D–0.1 M KCl, Tat-SF/pp140 was eluted with increasing salt concentrations and was detected mostly in 0.2 to 0.4 M KCl fractions. These fractions were pooled, dialyzed against buffer D–0.1 M KCl, and loaded onto a glutathione Sepharose (Pharmacia) column containing GST-Tat fusion proteins. After the column was washed with buffer D–0.4 M KCl, Tat-SF/pp140 was eluted from the column with buffer D containing 1.4 M KCl. The estimated overall purification after these steps was ~3000-fold. In the experiment shown in Fig. 3, the 0.2 to 0.4 M KCl heparin Sepharose fraction containing Tat-SF activity was subjected to fractionation through an Affi-Gel 10 matrix column (Bio-Rad) containing immobilized Tat. Tat-SF activity was eluted from the column with increasing salt concentrations. The 0.6 M KCl fraction was analyzed as described in Fig. 3.

10. T. O'Brien, S. Hardin, A. Greenleaf, J. T. Lis, *Nature* **370**, 75 (1994); M. E. Dahmus, *Biochim. Biophys. Acta.* **1261**, 171 (1995).
11. A. P. Rice and F. Carlotti, *J. Virol.* **64**, 1864 (1990).
12. The Tat-SF/pp140 fraction eluted from the GST-Tat column was subjected to SDS–polyacrylamide gel electrophoresis (PAGE), and the pp140 polypeptide was blotted onto a nitrocellulose membrane. Approximately 15 μg of pp140 were recovered from the membrane and subjected to digestion with lys-C. Six major peptides were obtained and microsequenced. One of the peptides (KMNAQETATGMAFEEPIDE) was contained in the sequence of EST60354 in the Washington University–Merck EST database. An Xho I–Eco RI fragment corresponding to the COOH-terminus of the Tat-SF1 gene and its 3′ untranslated region was labeled and used as a probe to screen a λZipLox (Gibco BRL) cDNA library prepared from human HL60 cells. Complementary DNAs were recovered from seven independent plaques in the autonomously replicating plasmid pZL1 as instructed by the manufacturer (Gibco BRL). The largest cDNA clone containing the full-length Tat-SF1 gene was named pZL-Tat-SF1-4b and was sequenced by dideoxy-DNA sequencing with T7 DNA polymerase.
13. D. R. Marshak and D. Carroll, *Methods Enzymol.* **200**, 134 (1991).
14. D. J. Kenan, C. C. Query, J. D. Keene, *Trends Biochem. Sci.* **16**, 214 (1991).
15. O. Delattre *et al.*, *Nature* **359**, 162 (1992); P. H. Sorensen *et al.*, *Nature Genet.* **6**, 146 (1994).
16. A. Crozat, P. Aman, N. Mandahl, D. Ron, *Nature* **363**, 640 (1993); T. H. Rabbitts, A. Forster, R. Larson, P. Nathan, *Nature Genet.* **4**, 175 (1993).
17. M. Ladanyi, *Diagn. Mol. Pathol.* **4**, 162 (1995); T. H. Rabbitts, *Nature* **372**, 143 (1994).
18. S. E. Harper, Y. Qiu, P. A. Sharp, *Proc. Natl. Acad. Sci. U.S.A.* **93**, 8536 (1996).
19. J. W. Lillie and M. R. Green, *Nature* **338**, 39 (1989).
20. H. Kato *et al.*, *Genes Dev.* **6**, 655 (1992); R. A. Marciniak and P. A. Sharp, *EMBO J.* **10**, 4189 (1991).
21. M. G. Izban and D. S. Luse, *Genes Dev.* **6**, 1342 (1992); D. Wang and D. K. Hawley, *Proc. Natl. Acad. Sci. U.S.A.* **90**, 843 (1993).
22. E. Bengal, O. Flores, A. Krauskopf, D. Reinberg, Y. Aloni, *Mol. Cell. Biol.* **11**, 1195 (1991); J. Greenblatt, J. R. Nodwell, S. W. Mason, *Nature* **364**, 401 (1993).
23. C. H. Herrmann and A. P. Rice, *J. Virol.* **69**, 1612 (1995).
24. N. A. McMillan *et al.*, *Virology* **213**, 413 (1995).
25. W. A. May *et al.*, *Mol. Cell. Biol.* **13**, 7393 (1993); H. Zinszner, R. Albalat, D. Ron, *Genes Dev.* **8**, 2513 (1994); D. D. Prasad, M. Ouchida, L. Lee, V. N. Rao, E. S. Reddy, *Oncogene* **9**, 3717 (1994).
26. P. J. Mitchell and R. Tjian, *Science* **245**, 371 (1989).
27. S. F. Altschul, W. Gish, W. Miller, E. W. Myers, D. J. Lipman, *J. Mol. Biol.* **215**, 403 (1990).
28. M. A. Truett *et al.*, *DNA* **4**, 333 (1985).
29. H. E. Gendelman *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **83**, 9759 (1986).
30. L. S. Tiley, P. H. Brown, B. R. Cullen, *Virology* **178**, 560 (1990).
31. J. R. Neumann, C. A. Morency, K. O. Russian, *BioTechniques* **5**, 444 (1987).
32. We are grateful to B. Pepinsky and Biogen for providing pure HIV Tat protein and Tat mutant TatΔC; to J. Borrow [Massachusetts Institute of Technology (MIT) Center for Cancer Research] for human cDNA libraries; and to R. Cook (MIT Biopolymers Laboratory) for peptide sequencing. We thank K. Luo, J. Borrow, and H. Kawasaki for valuable advice and discussions; and B. Blencowe, K. Cepek, G. Jones, K. Luo, and C. Query for helpful comments on the manuscript. We also thank M. Siafaca for secretarial support. Supported by grants from the National Institutes of Health (GM34277 and AI32486) to P.A.S., and partially supported by a National Cancer Institute Center core grant (CA14051). Q.Z. was supported by a postdoctoral fellowship of The Jane Coffin Childs Memorial Fund for Medical Research.

