

Foresight

Infectious Diseases: preparing for the future

OFFICE OF SCIENCE AND INNOVATION

S6: State-of-Science Review: Genomics and Bioinformatics

Julian Parkhill and Nicholas Thomson
The Sanger Institute, Wellcome Trust Genome Campus

This review has been commissioned as part of the UK Government's Foresight project, Infectious Diseases: preparing for the future. The views expressed do not represent the policy of any Government or organisation.

1 Introduction

2 Brief historical overview

3 Current and near-future technology

3.1 Sequencing and re-sequencing

3.2 Microarrays

3.3 SNP detection and PCR

3.4 Array-based re-sequencing

3.5 Bioinformatics

4 Applications

5 Conclusion

6 Glossary

7 References

1. Introduction

Until fairly recently, methods for detecting microbes were largely based on culturing the organisms on artificial media under laboratory conditions or on cellular staining and viewing through light microscopy. Coarse-level identification was similarly based on appearance after culture and differential culture on specific substrates. These are still the gold standard for microbial identification.

This technology has been in use for many decades, but suffers considerable drawbacks, including lack of speed, resolution and the fact that some disease agents simply cannot be grown in the lab. In addition, finer level resolution (typing) that differentiates strains within species has become essential for epidemiology. For these reasons, molecular techniques, based on direct detection and sequencing of genomic DNA are now becoming widely used. The genomic DNA contains the blueprint for the construction of the organism; amplification (by the Polymerase Chain Reaction; PCR) and direct readout of this via full sequencing, comparison with known sequences, or detection of small differences in the DNA (Single Nucleotide Polymorphisms; SNPs), enables highly sensitive and accurate detection and identification of microbial agents of disease. This Review focuses primarily on current and near-future technologies for DNA sequencing and identification using hybridisation, PCR or SNP-detection based systems.

This is a rapidly changing field, with novel technologies for high-throughput sequencing and hybridisation coming onto the market now, or likely to do so in the next few years. This trend is primarily driven by the need (medical and commercial) to identify rapidly and cheaply SNPs within individual human genomes for personalised medicine. However, the machines and technologies will also be suitable for detection and identification of microbial agents of disease. Such technologies can operate at three levels; factory scale, bench-top scale and on hand-held devices. Generally there is a trend for technologies to move down through these scales as they are developed, but each will continue to have specific uses. Factory-scale technologies will underlie the generation of the basic information that will support diagnostics and further research, bench-top devices will be used by researchers and reference laboratories, and hand-held devices will become important in the field.

2. Brief historical overview

Genetic information is encoded in DNA through the order of four chemicals - or bases - Adenine, Guanine, Cytosine, and Thymine along the double-stranded sugar-phosphate backbone. The model of DNA structure proposed in 1953 by Watson and Crick (Watson & Crick 1953), along with Franklin and Wilkins (Franklin & Gosling 1953; Wilkins et al. 1953), carried with it the inference that this was the likely method of encoding genes. The detailed specifics of the code were worked out in a series of seminal experiments over the next decade or more (see (Judson 1996) for a review).

Given this code, it was clear that genetic information could be read experimentally from DNA, and two technologies were devised to allow this to be done in a systematic way. Both methods involved producing a series of radio-actively labelled DNA strands each starting at the same place, but stopping at different occurrences of a specific base in the DNA. These strands could be separated by length using electrical fields in thin polyacrylamide gels (electrophoresis) and identified by autoradiography, producing a series of 'ladders' that could be read to give the sequence of bases in the original DNA.

The method of Maxam and Gilbert used chemical cleavage of DNA at specific or semi-specific bases to produce the ladders (Maxam & Gilbert 1977). The method devised by Sanger and colleagues used synthesis by naturally occurring DNA replicases which was terminated at specific bases by individual base analogues (Sanger et al. 1977). The Sanger method is more elegant, and easier to use, and proved to be more automatable, becoming the method of choice for DNA sequencing. It is still the basis of the methods used today, primarily due to its ability to be adapted to high-throughput protocols.

The fundamental importance of genetic information and the utility of DNA sequences in virtually every aspect of molecular biology drove the refinement and expansion of Sanger sequencing technology. This was initially at the laboratory scale, with developments including new polymerases, gel technology, cloning and sequencing vectors. Although these increased the throughput and ease of use of the system, it would still take months of work by an experienced scientist to generate a few thousand bases of DNA sequence.

Key breakthroughs enabled the truly industrial scale sequencing necessary to tackle the 3,000,000,000 bases of the human genome. These included:

- The development of fluorescent DNA dyes that could be read continuously by lasers to replace radioactive labelling and autoradiography (Ansorge et al. 1986; Prober et al. 1987; Smith et al. 1986).
- Replacement of thin layer polyacrylamide gels with individual polyacrylamide-filled capillaries (Heiger et al. 1990; Luckey et al. 1990; Swerdlow et al. 1990).

These technologies, with some subsequent refinement, form the basis for currently available sequencing machines. These machines automate the process of separating and detection of labelled DNA fragments, but the sequencing reactions themselves, and the preceding DNA extraction and amplification steps, still need to be done separately.

3. Current and near-future technology

3.1 Sequencing and re-sequencing

Current sequencing technologies are based on the chain-termination methods devised by Fred Sanger and colleagues in the 1970s. Throughput has been massively improved by the use of fluorescent dyes and capillary separation of DNA ladders, as illustrated in Figure 1. The levels of throughput currently

A further level of complexity arises from the fact that many genomes contain long, repeated sequences, and assemblies of random reads cannot resolve repeats longer than the read length. This is circumvented by simultaneously sequencing the ends of larger fragments of DNA. The linkage between pairs of reads at different distances gives the positional information required to assemble genomes. The process necessary to generate these fragments, known as cloning, is another intermediate stage between the DNA and the final sequence. Advances in bioinformatics have produced computer programs for assembly of sequence that take into account these paired-end reads.

With the advances described previously, the capacity to generate sequence data provided by these machines is large. A single ABI 3730, running 20 times a day, will generate roughly 1.25 million bases (Mb) of raw sequence, enough for one-fold coverage of a small bacterial genome. A set of 80 machines (roughly the number installed at the Sanger Institute, for example) can therefore generate 100 Mb of raw sequence per day - enough to cover three 4 Mb bacterial genomes at 8-fold redundancy each day or for 10-fold coverage of a human genome within a year. Even with 8-fold coverage, a number of gaps and errors will occur in the sequence, requiring a mixture of human and computer checking and re-sequencing. Certain criteria must be met before a sequence can be considered to be "finished": All bases must be covered by at least two clones, and at least one read in each direction, or with different sequencing chemistries. All gaps must be closed with sequence meeting the same criteria, and all repeats must be bridged with reads-pairs, or PCR products. Sequences that do not meet these criteria are referred to as "draft" sequences. If these criteria are met, the final accuracy can be very high. Indeed, an error rate of less than one base-calling error per Mb is achievable.

