Chromosome Research **13**: 225–235, 2005.

Elucidating genetic diversity with oligonucleotide arrays

Claire Kidgell[1] & Elizabeth A. Winzeler[1*,2]

[1] *Department of Cell Biology, The Scripps Research Institute, 10550 North Torrey Pines Road, ICND202, La Jolla, CA 92037, USA; Tel: 858 784 9468; Fax: 858 784 9860; E-mail: winzeler@scripps.edu;* [2] *Genomics Institute of the Novartis Research Foundation, San Diego, CA 92121, USA*

* *Correspondence*

Key words: comparative genomics, DNA hybridization, genetic diversity, oligonucleotide arrays

Abstract

DNA microarrays, initially designed to measure gene expression levels, also provide an ideal platform for determining genetic diversity. Oligonucleotide microarrays, predominantly high-density oligonucleotide arrays, have emerged as the principal platforms for performing genome-wide diversity analysis. They have wide-ranging potential applications including comparative genomics, polymorphism discovery and genotyping. The identification of inheritable genetic markers also permits the analysis of quantitative traits, population studies and linkage analysis. In this review, we will discuss the application of oligonucleotide arrays, in particular high-density oligonucleotide arrays for elucidating genetic diversity and highlight some of the directions that the field may take.

Elucidating the diversity of genes between and within individual species forms the basis of understanding the evolution and adaptation of an organism. The level of similarity (homogeneity) or difference (heterogeneity) within a species population indicates the diversity of the gene pool. This diversity can be assessed through the observation of expressed genetic traits (phenotype) or by determining the variants of individual genes (genotype). Such information can then be related to understanding the molecular basis of pathogenesis and disease transmission, for example, as well as deciphering the evolutionary history of a species. Understanding genetic variation is also important in the analysis of an organism's transcriptome, where a change in gene expression may result from coding sequence variability rather than differential gene regulation.

Genetic diversity can be introduced by either mutation or recombination. Mutation is a change in the DNA sequence of a gene within an organism. A single nucleotide change, or polymorphism, results in either a change in amino acid (non-synonymous) or a silent mutation, where no change in amino acid occurs (synonymous). Recombination produces different combinations of alleles as a result of the physical exchange of DNA between two different chromosomes in the case of higher organisms, or between another isolate or closely related species in the case of prokaryotic organisms. Other sources of genetic variation can be sequence rearrangements, genetic insertions and/or deletions. In all cases, the frequencies at which these events occur within a particular species are influenced by biological and ecological factors and help drive the evolutionary processes within the organism.

Molecular approaches to assess genetic diversity

The genetic analysis of organisms, both prokaryotic and eukaryotic, is a fast-moving and expanding scientific field. Population diversity can be measured by a variety of techniques, with the simplest method being an assessment of the observed phenotype of a population, such as eye colour or hair colour. However, for those species where the expressed phenotypes are not readily determined, biochemical and molecular approaches are needed for a more in-depth analysis of genetic relatedness. At the moment there is a wide-ranging number of approaches available for identifying genetic diversity or relatedness in a population. These include Pulse Field Gel Electrophoresis (PFGE) (Beadle *et al.* 2003), microsatellite mapping (Dearlove 2002), along with more recent techniques such as mass spectrometry-based genotyping (Pusch *et al.* 2002). However, while traditional techniques, such as PFGE, are highly discriminatory and have added greatly to the study of genetic diversity, they do not illustrate the total variability of the organism in question. The only way to show total variability is to sequence the genome. The advances in whole genome DNA sequencing within the last few years have permitted considerable progress towards the assessment of genetic diversity within certain organisms (Bentley & Parkhill 2004). Comparative genomics can facilitate a detailed catalogue of the biological similarities and differences between species, revealing fascinating insights into the genome evolution and biology of numerous organisms (Herrero *et al.* 2003, Nelson *et al.* 2004, McClelland *et al.* 2004). However, typically, only one strain or individual of a particular species is sequenced, thereby limiting the extent of analysis possible by this whole-genome approach.

