Molecular Ecology Resources (2009) **9** (Suppl. 1), 51–57 doi: 10.1111/j.1755-0998.2009.02632.x

BARCODING METHODOLOGY AND APPLICATIONS

Biological agent detection technologies

JOHN P. JAKUPCIAK* and RITA R. COLWELL†
*CosmosID, 5010 River Hill Road, Bethesda, MD 20816, USA, †University of Maryland Institute for Advanced Computer Studies, University of Maryland, College Park, MD 20742, USA

Abstract

The challenge for first responders, physicians in the emergency room, public health personnel, as well as for food manufacturers, distributors and retailers is accurate and reliable identification of pathogenic agents and their corresponding diseases. This is the weakest point in biological agent detection capability today.

There is intense research for new molecular detection technologies that could be used for very accurate detection of pathogens that would be a concern to first responders. These include the need for sensors for multiple applications as varied as understanding the ecology of pathogenic micro-organisms, forensics, environmental sampling for detect-to-treat applications, biological sensors for 'detect to warn' in infrastructure protection, responses to reports of 'suspicious powders', and customs and borders enforcement, to cite a few examples.

The benefits of accurate detection include saving millions of dollars annually by reducing disruption of the workforce and the national economy and improving delivery of correct countermeasures to those who are most in need of the information to provide protective and/or response measures.

Keywords: barcoding, biological agent, detection, identification, sequencing

Received 1 October 2008; revisions received 23 December 2008/23 January 2009; accepted 27 January 2009

Introduction

The availability of sensitive and cost-effective diagnostic methods is paramount to the success of biological agent detection systems for public health protection and across industry sectors. Currently available methods are generally costly and have evolved basically from diagnostic development and applications. These methods range from cell culture to antibody (Mason *et al.* 2003; Gessler *et al.* 2007), polymerase chain reaction (PCR; Christensen *et al.* 2006), microarray (Brodie *et al.* 2007), and sequencing approaches (Margulies *et al.* 2005). Partial genome sequencing and comparison with known sequence data (requires a priori knowledge of the bioagent) is not effective, particularly since bioengineering makes it possible to modify organisms to be more infectious, better at avoiding immune responses or resistant to medical countermeasures (Jackson *et al.* 2001). Critical to the success of biothreat surveillance is the ability to screen for and detect multiple agents rapidly in a single reaction with minimal sample processing (Cirino *et al.* 2004).

Traditional microbial typing technologies employed for characterization of pathogenic micro-organisms and monitoring their global spread are often difficult to standardize, are poorly portable, and lack sufficient ease of use, throughput, and automation.

A survey published in 2006, reported that emergency and primary care physicians and their local health care systems were not well prepared to respond to potential disease outbreaks or biological attacks and many believed that more resources should be allocated to equipping a response system. These findings highlight the importance of expanding bioterrorism preparedness efforts to improve the public health system (Alexander *et al.* 2006). Other analyses of bioterrorism preparedness have recommended modernization of public health response to emergencies and investigations (The Century Foundation 2005). Similar analyses of technology gaps by groups, such as the Department of Energy and the Heritage Foundation emphasize the need to improve emergency response, regional coordination and technologies (Heritage Foundation 2006).

Correspondence: John P. Jakupciak,
E-mail: jjakupciak@cosmosid.us.com

Table 1 Technology comparison

Technique	Sensitivity	Specificity	Target class
Microparticle immunoassay	Generally high (Ab quality control issues) (Ab continuous supply issues)	Depends on antibodies (Ab). Sometimes cross-species.	Proteins, sugars, DNA bacteria, viruses, toxins
ELISA	Moderate	Depends on antibodies. Sometimes cross-species.	Proteins, sugars, DNA bacteria, viruses, toxins
PCR	High	Very specific with certain primers. Phylogenetic assessment	DNA/RNA
DNA microarray	High	Extremely specific 100s–1000s of targets identifiable	DNA/RNA
MALDI-TOF	High for some molecules	Very specific for small molecules, less for large	Masses of biomarker proteins

Comparison

While antibody-based approaches are very widespread, there are many well-recognized limitations, for example, low quality control of antibody. Antibody production is dependent, in part on cell culture and the lack of understanding of the extent of biological community complexity attributes to the limited use of methods relying on culture. Nonculturable organisms can only be identified by molecular genetic methods. Singleplex, multiplex and reverse transcriptase–PCR (RT–PCR) approaches have become standard methods for most laboratories. Furthermore, nucleic acid-based methods (PCR and sequencing) are more sensitive than antibody-based detection systems (Lim *et al.* 2005).