19 June 1996; accepted 23 August 1996

Accessing Genetic Information with High-Density DNA Arrays

Mark Chee, Robert Yang, Earl Hubbell, Anthony Berno, Xiaohua C. Huang, David Stern, Jim Winkler, David J. Lockhart, Macdonald S. Morris, Stephen P. A. Fodor

Rapid access to genetic information is central to the revolution taking place in molecular genetics. The simultaneous analysis of the entire human mitochondrial genome is described here. DNA arrays containing up to 135,000 probes complementary to the 16.6-kilobase human mitochondrial genome were generated by light-directed chemical synthesis. A two-color labeling scheme was developed that allows simultaneous comparison of a polymorphic target to a reference DNA or RNA. Complete hybridization patterns were revealed in a matter of minutes. Sequence polymorphisms were detected with single-base resolution and unprecedented efficiency. The methods described are generic and can be used to address a variety of questions in molecular genetics including gene expression, genetic linkage, and genetic variability.

A central theme in modern genetics is the relation between genetic variability and phenotype. To understand genetic variation and its consequences on biological function, an enormous effort in comparative sequence analysis will need to be carried out. Conventional nucleic acid sequencing technologies make use of analytical separation techniques to resolve sequence at the single nucleotide level (*1*, *2*). However, the effort required increases linearly with the amount of sequence. In contrast, biological systems read, store, and modify genetic information by molecular recognition (*3*). Because each DNA strand carries with it the capacity to recognize a uniquely complementary sequence through base pairing, the process of recognition, or hybridization, is highly parallel, as every nucleotide in a large sequence can in principle be queried at the same time. Thus, hybridization can be used to efficiently analyze large amounts of nucleotide sequence. In one proposal, sequences are analyzed by hybridization to a set of oligonucleotides representing all possible subsequences (*4*). A second approach, used here, is hybridization to an array of oligonucleotide probes designed to match specific sequences. In this way the most informative subset of probes is used. Implementation of these concepts relies on recently developed combinatorial technologies to generate any ordered array of a large number of oligonucleotide probes (*5*).

Affymetrix, 3380 Central Expressway, Santa Clara, CA 95051, USA.

The fundamentals of light-directed oligonucleotide array synthesis have been described (*5*, *6*). Any probe can be synthesized at any discrete, specified location in the array, and any set of probes composed of the four nucleotides can be synthesized in a maximum of $4N$ cycles, where N is the length of the longest probe in the array. For example, the entire set of $\sim 10^{12}$ 20-nucleotide oligomer probes, or any desired subset, can be synthesized in only 80 coupling cycles. The number of different probes that can be synthesized is limited only by the physical size of the array and the achievable lithographic resolution (*7*).

An array consisting of oligonucleotides complementary to subsequences of a target sequence can be used to determine the identity of a target sequence, measure its amount, and detect differences between the target and a reference sequence. Many different arrays can be designed for these purposes. One such design, termed a 4L tiled array, is depicted in Fig. 1A. In each set of four probes, the perfect complement will hybridize more strongly than mismatched probes. By this approach, a nucleic acid target of length L can be scanned for mutations with a tiled array containing $4L$ probes. For example, to query the 16,569 base pairs (bp) of human mitochondrial DNA (mtDNA), only 66,276 probes of the possible $\sim 10^9$ 15-nucleotide oligomers need to be used.

The use of a tiled array of probes to read a target sequence is illustrated in Fig. 1C. A tiled array of 15-nucleotide oligomers varied

at position 7 from the 3′ end ($P^{15,7}$) was designed and synthesized for mt1, a cloned sequence containing 1311 bp spanning the control region of mtDNA (*8*–*11*). The upper panel of Fig. 1C shows a portion of the fluorescence image of an array hybridized with fluorescein-labeled mt1 RNA (*12*). The base sequence can be read by comparing the intensities of the four probes within each column. For example, the column for position 16,493 consists of the four probes, 3′-TGACATAGGCTGTAG, 3′-TGACATCGGCTGTAG, 3′-TGACATGGGCTGTAG, and 3′-TGACATTGGCTGTAG. The probe with the strongest signal is the probe with the A substitution (A, 49 counts; C, 8 counts, G, 15 counts, and T, 8 counts, where the background is 2 counts), identifying the base at position 16,493 as U in the RNA transcript. Continuing the process, the sequence at each position can be read directly from the hybridization intensities.

The effect on the array hybridization pattern caused by a single base change in the target is illustrated in Fig. 1B, and the detection of a single-base polymorphism is shown in the lower panel of Fig. 1C. The target was mt2, which differs from mt1 in this region by a T-to-C transition at position 16,493. Accordingly, the probe with the G substitution (third row) displays the strongest signal. Because the tiled array was designed to complement mt1, the hybridization intensities of neighboring probes that overlap position 16,493 are also affected by the change in target sequence. The hybridization signals of 15 probe sets of the 15-nucleotide oligomer tiled array are perturbed by a single base change in the target sequence. In the $P^{15,7}$ array, each probe querying the eight positions to the left and six positions to the right of the polymorphism contain at least one mismatch to the target. The result is a characteristic loss of signal or a "footprint" for the probes flanking a mutation position. Of the four probes querying each position, the loss of signal is greatest for the one designed to match mt1. We denote the subset of probes with zero mismatches to the reference sequence as P^0.