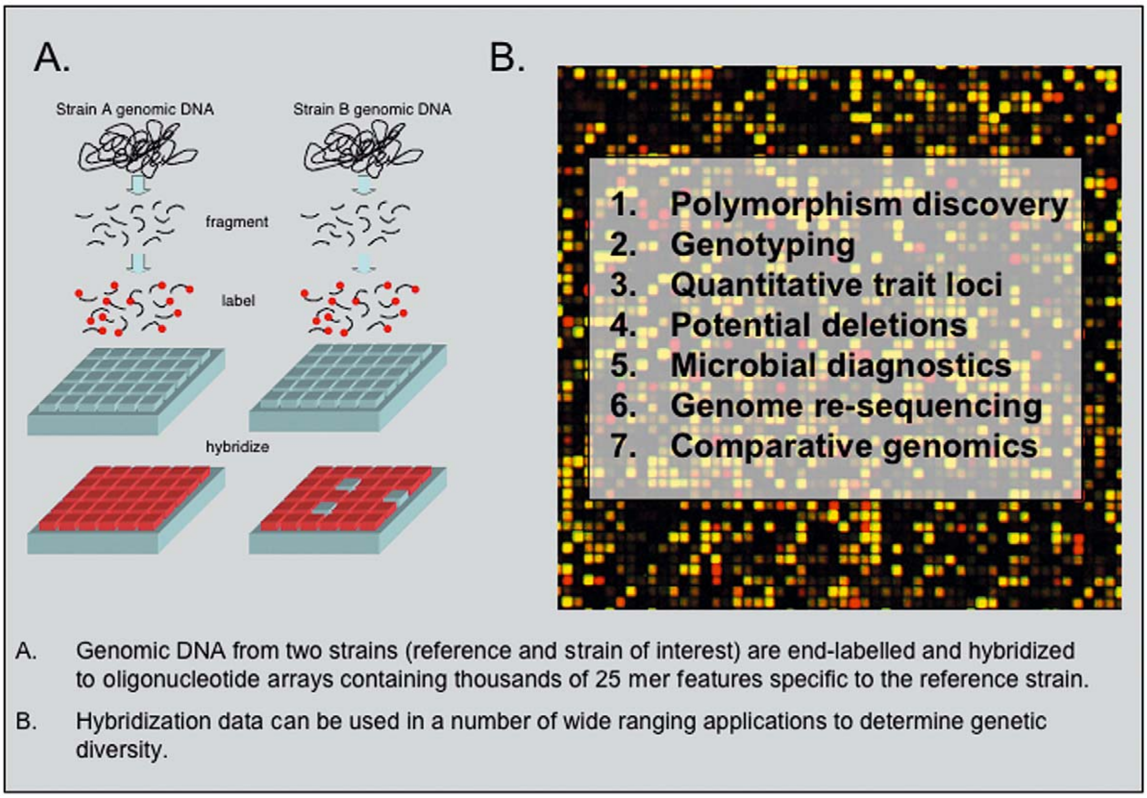
DNA microarrays

The availability of whole genome sequences from many organisms has directly influenced the development of DNA microarrays as rapid high-throughput molecular analysis platforms that are now commonplace in laboratory research. The microarray system is a powerful apparatus that has revolutionized functional and genomic analyses in a variety of species. A DNA microarray enables all or selected open reading frames from an annotated genome to be represented on a single microscope slide or window. Generally, there are two microarray technologies that dominate the field, the glass spotted DNA microarray and the high-density oligonucleotide array (Yauk *et al.* 2004).

Two methods can be used in the generation of the glass spotted microarray. Either PCR amplified open reading frames (ORFs) or cDNA clones from the genome of interest are robotically spotted onto poly-L-lysine-coated glass slides or long oligonucleotides (60–70 mers) are directly synthesized and subsequently also spotted onto poly-L-lysine slides. In contrast, a very different approach is taken in the manufacturing of high-density oligonucleotide arrays. Affymetrix (Affymetrix, Santa Clara, CA) have pioneered the production of the high-density array in which photolithographic chemistry is utilized to synthesize *in situ* short oligonucleotides, typically 25-mers directly onto the microarray platform (Lockhart & Winzeler 2000).

There are advantages and disadvantages associated with both techniques. Greater flexibility can be attained with the spotted microarray as specific PCR products can be easily generated and as such there is no fixed probeset. However, spotted PCR amplicons only facilitate a limited coverage of the genome and, due to the size of each PCR product representing an ORF, issues of cross hybridization can also arise. Inconsistencies in the hybridization properties of long oligonucleotide probes can also be an issue. Consequently, although the high-density oligonucletide array platform overcomes both of these issues, the photolithographic manufacturing process required to directly build the 25-mer feature onto the microarray platform significantly reduces the flexibility of this system. However, such an array produces better specificity, quality control and portability and the smaller feature size allows for significantly greater coverage of any genome. The increased number of probes for every ORF also permits novel observations that would be impossible using glass spotted microarrays.

Both the spotted DNA microarray and oligonucleotide array platforms can be successfully applied to genetically characterize a strain or species. However, the increased feature density and



1. Polymorphism discovery
2. Genotyping
3. Quantitative trait loci
4. Potential deletions
5. Microbial diagnostics
6. Genome re-sequencing
7. Comparative genomics

A. Genomic DNA from two strains (reference and strain of interest) are end-labelled and hybridized to oligonucleotide arrays containing thousands of 25 mer features specific to the reference strain.

B. Hybridization data can be used in a number of wide ranging applications to determine genetic diversity.

Figure 1. The hybridization of genomic DNA to high-density oligonucleotide arrays permits genome-wide observations of genetic diversity.

reproducibility offered by the high-density oligonucleotide arrays makes them the preferred platform for genomic analysis (Figure 1).

High-density oligonucleotide arrays: genomic analyses

Single nucleotide polymorphisms

The identification of single nucleotide polymorphisms (SNPs) within any genome can provide information as to the age and diversity of the organism in question. However, by only sampling particular regions of a genome, as in nucleotide sequencing, one can obtain a biased picture of the true extent of any such diversity. Although originally designed as a tool for gene expression analysis (Lockhart *et al.* 1996), hybridization of genomic DNA to an Affymetrix high-density oligonucleotide array can be used effectively for the genome-wide detection of variation in the alternative hybridizing strain relative to the reference strain.

Due to the short probe sequences (25-mer) used in the construction of the microarray, the alteration in hybridization signal caused by a single base change between the target and probe sequence can be readily identified. Consequently, because the exact location of each 25-mer probe in the genome is known, the position of these potential single feature polymorphisms (SFP) can be found. A SFP can represent a single nucleotide polymorphism, small insertion/deletion or full deletion. These genetic markers can then be used in population studies, the analysis of quantitative traits, genetic mapping and linkage analysis (Steinmetz *et al.* 2002).

