doi:10.1006/geno.2002.xxxx, available online at http://www.idealibrary.com on IDEAL

DEFOG: A Practical Scheme for Deciphering Families of Genes

Tania Fuchs,[1,*] Barbora Malecova,[2,*] Chaim Linhart,[3] Roded Sharan,[3] Miriam Khen,[1] Ralf Herwig,[2] Dmitry Shmulevich,[3] Rani Elkon,[4] Matthias Steinfath,[2] John K. O'Brien,[5] Uwe Radelof,[2] Hans Lehrach,[2] Doron Lancet,[1] and Ron Shamir[3,†]

*[1]Department of Molecular Genetics and the Crown Human Genome Center, The Weizmann Institute of Science, Rehovot **POSTAL CODE?**, Israel*
[2]Max-Planck Institut für Molekulare Genetik, Ihnestrasse 73, D-14195 Berlin, Germany
[3]School of Computer Science, Sackler Faculty of Exact Sciences, Tel-Aviv University, Tel Aviv 69978, Israel
[4]The David and Inez Laboratory for Genetic Research, Department of Human Genetics and Molecular Medicine, Sackler School of Medicine, Tel-Aviv University, Tel-Aviv 69978, Israel
[5]Department of Clinical Pharmacology, Royal College of Surgeons in Ireland, 123 St. Stephen's Green, Dublin 2, Ireland

**These authors contributed equally to this work.*

[†]To whom correspondence and reprint requests should be addressed. Fax: +972-3-6405384. E-mail: rshamir@post.tau.ac.il.

We developed a novel efficient scheme, DEFOG (for "deciphering families of genes"), for determining sequences of numerous genes from a family of interest. The scheme provides a powerful means to obtain a gene family composition in species for which high-throughput genomic sequencing data are not available. DEFOG uses two key procedures. The first is a novel algorithm for designing highly degenerate primers based on a set of known genes from the family of interest. These primers are used in PCR reactions to amplify the members of the gene family. The second combines oligofingerprinting of the cloned PCR products with clustering of the clones based on their fingerprints. By selecting members from each cluster, a low-redundancy clone subset is chosen for sequencing. We applied the scheme to the human olfactory receptor (OR) genes. OR genes constitute the largest gene superfamily in the human genome, as well as in the genomes of other vertebrate species. DEFOG almost tripled the size of the initial repertoire of human ORs in a single experiment, and only 7% of the PCR clones had to be sequenced. Extremely high degeneracies, reaching over a billion combinations of distinct PCR primer pairs, proved to be very effective and yielded only 0.4% nonspecific products.

Key Words: [AUTHORS: please include up to 10 key words]

INTRODUCTION

The study of large gene families allows a panoramic view of molecular evolution within and across species and contributes to the fields of functional and comparative genomics. A gene superfamily is a cluster of evolutionarily related sequences sharing a common ancestor [1], and consists of homologous gene families. The largest superfamily in human, and probably the largest in the genomes of all vertebrate species, is the olfactory receptor (OR) gene superfamily, which is a part of the G-protein coupled receptor (GPCR) hyperfamily [2,3].

OR genes encode seven-transmembrane-domain proteins [2,4,5], which are responsible for the recognition and differentiation of millions of odorants. OR genes were first characterized in rat a decade ago [2] and have since been detected in many vertebrate species [reviewed in 6]. So far, about 1400 OR genes and pseudogenes are known in 24 vertebrate species [7–11]. They are divided into 32 distinct families based on phylogenetic analysis [8].

Roughly 900 OR coding sequences were found in the first draft of the human genome [9,10]. As predicted [7,12], between 53% and 63% of them have frame disruptions and are therefore considered as pseudogenes. OR genes are found on almost all human chromosomes except chromosome 20 and Y [7,9,10,12,13]. Almost 80% of the ORs are organized in clusters of six or more genes [9,10]. This is in good agreement with previous fluorescence *in situ* hybridization (FISH) experiments and sequencing data [7,12,13].

The coding region of OR genes spans approximately 1 kb. This region lacks introns [2,12] and is preceded by a large

doi:10.1006/geno.2002.xxxx, available online at http://www.idealibrary.com on IDEAL

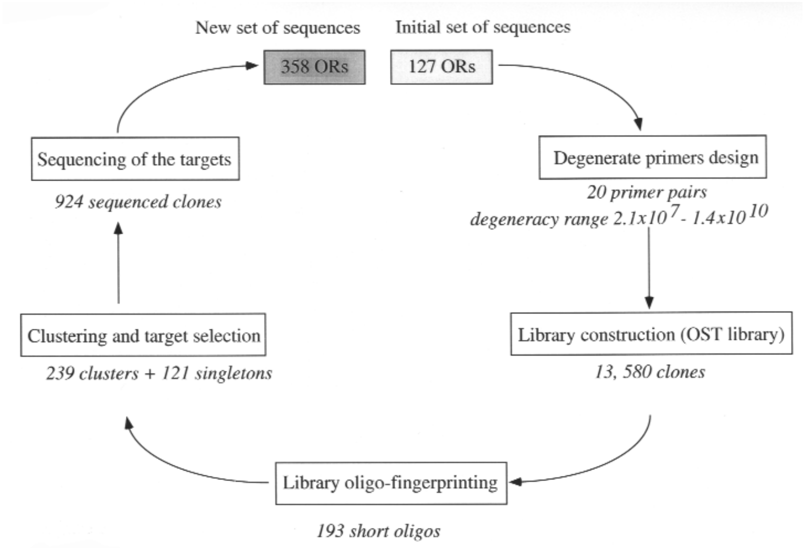
New set of sequences
Initial set of sequences
358 ORs
127 ORs
Degenerate primers design
20 primer pairs
degeneracy range 2.1x10^7 - 1.4x10^10
Library construction (OST library)
13, 580 clones
Library oligo-fingerprinting
193 short oligos
Clustering and target selection
239 clusters + 121 singletons
Sequencing of the targets
924 sequenced clones

