

Enrichment of super-sized resequencing targets from the human genome

Maynard Olson

Methods relying on dense arrays of synthetic oligodeoxynucleotides to target specific subsets of the human genome may enable routine resequencing of all human exons or multi-megabase-pair chromosomal regions.

DNA sequencing should be thought of as a process rather than a method. Different steps in the process have been rate-limiting at different times during the past 30 years. Presently, at least for resequencing applications, the central bottleneck is acquisition of sequencing templates from targeted, multi-megabase-pair subsets of the billions of base pairs of DNA present in the human and other large genomes. Three papers in this issue of *Nature Methods* raise the hope that this bottleneck is about to disappear^{1–3}.

The first decade after the introduction of efficient methods of DNA sequencing in 1977 was the ‘gene-cloning’ phase of genomics. Acquisition of the DNA from genomic regions of particular biological interest was rate-limiting. This situation began to change during the next decade as a result of two parallel developments: the invention of PCR and the introduction of automated sequencing instruments based on the four-color-fluorescence implementation of Sanger’s dideoxy sequencing technique. These developments laid the groundwork for the *de novo* sequencing of reference genomes and also allowed, for the first time, the large-scale resequencing of short, targeted segments of genomes for mutation-detection and characterization of natural genetic variation. As technical

improvements ensued, the cost of PCR and automated sequencing remained in rough balance. Throughout this period, sequence-based studies of genetic variation flourished without any major changes in the experimental paradigm.

The introduction of a new generation of sequencing instruments has suddenly disrupted this balance. Three ‘sequencing-by-synthesis’ instruments have been commercialized during the past two years and resequencing-by-hybridization methods have also become widely available⁴. Costs per nucleotide of raw-sequencing data have fallen by two orders of magnitude from $\sim \$10^{-3}$ to $\sim \$10^{-5}$. Furthermore, the new sequencing-by-synthesis instruments are best suited to applications that involve acquiring a large amount of raw sequence per sample ($\sim 10^8$ nt), in comparison with capillary-based instruments ($\sim 10^3$ nt).

Impressive as these improvements are, presently available instruments are still inadequate to allow routine resequencing of whole human genomes, an application that requires $\sim 10^{11}$ nt of raw data to achieve adequate sampling redundancy. Hence, exploitation of the full potential of the current instruments will require the ability to target resequencing on a megabase-pair scale. Three papers in this issue of *Nature Methods* forcefully address the

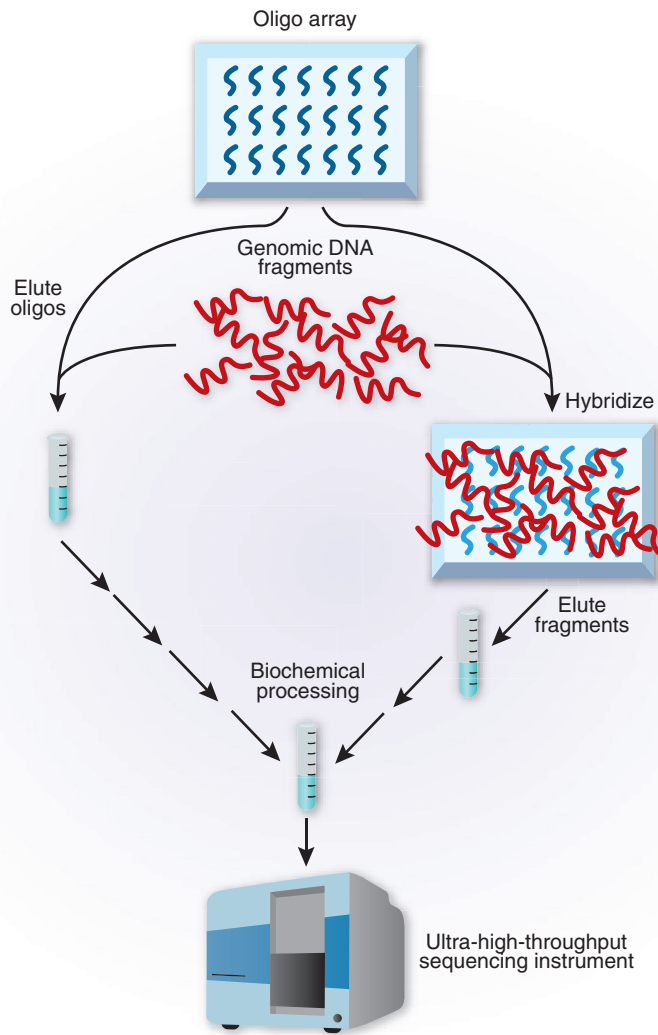
targeting problem with encouraging initial results (Fig. 1).