Despite this throughput, the demand for faster and higher-throughput sequencing is still growing strongly. The main reason for this is the perceived benefits of, and market for, 'the \$1,000 human genome'. A sequencing technology capable of this level of cost and throughput would open up wide-ranging possibilities for very large-scale human re-sequencing, leading to personalised medicine – the ability to tailor drug regimes to individuals and to rapidly identify disease susceptibilities. The company that can bring this to fruition will clearly benefit considerably. Of course, the fact that this technology is being driven by human genome sequencing will not prevent its use for research into, and detection and identification of, disease causing agents, whether viral, bacterial, protist or helminth. Potential applications in these areas are discussed below.

Further extensions of the capillary sequencing techniques using Sanger chemistry are still likely. Amersham have recently upgraded their top-of-range sequencer from 96 to 384 lanes, while a prototype 768-lane sequencing machine has been developed by a team at the Whitehead Institute in Massachusetts (Aborn et al. 2005), promising a potential 8-fold increase in throughput over the ABI3730 machine.

Beyond this, new technologies currently near-market will increase throughput by several orders of magnitude by removing altogether the need for

capillaries. Two companies appear to be currently leading the field in this. The first, 454 Technologies (<http://www.454.com>), has produced a highly-parallel method for automating a technique termed 'pyrosequencing', in which a template DNA is copied one base at a time. The addition of each base releases a pyrophosphate molecule which triggers a cascade resulting in the emission of light. The system includes an integrated bead-based DNA amplification system, and the chemistry and detection is performed in picolitre wells. The current technology generates around 300,000 100-200 bp reads per 4-hour run and claims to produce 20-30 Mb of raw sequence data, enough for an 8-10 fold coverage of an average bacterial genome. This is an increase in throughput of around two orders of magnitude from the ABI 3730. The 454 technology has already been used to identify point mutations within whole genomes of *Mycobacterium tuberculosis*, that were responsible for resistance to a novel antibiotic (Andries et al. 2005). Further increases in throughput are to be expected.

Promising an even higher throughput, but a little further from market, is a technology developed by Solexa (<http://www.solexa.com>), which is based on sequencing-by-synthesis. In this case, the DNA template is again copied enzymatically, but individually labelled fluorescent nucleotides are added, detected with a laser, and the fluorophore removed, allowing the addition of the next base. These reactions are performed on DNA fragments linked directly to glass slides, dispensing with the wells and allowing many more simultaneous reactions (up to 108 per cm²) albeit with shorter individual reads (30-50 bp). This implies a throughput of >3,000 Mb per cm² per run, bringing human re-sequencing within the range of these devices.

At present the advantages of both these new methodologies are balanced by reduced sequence length and lack of long-range information, due to the absence of a cloning step. This limits their usage primarily to comparative re-sequencing against reference sequences or random sample sequencing. However, this will allow very rapid identification of base-pair differences in non-repetitive regions of entire genomes, eventually up to the size of the human genome, and could well be extremely useful for sampling mixed populations in specific environments or niches, a process termed 'metagenomics' (Tyson et al. 2004; Venter et al. 2004).

Many more technologies are in development at many different levels, which may bring still greater advances in speed and cost in the longer term. Visigen (<http://www.visigenbio.com>) are developing a system that directly detects the incorporation of labelled nucleotides by a tethered polymerase. In this case, the incorporation is in real time, as opposed to the step-wise incorporation used by 454 and Solexa, opening up the possibility of reading long DNA sequences at the rate of polymerisation (about 1000 bp/second).

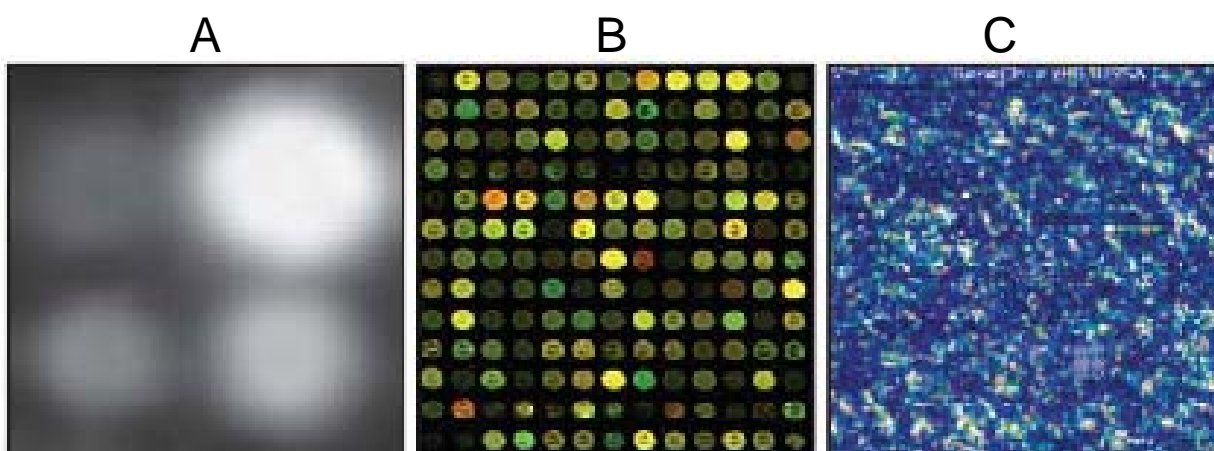
The projects recently funded by the US National Institutes of Health (NIH) under their Advanced Sequencing Technology Awards (<http://www.genome.gov/12513162>) indicate still further potential advances, some technical and relatively straightforward, and some scientific and more difficult. Improved pyrosequencing and sequencing by synthesis are projected, including integrated devices capable of performing every stage in

the standard sequencing process in a single micro-device. In addition to this, several potential methods for direct read-out sequencing of individual long DNA molecules are described. These include drawing DNA molecules through nanopores or nanogates and detecting individual bases by electrical or optical changes, and dragging a 'reading head' across DNA molecules with an atomic force microscope. Such technologies offer real possibilities for the very high-throughput, very rapid, integrated sequencing technologies that will be needed to realise some of the longer term aims discussed under the Foresight initiative.

3.2 Microarrays

Sequencing involves the detection and identification of DNA through direct read-out of the base sequence. However, DNA can also be detected and identified using the principle that a single-stranded nucleic acid molecule has the capacity to recognise a complementary strand through base-pairing. This process, known as hybridisation, allows the direct detection of DNA with a similar sequence to a given probe DNA, and a measure of the level of similarity through the strength of the interaction. Since the process of recognition and base-pairing is highly specific, many base-pairing reactions can be assayed simultaneously and in a complex mixture. This process has obvious utility in the detection and identification of disease agents.

The initial concept of DNA arrays originated from spotted or bacterial colony macroarrays where bacterial samples carrying different plasmid clones were spotted and lysed directly on nylon membranes, later to be challenged with labelled (usually radioactively labelled) DNA probes (Southern 2001). The spots would be manually applied to the membrane and spaced ~2mm apart and so a single membrane could hold tens of spotted samples. However, the small number of samples that can be tested on a single array, the difficulty in achieving the correct hybridisation conditions, the inherent experimental restrictions associated with using radioactivity, and the high background signal, through cross reaction, with the host bacterial DNA all reduce the usefulness of this technique.