FIG. 1. The DEFOG scheme. Starting from a known set of genes from the family of interest, degenerate primers are designed by an ad-hoc algorithm in order to amplify as many of these genes as possible, given a bound of the allowed degeneracy. Primer pairs are used to amplify genomic fragments in PCR reactions. The fragments are cloned, spotted on filters, and oligo-fingerprinted. The fingerprints are clustered using CLICK, a novel clustering algorithm, and representatives of the clusters are selected for sequencing. The numbers below each box indicate the actual results in each step in our study of human ORs. In particular, a single cycle of the DEFOG scheme increased the number of OR genes from 127 to 358.

intron and several short noncoding exons [15–19]. Inside the coding region there are several conserved segments [20] that allow easy amplification of the intronless coding region from genomic DNA by PCR assay. The PCR product is termed the "olfactory receptor sequence tag" (OST) [7].

We developed and experimentally tested a practical scheme for deciphering families of genes (DEFOG). The scheme provides a powerful means to obtain a sequence composition of a gene family in species for which high-throughput genomic sequencing data are unavailable. To validate DEFOG, we tested it on the human OR gene superfamily. Starting with a limited number of human ORs, it almost tripled the size of this set in a single experiment. We suggest that DEFOG can be successfully applied to ORs and other gene families in the genomes of various species.

Results

The DEFOG Scheme

The DEFOG scheme (Fig. 1) begins with the design of a set of degenerate primer pairs specific for the target gene family. Primers are called "degenerate" if they contain at some positions more than one possible nucleotide. The degeneracy of a primer is the number of sequences it would match perfectly. The design is performed by the degenerate primers design (DPD) algorithm, which aims at striking a balance between degeneracy and specificity of the primers. PCR reactions are then carried out with these primers on a template DNA. In the next step, an oligofingerprinting (OFP) process [21–25] is used to characterize clones by their patterns of hybridization with a series of up to 200 short oligonucleotides. The pattern of hybridization for each clone is called its fingerprint. The resulting clone fingerprints are then clustered using the CLICK algorithm [26], which groups the fingerprints to homogeneous and well-separated clusters, based on their pairwise similarity. Each cluster is then assumed to represent a single gene. Finally, representatives from each cluster are chosen for sequencing.

Deciphering the Human OR Gene Superfamily

We experimentally tested the DEFOG scheme on the human OR gene superfamily. Starting with an initial collection of 127 OR genes known at that time [7], we designed OR-specific degenerate primers with degeneracy ranging between 4600 and 442,000 (Table 1). The resulting effective primer pairs had combined degeneracy ranging between 2.1×10^7 and 1.4×10^{10} (Table 2). The intronless coding region of OR genes allows their amplification by PCR from genomic DNA. The primers were used in a series of PCR reactions on human genomic DNA [7]. We applied to the library of PCR products, consisting of 13,580 clones, an oligonucleotide fingerprinting procedure with a set of 193 8-mer oligonucleotide probes [21–25]. Cluster analysis of the clone fingerprints [26] revealed 239 clusters and 121 singletons (single clone clusters). Based on the clustering results, we selected 1058 clones for sequencing. Of 924 clones that we successfully sequenced, 4 did not belong to ORs and the remaining 920 clones represented 300 OR genes and pseudogenes. This procedure revealed a third of the entire human OR collection [9] in a single experiment. Of these 300 OR sequences, only 69 are encompassed in the training set of 127 ORs used for primer design. Thus, we almost tripled the size of the initial collection of human OR genes (from 127 to 358). The pseudogene proportion in the sequenced set was about 55%, similar to the ratio in the entire OR repertoire [9,10]. The family distribution (Fig. 2) shows that there was no cloning bias towards a certain closely related group of sequences. Low representation of families 51, 52, 55, and 56 is probably due to the low sequence similarity of these genes to the genes of the other OR families [9].

Design of Degenerate Primers

Given a set of genes from a gene family of interest, DPD provides a means to design degenerate primers that embody an efficient balance between degeneracy and specificity. The basic goal was to design oligonucleotides with high degen-

TABLE 1: Degenerate primers designed using the training set of 127 fully-known human OR genes

Side	Name	Primer sequence	Degeneracy	Two-mismatches coverage
5′	L5	CTNCAHWCNCCHATGTAYTTYYTYCT	4608	107 (84%)
	L9	ACNNTGVTNGGVAAYCTNCTCATYAT	9216	59 (46%)
	L10	CTBCAYDNNCCHATGTAYTTYTTBCT	10368	112 (88%)
	L20	CTYCANDVHCCCATGTAYYWYTTYBT	20736	110 (87%)
	L31	CTBCAYDNNCCHATGTAYTTBTTBYT	31104	114 (90%)
	L131	CTNCANWCNCCNATGTAYTTNYTNCTN	131072	110 (87%)
3′	R5	TTYCTCARRSTRTADATNADNGGGTT	4608	97 (76%)
	R20	TGKGABVHACANGTGBWRARRGCYTT	20736	79 (62%)
	R28	TTBCKNARRSTRTADATVARRGGRTT	27648	105 (83%)
	R73	TTBCKNARRSTRTADATNANRGGRTT	73728	109 (86%)
	R110	YNCAGDRCHCYYTTNAYDTCYYTRTT	110592	57 (45%)
	R147	RTTBCKNARNSTRTADATNARNGGGTT	147456	105 (83%)
	R442	TTBCKNARRSTRTADATNANDGRRYT	442368	113 (89%)

The last column specifies the number (percentage) of genes (out of 127) that each primer matches with up to two mismatched base pairs. L9 and R110 were designed at different positions (transmembrane segments TM1 and end of TM7) than the others (TM2 and TM7). L20 and R20 were designed on a subset of genes that were poorly matched by the rest of the primers.

eracy so that they would amplify a maximum number of novel genes. At the same time we strove to retain specificity, so that most of the amplified sequences will belong to the desired gene superfamily.