A comparison of P^0 hybridization signals from a target to those from a reference is ideally obtained by hybridizing both samples to the same array. We therefore developed a two-color labeling and detection scheme in which the reference is labeled with phycoerythrin (red), and the target with fluorescein (green) (*13*). By processing the reference and target together, experimental variability during the fragmentation, hybridization, washing, and detection steps is minimized or eliminated. In addition, during cohybridization of the reference and target, competition for binding sites results in a slight improvement in mismatch discrimination. Array hybridization is highly reproducible, and comparative analysis of data obtained from separate but identically synthesized arrays is also effective.

The two-color approach was tested by analyzing a 2.5-kb region of mtDNA that spans the $tRNA^{Glu}$, cytochrome b, $tRNA^{Thr}$, $tRNA^{Pro}$, control region, and $tRNA^{Phe}$ DNA sequences (*14*). A $P^{20,9}$ array (20-nucleotide oligomer probes varied at position 9 from the 3′ end) was designed to match the mt1 target (that is, P^0 sequence = mt1). The mt1 reference (red) and a polymorphic target sample (green) were pooled and hybridized simultaneously to the array. Differences between the target and reference sequences were identified by comparing the scaled red and green P^0 hybridization intensities (*15*). The marked decrease in target hybridization intensity, over a span of ~20 nucleotides, is shown for a single-base polymorphism at position 16,223 (Fig. 2A). The footprint is enlarged when two polymorphisms occur in close proximity (within ~20 nucleotides) (Fig. 2B). When polymorphisms are clustered, the size of the footprint depends on the number of polymorphisms and their separation (Fig. 2C).

To read polymorphisms accurately, we developed an algorithm that addresses the issue of multiple mismatches. The algorithm performs base identification but also flags regions of ambiguity caused by multiple mismatches. These regions are easily identified by the presence of a large footprint (Fig. 2, B and C) or by two or more bases identified as differing from P^0 within the span of a single probe. Discrepancies between base identifications and footprint patterns are also flagged for further analysis (for example, a P^0 footprint in which no polymorphism is identified; such a pattern is typical of a deletion). Thus, base identifications are valid only for unflagged regions. In flagged regions, the presence of sequence differences is detected, but no attempt is made to identify the sequence without further analysis.

Sequence analysis was carried out on the 2.5-kb target from 12 samples. A total of 30,582 bp containing 180 substitutions relative to mt1 were analyzed. Ninety-eight per-

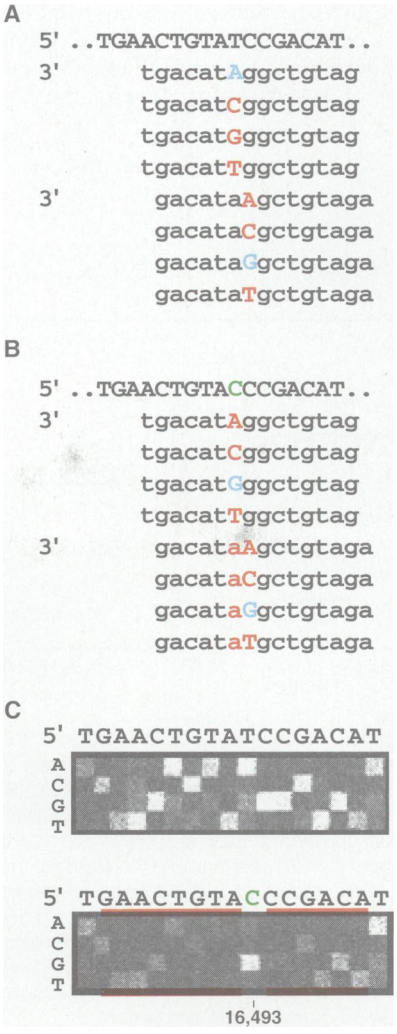

Fig. 1. (**A**) Design of a 4L tiled array. Each position in the target sequence (uppercase letters) is queried by a set of four probes on the chip (lowercase letters), identical except at a single position, termed the substitution position, which is either A, C, G, or T (blue indicates complementarity, red a mismatch). Two sets of probes are shown, querying adjacent positions in the target. (**B**) Effect of a change in the target sequence. The probes are the same as in (A), but the target now contains a single-base substitution (base C, shown in green). The probe set querying the changed base still has a perfect match (the G probe). However, probes in adjacent sets that overlap the altered target position now have either one or two mismatches (red) instead of zero or one, because they were designed to match the target shown in (A). (**C**) Hybridization to a 4L tiled array and detection of a base change in the target. The array shown was designed to the mt1 sequence. (Top) hybridization to mt1. The substitution used in each row of probes is indicated to the left of the image. The target sequence can be read 5′ to 3′ from left to right as the complement of the substitution base with the brightest signal. With hybridization to mt2 (bottom), which differs from mt1 in this region by a T→C transition, the G probe at position 16,493 is now a perfect match, with the other three probes having single-base mismatches (A 5, C 3, G 37, T 4 counts). However, at flanking positions, the probes have either single- or double-base mismatches, because the mt2 transition now occurs away from the query position.

