An Oligonucleotide Fingerprint Normalized and Expressed Sequence Tag Characterized Zebrafish cDNA Library

Matthew D. Clark,[1,5] Steffen Hennig,[1] Ralf Herwig,[1] Sandy W. Clifton,[3] Marco A. Marra,[3] Hans Lehrach,[1] Stephen L. Johnson,[2] and the WU-GSC EST Group[3,4]

[1]Max-Planck-Institut für Molekulare Genetik, 14195 Berlin, Germany; [2]Department of Genetics, Washington University, St. Louis, Missouri 63110, USA; [3]Washington University Genome Sequencing Center EST Lab, St. Louis, Missouri 63110, USA

The zebrafish is a powerful system for understanding the vertebrate genome, allowing the combination of genetic, molecular, and embryological analysis. Expressed sequence tags (ESTs) provide a rapid means of identifying an organism's genes for further analysis, but any EST project is limited by the availability of suitable libraries. Such cDNA libraries must be of high quality and provide a high rate of gene discovery. However, commonly used normalization and subtraction procedures tend to select for shorter, truncated, and internally primed inserts, seriously affecting library quality. An alternative procedure is to use oligonucleotide fingerprinting (OFP) to precluster clones before EST sequencing, thereby reducing the re-sequencing of common transcripts. Here, we describe the use of OFP to normalize and subtract 75,000 clones from two cDNA libraries, to a minimal set of 25,102 clones. We generated 25,788 ESTs (11,380 3′ and 14,408 5′) from over 16,000 of these clones. Clustering of 10,654 high-quality 3′ ESTs from this set identified 7232 clusters (likely genes), corresponding to a 68% gene diversity rate, comparable to what has been reported for the best normalized human cDNA libraries, and indicating that the complete set of 25,102 clones contains as many as 17,000 genes. Yet, the library quality remains high. The complete set of 25,102 clones is available for researchers as glycerol stocks, filters sets, and as individual EST clones. These resources have been used for radiation hybrid, genetic, and physical mapping of the zebrafish genome, as well as positional cloning and candidate gene identification, molecular marker, and microarray development.

[The sequence data described in this paper have been submitted to the dbEST/GenBank data library under accession nos. AA497144–AA497369, AA542435–AA542678, AA545709–AA545724, AI384176–AI384205, AI384761–AI384796, AI396646–AI396663, AI396733–AI396777, AI396895–AI396938, AI397015–AI397130, AI397219–AI397252, AI397388–AI397484, AI415743–AI416403, AI436854–AI437493, AI444118–AI444540, AI461280–AI461395, AI476823–AI478024, AI496677–AI497576, AI522337–AI522810, AI544445–AI546083, AI558267–AI558995, AI584192–AI585023, AI585025–AI585238, AI588088–AI588836, AI601277–AI601868, AI626134–AI626875, AI629052–AI629398, AI641018–AI641780, AI657549–AI658347, AI666929–AI667197, AI667264–AI667414, AI667488–AI667567, AI721460–AI721747, AI721839–AI721978, and AI722283–AI722483.]

Despite the recent interest in zebrafish for developmental biology and genetics, relatively few genomic resources have been available until recently. Expressed sequence tags (ESTs), the product of high-throughput, single-pass, cDNA sequence analysis, are often a useful part of the genomic characterization of human and model organisms, serving to enhance gene discovery, help annotate genomic sequence, and provide cDNA resources for mapping and further experimental analysis (Marra et al. 1998). A limiting reagent for EST projects is the availability of suitable normalized or nonnormalized libraries.

One method for library normalization is oligonucleotide fingerprinting (OFP) -based normalization, whereby cDNAs or other DNA sequences arrayed on filters are sequentially probed with hundreds of radiolabeled short oligonucleotides. The signals for each of these hybridizations are captured, generating a hybridization fingerprint for each clone, which is dependent upon its sequence. Clustering algorithms group clones with similar fingerprints (i.e., from the same gene or sequence) together (Fig. 1). Typically, one representative clone is rearrayed for further analysis, such as tag sequencing. OFP technology has been used previously in the mapping of the herpes simplex virus genome (Craig et al. 1990), the selection of clones for genomic shotgun sequencing (Radelof et al. 1998), and for the characterization and normalization of cDNA libraries (Meier-Ewert et al. 1993, 1998; Drmanac et al. 1996; Milosavljevic et al. 1996; Poustka et al. 1999).

Here, we have used OFP to characterize 75,000 cDNAs from zebrafish embryonic and adult liver cDNA libraries, and provide the most extensive validation of OFP to date. OFP analysis revealed a minimally redundant set of 25,102 cDNAs, which provided a suitable set of cDNAs for the recently initi-

[4]The WU-GSC Group is Deana Pape, Tammie Kucaba, Jennifer Bennett, Yvette E. Ritter, Irina Ronko, Brenda Theising, Rusudan Tsagearishivili, Todd Wylie, Rick Waterman, and Dan Fischer.
[5]Corresponding author.
E-MAIL clark@molgen.mpg.de; FAX ;49 30 84131380.
Article published on-line before print: *Genome Res.,* 10.1101/gr.186901. Article and publication are at http://www.genome.org/cgi/doi/10.1101/gr.186901.

11:1594–1602 ©2001 by Cold Spring Harbor Laboratory Press ISSN 1088-9051/01 $5.00; www.genome.org