We implemented a three-phase program, DPD, for designing effective degenerate primers. To find a good primer of length k and degeneracy d, DPD first extracts non-degenerate primer candidates from the input DNA sequences (the training set). It does so by scoring k-long subsequences appearing in the training set. For the 5′ (respectively, 3′) primer, we took the subsequences from the first (respectively, last) 300 bp of each OR gene. For each such subsequence, its best matching word (in terms of gapless local alignment) in every training sequence is located, and those words form a block, or a matrix, for which an information-based (entropy) score is computed. The subsequences with the best scoring blocks are selected as primer candidates. In the second phase, each candidate is expanded to a d-degenerate primer using an iterative procedure. In each iteration, this procedure adds one new nucleotide possibility at a single position. The nucleotide is chosen according to a score, which is based on the column distribution of the block induced by the candidate and on the number of additional genes the new primer matches. Other primers are generated using an opposite approach that starts with a completely degenerate primer candidate (with all four possible nucleotides at each position), and iteratively removes nucleotides at degenerate positions. In the third phase of DPD, a greedy hill climbing function improves the primers by repeatedly trying to replace each nucleotide with a different one until no replacement increases the number of genes the primer matches. Finally, the best primer found in the third phase is reported. Although there is no guarantee that DPD will find an optimal solution (that is, a primer of length k and degeneracy at most d that covers the largest number of genes), the results it produced in practice were quite satisfactory.

We executed DPD on a set of 127 known human OR genes and designed 13 primers: 6 for the 5′ side, and 7 for the 3′ side (Table 1). Primers were of length k = 26 or k = 27 and degeneracy was between 4608 and 442,368. DPD automatically selected most 5′ and 3′ primers in the regions corresponding to the transmembrane segments 2 and 7 of the protein, respectively. This fact reflects the significant conservation of OR protein sequence at these regions. Exceptions are four primers: two of which we deliberately designed at different positions (L9 in TM1 and R110 at the very end of TM7); and L20 and R20, which we designed on a subset of genes that were poorly matched by all the other primers.

Performance Analysis of Primers

To evaluate the quality of the primers, we used the following theoretical model. A primer pair was assumed to successfully amplify the genes it matched with no more than three errors, in both sides combined. This model ignores many important factors that influence PCR, such as the positions of the mismatches, but it provided a fairly good approximation: of 69 genes in the training set that we sequenced (data not shown), 68 genes matched the primers according to the above criterion.

According to the three-mismatches criterion, most of the primer pairs we designed covered 70–80% of the training set of 127 known OR genes. Because all but three of the OST sequences obtained by DEFOG correspond to the recently published full-length sequences [9], we could carry out an analysis for practically all the OSTs reported here. To this end, we used the HORDE human OR database (http://bioinformatics.weizmann.ac.il/HORDE), which contains 719 full-length coding regions of OR genes. We found that most primer pairs matched 50–60% of the 719 genes in this test set.

Figure 3A shows the theoretical coverage obtained by primer pairs, that is, the percentage of matched genes according to the three-mismatches model, as a function of their combined degeneracy (that is, the degeneracy of both primers multiplied). To perform an unbiased comparison, we only included primers of 26 bp that were designed on the whole training set at positions chosen automatically by DPD (that is, we used only primers L5, L10, L31, R5, R28, R73, and R442). As expected, primers with higher degeneracy match more

TABLE 2: Analysis of the 20 degenerate primer pairs

Primers	Degeneracy $\times10^6$	Coverage of 127	Coverage of 719	No. clones in library	No. clones sequenced	No. genes	Genes/ clones	No. new genes	% new genes	No. clusters
L5/R5	21	73%	50%	1730	173	98	0.57	73	74%	144
L10/R5	48	74%	51%	838	42	31	0.74	24	77%	76
L5/R28	127	74%	52%	901	75	50	0.67	36	72%	99
L9/R20	191	31%	13%	431	43	25	0.58	14	56%	52
L10/R28	287	74%	53%	740	57	39	0.68	28	72%	90
L5/R73	340	77%	60%	566	34	27	0.79	17	63%	82
L5/R110	510	51%	30%	598	31	22	0.71	19	86%	58
L31/R20	645	66%	47%	352	65	45	0.69	40	89%	71
L9/R110	1019	29%	11%	621	19	15	0.79	11	73%	29
L9/R147	1359	48%	21%	973	42	34	0.81	20	59%	56
L10/R147	1529	77%	55%	660	53	42	0.79	34	81%	91
L5/R442	2038	79%	63%	649	46	38	0.83	32	84%	89
L31/R73	2293	80%	62%	1033	27	25	0.93	18	72%	88
L20/R147	3058	77%	51%	747	67	43	0.64	34	79%	104
L31/R110	3440	55%	31%	426	25	21	0.84	19	90%	38
L131/R28	3624	76%	57%	181	14	12	0.86	11	92%	49
L9/R442	4077	54%	26%	748	28	20	0.71	14	70%	43
L10/R442	4586	80%	63%	691	46	37	0.80	26	70%	93
L31/R147	4586	78%	56%	564	28	26	0.93	18	69%	69
L31/R442	13759	82%	65%	131	9	8	0.89	6	75%	37

"Degeneracy" is the combined degeneracy, in millions, of the primer pair. The two "Coverage" columns specify the percentage of genes, out of the training set (127 genes) and the HORDE database (719 genes), respectively, that match the primer pair with up to three mismatched bases. "No. clones in library" is the number of clones we obtained in our OST library, and "No. clones sequenced" is the number of target clones that were successfully sequenced. "No. genes" is the number of distinct genes each primer pair yielded. "Genes/clones" is the sequencing redundancy, that is, the ratio between the number of distinct genes and the number of clones we sequenced. "No. new genes" is the number of unique new genes we found, that is, that were not in the training set, and "% new genes" is the percentage of this number out of the total number of genes (old and new) we obtained with the primer pair. "No. clusters" is the number of clusters that the clones obtained from a specific primer pair belong to.

genes, both in the training set and in the HORDE test set.