cent of the sequence was unambiguously assigned by a Bayesian base identification algorithm (*16*). Of this 98%, which contained both wild-type sequence and a high proportion of single-base footprints such as the example shown in Fig. 2A, 29,878 out of 29,879 bp were identified correctly (*17*). The remaining 2% of the sequence, which contained the multiple substitution footprints (such as those shown in Fig. 2, B and C), was flagged for further analysis. Of the 649 bp composing this 2%, 643 bp were located in or immediately adjacent to footprints (*18*). In all, 179 out of the 180 polymorphisms were unambiguously detected, 126 out of 127 were identified correctly in the unflagged regions, and 53 polymorphisms occuring in the flagged regions were detected as footprints. There were no unflagged false-positive base identifications, and only one false-positive footprint. These figures can be considered to be "worst case" estimates for the type of array and target used. The P^0 sequence represents a Caucasian haplotype, and our sample set included eight African samples having a large number of clustered differences to P^0. Furthermore, the variation in the hypervariable part of the control region is much higher than for the rest of the mitochondrial genome and for nuclear genes in general (Fig. 2 shows comparisons to African samples in this region).

The determination of a complete human mitochondrial DNA sequence more than 15 years ago has had a tremendous influence on studies of human origins and evolution and the role of mutations in degenerative diseases (*8, 10, 19*). Because of the cost and difficulty of conventional sequence analysis, most subsequent sequencing studies have focused only on two small hypervariable regions totaling ~600 bp (9). However, access to the entire genome is required for a full understanding of the governing genetics. We therefore designed a $P^{25,13}$ tiling array for the mitochondrial genome. The array contains a total of 136,528 synthesis cells, each ~35 μm by 35 μm in size (Fig. 3). In addition to a 4L tiling across the genome, the array contains a set of probes representing a single-base deletion at every position across the genome and sets of probes designed to match a range of specific mtDNA haplotypes. Using long-range polymerase chain reaction, we amplified the 16.6-kb mtDNA directly from genomic DNA samples (*20*). Labeled RNA targets were prepared by in vitro transcription and hybridized to the array. Genomic hybridization patterns were imaged in less than 10 min by a high-resolution confocal scanner (*21*).

The hybridization pattern of a 16.6-kb target to the mitochondrial genome chip is shown in Fig. 3. Although there are some regions of low intensity, most of the 25-nucleotide oligomer array hybridized efficiently: Simply by identifying the highest intensity in each column of four substitution probes, 99.0% of the mt3 sequence could be read correctly (P^0 sequence = mt3). The array was used to successfully detect three disease-causing mutations in a mtDNA sample from a patient with Leber's hereditary optic neuropathy (*22, 23*) (Fig. 3C). In addition, we detected a total of seven errors and new polymorphisms from previously unsequenced regions.

We then hybridized 10 genomes from African individuals to the array and unambiguously identified 505 polymorphisms. These were polymorphisms that could be clearly read and for which a confirmatory footprint was detected automatically. For the 10 samples, the 2.5-kb cytochrome b and control region sequences were known (*17*). No false positives were detected in the ~25 kb of sequence checked in this way. Additional clustered polymorphisms were detected by the presence of footprints but not read directly. A detailed analysis of the polymorphisms in these genomes, and others, will be presented elsewhere.

The throughput of a conventional gel-based sequencer, with an average read length of 400 nucleotides and 48 lanes that is run twice a day, might be two mitochondrial genomes a day at best. In contrast, the throughput of the nonoptimized system we describe is five chips per hour. Thus, 50 genomes can be read by hybridization in the time it takes to read two genomes conventionally. Furthermore, there are significant reductions in sample preparation requirements because the entire genome is labeled in a single reaction, so the cost is similar to that for a single sequencing reaction. Also, sequence reading at the level of data analysis is automated: The sequences can be read in a matter of minutes. No analytical separations or gel preparation is needed, which contributes to the speed of the experiment. Although the inability to read all possible sequences is a weakness of the 4L tiled array, it is not a major limitation, because in practice the small number of ambiguities can be checked by targeted conventional sequencing. In particular, highly repetitive sequences, such as long runs of a single base, are presently best analyzed with conventional technology. Finally, a clear advantage to the approach we describe is that it is highly scalable. The cost, effort, and time required to analyze the entire 16.6-kb mtDNA in a single experiment is virtually identical to that required to read 2.5 kb. This provides a clear path to further orders-of-magnitude improvements in efficiency.