PCR-based methods have critical limitations, since they depend on a priori knowledge of what sequence to detect in a sample further complicated by recent demonstrations of greater variability in genomic sequence than expected. In addition, PCR probe sequence resolution erodes (Table 1, Detection methods comparison).

A platform for genome identification of a specimen from any source must not only be sensitive and specific, but must also detect a variety of pathogens with high accuracy, including modified or previously uncharacterized agents, and this challenge is daunting when identification must be achieved using nucleic acids in a complex sample matrix. The available devices and instruments are severely limited in the length of time required for analysis, the complexity of the process employed, and the lack of a systems approach, that is, from extraction to identification without preconceived notion of what may be in the sample (hybridization surfaces, chips or microarrays). It is widely understood by microbial ecologists and more recently, medical microbiologists, that the microbial species encompasses significant variability with, perhaps, a core genome incorporated in a wider, more variable genome. Furthermore, the first described strain is designated the 'type species,' but it may not (and probably is not) the median strain to serve as the reference strain for sequencing comparison (Colwell & Liston 1961a, b, c).

Specialists from the Federal Bureau of Investigation, experts from hospitals and university research centres and key opinion leaders from the private sector agree that the most effective approach for comprehensive genetic variation discovery is by sequencing (Budowle *et al*. 2005). Rapid advances in biological engineering have dramatically impacted the design and capabilities of DNA sequencing tools. These have led to an increase in the number of base pairs sequenced per day by more than 500-fold while the costs have decreased three orders of magnitude.

Despite these advances, extension of sequencing technology should include the capacity to distinguish naturally occurring micro-organisms from intentionally distributed pathogens. This represents a considerable challenge because the diversity of the microbial world is effectively unknown. DNA sequencing has the capacity to address this need.

Sequencing

The emergence of non-Sanger sequencing as well as microarrays and flow cells has changed the DNA sequencing landscape. Innovation of a genome identification technology meeting the goal of a '$1000 genome' is in progress and remains to be discovered. Very likely, the engineering challenges and hurdles will be overcome and the basic science for DNA processing and detection will eventually be done. Given the trends of the past, a 'eureka' vision is highly probable that will introduce new concepts within a few years.

The build-out of genome identification DNA sequencing technology in the form of practical instrumentation will be achieved by incorporating the critical requirements for accurate long reads, without dependency for template amplification, capable of manipulating terabytes of data to provide reliable and useful identification of genetic sequences within any unknown sample, whether clinical, environmental, or other type of specimen.

With advancement of DNA sequencing technology, molecular typing methods based on nucleic acid fingerprints or

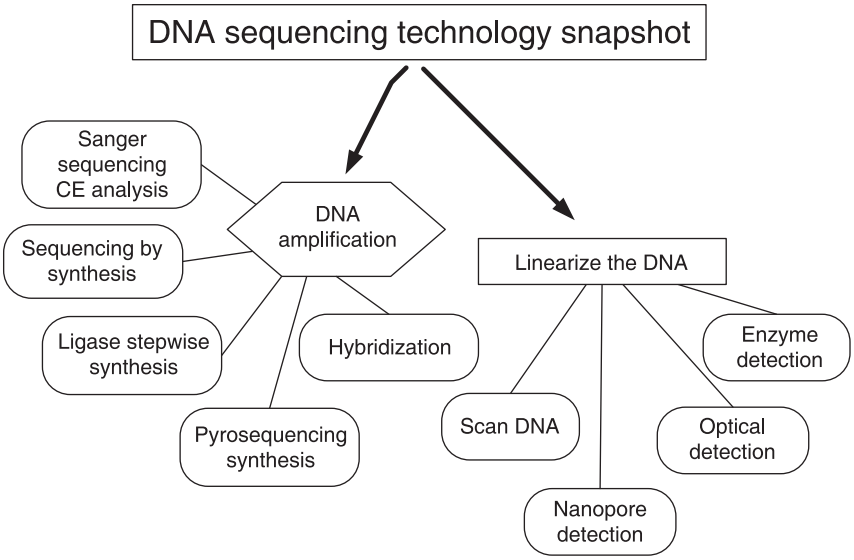

Fig. 1 DNA sequencing technology snapshot. CE, capillary electrophoresis.