Table 2 summarizes the empirical performance of the 20 primer pairs that we used in the experiment. Highly degenerate primers gave a very high proportion of new genes, but they sometimes yielded a small number of OST clones. By the sequencing efficacy of a pair of primers we mean the ratio between the number of different genes found and the number of clones successfully sequenced using them. For almost all primer pairs with combined degeneracy of over one billion, sequencing efficacy was 0.79–0.93, whereas for primers with lower degeneracy, it was 0.57–0.79. In comparison, the sequencing efficacy in the entire experiment was 0.32, demonstrating that the overlap in PCR products from distinct primer pairs is rather low. Figure 3B shows the sequencing efficacy as a function of the combined degeneracy of the primers (again, we only included primers of 26 bp that were designed on the whole training set at the default positions). These results also include 140 clones from six clusters, which we sequenced merely to obtain statistics for clustering analysis, so the real efficacy is even higher.

As an additional posterior analysis, we ran DPD on several training sets of different sizes. We ran DPD on 719 genes available from the HORDE database (data not shown), generating primer pairs with degeneracy ranging between 2.1×10^7 and 1.2×10^{10}. These primers cover 52–67% of the genes, depending on their degeneracy: only 2% more than primers with similar degeneracies designed on the 127 genes in the training set. Hence, the training set of 127 genes represented the full OR repertoire. Moreover, to estimate how well DEFOG would work on small training sets, we ran DPD on random subsets of 20, 30, and 40 OR genes, and computed the coverage of the obtained primers with respect to the full 719 set. On average, primers that were designed on subsets of size 20 matched 36–50% of the genes—15% fewer genes than primers with similar degeneracies that were constructed using the 127 genes training set. For subsets of size 30, the average difference was only 8%, and for size 40, only 5%. Thus, the DEFOG scheme does not require a large training set: in the case of the human OR subgenome, a random set of about 30 genes would have yielded very good results.

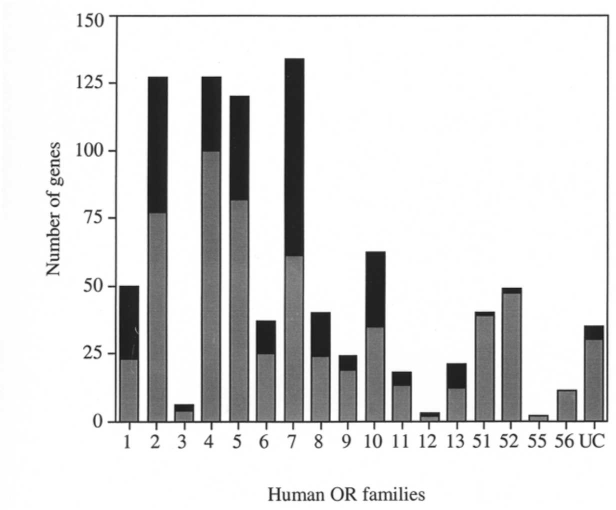

FIG. 2. Comparison of the number of ORs found by DEFOG with the total number of ORs known in each family **[AUTHORS: OK?]**. The partitioning into families is as previously published [8]. Bars show the total number of genes in each one of the 17 known families and unclassified (UC). The fraction detected by DEFOG is shown in black.

Clustering Evaluation

The clustering solution produced by the CLICK algorithm [26] contains 239 clusters and 121 singletons. About two-thirds of the clusters have a size of at most 20, and 86% have a size of at most 100. Assuming that this reflects the size distribution of gene clusters, it indicates that the designed primers were not biased towards any specific set of sequences.

Evaluation of the solution's quality was based on the annotation information for the sequenced clones. In total 920 sequenced clones had matches in the HORDE database, providing us with a true subclustering against which CLICK's solution could be compared. We carried out the evaluation in two steps.

We first computed the specificity and sensitivity of our solution with respect to the annotated clones. For a given clustering, we defined a pair of elements as mates if they belonged to the same cluster. Specificity is defined as the percentage of true mates (that is, mates in the true clustering) out of the total number of mates identified by our solution. Sensitivity is defined as the percentage of true mates identified by our solution out of the total number of true mates. For the 920 annotated clones, the clustering specificity was 0.57, and the sensitivity was 0.74. When computing these measures only for the 292 annotated clones that were closest to the centers of their clusters, the specificity increased to 0.87 and the sensitivity was 0.73. Since our goal was to discover as many genes as possible, we used stringent thresholds for the clustering, ensuring high specificity at the possible expense of splitting some true clusters.

Next, we examined the composition of fully annotated clusters. In choosing the clones for sequencing, we fully sequenced six medium-sized clusters. For each such cluster we could record the distribution of its members according to their annotation. We computed for each cluster its mean homogeneity, defined as the mean similarity (vector dot-product) between an element and the mean fingerprint of its assigned cluster (Fig. 4). Because these clusters were fully sequenced, we could also compute the purity of each cluster, that is, the fraction of its clones that belong to the most abundant annotation class in it. When comparing the purity of each of these clusters with its mean homogeneity (Fig. 4), a clear correlation can be observed: the higher the homogeneity, the greater the purity of the cluster. For small clusters that were also fully sequenced, we obtained the following statistics. Of 21 such clusters of size 3, 17 were pure (that is, their three members had the same annotation). Of 28 clusters of size 2 whose members were all annotated, 20 were pure.

