Direct haplotyping of kilobase-size DNA using carbon nanotube probes

Adam T. Woolley[1], Chantal Guillemette[2], Chin Li Cheung[1], David E. Housman[2]*, and Charles M. Lieber[1]*

*[1]Department of Chemistry and Chemical Biology, Harvard University, 12 Oxford Street, Cambridge, MA 02138. [2]Center for Cancer Research, Massachusetts Institute of Technology, Cambridge, MA 02139. *Corresponding authors (cml@cmliris.harvard.edu or dhousman@mit.edu).*

Received 18 February 2000; accepted 2 May 2000

We have implemented a method for multiplexed detection of polymorphic sites and direct determination of haplotypes in 10-kilobase-size DNA fragments using single-walled carbon nanotube (SWNT) atomic force microscopy (AFM) probes. Labeled oligonucleotides are hybridized specifically to complementary target sequences in template DNA, and the positions of the tagged sequences are detected by direct SWNT tip imaging. We demonstrated this concept by detecting streptavidin and IRD800 labels at two different sequences in M13mp18. Our approach also permits haplotype determination from simple visual inspection of AFM images of individual DNA molecules, which we have done on UGT1A7, a gene under study as a cancer risk factor. The haplotypes of individuals heterozygous at two critical loci, which together influence cancer risk, can be easily and directly distinguished from AFM images. The application of this technique to haplotyping in population-based genetic disease studies and other genomic screening problems is discussed.

Keywords: atomic force microscopy, haplotype, carbon nanotube, single-nucleotide polymorphism, DNA sequencing

The Human Genome Project is now providing massive amounts of genetic information that should revolutionize both the understanding and diagnosis of inherited diseases. In particular, the cataloging of single-nucleotide polymorphisms (SNPs) in gene coding and regulatory regions should lead to a greater comprehension of the genetic contribution to risk for common diseases such as cancer and heart disease[1–3]. To achieve maximum power, the haplotype of a subject—the specific alleles associated with each chromosome homolog—is a critical element in SNP mapping. However, the current methods for determining haplotypes have significant limitations that have prevented their use in large-scale genetic screening. For example, parental genotyping can be used to infer haplotypes in a family study[4,5], although in many cases it is impractical or impossible to obtain parental DNA. Furthermore, molecular techniques for determining haplotypes, such as allele-specific[6] or single-molecule polymerase chain reaction (PCR) amplification[7], are hampered by the need to optimize stringent reaction conditions and the potential for significant error rates.

Here we propose a simple, yet elegant new approach to haplotyping by direct visualization of polymorphic sites on individual DNA molecules. Our method utilizes atomic force microscopy (AFM)[8] with high-resolution single-walled carbon nanotube (SWNT) probes[9,10] to read directly multiple polymorphic sites in DNA fragments containing from ~100 to at least 10,000 bases. We have demonstrated this approach on specifically labeled sequences in an M13mp18 model system and have further applied our technique for determination of haplotypes on the UGT1A7 gene, which is under study for its role in cancer epidemiology[11]. The throughput of this method could be readily extended, and thus may become an important technique for probing the relationship of genetic variations to disease susceptibility and for population screening.

Our approach (Fig. 1) involves specific hybridization of labeled oligonucleotide probes to target sequences in DNA fragments, followed by direct reading of the presence and spatial locations of the labels by AFM. The oligonucleotide probes are designed such that under appropriate hybridization conditions binding does not occur in the presence of a single-base mismatch at polymorphic sites; i.e., labels are detected only at sequences fully complementary to the oligonucleotides. We utilize SWNT tips, which can be reproducibly prepared with tip radii less than 3 nm and ~10 base resolution[10,12], to enable high-resolution, multiplex detection of different labels. In these studies the oligonucleotides were labeled with either streptavidin or the fluorophore IRD800, which can be consistently distinguished from one another on the basis of size.

