# 15

# Sequencing by Aligning Mutated DNA Fragments (SAM)

## *A New Approach to Conventional Sequencing and SBH Sequencing*

Duncan Cochran[1,2,3], Gita Lala[1,2,3], Jonathan Keith[2,3], Peter Adams[2,3], Darryn Bryant[2,3], and Keith Mitchelson[1,3]
*[1]Australian Genome Research Facility, St Lucia 4072, Australia; [2]Department of Mathematics, University of Queensland, St Lucia 4072, Australia; [3]Combinomics Pty Ltd., St Lucia 4072, Australia*

**Abstract:** This paper discusses an original technique, *Sequencing by Aligning Mutants* (*SAM*), the purpose of which is to overcome sequencing difficulties caused by problematic genomic regions where local sequence characteristics hinder existing sequencing technologies. It involves forming a number of mutated copies of regions of the target DNA, cloning and sequencing each of these mutated fragments. The effectiveness of SAM technology is demonstrated by sequencing of problematic DNA elements. An application of SAM technology to chip-based sequencing-by-hybridization (SBH) is discussed.

**Key words:** Unclonable and problematic DNA, sequencing, DNA mutation, SBH.

## 1. Genome Sequence Assembly

All genome sequencing projects regularly encounter regions that yield no data with current sequencing strategies[1,2,3]. These gaps are present for several reasons including under-representation of sequences in libraries[4,5], the inability to assemble complex sequence regions correctly[6,7] and the presence of DNA motifs that are intractable to cycle sequencing[4]. The human genome project provides many examples of each of these different impediments to sequencing and serves to illustrate the measures that may be taken to overcome them.

### *1.1 Assembly of the Human Genome*

The haploid human genome consists of a total of 3.1 Giga base pairs (Gb). The International Human Genome Project (IHGP) commenced in 1994 as collaboration between medical research agencies, genome institutes and

laboratories to sequence the genome. Current achievements and progress[8,9] in this very large sequencing project illustrate the practical difficulty in determining all sequences completely within a genome, despite the enormous resources devoted to it.

The IHGP and other genome programs have established an international quality standard for "finished genomic sequence" to enable the comparison of genomic data from different programs and organisms, and also to define the standard that all genomic sequencing projects should seek to attain. "Finished sequence" is properly defined as the "complete sequence of a clone or genome, with an accuracy of at least 99.99% and no gaps"[10]. A more practical definition is that of "essentially finished sequence", meaning the complete sequence of a clone or genome, with an accuracy of at least 99.99% and no gaps, except those that cannot be closed by any current method. The standards also provide an "end point" for the completeness of genome data with any gap to be smaller than 150 kb and with more than 95% of the *euchromatic* regions as finished sequence. Sequence data that fall short of that benchmark but which can be positioned along the physical map of the chromosomes are termed 'draft'.

Following the publication of the initial drafts of the sequence[1,2] in 2001 in which approximately 80% was sequenced, the project moved deliberately towards the finishing process. Currently, the work is still a mosaic of finished and draft sequence. About 87% of the total genome is finished sequenced[9] and less than 13% is at the draft stage, although the gene-rich *euchromatic* parts of the genome (comprising approximately 2.95 Gb) are about 94% complete "finished" sequence, achieving the IHGP standard. In contrast, *heterochromatic* parts of the genome contain sequences that are more recalcitrant to current sequencing technologies. Additionally, the assembly of many large sequenced repetitive regions is incomplete, as algorithms cannot provide finished sequence at an acceptable standard. Consequently, many *heterochromatic* parts of the human genome must be considered as draft.

## *1.2 Shotgun Sequencing Strategies*

Several strategies for shotgun sequencing, such as "hierarchical" or "map-based shotgun sequencing", "whole-genome shotgun" sequencing and hybrid approaches[11], are still being evaluated for multimegabase- or gigabase-sized genomes of metazoans. The IHGP commenced as a map-based shotgun program, but recently has introduced some hybrid approaches. Perceptively, Waterston et al[5] note that a "great challenge arises in tackling complex genomes with a large proportion of repeat sequences that can give rise to misassembly". They argue that without reference to the IHGP builds, whole genome shotgun assembly methods pioneered by Celera would fail to assemble much larger regions containing repeated DNA motifs and that a whole genome sequencing approach alone would produce fewer assembled regions and more or larger gaps. This assertion highlights a problem common in the

assembly of repeated DNA regions, whether using shotgun data or hierarchical shotgun data – that repeated regions possess few "landmarks" that allow definition of the correct order of sub-repeats within the repeated DNA elements.

## *1.3 Gaps in the IHGP Working Draft Sequence*

To illustrate the regions of unfinished sequence of the human genome, Aach et al[4] compared the "gaps" in sequence data and the "gaps" between assembled finished sequence obtained by the IHGP[1] using "hierarchical shotgun sequencing" and Celera's[2] approach using "whole genome shotgun" sequencing. This comparison is shown in Figure 15.1.

