PRODUCT APPLICATION FOCUS

. . . a forum for manufacturers to describe the current and potential applications of new research reagents and products.

Set of Novel Tools for PCR Primer Design

V. Gorelenkov, A. Antipov, S. Lejnine, N. Daraselia, and A. Yuryev
InforMax, Bethesda, MD, USA

BioTechniques 31:1326-1330 (December 2001)

ABSTRACT

We have developed a new package of computer programs and algorithms for different PCR applications, including allele-specific PCR, multiplex PCR, and long PCR. The package is included in the upcoming VectorNTI® suite software and attempts to incorporate most of the current knowledge about PCR primer design. A wide range of primer characteristics is available for user manipulation to provide improved efficiency and increased flexibility of primer design. To accelerate the primer calculations, we have optimized algorithms using recent advances in computer science such as dynamic trees and lazy evaluation. Proper structural organization of input parameters provides further program acceleration. New VectorNTI primer design software allows calculations of primer pairs for long PCR amplification of 120-kb genomic DNA in 5 min under most stringent input parameters and clustering 435 primer pairs for multiplex PCR within 30 min on a standard Pentium III® PC. Our program allows the user to take advantage of molecule annotation by applying different kinds of filtering features during PCR primer design.

INTRODUCTION

Recent advances in PCR technology gave rise to a variety of PCR applications, which will undoubtedly grow in number in the near future. Long and accurate PCR (LA-PCR) applications include cloning of large genomic regions to search for polymorphism, viral genome amplifications (1), and amplification of sequencing gaps in genome sequencing projects (9). The efficiency and specificity of LA-PCR is very sensitive to primers used for amplification (2,8). Primers must be longer than those used in a standard PCR and have a higher melting temperature. The length of a product in LA-PCR slows down significantly the determination of primer uniqueness and necessitates the development of new accelerated algorithmic methods.

The goal of multiplex PCR (M-PCR) is to amplify several products in one reaction (10–12). Multiplex fold is a parameter used to describe a number of products amplified in one tube. Amplifications up to 14 multiplex fold have been reported (5). M-PCR is used for amplification of genetic markers for genotyping (13). The conventional genome-wide genotyping using microsatellite markers requires several hundred markers (7). The development of the single nucleotide polymorphism (SNP) markers for genotyping provides even more challenging scale for PCR amplification of at least 500 000 SNP products (14). Other M-PCR applications include the amplification of multiple viral and other pathogen strains for diagnostics (17) and the amplification of multiple cDNAs for gene expression analysis (20), which may require amplification of several thousand cDNAs, corresponding to every gene in the human genome.

This paper addresses the challenges of primer design for the PCR applications described above. The upcoming VectorNTI® software edition (InforMax, Bethesda, MD, USA) includes two new algorithms for LA-PCR and M-PCR and uses acceleration methods from computer science and incremental calculation of primer characteristics. The majority of genomic information comes in the form of annotated DNA molecules that must be analyzed by PCR. That is why we

Table 1. Comparison of VectorNTI Primer Design Efficiency with Primer3 Public Software Package (16)

	VectorNTI	Primer3	VectorNTI	Primer3
Product length	500–2000	500–2000	10 000–20 000	10 000–20 000
Calculation time (min)	0.3	1	2	Never finished
No. pairs found	More than 50	More than 50	12	0
No. similar primers between two programs	9	9	0	0

The λ phage sequence was used for primer calculations by both programs. The primer parameters were the same in both programs. Primer melting temperature: minimum 57°C, maximum 63°C. Primer length: minimum 18, maximum 27. Maximum complementarity 8.00 for Primer3 and maximum primer similarity 75% for VectorNTI were used. Maximum 3′ end complementarity 3.00 for Primer3 and maximum 3′ end consecutive match 8 bases for VectorNTI were used. No mispriming library was specified for Primer3, and no feature filtering was used in VectorNTI. Primer uniqueness was checked against the entire λ phage sequence in VectorNTI. Primers were considered similar if they overlapped by more than 80% of their sequence.

have introduced filtering features using molecule annotation, which adds another unique concept to our primer design program. The calculation of all primer characteristics is based on old VectorNTI code, which has been tested extensively by 20 000 VectorNTI customers worldwide. Thus, the quality of primer design has not been changed. Our innovations only affect the speed of primer calculations and introduce additional primer design options. The VectorNTI program for PCR design is identical to its old version if a user does not choose these new options. The primer quality can always be improved by stringing the primer design parameters.

