# Fingerprinting and characterisation of *Escherichia coli* isolates using DNA arrays.

*Antonia Cornelia van IJperen MSc*

A thesis submitted in partial fulfilment of the requirements of the University of Greenwich for the degree of Doctor of Philosophy

*July 2005*

This research project was carried out in collaboration with:

Health Protection Agency
Specialist and Reference Microbiology Division
61 Colindale Avenue
London NW9 5HT
United - Kingdom

To my parents for their
everlasting support and
unconditional love

## ABSTRACT

Two commercially available DNA whole genome *Escherichia coli* K12 arrays were compared to identify a subset of markers for typing. The arrays were identical in probe composition but different in substrate (membrane and glass slide arrays) and probe preparation (radio- and fluorescent-labelled). Labelled genomic *E. coli* DNA from five strains of the *E. coli* reference (ECOR) collection (ATCC35320 - ATCC35324) and *E. coli* K12 were hybridised against these arrays. A group of 1240 putative markers was identified on the membrane arrays and 649 were found on the glass slide arrays. Only a small proportion of these sequences (8%) was found through both platforms. Variability in the hybridisation signals from duplicate experiments made it difficult to identify useful markers.

In order to investigate whether this technology could be used for characterising or typing *E. coli* strains, an array for the detection of 29 pathogenicity markers in *E. coli* strains was produced. This array was used with eight reference strains, including different pathotypes, 72 strains from the ECOR collection, and 49 clinical isolates. A wide range of *E. coli* pathogenicity markers was detected. The pathogenicity markers that were most common include *chuA* and *iucC*, which are both involved in iron metabolism. Additionally, the clinical isolates were grouped into clusters different from groupings based on biochemical tests. This demonstrates that the use of pathogenicity array typing can complement diagnostic tests on clinical *E. coli* isolates.

An extended, second-generation, pathogenicity marker array containing 75 probes was made. The extended array successfully distinguished between ten closely related isolates from an outbreak of urinary tract infections, while previous tests were unable to do so. This array has the potential for providing a rapid and novel means of characterising pathogenic isolates.

## ACKNOWLEDGEMENTS

I would like to thank my principal supervisor, Dr. Jon Clewley for giving me the opportunity to undertake this thesis and for his supervision, encouragement and support during the course of this work. It has taken some time but at least it is finished (again). Thanks for all your hard work on correcting and proofreading. I would also like to thank my second supervisor, Prof. Nick Saunders and my laboratory director Dr. Philip Mortimer for their advice, discussions and guidance.

Without the support of my colleagues in the former Molecular Biology Unit and others, my time at the Health Protection Agency would not have been the same. Many friendships have been built over the last four and a half years and I will treasure them forever. Friends, in and outside work, have been invaluable. I will never forget their friendship, hospitality, words of encouragement, love and prayers. Thank you for just never giving up believing in me! That was the only force that kept me going all this time.

My collaborators helped me to realise this project and believed in the principle of this technology. I would like to thank Prof. Peter Kuhnert and his laboratory staff (for his practical help and enthusiasm while exploring the pathogenicity marker array), Dr. Rod Warren and Dr. Karen Howels (for their contribution of the BMEG collection and other isolates) and Dr. Henry Smith and Dr. Claire Jenkins (for their support in creating the second generation chip).

## Declaration

I certify that this work has not been accepted in substance for any degree, and is not concurrently submitted for any degree other than that of the Doctor of Philosophy (PhD) of the University of Greenwich. Data of some sections has been published as listed below. I also declare that this work is the result of my own investigation except where otherwise stated.

Signed by Candidate:

A. C. van IJperen

London, April 2005

Signed by Supervisor:

Dr. J.P. Clewley

## Publications arising from this work

**Van Ijperen, C., Kuhnert, P., Frey, J. & Clewley, J. P. (2002).** Virulence typing of *Escherichia coli* using microarrays. *Mol Cell Probes* **16**, 371-378.

**Van Ijperen, C. & Saunders, N. A. (2004).** Microarrays for bacterial typing: realistic hope or holy grail? *Methods Mol Biol* **266**, 213-227.

**Jenkins, C., van Ijperen, C., Chart, H., Dudley, E.G., Willshaw, G.A., Cheasty, T., Nataro, J. P. & Smith, H.R. (2005).** Analysis of the distribution of plasmid and chromosomal genes in strains of Enteroaggregative *Escherichia coli* using a DNA microarray. (In preparation)

## CONTENT

## List of Tables

## LIST OF FIGURES PAGE

## ABBREVIATIONS

| | |
|---|---|
| aaf/I | aggregative adherence fimbriae I |
| aafA | aggregative adherence fimbriae II |
| aafC | aggregative fimbriae II usher |
| aap | aggregative adherence protein (dispersin) |
| aat | leucyl/phenylalanyl-tRNA-proteintransferase |
| AFLP | amplified fragment length polymorphism |
| agg3A | aggregative fimbriae III |
| agg3C | aggregative fimbriae III usher |
| aggA | aggregative fimbriae I |
| aggC | aggregative fimbriae I usher |
| aggR | aggregative fimbrial regulator |
| API | analytical profile index |
| aspU | aspartyl tRNA (dispersin protein) |
| astA | aggregative heat-stable toxin |
| bfpA | bundle-forming pilus |
| BMEG | Bacterial Methods and Evaluation Group |
| BSA | bovine serum albumin |
| cAMP | cyclic adenosine monophosphate |
| CCD | charged-coupled device |
| CDSC | Communicable Diseases Surveillance Centre |
| cfa/I | colonisation factor antigen I |
| cfa/II | colonisation factor antigen II |
| chuA | hemin receptor |
| CLDT | cytholethal distending toxin |

| | |
|---|---|
| CLED | cystein lactose electrolyte deficient |
| cnf-1 | cytotoxic necrotising factor |
| cofA | colonisation factor antigen III |
| Cy3 | 1-ethyl-2-[(1*E*,3*E*)-5-(1-{6-[(2,5-dioxo-1-pyrrolidinyl)oxy]-6-oxohexyl}-3,3-dimethyl-5-sulfo-1,3-dihydro-2*H*-indol-2-ylidene]-1,3-propadienyl}-3,3-dimethyl-5-sulfo-3*H*-indolium |
| Cy5 | 1-ethyl-2-[(1*E*,3*E*)-5-(1-{6-[(2,5-dioxo-1-pyrrolidinyl)oxy]-6-oxohexyl}-3,3-dimethyl-5-sulfo-1,3-dihydro-2*H*-indol-2-ylidene]-1,3-pentadienyl}-3,3-dimethyl-5-sulfo-3*H*-indolium |
| DMSO | dimethylsulfoxide |
| DNA | deoxyribonucleic acid |
| dpm | disintegrations per minute |
| *E. coli* | *Escherichia coli* |
| EAggEC | enteroaggregative *E. coli* |
| eae | attaching and effacing protein (intimin) |
| EAF | enteropathogenic adherence factor |
| EAST | enteroaggregative heat stable toxin |
| ECF | enhancement chemifluorescence |
| ECOR | *E. coli* reference |
| EHEC | enterohaemorragic *E. coli* |
| EIEC | enteroinvasive *E. coli* |
| eltIA | ETEC heat-labile toxin 1 |
| eltIIA | ETEC heat-labile toxin 2 |

| | |
|---|---|
| ehxA | EHEC haemolysin |
| EPEC | enteropathogenic *E. coli* |
| ETEC | enterotoxigenic *E. coli* |
| ExPEC | extra intestinal pathogenic *E. coli* |
| F1C gene | Type 1 fimbriae |
| fhuA | ferrichrom iron receptor |
| fimA | type 1 fimbriae |
| gen DNA | genomic DNA |
| HC | haemorrhagic colitis |
| hlyA | alpha-haemolysin |
| HPA | Health Protection Agency |
| HPI | high pathogenicity island |
| HUS | haemolytic-uraemic syndrome |
| IDD | infectious intestinal disease |
| ipaH | invasion plasmid antigen |
| irp2 | iron regulator protein 2 |
| iucC | aerobactin |
| kDa | kiloDalton |
| kfiB | K5 capsule antigen |
| LEE | locus of enterocyte effacement |
| lngA | longus typeIV pilus |
| LT | heat-labile toxins |
| mg | milligram |
| ml | millilitre |
| mmol | millimolar |

| | |
|---|---|
| MLEE | multi locus enzyme electrophoresis |
| MLST | multi locus sequence typing |
| neuA,neuC | K1 capsule antigen |
| norm | normalisation |
| ORF | open reading frame |
| PAI | pathogenicity island |
| papA | P-fimbriae |
| PBS | phosphate buffered saline |
| PCR | polymerase chain reaction |
| pet | EAggEC plasmid encoded toxin |
| PFGE | pulse field gel electrophoresis |
| pg | picogram |
| pic | plasmid gene involved in colonisation |
| rDNA | ribosomal DNA |
| SD | standard deviation |
| SDS | sodium dodecyl sulphate |
| sec | seconds |
| sfaA | S-fimbriae A |
| sfaS | S-fimbriae S |
| shf | shigella flexneri like protein |
| SLT | Shiga-like toxins |
| SSC | saline sodium citrate |
| SSPE | saline sodium phosphate-EDTA |
| ST | heat-stable toxins |
| stxI | Shiga like toxin 1 |

| | |
|---|---|
| stxII | Shiga like toxin 2 |
| tia | putative invasion antigen |
| TTSS | type III secretion system |
| UPEC | uropathogenic *E. coli* |
| UPGMA | unweighted pair group method with arithmetic mean |
| UTI | urinary tract infection |
| µCi | microCurie |
| µg | microgram |
| µl | microlitre |
| VLA | Veterinary Laboratory Agency |
| VT | Vero cytotoxin |

Hvert menneskes liv er et eventyr skrevet av Guds finger

(Every man's life is a fairy tale written by God's finger)

H.C. Andersen 1805-1875

## *1.1 Escherichia coli*

*E. coli* is a member of the *Enterobacteriaceae*, and usually defined by the outcome of simple biochemical tests. These biochemical characteristics of the *Escherichia* genus include production of indole, the inability to liquefy gelatine, a negative Voges-Proskauer test and a positive methyl red reaction. Additionally they do not decompose urea and do not utilise ammonium citrate. They are also able to ferment carbohydrates, including lactose, glucose and mannitol. For example, rapid lactose fermentation is a well known property of *E. coli* strains (Kauffmann, 1954, Sussman, 1997a). By definition, the *Escherichia* genus is Gram-negative and oxidase-negative. Bacteria are shaped as small rods and they are between one and eight μm in length.

The intestinal organisms of neonates and breast-fed infants were first described by Theodor Escherich (1885) as *Bacterium coli commune* while studying the pathogenesis of enteric infections. *Bacterium coli* which was found early in his studies now bears his name as *E. coli* and is used by scientists all over the world as a model organism for bacterial genetics (Schembri *et al.*, 2004), population genetics (Dai and Zimmerly, 2002), evolutionary biology (Grana and Acerenza, 2001) and pathogenicity studies (Harel and Martin, 1999). In 1997 the genomic DNA from the laboratory strain *E. coli* K12 MG1655 was one of the first full length genomes to be sequenced (Blattner *et al.*, 1997). Although the genome sequence was of great value, it by no means provided all the answers to the many questions still remaining about *E. coli* biology. For example, over 38 % of the 4,290 identified open reading frames (ORFs) in the *E. coli* K12 genome sequence were of unknown function (Blattner *et al.*, 1997). Although there is much knowledge of the genetics, molecular biology, physiology and natural history of *E. coli*, there is enough still to be found to keep researchers intrigued for many years to come. This section describes the classification of *E. coli* strains and the developments in serological and molecular typing that aid their identification.

### 1.1.1 *E. coli* in the normal population

The genera of *Escherichia, Salmonella, Klebsiella, Yersinia and Shigella*, belongs to the family *Enterobacteriaceae* and show a high rate of similarity on the basis of phenotypic and genotypic characteristics (Ewing, 1953). Whole genome comparisons of *Enterobacteriaceae* have shown a similarity between them of 50 to 86%. For example, the similarity between *Shigella* and *Escherichia* genomes varies between 80 and 86%, but is up to 98 % for some individual genes (Fukushima *et al.*, 2002, Jin *et al.*, 2002, Zeigler, 2003). This allows genetic recombination to happen between *Shigella* and *Escherichia*, and therefore they may be considered as a biogroup rather than as separate genera.

The principal habitat of *E. coli* is the lower intestinal tract of birds and mammals, where they play an important role in the host metabolism by fermenting nutrient metabolites (Cummings and Macfarlane, 1997). *E. coli* is also present in the environment such as soil and surface waters through contamination by faeces (McFeters and Stuart, 1972, Ochman and Selander, 1984, Mühldorfer *et al.*, 1996). Normally *E. coli* coexists with the human host without being associated with disease in contrast to, for example, *Shigella* spp. or *Salmonella* spp. which are invasive pathogens in human hosts ( Sansonetti and Egile, 1998, Darwin and Miller, 1999).

As an aid to understand the biology of *E. coli* a set of 72 *E. coli* strains from the normal population was selected by Ochman and colleagues from 2,600 natural strains (Ochman and Selander, 1984). These 72 strains are referred to as the *E. coli* reference or "ECOR" collection, and they have been used in several typing studies (Miller and Hartl, 1986, Arnold *et al.*, 1999, Clermont *et al.*, 2001, Johnson *et al.*, 2001). The ECOR isolates have come from a variety of hosts and geographical locations, and continue to be used to study the variation and genetic structure of *E. coli*. They include strains isolated from healthy individuals as well as from patients suffering from urinary tract infections (UTI). The selection of strains for the collection was on the basis of three criteria: I) previously used in published studies; II) representative of the genotypic diversity based on their multi locus enzyme electrophoresis (MLEE) profiles (Herzer *et al.*, 1990); III) isolated from a wide variety of host species and geographical locations. On the basis of their MLEE profiles the ECOR collection was separated into four phylogenetic groups: A, B1, B2 and D. These four groups are considered to be a representative sample of the natural *E. coli* population, although slight changes in the normal population might have occurred due to clonal drift and antimicrobial pressure.

### 1.1.2 Detection and identification of *E. coli*

#### Selective bacterial growth and detection in cell culture

The identification and characterisation tools used for the investigation of *E. coli* isolates can be divided into three categories. Firstly those for detection; secondly those for confirmation of identity; and thirdly those for typing. *E. coli* is usually isolated and identified in stool, urine, blood or environmental samples. Detection and isolation is the first step towards the identification of any bacterium. For example, the presence of bacteria can be revealed by microscopic analysis of a sample. *E. coli* can be isolated easily by growth on selective media at 37°C under aerobic conditions. Different agar plates can be used for the selective growth of *Enterobacteriaceae*. An indication of the species is determined by the colour and/or appearance of the colonies on a given medium. For example, UTI isolates can be screened on cystine lactose electrolyte-deficient (CLED) agar ( Sandys, 1960, Mackey and Sandys, 1966, Fallon *et al.*, 2002). Other widely used selective media are MacConkey or methylene-blue agar on which they appear as pink and colonies with a green metallic sheen respectively.

*E. coli* pathotypes are strains causing a similar disease pattern in a host. Some of these pathotypes can be distinguished by the behaviour of the isolates in cell tissue culture, as is described for enteroaggregative *E. coli* (EAggEC) and enterohaemorragic *E. coli* (EHEC) strains (Konowalchuk *et al.*, 1977, Caprioli *et al.*, 1983, Nataro *et al.*, 1992). Molecular methods are usually preferred to cell culture studies, but they are not always decisive.

**Biochemical identification tests**

Simple tests for biochemical characteristics, such as the carbohydrate source and oxygen use or staining methods are usually sufficient for the confirmation of the identity of *Enterobacteriaceae*. Clinical laboratories often perform these biochemical tests using the analytical profile index (API) test manufactured by Biomérieux. The API is based on a series of biochemical tests carried out in a strip containing dehydrated substrates that initiate an enzymic colour reaction when inoculated with a diluted bacterial suspension (Penna *et al.*, 2002). Each individual test can be assigned a positive or negative result on the basis of the colour that develops. The results are then separated into groups of three, and the positive reactions give a score. The sum of the positive test scores gives a seven digit code which is unique to the subspecies level. Final identification is made by entering the data into API software.

**Serological identification tests**

Normally the biochemical characteristics described above are sufficient to identify the species, but strains of *E. coli* can be subdivided into many serotypes. These serotypes are based on surface antigens. In 1947, Kauffmann (1947) designed a scheme based on these antigens for the classification of *Enterobacteriaceae*, making detailed serotyping possible for the first time. The scheme was based on the O, H and K surface antigens. O antigens or somatic antigens relate to the lipopolysaccharide, H antigens are located on the flagella and K antigens correspond with antigenic determinants of capsular polysaccharides (see Figure 1.1). These antigens are all encoded by chromosomal genes and easily detected using agglutination tests (Kauffmann, 1947).

thermostable, and the bacteria are heated to inactivate capsular antigens that might interfere with the agglutination reaction. In a similar way cultures are tested for K and H antigens; these are not heated before agglutination. K antigens are confirmed using a gel diffusion assay (Orskov *et al.*, 1977, Gross and Rowe, 1985).

**Molecular identification tests for differential diagnosis**

Molecular tools, such as polymerase chain reaction (PCR), sequencing and DNA hybridisation, are nowadays widely used for species identification (Nataro *et al.*, 1992, Johnson, 2000, Holland *et al.*, 2000). The targeted sequences are considered to be species-specific, and the size of the amplified product or the similarity at the sequence level determined by hybridisation or sequencing is used as an identifier. Ribotyping is also widely used as a molecular tool (Grimont and Grimont, 1986). The conserved 16S or 23S ribosomal DNA (rDNA) sequences can be used to identify strains accurately as the ribosomal genes are conserved within species but diverge between species. Conventional ribotyping is done by the hybridisation of 16S and 23S rDNA probes to isolated genomic DNA. Recently a new application of rDNA for identification was described by Anthony and colleagues (2000). Amplified 23S rDNA genes from the isolates were labelled and hybridised to an array of 23S rDNA targets. Species and subspecies identification took place on the basis of hybridisation signals to the arrayed targets. Alternative methods of ribotyping involve sequencing of the ribosomal genes after amplification with universal 16S or 23S rDNA primers (Drancourt *et al.*, 2000).

Many variations on standard PCR amplification with gene specific primers in conserved chromosomal regions are used for the molecular identification of micro-organisms (Bou *et al.*, 2000, Hopkins and Hilton, 2001, Jinneman *et al.*, 2003). Specific genes are targeted for the identification of the bacterial pathotype (Hopkins and Hilton, 2001, Jinneman *et al.*, 2003). The problem with this approach is that some pathogenicity markers can be horizontally transferred and could occur in more than one pathotype. The detection of a gene would therefore not always lead to the correct identification of the pathotype. Increased specificity can be obtained by using two rounds of amplification (i.e. nested PCR). In nested PCR, a second set of primers for a second amplification are located within the sequence of the first round product. The products of the second round are then synthesised in high yield and primary products lacking the secondary primer annealing sites are excluded.

**Molecular identification test for typing**

PCR is used in molecular fingerprinting techniques that combine PCR with restriction enzyme digestion, sequencing and hybridisation (Smith *et al.*, 2000). These techniques are used both for the differential analysis of isolates and for epidemiological purposes. Some methods make use of the whole of the genome while other methods make use of a small region of the genome. Examples of such methods are pulse field gel electrophoresis (PFGE; Gordillo *et al.*, 1992, Noller *et al.*, 2003) and amplified fragment length polymorphism (AFLP, Arnold *et al.*, 1999, Velappan *et al.*, 2001) for whole genome sampling. In both MLEE (Ochman and Selander, 1984, Herzer et al., 1990) and multi locus sequence typing (MLST; Adiri *et al.*, 2003, Noller *et al.*, 2003) only certain regions of the genome are investigated. In general whole genome methods are used for a more detailed investigation into isolates while methods using regions of the genome only reveal specific allelic differences.

For example, in PFGE the whole genome DNA of an isolate is extracted and digested with restriction enzymes which cut DNA infrequently. The resulting large DNA fragments are separated by agarose gel electrophoresis, giving characteristic patterns. The advantage of such methods is that the whole genome is used for testing. Gordillo and colleagues (1992) used this method successfully to distinguish between enteroinvasive *E. coli* (EIEC) strains from an outbreak in Houston, USA. The outbreak strains were compared to strains of a similar serogroup (O143) and to non-EIEC strains. All EIEC outbreak strains had a similar restriction fragment length polymorphism pattern, and also showed a strong similarity to isolates previously associated with EIEC outbreaks. Non-EIEC strains causing diarrhoea and strains with the O143 serogroup that were not causing disease had very different electrophoresis patterns. This demonstrates that PFGE is not only useful to identify small differences between genomes, but also has the potential to be used on a larger epidemiological scale.

A second very precise method for discrimination between genomes on the basis of small differences that does not depend on sequencing of large tracts of the genome is AFLP (Velappan *et al.*, 2001). This technology is based on selective amplification of restriction enzyme fragments from a digest of genomic DNA. To achieve this, genomic DNA is usually digested with enzymes that cut frequently and occasionally to obtain suitable sized fragments that can be resolved on a polyacrylamide gel. These fragments are amplified using universal primers based on adaptor sequences ligated to the ends of the restriction enzyme fragments. With fluorescently labelled primers, automated DNA sequencers and laser detection instrumentation can be used for the detection of the amplified fragments (Arnold *et al.*, 1999). AFLP has the advantage that the whole genome is used and it is also a rapid method with a high throughput capacity. AFLP, whether performed with or without fluorescent primers, has been used successfully for typing isolates, especially in outbreak situations (Desai *et al.*, 1998, Arnold *et al.*, 1999, Smith *et al.*, 2000).

MLEE uses the relative electrophoretic mobility of intracellular enzymes to characterize and differentiate organisms by generating an electromorph profile. The MLEE profiles of all 72 isolates in the ECOR collection were initially prepared using 11 enzyme loci (Ochman and Selander, 1984). In time, this work was extended to include a total of 38 enzyme loci that were included in a cluster analysis, allowing the isolates to be placed into groups with identical phenotypic characteristics, which were called phenetic groups A, B1, B2, D (Herzer *et al.*, 1990). This method only samples limited loci in the genome and is therefore, like MLST, unsuitable for fine discrimination in outbreak investigations. Differences between conserved genes appear infrequently in closely related isolates. MLEE and MLST are therefore more applicable to understanding the genetics of bacterial populations as a whole (Maiden *et al.*, 1998).

MLST has also been used to characterise the ECOR collection. MLST is similar to MLEE in that several loci are compared, and in this case housekeeping genes are amplified and sequenced (Adiri *et al.*, 2003). For the *E. coli* MLST identification database these genes are: adenylate kinase, fumarate hydratase, DNA gyrase, isocitrate/isopropylamate dehydrogenase, malate dehydrogenase, shikimate dehydrogenase, adenylosuccinate dehydrogenase and the ATP/GTP binding motif (Chan and Aanensen, 2003). A second MLST scheme for pathogenic *E. coli* strains has been developed that uses several different housekeeping genes (Whittam, 2004). The sequences are then compared to database sequences, and are assigned an identification number to each of the loci, based on their nucleotide sequences. The identification numbers of all seven loci together form a unique code at the species and subspecies level. The limitation of this method is that, like MLEE, only a few loci (seven for MLST) are analysed and the epidemiological applications are therefore limited. The advantage of MLST over MLEE is that the method does not rely upon electrophoretic profiles, but is sequence based, therefore allowing easier comparison between different laboratories.

### 1.1.3 Pathogenic *E. coli*

*E. coli* is present in the host intestines as a harmless commensal bacterium. However, there are also pathogenic variants i.e. strains that can cause diseases of man or animals. These include UTIs, diarrhoea, septicaemia and more severe disorders such as haemorrhagic colitis (HC) or haemolytic-uraemic syndrome (HUS). *E. coli* has been associated with gastrointestinal diseases since Escherich first started to investigate the gut flora. The infectious intestinal disease (IID) study in England was initiated in 1992 (Tompkins *et al.*, 1999). This study indicated that 32.5% of disease cases were associated with *E. coli*, versus 11.9% in the control group (relative proportion [case/control] = 2.7). EAggEC was the most common pathotype isolated. Besides being responsible for many gastrointestinal infections, *E. coli* is also the most common organism isolated from hospital and community acquired UTIs, and is isolated in up to 90% of all cases (Farrell *et al.*, 2003).

Outbreaks of pathogenic *E. coli* in which large groups of the population are infected with the same *E. coli* serogroup are often caused by the consumption of contaminated foods or by direct or indirect contact with an infected host (O'Brien *et al.*, 2001, Brooks *et al.*, 2004). *E. coli* related infections are being reported with increasing frequency. Bacteremia infections caused by *E. coli* were detected 13,412 times in 2002 compared to only 7,880 times in 1992 by laboratories in England and Wales (Health Protection Agency, 2003a). Also EHEC infections continue to be a concern for public health. The Communicable Diseases Surveillance Centre (CDSC) of the Health Protection Agency (HPA) reports 595 laboratory confirmed cases of EHEC O157:H7 in 2002, but in 1997 and 1999 there were more than a 1,000 cases (Health Protection Agency, 2003b).

Pathogenic *E. coli* have been grouped into six different pathotypes: enteropathogenic *E. coli* (EPEC), enterotoxigenic *E. coli* (ETEC), EIEC, EAggEC, EHEC and extraintestinal pathogenic *E. coli* (ExPEC). They are distinguished on the basis of their mode of pathogenesis. Together they are responsible for a wide variety of diseases, including severe and persistent diarrhoea, UTIs and neonatal meningitis. It is understood that commensal bacteria may become pathogenic through the acquisition of pathogenic characteristics by horizontal gene transfer thereby enabling them to cause disease. The acquired genes include those involved in the production of adhesins, invasins, flagella, toxins, cell surface molecules, secretins or secretion systems.

**Enteropathogenic *E. coli***

The term EPEC was first used to describe a pathotype of *E. coli* associated with epidemic diarrhoea in newborn and young infants (Kauffmann and Dupont, 1950, Neter *et al.*, 1955). EPEC are non-invasive pathogens that frequently cause diarrhoea, fever and vomiting, often in children under two years old. Individuals at either end of the age spectrum are most susceptible to infections, but others may be affected during outbreaks (Neter, 1960). Infections are especially common in children and infants in the developing world (Trabulsi *et al.*, 2002). The serotyping scheme of Kaufmann made it possible to identify links between serogroup and pathotype. Only a limited number of serotypes are regularly associated with diarrhoeagenic diseases. In 1997, seventeen serogroups were found to be responsible for most of the EPEC outbreaks in humans of which O18, O20, O26, O44, O55, O111 and O158 were most frequently isolated from infections (Sussman, 1997a). Additional information about pathogenicity markers is sometimes required to confirm the pathotype of strains in different serogroups.

Multiple pathogenicity factors have been identified in EPEC strains including fimbriae, pili and toxins. These are encoded on both chromosomal and plasmid DNA. The pathogenic mechanism of EPEC is known as “attaching and effacing”. The genes responsible for this mechanism are all located on a relatively small region of the chromosome known as a pathogenicity island. This island is called the LEE-island, which is the abbreviation of the locus of enterocyte effacement. It was McDaniel and colleagues (1995) who first discovered that a cloned LEE-island caused ‘attaching and effacing’ in *E. coli* K12 strains. The LEE is made up of five operons (LEE 1-5) involving over 40 genes including genes needed to produce a type III secretion system (TTSS). This is an organelle that can transfer bacterial proteins into the host cell. Each operon promotor site is activated by a protein known as Ler, resulting in a transcription of DNA into mRNA. The structural proteins assemble the basal apparatus of the TTSS. Outside the cell the bacteria may use fimbriae to attach to the microvilli of gut epithelial cells, causing a signal transduction for the transcription of the secretion proteins on LEE4 (EspA/EspD/EspB). EspA polymerises to form a long hollow filament through which other proteins can be translocated. EspD and B form a pore in the membrane of the host cell through which the bacterial proteins can be transported. The translocated intimin receptor protein is phosphorylated and inserted into the eukaryotic host cell membrane where it binds intimately to the intimin protein in the bacterial outer membrane. The intimate association of the bacterial and eukaryotic host cells causes the the re-arrangment of the host cell cytoskeleton and the destruction of microvilli. The bacterium becomes partly embedded in the host cell membrane, but does not invade the host cell (Ulshen and Rollo, 1980, Rothbaum *et al.*, 1982). This process is shown in Figure 1.2.

Typical EPEC strains carry a plasmid of 60MDa called the EPEC adherence factor (EAF) plasmid (Nataro *et al.*, 1987b). The initial adherence is mediated by genes encoded on this plasmid, and the loss of it has direct implications for the adherence and pathogenic characteristics of the isolate. Strains that have lost this plasmid are called “atypical” EPEC (Levine *et al.*, 1985). Two loci on this plasmid have been identified as important pathogenicity markers: the *bfp* gene cluster and the *per* locus. The *bfp* gene cluster is involved in the production of bundle-forming pili. These pili are type IV fimbriae and form large bundles of adhesins of 50 - 500 nm able to bind other bacteria (Donnenberg *et al.*, 1992). The *bfpA* gene encodes the subunit of these

**Enteroinvasive *E. coli***

The EIEC are very closely related to *Shigella*, not only in their pathogenic mechanism, but also in the manifestation of mucous and bloody diarrhoea in the host coupled with severe inflammation (O'Brien *et al.*, 1982, Gilligan, 1999, Escobar-Paramo *et al.*, 2003). EIEC were first distinguished from *Shigella* spp. by DuPont *et al.* (1971) when investigating *E. coli* isolates from American soldiers in Vietnam. The guinea-pig eye model, also known as the Sereny test, together with cell and tissue culture tests, all showed the invasiveness of these *E. coli* strains. Shortly after this description the first known outbreak of *E. coli* related dysentery occurred, caused by serogroup O124 which is now recognised as EIEC (Tulloch *et al.*, 1973). As well as O124, 14 other serogroups have been associated with EIEC including O28, O112, O143 and O152 (Sussman, 1997a).

EIEC are classified as *E. coli* on the basis of their ability to ferment xylose and to produce gas from glucose, but they also resemble *Shigella* in their non-motility and lack of lysine decarboxylase. Strains are often unable to ferment lactose (Silva *et al.*, 1980). Detection of EIEC using biochemical or serological methods is therefore difficult, but specific EIEC DNA probes against plasmid encoded genes have proven to be a good approach (Sethabutr *et al.*, 1985).

The pathogenicity markers associated with EIEC, located on an important plasmid, are genes encoding for invasion plasmid antigens. These antigens help the bacteria to penetrate and multiply within the epithelial cells of the colon, leading to widespread cell destruction (Kim *et al.*, 1998). Without the plasmid, EIEC is unable to invade the host cells, but other pathogenicity markers on the bacterial genome may be present (Sansonetti *et al.*, 1982, 1983).

**Enterotoxigenic *E. coli***

The ETEC are an important cause of diarrhoea in infants and travellers to developing countries or to regions of poor sanitation (Jiang *et al.*, 2002). The clinical symptoms caused by this non-invasive pathogen vary from physical discomfort to cholera-like symptoms, occasionally with fever (Gorbach *et al.*, 1971, Brunton *et al.*, 1980). The effects of ETEC in humans were first described by Taylor and colleagues (1960) and were seen in children with diarrhoea. The ETEC toxins cause fluid secretion in ligated rabit intestinal loops. Smith *et al.* (1967) revisited and named the dilating substance involved enterotoxin. It was not until a few years later that two separate classes of toxins were identified (Sack *et al.*, 1971). ETEC are defined by the production of at least one type of enterotoxins. Serogroups associated with ETEC include O6, O8, O20, O78, O128, O148.

The enterotoxins are separated into the two classes of heat-labile toxin (LT) and heat-stable toxin (ST) (Smith and Gyles, 1970). The LT encoding genes are located on plasmids and resemble the toxin isolated from *V. cholerae*. They are easily inactivated by heating to 100°C (Clements *et al.*, 1980). The LT protein contains one subunit A of 27kDa and five subunits B of 11kDa. The B subunit binds to the gut mucosa and after binding the A subunit splits into two parts (Clements and Finkelstein, 1979, Clements *et al.*, 1980). Part of subunit A initiates the ADP-ribosylation of NAD resulting in increased levels of cAMP. Increased cAMP levels result in a net overflow of sodium ions, and a loss of chloride ions and water into the gut (Kantor *et al.*, 1974, Chart, 1998). Hence the main clinical symptom of ETEC infections is watery diarrhoea.

Two main antigenic variants of this toxin have been identified (Clements and Finkelstein, 1979, Holmgren *et al.*, 1982). The sequence encoding the LT-I protein is on a plasmid and the protein is neutralised by antibodies raised against the closely related cholera toxin. The LT-II protein is encoded on the chromosome and is antigenically distinct from cholera toxin as well as from LT-I (Pickett *et al.*, 1987). LT detection was classically performed using the rabbit ligated loop model, and later using tissue culture methods (Honda *et al.*, 1981b, 1982, Holmgren *et al.*, 1982). Immunological procedures that are easier to implement in the laboratory are the Biken test, ELISA and latex agglutination tests. In the Biken test LT producing ETEC strains and anti-cholera-toxin or anti-LT sera are placed in separate wells in an agar plate. LT positive strains will form a precipitin line in the agar between the wells after incubation (Honda *et al.*, 1981a).

The second class of ETEC enterotoxin genes (ST) encode a low-molecular weight secreted protein (2-5kDa) that alters the movement of fluid and electrolytes across the intestinal epithelium (Su and Brandt, 1995, Sussman, 1997b). The ST genes are encoded on plasmids. Unlike the LT protein complex, ST does not have different subunits, and it is not inactivated by heat. This class of enterotoxins is also divided into two subgroups on the basis of structure and function (Burgess *et al.*, 1978). ST-I binds to the extracellular domain of guanylate cyclase C, resulting in increased intracellular cGMP concentrations, and leading to fluid accumulation in the gut. ST-II has mainly been isolated from pigs, and the mechanism by which it operates is not clearly defined. It has been suggested that by binding to the appropriate receptors, ST-II leads to the activation of a GTP-binding regulatory protein resulting in increased levels of free cytosolic calcium (Okamoto and Yamanaka, 2000). Traditionally, ST-I detection was performed by injecting bacterial culture supernatants into the stomach of an infant mouse. The weight ratio of intestines to mouse carcass, four hours after injection, was used to determine ST-I expression (Giannella, 1976). ST-II detection tests using intestinal loops have also been developed (Burgess *et al.*, 1978). Currently detection of ETEC (both LT and ST) is determined with PCR and molecular probe hybridisation (Moseley *et al.*, 1980, Yavzori *et al.*, 1998).

ETEC are also identified by the presence of certain fimbriae called "colonisation factor antigens" that are essential for adherence to the host cell. They are filamentous structures of between two and seven nm diameter and are composed of one or more repeated protein subunit (Dougan and Morrissey, 1985, Gaastra and Svennerholm, 1996).

**Enteroaggregative *E. coli***

The EAggEC constitute a non-invasive pathotype, and are associated with acute and persistent diarrhoea in patients living in both developing and developed countries (Bhan *et al.*, 1989, Nataro *et al.*, 1992). The IID study in England showed that EAggEC was the most common isolated pathogenic *E. coli* group from patients with diarrhoea (Tompkins, 1995, Tompkins *et al.*, 1999). The first indication of this pathotype was described when certain pathogenic *E. coli* were found to adhere to HEp-2 cells, but did not express known adherence factors (Cravioto *et al.*, 1979).

Other reports soon confirmed this adherence of *E. coli* to HEp-2 cells, and different types of adherence were distinguished: localised, diffuse and aggregative (Scaletsky *et al.*, 1984, Nataro *et al.*, 1987a). The diffuse and aggregative adherence types were found in strains that did not carry adhesion factors previously identified in EPEC or ETEC isolates. EAggEC were named after the pattern of their adherence to HEp-2 cells, which was in an aggregative "stack-brick" pattern. However, this defines a heterogeneous group, and EAggEC have since been divided in two groups, typical and atypical (Scaletsky *et al.*, 2002, Zhou *et al.*, 2002).

