BIOINFORMATICS EDITORIAL

Vol. 20 no. 17 2004, pages 2893–2902
doi:10.1093/bioinformatics/bth347

# PICKY: *oligo microarray design for large genomes*

*Hui-Hsien Chou[1,2,*], An-Ping Hsia[3], Denise L. Mooney[1] and Patrick S. Schnable[3,4]*

*[1]Department of Genetics, Development and Cell Biology, [2]Department of Computer Science, [3]Department of Agronomy and [4]Center for Plant Genomics, Plant Sciences Institute, Iowa State University, Ames, IA 50011, USA*

Received on March 24, 2004; revised on April 29, 2004; accepted on May 14, 2004
Advance Access publication June 4, 2004

**ABSTRACT**

**Motivation:** Many large genomes are getting sequenced nowadays. Biologists are eager to start microarray analysis taking advantage of all known genes of a species, but existing microarray design tools were very inefficient for large genomes. Also, many existing tools operate in a batch mode that does not assure best designs.

**Results:** PICKY is an efficient oligo microarray design tool for large genomes. PICKY integrates novel computer science techniques and the best known nearest-neighbor parameters to quickly identify sequence similarities and estimate their hybridization properties. Oligos designed by PICKY are computationally optimized to guarantee the best specificity, sensitivity and uniformity under the given design constrains. PICKY can be used to design arrays for whole genomes, or for only a subset of genes. The latter can still be screened against a whole genome to attain the same quality as a whole genome array, thereby permitting low budget, pathway-specific experiments to be conducted with large genomes. PICKY is the fastest oligo array design tool currently available to the public, requiring only a few hours to process large gene sets from rice, maize or human.

**Availability:** PICKY is independent of any external software to execute, is designed for non-programmers to easily operate through a graphical user interface, and is made available for all major computing platforms (e.g. Mac, Windows and Linux) at http://www.complex.iastate.edu.

**Contact:** PICKY@www.complex.iastate.edu

**Supplementary information:** A short paper detailing the probability analysis for PICKY, a program that implements functions mentioned in the paper, and the output of the program are available online from the Publisher's website.

## INTRODUCTION

Recently, many genomes from various taxa have been sequenced, including some large genomes like *Drosophila* (Myers *et al*., 2000), *Arabidopsis* (Lin *et al*., 1999; Mayer, 1999; The Arabidopsis Genome Initiative, 2000), human (Lander *et al*., 2001; Venter *et al*., 2001), mouse (Waterston *et al*., 2002) and rice (Goff *et al*., 2002; Yu *et al*., 2002). Additional large genomes including maize (Chandler and Brendel, 2002), rat (Summers *et al*., 2001), chicken (Ren *et al*., 2003) and dog (Kirkness *et al*., 2003) may soon become available. Microarrays are among the most widely used methods to carry out research based on finished genomes. Technological advances allow microarray experiments to be conducted at higher throughput and greater convenience, but microarrays must be carefully designed for the data produced from them to be useful. Specifically, spots on the arrays for detecting the expression level of each gene must be unique to that gene (i.e. exhibit high specificity), must be able to detect that gene (i.e. exhibit high sensitivity), and must function optimally under the same melting temperature and other experiment conditions (i.e. exhibit high uniformity). Given that microarray terminologies are not always consistent in the literature, in this paper we explicitly use the terms 'probe' or 'oligo' to denote the spotted sensor on the microarray surface, and the terms 'target', 'non-target' or 'sequence' to denote the labeled single-strand DNA that are reverse-transcribed from mRNA.

Currently there are three major formats of microarrays, based on cDNA clones (DeRisi *et al*., 1997; Golub *et al*., 1999), lithographically synthesized short oligos (Lockhart *et al*., 1996; Zuo *et al*., 2002), and short or long synthesized oligos prescribed by users (Kane *et al*., 2000; Bosch *et al*., 2002). cDNA clones are usually available as a byproduct of whole genome or expressed sequence tag (EST) sequencing projects and are readily available for chip manufacturing, but they suffer from high cross-hybridization noise due to their inability to differentiate similar genes or gene families sharing long stretch of common subsequences. Different cDNA spots may have very different melting temperatures with their intended targets and it is impossible to obtain an optimal experiment temperature that can successfully separate all target hybridizations from all non-target hybridizations. Also, tracking and maintaining cDNA libraries have the potential to introduce additional errors into the experiments.

The Affymetrix lithographic arrays are the most popular short oligo (20–25 bp) array solution. They are manufactured

*To whom correspondence should be addressed.

*in situ* on silicon wafers and cannot be easily modified for ongoing genome projects without incurring the high cost of replacing fabrication masks. Short oligos as used in Affymetrix arrays cannot completely differentiate targets and non-targets in large genomes because these genomes tend to have the same or very similar 25mers showing up in many genes. Therefore, multiple match/mismatch oligo pairs must be used to confirm the detection of a specific gene (Mei *et al*., 2003).

The third type of microarrays uses longer oligos (50–70 bp) as spots. These arrays can be designed by end-users for all or a subset of the genes, and manufactured in small or large quantities. Since oligos used in these arrays can be individually updated throughout a project period, they allow quicker incorporation of new sequence data without incurring a high replacement cost. Owing to their longer length and single-strand nature, the oligos can be made very sensitive to their intended targets and therefore often only one oligo is needed to detect a gene (Relógio *et al*., 2002). In Addition, the oligos can be computationally optimized to achieve much greater specificity and uniformity, thereby reducing cross-hybridization noise. As only sequence data are required to design oligo arrays, clone tracking errors are also avoided. These factors make long oligo arrays most suitable for studying species under ongoing sequencing projects or species with lesser commercial interest (for which commercial chips may not be available).

Currently, only a few publicly available software tools deal with the optimal oligo array design problem, and none of them can efficiently handle large genomes like human, mouse or rice. Here, we introduce a powerful new program, PICKY, that facilitates oligo microarray design for large as well as small genomes. Having PICKY freely available to academic users lowers the cost of array design and makes this technology more accessible to all.

## SYSTEMS AND METHODS

### The oligo design parameters

Since PICKY is designed to compute the best oligo set for large genomes, the first question that must be answered is what makes the best oligo set? There have been several papers about this (Lockhart *et al*., 1996; Kane *et al*., 2000; Li and Stormo, 2001; Relógio *et al*., 2002), suggesting several criteria, including (a) base composition limit (no single base should constitute $>50\%$ of an oligo); (b) base distribution limit (no stretch of a continuous base should exceed 25% of the length of an oligo); (c) GC-content (best between 30 and 70%); (d) no secondary structure (i.e. oligos should not form dimers and hairpins or attempt to target sequence regions that may form dimers and hairpins); (e) length of continuous complementary match to non-targets (should ideally be $<15$ bp); and (f) the overall complementarity to non-targets (should ideally be $<75\%$). All these conditions are considered by PICKY and most are user adjustable parameters that can be modified depending on the characteristics of the input data. Conditions a, b and d are implicitly enforced by the other conditions, e.g. a single base region longer than 25% of the oligo size is over 15 bp of a 60 bp oligo, so it cannot be targeted by an oligo with condition e because the reverse-complement of each sequence is also considered as a non-target in PICKY's computation.