Results and discussion

Multiplexed sequence detection in M13mp18. We have tested this new method by identifying the spatial location of specific sequences with excellent discrimination from corresponding single-base mismatches in the M13mp18 plasmid using seven-base oligonucleotide probes. The essence of this experiment is captured in the AFM image of a DNA molecule that was marked with a streptavidin-labeled GGGCGCG sequence (Fig. 2A). This image shows a DNA fragment with a 2,200 nm contour length consistent with the 7,249 bp (ref. 13) of M13mp18, and a distinct streptavidin label 1,080 nm from one end of the *Bgl*II-digested DNA. Histogram summaries of results obtained from at least 15 streptavidin-labeled M13mp18 DNA molecules show clear peaks at 0.48 (3,512 bp) and 0.40 (2,893 bp) from the fragment ends, for samples cut with *Bgl*II and *Bam*HI, respectively (Fig. 2B, C). In contrast, histograms from control experiments with unlabeled oligonucleotides (data not shown) did not exhibit clusters of labels, indicating that the histogram peaks are due to specific detection of streptavidin. These results demonstrate two key points. First, based on the calculated distances of streptavidin labels from the two restriction site positions, the GGGCGCG site was determined to be at base 3,390 (Fig. 2D). This same site was calculated to be at base 3,402 from similar experiments with IRD800 labeling (data not shown); both

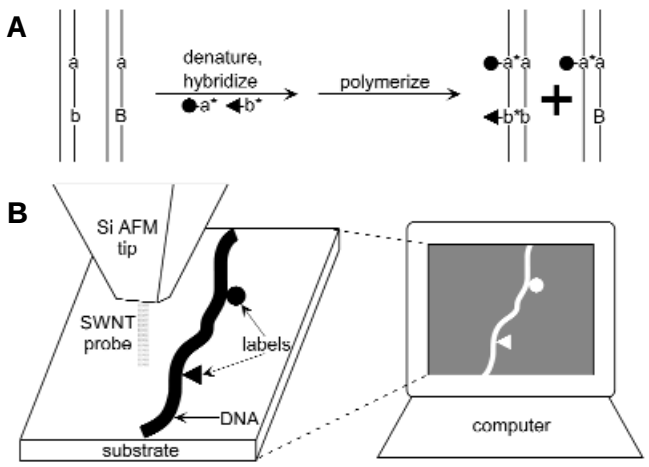

Figure 1. Schematic illustration of the method for labeling specific DNA sites and detection with SWNT AFM probes. (A) Labeled oligonucleotide probes (●-a* and ◀-b*) are specifically annealed to their complementary target sequences (a and b) but not to sequences with a single-base mismatch (A and B) in the ssDNA template. DNA polymerase and dNTPs are then used to synthesize complementary strands, generating dsDNA fragments specifically labeled at a and b with ● and ◀, respectively. (B) Labeled DNA molecules are deposited on freshly cleaved mica and imaged by AFM using SWNT probes. The presence and locations of the sequence-specific tags (● and ◀) can be readily observed in the AFM image.