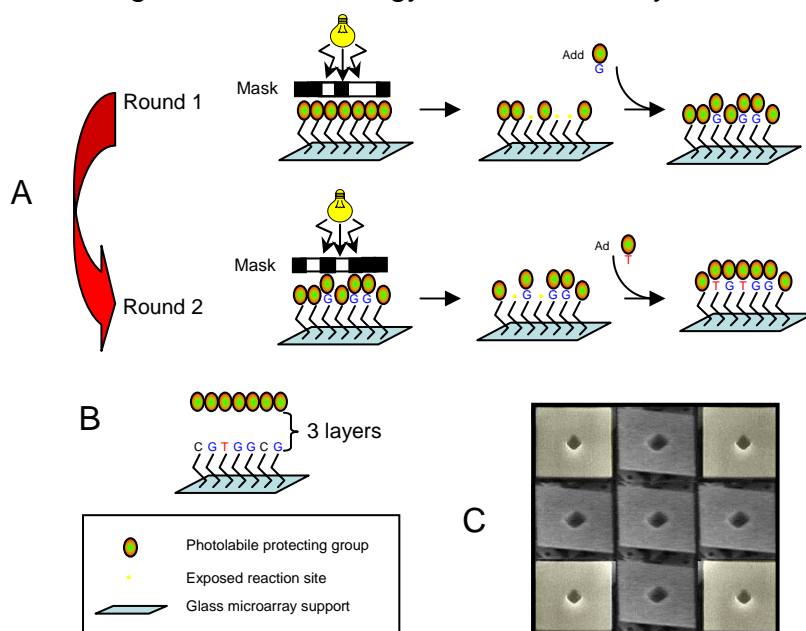


Three equivalently sized portions of three different DNA arrays: (A) four features (probes) on an original colony blot array (B) 169 features on a pen-tip spotted array and (C) 22,500 features on a photolithographic *in situ* synthesised Affymetrix GeneChip array. Taken from Stoughton (2004)

Although the principles of DNA-DNA based hybridisation are the same for modern arrays, the methods of manufacture, number of samples per unit area, resolution and actual size of the array are very different. Current arrays are approximately the size of a microscope slide (hence 'microarray') and can hold hundreds of thousands of DNA features (see Table 1; Figure 2). (For a comprehensive review see (Stoughton 2004)).

Technologies for manufacturing microarrays fall essentially into two genres: those for which the DNA probes are robotically spotted directly onto the support; and those where they are manufactured in-situ directly on the solid support.

Spotted microarrays require pre-manufactured probes which can be applied to an array in nano/picoliter volumes using needle/pen or inkjet and microjet deposition technology. The probes themselves are either covalently or non-covalently linked to the support material, usually glass or nylon. The advantage of this technology is its affordability and ease of use and, in the



case of inkjet delivery, its suitability for very high throughput (Southern 2001). However, there are drawbacks. DNA samples to be linked to the array must be pre-made, purified and, more importantly, stored prior to application on the microarray. In addition, the density of features that can be applied to an array falls very short of those that can be achieved by in-situ based techniques (Table 1; Figure 2).

Figure 3: High density *in situ* arrays.

(A) UV light shone through a photolithographic mask directs the localised removal of the photolabile protecting groups from linkers, and subsequently from blocked nucleotides. A pure solution of each protected nucleotide is added and chemically couples to the newly exposed reactive sites. This process is repeated in total four times, once for each of the four nucleotides, with a different photolithographic mask used for each nucleotide added. This entire process is repeated for every layer of the growing oligonucleotide chains (B). Typically oligonucleotides are between 25-70 nucleotides long. (C) In place of photolithographic masks the NimbleGen (www.nimblegen.com/) array technology uses a bank of programmable micro-mirrors to direct the light and de-protect different regions of the array. Adapted from (Lipshutz et al. 1999).

In-situ synthesised, high-density oligonucleotide arrays take several forms, perhaps the most mature technology being a UV light-based technique

developed by Fodor et al., (Fodor et al. 1991) and marketed by Affymetrix (<http://www.affymetrix.com/>). This relies on light directed oligonucleotide synthesis. In essence, synthetic linkers with photo-removable protecting groups are attached at precise locations, and in high density, to a glass slide. Activation by UV light removes the photo-labile protecting group and allows hydroxyl-protected nucleotides, which are washed over the array, to couple chemically with the de-protected synthetic linkers. This can then be repeated as the newly coupled nucleotides also possess a photo-labile protecting group. Successive rounds of de-protection and washing with hydroxyl-protected nucleotides extends the oligonucleotide chain in a precise and controlled manner.

However, in order to produce simultaneously many thousands of different oligonucleotides on a single array slide a photolithographic mask (an etched glass or chromium filter) is used. The UV light is passed through the mask which blocks out the light for specific areas of the array slide, leading to localised de-protection of the different oligonucleotides chains (Figure 3). A different mask is required for each of the four hydroxyl-protected nucleotides (A, C, G or T), and so four masks are required for each complete layer (or extension) of the oligonucleotide chains on the array. Thus, the pattern of the masks and the order with which the different hydroxyl-protected nucleotides are washed over the array precisely determines the sequence and length of the oligonucleotide chains being built.

The major draw back of the photolithographic mask-based technology is its lack of flexibility, with a new set of masks being required for any modification/addition to the array. Moreover, masks can take days to weeks to make and are expensive to manufacture. So there are inherently high start-up and running costs.

NimbleGen (<http://www.nimblegen.com>) has developed this in-situ light-directed chemical synthesis further, removing the need for the photolithographic mask. In place of a mask, NimbleGen's technology uses a solid-state array of 786,000 miniature aluminium micro-mirrors that reflect the light to precise locations on the array and direct the localised de-protection of the in-situ synthesised oligonucleotides on the array ((Singh-Gasson et al. 1999); Figure 3). Since the position of the moveable micro-mirrors can be individually programmed by computer, there is considerable flexibility, and high density arrays can be rapidly created at a much lower cost.

In addition to NimbleGen's 'maskless' array, there are several other techniques for producing high density arrays without the need for a mask including: the modified ink-jet technology used by Agilent (<http://www.agilent.com>); micro-fluidics developed by Febit (<http://www.febit.com>); and electrode-directed array manufacture marketed by CombiMatrix (<http://www.combimatrix.com>) and Nanogen (<http://www.nanogen.com>).

Agilent (a spin-off company from Hewlett-Packard) use inkjet printer technology to deliver single nucleotides very accurately in picoliter volumes in order to extend the oligonucleotide chains (Hughes et al. 2001). This system

uses standard, well tested technology and is highly flexible. Any changes to the design of the array simply involve changing the sequence file, analogous to a standard text file, to be printed. The main adaptations involved developing the carrier liquid and modifying standard oligonucleotide chain extension (phosphoramidite monomer based) chemistries. The limitations of this technology lie in the minimum feasible droplet size which may ultimately limit the maximum number of probes that could be placed on a single array slide.