Notably, whole genome shotgun sequencing produces many small assembled regions, punctuated by very many gaps frequently several Mb in size.

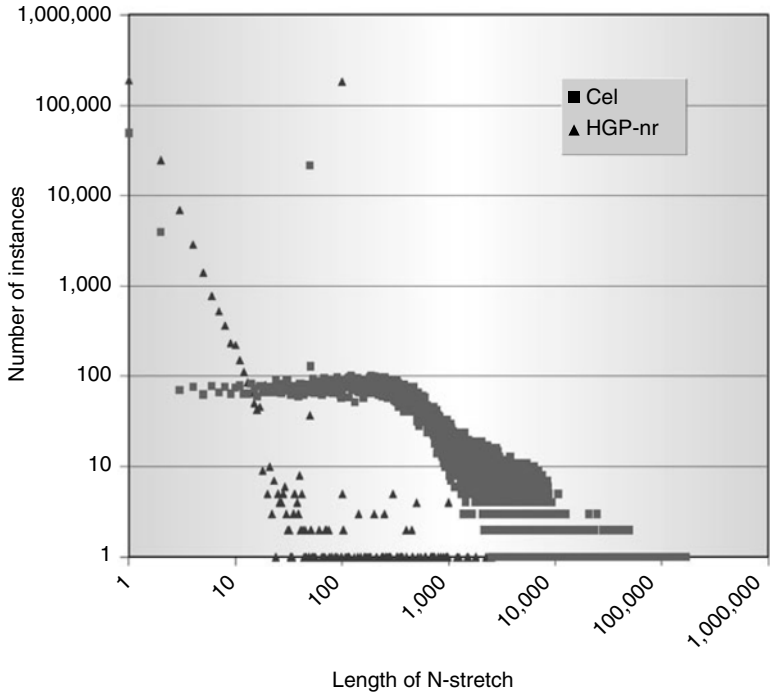


FIGURE 15.1. Size of continuous strings of Ns in the Celera (Cel) and Human Genome Project (HGP-nr) non-redundant genome assemblies. Long strings of Ns are used to represent gaps, but do not always represent gap size. The Celera assembly contained 169,779 stretches of Ns ranging in length from 1 to 168,735. The HGP-nr assembly contained 407,686 stretches of Ns ranging in length from 1 to 2,500. Reprinted with permission from Aach et al (2001) Nature 409, 856.

The hierarchical-shotgun sequencing approach produces much larger assembled regions, punctuated by numerous smaller gaps. These gaps appear in all sequenced regions of the human genome – the *euchromatic* parts, as well as in the *heterochromatic* parts. The gaps that occur in the IHGP working draft sequence fall into several general classes, which are:

i. Gaps within the sequence of ordered clones. Such gaps are mostly small (<200 bp) and represent unclonable regions or unsequenceable motifs within the sequencing libraries.
ii. Gaps between ordered clones and contigs. These are frequently larger gaps up to 0.3 Mb in size. The gaps again represent unclonable or under-represented regions that have not reached quality standards. In addition, the gaps may represent non-reconstructable regions, frequently highly repeated motifs, which despite being able to be sequenced cannot be reconstructed to acceptable quality standards.
iii. Gaps in large repeat regions. Whilst the primary sequence of large repeated regions may be able to be determined, the scale and subtle variant forms of the repeated sub-repeat motifs often prevents accurate assembly and ordering of these larger regions. This problem has been noted particularly for the assembly of megabase long repeated regions, such as telomere and centromere regions[6,12] and their flanking sub-repeated regions[13].
iv. Gaps may also be due to (almost identical) segmental duplications. The detection and correction of errors in assembly of segmental duplications will certainly be of increasing importance for key gene families and regions important for human medicine.

## *1.4 Segmental Duplications*

Eichler[6] and Bailey et al[7] considered large recent duplication events that fell well-below levels of draft sequencing error. Duplications of genomic regions (alignments 90%-98% similar and greater than or equal to 1 kb in length) comprise more than 3.6% of all human sequence. These duplications show clustering within the genome, and have up to 10-fold enrichment within peri-centromeric and sub-telomeric regions – regions of high mobility and sequence rearrangement. Duplicated sequences were found to be over-represented in unordered and unassigned contigs, indicating that duplications are difficult to assign to their correct position, even in recent assembly builds of the human genome. As shown in Figure 15.2, the under-representation or mis-assembly of duplicated sequences are also likely to be a major source of undetected error in current genome assemblies.

Correction of such errors will emerge from re-sequencing projects, from comparison with homologous genomic regions from closely related organisms, and through detailed re-examination of current genomic sequencing data[14].