High-density oligonucleotide arrays

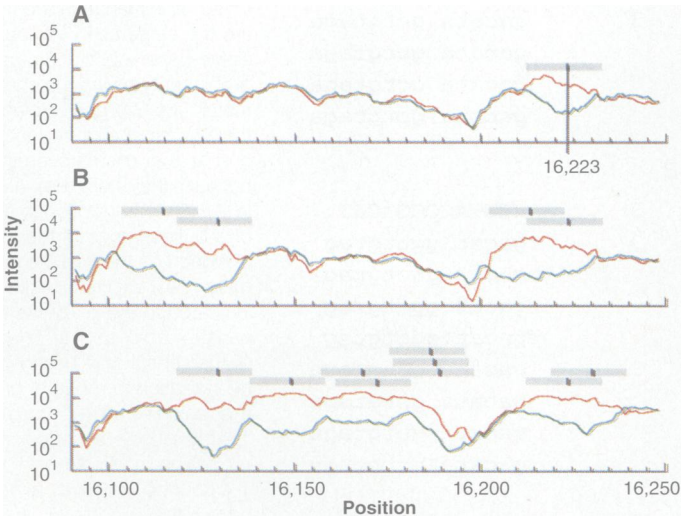

Fig. 2. Detection of base differences in a 2.5-kb region by comparison of scaled P^0 hybridization intensity patterns between a sample (green) and a reference (red) sequence. (**A**) Comparison of sequence ief007 to mt1. In the region shown, there is a single-base difference between the two sequences, located at position 16,223 (C in mt1, T in ief007). This results in a "footprint" spanning ~20 positions, 11 to the left and 8 to the right of position 16,223, in which the ief007 P^0 intensities are decreased by a factor of more than 10 on average relative to the mt1 intensities. The predicted footprint location is indicated by the gray bar, and the location of the polymorphism is shown by a vertical black line within the bar. The size of a footprint changes with probe length, and its relative position with substitution position (not shown). (**B**) Comparison of sequence ha001 to mt1. The ha001 target has four polymorphisms relative to mt1. The P^0 intensity pattern clearly shows two regions of difference between the targets. Each region contains two or more differences, because in both cases the footprints are longer than 20 positions and therefore are too extensive to be explained by a single-base difference. The effect of competition can be seen by comparing the mt1 intensities in the ief007 and ha001 experiments: The relative intensities of mt1 are greater in (B) where ha001 contains P^0 mismatches but ief007 does not. (**C**) The ha004 sequence has multiple differences to mt1, resulting in a complex pattern extending over most of the region shown. Thus, differences are clearly detected. Because hybridization intensities are extremely sequence-dependent, each of the mitochondrial sequences can also be identified simply by its hybridization pattern.

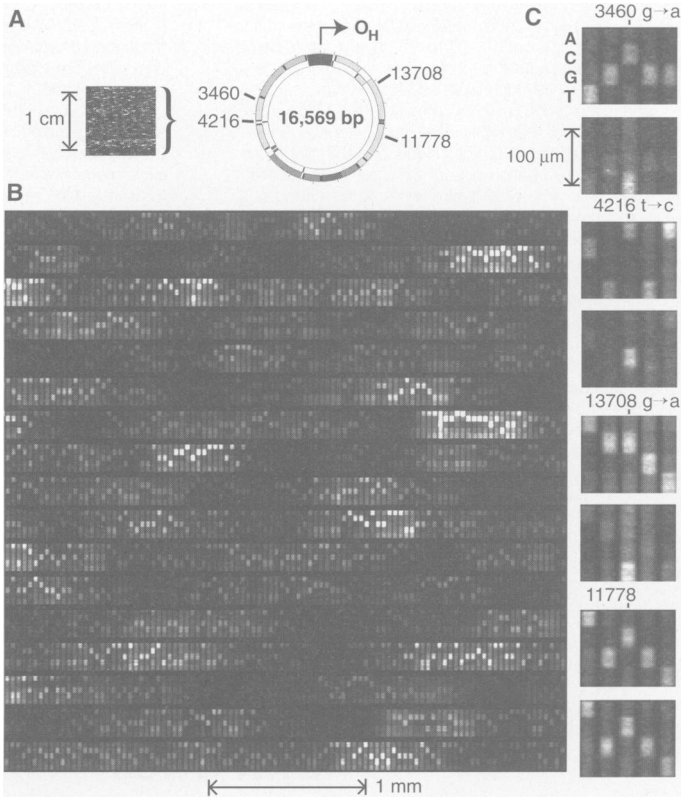

Fig. 3. Human mitochondrial genome on a chip. (**A**) An image of the array hybridized to 16.6 kb of mitochondrial target RNA (L strand). The 16,569-bp map of the genome is shown, and the H strand origin of replication (O_H), located in the control region, is indicated. (**B**) A portion of the hybridization pattern magnified. In each column there are five probes: A, C, G, T, and Δ, from top to bottom. The Δ probe has a single-base deletion instead of a substitution and hence is 24 instead of 25 bases in length. The scale is indicated by the bar beneath the image. Although there is considerable sequence-dependent intensity variation, most of the array can be read directly. The image was collected at a resolution of ~100 pixels per probe cell. (**C**) The ability of the array to detect and read single-base differences in a 16.6-kb sample is illustrated. Two different target sequences were hybridized in parallel to different chips. The hybridization patterns are compared for four different positions in the sequence. Only the $P^{25,13}$ probes are shown. The top panel of each pair shows the hybridization of the mt3 target, which matches the chip P^0 sequence at these positions. The lower panel shows the pattern generated by a sample from a patient with Leber's hereditary optic neuropathy (LHON). Three known pathogenic mutations, LHON3460, LHON4216, and LHON13708, are clearly detected. For comparison, the fourth panel in the set shows a region around position 11,778 that is identical in both samples.

provide the foundation for a powerful genetic analysis technology. The method can be used to characterize the spectrum of sequence variation in a population and can be applied to the analysis of many genes in parallel. In the case of human mtDNA, we simultaneously analyzed the control region, 13 protein coding genes, 22 tRNA genes, and 2 ribosomal RNA genes. The methods described here can be applied to other research areas in molecular genetics; for example, the ability to identify and sequence polymorphisms provides a basis for genetic mapping. The specificity of oligonucleotide hybridization and the scalability of the method suggests the possibility of a dedicated array that could be used to generate a high-resolution genetic map of an entire genome in a single experiment. Likewise, the concepts and techniques described here have been used to develop approaches for mRNA identification and the large-scale, parallel measurement of expression levels (*24*). Thus, the sequence of a gene, its spectrum of change in the population, its chromosomal location, and its dynamics of expression (all essential to a full understanding of function) can be determined with high-density probe arrays. The challenge now is to synthesize and read probe arrays at even higher density. For example, a 2 cm by 2 cm array, synthesized with probes occupying 1-μm synthesis sites in a 4L tiling, could query the entire coding content of the human genome, estimated at 100,000 genes.