In addition to these, PICKY also considers more sophisticated design parameters. PICKY accepts minimum and maximum oligo lengths instead of a fixed length. Within the specified range it can adjust the length of oligos to achieve greater specificity and uniformity among all oligos. Rather than requesting the melting temperature ($T_m$) as a parameter from users, PICKY takes the minimum separation temperature as a parameter, and ranks best oligo candidates by comparing their target and non-target melting temperatures. Only oligos that provide at least the minimum separation temperature will be considered in PICKY, and it is their joint temperature separations that finally determine the optimal $T_m$ suggested by PICKY for microarray experiments. PICKY also handles multiple target and non-target gene sets, where the non-target sets are used as a screening background while oligos are being designed for the target sets. This allows, for example, a small budget experiment to study a handful of genes from a large genome, but still guarantees the results will be as good as those obtained from a whole genome microarray.

Among the suggested criteria for selecting the best oligo set, we have found conditions e and f recommended in Kane *et al*.'s (2000) paper to be the most widely cited conditions (Li and Stormo, 2001; Rouillard *et al*., 2002; Xu *et al*., 2002). We call these Kane's first and second conditions. Since these two conditions are based on sequence similarity, they are not sufficient to determine good oligo candidates before in-depth thermodynamic calculation. Nevertheless, they provide the most efficient way to screen out bad oligo candidates. Although mismatches to non-targets do not ascertain good oligo candidates without thermodynamic comparisons, exact matches to non-targets do quickly identify oligos that should be avoided due to concerns of cross-hybridization. It will become clear in the following that Kane's two conditions play a pivotal role in the overall strategy of PICKY.

### Suffix array construction and search

The first step in PICKY's oligo computation is to construct a generalized suffix array that can quickly identify all substrings contained in all sequences and their complements (Gusfield, 1997). Suffix array is a very efficient and space-saving data structure that records in alphabetical order all possible suffixes and their locations in the input sequences. The theory of suffix array states that the longest common prefix (LCP) shared by any two non-adjacent suffixes must be equal or shorter than the LCP of any two neighboring suffixes between them in the suffix array (Manber and Myers, 1993). Thus, to determine if a particular sequence shares some substrings of

certain lengths with the other sequences, we can locate all its suffixes in the suffix array and scan both the left and right sides of each of its suffixes. The other suffixes encountered during the scan indicate shared substrings and their locations in the other sequences can be immediately identified. Since a sequence and its own complement are both represented in the suffix array, this scanning process also detects repetitive, low complexity, self-similar and self-complementary regions in the sequences. All of these are avoided as probe target regions by PICKY.

The Burkhardt–Kärkkäinen suffix array construction algorithm (Burkhardt and Kärkkäinen, 2003) is used in PICKY after some modification. We have surveyed several suffix array algorithms and found it to be the quickest and the most memory efficient one to construct the generalized suffix array. To facilitate locating suffixes in the suffix array, an inverse suffix array is constructed after the suffix array, which indices suffices in the suffix array from the perspective of each sequence. To avoid string comparison and to speed up suffix array scanning, a longest common prefix (LCP) array that records the length of the shared prefix for each neighboring suffix pair is pre-computed using the Kasai *et al*. (2001) algorithm. Altogether, PICKY requires 20 bytes to represent each DNA base in those arrays. If double-strand screening is turned off, the requirement drops to only 10 bytes per input base.

## Avoiding unnecessary computation

In the second step, PICKY uses the suffix array to guard against Kane's first condition, i.e. to make sure that no oligo target regions on any sequence share a common stretch equal or longer than the maximum match length of 15 bp (or a value specified by the users) with a non-target. A linear-time sweep across the suffix array looking at the two immediate neighbors of each suffix can quickly identify all regions on all input sequences that must be ruled out for violating this condition. This provides a dramatic boost in PICKY speed because it is not necessary to conduct the subsequent local alignment and thermodynamic calculation for these bad regions.

In addition to finding regions that violate Kane's first condition, for all remaining regions PICKY also records their average longest match length equal or higher than the minimum match length of 10 bp (or a value specified by the users). This information is used both to paint the onscreen sequences with different background colors to indicate cross-hybridization likelihood, and to prioritize regions when they are being considered for oligo selection. The average longest match length is used as the priority score in PICKY instead of the more elaborated average landscape adopted in the program ProbeSelect (Levy *et al*., 1998; Li and Stormo, 2001). The reason is that for large genomes, there is a greater chance of random short matches among sequences, and averaging these random matches does not help discriminating good and bad regions. For example, a 3460 maize gene set is analyzed in Figure 1, where the number of sequences that cannot have any probe-size region that is free of random short matches to the other sequences is drawn against the match size. When the minimum match size considered is 10 bp or above, none of the sequences can get any probe, and therefore are all bad. Averaging short matches at this size does not help identifying good and bad sequence regions. However, when the minimum match size is increased, the number of bad sequences drops sharply, and then levels off. This is when randomness ceases, and what remain are significant matches that should really be avoided. Therefore, by averaging the longest matches found in each region they can be more correctly prioritized.

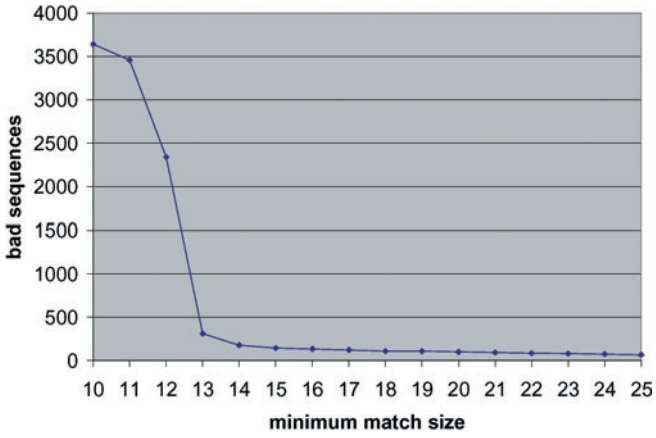


**Fig. 1.** Bad sequences versus minimum match size considered. A 3460 maize gene set is analyzed, where the number of sequences that cannot have any probe size region free of short matches to the other sequences is drawn against the minimum match size considered. Many sequences are considered bad due to short random matches, but only a few really have longer bad matches.

