

Sequencing by Cyclic Ligation and Cleavage (CycLiC) directly on a microarray captured template

Kalim U. Mir*, Hong Qi, Oleg Salata and Giuseppe Scozzafava

The Wellcome Trust Centre for Human Genetics, Oxford University, Roosevelt Drive, Oxford, OX3 7BN, UK

Received October 1, 2008; Revised October 25, 2008; Accepted October 28, 2008

ABSTRACT

Next generation sequencing methods that can be applied to both the resequencing of whole genomes and to the selective resequencing of specific parts of genomes are needed. We describe (i) a massively scalable biochemistry, Cyclical Ligation and Cleavage (CycLiC) for contiguous base sequencing and (ii) apply it directly to a template captured on a microarray. CycLiC uses four color-coded DNA/RNA chimeric oligonucleotide libraries (OL) to extend a primer, a base at a time, along a template. The cycles comprise the steps: (i) ligation of OLs, (ii) identification of extended base by label detection, and (iii) cleavage to remove label/terminator and undetermined bases. For proof-of-principle, we show that the method conforms to design and that we can read contiguous bases of sequence correctly from a template captured by hybridization from solution to a microarray probe. The method is amenable to massive scale-up, miniaturization and automation. Implementation on a microarray format offers the potential for both selection and sequencing of a large number of genomic regions on a single platform. Because the method uses commonly available reagents it can be developed further by a community of users.

INTRODUCTION

Methods that enable the high-throughput, low-cost sequencing of whole genomes or selected regions thereof will have wide impact across biology and medicine (1,2). Despite continuing efforts towards miniaturization (3,4), the electrophoresis-based implementation of the Sanger method cannot compete with the parallelism of surface-based platforms. Millions of sequencing reactions can be carried out simultaneously (5,6) on small areas on a surface. Massively parallel cyclic sequencing methods carried out on clonal templates generated via single-molecule

amplification on a bead or on a surface have been used successfully in the sequencing of bacterial genomes (5,7,8). Margulies *et al.* (7) used Pyrosequencing, a Sequencing-by-Synthesis (SBS) method (9), to perform sequencing in fiber-optic picolitre wells (454 sequencing) and compared to the Sanger method significantly reduced the time required to sequence a small bacterial genome; Shendure *et al.* (5) added ligation to a combinatorial sequence decoding strategy based on hybridization of partly degenerate oligonucleotides (10) and reduced sequencing cost 9-fold compared to the Sanger method; recently Pihak *et al.* (8) have made further reductions in time and cost using a Sequencing by Hybridization (SBH) scheme. Roche 454 sequencing has been used to sequence the genome of James Watson at a cost close to a million dollars, which is about 10 times lower than the cost of Sanger sequencing in large sequencing centers (11). Although commercial systems now claim to have brought the cost of human genome resequencing down to around \$100 000, innovation must continue towards sequencing human genomes for \$1000 or less. Until this goal is reached, because the current generation sequencing technologies do not include an integral means for selecting the molecules that are sequenced, their utility in human genetics is hindered. This is unless a separate technology is used first—adding significant cost and time to the process—for selecting sequences of interest.

In this article we describe Cyclical Ligation and Cleavage (CycLiC), a method which possesses elements of both SBS and SBL. CycLiC is used to demonstrate for the first time, steps of SBS/SBL on a template captured from solution on a microarray. With further development this 'direct to sequencing' microarray approach will enable a subset of molecules from different parts of the genome to be sequenced selectively or complete genomes to be sequenced systematically.

CycLiC uses oligonucleotide libraries (OL) in which all but one nucleotide is degenerate (5,10,12). The method involves iterative primer extension cycles, but instead of following chain elongation via the incorporation of nucleotides by DNA polymerase, the chain is grown base-by-base by successive ligation and detection steps

*To whom correspondence should be addressed. Tel: 01865 287652; Fax: 01865 287533; Email: kalim.mir@well.ox.ac.uk

using labeled oligonucleotides (ONs) followed by cleavage. The advantages of this approach are that, instead of having to develop highly modified and difficult to incorporate nucleotides, sequencing can be done with reagents that are already freely available. Moreover, with the availability of sufficient resolvable labels, chain extension could be done in sequence blocks. Although combinatorial encoding methods that would increase the repertoire of available labels are in development (13–15), sequence blocks of say, 5 and 6mer ONs would require a large number of resolvable fluorescent labels (1024 and 4096, respectively) which are not likely to be available for some time. Therefore we have developed a scheme in which the requirement for labels is simplified by coding for only one base in the ON length required for ligation and removing un-encoded nucleotides from the ON after ligation, before continuing with further synthesis. As technology becomes available for encoding/decoding progressively larger sets of ONs, the number of bases that are sequenced per cycle using this technology will incrementally increase.