Generally, the high-density arrays for SNP genotyping in humans contain thousands of allele-specific oligonucleotides for each SNP to be analysed. The probes contain all possible sequences at the site of the SNP, and multiplex hybridizations are undertaken (Hacia *et al.* 1998). In the Affymetrix GeneChip assay, a computer algorithm is then implemented to assign the genotype of each SNP. However, while SNP detection and genotyping in humans is now feasible on a microarray platform, mapping studies in which 400–500 SNPs were analysed using high-density allele-specific oligonucleotide microarrays showed that the assay failed to distinguish between heterozygous and homozygous SNP genotypes for a large fraction of the SNPs (Wang *et al.* 1998). Indeed, the PCR amplification step required to achieve sensitive and specific SNP genotyping is a principal factor that limits the use of high-throughput hybridization-based assays in human SNP genotyping (Syvanen 2001). To overcome such issues, approaches including DNA-polymerase assisted single nucleotide primer extension are now being implemented in order to perform parallel genotyping of SNPs on microarrays. In studies, this method provided a ten-fold better power of discrimination between genotypes in comparison with hybridization with allele-specific oligonucleotide probes (Syvanen 2001). Since a number of complex issues still remain with high-throughput microarray-based SNP genotyping in humans, in the remainder of this review, we will discuss the application of high-density oligonucleotide arrays to elucidate genetic diversity, with particular focus on studies undertaken with *Saccharomyces cerevisiae* (Winzeler *et al.* 2003), *Arabidopsis thalania* (Borevitz *et al.* 2003) and the pathogenic organisms, *Plasmodium falciparum* (Volkman *et al.* 2002) and *Mycobacterium tuberculosis* (Tsolaki *et al.* 2004). We will summarize by highlighting some of the directions that the field may take.

Malaria is responsible for at least 1.5 million deaths annually worldwide (Breman *et al.* 2001). Currently, there is no commercially available malaria vaccine and parasite resistance to the cheap yet effective anti-malarials is rapidly increasing. Despite the release of the complete genome sequence of the most lethal causative agent, the apicomplexan parasite *P. falciparum* (Gardner *et al.* 2002), relatively little is know about the extent of genetic diversity within this complex organism. Genetic diversity in the malaria parasite facilitates both its survival and propagation and therefore an understanding of such variation is critical if long-term control measures are to be implemented (Clark 2002).

Detecting variation, such as SNPs and putative deletions on a genome-wide scale using established molecular systems in this organism is technically challenging. However, it was recently reasoned that a custom-designed Affymetrix oligonucleotide array would facilitate such an analysis. In order to investigate this, a high-density oligonucleotide array based on the chromosome 2 sequence of *P. falciparum* was designed. This microarray consisted of 4167 unique single-stranded 25-mer probes, positioned approximately every 200 nucleotides to provide complete coverage of all 93,000 base-pairs of chromosome 2 (Volkman *et al.* 2002). The array also consisted of 395,833 probes for 80,000 cDNAs from human tissues so that the amount of host mRNA contamination occurring during the culture and extraction of the parasite DNA could be measured. The AT content of the malaria genome is extremely high (80%) (Gardner *et al.* 2002), yet this did not pose any problems in the subsequent analysis.

Although single nucleotide changes have previously been identified in *P. falciparum* (Clark 2002), the genome-wide analysis facilitated by hybridization of genomic DNA to the Affymetrix microarray identified significant differences in potential selection pressure across different gene families and locations within the chromosome (Volkman *et al.* 2002). A total of 981 SNPs was identified across the four strains, along with a large 42-kb pair deletion in the strain W2 isolated in Southeast Asia. Polymorphisms were predominantly located in those genes associated with varying the antigenic and adhesive character of the parasite. These gene families are of particular importance as currently they are being widely investigated as potential vaccine candidate epitopes due to their antigenicity. Elevated levels of genetic diversity were shown to be located in the subtelomeric ends of each chromosome, which is consistent with the extensive breakage and recombination previously reported in *Plasmodium* ssp. (Freitas-Junior *et al.* 2000) and yeast (Winzeler *et al.* 2003).

The broad genetic diversity observed within chromosome 2 warranted an analysis of the complete *P. falciparum* genome in order to elucidate the extent of genetic variation in this human pathogen. With this in mind, an Affymetrix custom oligonucleotide array containing 367 226

25-mer single-stranded probes specific to the coding and non-coding sequences of the entire *P. falciparum* genome was designed. The probes were placed on average every 150 bp on both strands (Le Roch *et al.* 2003). An additional 6000 human and mouse sequences corresponding to genes that are highly expressed in blood cells, 3602 Affymetrix controls and 2397 background controls were also included. To date, genomic DNA from several strains of *P. falciparum* have been hybridized to this array and initial results illustrate extensive diversity across and within the series of global isolates analysed. In addition, the ability of oligonucleotide arrays to identify areas within the malaria genome likely to be under selection pressure from the host's immune system may have a significant impact on the choice of future targets for vaccine and drug development.