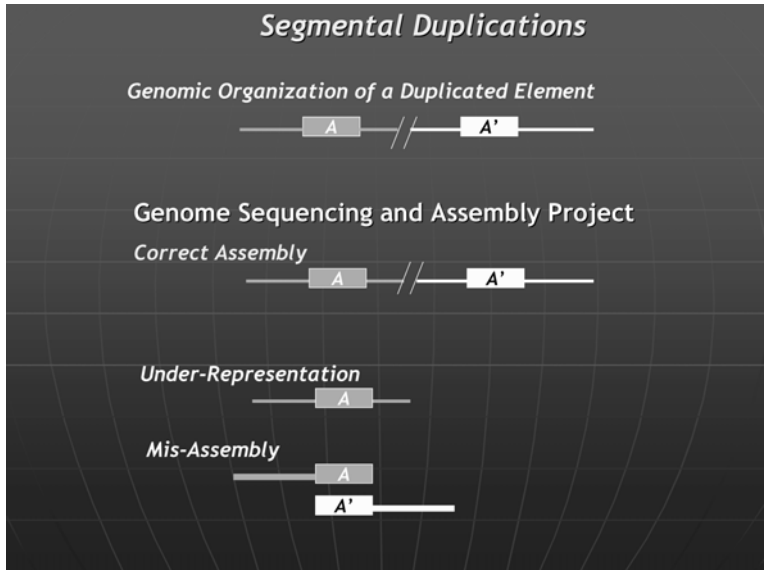


FIGURE 15.2. Errors in the identification and assembly of repeated regions – "segmental duplications" are likely to be a major source of undetected error in current genome assemblies. Correction of such errors will emerge from re-sequencing projects, from comparison with homologous genomic regions from related organisms and through detailed re-examination of current genomic sequencing data and assemblies. Reprinted with permission from Eichler (2001) *Genome Res*earch **11**, 653-656.

## 2. Sequencing by Aligning Mutants

This paper discusses a new and original approach to sequencing of difficult and repeated motifs, *"Sequencing by Aligning Mutants"* (*SAM*)[15-18]. As shown in Figure 15.3, SAM involves forming a number of randomly mutated copies of regions of the target DNA, then several of the mutated fragments are cloned and sequenced. The mutants should ideally be sufficiently altered that they no longer possess the characteristics that caused sequencing difficulties in the target, so existing sequencing techniques are applied more easily and successfully. However, the mutants must also be sufficiently similar to the target such that the original sequence can be inferred by analyzing the mutated sequences.

By aligning the sequenced mutated fragments to each other and to available pieces of the target sequence, mutation sites are identified and corrected for, enabling previously difficult-to-sequence regions of the target to be determined. The information content of the original sequence is not lost, but is merely distributed amongst multiple fragments. The randomness of the mutations is important for another reason: it enables sequence information lost in one mutant to be retained in most of the others.

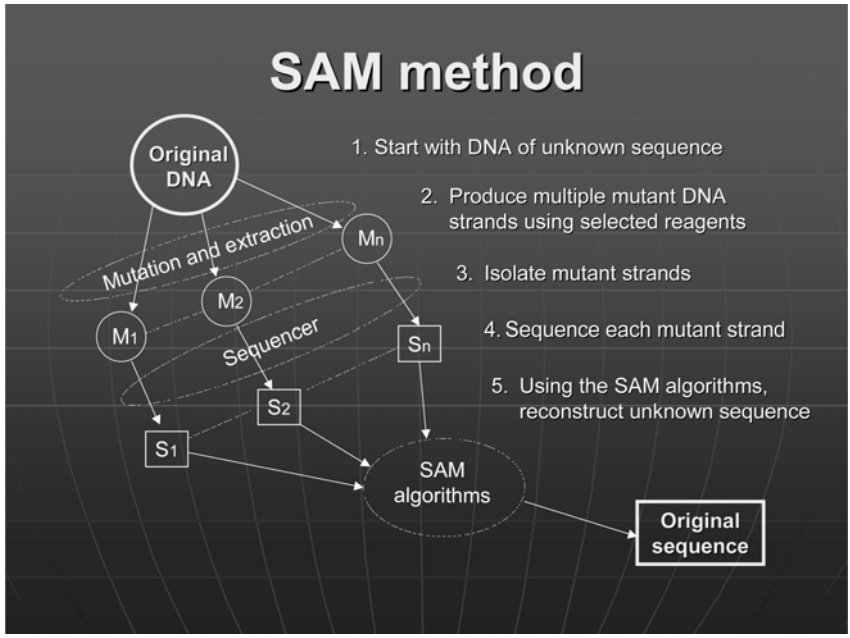


FIGURE 15.3. Sequencing aided by mutation (*SAM*) is a counter-intuitive overall approach to sequencing. Problematic motifs and regions are mutated sufficiently to overcome the obstructive cause to its cloning or sequencing. Random mutations make different altered copies of the region. The sequences determined from a low number of altered copies can be used to reconstruct the original wild-type sequence.

