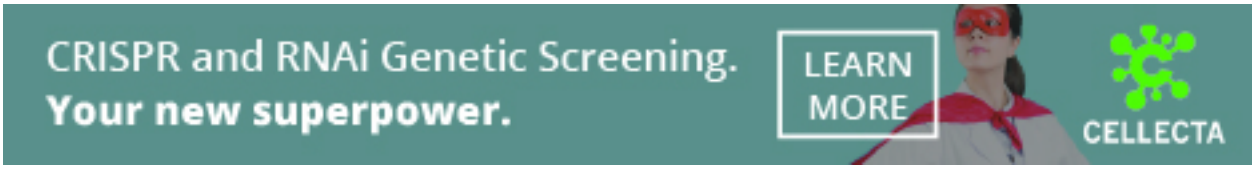
OSP: a computer program for choosing PCR and DNA sequencing primers.

L Hillier and P Green

Genome Res. 1991 1: 124-128
Access the most recent version at doi:10.1101/gr.1.2.124



OSP: A Computer Program for Choosing PCR and DNA Sequencing Primers

LaDeana Hillier and Philip Green

Genetics Department, Washington University School of Medicine
St. Louis, Missouri 63110

OSP (Oligonucleotide Selection Program) selects oligonucleotide primers for DNA sequencing and the polymerase chain reaction (PCR). The user can specify (or use default) constraints for primer and amplified product lengths, %(G+C), (absolute or relative) melting temperatures, and primer 3′ nucleotides. To help minimize nonspecific priming and primer secondary structure, OSP screens candidate primer sequences, using user-specifiable cutoffs, against potential base-pairing with a variety of sequences present in the reaction, including the primer itself, the other primer (for PCR), the amplified product, and any other sequences desired (e.g., repetitive element sequences in genomic templates, vector sequence in cloned templates, or other primer pair sequences in multiplexed PCR reactions). Base-pairing involving the primer 3′ end is considered separately from base-pairing involving internal sequences. Primers meeting all constraints are ranked by a "combined score," a user-definable weighted sum of any of the above parameters. OSP is being routinely and extensively used to select sequencing primers for the *Caenorhabditis elegans* genome sequencing project and human genomic PCR primer pairs for the Washington University Genome Center mapping project, with success rates exceeding 96% and 81%, respectively. It is available for research purposes from the authors, at no cost, in both text output and interactive graphics (X windows) versions.

DNA sequencing by the Sanger chain-termination method,[1] and the polymerase chain reaction (PCR),[2] are ubiquitous molecular biology tools that require oligonucleotides for priming DNA synthesis. Not all primers work, and those that do vary widely in their resistance to changes in reaction conditions, a situation that has spurred the development of several computer programs for choosing primers.[3,4]

The onset of major mapping and sequencing efforts as part of the Genome Initiative has increased the demand for efficient primer selection. OSP (Oligonucleotide Selection Program) was designed to satisfy the needs of two projects: the Washington University Genome Center's project to map human chromosomes 7 and X, which will require PCR assays to detect well over 1000 sequence-tagged sites (STSs)[5] in human genomic DNA, and the *Caenorhabditis elegans* genome sequencing project, which is presently using a mixed shotgun sequencing and walking strategy requiring selection of sequencing primers (roughly 100 per cosmid) for the walking steps. Major requirements for OSP were ease of use and flexibility in setting primer selection criteria, so that the program could evolve in the light of experience gained in both projects. In this paper we describe OSP, and summarize results obtained using it.

METHODS

OSP: Program Design and Operation

Input of sequences

For selecting PCR primers, the user provides either one sequence that includes the region to be amplified or two sequences flanking that region; for selecting sequencing primers, a single sequence must be provided. Either strand may be given, and ambiguous nucleotides within the sequence are permitted (although primers containing ambiguous nucleotides are rejected). Sequences may either be entered at run time (in the graphics version of the program, they may be "pasted" in from another window) or provided as a text file(s) in a variety of possible formats. The user may confine the search for candidate primers to particular regions of the input sequence. The primer locations may also be specified exactly, if the user wishes to compute scores or other characteristics for previously selected primers.