Although SNPs and deletions can be readily identified using Affymetrix high-density arrays, more complex types of genetic diversity may also be determined using this platform. Identifying inheritable markers permits the relatedness of different strains to be easily determined. The hybridization of fourteen strains of *S. cerevisiae* (laboratory and wild-type isolates) to an Affymetrix S98 oligonucleotide array containing 285 156 25-mer probes allowed approximately 16% of the yeast genome to be investigated. A total of 11 115 markers were derived that detected variation in at least one of the 14 strains. The criterion for determining relatedness was obtained by comparing each of the strains in a pair-wise fashion and determining whether each marker was present in both, one or neither of the strains. The genealogical relationship between all fourteen strains could then be easily plotted (Winzeler *et al.* 2003). Typically, phylogenetic trees are plotted from data derived from selected sequences of a particular genome. However, plotting relatedness by the use of high-density oligonucleotide arrays offers a significantly more refined and, through the increased number of markers, more in-depth approach to determining ancestral genetic relationships.

Quantitative trait loci

Traditionally, identifying the genes responsible for quantitative genetic traits in complex genomes is challenging and laborious (Lander & Schork 1994). In highly pathogenic organisms such as *Plasmodium* spp., genetic mapping and linkage analysis are essential to locate and verify genetic determinants involved in traits of drug resistance (Su *et al.* 1997), virulence (Wellems *et al.* 1987, Day *et al.* 1993) and transmission (Vaidya *et al.* 1995). Marker location data combined with information regarding recombination frequencies can subsequently be used as a basis for exploring the genetic structure and variation in populations.

High-density oligonucleotide arrays offer a novel approach to dissecting the genetic loci responsible for such phenotypes. For example, the Affymetrix yeast microarray (Winzeler *et al.* 1998) was successful in determining the quantitative trait responsible for the high temperature growth phenotype (Htg) common in clinical isolates of yeast. The hybridization of total genomic DNA from the reference strain S96 (Htg−) and the haploid strain, YJM789 (Htg+) identified a total of 3444 bi-allelic markers, with markers spaced approximately every 3504 bp across the genome. A series of Htg+ segregants was subsequently analysed in genome-wide scans and meotic recombination breakpoints were identified which permitted the genomic intervals inherited from the same parent to be identified (Steinmetz *et al.* 2002). This analysis indicated that a combination of both common and rare variants are likely to underlie quantitative traits and the number of genes responsible for each trait is generally far higher than anticipated. Consequently, genome-wide analysis of heritability permitted by oligonucleotide microarrays will probably be critical to effectively mapping quantitative traits in the future.

The mapping of recessive mutations by DNA hybridization has been demonstrated in the plant, *A. thaliana*. The *erecta* mutation is a recessive mutation in the *A. thaliana* Landsberg *erecta* strain (L*er*) and maps to a defined region on chromosome 2. Genomic DNA was extracted from pooled samples of either the reference strain Columbia (Col) and L*er* F2 plants showing the *erecta* phenotype or from the wild-type Col/L*er* F2 plants and hybridized to an oligonucleotide microarray consisting of 285 186 perfect match (PM) features specific to the reference strain, Col. A total of 3806 markers (SFPs) were scored following a single hybridization and the position of the *erecta* gene was mapped to within 12 cM from the exact position of the *erecta* gene, within a 95% confidence interval (Borevitz *et al.* 2003).

The ability to perform genetic analyses with oligonucleotide arrays offers renewed hope in complex organisms such as *Plasmodium* spp. where genetic and biochemical manipulation is severely limited due to the cultivation and life-cycle stages of this haploid eukaryote. The incidence of malaria in endemic countries is rising rapidly due to the appearance of multidrug-resistant parasites (White 2004) and insecticide-resistant mosquito vectors (Roberts & Andre 1994). However, one must be able to genetically characterize the drug-resistant phenotype in order to fully understand the basis of resistance and facilitate the implementation of novel therapies. The chloroquine resistance locus was mapped using microsatellite markers to a 36-kbp segment on chromosome 7 (Su *et al.* 1997) after typing 16 progeny strains from a cross between chloroquine-resistant and chloroquine-sensitive parental strains (Wellems *et al.* 1990). Although linkage analysis based on laboratory crosses has been highly successful in identifying the genetic basis of genetic traits such as drug resistance (Su *et al.* 1997) and parasite development (Guinet & Wellems 1997) traditional molecular analysis is difficult and limited by the number of specific markers available for each chromosome. In contrast, microarray hybridization has shown to be a valuable and highly efficient method to map inheritance markers and localize key genetic traits in a number of organisms (Steinmetz *et al.* 2002, Borevitz *et al.* 2003) and offers renewed optimism for genome-wide linkage analysis of quantitative traits in clinically important organisms.

Putative deletions

Whole genome comparisons can reveal extensive differences in gene content and genome organization between related organisms, which enable a better understanding of the genome function and evolution of a species. The loss of genetic material can be both deleterious and advantageous. For example, short-term evolutionary pressure from the immune system may favour the elimination of a gene that is a drug target, whereas long-term physiological requirements may be in place to maintain the gene in the population. The enteric bacteria *Salmonella enterica* serovar Typhimurium, the cause of human gastroenteritis, and *Salmonella enterica* serovar Typhi, which causes human typhoid, share 97.6% identity at the genome level (Edwards *et al.* 2002), yet 11% of *S. typhimurium* genes are lost or inactivated (pseudogene) in the *S. typhi* genome. This is widely thought to have contributed to the evolution from host generalist (*S. typhimurium*) to human-restricted variants (*S. typhi*; McClelland *et al.* 2001). Chromosomal gene deletions in humans can be attributed to a host of clinical outcomes. A large deletion in a region of the long arm of chromosome 7 is associated with the cause of Williams-Beuren syndrome and loss of the long arm of chromosome 4 has been identified previously as a common occurrence in adenocarcinomas of the oesophagus and gastro-oesophageal junction (Rumpel *et al.* 1999).