The schematic (Figure 1) shows the CycLiC strategy: only one of the nucleotides in the OL is defined (and must bind specifically) while the remaining nucleotide positions are made degenerate, so that any sequence in the target can be addressed. As an alternative, the degenerate positions could be taken up by or combined with universal nucleotide analogues. Ligation reactions could extend in the 3' to 5' or 5' to 3' direction; we chose to demonstrate the latter, which is compatible with standard ON arrays. In the experiments reported here the label is at the 5' terminus and codes for the 3' terminal base from

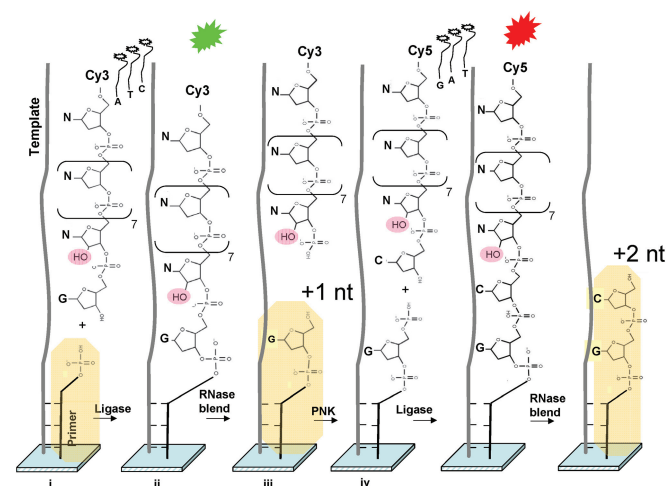


Figure 1. Four libraries of ONs are created, each with one defined terminal base, which is coded for by the label. After ligation of the primer with 'incoming' ON in a competitive reaction containing all four libraries (i), the incorporated degenerate ON is detected (ii). Then all but the first nucleotide occurring after the ligation junction are removed (iii). This shifts the register to the next base to be sequenced in the template. For the second cycle the same set of four libraries is added again to determine the identity of the next base; the primer is re-phosphorylated (iv) which allows ligation of the correct degenerate library from the mixture of all four to take place. The cycle is iterated to provide a contiguous sequence read. N denotes any base.

which it is separated by a string of degenerate nucleotide positions. The label also acts as terminator, to prevent more than one ON ligating along the chain during each reaction cycle.

To remove the un-encoded nucleotides from the ON, the approach requires a modification to be introduced into the ON sequence, namely a cleavable position. The modification must be compatible with ligation and, after cleavage, should provide a terminus that can be regenerated for further extension. In the experiments we present in this paper, of the degenerate positions, the first beyond the 3' terminal defined nucleotide is an RNA nucleotide, which is cleavable on its 3' side. We chose to focus on RNA as the cleavable linkage because it is readily available and a number of methods for its cleavage, that are compatible with further extension, are known (16–18). After cleavage a 5' OH is generated which requires addition of a 5' phosphate for further extension to proceed (reversing termination); this is readily achieved without additional reaction steps by supplementing the ligase reaction mix with Polynucleotide Kinase (PNK).

MATERIALS AND METHODS

Printing of arrays

Nexterion H slides from Schott Jena Glas (Jena, Germany) were printed using the Lucidea Spotter (Amersham Biosciences, Amersham, UK) at 55% relative humidity using the spotter's standard method with row and column pitch of 250 μ m. The spot diameter was \sim 140 μ m. The 3'-end aminated ONs were resuspended in Milli-Q water (Millipore, Billerica, MA, USA) and 2 \times oligo printing buffer (Full Moon Biosystems, Sunnyvale, CA, USA) to give a final ON concentration of 20 μ M in 1 \times ON printing buffer. After printing, the slides were placed for 2 h in a humid incubator set at 37°C followed by overnight storage under vacuum. The slides were blocked by submerging in blocking solution (50 mM ethanolamine/50 mM Tris pH 9.0) for 1 h at room temperature. The slides were then rinsed in Milli-Q water and dried ready for use.

Ligation/cleavage reactions on arrays

The surface of the microarray slide, containing surface attached primers, was conditioned by incubation with 10 μ l of 1 \times T4 ligase buffer (50 mM Tris-HCl, 10 mM MgCl₂, 10 mM DTT, 1 mM ATP, 25 μ g/ml BSA, pH 7.5, New England Biolabs, Beverly, MA, USA) at 37°C for 1 h with shaking at 300 rpm on a slide incubator (Thermomixer Comfort, Eppendorf, Hamburg, Germany). A 10 μ l ligation mix containing the following was prepared: 1 \times T4 ligase buffer, 20 pmol β -globin template, 1 μ l of 50% PEG 4000 (Fermentas, Vilnius, Lithuania), 1 μ l (400 U = 6 Weiss U) of T4 DNA ligase (NEB), 1 μ l (10 U) of T4 PNK kinase (NEB) and 50 pmol of 10mer degenerate ON library for each of the four colors (Cy3, Cy5, Fluorescein, Alexa594) (Supplementary Table 1). The mix was placed on the slide under a coverslip and incubated at 46°C for 1 h with acoustic wave mixing (ArrayBooster, Implen, Munich, Germany). The slide

was briefly dipped in wash 1 (0.2% SDS/1 × SSC), wash 2 (1 × SSC), wash 3 (0.1 × SSC) and wash 4 (Milli-Q water). All wash solutions were kept in a water bath set at 60°C. The slide was air dried and scanned on the ProscanArray 4-color scanner (Perkin Elmer, Waltham, MA, USA) and scans for each fluorophore were acquired.

The ligation products were cleaved by incubating the slide with 10 µl mix containing 1 × TED buffer (10 mM Tris pH 7.0, 5 mM EDTA, 2 mM DTT) and 2 µl RiboShredder RNase blend (Epicentre, Madison, WI, USA) at 37°C for 1 h. The slide was washed at 60°C using the same reagents and wash protocol described above.

For the second cycle of ligation new ligation mix containing all the reagents described above, minus the template, was prepared and placed onto the slide. After another round of incubation and washes the slide was scanned once more.

