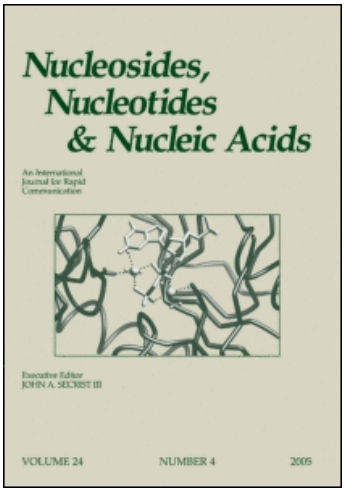

Nucleosides, Nucleotides and Nucleic Acids

Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t713597286

Dna Sequencing After the Human Genome Project

Charles R. Cantor[a]; Kai Tang[b]; Joel H. Graber[a]; Maryanne Maloney[b]; Dong Jing Fu[b]; Natalia E. Broude[a]; Fouad Siddiqi[a]; Hubert Koester[c]; Cassandra L. Smith[a]

[a] Center for Advanced Biotechnology and Departments of Biomedical Engineering, Biology, and Pharmacology and Experimental Therapeutics, Boston University, Boston, MA, USA [b] Center for Advanced Biotechnology and Departments of Biomedical Engineering, Biology, and Pharmacology and Experimental Therapeutics, Sequenom, Inc, San Diego [c] Department of Chemistry, University of Hamburg,

To cite this Article Cantor, Charles R. , Tang, Kai , Graber, Joel H. , Maloney, Maryanne , Fu, Dong Jing , Broude, Natalia E. , Siddiqi, Fouad , Koester, Hubert and Smith, Cassandra L.(1997) 'Dna Sequencing After the Human Genome Project', Nucleosides, Nucleotides and Nucleic Acids, 16: 5, 591 — 598

To link to this Article: DOI: 10.1080/07328319708002921

URL: http://dx.doi.org/10.1080/07328319708002921

NUCLEOSIDES & NUCLEOTIDES, 16(5&6), 591-598 (1997)

DNA SEQUENCING AFTER THE HUMAN GENOME PROJECT

Charles R. Cantor*, Kai Tang+, Joel H. Graber, Maryanne Maloney+, Dong Jing Fu+, Natalia E. Broude, Fouad Siddiqi, Hubert Koester#, and Cassandra L. Smith
Center for Advanced Biotechnology and Departments of Biomedical Engineering, Biology, and Pharmacology and Experimental Therapeutics, Boston University, Boston MA 02215 USA; +Sequenom, Inc, San Diego, #Department of Chemistry, University of Hamburg.

ABSTRACT In future DNA sequencing, gel electrophoresis, which is particularly effective for *de novo* sequencing, is likely to be replaced by sequencing by hybridization, mass spectrometry, or combinations of these two methods, which are particularly effective for comparative or diagnostic sequencing.

Today, most large scale DNA sequencing is carried out by automated gel electrophoresis of fluorescently-labeled samples. A typical instrument can analyze 10 to 36 samples in 2 to 10 hours, reading anywhere from 400 to 1000 bases from each sample. While this is extremely impressive when compared to manual sequencing rates of a few years ago, it is not really sufficient to allow for cost effective large scale diagnostic DNA sequencing or gene expression profiling. Gel electrophoresis is ideally suited for *de novo* DNA sequencing where accurate, long reads are required, especially for DNA from higher organisms where frequent interspersed repeats severely complicate the difficulty of assembly of individual sequence reads into finished sequence. However, it is not necessarily ideal for diagnostic DNA sequencing or expression profiling, where all that is needed is a comparison against a reference.

A number of alternatives to gel electrophoretic DNA sequencing have been proposed, and some of these are under intense investigation. At one extreme are methods that attempt to sequence single DNA molecules either by fluorescent detection of single nucleotides released by exonuclease digestion [1] or by some form of scanning probe microscopy. These approaches, if successful, might allow very long read lengths that would facilitate *de novo* DNA sequencing, but today they seem very far from successful execution. At the other extreme are sequencing by synthesis or other single-base addition or removal strategies such as Genetic Bit Analysis [2]. These work, and they are potentially

executable in a highly parallel form which could allow very high throughput. However, today the methods that seem to be the most prominent candidates for large scale application, predominantly in comparative DNA sequencing are the use of dense sample or probe arrays interrogated by hybridization or simple enzymatic assays, and various forms of mass spectrometry. These methods are quite distant from methods currently in use in typical DNA-based assays. However, they appear, separately and in concert, to have extraordinary promise for vastly increasing the speed and decreasing the cost of comparative DNA sequencing needed for mutation detection in diagnostic applications and for quantitation of mRNA populations for transcript imaging. The new methods are also potentially very user-friendly once the initial barrier of first exposure to them is overcome.