## *2.1 Local Sequencing Problems*

DNA sequence reads are frequently impaired by small motifs that may form secondary structures and other structures that impede the extension of DNA by processive polymerases[19,20]. Such recalcitrant motifs are often AT- or GC-rich repetitive regions that have either high thermal stability or other non-standard characteristics not found in mixed sequence DNA. Yet other small regions that cause sequence gaps may be unclonable or unstable in bacterial cells[3,21]. *SAM techniques* provide a novel method for obtaining sequence data from these intractable regions and could potentially enable genomic researchers to close the gaps present in the genome sequence maps.

Applying *SAM* with a sufficiently high mutation rate can potentially modify inverted repeats and prevent the formation of stem-and-loop structures by disrupting inverted repeat sequences. More generally, mutation can reduce repetition, raise or lower GC content and modify sequences that interfere with cloning host or vector functions or which inhibit DNA manipulations, rendering the recalcitrant fragment amenable to cloning and sequencing. Our calculations[17] indicate that with a substitution rate of 10%, it should be

possible to reconstruct the original sequence with an accuracy of less than one error per 10,000 bases using a small number of mutated fragments.

## 2.2 Base Content

Human chromosomes display wide variation in their regional nucleotide composition that in part reflects the classical *euchromatic* and *heterochromatic* regions. For example, each human chromosome has large swings in GC content:

- one stretch might have as much as 60 percent GC,
- while an adjacent stretch might have only 30 percent GC
- different GC content regions may cause sequencing difficulties for current sequencing technologies and approaches.

Some other organisms have even more extreme nucleotide bias than man. For example, the genome of *Dictylostelium discoideum* is approximately 70% AT, although its chromosomes also display local regions with still higher AT-rich motifs, as well as lower AT content[3]. DNA regions with extreme or high GC or AT composition may also cause sequencing difficulties, causing sequencing enzymes to fall off the DNA template prematurely, or to slip and jump bases[21].

Again, these motifs may interference with either cloning vector or cloning host functions, reducing the representation of the motif within cloned libraries. Indeed, efforts to sequence genomes with extreme AT or CG content are often confounded by a high level of failure to sequence clonable regions, as well as difficulties in gaining library representation of large portions of the genome[3]. Underrepresented regions create gaps, and as they are undefined may in some cases be difficult to isolate even with directed efforts to identify them. Potentially, the introduction of random mutations into whole genomes or into large fractions of genomes could lead to elimination some problem motifs, resulting in increased clonability and larger library representation of variant forms of these recalcitrant sequences.

# 3. Dye Terminator Cycle Sequencing

## 3.1 Repeat Sequences

Many patterns of bases create difficulties for *Taq* DNA polymerase cycle sequencing and DNA amplification. These patterns may be small, extending over some 30-300 bp, or may be larger, extending over 1 kb or up to many kilobases in length. Homopolymer tracts are one example of recalcitrant local motif (see Figure 15.4). Other larger repeated motifs such as microsatellite repeats (~ 3-6 bp repeat motif)[21], LINE and SINE elements[22] may also present barriers to sequencing. In some cases the physical order of the motifs

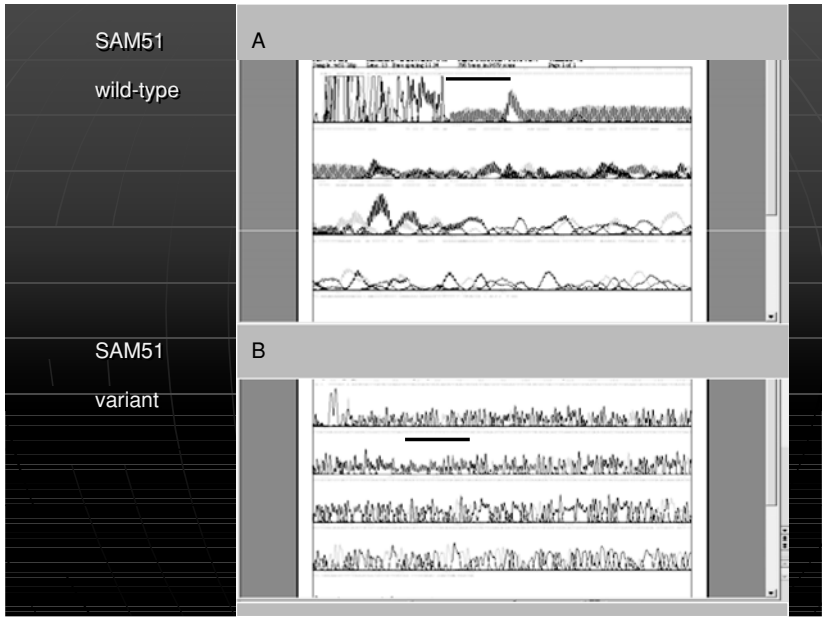


FIGURE 15.4. Comparison of ABI version 2.0 Big-Dye terminator sequencing of wild-type and mutant variant sequences. Bar indicates the problematic motif. A: the wild-type clone (SAM51) contains a homopolymer motif that prevents dye-terminator cycle sequencing. B: Introducing random substitution mutations can reduce the uniformity of the problem motif. One mutated variant of SAM51 is able to be readily sequenced using conventional sequencing technology.