Solution-based ligation/cleavage cycles

The first ligation reaction contained 250 pmol of template, 250 pmol of 3' fluorescein labeled primer, 375 pmol of 5' end-labeled complementary incoming ON 1 in 1 × T4 buffer, 2.5 µl of T4 PNK kinase and 2.5 µl of T4 DNA ligase in a final volume of 50 µl. The mix was incubated at 46°C for 1 h. The T4 DNA ligase was inactivated by incubation at 80°C for 10 min.

First cleavage reaction: Every 5 µl aliquot from the previous reaction, equivalent to 25 pmol of ligation product in 1 × T4 ligase buffer, was incubated with 3 U of Riboshredder RNase blend at 37°C for 1 h in a final volume of 20 µl. The ribonuclease enzymes were then inactivated: each 20 µl reaction aliquot was treated with 0.8 µl Proteinase K (ICN Biomedicals, Irvine, CA, USA) at 5 µg/µl at 37°C for 1 h followed by heat inactivation of Proteinase K enzyme by incubation at 99°C for 10 min.

Second ligation: the materials from the previous reaction (25 pmol aliquots) were supplemented with 25 pmol of fresh template, 37.5 pmol of complementary incoming ON 2, 0.5 µl of T4 PNK kinase and 0.5 µl of T4 DNA ligase. The mix was incubated at 46°C for 1 h. The T4 DNA ligase was inactivated by incubation at 80°C for 10 min.

Second cleavage reaction: This was carried out in the same manner described for first cleavage reaction.

Third and fourth ligation: These were carried out using the same protocols for second ligation replacing ON 2 with ON 3 and ON 4, respectively. Incoming ON 1, 2 and 3 were template specific 8mers, whereas incoming ON 4 was a degenerate 10mer with the first base at the 3' end being a specific nucleotide, C in this case. Table 1 in Supplementary Data gives the full list of incoming ONs.

Third and fourth cleavage reaction: Same as first cleavage reaction.

Five picomoles aliquots were loaded onto a 12.6% polyacrylamide/5.9 M Urea/33% Formamide/0.84 × TBE gel and run on a PowerEase 500 electrophoresis system (Invitrogen, Carlsbad, CA, USA) and imaged on a UV transilluminator table (Multimage Light cabinet, Alpha Innotec, San Leandro, CA, USA).

RESULTS AND DISCUSSION

Microarray capture and iterative ligation and cleavage

We chose to exemplify the method by using the sequence of a region of human β -globin gene as template. We spotted a probe complementary to a specific location on the template onto a microarray slide and conducted a first ligation cycle using all four coded OLs. Stringent washing was required to remove non-ligated molecules which had hybridized to the template. Following this, detection on a microarray scanner showed that the correct base was incorporated in a large majority (Figure 2A) and that the magnitude of the signal of the second brightest channel was negligible by comparison (base calling ratio, defined as ratio of brightest to second brightest channel, was 24:1). After detection, cleavage was conducted to remove the degenerate bases and label/terminator (Supplementary Data 1). We then proceeded with second base addition, by applying to the array a mixture containing PNK and T4 DNA ligase but no template. This again resulted in dominant signal from the expected ON, albeit at a reduced absolute fluorescence value (dropping from 12830 AFU to 1622 AFU).

In the microarray experiment shown in Figure 2, although sufficient signal was obtained on the array for two sequencing cycles, the overall signal was too low to determine base three. Some loss of signal can be attributed to loss of primer which on average is about 8% per cycle (Supplementary Data 2). However, the major reason for the decrease in signal is loss of template. It was found that under the wash conditions required to remove bound but non-ligating ONs, substantial amounts of template were also able to dissociate from the 18mer primer. The drop in

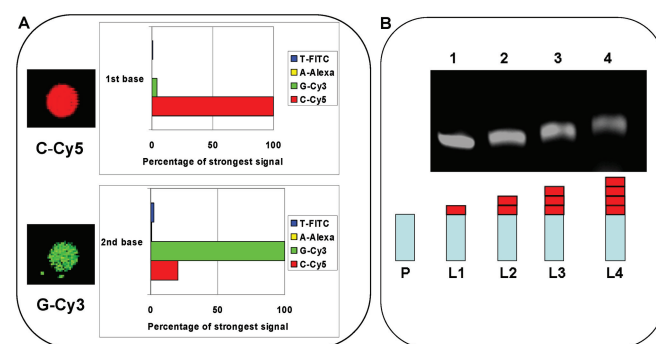


Figure 2. CycLiC Sequencing. (A) Top: Image of microarray spot (strongest signal) for first base addition and corresponding histogram showing the intensity of the four labeled degenerate ON libraries. Bottom: Image of microarray spot and histogram showing intensities for all four degenerate ON libraries, after cleavage and addition of the second base. NB data is normalized to strongest signal. Second base addition data is not corrected for template loss or for residual signal from cycle 1. (B) Polyacrylamide gel analysis of iterative cycles of ligation and cleavage carried out in solution. The first lane shows the product of first base ligation and cleavage. As the cycles of ligation and cleavage are repeated the size of the fragment increases base by base. The first three ligations were carried out with ONs specific to the sequence in the template while the fourth ligation was carried out with the OL corresponding to the complementary base in the template. The sketch below the gel illustrates the primer growing in length by one base at a time.

total fluorescence suggests an 8-fold loss of template going from first to second cycle. A Cy3-labeled template was used to assess the effects of wash steps on loss of template bound to surface-attached primer and corroborates this level of loss (Supplementary Data 8).