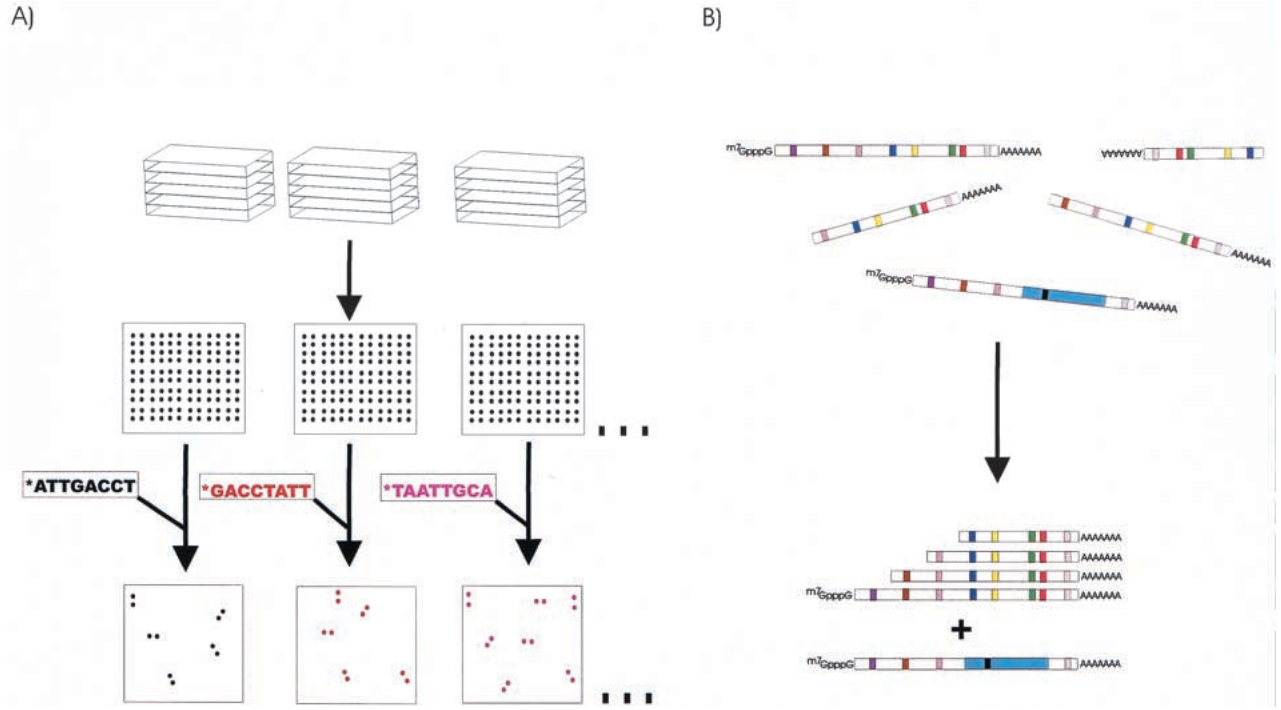

Figure 1 Oligofingerprint (OFP) analysis. (*A*) Inserts from a gridded library are PCR amplified, and identical arrays are generated by spotting the amplicons. The arrays are sequentially probed with ^{33}P-labeled, computer-selected, 8-mer oligonucleotides under high stringency conditions. In each hybridization, 5%–25% of the clones will be positive in each hybridization. (*B*) After image analysis quantifies the hybridization signals for each of the spots, a sequence fingerprint is generated for each clone based on which probe sequences are present and which are absent. Clustering algorithms then can group clones with significantly similar fingerprints (from the same transcript) together. cDNAs from the same gene, but different splice forms should be identifiable because they can give different fingerprints even if they share the same end sequences.

ated Washington University Zebrafish EST project (http://zfish.wustl.edu). Analysis of 10,654 3′ ESTs generated from this set revealed 7232 clusters (likely genes), corresponding to a 68% gene diversity and identifying 53% of all the genes identified by the Washington University Zebrafish EST project (66,158 ESTs and 13,685 3′ clusters by August 31, 2000). Using this sequence data, we have conducted the most thorough analysis of the OFP technique to date. ESTs and cDNA clones from this project already have enabled a variety of projects including large-scale radiation hybrid mapping (Geisler et al. 1999; Hukriede et al. 1999), analysis of human: zebrafish synteny relationships (Barbazuk et al. 2000), positional cloning projects (Kupperman et al. 2000), and the identification of molecular probes for gene-expression studies (Parichy et al. 2000).

RESULTS

To identify zebrafish cDNAs for an EST project, we first arrayed 66,000 cDNA clones from the late-somitogenesis library (RZPD lib. ICRFp524), and 20,000 cDNA clones from the adult-liver library (see Resource Availability). Fifty-five thousand clones from the ICRFp524 and 20,000 clones from the ICRFp532 libraries were selected for OFP analysis (Meier-Ewert et al. 1993, 1998; Drmanac et al. 1996; Milosavljevic et al. 1996; Clark et al. 1999). cDNA inserts first were PCR-amplified and spotted onto filters. Filters then were hybridized sequentially with 223 different ^{33}P-labeled short oligonucleotide probes (see http://www.molgen.mpg.de/~ag_zebrafish for oligonucleotides), and hybridization signals were captured using a phosphorimager. The resulting images were analyzed with the HFA image-analysis package, a hybridization fingerprint was generated for each clone, and fingerprints were grouped, based on similarity, to generate clusters of similar clones (Meier-Ewert et al. 1998; Clark et al. 1999). Using the results of this OFP clustering (hereafter referred to as OFP1), a single representative clone was selected from each cluster and rearrayed. By these means, the initial 75,000 clones from libraries ICRFp524 and ICRFp532 were reduced to a minimally redundant set of 25,102 clones, which we now refer to as library MPMGp609.

Normalization Assessment

One set of measures for success of library normalization techniques such as OFP is the degree of underclustering, where cDNAs from the same gene are placed in different clusters, and overclustering, where cDNAs from different genes are placed in the same cluster. Underclustering can be assessed by measuring the number of a few representative transcripts in pre- and postnormalization steps, or by large-scale assessment of transcript diversity, such as by EST analysis. Overclustering can be assessed by analysis of entire large- or medium-member clusters, or by rate of retention of representative cDNAs for poorly expressed genes following the normalization step.

Our initial analysis of library MPMGp609 showed that it was highly normalized. To assess normalization, we chose cDNA probes representing a range of various abundant, intermediate-level, and rare transcripts (Fig. 2). Hybridization of

Table 1. Comparison of Normalized Libraries

Library	MPMGp609	1NIB	1NFLS
Average insert size (kb)	1.9	1.5	0.9
Stringent protein hits (5′ reads)	42.6%	53.6%	15.2%
Coding Prediction 5′ reads	74.8%	43.8%	33.7%
Coding Prediction for 5′ reads with stringent protein hits	93.5%	70.3%	53.8%
Estimate of clones containing initiating methionine	33.9%	17.5%	11.6%
3′ ESTs with consensus poly-adenylation signal	76.0%	77.6%	47.6%
Sequence diversity (3′ clusters from 10,600 3′ reads)	68%	69%	80%

We compared the three normalized libraries: MPMGp609, 1NIB, and 1NFLS. Average insert size (based on at PCR products from at least 96 randomly-chosen clones) and stringent protein hits (WU-BLASTX score $e \geq -30$) are indicative of which transcript fragments are being examined (e.g., if the 5′ end of the insert falls in UTR, the protein coding region, or an artifactual unspliced non-coding region.) We searched for rarer, novel, or diverged transcripts with the gene prediction programs GenScan (Burge and Karlin 1997) and MZEF (Zhang 1997); a hit with either program was seen as indicative. The efficacy of these gene prediction programs on ESTs was tested using 5′ ESTs with stringent protein hits (i.e., known to be coding). The MPMGp609 library performed better than the other libraries, presumably because a zebrafish EST must have a much longer alignment with a mammalian protein than a human EST to produce a given score (here *e*-30) and thus contains a longer piece of coding sequence that is more likely to be detected. As a further quality measure we examined the number of 5′ ESTs with stringent protein hits, which matched the first methionine of their protein sequence hit. We also examined the proportion of 3′ ESTs containing the consensus poly-adenylation signals (AAUAAA or AUUAAA) present in 80% of eukaryotic 3′ UTRs (Pesole et al. 2000). To measure library complexity we clustered 10,600 3′ ESTs with at least 200 high-quality bases (as demarked in their dbEST entry) with PHRAP. The number of 3′ clusters (counting singletons as clusters) was taken as a measure of the number of transcripts identified.