## Local alignment and melting temperature estimation

The third step in PICKY's computation is to find, for each best candidate region of a sequence, all other sequences that are similar to it. Using the suffix array, these non-targets can be quickly gathered as stated earlier, and their melting temperatures with oligo candidates can then be estimated. There are several different models for predicting the melting temperatures of DNA hybridization. The model chosen in PICKY is the nearest neighbor (N-N) model (Breslauer *et al*., 1986; Rychlik *et al*., 1990; SantaLucia *et al*., 1996; Allawi and SantaLucia, 1997). The equation used in PICKY to estimate melting temperatures (°C) is as follows:

$$T_M = \frac{\Delta H}{\Delta S + R \ln (C/4)} + 12.0 \times \log_{10}[\mathrm{Na}^+] - 273.15,$$

where $\Delta H$ and $\Delta S$ are the accumulated enthalpy and entropy values based on the sequence content of an oligo and its target region using the updated N-N parameter tables (Allawi and

SantaLucia, 1997), $R$ is the molar gas constant 1.987, $C$ is the molar concentration of total oligonucleotides in the microarray experiment and $[Na^+]$ is the molar concentration of salt. The oligonucleotide concentration is generally unknown, so a value of $1 \times 10^{-6}$ M is used by default as suggested in the literature (Li and Stormo, 2001; Kaderali and Schliep, 2002; Rouillard *et al*., 2003). Salt has a stabilizing effect on oligonucleotide annealing, so a salt concentration term is added to the equation to correct for that temperature shift with a coefficient of 12.0 as suggested by an oligo vendor and the literature (SantaLucia *et al*., 1996; Devor *et al*., 2002). This is lower than some other studies (Rychlik *et al*., 1990; Nielsen *et al*., 2003). The initialization and dangling end effects (Bommarito *et al*., 2000) are also considered in PICKY. There are differences between hybridization in solution and on a microarray surface because one end of the oligos is fixed to the array surface, but currently there are no known surface based parameters. We are therefore using solution-based parameters as the best approximation and will start using surface based parameters when they become available.

The melting temperature between an oligo candidate and its target region is estimated first. Its melting temperatures with all potential non-targets are estimated next to prevent imperfectly matched cross-hybridization. Note that oligos with perfectly matched non-targets have already been screened out in the previously step. Generally, it is harder to estimate non-target melting temperatures given the limited knowledge of mismatch hybridization. Although there is only one perfect match to an oligo (i.e. its Watson–Crick complement), there are an enormous number of imperfect matches between an oligo and its non-targets. Fortunately, precise non-target melting temperatures are not necessary to simply tell if they can cause cross-hybridizations. Our strategy thus is to use Kane's second condition as a guideline, which states that any sequence similarity over 75% (or a value specified by the users) can potentially cause problems and must be examined. Using the suffix array, we uncover and align all non-targets with an oligo candidate up to the similarity level of Kane's second condition. Although sequence alignment generally takes quadratic time to compute, we have devised a novel local alignment method that works in linear time. The algorithm works by interleaving row and column computations of the alignment matrix and dynamically maintaining a viable alignment band until it ends. The alignment results are then used to estimate the probe/non-target melting temperatures using the same equation above. N-N parameters that accommodate simple mismatches have been experimentally determined. From the literature we have found parameters for single base mismatches (e.g. G·A, G·G or G·T) (Allawi and SantaLucia, 1997, 1998a, b and c; Peyret *et al*., 1999) and dangling end and gap mismatches (i.e. bulge or loop) (Zuker *et al*., 1999; Bommarito *et al*., 2000). The calculated melting temperatures of an oligo candidate with all its non-targets are then used to prioritize the oligo in the final step.

### Probe set selection and optimal experiment temperature determination

The last step is to compare target and non-target melting temperatures of all probe candidates from all sequences in order to find a subset that can detect each gene, has the least chance to cross-hybridize, shares a uniform temperature range, and maximizes the distance between the lowest target and the highest non-target melting temperatures of the chosen set. PICKY reports the optimal microarray experiment temperature after the probe set has been determined; this differs significantly from most oligo design tools that take the temperature as input to screen against probe candidates.

Selection of the optimal set of probe candidates resembles a non-integer knapsack problem (Martello and Toth, 1990), which is known to be NP-complete (i.e. it takes exponential time to optimally solve). However, we can limit ourselves to consider only the best, say, five non-overlapping probe candidates of each gene, and we can use an iterative algorithm to approximate the optimal selection of the probe set. First, the experiment temperature is set to the average mid-temperature between all target and non-target melting temperatures of all probe candidates, and then their deviations from this temperature are computed. For each gene, their probe candidates up to a user desired number are selected into a new set based on their deviations. A new experiment temperature will then be determined by the new set, and this iteration is repeated until it converges to an optimal set and an optimal temperature that no longer changes. The oligos designed are then presented in a GUI panel where they can be examined and subsequently saved to files.

It may seem that secondary structure screening is omitted from PICKY steps. That is not the case. As all sequences and their reverse-complements are represented in the suffix array and are included in cross-hybridization screening, self-similarity tests for secondary structures are conducted alongside mutual similarity tests for cross-hybridizations. Thus, no additional secondary structure screening step is necessary in PICKY. Furthermore, PICKY's inclusion of the reverse-complement of each input sequence ensures that oligos designed will function correctly even in the presence of anti-sense transcripts (Lehner *et al*., 2002; Shendure and Church, 2002; Carmichael, 2003; Osato *et al*., 2003; Yelin *et al*., 2003).

## RESULTS

### Effects of the minimum and maximum match parameters

PICKY's minimum match parameter determines its sensitivity level: exact matches found in the suffix array that are shorter than this are considered harmless and ignored. Conversely, the maximum match parameter determines its tolerance level: exact matches that are equal or longer than this are discarded immediately since they are considered unsafe

