EST-BASED ANALYSIS OF GENE EXPRESSION IN THE HUMAN COCHLEA

Irene S. Gabashvili, Richard J. Carter,
Peter Markstein
Hewlett-Packard Labs
{Irene.Gabashvili, Dick.Carter,
Peter.Markstein}@hp.com

Anne B.S. Giersch
Harvard University
agiersch@rics.bwh.harvard.edu

Hearing is one of the vital senses helping to perceive, reflect and communicate with the world around us. Genetics, developmental conditions, mechanical damage, infections, ototoxic medications, and aging are among the factors disabling or deteriorating this sense. Despite advances in genetic testing, linkage analysis and genomic/proteomic technologies, molecular-level understanding of the auditory system remains largely fragmented and incomplete. Hundreds of genes and proteins implicated in the process of hearing are known, but many more are yet to be discovered and characterized.

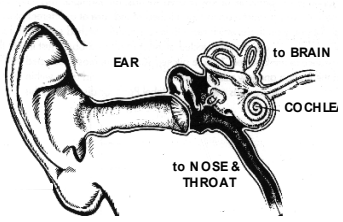

Fig.1 Diagram of the ear

One of the most fruitful approaches to identification of novel genes is analysis of tissue- and organ specific libraries. The cochlea is a sensory organ responsible for hearing (Fig. 1). Over 15,000 expressed sequence tags (ESTs) extracted from this organ (Morton fetal cochlear cDNA library, dbEST Library ID.371) had been previously clustered with other sequences and aligned to earlier versions of the human genome and known genes [1]. In this work, we show that even publicly available and relatively explored datasets need to be reanalyzed, based on better understanding of experimental procedures and potential sources of error. This is especially important as more genomic and phenotypic data becomes available almost on a daily basis.

To direct and control the process of EST mapping, we needed not only to align sequences ([2]-[5] and references therein), but also to check for a number of favorable and detrimental signals to identify the most likely mappings amongst many possibilities. This is realized by dynamic interaction of two in-house programs, *Enhancer2* and *BatchSearch*. *Enhancer2* is a 5000-line C++ program that finds exact matches of a number of input search patterns within a database of sequences. It is based on a Shift-AND algorithm to which we made novel enhancements: the ability to process the searched genome in chunks with little added overhead, and the ability to quickly abandon failing matches. The algorithm handles all IUPAC nucleotide codes with little additional overhead and is highly parallel.

BatchSearch is a 2500-line C++ program that interacts with *Enhancer2* by giving it search tasks and dynamically responding to its output. First, it trims the input EST of bases that are artifacts of the sequencing process. Second, a globally optimal set of high-entropy fragments is chosen from the EST using a dynamic programming algorithm. Then, the formulated exact-match search problem is passed to the waiting *Enhancer2* program. Depending on these results, *BatchSearch* can ask *Enhancer2* to refilter its search results, allowing for more widely dispersed clusters to be reported. Alternatively, *BatchSearch* can redo the genome search with smaller EST subsequences, in an effort to identify the most likely mapping. One search for six 20-nucleotide fragments using *Enhancer2* takes about 2.5 seconds on a 2.8 gHz Xeon CPU, and a dual-processor HP XW8000 PC workstation requires 5.5 hours to map the entire library of 15000 cochlear ESTs to the human genome.

We mapped over 98% of 15,049 ESTs in the Morton fetal cochlear library to specific regions in the human genome and genomes of laboratory organisms. Unmapped sequences (area 5 in Fig.2) are either ambigous or formed by nonspecific recombination events. Non-human contaminations in the dataset (area 4 in Fig.2) come from laboratory organisms – mainly yeast, E.coli, phages and cloning vectors, but there are also single occurences of such unexpected species as fish, worm and mouse. Many transcripts corresponding to ESTs present in the dataset might not be expressed