Experiments with a 40mer capture primer, where the primer-template duplex is longer and therefore more stable, gave sufficient signal to be able to determine a third base correctly (Supplementary Data 3). Stepwise yield appears to be lower when using a 40mer primer than an 18mer primer, but because more template is retained, a third base call can be made. It should be noted, that this experiment involved a chase reaction after the second cycle, to aid in completion of the second base reaction but the reaction times themselves were shorter due to the use of 'Quick' ligase which required just 20 min incubation.

Solution experiments show base by base primer extension

We investigated that the ligation-cleavage mechanism conforms to design such that the length of the primer increased one base per cycle. We demonstrated this by running extension products of stepwise ON ligation and cleavage cycles, carried out in solution, on a gel (Figure 2B). In contrast to the microarray experiments, washing could not be used to remove enzymes between steps and the original template remained in the reaction from cycle to cycle; moreover, additional template was added at each cycle. This is because where the microarray experiments were performed as a real sequencing reaction, the purpose of the solution experiments was specifically to investigate the extension of the primer through the cycles. A ladder of base-by-base extension products is clearly seen. The gel image shows that each length increment leaves no unextended bands (see also more sensitive exposure in Supplementary Data 4).

Cycle efficiency

The absence on the gel of any $n-1$ band below the extended ligation and cleavage product in each cycle suggests that the ligation efficiency must be very high in solution. This absence of out of phase signals suggests that in solution, reaction cycles must go close to completion. Although the first three cycles were conducted with high concentrations of specific ONs, the fourth cycle was conducted with a coded OL where the perfect match was at a much lower concentration; even here the primer appears to have completely extended. It should be noted that the ligation and cleavage product bands appear progressively weaker on the gel; we believe that this is primarily caused by hydrolysis due to repeated exposure to high temperature (which by contrast is not part of the microarray protocol).

What is the extent of completion in four-color sequencing cycles on the microarray? Base calling ratio between the first and second brightest channel decreased from 24:1 in cycle 1 to approximately 5:1 in cycle 2; the second brightest channel in cycle 2 coming from the label that gave the dominant signal in cycle 1. The reason for the increase in signal at cycle 2 from the second brightest channel can be attributed to three sources. First, a small

amount of the signal can be accounted for by mis-ligation. Secondly, some of the signal can be accounted for by primers molecules which did not extend at the first cycle but then went on to extend at the second cycle. These molecules are out of phase with the majority of molecules which had undergone extension at the first cycle. After a certain number of cycles (depending on the stepwise yield) the effects of this 'de-phasing' would be expected to build up until it is no longer possible to make a base call. Thirdly, a substantial amount of the signal can be accounted for by residual signal from molecules that did not undergo cleavage in the first round. Molecules may resist cleavage because they may have adhered to the surface in a conformation that hinders RNase action. When the second base signal has a substantial drop (as in our case due to template loss) the residual signal can represent a sizeable fraction of the second base signal.

In contrast to the solution experiments whose products can be analyzed on a gel, the presence of $n-1$ products cannot directly be visualized in a microarray format and therefore it is difficult to directly determine the stepwise yield. However, it can be estimated. When the level of signal in cycle 2 of the brightest channel in cycle 1, is corrected for residual signal, mis-ligation and the effect of loss of template from primer and loss of primer from the surface, an estimate for the extent of dephasing can be calculated. Supplementary Data 5 shows that by normalizing for these factors a stepwise yield of 99% is estimated for the ligation part of CycLiC. Solid-phase reactions are known not to be as efficient as their counterparts in solution but the choice of hydrogel substrate in our microarray experiments appears to provide solution-like yields for the ligation reaction. If the ligation reaction was the sole determinant of stepwise yield then 99% stepwise yield would support base-calling for up to around a 60-base read length. However, the yield of cleavage must also be taken into account. Taking an average 5% as the fraction of molecules that do not undergo cleavage (Supplementary Data 5), overall stepwise yield drops to around, 94% (95% of 99%), meaning that although 10 bases would still be readable, beyond this, the signal will begin to be overwhelmed by noise. It should be noted however, that it would not be necessary to achieve reaction completion at each cycle and dephasing would not be a problem if analysis was at the single molecule level. Single molecule detection would enable the asynchronous template-directed extension of each primer molecule within the spot to be monitored independently irrespective of its extent of extension; this would be followed by overlaying the single molecule reads to obtain a high quality consensus sequence (19); Harris *et al.* (20) have recently applied SBS at the single molecule level to sequence the M13 genome.

Specificity of ligation

We have tested fidelity of ligation at both sides of the ligation junction. Figure 2A shows that specificity is high on the incoming ON side (5') of the ligation junction. A microarray containing primers with all three terminal mismatches in addition to the correct match shows that

specificity is also high on the primer side (3') of the ligation junction (Supplementary Data 6). The CycLiC strategy takes advantage of this high specificity of DNA ligases at the ligation junction, requiring only the first base in the incoming ON to be a perfect match. It also takes advantage of the greater tolerance (21) to mismatch at sites distal to the ligation junction. In our microarray experiments, although the full complement to any given target sequence is a very small proportion of the total degenerate pool within a particular 10mer library, the strength of the fluorescent signal suggests that a significantly larger fraction of the degenerate pool than can be accounted for by the perfect match is able to ligate. The perfect match would be at a concentration of ~ 0.2 fmol within the degenerate pool (one specific, nine randomized bases equates to one in 262 144 of 50 pmol), which without some form of signal amplification would not be expected to give the level of fluorescent signals we obtain on the microarray. This suggests that substantial number of bases beyond the ligation junction must be tolerated despite being unable to form canonical Watson–Crick base pairs. This is a significant insight and confers to CycLiC a distinct advantage over other ligation-based methods (discussed below) which require mismatch discrimination to extend well beyond the ligation junction; the CycLiC strategy reads and therefore requires a perfect match only at the first base beyond the ligation junction, a point at which discrimination by DNA ligase between correct and incorrect base pairing is highly exacting.

