Creation of a minimal tiling path of genomic clones for *Drosophila*: provision of a common resource

Volker Hollich[1], Eric Johnson[2], Eileen E. Furlong[3], Boris Beckmann[1], Joseph Carlson[4], Susan E. Celniker[4], and Jörg D. Hoheisel[1]

[1]Deutsches Krebsforschungszentrum, Heidelberg, Germany, [2]University of Oregon, Eugene, OR, USA, [3]EMBL, Heidelberg, Germany, and [4]Lawrence Berkeley National Laboratory, Berkeley, CA, USA

BioTechniques 37:282-284 (August 2004)

On the basis of shotgun subclone libraries used in the sequencing of the Drosophila melanogaster *genome, a minimal tiling path of subclones across much of the genome was determined. About 320,000 shotgun clones for chromosomes X(12–20), 2R, 2L, 3R, and 4 were available from the Berkeley Drosophila Genome Project. The clone inserts have an average length of 3.4 kb and are amenable to standard PCR amplification. The resulting tiling path covers 86.2% of chromosome X(12–20), 86.2% of chromosomal arm 2R, 79.0% of 2L, 89.6% of 3R, and 80.5% of chromosome 4. In total, the 25,135 clones represent 76.7 Mb—equivalent to about 67% of the genome —and would be suitable for producing a microarray on a single slide.*

INTRODUCTION

Representing the entire genome of an organism on DNA microarrays rather than the coding regions only is prerequisite to various functional analyses, such as chromatin immunoprecipitation experiments (1). But even for transcriptional profiling analyses, it could be advantageous, since a comprehensive coverage would by definition represent a complete and normalized gene repertoire irrespective of the status of sequence annotation. In order to produce a genomic tiling path, typically, a large set of PCR primers is designed on the basis of the genome sequence. A recent publication (2) reports on experiments performed on a relatively small set of such fragments that represent in total about 3 Mb of the *Drosophila* chromosomes 2 and 3. However, this approach is rather time-consuming and expensive. For coverage of the entire 115-Mb *Drosophila* sequence with 3-kb non-overlapping fragments, more than 76,000 primer molecules would be needed. Alternatively, the very DNA fragments on which the sequencing process was performed could be utilized to such an end. Since usually shotgun clones form the basis of large-scale sequencing projects, all fragments could be readily amplified with a single primer pair, thus creating enormous savings in time and expense. Slightly disadvantageous is the fact that the fragments cannot be placed end-to-end, but would overlap in part. Thus, slightly more fragments would be needed to cover a genome. However, a certain degree of redundancy in coverage may prove to be beneficial for analytical purposes. Adopting the latter strategy, we set out to cover the genome of *Drosophila melanogaster* by selecting a minimal tiling path across the entire genome from the bacterial artificial chromosome (BAC)-based subclone libraries used in the sequencing project (3).

MATERIALS AND METHODS

Clone Selection

Based on the sequencing data, a minimal tiling path was calculated for each subclone contig. This was accomplished by construction of a directed acyclic graph for every contig. Within this graph, each clone is represented by a vertex, and the set of vertices within the contig is called V. An edge between two vertices is introduced whenever two clones overlap. The set of edges is defined by $E = [(v_1,v_2) \mid v_2$ overlaps v_1, start(v_1) <= start(v_2)]. Each edge is denoted by a distance. We followed two different approaches, allowing us to minimize either the total number of clones or the overlap between clones. For the former, a constant number is assigned to each edge as distance; for the latter, the degree of overall overlap is taken as measure. Seeking the shortest path between the first and the last clone will deliver the desired minimal tiling path. Several solutions for the shortest path problem exist. We chose the Dijkstra algorithm (4). For the actual calculation, we used C++ with the extension of the Graph Template Library (GTL) by Forster and colleagues (http://infosun.fmi.uni-passau.de/GTL/), which includes an implementation of the shortest path algorithm.

PCR Amplification

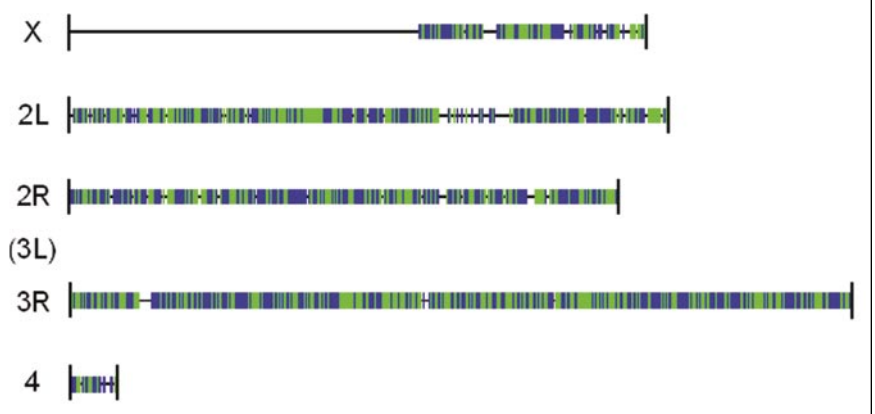
PCR amplification of the clone inserts was done in 96-well plates. Each 50-µL reaction contained 0.2 µM each of T7 and PM001 primers, 50 µM each of the four deoxynucleotide triphosphates, and 0.02 U *Taq* DNA polymerase (Qiagen, Hilden, Germany) in buffer provided by the enzyme supplier (1.5 mM $MgCl_2$). Plasmid DNA was added by transferring with a 96-pin tool a small amount of bacterial suspension from the library growth plates. The reactions were incubated at 94°C for 2 min, followed by 10 cycles of denaturation at 94°C for 10 s, annealing at 60°C for 30 s, and elongation at 68°C for 3 min. In another 20 cycles, the elongation time was prolonged by 10 s per cycle. Finally a final 7-min elongation step was performed at 68°C. Run on agarose gels, the average insert size was found to be 3.4 kb with a standard deviation of 0.43 kb.