MATERIALS AND METHODS

A 1-GGz computer (DELL® Dimension 4100) with 512 MB RAM was used for all calculations. Software code was written on STL C++ programming language using Microsoft® Visual Studio® 6.0. DNA sequences of yeast, human genome, and λ phage were obtained from the public National Center for Biotechnology Information database (http://www.ncbi.nlm.nih.gov).

RESULTS

Parameters for PCR Primer Design

We have introduced a large set of parameters for PCR primer design in an attempt to give the user the ability to manipulate and control all possible primer and product properties. The following standard primer structure properties were used: range of melting temperature calculated by the nearest neighbor method (4,18), range of primer GC content and length, longest allowed homo-polymeric run, maximum allowed palindrome length, and maximum allowed stability of hairpin loop structures (defined by loop length and free energy). In addition, we put special emphasis on primer 3′ end structure, which is crucial for PCR specificity (6,15). VectorNTI allows the user to specify the length of the 3′ end to be included in the analysis, the desired threshold of 3′ end melting free energy, and the desired nucleotides in the last four positions of a primer.

The following are the options for primer sequence uniqueness calculations between primer and PCR template in our program: maximum percent similarity, maximum allowed consecutive match, maximum allowed consecutive match of the primer 3′ end, and maximum allowed percent similarity for the primer 3′ end. Because PCR amplification efficiency favors shorter products, it may be sufficient in many cases to check the uniqueness only within primer binding sites. Depending on the product length and the overall length of the molecule, the introduction of this option accelerated primer calculations 2–10 times compared to calculating primer uniqueness in the entire molecule.

The product region that immediately follows the 3′ end of a primer was shown to influence the amplification efficiency (15). To allow the user control over product properties, the following product parameters were introduced: the range of the product GC content and the GC content of the product regions immediately adjacent to primer 3′ end. The user can also specify the first bases of a product immediately following the primer 3′ end.

We have arranged the calculation of primer properties in the reciprocal order of the processor time required for their calculation (Figure 1A). Parameters requiring faster times are calculated first, so fewer primers are used as an input for slower steps. VectorNTI calculates qualities for every primer as deviations from the required range of parameters. Primer qualities are then recalculated in primer acceptance/rejection criterion based on the importance of each parameter as specified by the user (Figure 1B). Output is sorted by primer qualities before presenting it to the user.

Use of DNA Molecule Annotation for PCR Primer Design

We have taken advantage of the existing capabilities of InforMax VectorNTI software, which allow the user to manually annotate any DNA or protein molecule by creating differ-

ent kinds of features. The molecule annotation provides an opportunity to use the filtering features for primer design. For example, using our "Feature reject" filter and sequence annotated with "Repetitive region" features one can omit repetitive regions from PCR primer design analysis. Conversely, if only cDNAs are needed for amplification, our "Feature include" filter will limit the primer design only to the molecule regions containing annotated cDNAs (Figure 1C). It also allows the user to design primers for allele-specific PCR when DNA molecule is annotated with "Variation" features and to design primers for nested multiplex PCR from a single cDNA annotated with exons. We have accelerated the filtering features using dynamic mapping of DNA template position to a list of features. Such dynamic mapping allows direct access during the program run only to those features that are present in the region of analysis, omitting features that are located outside. Filtering features are incorporated in all VectorNTI PCR primer design algorithms.

Primer Design Algorithm for Multiplex PCR

We have developed a new algorithm for M-PCR primer design that presents the user with calculated product clusters that can be amplified in one PCR tube. The user must specify multiplex fold, which corresponds to a minimum allowed size of product clusters. The basic requirements for M-PCR primers and product compatibility are as follows. (*i*) The primers used in one M-PCR must not form dimers with each other, and (*ii*) primer pairs used in one M-PCR must anneal only to their product. The latter condition can be stringed even further by requiring that primers should not anneal to any PCR template in M-PCR except its own. This makes primers template specific, rather than product specific. Product length difference was introduced as an additional parameter to separate PCR products by gel electrophoresis after amplification. VectorNTI calculates primers and PCR products for every product according to a standard set of user-defined requirements for