can cause problems for DNA stability and problems for sequencing enzymes. For example, direct repeats may cause problems with PCR, causing primers to mis-prime at multiple sites. Inverted repeats cause hairpins and other complex DNA structures which may be incompatible with plasmid stability or sequencing systems[19-21].

## *3.2 Improved Sequencing of Simple-Repeats Using SAM Techniques*

Figures 15.4 and 15.5 illustrate the use of SAM technologies: homopolymer A tracts that cause difficulties for commercial sequencing kits were used as test molecules.

Figure 15.4A shows the wild-type element in which the sequence could not be determined, with harmonic stutter caused by polymerase slippage and restarts obliterating the usual chromatograph pattern. Following mutation of the region, the sequence of a representative variant shown in Figure 15.4B could be read directly using the same commercial sequencing kit. The

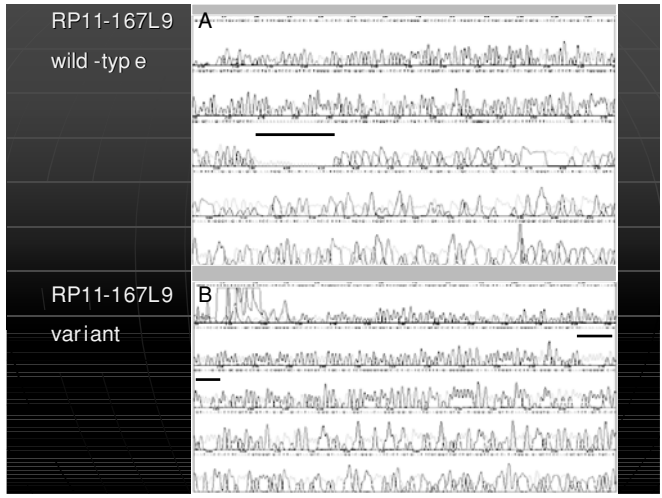


FIGURE 15.5. Comparison of ABI version 2.0 Big-Dye terminator sequencing of wild-type human BAC RP11-167L9, and mutated variants. Bar indicates a problematic polyA motif. A: a region of the wild-type clone contains a problematic motif that prevents dye-terminator cycle sequencing. B: one mutated variant of RP11-167L9 is able to be readily sequenced using conventional technology.

introduction of random substitution mutations reduced the uniformity of the problem motif and allowed *Taq* polymerase to extend through the region. The elimination of stutter bands created a well-defined chromatogram with uniform peaks and good peak separations.

Figure 15.5A displays another wild-type simple repeat region, the sequence trace through the repeat motif is poor, with a weak signal causing the miscalling of some bases and a miscalling of the unit size of the repeat. Although the chromatogram is readable beyond the repeat motif, peaks are broad and potentially miscalling could occur here. Following the mutation of the region, the sequence of a representative variant shown in Figure 15.5B was strong and could be read directly using the same commercial sequencing kit. Again, the introduction of random mutations reduced the uniformity of the motif and allowed *Taq* polymerase to extend through the region and beyond with uniform peaks and good peak separations.

Figure 15.6 shows a Clustal W alignment of sequences from three mutated variant clones of the target region illustrated in *Figure 15.5*, along with the wild-type element read (*polyAwt03_2*) and the published *estimation* of an intractable polyA tract (*target.txt*) of ~36 A residues within the wild-type target. Analysis of fewer than 10 variant sequence reads using *SAM algorithms* recovered the correct sequence and determined the correct size of the repeat as 22 residues.