these probes to filters spotted with the original (ICRFp524 and ICRFp532) and the rearrayed (MPMGp609) libraries revealed that highly expressed genes, such as *β-actin* or *cytochrome C oxidase I* that were present at more than 1% in nonnormalized libraries were reduced twofold to threefold in abundance in the rearrayed, normalized library. In contrast, clones for rare transcripts such as *kox21* or *KIAA0235* are retained consistently in the rearrayed library. However, hybridization of such cDNA probes back to library arrays may underestimate the level of normalization, as splice variants, paralogous, and closely related genes all will be positive with a long hybridization probe, but should be assigned to different OFP clusters. For example, there appear to be at least three highly similar zebrafish Tubulin *β-2* genes: unigene clusters Dr.740, Dr.1359, and Dr.2813.

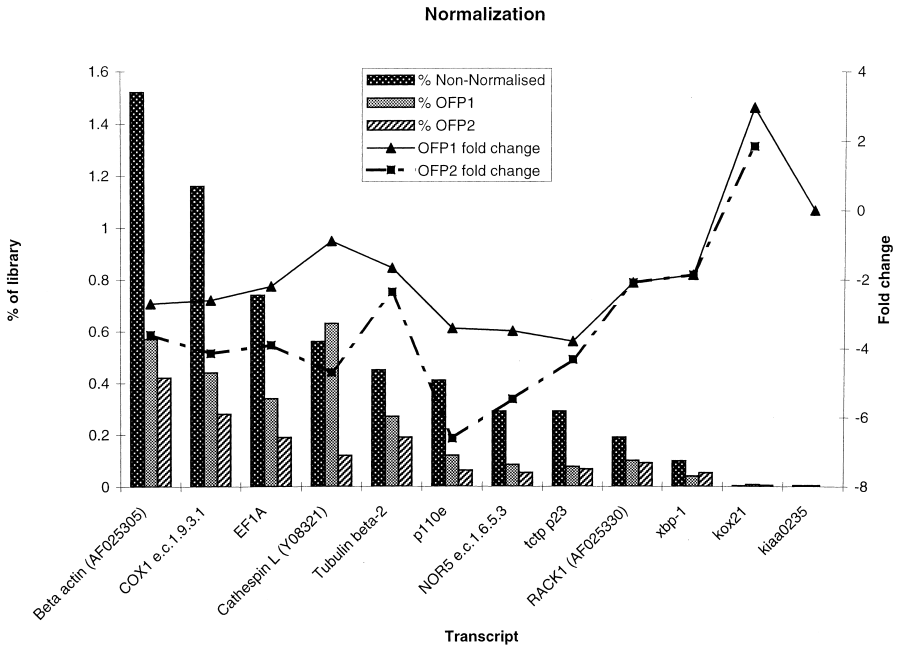

Figure 2 Library normalization. cDNAs of high (1%), medium (0.1%), and low (<0.01%) level transcripts were hybridized back to the starting (nonnormalized) libraries and normalized MPMGp609 library. The level of many prevalent transcripts was reduced by twofold to eightfold, and that of lower-level transcripts increases twofold–threefold. OFP2 is a more recent clustering carried out using a different algorithm (Herwig et al. 1999) and has an even higher normalization rate with only 20,883 clone clusters, and clearly performs better at clustering problematic transcripts such as Cathespin L.

After this initial analysis, we generated 25,788 ESTs (11,380 3′ and 14,408 5′) from over 16,000 of these clones. Clustering 10,654 high-quality 3′ ESTs from this set identified 7232 clusters (likely genes). This corresponds to a 68% gene diversity rate, greater than that observed for any of the other zebrafish cDNA libraries, and similar to the normalization rate from two of the most extensively sampled libraries from the WashU-Merck Human EST project (Bonaldo et al. 1996; Hillier et al. 1996). This gene diversity rate (68%) suggests the complete set of 25,102 clones may contain as many as 17,000 genes.

Using the ESTs from the MPMGp609 library, we attempted to improve our clustering procedures and found the procedure of Herwig et al. (1999), which uses a k-means–based algorithm, and sequential rounds of clustering to reduce underclustering (Fig. 2). The

OFP2 clustering gives higher normalization rates, especially for the more problematic transcripts such as Cathepsin L (see Discussion), but removes both the kiaa0235-positive clones.

We examined the purity of several OFP clusters by tag sequencing all their clones. A total of 102 5′ tag sequences of clones from four clusters of varying size passed automated quality control. This consists of a 96-member cluster (49 of 56 clones, 87.5% pure), two 34-member clusters (28 of 28 clones, 100% pure, and 14 of 15 clones, 93% pure), and one six-member cluster (three of three clones, 100% pure). Overall, 92.2% of the clones from a given OFP cluster were from the same transcript. Further examination showed that the false clones were typically in the largest clusters and assigned a low confidence value within their cluster, and thus were not selected for rearrayed MPMGp609 clone set (which contains the highest confidence clone from each cluster). These results were similar to previous studies (Meier-Ewert et al. 1998; Poustka et al. 1999), which also found 90% or greater cluster purity. Because OFP clusters contain additional clones that may represent full-length or splice variants, researchers can obtain the complete clone set of any OFP cluster from the Resource Centre (RZPD) of the German Human Genome Project (see http://www.molgen.mpg.de/~ag_zebrafish/).

Library Quality

However, the degree of normalization is only one measure of library quality. Other important attributes include a large average-insert size and high coding-sequence content, which is essential for homology and comparative studies. Furthermore, the presence of intact and processed 3′ untranslated regions (UTRs) (as assessed by correctly positioned canonical poly-adenylation signal and poly-A tails) and a large fraction of clones with intact initiating methionine codons may reflect the proportion of full-length cDNAs.

To assess insert size, we sized more than 200 inserts from the MPMGp609 library on agarose gels revealing an average insert size of 1.9 kb/clone, similar to the median mRNA size reported for humans. This insert size greatly exceeds the 1.5 kb and 0.9 kb average insert for the 1NIB and 1NFLS libraries, two of the most characterized, normalized libraries from the WashU-Merck Human EST project (Hillier et al. 1996).