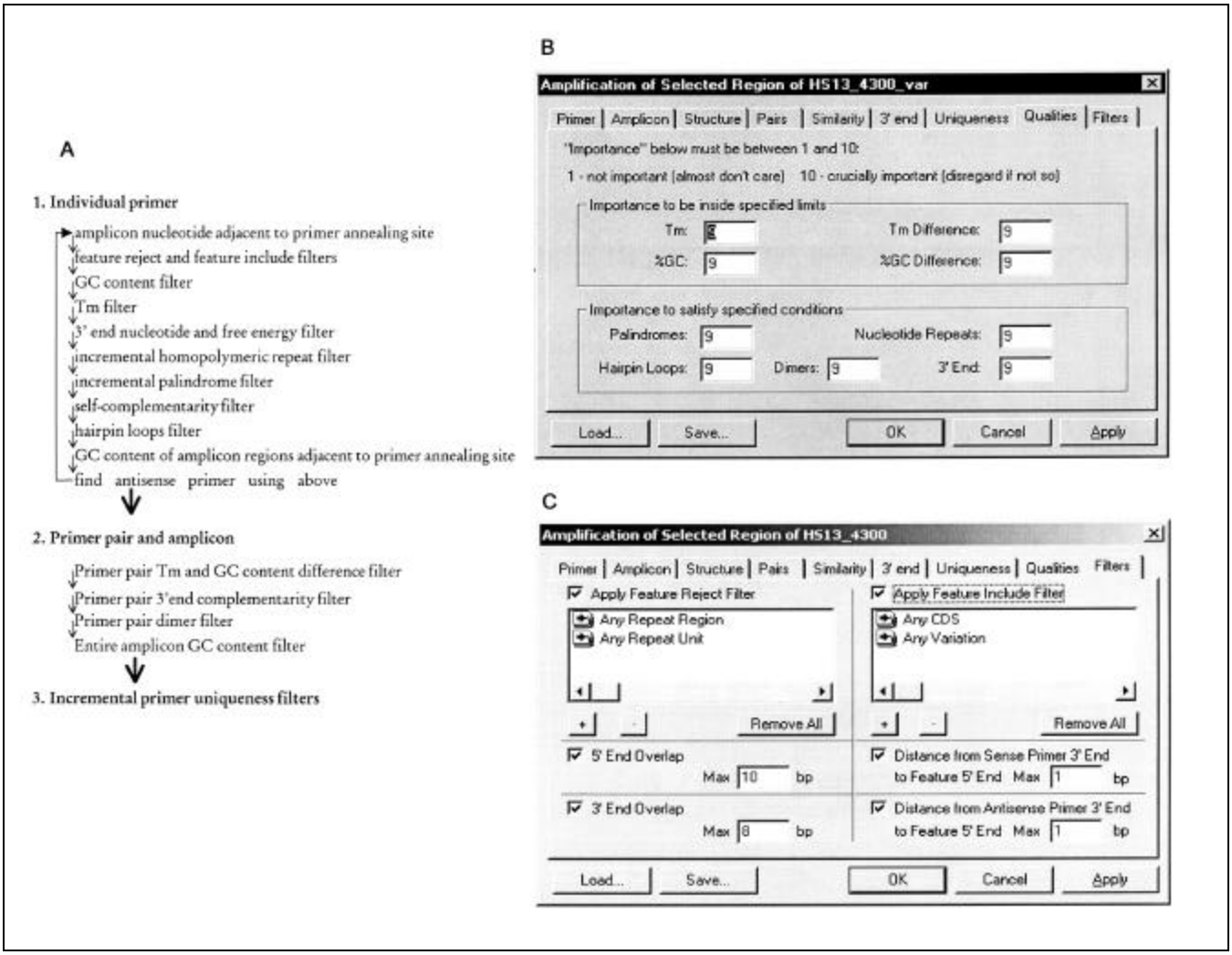

Figure 1. Primer design in VectorNTI. (A) Flowchart of consecutive filters used in our standard PCR primer design protocol. The filters that require smaller calculation time are used first. Incremental filters calculate primer properties using properties calculated for a primer from the previous program cycle, which differs only by one nucleotide. (B) A screen shot of the VectorNTI dialog for assigning importance levels for each PCR primer design parameter. (C) A screen shot of the VectorNTI dialog for the filtering features. Current settings will design allele-specific primers in coding regions (CDS), omitting repetitive regions.

PCR primer (Figure 2B). PCR products are clustered using the above product compatibility parameters. Two products are compatible if primers from one product satisfy the uniqueness requirements in a second product and vice versa.