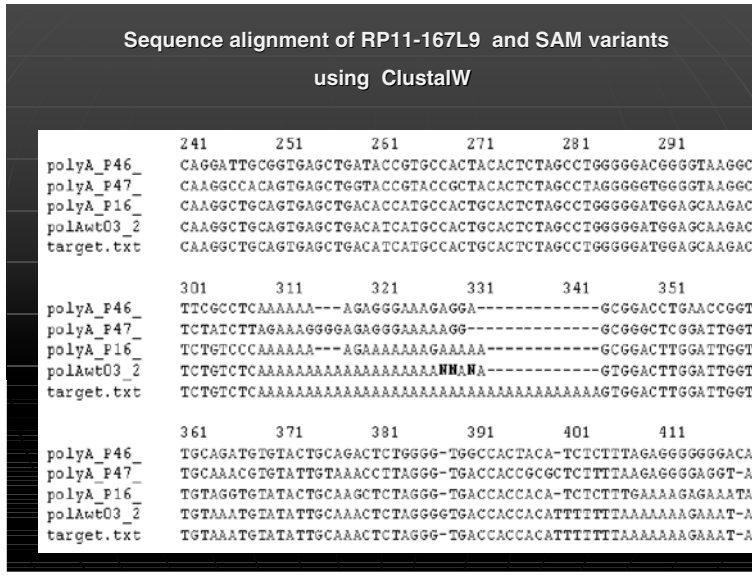


FIGURE 15.6. Clustal W alignment of region of wild-type human BAC RP11-167L9 (polAwt), and mutated variants compared to the published estimation of the polyA tract (target). The wild-type clone contains a problematic motif that prevents accurate dye-terminator cycle sequencing (variable length reads, unreadable N's present), whereas randomly mutated variants of the clone display consistently reduced polyA tract length.

# 4. Sequencing by Hybridization (SBH)

## *4.1 The Limitations of Conventional SBH*

"Sequencing by hybridization" (SBH) is a potentially powerful sequencing technology that analyses sequence reads in massively parallel manner by the hybridization of short target fragments to an array of "all" possible oligonucleotide probes. SBH is based on a relatively simple concept, in which a non-redundant array of oligonucleotides (of length $p$) is arranged on a solid support (typically a silica or glass slide)[23-25]. If a target fragment hybridizes to particular probes in the array, a representative signal is obtained from each of those probes. Signals from each element in the array that hybridizes to each complementary region of the target fragment collectively constitute the "SBH spectrum", which may be analysed by reconstruction algorithms to generate a sequence of the entire target region. Potentially, SBH can provide megabase-scale simultaneous sequencing capacity if practical hybridization and data reconstruction problems can be overcome. Although SBH has been used in detailed local mapping, SNP detection and re-sequencing of relatively small regions[24], doubts remain as to whether SBH can be used for *de novo* sequencing, particularly for megabase scale.

Southern and colleagues[26-28] have demonstrated that steric factors, sequence motifs and differential nucleotide hybridization stabilities each contribute to non-uniformity in hybridization of the target to anchored oligonucleotide probes. Potentially, the breakage of target DNA into small pieces could disrupt the sequence runs and structural barriers that prevent hybridization to probe oligonucleotides. Although some physical intra-strand interactions may be reduced, inter-strand interactions between the smaller target element fragments could be expected to increase.

SBH is a qualitative hybridization technology and many practical barriers to its use for *de novo* sequencing have been identified:

- Hybridization is non-quantitative and non-representative
- Some sequence motifs provide no signals, yet others provide strong signals
- Foldback and other intra-strand hybridizations can limit availability of target regions for hybridization to oligonucleotide probes
- Short repeated regions provide ambiguous reconstruction paths
- AT and CG rich targets have different hybridization efficiency

SBH cannot be used for sequencing simple sequence DNA, such as poly A, (GT)n, repeats, etc. As it cannot quantify the number of repeat copies present in the target, it cannot easily determine the length of the repeat region. SBH also cannot determine the order of repeated sequence elements: these becomes ambiguous if repeat elements longer than length $p$ are present in the target. Practically, SBH is limited to analysis of targets that *lack* 3 or more copies of repeats longer than ($p$-*1*). The probability of ambiguity increases exponentially as probe length decreases, suggesting longer oligonucleotide probe lengths as a practical solution. However, as probe length increases the complexity and number of probes located on the SBH array necessarily also increases exponentially. Other informative techniques must be used to supplement SBH data to overcome these ambiguities[25]. These contradictory physical and mathematical problems require novel solutions – SAM technology can provide these solutions.

## *4.2 The Use of SAM Techniques for SBH Sequencing*

*SAM techniques* can resolve ambiguities in SBH reconstruction. Mutant variants of a target sequence are analysed by SBH, each variant generating a unique SBH spectrum, then each variant spectrum is reconstructed using *SAM algorithms*. Mutant fragments contain *unambiguously reconstructable regions* that span repeated ($p$-*1*)-mers in the target. These reconstructed regions of mutants are then used in combination as templates to resolve ambiguities in reconstruction of the target sequence spectrum. Computer simulations have shown that analysis using *SAM algorithms* can often get the overall order of maximal sub-strings correct, even when a small number of short sub-strings are misplaced.

Interestingly, Southern and Nguyen[29] and Nguyen et al[30] have recently suggested the introduction of nucleotide analogues into the oligonucleotide probes and use of analogues with chaotropic agents respectively as methods for reducing differential hybridization potential between AT- and CG-rich targets and targets with structural motifs. These methods however do not address the problems of ambiguous reconstruction caused by repeated motifs, they address the differential hybridization stabilities of probes of different base composition.