Discussion

The analysis of gene families in one species, as well as across species, provides a powerful tool for molecular evolution studies [27–31]. The common approach for sequencing a gene family is to carry out PCR on genomic DNA or a cDNA library. Obviously, in this case one should use degenerate

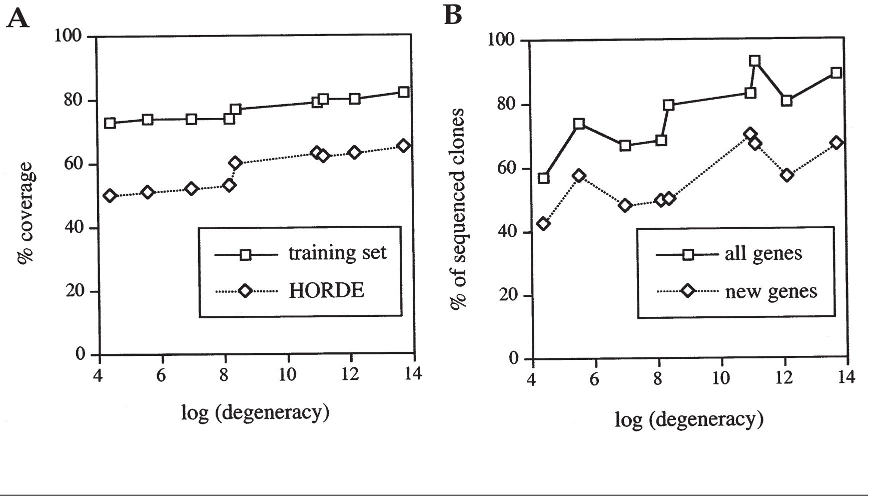

FIG. 3. Theoretical and empirical performance of primer pairs as a function of their combined degeneracy (only primers of length 26 bp designed on transmembrane domains TM2 and TM7 using the whole training set are included). (A) The percentage of genes matching each primer pair with up to three errors in the training set of 127 genes and in the HORDE test set of 719 genes. (B) Sequencing efficacy: the percentage of different genes and of different new genes that were found using each primer pair, out of the total number of clones sequenced for that pair.

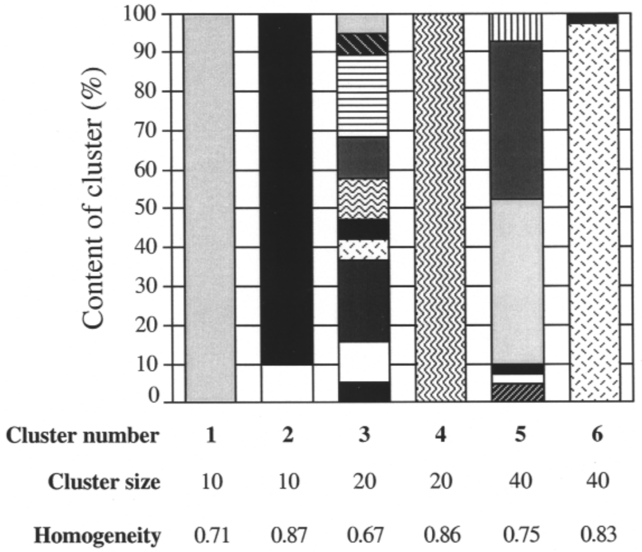

FIG. 4. Composition and homogeneity of the six medium-sized fully sequenced clusters. The composition of each cluster is displayed by a corresponding bar, in which each segment represents a distinct OR gene, and the size of the segment represents the percentage of clones corresponding to that gene. Below each are bar shown the size and mean homogeneity of the cluster (the mean similarity between the fingerprint of an element and the average fingerprint of the cluster).

primers to minimize the number of PCR reactions [32]. However, the use of such highly degenerate primers increases the probability of nonspecific matches during the PCR reaction. Furthermore, PCR-based libraries, as well as cDNA and shotgun libraries, usually have high redundancy (on average about 140-fold [24,33]), which makes a straightforward exhaustive sequencing of them inefficient and expensive.

The DEFOG scheme, which we developed and tested on the human OR gene superfamily, aims to overcome the aforementioned difficulties as follows. First, we use a sophisticated primer design procedure for constructing degenerate primers with a good balance between degeneracy and specificity. Second, we combine OFP technology and cluster analysis to obtain many different genes by low-redundancy sequencing.

The degeneracy of the primer pairs used in the described experiments ranged between 2.1 3 10^7 and 1.4 × 10^{10}—perhaps the highest degeneracy ever used successfully in PCR reactions [33]. Nevertheless, only 0.4% (4 of 924) of the sequenced clones carried genes other than ORs. Moreover, the primers were not strongly biased towards preferentially amplifying specific genes as shown by their phylogenetic family distribution (Fig. 2). Moreover, two-thirds of the gene clusters we obtained included no more than 20 clones, and 300 different genes were detected.

Obviously, primer pairs with very high degeneracy could match many genes both theoretically and practically (Figs. 3A and 3B). However, they yield a smaller number of products in PCR reactions. This is probably due to the low concentration of each single primer in the mixture during the PCR amplification. Indeed, the most degenerate primer pair we used (L31/R442, with redundancy of 1.4 × 10^{10}) yielded the smallest number of OST clones: 131 versus an average of 724 per primer pair in the whole experiment (Table 2). We experimented with primers with even higher degeneracies, but they usually gave very poor yield (data not shown). In addition to the higher coverage, primers with higher degeneracy also allow higher sequencing efficacy: for most primer pairs with very high degeneracy (109 or more), the ratio between the number of different genes found and the number of clones sequenced was 0.79 to 0.93, whereas for primers with lower degeneracy, this ratio was 0.57 to 0.79.