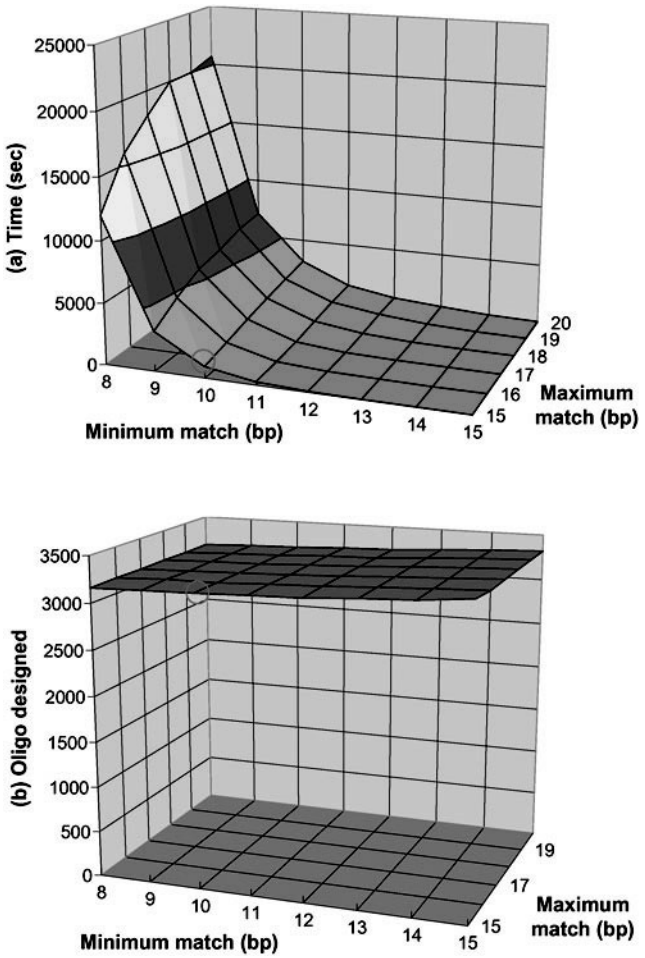


**Fig. 2.** Effect of the minimum and maximum matches on speed and oligos. (**a**) running time and (**b**) oligos designed for a 3460 maize gene set are drawn against different combinations of the minimum and maximum match values. Red circles indicate the default settings in PICKY.

by Kane's first condition. Only matches whose length falls between these two limits will trigger the time-consuming local alignment, non-target melting temperature estimation and oligo candidate validation process. Therefore, the minimum and maximum matches determine the number of non-targets screened by PICKY, and its speed. Different combinations of these two parameters were used to process a 3460 maize gene set to reveal their effects on PICKY speed and the number of oligos it can design. The results are shown in Figure 2.

As seen in Figure 2a, when the minimum match is shortened, a lot of random short matches are admitted into the validation pipeline and that considerably increases the runtime of PICKY. However, they do not have much effect on the number of oligos designed (Fig. 2b). For example, reducing minimum match from 10 to 9 bp increases PICKY runtime from 930 to 3225 s but only reduces oligos designed from 3184 to 3174. Despite the greatly increased workload to screen short random

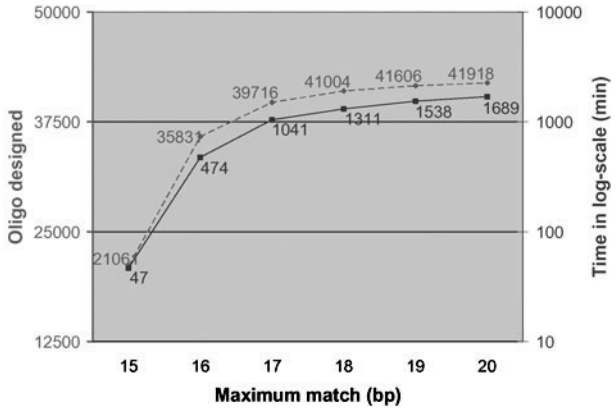


**Fig. 3.** Effect of the maximum match on speed and oligos for large genomes. The running time in log-scale (blue) and oligos designed for a 57 721 rice gene set (red) are drawn against different maximum match values. Here the minimum match is unchanged at the default of 10 bp.

matches, only a few slightly unfavorable oligos are removed from the set. This substantiates again the reasoning earlier in Avoiding unnecessary computation that short random matches are not good indicators of region uniqueness. Alternatively, increasing minimum match cannot improve PICKY speed but will admit bad oligos into the set, so a reasonable default value for this parameter is set to 10 bp (indicated by red circles in this figure). Rigorous probability modeling and computer simulation also suggest this default setting (see the Supplementary information).

The effect of maximum match is less obvious in the previous example because this parameter only controls sequences that are highly similar to each other, and these tend to be rare in smaller gene sets. To see its effect on larger gene sets, we ran PICKY on a rice gene set containing 57721 sequences. This gene set not only is large but contains lots of highly similar sequences. In fact, the largest gene family in this set contains over 9399 transposon sequences that are more than 90% similar to each other over 90% of their entire length (see Benchmark oligo sets). The results under different maximum matches are shown in Figure 3 with time drawn in log-scale. It can be seen that for large genomes it may be necessary to set a higher maximum match tolerance in order to obtain more usable oligos. For example, increasing maximum match from 15 to 16 bp significantly increases the number of oligos designed, from 21 061 to 35 831. Extending the maximum match beyond 17 bp, however, results in diminishing oligo return and a much greater chance of cross-hybridization. Therefore, users have to decide a balance between the quality and the number of oligos for their projects. Interestingly, the running time of PICKY in log-scale corresponds precisely with the number of oligos designed. Initially when the maximum match is increased from the default of 15 bp, a lot of oligo candidates that were previously excluded are entered into

PICKY's screening pipeline, causing exponential time growth. Nonetheless, the growth quickly levels off after most random short matches are admitted into the computation pipeline and the remaining long matches no longer get added by simply extending the maximum match.

## Temperature landscapes and the minimum separation

Despite the explanation in Systems and methods, PICKY does not enumerate all oligo candidates of a target region to individually compare them with all potential non-targets. Instead, all non-targets discovered by the suffix array are aligned and compared to the whole target region just once, for each gene. Two temperature landscapes are produced during this process. One is computed from the target region itself and includes all oligo candidates within the allowable lengths at different locations. For example, in Figure 4a, the target temperature landscape of a sequence region is shown. Oligo candidates targeting between base location 773–862 and with variable length 50–70 bp can hybridize with this region at different melting temperatures. Only locations targeted by valid oligos have a non-zero temperature. Hence, temperatures for longer oligos start dropping off beyond location 792 ($862 - 70\text{mer} = 792$) and the last location in this region that can be targeted is 812 ($862 - 50\text{mer} = 812$).

The non-target temperature landscape for the same region (Fig. 4b) is computed by aligning each potential non-target with the target region and estimating its melting temperatures with all oligo candidates for the same region. Since the goal is to avoid cross-hybridizations, only the highest non-target temperature discovered at each location is recorded in the non-target temperature landscape. Hence, this landscape is relatively flat because the highest temperature found in a location is shared by all oligos overlapping the same location. In batch-mode style processing a program might have selected oligos with the highest target temperatures, e.g. the 70mer oligo targeting location 778 indicated by the red circle in Figure 4a. In PICKY's integrated approach instead, both the target and non-target landscapes are considered, and it is their difference that determines the oligo selection. The temperature difference between the two landscapes is shown in Figure 4c. The best oligo candidate for this region of the gene should actually be the 64mer targeting location 796 indicated by the red circle in Figure 4c because a greater difference in melting temperature translates to both a greater specificity and a lower background. The oligo for a gene is later selected among the best candidates of each valid region of the gene by further comparing their temperatures. This example also highlights the benefit of varying oligo lengths to achieve a higher difference between target and nontargets melting temperatures.