Porreca and colleagues use one custom-synthesized 100-mer as a capture probe for targeting each exon-sized segment (~ 100 bp) of the human genome by a technique that is related to PCR but far better suited to multiplexing. More important than the details of the protocol is its capacity: starting with a single commercially available array of 5.5×10^4 custom-designed oligodeoxynucleotides, they were able to target, in a single-tube reaction, 13% of the protein-coding sequences in the human genome (6.7 Mbp)¹. The main weakness in this study is that the success of the targeting was much less uniform than a random model would predict (that is, some targets were recovered hundreds of times whereas many others were missed altogether). Because of this nonuniformity, extremely deep sequencing of the pooled targets would be required to assess what proportion of targets can be recovered in experiments of realistic scale. The results are nonetheless impressive, particularly because the background of nontargeted sequences was extremely low.

Okou *et al.*² and Albert *et al.*³ describe success with protocols that are similar to each other but quite different than that presented in Porreca *et al.*¹ In the former two papers, targeting is achieved by direct hybridization capture of segments of the human genome onto commercially available arrays containing 3.9×10^4 custom-synthesized oligodeoxynucleotides ranging in length from 50 to 93 nt. Okou and colleagues’ most ambitious experiment targeted 304 kbp of human DNA², whereas Albert and colleagues targeted up to ~ 5 Mbp³.

Although all three teams of investigators clearly achieved their proof-of-principle goals, each of the protocols will require additional testing and refinement. At present, it is impossible to pick a winner because each paper presents a different set of success metrics. One limitation that they all have is that none can target completely arbitrary subsets of the human genome: the repeated sequences that make

Maynard Olson is at the University of Washington, Department of Genome Science and Medicine, Seattle, Washington 98195, USA.
e-mail: mvo@u.washington.edu



designed for highly multiplexed genotyping. The methods described in Okou *et al.*² and Albert *et al.*³ descend from a long lineage of hybridization-based capture systems that started before the development of recombinant-DNA techniques that largely put them on hold in the 1970s.

Driven by the same goals that motivated the work described in the papers in this issue, hybridization capture has enjoyed a recent revival. Important papers include one describing capture of small genomic fragments on BAC arrays⁷ and another describing haplotype-specific capture by short oligodeoxynucleotides of genomic fragments hundreds of kilobase pairs long⁸. Nonetheless, the papers in this issue of *Nature Methods* are the first that have the simplicity and apparent power to attract widespread adoption.

At least four applications can be readily envisioned: (i) scanning of all human exons for relevant variation in case-control studies of medically important phenotypes; (ii) routine scanning of megabase pair-sized candidate regions that have been implicated in disease phenotypes by linkage or association for potentially causal variants; (iii) scanning of hot spots of structural variation and of recurrent mutation for mutations that may contribute to behavioral, neurological or other phenotypes frequently associated with unstable genomic regions; and (iv) studies of genotype-phenotype correlations, based on complete resequencing of protein-coding regions, in the 'normal' human population. Collectively, these and other innovative applications of targeted resequencing of large segments of the human genome are likely to change the face of human genetics during the years immediately ahead.

1. Porreca, J.G. *et al. Nat. Methods* **4**, 931–936 (2007).
2. Okou, T.D. *et al. Nat. Methods* **4**, 907–909 (2007).
3. Albert, T.J. *et al. Nat. Methods* **4**, 903–905 (2007).
4. Shendure, J. *et al. Nat. Rev. Genet.* **5**, 335–344 (2004).
5. Nilsson, M. *et al. Science* **265**, 2085–2088 (1994).
6. Hardenbol, P. *et al. Nat. Biotechnol.* **21**, 673–678 (2003).
7. Bashirdes, S. *et al. Nat. Methods* **2**, 63–69 (2005).
8. Guo, Z. *et al. Proc. Natl. Acad. Sci. USA* **103**, 6964–6969 (2006).

Figure 1 | Two ways to use high-density arrays of custom-synthesized oligodeoxynucleotides to target subsets of the human genome. On the left, as in Porreca *et al.*¹, the oligos are washed off the array (dark blue array) and combined with sheared genomic DNA. Each oligo captures a particular genomic target in a form that can be readily converted to the input sample for an ultra-high-throughput sequencing instrument. On the right, as in Okou *et al.*² and Albert *et al.*³, targeted genomic fragments are captured directly on the array (light blue array) by DNA-DNA hybridization and then eluted and processed into a form suitable for sequencing.

up ~50% of the genome pose serious challenges for any targeting method that uses hybridization to capture short fragments of genomic DNA.

Nonetheless, these methods may be expected to pave the way for a huge variety of resequencing applications that focus primarily on single-copy DNA. The real test of whether any of them are ready for prime time will be the extent to which they prove themselves in actual applications.

For example, investigators will rapidly abandon PCR-based resequencing of large genes or sets of genes in search of mutations if and only if these methods deliver comparable false-positive and false-negative rates under realistic conditions of use.

None of these papers appear 'out of the blue'. The system presented in Porreca *et al.*¹ is in a lineage that started with the development of 'padlock probes'⁵ followed by 'molecular inversion probes'⁶ that were

Kim Caesar