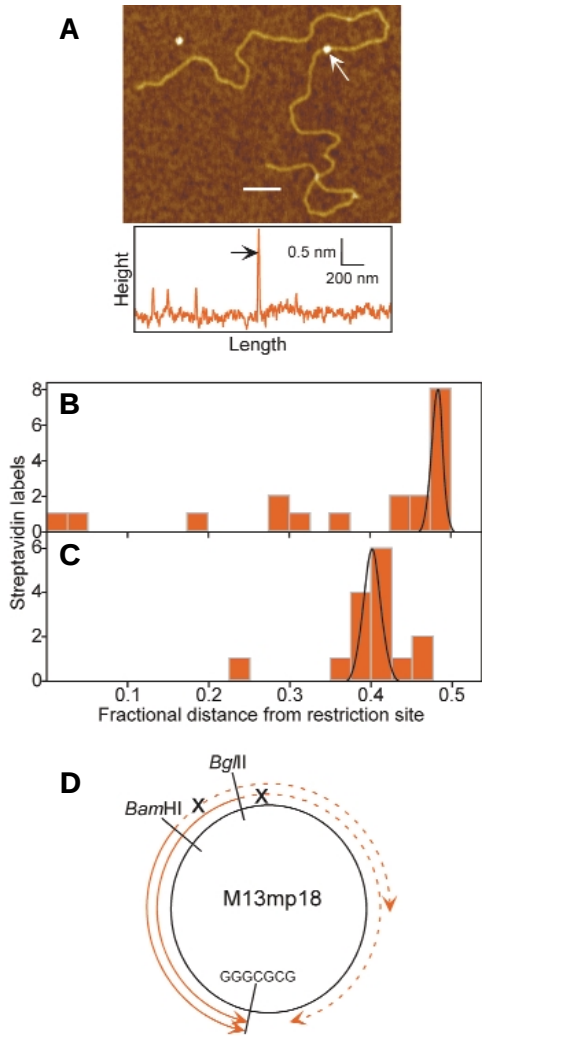

Figure 2. Detection of labeled DNA sites with nanotube tips. (A) SWNT tip AFM image and height profile along DNA, obtained with streptavidin-labeled GGGCGCG in M13mp18 digested with *Bgl*II; the arrow points to the streptavidin tag. Places where DNA strands cross each other (left side of height profile) are easily differentiated from labels. The image height scale is 3 nm, and the white bar corresponds to 100 nm. (B) Histogram of number of streptavidin tags as a function of distance from the *Bgl*II restriction site, obtained from height plots along M13mp18 labeled at GGGCGCG. (C) Same as (B), except sample was digested with *Bam*HI and distances were measured from this site. (D) Map of M13mp18 shows the location of GGGCGCG calculated from the histograms in (B) and (C). Arrowheads indicate possible positions of the target sequence, based on the calculated distance from the restriction sites; the labeled sequence occurs where two arrowheads meet (one from each digest). Solid arcs indicate the correct paths, whereas incorrect paths are shown as dashed arcs with an "×" through them.

results are in excellent agreement with the known location (base 3,405). Second, we see no evidence for labeling at the single-base mismatch sites located at 1,115 and 3,595, thus demonstrating the specificity of labeling and the potential for SNP detection.

Because the streptavidin and IRD800 molecules can be readily distinguished on the basis of their heights and shapes (e.g., average measured height of streptavidin labels was 1.7 ± 0.5 nm vs. 0.7 ± 0.3 nm for IRD800) using SWNT tips, simultaneous detection of two or more distinct sites should be feasible. To test this concept, we prepared M13mp18 labeled at GGGCGCG with IRD800 and at TCTCAGC with streptavidin and then imaged these fragments using SWNT probes. Histograms similar to Figure 2B and C were generated for streptavidin (>1 nm) and IRD800 (<1 nm) peaks detected in surface plots along imaged DNA fragments for *Bgl*II and *Alw*NI digests (data not shown). From these histograms we calculated that TCTCAGC occurs at bases 2,024 and 4,059, in good agreement with its known positions at 2,013 and 4,077, and that GGGCGCG is at base 3,422, corresponding well with the expected value of 3,405. These results demonstrate clearly the potential for multiplexed sequence detection in large DNA strands and open the possibility for profiling multiple polymorphic sites on DNA fragments in the 10 kilobase or larger size range.