Comparative genomics, although a vital tool for identifying potential deletions between two genomes is limited by the number of complete genome sequences available. Following the hybridization of genomic material to Affymetrix gene expression arrays, clusters of SFPs that exhibit a low hybridization signal can be considered potential deletions (Borevitz *et al.* 2003). This approach provides a means by which to interrogate hundreds of unsequenced genomes and gain information as to the diversity of a particular population.

The comparison of complete genome sequences of two strains of the causal agent of tuberculosis in humans, *M. tuberculosis*, identified large sequence polymorphisms (LSP) and SNPs but clues as to the molecular basis of virulence and pathogenicity remained unresolved (Fleischmann *et al.* 2002). In contrast, the hybridization of 100 epidemiologically well-characterized clinical isolates of *M. tuberculosis* to an Affymetrix *M. tuberculosis* high-density oligonucleotide array facilitated an improved understanding of genomic deletions (Tsolaki *et al.* 2004). Approximately 5.5% of the *M. tuberculosis* reference genome H37Rv was found to be completely or partially absent from the clinical isolates investigated. Gene deletions were observed in many functional classes of genes; however, the findings also suggested that these deletions did not have a strong effect on the isolate phenotype. Deletions were not found to be evenly distributed throughout the genome, with certain closely related isolates exhibiting distinct deletions suggesting genomically disruptive processes specific to an individual mycobacterial lineage (Tsolaki *et al.* 2004).

Oligonucleotide arrays: microbial diagnostics

The specificity of oligonucleotide arrays in detecting single base-pair variations suggests that microarrays could also play a role in the detection and genotyping of viral and bacterial pathogens. Multiple pathogens and variable sequences could be detected in a single assay, which would revolutionize current typing and detection methods. Although the Affymetrix array platform offers the highest probe density and resolution, the high price and lack of flexibility with this technology currently limits their application in microbial diagnostics. To date, there are few examples of such technology being used in the clinical setting but preliminary studies suggest that oligonucleotide arrays would offer a fast high-throughput alternative for the parallel detection of organisms, which would overcome some of limitations encountered when using traditional molecular-based techniques (Bodrossy & Sessitsch 2004).

In 2004, a study was published describing the use of oligonucleotide arrays in the detection of common pathogenic bacteria causing foodborne infections (Hong *et al.* 2004). Twenty-one species-specific oligonucleotide probes of the 23 S rRNA gene from sixteen bacterial species were synthesized and spotted onto nylon membranes. The results were extremely encouraging, showing that the custom oligonucleotide array was successful in distinguishing between nine of the pathogenic bacteria. However, the results also highlighted a potential drawback with the approach in that closely related bacterial strains may not be so easily distinguished. Indeed, the widely used 16 S rRNA markers do not facilitate resolution to below the species level and, in cases of the *Enterobacteriaceae*, do not allow even species differentiation. Therefore, a wide range of highly validated markers that allow resolution at a number of taxonomic levels would be required in order for microarrays to be used in clinical applications.

A more extensive study of the discriminatory power of oligonucleotide arrays was undertaken in the detection of viral pathogens. A series of overlapping 70-mer oligonucleotides specific to the most highly conserved sequences within a viral gene family from 140 sequenced viral genomes were represented on the microarray. This custom-designed array is capable of detecting hundreds of viruses (Wang *et al.* 2002). Following a random sequence independent amplification step, a diverse set of viruses such as rhinovirus and para-influenza virus could be detected from human respiratory samples. Differential patterns of hybridization also enabled discrimination between viral sero-types, greatly increasing the ease at which viruses can be typed. Consequently, this information may also form the basis for the re-assignment of established viral subtypes. The use of the most conserved sequences also permits unsequenced, unidentified or newly evolved viral family members to be detected as hybridization of DNA isolated from an unsequenced/uncharacterized virus to the conserved regions within a viral gene family suggests a common lineage (Wang *et al.* 2002).

In conclusion, although oligonucleotide arrays are currently not routinely used in molecular diagnostics, preliminary studies are encouraging (Wang *et al.* 2002, Hong *et al.* 2004) and suggest that microarrays can be used as a rapid, accurate and efficient approach for identifying diversity within a species. However, issues with specificity and sensitivity still have to be resolved.

Pitfalls and disadvantages

While the application of oligonucleotide arrays in current scientific research is vast, there are also disadvantages and pitfalls associated with such a technique (Table 1). Cost considerations and the requirement for high-throughput screening equipment currently limit the use of this technology to those laboratories that have the financial resources to fund such experiments. With regard to genetic diversity analyses, an important practical limitation is the number of specific probe features that can be synthesized on a single array platform, which restricts the extent of variability

Table 1. Advantages and disadvantages associated with elucidating genetic diversity by high-density oligonucleotide arrays.