The pseudogene proportion in the sequenced set was about 55%, which is in good agreement with the ratio shown for the entire OR repertoire [9,10]. This suggests that the primer selection method did not create a bias of genes versus pseudogenes, which could result from higher sequence conservation in intact genes.

In PCR-based libraries, many clones may contain the same gene. To overcome such redundancy, we applied oligonucleotide fingerprinting (OFP). This technique was previously used successfully on cDNA and shotgun libraries [21–25]. In OFP, each clone of the library obtains a unique fingerprint of its hybridization pattern with a set of short oligonucleotides. Then, cluster analysis [26] was used to group clones with similar fingerprints. Each cluster ideally contains the clones corresponding to a single gene. We next chose representatives from each cluster for sequencing. In total, we sequenced 924 clones, spanning 300 different OR genes, implying a very low sequencing redundancy of three sequences per gene, on average. To obtain these 300 genes by random sequencing of clones from the library, almost all 13,580 clones should have been sequenced.

Cluster analysis was the main tool in reducing the sequencing redundancy, by allowing the selection of a relatively small number of sequences that represent (almost) all amplified genes. The evaluation of the clustering solution based on the annotation of sequenced clones indicated that the computed clusters were homogeneous, in spite of the high similarity among OR genes. In particular, out of six medium-sized fully sequenced clusters, four were pure or almost pure, and two were composed of two major classes, each containing clones corresponding to a certain gene (Fig. 4). To further understand the reason for the two impure clusters, we compared the mean fingerprints of the two major classes in each of these clusters. The comparison gives a partial explanation to the mixed clusters: the mean fingerprints of the two classes in each cluster are very similar (their dot product was 0.7 out of a maximum of 1), which is in part due to sequence similarity (data not shown).

The DEFOG experiment revealed 300 OR genes and pseudogenes, which is over 40% of the recently published full human OR repertoire [9,10]. Of these only 69 were encompassed in the training set of 127 ORs used for the primers design. Therefore, DEFOG enlarged the training set by 180% (from 127 to 358). This implies that in species lacking large-scale sequencing projects (like the Human Genome Project), our scheme can be a good solution for sequencing gene families.

We suggest that the DEFOG scheme can be successfully applied to other gene families. First and foremost, OR repertoires of various species such as mouse, cat, or dog can be revealed, which would allow a scrupulous study of the evolution of this prominent superfamily. In light of the high prevalence of pseudogenes in the human OR repertoire, determining the mRNA expression levels of different human OR genes is of interest. DEFOG can be easily applied to the cDNA libraries from olfactory epithelium of human, as well as other species. The scheme can also be applied to other gene families that exhibit sequence conservation (for example, G-protein coupled receptors, protein kinases, aldehyde dehydrogenase (ALDH) gene superfamily, etc.), as well as to families of genes sharing a common domain. The DPD algorithm can also be used to amplify a specific gene in different organisms, based on a set of known homologous sequences.

The DEFOG scheme can be improved in several ways. The degenerate primer design algorithm can be extended to construct several primer pairs in a single run, rather than just one, in order to fully cover the training set (that is, amplify all the known genes). Another extension could be to improve the primer–gene matching model, by assigning different weights to mismatches at different positions (for example, a mismatch at the 3′ end of a 5′ primer is more destructive than a mismatch at the 5′ end) [27]. The OFP process can be improved by choosing specific oligonucleotides, which differentiate between the studied sequences [34]. In addition, we can improve the process of selecting target clones for sequencing by choosing fewer targets from highly homogeneous clusters, and more targets from clusters with low homogeneity, in order to span more genes and save on sequencing.

Methods

Degenerate primers design. We designed degenerate primers using a novel algorithm we developed for this purpose. The design was based on 127 full-length human OR coding regions known at that time [7]. We designed 13 primers: 6 for the 5′ side and 7 for 3′ side. The length of the primers was 26 bp or 27 bp and their degeneracy ranged from 4608 to 442,368 (Table 1). The DPD program and all primer sequences used in this study are available on request (chaiml@post.tau.ac.il).

Generation of OST libraries. We performed PCRs with 20 different combinations of primer pairs (Table 2). (We plated, subcloned, and sequenced individually 20 libraries.) Reactions were carried out in a total volume of 25 µl, containing 0.2 mM concentration of each deoxynucleotide (Promega, Madison, WI), 50 pmol of each primer, PCR buffer containing 1.5 mM $MgCl_2$, 50 mM KCl, 10 mM Tris, pH 8.3, 1 unit of *Taq* DNA polymerase (Boehringer Mannheim, Mannheim, Germany), and 50 ng genomic DNA. PCR conditions were as follows: 35 cycles of 1 minute at 94°C, 1 minute at 55°C, and 1 minute at 72°C. The first step of denaturation and the last step of extension were each 3 minutes at 94°C and 72°C, respectively.

Primers used were modified for subsequent subcloning into the pAMP1 vector. The PCR products were subcloned into the pAMP1 vector, without prior purification, using Clone Amp System (Gibco BRL) and DH5 bacterial competent cells (Gibco BRL).

The bacterial suspension was plated out on 22 cm × 22 cm LB-agar plates containing ampicillin, X-gal, and IPTG for blue/white selection of recombinant clones. Well-separated, white colonies were picked by a robotic picking system [35,36]. In total 13,580 colonies were transferred into 384-well microtiter plates (Genetix) containing 2YT medium, ampicillin, and 7.5% glycerol. After incubation at 37°C overnight, plates were replicated and stored at –80°C for further use.