Direct haplotype determination in UGT1A7. Our experiments suggest that this approach is ideally suited for identifying specific haplotypes associated with genetic disorders. To illustrate this critical point, we have determined haplotypes on a UDP-glucuronosyltransferase gene, UGT1A7 (ref. 14), the enzyme product of which is involved in inactivation of carcinogens such as benzo[*a*]pyrene metabolites[11]. This gene has two polymorphic sites (separated by 233 bp) that determine four alleles, each specifying different polypeptide chains (Fig. 3A). Importantly, individuals who are heterozygous at both sites have a single genotype, but one of two haplotypes—(*1/*3) or (*2/*4)—which cannot be differentiated using conventional methods. This ambiguity is crucial in screening, since each allele exhibits substantially different enzymatic activity toward targeted carcinogens[11]. To distinguish these haplotypes using SWNT tips, allele-specific probes were hybridized to DNA samples (Fig. 3A). The probes were chosen such that the (*1/*3) haplotype would show singly labeled fragments whereas the (*2/*4) haplotype would exhibit DNA with two or no labels. Significantly, subject samples could be unambiguously shown to exhibit the (*1/*3) haplotype by direct inspection of AFM images (Fig. 3B, C); that is, DNA molecules were only end-labeled with the streptavidin or the IRD800 probes. Representative images of the two types of end-labeled DNA fragments show a 140 nm, streptavidin-tagged DNA (Fig. 3B) and a 210 nm, IRD800-labeled fragment (Fig. 3C), which are characteristic of the *3 and *1 alleles, respectively. Control experiments on samples homozygous for the *2 allele showed no specific labeling (data not shown); because of the low occurrence frequency of the *4 allele[11], we did not test any samples known to carry this allele.