### *4.3 Computer Simulations Using SAM for SBH Reconstruction*

For simulation studies, random 100 kb pieces of human genome sequence were used as a source of "real" DNA sequence. Randomly mutated variant sequences were generated at different mutation intensities, allowing each base equal probability of variation. Computer simulations of SBH reconstructions of mutant variants of target DNA elements were undertaken using SAM algorithms to direct the build as outlined in *Section 4.2*. The results can be compared to standard SBH reconstructions of the original wild-type human target sequences. Randomly selected 5 kb fragments of human genomic DNA can be *completely reconstructed* in 99.9% of attempts with fewer than 1 error per 1000 bases (97.8% perfectly correct) using 9 mutants and probes of length 13. In contrast, fewer than 1% of reconstructions of 5 kb fragments using standard SBH spectra were correct, even allowing 0.1% error.

The size of DNA fragments that can be reconstructed correctly when SAM techniques are used also increases markedly. Human DNA fragments up to 30 kb long were successfully reconstructed during simulated SAM experiments. In contrast, no fragments of this length could be reconstructed applying conventional builds of the SBH spectrum. There is scope for improving the performance of the current SAM reconstruction algorithms for SBH. For example, the current algorithms assume equal probabilities of mutation of each nucleotide base. Algorithms could be modified to incorporate the probabilities of particular mutation events, which are determined empirically for particular mutation protocols[16,17].

## 5. Conclusions

This paper discusses the novel approach to sequencing which we have developed called *Sequencing by Aligning Mutants* (*SAM*). It was developed with the purpose of providing a simple and effective method of sequencing DNA motifs that cannot be sequenced by other current techniques. *SAM technologies* include methods to achieve highly controlled levels of mutation in target DNA elements. Preferably these mutations are simple substitution mutations.

The methods also include advanced assembly and reconstruction algorithms to recover original sequence from a small number of altered versions of the target[15-18].

We have shown that improved sequence reads can be obtained using SAM techniques and conventional Dye-terminator cycle sequencing from several model DNA's which contain "difficult to sequence motifs". Although not displayed here for sake of space, the reconstructed sequence recovered from the model targets is accurate even using fewer than 10 variants. The protocols are repeatable and are readily modifiable for different DNA sequence motifs. Although the overall approach is novel, the technologies were developed with the view that they are compatible with use of standard laboratory processes and equipment, and thus available for conventional molecular biological laboratories which may be lacking sophisticated genomic analysis equipment.

The intention of our laboratory is to develop and release portfolios of methods and reagents as well as portfolios of advanced assembly algorithms. Versions of the algorithms may have additional applications for improved sequence comparison. Together these developments are intended to form the basis for several different genomic software tools to be applied along with conventional sequencing kits.

The application of SAM technologies to SBH based sequencing is a new area of development, as the potential for repetitive motifs and other structural motifs that interfere with target::probe hybridization could be diminished within variant target molecules. New array chemistries such as PNA oligonucleotide probes[31] could be used to further alter target::probe interactions and provide a broader spectrum of hybridization signals than conventional DNA::DNA arrays.

## *References*

1. The International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature* **409**, 860-921 (2001).
2. Venter, C. et al. The sequence of the human genome. *Science* **291**, 1304-1351 (2001).
3. *Dictyostelium* Genome Sequencing Consortium. Sequence and analysis of chromosome 2 of *Dictylostelium discoideum. Nature* **418**, 79-85 (2002).
4. Aach, J., Bulyk, M.L., Church, G.M., Comander, J., Derti, A. & Shendure, J. Computational comparison of two draft sequences of the human genome. Nature **409**, 856-859 (2001).
5. Waterston, R.H., Lander, E.S. & Sulston, J.E. On sequencing the human genome. *Proc. Natl. Acad. Sci. USA* **99**, 3712-3716 (2002).
6. Eichler, E.E. Segmental duplications: what's missing, misassigned, and misassembled—and should we care? *Genome Res*. **11**, 653-656 (2001).
7. Bailey, J.A., Yavor, A.M., Massa, H.F., Trask, B.J. & Eichler, E.E. Segmental duplications: organization and impact within the current human genome project assembly. *Genome Res*. **11**, 1005-1017 (2001).