PCR amplification of OST clones. The hybridization of short oligonucleotides requires large amounts of high-purity target DNA. We carried out PCR amplifications in 384-well microtiter plates (Perkin Elmer). Using disposable plastic 384-pin inoculation devices (Genetix), a small amount of the bacterial suspension was added to a 25 µl reaction volume containing 50 mM KCl, 15 mM Tris-HCl (pH 8.5), 35 mM Tris-Base, 1.5 mM $MgCl_2$, 0.1% Tween 20, 0.015 mM Cresol red, 200 µM dNTPs, 7.5 pmol of each PCR primer (primer I, 20-mer, 5′-AAGCTTGGATCCTCTAGAGC-3′; primer II, 18-mer, 5′-CTGCAGGTACCGGTCCGG-3′), and 0.5 U *Thermus aquaticus* (*Taq*) DNA polymerase. PCRs were performed for 30 cycles consisting of 1 minute at 94°C, 1 minute at 55°C, and 1 minute at 72°C in 384-well PCR machines (MJ Research).

Arraying of PCR products. We generated robotically high-density filter arrays of the PCR-amplified OST clones as described [23]. All OST clones were spotted as quadruplets onto 22 cm × 22 cm nylon membranes. Each membrane contains 54,320 OST clone spots and 2304 spots of genomic salmon sperm DNA. The latter spots yield signals in every oligonucleotide hybridization experiment and are necessary to guide the automated image analysis. We prepared 30 filter copies for parallel hybridization experiments.

Oligonucleotide hybridization and OST clones back-hybridization. A set of 193 8-mer oligonucleotides was used in hybridization of the OST library. The oligonucleotides were labeled at the 59-end by a kinase reaction using [(α^{33}P]dATP (Amersham International) and T4 polynucleotide kinase (New England Biolabs). We carried out the hybridizations as described [22,33]. The intensities of the hybridization signals were measured by phosphor storage autoradiography (Fuji).

As a control for the clustering of clones, PCR products of nine randomly chosen OST clones were hybridized back to OST library in nine separate hybridization experiments. We labeled 200 ng (0.44 pmol) of each probe in a random hexamer priming reaction using [α^{33}P]dGTP (Amersham International) and Klenow polymerase (New England Biolabs). Each probe was used in a separate hybridization experiment. The hybridizations were performed overnight at 65°C in hybridization bottles containing 10 ml of 0.25 M sodium phosphate, pH 7.2, 5% sodium dodecylsulfate, and 1.25 mM EDTA, with a probe concentration of 20 ng/ml (44 pM). Filters were washed in the same buffer at 65°C for 3 hours and intensities of the hybridization signals were measured by phosphor storage autoradiography (Fuji).

Automated image analysis and data normalization. We analyzed the hybridization images obtained from the BAS-1800 scanner (Fuji) using a program developed in-house.

We normalized the raw data using double-ranking [33] to compensate for different overall hybridization intensities from different probes, and different amount of DNA of different clones.

Cluster analysis. For the cluster analysis of the OFP data we used CLICK, a novel clustering algorithm [26]. The algorithm represents the input data as a weighted graph, where vertices correspond to elements and edge weights reflect pairwise similarity between the corresponding elements. Under certain probabilistic assumptions, the weight of an edge reflects the likelihood that its endpoints originate from the same cluster. The clustering process can be described recursively as follows: In each step the algorithm handles some connected component of the subgraph induced by the yet unclustered elements. If the likelihood that the component should be further partitioned is below a threshold, it is considered a kernel of some cluster. Otherwise, a minimum weight cut is computed, and the component is split into its two most loosely connected pieces according to this cut. After the above process terminates, an adoption procedure enriches kernels by adding to them singletons whose fingerprints are highly similar to the mean fingerprint of the kernel. Finally, a merging procedure merges similar clusters.

To calibrate the clustering process and tune CLICK's running parameters, we used the information obtained by the back-hybridization experiments. These experiments yield a highly reliable subclustering of a subset of the elements, as each experiment pinpoints in principle all clones matching a single gene. Using that subclustering as a "true" (partial) solution, we compared solutions produced by CLICK with several parameter choices. The comparison was based on computing a Jaccard score [37] for each solution and picking the solution with the best score. To further validate this solution, we reclustered the data using a second clustering method, which is a variant of K-means [33]. We obtained similar results (data not shown).

Target selection. We selected representatives from each cluster of clones for sequencing. To maximize the number of different genes we expect to discover, we picked more targets from larger clusters, as such clusters are more likely to contain several genes. We fully sequenced small clusters with up to three clones (including singletons), and sampled larger clusters proportionally to their size. From each cluster, the first two targets we selected were centrals, that is, with fingerprints closest to the average fingerprint of the cluster; the next three were outliers, whose fingerprints were farthest from the mean; and the rest were chosen randomly. In addition, we fully sequenced six medium-sized clusters with various homogeneity scores for a more careful analysis of the clustering performance.

Re-arraying of OST clones and sequencing. A total of 1058 selected target clones were re-arrayed using a robotic device. We performed the sequencing on automatic DNA sequencers ABI PRISM 377 or ABI PRISM 3700 (Perkin Elmer Applied Biosystems) using Big Dye Terminator mix (Perkin Elmer) and primers 5′-AAGCTTGGATCCTCTAGAGC-3′ and 5′-CTGCAGGTACCG-GTCCGG-3′ for sequencing in forward and reverse directions, respectively, under the recommended conditions (Perkin Elmer).

Sequence analysis. Base calling was further edited using Sequencher program (GeneCodes Corp., Version 3.0). Sequence identification was performed by BLAST [38] searches against either human or specific database (HORDE, http://bioinformatics.weizmann.ac.il/HORDE), or the working draft of the human genome (http://www.ncbi.nlm.nih.gov/genome/guide/human).

Acknowledgments

This work was partially supported by GIF grant G-0506-183.0396. R.S. was supported by an Eshkol fellowship from the Ministry of Science, Israel.

RECEIVED FOR PUBLICATION DECEMBER 26, 2001;
ACCEPTED JUNE 18, 2002.