Although the highest temperature in the difference landscape suggests an oligo candidate, it may still be rejected if its value is lower than the minimum temperature separation. PICKY's default setting is to ignore oligo candidates

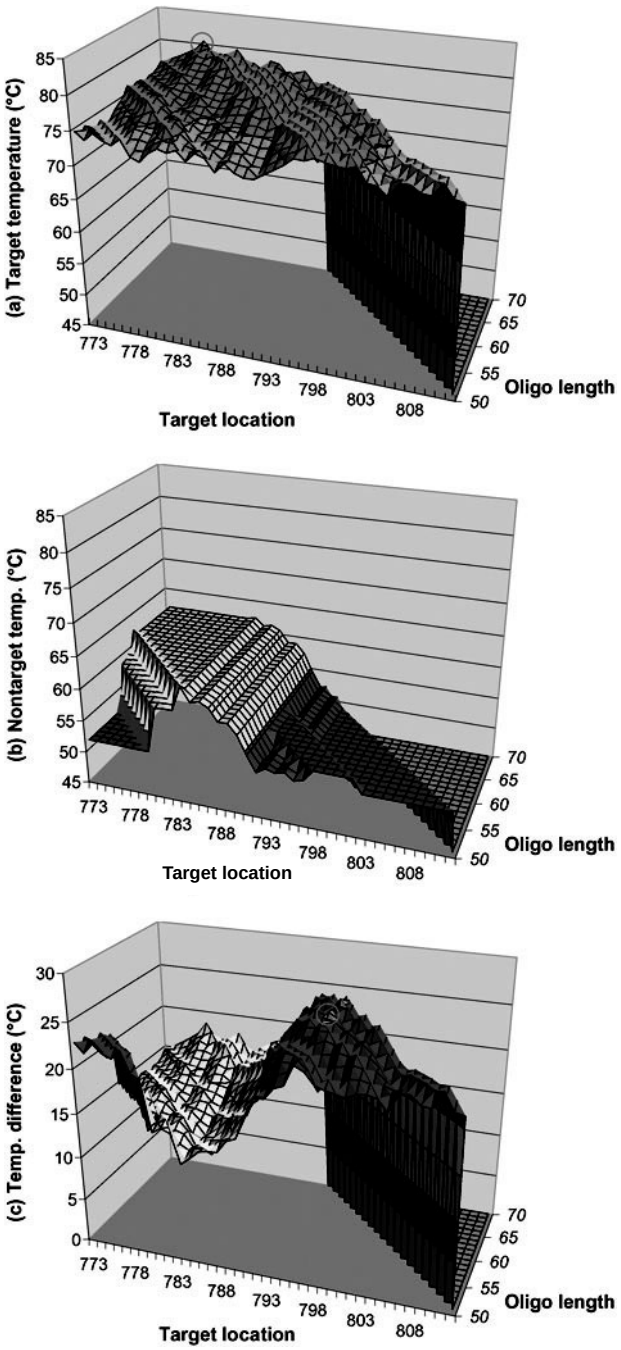


**Fig. 4.** Temperature landscapes and the selection of oligo target regions. (**a**) target, (**b**) non-target and (**c**) difference melting temperature landscapes are drawn against different sequence locations and oligo lengths. Red circles indicate oligos that might be chosen under different design strategies.

whose target and non-target melting temperature difference is lower than 20°C. This parameter can be adjusted by the users, but we do not recommend setting it lower than 10°C because the non-target melting temperature is only an estimate that is likely to vary from the true value, potentially admitting

**Table 1.** Summary of oligos designed for rice gene families

| Similarity and length | Total no. of families | Total no. of genes | Designed oligos (%) | Largest gene family size | Remaining genes | Remaining oligos (%) |
|---|---|---|---|---|---|---|
| • 90% S | | | | | | |
| 90% L | 3579 | 19419 | 716 (4%) | 9399 | 10020 | 416 (4%) |
| • 80% S | | | | | | |
| 80% L | 4942 | 26440 | 2756 (10%) | 12405 | 14035 | 2098 (15%) |
| • 80–90% S | | | | | | |
| 80–90% L | 1363 | 7021 | 2040 (29%) | 3006 | 4015 | 1682 (42%) |

cross-hybridizations. On the other hand, a higher separation can reduce the total number of valid oligos found. Therefore, based on our experience 15–20°C seems to be the most suitable range for this parameter. Once an oligo candidate is selected, the other candidates overlapping it will not be considered further, but non-overlapping candidates will continue to be selected while they still satisfy the minimum temperature separation. This selection is independent of the order of the input sequences since genes are considered individually at this stage.

## Comparing oligo and cDNA arrays

Many large-genome species underwent an allotetraploidization event during their evolution (Helentjaris *et al*., 1988; Gaut and Doebley, 1997). Owing to this, these genomes contain many paralogs. Over time paralogs typically undergo one of two fates: (1) decay through mutation resulting in a pseudogene or (2) functional divergence (reviewed by Wendel, 2000). Functional divergence can occur either via alterations of gene function or expression pattern. The hypothesis has been that members of gene families exhibit cell-type-specific expression patterns. Given the high degree of nucleotide similarity, it is impossible to study the expression patterns of individual paralogs using cDNA arrays. PICKY can design oligo arrays that differentiate the expression patterns of certain gene family members. We analyzed PICKY oligo design for a 57 721 rice gene set and summarize the result in Table 1.

Rice genes are clustered into families using two stringency settings: >90% similarity over 90% of sequence length, and >80% similarity over 80% of sequence length. For the first cluster 3579 families are identified, containing a total of 19 419 genes. For these, PICKY can only design 716 oligos. Obviously, at this level of similarity it becomes thermodynamically impossible to differentiate paralogous genes. With the 80% cluster, PICKY can design 2756 oligos for a total of 26 440 genes, or ∼10%. It is still low, because the 90% cluster is a subset of the 80% cluster and includes all highly similar genes. If we consider only sequences that are added to the 80% cluster, i.e. those whose similarity is between 80 and 90% , then PICKY can design 2040 oligos for 7021 sequences. Actually, in both clusters most gene families (>99%) have less than 11 members but the largest gene family is made of many transposon sequences and consists ∼48% total clustered genes. By removing this family, PICKY can design 1682 oligos for 4015 sequences at the 80–90% similarity level with a 42% success rate. PICKY design parameters can be further relaxed to increase these numbers, albeit the oligos may have a lower quality.