In sequencing by hybridization (SBH) as originally proposed [3-7], a complete array of all probes of a given length would be constructed on a surface. Usually it was considered that octamers would be used, so a total of 65,336 probes would be needed. Fragments of the target sample would be allowed to hybridize to the immobilized probes, and the pattern of hybridization would reveal the sequence of the target. A number of problems interfere with this initial design. First, hybridization of short DNA fragments is not stringent for correct base pairing at the ends of the duplex. Second, it is impossible to find single experimental conditions that allow highly discriminatory hybridizations across the full range of base composition. Third, target secondary structure strongly competes for hybridization to short probes and, as a result, some target sequences cannot be read at all. Fourth, typical hybridizations are not quantitative enough to discern accurately the number of target molecules hybridized to a particular probe. Thus, a sequence that is repeated in the target can easily be miscalled as a single copy. Fifth, with n-base probes, sequence repeats of length n-1 lead to ambiguities in the reconstruction of the full sequence. For 8-letter words these so-called branch point ambiguities limit effective reconstruction lengths to around 250. The same problems interfere with an alternate form of SBH in which an array of targets is built and interrogated by successive hybridization with single short probes or mixtures of probes.

A number of variations have been proposed to circumvent the difficulties encountered in early SBH studies (Figure 1).

One approach is to use much longer probes [8]. This allows end effects to be ignored or suppressed and also reduces the chance that an extreme in base composition will render the hybridization inexact or uninformative. Longer probes also compete better with secondary structures in the target. However, today it is not practical to make complete distributions of

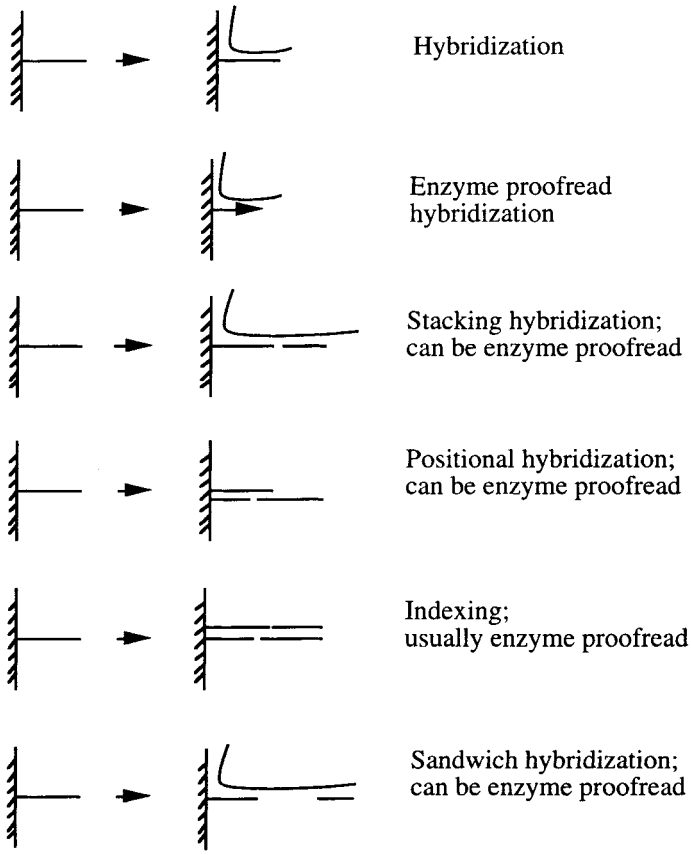

Figure 1. Modes of DNA Capture

all 416 16-mers. Hence this approach is limited to sequence checking or comparisons. Problems of quantitation remain. An alternative approach is positional sequencing by hybridization [9]. Here, duplex probes with single-stranded overhangs are used to capture the ends of complementary target fragments. Stacking hybridization or ligation can be used to ensure that the base-pairing at the end of the target is accurate, while DNA polymerase catalyzed extension of the 3' end of the probe using the bound target as a template can be used to ensure that the base-pairing of the end of the probe is accurate. Together these steps allow the full range of base compositions to be analyzed using a single experimental condition, and the additional information inherent in the knowledge that one is at the end of the target fragment helps overcome many branch point ambiguities [10].