Table 1: Mechanisms of microarray manufacture.

Process	Manufacturer ^a	Substrate	Density/resolution
Pre-synthesised oligos			
Pen tip deposition	Clontech	Nylon	<10 features/mm ²
Ink-jet deposition	GE Healthcare	Glass or polyacrylamide	100 features/mm ²
Electro-phoretically driven deposition	Nanogen	Silicon	100 features/mm ²
<i>In-situ</i> based technologies			
Photolithographic mask	Affymetrix	glass	8200 features/mm ²
Micromirror virtual masks	NimbelGen -Systems	Glass	~1000 features/mm ²
Micromirror and microfluidics	FeBit	Glass/silicon	~1000 features/mm ²
Ink-Jet printer	Agilent – Technologies	Glass	100 features/mm ²
Electrode-directed	CombiMatrix	Silicon	<100 features/mm ²

Adapted from Stoughton (2005).

^a See Stoughton (2005) for a more exhaustive list of manufacturers.

Miniaturisation is also a current concern for the electrode-directed array system used by Nanogen and CombiMatrix. Their technology piggybacks on the microelectronics industry and is therefore certain to improve greatly in the near future. It uses an integrated circuit with arrays of microelectrodes that generate a reconfigurable electric field over the surface of the array, which can direct the rapid electrophoretic transport of DNA molecules across its surface. This is useful both for the construction of the array itself and for the rapid transport of test material onto the array. The electric fields can also be used to alter the stringency with which hybridisation to the array can occur.

Currently Nanogen commercially produces a 400-test-site chip array (i.e. 400 independent positions for hybridisation) that is ~1cm² in size. However, if you compare the size of each test site with the size of the equivalent feature on an Affymetrix array, these microelectronic arrays are somewhat larger. Each

electrode reaction site is 80-94 microns in diameter which is ~20 x larger than the individual oligonucleotide features on the current Affymetrix Genechip array (5 microns in diameter). This greatly limits the number of samples that can be tested simultaneously (see Table 1). Considering the rate at which microelectronics has been progressing, current limitations hindering electrode-directed array technology are unlikely to apply in the near future. In a similar vein, it is also anticipated that, using photolithographic and micromirror array construction techniques, it should be possible to reduce the feature size to ~1 micron, dramatically increasing the number of probes per array slide, further reducing cost and sample size.

There has also been rapid progress in developing microarray reading technologies. Currently, all commercial technologies use either white light or laser confocal-based array scanning systems to detect the fluorescent dyes used to label samples. The latter offers higher sensitivity and resolution. However, current advances in array reading methodology include other, less mature, technology such as surface plasmon resonance which enables direct detection of the molecular interactions between the probe and sample on the array surface (Schultz 2003). This negates the need for fluorescent or biotin etc. labelling, and also offers the possibility of being able to detect single molecules. This will overcome some of the problems of signal amplification and perhaps reduce the amount of biological sample required for testing. This technology also promises a reduction in size for the microarray detection system as a whole. Bearing in mind that most of the current microarray systems discussed above are 'bench top' scale, this is essential if a hand-held device is to be realised.

Currently, many of the high density array technologies described above are not widely used, partly due to lack of synthesis machines on the market, and partly due to the plethora of Affymetrix patents that govern this area and that have been hotly contested (Dickson 2000).

3.3 SNP detection and PCR

Detection and identification protocols are also likely to include technologies such as SNP detection or real-time PCR.

Single nucleotide polymorphisms (SNPs) are essentially regions where individual organisms differ by a single base-change in an otherwise identical stretch of DNA. Their primary use is in human genetics, where they form the basis of many mapping, localisation and gene-association strategies. It should be noted that SNP identification is only a method of inferring genotypes – the SNPs themselves do not form the genotype. An international public/private consortium (HapMap; <http://www.hapmap.org/>) was formed to identify SNPs in the human genome (Hapmap_Consortium 2003) and to study their association in low-recombination blocks (haplotypes) as an aid to large-scale gene-association and gene-discovery programs. This consortium has, so far, identified and mapped around 1 million SNPs, while a private company, Perlegen Sciences, has publicly released data on 1.6 million SNPs (Hinds et al. 2005).

Given these large numbers of known SNPs, gene association studies in humans require very high throughput determination of SNP sequences (genotyping). As with high-throughput sequencing, the commercial sector has stepped in with the required SNP assay systems. These can be based on a number of technologies, including high-stringency hybridisation or single-base extension assays on oligonucleotide micro-arrays (see above), through a number of different oligonucleotide extension/ligation/detection systems (e.g. ABI/SNPlex (De la Vega et al. 2005), Illumina (Shen et al. 2005)) to DNA cleavage/Mass spectrometry technologies (e.g. Sequenom (Jurinke et al. 2002)). Many of these are multiplexed and automated to allow very high-throughput, producing 108 to 109 genotypes per day.

Again, it is likely that in the future these technologies will be further expanded and miniaturised, thereby making an impact on the detection and identification of infectious disease agents. Molecular techniques based on single base changes are often used to discriminate within near-clonal species of bacteria for epidemiological reasons. Examples include *Bacillus anthracis* (Easterday et al. 2005) and *Yersinia pestis* (Achtman et al. 2004). In addition, sequence-based techniques such as multi-locus sequence typing (MLST)(Urwin & Maiden 2003) are increasingly used for epidemiological studies on more diverse species. It is likely that some of these techniques will be adaptable to MLST-type studies. Indeed, Sequenom are already investigating the use of their mass-spec technology in this way.

Polymerase chain reaction (PCR) (Saiki et al. 1988) is a method for amplifying defined stretches of DNA, which is used in some form in almost all the DNA-based detection and identification systems described here. Although its adoption in molecular biology has been near universal, and it is almost impossible to overestimate the lack of progress that would have been achieved without it, PCR does have some drawbacks. Among these is the fact that it is not quantitative in its output and that it requires accurate and rapid thermal cycling, both of which could limit its ability to be automated within integrated systems.

The first of these problems has been tackled with real-time PCR systems, where the amplification process is monitored throughout, rather than just at the end of the process. Relatively highly parallel bench-top real-time PCR machines are now available (e.g. AppliedBiosystems 7900HT which can perform 384 reactions simultaneously), and hand-held real-time PCR devices have been produced for pathogen detection (Higgins et al. 2003) (<http://www.idahotech.com/razor/>). This is an area that is likely to be developed further in the future.