RESULTS AND DISCUSSION

Of the libraries used in the sequencing project, about 320,000 shotgun clones for chromosomes X(12–20), 2R, 2L, 3R, and 4 were still available at Lawrence Berkeley National Laboratory (Berkeley, CA, USA). The clones used for sequencing in the other centers involved in the sequenc-

ing project—sublibraries covering regions 1–11 of chromosome X and all of the left arm of chromosome 3 as well as a global shotgun library—had been destroyed prior to the start of this initiative. To construct the tiling path, we initially determined the sequence positions of the subclones within the regions that are defined by 638 BAC clones (5). This included not only subclones, which had been produced from the respective BAC, but also subclones derived from P1 clones generated during an earlier phase of the sequencing project. The *D. melanogaster* chromosome arms of euchromatic sequence Release 3 (6) had been constructed by joining the individual sequences that represent the BAC clone inserts. As a result, the location of each BAC within an arm is known precisely. As a control, we compared the distance of the BAC end sequences within the genomic sequence and the actual length of each BAC insert used in our analysis. On the template of overlapping BAC sequences, the position of the shotgun clones was extracted from the Phrap sequence assembly, thus defining the start and end of each subclone insert. Subsequently, overlapping subclones were combined into contigs. Because of both unfinished BACs and missing shotgun clones, however, 2641 gaps remained in addition to the absent X(1–11) and 3L areas (Figure 1). These gaps could not be filled with 2-kb clones from the whole genome shotgun approach, since these clones were not available either.

Since the tiling path is based on randomly produced fragments, there is bound to be some overlap between them. However, as known from earlier analyses (e.g., References 2 and 7), this is rather an advantage (e.g., increasing resolution and providing some degree of redundancy). In the selection process of a minimal path, minimizing the degree of overlap between clones on 2L gave rise to 1.1% more clones, whereas aiming at a minimal total of clones resulted in 39.2% more overlap. This is due to the variation in clone lengths. Thus, the clone with the least overlap might be shorter than another clone, which spans further. Analyses on the other arms led to similar results. As the overlap-optimized path has only a small percentage of additional clones, we decided to base our minimal tiling path on this selection process, resulting in a set of 25,135 clones. In Figure 1, the coverage of the chromo-



	2L	2R	3R	4	X (12-20)
Size	22,217,931 bp	20,329,590 bp	27,892,406 bp	1,237,864 bp	8,921,907 bp
Minimal tiling path	17,558,095 bp	17,523,944 bp	24,980,170 bp	997,083 bp	7,691,306 bp
Missing sequence	4,659,836 bp	2,805,646 bp	2,912,236 bp	240,781 bp	1,230,601 bp
Largest single gap	576,596 bp	264,401 bp	548,252 bp	240,739 bp	178,701 bp
Coverage	79.0%	86.2%	89.6%	80.5%	86.2%
No. of clones	6,319	6,409	9,286	363	2,758

Figure 1. Representation of the tiling path's genome coverage. Horizontal lines indicate the chromosomes of the 115-Mb *Drosophila* genome. The genomic regions that are covered by the minimal tiling path of 25,135 shotgun clones are represented as blue and green colored areas. Interruption of the coloring depicts large gaps. Any change in color from blue to green or vice versa indicates the existence of a gap that is too small to be visible. The table presents the relevant numbers.

somal arms is shown, indicated by the green and blue colored areas. Small gaps, which are not visible at this scale, are marked by changes in the coloring. Detailed representations providing the exact clone positions, clone names, the degree of overlap, positions of genes, and the length of the gaps can be accessed online (http://www.dkfz.de/funct_genome/Drosophila/ArmAnzeigeApplet.html and http://flycompute.uoregon.edu/cgi-bin/subclone_ranger.pl). In addition, tab-delimited files listing the name and exact position of each clone in the tiling path are available from http://www.dkfz.de/funct_genome/Drosophila/.

We performed PCR amplifications of clone inserts at two sites independently. Even without addition of a proofreading polymerase, the success rate of amplification was 93%. This is similar to results obtained with primers that produced from genomic DNA template exon-specific PCR products of 300 to 500 bp in length (8) and in agreement with experiences made on shotgun clone libraries, which were used to produce genomic microarrays for other organisms, such as *Pseudomonas putida* (7). To ensure proper tracking of genomic clones during rearraying, each PCR product was compared to the known size of the genomic clone insert. In addition, the identity of several randomly picked clones was confirmed by resequencing.