Unfortunately, the PSBH protocols are complex, and the method has not yet been successfully used to analyze more than short model targets. However, the same protocols developed for PSBH form an excellent procedure to prepare DNA for sequencing by other methods [11-13]. In high throughput DNA sequencing making samples is often the major rate and cost limiting step. This is largely true because samples are typically made one at a time, that is, one per tube or microtitre plate well. However, with PSBH protocols a mixture of samples can be made in a single tube and then fractionated by collection onto surface-immobilized specific primers. This works very well for conventional gel electrophoretic DNA sequencing as well as for more novel approaches. A five-base overhang is quite effective; thus, only 1024 different primers would suffice for all DNA sequencing. In a modification of the PSBH procedure, a partially single-stranded probe can be used to capture the end of a duplex DNA fragment, containing a complementary overhang. Type IIS restriction enzymes which cut outside their recognition sequences can be used very conveniently to subdivide a target into an unique set of such fragments. DNA sequence is then read by nick translation or strand displacement by DNA polymerase in the presence of the four normal dpppN's and a trace amount of one dideoxy-pppN. This approach to DNA sequencing might be very effective for skimming where the goal is to obtain a single pass read over a non-redundant set of DNA fragments. It is a close analog of the method of DNA indexing that has been developed for capturing specific genomic DNA fragments [14], and has also found use for profiling gene expression [15, 16].

Other variations of SBH still remain to be developed more fully and evaluated. Among the attractive possibilities are using enzymes to proofread hybridization to conventional single-stranded DNA probes, and using stacking hybridization [5], in which a captured target is used for a second-step hybridization reaction with a probe adjacent to the initial duplex. This lengthens the read of the target sequence. In more complex schemes, a captured set of targets could be subjected to additional hybridizations anywhere along their sequences, and these additional hybridizations could also be enhanced by enzymatic steps. Other approaches include the use of enzymes to cleave mismatched target-probe complexes to enhance the accuracy of the sequence read [17]. All of these variations still suffer from one basic limitation - the read-out is usually a single intensity of hybridization at each probe location. This often results in a rather inefficient utilization of the total array of probes.

Mass spectrometry (MS) is another promising method for novel DNA analysis. Two types of instruments can access the mass range needed to look at DNA molecules up to 100 bases in length: time of flight (TOF, [18]) and fourier transform ion cyclotron resonance (FT, [19]). Two different methods have been successful in placing DNA into the vapor

phase without too severe fragmentation. In electrospray (ES), a thin stream of liquid containing the molecules of interest is shot into the vacuum chamber of an MS instrument in the presence of a high electrical field. This method combines nicely with fractionation tools like capillary electrophoresis. In matrix assisted laser desorption ionization (MALDI) the DNA samples are mixed with organic materials like picolinic acid which form a crystalline matrix as the solvent evaporates. The resulting surfaces are placed in the vacuum chamber of the MS and struck by a laser pulse which vaporizes small bits of the matrix, and DNA fragments are carried aloft along with matrix molecules. The different vaporization and detection methods can be combined to form alphabetic nightmares like MALDI-TOF MS, ES-FT MS and so on. ES typically produces large numbers of different charged species so the resulting spectra are quite complex. MALDI produces mostly singly-charged species, and MALDI-TOF is the method in most widespread use today.

Several groups have successfully sequenced short DNA fragments by performing traditional Sanger, solid state, or PSBH-based sample preparations and then analyzing the resulting mixture of fragments by MALDI-TOF MS. In essence, what is being done is electrophoresis in the vapor phase. In MS sequencing, artifacts like compressions and polymerase pauses do not interfere because secondary structures do not affect velocities in vacuum, and even if pauses occur, the resulting fragment masses are still indicative of the sequence. At present, sequences with a total length of up to about 80 bases are readable in this way [20, 21]. While this is a far cry from current electrophoretic sequence reads, it is much more informative than a single SBH result, and it can be carried out on large arrays of targets in a manner similar to the use of arrays in SBH. Unlike conventional sequencing, with current MS the length of the primer is a significant consideration since it subtracts from the total length of new sequence read. Thus, it will be worth exploring methods to shorten the primers or remove them before reading the sequence by fragment mass analysis.