Constraints

Primer and amplified product characteristics that may be constrained are listed in Table 1, along with OSP's default values, which have been found to work well in several mapping and sequencing projects (see Results). The constraint criteria are those thought (largely on the basis of anecdotal evidence) to be important to ensure sensitive and specific priming: for example, if G+C content is too high, the primer may be more prone to adopt secondary structure or to anneal nonspecifically to GC-rich regions of the template DNA, while if it is too low the primer may not anneal to its target sequence. If the user does not wish to use the defaults, modified constraint values for any or all parameters may be provided in a "constraint file," and/or entered directly at run time (the constraints for any parameter can also be disabled).

"Annealing scores" are intended to provide simple measures of base-pairing propensity of the primer with

1:124–128©1991 by Cold Spring Harbor Laboratory Press ISSN 1054-9803/91 $3.00

TABLE 1 User-constrainable Parameters, Default Limits, and Default Weights

	Minimum	Maximum	Weight
Single primer parameters			
3′ nucleotides[a]	S[a]	—[b]	
Primer length (nucleotides)[c]	18	22	0.0
Primer G+C content (%)	40.0	55.0	0.0
Primer T_m (Celsius)[d]	50.5	55.0	0.0
Annealing scores[e]			
Primer–self internal	—	14.0	1.0
Primer–self 3′	—	8.0	2.0
Primer–other internal	—	NS[f]	0.0
Primer–other 3′	—	NS	0.0
Primer pair parameters			
Product length (bp)[g]	100	300	0.0
Product G+C content (%)	40.0	55.0	0.0
Product T_m (Celsius)	70.0	90.0	0.0
Primer T_m difference (Celsius)	—	2.0	0.0
Annealing scores			
Primer–primer internal	—	14.0	1.0
Primer–primer 3′	—	8.0	2.0
Primer–product internal	—	NS	0.0
Primer–product 3′	—	NS	0.0

[a]Any string of 3′ terminal nucleotide(s) for the primers may be specified, with all standard ambiguity codes allowed. The default is a single nucleotide S, which matches either C or G.
[b](—)Not meaningful.
[c]Values shown are for PCR primers. For sequencing primers, defaults are min. = 17, max. = 18.
[d]T_m (melting temperature) is calculated as: 62.3 + 0.41(%(G+C)) − 500/length.[(6)] Any ambiguous nucleotides are ignored.
[e]See text for explanation.
[f](NS) Not set (in OSP defaults).
[g]For sequencing primers, product length is the distance from the end of the sequence. Default constraints are min. = 40, max. = 100.

the various nontarget sequences present in the reaction. They are computed as follows: The user may define a score for an A/T base pair (default value, 2) and for a C/G base pair (default value, 4). For any alignment of two sequences (with the first sequence written 5′ to 3′, and the second one 3′ to 5′) scores are summed over contiguous complementary nucleotide pairs; the maximum such score, taken over all contiguous blocks in all alignments, is the internal annealing score. The 3′ annealing score (which is computed and weighted separately because it likely has a greater influence on nonspecific priming) is based only on contiguous complementary matches that include the 3′ end of the primer. Annealing scores are calculated for the primer with itself, with the other primer, with the amplified product, and with a user supplied set of "other" sequences. At a minimum, the latter would typically include repetitive sequences when the template is genomic DNA, or vector sequence when the template is cloned DNA. It could also include other primer sequences if one is interested in "multiplexing" multiple PCR primer pairs in the same reaction tube.

For example, the alignment

```
5′-ACGGATTGTGCCGTATTG-3′
3′-      TGACACGTGTAGACGAGG-5′
```

of two candidate primers would yield an internal annealing score of 16 (using the default base-pair scores), corresponding to the pairing of TGTGC on the top strand with ACACG on the bottom, and a 3′ annealing score of 6, corresponding to the pairing of TG at the 3′ end of the first primer with AC. Since the default internal annealing cutoff (14.0) is exceeded, this pair would be rejected by OSP.