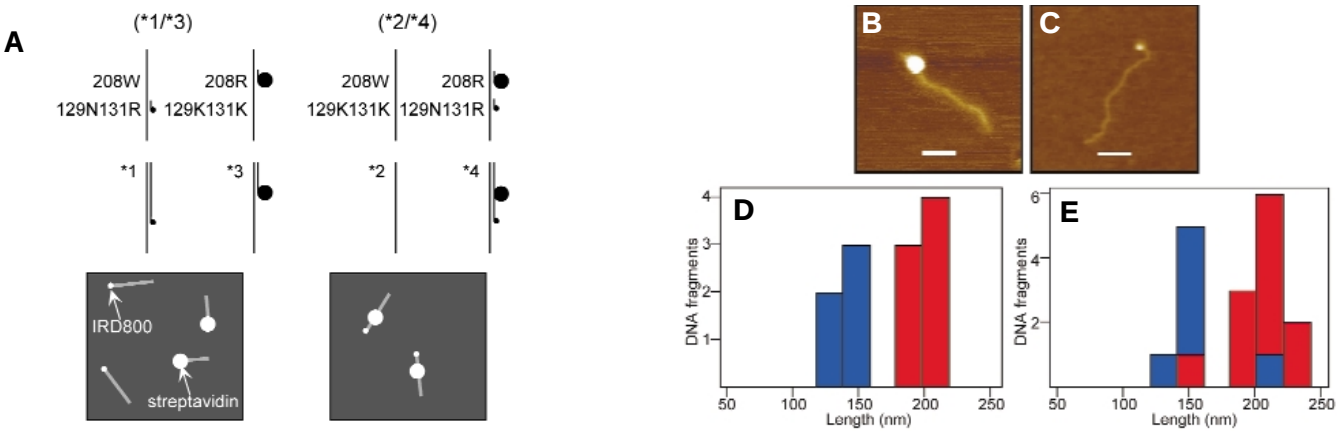

Figure 3. Direct haplotyping of UGT1A7 using SWNT probes. (A) Schematic showing haplotypes, alleles, genotypes, and locations of probes in samples analyzed. The (*1/*3) and (*2/*4) haplotypes, which have the same genotype (heterozygous at both loci), are specifically labeled at the 129N131R and 208R sites with IRD800 (small filled circle) and streptavidin (large filled circle), respectively. AFM images of a (*1/*3) sample should have an approximately equal number of fragments that are ~210 nm (663 bp) long with IRD800 at one end, or ~140 nm (430 bp) long with streptavidin at one end (the random-coil structure of the ssDNA tail should not contribute significantly to the length). In contrast, a (*2/*4) sample should contain ~210 nm fragments with IRD800 at one end and streptavidin ~70 nm (233 bp) distant. (B) Representative SWNT tip AFM height image of the *3 allele (streptavidin end-labeled, ~140 nm DNA) detected in a sample that was heterozygous at both loci. (C) SWNT probe image of the *1 allele (IRD800 end-labeled, ~210 nm DNA) detected in the same sample as (B). The height scale is 3 nm, and the white bar corresponds to 50 nm in both images. (D) Histogram showing number of streptavidin (blue) and IRD800 (red) end-labeled fragments vs. DNA length for a sample known to have either the (*1/*3) or (*2/*4) haplotype. The cluster of streptavidin-tagged DNA at ~140 nm is typical of the *3 allele, and the grouping of IRD800 labeled fragments at ~210 nm indicates the *1 allele. (E) Histogram same as in (D), but from a different individual; both samples were determined to have the (*1/*3) haplotype.

To further substantiate that the haplotype determined by image inspection was indeed the consequence of specific probe hybridization to the expected sites on the target DNA, detailed length measurements were carried out. Histograms plotting number of streptavidin and IRD800 end-labeled DNA fragments as a function of length, for two different samples (Fig. 3D, E), each show a grouping for streptavidin around 140 nm and for IRD800 near 210 nm. These fragment distributions for both samples are in agreement with that expected for the (*1/*3) haplotype (Fig. 3A). Furthermore, no observed fragments matched the predicted profile of the *4 allele (Fig. 3A); hence, we conclude that both of these samples were of the haplotype (*1/*3).

We believe that direct haplotyping using SWNT AFM probes represents a significant advance over conventional approaches and could facilitate the use of SNPs for association and linkage studies of inherited diseases and genetic risk[15,16]. Our studies of M13mp18 show that this methodology could be used to detect multiple SNPs in 10-kb samples with a resolution of approximately 10 bases, and moreover, should be extendable to 100 kb samples with similar resolution. The large DNA sizes that can be directly haplotyped are unique to our technique and will be useful independent of the sample throughput. In addition, the simplicity and distinctiveness of the AFM images of alternative haplotypes indicate that automated analysis may also be feasible. The current throughput for an instrument imaging with a single SWNT tip could be greater than 200 samples per day with a redundancy of 10 independent images per sample. In addition, our methodology could be extended from single sample analysis to a very high-throughput parallel technique by exploiting multiple tip arrays, which have now been made as large as 32×32 for ultrahigh-density hard-disk storage[17]. The implementation of these technical improvements would allow haplotyping of over 200,000 samples per day with a single instrument using our technique. Finally, the recent synthesis of carbon nanotubes with 0.25 nm radii[18], which are smaller than the spacing between DNA bases, indicates that further improvements in nanotube probes[12], nanotube end labeling[9], and/or DNA labeling methods could enable direct reading of the DNA sequence of fragments that are tens of kilobases in size.

Experimental protocol

DNA sequence labeling in M13mp18. Biotin (Operon, Alameda, CA) or IRD800 (Li-Cor, Omaha, NE)-labeled CGCGCCC (8 pmol) was annealed to 80 fmol M13mp18 in 1× EcoPol buffer (New England Biolabs, Beverly, MA) with 100 µM dNTPs at 25°C. Klenow Fragment exo$^-$ (New England Biolabs, 10 U) was added, allowed to sit at 25°C for 5 min, and then warmed to 37°C for 30 min. Restriction digestion was carried out on ~30 fmol DNA with 5 U of *Bgl*II, *Bam*HI, or *Alw*NI in the recommended digestion buffer (New England Biolabs) at 37°C for 60 min. Digesting two separate aliquots with a different restriction enzyme is necessary because DNA fragment ends are indistinguishable by AFM. After digestion, biotin labels were conjugated with streptavidin (7.5 pmol) at room temperature for 10 min in restriction buffer. Samples were ethanol-precipitated and resuspended in 10 mM Tris, 1 mM EDTA (TE), pH 8.0. Multiplex labeling at the sequence GCTGAGA was performed the same as described above, except the annealing was carried out at 15°C.