'mini-sequencing' are currently being replaced by more sensitive genome-wide single nucleotide polymorphism-based methods, as exemplified by anthrax bacilli (Patil *et al.* 2001). *De novo* genome sequence determination is not the maximum capability of sequencing technology. The technology forecast includes extending the capability to perform metagenomic (environmental) sequencing. Rather than focusing on an individually isolated genome, environmental sequencing is aimed at sampling populations of genomes as found, for example, in bodies of water or different types of soil and within or on the surface of the human body. In this way, an analysis of sequence and gene diversity can be obtained from organisms that cannot be cultured using conventional techniques. Genome sequencing capability will facilitate the evaluation of 'deep' resequencing methods which compare different sources of DNA across one or a few genes. For example, K-ras gene mutations associated with cancer.

Types of sequencing

The most widely used technology for DNA sequencing is a capillary electrophoresis (CE)-based system employing the Sanger method. Although chemistries and automation advances have made Sanger-based DNA sequencing easier and faster, the basic technology remains the same. An immense challenge is that of managing the variety and complexity of data types, the hierarchy of biology, and the inevitable need to acquire data by a wide variety of modalities. Inexpensive DNA sequencing will revolutionize medicine by making personalized treatments possible. Rapid genome sequencing is regarded as the next great frontier for science that will allow doctors to determine individual susceptibility to disease and the genetic links to cancer or cardiovascular disease.

Sanger sequencing is a method introduced by Frederick Sanger (Sanger *et al.* 1977). It is remarkable that Sanger sequencing-based methods are so ubiquitous and long-lived. The flexibility of the method has been its strongest asset. Furthermore, the utility of using genomic DNA directly, that is, cloneless libraries greatly accelerated DNA analyses. Over the years, instrumentation for DNA sequencing has improved dramatically in terms of read length and throughput (Fig. 1).

DNA sequencing methods reaching the market include bead-based, microfluidic-based and microarray-based approaches as well as emerging concepts for sequencing, for example, nanopore-based sequencing (Rhee & Burns 2007). Unlike Sanger sequencing, the use of pyrosequencing for sequencing-by-synthesis does not require fluorescent labelling. Incorporation of each dNTP is accompanied by release of pyrophosphate, which is converted by sulfurylase into ATP, which leads to the release of light from the conversion of luciferase to oxyluciferin. However, asynchronistic extensions may occur because of slight variations in dispensing order of the dNTPs.

Bridge amplification in a flow cell represents another alternative to Sanger sequencing. It employs four-colour tagging, but uses forward-thinking, acid-labile reversible terminator chemistry.

Synthesis-by-hybridization defines a resequencing method. DNA microarrays are the modern, massively parallel version of classic molecular biology hybridization techniques. The technique permits analysis of genetic material (DNA) and monitoring of expression changes (RNA, really based on cDNA) occurring in a biological sample under various conditions. Microarrays have been used successfully in various research areas including DNA sequencing. While microarray-based approaches enable high-throughput, they are limited to genes present in reference databases and hence require pre-selected sequence information limiting what