Primer Selection Algorithm

OSP scans each strand in a 5′ to 3′ direction for possible primers. For each position within the sequence, it tests candidate primers having that position as the 3′ terminus, examining in order the criteria in the upper half of Table 1. Primers containing an ambiguous nucleotide are automatically rejected. Primers with the same 3′ end are considered in order of increasing length; this allows consideration of longer primers to be aborted when "monotonically increasing" parameters (e.g., primer-self or primer-other annealing) exceed their constraint values. If all criteria are satisfied, the primer is saved in a list of candidates for that strand.

PCR Primer Pair Selection Algorithm

Following construction of lists of candidate primers for each strand, candidate primer pairs (one from each list) are considered. If several pairs have the same 3′ ends for both primers, at most one pair (the shortest) is accepted. For each candidate pair, constraints involving the amplified product and primer pair (lower half of Table 1) are tested, and any pair meeting each applicable constraint is added to the list of valid pairs.

Once all possible primer pairs have been considered, they are ranked based on their "combined score," which is a user-definable weighted sum of any of the parameters in Table 1. The default combined score (which uses the default weights in Table 1) is based on the primer-self and primer-primer annealing scores, and weighs 3′ annealing twice as much as internal annealing. The user may supply different weights in the constraint file.

Program Output

OSP notifies the user of the number of primers (or primer pairs) accepted, and provides a table listing the number of pairs rejected for each constraint. (This is often useful in cases where no pair meets all constraints.) The user can then choose to save (to a file) a ranked list of as many of the primers as desired, along with their characteristics. The output file also records the constraint values used and the numbers of primers or pairs accepted and rejected.

Graphic Interface

A menu-driven "point and click"