The typical EAggEC are characterised by the presence of a number of genes whose role in pathogenicity is well-defined. EAggEC fimbrial proteins are encoded on a 60MDa plasmid and can form fimbriae that can extend over a long distance (10μm) (Nataro *et al.*, 1992, Czeczulin *et al.*, 1997, Bernier *et al.*, 2002). Other genes encoded on the plasmid are *agg*R, a transcriptional regulator important for the transcription of the fimbrial genes, the *aap* gene that encodes a protein called dispersin, which facilitates the dispersal of EAggEC across the surface of the gut, and the *aat* gene, necessary for the transport of dispersin. Typical EAggEC also produce the heat-stable toxin EAST, a homologue of the ETEC ST.

Although typical EAggEC can be detected by cell culture, or by the molecular detection of their fimbrial genes using either PCR or DNA-hybridisation, the atypical EAggEC are more difficult to define. They do not always adhere the HEp-2 cells in a stack-brick formation and the fimbrial genes may not be present. Other characteristic pathogenicity markers for this pathotype have not yet been defined, making detection and diagnosis difficult (personal communication Dr. C. Jenkins, HPA). Further investigation of putative pathogenicity factors will become easier after the completion of the current EAggEC genome sequencing project (Chaudhuri and Pallen, 2004). The prototype strain used for this project is 042 (serotype O44:H18).

**Enterohemorrhagic *E. coli***

The EHEC is the most recently defined pathotype of *E. coli*. In 1983, a then rare verocytotoxin (VT) producing serotype (O157:H7) was isolated from patients with an unusual gastrointestinal disease in Oregon and Michigan, USA. All cases were linked to the ingestion of inadequately cooked meat products from fast-food restaurants. The symptoms of disease were severe abdominal pain, watery diarrhoea developing into haemorrhagic colitis, and little or no fever (Riley *et al.*, 1983). Since then this EHEC serotype has been associated with other outbreaks of severe gastrointestinal disease, haemorrhagic colitis, and even the life threatening kidney disorder haemolytic uraemic syndrome (Ryan *et al.*, 1986). The resulting disease is characterised by haemolytic anaemia, thrombocytopenia, renal failure and central-nervous-system manifestations (Riley, 1987). Two isolates from this serotype have been sequenced (Hayashi *et al.*, 2001, Perna *et al.*, 2001)

EHEC are distinguished from EPEC by the production of VT (O'Brien *et al.*, 1982, Levine *et al.*, 1987). They also carry the LEE region, as described for EPEC strains for attachment to host tissues. The first indication of *E. coli* strains producing a VT came from Konowalchuk and colleagues (1977) who reported that certain strains of *E. coli* produced a heat-labile toxin different from the LT detected in ETEC strains that was cytotoxic for Vero cells. It is also known as shiga-like toxin (SLT) due to structural similarity with this toxin. Antibodies against the known ETEC LT did not neutralise the cytotoxic activity. The isolates that produce SLT are mostly associated with diarrhoea and belonged to a variety of serogroups. Serogroups most commonly associated with EHEC are O26, O55, O111, O103, O128 and O157. Understandably, there is a similarity between these compared to the EPEC strains. It is the production of SLT that defines these isolates as EHEC, and therefore separation of EPEC and EHEC isolates cannot be based on the serogroup alone.

The EHEC SLT are members of a large family of subunit toxins that share a common mode of action: they inhibit protein synthesis in the host cells by the removal of an adenine residue from the 28S ribosomal subunit, and they function as an enterotoxin leading to the induction of fluid secretion, which may eventually lead to death (Endo *et al.*, 1988). The toxins have a structure similar to the ETEC LT. They are composed of one A subunit of 32 kDa and five B subunits of 7.7 kDa. The B subunit binds to a glycolipid receptor on the surface of eukaryotic cells and subsequently part of the A subunit binds to the 28S ribosomal subunit (Endo *et al.*, 1988).

Other pathogenicity markers like intimin, enterohemolysin and secreted serine proteases, may also be present (Welinder-Olsson *et al.*, 2002). These pathogenicity markers also occur in other non-O157 VT producing EHEC (O'Brien *et al.*, 1982, Khan *et al.*, 2002).

**Extraintestinal pathogenic *E. coli***

Several non-intestinal diseases are caused by a group of *E. coli* defined as the ExPEC (Johnson and Russo, 2002). There are two main types of illness mediated by ExPEC strains: UTI and neonatal meningitis. Although these infections are of great medical importance, they only occasionally occur in outbreak situations, and have therefore not captured public attention like the intestinal strains. ExPEC strains often carry multiple antibiotic resistance determinants, complicating treatment (Phillips *et al.*, 1988). Most ExPEC can be classified into two of the four phylogenetic *E. coli* groups (B2 and D) as described in section 1.1.1 and have been identified with specific pathogenicity markers (Clermont *et al.*, 2001, Johnson *et al.*, 2002, 2003). ExPEC are one of the principle causes of morbidity and mortality arising from community- and hospital-acquired extraintestinal infections in human, of which UTI is most commonly observed (Donnenberg and Welch, 1996).

**Uropathogenic *E. coli***

Uropathogenic *E. coli* (UPEC) are responsible for almost 80% of all UTIs occurring in women and the elderly (Farrell *et al.*, 2003). Infections may be asymptomatic, but can develop into cystitis or pyelonephritis. Both of these conditions need immediate medical attention. Symptoms of cystitis include dysuria, urinary urgency and frequency. A more serious UTI results in pyelonephritis caused by organisms ascending to the kidneys. Patients suffer fever, flank pain, bacteriuria, abdominal or groin pain and vomiting. Infections can spread beyond the urinary tract and enter the bloodstream. UTI can lead to the development of bacteremia as a secondary disease because of bacteria entering the bloodstream.

The pathogenicity markers found in UPEC strains include a number of different fimbriae, among which the P fimbriae was the first to be identified with an important role in cell attachment (Svanborg Eden and Hansson, 1978, Hagberg *et al.*, 1981). A further understanding of P fimbriae biogenesis and action arrived when Hull and colleagues (1981) cloned the genes encoding it. Cells expressing the cloned P fimbria showed acquired adherence properties in haemagglutination tests. Other fimbriae include F1C, S and Dr and type 1 fimbriae (Johnson and Stell, 2000). Certain serogroups like O1, O2, O4, O6, O7, O18 and O75 have been strongly associated with UTI as have capsule antigens such as K1 and K5 (Mabeck *et al.*, 1971, Glode *et al.*, 1977, Kaijser *et al.*, 1977, Kaijser and Jodal, 1984, Petit *et al.*, 1995).

**Neonatal meningitis *E. coli* strains**

*E. coli* are responsible for 40% of neonatal meningitis cases, often those that are life threatening. Isolates usually carry the K1 antigen as a pathogenicity marker (Glode *et al.*, 1977). Serogroups associated with meningitis include O1, O6, O7 and O18. A study carried out in Japan indicates that ExPEC isolates containing this K1 antigen are present at a high prevalence in pregnant women (Obata-Yasuoka *et al.*, 2002). It was concluded that isolates causing neonatal meningitis could be transmitted during natural childbirth.

**Nosocomial infections**

*Enterobacteriaceae*, predominantly *E. coli*, *Klebsiella* and *Enterobacter* spp., cause over a third of infections acquired during hospitalisation. Infection can be acquired during or after operations through the use of medical equipment and through wound infections (Vincent *et al.*, 1995). Immunocompromised patients are particularly susceptible to infection. Many isolates carry antimicrobial resistance genes on chromosomal or plasmid DNA, conferring resistance to one or more antimicrobials (Vincent *et al.*, 1995). Multiple resistance, i.e. resistance against four or more antimicrobials, can cause severe problems in patient treatment (Livermore *et al.*, 2000). The patient may suffer if ineffective antibiotic treatment is given, but on the other hand, antibiotic resistance may be encouraged if unnecessary antibiotics are used. Consequently, it is helpful to detect the pathogen and its antimicrobial resistance profile at an early stage.

### 1.1.4 *E. coli* whole genome sequences

The comparison of clinical isolates at the nucleotide level is the most comprehensive method for phylogenetic and evolutionary investigations of *E. coli*. Even with current technology, the sequencing of the genome for every isolate is too expensive and labour intensive to be realistic. As of early 2005, five complete *E. coli* genomes have been sequenced. These include two *E. coli* K12 strains (MG1655, W3110), two O157:H7 EHEC strains (EDL933 and RIMD 0509952) and the UPEC strain CFT073 (serotype O6:H1) (Blattner *et al.*, 1997, Hayashi *et al.*, 2001, Perna *et al.*, 2001, Welch *et al.*, 2002). Other strains that are in the process of being sequenced are the prototype EAggEC strain 042 (serotype O44:H18), an EPEC strain E2348/69 (serotype 127:H6) and *E. coli* strain DH10B (Chaudhuri and Pallen, 2004).

The few whole genomes from different pathotypes that have been sequenced to date have revealed very useful comparative data, which have led to some interesting discoveries. For example, 1,387 novel genes were found in *E. coli* O157 compared to *E. coli* K12. These are expected to represent functions that *E. coli* O157 has acquired. It is likely that they will include pathogenicity markers which will be target genes for the investigation of pathogenicity in general, and also for studying host-pathogen interactions and infection mechanisms (Hayashi *et al.*, 2001).

A three way comparison of the genomes of *E. coli* strains MG1655 (K12), EDL933 (EHEC) and CFT073 (UPEC) revealed that they only have 39.2% of their predicted protein ORFs in common (Welch *et al.*, 2002). This 39.2% may represent an ancestral backbone sequence passed on in bacterial duplication. Genes that are only present in one or two of the strains may have been either acquired by horizontal gene transfer or deleted from a common ancestor. Such genes could have been acquired from closely related species in the different ecological niches the strains occupy (Lawrence and Ochman, 1998). It is unlikely that all these changes have been obtained through horizontal transfer, or the mutation of existing genes. A very high mutation rate would have been required in order to generate the level of diversity seen between these strains. LeClerc and colleagues discovered defective genes in O157:H7 involved in repairing DNA mismatches (LeClerc *et al.*, 1996). Defects in these mechanisms are also present in non-pathogenic strains and therefore unlikely to be the only cause of this genetic diversity. Eisen (2001) has suggested that it is possible that genes that are not present in all strains originated in a common ancestor and have been deleted over time.

## *1.2* E. coli *pathogenicity*

Virulence is defined as "the degree of pathogenicity (capability of causing disease) of a micro-organism" (Micropaedia, 1974). Harmless micro-organisms can acquire pathogenicity factors through horizontal gene transfer. The acquired sequences include toxin, adhesin, capsule and iron acquisition genes (Sussman, 1997b). Gene acquisition in *E. coli* occurs by horizontal transfer of DNA through transformation, conjugation or transduction (Roy, 1999). Transformation of the bacterial cell involves naked DNA from the surroundings being taken up into the bacterial cell through the cell wall. Chromosomal or plasmid DNA can also be exchanged via pili, in a process known as conjugation. Finally, DNA can be transferred through vectors such as bacteriophages. All acquired DNA can be integrated into the genome of the infected host and replicated through the host's own replication system.

### 1.2.1 Pathogenicity markers

Pathogenicity markers can be separated into different classes depending on their involvement in pathogenesis. Genes that are involved in the attachment of the bacterial cell are called adhesins. Pathogenic *E. coli* can produce several types of adhesin genes associated with pathogenesis including fimbriae. Fimbriae are hair-like structures of around seven nm and are characterised by their ability to bind to surfaces through, for example, D-mannose-containing residues. Fimbriae play an important role in pathogenesis by attaching to the host cell and are also involved in the evasion of attacking phagocytotic white blood cells (Mattick, 2002). A single *E. coli* isolate can express multiple fimbrial types. Their main function is adherence to the host cell and they are generally not involved in movement of the cell. The exception to this rule is type IV fimbriae that can twitch slightly and, therefore might have a role in the movement of the cell (Mattick, 2002). Type IV fimbriae include bundle forming pili produced by EPEC and the longus type IV pilus and colonisation factor antigens produced by ETEC (Donnenberg *et al.*, 1992, Gomez-Duarte *et al.*, 1999).

The words fimbriae and pili are now used interchangeably, but pilus was originally reserved for specialised adhesins that were involved in DNA transfer in bacterial conjugation, also known as the F-factor or sex pilus. Pili are differentiated from fimbriae by the formation of a helical tube-like structure and, therefore, typically have a slightly wider diameter. Their ability to conjugate may well play a role in the assembly of pathogenic markers. The fimbriae of ETEC are called colonisation factor antigens as they assist the colonisation process in the host gut or urinary tract, and are therefore strongly associated with pathogenesis (Gaastra and Svennerholm, 1996).

Fimbriae are grouped into different types: I-VII. Type I fimbriae are present on many bacteria, whilst other fimbriae are associated with specific pathotypes or serogroups of *E. coli*, while others are host related and only appear on strains from a common source (Klemm, 1984, Boylan *et al.*, 1988, Nataro *et al.*, 1992, Frydendahl *et al.*, 2001, Johnson *et al.*, 2002). Colonisation factor antigens are present in ETEC strains (Taniguchi *et al.*, 1995). P and S fimbriae are pathogenicity markers that are frequently associated with UPEC (Johnson and Stell, 2000, Oelschlaeger *et al.*, 2002b). EAggEC has specific adherence fimbriae that are involved in the formation of the stack-brick adhesion pattern to host cells (Bernier *et al.*, 2002, Elias *et al.*, 2002). Examples of origin specific fimbriae are K88 in porcine hosts, CS31A in bovine hosts and Pap in human hosts.

Another large group of pathogenicity markers codes for bacterial toxins and their secretion systems. The toxins are composed of secreted proteins that damage host cells through a variety of mechanisms: they can damage surrounding tissue, lyse the host cell, block protein synthesis, or interfere with host cell functions. Various toxins have been identified in *E. coli* (Abe *et al.*, 1990, Bebora, 1997, Call *et al.*, 2001). The enterotoxins in ETEC and the SLT in EHEC are the most well known and have been described in section 1.1.3.

Other toxins have been found in pathogenic *E. coli* strains. Haemolysins found in EHEC and EPEC strains are capable of destroying the host erythrocytes. It is speculated that iron released from the erythrocytes assists in bacterial survival. *E. coli* strains producing haemolysin were first described by Kayser (1903). An association between haemolysin producing strains and their pathogenic effects has been recognised especially in ExPEC (Welch *et al.*, 1981). The most common haemolysin is α-haemolysin. The importance of α-haemolysin in pathogenesis was shown by Welch and colleagues (1981) when a cloned haemolysin gene introduced in a non-pathogenic strain resulted in a pathogenic phenotype. The haemolysin toxin is a large 110kDa protein and its encoding sequences are positioned near other pathogenicity markers on the chromosome. Haemolysin is secreted via a type I secretion system and binds the host erythrocytes.

The cytotoxic necrotising factor (cnf-1) is commonly found in UPEC strains, but it was originally observed in strains isolated from children with enteritis. cnf-1 is recognised by its pathogenic ability to cause lesions in rabbits and morphological changes in *in vitro* cell culture (Caprioli *et al.*, 1983). The 110kDa CNF1 protein is encoded by a single gene on the chromosome. In ExPEC strains cnf-1 is linked to the presence of haemolysin and P-fimbriae that is located on the same pathogenicity island.

EAggEC strains also produce a heat-stable toxin (EAST) which is different in size from previously defined heat-stable toxins, but which seems to have a similar mode of action (Savarino *et al.*, 1993). It is predicted that EAST1 stimulates the production of guanylate cyclase through the same receptor-binding region as ST and guanylin leading to increased intracellular cGMP concentrations and fluid accumulation in the gut.

The third group of pathogenicity markers are capsules which protect bacteria from host defence mechanisms (Schembri *et al.*, 2004). The capsules also help the bacterial cell to attach to the surface and support biofilm formation (Danese *et al.*, 2000). Approximately 90 different capsule genes have been identified. They consist of acid polysaccharides made up from repeating oligosaccharide units. Capsule antigens can be classified using specific antisera as described by Kaufmann (1947), and may be observed by light microscopy after India ink staining as a "halo", or by immunoelectron microscopy. Two groups of capsule genes have been identified. Group I antigens are expressed at all temperatures, and group II are only expressed at temperatures higher than 25°C (Cieslewicz and Vimr, 1996). Only a few of the capsules in group II are frequently associated with pathogens, these include K1 and K5. Capsule K1 deactivates the complement system by binding to the components resulting in bacteria escaping phagocytosis. K1 is associated with the majority of isolates causing neonatal meningitis (Glode *et al.*, 1977). Humans and animals seem unable to produce specific antibodies against capsular antigen K5 because of its structural identity with desulphoheparin; an intermediate in heparin biosynthesis. (Kaijser and Jodal, 1984, Kroncke *et al.*, 1990, Finke *et al.*, 1991). Group II capsules are often involved in UTI infections.

### 1.2.2 Pathogenicity islands (PAIs) and plasmids

Pathogenicity markers are often located at specific sites in the chromosome (also known as pathogenicity islands (PAIs)), or on plasmids. PAIs are large DNA regions (10-200 kb) that carry pathogenicity markers and mobility genes. They are absent from non-pathogenic members of the same and closely related species. These regions are often enclosed by inverted repeats and are characterised by a difference in G+C content, an atypical codon usage and are also frequently associated with tRNA genes (Dozois and Curtiss, 1999). These characteristics suggest that the genes in this region are acquired through horizontal gene transfer, as their characteristics differ strongly from DNA on either side of the PAI. For example, as genes representing mobility factors such as integrases, transposases or parts of insertion elements are often encoded on the PAIs. PAIs have also been detected in *Salmonella* spp., *Helicobacter* and *Yersinia* spp. Hacker and colleagues have studied the prototype UPEC strain (536), and have identified five pathogenicity islands (Oelschlaeger *et al.*, 2002a, Oelschlaeger *et al.*, 2002b, Hacker *et al.*, 2003). Pathogenicity markers on these islands are: α-haemolysin (PAI I & PAI II); P-fimbriae (PAI II); S-fimbriae (PAI III); genes with homology to the iron uptake systems as described for *Yersinia* species (PAI IV); and capsular polysaccharide (PAI V). Mutant strains lacking one or more pathogenicity island have a phenotype that not only lacks the function of the genes on that island, but also does not produce products encoded on the other islands. Thus the pathogenicity islands appear to have a regulatory apparatus for the global control of pathogenesis.

Another pathogenicity island that is found in EPEC and EHEC strains is known as the LEE-island, which carries the genes encoding the adherence factor intimin (McDaniel *et al.*, 1995). Other pathogenicity markers that have been found on pathogenicity islands are type III and type IV secretion systems, toxin genes and capsular polysaccharides (Kaper *et al.*, 1997, Boyd and Hartl, 1998, Bingen-Bidois *et al.*, 2002). One particular PAI is called the high pathogenicity island (HPI) which was originally found in *Yersinia* spp. and includes yersiniabactin genes; homologues have also been detected in many other enterobacterial species pathogenic to humans (Koczura and Kaznowski, 2003). HPI-positive strains are also found in non-pathogenic *E. coli* strains isolated from humans but not in environmental strains. Therefore, the contribution of the HPI to pathogenesis remains unclear, but it has been speculated that the HPI assists in the adaptation to human hosts.

In addition to these chromosomal regions, a number of plasmids that carry pathogenicity markers have been identified in pathogenic *E. coli* (Bebora, 1997). Plasmids are easily replicated and transferred, but are often unstable during cell division. Pathogenicity-associated genes might therefore be lost during replication. Genes identified on plasmids encode for toxins and adherence factors produced by ETEC, bundle-forming pili specific to EPEC and EHEC strains, and the heat-stable toxin produced by EAggEC (Savarino *et al.*, 1993, Bebora, 1997, Kaper *et al.*, 1997, Gomez-Duarte *et al.*, 1999). Certain pathogenicity plasmids, especially those that encode an aerobactin-mediated iron uptake system may also contain antibiotic resistance genes, which can provide selective advantage (Johnson *et al.*, 1988, Phillips *et al.*, 1988). Although these plasmids are very important in pathogenicity, it seems unlikely that the acquisition of just one plasmid could cause a microbe to become a pathogen as many contributing factors, both chromosomal and environmental, are necessary for the expression of plasmid genes.

### 1.2.3 Pathogenicity marker expression

The expression of pathogenicity markers responds to environmental factors, as well as to other expressed pathogenicity markers and internal cell signals (Harel and Martin, 1999). Not all pathogenicity markers are produced constitutively as this may be disadvantageous; for example, an adhesin can assist the colonisation process in the gut, but is a disadvantage whilst the pathogen is being transported through the bloodstream to the site of infection.

Although the specific host signals are not yet clearly understood, environmental signals that influence pathogenicity marker regulation have been identified. For example pH, temperature and iron concentration all affect gene regulation of pathogenesis *in vivo*. The host digestive system is susceptible to large pH changes, and *E. coli* isolates causing infections in this environment need to adjust to conditions of low pH. To resist pH stress, bacteria have evolved several mechanisms, including regulatory networks that control several genes involved in acid tolerance. For example, *E. coli* O157:H7 strains have been found to be more acid-resistant than generic strains (Conner and Kotrola, 1995). Outbreaks in Canada, spread by contaminated apple cider, suggested that this serotype might be more resistant to lower pH. Conner (1995) showed that *E. coli* O157:H7 is able to survive a range of different acid stresses at a variety of temperatures above 10°C. Also, certain flagella genes have been shown to be regulated by acid responses (Soutourina *et al.*, 2002). It is therefore likely that changes in pH initiate gene responses for the expression of other pathogenicity markers that are not directly involved in acid resistance.

Some capsular genes are temperature sensitive and are only expressed at and above 25°C. The production of these capsules *in vitro* can be studied by the manipulation of the incubation temperature during bacterial growth. Another thermoregulated process described by Umanski and colleagues (2002), shows that the regulation of the LEE operon is repressed below 27°C and expressed at 37°C. This operon initiates the transcription of many genes involved in pathogenesis, including a type III secretion system. Hence the activation of this operon increases the virulence of the pathogen.

Hosts are more susceptible to infection by bacteria when iron is freely available as iron promotes bacterial growth (Bullen *et al.*, 1991). Neonatal meningitis strains are almost always associated with iron associated pathogenesis markers (Negre *et al.*, 2004). To acquire iron from the host iron-binding proteins, *E. coli* has developed several mechanisms. One of these mechanisms uses a small iron-binding molecule called a siderophore which has a higher affinity for iron than host proteins such as transferrin and lactoferrin (Crosa, 1989). Two siderophore systems are most frequently found in *E. coli*: I) the enterobactin system and II) the aerobactin system. The iron-chelator enterobactin is produced when iron restrictions are present *in vitro*. It can remove iron from other iron-binding proteins due to the extremely high formation constant of $10^{52}$. Although highly effective in iron acquisition, enterobactin is only used once as after cleavage of the iron ion, the resulting molecule is discarded. Aerobactin has a formation constant of $10^{22.9}$ and is therefore less effective in binding iron than enterobactin, but still highly compatible with other iron binding proteins.

Furthermore, after the cleavage of iron from the chelator it can be reused and uses therefore less energy in the formation of the chelator compared to the enterobactin system. The aerobactin system is often detected in isolates obtained from blood and could therefore be associated with septicaemia. The yersiniabactin system located on the HPI plays only a very small role in iron uptake in *E. coli*. Other iron acquisition genes widely distributed in pathogenic *E. coli* also play a role in susceptibility to infection (e.g. *chuA*, *iucC* and *fhuA*) (Coulton *et al.*, 1986, Martinez *et al.*, 1994, Torres and Payne, 1997). The genetic locus for the ferric uptake regulation (*fur*) is linked with iron associated cell processes (Schäffer *et al.*, 1985). This locus encodes a 17kDa protein, which acts as a transcriptional repressor of genes involved in the iron assimilation pathways. It can bind to specific sequence called the "iron box", which can be found in the promoter region of Fur-regulated genes.

One approach to effectively achieve rapid identification and characterisation of *E. coli* pathotypes is the use of DNA arrays. This technology may be able to detect multiple potential pathogenicity markers simultaneously, and can also be used to monitor their expression.

## 1.3 Arrays

In the last decade, arrays have been used to address many biological questions. They are now used intensively in virtually all areas of research in the biological sciences (Schena *et al.*, 1996, Behr *et al.*, 1999, Alizadeh *et al.*, 2000, Porwollik *et al.*, 2002, van Ijperen *et al.*, 2002). The term "microarray" was first used by Schena and colleagues (1995) to identify and study a subset of *Arabidopsis* genes. They studied plant transcription factors in the early 1990's and adapted the Affymetrix yeast array concept to create the first quantitative DNA microarray, using a two-colour fluorescence hybridisation method. Also, Southern and colleagues were early pioneers in array technology, successfully creating arrays for use in resequencing and studying DNA interactions (Maskos and Southern, 1993, Southern *et al.*, 1994). Since then, due to the relatively rapid completion of whole genome sequencing projects of both prokaryotes and eukaryotes, microarray technology advanced rapidly (Salama *et al.*, 2000, Dorrell *et al.*, 2001, Smoot *et al.*, 2002, Wang *et al.*, 2002a). The technology has proved not only to be very useful in gene expression patterns, but has also been rapidly adapted for genotyping (Behr *et al.*, 1999, Salama *et al.*, 2000) and resequencing (Saiki *et al.*, 1989, Cronin *et al.*, 1996). This section will describe in more detail various aspects of the arraying process, including methods of data analysis.

### 1.3.1 Array technology overview

The term array is applied to variants of a technology in which the common feature is that they all comprise a set of defined nucleic acid sequences (probes), placed at specific X, Y co-ordinates on a solid support surface (e.g. coated glass microscope slides or nylon membranes). Arrays can differ in the composition, length and density of the arrayed probes, the structure of the solid support and the nature of the target DNA. The technology is based on hybridisation of labelled target DNA against a large number of probes (up to 40,000 per microarray slide) attached to a solid support. The most powerful characteristic of array technology is that thousands of genes from different samples can relatively easily be analysed, e.g. two different samples can be compared in the same experiment by the labelling of the targets with different fluorescent dyes (Richter *et al.*, 2002, t Hoen *et al.*, 2003). Understandably, the potential for this tool is large in the biosciences, especially in oncology, pharmacology and biochemistry.

Rapid developments in technology have lead to the manufacture of equipment to assist the practicalities of the arraying process. A variety of robotics, including liquid handling and spotting robots, as well as microscope scanners and automated hybridisation equipment are now commercially available. Also, software programs for the analysis of the hybridisation data are being continually improved and updated, including those for standardisation, normalisation and clustering functions of the array data.

Figure 1.3 depicts the three separate parts of the arraying process:

**a)** array printing, including probe preparation, arraying of the probes onto a solid support and robotic tools;

**b)** array hybridisation, including target preparation and pre- and post-hybridisation steps;

amplification for array use in which standard plasmid sequencing primers were used to amplify cloned, sequence-defined fragments used in the *Campylobacter jejuni* NCTC 11168 genome sequencing project. The advantage of this approach is the low cost of the amplification primers, however the drawback is that the PCR products do not necessarily represent single ORFs.

The purity and identity of PCR amplified probes needs to be confirmed before arraying them onto the solid phase, which is a laborious procedure. The most common method is to analyse the products of each PCR by agarose gel electrophoresis to check for the presence of a single amplicon of the expected molecular size. A representative sample of the products, depending on the resources available, is then sequenced to provide a positive identity check prior to arraying (Taylor *et al.*, 2001). The PCR products are optimally about 400-1500 base pairs, and preferably represent the longest possible specific region of the ORF. In general, similar sized probes give a more evenly spread hybridisation signal, which assists in the normalisation process.

Oligonucleotides ordered from commercial companies are purified and are ready to print as soon as they arrive in the laboratory. This saves the time and resources required to run PCR reactions, analyse the products, and redesign any PCRs that do not give the expected product. The length of the oligonucleotide probes spotted with a printing robot is generally 20 to 70 bases. Short oligonucleotides (20-50mer) are best bound to the glass surface via an aminolink group, but this increases the cost of synthesis (see also Figure 1.4 , page 45) (Kane *et al.*, 2000). Often, the genes of interest on oligoarrays are covered by multiple oligonucleotides to exclude inter-array variation, and increase specificity and sensitivity. Oligonucleotides can be designed using specific software packages. To find specific oligonucleotides covering an ORF the melting temperature, specificity, secondary structure and the length of the oligonucleotide are taken into consideration. For a small number of probes this is relatively easy as there will be many suitable sequences, but once the required number of probes becomes larger there will be less suitable specific sequences that do not interact in so many ways. For example, the selection of oligonucleotides for printing microarrays might result in unspecific probes that cause cross-hybridisation. Either the single most suitable oligonucleotide can be chosen to amplify an ORF, or less suitable multiple oligonucleotides with little predicted sequence similarity can be used to maximise the detection of the ORF of interest.

Oligonucleotides can also be directly synthesised on the slide through *in situ* synthesis using photolithography (Pease *et al.*, 1994). Arrays manufactured by photolithography, as produced by Affymetrix, have a 100 to 200 times higher density than arrays produced with a printing robot. Photolithography uses masks to isolate the site of activation and elongation of the oligonucleotide, and is initiated by light. Microarrays made by photolithography are limited to oligonucleotide probes of 20-25 bases.

In comparing oligonucleotide arrays with arrays made of longer PCR products it has become clear that both can be used for similar applications, but they may not give identical results (Li *et al.*, 2002). Furthermore, small differences in sequence can only be detected using oligoarrays. A pilot study can confirm whether oligonucleotides or PCR products will give the most accurate results. Overall signal variation has been found to be less for oligonucleotide arrays as all probes have approximately the same length, which may not be the case with PCR products (Kuo *et al.*, 2002). When using oligonucleotide arrays it is possible to detect sequence differences at the single base level (Cronin *et al.*, 1996). In contrast, arrays made from PCR products can detect similar genes with partial sequence similarity to the gene of interest, which may be useful for the detection of homologous genes.

**Pre-arraying treatment**

To make an array, PCR products or oligonucleotides are first redissolved in a volatile spotting solution to ensure rapid drying of the gene products (probes), but excessive evaporation of the sample during the printing process must be avoided. This is influenced by the local temperature and humidity. A common spotting solution is 50% dimethyl-sulphoxide (DMSO), which is normally used on non-activated surfaces with PCR products (Rickman *et al.*, 2003). DMSO can interfere with some active groups on the slide surface, and it is therefore less suitable for printing oligonucleotides than salt-based spotting buffers. DMSO cannot be used for printing on membranes as it causes perforation. Salt-based spotting buffers are better used for these applications, but may cause blocking or erosion of the capillary pins used for the deposition of probes onto the surface. Changes in the volumes of the probe solutions dispensed into microtitre plate wells for arraying inevitably occur over time. The probes can be dried down fully and redissolved in the original volume of solvent.

**Printing of the arrays**

Arrays can be printed either on membranes or glass slides. Nylon membrane arrays are generally not used at high densities as the probes are large due to the absorption of the probe solution into the membrane (Kuhnert *et al.*, 1997, Anthony *et al.*, 2000). One reason for printing probes further apart (greater pitch) on membranes is to accommodate variations in the size of the probes formed. When radioactive detection methods are used extra space must be allowed between probes to prevent the merging of signals from adjacent probes due to the scattering of the radioactive signal. For example, on the commercially available Panorama membrane arrays (Sigma-Genosys), the probes are printed five millimetres apart.

Glass microscope slides used for the printing of microarrays can be coated with reactive groups to allow covalent binding of the DNA probes. Robotics arms controlled by stepper motors are used to position the probes micrometers apart (<200 μm). The density of probes arrayed with printing robots can vary between 20,000 - 70,000 probes per microarray slide (Wrobel *et al.*, 2003). Slides with different coatings are available, and are used depending on the nature of the probes (Taylor *et al.*, 2003). There are two classes of slides that are widely used for arraying either PCR products or oligonucleotides. Slides for arraying PCR products are usually coated with polylysine or aminosilane. These surfaces bind unmodified DNA covalently via negatively charged phosphate groups (Figure 1.4a). The second class of slides, which are modified with aldehyde (Figure 1.4b) or epoxy groups (Figure 1.4c), are usually recommended for printing oligonucleotides that have an aminolinker. These "active" binding surfaces have the advantage that all of the molecules are bound via at the same molecule and have the same orientation on the slide e.g. the aminogroup of the modified oligonucleotide binds to the active groups on the slide. This is important for relatively short oligonucleotide sequences, where steric hindrance can significantly affect the specificity and stability of target

susceptible than humans to error. Also, because most systems are now fully automated, these can save the scientist repetitive and laborious PCR work.

Other robotics are used for the precise placements of the probe onto the solid surface (Schena, 1996, Cheung *et al.*, 1999, Thompson *et al.*, 2001). Laboratory arrayers use small pins for the deposition of the probes onto glass slides or membranes. There are various printing systems, but most arrayers use split-pin technology (Figure 1.5A). These pins are capable of producing microarrays comprised of up to 40,000 probes on a single microscope slide. The split-pin is a tiny capillary with a total volume of several nanolitres. Picolitres of probe are deposited each time the pin touches the surface of the array.

Alternative technologies like the ring and pin system and inkjets are available for producing arrays. Ring and pin technology uses multiple (usually eight) pin/ring pairs; the pin and ring of each pair can be moved up and down separately but are kept constant in relation to one another with respect to the XY plane. The ring is immersed in a sample well so that an aliquot is held in the ring's centre via surface tension. A spring-loaded pin is then driven through the ring effecting probe deposition (see Figure 1.5B). The system can probe fluids with significantly different viscosities on a variety of flat-surface substrates (Sinclair, 1999, Holloway *et al.*, 2002). The arrays are composed of probes with a small diameter and a good morphology. Arrayers using piezoelectric (inkjet) technology are non-contact printers and can therefore be used on any microarray surface. Nanolitres of probe solution are "fired" onto the surface from a very small distance. Arrays created using this method do not have

Target labelling is done by the direct or indirect incorporation of suitable labels (i.e. fluorescent, radioactive or biochemical). The label that is most suitable is determined by the nature and characteristics of the surface on which the array is printed. Targets used for hybridisation on membrane arrays are not usually labelled with fluorescent markers, because the auto-fluorescent characteristics of the membrane make the detection of a fluorescent probe problematic (see also section 3.2.2, page 97). Instead, targets used for hybridisation against membranes are more commonly labelled using radioactive tracers (e.g. radioactive phosphor) or hapten labels (e.g. digoxigenin) (Kuhnert *et al.*, 1997, Amon and Ivanov, 2003). Target nucleic acids for glass slide hybridisation are produced using fluorescently labelled nucleotides. The fluorescent labels attached to these nucleotides are usually members of the cyanine series that are large aromatic molecules. Two of these fluors that are intensively used in array technology are 1-ethyl-2-[(1*E*,3*E*)-5-(1-{6-[(2,5-dioxo-1-pyrrolidinyl)oxy]-6-oxohexyl}-3,3-dimethyl-5-sulfo-1,3-dihydro-2*H*-indol-2-ylidene]-1,3-propadienyl}-3,3-dimethyl-5-sulfo-3*H*-indolium (Cy3) and 1-ethyl-2-[(1*E*,3*E*)-5-(1-{6-[(2,5-dioxo-1-pyrrolidinyl)oxy]-6-oxohexyl}-3,3-dimethyl-5-sulfo-1,3-dihydro-2*H*-indol-2-ylidene]-1,3-pentadienyl}-3,3-dimethyl-5-sulfo-3*H*-indolium (Cy5). The differences in the trade names (Cy3 and Cy5) refer to the number of C-atoms in the single/double bond chain between the large aromatic groups. The emission maxima of these two fluors are well separated so that interference of signal during data acquisition is avoided, and these fluors can therefore be used in the same hybridisation experiment to compare results from two individual samples (Richter *et al.*, 2002, t Hoen *et al.*, 2003).

Direct labelling methods use transcription or replication to incorporate nucleotides with a label attached directly into the amplified DNA. The large Cy dyes which are attached to one of the nucleotide bases can cause problems for efficient incorporation, as steric hindrance impairs the activity of some polymerases. Indirect labelling methods firstly incorporate nucleotides carrying a small reactive group, to which a label is attached in a subsequent chemical reaction. The advantage of this approach for the labelling of target DNA is that the modified and natural nucleotides used in the amplification are incorporated with approximately equal efficiency by a wide range of polymerases. The most widely used method for labelling targets indirectly for microarray hybridisation uses amino-allyl-modified dNTPs. The small amino-allyl dNTPs are incorporated into the target nucleic acid at a similar rate to unmodified nucleotides (Richter *et al.*, 2002). The Cy dye is then bound to the amino-allyl dNTPs in the target by ester bonding. This approach gives good yields of highly labelled target, and so minimises experimental variation.