Cycle times

Time-course data show that there is potential for the duration of ligation and cleavage reactions to be shortened (Supplementary Data 7). Using a commercial 'Quick' ligase (NEB) at room temperature we have successfully used a ligation time of 20 min (Supplementary Data 3); according to manufacturer information around 10–15 min may be possible too. Time-course analysis (Supplementary Data 7.) suggests that ligation with standard T4 DNA ligase could potentially also be shortened to around 20 min at room temperature because most of the reaction appears to be over by this point; however, this does not necessarily mean that the primers have all extended, just that relatively little further reaction is taking place. Time-course analysis for the cleavage reaction suggests that this step could be shortened to 1 min; although the caveat is that without further optimization a proportion of molecules may remain recalcitrant to cleavage and may drop out of subsequent cycles.

Further optimization

We are currently undertaking further development and optimization of the CycLiC biochemistry on templates captured from solution. Most importantly, loss of primer and template needs to be mitigated. A robust primer attachment chemistry has been described in the literature (22). Template loss may be reduced by lengthening primer-template duplex, as suggested by the results we present herein using 18mer and 40mer probes and/or, by modifying primers to form more stable duplex.

Another approach would be to lock the template and primer together; preliminary data show that template is preferentially retained after crosslinking (Supplementary Data 8). Also, when available in sufficient quantity the template could be replenished at each cycle. There are several parameters that should be addressed with regard to ligation efficiency. For example, although in our microarray experiments we chose to use 10mer coded OLs, because we reasoned that they may provide good stability, T4 DNA ligase supports ligation down to 5mers and the optimal ON length remains to be fully studied. Nucleotide modifications with favorable features, from the vast available toolkit (e.g. see www.glenres.com), should be explored; this includes the testing of universal bases and modifications that equalize AT with GC base pairs. The extent of cleavage needs to be brought to a consistent maximum level, so that close to 100% overall stepwise yields can be achieved; apart from optimization of the RNA cleavage system described herein, other cleavable systems that have been described in the literature could be adapted (23). Finally, analysis of a large set of loci in a single experiment on a microarray will allow the further optimization of ligation specificity and to determine any sequence specific effects on ligation efficiency.

Comparison with related sequencing methods

Landegren *et al.* (24) were the first to show ligation of two ONs for interrogating bases in template DNA. Ligation of degenerate or semi-degenerate OLs has been used in several sequencing systems described in the literature. Completely degenerate oligonucleotide libraries (DOLs) have been used in SBH schemes enhanced by ligation (25,26). In these methods the sequence is read only by the array ONs, the DOLs in solution serving only to facilitate SBH and label the reaction. Recently, Shendure *et al.* carried out cyclical sequencing by ligation (SBL) on microbead arrays using semi-degenerate OLs (5). CycLiC has features distinct from SBL which could offer potential advantages: Where SBL requires 36 coded ON libraries, CycLiC requires only four; CycLiC's sequence interrogation at the base proximal to the ligation site is likely to provide greater accuracy than SBL which requires discrimination up to six and seven bases beyond the ligation junction (21,27); Compared to SBL, CycLiC has the potential for sequence reads beyond the footprint of the OL; CycLiC is compatible with the microarray approach because in contrast to SBL it does not require stripping of primer from template at each cycle.

Two commercial systems add a cleavage step in the sequencing cycle but are rather complex compared to CycLiC. Although now commercially discontinued, Massively Parallel Signature Sequencing (MPSS) (28) was the first of the methods to involve ligation and cleavage. MPSS involved ligation of ONs comprising a sequence decoding element and a recognition sequence for a distal cleaving Type II restriction enzyme in a process whereby a double-stranded template is iteratively shortened as bases from the terminus of one strand are decoded. CycLiC is distinct from MPSS: it involves net synthesis rather than net degradation; cleavage occurs

within the ligated oligonucleotide rather than upstream of it; a recognition sequence for a Type II restriction enzyme does not need to be engineered into the decoding ONs. In parallel with our work, Applied Biosystems (Foster City, USA) have introduced a commercial sequencing system, Supported Oligonucleotide Ligation and Detection (SOLID) which shares the concept with CycLiC of net synthesis by ligation and cleavage and four colour probe sets (29). However, in SOLID a base call cannot be made directly from a single cycle but must be decoded from the gathered data set; this is due to a semi-ambiguous di-base encoding of the first and second nucleotide after the ligation junction (30). Cleavage in SOLID is between the fifth and sixth position; in CycLiC, ligation is equally effective whether cleavable location is between the first and second position or the fifth and sixth (data not shown). Because the site of cleavage is distal from the encoded di-base, primer annealing must be reset (existing primer removed and replaced with offset primer) before further reaction cycles can be carried out to determine the bases in between. By comparison, CycLiC is a more literal reading of the sequence; one cycle equals one base call and consecutive bases are directly determined. Hence there is no need to compile the sequence read algorithmically or to align sequences by using 'color-space' (30). The resetting in SOLID which enables a longer read-length than possible by their ligation-cleavage mechanism itself, can be matched in the microarray approach by tiling capture primers every fifth base in the reference sequence; as read-lengths improve the distance separating tiles can be increased.