REFERENCES AND NOTES

1. F. Sanger, S. Nicklen, A. R. Coulson, *Proc. Natl. Acad. Sci. U.S.A.* **74**, 5463 (1977).
2. A. M. Maxam and W. Gilbert, *ibid.*, p. 560.
3. J. D. Watson and F. H. C. Crick, *Nature* **171**, 737 (1953).
4. W. Bains and G. C. Smith, *J. Theor. Biol.* **135**, 303 (1988); Y. P. Lysov *et al.*, *Dokl. Akad. Nank. SSSR* **303**, 1508 (1988); R. Drmanac, I. Labat, I. Brukner, R. Crkvenjakov, *Genomics* **4**, 114 (1989); E. Southern, U. Maskos, R. Elder, *ibid.* **13**, 1008 (1992); see also R. B. Wallace *et al.*, *Nucleic Acids Res.* **6**, 3543 (1979).
5. S. P. A. Fodor *et al.*, *Science* **251**, 767 (1991).
6. A. C. Pease *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **91**, 5022 (1994).
7. In the present format, we can routinely achieve a density of 409,600 synthesis sites in a 1.28 cm by 1.28 cm array. Each 20 μm by 20 μm site contains ~4 × 10^6 functional copies of a specific probe, which corresponds to a mean distance of about 100 Å between probes (M. O. Trulson, D. Stern, R. P. Rava, unpublished results).
8. S. Anderson *et al.*, *Nature* **290**, 457 (1981).
9. The control region of mtDNA is characterized by high amounts of sequence polymorphism concentrated in two hypervariable regions [B. D. Greenberg, J. E. Newbold, A. Sugino, *Gene* **21**, 33 (1983); C. F. Aquardo and B. D. Greenberg, *Genetics* **103**, 287 (1983)].
10. R. L. Cann, W. M. Brown, A. C. Wilson, *Genetics* **106**, 479 (1984).
11. The mt1 and mt2 sequences were cloned from amplified genomic DNA extracted from hair roots [P. Gill, A. J. Jeffreys, D. J. Werrett, *Nature* **318**, 577 (1985); R. K. Saiki *et al.*, *Science* **239**, 487 (1988)]. The clones were sequenced conventionally (*1*). Cloning was performed only to provide a set of pure reference samples of known sequence. For templates for fluorescent labeling, DNA was reamplified from the clones with primers bearing bacteriophage T3 and T7 RNA polymerase promoter sequences (bold; mtDNA sequences uppercase): L15935-T3, 5′-ctcgg**aattaaccctcactaaagg**AAACCTTTTTCC-AAGGA and H667-T7, 5′-**taatacgactcactatagg**gagAGGCTAGGACCAAACCTATT.
12. Labeled RNAs from the two complementary mtDNA strands [designated L and H (*8*)] were transcribed in separate reactions from a promoter-tagged polymerase chain reaction (PCR) product. Each 10-μl reaction contained 1.5 mM each of the triphosphate nucleotides ATP, CTP, GTP, and UTP; 0.24 mM fluorescein-12–CTP (Du Pont); 0.24 mM fluorescein-12–UTP (Boehringer Mannheim); ~1 to 5 nM (1.5 μl) crude unpurified 1.3-kb PCR product; and T3 or T7 RNA polymerase (1 U/μl) (Promega) in a reaction buffer supplied with the enzyme. The reaction was carried out at 37°C for 1 to 2 hours. RNA was fragmented to an average size of <100 nucleotides by adjusting the solution to 30 mM $MgCl_2$, by the addition of 1 M $MgCl_2$, and heating at 94°C for 40 min. Fragmentation improved the uniformity and specificity of hybridization (M. Chee *et al.*, data not shown). The extent of fragmentation is dependent on the magnesium ion concentration [J. W. Huff, K. S. Sastry, M. P. Gordon, W. E. C. Wacker, *Biochemistry* **3**, 501 (1964); J. J. Butzow and G. L. Eichorn, *Biopolymers* **3**, 95 (1965)]. Good hybridization results have been obtained with both DNA and RNA targets prepared with a variety of labeling schemes, including incorporation of fluorescent and biotinylated deoxynucleoside triphosphates by DNA polymerases, incorporation of dye-labeled primers during PCR, ligation of labeled oligonucleotides to fragmented RNA, and direct labeling by photo–cross-linking a psoralen derivative of biotin directly to fragmented nucleic acids (L. Wodicka, personal communication).
13. For two-color detection experiments, the reference and unknown samples were labeled with biotin and fluorescein, respectively, in separate transcription reactions. Reactions were carried out as described (*12*) except that each contained 1.25 mM of ATP, CTP, GTP, and UTP and 0.5 mM fluorescein-12–UTP or 0.25 mM biotin-16–UTP (Boehringer Mannheim). The two reactions were mixed in the ratio 1:5 (v/v) biotin:fluorescein and fragmented (*12*). Targets were diluted to a final concentration of ~100 to 1000 pM in 3M TMACl [W. B. Melchior Jr. and P. H. von Hippel, *Proc. Natl. Acad. Sci. U.S.A.* **70**, 298 (1973)], 10 mM tris-HCl, pH 8.0, 1 mM EDTA, 0.005% Triton X-100, and 0.2 nM control oligonucleotide labeled at the 5′ end with fluorescein (5′-CTGAACGGTAG-CATCTTGAC). Samples were denatured at 95°C for 5 min, chilled on ice for 5 min, and equilibrated to 37°C. A volume of 180 μl of hybridization solution was then added to the flow cell [R. Lipshutz *et al.*, *Biotechniques* **19**, 442 (1995)] and the chip incubated at 37°C for 3 hours with rotation at 60 rpm. The chip was washed six times at room temperature with 6× SSPE (0.9 M NaCl, 60 mM NaH_2PO_4, 6 mM EDTA, pH 7.4), 0.005% Triton X-100. Phycoerythrin-conjugated streptavidin (2 μg/ml in 6× SSPE, 0.005% Triton X-100) was added and incubation continued at room temperature for 5 min. The chip was washed again