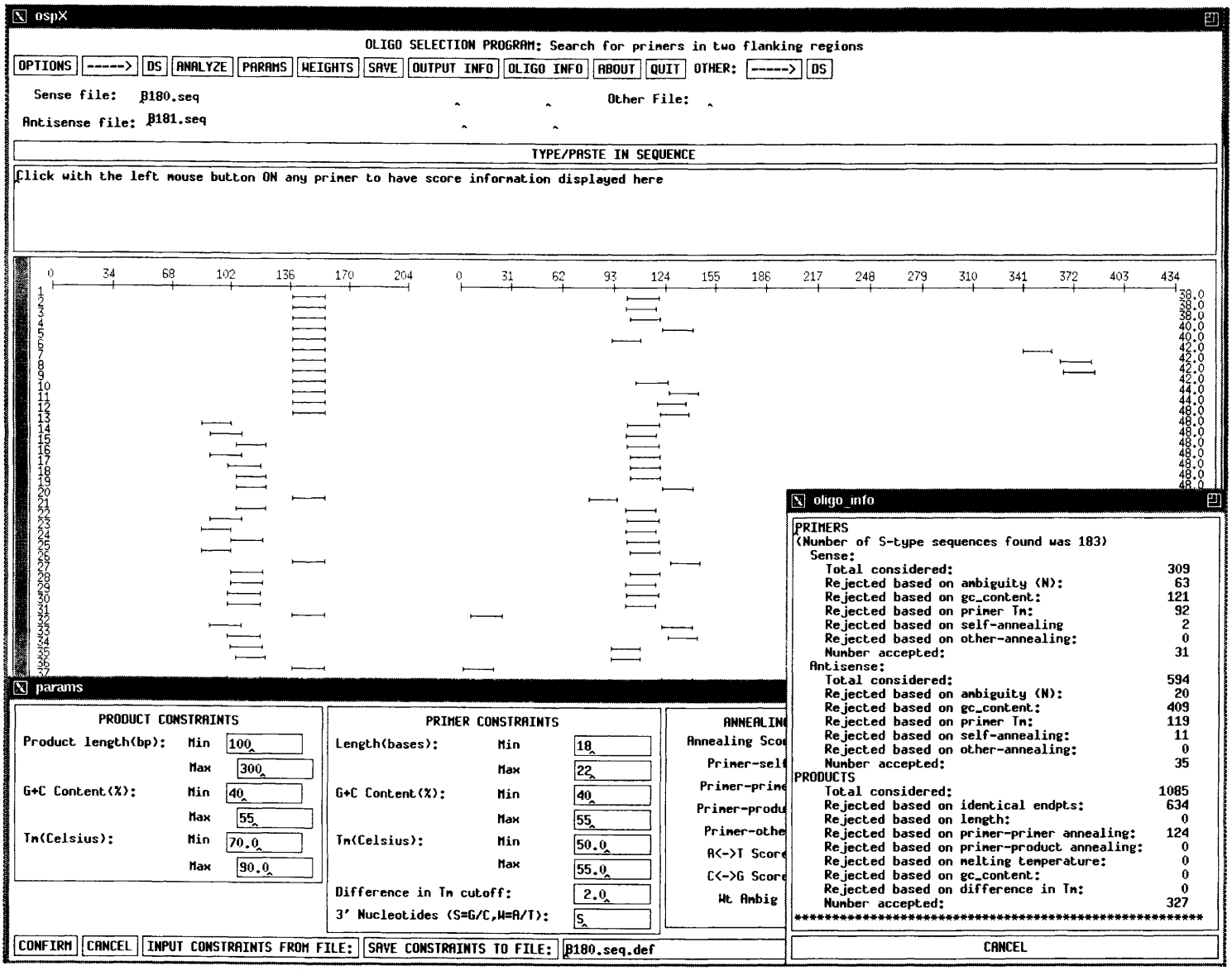


FIGURE 1 Screen dump of OSP display. See text for explanation.

graphical display version of OSP has also been developed, which runs under X windows (Fig. 1). Clicking on the OPTIONS box brings up a menu with the program options. The sequence(s) are entered either by specifying a file name(s) or by typing or "pasting" sequence into a sequence window. Any constraint or weight may be altered by entering values in a "constraints" or "weights" text window. The user may alter the default values displayed in the constraints window by specifying a "constraint file" name on the command line.

Analysis is initiated by clicking on the ANALYZE box. The graphic results window displays a line representing the input sequence(s), with locations of the ranked primers displayed beneath it. Clicking on a primer causes information concerning it to be displayed in the text results window. The number of pairs accepted, along with a breakdown by constraint of the number rejected, can be displayed by clicking on OLIGO INFO. Any constraint can be changed and the program rerun by clicking on the appropriate buttons. Clicking on the SAVE box outputs a specified number of primer pairs to a file.

If the sequence file is accompanied by a trace data file from a fluorescent sequencing machine (ABI 373A or Pharmacia A.L.F.), clicking above the line representing the sequence calls up the display of the trace (using the ted trace editor[7]) for that portion of the sequence. This feature has proven useful for verifying accuracy of the sequence in the vicinity of candidate primers, in the *C. elegans* sequencing project.

RESULTS

OSP has been used to choose over 365 sequencing primers and over 300 PCR primer pairs. Table 2 gives the "success rates" (proportion of primers successfully used in sequencing reactions) for four cosmids in the *C. elegans* genome sequencing project.[8] This project involves a two-stage strategy for sequenc-

TABLE 2 OSP-selected *C. elegans* Sequencing Primer Success Rates

Project(a)	Template DNA	Number successful/ total(%)[b]	Data source
B0303	plasmid	95/102 (93.1)	R. Wilson
ZK370	plasmid	44/45 (97.8)	R. Wilson
ZK643	plasmid/M13	109/110 (99.1)	J. Sulston
F59B2	plasmid/M13	98/100 (98.0)	T. Hawkins

[a]Cosmids from *C. elegans* physical map. OSP default constraint settings were used, with the following exceptions: min. primer T_m = 50, max. = 60; for ZK643 and F59B2, primer length constraints were min. = 17, max. = 23, and min. primer G+C content = 30%. For sequencing reaction conditions, see ref. 9.
[b]For 20 of the ZK643 and F59B2 primers, it was necessary to rerun the reactions at a higher temperature to obtain specific priming; these are included among the successes.

ing cosmids from the *C. elegans* physical map: an initial "shotgun" stage, in which a single sequence tract is generated from each of several hundred large-insert M13 or plasmid subclones of the cosmid and assembled into contigs, followed by a "walking" stage to fill gaps between contigs, to obtain double-stranded coverage in all regions, and to resolve any remaining sequence ambiguities. OSP is used to choose the walking primers.