The power of current MS in discriminating DNA fragment masses is impressive. In TOF, mass accuracies of 1 Da and resolutions of a thousand are achievable for fragments in the 30 to 60 base range. With FT the resolution and mass accuracies are a hundred times higher. This encourages the development of novel approaches which combine hybridization and mass measurement for extremely high throughput DNA analysis. For example, in most comparative or diagnostic DNA sequencing the question of real interest is whether a given target is identical to a previously known sequence or differs anywhere by a single base or more. Full sequencing is a most inefficient way to extract this meager bit of information. Various types of heteroduplex analyses are being developed to confront this problem [17,

22-25], but it is not yet clear how robust or automatable most of these will be. With MS, the problem is easily addressed by high resolution restriction mapping. The target can be cut to completion with a set of frequently-cutting nucleases and the resulting fragments examined as a mixture by MALDI MS. The smallest possible single base change, an A to a T, is 9 Da which is easily detectable in fragments of length 20 to 30. Deletions or insertions, which are usually more important in their pathophysiological consequences, are trivial to detect by MS.

MS also offers powerful approaches to analyses of complex DNA mixtures by indexing procedures. For example, in transcript imaging the goal is to develop a quantitative measure for the frequency of occurrence of all mRNAs in a sample. Two very different schemes for accomplishing such analyses efficiently have been proposed. Both first convert the sample to cDNA. Competitive hybridization to an array or probes is one approach [26]. This is essentially a variant of SBH. It requires PCR of the full cDNA library, and distortions due to this PCR are a concern. The other approach, called serial analysis of gene expression [SAGE, 27], captures unique fragments from each cDNA prior to amplification. These are each converted to single, short (10 to 15 base) identifiers which are ligated together prior to PCR to ensure more uniform amplification. The resulting tandem array of fragments is sequenced. MS would appear to offer an ideal way to combine the power of both of these methods. The short index sequence is an ideal target for ms. Instead of tandem ligation, it may be possible to amplify the short pieces ligated between known adapter sequences. Hybridization capture could be used to reveal information about sequences at one end of each index fragment, while direct mass analysis of the captured fragment or MS detection of a mass-labeled second probe ligated to the other end of the index fragment would provide additional sequence information. Pilot studies indicate the feasibility of this approach, but a full scale test has not yet been attempted.

In the example just described, and in other protocols that can be imagined, if MS is used as a hybridization detector, the effective single color limitation in conventional SBH is removed dramatically. The number of different masses for targets of length n increase as n^3. Thus, if one used methods to cut the target into discrete fragments prior to hybridization, so long as these fell within the mass range accessible to MS, most could be identified uniquely. Thus, analyses should be possible even if large numbers of different DNA targets hybridize to the same probe location. This added dimension in the analysis encourages considerations of degenerate probes to increase further the efficiency and throughput of the whole process. For example, binary probes (e.g. each position is R or Y; [28]) substantially decrease the size needed for a complete array of all n-mers. They also

lead to quite degenerate hybridizations. However, if the different targets are distinguishable by MS, the much smaller sample array might have the same sequencing power as a larger one.

If the cost of DNA sequencing can be reduced by several orders of magnitude, it will open the way for large scale diagnostic DNA sequencing, for the quantitative analysis of patterns of gene expression, and for thorough surveys of the diversity of mankind and the diversity of life on this planet. Inexpensive large scale DNA sequencing will also facilitate much more informative analysis of the environment and the effect of man's activities on our ecosystems. Such sequencing ability will also allow DNA to be used as a universal additive or bar code to track almost all manufactured products.

REFERENCES

1. Ambrose, W. P., Goodwin, P. M., Jett, J. H., Johson, M. E., et al. (1993). *Intl J. Phys. Chem. 1993*, 1535-1542.

2. Nikiforov, T. T., Rendle, R. B., Goelet, P., Rogers, Y. H., Kotewicz, M. L., Anderson, S., Trainor, G. L., and Knapp, M. R. (1994) *Nucl. Acids. Res. 22*, 4167-4175.