To assess coding sequence content, we analyzed 5′ ESTs from each of the three libraries by BLASTX searches to the SWISS-PROT and TREMBL protein databases (Bairoch and Apweiler 2000) to identify ESTs containing significant homology to known proteins, as assessed by a BLAST e-score $\leq e^{-30}$. By these criteria, we found 42.6% of the MPMGp609 5′ ESTs contain protein-coding sequence. In contrast, these criteria indicate 53.6% of the 1NIB library and only 15.2% of the 1NFLS libraries contain protein-coding sequence.

However, the criteria we used may underestimate the proportion of zebrafish ESTs contain coding sequence compared to human ESTs, as the protein databases are relatively rich in mammalian proteins, but poor in teleost proteins. This idea is supported by our examination of the BLAST searches, which show that many of the human ESTs had short but nearly perfect hits to human or mammalian proteins, while more often the zebrafish ESTs had longer but lower similarity matches to mammalian, and few had matches to teleost protein sequences. We also found that 74.8% of 5′ ESTs from MPMGp609 were predicted to contain sequence coding by either the Genscan (Burge and Karlin 1997) or MZEF (Zhang 1997) packages. In contrast, only 43.8% of 5′ ESTs from the 1NIB library, and only 33.7% of 5′ ESTs from the 1NFLS were predicted to contain coding sequence. Our estimate for the short-insert 1NFLS library is apparently inflated with respect to other libraries by the large proportion of 5′ reads that contain 3′ UTRs and polyadenylation signals, which MZEF also calls as coding sequence by our criteria. Many of the 1NFLS inserts appear to consist wholly of 3′ UTR, an increased rate of short and truncated inserts being a well-known drawback of many normalization procedures.

The MPMGp609 library also has a high level of intact initiating methionine codons. Of 5′ ESTs with stringent protein hits, 33.9% aligned with the initiating methionine of the database sequence, suggesting that these zebrafish cDNA clones contain intact 5′ UTRs. Presumably, an equally high fraction of 5′ ESTs from this library with predicted coding sequence (but no stringent protein hit) also may contain an intact 5′ UTR. In contrast, only 17.5% of 5′ ESTs with stringent protein hits from the 1NIB and 11.6% from the 1NFLS libraries align with the initiating methionine of the database entry, and thus may contain an intact 5′ end. We believe that this analysis may underestimate the number of clones with intact 5′ UTRs, because some genes may have 5′ UTRs longer than the 5′ ESTs read length, or may not be highly conserved at their amino terminus.

Often, library clones represent cDNAs that are primed internally within a transcript, represent pre-mRNA molecules, or are genomic contaminants that may be as much as 15%–20% of 3′ ESTs in some libraries (Gautheret et al. 1998). As another measure of library quality, we examined the level of consensus poly-adenylation signals in the 3′ reads of the three libraries. We found similar levels, 76% and 78%, of consensus polyadenylation signals in 3′ ESTs of the MPMGp609 and 1NIB libraries respectively, close to the 80% level found among 3′ UTRs in UTRDB (Pesole et al. 1996). Thus, most of the clones in MPMGp609 likely have intact 3′ UTRs. Taken together, our findings of large average-insert size, high proportion of coding sequence, and high proportion of full-length clones helps indicate the suitability of the OFP-normalized MPMGp609 library for genomic uses and its usefulness to the wider zebrafish community for functional analyses.

Sequence Analysis

Because ESTs may not represent genes or even contain coding sequence, we also compared the ESTs from this project to consensus sequences of the Human UniGene clusters (Haas et al. 2000; http://www.dkfz-heidelberg.de/tbi/services/GeneNest/index). We obtained 6654 stringent (e-score $\leq {}^{-}20$) TBLASTX hits from 30,707 Zebrafish ESTs (5′ and 3′ reads). Of these, 4769 matched 2036 UniGenes with assigned map positions from the human gene map project (Deloukas et al. 1998), this includes 288 possible zebrafish orthologs of cloned human-disease genes (P. Aanstad, pers. comm.).

Given the interest in using the comparison of the human and zebrafish genomes to annotate both genomes and aid in the cloning of zebrafish mutations, we have made this data set available on the Web (see http://www.molgen.mpg.de/~ag_zebrafish/). Because even the human EST set is far from complete, we also searched a 200-MB section of finished human genomic sequence. We further identified 74 zebrafish ESTs that had significant matches to this region, but no significant human UniGene hits; several of these genomic re-

gions match exons of in silico-predicted genes that now have been confirmed by RT-PCR (R. Sudbrak, unpubl.). Thus, as the human-genome sequence is completed, we might expect this initial set of zebrafish ESTs to provide support for as many as 1000 (3.2/0.2 × 74) human genes that are not currently supported by human ESTs.

Differential Expression

Differential expression technologies, such a EST projects on nonnormalized libraries (Okubo et al. 1992) and SAGE (Velculescu et al. 1995), count the number of clones, or tags, in a library from different transcripts and provide a means of quantifying expression levels in different tissues. Such digital expression profiles can be used as electronic or virtual Northern blots and are very useful in searching for molecular markers or identifying candidate genes for mutations.

OFP also can identify transcripts differentially expressed between the two source materials (Meier-Ewert et al. 1998). We were interested to see if we could use our OFP results to identify transcripts specific to the embryonic library for further analysis. Assuming each OFP cluster represents a transcript, and using the number of clones from each library as a digital measure of expression level for each of the starting materials, we highlighted those transcripts highly likely to be differentially expressed (Fig. 3). Examining ESTs from OFP clusters (transcripts) rich in embryonic clones, we found that these were highly enriched for muscle-specific proteins such as *myosins*, *tropomyosin*, and *skeletal actin*. In contrast, OFP clusters rich in liver clones were highly enriched for liver-specific proteins such as *trypsin*, *chymotrypsin*, *serine-protease inhibitor*, and *vitellogenin*, the predominant yolk protein that is made in the liver and excreted into the blood for absorption by the ovaries. We further tested the OFP-based expression profiles by choosing six clusters for which we had at least one EST (not coding for known liver or embryo-specific genes). We designed "overgo" oligonucleotide probes (McPherson et al. 2001; http://genome.wustl.edu/gsc/overgo/overgo.html) from these ESTs and hybridized them back to the full embryo and liver libraries. After scoring the hybridizations, we compared the number of clones in each library with the number of clones screened for each transcript, using the method of Audic and Claverie (1997) (http://igs-server.cnrs-mrs.fr) to calculate the probability of differential expression. Five of the six clusters were judged to be differentially expressed (Audic and Claverie probabilities >0.97), and the remaining cluster had a lower probability (>0.7). This last cluster also had the lowest scores of the six clusters based on our OFP analysis (1.1×10^{-3}). The complete list of digital-expression profiles of the embryo and liver based on our OFP analysis is available at http://www.molgen.mpg.de/~ag_zebrafish/.

These results indicate the ability of the OFP system to identify differentially expressed genes. Thus, as we continue to OFP further cDNA libraries, we will be able to generate an overview of expression throughout embryonic development and the adult body.