8. International Human Genome Sequencing Consortium. (2002). Current sequencing status.*http://www.ncbi.nlm.nih.gov/genome/seq/*page.cgi?F=HsProgress. shtml&& ORG=Hs
9. Wolfsberg, T.G., Wetterstrand, K.A., Guyer, M.S., Collins, F.S. & Baxevanis, A.D. Introduction: putting it together. *Nature Genet*. **32** Suppl, 5–8, (2002).
10. Collins, F.S. & McKusick, V.A. Implications of the Human Genome Project for medical science. *J. Am. Med. Assoc*. **285**, 540-544 (2001).
11. Green, E.D. Strategies for the systematic sequencing of complex genomes. *Nature Rev. Genet*. **2**, 573-583 (2001).
12. Horvath, J.E., Schwartz, S. & Eichler, E.E. The mosaic structure of human pericentromeric DNA: a strategy for characterizing complex regions of the human genome. *Genome Res*. **10**, 839-852 (2000).
13. Horvath, J.E., Viggiano, L., Loftus, B.J., et al. Molecular structure and evolution of an alpha satellite/non-alpha satellite junction at 16p11. *Hum. Mol. Genet*. **9**, 113-123 (2000).
14. Bailey, J.A., Gu, Z., Clark, R.A., Reinert, K., et al. Recent segmental duplications in the human genome. *Science* **297**, 1003-1007 (2002).
15. Keith, J.M., Adams, P., Bryant, D., Mitchelson, K.R., Cochran, D.A.E., & Lala, G.H. Inferring an original sequence from erroneous copies: a Bayesian approach. *Proceedings of the 1st Asia-Pacific Bioinformatics Conference* (APBC 2003). (ed. Chen, Y-P.P) **19**, 23-28 (2003).
16. Keith, J.M., Adams, P., Bryant, D., Cochran, D.A.E., Lala, G.H. & Mitchelson, K.R. Inferring an original sequence from erroneous copies: two approaches. *Asia-Pacific BioTech News* **7**, 107-114 (2003).
17. Keith, J.M., Adams, P., Bryant, D., Cochran, D.A.E., Lala, G.H. & Mitchelson, K.R. Algorithms for sequencing aided by mutagenesis. *Bioinformatics* **20**, 2401-2410 (2004).
18. Keith, J.M., Adams, P., Bryant, D., Kroese, D.P., Mitchelson, K.R., Cochran, D.A.E. & Lala, G.L. A simulated annealing algorithm for finding consensus sequence. *Bioinformatics* **18**, 1494-1499 (2002).
19. Razin, S.V., Ioudinkova, E.S., Trifonov, E.N. & Scherrer, K. Non-clonability correlates with genomic instability: a case study of a unique DNA region. *J. Mol. Biol*. 307, 481-486 (2001).
20. Kang, H.K. & Cox, D.W. Tandem repeats 3′ of the IGHA genes in the human immunoglobulin heavy chain gene cluster. *Genomics* **35**, 189-195 (1996).
21. Baran, N, Lapidot, A. & Manor, H. Formation of DNA triplexes accounts for arrests of DNA synthesis at d(TC)n and d(GA)n tracts. *Proc. Natl. Acad. Sci. USA* **88**, 507-511 (1991).
22. Mallon, A.M., Platzer, M., Bate, R., Glöckner, G. et al. Comparative genome sequence analysis of the Bpa/Str region in mouse and man. *Genome Res*. **10**, 758-775 (2000).
23. Southern, E.M. DNA microarrays: History and overview. *Methods Mol. Biol*. **170**, 1-15 (2001).
24. Chechetkin, V.R., Turygin, A.Y., Proudnikov, D.Y., et al. Sequencing by hybridization with the generic 6-mer oligonucleotide microarray: an advanced scheme for data processing. *J. Biomol. Struct. Dyn*. **18**, 83-101 (2000).
25. Drmanac, R., Drmanac, S., Chui, G., et al. Sequencing by hybridization (SBH): advantages, achievements, and opportunities. *Adv. Biochem. Eng. Biotechnol*. **77**, 75-101 (2002).

26. Mir, K.U. & Southern, E.M. Determining the influence of structure on hybridization using oligonucleotide arrays. *Nat. Biotechnol*. **17**, 788-792 (1999).
27. Southern, E., Mir, K. & Shchepinov, M. Molecular interactions on microarrays. *Nat. Genet*. **21** (Suppl.), 5-9 (1999).
28. Shchepinov, M.S., Case-Green, S.C. & Southern, E.M. Steric factors influencing hybridisation of nucleic acids to oligonucleotide arrays. *Nucleic Acids Res*. **25**, 1155-1161 (1997).
29. Nguyen, H.K. & Southern, E.M. Minimising the secondary structure of DNA targets by incorporation of a modified deoxynucleoside: implications for nucleic acid analysis by hybridisation. *Nucleic Acids Res*. **28**, 3904-3909 (2000).
30. Nguyen, H.K., Fournier, O., Asseline, U., Dupret, D. & Thuong, N.T. Smoothing of the thermal stability of DNA duplexes by using modified nucleosides and chaotropic agents. *Nucleic Acids Res*. **27**, 1492-1498 (1999).
31. Weiler, J., Gausepohl, H., Hauser, N., Jensen, O.N. & Hoheisel, J.D. Hybridisation based DNA screening on peptide nucleic acid (PNA) oligomer arrays. *Nucleic Acids Res*. **25**, 2792-2799 (1997).