As the label on the 5' end in CycLiC's OLs is able to terminate extension at each ligation step, the homopolymer problem which is a feature of Roche 454 sequencing is eliminated. Also the addition of all four bases simultaneously ensures fewer addition and wash steps per cycle and lessens the chance of mis-incorporation of bases compared to approaches where each base is added sequentially (7,20).

The case for sequencing on microarrays

While random array approaches (5,7,20,29) offer high density, they are not tailored to targeted resequencing of selected regions of the genome. By contrast, microarrays enable complex mixtures (e.g. mRNA populations) to be resolved by sorting of specific sequences by hybridization to unique spatial addresses on a surface. This allows microarrays to systematically address whole chromosomes (31) or to select multiple specific genomic regions of choice (32,33). However, to date, sequencing on microarrays has been limited to obtaining a single base or less of information per feature (31). In contrast, SBS on arrays has the potential to read 100s of bases of sequence from every array feature.

Where random arrays use the reference sequence to compile the sequence, a microarray approach uses the reference sequence at the outset to design a target specific array. Therefore, in order to be informative, the random array approach requires 20–50 bases (34) to anchor the sequence read to a location in the genome. On the other hand, microarray-based resequencing starts with a known

location and therefore the sequence read is informative right from the start; sequence assembly is not required, only base calling.

Ju *et al.* (35) have reported using a mutant 9°N DNA polymerase and reversible nucleotide terminators to sequence 13 bases of a template spotted on a microarray. The template, however, was self-priming and target capture was not demonstrated. When this or other SBS chemistries (36), including our own method is coupled to target capture on a microarray, as we have shown in principle here, the intrinsic scalability of microarray technology promises massive sequencing throughput. A 2007 GeneChip™ study employed 180 million probe features on a single whole wafer array (6). Moore's law, which predicts the rate of miniaturization achievable by the underlying photolithography technology, would suggest, that by 2009 we may have 360 million features. It is not then inconceivable, that on such an array a read as short as 10 bases could be applied to the systematic resequencing of a whole human genome of 3.2 billion bases.

Present next generation sequencing technologies, due to their random 'shotgun' arrangement were not developed with selective sequencing in mind. Therefore selection must be done before the sample is applied to the platform. This typically involves multi-step manipulations based around locus-specific PCR, microarray affinity selection (37,38) or microarray generated selectors used in solution (39,40). We have shown by contrast that the same array ON molecules can be used first as a probe to capture a template from solution and can then act as a primer to extend DNA bases cyclically. This suggests the possibility of selection and sequencing on a single platform; support for this contention comes from studies (41), which have shown that single base extension at ~500K genomic locations from primers attached to a surface can be done directly on whole genome amplified DNA, without the need for locus specific PCR. This suggests that direct capture and sequencing of generically amplified genomic DNA should be possible. This will make sequencing of selected genomic regions streamline and cost-effective. If selective sequencing can be applied seamlessly to a large number of individuals it would find wide application in human genetics, for example genomic regions pinpointed by high-density SNP scans of case and control populations (42) could be further investigated by selective sequencing in individuals to find causative mutations. Indeed, it would be possible to start genetic studies with the sequencing of all exons and control regions if not whole human genomes.

CONCLUSIONS

We have shown proof-of-principle of a hybrid SBS/SBL biochemistry called CycLiC. Gel experiments show that the ligation and cleavage mechanism is effective and can lead to high reaction completion. Notably, we have also shown that a microarray probe can be used to capture a template which can then be subjected to CycLiC to read up to three contiguous bases. A range of further optimization studies should be conducted, including how to retain the template-primer duplex in place after capture in order to extend the sequence read further. CycLiC on

microarrays uses 'generic' commercially available enzymes, common oligonucleotide modifications and standard microarrays and because the read will be from a targeted genomic location, sequence re-assembly algorithms will not be necessary. The method will therefore be amenable to adaptation and improvement by a community of users for sequencing as much or as little as desired from any genome for which a reference sequence is available.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We acknowledge Misha Shchepinov for intellectual input. We are grateful to Tim Watts for technical assistance and David L.V. Bauer for help with drawings. We also would like to thank Jiannis Ragoussis and Anthony Monaco for encouraging this work.

FUNDING

The research leading to these results has received funding from the following awards to K.U. Mir: Wellcome Trust [grant number WT072963/Z/03/Z]; Biotechnology and Biological Sciences Research Council (BBSRC) [grant number BB/E025013/1]; and the European Community's Seventh Framework Programme [FP7/2007-2013] under grant agreement number [HEALTH-F4-2008-201418] entitled READNA. Funding for open access charge: The Wellcome Trust; and the European Community's Seventh Framework Programme [grant number HEALTH-F4-2008-201418].

Conflict of interest statement. None declared.