**Table 2.** Comparison of PICKY with the other oligo design tools

| | Maize 250 gene set 168 585 bp | | Maize 3460 gene set 2 340 756 bp |
|---|---|---|---|
| PICKY 1.0 | Running time: | 15 s | 15 min |
| | Oligos designed: | 249 | 3183 |
| OligoWiz 1.0 | Running time: | 14 min | 2 h 48 min |
| (Nielsen *et al*., 2003) | Oligos designed: | 250 | 3460 |
| ArrayOligo Selector 3.5 | Running time: | 12 min | 3 h 38 min |
| (Bozdech *et al*., 2003) | Oligos designed: | 250 | 3460 |
| ProMide | Running time: | 34 min | 88 h 38 min |
| (Rahmann, 2002) | Oligos designed: | 250 | 3460 |
| OligoArray 2.1 | Running time: | 17 h 35 min | 261 h 59 min |
| (Rouillard *et al*., 2003) | Oligos designed: | 250 | 3459 |

## Comparing Picky with the other oligo design tools

We compared PICKY with several existing oligo design tools and tabulated the results in Table 2. Programs that are not available to us or that cannot handle our input are not included. Initially four gene sets ranging from 250 to 57 721 sequences were to be considered, but only the smaller two sets can be tested on time for publication. The execution time and the number of oligos designed by each tool are listed.

Probably the most distinctive benefit of PICKY is its speed. We ran all oligo design tools with the same gene sets on the same dual 500 MHz Pentium III Dell with 512 MB memory. Since PICKY does not depend on external software to execute,

**Table 3.** Benchmark oligo sets for large-genome species

| Yeast | *P. falciparum* | *Drosophila* | *Arabidopsis* | Rice | Mouse | Human |
|---|---|---|---|---|---|---|
| 6343 genes | 9616 genes | 18 513 genes | 25 447 genes | 57 721 genes | 25 392 genes | 26 088 genes |
| 8 959 501 bp | 10 783 114 bp | 30 348 312 bp | 37 829 502 bp | 85 714 244 bp | 52 701 574 bp | 64 875 065 bp |
| 48 min | 1 h 12 min | 2 h 9 min | 4 h 25 min | 14 h 38 min | 15 h 38 min | 16 h |
| 5660 oligos | 6117 oligos | 12 417 oligos | 23 516 oligos | 35 831 oligos | 21 463 oligos | 20 053 oligos |

it completely avoids the overhead of loading and executing external programs. PICKY uses suffix array to uncover similarities, and conducts slower melting temperature calculation only after all highly similar regions have been ruled out. Ironically, these highly similar regions would require the longest time to align and estimate for melting temperatures. A lot of time might have been wasted in processing them in the other tools. PICKY is very efficiently implemented as outlined in this paper, and also employs sophisticated computer science techniques like multithreading, template libraries and combinatorial optimization.

Another noticeable difference is that PICKY only designed oligos that can unambiguously identify each input sequence. The other tools designed oligos for almost all input sequences despite the fact that some redundant input sequences are too similar to be individually recognized by oligo arrays. For example, Contig84 is a proper subsequence of Contig2138 in the set and cannot have any unique oligo designed for it. However, all the other tools designed an oligo for it that either perfectly matches with Contig2138 or is targeting the poly(T)region, but only OligoArray 2.1 indicates this fact to its users. Finally, PICKY seems to be the only tool that also considers the reverse strand of each input sequence and can protect against anti-sense transcripts with its design.

### Benchmark oligo sets and PICKY availability

We have used PICKY to design oligo sets for several large-genome species (Table 3). These *benchmark* oligo sets can be downloaded directly from our website (see Availability) for interested scientists to study and compare. These sets are meant for evaluation only since new features added to PICKY later may not be reflected in the archived oligo sets. Interested scientists should contact us before ordering oligos based on our design. Oligo sets for more species may be added later if people request them and we can obtain their complete gene sets for design. Note that for each species processed, PICKY's parameters can be attenuated to achieve either a higher quality but smaller oligo set, or vice versa. The various parameter settings and the resulted oligo sets are all documented at our website.

Users can also license PICKY directly from us and run it on their own gene collections. The added benefit of running PICKY directly is that it can be used for other research purposes, such as sequence comparison, function annotation, gene family clustering and phylogenetic study. PICKY's interactive display of detected gene similarities and multiple sequence alignments allows users to explore its results dynamically and efficiently without cluttering their screen. A standard version of PICKY is available to academic users without charge. For commercial users, or users who request special alternations to PICKY for their needs, a licensing fee may be requested to support the continuous development of PICKY.

## DISCUSSION

A pipeline of screening is typically conducted in the other oligo design tools, e.g. BLAST (Altschul *et al*., 1990) or a hashing method is used to select unique oligo candidates based on similarity, then MFold (Zuker, 2003) or some other tools are used to estimate thermodynamic properties (Xu *et al*., 2002; Nielsen *et al*., 2003; Rouillard *et al*., 2003; Wang and Seed, 2003). In our opinion, a batch-mode design cannot produce the optimal oligo sets. The most critical problem of this method is that the size of the oligos and the experiment temperature must be given a priori as parameters to get the design pipeline started, but our research suggests that these parameters should instead be determined by the chosen oligo set after it has been selected. Similar to the other people's observation (Nielsen *et al*., 2003; Rouillard *et al*., 2003), we also noticed that the best oligo set should contain oligos of varying sizes, i.e. non-uniformity in oligo lengths can achieve greater uniformity in the melting temperatures of all oligos and reduce the chance of cross-hybridization. To sum up, we believe PICKY achieves the global optimization that is better than a batch-mode design output.

PICKY has an easy-to-use graphical user interface (GUI). To avoid the overhead of having to maintain different code when deploying PICKY to different computing platforms, a nice cross-platform GUI library wxWindows (http://www.wxwindows.org) is used. It allows the same PICKY source code to compile on all supported platforms, including Mac, Windows and Linux. PICKY currently contains over 10 000 lines of C++ code and is expected to grow even bigger when more powerful new features not covered in this paper are added.