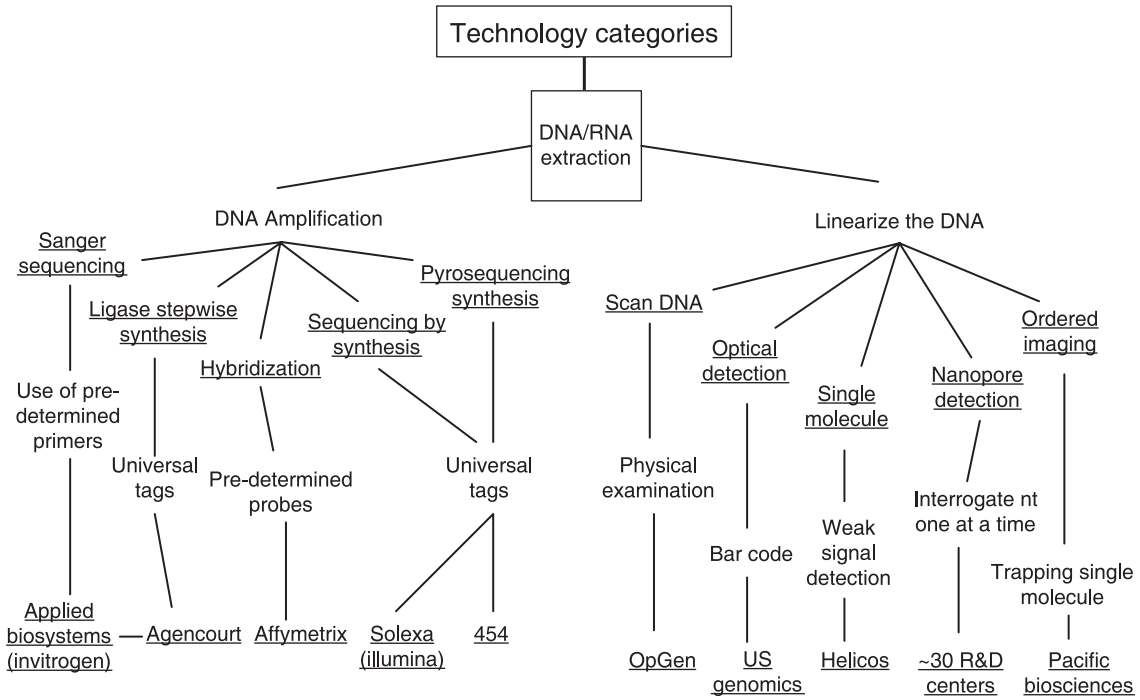

Fig. 2 Technology categories.

DNA samples can be interrogated (Fig. 2, DNA sequencing methods comparison).

Annotation

The genomes of most organisms, from the simplest unicellular organism to more complex species, consist of a variety of genomic landscapes. Each with a unique profile of genetic content, GC-richness, high-copy repeats and low-copy repeats (Bacolla *et al.* 2006). Particularly near regions of structurally important sequences, such as the centromere, the genomic landscape can become quite problematic for analysis. The primary factor contributing to the difficulty in studying these areas within the genome is that clusters of genomic duplications and unusual repeat structures often lie in close proximity. Consequently, sequence similarity-based methods of global genome assembly fail to properly assign the correct positions of duplicated sequences. Annotation and interpretation of data from current DNA sequencing technologies can be difficult, because artificial overlaps form, significant warping of working draft sequences occurs, and numerous gaps appear in the assembly, all which, reduce overall quality and relevance of the assembly. These effects are further compounded by the absence of unique sequence-tagged sites (STS) or DNA landmarks within such regions and a general under-representation of such areas in clone-by-clone sequencing. Thus, current DNA sequencing instruments are challenged by the presence of large spans of duplicated sequence, which interfere with genome analysis.

With faster methods to collect data, more attention has turned to sequence annotation. Researcher can quickly generate sequence information and want to annotate their data. An annotated sequence provides a wealth of information about the organism not directly obvious from the sequence. It also acts as a standard, giving investigators the ability to work on the same basic gene structures and to compare findings. The annotation process is a challenging task especially in light of the limited infrastructure and expertise.

New opportunities to develop faster and less expensive methods for sequencing DNA are pushing genome sequencing technology to include answers to:

- information on the nature and source of a sample
- effective data collection for comparison of samples (from known and probable locations)
- confidence in data comparison
- sample divergence from a common ancestor (the mechanism of the variation or the heterotachy of the mutation)
- genome-wide information analysis, including evolutionary distance to enable the measurement of the variation
- evidence of genetic engineering
- all or partial data exclusion as a contaminant or failure in sample handling.