The 25,135 clones that make up the minimal tiling path covering the *Drosophila* genome, with the exception of the 3L and X(1–11) regions, will be a valuable tool for functional analyses in *Drosophila* and especially so in combination with other genome-wide tools such as a comprehensive RNA interference (RNAi) library (9). The clone inserts can be produced by standard PCR amplification and would be suitable for arraying onto a single microscopic slide. Overall coverage is 85.3%. Taking into account also X(1–11) and 3L, this value drops to 67%. Thus, two thirds of the entire genome is available in short and defined fragments made from the very DNA used for sequencing.

As for *Drosophila*, other ongoing genome projects could create an added value by extending the use of their shotgun clone libraries beyond sequencing to the selection and distribution of a minimal tiling path. Thereby, the entire sequence would be available not only as an electronic file but also be accessible physically. Still the majority of clones used in sequencing can be discarded eventually, with freezer space frequently being a limiting factor. As an extention from studies of individual genomes, the approach of using sequencing shotgun clones for subsequent analyses would also allow the global transcriptional analysis of metagenomes (e.g., picking only fragments of unique sequence) (10). Since the majority of sequence would be from microbial organisms, it would mostly consist of coding sequence.

The library of the *D. melanogaster* minimal tiling path reported here will be distributed in microtiter plates to interested parties by the laboratories involved in this study.

ACKNOWLEDGMENTS

We thank Steve Russell, who was instrumental in getting the project started and Karl-Heinz Glatting for discussions on how to calculate the tiling path. We are also grateful to Jessica Werdermann, Pascale M. Voelker, Jason Carriere, and Ken Wan for technical help. This work was funded in part by the German Human Genome Project (DHGP) program of the German Federal Ministry of Education and Science (BMBF).

COMPETING INTERESTS STATEMENT

The authors declare that they have no competing interests.

REFERENCES

1. **Hecht, A. and M. Grunstein.** 1999. Mapping DNA interaction sites of chromosomal proteins using immunoprecipitation and polymerase chain reaction. Methods Enzymol. *304*:399-414.
2. **Sun, L.V., L. Chen, F. Greil, N. Negre, T.-R. Li, G. Cavalli, H. Zhao, B. van Steensel, and K.P. White.** 2003. Protein-DNA interaction mapping using genomic tiling path microarrays in *Drosophila*. Proc. Natl. Acad. Sci. USA *100*:9428-9433.
3. **Adams, M.D., S.E. Celniker, R.A. Holt, C.A. Evans, J.D. Gocayne, P.G. Amanatides, S.E. Scherer, P.W. Li, et al.** 2000. The genome sequence of *Drosophila melanogaster*. Science *287*:2185-2195.
4. **Dijkstra, E.W.** 1959. A note on two problems in connection with graphs. Numer. Math. *1*:269-271.
5. **Hoskins, R.A., C.R. Nelson, B.P. Berman, T.R. Laverty, R.A. George, L. Ciesiolka, M. Naeemuddin, A.D. Arenson, et al.** 2000. A BAC-based physical map of the major autosomes of *Drosophila melanogaster*. Science *287*:2271-2274.
6. **Celniker, S.E., D.A. Wheeler, B. Kronmiller, J.W. Carlson, A. Halpern, S. Patel, M. Adams, M. Champe, et al.** 2002. Finishing a whole-genome shotgun: release 3 of the *Drosophila melanogaster* euchromatic genome sequence. Genome Biol. *3*:research0079.1-research0079.14.
7. **Stjepandic, D., C. Weinel, H. Hilbert, H.L. Koo, F. Diehl, K.E. Nelson, B. Tümmler, and J.D. Hoheisel.** 2002. The genome structure of *Pseudomonas putida*; high-resolution mapping and microarray analysis. Environ. Microbiol. *4*:819-823.
8. **Hild, M., B. Beckmann, S.A. Haas, B. Koch, V. Solovjev, C. Busold, K. Fellenberg, M. Boutros, et al.** 2003. An integrated gene annotation and transcriptional profiling approach towards the full gene content of the *Drosophila* genome. Genome Biol. *5*:R3.
9. **Boutros, M., A.A. Kiger, S. Armknecht, K. Kerr, M. Hild, B. Koch, S.A. Haas, B. Beckmann, et al.** 2004. Genome-wide RNAi analysis of growth and viability in *Drosophila* cells. Science *303*:382-385.
10. **Venter, J.C., K. Remington, J.F. Heidelberg, A.L. Halpern, D. Rusch, J.A. Eisen, D. Wu, J. Paulsen, et al.** 2004. Environmental genome shotgun sequencing of the Saragosso Sea. Science *304*:66-74.

Received 23 March 2004; accepted 7 May 2004.

Address correspondence to Jörg D. Hoheisel, Functional Genome Analysis, Deutsches Krebsforschungszentrum, Im Neuenheimer Feld 580, 69120 Heidelberg, Germany. e-mail: j.hoheisel@dkfz.de