Sequencing with Microarray Technology—A Powerful New Tool For Molecular Diagnostics

Nucleic acid sequencing is a fundamental technique that was recognized with the 1980 Nobel Prize in Chemistry. The method allows delineation of DNA sequences with extraordinary accuracy and, since its introduction in the 1970s, has undergone many important modifications and improvements. Among these are the achievement of long reads (up to ~1000 bp per analysis), better accuracy (related to the discovery of highly versatile and thermostable sequencing enzymes), improved sensitivity with thermocycling protocols (linear amplifications), full automation, higher speed (related to the introduction of thin gels), and substitution of radioactivity with fluorescent and other probes. All of these improvements have allowed scientists to attempt something that was unthinkable 15–20 years ago, i.e., delineation of the complete sequence of the human genome (~3×10^9 bp) and genomes of other organisms. We have already witnessed, over the last few years, the release of the complete sequence of simple organisms as well as of more complex ones, the latest being the *Drosophila* genome *(1)*. We are now very close to the completion of the entire Human Genome Project *(2)*. These achievements represent a triumph of the DNA sequencing methodology.

Now that we (almost) know the complete DNA sequences of these organisms and humans, the question arises. How are we to use this information? The next step will be the complete annotation of the human genome, which will include classification of the raw DNA sequence into well-defined gene structures. We will then need to predict and experimentally verify the encoded proteins and their possible biological functions (physiology). Once this is done, we can begin to ask questions about how genomic variation in certain genes (polymorphisms, mutations) can cause or predispose to specific human diseases (pathophysiology). We already have many examples of subtle genetic changes that can cause very serious human diseases, including cystic fibrosis, various forms of anemias, premature atherosclerosis, cancers, neurodegenerative diseases, autoimmune and immunodeficiency syndromes, and other conditions. The number of known disease-related variations will surely increase dramatically with our increasing knowledge of the sequence of the human genome. Already, many researchers and companies are trying to identify all single nucleotide polymorphisms within the human genome.

One could then envision diagnostic, prognostic, preventive, and therapeutic approaches based on determining specific DNA variations within the human genome. How are we going to achieve these goals? Are the current sequencing technologies adequate for these tasks? Will we be able to scan, quickly and cheaply, the whole human genome of each individual for the purpose of identifying genetic variations associated with human diseases?

Let us consider a simple example to examine whether the current sequencing technologies, based on the classical dideoxy (chain termination) method and separation of DNA fragments on gels or capillaries, could meet these future requirements. Imagine that you wish to examine a 50-kb region of the human genome to identify mutations or polymorphic sites. Please remember that this 50-kb region represents only 1/60 000th of the whole human genome. Furthermore, imagine that you want to check 1000 different individuals for this particular region of DNA. To do this, you will need to sequence a total of 50 Mb of DNA (50 kb × 1000, or 5×10^7 bp). If you have a sequencer that can simultaneously process 30 lanes and you can obtain ~500 bases of sequence per lane (these numbers are reasonable, based on the capabilities of current machines), you will be able to obtain 15 kb of sequence per day (time would include PCR amplification, sequencing reactions, and loading and running the machine). To finish the task mentioned above, you would need ~10 working years. Clearly, such methodology would not be appropriate for high-throughput, low-cost sequencing of large amounts of DNA.

A solution to this problem could include methods that can look at the whole 50-kb sequence at once. For example, if you have a device that can identify changes in the 50-kb sequence in one experiment and you can run many devices per day (e.g., 50–100), then the task can be completed within a few days.