The success rates for PCR primers are given in Table 3. The bulk of these primers were selected for the *C. elegans* physical mapping project, and for the 7 and X chromosome mapping projects in the Washington University Genome Center. For the latter (in which sequences derived from λ or M13 genomic clones were used to choose the primers) we include as "successes" all cases in which a band of the anticipated size was obtained from amplification of the genomic DNA source from which the clone library was prepared, whether or not the sequence ultimately turned out to be single copy in the genome or mapped to the desired human chromosome.

The success rates in these tables are likely to be underestimates of OSP's effectiveness at choosing primers, since failures may reflect use of incorrect sequence, oligonucleotide synthesis errors, reaction failures, or (in the genomic PCR cases) sequencing from clones which may not be faithful replicas of the genome, rather than poor choices by OSP. In particular, cloning artifacts almost certainly account for a substantial fraction of the PCR failures. In the chromosome 7 project, most of the failed primer pairs tested were found to amplify the clone DNA successfully, even when it was mixed with genomic DNA such that the clone sequence was present at the same concentration as single-copy genomic sequences (E.D. Green, pers. comm.). Thus, these failures more likely result from a fraction of the clones that contain rearranged or foreign DNA, rather than from nonrobust primers. In this light, it is interesting that the PCR success rates appear to depend significantly on the method used to construct the clones or subclones from which the sequence was determined: the overall success rate was only 76.2% for sequences inferred from M13 subclones of flow-sorted human chromosomes (lines 1, 3, and 5 in Table 3), while for sequences obtained by other methods (lines 2 and 4 in Table 3), it was 88.8%. This suggests that, at least in these experiments, M13 cloning from small amounts of source DNA may have been more prone to rearrangements or inclusion of contaminating DNA, and that OSP's success rate at choosing human genomic PCR primers may approach 90% when the genomic sequence is accurately known.

The variation in success rates among the mapping projects probably also reflects differing efforts to optimize the PCR conditions; for example, a larger variety of temperature regimes and buffers were tried in the chromosome 7 project (cf. ref. 10) than in the X chromosome project, which may account for its higher success rate.

It should also be noted that these success rates will likely be improved by OSP's testing for base-pairing with sequences in the "other sequence" file, which was not available at the time these experiments were done.

TABLE 3 OSP-selected PCR Primer Pair Success Rates

Project	Template DNA	Number successful/ total(%)	Data source
Chromosome 7[a]	human genomic	63/80 (78.8)[b]	E.D. Green
		55/65 (84.6)[c]	E.D. Green
Chromosome 13	human genomic	14/17 (82.4)[b]	P. Kwok
		24/24 (100.0)[d]	P. Kwok
Chromosome X	human genomic	19/29 (65.5)[b]	J. Kere
C. elegans physical map	*C. elegans* cosmid	46/46 (100.0)	Y. Kozono
C. elegans physical map	*C. elegans* YAC	41/44 (93.2)	Y. Kozono

[a]Default OSP constraint settings were relaxed as follows: min. product length, 60; max. product T_m, 82; max. G+C content, 60%; max. difference between primer T_m's, 3. Sequences for which OSP found no primer pairs meeting all constraints were skipped. See ref. 10 for a description of the PCR conditions used.
[b]Sequences derived from M13 clones of flow-sorted human chromosomes.
[c]Sequences derived from "bubble PCR" of λ clones.
[d]Sequences derived from bubble PCR of YAC ends. Many of these primers were also used for sequencing (successfully, in all cases).

DISCUSSION

The detailed kinetics of PCR and sequencing reactions, in particular primer–template interactions, are not well understood, and as a result the criteria currently used to choose primers are largely empirical. Primer T_m is of obvious relevance for the temperature cycling protocols. Primer %G+C content is also thought to be important, although it is unclear whether this is because of its influence on secondary structure or nonspecific annealing to template DNA. Specificity of the priming reaction should be enhanced, and primer secondary structure and primer–primer interactions

largely avoided, by minimizing potential primer base-pairing with nontarget sequences in the reaction mixture, particularly those present in high copy number. Thus, it is desirable to ensure low annealing of the primer with itself; with the other primer (in PCR), with (nontarget) sites in the amplified product; with repetitive elements, when the template is genomic DNA; and additionally, when the template is cloned DNA, with any nontarget sequences known to be present in the clone (such as the vector sequence). OSP allows the calculation of a simple base-pairing or "annealing" score for the primer with each of these types of sequences.