After purification, the labelled targets are redissolved in buffered solutions. Some array applications require the presence of competitor DNA during the hybridisation reaction. For example, addition of $C_0t$-1 DNA to the hybridisation solution reduces the reannealing of repetitive elements by binding to the target sequences. Similarly, tRNA acts as a blocker of non-specific hybridisation (Pollack, 2003). The use of formamide in the hybridisation solution has the advantage that the DNA is denatured at lower temperatures (Ideker *et al.*, 2003). Hybridisation at a lower temperature reduces evaporation, which avoids the high background signals associated with drying and excessive probe concentrations.

### Pre hybridisation treatment

Most membranes and slides are pre-treated before hybridisation to prevent non-specific binding of target DNA to the array surface. Membranes are blocked by incubation in blocking buffer containing bovine serum albumin (BSA) (Taylor *et al.*, 2003). When glass slides are used, blocking can be done either with blocking buffers similar to the ones used for membrane blocking, or by incubation in humidity chambers causing the active binding groups on the slide surface to be deactivated (see section 1.3.2) (Chiu *et al.*, 2003).

### Hybridisation

Membranes are treated as traditional Southern blots during hybridisation. Incubation at the desired hybridisation temperature is done in roller bottles, plastic bags or containers in a hybridisation oven or waterbath. For microarray slides a similar approach is possible. Special water tight chambers (Corning or Genetix Limited) that fit single or multiple microarray slides are used for hybridisation (see also Figure 2.6, page 67). The chambers prevent the slide from drying out and can be used for incubation in either a hybridisation oven or water bath. The target is incubated under a coverslip so that several microlitres (µl) of target are evenly spread over the array. Automatic hybridisation stations are also commercially available (Cortese, 2000, Holloway *et al.*, 2002). In these systems, hybridisation takes place in a low-volume chamber with access ports for the addition of buffers. An advantage of this equipment is that the hybridisation process is dynamic, as a pump moves the target solution around over the surface of the slide. This increases signal intensities and ensures that all positions on the array are equally exposed to the target (Holloway *et al.*, 2002). The disadvantages are that the automated systems require a larger target volume than used in coverslip hybridisations, and they are less suitable for large numbers of probes as the ring sealing of the hybridisation area restricts the surface area on which probes can be printed.

**Post-hybridisation**

Washes of increasing stringency follow the hybridisation procedure to remove any non-specifically bound target. The stringency of the post-hybridisation washes affects the strength of the hydrogen bonding between the probe and target and can be adjusted by changing the washing temperature, salt concentration and by adding denaturing agents. A higher stringency is obtained at higher wash temperatures and lower salt concentrations (Sambrook *et al.*, 2001). For the post-hybridisation of glass slides arrays, sodium dodecyl sulphate (SDS) should be avoided in the last wash as this can cause interference with data acquisition because of its auto-fluorescent nature (Massimi *et al.*, 2003).

### 1.3.4 Data acquisition and interpretation

The third and final part of the arraying process is depicted in Figure 1.3C. It includes the data acquisition and analysis and is by far the longest part of the arraying process. When whole genome arrays are used, one experiment often results in thousands of data points, which may require individual analysis. Data analysis software assists this process, but it is still labour intensive. Interaction between scientists, computer analysts and statisticians are important for the correct interpretation of the data.

**Data acquisition**

The results of membrane experiments can be visualised using either X-ray film or a variable mode imager. Membranes can be exposed to X-ray films in the dark and hybridisation signals revealed after photographic development. Variable mode imagers can be used to detect the signals on the hybridised membrane by laser scanning. Unlike exposure to X-ray film, this data acquisition is direct and not cumulative. Therefore hybridisation signals need to be very strong for the imager to be able to detect them (Bertucci *et al.*, 1999).

A variety of fluorescent scanners are now commercially available for data acquisition from glass slide microarrays. The Affymetrix 428, the Axon Genepix 4000 and the Perkin Elmer ScanArray scanners are most commonly used, but others are listed by Holloway and colleagues (2002). All of these employ a similar technology for data acquisition. High resolution scanners contain one or more lasers for excitation of the dyes most commonly used for target labelling. For example, Cy3 is excited at 532 nm and fluoresces maximally at 570 nm while Cy5 is excited at 635 nm and fluoresces maximally at 670 nm. The detector device measures the signal intensity every 10 to 4 μm depending on the scanner. The overall hybridisation signal from each probe is visualised through the sum of all the pixels for that probe in a digital image. A function known as pseudocolouring is often used to aid visual assessment of scans. The faintest probes are usually shown in blue and the stronger probes are shown in increasingly hot colours towards red and finally in white. Monocolour images, usually red for Cy5 and green for Cy3 signal, are used for easy visual comparison of overlaid images. This results in images of the array with either green, red or yellow probes indicating the hybridisation of Cy3, Cy5 or both targets respectively.

### Intensity measurement

A variety of software packages are available for microarray data analysis. Although there are differences, most programs have basic functions to allow the measurement of signal intensities and to present the data in convenient formats (e.g. histograms, scatterplots etc). The available programs are under continual development, and upgrades of them often include improved tools for normalisation and statistical analysis.

The array analysis process is a multi-step process. A customised grid is positioned on top of the digital image by software manipulation, whether the image is derived from a membrane or a glass slide array. This grid includes information about gene identification, size of the probe, background substraction, reference probes etc. For the identified probe size a measurement of the signal intensity is taken. These data are used for the interpretation of the results. Microarray experiments on a whole genome array result in thousands of data probes that are complicated to interpret because of the influence of the many experimental parameters. The data have to be adjusted in such a way that only biological differences contribute to changes in signal intensity (Kroll and Wolfl, 2002, Quackenbush, 2002). This process is called normalisation, and it is one of the most discussed and important aspects of microarraying.

### Normalisation

Array experiments are susceptible to manual and experimental variation, caused by, for example, unequal quantities of starting material, and differences in labelling and detection efficiencies. It is necessary to exclude any variation arising during the arraying process for the correct interpretation of any biologically significant results.

This is most important in gene expression analysis studies as quantitative differences are measured, however, all data obtained from array experiments need to be carefully adjusted. Many different approaches to data normalisation are currently used and there is debate about which method is best. Two reviews have described and compared the different methods (Kroll and Wolfl, 2002, Quackenbush, 2002).

Background correction is the first normalisation step. By subtracting the background from the signal intensities the first variable between arrays is excluded. After background correction, total intensity or 'global' normalisation methods are used. These methods rely on two assumptions. Firstly, that equal amounts of template DNA have been labelled and, secondly, that the arrayed elements include only genes that give a representative range of signal intensities (Quackenbush, 2002). Normalisation factors are calculated by summing the overall signal intensities from both scans separately, then assuming that the sum values should be identical, each individual data point is transformed using this factor. Variations on this ‘total intensity’ approach use mean or median value of all or a selected subset of probes to achieve global normalisation. All these methods are designed to minimise experimental variation. However, each method will be more or less suitable for normalisation of particular data sets as discussed by Kroll and Wolfl (2002). To summarise, all total intensity normalisation approaches are applicable to a wide variety of gene sets and are easily computed and applied. All are good for noise reduction in the data. Mean normalisations are influenced strongly by non-linear behaviour of high intensity genes and strong signals from single genes. Median normalisation is less often applicable when many genes have low signal intensities, as it skews the data. Additionally, other more complicated normalisation algorithms have been designed for array data and these are mostly used in expression analysis. One example is locally weighted linear regression (Lowess) normalisation which is used to remove intensity dependent effects in the ratio value (Cleveland, 1979, Yang *et al.*, 2002). In contrast to global normalisation methods, this is a non-linear adjustment of the data. The relative error increases for probes with a lower intensity, and therefore the signal intensity is adjusted differently for low and high signal intensities. In general these algorithms are more difficult to compute and apply, but are sometimes more precise in filtering out experimental differences.

An acceptable standard needs to be determined for the interpretation of microarray data, whether it concerns the presence or absence of the genes on the array, or measurement of over/under expressed genes. For genotyping applications, this can be done by the hybridisation of strains with a previously defined hybridisation pattern. The minimal signal for known positive probes is used as a threshold to score all other genes on the array. For gene expression experiments a ratio value is set to determine over and under expressed genes. Data are usually transposed to log-ratio format for statistical purposes (Quackenbush, 2002). When looking at expression data a minimal log-ratio fold change of two is seen as an up-regulation and in the same way down-regulation of a gene is indicated by probes with a log-ratio lower than 0.5 (Schena, 1996).

Normalisation will continue to be a much discussed topic in microarray technology. Standardisation of normalisation methods and the close collaboration between biologists, computer scientists and statisticians will hopefully allow a better understanding of the normalisation of array data, and establish easy-to-apply algorithms for all types of microarray data as comparing different data sets currently proves rather difficult.

**Data mining**

Microarray experiments generate a large amount of data even before normalisation and transformation of the results. Not only the scanned digital microarray image of several 100 Mb but also the intensity measurements have to be stored. Additionally, the raw data are normalised and, when appropriate, further transformed using other algorithms, causing at least a doubling in size of the data file. Guidelines have now been established for the additional information that needs to be stored together with the data, so that microarray data can be exchanged between different laboratories (Brazma *et al.*, 2001, Spellman *et al.*, 2002). The most appropriate method for storing this type of data is in databases. Although any database software can be amended with scripts for the handling and filtering of the data, specialised microarray packages are more suitable (e.g. Genespring, Silicon genetics; GeneTraffic, Iobion; BioNumerics & GeneMaths, Applied Maths).

## 1.4 Aims

The main hypothesis behind this work is that a subset of genes derived from the genome of a bacterium can be used to assess whether or not a strain of *E. coli* carries these genes and can characterise the strain. A general hypothesis was that DNA arrays can also be used for typing organisms without the need for whole genome sequencing.

In order to achieve this, specific objectives of the work described in this thesis were:

- To determine which type of commercial whole genome arrays would produce the most accurate results for genotyping bacterial strains. The two types of whole genome arrays that were available were composed of PCR amplicons printed on membranes or oligonucleotide probes on glass slides.
- To determine whether it is possible to identify candidate genes characteristic for the individual *E. coli* strains that could be used as sub-typing markers for DNA 'fingerprinting' of strains. The work was designed to test the hypothesis that interpretation of array data would lead to the identification of suitable of marker genes of *E. coli* that could be used on a subtyping array.
- To determine whether a custom made pathogenicity marker array would show the distribution of pathogenicity markers in *E. coli* isolates. The work was designed to test the hypothesis that certain pathogenicity markers are widely distributed in all isolates. Such an array was produced and then used to investigate selected *E. coli* strains.

- To determine whether custom made pathogenicity marker arrays could be used for the identification of different pathotypes in individual clinical specimens.
- To determine whether an extended array could be used to distinguish between closely related isolates from outbreak situations. The work was designed to test the hypothesis that an array including a wider selection of pathogenicity markers and sequences could characterise clinical isolates in greater detail.

The BMEG collection was established at the Shrewsbury Public Health Laboratory over a period of two weeks in January and February 2000. All incoming urine samples with a white blood cell count higher than 100 x $10^6$/litre were selected for the evaluation of different types of agar for rapid bacterial identification (Fallon *et al.*, 2002). *E. coli* isolates from this collection, identified by API, comprise a unique collection for the investigation of the distribution of pathogenicity markers and the relationship between strains collected in the same geographical area.

Ten UPEC isolates with multiple antimicrobial resistance patterns were isolated from individuals infected with an outbreak of urinary tract infection. These strains were previously indistiguisable using API and PFGE and were challenged against the pathogenicity marker array to find any differences between them.

All isolates were grown first on agar plates (Luria Broth (LB) or Cystine Lactose Electrolyte-Deficient (CLED)) overnight at 37°C, and where necessary single colonies were grown in LB cultures overnight at 37°C. Cultures were stored on beads (Technical Service Consultants Ltd.) at -70°C.

## 2.2 Antimicrobial susceptibility testing

Antimicrobial susceptibility testing of the *E. coli* isolates in the BMEG collection was carried out using Oxoid antibiotic disks. Isosensitest agar plates (Chester Media Service) were inoculated with bacterial cultures of low cell density (McFarland scale 0.5). Antimicrobial resistance disks for trimethoprim, ampicillin, cephalexin, norfloxacin, augmentin and nitrofurantoin were then placed firmly on the plates.

Cells were harvested from plates with a 1 µl culture loop or from suspensions by 1 minute centrifugation at 10,000 x g and resuspended in 180 µl Qiagen ATL buffer. Cells were incubated at 55°C after addition of 20 µl proteinase K (20 mg/ml) and vortexed occasionally until lysis was complete. Two hundred µl Qiagen AL buffer were added and the sample was immediately mixed thoroughly before incubation at 70°C for 10 min. After addition of 200 µl 96-100% ethanol (Sigma) homogeneous solutions were transferred into the DNeasy column. DNA was bound to the column by a 1 minute centrifugation at 10,000 x g. The column was washed twice by centrifugation with 500 µl wash buffers Qiagen AW1 and AW2 respectively. The dry column was placed in a clean 2 ml collection tube and the DNA was eluted in 400 µl nuclease-free $dH_2O$ (Promega). DNA was stored at -20°C.

### 2.3.2 MagNA Pure Automated DNA extraction

The MagNA Pure automated extraction robot (Roche) also uses proteinase K to lyse the cells and then captures the extracted DNA on magnetic beads. Genomic DNA was extracted from single colonies grown overnight in 1 ml suspension in LidBac culture tubes (Eppendorf). Cells were washed and resuspended in 100 µl PBS and together with the necessary solutions loaded onto the machine. Genomic DNA was extracted using the DNA I high performance protocol. The robot is able to extract 32 samples in one extraction run in approximately 1.5 hours. After extraction of the genomic DNA, all samples were treated with RNase A (10 mg/ml, Sigma) for 1 hour at 37°C. DNA was stored at -20°C.

## 2.4 Quantification and Qualification of DNA

### 2.4.1 Spectrophotometrical analysis

The quantity of genomic DNA extracted was spectrophotometrically determined by measuring the UV absorbance. Extracted genomic DNA was diluted in nuclease free $dH_2O$ and read in an Eppendorf Biophotometer (Eppendorf) against nuclease free $dH_2O$. DNA dilutions were adjusted if the reading taken at 260 nm ($A_{260}$) was not within the 0.1 – 1 linear range. A clean DNA solution with an $A_{260}$ of 1.0 has a concentration of 50 µg/ml. The concentration of an unknown sample is calculated by:

**DNA conc (µg/ml) = [$A_{260}$[[50][dilution factor]**

The UV absorbance at 280 nm ($A_{280}$) was also measured to determine the purity of the DNA. A clean DNA sample of good quality will have an $A_{260}/A_{280}$ ratio of 1.8. Impurities such as proteins will decrease this value as their absorption will cause an increase in $A_{280}$, indicating a less pure sample and leading to an incorrect quantification.

### 2.4.2 Agarose gel electrophoresis

DNA in extracts was separated by electrophoresis on 1 % w/v agarose gels to check their integrity. Agarose gels (SeaKem® LE Agarose, Cambrex) were prepared in 1 x TBE buffer (Invitrogen) and heated in a microwave oven until dissolved. Gels were cooled and allowed to set for at least 30 minutes at room temperature before 3 µl of sample, mixed with 2 µl Orange G loading buffer (Severn Biotech) and 5 µl $dH_2O$, were loaded onto the gel. Samples were electrophoresed at 8 V/cm until the orange loading dye migrated to the bottom of the gel. Gels were stained in ethidium bromide (5 µg/ml) in distilled water. Visualisation of the gel was with UV light (312 nm), and images were captured on Polaroid film or with the use of a CCD camera in Geldoc system (Biorad).

## 2.5 DNA digestion

Genomic DNA samples with an $A_{260}/A_{280}$ ratio between 1.77-1.83, were digested before labelling. Only the labelling with Cy dyes was done directly on heat denatured genomic DNA extracts. For the digestion approximately 1.5 µg of DNA was mixed with 10 µl 10 x digestion buffer and one unit restriction enzyme (*EcoR*I and *Mse*I; New England Biolabs). The total volume of the reaction was adjusted to 100 µl with nuclease free water. Digests were incubated for at least one hour at 37°C and samples were analysed by electrophoresis on a 1% w/v agarose gel as described in section 2.4.2 to determine digestion efficiency.

## *2.6 DNA labelling*

### 2.6.1 Digoxigenin labelling

To generate hybridisation probes the smaller *Mse*I digested DNA fragments were labelled with digoxigenin using the DIG High Prime DNA Labelling kit (Roche). DNA was heat denatured by a 5 minutes incubation at 95°C followed immediately by 5 minutes on ice. Each labelling reaction contained 1 µg heat denatured DNA and 4 µl DIG-High Prime mixture containing random hexamer primers, nucleotides, DIG-dUTP, Klenow enzyme and buffer components in a total volume of 20 µl. The reaction tube was incubated at 37°C for 1 hour and the reaction was stopped by the addition of 2 µl 0.2 M EDTA pH 8.0 (Invitrogen). The labelled DNA preparations were used as hybridisation probes, and were cleaned from unincorporated nucleotides and enzyme using Microcon spin columns (YM-100, Millipore). The signals from a dot-blot experiment of labelled and control DNA were compared to define the incorporation efficiency. Seven serial dilutions were made ranging from10 pg/µl to 0.01 pg/µl and spotted onto Hydrobond-N$^+$ membrane (Amersham Biosciences), crosslinked and washed for 2 minutes in 20 ml maleic acid buffer (Roche). Membranes were incubated for 30 minutes in 10 ml blocking solution followed by 30 minutes in 10 ml antibody solution, and washed 2 times for 15 minutes in 10 ml washing buffer. Membranes were equilibrated in 10 ml detection buffer for 2 minutes before visualisation with the chemiluminescence substrate ‘CSPD ready-to-use’ (Amersham Biosciences) that was applied to the membrane and incubated for 5 minutes at room temperature. The excess liquid was removed and the membranes were exposed to ECL-Hyperfilm (Amersham Biosciences) for 15-25 minutes in the dark (details of development in 2.7). Signal intensities were compared to the control material to calculate the amount of DIG-labelled DNA. Probes for membrane hybridisation were heat denatured for 5 minutes at 95°C and cooled on ice for 5 minutes before overnight hybridisation.

### 2.6.2. Fluorescent labelling

Genomic DNA was fluorescently labelled using an enhancement chemifluorescence (ECF) random prime labelling kit. (Amersham Biosciences). Up to 2 µg of *Mse*I digested, heat denatured DNA in a maximum volume of 34 µl was mixed with 10 µl nucleotide mix, 5 µl primer solution and 1 µl enzyme solution supplied by the manufacturer to make a total reaction volume of 50 µl. Reactions were incubated between 1 to 3 hours at 37°C and terminated by the addition of 2 µl 0.2 M EDTA (pH 8.0). Hybridisation probes were cleaned from unincorporated nucleotides and enzyme using Microcon spin columns (YM-100, Millipore) and semi-quantified before use as described below.

The signals from a dot-blot experiment of labelled and control DNA were compared to define the incorporation efficiency. Seven serial dilutions in TE buffer of the 5 x nucleotide mix, ranging from 1:5 to 1:500, were spotted onto Hydrobond-N$^+$ membrane (Amersham Biosciences). A second filter was made with the 5 µl labelled probe and a negative control. Both filters were placed on filter paper moistened with TE buffer. Signals were viewed using UV light. Labelled DNA with a comparable intensity to the 1/100 control dilution was considered to be sufficiently labelled for hybridisation experiments. Probes were heat denatured for 5 minutes at 95°C and cooled on ice for 5 minutes before use.

### 2.6.3 $^{33}$P labelling

For the labelling of probe DNA with $^{33}$P, a commercial random prime kit was used (Rediprime II, Amersham Biosciences). Up to 1 µg of *Mse*I digested DNA in 45 µl TE buffer was denatured for 5 minutes at 95°C and cooled on ice for 5 minutes. The denatured DNA was transferred to a ready made reaction tube containing a buffered solution of dATP, dGTP, dTTP, exonuclease-free Klenow enzyme and random primers in a dried stabilised form with a light blue colour. Five µl of radioactive labelled dCTP (Redivue $^{33}$P dCTP, Amersham Biosciences) were added and the contents of the tube were mixed until the solution appeared purple (following mixing of the pink coloured Redivue $^{33}$P dCTP with the blue stabilised pellet of reaction mix). The reaction was incubated at 37°C between 1 and 3 hours and stopped by the addition of 5 µl 0.2 M EDTA. For the use in hybridisation experiments the DNA was heat denatured for 5 minutes at 95°C and cooled on ice before use. The specific activity was estimated by the theoretical incorporation as described by the manufacturer, or measured using a scintillation counter; both described below.

The specific activity of the probe was calculated using the manufacturer handbook. Firstly, the total amount of DNA at the end of the reaction has to be determined using the following formula:

$$\textbf{Mass of DNA (ng)} = \frac{[\mu\text{Ci added}][13.2][\%\text{incorporation}]}{\text{Specific activity of }[^{33}\text{P}]\text{dCTP}} + \textbf{starting template (ng)}$$

Secondly the amount of radioactivity incorporated in disintegrations per minute (dpm) can be calculated using the following formula:

$$\textbf{Activity incorporated (dpm)} = [\mu\text{Ci added}][2.2 \times 10^4][\%\text{incorporation}]$$

The specific activity is calculated by a division of these two:

$$\textbf{Specific activity dpm/}\mu\textbf{g} = \frac{[\text{dpm incorporated}][10^3]}{\text{Mass of DNA (ng)}}$$

Radioactive probes were also measured for the incorporation of radioactive phosphor with a scintillation counter. One µl of labelled DNA, radioactive label and a negative control were spotted onto individual pieces of DE81 filter paper (Whatman) and washed 3 times for 5 minutes in 10 ml 5 % $Na_2HPO_4 \cdot 12H_2O$, 5 minutes in deionised water and 2 minutes in industrial methylated spirit. Filters were dried on filter paper and placed into a scintillation vial. Three ml scintillation fluid was added (Ecoscint A, National Diagnostics) to the vials and they were counted in a Beckman LS 1801 scintillation counter (Beckman Instruments). Unwashed and blank filters were included for baseline measurement.

Probes that were not used immediately were purified using NAP$^{TM}$-5 columns (Amersham Biosciences). The excess liquid was poured off the top of the column and after removal of the bottom seal the column was equilibrated using 10 ml TE buffer pH 7.0. The sample in a maximum volume of 50 µl was applied to the column and absorbed in the gel bed. DNA was eluted in 1 ml TE buffer. The purified probes were stored at -20°C.

### 2.6.4 Fluorescent Cy labelling

Fluorescent probes for hybridisation to the glass arrays were made using a modification of a random labelling kit (Bio-Prime, Invitrogen). Heat denatured genomic DNA (1.5 - 2 μg) was used as template. Each reaction contained 20 μl random primers (750 μg/ml), 5 μl dNTPs (1.2 mM each of dATP, dCTP, dGTP and 0.6 mM dTTP), 1 μl Klenow enzyme (40 units/μl) and 1 μl Cy labelled dUTP (25 ng/μl) in a total volume of 50 μl. The reaction was incubated at 37°C in the dark for at least 3 hours. After labelling, QIAquick PCR purification columns were used to remove unincorporated dyes. Samples were mixed with 5 volumes buffer PB and applied to a QIAquick spin column. DNA was bound to the column by centrifugation for 1 minute at 10,000 x g. Filters were washed with 700 μl buffer PE. The flow-through liquid was discarded and the column was dried by an additional spin for 1 minute at 10,000 x g. Targets were eluted in 50 μl nuclease free water. A sample of 1 to 2 μl was run on a 1 % w/v agarose gel and scanned with the Typhoon scanner to investigate incorporation of the Cy dye.

The specific activity of the incorporation of the fluorescent probes was determined as described by Murray *et al.* (2001). The formula given below calculates the total number of nucleotides divided by the Cy5 dye labelled nucleotides i.e. the lower value the more fluorescent dye is incorporated.

$$\textbf{Specific Activity (ng/pmol)} = \frac{[\textbf{Labelled target (ng)}]\ [\textbf{1000}]}{[\textbf{Incorporated Cy5 (pmol)}]\ ]\textbf{324.5}]}$$

in which:

$$\textbf{Labelled target (ng)} = [\textbf{A}_{\textbf{260}}]\ [\textbf{50}]\ [\textbf{volume} (\mu\textbf{l})][\textbf{1000}]$$

$$\textbf{Incorporated Cy5 (pmol)} = \frac{[\textbf{A}_{\textbf{650}}]\ [\textbf{volume} (\mu\textbf{l})]}{[\textbf{0.25}]}$$

For Cy3 labelled targets the following formulas apply.

$$\textbf{Specific Activity (ng/pmol)} = \frac{[\textbf{Labelled target (ng)}]\ [\textbf{1000}]}{[\textbf{Incorporated Cy3 (pmol)}[\ [\textbf{324.5}]}$$

in which:

$$\textbf{Labelled target (ng)} = [\textbf{A}_{\textbf{260}}]\ [\textbf{50}]\ [\textbf{volume} (\mu\textbf{l})][\textbf{1000}[$$

$$\textbf{Incorporated Cy3 (pmol)} = \frac{[\textbf{A}_{\textbf{550}}]\ [\textbf{volume} (\mu\textbf{l})]}{]\textbf{0.15}[}$$

After spectrophotometrical analysis of the targets these were dried under vacuum in a Speed Vac (Savant) and redissolved in 30 µl 1 x hybridisation buffer (see Appendix I).

## *2.7 Southern blotting and hybridisation*

Genomic DNA and restriction digests were separated by agarose gel electrophoresis on 1 % w/v gels and were transferred onto Hydrobond-N$^+$ membrane (Amersham Biosciences). The gel was depurinated for 30 minutes in 0.25M HCl. This step partially depurinates DNA into smaller fragments that are more easily transferred to the membrane. This was followed by denaturation in 1.5M NaCl/0.5M NaOH twice for 20 minutes causing the double stranded DNA to denature. Finally, the gel was neutralised in 1.5M NaCl/0.5M Tris HCl twice for 20 minutes. This final step raises the pH of the gel as DNA will bind less efficient to the membrane at a lower pH. The gel was very fragile after these washes and had to be handled with great care while building the blotting stack tower. The transfer buffer (20 x saline sodium citrate (SSC), Invitrogen) in the reservoir transferred the DNA to the membrane by upwards capillary action through the filter wick. The gel was placed on top of the wick and a plastic seal was placed around the gel. The pre-wetted membrane was placed on top of the gel. Filter paper and tissues were used to absorb the transfer buffer. The tower was completed with a glass plate to distribute the pressure of a heavy weight. After overnight blotting, the blotting tower was disassembled and the reduced gel was checked by UV light for adequate transfer, i.e. when no significant ethidium bromide stained DNA was detectable in the gel, the transfer had been completed. Positions of the lanes were indicated on the membrane by pencil marks for orientation purposes. Membranes were UV-crosslinked for one min on a UV light source (wavelength 312 nm) for covalent binding of the DNA to the membrane as shown in Figure 2.2. Membranes were used directly for hybridisation or wrapped in plastic and kept at 4°C.

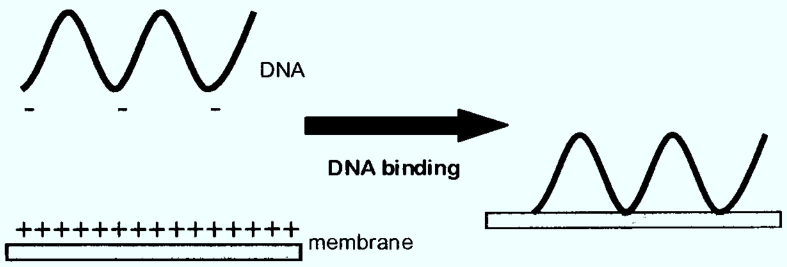

Figure 2.2 Covalent DNA binding to Hydrobond-N$^+$ membrane.
The negative phosphate backbone binds to the positively charged membrane.

washed. A chemiluminescent substrate to the antibody conjugate was added before exposure and visualisation of the membrane.

Hybridisation signals were detected by phosphor imaging on a Typhoon 8600 variable mode imager (see Figure 2.10) or on film. Blots were exposed to ECL Hyperlink film (Amersham Biosciences) for 25 min to 1 hour and then developed and fixed using Kodak solutions. One in 4 dilutions of X-ray developer and fixer were prepared in tap water. The film was immersed in developer and agitated until signal became visible. Films were immersed in the fixing solution for twice the clearing time and air-dried. Radioactive labelled probes hybridised to Southern blots were visualised on X-ray film or phosphor storage screens (see section 2.10.1)

## *2.8 Whole genome membrane arrays*

Commercial membrane arrays were used for full genome analysis of radioactive probes derived from *Mse*I digested DNA as described in 2.5.2. Each of the membranes (Sigma-Genosys) contains 4,290 PCR amplified ORFs of *E. coli* K12. The majority of ORFs had been amplified from start to stop codon. All 4,290 ORFs were printed in duplicate at 10 ng per probe onto positively charged nylon membranes. DNA was bound covalently onto the membranes by cross-linking using UV-light. The array consists of three fields, each field has a primary grid composed of 16 rows (A-P) and 24 columns (1-24) and a secondary grid with four genes printed in duplicate in a staggered formation. The corners of each field (A1, A24, P1, P24) contain genomic *E. coli* K12 DNA. These probes act as positive control orientation probes and can be used to normalise data between replicate arrays. A layout of the

Before first use the arrays were washed in 50 ml 2 x saline sodium phosphate-EDTA (SSPE) for 5 min. Prehybridisation buffer consisting of hybridisation solution (Sigma-Aldrich) supplemented with salmon testes DNA (Sigma) to a final concentration of 100 µg/ml was prewarmed to 65°C. Incubation in 5 ml buffer was performed in a hybridisation oven in roller bottles at 65°C between 1 to 2 hours. Radioactive labelled probe was heat denatured at 95°C for 10 min in 3 ml hybridisation solution and cooled on ice. The prehybridisation buffer was replaced by the denatured labelled target DNA in fresh hybridisation buffer before overnight hybridisation at 65°C. Post hybridisation membranes were washed 3 times for 2 minutes at room temperature in 50 ml wash buffer (0.5x SSPE and 0.2% SDS) and 3 times 20 min at 65°C in 80-100 ml wash buffer. Arrays were sealed in plastic and exposed to a Kodak Low Energy Storage Phosphor screen for 24 hours before visualisation using a Typhoon 8600 scanner (see Figure 2.10).

To remove the radioactive label from the membranes they were incubated in preboiled stripping buffer for 20 minutes. After draining the excess solution the membranes were again exposed to the storage phosphor screens to confirm that no signal residue was still present, before reusing the arrays. During hybridisation, stripping and storage the membranes were not allowed to dry completely as this makes stripping and reprobing less efficient.

### 2.9.2 First generation pathogenicity marker array

A subset of PCR amplified pathogenicity markers was spotted onto aminosilane coated glass slides (CMT-GAPII, VWR) for the screening of pathogenicity markers in *E. coli*. Probes for arraying were prepared from clones by Kuhnert and colleagues (1997, 2000) The genes were selected on the basis of involvement in pathogenesis of UTI strains and also included several well known pathogenicity markers from other pathotypes. Pathogenicity markers were previously cloned into pBluescript plasmids after amplification from pathogenic *E. coli* strains using gene specific primers that are listed in Appendix II.

**Plasmid extraction**

For probe preparation, plasmids containing pathogenicity marker genes were extracted from 100 ml broth cultures using a plasmid extraction kit (HiSpeed Plasmid Midi Kit, Qiagen) following the manufacturer's instructions. Clones were grown in 250 ml flasks in 100 ml LB medium at 37°C overnight under continuous agitation at 225 rpm (imMedia Amp; Invitrogen). Cells were harvested by centrifugation for 15 minutes at 6000 x g. Cell pellets were resuspended in 6 ml buffer P1 containing RNase A and lysed by the addition of 6 ml of buffer P2. Samples were incubated at room temperature for 5 minutes. After addition of 6 ml buffer P3 and mixing by inverting the tube several times, the lysate was directly transferred to the QIAfilter Cartridge and incubated for 10 minutes. During this incubation the Hi Speed Midi Tip was equilibrated, allowing 4 ml QBT buffer to pass through the tip by gravity flow. The lysate was passed through the QIAfilter cartridge into the Hi Speed Midi Tip by gentle pressure from the plunger and allowed to pass through the tip by gravity. The tip was washed with 20 ml buffer QC. Plasmid DNA was eluted in 5 ml buffer QF and precipitated using the QIAPrecipitator Midi module. After addition of 3.5 ml isopropanol (Sigma) and 5 minutes incubation at room temperature the eluate/isopropanol mixture was filtered through the QIAprecipitator. The filter was washed twice with 2 ml 70% v/v ethanol and dried. Plasmid DNA was eluted in 1 ml nuclease free water (Promega).

DNA fragments were digested from the plasmids with restriction enzymes selected during the primer design (shown in Appendix II), using the method described in section 2.5. Products were gel-purified before amplification for the complete removal of vector sequences using the electrophoresis method described in section 2.4.2. Further purification was performed using the Qiagen gel extraction kit. In brief, the DNA bands were excised from the agarose gel and weighed. Twice the volume of the excised agarose gel was added in resuspension buffer and tubes were incubated at 50°C until all agarose was dissolved. DNA was purified through silica columns as described for the PCR purification method (section 2.6.4).

**DNA fragment amplification**

Genes encoding pathogenicity markers were amplified from purified plasmid inserts. One μl of purified fragment was added to 10 μl 10 x PCR buffer, 10 μl 2 mM dNTPs, 3 μl of each primer (20 pmol/μl), one unit *Taq* in a total volume of 100 μl. Amplification was carried out in GeneAmp PCR system 9700 thermal cycler (Applied Biosiences) using the following PCR cycling program. An initial denaturing step for 4 minutes at 95°C was followed by a three step cycle of 30 sec at 95°C, 30 sec at the annealing temperature ($T_A$), 30 sec at 72°C for 30 cycles and a final elongation 7 minutes at 72°C. Products were held at 4°C after completion of the programme. $T_A$ values used for amplification are given in Appendix II. PCR products were cleaned using the QIAquick PCR purification kit (see section 2.5.4) and quantified using spectrophotometrometrical and agarose gel electrophoreisis analysis (see section 2.4).

Equal amounts of amplified product (1.5 µg) were dried and resuspended in 20 µl 50% v/v DMSO and allocated a position in a 384-well plate before arraying.

**Microarray construction**

Arrays were made on a Microgrid II arrayer (Biorobotics) at 25°C and 40 % relative humidity. An image of the instrument is shown in Figure 2.7A. A program was designed for the allocation of the probes on the slide. Four capillary pins, shown in Figure 2.7B, were used for the arraying process. These small pins, 100 µm in diameter, print probes of 180 µm only millimetres apart. The reservoir contained up to 55 nl of PCR product and deposited 50 pl per target visit. Array patterns were designed in such a way that positive and negative controls were used on easy identifiable places for orientation purposes (Figure 2.7C). Products were put on the CMT-GAPII slide (Corning) in random positions to cut out inter-slide variability. Pins were washed before each new 384 well source visit in distilled water, once in both circulating water baths for 2 sec and once in the main wash station for 4 sec, followed by drying of the pins under vacuum. Each slide contained two identical arrays. Each pathogenicity marker array contained 144 probes representing PCR products from *E. coli* pathogenicity marker genes and controls. All probes were printed at least in triplicate. In total there were 30 positive control probes, and 15 negative control probes. The positive controls included a 16S rDNA dilution series and genomic DNA probes. Negative control probes were water and spotting solution.