Although various technical improvements in current sequencing methods have been introduced, including ultra-thin gel technology *(3)* and capillary electrophoresis *(4)*, the throughputs of these technologies are still limited. At present, the most efficient ways of performing massive parallel sequencing include, in principle, sequencing by hybridization on miniature devices known as microarrays *(5, 6)*. These devices are already used extensively for deriving information about gene expression patterns *(7–12)*. For sequencing, the method involves immobilization of the target to be sequenced in a microarray format, hybridization of this target with a very large set of short, labeled probes (e.g., 10 000 probes, each seven nucleotides long), and then examination of the pattern of hybridization and computation of the original DNA sequence *(13)*. Another approach involves exactly the reverse, i.e., immobilization of thousands of short probes (~18–20 nucleotides) in a microarray and then hybridization of these short probes with the target of interest, which has been labeled beforehand with a fluorescent probe. Again, from the pattern of hybridization, one can decipher the DNA sequence of interest in one analysis *(14)*.

In the approach that uses immobilization of short oligonucleotides on an array for sequencing, let us consider an example. Imagine that you want to detect a possible base variation around codon 248 of the *p53* gene (this codon is in exon 7). The *p53* gene is a valuable model system because it is mutated at random positions throughout its entire ~1-kb coding sequence in many

forms of cancer *(15)*. At any particular area of the chip, there are five immobilized oligonucleotides, four differing by one base—A, C, G, or T—plus one that has the base deleted. After hybridization and scanning of the particular area, only one of the five oligonucleotides will form a perfect match with the target and thus reveal the sequence of the DNA in the sample. If there is a heterozygous or homozygous mutation in this position, the oligonucleotide probes will hybridize differently from the wild-type sequence and reveal a signal that detects this particular base change. In one example, we were able to detect either the wild-type sequence (CGG) or a mixture of wild-type and mutant alleles (heterozygous mutation, CA/GG). With this particular technology, every nucleotide of the *p53* gene coding sequence can be tested in areas comprising five oligonucleotides each. If we are interested in scanning the coding sequence of the whole gene (~1262 bp), we will need ~6000–7000 different oligonucleotide probes (gene length × five oligomers per base). For double-stranded (sense and antisense) sequencing, this number should be doubled. With the current technology (photolithography), >50 000 oligonucleotides have been immobilized on a 1.28 × 1.28 cm area of a chip, and densities can go up to 400 000 oligonucleotides. Because the method can be automated, many samples can be run in parallel. From this analysis, it can be concluded that the technology currently exists to produce parallel massive sequencing data using miniature microchip technology.

In this issue of *Clinical Chemistry*, Wikman et al. *(16)* evaluate the performance of the p53 GeneChip® (Affymetrix) using bladder cancer tissue samples previously sequenced with the classic technology. I will not focus on the actual clinical importance of identifying *p53* gene mutations but will use this example as a model to investigate whether the current microarray sequencing by technology works and to what extent.

The message from the report by Wikman et al. *(16)* is very clear. The chip has important advantages as a sequencing method, regarding speed and amount of data generated, and it is also almost free from interference by mixtures of templates from nonpathological and pathological tissue. The authors demonstrated that even 1% content of target from the diseased tissue can be detected in the presence of 99% wild-type content. This is in contrast to the capabilities of classical sequencing methodologies, which usually require at least 30% pathological tissue for accurate detection of mutations *(3)*. The improved sensitivity reflects the fact that with the microchip sequencing technology, each spot (or "cell") coated with a specific oligonucleotide is testing for a unique sequence, which could be either wild-type or mutant. Thus, the wild-type contamination does not interfere with the mutant-type detection and vice versa.

Wikman et al. *(16)* discovered a need for modifications of the chip method. For example, they were forced to modify the hybridization oven to provide constant agitation of the chip to achieve more uniform staining. In addition, evaluation of the data with either the GeneChip report or by a fixed cutoff value frequently gave erroneous results (false positives and false negatives). The authors attributed this problem to the fact that every cell in the GeneChip should be considered a unique area because it contains an oligonucleotide of a specific sequence. The background signals generated by these oligonucleotides may not be the same across the whole chip. Consequently, use of a single signal cutoff may not be appropriate for every cell. The authors solved this problem by performing multiple experiments and determining (where possible) a more statistically sound background value for each cell and then using this cutoff to make decisions concerning the presence or absence of the mutation. This approach improved the specificity of the GeneChip to 86%, which is similar to the specificities observed in other studies. In a previous study, Ahrendt et al. *(17)* used with the same gene chip for lung cancer *p53* gene sequence analysis and found that the specificity was almost 100%, even when a fixed cutoff value was used. The difference between the study by Wikman et al. *(16)* and the study by Ahrendt et al. *(17)* is not obvious, but there may be batch-to-batch differences in these devices.