REFERENCES

- Mir, K.U. and Southern, E.M. (2000) Sequence variation in genes and genomic DNA: methods for large-scale analysis. *Annu. Rev. Genomics Hum. Genet.*, **1**, 329–360.
- Collins, F.S., Green, E.D., Guttmacher, A.E. and Guyer, M.S. (2003) A vision for the future of genomics research. *Nature*, **422**, 835–847.
- Blazek, R.G., Kumaresan, P. and Mathies, R.A. (2006) Microfabricated bioprocessor for integrated nanoliter-scale Sanger DNA sequencing. *Proc. Natl Acad. Sci. USA*, **103**, 7240–7245.
- Fredlake, C.P., Hert, D.G., Mardis, E.R. and Barron, A.E. (2006) What is the future of electrophoresis in large-scale genomic sequencing? *Electrophoresis*, **27**, 3689–3702.
- Shendure, J., Porreca, G.J., Reppas, N.B., Lin, X., McCutcheon, J.P., Rosenbaum, A.M., Wang, M.D., Zhang, K., Mitra, R.D. and Church, G.M. (2005) Accurate multiplex polony sequencing of an evolved bacterial genome. *Science*, **309**, 1728–1732.
- Frazier, K.A., Eskin, E., Kang, H.M., Bogue, M.A., Hinds, D.A., Beilharz, E.J., Gupta, R.V., Montgomery, J., Morensoni, M.M., Nilsen, G.B. *et al.* (2007) A sequence-based variation map of 8.27 million SNPs in inbred mouse strains. *Nature*, **448**, 1050–1053.
- Margulies, M., Egholm, M., Altman, W.E., Attiya, S., Bader, J.S., Bembien, L.A., Berka, J., Braverman, M.S., Chen, Y.J., Chen, Z. *et al.* (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, **437**, 376–380.
- Pihlak, A., Baurén, G., Hersoug, E., Lönnerberg, P., Metsis, A. and Linnarsson, S. (2008) Rapid genome sequencing with short universal tiling probes. *Nat. Biotechnol.*, **26**, 676–684.
- Metzker, M.L. (2005) Emerging technologies in DNA sequencing. *Genome Res.*, **15**, 1767–1776.
- Epstein, J.R., Ferguson, J.A., Lee, K.H. and Walt, D.R. (2003) Combinatorial decoding: an approach for universal DNA array fabrication. *J. Am. Chem. Soc.*, **125**, 13753–13759.
- Wheeler, D.A., Srinivasan, M., Egholm, M., Shen, Y., Chen, L., McGuire, A., He, W., Chen, Y.J., Makhijani, V., Roth, G.T. *et al.* (2008) The complete genome of an individual by massively parallel DNA sequencing. *Nature*, **452**, 872–876.
- Ecker, D.J., Vickers, T.A., Hanecak, R., Driver, V. and Anderson, K. (1993) Rational screening of oligonucleotide combinatorial libraries for drug discovery. *Nucleic Acids Res.*, **21**, 1853–1856.
- Stavis, S.M., Edel, J.B., Li, Y.G., Samiee, K.T., Luo, D. and Craighead, H.G. (2005) Detection and identification of nucleic acid engineered fluorescent labels in submicrometre fluidic channels. *Nanotechnology*, **16**, S314–S323.
- Li, Y., Cu, Y.T. and Luo, D. (2005) Multiplexed detection of pathogen DNA with DNA-based fluorescence nanobarcodes. *Nat. Biotechnol.*, **23**, 885–889.
- Shchepinov, M.S., Chalk, R. and Southern, E.M. (1999) Trityl mass-tags for encoding in combinatorial oligonucleotide synthesis. *Nucleic Acids Symp. Ser.*, **42**, 107–108.
- Mauger, F., Jaunay, O., Chamblain, V., Reichert, F., Bauer, K., Gut, I.G. and Gelfand, D.H. (2006) SNP genotyping using alkali cleavage of RNA/DNA chimeras and MALDI time-of-flight mass spectrometry. *Nucleic Acids Res.*, **34**, e18.
- Tian, J., Gong, H., Sheng, N., Zhou, X., Gulari, E., Gao, X. and Church, G. (2004) Accurate multiplex gene synthesis from programmable DNA microchips. *Nature*, **432**, 1050–1054.
- Lee, H.J., Wark, A.W., Li, Y. and Corn, R.M. (2005) Fabricating RNA microarrays with RNA-DNA surface ligation chemistry. *Anal. Chem.*, **77**, 7832–7837.
- Mir, K.U. (2006) Ultrasensitive RNA profiling: counting single molecules on microarrays. *Genome Res.*, **16**, 1195–1197.
- Harris, T.D., Buzby, P.R., Babcock, H., Beer, E., Bowers, J., Braslavsky, I., Causey, M., Colonell, J., Dimeo, J., Efcavitch, J.W. *et al.* (2008) Single-molecule DNA sequencing of a viral genome. *Science*, **320**, 106–109.
- Deng, J.Y., Zhang, X.E., Mang, Y., Zhang, Z.P., Zhou, Y.F., Liu, Q., Lu, H.B. and Fu, Z.J. (2004) Oligonucleotide ligation assay-based DNA chip for multiplex detection of single nucleotide polymorphism. *Biosens. Bioelectron.*, **19**, 1277–1283.
- Fedurco, M., Romieu, A., Williams, S., Lawrence, I. and Turcatti, G. (2006) BTA, a novel reagent for DNA attachment on glass and efficient generation of solid-phase amplified DNA colonies. *Nucleic Acids Res.*, **34**, e22.
- Shchepinov, M.S., Denissenko, M.F., Smylie, K.J., Wörl, R.J., Leppin, A.L., Cantor, C.R. and Rodi, C.P. (2001) Matrix-induced fragmentation of P3'-N5' phosphoramidate-containing DNA: high-throughput MALDI-TOF analysis of genomic sequence polymorphisms. *Nucleic Acids Res.*, **29**, 3864–72.
- Landegren, U., Kaiser, R., Sanders, J. and Hood, L. (1988) A ligase mediated gene detection technique. *Science*, **241**, 1077–80.
- Cowie, S., Drmanac, S., Swanson, D., Delgrosso, K., Huang, S., du Sart, D., Drmanac, R., Surrey, S. and Fortina, P. (2004) Identification of APC gene mutations in colorectal cancer using universal microarray-based combinatorial sequencing-by-hybridization. *Hum. Mutat.*, **24**, 261–271.
- Gunderson, K.L., Huang, X.C., Morris, M.S., Lipshutz, R.J., Lockhart, D.J. and Chee, M.S. (1998) Mutation detection by ligation to complete n-mer DNA arrays. *Genome Res.*, **8**, 1142–53.
- Housby, J.N. and Southern, E.M. (1998) Fidelity of DNA ligation: a novel experimental approach based on the polymerisation of libraries of oligonucleotides. *Nucleic Acids Res.*, **26**, 4259–4266.
- Brenner, S., Johnson, M., Bridgham, J., Golda, G., Lloyd, D.H., Johnson, D., Luo, S., McCurdy, S., Foy, M., Ewan, M. *et al.* (2000) Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. *Nat. Biotechnol.*, **18**, 630–634.
- Valouev, A., Ichikawa, J., Tonthat, T., Stuart, J., Ranade, S., Peckham, H., Zeng, K., Malek, J.A., Costa, G., McKernan, K. *et al.* (2008) A high-resolution, nucleosome position map of *C. elegans*