as proteins, but instead are degraded by nonsense-mediated mRNA decay or other cell surveillance mechanisms. We found a number of incomplete, truncated mRNAs in the library, confirming this possibility. About 20% of all ESTs are potential genomic contaminations. Almost 80% of the Morton ESTs (Fig.2, areas 1 and 2) are annotated in the latest build of Unigene, although about 8% of these annotations remain hypothetical. Of our gene assignments, 99% are the same as in Unigene. Our "new gene" annotations often correspond to Unigene's "transcribed loci" and most discrepancies in gene names are solely due to the different naming of same gene. For example, "ecotropic viral integration site 2A" is the same gene as "neurofibromin 1 (neurofibromatosis, von Recklinghausen disease, Watson disease.)" Less than 1% of our EST mappings do not correspond to Unigene assignments. In half of these cases our results might be better. There are also examples of old Unigene annotations being better than the latest ones.

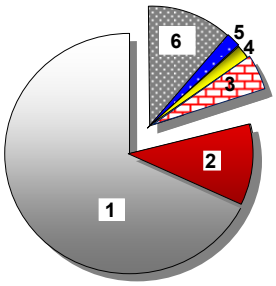

Fig.2 Distribution of our mappings of cochlear ESTs. Areas 1 and 2 are classified by Unigene. See text.

In addition to the 4,058 Unigene clusters, we have determined almost 1,000 new loci. Many of these might represent novel genes (area 3 in Fig. 2.) Others are isoforms which align within known genes, often within introns (area 6.)

Comparison of our mappings to alignments produced by other tools, including BLAST [2] and BLAT [4], shows that our solutions are essentially the same. Our tool provides additional post-processing capabilities and is faster. Most of the novel genes found in this work are being incorporated in the new build of the human genome [6], but we were the first to analyze tissue-specificity of these genes. We have further narrowed down the list of novel genes by filtering out genomic contaminations and highly repetitive sequences. The candidate genes include a possible transcription factor, a motor protein, a collagen and a transmembrane protein. The findings are currently being verified by independent approaches.

Among about 5,000 genes identified, almost 2,000 genes are represented by single ESTs. Less than 200 genes are supported by 10 or more sequences. The difference between cochlear and other existing libraries is statistically significant for a very small number of relatively highly expressed genes. These genes maintain the shape of acoustic resonators in the ear. Mutations of these genes are often associated with syndromic deafness (e.g., osteogenesis imperfecta caused by defects in collagen). Low expressed genes specific for cochlea include regulatory proteins potentially responsible for nonsyndromic hearing loss.

Many crucial processes of life, hearing being one of them, are only partially understood at the molecular level. Large-scale sequencing of tissue-specific genes and fast yet reliable mapping of sequences will help to identify key components of sound transduction and speed up progress in hearing research.

REFERENCES

[1] A.B. Skvorak, Z. Weng, A. J. Yee, N. G. Robertson, C. C. Morton, "Human cochlear expressed sequence tags provide insight into cochlear gene expression and identify candidate genes for deafness", Hum Mol Genet. (1999), **8**, 439-452.

[2] S.F. Altschul, W. Gish, W. Miller, E.W. Myers, D.J. Lipman, "Basic local alignment search tool." J. Mol. Biol. (1990), **215**, 403-410.

[3] C. Gemund, C. Ramu, B. Altenberg-Greulich, T.J. Gibson, "Gene2EST: a BLAST2 server for searching expressed sequence tag (EST) databases with eukaryotic gene-sized queries", *Nucleic Acids Res* (2001), **29**, 1272–1277.

[4] W.J. Kent, "BLAT—the BLAST-like alignment tool", Genome Res (2002), **12**, 656–664

[5] J. Krüger, A. Sczyrba, S. Kurtz and R. Giegeri **"**e2g: an interactive web-based server for efficiently mapping large EST and cDNA sets to genomic sequences". Nucleic Acids Res. (2004), **32** (Web Server issue), W301-304

[6] D. Thierry-Mieg, J-T. Thierry-Mieg, M. Potdevin, M. Sienkiewicz. "Identification and functional annotation of cDNA-supported genes in higher organisms using AceView, unpublished. http://www.aceview.org/