The requirement for thermal cycling is a more serious drawback for miniaturised integrated devices. Several approaches have been developed to circumvent this, including: loop-mediated isothermal amplification (Notomi et al. 2000); helicase-dependant amplification (Vincent et al. 2004); rolling-circle amplification (Detter et al. 2002); and others (Andras et al. 2001). It is likely that such methods and developments from them will lead to the rapid and specific DNA isothermal amplification that will be required for small-scale integrated DNA identification devices

3.4 Array-based re-sequencing

Array-based re-sequencing combines several of the aspects of sequencing, SNP detection and microarrays. There are essentially three different approaches; the first two rely on the fact that even single base pair mismatches between the probe and the sample DNA can be detected as variations in the hybridisation signal observed on a microarray. With this in mind there are essentially two methodologies (reviewed in more detail in (Hacia 1999; Pastinen et al. 1997)):

The first approach (gain-of-signal) relies on knowing that a specific region of DNA varies (insertions/deletions/substitutions) by a single base or a small number of bases. Synthetic probes are designed such that all possible variations of that specific region of interest are represented on the array. Consequently, the array probe representing the correct sequence variation will display a heightened signal relative to those with mismatches. This technique is useful in detecting variants with single base-pair differences in small sequences. But, if there is a more complex series of base polymorphisms or when looking for multiple sequence variations, for example in a whole genome, the number of probe variants required becomes prohibitive. Re-sequencing arrays of this type have been used by Affymetrix for bacterial genomes.

The second (loss-of-signal) approach scores the observed signal of the sample DNA compared to the control DNA which is identical to the probes. However, a reduction in signal observed for a 20 bp probe would only show that a base difference was within that 20 bp region. Consequently, a series of overlapping probes are designed to cover the entire genome/region of interest, ideally such that each probe is overlapped by the adjacent probe by all but 1 bp. The pattern of reduced signal for overlapping probes allows the position of SNPs to be identified. It also has the added advantage that the site of possible SNPs does not have to be known a priori.

A combination of the loss-of-signal and gain-of-signal approaches is currently commercially exploited by NimbleGen (<http://www.nimblegen.com>) for both SNP detection and whole genome re-sequencing of bacterial genomes which are > 99% identical to a known reference genome. The first round of analysis uses loss of signal for a whole genome tiling path with probes that overlap. This identifies the general region of variation to within ~10bp. The effect of reducing the degree of overlap is to significantly reduce the number of probes, and therefore the cost, required to cover the sequence. For the second round of the analysis, a much smaller array is designed with probes representing all the possible sequence variations for that limited region. Gain-of-signal is then used to identify the specific base changes.

Alternatively, it is possible to use array-based mini-sequencing to identify SNPs. Specific probes adjacent to specific sites for potential nucleotide variation are coupled to the array by the 5' end. The DNA sample hybridises to the array probes, and DNA polymerase with fluorescent nucleotides is used to extend the probe sequence by one base-pair using the unlabelled DNA sample as a template. Since the position of each probe on the array is known

and the colour of each fluorescent nucleotide is different, the nature of the added base and, therefore, any possible sequence variation, can be determined. The strength of this technique is that it is possible to assay a complex mixture of base variations concurrently.

3.5 Bioinformatics

Bioinformatics - the process of storing, accessing and analysing biological data using mathematical tools - is most commonly applied to sequence data. It is a large and rapidly expanding field, and will fundamentally underpin most of the technologies for detection and identification of disease agents that are discussed in this review.

At the most basic level, much sequence analysis consists of more or less sophisticated ways to compare unknown sequences against databases of known sequences in order to identify them and/or determine their function. This is used in disease identification mainly through sequencing of regions of DNA that are common to large groups of organisms and comparing these to known sequences. The most common example is the 16S ribosomal RNA gene, a central component of the cellular machinery in all bacteria. The sequence is conserved enough to be amplified with a standard set of degenerate PCR primers, but variable enough to differentiate at the species level. The 16S sequence is therefore used as a molecular "bar code" to identify bacterial species. Search programs, such as BLAST (Altschul et al. 1990), are used to compare sequences to databases in order to identify species (<http://rdp.cme.msu.edu/>; (Cole et al. 2005)). New sequences of identified or unidentified organisms are used to build up these databases, which contain very large numbers of sequences; the 16S database currently contains over 200,000 sequences.

Several of the advances discussed here necessitate more complex tools. For example, large scale 'shotgun' 16S analysis of populations (Eckburg et al. 2005) and survey sequencing of unselected chromosomal sequences from bacterial populations (Tringe et al. 2005) produce a mixture of known and unknown sequences that require much more than a simple comparison to extract data on members of the population. Approaches using unsupervised clustering can be used to detect patterns in this kind of data, which may enable diagnostic information to be extracted without a full knowledge of all the underlying data (Boldrick et al. 2002; Rubins et al. 2004).

Identifying organisms based on random sequences requires searches against the general sequence databases, which are very large, and undergoing exponential growth. The three main databases; GenBank (US), EMBL (EU) and DDBJ (Japan), exchange data daily and currently contain over 123,000,000,000 bases in 67,000,000 records. Sensitive searching of such databases in a reasonable time requires high-performance hardware and software, and continued use of these databases will require continued increases in both. Much bioinformatics research is driven by this need to handle continuously increasing data sets.

Other applications discussed here require conceptual advances in data interpretation, as well as data handling. For example, bacterial populations are not strictly clonal; different groups of bacteria exchange DNA to greater or lesser extents, and understanding bacterial evolution and epidemiology requires the development of specialised methods to analyse bacterial population genetics (Feil 2004).

In addition, bioinformatics enables data exchange and integration. Many technologies, such as MLST (Urwin & Maiden 2003), require integration across countries and groups, and this is often driven by advances in web and network technologies. Further development of these integration mechanisms will be of fundamental importance in realising the promise of genomic detection and identification in the field.

4. Applications

Most early typing techniques do not rely on genomic information, except in the most peripheral way. These usually detect variations/differences in cell surface structures, often using specific antibodies (serotyping) or bacterial viruses (phage typing). Such techniques can be difficult to use reproducibly, both within and between laboratories. Recent advances in typing strategies have predominantly followed a path of getting closer to the genome sequence itself. Predominant among these are techniques based on cleaving the genomic DNA with restriction enzymes which cut the DNA at specific sites, and separating the resulting fragments by electrophoresis (pulsed-field gel electrophoresis: PFGE). Changes in the genome are reflected in changes in the pattern of lengths of fragments (restriction-fragment length polymorphism: RFLP), allowing strains to be differentiated. In some cases, fragments containing specific genes are labelled radioactively. Detecting fragments containing ribosomal genes is a common technique termed 'ribotyping'.

Detection and differentiation of microbial agents of disease can both be considerably improved by PCR. This can be used in many ways, both to identify particular organisms, by using specific PCR reactions, and to differentiate between strains or organisms, using discriminatory reactions. For example, specific reactions can be used to identify *Staphylococcus aureus* (SA), and more discriminatory PCR reactions used to detect which of these are Methicillin Resistant (MRSA). Real-time PCR enables detection to be more quantitative and accurate. For finer typing, a system has been developed whereby seven specific genes are PCR-amplified and then sequenced. Small variations within these sequences allow individual strains to be identified and tracked. This process, known as Multi-locus Sequence Typing (MLST; (Urwin & Maiden 2003)), allows identification to be exact and consistent, and it can be used world-wide, with information exchanged on the internet. Such a standardised system allows truly global strain identification, emphasising the requirement for a unified nomenclature for organisms and genes.