In comparison to other primer selection programs we have examined, OSP is unusual in the range of user-settable parameters employed (for example, base-pairing involving the internal primer sequence can be scored differently from base-pairing at the 3′ end, which is presumably more likely to lead to nonspecific priming), in allowing prescreening against other sequences, and in the flexibility allowed the user in setting constraints and ranking candidate primers. Any criterion can be altered to find primer-pairs in a particularly difficult sequence or to meet different experimental conditions. This flexibility allows the user's selection criteria to evolve in the light of laboratory experience.

As yet, no method for choosing primers has been tested in a controlled experiment, and OSP is no exception. Since the criteria it uses are much the same as those used by investigators without such a program available to them, its success rates quite possibly do not exceed those attainable by an expert without computer assistance. Nonetheless, at a minimum OSP allows the choice to be made rapidly and automatically, and has success rates that exceed those often reported anecdotally.

AVAILABILITY

C language source code for OSP is available (for research purposes only) at no cost from the authors, in either the text output version (tested for VAX/VMS, PC, MAC, and SUN Sparcstations), or interactive X windows graphics version (tested for SUN Sparcstations).

ACKNOWLEDGMENTS

We thank Pui Kwok, John Sulston, Eric Green, Trevor Hawkins, Juha Kere, Bob Waterston, Rick Wilson, and Yuko Kozono for suggestions to improve the program's ease of use, and for providing data concerning their primer success rates. This work was supported by National Institutes of Health grants P50-HG00201 and R01-HG00136, and by a New Faculty Award from the Lucille P. Markey Charitable Trust.

REFERENCES

1. Sanger, F., S. Nicklen, and A.R. Coulson. 1977. DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci.* **74:** 5463–5467.
2. Mullis, K.B. and F.A. Faloona. 1987. Specific synthesis of DNA in vitro via a polymerase-catalyzed chain reaction. *Methods Enzymol.* **155:** 335–350.
3. Lowe, T., J. Sharefkin, S.Q. Yang, and C.W. Dieffenbach. 1990. A computer program for selection of oligonucleotide primers for polymerase chain reactions. *Nucleic Acids Res.* **18:** 1757–1761.
4. Rychlik, W. and R.E. Rhoads. 1989. A computer program for choosing optimal oligonucleotides for filter hybridization, sequencing, and *in vitro* amplification of DNA. *Nucleic Acids Res.* **17:** 8543–8551.
5. Olson, M., L. Hood, C. Cantor, and D. Botstein. 1989. A common language for physical mapping of the human genome. *Science* **245:** 1434–1435.
6. Bolton, E.T. and B.J. McCarthy. 1962. A general method for the isolation of RNA complementary to DNA. *Proc. Natl. Acad. Sci.* **48:** 1390–1394.
7. Gleeson, T. and L. Hillier, in preparation.
8. Sulston, J., R. Ainscough, M. Berks, M. Craxton, A. Coulson, S. Dear, Z. Du, R. Durbin, P. Green, N. Halloran, T. Hawkins, L. Hillier, C. Huynh, Y. Kozono, C. Lee, B. Lutterbach, M. Metzstein, Q. Qiu, R. Shownkeen, R. Staden, J. Thierry-Mieg, K. Thomas, R. Wilson, and R. Waterston, in preparation.
9. Hawkins, T.L. and J.E. Sulston. 1990. Automated fluorescent primer walking. *Techniques* **2:** 307–310.
10. Green, E.D., R.M. Mohr, J.R. Idol, M. Jones, J.M. Buckingham, L.L. Deaven, R.K. Moyzis, and M.V. Olson. 1991. Systematic generation of sequence-tagged sites (STSs) for physical mapping of human chromosomes: Application to the mapping of human chromosome 7 using yeast artificial chromosomes. *Genomics* (in press).

Received August 1, 1991; accepted September 17, 1991.