### 2.9.3 Second generation pathogenicity marker array

In collaboration with Dr. Henry Smith's group at the Central Public Health Laboratory in Colindale a second-generation pathogenicity marker array was made. This array was made to characterise typical and atypical EAggEC and to develop new detection methods targeting genes characteristic of both groups of EAggEC. It included all sequences from the first generation array described in section 2.9.2 and gene sequences amplified directly from EAggEC strains (listed in Appendix III) that could potentially be involved in pathogenesis. This last group of genes was identified by sequence comparison of the genome and plasmid sequence of EAggEC strain 042 to the Genbank database by Dr. Dudley, Baltimore. PCR products were resuspended at a concentration of 75 ng/µl in 50 % v/v DMSO and printed from a 384-well plate onto CMT-GAPII slides. All products were printed in triplicate on each array, which contained a total of 81 probes representing pathogenicity sequences and controls. The probes on the slides were covalently bound to the slides by crosslinking at 2000 x 100 mJ in a UV Stratalinker. The array consists of a primary grid with 2 rows and 2 columns and a secondary grid with 10 columns and 11 rows. A layout is shown in Figure 2.8.

least one of the genes was above the minimum detection value. Genes were filtered using the following MSExcel software function: =IF(AND(OR(A>50000,B>50000),C>3),"pos in strain 1",IF(AND(OR(A>50000,B>50000),C<0.3333),"pos in K12","neg")) in which A = raw intensity data for the probe in strain 1, B = raw intensity data for the probe in K12 , C = ratio of the normalised intensity values.

For the pathogenicity marker arrays, a threshold was calculated by analysing hybridisation patterns of strains with a known pathotype. Probes with a normalised intensity higher than the threshold were called present and probes with an intensity lower than the threshold were called absent. An MSExcel software macro was used to assist in this analysis and is included on the enclosed CD-rom.

## *2.11 Data mining*

Data was stored in a Bionumerics database (Applied Maths). Experiments (e.g. numerical values or assay results) were linked to the entries (e.g. bacterial strains) by a unique key. The three major experiment types in Bionumerics are fingerprint type (e.g. a densitometric result), character type (e.g. an array of well determined values including binary data) and sequence type (e.g. sequence data). The character type was used for the storage and analysis of the array data.

## *3.1 Genomic DNA extraction*

To determine the most appropriate method for recovering intact high molecular weight genomic DNA from bacterial cultures, semi-manual and automated extraction procedures were compared (see section 2.3, page 64). The semi-manual method using silica gel membrane technology in which DNA from a lysate is bound to a filter and eluted after several wash steps (Qiagen method), whereas the automated method used magnetic beads to capture the genomic DNA from a lysate (MagNA pure method; see section 2.3.2). Both methods gave similar yields of genomic *E. coli* DNA (40 µg/ml) from overnight cultures prepared as decribed in section 2.1. When the integrity of the DNA was checked by electrophoresis on agarose gels, the manual extracts showed smears of DNA (see Figure 3.1), whereas the automated extracts showed clear single bands of high molecular DNA (see Figure 3.2). The automated extraction method also recovered rRNA, whereas the manual extraction method had an RNase treatment step incorporated in the extraction procedure. Therefore, DNA recovered by the automated method had to be treated with RNase after extraction to remove RNA which would interfere with subsequent labelling and hybridisation of the DNA. The manual extraction method was done on individual samples and was relatively laborious. In comparison, the MagNA pure instrument could be used to extract 32 samples simultaneously, and was fully automated. DNA was extracted rapidly and there was minimal hands-on time involved in setting up the robot.

efficient method, but hazardous waste is produced. $^{33}$P, which was used in this study, has a shorter half-life than the more widely used $^{32}$P isotope (Amersham Biosciences, 2004). By testing DNA labelled in different ways on similar membranes, the methodologies involved and the specificity of the hybridisation were investigated.

### 3.2.1 Digoxigenin labelling of the hybridisation target

DNA was labelled with digoxigenin by random hexamer amplification of *Mse*I digested genomic DNA. The incorporation efficiency of the digoxigenin labelled dUTP into the genomic DNA was assessed by comparison of the labelled probe and a control probe (provided by manufacturer) on the same membrane. After development of the exposed film, the signals were compared and the concentration of the labelled product calculated. Typical signal intensities are shown in Figure 3.4. By comparing the signal intensity of the labelled DNA probe with a dilution series of labelled control DNA of known concentration, an estimation of the probe concentration was made. In this example, the control DNA and the labelled *E. coli* DNA showed identical signal intensities. The signal of the 1:330 *E. coli* DNA probe that was prepared, was equivalent to the signal of the control probe with a concentration 5 pg/µl (See Figure 3.4). Therefore the approximate concentration of the labelled probe was estimated at 5 pg/µl x 330 = 1650 pg/µl = 1.7 ng/µl.

The average of the duplicate measurements of probe readings was compared to the unwashed sample to determinate the incorporation efficiency (%):

$$\%\text{ efficiency} = \frac{\text{average sample reading}}{\text{unwashed reading}} \times 100\%$$

$$= \frac{(1.35 \times 10^7 + 1.40 \times 10^7)/2}{1.48 \times 10^7} \times 100\% = 93\%$$

The specific activity was calculated by:

$$\text{Specific activity} = \frac{(\text{CPM incorporated} - \text{background})/\text{efficiency}}{\text{Mass DNA } (\mu g)}$$

$$= \frac{((1.35 \times 10^7 + 1.40 \times 10^7)/2 - 152)/0.93}{1} = \mathbf{1.6 \times 10^7}\text{ dpm}/\mu g$$

Both methods showed that the probes were efficiently labelled. The manufacturer's manual gives a formula that allows an estimation of the specific activity, but for an accurate measurement a scintillation counter is necessary. Only labelled probes with a specific activity higher than 1 x $10^7$ dpm/µg were considered to be labelled sufficiently for use in hybridisation reactions.

Membranes prepared from agarose gels by Southern blotting were hybridised overnight, processed and then visualised by autoradiography using X-ray film, or by phosphor imaging after exposure to a storage screen. Examples of these are shown in Figure 3.7 (autoradiography) and Figure 3.8 (phosphor imaging). The use of the Typhoon instrument for detection clearly shows an improvement in the detection of the hybridisation signal. Both undigested and digested DNA targets on the membranes were strong and clearly detectable after hybridisation with the radio-labelled probe.

Specific activity for Cy3 $= \frac{(2.39 \times 10^{3}) \times 1000}{61*324.5} = 120$ ng/pmol

For Cy5 as well as Cy3 hybridisation DNA targets with a "specific activity" below 175 ng/pmol were used for hybridisation.

### 3.3 Discussion DNA extraction and labelling procedures

Molecular characterisation methods used for bacterial typing generally require good quality, pure DNA. The extraction procedures used to obtain such DNA range from manual phenol-chloroform methods and commercial extraction kits (often based on binding DNA to a resin) to automated systems for (semi-) high-throughput screening (Schmidt *et al.*, 1995b, Sambrook *et al.*, 2001, Mygind *et al.*, 2003, Smith *et al.*, 2003). For example, Sambrook and colleagues (2001) describe a phenol-chloroform method for extracting DNA from micro-organisms for use in standard molecular techniques. Smith *et al.* (2003) compared five high throughput extraction kits using DNA binding, DNA filter plates, or metallic beads, and found 96-well plate methods, such as the Montage plasmid Miniprep$_{96}$ kit (Millipore), easiest to use. Mygind and colleagues (2003) also evaluated kits for bacterial DNA extraction, including the Qiagen DNeasy Tissue and the MagNA Pure extraction kit, which were also used in this study. They concluded that DNA extracted with the MagNA Pure was of the highest concentration and purity. Automated methods are increasingly favoured for their consistency of DNA quality and purity, and for their ease of use (Mygind *et al.*, 2003). Manual extractions are more laborious, and RNA and inhibitors of PCR need to be removed before the extracted nucleic acid can be used in downstream applications. DNA extraction kits generally shorten labour time, and may increase the purity of the DNA compared to manual extractions using standard molecular techniques. DNA prepared by automated systems can often be used directly in downstream applications and, may give more reproducible results.

The Qiagen extraction kit and the Roche MagNA Pure automated extraction robot were compared as part of this study. Genomic DNA extracted with the MagNA Pure instrument was visible as a strong clear band following gel electrophoresis. The Qiagen extraction kit showed smears of DNA (see Figure 3.1 and 3.2). This may be because the DNA was not forced through a filter during the MagNA Pure process unlike the Qiagen extraction method. DNA sample variation, in terms of yield and $A_{260}/A_{280}$ ratio, was better for DNA prepared with the MagNA Pure than with the Qiagen extraction method. Similar results were found in other studies (personal communication, Dr. J. Logan and Dr. K Edwards, HPA) and are also acknowledged by Roche.

The DNA obtained by any extraction method has to be labelled in a uniform, efficient and reproducible manner if it is to be suitable for hybridisation against whole genome arrays. The labelling of DNA is a well established technique and can be performed either directly, incorporating the labelled nucleotide into the DNA, or indirectly, by incorporating a modified nucleotide into the DNA and then attaching a label in a second reaction (Richter *et al.*, 2002). Various procedures, including fluorescence, digoxigenin and radioactive tracers, efficiently incorporate the tracer into the DNA, and the differing characteristics of the incorporated molecular structures make them more or less useful for different applications (Wang *et al.*, 2002c). Genomic DNA hybridisations were used to identify a suitable way for the hybridisation and detection of targets with the Panorama membrane microarrays. Colourimetric labelling methods are often used for membrane hybridisations (Bertucci *et al.*, 1999). These are easy to use and do not have the drawbacks of stability and safety associated with radioactive probes. The experiments described in section 3.2.1 show that digoxigenin labelled DNA gave only weak hybridisation signals for the detection of digested and undigested genomic DNA on Southern blots (see Figure 3.5). A hybridisation signal was only observed after increasing the probe concentration from the advised 25 μg/ml to 100 μg/ml of hybridisation solution. These observations may have been because the structure of the DNA did not allow efficient labelling, or they may have been due to impurities in the template DNA. The specific activity of the labelled DNA was determined before hybridisation, and it was considered to be sufficient for hybridisation. Impurities in the DNA preparations were not observed in either gel or spectrophotometric analyses.

The use of fluorescent molecules for labelling DNA was investigated to determine if this approach would improve the sensitivity of the signal obtained following hybridisation of the DNA to the arrays. Fluorescent labels are widely used in biological applications (e.g. sequencing or fluorescent electron microscopy or fluorescent-activated cell sorting) (Kaiser *et al.*, 1989, Knutton *et al.*, 1997, Tung *et al.*, 2004). The advantage of using a laser scanner for the detection of the fluorescent labels is that more than one label can be detected in each experiment. Although fluorescent labelled DNA is a good for hybridisations against glass slides, using such DNA for membrane hybridisations leads to problems of detection, such as auto-fluorescence of the nylon membrane, which interferes with data acquisition. Indeed, the Typhoon variable mode imager used in this study was unable to detect the hybridisation signals directly from membrane arrays probed with fluorescently labelled DNA. The signal was amplified by the use of an antifluorescein, bound to alkaline phosphatase. Even after amplification the chemiluminescent signal was barely detectable after exposure of these membranes to film (results not shown). Fluorescent labelling, therefore, was not used for the hybridisation experiments with the membrane arrays.

During the initial hybridisation experiments it was found that when $^{33}$P labelled DNA gave a good signal when visualised with autoradiography (e.g. Figure 3.7). Following overnight exposure of the hybridised blots to a storage phosphor screen the sensitivity of detection increased (e.g. Figure 3.8). The disadvantages of radio-labelling and the longer exposure of the blots to the storage phosphor screens were negligible compared to the increase in sensitivity. The drawback of using radio-labelled probes and the storage phosphor screen was that the detection process was time consuming and would not be applicable for high throughput screening. The membrane arrays would be hybridised with radioactively labelled targets and the glass slides would be hybridised with Cy3/Cy5 labelled targets. Both methods would use random prime amplification for the incorporation of the label. The membrane arrays would be visualised using the storage phosphor screens and the Typhoon instrument. The glass slides would be scanned directly in the Affymetrix 428 microarray scanner.

## 4.1 Introduction

This chapter describes a comparison of two commercial whole genome DNA arrays for typing and comparing *Escherichia coli* (*E. coli*) strains. A whole genome array prepared from all PCR amplified ORFs on nylon membranes (Panorama Array, Sigma-Genosys) was compared with a whole genome array prepared from ORF specific 70-mer oligonucleotides on glass slides (Pan Array, MWG). Both arrays were designed to cover all ORFs in the *E. coli* K12 gene sequence. To determine which array would produce the most accurate results, each type of array was hybridised with the same strains from the ECOR collection. The array data were interpreted to identify possible candidate genes characteristic for individual *E. coli* strains that could be used as sub-typing markers for DNA 'fingerprinting' of the strains.

It was anticipated that the commercial arrays would display strain-specific hybridisation patterns when hybridised with DNA prepared from different *E. coli* strains. This is illustrated in Figure 4.1, which shows how two hypothetical ECOR strains might compare with *E. coli* K12. Some ECOR strains might have multiple copies of genes present in *E. coli* K12 (e.g. the blue gene for ECOR x and the green gene for ECOR y in Figure 4.1A). Therefore the hybridisation signal for these genes would be stronger for those isolates compared to the signal intensity of the same gene in *E. coli* K12. Other genes might be deleted or replaced with a completely new gene (the "yellow" genes in Figure 4.1A) and a hybridisation signal for the original gene in *E. coli* K12 would not be detected. Interestingly, the replacement genes might have been acquired through horizontal transfer and could potentially be genes involved in

## 4.2 Panorama Arrays

### 4.2.1 Whole genome Panorama membrane hybridisations

Initially, the applicability of the Panorama whole-genome membranes (Sigma-Genosys) for a detailed characterisation of *E. coli* strains was investigated. Radio-labelled probes were prepared from genomic DNA isolated from *E. coli* K12 and five strains from the ECOR collection (in the data tables these ECOR strains are labelled ECOR1-ECOR5). Hybridisations against *E. coli* K12 were carried out in duplicate to test reproducibility (in the data tables these two hybridisations are labelled K12_1 and K12_2). Typical results of just one field of the array are shown in Figure 4.2. Every membrane had three of these fields all with different probes. The signal intensities of the probes on the membranes were measured using the specialised microarray software Arrayvision 6.0 (Imaging Research Inc.). A defined grid template was positioned over the digital image and measurements were returned in a large data table including gene identities, raw data and background measurements. This process is illustrated in Figure 4.3. MS Excel software was used for the calculation and interpretation of the data. All raw data can be found on the enclosed CD-rom under Chapter 4/Raw data/Panorama Arrays.

### 4.2.2 Array normalisation

Normalisation is the computational process by which data from different arrays are equalised before analysis (Schena, 2003c). The most appropriate normalisation factor to use was investigated with the data acquired from the membrane arrays. Data were normalised separately using: (a) the average signal of the genomic DNA reference probes; (b) the median value of all probes on the membrane; and (c) the mean value of all probes on the membrane. The appropriate normalisation factors were calculated and all measurements of the probe intensities for that hybridisation were adjusted using these factors. Part of a complete datasheet for the normalisation of one experiment is given in Table 4.1. The full datasheets for normalisation of all hybridisations can be found on the enclosed CD-rom under Chapter 4/Normalised data/Panorama Arrays.

### 4.2.4 Removal of hybridisation target from the membranes

One advantage of membrane arrays compared to glass slides is that the hybridisation probes can be removed from the membrane and the membrane can be reused in a second hybridisation reaction. To investigate whether all detectable signal was removed from the membranes, a "stripping" protocol using SDS buffer was used (see section 2.8). Most of the signal was removed to satisfactory levels as over 4200 probes were stripped of more than 99% of their signal. This included all the genomic DNA positive control probes. The difficulty was that even after stripping, the intensity of the brightest probes was still significantly higher than the background. Therefore, in subsequent experiments, when an unknown sample was to be hybridised with these membranes, probes absent in the tested strain might still give a signal higher than the background. Because of this membranes often had to be stripped more than once to leave intensities well under the background levels. Stripping of the membrane caused unwanted loss of membrane bound DNA that resulted in decreased signal values.

### 4.2.5 Identification of unique hybridisation patterns of *E. coli* strains

Although the results obtained from probing the membranes with *E. coli* K12 DNA were not reproducible, further experiments to shed more light on this, using the ECOR strains, were performed. For example, to investigate the ECOR strains, their hybridisation patterns were compared to the equivalent pattern obtained with the reference strain *E. coli* K12. Ratios of the signal intensities of the tested strain versus the reference strain were calculated after mean normalisation. On the basis of the ratio values probes were divided into three groups. **Group 1**: Probes with a high ratio were those with a higher intensity in the test strain compared with *E. coli* K12. These probes were called positive in the test strain. **Group 2**: Probes with a low ratio were those with a higher intensity in the *E. coli* K12 strain compared with the test strain. These were called positive in K12. **Group 3**: Probes with ratios close to one were considered to be for genes that were either both present or both absent in the reference and test strains.

The cut-off value of interest was set for probes with a ratio higher than three or lower than one third, meaning that the intensity in one strain must be three times the intensity level in the other strain for it to be called a positive in the strain where it was brightest. Probes that had a low intensity value in both arrays were filtered out of the list of potential informative genes, as these were likely to cause unreliable high or low ratios that could skew the data and lead to misinterpretation of the analysis. The analysis described above was performed using MS Excel software using the function described in section 2.10.2. Venn diagrams, shown in Figure 4.7, were used to get an initial overview of the size of the three different groups.

the mean intensity of all ORF representing probes or the median intensity of all ORF representing probes.

Normalisation against the average signal intensity value of the genomic DNA reference probes led to two problems. Firstly, the signal intensities of the genomic DNA probes were very different in value across the membrane. It would not have been appropriate to simply average these values and use the mean for normalisation since the use of this value would have introduced systematic errors. The membrane arrays were normalised per field using the average value of the signal intensity of the genomic DNA reference probe in that field. Analysis of the data in this way may have lead to misleading interpretations. Secondly, normalisation against the genomic DNA reference probes was not possible for the Pan arrays used later in this project, as these did not contain comparable reference probes. The median of the overall intensity values can easily be influenced by extremely low hybridisation signals (Kroll and Wolfl, 2002), which appeared in the Panorama arrays. The gene representation printed on the Panorama and Pan array was identical, with the exception of the positive and negative control genes. It was therefore concluded that using the mean normalisation factor method was most appropriate to compare normalised data from the Panorama and Pan arrays.

Microarray gene expression experiments require a high number of replicates for the data to be reliable (Churchill, 2002). There is a high level of variability that can occur during the experiment, and the large number of probes on microarrays cause normally acceptable levels of false positives or negatives to lead to misidentification of many genes (Lee *et al.*, 2000). The work described using membrane arrays gave results that were not satisfactory from the point of view of reliability and reproducibility. It may not be completely unexpected to find a poor reproducibility in the membrane hybridisations from just one replication experiment. One solution would be to repeat the experiments multiple times and thereby remove outliers and average the reproducible data. The variation in signal intensities of a repeated experiment can be the result of many aspects of the labelling and hybridisation process. In the experiments described in section 4.2.3, Figures 4.5 and 4.6, labelled target DNA was made from one genomic DNA extract but labelled in separate reactions. It is possible that the lack of reproducibility observed was because incorporation of the label into the DNA was not comparable for the different targets. When the incorporation of label was measured the specific activities were similar. If targets were not labelled with the same efficiency there should not be a great effect on the reproducibility after normalisation has taken place. Possible variability introduced during the (post) hybridisation reaction should have been eliminated by the normalisation of the signal intensities.

Another possible reason for the lack of reproducibility in probe signal intensities is the random nature of the labelling reaction. Radio-labelled nucleotides may have incorporated into different regions of the genome with different efficiencies leading to variation in hybridisation signal. This might explain the results of the membrane hybridisation with labelled genomic *E. coli* K12 DNA shown in Figure 4.2. All the genes on the array are amplified from the *E. coli* K12 strain and should give probes with constant signal intensities significantly above background after hybridisation with the *E. coli* K12 labelled genomic DNA target. When an array hybridised in this way was examined it was noticed that the signal intensities were inconsistent. Some probes were very intense while others were only just detectable. Although the random nature of the labelling reaction is a possible explanation of why replicated experiments were not reproducible when membranes were used, similar problems were not observed with the fluorescently labelled targets hybridised against the glass slide arrays, which were also labelled using a random amplification. Another explanation is that the inconsistency in the signal intensity of the probes is caused by differences in concentration of spotted PCR product during the production of the membranes by the manufacturer. For example, there could be variability in either the quantity of probe delivered to the membrane or in the efficiency of binding to the membrane surface. Although this remains the most likely explanation for the poor quality of the results obtained on the membranes no further information is available to support this hypothesis.

Even though the membrane hybridisation data lacked reproducibility, the results derived from the hybridisations were analysed. It was anticipated that the number of probes giving a different hybridisation signal for *E. coli* K12 and the ECOR test strains would differ to a greater extent than when replicate K12 target preparations were compared. There were 548 variable probes (258 and 290; see Figure 4.7A) identified from the *E. coli* K12 hybridisation experiments.

After normalisation of the signal intensity against the mean value, ratios of the tested ECOR strain versus *E. coli* K12 were calculated as described in section 2.9.2. On the basis of the value of that ratio, genes were divided into the three groups described in section 4.2.5. Most genes were part of the third group (i.e. their signal intensity did not differ significantly between the ECOR and *E. coli* K12 strain). Probes in group 3 can be described as core genes of *E. coli* (Welch *et al.*, 2002, Anjum *et al.*, 2003, Smalley *et al.*, 2003). Genes with very different hybridisation patterns were also detected. Genes identified in the second group, with a lower signal intensity than found in the hybridisation of *E. coli* K12 DNA, can be explained in three ways. Firstly, the genes may be absent from the ECOR strain. Secondly, fewer copies of the gene may be present in the ECOR strain compared to the *E. coli* K12 strain. Thirdly, the gene may have a sequence similarity less than 100% with the gene in *E. coli* K12, and therefore the hybridisation of genomic DNA from ECOR strain is less efficient compared to the signal obtained from the hybridisation with *E. coli* K12 genomic DNA. The genes in the first group, with a higher intensity in the tested ECOR strain can be explained by genes that are present in multiple copies in the ECOR strain compared to the *E. coli* K12 reference strain. Some of these genes were unique to just one of the ECOR strains and could therefore be probes with potential to be used for the characterisation of the individual strains. These should only be considered in combination with the missing genes as these genes were amplified from the *E. coli* K12 genome and can therefore not be characteristic for just that ECOR strain.

The *E. coli* K12 self-hybridisation experiments on the membrane arrays showed 258 and 290 differences (see Figure 4.7). A student t-test revealed that the two data sets were significantly different. When assuming that both data sets obtained in the K12 duplicate experiment are normally distributed, a paired t-test showed a probability of P = 0.0001 for the comparison of the mean of the replicated data. The variance of both distributions was also significantly different (P = 0.0026). It was expected that a strain different from K12 would have an increase in genes with a different hybridisation pattern. Therefore, it was expected that the investigation of the hybridisation patterns of the ECOR strains against K12 would indicate more than 290 differences. However, analysis of the results of the hybridisation patterns of five ECOR strains showed that this is not the case (see section 4.2.5). Only the ECOR1 strain gave slightly more than 290 gene differences. In total, 313 genes gave a more intense signal in this strain compared to the signals from the *E. coli* K12 hybridisation. Statistical tests of the tested strains and the *E. coli* K12 on the different groups as presented in Figure 4.7 show a significant difference between the expected values for *E. coli* K12 as detected through the replication experiment and observed values in the tested strains. The Chi-square probabilities for the tested strains in comparison to the duplicated *E. coli* K12 experiment are all lower than 0.04. The groups of present and absent probes contain significantly different numbers from what was expected on the basis of the reproducibility experiment. Also the probabilities of the tested strains mean are significantly different compared to the duplicated K12 experiment ($P<<0.0001$). So although the distributions are different, this has not reflected in an increase of possible markers. It is therefore challenging to identify the genes that are truly different between K12 and other *E. coli* by this membrane-based, whole genome array hybridisation approach.

The genes belonging to group 1 and 2 were analysed according to their function. All groups had at least one difference in probe hybridisation signal. Some of the functional group (i.e. fatty acid and phospholipids metabolism and membrane proteins) had only few differences suggesting that these functional groups are very well conserved in *E. coli.* Many of the potentially interesting genes are included in the group of hypothetical, unclassified and unknown genes, which are by far the largest category of genes on the array (see Table 4.2). The genome sequences of more

*E. coli* isolates and other closely related species arising from whole-genome sequencing projects will most likely shed more light on the function of these genes. It needs to be taken into account that only genomic DNA was hybridised, and that no conclusions can be drawn from these experiments as to whether this results in an alteration of expression levels. However, genes that appear to be absent in the tested strains cannot be expressed.

In total, 1240 genes had different probe intensities in the ECOR strains compared to *E. coli* K12. From these 1240 genes, 535 were identified as being more intense and 722 were identified being less intense in any of the ECOR strains. These numbers may appear to be misleading but some genes were more intense in one strain while less intense in another. So there is a small overlap between the two categories. From the 290 and 258 probes that gave irreproducible results in the *E. coli* K12 reproducibility experiment (see Figure 4.7), almost all (494) appeared in the list of candidate genes. The reliability of these 494 genes should therefore be questioned and not be the focus of further investigations to use these genes as typing markers.

For example, one of the genes with high probe intensity in all strains was *entB* (b0595), which plays a role in enterobactin assembly which is an important mediator for iron transport in *E. coli*, and is detected in the majority of strains. Other genes include gatY (b2096) as well as its operon (b2087), giving confidence that the genes are not just different by chance, but that closely related probes appear on the list of potential informative genes. The genes for galactitol and ribitol utilisation are are mutually exclusive although their sequence is not similar. Previous reports show that a positive Gat phenotype is easily lost at the transduction of genes for ribitol utilisation (Woodward and Charles, 1983). The ECOR strains could have lost the *gat* genes but still be able to metabolise galactitol via the alternative pathway.

Genes with a low probe intensity in all ECOR strains include a large group ORF defined as phage, plasmid or transposons and hypothetical unclassified and unknown. ORFs that were well characterised in this group of genes missing from all five ECOR stains include a cluster of genes involved in lipopolysaccharide biosynthesis (b3624, b3627, b3629, b3630). The difference in hybridisation signal of these genes cannot be caused by a membrane effect as these genes are widely spread over the membrane. This region in the chromosome could therefore be interesting for further investigation to see whether these genes are replaced with other genes or are lost completely. It could also indicate a region that might be of interest for molecular serotyping. Although *E. coli* K-12 does not express LPS these genes were still identified on the chromosome. The ECOR strains might have different LPS and therefore not give a hybridisation signal for these particular probes. These probes could still be of interest as missing a gene could be just as characteristic for a strain as genes that are present.

The protocol provided by the manufacturer of the Panorama array (Sigma-Genosys) gives a method for stripping the membranes (up to ten times) so that they can be re-used, but warns that the signal might be reduced. This method decreased signal intensities of most probes by over 99% after stripping, however, some of the remaining signals were still more intense than the background measurement of the previous hybridisation. Therefore, if these stripped membranes were to be re-used immediately, erroneous conclusions might be drawn. Although great care was taken during the stripping of the membranes, and the stripped membranes were re-exposed to the storage phosphor screens before re-hybridisation, re-use may have affected the reproducibility of the results. High-throughput screening would not be achieved using this methodology because of the time-consuming characteristic of stripping and data acquisition. Reusing membrane arrays for the high-throughput testing of isolates was therefore not considered acceptable for fingerprinting analyses and it was anticipated that non-reusable glass slide microarrays would give more reliable results.

The Panorama membrane arrays were large and not easy to handle. Radio-labelling of the hybridisation target made data acquisition time consuming. Stripping the membrane was possible, but complications arose for probes with a high intensity level in relation to the background. The reproducibility of the repeated experiments on the stripped membranes was poor. The genes showing a difference in signal intensity levels were abundant, but did not exceed the number of genes observed to be different in the *E. coli* K12 duplication experiment. The Panorama arrays were therefore not considered to be suitable for the development of a fingerprinting or typing method. Glass slide arrays and fluorescent labelling of the target were therefore investigated to determine if they would be free of the problems encountered when using the Panorama membrane arrays.

## 4.3 Pan Arrays

### 4.3.1 Whole genome Pan glass slide array hybridisations

To investigate the application of the Pan whole-genome glass arrays (MWG-Biotech) for genomic typing of *E. coli*, glass slides were hybridised with fluorescently labelled probes prepared from *E. coli* K12 and five ECOR strains. These were identical to the strains used in the Panorama array experiments and are numbered ECOR1-ECOR5 in this results section. Six whole genome glass slide arrays were simultaneously hybridised with *E. coli* K12 DNA labelled with Cy3 and DNA extracted from one of the bacterial strains in the ECOR collection labelled with Cy5. Hybridisations for the ECOR1 strains and *E. coli* K12 were carried out in duplicate to test reproducibility. This replicated data are labelled ECOR1a and ECOR1b. Data were acquired by scanning the hybridised slides using an Affymetrix 428 microarray scanner (see Figure 4.13). Signal intensities of the probes on the glass slides were measured using Imagene 4.0 (Biodiscovery) microarray software. A defined grid template was positioned over the digital image and measurements were returned in a large data table including gene identities, raw data and background measurements. This process is illustrated in Figure 4.14. MS Excel software was used for the calculation and interpretation of the data. All raw data can be found on the enclosed CD-rom under Chapter 4/Raw data/Pan Arrays.

### 4.3.4 Identification of strain-unique hybridisation patterns

Similar to the analysis for the Panorama arrays the hybridisation patterns of the ECOR DNA were compared with the patterns of the reference strains *E. coli* K12. Ratios of the signal intensities from the tested strain versus the reference strain were calculated after mean normalisation of the signal intensities. On the basis of those ratio values, probes were divided into three groups similar to those described previously (see section 4.2.5). Probes with a low raw intensity value were filtered out of the data set using the same MS Excel software filtering function as was used for the Panorama Arrays (see also section 2.10.2). Venn diagrams, shown in Figure 4.17, were used to get an initial overview of the size of the three different groups.

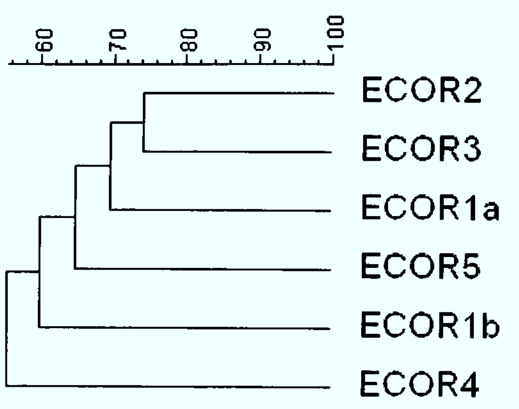

Figure 4.20 Dendogram of binary data obtained from absent genes in ECOR strains hybridised against the Pan arrays. Dendrogram was created using categorical clustering and the UPGMA algorithm.

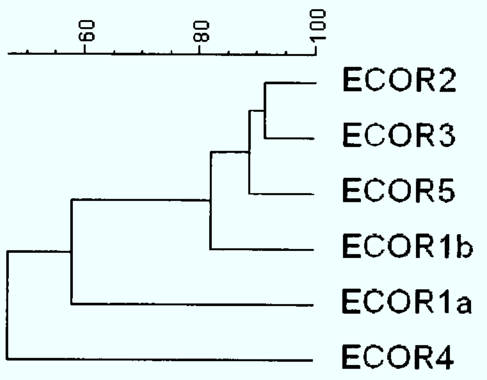

Figure 4.21 Dendogram of binary data obtained from present genes in ECOR strains hybridised against the Pan arrays Dendrogram was created using categorical clustering and the UPGMA algorithm.

### 4.3.5 Discussion whole genome glass slide microarray for typing *E. coli* strains

There has been a rapid increase in commercial availability of *E. coli* arrays which are available from Operon, Affymetrix, Clonetech and MWG. The first generation of a MWG oligonucleotide glass slide microarray for *E. coli* K12 was available shortly after the start of this thesis project. This array comprised of 70-mer oligonucleotides and was used to compare to the results and experiences obtained with the membrane based array of amplified PCR products of *E. coli* K12 ORFs. One aim was to determine if glass slide technology would overcome any of the disadvantages seen in the Panorama array experiments.

Hybridisation targets for microarrays are often labelled indirectly as the Cy dyes are large and are difficult to incorporate directly. Although labelling DNA indirectly is more time consuming, it might give a more effectively labelled DNA product (Richter *et al.*, 2002). Because random prime labelling with direct incorporation of the radioactive tracers was used in target preparation for the membrane hybridisation, a similar approach was used for Pan arrays hybridisation. While testing glass slide substrates for hybridisation with targets labelled in a direct manner with Cy dyes, hybridisation signals were easily detected using a confocal laser microscope scanner. The reproducibility of the glass slide array hybridisations was better than for the membrane arrays. The Bland-Altman plot shows that more than 96% of the probes fell within two standard deviations of the mean. One general advantage of using glass slides for arraying experiments is that competitive hybridisation with two differently labelled DNA preparations is possible. In this work, arrays were hybridised with genomic DNA from the strain of interest labelled with Cy5, and with *E. coli* K12 genomic DNA labelled with Cy3. Using the *E. coli* K12 signals as a hybridisation reference on each array gave the ratios a higher reliability, as inter array variation and variability in the hybridisation reaction were excluded.

Re-using glass slides arrays after stripping the signal from them was not possible as the slides need to be dried before data acquisition could take place, and the target binding was irreversible after it had dried onto the slide. Therefore, each experiment required a separate microarray. For reproducibility purposes, the arrays for a complete experiment were all printed in the same print run to minimise variation.

The mean normalised signal intensities were converted into ratios and classified in the three different groups previously described above for the membrane arrays (section 4.2.5). Compared to the membrane arrays, fewer genes differed for the *E. coli* K12 hybridisation signals (e.g. the blue numbers in Figure 4.17), suggesting that the variability with the glass slide arrays is less than with the membrane hybridisations. Also the means of the data distributions of the replicated experiment was only just significantly different (T-test: P = 0.05), and their variance was very different (F-test: P = 0). From the 290 and 258 probes that gave irreproducible results in the *E. coli* K12 reproducibility experiment using the Panorama arrays (see Figure 4.7) only 70 probes were also detected in duplicated Pan array hybridisations. A total of 509 genes were shown to give an irreproducible result in a duplicated Pan array experiment.

After the competitive hybridisation of Cy3 labelled *E. coli* K12 and Cy5 labelled ECOR DNA, a large number of genes had different probe signal intensities. The genes belonging to group 1 and 2 were analysed according to their function. Most of the functional groups had only few differences suggesting that these functional groups are very well conserved in *E. coli*. By far most differences were detected in the category of "hypothetical, unclassified and unknown genes", which are by far the largest category of genes on the array.

In total, 649 genes had different probe intensities in the ECOR strains. From these, 70 were identified as being more intense and 586 as less intense in any of the ECOR strains. There is a small overlap between these two categories, which explains the small discrepancy when numbers are added. The number of probes that appear less intense (586) is similar to those detected in the Panorama array experiment. This might indicate that the data for the less intense probes is more reliable than data for probes that had a higher intensity in the ECOR strains compared to the *E. coli* K12 data.

Two of the genes that appeared more than once in the list of present probes were b2001 and b3913. Both these ORFs belong to the hypothetical, unclassified and unknown genes category and the reason why these genes have a higher signal intensity remains uncertain. There were no genes that had a high intensity in all of the ECOR strains, as seen in the membrane array experiments.

Genes with a low probe intensity in all ECOR strains include a large group of genes classified as Phage, transposon or plasmid (13) and as Hypothetical, unclassified and unknown (22). Similar to the Panorama arrays, probes involved in lipopolysaccharide biosynthesis included on this array also show a low hybridisation signal (b2031 and b2033). Other genes are *perR* (b0254) a gene involved in peroxidase resistance in the stationary phase and another putative gene involved in iron-uptake (b0263).