and scanned at a resolution of ~74 pixels per probe cell. Two scans were collected: a fluorescein scan was obtained with a 515- to 545-nm band-pass filter, and a phycoerythrin scan with a 560-nm long-pass filter. Signals were separated to remove spectral overlap and average counts per cell determined.

14. Each 2.5-kb target sequence was PCR-amplified directly from genomic DNA with the primer pair L14675-T3 (5′-aattaaccctcactaaagggATTCTCGCACGGACTACAAC) and H667-T7 (*11*).
15. To scale the sample to the reference intensities, we constructed a histogram of the base 10 logarithm of the intensity ratios for each pair of probes. The histogram had a mesh size of 0.01 and was smoothed by replacing the value at each point with the average number of counts over a five-point window centered at that point. The highest value in the histogram was located, and the resulting intensity ratio was taken to be the most probable calibration coefficient.
16. Base identification was accomplished with a Bayesian classification algorithm based on variable kernel density estimation. The likelihood of each identification associated with a set of hybridization intensity values was computed by comparing an unknown set of probes to a set of example cases for which the correct base identification was known. The resulting four likelihoods were then normalized so that they summed to 1. Data from both strands were combined by averaging the values. If the most likely base identification had an average normalized likelihood greater than 0.6, it was called, otherwise the base was called as an ambiguity. The example set was derived from two different samples, ib013 and ief005, which have a total of 35 substitutions relative to mt1, of which 19 are shared with the 12 samples analyzed and 16 are not. Identification performance was not sensitive to the choice of examples.
17. To provide an independently determined reference sequence, each 2.5-kb PCR amplicon was sequenced on both strands by primer-directed fluorescent chain–terminator cycle sequencing with an ABI 373A DNA sequencer and assembled and manually edited with Sequencher 3.0. The analysis presented here assumes that the sequence amplified from genomic DNA is essentially clonal [R. J. Monnat and L. A. Loeb, *Proc. Natl. Acad. Sci. U.S.A.* **82**, 2895 (1985)] and that its determination by gel-based methods is correct. A frequent length polymorphism at positions 303 to 309 was not detected by hybridization under the conditions used. It was excluded from analysis and is not part of the set of 180 polymorphisms discussed in the text. However, polymorphisms at this site have previously been differentiated by oligonucleotide hybridization [M. Stoneking, D. Hedgecock, R. G. Higuchi, L. Vigilant, H. A. Erlich, *Am. J. Hum. Genet.* **48**, 370 (1991)].
18. The P^0 intensity footprints were detected in the following way: The reference and sample intensities were normalized (*15*), and *R*, the average of $\log(P^0_{reference}/P^0_{sample})$ over a window of five positions, centered at the base of interest, was calculated for each position in the sequence. Footprints were detected as regions having at least five contiguous positions with a reference or sample intensity at least 50 counts above background and an *R* value in the top 10th percentile for the experiment. At 205 polymorphic sites, where the sample was mismatched to P^0, the mean *R* value was 1.01, with a standard deviation of 0.57. At 35,333 nonpolymorphic sites (that is, where both reference and sample had a perfect match to P^0) the mean value was −0.05, with a standard deviation of 0.25.
19. R. L. Cann, M. Stoneking, A. C. Wilson, *Nature* **325**, 31 (1987); M. Zeviani *et al.*, *Am. J. Hum. Genet.* **47**, 904 (1990); D. C. Wallace, *Annu. Rev. Biochem.* **61**, 1175 (1992); S. Horai, K. Hayasaka, R. Kondo, K. Tsugane, N. Takahata, *Proc. Natl. Acad. Sci. U.S.A.* **92**, 532 (1995); T. Hutchin and G. Cortopassi, *ibid.*, p. 6892.
20. Long-range PCR amplification was carried out on genomic DNA with Perkin-Elmer GeneAmp XL PCR reagents according to the manufacturer's protocol. Primers were L14836-T3 (5′-aattaaccctcactaaagggATGAAACTTCGGCTCACTCCTTGGCG) and RH1066-T7 (5′-taatacgactcactatagggaTTTCATCATGCGGAGATGTTGGATGG), based on RH 1066 [S. Cheng, R. Higuchi, M. Stoneking, *Nature Genet.* **7**, 350 (1994)]. Each 100-μl reaction contained 0.2 μM concentration of each primer and ~10 to 50 ng of total genomic DNA. Transcription reactions were carried out in 10 μl with Ambion MAXIscript kit according to the manufacturer's protocol. The concentration of the 16.6-kb PCR template was ~2 nM, and the reaction contained Ambion 1× biotin-14-CTP/NTP mix and 0.2 mM biotin-16–UTP. Incubation was at 37°C for 2 hours. Fragmentation and hybridization were as described (*13*), except that 3.5 M TMACl and the biotin-labeled oligonucleotide 5′-CTGAACGGTAGCATCTTGAC were used in the hybridization buffer, which also contained fragmented baker's yeast RNA (100 μg/ml) (Sigma). Hybridization was carried out at 40°C for 4 hours.
21. A custom telecentric objective lens with a numerical aperture of 0.25 focuses 5 mW of 488-nm argon laser light to a 3-μm-diameter spot, which is scanned by a galvanometer mirror across a 14-mm field at 30 lines per second. Fluorescence collected by the objective is descanned by the galvanometer mirror, filtered by a dichroic beamsplitter (555 nm) and a band-pass filter (555 to 607 nm), focused onto a confocal pinhole, and detected by a photomultiplier. Photomultiplier output is digitized to 12 bits. A 4096 by 4096 pixel image is obtained in less than 3 min. Pixel size is 3.4 μm. The data from four sequential scans were summed to improve the signal-to-noise ratio.
22. M. D. Brown, A. S. Voljavec, M. T. Lott, I. MacDonald, D. C. Wallace, *FASEB J.* **6**, 2791 (1992).
23. Mitochondrial DNA populations can contain more than one sequence type, in a condition known as heteroplasmy. The LHON mutations shown in Fig. 3C were characterized as being homoplasmic by conventional sequencing and restriction endonuclease digestion (M. Brown, personal communication). In controlled mixing experiments, we have shown that sequences present at the level of 10% can easily be detected by hybridization (M. Chee and R. Yang, unpublished results; N. Shen, personal communication). The sensitivity of detection is sequence dependent. Importantly, hybridization can be used to detect heterozygous nuclear DNA sequences (J. Hacia *et al.*, in preparation).
24. D. J. Lockhart *et al.*, *Nature Biotech.*, in press.