UGT1A7 alleles. The alleles of UGT1A7 with their GenBank accession numbers and genotypes at the polymorphic loci are UGT1A7*1, HSU39570 and HSU89507 (129N131R208W); UGT1A7*2, AF110191 (129K131K208W); UGT1A7*3, AF110192 (129K131K208R); UGT1A7*4, AF110193 (129N131R208R). The numbers and capital letters in the genotypes (i.e., 208R) correspond to the numbers and types of amino acids in the protein encoded by the UGT1A7 gene at the polymorphic site.

PCR amplification of UGT1A7 samples. PCR primers (forward 5′-CTATCTGTACTTCTTCCACTTAC and reverse 5′-ACTTACATATCAA-CAAGAGCTGC) were designed to amplify a fragment that encompasses both polymorphic sites in the UGT1A7 first exon (from nucleotide -76 to 1048 in the sequence corresponding to GenBank accession number HSU39570). PCR was performed on 20 ng of genomic DNA in 50 µl aliquots containing 20 pmol of each primer, 1× reaction buffer (50 mM KCl, 1.5 mM $MgCl_2$, and 10 mM Tris pH 8.5), 100 µM dNTPs, 4% dimethyl sulfoxide, and 2 U *Taq* DNA polymerase (PE Applied Biosystems, Branchburg, NJ). The amplification conditions were as follows: denaturation at 94°C for 5 min, 5 cycles each consisting of 60 s at 94°C, 45 s at 62°C, 90 s at 72°C, followed by 30 cycles each consisting of 60 s at 94°C, 45 s at 56°C, 90 s at 72°C, followed by 7 min at 72°C. PCR products were purified using Qiagen Quick Columns (Qiagen, Santa Clarita, CA) to remove the primers and then dissolved in water.

Labeling UGT1A7 alleles. PCR amplicons (~100 fmol) were denatured at 95°C for 10 min, and then oligonucleotides (4 pmol) complementary to the 129N131R (IRD800-AATGACCGA) and 208R (biotin-AGTACGGAA) loci were annealed at 24°C in 1× EcoPol buffer with 50 µM dNTPs. Klenow

Fragment exo$^-$ (2.5 U) was added, and the mixture was maintained at 24°C for 2 min, followed by heating to 37°C for 30 min to extend the primed strands. Samples were purified with Concert PCR purification systems (Gibco BRL-Life Technologies, Grand Island, NY) to remove excess primers, and the DNA was resuspended in TE, pH 7.0. Streptavidin (0.5 pmol) was conjugated to biotinylated DNA in TE with 0.1 M NaCl for 2 h at room temperature.

Sample deposition and AFM imaging. DNA was diluted to ~100 pg μl^{-1} in 10 mM $MgCl_2$ and deposited onto freshly cleaved mica for 5 min. Then the surface was rinsed several times with water and dried gently under a stream of nitrogen gas before AFM imaging[10]. Images were recorded under ambient conditions in tapping-mode at 1.5–2 Hz with a tip resonance frequency of 60–70 kHz and amplitudes of 15–40 nm using a Digital Instruments (Santa Barbara, CA) Multimode Nanoscope IIIa. In contrast to previous contact-mode studies with microfabricated tips[19], the relative humidity was found to have a minimal influence on measured DNA heights in our experiments, and thus no efforts were made to control ambient humidity. The insensitivity to humidity is due to the very small cylindrical structure of the nanotube probes[9,10,12,20]. SWNT ropes were mounted on Au-coated force modulation etched silicon probe cantilevers (Digital Instruments, k = 1–5 N m^{-1}) using micromanipulators under the direct view of an optical microscope as described previously[9,20].

Acknowledgments

C.M.L. acknowledges support of this work from AFOSR. A.T.W. and C.G. acknowledge fellowship support from the Cancer Research Fund of the Damon Runyon-Walter Winchell Foundation and the Medical Research Council of Canada, respectively.