DISCUSSION

Oligonucleotide Hybridization Fingerprint Analysis

We identified several areas for improvement of OFP in zebrafish. First, the set of oligonucleotides used for hybridization is suboptimal for zebrafish. Our experience shows that

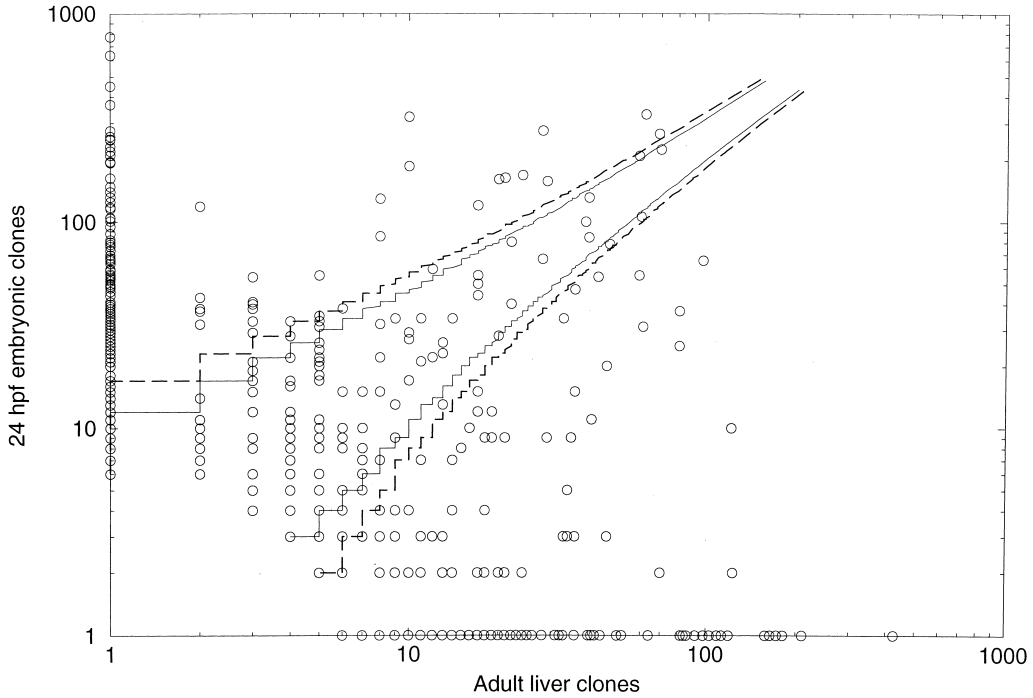

Figure 3 We used the number of clones in a cluster from each library as a measure of the level of the transcripts in the starting materials and plotted the adult liver and late somitogenesis on the *X* and *Y*-axis, respectively. Each circle represents an OFP2 clone cluster (transcript). The lines correspond to the bands of the 90% (filled line) and 95% (broken line) confidence intervals of the two-sided binomial test. Clusters outside of these bands are significantly differentially expressed according to the respective significance levels. Many of the clusters are composed solely of embryo or liver clones.

many hybridize with zebrafish cDNAs with lower rates than with human, reducing their potential informativeness. With the availability of additional zebrafish cDNA sequences, including EST sequencing supported by this project, we have designed a new set of zebrafish-specific oligonucleotides, which should improve future OFP efforts. The apparently poorer level of normalization achieved for some transcripts (e.g., Cathepsin L in OFP1, see Fig. 2) appears to be from the low number of the oligo sequences the mRNA contains (seven out of a total of 223). There are at least two very similar Cathepsin L genes in zebrafish (Unigene clusters Dr. 223 and Dr. 686) that may crosshybridize, but which OFP and EST analysis resolved. Increasing the average number of oligos per cDNA will improve both the possible normalization rate, and resolving power of the technique, clustering cDNAs from the same gene, while resolving closely related genes.

A second consideration for improved performance of OFP is the method used for clustering. The clustering procedure we used here tends to produce multiple clusters for many highly expressed genes (underclustering) because the algorithm tends to cluster together cDNAs of similar length from the same gene, and then fails to cluster overlapping cDNAs of different lengths from these genes (Poustka et al. 1999). While this effect reduces the efficiency of the normalization, it can provide a tiled path of 5′ ESTs along the transcript that then can be used to assemble virtual transcripts and to better annotate genomic sequence. The underclustering problem can be alleviated partially by the superior algorithm of Herwig et al. (1999), which uses a k-means–based algorithm, and sequential rounds of clustering to reduce underclustering. Because length variance is the predominant reason for underclustering, this effect should be minimized for stringently size-selected cDNA libraries, or from libraries of full-length cDNAs (for instance, oligo-dT primed and cap-selected libraries [Suzuki et al. 1997]). Additional sources of apparent underclustering may be alternative splice forms that represent different mRNA species and alter clone fingerprints. These would be correctly split into different OFP clusters, yet be from the same gene. In assembling our ESTs from different clusters, we saw features that appear to be alternate use of exons and sequence polymorphisms. Examination of more clones from the OFP clusters corresponding to such ESTs could be a rich source of such data. Instructions for ordering these clones from the RZPD is available through our web site (http://www.molgen.mpg.de/~ag_zebrafish/). Neither of the libraries were made from inbred strains, but were constructed with material from several individuals. Polymorphisms and alternative splicing may account for the large number of vitellogenin ESTs from library MPMGp609. At least seven vitellogenin genes have been reported in zebrafish (Wang et al. 2000); most are long (at least 4 kb) and several appear to produce a number of splice variants. In addition, six of the seven vitellogenin genes contain a serine-rich phosphovitin region rich in single-nucleotide polymorphisms.

Comparison to Other Normalization Methods

There are several methods of library normalization that an EST project could use. For a small project, a common approach is to hybridize common cDNAs or a complex cDNA probe and then to select cDNAs that do not hybridize. This is a rapid technique, which should eliminate many of the most common cDNAs. However, it also selects for clones with no or small inserts, and can eliminate related or repeat containing cDNAs. This approach has been previously used in some human and zebrafish EST projects (Adams et al. 1995; Gong et al. 1997) with some success. However, as we have seen, this approach could well have lost some of the closely related genes such as the zebrafish Tubulin β-2 and Cathepsin L genes.

For larger and more recent projects, the reassociation kinetic-based approach (Soares et al. 1994) has been highly successful. However, this technique must be carried out in a careful and controlled manner to avoid its tendency to select short, truncated, or internally primed inserts (Bonaldo et al. 1996), library artifacts that can artificially inflate the apparent diversity of the library, and number of genes identified (Ewing and Green 2000). The experience and care needed in constructing such high-quality, normalized libraries has precluded its more widespread use despite the seemingly easy principles belying the technique. Projects seeking to identify full-length cDNAs also have excluded this technique as a result of its tendency to reduce the average insert size, and instead have sequenced high-quality, nonnormalized cDNA libraries, paying the price of redundancy (Rubin et al. 2000).