Dendrograms from all membrane experiments were compared. A dendrogram was obtained from the hybridisation results on the membrane arrays (see Figure 4.11 & 4.12), but it does not show any similarity with the dendrogram of the ECOR strains determined by MLEE (Herzer *et al.*, 1990). The dendrogram drawn from the Pan array results shows a different relationship among the ECOR strains compared to that seen with the Panorama arrays (compare Figures 4.11 and 4.19 and Figures 4.12 and 4.20). The grouping of the five isolates as seen in clustering of the present probes identified in the Pan array array experiments show the best similarity to the dendrogram of the MLEE (shown in Figure 4.10) (Herzer *et al.*, 1990). This may indicate that the reduction of noise in the Pan arrays increases the reliability of the array results.

Compared to the membrane arrays, the glass slide arrays were relatively easy to handle. The characteristics of the fluorescent labelling mean that the target DNA had to be protected from light; hence target preparation and hybridisation were carried out under limited lighting. Each experiment was done on a single printed glass slide array so stripping complications did not arise. The hybridisation signals from the Pan array were all of a similar lower intensity, compared to the Panorama membranes. This is likely to be due to the probes being of constant length (70-mers oligonucleotides), giving a smaller standard deviation. Overall fainter signals resulted in the easier detection of high background signals, and oversaturated probes were not observed. The reproducibility was better with the glass slide arrays compared to the membrane arrays due to the co-hybridisation of *E. coli* K12 DNA as an internal control. Only a few genes had a higher signal intensity in an ECOR strain than in the *E. coli* K12 control.

Of the two objectives regarding this part of the project the first one, to determine which type of commercial whole genome array would produce the most accurate results for genotyping bacterial strains has revealed that glass slide arrays are preferred. The slides are easier to handle and give more accurate hybridisation results because of the advantage of an internal control labelled with a different Cy dye, that can be co-hybridised in every experiment. Also data acquisition is more rapid and allows high throughput testing. The second objective, to identify candidate genes for individual *E. coli* strains that could be used as sub-typing markers for DNA 'fingerprinting' has identified 1240 genes through the membrane hybridisation experiments and 649 genes through the glass slide hybridisation experiments. In the group of present genes only 15 probes were found that were identified in both Panorama and Pan array hybridisation experiments. Only 6 of those did not appear in the list of irreproducible genes found in a replicated experiment. In the group of absent genes 198 probes were found in both Panorama and Pan array hybridisation experiments. Only 98 of these did not appear in the list of irreproducible genes. These 98 and 6 probes probably contain the most valuable information for the characterisation of strains.

Further research on these genes identified using whole genome arrays would be time consuming and would involve cloning, sequencing and mapping of these candidates. Perhaps this would not be very informative because of the low reproducibility of the experiments. Before deciding whether it was worthwhile undertaking such studies, the use of arrays bearing specific genes coding for pathogenicity markers was investigated. The experiments described in this chapter have indicated how arrays might be used for fingerprinting, once several technical problems associated with reproducibility are overcome. Further increase in reproducibility through the optimisation of DNA labelling, adjustment of the hybridisation conditions, and consequent improvement of the signal to noise ratio would be advantageous. This would allow potentially interesting genes to be identified more precisely. On the positive side, these experiments gave insight into array technology and data analysis, and allowed efficient design and use of the custom made pathogenicity marker array described in the next chapter.

## 5.1 Introduction

This chapter describes the preparation, validation and use of a custom made pathogenicity marker array containing a small number of *E. coli* pathogenicity sequences obtained from several *E. coli* pathotypes. The pathogenicity sequences on the array included adhesin, capsule, toxin, invasion and iron acquisition genes, some of which were specific for certain *E. coli* pathotypes. It was expected that the genes on the array would make it possible to distinguish between different pathotypes and so allow characterisation of the *E. coli* strains, as illustrated in Figure 5.1. Fluorescently labelled genomic DNA from *E. coli* strains of known pathotype were hybridised against the custom-made array to determine the presence of genes associated with pathogenicity after hybridisation. Subsequently, all the *E. coli* isolates in the ECOR collection were screened for the presence or absence of these pathogenicity sequences, to investigate the distribution of pathogenicity markers within the collection of isolates. The array results were displayed as a dendrogram of *E. coli* pathotypes to identify groups of strains with similar characteristics. Finally, a group of clinical *E. coli* isolates was tested on the same array to examine the clonality and classify which pathotype.

sequences and the *E. coli* sequences on the array. Pre-purification of the target sequence was necessary since previous experiments had shown that amplification directly from the plasmid preparation resulted in the generation of non-specific hybridisation signals (Kuhnert *et al.*, 1997). The enzyme-restricted fragments were purified by gel electrophoresis to remove PCR inhibitors (e.g. restriction enzymes and high salt buffer) and to isolate the PCR template. The pathogenicity marker sequences were amplified using gene specific primers, and a small amount of the PCR product was electrophorised to verify the size of the product (Fig 5.4).

Initially, all fragments were amplified at the same $T_A$ (58°C), but this did not give equal amounts of PCR product. For example, the yield of PCR product amplified from the *aafA* sequence, in Figure 5.4 lane 6 was much lower than that of the product amplified from the *eltIA* sequence in lane 4. To increase the yield of the weaker PCR products, amplifications were repeated with an adjusted $T_A$ as listed in Appendix II. The new $T_A$ values were re-calculated on the basis of the sequence of the primers, using a world wide web based tool for the calculation of the properties of the oligonucleotide (Kibbe *et al.*, 2000). The sequences that required re-amplification were then grouped by similar $T_A$ to reduce the number of PCR reactions that were needed. The amount of PCR product was measured as described in section 2.3.1. One and a half µg of purified DNA was dried and redissolved in 20 µl 50 % DMSO. The PCR amplicons were stored frozen in a 384-well plate and used for printing microarray glass slides with the MicrogridII arrayer.

## 5.3 Validation of the pathogenicity marker array

### 5.3.1 Hybridisation of genomic DNA from pathotype reference strains labelled with Cy dye

The labelled genomic DNA prepared from eight *E. coli* strains were used as hybridisation targets to validate the first generation pathogenicity marker array. These eight strains included five reference strains (UPEC, EPEC, ETEC, EHEC, EAggEC; see table 2.1) used for the amplification of some of the pathogenicity sequences, and two clinical isolates previously tested for the presence of the same pathogenicity markers by dot blotting. *E. coli* K12 DNA was used as a negative control. Genomic DNA of these strains was labelled by random amplification incorporation of fluorescent Cy labelled dUTPs as described in section 2.5.4. The arrays were incubated overnight in individual hybridisation chambers with two DNA targets labelled with different fluorescent dyes. The test strains were labelled with Cy5 and the *E. coli* K12 DNA was labelled with Cy3. To investigate reproducibility, three probes were prepared from DNA isolated from the UPEC strain and three identical hybridisations were done using this DNA. (Data from the three separate hybridisations can be recognised in the data tables as UPEC_1, UPEC_2 and UPEC_3). Wash buffer containing SSC and SDS was used for post-hybridisation washes as described in section 2.9.2. One array was used per hybridisation for each of the test strains.

Arrays were scanned using a confocal microscope laser array scanner. The resulting images were imported into Arrayvision software and signal intensities were measured using a grid overlay that compensated for background signals. Further data processing was done using MS Excel software. All raw data can be found on the enclosed CD-rom under Chapter 5/Raw data pathogenicity marker array/Validation. The signal intensities of the probes were corrected for background, and were mean normalised (see also section 1.3.4). The normalisation datasheet is given in Table 5.2, and can also be found on the enclosed CD-rom under Chapter 5/Normalised data pathogenicity marker array/Validation.

the array scan. This is most likely an effect of the high signals from the positive genes in the EPEC reference strain. All but one of the positive pathogenicity markers detected in this strain had larger PCR fragments representing the gene, which may have caused variability in signal. This affected the normalisation factor and therefore the normalised data. The image of the actual array can in these cases confirm the presence of the gene. All pathogenicity sequences that were identified as positive in any of the strains are listed in Table 5.3. Pathogenicity markers unique to one of the pathotypes are indicated in red, while those positive in more than one of the reference strains are indicated in black.

markers. Also, previous membrane studies showed that a wider variety of strains did give a hybridisation signal for all gene products (Kuhnert *et al.*, 1997). All the genes expected to be present in the reference strains were detected and no additional ones were found. It was therefore concluded that genes without showing a positive signal in the validation of this array would hybridise if they were present in the test strains.

Data were stored in the program Bionumerics, and a dendrogram calculated using categorical clustering with the UPGMA algorithm. Results of the clustering of the different pathotypes are displayed in Figure 5.9

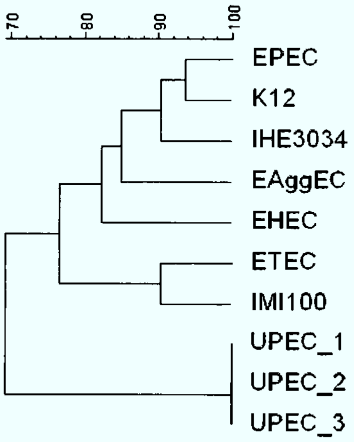

Figure 5.9 Dendogram of binary data obtained from *E. coli* reference strains hybridised against the pathogenicity marker array.
Dendrogram was created using categorical clustering methods and the UPGMA algorithm.

The hybridisation results obtained using the pathogenicity marker array were compared to the hybridisation results previously obtained using membrane hybridisations, which included a similar set of genes (Kuhnert *et al.*, 1997). Genes were more easily detected on the glass slides because of a lower signal to noise ratio.

### 5.3.5 Discussion of the validation of the pathogenicity marker array

The *E. coli* pathogenicity marker array was made with PCR amplified probes, which were spotted onto aminosilane coated glass slides. Only gel purified plasmid DNA inserts were used for the amplification of the pathogenicity marker sequences to prevent cross contamination with *E. coli* K12 vector sequences. This is a complex method for obtaining the fragments, and in theory it would have been quicker to amplify the pathogenicity markers directly from genomic DNA using the gene specific primers, as was done when the sequences were originally cloned. The advantage of having strains that carry the pathogenicity markers in cloned plasmids was that they were non-pathogenic themselves and can be precessed and transported safely. Also, amplification from the isolated plasmid DNA decreased the chance of non-specific amplification.

The array was validated using labelled genomic DNA from seven strains with known pathogenic marker profiles, as confirmed by membrane hybridisation or PCR amplification (Kuhnert *et al.*, 1997). Probes covering the same gene sequence were present on the pathogenicity marker array and the original membrane arrays. This made validation of the array easier as positive signals previously seen by membrane hybridisation were expected to appear on the glass slide microarrays. False positives were not detected after hybridisation of the labelled genomic DNA from these reference strains and less than 1% were false negative results. The results of the two assays were therefore in good agreement.

Fluorescent dyes were incorporated into the target DNA by random amplification. Together with the use of glass slides, this improved the sensitivity of the method by giving a lower background compared to that observed in membrane hybridisation experiments. For example, five *E. coli* strains were tested on membranes as well as glass slides. Some of the strains gave a high background signal on the membrane system, making observation of the positive probes difficult. In contrast, all of the strains were relatively easy to analyse using the glass slide system (e.g. see Figure 5.5). There was a clear separation between the signal intensity of present and absent genes using the glass slides. Additionally, the time spent processing samples was significantly less than for experiments involving membrane arrays, mainly due to the shorter post-hybridisation washes and antibody incubation times. The confocal microscope scanner was sensitive enough to detect the emission signal directly, and there was therefore no need for the amplification of the signal.

Multiplex PCR has previously been used for the detection of *E. coli* pathogenicity markers (Pass *et al.*, 2000, Call *et al.*, 2001, Chizhikov *et al.*, 2001, Wang *et al.*, 2002b). In these studies the multiplex PCR products were analysed on gels or hybridised to microarrays. The number of genes that could be detected in one multiplex PCR was limited. Also sequence variation in the primer region could lead to false negative results for some of the pathogenicity markers. In contrast, random amplification of genomic DNA, as described here, was a more rapid way of creating a broad-range hybridisation target. The results (see Figure 5.5) showed that the use of genomic DNA and rapid random amplification did not affect the sensitivity of the method, and that individual genes were still detectable. The positive signals were strong and there was no background hybridisation when DNA from control strains was tested.

## 5.4 Investigation of the distribution of pathogenicity gene sequences in the ECOR collection

To investigate the distribution of pathogenicity markers within the *E. coli* strains, labelled genomic DNA from all 72 strains in the ECOR collection was hybridised to the custom made pathogenicity marker array. *E. coli* K12 DNA was used as a negative control. One array was used per hybridisation of every test strain, and all hybridisations were performed in duplicate.

### 5.4.1 Hybridisation of genomic DNA from ECOR strains labelled with Cy dyes

To investigate the presence of pathogenicity markers in the ECOR strains, fluorescently labelled genomic DNA was hybridised against the first generation pathogenicity marker array. Data were analysed further with MS Excel software using a macro for normalisation and calling of the presence or absence of each pathogenicity marker. All raw data can be found on the enclosed CD-rom under Chapter 5/Raw data pathogenicity marker array/ECOR collection.

### 5.4.3 Discussion of the investigation of the distribution of pathogenicity markers in ECOR strains

Strains carrying pathogenicity markers have the potential to cause disease, but *E. coli* isolated from healthy individuals may carry a variety of pathogenicity markers. This was evident form the microarray analysis of the ECOR strains (see table 5.4), and this result confirms previous findings of genes involved in pathogenesis in *E. coli* from water and stool samples of healthy individuals (Mühldorfer *et al.*, 1996).

The ECOR collection was representative of different strains of *E. coli* from different sources (Ochman and Selander, 1984). Certain pathogenicity markers (e.g. *chuA*, F1C gene, *papA*, *sfaA*, *sfaS*) appear very frequently within this collection. They are located in pathogenicity 'islands' and can often be acquired by horizontal gene transfer (Roy, 1999). It is likely that these genes have spread throughout the *E. coli* species. Work presented by Vieira and colleagues describes the detection of pathogenicity markers in isolates with pathotypes other than expected (Vieira *et al.*, 2001). They report in their study the presence of DNA sequences related to pathogenicity in EPEC, EHEC and other pathogenic categories in a collection of 59 non-EPEC serogroups. These 59 *E. coli* strains carried the *eae* gene, but lacked other EPEC and EHEC related sequences such as EAF or the EHEC related SLT gene probes and were therefore defined non-EPEC strains. There was a high rate of LEE associated and *hly* sequences (associated with EPEC and EHEC strains), while other putative pathogenicity associated sequences were detected at a lower level. Their findings on the combination of pathogenicity markers showed strains that were, for example, potential UPEC strains that carried LEE associated sequences not normally associated with UPEC strains. The work presented in this thesis confirms the results that pathogenicity markers are widely distributed in *E. coli*. Thus, the separation of pathotypes might become less distinct as horizontal transfer continues to occur.

The binary results of the presence or absence of pathogenicity markers in the ECOR collection were analysed using the categorical clustering algorithm in Bionumerics. The result did not show some similarity to previous clustering of these 72 *E. coli* strains in their phenetic groups as determined by Herzer and colleagues (1990), but no distinct groups were defined.

In the dendrogram shown in Figure 5.10 a few clusters were observed that included one of the reference strains. For example, one large group showed an identical hybridisation patterns to the *E. coli* K12 strain in which no pathogenicity markers were detected (Figure 5.10, purple cluster). Two of the ECOR strains in this group were originally from Swedish female patients suffering from symptomatic UTI. It would have been expected that these isolates would have had a hybridisation pattern comparable to that of the UPEC reference strain used in this study. It is most likely that the isolates from these patients did not cause the infection, but represent a commensal *E. coli* isolate. As no additional clinical information was available, it was not possible to draw any definitive conclusions. There did not seem to be any relationship between the isolates of this cluster regarding their group, serotype, host or place of isolation. This suggests that horizontal transfer of pathogenicity associated sequences might not be localised or specific to certain phenetic groups.

The second largest group of ECOR strains clustered closely around the UPEC reference isolates (Figure 5.10 dark blue cluster). This group mainly contained isolates from *E. coli* group B2, and was a mixture of strains isolated from both healthy poeple and UTI patients. It has been reported that isolates belonging to groups B and C lacked pathogenic associated sequences (Kuhnert *et al.*, 1997). In contrast, a large group of extraintestinal strains belonging to phylogenetic group B2 and D contained most of the genes encoding adhesion fimbriae, toxins and iron acquisition mechanisms, for which they were tested (Bingen-Bidois *et al.*, 2002). Using the pathogenicity marker array most of the strains that clustered around the UPEC reference strain were from the B2 group. No other obvious groups were detected that were closely related to any particular pathotype. This suggests that there was no relationship between the phenetic groups and the pathogenicity marker groups. The information from the pathogenicity marker array may be inadequate for detecting phenetic relationships, as the number of genes on the array is limited. To investigate further whether the pathogenicity marker array was an appropriate tool for the identification of these relationships, a group of clinical isolates was tested.

## *5.5 Pathogenicity marker screening of clinical E. coli isolates from urinary tract infections.*

To investigate the distribution of pathogenicity marker genes and to identify clonal groups in clinical *E. coli* isolates obtained from patients with urinary tract infections, a collection of strains was screened using the pathogenicity marker array. This collection was obtained from the Shrewsbury Public Health Laboratory in a collaborative study with Professor R. E. Warren's group. The isolates had been

### 5.5.2 Hybridisation of genomic DNA from clinical isolates labelled with Cy label

To investigate the distribution of pathogenicity markers and the clonal grouping of these isolates, 49 isolates from the Shrewsbury collection were initially used for array hybridisation. Data were imported into MS Excel software for further analysis. Raw data can be found on the enclosed CD-rom under Chapter 5/Raw data pathogenicity marker array/UTI *E. coli.*

### 5.5.3 Investigation of distribution of pathogenicity markers in the ECOR collection

The MS Excel software macro described in section 5.4.2. was used to identify the pathogenicity markers present in the Shrewsbury collection of isolates. The normalised data of all the tested *E. coli* strains from patients with urinary tract infections can be found on the enclosed CD-rom under Chapter 5/Normalised data pathogenicity marker array/UTI *E. coli.* Some typical results are listed in Table 5.6, showing the positive markers as black entries for the strains in which they were detected. The complete list for all 49 strains can be found on the enclosed CD-rom under Chapter 5/Pathogenicity markers in UTI *E. coli* and includes all pathogenicity markers included on the first generation pathogenicity marker array.

### 5.5.4 Discussion of the investigation of the distribution of pathogenicity markers in clinical *E. coli* isolated from patients with UTI infections

Strains of *E. coli* from the Shrewsbury collection were investigated using the first generation pathogenicity marker array containing 29 pathogenicity markers. These strains were previously characterised using API strips (Fallon *et al.*, 2002). *E. coli* is the most frequently isolated uropathogen, and multiple antimicrobial resistance patterns within these isolates are an increasing problem (Threlfall *et al.*, 2000, Fallon *et al.*, 2002, Farrell *et al.*, 2003). The antimicrobial resistance patterns of strains isolated from UTI patients were determined using a disk diffusion test (see section 5.5.1). The outcome of these tests compared very well to the antimicrobial resistance patterns found in *E. coli* isolated from UTIs (Farrell *et al.*, 2003). It could therefore be concluded that this collection represented *E. coli* UTI isolates as seen elsewhere. Furthermore, the resistance data gave yet another method of placing these strains into groups with similar characteristics using the Bionumerics clustering function. Finally, the antimicrobial resistance patterns could be used as markers on the array (e.g. trimethoprim, ampicillin and cephalexin) and tested along side other pathogenicity markers for a more detailed characterisation of *E. coli.*

The most common single antimicrobial resistance in members of the Shrewsbury collection was ampicillin resistance, 49% of strains exhibited resistance to this antimicrobial (see section 4.5.1). Multiple resistance patterns, seen in 14% of the strains, most commonly included a resistance against either ampicillin, trimethoprim and/or augmentin. As well as the testing of the antimicrobial resistance patterns, the pathogenicity markers of these isolates were identified by the hybridisation of Cy-labelled genomic DNA against the pathogenicity marker array.

When results from the pathogenicity testing of the first 49 clinical isolates were compared to previous dendrograms from API profiles or antimicrobial resistance patterns, obvious similarities were not detected (see Figure 5.11). This could have arisen because there were only 29 pathogenicity markers on the array. The grouping of these 49 isolates on the basis of just the pathogenicity markers on the first generation pathogenicity marker array might not be similar to any previously seen patterns. Eleven of these 29 pathogenicity markers were not detected in any of the 49 clinical isolates. Three out of the 18 probes detected in the isolates were positive controls, leaving only 15 genes as markers for detecting relationships among the strains and pathotypes. To address this problem a substantial addition of markers onto the array is necessary. A group of seven isolates showed high similarity to the J96 UPEC reference strain, but none of the other isolates tested were very closely related to any of the pathotypes. It was expected that more isolates from UTI patients would cluster with the UPEC reference strain because of the source of the isolates. Analysis of the clinical isolates did not reveal all the pathogenicity markers normally associated with the UPEC pathotype. It could be that the isolates investigated were not the isolates causing the UTI or that a certain combination of pathogenicity factors is essential to produce the clinical features of UTI. For example, over 50% of the strains carry the *chuA* and *iucC* markers, both these genes are associated with iron metabolism. No obvious patterns or groupings were apparent in this group of 49 strains. Comparison with results obtained through API profiles and resistance testing did not show any similarity to grouping of the strains using the array data. It was therefore not considered worthwhile investigating any of the other 324 strains. Instead, a set of clonal strains with an identical API pattern from a potential outbreak of UTIs were investigated, to determine whether an extended array could distinguish them on the basis of a wider range of pathogenicity markers.

## 6.1 Introduction

After the successful use of the first generation array to detect pathogenicity markers in *E. coli*, as described in the previous chapter, other genes were considered as candidates to be included on the pathogenicity marker array. In collaboration with the HPA Laboratory of Enteric Pathogens a second generation array was prepared, including potential pathogenicity markers for the EAggEC pathotype identified from literature and additionally from sequence information that became available during the whole genome sequencing of EAggEC. Fifty-one sequences with high sequence similarity to other well known pathogenicity markers as well as a recently identified PAI in EAggEC were included.

Although the typical EAggEC phenotype is the adhesion to HEp-2 cells in an aggregative "stacked brick" pattern, atypical EAggEC fail to do so. Thus the heterogeneous nature of this pathotype makes identification difficult, and no single PCR target has been found for the identification of both typical and atypical EAggEC. The initial aim of this study was to use microarray technology to determine whether any common targets could be identified. The second generation array was also used to investigate ten multiple resistant UPEC isolates. These strains were isolated from individuals infected during an outbreak of urinary tract infection and these isolates could not be distinguished by PFGE. The Shrewsbury Public Health Laboratory had confirmed that the API profiles of these strains were identical.

## *6.2 Construction of the second generation pathogenicity marker array*

### 6.2.1 Identification of pathogenicity associated sequences

The first generation array was extended with 51 genes for which sequence information had been revealed during the EAggEC sequencing project (personal communications Dr. E. Dudley, Baltimore). These genes included genes that were well characterised as well as others with unknown function, selected on the basis of their sequence similarity with genes involved in pathogenicity of the same and closely-related species. The selected genes were either chromosomal or plasmid-encoded. Many were associated with type III secretion systems, which are present in EPEC and EHEC strains and may also be present in EAggEC. The additional genes added to the first generation array are listed in Table 6.1. The second generation array included a total of 75 putative pathogenicity factors in *E. coli*, as well as 3 positive control genes: 16S rDNA, *fhuA*, and *fimA*; and 3 negative controls: a synthetic oligonucleotide, spotting solution and water.

### 6.2.2 Amplification of pathogenicity associated sequences

When the first generation array was expanded, the direct approach of amplifying genes from genomic DNA using gene-specific primers was adopted in order to save time. The pathogenicity markers listed in Table 6.1 were amplified directly from the genomic DNA of two well-described strains of EAggEC, 042 (O44:H18) and 17-2 (O3: H2), using gene specific primers. The primers, amplification conditions and bacterial isolates for the amplification of genes previously described are referenced in Table 6.1. Amplification primer sequences for the genes not described elsewhere are shown in Appendix III. The amplified PCR products were purified using QIAquick PCR columns (see section 2.5.4) and quantified by spectrophotometry. The arrays were printed in a similar way as described for the first generation pathogenicity marker array.

## *6.3 Validation of the second generation pathogenicity marker array*

### 6.3.1 Hybridisation of genomic DNA from EAggEC strains labelled with Cy dye

The two strains from which the potential pathogenicity markers of EAggEC were amplified were used for the validation of the array. Although not all pathogenicity markers from these strains were known, the markers added to the first generation array could be validated and tested for a positive signal. *E. coli* K12 DNA was used as a negative control. Two non-EAggEC strains (EHEC O157:H7 and EPEC O127:H7) were also used as controls to detect non-EAggEC specific hybridisation signals. Genomic DNA from all these strains was labelled by random amplification incorporation of fluorescent Cy labelled dUTPs, as described in section 2.6.4. One array was used per hybridisation for each of the tested strains. Arrays were scanned and signal intensities were measured using Imagene 4.0 as previously described. All raw data can be found on the enclosed CD-rom under Chapter 6/Raw data second generation array/Validation. The normalised data can be found on the enclosed CD-rom under Chapter 6/Normalised data second generation array/Validation. The interpretation of the data as to whether a gene was present in the tested strain was adjusted, for two reasons. Firstly, the overall signal intensity had increased as the second generation pathogenicity marker array had additional EAggEC potential pathogenicity markers. Secondly, Imagene measures signal intensity levels differently from Arrayvision leading to lower raw data values. In an identical approach to that described in Chapter 5 the threshold was calculated at 0.15 through the analysis of strains with a known hybridisation pattern. The distribution of positive and negative signals was distinct. The threshold of 0.15 was also estimated in consultation with a statistician (CDSC Statistic Unit) to optimise sensitivity and specificity. The binary data of present and absent genes for the validation experiment is listed in Table 6.2. This table shows the positive markers detected in the control strains as black entries. Gene probes on the array that did not hybridise to any of the strains investigated have been excluded from the table but are listed in the footnotes.

and type III fimbriae sequences of which were amplified from this control strain. EAggEC 17-2 DNA also hybridised to most of the putative EAggEC genes and to two of the iron acquisition genes. Classical strains of EPEC, such as E2348/69, harbour the *bfp* and *eae* genes and DNA from the EPEC control strain hybridised with these two probes, as well as the gene designated 54f03-02 encoding a flagella biosynthesis protein, *chu*A and a gene of unknown function designated 13-1. DNA from the EHEC control strain hybridised with *sltI*, *hly*A and the gene encoding EAST, but failed to hybridise with *sltII*. Like the EPEC strain EHEC DNA also hybridised to the gene of unknown function, 13-1 and the flagella biosynthesis associated gene 54f03-2, indicating that these genes are not EAggEC-specific, but were also distributed among other *E. coli* pathotypes. Some of the chromosomal genes identified in EAggEC DNA sequence as putative marker genes, showed a sequence homology to EHEC genes associated with the type III secretion system. These genes were also detected after EHEC O157:H7 DNA hybridisation. After hybridisation of DNA from other EAggEC strains specific markers for the characterisation of typical as well as atypical EAggEC were not identified. This second generation array was then printed for standard use for investigation of other isolates and proved to be very useful.

## *6.4 Pathogenicity marker screening of clinical* E. coli *isolates from an outbreak of urinary tract infections in a residential home*

### 6.4.1 Hybridisation of genomic DNA from clonal UTI strains labelled with Cy dye

Ten UPEC isolates, with multiple antimicrobial resistance, from individuals infected during an outbreak of urinary tract infections were obtained from the Shrewsbury Public Health Laboratory, and were characterised by hybridisation against the second generation pathogenicity marker array. These strains were not distinguishable by PFGE and had identical API profiles. Genomic DNA was isolated using the automated extraction procedure described in section 2.2.2. Extracted DNA was fluorescently labelled by the incorporation of Cy5 labelled dUTP by random amplification as described in section 2.5.4. The labelled DNA was hybridised overnight at 42°C against the second generation pathogenicity marker arrays. Arrays were scanned after hybridisation and signal intensities were measured using Imagene software. Normalised signal intensities were interpreted using MS Excel software. Raw data can be found on the enclosed CD-rom under Chapter 6/Raw data second generation array/UTI outbreak and normalised data can be found under Chapter 6/Normalised data second generation array/UTI outbreak. The binary data of present and absent genes for the validation experiment is listed in Table 6.3. This table shows the positive markers detected in the UTI outbreak isolates as black entries. Gene probes on the array that did not hybridise to any of these strains have been excluded from the table but are listed in the footnotes.

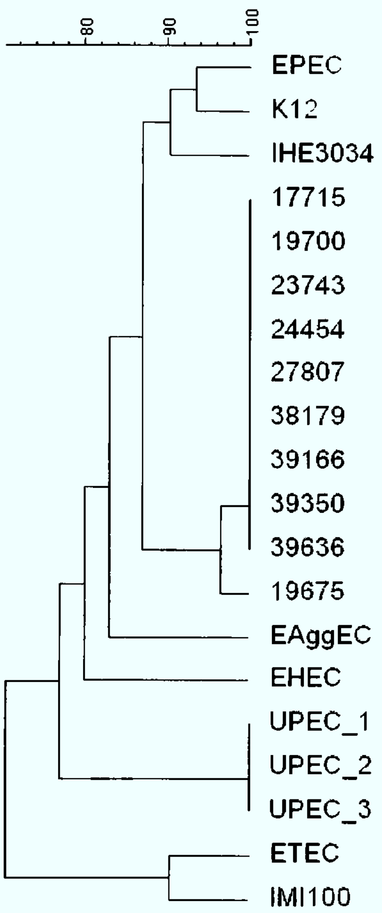

Figure 6.1 Dendrogram of binary data obtained from UTI isolates and *E. coli* reference strains hybridised against first generation pathogenicity marker array.

Dendrogram was created using categorical clustering methods and the UPGMA algorithm.

In Figure 6.1, all but one UTI isolate (19675) showed an identical hybridisation pattern. Distinguishing the strains from each other was therefore not possible. After including all data obtained in the hybridisation experiment using the second generation array, the UTI isolates showed clear differences (see Figure 6.2).

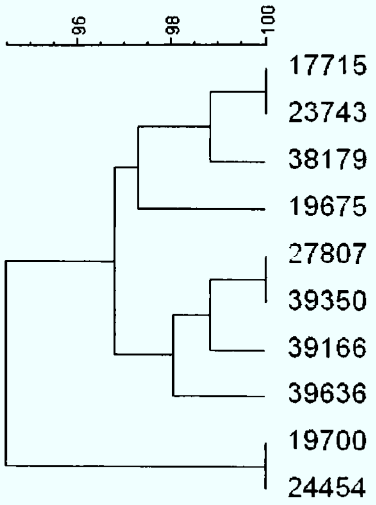

Figure 6.2 Dendrogram of binary data obtained from UTI isolates hybridised against second generation pathogenicity marker array.

Dendrogram was created using categorical clustering methods and the UPGMA algorithm.

### 6.4.3 Discussion

In the validation of the second generation pathogenicity marker array, all the probes gave the expected pattern after hybridisation of 042 EAggEC strain DNA, with the exception of the 18-3 probe, which gave a false-negative result. This result suggests that the probe could have been wrongly amplified, or that the probe DNA was lost in the process of spotting the arrays. The hybridisation pattern of the strain tested for type I and type II fimbriae, 17-2, was not as expected. Type I and type II fimbriae were amplified from this strain so therefore should have given a positive hybridisation signal after hybridisation with 17-2 DNA. Both fimbrial genes were detected but the molecules that ensure that the fimbriae arrive at the correct location in the cell (usher) were not present. Also one of the type II fimbrial genes gave a positive signal. A possible explanation was the high sequence similarity of these probes and the resulting cross-hybridisation. Furthermore, in the hybridisation of the clonal UTI DNA, only the type II fimbrial genes gave a positive signal and type I or III fimbrial genes were not detected. The DNA from the *E. coli* K12 reference strain only hybridised with the positive controls, as previously seen in the first generation array. DNA from the other two pathotypes (EPEC and EHEC) used as hybridisation targets, identified various markers associated with pathogenicity. Eight of the added putative EAggEC pathogenicity markers (i.e. *13-1, 174d11-1, 174d11-2, 54f03-2, 6-1, 7-1, 7-2* and *90g04-1*) gave a positive hybridisation signal. Six of these probes (i.e. *13-1, 174d11-1, 174d11-2, 6-1, 7-1* and *7-2*) were expected as there was a high sequence similarity with EHEC genes, as shown in Table 6.1. The other two positive probes (i.e. *54f03-2* and *90g04-1)* indicate that that these probes are not specific for EAggEC strains and therefore were not suitable as markers for the detection of just typical and atypical EAggEC.

To investigate whether this array was an appropriate tool to distinguish between closely related isolates from UTI patients, DNA from ten clinical isolates from an outbreak of UTI were investigated. With the exception of one strain (i.e. 19675) that had an additional capsule antigen gene, the 29 pathogenicity markers present on the first generation array failed to distinguish between the ten isolates. All strains clustered together and did not show any obvious similarity with any of the reference strains. This supports previous findings that relationships between the ECOR and other clinical isolates on the basis of the first generation array were difficult to detect. When the additional genes from the second generation array were taken into account, the strains could be distinguished on the basis of the presence and absence of various pathogenicity markers.

The differences between the hybridisation signals of the ten UTI outbreak strains on the second generation arrays were in genes encoding invasion, capsule and fimbriae prepared from the EAggEC prototype 042 genomic DNA. These probes added unto the second generation pathogenicity marker array were selected to identify markers for the detection of EAggEC strains. To see these genes in UTI strains was therefore unexpected, but suggests that they were not pathotype-specific for EAggEC. This confirms previous results and supports other findings that pathogenicity markers may be horizontally transferred between isolates and not associated with just one pathotype (Vieira *et al.*, 2001). It also demonstrates that the inclusion of more genes on the array increases the likelihood of distinguishing strains not separable on the basis of genes included on the first generation array. The separation of the isolates into different groups on the basis of the larger second generation array may therefore be useful, although it would be advantageous to include more putative markers on the array.

Rapid screening systems using many genes on one array may aid or possibly even replace other routine diagnostic tests such as, for example serotyping (Bekal *et al.*, 2003). Gene probes from many pathotypes can be included on one array, which makes this technology very powerful. Array hybridisation revealed not only the presence or absence of a target gene, as in PCR or single colony hybridisation, but also gave an indication of the pathotype. This could speed up the characterisation process. Pathotype determination on the basis of just the array results needs to be assessed carefully, as not all markers associated with that pathotype will be detected in all hybridisations.

# 7. General discussion

Microarray technology has developed rapidly over the past decade (Schena *et al.*, 1995, Schena, 2003a, Mantripragada *et al.*, 2004). The literature describes various applications, from gene expression (Arfin *et al.*, 2000) and drug development (reviewed in (Debouck and Goodfellow, 1999)) to genotyping (Anthony *et al.*, 2000, Wang *et al.*, 2002a) and comparative genomic hybridisation (Dorrell *et al.*, 2001, Anjum *et al.*, 2003). The advantage of array technology in comparison with standard molecular techniques is that thousands of genes can be tested in a single experiment using a sample volume of just several microlitres. The results obtained from a single microarray experiment can be far more informative than the results from the best multiplex or real-time PCR reactions currently available. For example, depending on the array, information may be obtained on genetic defects and patient profiles. Array technology also has the potential to be used in hereditary screening and drug treatment of patients (Schena, 2003b).

The availability of whole genome sequences has lead to a rapid expansion in the production of chips and arrays, mostly for expression and genotyping experiments (Gingeras *et al.*, 1998, Alizadeh *et al.*, 2000, Call *et al.*, 2001, Detweiler *et al.*, 2001). Some arrays comprise whole genomes, whilst others carry a subset of genes related to specific diseases (Firoved and Deretic, 2003), to cell processes (Nakamura, 2004) or to groups of species and subspecies (Wang *et al.*, 2002a).

The main hypothesis of this work was to investigate whether a subset of known markers or ORFs could be identified and spotted onto an array in order to assess whether or not a strain of *E. coli* carried these genes and whether this information may then be used in characterising the strain for epidemiological purposes. A general hypothesis was to determine whether DNA arrays could also be used for typing organisms without the need for whole genome sequencing. The first step was therefore to use the whole genome arrays to determine whether: I) they could themselves be used for typing and II) whether specific markers could be identified to add to a smaller selection of pathogenicity markers for arraying onto glass slides. The second step was to produce and test such a smaller, more specific array, using various culture collections.