A major obstacle to identification of micro-organisms is having a reference or comparison strain. For all named microbial species, there is a requirement for deposition of the culture in a collection so that others may have the reference

strain. Research over the last 50 years has shown that the reference strain rarely is the 'median' strain and often is an outlier to the species. Furthermore, recent genome sequence data show that strains passaged several times in media will display single nucleotide polymorphism (SNP) differences and multiple isolates of a single species will not be identical base pair by base pair. This reality must be dealt with in identifying pathogens, especially isolates or even the nucleic acid from a natural environment.

The high-throughput nature of the sequencing device inevitably produces sequencing errors and limits data quality. The errors in standard genome sequencing projects can be reduced by applying efficient genome assembly techniques. Potential sequencing errors can also be further minimized by posterior computational processes (Gajer *et al.* 2004; Huse *et al.* 2007).

Discussion

Detection examples

In October 2001, the use of the US mail system as a method for disseminating a weapon set off a national incident for biological agent detection. Traditional microbiological laboratory methods have served as the standard for identification of viable pathogens, but are time-consuming and labour intensive. The sheer number of pathogens and their complex biology, diversity and capacity to exchange genetic material complicates interpretation of current constrained data collection. Moreover, it is estimated that 99% of microbial species cannot be cultured (Torsvik *et al.* 1990; Amann *et al.* 1995; Zak & Visser 1996; Ranjard *et al.* 2000; Bridge & Spooner 2001; Anderson & Cairney 2004; Harayama *et al.* 2004). Even when culturing is effective, the long process is a limitation for investigators who need rapid sample analysis-to-answer.

To accomplish this, data acquisition must distinguish individual isolates from similar samples to the most precise level possible (ideally to a single source). An unexpected and deadly pathogen is severe acute respiratory syndrome (SARS) virus. The natural hosts of the virus are thought to be wild civets and bats (Guan *et al.* 2003; Li *et al.* 2005). The epidemic of SARS appears to have originated in Guangdong Province, China in November 2002, although the Chinese government did not disclose the information outside China nor informed World Health Organization of the outbreak of hitherto unknown infectious disease. The lack of a proper diagnostic method for SARS resulted in public health crisis in 2003 (Bloom 2003). Right after SARS was recognized as a potential threat to global health, several leading laboratories were brought to identify the causal agent. Initial electron microscopic examination in Hong Kong and Germany found viral particles with structures suggesting paramyxovirus in respiratory secretions of SARS patients (Hassler *et al.* 2003). In contrast, Chinese researchers reported that a chlamydia-like disease may be behind SARS.

Biodiversity

The mutation rates of rapidly evolving microbial genomes can be up to 1 in 100 000 bp. The production rate of viruses being as rapid as 10^{12} virions per day indicates a very high genetic diversity (Neumann *et al.* 1998) measurable because viral samples duplicated closely related in time are unlikely to harbour identical genomes! Furthermore, because of recombination, insertion sequences, rearrangements or gene duplications, the genome size of isolates from the same species can be different (Swiecicka 2003).

Following on the heels of the publication of the sequence analysis of human DNA comes the realization that there is more-than-expected variation (Sutton *et al.* 2007). This observation combined with the recent demonstration from the J. Craig Venter Institute (JCVI) of the first transplant of a bacterial genome heightens the need for genome identification technology. The technology must be able to distinguish human variation and to detect engineered organisms.

Genetic material transplant is not just gene transfer, but the transfer of an entire chromosome. Thus, the outer shell of the organism no longer represented its genomic content, the Trojan Horse of micro-organism population genetics. The surface appearance cloaks the hidden genetic content. In the case of SARS, detection was based on the physical appearance of the unknown biothreat, but today and in the future, organism identification will require genome identification. Genome transplantation is an essential enabling step in the field of synthetic genomics as it is a key mechanism by which chemically synthesized chromosomes can be activated within a cell. The ability to transfer naked DNA isolated from one species into a second microbial species paves the way for subsequent experiments to transplant a fully synthetic bacterial chromosome into a living organism.

The way forward

Current genome identification strategies are based on the use of a reference to make a detection of the next unknown. Forward-looking strategies encompass not just the detection of a target or key signature, but also the characterization of the population and the identification of genomes in a mixture in an environment. Approaches based on antibodies, PCR-probes, microarray-probes can only capture information on predicted answers.