In the studies by Ahrendt et al. *(17)* and Wikman et al. *(16)*, it was concluded that the GeneChip does not detect frameshift mutations (from insertions or deletions in the *p53* gene). This is certainly an important limitation because it has already been established that such frameshift mutations represent 10–20% of all known *p53* mutations *(15)*. The reason that the chip fails to detect such mutations is that it is not designed to do so, with the exception of single-base deletions.

The GeneChip sequencing methodology and other similar devices provide an effective conceptual solution to the need for detecting sequence variability between individuals, over large genomic regions, with reasonable ease, speed, and cost. Clearly, these methods are promising, and they will likely become routine over the next few years. Microarray technologies are already used in various formats for gene expression studies. If diagnostic and therapeutic decisions are to be based on such results, the methods must be shown to be highly reliable and reproducible. The current technologies lack robustness, and the commercial products seem to require modification of the assay protocols to obtain reasonable results. With appropriate maturation and further development, these techniques may soon become the workhorses of clinical molecular diagnostics.

In a second report in this issue of *Clinical Chemistry*, Inganäs et al. *(18)* address a slightly different problem. Again, the authors selected *p53* as a model system. As indicated, the gold standard for detecting *p53* gene mutations is sequencing of the whole gene. However, if one deals with large numbers of samples, it is not efficient to sequence samples or gene regions that likely do not have any mutations. How would I know whether a specific sample has mutations before beginning the time-consuming and expensive procedure of sequencing the whole gene? This is where gene-scanning methodologies, which allow researchers or clinicians to determine whether there is a mutation somewhere in a gene, can help. These

methods do not indicate the precise location or the type of mutation, but they do provide information about the region of the gene in which it is located. If this finding is followed with DNA sequencing, will enables detection of the exact position and type of the mutation. Scanning techniques that are already widely used include heteroduplex analysis, single-strand conformation polymorphism, denaturing gradient and gel electrophoresis, and chemical cleavage methods.

Inganäs et al. *(18)* evaluated a newly developed method, known under the trade name PASSPORT™ Mutation Scanning. The method involves PCR amplification, heteroduplex formation by hybridization of the PCR product with a wild-type reference DNA (this is accomplished by simple mixing and heating), and then enzymatic cleavage of any heteroduplex that may form (if there are mutations) by the enzyme T4 endonuclease VII. This enzyme has the ability to cleave both DNA strands around the mismatch, usually within 3–6 bp of the mismatch. The cleavage products are then detected by electrophoresis. The number of cleavages detected corresponds to the number of possible mutations (up to five can be detected in a single heteroduplex).

Inganäs et al. *(18)* addressed the issues of sensitivity, specificity, and positional accuracy of this method through the use of *p53* as a model and colorectal cancer DNA samples. For resolution of the fragments, they used manual methods based on the detection of radioactivity as well as fluorescence-based detection on automated DNA sequencers. Their results are very encouraging in that the sensitivity and specificity are quite high compared with direct DNA sequencing data, which should also be interpreted with caution because the sensitivity and specificity of this method are not 100%. The authors demonstrated this deficiency of the standard method by repeat analysis and reclassification of some samples from wild-type to mutant. They report a sensitivity of 92–100% and a specificity of 85–91% for the PASSPORT methods. These are impressive numbers, considering that the most popular scanning method, single-strand conformation polymorphism, is neither as sensitive nor as specific for detecting such mutations.