The clustering criteria can be presented as a logical conjunction of two steps. The first step calculates the product length difference and the intersection of products from the same molecule and demands little calculation efforts. The second step calculates primer cross-complementarily and cross-uniqueness and needs more computation time. The algorithm uses the first criteria to build an initial cluster approximation, and the cluster is then refined using the second criteria (Figure 2A). During the clustering step, the program builds square Boolean compatibility matrix, which holds all possible PCR product pairs. If two products do not satisfy the compatibility requirements, the element in the matrix is set FALSE; otherwise, it is TRUE. The compatibility matrix is symmetrical. Two symmetrical elements correspond to both product rows and columns. Matrix calculations were optimized using the lazy evaluation programming technique (19). This technique allowed us to avoid unnecessary calculation of the elements. The elements are calculated only if it is necessary to compare (or evaluate) the matrix element during the calculation.

The unique features of our M-PCR algorithm are the calculation of primer cross-uniqueness with other products and the speed of calculations achieved through the use of the lazy evaluation technique. Using the VectorNTI program for M-PCR, we were able to clusterize 849 yeast cDNAs from yeast chromosome three and four in 1 h on the 1-GGz computer.

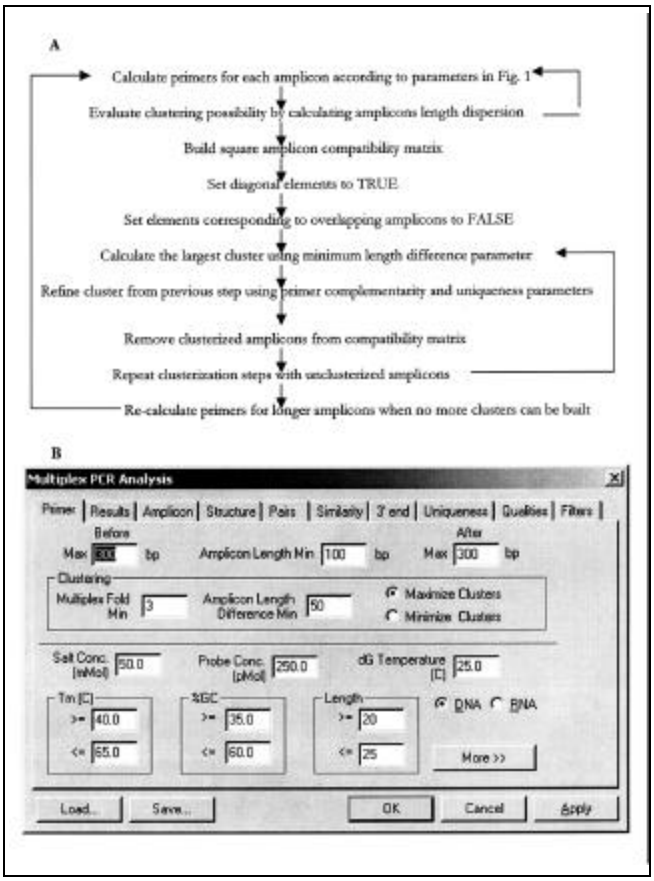

Figure 2. PCR multiplexing in VectorNTI. (A) Algorithmic steps of VectorNTI multiplex PCR primer design. If user requests the product length difference, then the program first evaluates clustering possibility before every clusterization from the point of view of this parameter. Then, an approximate product cluster is calculated according to the length compatibility requirements. Other compatibility requirements: primer-primer complementarity and primer uniqueness are calculated only within this cluster because they require longer calculation times. After the largest possible cluster is calculated, its products are removed from the matrix and unclusterized products are clusterized again. After all possible clusterization attempts, the remaining products are directed for a new step of the primer design with increased product length, and another round of clusterization follows. The program terminates when the number of unclusterized products is smaller than the required multiplex fold or when all re-amplifications attempts are made. When the "Minimize cluster" option is used, the number of products exactly equal to multiplex fold is removed from the matrix after each clusterization attempt. This leads to longer clusterization times than in the original algorithm. (B) A screen shot of the VectorNTI dialog for multiplex PCR primer design.