Our comparison of the clones and sequences from the MPMGp609 and from two human libraries normalized by the Soares method shows the well-known drawbacks of this technique. The average insert size of the 1NIB and 1NFLS libraries are quite low compared to the MPMGp609 library. Many clones in these libraries consist wholly of 3′ UTR, and a noticeable proportion appear to be internally primed.

More recently, Carninci et al. (2000) have published a method for the normalization of full-length, enriched, cDNA libraries that were used in the RIKEN Mouse Gene Encyclopaedia Project (Kawai et al. 2001). The method used is similar to method 2 in Bonaldo et al. (1996), but uses trehalose thermostabilized reverse transcription and cap-trapped, first-strand cDNA to increase the quality of the starting cDNA; avoids the library amplification steps; and uses biotinylation and streptavidin binding instead of hydroxyapatite columns for hybrid capture. With the cap-trapping technique, the majority of the cDNAs appear to contain the 5′ end of the mRNA, adding important data for identifying the first exons of genes in the vertebrate genome. However, the high numbers of clusters (128,600 3′ clusters from 930,000 3′ ESTs) and the ability to collapse these clusters further after full-insert sequencing (the authors estimate that 21,076 selected clones may represent 12,890 genes) suggest that there still may be some problems with internally primed, truncated, or other problematic clones. The extent of any possible drawbacks of the technique are unclear, as any problems are likely to have been exacerbated by their choice of the shortest cDNAs (possibly truncated or internally primed) for easier full-insert sequencing.

Summary of Resource

Here we have used OFP as a prescreen for an EST project. This approach combines the strengths of both strategies. The nature of the OFP strategy allows deeper analysis of libraries and, for those transcripts for which a sequence is available, to match the cluster to the sequence based on its fingerprint, while EST sequencing allows homology searches and the design of PCR assays. Our combined use of the OFP and EST techniques allowed us to isolate the majority of all identified zebrafish genes as ESTs, without the sacrifices of library quality commonly associated with other normalization and subtraction techniques.

The identification of thousands of zebrafish genes, in-

cluding orthologs of human disease genes, opens new possibilities for zebrafish researchers and those wishing to use the power of the zebrafish system. The MPMGp609 library contains sequenced clones for the majority of identified zebrafish genes, and because ~30% of the sequencing reactions for MPMGp609 library failed, there must remain many as yet unidentified transcripts in the library. Thus, the MPMGp609 library is an essential tool for many projects, such as positional cloning, in situ screens, and microarray development.

Conclusion

With our extensive sequence data from the OFP-selected clones, we have been able to carry out the most thorough examination of the OFP technique to date, both for normalization success and any effects on the quality of the library. Our sequence data also allowed us to directly compare the MPMGp609 library with two of the most extensively characterized cDNA libraries from the Human Genome Project, normalized using a different technique. Our analysis shows that both the normalization and quality of the MPMGp609 library are high; the average insert size of clones from the human libraries are lower than the MPMGp609 library, and many more of the human clones consist wholly of 3′ UTR, with a noticeable proportion that appear to be internally primed. Our sequence analysis suggests that the high normalization and quality of the MPMGp609 library allows large numbers of ESTs to be identified as apparent zebrafish orthologs of human genes. Aligning these zebrafish ESTs to human genomic sequence also has shown that sometimes these are the only ESTs available for whole in silico-predicted genes, or at least the more 5′ exons.

With the need to verify in silico-predicted genes from the human genome with full-length mRNAs or ESTs, and with estimates that only 30% of human exons are covered by existing ESTs (Sunyaev et al. 1999), our approach offers several advantages. OFP allows the efficient screening of millions of clones from high-quality libraries, matching cDNA clones to identified or predicted mRNAs using their theoretical fingerprints (Meier-Ewert et al. 1998). Subsequent EST sequencing should uncover new exons and full-length cDNAs for further analysis.

METHODS

Oligonucleotide Probe Selection

At the inception of this project, the largest vertebrate data set of coding sequences was for human, while there was a low number of zebrafish sequences. Oligonucleotide probes were selected from a list of all possible 7- and 8-mer sequences, using an algorithm designed to generate the maximum information content for each hybridization, and best partition the human mRNA sequences from GenBank based on their theoretical fingerprints (Herwig et al. 2000). Impractical oligonucleotides were removed from the starting set by considering G/C and A/T content and palindromic and other problematic sequences.

Oligonucleotide Fingerprinting

OFP was carried out on high-density grids of 75,000 clones from a whole late somitogenesis embryonic cDNA library (RZPD library ICRFp524) and an adult liver cDNA library (RZPD library ICRFp532) with 223 different oligonucleotide probes of 7–11 bp. Each of the probes was hybridized separately under stringent conditions that allow single-base discrimination (Drmanac et al. 1990). The previously described steps involved in OFP include: cDNA library construction, arraying, cDNA array generation, oligonucleotide hybridization, data capture, image analysis, and clustering (Meier-Ewert et al. 1998; Clark et al. 1999).

Clone Clustering by Oligonucleotide Fingerprinting

Normalized hybridization data from high-quality hybridizations (assessed using 1920 sequenced M13 genomic shotgun clones) was used for clustering (Clark et al. 1999). For OFP1, the fingerprint of every clone was compared to every other clone, and clones with similarity scores above six standard deviations from the mean were clustered as cliques. Cliques then were merged into clusters if their member lists overlapped by 60% or more, similar to methods described Meier-Ewert et al. (1998). For OFP2, we used the method of Herwig et al. (1999).

Cluster Confirmation

Clusterings were checked by hybridizing cDNA inserts back onto the filters, scoring the results, and assessing whether clones from the same genes were clustered together. cDNAs representing the highly expressed (~1%), medium expressed (~0.1%), and lowly expressed (~0.01% or less) genes were used. In addition, some clusters were randomly chosen and all (or at least 20 randomly chosen) members were tag sequenced and assembled.

Rearraying

Clones in each cluster were ranked according to their similarity to the cluster's consensus fingerprint, and the clone with the highest similarity was chosen for rearraying. Rearraying was performed with a custom-made robot as previously described (Radelof et al. 1998). The resulting clone set is RZPD library MPMGp609 plates 1 through 64. We checked fidelity of the rearraying by picking one clone per quadrant (96 well section) from each 384 well plate of the rearrayed library and the corresponding original clone. These pairs of clones, original and rearrayed, were PCR amplified and the products run side by side on 1.2% agarose gels. Only four clone pairs from the 256 pairs analyzed (1.6%) gave different results: one clone pair (0.4%) had different sizes, while three clones (1.2%) had additional bands. Therefore, we judge the fidelity of rearraying to be at least 98%, though some of the disagreements could be in the handling of these clones for the comparison, rather than the fully automated step of rearraying.