25. We thank M. Brown and D. Wallace for the gift of the LHON sample and R. Ward for the 10 African samples, M. Trulson for assistance in two-color hybridization, P. Fiekowsky for image analysis, and P. Berg and E. Lander for comments on the manuscript. R. Davis contributed to the initial concepts in oligonucleotide tiling. We especially thank L. Stryer for his incessant and persistent encouragement. Supported in part by Human Genome grant 5RO1HG00813 from NIH (S.P.A.F.).

5 April 1996; accepted 26 July 1996

An Asymmetric Model for the Nucleosome: A Binding Site for Linker Histones Inside the DNA Gyres

Dmitry Pruss, Blaine Bartholomew, Jim Persinger, Jeffrey Hayes, Gina Arents, Evangelos N. Moudrianakis, Alan P. Wolffe*

Histone-DNA contacts within a nucleosome influence the function of trans-acting factors and the molecular machines required to activate the transcription process. The internal architecture of a positioned nucleosome has now been probed with the use of photoactivatable cross-linking reagents to determine the placement of histones along the DNA molecule. A model for the nucleosome is proposed in which the winged-helix domain of the linker histone is asymmetrically located inside the gyres of DNA that also wrap around the core histones. This domain extends the path of the protein superhelix to one side of the core particle.

The nucleosome has an active role in gene regulation. Mutations of the core histones have specific consequences for the transcription of particular genes (*1*). The specificity of these effects can be explained both by the positioning of histones with respect to DNA sequence (*2*) and the potential targeting of histone modifications to particular nucleosomes (*3*). Thus, an understanding of nucleosomal architecture is central to understanding the transcription process.

The nucleosome contains two molecules of each of the four core histones (H2A, H2B, H3, and H4), a single molecule of a linker histone (H1, H1°, or H5), and ~180 base pairs (bp) of DNA (*4*). In isolation, the core histones assemble into an octameric complex (*5*), whose structure has been determined at 3.1Å resolution (*6–8*). The exact path of DNA on the surface of the histone octamer, the position of the linker histone molecule within the nucleosome, and the path of linker DNA between adjacent nucleosomes (*9–11*) remain to be determined.

We used positioned nucleosomes containing the *Xenopus borealis* somatic 5S ribosomal RNA (rRNA) gene to examine

D. Pruss and A. P. Wolffe, Laboratory of Molecular Embryology, National Institute of Child Health and Human Development, National Institutes of Health, Building 6, Room B1A-13, Bethesda, MD 20892–2710, USA.
B. Bartholomew and J. Persinger, Department of Medical Biochemistry, Southern Illinois University at Carbondale, School of Medicine, Carbondale, IL 62901–4413, USA.
J. Hayes, Department of Biochemistry, School of Medicine and Dentistry, University of Rochester, Rochester, NY 14642, USA.
G. Arents and E. N. Moudrianakis, Department of Biology, Johns Hopkins University, Baltimore, MD 21218, USA.

*To whom correspondence should be addressed. E-mail: awlme@helix.nih.gov