Sequencing based on a metagenomic approach has established the capability to correlate the traceability or identification of an organism as a causative agent (Constantin *et al.* 2008) and can further distinguish between nontoxicgenic and pathogenic versions (Mohapatra *et al.* 2008). On account of biodiversity being significantly large between and within

species, sequencing enables the solution for bioweapon detection, pathogen identification, and predictions of disease outbreaks; true personalized medicine. At first glance, the extent of biodiversity could seem to be too much of a challenge to fully characterize; however, sequencing technologies can measure that diversity and accurately bin and distinguish organisms based on population genetics principles, mutation rates, genome stability, host interactions, genetic mobility, microbial ecology and population dynamics. Comparative genomics will provide genome identification because it is based on identification of populations and not on a technique that tries to find the best match to a pre-defined reference.

Conclusion

The genome revolution leads far beyond an instrument to meet the National Human Genome Research Institute goal of a complete human genome for $1000. This goal challenges the technical community and is machine independent, and does not include the more challenging specifications required of the device to identify micro-organisms. It is likely, that recent market trends that have resulted in a large number of diverse approaches to sequencing represented by a variety of commercial and academic research centres, which are faced with dedicating teams of scientists/technicians and engineers to accomplish their singular goal, will move in the opposite direction. Nevertheless, we can expect to consolidate the independent efforts into large collaborative efforts over the next several years with sharper focus on the identification of all life forms and characterization of their populations.

Conflict of interest statement

The authors have no conflict of interest to declare and note that the funders of this research had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

References

Alexander GC, Larkin GL, Wynia MK (2006) Physicians preparedness for bioterrorism and other public health priorities. *Academic Emergency Medicine*, **13**, 1238–1241.

Amann RI, Ludwig W, Schleifer KH (1995) Phylogenetic identification and in situ detection of individual microbial cells without cultivation. *Microbial Reviews*, **59**, 143–169.

Anderson IC, Cairney JWG (2004) Diversity and ecology of soil fungal communities. *Environmental Microbiology*, **6**, 769–779.

Bacolla A, Collins JR, Gold B *et al.* (2006) Long homopurine homopyrimidine sequences are characteristic of genes expressed in brain and the pseudoautosomal region. *Nucleic Acids Research*, **34**, 2663–2675.

Bloom BR (2003) Lessons from SARS. *Science*, **300**, 701.

Bridge P, Spooner B (2001) Soil fungi: diversity and detection. *Plant and Soil*, **232**, 147–154.

Brodie EL, DeSantis TZ, Parker JP, Zubietta IX, Piceno YM, Andersen GL (2007) Urban aerosols harbor diverse and dynamic bacterial populations. *Proceedings of the National Academy of Sciences, USA*, **104**, 299–304.

Budowle B, Johnson MD, Fraser CM, Leighton TJ, Murch RS, Charkraborty R (2005) Genetic analysis and attribution of microbial forensics evidence. *Critical Reviews in Microbiology*, **31**, 233–254.

Christensen DR, Hartman LJ, Loveless BM *et al.* (2006) Detection of biological threat agents by real-time PCR: comparison of assay performance on the RAPID, the Lightcycler, and the Smart cycler platforms. *Clinical Chemistry*, **52**, 141–145.

Cirino NM, Musser KA, Egan C (2004) Multiplex diagnostic platforms for detection of biothreat agents. *Expert Review of Molecular Diagnostics*, **4**, 841–857.

Colwell RR, Liston J (1961a) Taxonomy of Xanthomonas and pseudomonas. *Nature*, **191**, 617–619.

Colwell RR, Liston J (1961b) Taxonomic analysis with the electronic computer of some Xanthomonas and pseudomonas species. *Journal of Bacteriology*, **82**, 913–919.

Colwell RR, Liston J (1961c) Taxonimic relationships among the pseudomonads. *Journal of Bacteriology*, **82**, 1–14.

Constantin de Magny G, Murtugudde R, Sapiano MR *et al.* (2008) Environmental signatures associated with cholera epidemics. *Proceedings of the National Academy of Sciences, USA*, **105**, 17676–17681.

Gajer P, Schatz M, Salzberg SL (2004) Automated correction of genome sequence errors. *Nucleic Acids Research*, **32**, 562–569.