The two whole genome arrays in this study were commercially available and were used for the determining the type of array most suitable for genotyping. The differences between the two arrays were their probe composition (oligonucleotides or PCR-amplified DNA probes) and their printing substrate (nylon membrane or coated glass slide). From a practical perspective the glass slides were practically superior and quicker to process. More importantly, the results obtained with the glass slide arrays had a narrower distribution and gave more even signal intensity levels.

There are two approaches for the use of DNA arrays in screening bacteria for typing. Firstly, isolates may be compared by the presence of similar genes (Dobrindt *et al.*, 2003). Secondly, genes or regions in the genomic DNA that are absent in one isolate but present in another may be compared (Anjum *et al.*, 2003). Thus, either the presence of the genes (i.e. similarity) or the absence of the genes (i.e. dissimilarity) can form the basis for the typing comparisons. To find appropriate candidate markers for characterisation, genomic DNA was extracted from various *E. coli* strains and was hybridised to the commercial arrays. Probes giving both lower and higher hybridisation signals were taken into account when identifying typing markers. A large group of putative markers was identified, but the results were not found to be reproducible. Statistical analysis showed that the mean and variance of these data distributions differed significantly. Due to time constraints and this lack of reproducibility further investigation into the genomic regions identified by this approach was not carried out. Instead, a small customised array was prepared to demonstrate that arrays could be used for the characterisation of isolates.

For further research into using whole genome arrays for fingerprinting, it would firstly require to investigate the reproducibility issues. It would be necessary to do multiple experiments for each array hybridisation. Also, alterations in the hybridisation conditions could affect sensitivity and specificity of the hybridisation and produce more reliable data. If a new array was to be created it should contain all the genes from sequenced *E. coli* genomes available in the GenBank database (Blattner *et al.*, 1997, Hayashi *et al.*, 2001, Perna *et al.*, 2001, Welch *et al.*, 2002, Chaudhuri and Pallen, 2004, NIH, 2004). As only genes of known sequence can be included on the array, the technology is, at present, limited due to incomplete knowledge of the genes carried by the test strains.

Currently different pathotypes of *E. coli* are identified by the disease they cause whilst supplemented by serological or molecular tools. Some of these are laborious, or have limited potential to distinguish between the different pathotypes (Orskov *et al.*, 1977, Fallon *et al.*, 2002, Jenkins *et al.*, 2002, Jothikumar and Griffiths, 2002, Osek, 2002). For example, testing to distinguish between EAggEC strains is done by investigating the attachment of the bacteria to HEp-2 cells (Nataro *et al.*, 1992). EAggEC strains show a stacked brick formation when adhering to the cells. This method gives no further indication of the pathotype when a stacked brick formation is not observed.

Molecular tools used for identification include PCR-based detection methods, a well established technique, but which will only provide limited information regarding the pathotype (Schmidt *et al.*, 1995b). It is also limited by the number of targets that can be detected in one amplification reaction, and by the detection of genes with sequence variation. In contrast to the whole genome arrays, small arrays containing a subset of genes of interest, including chromosomal and plasmid genes from different strains have also been designed. The genes were either identified from whole genome hybridisation experiments, (Dobrindt *et al.*, 2003) or from well characterised genes described in the literature (Kuhnert *et al.*, 1997, Anthony *et al.*, 2000). Using this approach, this work revealed a wide distribution of pathogenicity markers in the ECOR strains and clinical isolates. Genes that appeared in more than 20% of the ECOR strains were *chuA* (42%), *iucC* (24%), *papA* (31%) and, as expected the positive control signals. In addition clinical isolates contained pathogenicity markers *chuA*, *cnfI*, F1C gene, *iucC*, *neuA/neuC*, *papA*, *hlyA*, *sfaS*, with a frequency of 20% or more.

Nine clinical isolates from patients with UTI did group with the UPEC reference pathotype in the dendrogram, and could therefore be classified as being of this pathotype. Because of the source of the isolates, more of the clinical isolates were expected to cluster with the UPEC reference strain. None of the strains clustered with any other pathotype. This genomic screening for genes is potentially of great value in the diagnostic testing of bacterial isolates. In one array hybridisation patterns of many genes can be examined. Current developments make it possible to print the arrays in single tubes or in a 96-well plate format, which may therefore make the technology more accessible to diagnostic laboratories (Perrin *et al.*, 2003). The array developed in the presented work will have to be challenged with strains from a wider source to cover all pathotypes within the test strains, but the initial experiments have demonstrated that indication of the pathotype can be achieved with the array.

After extension of the first generation pathogenicity array with markers for EAggEC, this array was challenged with clonal UTI isolates. The results demonstrated that these strains could be differentiated. There was an increased level of separation compared with the first generation pathogenicity marker array, with which these isolates gave identical hybridisation patterns. It would be beneficial to extend this relatively small second generation array with other biomarkers from sequenced *E. coli* strains and related species, including all relevant pathogenicity markers and antimicrobial resistance genes. This would make pathotype investigations more precise and also provide additional insights on strain characterisation and horizontal transfer of genes, probably even across species borders. As well as many promising clinical applications, arrays allow pathogenicity testing of the 'non pathogenic' *E. coli* strains that are used in biotechnological applications, and would allow testing for food safety purposes (Bekal *et al.*, 2003, Kuhnert *et al.*, 2000).

Although the presence of genes is very useful information for the identification of the pathotype, information about the expression of these genes is lacking. It would therefore be interesting to examine mRNA expression patterns of the pathogenicity markers on the array. The information about expression patterns would be enhanced with the inclusion of other probes on the array, with sequences of upstream elements of the pathogenicity markers, such as operon and leader sequences. These would act as internal controls and provide further information about transcription. Such an array is currently being manufactured and validated for the investigation of closely related *Enterobacteriaceae* species (personal communication with Dr. M. Anjum, VLA). The array will also carry genes used for biochemical speciation (API typing system, see section 1.1.2) and antimicrobial resistance markers. Such an array will help provide a clearer understanding of pathogenesis and host-pathogen interactions.

Microarray technology has already contributed enormously to the current knowledge of genomics. Whole genome arrays, and arrays with specific subsets of genes are now produced and validated for many organisms. The arrays described in this thesis were constructed, used and validated and thereby gave insight into the potential use of array technology in both research and diagnostic laboratories. The arrays and methods developed have already been shared with collaborators in other laboratories (Dr. Steve Green, HPA Southampton and Dr. Claire Jenkins, HPA Colindale). From this and related work, there is no doubt that bacterial diagnostic and typing arrays will come to be used widely for diagnosis, surveillance, reference and research.

**Abe, A., Komase, K., Bangtrakulnonth, A., Ratchtrachenchat, O. A., Kawahara, K. and Danbara, H.** (1990) Trivalent heat-labile- and heat-stable-enterotoxin probe conjugated with horseradish peroxidase for detection of enterotoxigenic *Escherichia coli* by hybridization. *J Clin Microbiol,* **28:** 2616-20.

**Adiri, R. S., Gophna, U. and Ron, E. Z.** (2003) Multilocus sequence typing (MLST) of *Escherichia coli* O78 strains. *FEMS Microbiol Lett,* **222:** 199-203.

**Alizadeh, A. A., Eisen, M. B., Davis, R. E., Ma, C., Lossos, I. S., Rosenwald, A., Boldrick, J. C., Sabet, H., Tran, T., Yu, X., Powell, J. I., Yang, L., Marti, G. E., Moore, T., Hudson, J., Jr., Lu, L., Lewis, D. B., Tibshirani, R., Sherlock, G., Chan, W. C., Greiner, T. C., Weisenburger, D. D., Armitage, J. O., Warnke, R., Levy, R., Wilson, W., Grever, M. R., Byrd, J. C., Botstein, D., Brown, P. O. and Staudt, L. M.** (2000) Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature,* **403:** 503-11.

**Amersham Biosciences** (2004), Online technical support. [Online], Available from <http://www4.amershambiosciences.com/aptrix/upp01077.nsf/Content/Products?OpenDocument&parentid=74833&moduleid=74980&zone=Radiochemicals> [Accessed 03/07/2004]

**Amon, P. and Ivanov, I.** (2003) Genomic DNA labeling for hybridization with DNA arrays. *Biotechniques,* **34:** 700-2, 704.

**Andrews, J. M.** (2001) BSAC standardized disc susceptibility testing method. *J Antimicrob Chemother,* **48 Suppl 1:** 43-57.

**Anjum, M. F., Lucchini, S., Thompson, A., Hinton, J. C. and Woodward, M. J.** (2003) Comparative genomic indexing reveals the phylogenomics of *Escherichia coli* pathogens. *Infect Immun,* **71:** 4674-83.

**Anthony, R. M., Brown, T. J. and French, G. L.** (2000) Rapid diagnosis of bacteremia by universal amplification of 23S ribosomal DNA followed by hybridization to an oligonucleotide array. *J Clin Microbiol,* **38:** 781-8.

**Arfin, S. M., Long, A. D., Ito, E. T., Tolleri, L., Riehle, M. M., Paegle, E. S. and Hatfield, G. W.** (2000) Global gene expression profiling in *Escherichia coli* K12. The effects of integration host factor. *J Biol Chem,* **275:** 29672-84.

**Arnold, C., Metherell, L., Willshaw, G., Maggs, A. and Stanley, J.** (1999) Predictive fluorescent amplified-fragment length polymorphism analysis of *Escherichia coli*: high-resolution typing method with phylogenetic significance. *J Clin Microbiol,* **37:** 1274-9.

**Bebora, L. C.** (1997) Role of plasmids in the virulence of enteric bacteria. *East Afr Med J,* **74:** 444-6.

**Behr, M. A., Wilson, M. A., Gill, W. P., Salamon, H., Schoolnik, G. K., Rane, S. and Small, P. M.** (1999) Comparative genomics of BCG vaccines by whole-genome DNA microarray. *Science,* **284:** 1520-3.

**Bekal, S., Brousseau, R., Masson, L., Prefontaine, G., Fairbrother, J. and Harel, J.** (2003) Rapid identification of *Escherichia coli* pathotypes by virulence gene detection with DNA microarrays. *J Clin Microbiol,* **41:** 2113-25.

**Bernier, C., Gounon, P. and Le Bouguenec, C.** (2002) Identification of an aggregative adhesion fimbria (AAF) type III-encoding operon in enteroaggregative *Escherichia coli* as a sensitive probe for detecting the AAF-encoding operon family. *Infect Immun,* **70:** 4302-11.

**Bertucci, F., Bernard, K., Loriod, B., Chang, Y. C., Granjeaud, S., Birnbaum, D., Nguyen, C., Peck, K. and Jordan, B. R.** (1999) Sensitivity issues in DNA array-based expression measurements and performance of nylon microarrays for small samples. *Hum Mol Genet,* **8:** 1715-22.

**Bhan, M. K., Raj, P., Levine, M. M., Kaper, J. B., Bhandari, N., Srivastava, R., Kumar, R. and Sazawal, S.** (1989) Enteroaggregative *Escherichia coli* associated with persistent diarrhea in a cohort of rural children in India. *J Infect Dis,* **159:** 1061-4.

**Bingen-Bidois, M., Clermont, O., Bonacorsi, S., Terki, M., Brahimi, N., Loukil, C., Barraud, D. and Bingen, E.** (2002) Phylogenetic analysis and prevalence of urosepsis strains of *Escherichia coli* bearing pathogenicity island-like domains. *Infect Immun,* **70:** 3216-26.

**Bland, J. M. and Altman, D. G.** (1986) Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet,* **1:** 307-10.

**Bland, J. M. and Altman, D. G.** (1999) Measuring agreement in method comparison studies. *Stat Methods Med Res,* **8:** 135-60.

**Blattner, F. R., Plunkett, G., 3rd, Bloch, C. A., Perna, N. T., Burland, V., Riley, M., Collado-Vides, J., Glasner, J. D., Rode, C. K., Mayhew, G. F., Gregor, J., Davis, N. W., Kirkpatrick, H. A., Goeden, M. A., Rose, D. J., Mau, B. and Shao, Y.** (1997) The complete genome sequence of *Escherichia coli* K-12. *Science,* **277:** 1453-74.

**Bou, G., Cervero, G., Dominguez, M. A., Quereda, C. and Martinez-Beltran, J.** (2000) PCR-based DNA fingerprinting (REP-PCR, AP-PCR) and pulsed-field gel electrophoresis characterization of a nosocomial outbreak caused by imipenem- and meropenem-resistant *Acinetobacter baumannii. Clin Microbiol Infect,* **6:** 635-43.

**Boyd, E. F. and Hartl, D. L.** (1998) Chromosomal regions specific to pathogenic isolates of *Escherichia coli* have a phylogenetically clustered distribution. *J Bacteriol,* **180:** 1159-65.

**Boylan, M., Smyth, C. J. and Scott, J. R.** (1988) Nucleotide sequence of the gene encoding the major subunit of CS3 fimbriae of enterotoxigenic *Escherichia coli. Infect Immun,* **56:** 3297-300.

**Brazma, A., Hingamp, P., Quackenbush, J., Sherlock, G., Spellman, P., Stoeckert, C., Aach, J., Ansorge, W., Ball, C. A., Causton, H. C., Gaasterland, T., Glenisson, P., Holstege, F. C., Kim, I. F., Markowitz, V., Matese, J. C., Parkinson, H., Robinson, A., Sarkans, U., Schulze-Kremer, S., Stewart, J., Taylor, R., Vilo, J. and Vingron, M.** (2001) Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. *Nat Genet,* **29:** 365-71.

**Brokx, S. J., Ellison, M., Locke, T., Bottorff, D., Frost, L. and Weiner, J. H.** (2004) Genome-wide analysis of lipoprotein expression in *Escherichia coli* MG1655. *J Bacteriol,* **186:** 3254-8.

**Brooks, J. T., Bergmire-Sweat, D., Kennedy, M., Hendricks, K., Garcia, M., Marengo, L., Wells, J., Ying, M., Bibb, W., Griffin, P. M., Hoekstra, R. M. and Friedman, C. R.** (2004) Outbreak of Shiga toxin-producing *Escherichia coli* O111:H8 infections among attendees of a high school cheerleading camp. *Clin Infect Dis,* **38:** 190-8.

**Brunton, J., Hinde, D., Langston, C., Gross, R., Rowe, B. and Gurwith, M.** (1980) Enterotoxigenic *Escherichia coli* in central Canada. *J Clin Microbiol,* **11:** 343-8.

**Bullen, J. J., Ward, C. G. and Rogers, H. J.** (1991) The critical role of iron in some clinical infections. *Eur J Clin Microbiol Infect Dis,* **10:** 613-7.

**Burgess, M. N., Bywater, R. J., Cowley, C. M., Mullan, N. A. and Newsome, P. M.** (1978) Biological evaluation of a methanol-soluble, heat-stable *Escherichia coli* enterotoxin in infant mice, pigs, rabbits, and calves. *Infect Immun,* **21:** 526-31.

**Call, D. R., Brockman, F. J. and Chandler, D. P.** (2001) Detecting and genotyping *Escherichia coli* O157:H7 using multiplexed PCR and nucleic acid microarrays. *Int J Food Microbiol,* **67:** 71-80.

**Calmette, A. and Guerin, C.** (1920) Nouvelles recherches expérimentales sur la vaccination des bovidés contre la tuberculose. *Ann Inst Pasteur Microbiol,* **34:** 553-7.

**Caprioli, A., Falbo, V., Roda, L. G., Ruggeri, F. M. and Zona, C.** (1983) Partial purification and characterization of an *Escherichia coli* toxic factor that induces morphological cell alterations. *Infect Immun,* **39:** 1300-6.

**Carbon P, Ehresmann, C., Ehresmann, B. and Ebel, J. P.** (1979) The complete nucleotide sequence of the ribosomal 16-S RNA from *Escherichia coli.* Experimental details and cistron heterogeneities. *Eur J Biochem,* **100:** 399-410.

**Chan, M. and Aanensen, D.** (2003), Multi Locus Sequence Typing [Online], Available from <www.mlst.net> [Accessed 28/02/2005]

**Chart, H.** (1998) Toxigenic *Escherichia coli. Symp Ser Soc Appl Microbiol,* **27:** 77S-86S.

**Chaudhuri, R. and Pallen, M.** (2004) *coli*Base [Online], Available from <http://www.colibase.bham.ac.uk> [Accessed 05/05/2005]

**Cheung, V. G., Morley, M., Aguilar, F., Massimi, A., Kucherlapati, R. and Childs, G.** (1999) Making and reading microarrays. *Nat Genet,* **21:** 15-9.

**Chiu, S. K., Hsu, M., Ku, W. C., Tu, C. Y., Tseng, Y. T., Lau, W. K., Yan, R. Y., Ma, J. T. and Tzeng, C. M.** (2003) Synergistic effects of epoxy- and amine-silanes on microarray DNA immobilization and hybridization. *Biochem J,* **374:** 625-32.

**Chizhikov, V., Rasooly, A., Chumakov, K. and Levy, D. D.** (2001) Microarray analysis of microbial virulence factors. *Appl Environ Microbiol,* **67:** 3258-63.

**Churchill, G. A.** (2002) Fundamentals of experimental design for cDNA microarrays. *Nat Genet,* **32 Suppl:** 490-5.

**Cieslewicz, M. and Vimr, E.** (1996) Thermoregulation of kpsF, the first region 1 gene in the kps locus for polysialic acid biosynthesis in *Escherichia coli* K1. *J Bacteriol,* **178:** 3212-20.

**Clements, J. D. and Finkelstein, R. A.** (1979) Isolation and characterization of homogeneous heat-labile enterotoxins with high specific activity from *Escherichia coli* cultures. *Infect Immun,* **24:** 760-9.

**Clements, J. D., Yancey, R. J. and Finkelstein, R. A.** (1980) Properties of homogeneous heat-labile enterotoxin from *Escherichia coli. Infect Immun,* **29:** 91-7.

**Clermont, O., Cordevant, C., Bonacorsi, S., Marecat, A., Lange, M. and Bingen, E.** (2001) Automated ribotyping provides rapid phylogenetic subgroup affiliation of clinical extraintestinal pathogenic *Escherichia coli* strains. *J Clin Microbiol,* **39:** 4549-53.

**Cleveland, W. S.** (1979) Robust locally weighted regression and smoothing scatterplots. *J Amer Stat Assoc,* **74:** 829-39.

**Conner, D. E. and Kotrola, J. S.** (1995) Growth and survival of *Escherichia coli* O157:H7 under acidic conditions. *Appl Environ Microbiol,* **61:** 382-5.

**Cortese, J. D.** (2000) Array of options; instrumentation to exploit the DNA microarray explosion. *The Scientist,* **14:** 26.

**Coulton, J. W., Mason, P., Cameron, D. R., Carmel, G., Jean, R. and Rode, H. N.** (1986) Protein fusions of beta-galactosidase to the ferrichrome-iron receptor of *Escherichia coli* K-12. *J Bacteriol,* **165:** 181-92.

**Cravioto, A., Gross, R. J., Scotland, S. M. and Rowe, B.** (1979) An adhesive factor found in strains of *Escherichia coli* belonging to the traditional infantile enteropathogenic serotypes. *Curr Micro,* **3:** 95-9.

**Cronin, M. T., Fucini, R. V., Kim, S. M., Masino, R. S., Wespi, R. M. and Miyada, C. G.** (1996) Cystic fibrosis mutation detection by hybridization to light-generated DNA probe arrays. *Hum Mutat,* **7:** 244-55.

**Crosa, J. H.** (1989) Genetics and molecular biology of siderophore-mediated iron transport in bacteria. *Microbiol Rev,* **53:** 517-30.

**Cummings, J. H. and Macfarlane, G. T.** (1997) Role of intestinal bacteria in nutrient metabolism. *J Parenter Enteral Nutr,* **21:** 357-65.

**Czeczulin, J. R., Balepur, S., Hicks, S., Phillips, A., Hall, R., Kothary, M. H., Navarro-Garcia, F. and Nataro, J. P.** (1997) Aggregative adherence fimbria II, a second fimbrial antigen mediating aggregative adherence in enteroaggregative *Escherichia coli. Infect Immun,* **65:** 4135-45.

**Dai, L. and Zimmerly, S.** (2002) The dispersal of five group II introns among natural populations of *Escherichia coli. RNA,* **8:** 1294-307.

**Danese, P. N., Pratt, L. A. and Kolter, R.** (2000) Exopolysaccharides production is required for the development of *Escherichia coli* K-12 Biofilm Architecture. *J Bacteriol,* **182:** 3593-6.

**Darwin, K. H. and Miller, V. L.** (1999) Molecular basis of the interaction of *Salmonella* with the intestinal mucosa. *Clin Microbiol Rev,* **12:** 405-28.

**Debouck, C. and Goodfellow, P. N.** (1999) DNA microarrays in drug discovery and development. *Nat Genet,* **21:** 48-50.

**Desai, M., Tanna, A., Wall, R., Efstratiou, A., George, R. and Stanley, J.** (1998) Fluorescent amplified-fragment length polymorphism analysis of an outbreak of group A streptococcal invasive disease. *J Clin Microbiol,* **36:** 3133-7.

**Detweiler, C. S., Cunanan, D. B. and Falkow, S.** (2001) Host microarray analysis reveals a role for the *Salmonella* response regulator phoP in human macrophage cell death. *Proc Natl Acad Sci U S A,* **98:** 5850-5.

**Dobrindt, U., Agerer, F., Michaelis, K., Janka, A., Buchrieser, C., Samuelson, M., Svanborg, C., Gottschalk, G., Karch, H. and Hacker, J.** (2003) Analysis of genome plasticity in pathogenic and commensal *Escherichia coli* isolates by use of DNA arrays. *J Bacteriol,* **185:** 1831-40.

**Donnenberg, M. S., Giron, J. A., Nataro, J. P. and Kaper, J. B.** (1992) A plasmid-encoded type IV fimbrial gene of enteropathogenic *Escherichia coli* associated with localized adherence. *Mol Microbiol,* **6:** 3427-37.

**Donnenberg, M. S. and Welch, R. A.** (1996) Virulence determinants of uropathogenic *Escherichia coli,* in: Molbley, H. L. T. and Warren, J. W. (eds) *Urinary tract infections. Molecular pathogenesis and clinical management* ASM Press, Washington D.C., pp. 135-74.

**Dorrell, N., Mangan, J. A., Laing, K. G., Hinds, J., Linton, D., Al-Ghusein, H., Barrell, B. G., Parkhill, J., Stoker, N. G., Karlyshev, A. V., Butcher, P. D. and Wren, B. W.** (2001) Whole genome comparison of *Campylobacter jejuni* human isolates using a low-cost microarray reveals extensive genetic diversity. *Genome Res,* **11:** 1706-15.

**Dougan, G. and Morrissey, P.** (1985) Molecular analysis of the virulence determinants of enterotoxigenic *Escherichia coli* isolated from domestic animals: applications for vaccine development. *Vet Microbiol,* **10:** 241-57.

**Dozois, C. M. and Curtiss, R., 3rd** (1999) Pathogenic diversity of *Escherichia coli* and the emergence of 'exotic' islands in the gene stream. *Vet Res,* **30:** 157-79.

**Drancourt, M., Bollet, C., Carlioz, A., Martelin, R., Gayral, J. P. and Raoult, D.** (2000) 16S ribosomal DNA sequence analysis of a large collection of environmental and clinical unidentifiable bacterial isolates. *J Clin Microbiol,* **38:** 3623-30.

**DuPont, H. L., Formal, S. B., Hornick, R. B., Snyder, M. J., Libonati, J. P., Sheahan, D. G., LaBrec, E. H. and Kalas, J. P.** (1971) Pathogenesis of *Escherichia coli* diarrhea. *N Engl J Med,* **285:** 1-9.

**Eisen, J. A.** (2001) Gastrogenomics. *Nature,* **409:** 463, 465-6.

**Elias, W. P., Uber, A. P., Tomita, S. K., Trabulsi, L. R. and Gomes, T. A.** (2002) Combinations of putative virulence markers in typical and variant enteroaggregative *Escherichia coli* strains from children with and without diarrhoea. *Epidemiol Infect,* **129:** 49-55.

**Endo, Y., Tsurugi, K., Yutsudo, T., Takeda, Y., Ogasawara, T. and Igarashi, K.** (1988) Site of action of a Vero toxin (VT2) from *Escherichia coli* O157:H7 and of Shiga toxin on eukaryotic ribosomes. RNA N-glycosidase activity of the toxins. *Eur J Biochem,* **171:** 45-50.

**Escherich, T.** (1885) Die Darmbakterien des neugeboren und sauglings. *Fortschritte der Medizin,* **3:** 515-522, 547-554.

**Escobar-Paramo, P., Giudicelli, C., Parsot, C. and Denamur, E.** (2003) The evolutionary history of *Shigella* and enteroinvasive *Escherichia coli* revised. *J Mol Evol,* **57:** 140-8.

**Ewing, W. H.** (1953) Serological relationships between shigella and coliform cultures. *J Bacteriol,* **66:** 333-40.

**Falbo, V., Pace, T., Picci, L., Picci, E. and Capriolo, A.** (1993) Isolation and nucleic acid sequence of the gene encoding cytotoxic necrotizing factor 1 of *Escherichia coli. Infect Immun,* **61:** 4909-14.

**Fallon, D., Andrews, N., Frodsham, D., Gee, B., Howe, S., Iliffe, A., Nye, K. J. and Warren, R. E.** (2002) A comparison of the performance of cystine lactose electrolyte deficient (CLED) agar with Oxoid chromogenic urinary tract infection

(CUTI) medium for the isolation and presumptive identification of organisms from urine. *J Clin Pathol,* **55:** 524-9.

**Farrell, D. J., Morrissey, I., De Rubeis, D., Robbins, M. and Felmingham, D.** (2003) A UK multicentre study of the antimicrobial susceptibility of bacterial pathogens causing urinary tract infection. *J Infect,* **46:** 94-100.

**Finke, A., Bronner, D., Nikolaev, A. V., Jann, B. and Jann, K.** (1991) Biosynthesis of the *Escherichia coli* K5 polysaccharide, a representative of group II capsular polysaccharides: polymerization in vitro and characterization of the product. *J Bacteriol,* **173:** 4088-94.

**Firoved, A. M. and Deretic, V.** (2003) Microarray analysis of global gene expression in mucoid *Pseudomonas aeruginosa. J Bacteriol,* **185:** 1071-81.

**Frydendahl, K., Imberechts, H. and Lehmann, S.** (2001) Automated 5' nuclease assay for detection of virulence factors in porcine *Escherichia coli. Mol Cell Probes,* **15:** 151-60.

**Fukushima, M., Kakinuma, K. and Kawaguchi, R.** (2002) Phylogenetic analysis of *Salmonella*, *Shigella*, and *Escherichia coli* strains on the basis of the *gyrB* gene sequence. *J Clin Microbiol,* **40:** 2779-85.

**Gaastra, W. and Svennerholm, A. M.** (1996) Colonization factors of human enterotoxigenic *Escherichia coli* (ETEC). *Trends Microbiol,* **4:** 444-52.

**Giannella, R.** (1976) Suckling mouse model for detection of heat-stable *Escherichia coli* enterotoxin: characteristics of the model. *Infect Immun,* **14:** 95-9.

**Gilligan, P. H.** (1999) *Escherichia coli.* EAEC, EHEC, EIEC, ETEC. *Clin Lab Med,* **19:** 505-21.

**Gingeras, T. R., Ghandour, G., Wang, E., Berno, A., Small, P. M., Drobniewski, F., Alland, D., Desmond, E., Holodniy, M. and Drenkow, J.** (1998) Simultaneous genotyping and species identification using hybridization pattern recognition analysis of generic *Mycobacterium* DNA arrays. *Genome Res,* **8:** 435-48.

**Glode, M. P., Sutton, A., Robbins, J. B., McCracken, G. H., Gotschlich, E. C., Kaijser, B. and Hanson, L. A.** (1977) Neonatal meningitis due of *Escherichia coli* K1. *J Infect Dis,* **136 Suppl:** S93-7.

**Gomez-Duarte, O. G., Ruiz-Tagle, A., Gomez, D. C., Viboud, G. I., Jarvis, K. G., Kaper, J. B. and Giron, J. A.** (1999) Identification of *lngA*, the structural gene of longus type IV pilus of enterotoxigenic *Escherichia coli. Microbiology,* **145:** 1809-16.

**Gorbach, S. L., Banwell, J. G., Chatterjee, B. D., Jacobs, B. and Sack, R. B.** (1971) Acute undifferentiated human diarrhea in the tropics. I. Alterations in intestinal micrflora. *J Clin Invest,* **50:** 881-9.

**Gordillo, M. E., Reeve, G. R., Pappas, J., Mathewson, J. J., DuPont, H. L. and Murray, B. E.** (1992) Molecular characterization of strains of enteroinvasive *Escherichia coli* O143, including isolates from a large outbreak in Houston, Texas. *J Clin Microbiol,* **30:** 889-93.

**Grana, M. and Acerenza, L.** (2001) A model combining cell physiology and population genetics to explain *Escherichia coli* laboratory evolution. *BMC Evol Biol,* **1:** 12.

**Grimont, F. and Grimont, P. A.** (1986) Ribosomal ribonucleic acid gene restriction patterns as potential taxonomic tools. *Ann Inst Pasteur Microbiol,* **137B:** 165-75.

**Gross, R. and Rowe, B.** (1985) Serotyping of *Escherichia coli*, in: Sussman, M. (ed) *The virulence of Escherichia coli,* Cambridge University Press, Cambridge, pp. 345-60.

**Haas, S. A., Hild, M., Wright, A. P., Hain, T., Talibi, D. and Vingron, M.** (2003) Genome-scale design of PCR primers and long oligomers for DNA microarrays. *Nucleic Acids Res,* **31:** 5576-81.

**Hacker, J., Blum-Oehler, G., Hochhut, B. and Dobrindt, U.** (2003) The molecular basis of infectious diseases: pathogenicity islands and other mobile genetic elements. A review. *Acta Microbiol Immunol Hung,* **50:** 321-30.

**Hagberg, L., Jodal, U., Korhonen, T. K., Lidin-Janson, G., Lindberg, U. and Svanborg Eden, C.** (1981) Adhesion, hemagglutination, and virulence of *Escherichia coli* causing urinary tract infections. *Infect Immun,* **31:** 564-70.

**Harel, J. and Martin, C.** (1999) Virulence gene regulation in pathogenic *Escherichia coli. Vet Res,* **30:** 131-55.

**Hartman, A. B., Venkatesan, M., Oaks, E. V. and Buysse, J. M.** (1990) Sequence and molecular characterization of a multicopy invasion plasmid antigen gene, *ipaH,* of *Shigella flexneri. J Bacteriol,* **172:** 1905-15.

**Hayashi, T., Makino, K., Ohnishi, M., Kurokawa, K., Ishii, K., Yokoyama, K., Han, C. G., Ohtsubo, E., Nakayama, K., Murata, T., Tanaka, M., Tobe, T., Iida, T., Takami, H., Honda, T., Sasakawa, C., Ogasawara, N., Yasunaga, T., Kuhara, S., Shiba, T., Hattori, M. and Shinagawa, H.** (2001) Complete genome sequence of enterohemorrhagic *Escherichia coli* O157:H7 and genomic comparison with a laboratory strain K-12. *DNA Res,* **8:** 11-22.

**Health Protection Agency** (2003a), Topics of infection A-Z [Online], Available from <www.hpa.org.uk/ionfection/topics_az/bacteraemia/gram_neg.htm> [Accessed 28/02/2005]

**Health Protection Agency** (2003b), Topics of infection A-Z [Online], Available from < www.hpa.org.uk/ionfection/topics_az/ecoli/O157/menu.htm> [Accessed 28/02/2005]

**Hegde, P., Qi, R., Abernathy, K., Gay, C., Dharap, S., Gaspard, R., Hughes, J. E., Snesrud, E., Lee, N. and Quackenbush, J.** (2000) A concise guide to cDNA microarray analysis. *Biotechniques,* **29:** 548-50, 552-4, 556.

**Herzer, P. J., Inouye, S., Inouye, M. and Whittam, T. S.** (1990) Phylogenetic distribution of branched RNA-linked multicopy single-stranded DNA among natural isolates of *Escherichia coli. J Bacteriol,* **172:** 6175-81.

**Hess, J., Wels, W., Vogel, M. and Goebel, W.** (1986) Nucleotide sequence of a plasmid-encoded hemolysin determinant and its comparison with a corresponding chromosomal hemolysin sequence. *FEMS Microbiol Lett,* **34:** 1-11.

**Holland, J. L., Louie, L., Simor, A. E. and Louie, M.** (2000) PCR detection of *Escherichia coli* O157:H7 directly from stools: evaluation of commercial extraction methods for purifying fecal DNA. *J Clin Microbiol,* **38:** 4108-13.

**Holloway, A. J., van Laar, R. K., Tothill, R. W. and Bowtell, D. D.** (2002) Options available--from start to finish--for obtaining data from DNA microarrays II. *Nat Genet,* **32 Suppl:** 481-9.

**Holmgren, J., Fredman, P., Lindblad, M., Svennerholm, A. M. and Svennerholm, L.** (1982) Rabbit intestinal glycoprotein receptor for *Escherichia coli* heat-labile enterotoxin lacking affinity for cholera toxin. *Infect Immun,* **38:** 424-33.

**Honda, T., Akhtar, Q., Glass, R. I. and Kibriya, A. K.** (1981a) A simple assay to detect *Escherichia coli* producing heat labile enterotoxin: results of a field study of the Biken tests in Bangladesh. *Lancet,* **2:** 609-10.

**Honda, T., Arita, M., Takeda, Y. and Miwatani, T.** (1982) Further evaluation of the Biken test (modified Elek test) for detection of enterotoxigenic *Escherichia coli* producing heat-labile enterotoxin and application of the test to sampling of heat-stable enterotoxin. *J Clin Microbiol,* **16:** 60-2.

**Honda, T., Taga, S., Takeda, Y. and Miwatani, T.** (1981b) Modified Elek test for detection of heat-labile enterotoxin of enterotoxigenic *Escherichia coli. J Clin Microbiol,* **13:** 1-5.

**Hopkins, K. L. and Hilton, A. C.** (2001) Use of multiple primers in RAPD analysis of clonal organisms provides limited improvement in discrimination. *Biotechniques,* **30:** 1262-4, 1266-7.

**Hull, R. A., Gill, R. E., Hsu, P., Minshew, B. and Falkow, S.** (1981) Construction and expression of recombinant plasmids encoding type I or D-mannose-resistant pili from a urinary tract infection *Escherichia coli* isolate. *Infect Immun,* **33:** 933-8.

**Ideker, T., Ybarra, S. and Grimmond, S.** (2003) Hybridization and posthybridization washing, in: Bowtell, D. D. and Sambrook, J. (eds) *DNA Microarrays; a molecular cloning manual,* Cold Spring Harbor Laboratory Press, New York, pp. 228-88.

**Jenkins, C., Chart, H., Cheasty, T., Willshaw, G. A., Pearce, M. C., Foster, G., Gunn, G. J., Smith, H. R., Dougan, G., Synge, B. A. and Frankel, G.** (2002) Verocytotoxin-producing *Escherichia coli* (VTEC) other than serogroup O157 from Scottish cattle. *Vet Rec,* **151:** 58-60.

**Jerse, A. E., Yu, J., Tall, B. D. and Kaper, J. B.** (1990) A genetic locus of enteropathogenic *Escherichia coli* necessary for the production of attaching and effacing lesions on tissue culture cells. *Proc Natl Acad Sci U S A,* **87:** 7839-43.

**Jiang, Z. D., Lowe, B., Verenkar, M. P., Ashley, D., Steffen, R., Tornieporth, N., von Sonnenburg, F., Waiyaki, P. and DuPont, H. L.** (2002) Prevalence of enteric pathogens among international travelers with diarrhea acquired in Kenya (Mombasa), India (Goa), or Jamaica (Montego Bay). *J Infect Dis,* **185:** 497-502.