## ACKNOWLEDGEMENTS

We thank Drs Robin Buell, Shu Ouyang and Qiaoping Yuan of TIGR for providing the rice gene set and allowing us to include in this study their gene family clustering results, Dr Chih-Hsien Chou for conducting the probability analysis provided in the Supplementary information, ISU Complex Computation Lab staff Wesley Chuang, Sunyoung Park, Kazuya Suzuki, Satish Vemula, Kent Weber and Annie Zhao for data collection, graphics design and beta-testing, Schnable lab staff Karthik Viswanathan for data collection, authors of the Burkhardt–Kärkkäinen algorithm and the wxWindows library for making their code freely available, and all members of the NSF rice oligo array project (http://www.ricearray.org) for helpful discussions. A portion of the validation of Picky was supported by the NSF Plant Genome Project: DBI-0313887 (Pam Ronald, UC Davis, principal investigator).

## REFERENCES

Allawi,H.T. and SantaLucia,J.J. (1997) Thermodynamics and NMR of internal G*T mismatches in DNA. *Biochemistry*, **36**, 10581–10594.

Allawi,H.T. and SantaLucia,J.J. (1998a) Nearest neighbor thermodynamic parameters for internal G*A mismatches in DNA. *Biochemistry*, **37**, 2170–2179.

Allawi,H.T. and SantaLucia,J.J. (1998b) Nearest-neighbor thermodynamics of internal A*C mismatches in DNA: sequence dependence and pH effects. *Biochemistry*, **37**, 9435–9444.

Allawi,H.T. and SantaLucia,J.J. (1998c) Thermodynamics of internal C*T mismatches in DNA. *Nucleic Acids Res.*, **26**, 2694–2701.

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.

Bommarito,S., Peyret,N. and SantaLucia,J.J. (2000) Thermodynamic parameters for DNA sequences with dangling ends. *Nucleic Acids Res.*, **28**, 1929–1934.

Bosch,J.T., Seidel,C., Batra,S., Lam,H., Tuason,N., Saljoughi,S. and Saul,R. (2002) Validation of sequence-optimized 70 base oligonucleotides for use on DNA microarrays. http://www.westburg.al/download/arrayposter.pdf

Bozdech,Z., Zhu,J., Joachimiak,M.P., Cohen,F.E., Pulliam,B. and DeRisi,J.L. (2003) Expression profiling of the schizont and trophozoite stages of *Plasmodium falciparum* with a long-oligonucleotide microarray. *Genome Biol.*, **4**, R9.

Breslauer,K.J., Frank,M., Blocker,H. and Marky,L.A. (1986) Predicting DNA duplex stability from the base sequence. *Biochemistry*, **83**, 3746–3750.

Burkhardt,S. and Kärkkäinen,J. (2003) Fast lightweight suffix array construction and checking. 14th Annual Symposium, CPM 2003, Morelia, Michocán, Mexico. Lecture Notes in Computer Science, Vol. 2676, pp. 55–69. Springer-Verlag, Berlin.

Carmichael,G.G. (2003) Antisense starts making more sense. *Nat. Biotechnol.*, **21**, 371–372.

Chandler,V.L. and Brendel,V. (2002) The maize genome sequencing project. *Plant Physiol.*, **130**, 1594–1597.

DeRisi,J.L., Iyer,V.R. and Brown,P.O. (1997) Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science*, **278**, 680–686.

Devor,E.J., Huang,L. and Owczarzy,R. (2002) Calculation of $T_m$ for oligonucleotides. IDT technical bulletins. http://www.idtdra.com/program/techbulletins/techbulletins.asp

Gaut,B.S. and Doebley,J.F. (1997) DNA sequence evidence for the segmental allotetraploid origin of maize. *Proc. Nat. Acad. Sci., USA*, **94**, 6809–6814.

Goff,S.A., Ricke,D., Lan,T.H., Presting,G., Wang,R., Dunn,M., Glazebrook,J., Sessions,A., Oeller,P., Varma, H. *et al*. (2002) A draft sequence of the rice genome (*Oryza sativa* L. ssp. *japonica*). *Science*, **296**, 92–100.

Golub,T.R., Slonim,D.K., Tamayo,P., Huard,C., Gaasenbeek,M., Mesirov,J.P., Coller,H., Loh,M.L., Downing,J.R., Caligiuri,M.A. *et al*. (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, **286**, 531–537.

Gusfield,D. (1997) *Algorithms on Strings, Trees and Sequences*. Cambridge University Press, Cambridge, UK.

Helentjaris,T., Weber,D. and Wright,S. (1988) Identification of the genomic locations of duplicate nucleotide sequences in maize by analysis of restriction fragment length polymorphisms. *Genetics*, **118**, 353–363.

Kaderali,L. and Schliep,A. (2002) Selecting signature oligonucleotides to identify organisms using DNA arrays. *Bioinformatics*, **18**, 1340–1349.

Kane,M.D., Jatkoe,T.A., Stumpf,C.R., Lu,J., Thomas,J.D. and Madore,S.J. (2000) Assessment of the sensitivity and specificity of oligonucleotide (50mer) microarrays. *Nucleic Acids Res.*, **28**, 4552–4557.

Kasai,T., Lee,G., Arimura,H., Arikawa,S. and Park,K. (2001) Linear-time longest-common-prefix computation in suffix arrays and its applications. *Combinatorial Pattern Matching, 12th Annual Symposium*, in Jerusalem, Israel. Amir,A. and Landau,G.M. (ed.), Springer-Verlag, Berlin.

Kirkness,E.F., Bafna,V., Halpern,A.L., Levy,S., Remington,K., Rusch,D.B., Delcher,A.L., Pop,M., Wang,W., Fraser,C.M. *et al*. (2003). The dog genome: survey sequencing and comparative analysis. *Science*, **301**, 1898–1903.

Lander,E.S., Linton,L.M., Birren,B., Nusbaum,C., Zody,M.C., Baldwin,J., Devon,K., Dewar,K., Doyle,M., FitzHugh, W. *et al*. (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.

Lehner,B., Williams,G., Campbell,R.D. and Sanderson,C.M. (2002) Antisense transcripts in the human genome. *Trends Genet.*, **18**, 63–65.

Levy,S., Compagnoni,L., Myers,E.W. and Stormo,G.D. (1998) Xlandscapes: the graphical display of word frequencies in sequences. *Bioinformatics*, **14**, 74–80.

Li,F. and Stormo,G.D. (2001) Selection of optimal DNA oligos for gene expression arrays. *Bioinformatics*, **17**, 1067–1076.

Lin,X., Kaul,S., Rounsley,S., Shea,T.P., Benito,M.-I., Town,C.D., Fujii,C.Y., Mason,T., Bowman,C.L., Barnstead,M. *et al*. (1999) Sequence and analysis of chromosome 2 of the plant *Arabidopsis thaliana*. *Nature*, **402**, 761–768.