3. Drmanac, R., Drmanac, S., Labat, I., Crkvenjakov, R., Vicentic, A., and Gemmell, A. (1992) *Electrophoresis 13*, 566-573.

4. Strezoska, Z., Paunesku, T., Rodasavljevic, D., Labat, I., Drmanac, R., and Crkvenjakov, R. (1991) *Proc. Natl. Acad. Sci. USA 88*, 10089-10093.

5. Khrapko, K. R., Lysov, Y. P., Khorlin, A. A., Ivanov, I. B., Yershov, G. M., Vasilenko, S. K., Florentiev, V. L., and Mirzabekov, A. D. (1991) *J. DNA Sequencing Mapping 1*, 375-388.

6. Bains, W. (1991) *Genomics 11*, 294-301.

7. Southern, E. M., Maskos, U., and Elder, J. K. (1992) *Genomics 13*, 1008-1017.

8. Chee, M., Yang, R., Hubbell, E., Berno, A., Huang, X. C., Stern, D., Winkler, J., Lockhart, D. J., Morris, M. S., Fodor, S. P. A. (1996) *Science 274*, 610-614.

9. Broude, N. E., Sano, T., Smith, C. L., and Cantor, C. R. (1994) *Proc. Natl. Acad. Sci. USA 91*, 3072-3076.

10. Pevzner, P. A. (1994) *Computers Chem. 19*, 221-223.

11. Fu, D.-J., Broude, N. E., Köster, H., Smith, C. L., and Cantor, C. R. (1995) *Proc. Nat. Acad. Sci. USA 92*, 10162-10166.

12. Fu, D.-J., Broude, N. E., Köster, H., Smith, C. L., and Cantor, C. R. (1995) *Genetic Analysis (Biomolecular Engineering) 12*, 137-142.

13. Köster, H., Tang, K., Fu, D.-J., Braun, A., van den Boom, D., Smith, C.L., Cotter, R. J., and Cantor, C. R. (1996) *Nature Biotech. 14*, 1123-1128

14. Unrau, P. and Deugau, K. V. (1994) *Gene 145*, 163-169.

15. Kato, K. (1995) *Nucleic Acids Res. 23*, 3685-3690.

16. Kato, K. (1996) *Nucleic Acids Res. 24*, 394-395

17. Pottmeyer, S., and Kemper, B. (1992) *J. Mol. Biol. 223*, 607-615.

18. Tang, K., Fu, D.-J., Cotter, R. J., Cantor, C. R., and Köster, H. (1995) *Nuc. Acids. Res. 23*, 3126-3131.

19. Little, D. P., Thannhauser, T.W., and McLafferty, F. W. (1995) *Proc. Nat. Acad. Sci. USA 92*, 2318-2322.

20. Roskey, M. T., Juhasz, P., Smirnov, I. P., Takach, E. J., Martin, S. A., and Haff, L. A. (1996) *Proc. Nat. Acad. Sci. USA 93*, 4724-4729.

21. Köster, H., Tang, K., Fu, D.-J, Braun, A., van den Boom, D., Smith, C. L., Cotter, R. J., and Cantor, C. R. (1996) *Nature Biotech. 14*, 1123-1128.

22. Babon, J. J., Youil, R., and Cotton, R.G. (1995) *Nucl. Acids. Res. 23*, 5082-5084.

23. Youil, R., Kemper, B.W., and Cotton, R.G. (1995) *Proc. Nat. Acad. Sci. USA 92*, 87-91.

24. Cotton, R. G. H. (1991) *Methods Mol. Biol. 9*, 39-49.

25. Soto, D., and Sukamar, S. (1992) *PCR Meth. and Appl. 2*, 96-98.

26. Schena, M., Shalon, D., Davis, R. W., and Brown, P.O. (1995) *Science 270*, 467-70

27. Velculescu, V. E., Zhang, L., Vogelstein, B., and Kinzler, K. W. (1995) *Science 270*, 484-7.

28. Pevzner, P. A., and Lipshutz, R. J. (1994). *Lecture Notes in Computer Science 841*, 143-158.