Such sequenced-based detection and identification, currently in use in the laboratory and beginning to expand into clinical practice, will form the platform

for many of the high-throughput technologies described above. Advances in miniaturisation and integration of PCR systems, along with increases in throughput (and decrease in cost) of Sanger-type sequencing will allow sequenced-based detection and identification to become the standard mechanism for clinical investigation.

However, the economics of medical intervention are complex. The need to train medics and medical scientists, and to introduce new hardware throughout the system, means that such advances take time to become generally and consistently utilised. It could, therefore, be envisaged that these technologies might come on stream within the next 5-15 years.

It is likely that hybridisation (microarray) and PCR-based technology will supplement or replace sequencing technologies for identification and characterisation of known agents in the medium term (perhaps 15-25 years) in clinical settings. These technologies require information on the sequence of known agents, but are likely to be more readily produced in kit form for standardised machines, or in hand-held devices. Initial steps towards this have been taken at the laboratory level, with specific microarrays used for identification of viruses in the central nervous system (Boriskin et al. 2004), and broad 16S arrays used for identification of microbial species in mixed populations (Palmer et al. 2006). Greater use of sequencing, re-sequencing and microarray-based comparative genomic hybridisation in the research arena will massively increase knowledge of microbial variation and epidemiology. This will allow the design of detection and identification systems based on such prior knowledge.

Beyond this, novel sequencing technologies, based on very-high-throughput random sequencing or single-molecule sequencing, may well become the method of choice for identifying and classifying disease agents. These require no prior knowledge of individual sequences and should allow unusual or completely novel agents to be identified, irrespective of diagnosis. Incorporating this technology into clinical laboratories or field devices would be a great advance for both medicine and research.

The continuation of large-scale microbial genome sequencing, particularly within-species comparative sequencing, has begun to identify some general evolutionary pathways followed by certain recently-evolved human pathogens.

Current studies have identified plasmids and other mobile chromosomal elements that are associated with acquired drug resistance, adaptation to specific pathogenic niches and a heightened pathogenic potential. While it may be some time before this becomes predictive, it can be envisaged that such sequences could be monitored for their occurrence in novel chromosomal contexts, initially through broad microarray systems that contained core genes from a number of target microbes, coupled with pathogenicity and resistance genes. For example, specific combinations of hybridisation signals would identify known virulence genes in bacteria not previously seen to contain them. However, it should be emphasised that microbes are extremely variable; a recent study of several strains of *Streptococcus agalactiae* suggested that, while the core genes of the genome

are conserved across strains, the number of different accessory (variably present) genes within the whole species may be effectively unlimited (Tettelin et al. 2005).

The increase in whole genome sequencing suggested above would allow much more than just individual diagnoses, although it would be essential for detection devices to be connected to a centralised database, ideally in real-time. This would allow up-to-date diagnosis, but, more importantly, collecting and monitoring the genome sequences of organisms in the field should facilitate the study of evolutionary processes as they occur.

A globally connected system would allow the construction of the very large database of base-line variation from which significant deviations could be detected, such as those driven by the introduction of a new therapeutic agent or the cross species jump of a new or potential biological threat. A possible application where continuous and, perhaps, even fully automated monitoring would be useful is in identifying new variants or recombinants of the influenza virus. In this way, it might be possible to provide advanced warning of epidemics of novel or drug-resistant diseases. Initial steps toward this have again been taken with microarray systems; for example a contribution to the initial identification of SARS as a new coronavirus was made by hybridisation of the DNA against a microarray containing sequences from many different viral families (Ksiazek et al. 2003; Wang et al. 2003).

One aspect that will require careful consideration is that of the normal human microbiota and the effect this has on sampling. While some sites in the human body, such as blood or urine, are nominally sterile, many, such as the gut or oral cavity, carry large numbers of bacteria. Current estimates suggest that the mouth may carry upwards of 800 bacterial species (William Wade, pers comm), and that the gut is home to around 7×10^{13} bacteria (Whitman et al. 1998). In general, these bacteria are not harmful. In fact many of the species in the gut are beneficial for much of its function. Current techniques involving culture or PCR are discriminatory, however; they will isolate or identify relevant organisms from amongst this heterogeneous milieu.

The potential technologies discussed here, involving direct sequencing of samples, may need to encompass some form of selection or purification in order to identify the signal from the noise. Sampling sterile fluids (blood or urine) may solve this problem for some diseases, but many pathogens cause acute disease while remaining undetectable in these bodily fluids. This does present a potential problem, as any method of selection or purification will only serve to identify known organisms and will fail to detect novel or unusual isolates that could be identified with a more holistic methodology. In the end, this may perhaps mean a more brute-force approach with increased throughput and decreased cost, potentially allowing identification of all organisms existing in any given person/environment, and the correlation of this knowledge with all aspects of health and disease.

5. Conclusion

Molecular techniques for detection and identification of infectious disease agents based on knowledge of genomic sequences, and using techniques to directly access DNA sequence, are common now, and are likely to become standard in the near future. Technological advances in the techniques and the hardware will drive the increases in throughput, automation and miniaturisation that will enable these technologies to be used and interconnected at all levels from the laboratory to the field.

6. Glossary

BLAST	Basic Local Alignment Search Tool – a program for comparing DNA sequences
clone	A piece of foreign DNA spliced into a plasmid for sequencing or manipulation
DNA	Deoxyribose Nucleic Acid – the storage mechanism for genetic information
MLST	Multi Locus Sequence Typing – a method to differentiate within bacterial species by sequencing short regions of specific genes
PCR	Polymerase Chain Reaction – a method for amplifying specific segments of DNA
PFGE	Pulse Field Gel Electrophoresis – a process for separating DNA fragments on agarose gels
plasmid	A small self-replicating DNA molecule that grows within bacterial cells
RFLP	Restriction Fragment Length Polymorphism – a method for identifying differences in fragmented bacterial chromosomes
RNA	Ribose Nucleic Acid – a form of nucleic acid used for structural and messenger functions in the cell.
SNP	Single Nucleotide Polymorphism – single base differences in DNA between individuals

7. References

Aborn, J. H., El-Difrawy, S. A., Novotny, M., Gismondi, E. A., Lam, R., Matsudaira, P., McKenna, B. K., O'Neil, T., Streechon, P. & Ehrlich, D. J. (2005) *A 768-lane microfabricated system for high-throughput DNA sequencing*. *Lab Chip* 5, 669-74.