**Jin, Q., Yuan, Z., Xu, J., Wang, Y., Shen, Y., Lu, W., Wang, J., Liu, H., Yang, J., Yang, F., Zhang, X., Zhang, J., Yang, G., Wu, H., Qu, D., Dong, J., Sun, L., Xue, Y., Zhao, A., Gao, Y., Zhu, J., Kan, B., Ding, K., Chen, S., Cheng, H., Yao, Z., He, B., Chen, R., Ma, D., Qiang, B., Wen, Y., Hou, Y. and Yu, J.** (2002) Genome sequence of *Shigella flexneri* 2a: insights into pathogenicity through comparison with genomes of *Escherichia coli* K12 and O157. *Nucleic Acids Res,* **30:** 4432-41.

**Jinneman, K. C., Yoshitomi, K. J. and Weagant, S. D.** (2003) Multiplex real-time PCR method to identify Shiga toxin genes stx1 and stx2 and *Escherichia coli* O157:H7/H- serotype. *Appl Environ Microbiol,* **69:** 6327-33.

**Johnson, J. R.** (2000) Development of polymerase chain reaction-based assays for bacterial gene detection. *J Microbiol Methods,* **41:** 201-9.

**Johnson, J. R., Delavari, P., Kuskowski, M. and Stell, A. L.** (2001) Phylogenetic distribution of extraintestinal virulence-associated traits in *Escherichia coli. J Infect Dis,* **183:** 78-88.

**Johnson, J. R., Moseley, S. L., Roberts, P. L. and Stamm, W. E.** (1988) Aerobactin and other virulence factor genes among strains of *Escherichia coli* causing urosepsis: association with patient characteristics. *Infect Immun,* **56:** 405-12.

**Johnson, J. R., Murray, A. C., Gajewski, A., Sullivan, M., Snippes, P., Kuskowski, M. A. and Smith, K. E.** (2003) Isolation and molecular characterization of nalidixic acid-resistant extraintestinal pathogenic *Escherichia coli* from retail chicken products. *Antimicrob Agents Chemother,* **47:** 2161-8.

**Johnson, J. R., Oswald, E., O'Bryan, T. T., Kuskowski, M. A. and Spanjaard, L.** (2002) Phylogenetic distribution of virulence-associated genes among *Escherichia coli* isolates associated with neonatal bacterial meningitis in the Netherlands. *J Infect Dis,* **185:** 774-84.

**Johnson, J. R. and Russo, T. A.** (2002) Extraintestinal pathogenic *Escherichia coli*: "the other bad *E coli*". *J Lab Clin Med,* **139:** 155-62.

**Johnson, J. R. and Stell, A. L.** (2000) Extended virulence genotypes of *Escherichia coli* strains from patients with urosepsis in relation to phylogeny and host compromise. *J Infect Dis,* **181:** 261-72.

**Johnson, W. M. and Lior, H.** (1988) A new heat-labile cytolethal distending toxin (CLDT) produced by *Escherichia coli* isolates from clinical material. *Microb Pathog,* **4:** 103-13.

**Jothikumar, N. and Griffiths, M. W.** (2002) Rapid detection of *Escherichia coli* O157:H7 with multiplex real-time PCR assays. *Appl Environ Microbiol,* **68:** 3169-71.

**Kaijser, B., Hanson, L. A., Jodal, U., Lidin-Janson, G. and Robbins, J. B.** (1977) Frequency of *E. coli* K antigens in urinary-tract infections in children. *Lancet,* **1:** 663-6.

**Kaijser, B. and Jodal, U.** (1984) *Escherichia coli* K5 antigen in relation to various infections and in healthy individuals. *J Clin Microbiol,* **19:** 264-6.

**Kaiser, R. J., MacKellar, S. L., Vinayak, R. S., Sanders, J. Z., Saavedra, R. A. and Hood, L. E.** (1989) Specific-primer-directed DNA sequencing using automated fluorescence detection. *Nucleic Acids Res,* **17:** 6087-102.

**Kane, M. D., Jatkoe, T. A., Stumpf, C. R., Lu, J., Thomas, J. D. and Madore, S. J.** (2000) Assessment of the sensitivity and specificity of oligonucleotide (50mer) microarrays. *Nucleic Acids Res,* **28:** 4552-7.

**Kantor, H. S., Tao, P. and Wisdom, C.** (1974) Action of *Escherichia coli* enterotoxin: adenylate cyclase behaviour of intestinal epithelial cells in culture. *Infect Immun,* **9:** 1003-10.

**Kaper, J. B., McDaniel, T. K., Jarvis, K. G. and Gomez-Duarte, O.** (1997) Genetics of virulence of enteropathogenic *E. coli. Adv Exp Med Biol,* **412:** 279-87.

**Karjalainen, T. M., Evans, D. G., So, M. and Lee, C. H.** (1989) Molecular cloning and nucleotide sequence of the colonization factor antigen I gene of *Escherichia coli. Infect Immun,* **57:** 1126-30.

**Kauffmann, F.** (1947) The serology of the coli group. *J Immunol,* **57:** 71-100.

**Kauffmann, F.** (1954) *Enterobacteriaceae,* Ejnar Munksgaard Publisher, Copenhagen.

**Kauffmann, F. and Dupont, A.** (1950) *Escherichia* strains from infantile epidemic gastro enteritis. *Acta Pathol Microbiol Scand,* **27:** 552-64.

**Kayser, H.** (1903) Ueber Bakterienhamolysine, im Besonderen das Colilysin. *Zeit Hyg Infekt,* **42:** 118-38.

**Khan, A., Yamasaki, S., Sato, T., Ramamurthy, T., Pal, A., Datta, S., Chowdhury, N. R., Das, S. C., Sikdar, A., Tsukamoto, T., Bhattacharya, S. K., Takeda, Y. and Nair, G. B.** (2002) Prevalence and genetic profiling of virulence determinants of non-O157 Shiga toxin-producing *Escherichia coli* isolated from cattle, beef, and humans, Calcutta, India. *Emerg Infect Dis,* **8:** 54-62.

**Kibbe, W. A., Qing Cao, M. S., Buehler, E. and Somers, B.** (2000), Oligonucleotide properties calculator [Online], Available from <www.basicnorthwestern.edu/biotools/oligocalc.html> [Accessed 16/04/2005]

**Kim, J. M., Eckmann, L., Savidge, T. C., Lowe, D. C., Witthoft, T. and Kagnoff, M. F.** (1998) Apoptosis of human intestinal epithelial cells after bacterial invasion. *J Clin Invest,* **102:** 1815-23.

**Klemm, P.** (1984) The fimA gene encoding the type-1 fimbrial subunit of *Escherichia coli.* Nucleotide sequence and primary structure of the protein. *Eur J Biochem,* **143:** 395-9.

**Knutton, S., Adu-Bobie, J., Bain, C., Phillips, A. D., Dougan, G. and Frankel, G.** (1997) Down regulation of intimin expression during attaching and effacing enteropathogenic *Escherichia coli* adhesion. *Infect Immun,* **65:** 1644-52.

**Koczura, R. and Kaznowski, A.** (2003) The *Yersinia* high-pathogenicity island and iron-uptake systems in clinical isolates of *Escherichia coli. J Med Microbiol,* **52:** 637-42.

**Konowalchuk, J., Speirs, J. I. and Stavric, S.** (1977) Vero response to a cytotoxin of *Escherichia coli. Infect Immun,* **18:** 775-9.

**Kroll, T. C. and Wolfl, S.** (2002) Ranking: a closer look on globalisation methods for normalisation of gene expression arrays. *Nucleic Acids Res,* **30:** e50.

**Kroncke, K. D., Boulnois, G., Roberts, I., Bitter-Suermann, D., Golecki, J. R., Jann, B. and Jann, K.** (1990) Expression of the *Escherichia coli* K5 capsular antigen: immunoelectron microscopic and biochemical studies with recombinant *E. coli. J Bacteriol,* **172:** 1085-91.

**Kuhnert, P., Boerlin, P. and Frey, J.** (2000) Target genes for virulence assessment of *Escherichia coli* isolates from water, food and the environment. *FEMS Microbiol Rev,* **24:** 107-17.

**Kuhnert, P., Hacker, J., Muhldorfer, I., Burnens, A. P., Nicolet, J. and Frey, J.** (1997) Detection system for *Escherichia coli*-specific virulence genes: absence of virulence determinants in B and C strains. *Appl Environ Microbiol,* **63:** 703-9.

**Kuo, W. P., Jenssen, T. K., Butte, A. J., Ohno-Machado, L. and Kohane, I. S.** (2002) Analysis of matched mRNA measurements from two different microarray technologies. *Bioinformatics,* **18:** 405-12.

**Lan, C. Y., Newport, G., Murillo, L. A., Jones, T., Scherer, S., Davis, R. W. and Agabian, N.** (2002) Metabolic specialization associated with phenotypic switching in *Candida albicans. Proc Natl Acad Sci U S A,* **99:** 14907-12.

**Lawrence, J. G. and Ochman, H.** (1998) Molecular archaeology of the *Escherichia coli* genome. *Proc Natl Acad Sci U S A,* **95:** 9413-7.

**LeClerc, J. E., Li, B., Payne, W. L. and Cebula, T. A.** (1996) High mutation frequencies among *Escherichia coli* and *Salmonella* pathogens. *Science,* **274:** 1208-11.

**Lee, M. L., Kuo, F. C., Whitmore, G. A. and Sklar, J.** (2000) Importance of replication in microarray gene expression studies: statistical methods and evidence from repetitive cDNA hybridizations. *Proc Natl Acad Sci U S A,* **97:** 9834-9.

**Levine, M. M., Nataro, J. P., Karch, H., Baldini, M. M., Kaper, J. B., Black, R. E., Clements, M. L. and O'Brien, A. D.** (1985) The diarrheal response of humans to some classic serotypes of enteropathogenic *Escherichia coli* is dependent on a plasmid encoding an enteroadhesiveness factor. *J Infect Dis,* **152:** 550-9.

**Levine, M. M., Xu, J. G., Kaper, J. B., Lior, H., Prado, V., Tall, B., Nataro, J., Karch, H. and Wachsmuth, K.** (1987) A DNA probe to identify enterohemorrhagic *Escherichia coli* of O157:H7 and other serotypes that cause hemorrhagic colitis and hemolytic uremic syndrome. *J Infect Dis,* **156:** 175-82.

**Li, J., Chen, S. and Evans, D. H.** (2001) Typing and subtyping influenza virus using DNA microarrays and multiplex reverse transcriptase PCR. *J Clin Microbiol,* **39:** 696-704.

**Li, J., Pankratz, M. and Johnson, J. A.** (2002) Differential gene expression patterns revealed by oligonucleotide versus long cDNA arrays. *Toxicol Sci,* **69:** 383-90.

**Livermore, D. M., Threlfall, E. J., Reacher, M. H., Johnson, A. P., James, D., Cheasty, T., Shah, A., Warburton, F., Swan, A. V., Skinner, J., Graham, A. and Speller, D. C.** (2000) Are routine sensitivity test data suitable for the surveillance of resistance? Resistance rates amongst *Escherichia coli* from blood and CSF from 1991-1997, as assessed by routine and centralized testing. *J Antimicrob Chemother,* **45:** 205-11.

**Mabeck, C. E., Orskov, F. and Orskov, I.** (1971) *Escherichia coli* serotypes and renal involvement in urinary-tract infection. *Lancet,* **1:** 1312-4.

**Mackey, J. P. and Sandys, G. H.** (1966) Diagnosis of Urinary Infections. *BMJ*: 1173.

**Maiden, M. C., Bygraves, J. A., Feil, E., Morelli, G., Russell, J. E., Urwin, R., Zhang, Q., Zhou, J., Zurth, K., Caugant, D. A., Feavers, I. M., Achtman, M. and Spratt, B. G.** (1998) Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. *Proc Natl Acad Sci U S A,* **95:** 3140-5.

**Mantripragada, K. K., Buckley, P. G., de Stahl, T. D. and Dumanski, J. P.** (2004) Genomic microarrays in the spotlight. *Trends Genet,* **20:** 87-94.

**Marklund, B. I., Tennent, J. M., Garcia, E., Hamers, A., Baga, M., Lindberg, F., Gaastra, W. and Normark, S.** (1992) Horizontal gene transfer of the *Escherichia coli* pap and prs pili operons as a mechanism for the development of tissue-specific adhesive properties. *Mol Microbiol,* **6:** 2225-42.

**Martinez, J. L., Herrero, M. and de Lorenzo, V.** (1994) The organization of intercistronic regions of the aerobactin operon of pColV-K30 may account for the differential expression of the *iucABCD iutA* genes. *J Mol Biol,* **238:** 288-93.

**Maskos, U. and Southern, E. M.** (1993) A study of oligonucleotide reassociation using large arrays of oligonucleotides synthesised on a glass support. *Nucleic Acids Res,* **21:** 4663-9.

**Massimi, A., Harris, T., Childs, G. and Somerville, S.** (2003) Printing on glass slides, in: Bowtell, D. D. and Sambrook, J. (eds) *DNA Microarrays; a molecular cloning Manual* Cold Spring Harbor Laboratory Press, New York, pp. 71-8.

**Mattick, J. S.** (2002) Type IV pili and twitching motility. *Annu Rev Microbiol,* **56:** 289-314.

**McDaniel, T. K., Jarvis, K. G., Donnenberg, M. S. and Kaper, J. B.** (1995) A genetic locus of enterocyte effacement conserved among diverse enterobacterial pathogens. *Proc Natl Acad Sci U S A,* **92:** 1664-8.

**McFeters, G. A. and Stuart, D. G.** (1972) Survival of coliform bacteria in natural waters: field and laboratory studies with membrane-filter chambers. *Appl Microbiol,* **24:** 805-11.

**Mellies, J. L., Elliott, S. J., Sperandio, V., Donnenberg, M. S. and Kaper, J. B.** (1999) The Per regulon of enteropathogenic *Escherichia coli* : identification of a regulatory cascade and a novel transcriptional activator, the locus of enterocyte effacement (LEE)-encoded regulator (Ler). *Mol Microbiol,* **33:** 296-306.

**Micropaedia** (1974) *The new encyclopaedia brittanica,* William Benton, London.

**Miller, R. D. and Hartl, D. L.** (1986) Biotyping confirms a nearly clonal population structure in *Escherichia coli. Evolution,* **40:** 1-12.

**Moseley, S. L., Huq, I., Alim, A. R. M. A., So, M., Samadpour-Motalebi, M. and Falkow, S.** (1980) Detection of enterotoxigenic *Escherichia coli* by DNA colony hybridisation. *J Infect Dis,* **142:** 892-8.

**Mühldorfer, I., Blum, G., Donohue-Rolfe, A., Heier, H., Olschlager, T., Tschape, H., Wallner, U. and Hacker, J.** (1996) Characterization of *Escherichia coli* strains isolated from environmental water habitats and from stool samples of healthy volunteers. *Res Microbiol,* **147:** 625-35.

**Murray, A. E., Lies, D., Li, G., Nealson, K., Zhou, J. and Tiedje, J. M.** (2001) DNA/DNA hybridization to microarrays reveals gene-specific differences between closely related microbial genomes. *Proc Natl Acad Sci U S A,* **98:** 9853-8.

**Mygind, T., Ostergaard, L., Birkelund, S., Lindholt, J. S. and Christiansen, G.** (2003) Evaluation of five DNA extraction methods for purification of DNA from atherosclerotic tissue and estimation of prevalence of *Chlamydia pneumoniae* in tissue from a Danish population undergoing vascular repair. *BMC Microbiol,* **3:** 19.

**Nagy, G., Dobrindt, U., Kupfer, M., Emody, L., Karch, H. and Hacker, J.** (2001) Expression of hemin receptor molecule *chuA* is influenced by *RfaH* in uropathogenic *Escherichia coli* strain 536. *Infect Immun,* **69:** 1924-8.

**Nakamura, Y.** (2004) Isolation of p53-target genes and their functional analysis. *Cancer Sci,* **95:** 7-11.

**Nataro, J. P., Deng, Y., Maneval, D. R., German, A. L., Martin, W. C. and Levine, M. M.** (1992) Aggregative adherence fimbriae I of enteroaggregative *Escherichia coli* mediate adherence to HEp-2 cells and hemagglutination of human erythrocytes. *Infect Immun,* **60:** 2297-304.

**Nataro, J. P., Kaper, J. B., Robins-Browne, R., Prado, V., Vial, P. and Levine, M. M.** (1987a) Patterns of adherence of diarrheagenic *Escherichia coli* to HEp-2 cells. *Pediatr Infect Dis J,* **6:** 829-31.

**Nataro, J. P., Maher, K. O., Mackie, P. and Kaper, J. B.** (1987b) Characterization of plasmids encoding the adherence factor of enteropathogenic *Escherichia coli. Infect Immun,* **55:** 2370-7.

**Negre, V. L., Bonacorsi, S., Schubert, S., Bidet, P., Nassif, X. and Bingen, E.** (2004) The siderophore receptor *IroN,* but not the high-pathogenicity island or the hemin receptor *ChuA*, contributes to the bacteremic step of *Escherichia coli* neonatal meningitis. *Infect Immun,* **72:** 1216-20.

**Neter, E.** (1960) Enteropathogenic *Escherichia coli* enteritis. *Pediatr Clin North Am,* **7:** 1015-24.

**Neter, E., Westphal, O., Luderitz, O., Gino, R. M. and Gorzynski, E. A.** (1955) Demonstration of antibodies against enteropathogenic *Escherichia coli* in sera of children of various ages. *Pediatrics,* **16:** 801-8.

**NIH** (2004), GenBank [Online], Available from <http://www.ncbi.nlm.nih.gov/genomes/static/eub_p.html> and <http://www.ncbi.nlm.nih.gov/genomes/static/eub_g.html> and <http://www.ncbi.nlm.nih.gov/genomes/static/a_u.html> [Accessed 22/05/2004]

**Noller, A. C., McEllistrem, M. C., Stine, O. C., Morris, J. G., Jr., Boxrud, D. J., Dixon, B. and Harrison, L. H.** (2003) Multilocus sequence typing reveals a lack of diversity among *Escherichia coli* O157:H7 isolates that are distinct by pulsed-field gel electrophoresis. *J Clin Microbiol,* **41:** 675-9.

**Obata-Yasuoka, M., Ba-Thein, W., Tsukamoto, T., Yoshikawa, H. and Hayashi, H.** (2002) Vaginal *Escherichia coli* share common virulence factor profiles, serotypes and phylogeny with other extraintestinal *E. coli. Microbiology,* **148:** 2745-52.

**O'Brien, A. D., LaVeck, G. D., Thompson, M. R. and Formal, S. B.** (1982) Production of *Shigella dysenteriae* type 1-like cytotoxin by *Escherichia coli. J Infect Dis,* **146:** 763-9.

**O'Brien, S. J., Murdoch, P. S., Riley, A. H., King, I., Barr, M., Murdoch, S., Greig, A., Main, R., Reilly, W. J. and Thomson-Carter, F. M.** (2001) A foodborne outbreak of Vero cytotoxin-producing *Escherichia coli* O157:H-phage type 8 in hospital. *J Hosp Infect,* **49:** 167-72.

**Ochman, H. and Selander, R. K.** (1984) Standard reference strains of *Escherichia coli* from natural populations. *J Bacteriol,* **157:** 690-3.

**Oelschlaeger, T. A., Dobrindt, U. and Hacker, J.** (2002a) Pathogenicity islands of uropathogenic *E. coli* and the evolution of virulence. *Int J Antimicrob Agents,* **19:** 517-21.

**Oelschlaeger, T. A., Dobrindt, U. and Hacker, J.** (2002b) Virulence factors of uropathogens. *Curr Opin Urol,* **12:** 33-8.

**Okamoto, K. and Yamanaka, H.** (2000) Properties and actions of heat-stable enterotoxin of *Escherichia coli. J Nat Toxins,* **9:** 213-29.

**Orskov, I., Orskov, F., Jann, B. and Jann, K.** (1977) Serology, chemistry, and genetics of O and K antigens of *Escherichia coli. Bacteriol Rev,* **41:** 667-710.

**Osek, J.** (2002) Rapid and specific identification of Shiga toxin-producing *Escherichia coli* in faeces by multiplex PCR. *Lett Appl Microbiol,* **34:** 304-10.

**Pandey, M., Khan, A., Das, S. C., Sarkar, B., Kahali, S., Chakraborty, S., Chattopadhyay, S., Yamasaki, S., Takeda, Y., Nair, G. B. and Ramamurthy, T.** (2003) Association of cytolethal distending toxin locus *cdtB* with enteropathogenic *Escherichia coli* isolated from patients with acute diarrhea in Calcutta, India. *J Clin Microbiol,* **41:** 5277-81.

**Park, T., Yi, S. G., Kang, S. H., Lee, S., Lee, Y. S. and Simon, R.** (2003) Evaluation of normalization methods for microarray data. *BMC Bioinformatics,* **4:** 33.

**Pass, M. A., Odedra, R. and Batt, R. M.** (2000) Multiplex PCRs for identification of *Escherichia coli* virulence genes. *J Clin Microbiol,* **38:** 2001-4.

**Paton, A. W., Paton, J. C., Goldwater, P. N., Heuzenroeder, M. W. and Manning, P. A.** (1993) Sequence of a variant Shiga-like toxin type-I operon of *Escherichia coli* O111:H$^-$. *Gene,* **129:** 87-92.

**Pease, A. C., Solas, D., Sullivan, E. J., Cronin, M. T., Holmes, C. P. and Fodor, S. P.** (1994) Light-generated oligonucleotide arrays for rapid DNA sequence analysis. *Proc Natl Acad Sci U S A,* **91:** 5022-6.

**Penna, V. T., Martins, S. A. and Mazzola, P. G.** (2002) Identification of bacteria in drinking and purified water during the monitoring of a typical water purification system. *BMC Public Health,* **2:** 13.

**Perna, N. T., Plunkett, G., 3rd, Burland, V., Mau, B., Glasner, J. D., Rose, D. J., Mayhew, G. F., Evans, P. S., Gregor, J., Kirkpatrick, H. A., Posfai, G., Hackett, J., Klink, S., Boutin, A., Shao, Y., Miller, L., Grotbeck, E. J., Davis, N. W., Lim, A., Dimalanta, E. T., Potamousis, K. D., Apodaca, J., Anantharaman, T. S., Lin, J., Yen, G., Schwartz, D. C., Welch, R. A. and Blattner, F. R.** (2001) Genome sequence of enterohaemorrhagic *Escherichia coli* O157:H7. *Nature,* **409:** 529-33.

**Perrin, A., Duracher, D., Perret, M., Cleuziat, P. and Mandrand, B.** (2003) A combined oligonucleotide and protein microarray for the codetection of nucleic acids and antibodies associated with human immunodeficiency virus, hepatitis B virus, and hepatitis C virus infections. *Anal Biochem,* **322:** 148-55.

**Petit, C., Rigg, G. P., Pazzani, C., Smith, A., Sieberth, V., Stevens, M., Boulnois, G., Jann, K. and Roberts, I. S.** (1995) Region 2 of the *Escherichia coli* K5 capsule gene cluster encoding proteins for the biosynthesis of the K5 polysaccharide. *Mol Microbiol,* **17:** 611-20.

**Phillips, I., Eykyn, S., King, A., Gransden, W. R., Rowe, B., Frost, J. A. and Gross, R. J.** (1988) Epidemic multiresistant *Escherichia coli* infection in West Lambeth Health District. *Lancet,* **1:** 1038-41.

**Pickett, C. L., Weinstein, D. L. and Holmes, R. K.** (1987) Genetics of type Ila heat-labile enterotoxin of *Escherichia coli*: operon fusions, nucleotide sequence, and hybridization studies. *J Bacteriol,* **169:** 5180-7.

**Pollack, J. R.** (2003) Comparative genomic hybridisation using cDNA microarrays, in: Bowtell, D. D. and Sambrook, J. (eds) *DNA Microarrays; a molecular cloning Manual*, Cold Spring Harbor Laboratory Press, New York, pp. 363-9.

**Porwollik, S., Wong, R. M. and McClelland, M.** (2002) Evolutionary genomics of *Salmonella*: gene acquisitions revealed by microarray analysis. *Proc Natl Acad Sci U S A,* **99:** 8956-61.

**Quackenbush, J.** (2002) Microarray data normalization and transformation. *Nat Genet,* **32 Suppl:** 496-501.

**Quantrell, R. J., Naylor, S. W., Roe, A. J., Spears, K. and Gally, D. L.** (2004) EHEC O157:H7 getting to the bottom of the burger bug. *Micro Today,* **31:** 126-8.

**Richter, A., Schwager, C., Hentze, S., Ansorge, W., Hentze, M. W. and Muckenthaler, M.** (2002) Comparison of fluorescent tag DNA labeling methods used for expression analysis by DNA microarrays. *Biotechniques,* **33:** 620-8, 630.

**Rickman, D. S., Herbert, C. J. and Aggerbeck, L. P.** (2003) Optimizing spotting solutions for increased reproducibility of cDNA microarrays. *Nucleic Acids Res,* **31:** e109.

**Riley, L. W.** (1987) The epidemiologic, clinical and microbiologic features of hemorrhagic colitis. *Ann Rev Microbiol,* **41:** 383-407.

**Riley, L. W., Remis, R. S., Helgerson, S. D., McGee, H. B., Wells, J. G., Davis, B. R., Hebert, R. J., Olcott, E. S., Johnson, L. M., Hargrett, N. T., Blake, P. A. and Cohen, M. L.** (1983) Hemorrhagic colitis associated with a rare *Escherichia coli* serotype. *N Engl J Med,* **308:** 681-5.

**Rothbaum, R., McAdams, A. J., Giannella, R. and Partin, J. C.** (1982) A clinicopathologic study of enterocyte-adherent *Escherichia coli*: a cause of protracted diarrhea in infants. *Gastroenterology,* **83:** 441-54.

**Roy, P.** (1999) Horizontal transfer of genes in bacteria. *Micro Today,* **26:** 168-70.

**Ryan, C. A., Tauxe, R. V., Hosek, G. W., Wells, J. G., Stoesz, P. A., McFadden, H. W., Jr., Smith, P. W., Wright, G. F. and Blake, P. A.** (1986) *Escherichia coli* O157:H7 diarrhea in a nursing home: clinical, epidemiological, and pathological findings. *J Infect Dis,* **154:** 631-8.

**Sack, R. B., Gorbach, S. L., Banwell, J. G., Jacobs, B., Chatterjee, B. D. and Mitra, R. C.** (1971) Enterotoxigenic *Escherichia coli* isolated from patients with severe cholera-like disease. *J Infect Dis,* **123:** 378-85.

**Saiki, R. K., Walsh, P. S., Levenson, C. H. and Erlich, H. A.** (1989) Genetic analysis of amplified DNA with immobilized sequence-specific oligonucleotide probes. *Proc Natl Acad Sci U S A,* **86:** 6230-4.

**Salama, N., Guillemin, K., McDaniel, T. K., Sherlock, G., Tompkins, L. and Falkow, S.** (2000) A whole-genome microarray reveals genetic diversity among *Helicobacter pylori* strains. *Proc Natl Acad Sci U S A,* **97:** 14668-73.

**Sambrook, J., Russell, D. W. and Sambrook, J.** (2001) *Molecular Cloning: A Laboratory Manual,* Cold Spring Harbor Laboratory, New York.

**Sandys, G. H.** (1960) A new method of preventing swarming of *Proteus spp.* with a description of a new medium suitable for use in routine laboratory practice. *J Med Lab Technol,* **17:** 224-33.

**Sansonetti, P. J., d'Hauteville, H., Ecobichon, C. and Pourcel, C.** (1983) Molecular comparison of virulence plasmids in *Shigella* and enteroinvasive *Escherichia coli. Ann Microbiol (Inst Pasteur),* **134A:** 295-318.

**Sansonetti, P. J., d'Hauteville, H., Formal, S. B. and Toucas, M.** (1982) Plasmid-mediated invasiveness of shigella-like *Escherichia coli. Ann Microbiol (Inst Pasteur),* **132A:** 351-55.

**Sansonetti, P. J. and Egile, C.** (1998) Molecular bases of epithelial cell invasion by Shigella flexneri. *Antonie Van Leeuwenhoek,* **74:** 191-7.

**Savarino, S. J., Fasano, A., Watson, J., Martin, B. M., Levine, M. M., Guandalini, S. and Guerry, P.** (1993) Enteroaggregative *Escherichia coli* heat-stable enterotoxin 1 represents another subfamily of *E. coli* heat-stable toxin. *Proc Natl Acad Sci U S A,* **90:** 3093-7.

**Scaletsky, I. C., Fabbricotti, S. H., Aranda, K. R., Morais, M. B. and Fagundes-Neto, U.** (2002) Comparison of DNA hybridization and PCR assays for detection of putative pathogenic enteroadherent *Escherichia coli. J Clin Microbiol,* **40:** 1254-8.

**Scaletsky, I. C., Silva, M. L. and Trabulsi, L. R.** (1984) Distinctive patterns of adherence of enteropathogenic *Escherichia coli* to HeLa cells. *Infect Immun,* **45:** 534-6.

**Schäffer, S., Hantke, K. and Braun, V.** (1985) Nucleotide sequence of the iron regulatory gene fur. *Mol Gen Genet,* **200:** 110-13.

**Schembri, M. A., Dalsgaard, D. and Klemm, P.** (2004) Capsule shields the function of short bacterial adhesins. *J Bacteriol,* **186:** 1249-57.

**Schena, M.** (1996) Genome analysis with gene expression microarrays. *Bioessays,* **18:** 427-31.

**Schena, M.** (2003a) *Microarray analysis,* John Wiley & Sons, Hoboken, New Jersey.

**Schena, M.** (2003b) *Microarray analysis,* John Wiley & Sons, Hoboken, New Jersey, pp. 443-52.

**Schena, M.** (2003c) *Microarray analysis,* John Wiley & Sons, Hoboken, New Jersey, pp. 307.

**Schena, M., Shalon, D., Davis, R. W. and Brown, P. O.** (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science,* **270:** 467-70.

**Schena, M., Shalon, D., Heller, R., Chai, A., Brown, P. O. and Davis, R. W.** (1996) Parallel human genome analysis: microarray-based expression monitoring of 1000 genes. *Proc Natl Acad Sci U S A,* **93:** 10614-9.

**Schmidt, H., Beutin, L. and Karch, H.** (1995a) Molecular analysis of the plasmid-encoded hemolysin of *Escherichia coli* O157:H7 strain EDL 933. *Infect Immun,* **63:** 1055-61.

**Schmidt, H., Knop, C., Franke, S., Aleksic, S., Heesemann, J. and Karch, H.** (1995b) Development of PCR for screening of enteroaggregative *Escherichia coli. J Clin Microbiol,* **33:** 701-5.

**Schmoll, T., Hacker, J. and Goebel, W.** (1987) Nucleotide sequence of the *sfaA* gene coding for the S-fimbrial protein subunit of *Escherichia coli. FEMS Microbiol Lett,* **41:** 229-35.

**Schmoll, T., Hoschutzky, H., Morschhauser, J., Lottspeich, F., Jann, K. and Hacker, J.** (1989) Analysis of genes coding for the sialic acid-binding adhesin and two other minor fimbrial subunits of the S-fimbrial adhesin determinant of *Escherichia coli. Mol Microbiol,* **3:** 1735-44.

**Sethabutr, O., Hanchalay, S., Echeverria, P., Taylor, D. N. and Leksomboon, U.** (1985) A non-radioactive DNA probe to identify *Shigella* and enteroinvasive *Escherichia coli* in stools of children with diarrhoea. *Lancet,* **2:** 1095-7.

**Shchepinov, M. S., Case-Green, S. C. and Southern, E. M.** (1997) Steric factors influencing hybridisation of nucleic acids to oligonucleotide arrays. *Nucleic Acids Res,* **25:** 1155-61.

**Sheikh, J., Czeczulin, J. R., Harrington, S., Hicks, S., Henderson, I. R., Le Bouguenec, C., Gounon, P., Phillips, A. and Nataro, J.** (2002) A novel dispersin protein in enteroaggregative *Escherichia coli. J Clin Invest,* **110:** 1329-37.

**Silva, R. M., Toledo, M. R. and Trabulsi, L. R.** (1980) Biochemical and cultural characteristics of invasive *Escherichia coli. J Clin Microbiol,* **11:** 441-4.

**Sinclair, B.** (1999) Everything's great when it sits on a chip. *The scientist,* **13:** 18.

**Smalley, D. J., Whiteley, M. and Conway, T.** (2003) In search of the minimal *Escherichia coli* genome. *Trends Microbiol,* **11:** 6-8.

**Smith, D., Willshaw, G., Stanley, J. and Arnold, C.** (2000) Genotyping of verocytotoxin-producing *Escherichia coli* O157: comparison of isolates of a prevalent phage type by fluorescent amplified-fragment length polymorphism and pulsed-field gel electrophoresis analyses. *J Clin Microbiol,* **38:** 4616-20.

**Smith, H. W. and Gyles, C. L.** (1970) The relationship between two apparently different enterotoxins produced by enteropathogenic strains of *Escherichia coli* of porcine origin. *J Med Microbiol,* **3:** 387-401.

**Smith, H. W. and Halls, S.** (1967) Studies on *Escherichia coli* Enterotoxin. *J Path Bact,* **93:** 531-43.

**Smith, K., Diggle, M. A. and Clarke, S. C.** (2003) Comparison of commercial DNA extraction kits for extraction of bacterial genomic DNA from whole-blood samples. *J Clin Microbiol,* **41:** 2440-3.

**Smoot, J. C., Barbian, K. D., Van Gompel, J. J., Smoot, L. M., Chaussee, M. S., Sylva, G. L., Sturdevant, D. E., Ricklefs, S. M., Porcella, S. F., Parkins, L. D., Beres, S. B., Campbell, D. S., Smith, T. M., Zhang, Q., Kapur, V., Daly, J. A., Veasy, L. G. and Musser, J. M.** (2002) Genome sequence and comparative microarray analysis of serotype M18 group A *Streptococcus* strains associated with acute rheumatic fever outbreaks. *Proc Natl Acad Sci U S A,* **99:** 4668-73.

**Southern, E. M., Case-Green, S. C., Elder, J. K., Johnson, M., Mir, K. U., Wang, L. and Williams, J. C.** (1994) Arrays of complementary oligonucleotides for analysing the hybridisation behaviour of nucleic acids. *Nucleic Acids Res,* **22:** 1368-73.

**Soutourina, O. A., Krin, E., Laurent-Winter, C., Hommais, F., Danchin, A. and Bertin, P. N.** (2002) Regulation of bacterial motility in response to low pH in *Escherichia coli*: the role of the H-NS protein. *Microbiology,* **148:** 1543-51.

**Spellman, P. T., Miller, M., Stewart, J., Troup, C., Sarkans, U., Chervitz, S., Bernhart, D., Sherlock, G., Ball, C., Lepage, M., Swiatek, M., Marks, W. L., Goncalves, J., Markel, S., Iordan, D., Shojatalab, M., Pizarro, A., White, J., Hubley, R., Deutsch, E., Senger, M., Aronow, B. J., Robinson, A., Bassett, D., Stoeckert, C. J., Jr. and Brazma, A.** (2002) Design and implementation of microarray gene expression markup language (MAGE-ML). *Genome Biol,* **3:** research0046.1-0046.9.

**Spicer, E. K. and Noble, J. A.** (1982) *Escherichia coli* heat-labile enterotoxin. Nucleotide sequence of the A subunit gene. *J Biol Chem,* **257:** 5716-21.

**Su, C. and Brandt, L. J.** (1995) *Escherichia coli* O157:H7 infection in humans. *Ann Intern Med,* **123:** 698-714.

**Sussman, M.** (1997a) *Escherichia coli* and human disease, in: Sussman, M. (ed) *Escherichia coli: mechanisms of virulence* Cambridge University Press, Cambridge, pp. 3-48.

**Sussman, M.** (1997b) *Escherichia coli: mechanisms of virulence,* Cambridge University Press, Cambridge.

**Svanborg Eden, C. and Hansson, H. A.** (1978) *Escherichia coli* pili as possible mediatiors of attachment to human urinary tract epithelial cells. *Infect Immun,* **21:** 229-37.

**t Hoen, P. A., de Kort, F., van Ommen, G. J. and den Dunnen, J. T.** (2003) Fluorescent labelling of cRNA for microarray applications. *Nucleic Acids Res,* **31:** e20.