Lockhart,D.J., Dong,H., Byrne,M.C., Follettie,M.T., Gallo,M.V., Chee,M.S., Mittmann,M., Wang,C., Kobayashi,M., Horton,H. *et al*. (1996) Expression monitoring by hybridization to

high-density oligonucleotide arrays. *Nat. Biotechnol.*, **14**, 1675–1680.

Manber,U. and Myers,E.W. (1993) Suffix arrays: a new method for on-line string searches. *SIAM J. Comput.*, **22**, 935–948.

Martello,S. and Toth,P. (1990) *Knapsack Problems: Algorithms and Computer Implementations*. John Wiley & Sons, New York.

Mayer,K. (1999) Sequence and analysis of chromosome 4 of the plant *Arabidopsis thaliana. Nature*, **402**, 769–777.

Mei,R., Hubbell,E., Bekiranov,S., Mittmann,M., Christians,F.C., Shen,M.M., Lu,G., Fang,J., Liu,W.M., Ryder,T. *et al*. (2003) Probe selection for high-density oligonucleotide arrays. *Proc. Natl. Acad. Sci., USA*, **100**, 11237–11242.

Myers,E.W., Sutton,G.G., Delcher,A.L., Dew,I.M., Fasulo,D.P., Flanigan,M.J., Kravitz,S.A., Mobarry,C.M., Reinert,K.H.J., Remington,K.A. *et al*. (2000) A whole-genome assembly of *Drosophila*. *Science*, **287**, 2196–2204.

Nielsen,H.B., Wernersson,R. and Knudsen,S. (2003) Design of oligonucleotides for microarrays and perspectives for design of multi-transcriptome arrays. *Nucleic Acids Res.*, **31**, 3491–3496.

Osato,N., Yamada,H., Satoh,K., Ooka,H., Yamamoto,M., Suzuki,K., Kawai,J., Carninci,P., Ohtomo,Y., Murakami,K. *et al*. (2003) Antisense transcripts with rice full-length cDNAs. *Genome Biol.*, **5**, R5.

Peyret,N., Seneviratne,P.A., Allawi,H.T. and SantaLucia,J.J. (1999) Nearest-neighbor thermodynamics and NMR of DNA sequences with internal A*A, C*C, G*G, and T*T mismatches. *Biochemistry*, **38**, 3468–3477.

Rahmann,S. (2002) Rapid large-scale oligonucleotide selection for microarrays. *Algorithms in Bioinformatics*, Vol. 2452, pp. 434–434, Springer-Verlag, Berlin.

Relógio,A., Schwager,C., Richter,A., Ansorge,W. and Valcarcel,J. (2002) Optimization of oligonucleotide-based DNA microarrays. *Nucleic Acids Res.*, **30**(11) E51.

Ren,C., Lee,M.K., Yan,B., Ding,K., Cox,B., Romanov,M.N., Price,J.A., Dodgson,J.B. and Zhang,H.B. (2003) A BAC-based physical map of the chicken genome. *Genome Res.*, **13**, 2754–2758.

Rouillard,J.-M., Herbert,C.J. and Zuker,M. (2002) OligoArray: genome-scale oligonucleotide design for microarrays. *Bioinformatics*, **18**, 486–487.

Rouillard,J.M., Zuker,M. and Gulari,E. (2003) OligoArray 2.0: design of oligonucleotide probes for DNA microarrays using a thermodynamic approach. *Nucleic Acids Res.*, **31**, 3057–3062.

Rychlik,W., Spencer,W.J. and Rhoads,R.E. (1990) Optimization of the annealing temperature for DNA amplification *in vitro*. *Nucleic Acids Res.*, **18**, 6409–6412.

SantaLucia,J.J., Allawi,H.T. and Seneviratne,P.A. (1996) Improved nearest-neighbor parameters for predicting DNA duplex stability. *Biochemistry*, **35**, 3555–3562.

Shendure,J. and Church,G.M. (2002) Computational discovery of sense-antisense transcription in the human and mouse genomes. *Genome Biol.*, **3**, Research0044.

Summers,T.J., Thomas,J.W., Lee-Lin,S.Q., Maduro,V.V., Idol,J.R. and Green,E.D. (2001) Comparative physical mapping of targeted regions of the rat genome. *Mamm. Genome*, **12**, 508–512.

The Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana. Nature*, **408**, 796–815.

Venter,J.C., Adams,M.D., Myers,E.W., Li,P.W., Mural,R.J., Sutton,G.G., Smith,H.O., Yandell,M., Evans,C.A., Holt,R.A. *et al*. (2001) The sequence of the human genome. *Science*, **291**, 1304–1351.

Wang,X. and Seed,B. (2003) Selection of oligonucleotide probes for protein coding sequences. *Bioinformatics*, **19**, 796–802.

Waterston,R.H., Lindblad-Toh,K., Birney,E., Rogers,J., Abril,J.F., Agarwal,P., Agarwala,R., Ainscough,R., Alexandersson,M., An,P. *et al*. (2002) Initial sequencing and comparative analysis of the mouse genome. *Nature*, **420**, 520–562.

Wendel,J.F. (2000) Genome evolution in polyploids. *Plant Mol. Biol.*, **42**, 225–249.

Xu,D., Li,G., Wu,L., Zhou,J. and Xu,Y. (2002) PRIMEGENS: robust and efficient design of gene-specific probes for microarray analysis. *Bioinformatics*, **18**, 1432–1437.

Yelin,R., Dahary,D., Sorek,R., Levanon,E.Y., Goldstein,O., Shoshan,A., Diber,A., Biton,S., Tamir,Y., Khosravi,R. *et al*. (2003) Widespread occurrence of antisense transcription in the human genome. *Nat. Biotechnol.*, **21**, 379–86.

Yu,J., Hu,S., Wang,J., Wong,G.K., Li,S., Liu,B., Deng,Y., Dai,L., Zhou,Y., Zhang,X. *et al*. (2002) A draft sequence of the rice genome (*Oryza sativa* L. ssp. *indica*). *Science*, **296**, 79–92.

Zuker,M. (2003) MFold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res.*, **31**, 3406–3415.

Zuker,M., Mathews,D.H. and Turner,D.H. (1999) Algorithms and thermodynamics for RNA secondary structure predictions. *A Practical Guide in RNA Biochemistry and Biotechnology*. Barciszewski,J. and Clark,B.F.C. (ed.) Kluwer Academic Publishers, Dordrecht.

Zuo,F., Kaminski,N., Eugui,E., Allard,J., Yakhini,Z., Ben-Dor,A., Lollini,L., Morris,D., Kim,Y., DeLustro, B. *et al*. (2002) Gene expression analysis reveals matrilysin as a key regulator of pulmonary fibrosis in mice and humans. *Proc. Natl Acad. Sci., USA*, **99**, 6292–6297.