- Achtman, M., Morelli, G., Zhu, P., Wirth, T., Diehl, I., Kusecek, B., Vogler, A. J., Wagner, D. M., Allender, C. J., Easterday, W. R., Chenal-Francois, V., Worsham, P., Thomson, N. R., Parkhill, J., Lindler, L. E., Carniel, E. & Keim, P. (2004) *Microevolution and history of the plague bacillus, Yersinia pestis*. Proc Natl Acad Sci U S A 101, 17837-42.
- Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990) *Basic local alignment search tool*. J Mol Biol 215, 403-10.
- Andras, S. C., Power, J. B., Cocking, E. C. & Davey, M. R. (2001) *Strategies for signal amplification in nucleic acid detection*. Mol Biotechnol 19, 29-44.
- Andries, K., Verhasselt, P., Guillemont, J., Gohlmann, H. W., Neefs, J. M., Winkler, H., Van Gestel, J., Timmerman, P., Zhu, M., Lee, E., Williams, P., de Chaffoy, D., Huitric, E., Hoffner, S., Cambau, E., Truffot-Pernot, C., Lounis, N. & Jarlier, V. (2005) *A diarylquinoline drug active on the ATP synthase of Mycobacterium tuberculosis*. Science 307, 223-7.
- Ansorge, W., Sproat, B. S., Stegemann, J. & Schwager, C. (1986) *A non-radioactive automated method for DNA sequence determination*. J Biochem Biophys Methods 13, 315-23.
- Boldrick, J. C., Alizadeh, A. A., Diehn, M., Dudoit, S., Liu, C. L., Belcher, C. E., Botstein, D., Staudt, L. M., Brown, P. O. & Rielman, D. A. (2002) *Stereotyped and specific gene expression programs in human innate immune responses to bacteria*. Proc Natl Acad Sci U S A 99, 972-7.
- Boriskin, Y. S., Rice, P. S., Stabler, R. A., Hinds, J., Al-Ghusein, H., Vass, K. & Butcher, P. D. (2004) *DNA microarrays for virus detection in cases of central nervous system infection*. J Clin Microbiol 42, 5811-8.
- Cole, J. R., Chai, B., Farris, R. J., Wang, Q., Kulam, S. A., McGarrell, D. M., Garrity, G. M. & Tiedje, J. M. (2005) *The Ribosomal Database Project (RDP-II): sequences and tools for high-throughput rRNA analysis*. Nucleic Acids Res 33, D294-6.
- De la Vega, F. M., Lazaruk, K. D., Rhodes, M. D. & Wenz, M. H. (2005) *Assessment of two flexible and compatible SNP genotyping platforms: TaqMan SNP Genotyping Assays and the SNPLex Genotyping System*. Mutat Res 573, 111-35.
- Detter, J. C., Jett, J. M., Lucas, S. M., Dalin, E., Arellano, A. R., Wang, M., Nelson, J. R., Chapman, J., Lou, Y., Rokhsar, D., Hawkins, T. L. & Richardson, P. M. (2002) *Isothermal strand-displacement amplification applications for high-throughput genomics*. Genomics 80, 691-8.
- Dickson, D. (2000) *Affymetrix loses first round of patent battle*. Nature 404, 697.
- Easterday, W. R., Van Ert, M. N., Zanecki, S. & Keim, P. (2005) *Specific detection of Bacillus anthracis using a TaqMan mismatch amplification mutation assay*. Biotechniques 38, 731-5.

Eckburg, P. B., Bik, E. M., Bernstein, C. N., Purdom, E., Dethlefsen, L., Sargent, M., Gill, S. R., Nelson, K. E. & Relman, D. A. (2005) *Diversity of the human intestinal microbial flora*. Science 308, 1635-8.

Feil, E. J. (2004) *Small change: keeping pace with microevolution*. Nat Rev Microbiol 2, 483-95.

Fodor, S. P., Read, J. L., Pirrung, M. C., Stryer, L., Lu, A. T. & Solas, D. (1991) *Light-directed, spatially addressable parallel chemical synthesis*. Science 251, 767-73.

Franklin, R. E. & Gosling, R. G. (1953) *Molecular configuration in sodium thymonucleate*. Nature 171, 740-1.

Hacia, J. G. (1999) *Resequencing and mutational analysis using oligonucleotide microarrays*. Nat Genet 21, 42-7.

Hapmap_Consortium (2003) *The International HapMap Project*. Nature 426, 789-96.

Heiger, D. N., Cohen, A. S. & Karger, B. L. (1990) *Separation of DNA restriction fragments by high performance capillary electrophoresis with low and zero crosslinked polyacrylamide using continuous and pulsed electric fields*. J Chromatogr 516, 33-48.

Higgins, J. A., Nasarabadi, S., Karns, J. S., Shelton, D. R., Cooper, M., Gbakima, A. & Koopman, R. P. (2003) *A handheld real time thermal cycler for bacterial pathogen detection*. Biosens Bioelectron 18, 1115-23.

Hinds, D. A., Stuve, L. L., Nilsen, G. B., Halperin, E., Eskin, E., Ballinger, D. G., Frazer, K. A. & Cox, D. R. (2005) *Whole-genome patterns of common DNA variation in three human populations*. Science 307, 1072-9.

Hughes, T. R., Mao, M., Jones, A. R., Burchard, J., Marton, M. J., Shannon, K. W., Lefkowitz, S. M., Ziman, M., Schelter, J. M., Meyer, M. R., Kobayashi, S., Davis, C., Dai, H., He, Y. D., Stephaniants, S. B., Cavet, G., Walker, W. L., West, A., Coffey, E., Shoemaker, D. D., Stoughton, R., Blanchard, A. P., Friend, S. H. & Linsley, P. S. (2001) *Expression profiling using microarrays fabricated by an ink-jet oligonucleotide synthesizer*. Nat Biotechnol 19, 342-7.

Judson, H. F. (1996). *The eighth day of creation: makers of the revolution in biology*. Expanded edit, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York.

Jurinke, C., van den Boom, D., Cantor, C. R. & Koster, H. (2002) *Automated genotyping using the DNA MassArray technology*. Methods Mol Biol 187, 179-92.

Ksiazek, T. G., Erdman, D., Goldsmith, C. S., Zaki, S. R., Peret, T., Emery, S., Tong, S., Urbani, C., Comer, J. A., Lim, W., Rollin, P. E., Dowell, S. F., Ling, A. E., Humphrey, C. D., Shieh, W. J., Guarner, J., Paddock, C. D., Rota, P., Fields, B., DeRisi, J., Yang, J. Y., Cox, N., Hughes, J. M., LeDuc, J. W.,