- reveals a lack of universal sequence-dictated positioning. *Genome Res.*, **18**, 1051–63.
30. Holt, R.A. and Jones, S.J. (2008) The new paradigm of flow cell sequencing. *Genome Res.*, **18**, 839–46.
31. Patil, N., Berno, A.J., Hinds, D.A., Barrett, W.A., Doshi, J.M., Hacker, C.R., Kautzer, C.R., Lee, D.H., Marjoribanks, C., McDonough, D.P. *et al.* (2001) Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. *Science*, **294**, 1719–1723.
32. Matsuzaki, H., Dong, S., Loi, H., Di, X., Liu, G., Hubbell, E., Law, J., Berntsen, T., Chadha, M., Hui, H. *et al.* (2004) Genotyping over 100,000 SNPs on a pair of oligonucleotide arrays. *Nat. Methods*, **1**, 109–111.
33. Wang, D.G., Fan, J.B., Siao, C.J., Berno, A., Young, P., Sapolsky, R., Ghandour, G., Perkins, N., Winchester, E., Spencer, J. *et al.* (1998) Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. *Science*, **280**, 1077–1082.
34. Whiteford, N., Haslam, N., Weber, G., Prugel-Bennett, A., Essex, J.W., Roach, P.L., Bradley, M. and Neylon, C. (2005) An analysis of the feasibility of short read sequencing. *Nucleic Acid Res.*, **33**, e171.
35. Ju, J., Kim, D.H., Bi, L., Meng, Q., Bai, X., Li, Z., Li, X., Marma, M.S., Shi, S., Wu, J. *et al.* (2006) Four-color DNA sequencing by synthesis using cleavable fluorescent nucleotide reversible terminators. *Proc. Natl Acad. Sci. USA*, **103**, 19635–19640.
36. Turcatti, G., Romieu, A., Fedurco, M. and Tairi, A.P. (2008) A new class of cleavable fluorescent nucleotides: synthesis and optimization as reversible terminators for DNA sequencing by synthesis. *Nucleic Acids Res.*, **36**, e25.
37. Albert, T.J., Molla, M.N., Muzny, D.M., Nazareth, L., Wheeler, D., Song, X., Richmond, T.A., Middle, C.M., Rodesch, M.J., Packard, C.J. *et al.* (2007) Direct selection of human genomic loci by microarray hybridization. *Nat. Methods*, **4**, 903–905.
38. Hodges, E., Xuan, Z., Balija, V., Kramer, M., Molla, M.N., Smith, S.W., Middle, C.M., Rodesch, M.J., Albert, T.J., Hannon, G.J. *et al.* (2007) Genome-wide in situ exon capture for selective resequencing. *Nat. Genet.*, **39**, 1522–7.
39. Dahl, F., Gullberg, M., Stenberg, J., Landegren, U. and Nilsson, M. (2005) Multiplex amplification enabled by selective circularization of genomic DNA fragments. *Nucleic Acid Res.*, **33**, e71.
40. Porreca, G.J., Zhang, K., Li, J.B., Xie, B., Austin, D., Vassallo, S.L., LeProust, E.M., Peck, B.J., Emig, C.J., Dahl, F. *et al.* (2007) Multiplex amplification of large sets of human exons. *Nat. Methods*, **4**, 931–936.
41. Steemers, F.J., Chang, W., Lee, G., Barker, D.L., Shen, R. and Gunderson, K.L. (2006) Whole-genome genotyping with the single-base extension assay. *Nat. Methods*, **3**, 31–33.
42. Wellcome Trust Case Control Consortium. (2007) Genome-wide association study of 14 000 cases of seven common diseases and 3000 shared controls. *Nature*, **447**, 661–678.