Advantages	Disadvantages
• Genome-wide analysis.	• All data derived is relative to the reference strain (generally the sequenced strain).
• Extremely rapid and reproducible.	• Cost prohibitive (arrays and expensive laboratory equipment are required to undertake experiments).
• Large amounts of data (potential deletions and single nucleotide polymorphisms) can be obtained from a single hybridization.	• Diversity resulting from insertions, deletions and rearrangements is not readily determined.
• High resolution (up to 1 bp).	• The alternative allele and position of a SNP can only be determined with resequencing arrays or traditional methods.
• Inheritable markers can be used to easily map quantitative traits such as drug resistance.	• The Affymetrix array platform is currently not suitable for microbial genotyping and identification due to the lack of specific species markers.
• High-throughput.	• Low flexibility.
• Low false-positive rate of SFP identification.	• False-negative rate of SFP identification must be considered.

that can be discovered. For example, only 16% of the yeast genome *S. cerevisiae* was represented on a single Affymetrix chip (Winzeler *et al.* 2003). Although recent array designs have enabled up to 500 000 unique features to be represented (Le Roch *et al.* 2003), this still does not facilitate complete genome coverage for many organisms. Consequently, although the false discovery rate for SFP detection is low (Borevitz *et al.* 2003), the false negative discovery rate must also be considered.

Secondly, the detection of SFPs using a 25-mer oligonucleotide array only permits the resolution of polymorphisms within 25 base pairs and does not provide the actual sequence of the alternative allele. This is a disadvantage in terms of detecting novel drug resistant or disease-related alleles where, in some cases, a single point mutation may be all that is required to confer the phenotype. In population biology, this level of resolution does not permit the assignment of a synonymous or non-synonymous nucleotide change and consequently limits the extent of evolutionary analysis that can be undertaken. It has also been shown that the location of the SFP within the 25-mer sequence is important (Borevitz *et al.* 2003, Winzeler *et al.* 2003). Polymorphisms located near the central base of the 25-mer feature are more likely to be detected than those near to the 3′ or 5′ end.

Frequently the features on an oligonucleotide array are specific to one particular strain or individual of a species. Consequently, all subsequent hybridization analysis and comparisons are relative to this reference strain. Although a higher intensity of hybridization signal at a particular locus may be characteristic of a gene duplication, insertions and re-arrangements within the hybridizing strain will not be identified with this approach thereby limiting the extent to which genetic diversity can be characterized.

Future applications

Whole genome tiling arrays offer a redefined approach for elucidating genetic diversity and SFP discovery. In contrast to the Affymetrix expression arrays where only a proportion of the genome of interest is represented, tiling arrays facilitate significantly higher genome coverage. Whole genome tiling arrays interrogate the whole genome in an unbiased approach allowing various features of the genome to be investigated (Mockler & Ecker 2005). Non-overlapping or partially overlapping probes are designed that tile the whole genome from end to end. This not only enables the identification of many more SFPs within a particular genome but overlapping probes increase the resolution of SFP discovery. For the 125-Mbp *Arabidopsis* genome, only 12 chips were required to tile both strands of the genome at a 25-base-pair resolution. Furthermore, at an 8-base-pair resolution, only 36 chips were needed to cover the complete genome (Yamada *et al.* 2003). Although such an approach is viable for the smaller eukaryote and prokaryotic genomes, the cost and technical practicalities of this method mean it may not be feasible for the analysis of larger eukaryotic genomes.

Resequencing arrays offer the highest resolution and discriminatory power in array technology to date, essentially resequencing a genome or portion of a genome relative to the reference sequence. The resequencing array is designed with a set of tiled probes at an ultra-high resolution (1 bp). Essentially, four probes are designed for each base pair in any given sequence. One probe is specific for the reference sequence while the three remaining probes vary at the central base and code for the three alternative nucleotides at that position (Figure 2). Using the same principles as the Affymetrix expression arrays for detecting mismatches (SFP), following the labelling and hybridization of genomic DNA, three out of the four probes should show a decrease in hybridization signal between the target and probe sequence whereas the fourth probe, containing the correct nucleotide at the central position, should show a stronger hybridization signal.

Using this approach, an entire genome can be resequenced in a single hybridization experiment (Wong *et al.* 2004), facilitating in-depth genome-wide analysis of diversity. The recent use of resequencing arrays in decoding the 30-kb genome of several clinical isolates of the SARS-Coronavirus (SARS-CoV) demonstrates the application of this array-based technology as a rapid and reliable approach for assessing genetic diversity within a species population and undertaking epidemiological studies of outbreaks of disease (Wong *et al.* 2004). The advantage of resequencing arrays in determining the alternative alleles present at each nucleotide mismatch has important implications for population biology and diversity analysis within a species. While sequence rearrangements and novel sequences or insertions will still not be identifiable by such an approach, the possibilities