References

1. Dayhoff, M. O. (1976). The origin and evolution of protein superfamilies. *Fed. Proc.* **35:** 2132–2138.
2. Buck, L., and Axel, R. (1991). A novel multigene family may encode odorant receptors: a molecular basis for odor recognition. *Cell* **65:** 175–187.
3. Buck, L. B. (1992). The olfactory multigene family. *Curr. Biol.* **2:** 467–473.
4. Lancet, D., and Pace, U. (1987). The molecular basis of odor recognition. *Trends Biochem. Sci.* **12:** 63–66.
5. Reed, R. R. (1990). How does the nose know? *Cell* **60:** 1–2.
6. Mombaerts, P. (1999). Molecular biology of odorant receptors in vertebrates. *Annu. Rev. Neurosci.* **22:** 487–509.
7. Fuchs, T., Glusman, G., Horn-Saban, S., Lancet, D., and Pilpel, Y. (2000). The human olfactory subgenome: from sequence to structure and evolution. *Hum. Genet.* **108:** 1–13.
8. Glusman, G., *et al.* (2000). The olfactory receptor gene superfamily: data mining, classification and nomenclature. *Mamm. Genome* **11:** 1016–1023.
9. Glusman, G., Yanai, I., Rubin, I., and Lancet, D. (2001). The complete human olfactory subgenome. *Genome Res.* **11:** 685–702.
10. Zozulya, S., Echeverri, F., and Nguyen, T. (2001). The human olfactory receptor repertoire. *Genome Biol.* **2:** RESEARCH0018.
11. Zhang, X., and Firestein, S. (2002). The olfactory receptor gene superfamily of the mouse. *Nat. Neurosci.* **5:** 124–133.
12. Rouquier, S., *et al.* (1998). Distribution of olfactory receptor genes in the human genome. *Nat. Genet.* **18:** 243–250.
13. Trask, B. J., *et al.* (1998). Members of the olfactory receptor gene family are contained in large blocks of DNA duplicated polymorphically near the ends of human chromosomes. *Hum. Mol. Genet.* **7:** 13–26.
14. Nef, P., *et al.* (1992). Spatial pattern of receptor expression in the olfactory epithelium. *Proc. Natl. Acad. Sci. USA* **89:** 8948–8952.
15. Asai, H., *et al.* (1996). Genomic structure and transcription of a murine odorant receptor gene: differential initiation of transcription in the olfactory and testicular cells. *Biochem. Biophys. Res. Commun.* **221:** 240–247.
16. Glusman, G., Clifton, S., Roe, R., and Lancet, D. (1996). Sequence analysis in the olfactory receptor gene cluster on human chromosome 17: recombinatorial events affecting receptor diversity. *Genomics* **37:** 147–160.
17. Walensky, L. D., *et al* (1998). Two novel odorant receptor families expressed in spermatids undergo 5′-splicing. *J. Biol. Chem.* **273:** 9378–9387.
18. Qasba, P., and Reed, R. R. (1998). Tissue and zonal-specific expression of an olfactory receptor transgene. *J. Neurosci.* **18:** 227–236.
19. Sosinsky, A., Glusman, G., and Lancet, D. (2000). The genomic structure of human olfactory receptor genes. *Genomics* **70:** 49–61.
20. Pilpel, Y., and Lancet, D. (1999). The variable and conserved interfaces of modeled olfactory receptor proteins. *Protein Sci.* **8:** 969–977.
21. Hoheisel, J. D., Lennon, G. G., Zehetner, G., and Lehrach, H. (1991). Use of high coverage reference libraries of Drosophila melanogaster for relational data analysis. A step towards mapping and sequencing of the genome. *J. Mol. Biol.* **220:** 903–914.
22. Radelof, U., *et al.* (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.
23. Meier-Ewert, S., *et al.* (1998). Comparative gene expression profiling by oligonucleotide fingerprinting. *Nucleic Acids Res.* **26:** 2216–2223.
24. Poustka, A. J., *et al.* (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.
25. 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.
26. Sharan, R., and Shamir, R. (2000). CLICK: a clustering algorithm with applications to gene expression analysis. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **8:** 307–316.
27. Henikoff, S., *et al.* (1997). Gene families: the taxonomy of protein paralogs and chimeras. *Science* **278:** 609–614.
28. Tatusov, R. L., Koonin, E., and Lipman, D. J. (1997). A genomic perspective on protein families. *Science* **278:** 631–637.
29. Gogarten, J. P., and Olendzenski, L. (1999). Orthologs, paralogs and genome comparisons. *Curr. Opin. Genet. Dev.* **9:** 630–636.
30. Lander, E. S., *et al.* (2001). Initial sequencing and analysis of the human genome. *Nature* **409:** 860–921.
31. Venter, J. C., *et al.* (2001). The sequence of the human genome. *Science* **291:** 1304–1351.
32. Rose, T. M., *et al.* (1998). Consensus-degenerate hybrid oligonucleotide primers for amplification of distantly related sequences. *Nucleic Acids Res.* **26:** 1628–1635.
33. Herwig, R., *et al.* (1999). Large-scale clustering of cDNA-fingerprinting data. *Genome Res.* **9:** 1093–1105.
34. Herwig, R., *et al.* (2000). Information theoretical probe selection for hybridisation experiments. *Bioinformatics* **16:** 890–898.
35. Maier, E., Meier-Ewert, S., Ahmadi, A. R., Curtis, J., and Lehrach, H. (1994). Application of robotic technology to automated sequence fingerprint analysis by oligonucleotide hybridisation. *J. Biotechnol.* **35:** 191–203.
36. Maier, E., Meier-Ewert, S., Bancroft, D., and Lehrach, H. (1997). Automated array technologies for gene expression profiling. *Drug Discov. Today* **2:** 315–324.
37. Everitt, B. (1993). *Cluster Analysis*. Edward Arnold, London.
38. Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.* **215:** 403–410.