Expressed Sequence Tag Sequencing

DNA preparation, data generation, and processing have been previously described (Hillier et al. 1996). All data was submitted to public databases within 24 h.

Expressed Sequence Tag Clustering and Sequence Analysis

ESTs (submitted to dbEST by August 8, 2000) with >200 bp of high-quality sequence were selected and masked with `RepeatMasker` (A. Smit, unpubl.; http://repeatmasker.genome.washington.edu/cgi-bin/RepeatMasker) before further analysis. 3′ ESTs were clustered with `Phrap` (http://www.phrap.org/), which has been used successfully for EST clustering (Ewing and Green 2000; Rubin et al. 2000). We used the following settings "-penalty -3 -gap_ext -3 -minmatch 70 -minscore 50 -maxgap 10 -confirm_length 50", which we empirically determined to be optimal settings reducing overclustering from unmasked or unknown repeats and very similar sequences such as recently duplicated paralogous genes.

Masked ESTs were searched against databases with WU-BLAST2 (W. Gish, unpubl.; http://blast.wustl.edu). Searches and databases were BLASTN to GenEmbl, BLASTX to SWISS-PROT+TREMBL, BLASTN to dbEST, and TBLASTX to Human UniGene consensus sequences. An *e*-value of −30 was used to indicate a stringent match against the SWISS-PROT+TREMBL databases. For the human unigene consensus and GenBank databases, an *e*-value of −20 was used to indicate a good match, and a possible ortholog if this was the best hit in the entire Human UniGene database.

Resource Availability

Libraries ICRFp532, ICRFp524, and the rearrayed library MPMGp609 are available from the RZPD (http://www.rzpd.de) as glycerol stocks in 384 well microtiter plates, spotted filters, or individual clones. DNA pools are available for PCR screening of the MPMGp609 library. Individual clones can be ordered by specifying the WashU clone name (e.g., fb16f02), or by BLAST searching a database of available clones. Many of the EST clones from the MPMGp609 library are part of an OFP cluster, and the additional clones from the cluster can also be ordered from the RZPD (see http://www.molgen.mpg.de/~ag_zebrafish/).

EST clones soon will be available through the IMAGE consortium distribution network using the IMAGE clone ID number attached to each dbEST entry.

ACKNOWLEDGMENTS

We thank Sebastian Meier-Ewert, Leo Schalkwyk, Michael van Wiles, Pia Aanstad, Georgia Panopoulou, and Albert Poustka for technical advice and helpful discussions. We also thank Dirk Schüdde, Elisabeth Brünke, and Pierre Emesberger for technical assistance, and Mario Drungowski for data analysis. This work was funded by the Imperial Cancer Research Fund, Max-Planck Society, German Human Genome Project (DHGP) and the NIH (grant RO1DK55379).

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

REFERENCES

Adams, M.D., Kerlavage, A.R., Fleischmann, R.D., Fuldner, R.A., Bult, C.J., Lee, N.H., Kirkness, E.F., Weinstock, K.G., Gocayne, J.D., White, O., et al. 1995. Initial assessment of human gene diversity and expression patterns based upon 83 million nucleotides of cDNA sequence. *Nature* **377:** 3–174.

Audic, S. and Claverie, J.M. 1997. The significance of digital gene expression profiles. *Genome Res.* **7:** 986–995.

Bairoch, A. and Apweiler, R. 2000. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res.* **28:** 45–48.

Barbazuk, W.B., Korf, I., Kadavi, C., Heyen, J., Tate S., Wun, E., Bedell, J.A., McPherson, J.D., and Johnson, S.L. 2000. The syntenic relationship of the zebrafish and human genomes. *Genome Res.* **10:** 1351–1358.

Bonaldo, M.F., Lennon, G., and Soares, M.B. 1996. Normalization and subtraction: Two approaches to facilitate gene discovery. *Genome Res.* **6:** 791–806.

Burge, C. and Karlin, S. 1997. Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.* **268:** 78–94.

Carninci, P., Shibata, Y., Hayatsu, N., Sugahara, Y., Shibata, K., Itoh, M., Konno, H., Okazaki, Y., Muramatsu, M., and Hayashizaki, Y. 2000. Normalization and subtraction of cap-trapper-selected cDNAs to prepare full-length cDNA libraries for rapid discovery of new genes. *Genome Res.* **10:** 1617–30.

Clark, M.D., Panopoulou, G.D., Cahill, D.J., Bussow, K., and Lehrach, H. 1999. Construction and analysis of arrayed cDNA libraries. *Methods Enzymol.* **303:** 205–233.

Craig, A.G., Nizetic, D., Hoheisel, J.D., Zehetner, G., and Lehrach, H. 1990. Ordering of cosmid clones covering the herpes simplex virus type I (HSV- I) genome: A test case for fingerprinting by hybridisation. *Nucleic Acids Res.* **18:** 2653–2660.

Deloukas, P., Schuler, G.D., Gyapay, G., Beasley, E.M., Soderlund, C., Rodriguez-Tome, P., Hui, L., Matise, T.C., McKusick, K.B., Beckmann, J.S., et al. 1998. A physical map of 30,000 human genes. *Science* **282:** 744–746.

Drmanac, R., Strezoska, Z., Labat, I., Drmanac, S., and Crkvenjakov, R. 1990. Reliable hybridization of oligonucleotides as short as six nucleotides. *DNA Cell Biol.* **9:** 527–534.

Drmanac, S., Stavropoulos, N.A., Labat, I., Vonau, J., Hauser B., Soares, M.B., and Drmanac, R. 1996. Gene-representing cDNA clusters defined by hybridization of 57,419 clones from infant brain libraries with short oligonucleotide probes. *Genomics* **37:** 29–40.

Ewing, B. and Green, P. 2000. Analysis of expressed sequence tags indicates 35,000 human genes. *Nat. Genet.* **25:** 232–234.

Gautheret, D., Poirot, O., Lopez, F., Audic, S., and Claverie, J.M. 1998. Alternate polyadenylation in human mRNAs: A large-scale analysis by EST clustering. *Genome Res.* **8:** 524–530.

Geisler, R., Rauch, G.J., Baier, H., van Bebber, F., Brobeta, L., Dekens, M.P., Finger, K., Fricke, C., Gates, M.A., Geiger, H., et al. 1999. A radiation hybrid map of the zebrafish genome. *Nat Genet.* **23:** 86–89.

Gong, Z., Yan, T., Liao, J., Lee, S.E., He, J., and Hew, C.L. 1997. Rapid identification and isolation of zebrafish cDNA clones. *Gene* **201:** 87–98.