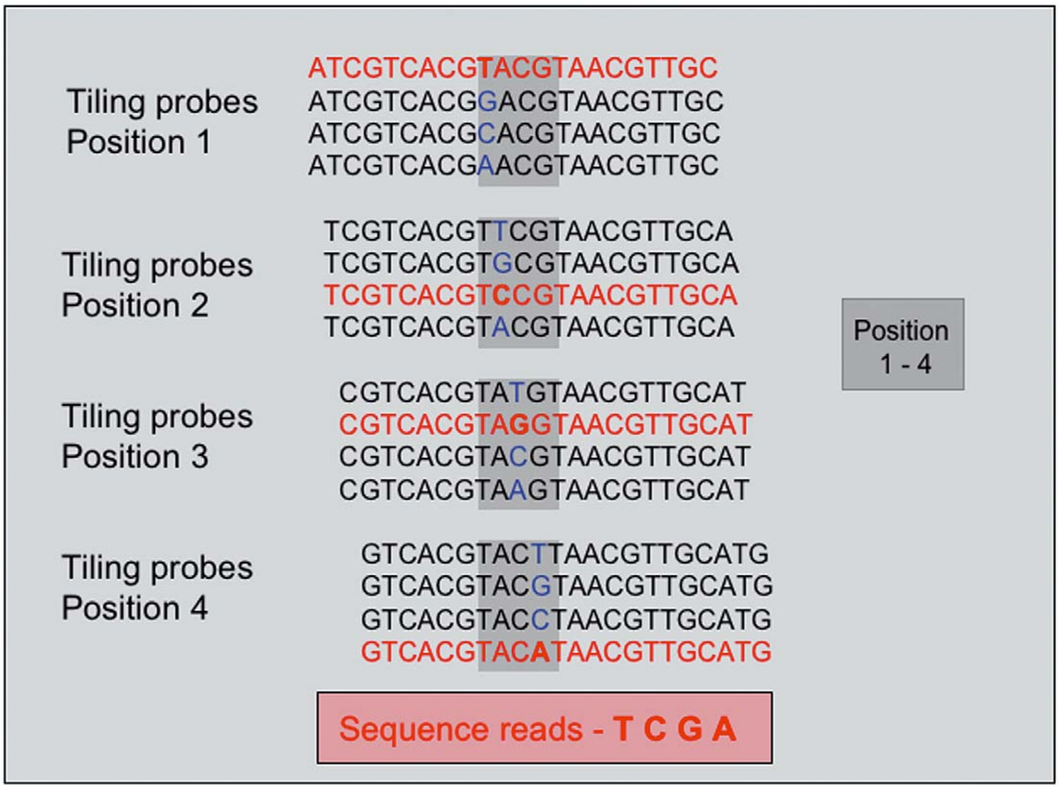


Figure 2. Tiling re-sequencing array design. Four probes per strand are tiled for each base in the reference genome. Each 25-mer probe varies at the central position to incorporate all four possible nucleotides (A, T, C, G).

offered by this technology for future analysis of genetic diversity are immense.

Conclusions

There is no doubt that the DNA microarray is a powerful tool that has revolutionized the field of genetic diversity. A single hybridization can quickly yield vast quantities of data on the relatedness of one strain to another, at a single base-pair resolution. While traditional typing techniques suffer from severe limitations, microarrays offer a comprehensive and unbiased approach to analysing diversity and permit observations that would be overlooked with established techniques where only small regions of a given genome are investigated. In today's society, accurately distinguishing between closely related strains is essential in cases of pathogens associated with food-borne diseases or bioterrorism. Indeed, microarrays played a critical role in epidemiological studies concerned with the emergence of the deadly new pathogen, SARS. We also foresee the application of microarrays in the typing and classification of organisms such as bacteria and viruses where traditional antibody typing can sometimes fall short. In summary, the application of oligonucleotide arrays in elucidating genetic diversity is immense and will continue to make a significant impact in future population and evolutionary genetic studies.

References

Beadle J, Wright M, McNeely L, Bennett JW (2003) Electrophoretic karyotype analysis in fungi. *Adv Appl Microbiol* **53**: 243–270.

Bentley SD, Parkhill J (2004) Comparative genomic structure of prokaryotes. *Annu Rev Genet* **38**: 771–791.

Bodrossy L, Sessitsch A (2004) Oligonucleotide microarrays in microbial diagnostics. *Curr Opin Microbiol* **7**: 245–254.

Borevitz JO, Liang D, Plouffe D *et al.* (2003) Large-scale identification of single-feature polymorphisms in complex genomes. *Genome Res* **13**: 513–523.

Breman JG, Egan A, Keusch GT (2001) The intolerable burden of malaria: a new look at the numbers. *Am J Trop Med Hyg* **64**(1–2 Suppl): iv–vii.

Clark AG (2002) Population genetics: malaria variorum. *Nature* **418**: 283–285.

Day KP, Karamalis F, Thompson J *et al.* (1993) Genes necessary for expression of a virulence determinant and for transmission of *Plasmodium falciparum* are located on a 0.3-megabase region of chromosome 9. *Proc Natl Acad Sci USA* **90**: 8292–8296.

Dearlove AM (2002) High throughput genotyping technologies. *Brief Funct Genomic Proteomic* **1**: 139–150.

Edwards RA, Olsen GJ, Maloy SR (2002) Comparative genomics of closely related salmonellae. *Trends Microbiol* **10**: 94–99.

Fleischmann RD, Alland D, Eisen JA *et al.* (2002) Whole-genome comparison of *Mycobacterium tuberculosis* clinical and laboratory strains. *J Bacteriol* **184**: 5479–5490.

Freitas-Junior LH, Bottius E, Pirrit LA *et al.* (2000) Frequent ectopic recombination of virulence factor genes in telomeric chromosome clusters of *P. falciparum*. *Nature* **407**: 1018–1022.

Gardner MJ, Hall N, Fung E *et al.* (2002) Genome sequence of the human malaria parasite *Plasmodium falciparum*. *Nature* **419**: 498–511.

Guinet F, Wellems TE (1997) Physical mapping of a defect in *Plasmodium falciparum* male gametocytogenesis to an 800 kb segment of chromosome 12. *Mol Biochem Parasitol* **90**: 343–346.

Hacia JG, Sun B, Hunt N *et al.* (1998) Strategies for mutational analysis of the large multiexon ATM gene using high-density oligonucleotide arrays. *Genome Res* **8**: 1245–1258.