1. Brookes, A.J. The essence of SNPs. *Gene* **234**, 177–186 (1999).
2. Cargill, M. et al. Characterization of single-nucleotide polymorphisms in coding regions of human genes. *Nat. Genet.* **22**, 231–238 (1999).
3. Halushka, M.K. et al. Patterns of single-nucleotide polymorphisms in candidate genes for blood-pressure homeostasis. *Nat. Genet*. **22**, 239–247 (1999).
4. Sobel, E. & Lange, K. Descent graphs in pedigree analysis: applications to haplotyping, location scores, and marker-sharing statistics. *Am. J. Hum. Genet.* **58**, 1323–1337 (1996).
5. Hodge, S.E., Boehnke, M. & Spence, M.A. Loss of information due to ambiguous haplotyping of SNPs. *Nat. Genet.* **21**, 360–361 (1999).
6. Ruano, G. & Kidd, K.K. Direct haplotyping of chromosomal segments from multiple heterozygotes via allele-specific PCR amplification. *Nucleic Acids Res.* **17**, 8392 (1989).
7. Ruano, G., Kidd, K.K. & Stephens, J.C. Haplotype of multiple polymorphisms resolved by enzymatic amplification of single DNA molecules. *Proc. Natl. Acad. Sci., USA* **87**, 6296–6300 (1990).
8. Binnig, G., Quate, C.F. & Gerber, C. Atomic force microscope. *Phys. Rev. Lett.* **56**, 930–933 (1986).
9. Wong, S.S., Woolley, A.T., Joselevich, E., Cheung, C.L. & Lieber, C.M. Covalently-functionalized single–walled carbon nanotube probe tips for chemical force microscopy. *J. Am. Chem. Soc.* **120**, 8557–8558 (1998).
10. Wong, S.S. et al. Single-walled carbon nanotube probes for high-resolution nanostructure imaging. *Appl. Phys. Lett.* **73**, 3465–3467 (1998).
11. Guillemette, C., Ritter, J.K., Auyeung, D.J., Kessler, F.K. & Housman, D.E. Structural heterogeneity at the UGT1A loci: functional consequences of three novel missense mutations in the human UGT1A7 gene. *Pharmacogenetics*, **in press** (2000).
12. Hafner, J.H., Cheung, C.L. & Lieber, C.M. Direct growth of single-walled carbon nanotube scanning probe microscopy tips. *J. Am. Chem. Soc.* **121**, 9750–9751 (1999).
13. Fang, Y. et al. Solid-state DNA sizing by atomic force microscopy. *Anal. Chem.* **70**, 2123–2129 (1998).
14. Mackenzie, P.I. et al. The UDP glycosyltransferase gene superfamily: recommended nomenclature update based on evolutionary divergence. *Pharmacogenetics* **7**, 255–269 (1997).
15. Risch, N. & Merikangas, K. The future of genetic studies of complex human diseases. *Science* **273**, 1516–1517 (1996).
16. Collins, F.S., Guyer, M.S. & Chakravarti, A. Variations on a theme—cataloging human DNA sequence variation. *Science* **278**, 1580–1581 (1997).
17. Vettiger, P. et al. Ultrahigh density, high-data-rate NEMS-based AFM data storage system. *Microelectronic Engineering* **46**, 11–17 (1999).
18. Sun, L. F. et al. Creating the narrowest carbon nanotubes. *Nature* **403**, 384 (2000).
19. Thundat, T. et al. Atomic force microscopy of deoxyribonucleic acid strands adsorbed on mica: the effect of humidity on apparent width and image contrast. *J. Vac. Sci. Technol. A* **10**, 630–635 (1992).
20. Wong, S.S., Harper, J.D., Lansbury, P.T., Jr. & Lieber, C.M. Carbon nanotube tips: high resolution probes for imaging biological systems. *J. Am. Chem. Soc.* **120**, 603–604 (1998).