Haas, S., Beissbarth, T., Rivals, E., Krause, A., and Vingron, M. 2000. Nov GeneNest: Automated generation and visualization of gene indices. *Trends Genet.* **16:** 521–523.

Herwig, R., Poustka, A.J., Muller, C., Bull, C., Lehrach, H., and O'Brien, J. 1999. Large-scale clustering of cDNA-fingerprinting data. *Genome Res.* **9:** 1093–1105.

Herwig, R., Schmitt, A.O., Steinfath, M., O'Brien, J., Seidel, H., Meier-Ewert, S., Lehrach, H., and Radelof, U. 2000. Information theoretical probe selection for hybridisation experiments. *Bioinformatics* **16:** 890–898.

Hillier, L.D., Lennon, G., Becker, M., Bonaldo, M.F., Chiapelli, B., Chissoe, S., Dietrich, N., DuBuque, T., Favello, A., Gish, W., et al. 1996. Generation and analysis of 280,000 human expressed sequence tags. *Genome Res.* **6:** 807–828.

Hukriede, N.A., Joly, L., Tsang, M., Miles, J., Tellis, P., Epstein, J.A., Barbazuk, W.B., Li, F.N., Paw, B., Postlethwait, J.H., et al. 1999. Radiation hybrid mapping of the zebrafish genome. *Proc. Natl. Acad. Sci.* **96:** 9745–9750.

Kawai, J., Shinagawa, A., Shibata, K., Yoshino, M., Itoh, M., Ishii, Y., Arakawa, T., Hara, A., Fukunishi, Y., Konno, H., et al. 2001. Functional annotation of a full-length mouse cDNA collection. *Nature* **409:** 685–690.

Kupperman, E., An, S., Osborne, N., Waldron, S., and Stainier, D.Y. 2000. A sphingosine-1-phosphate receptor regulates cell migration during vertebrate heart development [see comments]. *Nature* **406:** 192–195.

Marra, M.A., Hillier, L., and Waterston, R.H. 1998. Expressed sequence tags—ESTablishing bridges between genomes. *Trends Genet.* **14:** 4–7.

McPherson, J.D., Marra, M., Hillier, L., Waterston, R.H., Chinwalla, A., Wallis, J., Sekhon, M., Wylie, K., Mardis, E.R., Wilson, R.K., et al. 2001. A physical map of the human genome. *Nature* **409:** 934–941.

Meier-Ewert, S., Lange, J., Gerst, H., Herwig, R., Schmitt, A., Freund, J., Elge, T., Mott, R., Herrmann, B., and Lehrach, H. 1998. Comparative gene expression profiling by oligonucleotide fingerprinting. *Nucleic Acids Res.* **26:** 2216–2223.

Meier-Ewert, S., Maier, E., Ahmadi, A., Curtis, J., and Lehrach, H. 1993. An automated approach to generating expressed sequence catalogues. *Nature* **361:** 375–376.

Milosavljevic, A., Zeremski, M., Strezoska, Z., Grujic, D., Dyanov, H., Batus, S., Salbego, D., Paunesku, T., Soares, M.B., and Crkvenjakov, R. 1996. Discovering distinct genes represented in 29,570 clones from infant brain cDNA libraries by applying sequencing by hybridization methodology. *Genome Res.* **6:** 132–141.

Okubo, K., Hori, N., Matoba, R., Niiyama, T., Fukushima, A., Kojima, Y., and Matsubara, K. 1992. Large scale cDNA sequencing for analysis of quantitative and qualitative aspects of gene expression. *Nat. Genet.* **2:** 173–179.

Parichy, D.M., Ransom, D.G., Paw, B., Zon, L.I., and Johnson, S.L. 2000. An orthologue of the kit-related gene fms is required for development of neural crest-derived xanthophores and a subpopulation of adult melanocytes in the zebrafish, *Danio rerio*.

Development **127:** 3031–3044.
Pesole, G., Grillo, G., and Liuni, S. 1996. Databases of mRNA untranslated regions for metazoa. *Comput. Chem.* **20:** 141–144.
Pesole, G., Liuni, S., Grillo, G., Licciulli, F., Larizza, A., Makałowski, W., and Saccone, C. 2000. UTRdb and UTRsite: Specialized databases of sequences and functional elements of 5′ and 3′ untranslated regions of eukaryotic mRNAs. *Nucleic Acids Res.* **28:** 193–196.
Poustka, A.J., Herwig, R., Krause, A., Hennig, S., Meier-Ewert, S., and Lehrach, H. 1999. Toward the gene catalogue of sea urchin development: The construction and analysis of an unfertilized egg cDNA library highly normalized by oligonucleotide fingerprinting. *Genomics* **59:** 122–133.
Radelof, U., Hennig, S., Seranski, P., Steinfath, M., Ramser, J., Reinhardt, R., Poustka, A., Francis, F., and Lehrach, H. 1998. Preselection of shotgun clones by oligonucleotide fingerprinting: an efficient and high throughput strategy to reduce redundancy in large-scale sequencing projects. *Nucleic Acids Res.* **26:** 5358–5364.
Rubin, G.M., Hong, L., Brokstein, P., Evans-Holm, M., Frise, E., Stapleton, M., and Harvey, D.A. 2000. A *Drosophila* complementary DNA resource. *Science* **287:** 2222–2224.
Soares, M.B., Bonaldo, M.F., Jelene, P., Su, L., Lawton, L., and Efstratiadis, A. 1994. Construction and characterization of a normalized cDNA library. *Proc. Natl. Acad. Sci.* **91:** 9228–9232.
Sunyaev, S., Hanke, J., Aydin, A., Wirkner, U., Zastrow, I., Reich, J., and Bork, P. 1999. Prediction of nonsynonymous single nucleotide polymorphisms in human disease-associated genes. *J. Mol. Med.* **77:** 754–760.
Suzuki, Y., Yoshitomo-Nakagawa, K., Maruyama, K., Suyama, A., and Sugano, S. 1997. Construction and characterization of a full length-enriched and a 5′-end-enriched cDNA library. *Gene* **200:** 149–156.
Velculescu, V.E., Zhang, L., Vogelstein, B., and Kinzler, K.W. 1995. Serial analysis of gene expression. *Science* **270:** 484–487.
Wang, H., Yan, T., Tan, J.T., and Gong, Z. 2000. A zebrafish vitellogenin gene (vg3) encodes a novel vitellogenin without a phosvitin domain and may represent a primitive vertebrate vitellogenin gene. *Gene* **256:** 303–310.
Zhang, M.Q. 1997. Identification of protein coding regions in the human genome by quadratic discriminant analysis. *Proc. Natl. Acad. Sci.* **94:** 565–568.

Received March 1, 2001; accepted in revised form June 12, 2001.