Herrero E, de la Torre MA, Valentin E (2003) Comparative genomics of yeast species: new insights into their biology. *Int Microbiol* **6**: 183–190.

Hong BX, Jiang LF, Hu YS, Fang DY, Guo HY (2004) Application of oligonucleotide array technology for the rapid detection of pathogenic bacteria of foodborne infections. *J Microbiol Meth* **58**: 403–411.

Lander ES, Schork NJ (1994) Genetic dissection of complex traits. *Science* **265**: 2037–2048.

Le Roch KG, Zhou Y, Blair PL *et al.* (2003) Discovery of gene function by expression profiling of the malaria parasite life cycle. *Science* **301**: 1503–1508.

Lockhart DJ, Winzeler EA (2000) Genomics, gene expression and DNA arrays. *Nature* **405**: 827–836.

Lockhart DJ, Dong H, Byrne MC *et al.* (1996) Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nat Biotechnol* **14**: 1675–1680.

McClelland M, Sanderson KE, Spieth J *et al.* (2001) Complete genome sequence of *Salmonella enterica* serovar Typhimurium LT2. *Nature* **413**: 852–856.

McClelland M, Sanderson KE, Clifton SW *et al.* (2004) Comparison of genome degradation in Paratyphi A and Typhi, human-restricted serovars of *Salmonella enterica* that cause typhoid. *Nat Genet* **36**: 1268–1274.

Mockler TC, Ecker JR (2005) Applications of DNA tiling arrays for whole-genome analysis. *Genomics* **85**: 1–15.

Nelson KE, Fouts DE, Mongodin EF *et al.* (2004) Whole genome comparisons of serotype 4b and 1/2a strains of the food-borne pathogen *Listeria* monocytogenes reveal new insights into the core genome components of this species. *Nucl Acids Res* **32**: 2386–2395.

Pusch W, Wurmbach JH, Thiele H, Kostrzewa M (2002) MALDI-TOF mass spectrometry-based SNP genotyping. *Pharmacogenomics* **3**: 537–548.

Roberts DR, Andre RG (1994) Insecticide resistance issues in vector-borne disease control. *Am J Trop Med Hyg* **50**(6 Suppl): 21–34.

Rumpel CA, Powell SM, Moskaluk CA (1999) Mapping of genetic deletions on the long arm of chromosome 4 in human esophageal adenocarcinomas. *Am J Pathol* **154**: 1329–1334.

Steinmetz LM, Sinha H, Richards DR (2002) Dissecting the architecture of a quantitative trait locus in yeast. *Nature* **416**: 326–330.

Su X, Kirkman LA, Fujioka H, Wellems TE (1997) Complex polymorphisms in an approximately 330 kDa protein are linked to chloroquine-resistant *P. falciparum* in Southeast Asia and Africa. *Cell* **91**: 593–603.

Syvanen AC (2001) Accessing genetic variation: genotyping single nucleotide polymorphisms. *Nat Rev Genet* **2**: 930–942.

Tsolaki AG, Hirsh AE, DeRiemer K (2004) Functional and evolutionary genomics of *Mycobacterium tuberculosis*: insights from genomic deletions in 100 strains. *Proc Natl Acad Sci USA* **101**: 4865–4870.

Vaidya AB, Muratova O, Guinet F, Keister D, Wellems TE, Kaslow DC (1995) A genetic locus on *Plasmodium falciparum* chromosome 12 linked to a defect in mosquito-infectivity and male gametogenesis. *Mol Biochem Parasitol* **69**: 65–71.

Volkman SK, Hartl DL, Wirth DF *et al*. (2002) Excess polymorphisms in genes for membrane proteins in *Plasmodium falciparum*. *Science* **298**: 216–218.

Wang DG, Fan JB, Siao CJ *et al*. (1998) Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. *Science* **280**: 1077–1082.

Wang D, Coscoy L, Zylberberg M *et al*. (2002) Microarray-based detection and genotyping of viral pathogens. *Proc Natl Acad Sci USA* **99**: 15687–15692.

Wellems TE, Walliker D, Smith CL *et al*. (1987) A histidine-rich protein gene marks a linkage group favored strongly in a genetic cross of *Plasmodium falciparum*. *Cell* **49**: 633–642.

Wellems TE, Panton LJ, Gluzman IY *et al*. (1990) Chloroquine resistance not linked to mdr-like genes in a *Plasmodium falciparum* cross. *Nature* **345**: 253–255.

White NJ (2004) Antimalarial drug resistance. *J Clin Invest* **113**: 1084–1092.

Winzeler EA, Richards DR, Conway AR *et al*. (1998) Direct allelic variation scanning of the yeast genome. *Science* **281**: 1194–1197.

Winzeler EA, Castillo-Davis CI, Oshiro G *et al*. (2003) Genetic diversity in yeast assessed with whole-genome oligonucleotide arrays. *Genetics* **163**: 79–89.

Wong CW, Albert TJ, Vega VB *et al*. (2004) Tracking the evolution of the SARS coronavirus using high-throughput, high-density resequencing arrays. *Genome Res* **14**: 398–405.

Yamada K, Lim J, Dale JM *et al*. (2003) Empirical analysis of transcriptional activity in the *Arabidopsis* genome. *Science* **302**: 842–846.

Yauk CL, Berndt ML, Williams A, Douglas GR (2004) Comprehensive comparison of six microarray technologies. *Nucl Acids Res* **32**: e124.