Gessler F, Pagel-Wieder S, Avondet MA, Bohnel H (2007) Evaluation of lateral flow assays for the detection of botulinum neurotoxin type A and their application in lab diagnosis of botulism. *Diagnostic Microbiology and Infectious Disease*, **57**, 243–249.

Guan Y, Zheng BJ, He YQ *et al.* (2003) Isolation and characterization of viruses related to the SARS coronavirus from animals in southern China. *Science*, **302**, 276–278.

Harayama S, Kasai Y, Hara A (2004) Microbial communities in oil-contaminated seawater. *Current Opinions in Biotechnology*, **15**, 205–214.

Hassler D, Schwarz TF, Braun R (2003) SARS: a new paramyxovirus or coronavirus? *Deutsche Medizinische Wochenschrift*, **128**, 786.

Heritage Foundation. Empowering America: A proposal for enhancing regional preparedness (2006) Apr. 7. http://www.heritage.org/Research/HomelandSecurity/SR06.cfm (accessed December 22, 2008).

Huse SM, Huber JA, Morrison HG, Sogin ML, Welch MD (2007) Accuracy and quality of massively-parallel DNA pyrosequencing. *Genome Biology*, **8**, R143.

Jackson RJ, Ramsay AJ, Christensen CD, Beaton S, Hall DF, Ramshaw IA (2001) Expression of mouse interleukin by a recombinant ectromelia virus suppresses cytolytic lymphocyte responses and overcomes genetic resistance to mousepox. *Journal of Virology*, **75**, 1205–1210.

Li W, Shi Z, Yu M *et al.* (2005) Bats are natural reservoirs of SARS-like coronaviruses. *Science*, **310**, 676–679.

Lim DV, Simpson JM, Kearns EA, Kramer MF (2005) Current and developing technologies for monitoring agents of bioterrorism and biowarfare. *Clinical Microbiology Reviews*, **18**, 583–607.

Margulies M, Egholm M, Altman WE *et al.* (2005) Genome sequencing in open microfabricted high density picoliter reactions. *Nature*, **15**, 376–380.

Mason HY, Lloyd C, Dice M, Sinclair R, Ellis Jr W, Powers L (2003) Taxonomic identification of microorganisms by capture and intrinsic fluorescence detection. *Biosensors and Bioelectronics*, **18**, 521–527.

Mohapatra SS, Ramachandran D, Mantri CK, Colwell RR, Singh DV (2008) Determination of relationships among non-toxigenic Vibrio cholerae O1 biotype El Tor strains from housekeeping gene sequences and ribotype patterns. *Research in Microbiology*, (online early article).

Neumann AU, Lam N, Dahari H *et al*. (1998) Hepatitis C viral dynamics In Vivo and the anti-viral efficacy of interferon therapy. *Science*, **282**, 103–107.

Patil N, Berno AJ, Hinds DA *et al*. (2001) Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. *Science*, **294**, 1719–1723.

Ranjard L, Poly F, Nazaret S (2000) Monitoring complex bacterial communities using culture-independent molecular techniques. *Research Microbiology*, **151**, 167–177.

Rhee M, Burns MA (2007) Nanopore sequencing technology: nanopore preparations. *Trends in Biotechnology*, **25**, 174–181.

Sanger F, Nicklen S, Coulson AR (1977) DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences, USA*, **74**, 5463–5467.

Sutton G, Ng P, Feuk L *et al*. (2007) The diploid genome sequence of an individual human. *PLoS Biology*, **5**, e254.

Swiecicka I (2003) Molecular typing by pulsed-field gel electrophoresis of *Bacillus thuringiensis* from root voles. *Current Microbiology*, **46**, 256–260.

The Century Foundation 'Are bioterrorism dollars making us safer?' (2005) Jan 13. www.tcf.org. http://centuryfoundation.org/print.asp?type=PR&pubid=44 (accessed December 22, 2008)

Torsvik V, Goksoyr J, Daae FL (1990) High diversity in DNA of soil bacteria. *Applied and Environmental Microbiology*, **56**, 782–787.

Zak JC, Visser S (1996) An appraisal of soil fungal biodiversity. *Biodiversity and Conservation*, **5**, 169–183.