 10 ⁻¹⁸⁰	1132.4
Human papillomavirus type 16	3	3	2	519	98.65%	< 10 ⁻¹⁸⁰	963.9
Kaposi sarcoma herpesvirus	2	2	2	140	100.00%	85 ⁻⁷³	278.0
<i>Autographa californica</i> nucleopolyhedrovirus (baculovirus)	1	1	1	352	99.43%	< 10 ⁻¹⁸⁰	682.4
Hepatitis C virus	1	1	1	286	92.66%	40 ⁻¹⁰⁸	397.0
Epstein-Barr virus	1	1	1	617	97.73%	< 10 ⁻¹⁸⁰	1112.6

ESTs passing all seven filters were compared to the nucleotide (NT) database with the MEGABLAST algorithm. Alignments to viral sequences in the NT database with a bit score of greater than 250 (equivalent to an expectation value (E value) of ~10⁻⁶⁴) are reported above.

known human or *E. coli* sequences. Two of these 22 ESTs are derived from HPV-18.

To determine whether nonspecific subtracted sequences can be eliminated experimentally, we amplified by PCR the 22 non-matching sequences from genomic DNA samples, including HeLa and negative human controls (Web Note I and Web Fig. A). As predicted, the 20 non-specific sequences were excluded, as 11 primer pairs amplified all samples and 9 amplified no samples. In contrast, the two HPV-18 primer pairs specifically amplified only HeLa cell DNA. If HPV-18 were a new candidate infectious agent, we would need to verify its association with cervical carcinoma further. However, HPV-18 is already a known cervical carcinoma agent, present in 20% of cases¹⁴.

Here we show that sequence-based computational subtraction is able to identify known viral pathogens in infected tissues using existing EST libraries. However, these libraries are not ideal for microbial discovery, for two reasons: (i) they are focused on normal tissues and the most common cancers and (ii) the methods used for library generation often lead to sequence contamination.

We therefore propose to generate, sequence and filter cDNA libraries from tissues of diseases such as systemic lupus erythematosus and extraintestinal Crohn disease, which are candidates for infectious

etiology¹. In light of the imminent completion of the human genome sequence and measures to limit microbial and molecular contamination, the filtered sequences should contain almost exclusively foreign genes. The percentages of pathogenic sequences in hepatitis B virus-infected liver (0.1%) and in HPV-18-infected cervical carcinoma cell (0.03%) libraries suggest that sampling fewer than 10,000 sequences should be sufficient to discover viral genes. Further experiments could then distinguish sequences of benign commensal organisms from pathogen sequences by assessing the strength of their association with disease.

URL. A searchable database of the 65,839 subtracted ESTs and corresponding alignments is available at <http://research.dfc.harvard.edu/meyersonlab/compsub.html>.

Note: Supplementary information is available on the Nature Genetics web site (http://genetics.nature.com/supplementary_info/).

Acknowledgments

We thank P. Vaglio, M. Vidal and W. Wong for assistance with computational platforms; R. DePinho, W. Kaelin, K. Polyak and M. Vidal for discussion and the National Cancer Institute (D.T.) and the Pew Scholars in the Biomedical Sciences (M.M.) for support. M.M. is a Claudia Adams Barr investigator.

Griffin Weber^{1-3*}, Jay Shendure^{4*}, David M. Tanenbaum^{1,2}, George M. Church⁴ & Matthew Meyerson^{1,2}

**These authors contributed equally to this work.*

¹Department of Adult Oncology, Dana-Farber Cancer Institute, 44 Binney Street, Boston, Massachusetts 02115, USA.

²Department of Pathology, Harvard Medical School, Boston, Massachusetts 02115, USA.

³Decision Systems Group, Brigham and Women's Hospital, 75 Francis Street, Boston, Massachusetts 02115, USA. ⁴Department of Genetics, Harvard Medical School, Boston, Massachusetts 02115, USA. Correspondence should be addressed to M.M. (e-mail: matthew_meyerson@dfci.harvard.edu).

Received 23 August; accepted 5 December 2001.

1. Relman, D.A. *Science* **284**, 1308–1310 (1999).
2. Relman, D.A., Loutit, J.S., Schmidt, T.M., Falkow, S. & Tompkins, L.S. *N. Engl. J. Med.* **323**, 1573–1580 (1990).
3. Relman, D.A., Schmidt, T.M., MacDermott, R.P. & Falkow, S. *N. Engl. J. Med.* **327**, 293–301 (1992).
4. Lisitsyn, N. & Wigler, M. *Science* **259**, 946–951 (1993).
5. Chang, Y. *et al. Science* **266**, 1865–1869 (1994).
6. Lander, E.S. *et al. Nature* **409**, 860–892 (2001).
7. Venter, J.C. *et al. Science* **291**, 1304–1305 (2001).
8. Adams, M.D. *et al. Science* **252**, 1651–1656 (1991).
9. Boguski, M.S., Lowe, T.M. & Tolstoshev, C.M. *Nature Genet.* **4**, 332–333 (1993).
10. Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. *J. Mol. Biol.* **215**, 403–410 (1990).
11. Zhang, Z., Schwartz, S., Wagner, L. & Miller, W. *J. Comput. Biol.* **7**, 203–214 (2000).
12. Befeler, A.S. & Di Bisceglie, A.M. *Infect. Dis. Clin. North Am.* **14**, 617–632 (2000).
13. Boshart, M. *et al. EMBO J.* **3**, 1151–1157 (1984).
14. Burger, R.A. *et al. J. Natl Cancer Inst.* **88**, 1360–1368 (1996).