- Bellini, W. J. & Anderson, L. J. (2003) *A novel coronavirus associated with severe acute respiratory syndrome*. N Engl J Med 348, 1953-66.
- Lipshutz, R. J., Fodor, S. P., Gingeras, T. R. & Lockhart, D. J. (1999) *High density synthetic oligonucleotide arrays*. Nat Genet 21, 20-4.
- Luckey, J. A., Drossman, H., Kostichka, A. J., Mead, D. A., D'Cunha, J., Norris, T. B. & Smith, L. M. (1990) *High speed DNA sequencing by capillary electrophoresis*. Nucleic Acids Res 18, 4417-21.
- Maxam, A. M. & Gilbert, W. (1977) *A new method for sequencing DNA*. Proc Natl Acad Sci U S A 74, 560-4.
- Notomi, T., Okayama, H., Masubuchi, H., Yonekawa, T., Watanabe, K., Amino, N. & Hase, T. (2000) *Loop-mediated isothermal amplification of DNA*. Nucleic Acids Res 28, E63.
- Palmer, C., Bik, E. M., Eisen, M. B., Eckburg, P. B., Sana, T. R., Wolber, P. K., Relman, D. A. & Brown, P. O. (2006) *Rapid quantitative profiling of complex microbial populations*. Nucleic Acids Res 34, e5.
- Pastinen, T., Kurg, A., Metspalu, A., Peltonen, L. & Syvanen, A. C. (1997) *Minisequencing: a specific tool for DNA analysis and diagnostics on oligonucleotide arrays*. Genome Res 7, 606-14.
- Prober, J. M., Trainor, G. L., Dam, R. J., Hobbs, F. W., Robertson, C. W., Zagursky, R. J., Cocuzza, A. J., Jensen, M. A. & Baumeister, K. (1987) *A system for rapid DNA sequencing with fluorescent chain-terminating dideoxynucleotides*. Science 238, 336-41.
- Rubins, K. H., Hensley, L. E., Jahrling, P. B., Whitney, A. R., Geisbert, T. W., Huggins, J. W., Owen, A., Leduc, J. W., Brown, P. O. & Relman, D. A. (2004) *The host response to smallpox: analysis of the gene expression program in peripheral blood cells in a nonhuman primate model*. Proc Natl Acad Sci U S A 101, 15190-5.
- Saiki, R. K., Gelfand, D. H., Stoffel, S., Scharf, S. J., Higuchi, R., Horn, G. T., Mullis, K. B. & Erlich, H. A. (1988) *Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase*. Science 239, 487-91.
- Sanger, F., Nicklen, S. & Coulson, A. R. (1977) *DNA sequencing with chain-terminating inhibitors*. Proc Natl Acad Sci U S A 74, 5463-7.
- Schultz, D. A. (2003) *Plasmon resonant particles for biological detection*. Curr Opin Biotechnol 14, 13-22.
- Shen, R., Fan, J. B., Campbell, D., Chang, W., Chen, J., Doucet, D., Yeakley, J., Bibikova, M., Wickham Garcia, E., McBride, C., Steemers, F., Garcia, F., Kermani, B. G., Gunderson, K. & Oliphant, A. (2005) *High-throughput SNP genotyping on universal bead arrays*. Mutat Res 573, 70-82.

- Singh-Gasson, S., Green, R. D., Yue, Y., Nelson, C., Blattner, F., Sussman, M. R. & Cerrina, F. (1999) *Maskless fabrication of light-directed oligonucleotide microarrays using a digital micromirror array*. *Nat Biotechnol* 17, 974-8.
- Smith, L. M., Sanders, J. Z., Kaiser, R. J., Hughes, P., Dodd, C., Connell, C. R., Heiner, C., Kent, S. B. & Hood, L. E. (1986) *Fluorescence detection in automated DNA sequence analysis*. *Nature* 321, 674-9.
- Southern, E. M. (2001) *DNA microarrays. History and overview*. *Methods Mol Biol* 170, 1-15.
- Stoughton, R. B. (2004) *Applications of DNA Microarrays in Biology*. *Annu Rev Biochem*.
- Swerdlow, H., Wu, S. L., Harke, H. & Dovichi, N. J. (1990) *Capillary gel electrophoresis for DNA sequencing. Laser-induced fluorescence detection with the sheath flow cuvette*. *J Chromatogr* 516, 61-7.
- Tettelin, H., Massignani, V., Cieslewicz, M. J., Donati, C., Medini, D., Ward, N. L., Angiuoli, S. V., Crabtree, J., Jones, A. L., Durkin, A. S., Deboy, R. T., Davidsen, T. M., Mora, M., Scarselli, M., Margarit y Ros, I., Peterson, J. D., Hauser, C. R., Sundaram, J. P., Nelson, W. C., Madupu, R., Brinkac, L. M., Dodson, R. J., Rosovitz, M. J., Sullivan, S. A., Daugherty, S. C., Haft, D. H., Selengut, J., Gwinn, M. L., Zhou, L., Zafar, N., Khouri, H., Radune, D., Dimitrov, G., Watkins, K., O'Connor, K. J., Smith, S., Utterback, T. R., White, O., Rubens, C. E., Grandi, G., Madoff, L. C., Kasper, D. L., Telford, J. L., Wessels, M. R., Rappuoli, R. & Fraser, C. M. (2005) *Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: implications for the microbial "pan-genome"*. *Proc Natl Acad Sci U S A* 102, 13950-5.
- Tringe, S. G., von Mering, C., Kobayashi, A., Salamov, A. A., Chen, K., Chang, H. W., Podar, M., Short, J. M., Mathur, E. J., Detter, J. C., Bork, P., Hugenholtz, P. & Rubin, E. M. (2005) *Comparative metagenomics of microbial communities*. *Science* 308, 554-7.
- Tyson, G. W., Chapman, J., Hugenholtz, P., Allen, E. E., Ram, R. J., Richardson, P. M., Solovyev, V. V., Rubin, E. M., Rokhsar, D. S. & Banfield, J. F. (2004) *Community structure and metabolism through reconstruction of microbial genomes from the environment*. *Nature* 428, 37-43.
- Urwin, R. & Maiden, M. C. (2003) *Multi-locus sequence typing: a tool for global epidemiology*. *Trends Microbiol* 11, 479-87.
- Venter, J. C., Remington, K., Heidelberg, J. F., Halpern, A. L., Rusch, D., Eisen, J. A., Wu, D., Paulsen, I., Nelson, K. E., Nelson, W., Fouts, D. E., Levy, S., Knap, A. H., Lomas, M. W., Nealson, K., White, O., Peterson, J., Hoffman, J., Parsons, R., Baden-Tillson, H., Pfannkoch, C., Rogers, Y. H. & Smith, H. O. (2004) *Environmental genome shotgun sequencing of the Sargasso Sea*. *Science* 304, 66-74.

Vincent, M., Xu, Y. & Kong, H. (2004) *Helicase-dependent isothermal DNA amplification*. EMBO Rep 5, 795-800.

Wadman, M. (1998) *Human genome deadline cut by two years*. Nature 395, 207.

Wang, D., Urisman, A., Liu, Y. T., Springer, M., Ksiazek, T. G., Erdman, D. D., Mardis, E. R., Hickenbotham, M., Magrini, V., Eldred, J., Latreille, J. P., Wilson, R. K., Ganem, D. & DeRisi, J. L. (2003) *Viral discovery and sequence recovery using DNA microarrays*. PLoS Biol 1, E2.

Watson, J. D. & Crick, F. H. (1953) *Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid*. Nature 171, 737-8.

Whitman, W. B., Coleman, D. C. & Wiebe, W. J. (1998) *Prokaryotes: the unseen majority*. Proc Natl Acad Sci U S A 95, 6578-83.

Wilkins, M. H., Stokes, A. R. & Wilson, H. R. (1953) *Molecular structure of deoxypentose nucleic acids*. Nature 171, 738-40.

All the reports and papers produced within the Foresight project 'Infectious Diseases: preparing for the future,' may be downloaded from the Foresight website (www.foresight.gov.uk). Requests for hard copies may also be made through this website.

First published April 2006. Department of Trade and Industry. www.dti.gov.uk

© Crown copyright