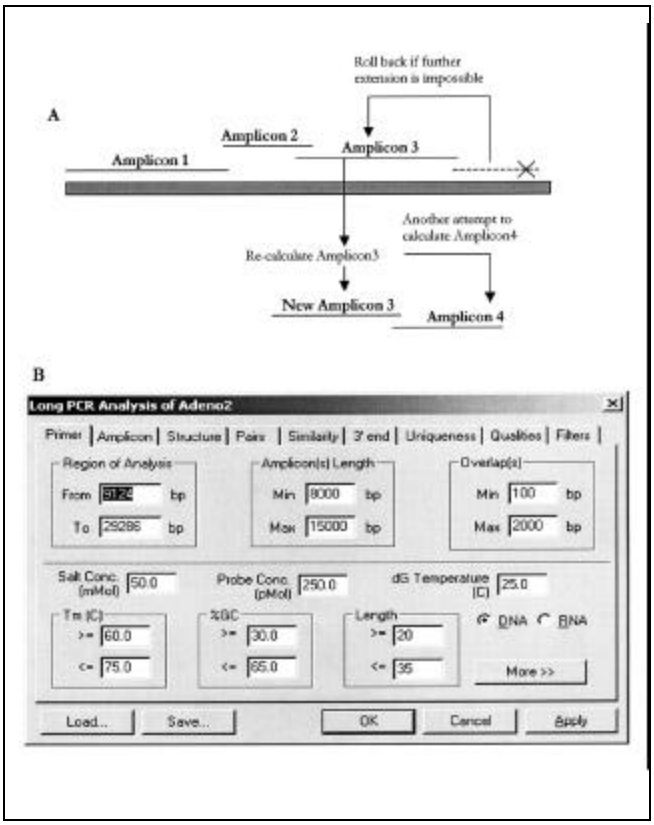

Figure 3. Long PCR primer design in VectorNTI. (B) Illustration of VectorNTI algorithm for LA-PCR primer design. The algorithm starts to calculate PCR primers at the beginning of a region selected for amplification according to user-specified parameters for products and primers. It then proceeds to the calculation of PCR primers for the second product, which overlaps with the first one by a user-defined overlap. The iterations continue until the entire region is completely covered by products. If during any iteration step it is impossible to find primers satisfying input parameters, then the algorithm will roll back to calculate another primer pair at a previous step for a shorter product. This rollback allows the user to consider another region for primer design at failed iteration. The consecutive rollbacks may be required if every rollback step fails to produce primers until they reach the very first product. The process stops if no primers can be recalculated at the first iteration. (B) A screen shot of the VectorNTI dialog for long PCR primer design.

About 100 Mb RAM were used during the program run. The calculation time grows as the square of multiplex fold in our algorithm.

Primer Design for Long PCR

Current LA-PCR technology can amplify up to 42–50 kb products from a homogeneous template, such as λ phage (1) and up to 22–30 kb fragments from a heterogeneous mixture of genomic DNA (8). The amplification of longer regions of 100–1000 kb still remains impossible. Such amplification would be desirable during positional cloning projects, in molecular haplotyping, and in the search for polymorphism in individual genomes. The practical solution to this problem is covering big genomic regions with a set of overlapping LA-PCR products of shorter length (Figure 3A). The presence of repetitive sequences, primer melting temperature, and other limitations on primers complicates LA-PCR primer calculations. We have developed an algorithm that allows the design of such fragments and presents the user with a set of products covering an entire DNA region (Figure 3B). To the best of our knowledge, there is no algorithm that finds primers for such "product walking" to cover large DNA regions. The high speed of calculations for LA-PCR also surpasses any other programs for primer design. Primers to amplify 100 kb human genomic DNA were determined by our algorithm in less then 10 min using the 1-GGz computer. Algorithm allows user to specify the range of product lengths and the range of product overlaps. Table 1 contains the comparison of our primer design software with public Primer3™ package from the Massachusetts Institute of Technology (16).

DISCUSSION

Our new software for PCR primer design allows further automation of the laborious task of PCR primer design. It gives molecular biologists a rich set of tools for choosing primer characteristics. New algorithms for M-PCR and LA-PCR are implemented in the user-friendly interface of VectorNTI software. The interface provides easy handling of a large number of molecules, convenient feature annotation of DNA molecules, and specification of desired PCR primer parameters. Because of a large set of primer parameters, we have provided an option to save and load favorite parameters from the external file. Scientists can determine experimentally the best parameters for a certain type of PCR application and save them for future use or exchange them with collaborators. The sets of molecules repeatedly used for M-PCR can also be saved in a specific file.

The design of our M-PCR clusterization algorithm allows the further incorporation of parameters for primer and product compatibility. For example, the appearance of chimeric molecules during the last rounds of PCR amplification is a well-known PCR artifact (3). It appears when incomplete PCR product anneals to another product and allows polymerase to extend using it as a template. The obvious solution to this problem is to avoid products in the M-PCR cluster that have homology regions.