In conclusion, the two reports in this issue of *Clinical Chemistry* offer new ways of performing genetic analysis. One method improves the efficiency with which genetic scanning is performed and may replace other, less efficient methodologies currently in use. The GeneChip enables quick examination of large genomic regions for presence of mutations. Additional improvements in the GeneChip methodology and reduction of cost will likely facilitate its use on a routine basis.

References

1. Myers EW, Sutton GG, Delcher AL, Dew IM, Fasulo DP, Flanigan MJ, et al. A whole-genome assembly of *Drosophila*. Science 2000;287:2196–204.
2. Pennisi E. DOE team sequences three chromosomes. Science 2000;288: 417–18.
3. Bharaj BS, Angelopoulou K, Diamandis EP. Rapid sequencing of the *p53* gene with a new automated DNA sequencer. Clin Chem 1998;44:1397–403.
4. Carrilho E. DNA sequencing by capillary array electrophoresis and microfabricated array systems. Electrophoresis 2000;21:55–65.
5. McKenzie SE, Mansfield E, Rappaport E, Surrey S, Fortina P. Parallel molecular genetic analysis. Eur J Hum Genet 1998;6:417–29.
6. Kruglyak S. Multistage sequencing by hybridization. J Comput Biol 1998;5: 165–71.
7. Lee PS, Lee KH. Genomic analysis. Curr Opin Biotechnol 2000;11:171–5.
8. Thompson M, Furtago LM. High density oligonucleotide and DNA probe arrays for the analysis of target DNA. Analyst 1999;124:1133–6.
9. Carulli JP, Artinger M, Swain PM, Root CD, Chee L, Tulig C, et al. High throughout analysis of differential gene expression. J Cell Biochem Suppl 1998;30–31,286–96.
10. Epstein CB, Butow RA. Microarray technology-enhanced versatility, persistent challenge. Curr Opin Biotechnol 2000;11:36–41.
11. Debouck C, Goodfellow PN. DNA microarrays in drug discovery and development. Nat Genet 1999;21:48–50.
12. Gerhold D, Rushmore T, Caskey CT. DNA chips: promising toys have become powerful tools. Trends Biochem 1999;24:168–73.
13. Drmanac S, Kita D, Labat I, Hauser B, Schmidt C, Burczak JD, Drmanac R. Accurate sequencing by hybridization for DNA diagnostics and individual genomics. Nat Biotechnol 1998;16:54–8.
14. Pease AC, Solas D, Sullivan EJ, Cronin MT, Holmes CP, Fodor SP. Proc Natl Acad Sci U S A 1994;91:5022–6.
15. Hollstein M, Hergenhahn M, Yang Q, Bartsch H, Wang ZQ, Hainaut P. New approaches to understanding p53 gene tumor mutation spectra. Mutat Res 1999;431:199–209.
16. Wikman FP, Lu M-L, Thykjaer T, Olesen SH, Andersen LD, Cordon-Cardo C, Ørntoft TF. Evaluation of the performance of a p53 sequencing microarray chip using 140 previously sequenced bladder tumor samples. Clin Chem 2000;46:1555–61.
17. Ahrendt SA, Halachmi S, Chow JT, Wu L, Halachmi N, Yang SC, et al. Rapid p53 sequence analysis in primary lung cancer using an oligonucleotide probe array. Proc Natl Acad Sci U S A 1999;96:7382–7.
18. Inganäs M, Byding S, Eckersten A, Eriksson S, Hultman T, Jorsback A, et al. Enzymatic mutation detection in the *P53* gene. Clin Chem 2000;46:1562–73.

Eleftherios P. Diamandis

Department of Pathology and Laboratory Medicine
Mount Sinai Hospital
600 University Ave.
Toronto, Ontario M5G 1X5, Canada
and
Department of Laboratory Medicine and Pathobiology
University of Toronto
Toronto, Ontario, Canada
Fax 416-586-8628
E-mail ediamandis@mtsinai.on.ca