REFERENCES

1. **Akin, A., C.C. Wu, and T.L. Lin.** 1999. Amplification and cloning of infectious bursal disease virus genomic RNA segments by long and accurate PCR. J. Virol. Methods *82*:55-61.
2. **Barnes, W.M.** 1994. PCR amplification of up to 35-kb DNA with high fidelity and high yield from lambda bacteriophage templates. Proc. Natl. Acad. Sci. USA *91*:2216-2220.
3. **Brakenhoff, R.H., J.G. Schoenmakers, and N.H. Lubsen.** 1991. Chimeric cDNA clones: a novel PCR artifact. Nucleic Acids Res. *19*:1949.
4. **Breslauer, K.J., R. Frank, H. Blocker, and L.A. Marky.** 1986. Predicting DNA duplex stability from the base sequence. Proc. Natl. Acad. Sci. USA. *83*:3746-3750.
5. **Broude, N.E., L. Zhang, K. Woodward, D. Englert, and C.R. Cantor.** 2001. Multiplex allele-specific target amplification based on PCR suppression. Proc. Natl. Acad. Sci. USA. *98*:206-211.
6. **Buck, G.A., J.W. Fox, M. Gunthorpe, K.M. Hager, C.W. Naeve, R.T. Pon, P.S. Adams, and J. Rush.** 1999. Design strategies and performance of custom DNA sequencing primers. BioTechniques *27*:528-536.
7. **Buetow, K.H., J.L. Weber, S. Ludwigsen, T. Scherpbier-Heddema, G.M. Duyk, V.C. Sheffield, Z. Wang, and J.C. Murray.** 1994. Integrated human genome-wide maps constructed using the CEPH reference panel. Nat. Genet. *6*:391-393.
8. **Cheng, S., C. Fockler, W.M. Barnes, and R. Higuchi.** 1994. Effective amplification of long targets from cloned inserts and human genomic DNA. Proc. Natl. Acad. Sci. USA *91*:5695-5699.
9. ***C. elegans* Sequencing Consortium.** 1998. Genome sequence of the nematode *C. elegans*: a platform for investigating biology. Science *282*:2012-2018.
10. **Doi, K. and H. Imai.** 1999. A greedy algorithm for minimizing the number of primers in multiple PCR experiments. Genome Inform Ser Workshop Genome Inform. *10*:73-82.
11. **Edwards, M.C. and R.A. Gibbs.** 1994. Multiplex PCR: advantages, development, and applications. PCR Methods Appl. *3*:S65-S75.
12. **Elnifro, E.M., A.M. Ashshi, R.J. Cooper, and P.E. Klapper.** 2000. Multiplex PCR: optimization and application in diagnostic virology. Clin. Microbiol. Rev. *13*:559-570.
13. **Fujii, K., H. Senju**, K. **Yoshida, K. Sekiguchi, K. Imaizumi, K. Kasai, and H. Sato.** 2000. Multiplex PCR amplification of TH01, D9S304, and D3S1744 loci. J. Hum. Genet. *45*:303-304.
14. **Kruglyak, L.** 1999. Prospects for whole-genome linkage disequilibrium mapping of common disease genes. Nat. Genet. *2*:139-144.
15. **McGrath, A., D.G. Higgins, and T.V. McCarthy.** 1998. Sequence analysis of DNA randomly amplified from the *Saccharomyces cerevisiae* genome. Mol. Cell Probes *12*:397-405.
16. **Rozen, S. and H. Skaletsky.** 2000. Primer3 on the WWW for general users and for biologist programmers. Methods Mol. Biol. *132*:65-86.
17. **Storch, G.A.** 2000. Diagnostic virology. Clin. Infect. Dis. *31*:739-751.
18. **Sugimoto, N., S. Nakano, M. Katoh, A. Matsumura, H. Nakamuta, T. Ohmichi, M. Yoneyama, and M. Sasaki.** 1995. Thermodynamic parameters to predict stability of RNA/DNA hybrid duplexes. Biochemistry *34*:11211-11216.
19. **Wadler, P.** 1996. Lazy *vs.* strict. ACM Computing Surveys *28*:318-320.
20. **Wikman, F.P., M.L. Lu, T. Thykjaer, S.H. Olesen, L.D. Andersen, C. Cordon-Cardo, and T.F. Orntoft.** 2000. Evaluation of the performance of a p53 sequencing microarray chip using 140 previously sequenced bladder tumor samples. Clin. Chem. *46:*1555-1561.

Address correspondence to:

Dr. Anton Yuryev
Informatics Department
Orchid BioSciences, Inc.
303 College Road East
Princeton, NJ 08540
e-mail: ayuryev@orchid.com