**Taniguchi, T., Fujino, Y., Yamamoto, K., Miwatani, T. and Honda, T.** (1995) Sequencing of the gene encoding the major pilin of pilus colonization factor antigen III (CFA/III) of human enterotoxigenic *Escherichia coli* and evidence that CFA/III is related to type IV pili. *Infect Immun,* **63:** 724-8.

**Taylor, E., Cogdell, D., Coombes, K., Hu, L., Ramdas, L., Tabor, A., Hamilton, S. and Zhang, W.** (2001) Sequence verification as quality-control step for production of cDNA microarrays. *Biotechniques,* **31:** 62-5.

**Taylor, J., Wilkins, P. and Payne, J. M.** (1960) Relation of rabit gut reaction to enteropathogenic *Escherichia coli. Br J Exp Pathol,* **42:** 43-52.

**Taylor, S., Smith, S., Windle, B. and Guiseppi-Elie, A.** (2003) Impact of surface chemistry and blocking strategies on DNA microarrays. *Nucleic Acids Res,* **31:** e87.

**Thompson, A., Lucchini, S. and Hinton, J. C.** (2001) It's easy to build your own microarrayer! *Trends Microbiol,* **9:** 154-6.

**Threlfall, E. J., Ward, L. R., Frost, J. A. and Willshaw, G. A.** (2000) The emergence and spread of antibiotic resistance in food-borne bacteria. *Int J Food Microbiol,* **62:** 1-5.

**Tompkins, D.** (1995) The infectious intestinal disease (IID) in England study: interim report. *PHLS Microbiology Digest,* **13:** 84-5.

**Tompkins, D. S., Hudson, M. J., Smith, H. R., Eglin, R. P., Wheeler, J. G., Brett, M. M., Owen, R. J., Brazier, J. S., Cumberland, P., King, V. and Cook, P. E.** (1999) A study of infectious intestinal disease in England: microbiological findings in cases and controls. *Commun Dis Public Health,* **2:** 108-13.

**Torres, A. G. and Payne, S. M.** (1997) Haem iron-transport system in enterohaemorrhagic *Escherichia coli* O157:H7. *Mol Microbiol,* **23:** 825-33.

**Trabulsi, L. R., Keller, R. and Tardelli Gomes, T. A.** (2002) Typical and atypical enteropathogenic *Escherichia coli. Emerg Infect Dis,* **8:** 508-13.

**Tulloch, E. F., Jr., Ryan, K. J., Formal, S. B. and Franklin, F. A.** (1973) Invasive enteropathic *Escherichia coli* dysentery. An outbreak in 28 adults. *Ann Intern Med,* **79:** 13-7.

**Tung, J. W., Parks, D. R., Moore, W. A., Herzenberg, L. A. and Herzenberg, L. A.** (2004) New approaches to fluorescence compensation and visualization of FACS data. *Clin Immunol,* **110:** 277-83.

**Ulshen, M. H. and Rollo, J. L.** (1980) Pathogenesis of *Escherichia coli* gastroenteritis in man--another mechanism. *N Engl J Med,* **302:** 99-101.

**Umanski, T., Rosenshine, I. and Friedberg, D.** (2002) Thermoregulated expression of virulence genes in enteropathogenic *Escherichia coli. Microbiology,* **148:** 2735-44.

**van Die, I., van Geffen, B., Hoekstra, W. and Bergmans, H.** (1985) Type 1C fimbriae of a uropathogenic *Escherichia coli* strain: cloning and characterization of the genes involved in the expression of the 1C antigen and nucleotide sequence of the subunit gene. *Gene,* **34:** 187-96.

**van Ijperen, C., Kuhnert, P., Frey, J. and Clewley, J. P.** (2002) Virulence typing of *Escherichia coli* using microarrays. *Mol Cell Probes,* **16:** 371-8.

**Velappan, N., Snodgrass, J. L., Hakovirta, J. R., Marrone, B. L. and Burde, S.** (2001) Rapid identification of pathogenic bacteria by single-enzyme amplified fragment length polymorphism analysis. *Diagn Microbiol Infect Dis,* **39:** 77-83.

**Vieira, M. A., Andrade, J. R., Trabulsi, L. R., Rosa, A. C., Dias, A. M., Ramos, S. R., Frankel, G. and Gomes, T. A.** (2001) Phenotypic and genotypic characteristics of *Escherichia coli* strains of non-enteropathogenic *E. coli* (EPEC) serogroups that carry EAE and lack the EPEC adherence factor and Shiga toxin DNA probe sequences. *J Infect Dis,* **183:** 762-72.

**Vincent, J. L., Bihari, D. J., Suter, P. M., Bruining, H. A., White, J., Nicolas-Chanoin, M. H., Wolff, M., Spencer, R. C. and Hemmer, M.** (1995) The prevalence of nosocomial infection in intensive care units in Europe. Results of the European Prevalence of Infection in Intensive Care (EPIC) Study. EPIC International Advisory Committee. *Jama,* **274:** 639-44.

**Wang, D., Coscoy, L., Zylberberg, M., Avila, P. C., Boushey, H. A., Ganem, D. and DeRisi, J. L.** (2002a) Microarray-based detection and genotyping of viral pathogens. *Proc Natl Acad Sci U S A,* **99:** 15687-92.

**Wang, G., Clark, C. G. and Rodgers, F. G.** (2002b) Detection in *Escherichia coli* of the genes encoding the major virulence factors, the genes defining the O157:H7 serotype, and components of the type 2 Shiga toxin family by multiplex PCR. *J Clin Microbiol,* **40:** 3613-9.

**Wang, R. F., Kim, S. J., Robertson, L. H. and Cerniglia, C. E.** (2002c) Development of a membrane-array method for the detection of human intestinal bacteria in fecal samples. *Mol Cell Probes,* **16:** 341-50.

**Watson, A., Mazumder, A., Stewart, M. and Balasubramanian, S.** (1998) Technology for microarray analysis of gene expression. *Curr Opin Biotechnol,* **9:** 609-14.

**Weinstein, D. L., Jackson, M. P., Samuel, J. E., Holmes, R. K. and O'Brien, A. D.** (1988) Cloning and sequencing of a shiga-like toxin type II variant from an *Escherichia coli* strain responsible for edema disease of swine. *J Bacteriol,* **170:** 4223-30.

**Welch, R. A., Burland, V., Plunkett, G., Redford, P., Roesch, P., Rasko, D., Buckles, E. L., Liou, S. R., Boutin, A., Hackett, J., Stroud, D., Mayhew, G. F., Rose, D. J., Zhou, S., Schwartz, D. C., Perna, N. T., Mobley, H. L., Donnenberg, M. S. and Blattner, F. R.** (2002) Extensive mosaic structure revealed by the complete genome sequence of uropathogenic *Escherichia coli. Proc Natl Acad Sci U S A,* **99:** 17020-4.

**Welch, R. A., Dellinger, E. P., Minshew, B. and Falkow, S.** (1981) Hemolysin contributes to virulence of extraintestinal *Escherichia coli*. *Mol Gen Genet,* **175:** 343-50.

**Welinder-Olsson, C., Badenfors, M., Cheasty, T., Kjellin, E. and Kaijser, B.** (2002) Genetic profiling of enterohemorrhagic *Escherichia coli* strains in relation to clonality and clinical signs of infection. *J Clin Microbiol,* **40:** 959-64.

**Whittam, T. S.** (2004), Multi Locus Sequence Typing (MLST) of pathogenic *Escherichia coli* [Online], Available from <www.shigatox.net/cgi-bin/mlst7/index.html> [Accessed 28/02/2005]

**Woodward, M. J. and Charles, H. P.** (1983) Polymorphism in *Escherichia coli*: rtl, atl and gat regions behave as chromosomal alternatives. *J Gen Microbiol,* **129:** 75-84.

**Wrobel, G., Schlingemann, J., Hummerich, L., Kramer, H., Lichter, P. and Hahn, M.** (2003) Optimization of high-density cDNA-microarray protocols by 'design of experiments'. *Nucleic Acids Res,* **31:** e67.

**Yamamoto, T. and Echeverria, P.** (1996) Detection of the enteroaggregative *Escherichia coli* heat-stable enterotoxin 1 gene sequences in Enterotoxigenic *E. coli* strains pathogenic in humans. *Infect Immun,* **64:** 1441-5.

**Yang, Y. H., Dudoit, S., Luu, P., Lin, D. M., Peng, V., Ngai, J. and Speed, T. P.** (2002) Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Res,* **30:** e15.

**Yavzori, M., Porath, N., Ochana, O., Dagan, R., Orni-Wasserlauf, R. and Cohen, D.** (1998) Detection of enterotoxigenic *Escherichia coli* in stool specimens by polymerase chain reaction. *Diagn Microbiol Infect Dis,* **31:** 503-509.

**Zapata, G., Vann, W. F., Aaronson, W., Lewis, M. S. and Moos, M.** (1989) Sequence of the cloned *Escherichia coli* K1 CMP-N-acetylneuraminic acid synthetase gene. *J Biol Chem,* **264:** 14769-74.

**Zeigler, D. R.** (2003) Gene sequences useful for predicting relatedness of whole genomes in bacteria. *IJSEM,* **53:** 1893-900.

**Zhou, Z., Ogasawara, J., Nishikawa, Y., Seto, Y., Helander, A., Hase, A., Iritani, N., Nakamura, H., Arikawa, K., Kai, A., Kamata, Y., Hoshi, H. and Haruki, K.** (2002) An outbreak of gastroenteritis in Osaka, Japan due to *Escherichia coli* serogroup O166:H15 that had a coding gene for enteroaggregative *E. coli* heat-stable enterotoxin 1 (EAST1). *Epidemiol Infect,* **128:** 363-71.

## APPENDIX I: BUFFERS AND SOLUTIONS

**Water**
For PCR applications: Promega Nuclease free water
For preparation/dilution of solutions: MilliQ filtered water

**Qiagen DNeasy Tissue Kit** (Qiagen, Crawley, UK)

| | |
|---|---|
| Buffer ATL | resuspension buffer |
| Buffer AL | lysis buffer, chaotropic salt solution |
| Buffer AW1 | wash buffer 1, chaotropic salt solution |
| Buffer AW2 | wash buffer 2, sodium azide solution |

**MagNa Pure LC DNA Isolation kit** (Roche, Lewes, UK)

| | |
|---|---|
| Wash buffer I | wash buffer I, chaotropic salt solution for removing PCR inhibitors |
| Wash buffer II | wash buffer II for removing salts and proteins |
| Lysis/Binding buffer | for cell lysis and binding of DNA |
| Proteinase K | for digestion of proteins |
| Magnetic glass particles | for binding DNA |
| Elution buffer | 10mM Tris-HCl, pH 8.0 for elution of DNA |

**PBS** (CPHL Media service, London UK)
0.1 M Phosphate buffered saline pH 7.4
80 mM sodium phosphate
15 mM potassium phosphate
27 mM KCl
1.37M NaCl

**10x TBE** (Invitrogen, Paisley, UK)
890 mM Tris
890 mM boric acid
20 mM EDTA, pH 8.4

**Orange G gel loading buffer** (Severn Biotech, Kidderminster, UK)
0.25% Orange G
10% Ficoll in TE

**DIG High Prime DNA Labeling and Detection Starter Kit II** (Roche, Lewes, UK)

| | |
|---|---|
| DIG-High Prime | labelling mixture containing random primers, nucleotides, DIG-dUTP and Klenow enzyme |
| DIG-labelled Control DNA | 5 µg/ml of *Bam* HI pBR328 DNA |
| DNA Dilution buffer | 50 µg/ml herring sperm DNA in 10mM Tris-HCl, 1mM EDTA pH 8.0 |
| Anti-Digoxigenin-AP Conjungate | 750U/ml Sheep Fab fragment conjungated to alkaline phosphatase |
| CSPD ready to use | chemiluminescent substrate for alkaline phosphatase |
| Blocking solution | |

| | |
|---|---|
| DIG Easyhyb granules | to be dissolved and used as hybridisation solution |

**ECF random prime labelling module** (Amersham, Amersham, UK)

| | |
|---|---|
| Nucleotide mix | Fluorescein-11-dUTP, dATP, dGTP and dTTP in Tris-HCl pH 7.8, 2-mercaptoethanol and $MgCl_2$ |
| Primers | Random nonamers |
| Enzyme solution | 5U/µl exonuclease free Klenow in buffer pH6.5 |
| Control unlabelled DNA | 10ng/ml *Hind* III lambda DNA |
| Control fluorescein-labelled DNA | 50pg/ml fluorescein-labelled *Hind* III lambda DNA in 50ng/ml herring sperm carrier DNA |
| Liquid block | Blocking solution |

**ECF signal amplification module** (Amersham, Amersham, UK)
Anti-fluorescein alkaline phosphatase (AP) conjungate
ECF substrate
ECF substrate dilution buffer

**10 x TE buffer** (Invitrogen Paisley, UK)
100mM Tris pH 7.4
10mM EDTA

**Rediprime II random prime labelling system** (Amersham, Amersham, UK)

| | |
|---|---|
| Labelling reaction containing | Buffered solution of dATP, dGTP, dTTP<br>exonuclease free Klenow enzyme<br>Random primers in a dried and stabilised form |
| Control DNA | 300ng of lambda *Hind* III DNA in a dried and stabilised form |

**BioPrime Labelling kit** (Invitrogen, Paisley, UK)
DNA Polymerase (Klenow fragment)
2.5 x random prime solution
(other kit components are not used)

**Qiaquick PCR Purification Kits** (Qiagen, Crawley, UK)

| | |
|---|---|
| Buffer PB | chaotropic salt solution |
| Buffer PE | wash buffer |
| Buffer EB | elution buffer |

**Hybridisation**

<u>Southern blot</u>
**20 x SSC** (Invitrogen, Paisley, UK)
3.0 M NaCl
0.3 M sodium citrate
pH 7.0

**10% SDS** (Invitrogen, Paisley, UK)

**BSA** (Sigma, Poole, UK)

Panorama arrays
**20 x SSPE** (Invitrogen, Paisley, UK)
3.0 M NaCl
0.2 M $NaH_2PO_4$
0.02 M EDTA
pH 7.4

**stripping buffer**
10mM Tris, pH 7.5
1mM EDTA
1% SDS

Pan Arrays and custom virulence chip

**Prehybridisation buffer**
4 x SSC
0.5%SDS
1% BSA

**1 x hybridisation buffer**
50% formamide
5 x SSC
0.1%SDS

**wash buffer A**
1 x SSC
0.5%SDS

**wash buffer B**
0.06 x SSC

**HiSpeed Plasmid Purification**
Buffer P1
Buffer P2
Buffer P3
Buffer QBT
Buffer QC
Buffer QF
Buffer TE

**ethanol** (AnalR grade, Sigma, Poole, UK)

**nuclease free $dH_2O$** (Promega, Southampton, UK).

## APPENDIX II: PLASMIDS CARRYING PATHOGENICITY MARKER SEQUENCES AND AMPLIFICATION PRIMER PAIRS FOR PROBE PREPARATION

| Name | Gene | Position on Acc. No. | Primers | | Length | $T_A$ | Restriction enzymes |
|---|---|---|---|---|---|---|---|
| pJFFEC1 | *stxI* | 3464-3987 on L04539 | EC1/2-L | 5′-GC**TCTAGA**TTGAACGAAATAATTTATATG-3′ | 530 | 58 | *Xba I* |
| | | | EC1-R | 5′-GC**TCTAGA**TGATGATGACAATTCAGTAT-3′ | | | |
| pJFFEC2 | *stxII* | 518-1036 on M21534 | EC1/2-L | 5′-GC**TCTAGA**TTGAACGAAATAATTTATATG-3′ | 520 | 58 | *Xba I + Bam HI* |
| | | | EC2-R | 5′-GC**GGATCC**ATGATGGCAATTCAGTAT-3′ | | | |
| pJFFEC3 | *eltIA* | 159-848 on V00275 | EC3-L | 5′-AC**GGATCC**TACCGTGCTGACTCTAGACC-3′ | 680 | 58 | *Bam HI +EcoR I* |
| | | | EC3-R | 5′-CGC**GAATTC**TGTTATATATGTCAAC-3′ | | | |
| pJFFEC4 | *eltIIA* | 107-770 on M17894 | EC4-L | 5′-TC**GAATTC**AGCAAACGATTTCTTTAGAG-3′ | 665 | 58 | *EcoR I + Kpn I* |
| | | | EC4-R | 5′-AA**GGTACC**CCTGCGTTTTAAGAGTTTTT-3′ | | | |
| pJFFECAAF | *aaf/I* | 53-643 on X81423 | ECAAF-L | 5′-TCT**GAATTC**GACACAGACTCTGGCGAAAG-3′ | 590 | 60 | *EcoR I* |
| | | | ECAAF-R | 5′-TGT**GAATTC**TGGGATTGCACTCTCAGGA-3′ | | | |
| pJFFECAER | *fhuA* | 1039-2036 on M12486 | ECAER-L | 5′-ATG**GAATTC**CCGGTTTCCGTGCTTTA-3′ | 1000 | 58 | *EcoR I* |
| | | | ECAER-R | 5′-CGG**GAATTC**CGGCAACGCGGTTAA-3′ | | | |
| pJFFECBFP | *bfpA* | 32-599 on Z12295 | ECBFP-L | 5′-CCT**GAATTC**ACGGGGGTTTTATAAGGAAA-3′ | 570 | 58 | *EcoR I* |
| | | | ECBFP-R | 5′-TCA**GAATTC**TTACATGCAGTTGCCGCTTC-3′ | | | |
| pJFFECCFA | *cfa/I* | 1403-2063 on M55661 | ECCFA-L | 5′-AAT**ATCGAT**GATAACTGTGTAAAAA-3′ | 650 | 58 | *Cla I + Pst I* |
| | | | ECCFA-R | 5′-GTTTC**CTGCAG**TTGGGGCGGTAC-3′ | | | |
| pJFFECCNF | *cnf-1* | 1321-2229 on X70670 | ECCNF-L | 5′-TTT**AAGCTTT**TACTAAAAAATTATTA-3′ | 910 | 50 | *Hind III* |
| | | | ECCNF-R | 5′-TTT**AAGCTT**AACGTCTAACAAATT-3′ | | | |
| pJFFECCS3 | *cfa/II* | 11-556 on M35657 | ECCS3-L | 5′-GTA**GAATTC**CAGGTACGTATACTGTTGG-3′ | 540 | 58 | *EcoR I* |
| | | | ECCS3-R | 5′-TAT**GAATTC**ACGGTAATTACCTGAAACT-3′ | | | |
| pJFFECEAE | *eae* | 2565-3241 on M58154 | ECEAE-L | 5′-GGC**GAATTC**CGCATGAGCGGCTG-3′ | 680 | 58 | *EcoR I* |
| | | | ECEAE-R | 5′-ATT**GAATTC**ATAGGCGCGAGCCGTCAC-3′ | | | |
| pJFFECF1C | F1C | 185-704 on M13053 | ECF1C-L | 5′-GC**GAATTC**ATCTCCATGGCTGTA-3′ | 520 | 58 | *EcoR I* |
| | gene | | ECF1C-R | 5′-GC**GAATTC**ACTTTAAAGGTGGCGTCG-3′ | | | |
| pJFFECFIM | *fimA* | 623-1130 on X00981 | ECFIM-L | 5′-GGC**GAATTC**TGTTCTGTCGGCTCTGTC-3′ | 510 | 58 | *EcoR I* |
| | | | ECFIM-R | 5′-TTG**GAATTC**AACCTTGAAGGTCGCATC-3′ | | | |
| pJFFECIPA | *ipaH* | 991-1676 on M76445 | ECIPA-L | 5′-TCC**GAATTC**CTTGACCGCCTTT-3′ | 690 | 60 | *EcoR I* |
| | | | ECIPA-R | 5′-TTC**GAATTC**ACGCATCACCTGTGCA-3′ | | | |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| pJFFECIUC | *iucC* | 2829-3635 on X76100 | ECIUC-L | 5′-GC**GAATTC**GGCGATGACCGCTACTG-3′ | 810 | 58 | *EcoR I* |
| | | | ECIUC-R | 5′-GC**GAATTC**CAGCGTGAAGCCAGTG-3′ | | | |
| pJFFECK1 | *neuA* | from 786 on J05023 | ECK1B-L | 5′-GCG**GAATTC**ATTGGACACTCGCTGTTTG-3′ | 840 | 60 | *EcoR I* |
| | *neuC* | to 499 on M84026 | ECK1B-R | 5′-GCG**GAATTC**GCATTGATGCTGCGATAG-3′ | | | |
| pJFFECK5 | *kfiB* | 5120-5719 on X77617 | ECK5-L | 5′-TCT**GAATTC**GACTACCTCCCATAATG-3′ | 600 | 58 | *EcoR I* |
| | | | ECK5-R | 5′-CGC**GAATTC**CGGGTGGGCAGATCCATCT-3′ | | | |
| pJFFECPAP | *papA* | 1775-2300 on X61239 | ECPAP-L | 5′-ATT**GAATTC**GTTATTGCCGGTGCGGTA-3′ | 530 | 58 | *EcoR I* |
| | | | ECPAP-R | 5′-TCA**GAATTC**AATTCGCAACTGCTGAGA-3′ | | | |
| pJFFECSFS | *sfaS* | 626-1195 on X16664 | ECSFS-L | 5′-GC**GAATTC**TTATATTGGCCACCGGTC-3′ | 570 | 60 | *EcoR I* |
| | | | ECSFS-R | 5′-GC**GAATTC**AACAATGCAAACGATGGC-3′ | | | |
| pJFFECST | *stIa/stIb* | part of pKAD008 | ECST-L | 5′-ATC**TCTAGA**GATCGAATTCCCG-3′ | 680 | 60 | *Xba I* |
| | | | ECST-R | 5′-ATG**TCTAGA**CCCAGAATCTGAGCACA-3′ | | | |
| pJFFECSFA | *sfaA* | 176-706 on X17420 | ECSFA-L | 5′-TTA**GAATTC**ATCTCCATGGCTGTA-3′ | 530 | 55 | *EcoR I* |
| | | | ECSFA-R | 5′-TGA**GAATTC**TGGTACTGAACTTTAAAGGT-3′ | | | |
| pJFFRTX6II | *hlyA* | 3420-4093 on M14107 | RTX6II-L | 5′-TAT**GAATTC**ACTCATATCAATGG-3′ | 680 | 55 | *EcoR I* |
| | | | RTX6II-R | 5′-TCT**GAATTC**TGATTAGAGATATCACCTGACTC-3′ | | | |
| pJFFRTX7 | *elyA* | 2140-2922 on X79839 | RTX7-L | 5′-GAT**GAATTC**AAAGGCGGTAA-3′ | 780 | 58 | *EcoR I* |
| | | | RTX7-R | 5′-TAA**GAATTC**ATCACCTGAATCGAAC-3′ | | | |
| pJFFECAST | *astA* | 133-393 on S81691 | ECAST-L | 5′-CGC**GAATTC**TGCCATCAACACAGTATA-3′ | 260 | 58 | *EcoR I* |
| | | | ECAST-R | 5′-CGC**GGATCC**GTTGGATAAGCGAAGAAC-3′ | | | |
| pJFFECCOF | *cofA* | 386-1000 on D37957 | ECCOF-L | 5′-CG**GAATTC**TGGAAGTCATCATCGTT-3′ | 615 | 58 | *EcoR I* |
| | | | ECCOF-R | 5′-CG**GAATTC**GGCTCGCCAAAGTAATAGAG-3′ | | | |
| pJFFECLNG | *lngA* | 61-651 on AF004308 | ECLNG-L | 5′-CG**GAATTC**CGTGTATAACCGGACACA-3′ | 590 | 58 | *EcoR I* |
| | | | ECLNG-R | 5′-CG**GAATTC**GGCGGCCACAGACATATCTA-3′ | | | |
| pJFFECCHU | *chuA* | 1661-2419 on U67920 | ECCHU-L | 5′-GCG**GAATTC**GCTATGACAGTTATCGC-3′ | 760 | 64 | *EcoR I* |
| | | | ECCHU-R | 5′-GCG**GAATTC**TTGCGGCGACCAGTACT-3′ | | | |
| | 16S rRNA | | 16S-L | 5′-AGAGTTTGATCATGGCTCAG-3′ | 1500 | 55 | n/a |
| | | | 16S-R | 5′-GTGTGACGGGCGGTGTGTAC-3′ | | | |

*Molecular and Cellular Probes* (2002) **16**, 371–378
doi:10.1006/mcpr.2002.0437, available online at http://www.idealibrary.com on IDEAL®

# Virulence typing of *Escherichia coli* using microarrays

C. van Ijperen,[1] P. Kuhnert,[2] J. Frey[2] and J. P. Clewley*[1]

[1]*Molecular Biology Unit, Central Public Health Laboratory, 61 Colindale Avenue, London, NW9 5HT, UK and* [2] *Institute of Veterinary Bacteriology, University of Bern, 122 Langasse, CH-3012, Bern, Switzerland*

*(Received 24 July 2002; Accepted for publication 13 August 2002)*

We describe a microarray based broad-range screening technique for *Escherichia coli* virulence typing. Gene probes were amplified by PCR from a plasmid bank of characterised *E. coli* virulence genes and were spotted onto a glass slide to form an array of capture probes. Genomic DNA from *E. coli* strains which were to be tested for the presence of these virulence gene sequences was labelled with fluorescent cyanine dyes by random amplification and then hybridised against the array of probes. The hybridisation, washing and data analysis conditions were optimised for glass slides, and the applicability of the method for identifying the presence of the virulence genes was determined using reference strains and clinical isolates. It was found to be a sensitive screening method for detecting virulence genes, and a powerful tool for determining the pathotype of *E. coli*. It will be possible to expand and automate this microarray technique to make it suitable for rapid and reliable diagnostic screening of bacterial isolates. © 2002 Elsevier Science Ltd. All rights reserved.

***KEYWORDS:*** pathogenic *Escherichia coli*, microarray, virulence typing, probe hybridisation, genetic diversity.

## INTRODUCTION

DNA arrays are available in formats ranging from high density chips with over a hundred thousand gene probes, to low density membranes with several genes only.[1,2] Probes immobilised on a solid support may be open reading frames (ORFs) which have been amplified by PCR or oligonucleotides representing a conserved region of a gene.[3] Two-colour hybridisations allow the comparison of test and control samples on the same slide and give more reproducible and reliable data than approaches using single labels and membrane based formats.[4] There is also a limit to the number of probes that can be applied to a semi-porous surface compared to a glass slide. High density oligonucleotide chips are used for gene expression studies, mutation detection, resequencing (reviewed by Hacia *et al.*[5]) and genotypic analysis (reviewed by Lipshutz *et al.*[1] and Gingeras *et al.*[6]).

The use of arrays in bacteriology has included *Salmonella*, *Helicobacter* and *Campylobacter* typing,[7–9] and Chizhikov and colleagues have recently described a multiplex PCR with a microarray hybridisation step for the detection of some *Escherichia coli* virulence genes.[10] The species *E. coli* consists of many different genotypes and phenotypes. These strains range from the highly pathogenic *E. coli* O157:H7 to non-pathogenic isolates which form normal intestinal flora and are often used as safe laboratory strains.[11] Pathogenic *E. coli* strains regularly cause disease in people exposed to contaminated food.[12] *Escherichia coli* infection is a common cause of diarrhoea in infants in developing countries, and can manifest as haemorrhagic colitis and haemolytic-uremic syndrome.[13] These diseases are a result of the virulence factors for colonisation and/or invasion of the host encoded by genes of pathogenic *E. coli* strains. Several of these

* Author to whom correspondence should be addressed at: Dr. J. P. Clewley, SBVL, Central Public Health Laboratory, 61 Colindale Avenue, London NW9 5HT, UK. Tel: +44 20 8200 4400; Fax: +44 20 8200 1569; E-mail: jclewley@phls.org.uk

0890–8508/02/$ – see front matter

NZ1470-95, the enteropathogenic *E. coli* (EPEC) NZ1743-95, the enterohaemorrhagic *E. coli* (EHEC) NZ2168-97 and the enterotoxigenic (ETEC) NZ3211-94. Two clinical isolates IHE3034 and IMI100 were also used in this study. *Escherichia coli* K-12 was used as a negative control. Genomic DNA from these was extracted as previously described by Kuhnert *et al.*[14] Five *E. coli* strains from the ECOR collection (ATCC35320, ATCC35321, ATCC35322, ATCC35323 and ATCC35324) were included in this study. Genomic DNA was extracted from overnight cultures of these using a MagNA Pure extraction robot (Roche, Lewes, UK). All genomic DNA extracts were RNase treated and quantified using a Genequant instrument (Amersham Biosciences, Little Chalfont, UK).

## Array constructions

Twenty-four of 28 probes used to construct the array have previously been described by Kuhnert *et al.*[14] In addition four constructs were made for the detection of the genes for the heat-stable enterotoxin 1 (*astA*), colonization factor antigen III (*cofA*), longus type IV pilus (*lngA*) and the haemin receptor (*chuA*). Primers and bacterial isolates used for amplification of the genes are listed in Table 2. Amplified genes were cloned into pBluescript SK- (Stratagene, Amsterdam, The Netherlands) and the sequences verified. Details of the preparation of the probes are given in the Results section. Restriction enzyme digestion was used for the isolation of the virulence genes from the plasmids. The gene inserts were gel purified before PCR amplification. The length of the 28 amplified products varied between 250 and 1400 bases. PCR products were purified using HighPure PCR purifications columns (Roche, Lewes, UK). Membrane macroarrays were prepared as described previously.[14] For the preparation of the microarrays 1·5 µg of PCR product was resuspended in 20 µl 50% DMSO and transferred to a 384-well microtiter plate. CMT-GAP II microarray slides (Corning, Schiphol-Rijk, The Netherlands) were used for printing of the probes using a Microgrid II (BioRobotics Ltd, Cambridge, UK) arrayer at 35% humidity and 23°C. Each PCR product was spotted six times on each microarray. After spotting the slides were crosslinked in a UV-linker at 60 mJ and denatured for 3 min in 95°C water just before hybridisation, followed by an isopropanol wash for fixing.

## Array hybridisation

Genomic DNA used for hybridisation of membranes was labelled with digoxigenin using a DIG High Prime labelling kit (Roche, Lewes, UK) following the manufacturer's instructions. Labelled genomic DNA was used without purification in an overnight hybridisation at 68°C as described previously.[14] After incubation with antibody and CPD Star (Roche, Lewes, UK), signal detection of the membrane based macroarrays was done on X-ray film.

Genomic DNA used for hybridisation of microarrays was fluorescently labelled. The fluorescent cyanine dyes Cy3-dUTP and Cy5-dUTP (Amersham, Little Chalfont, UK) were incorporated into 2·5 µg *E. coli* genomic DNA by random amplification as described by Murray *et al.*[17] After incorporation of the Cy dyes, DNA was purified with Qiaquick purification columns (Qiagen, Crawley, UK) and isopropanol precipitated. Labelled products were dried under vacuum and resuspended in 30 µl hybridisation buffer containing 50% formamide, 5 × SSC and 0·1% SDS. After at least 1 h prehybridisation in 4 × SSC, 1% SDS and 1% BSA, half of the fluorescently labelled probe was used in an overnight hybridisation at 42°C. Post-hybridisation washes were carried out for 2 × 5 min in 2 × SSC, 0·05% SDS followed by 2 × 15 min in 0·06 × SSC. Microarrays were transferred to centrifugation tubes, dried at 2000 rpm for 4 min and scanned using

**Table 2.** Primer sequences and *E. coli* isolates used for amplification of virulence probes

| Gene | Primer | Sequence | Strain |
|---|---|---|---|
| *astA* | ECAST-L | CGCGAATTCTGCCATCAACACAGTATA | NZ1470-95 |
| | ECAST-R | CGCGGATCCGTTGGATAAGCGAAGAAC | |
| *cofA* | ECCOF-L | CGGAATTCTGGAAGTCATCATCGTT | DS15-1 |
| | ECCOF-R | CGGAATTCGGCTCGCCAAAGTAATAGAG | |
| *lngA* | ECLNG-L | CGGAATTCCGTGTATAACCGGACACA | DS15-1 |
| | ECLNG-R | CGGAATTCGGCGGCCACAGACATATCTA | |
| *chuA* | ECCHU-L | GCGGAATTCGCTATGACAGTTATCGC | 536 |
| | ECCHU-R | GCGGAATTCTTGCGGCGACCAGTACT | |

**Table 3.** Distribution of virulence genes in *E. coli* strains

| Gene ID[1] | K12 | UPEC | EAggEC | EPEC | ETEC | EHEC | IM100 | IHE 3034 | ATCC 35320 | ATCC 35321 | ATCC 35322 | ATCC 35323 | ATCC 35324 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *aafI* | – | – | ++ | – | – | – | – | – | – | – | – | – | – |
| *bfpA* | – | – | – | + | – | – | – | – | – | – | – | – | – |
| *cfaI* | – | – | – | – | – | – | ++ | – | – | – | – | – | – |
| *cfaII* | – | – | – | – | ++ | – | – | – | – | – | – | – | – |
| *cfaIII* | – | – | – | – | – | – | – | – | – | – | – | – | – |
| *eae* | – | – | – | ++ | – | ++ | – | – | – | – | – | – | – |
| *F1C* gene | – | ++ | – | – | – | – | – | – | – | – | – | – | |
| *lngA* | – | – | – | – | + | – | – | – | – | – | – | – | – |
| *papA* | – | ++ | – | – | – | – | – | – | – | + | – | – | – |
| *sfaA* | – | ++ | – | – | – | – | – | + | – | – | – | – | – |
| *sfaS* | – | ++ | – | – | – | – | – | + | – | – | – | – | – |
| *neuA, neuB* | – | – | – | – | – | – | – | ++ | – | – | – | – | – |
| *kfiB* | – | – | – | – | – | – | – | – | – | – | – | – | – |
| *ipaH* | – | – | – | – | – | – | – | – | – | – | – | – | – |
| *chuA* | – | ++ | + | – | – | + | – | ++ | – | – | – | – | – |
| *iucC* | – | – | – | – | – | – | – | – | – | ++ | – | – | ++ |
| *astA* | – | – | + | – | + | + | ++ | – | ++ | – | – | – | ++ |
| *cnfI* | – | ++ | – | – | – | – | – | – | – | – | – | – | – |
| *ehxA* | – | – | – | – | – | + | – | – | – | – | – | – | – |
| *eltIA* | – | – | – | – | ++ | – | ++ | – | – | – | – | – | – |
| *eltIIA* | – | – | – | – | – | – | – | – | – | – | – | – | – |
| *hlyA* | – | ++ | – | – | – | – | – | – | – | – | – | – | – |
| *sltI* | – | – | – | – | – | ++ | – | – | – | – | – | – | – |
| *sltII* | – | – | – | – | – | ++ | – | – | – | – | – | – | – |
| *stIA, stIB* | – | – | – | – | ++ | – | ++ | – | – | – | – | – | – |
| *rrs*[2] | ++ | ++ | ++ | ++ | ++ | ++ | ++ | ++ | ++ | ++ | ++ | ++ | ++ |
| *fhuA*[2] | ++ | ++ | ++ | ++ | ++ | ++ | ++ | ++ | ++ | ++ | ++ | ++ | ++ |
| *fimA*[2] | ++ | ++ | – | ++ | – | ++ | + | ++ | ++ | ++ | ++ | + | ++ |

[1] Genes are in the same order as Table 1.

[2] Genes are positive controls for hybridisation and not relevant for the virulence screening.

achieved by expanding and adapting a membrane based system for virulence screening to a microarray format. The methods involved were little more complicated than those used for the membrane based system.

Fluorescent dyes were incorporated by random amplification of the genomic DNA and their use with slides improved the sensitivity of the method by giving a lower background compared to that observed in membrane hybridisation experiments. For example, some of the strains gave a high background signal on the membrane system, making identification of the positive spots difficult. In contrast, all of the strains were relatively easy to analyse using the glass slide system. On these slides there was a clear separation in signal intensity when genes were present compared to when they were absent. Additionally, the time spent processing samples was significantly less. This saving was mainly due to the shorter post hybridisation washes and antibody incubation time. The use of fluorescence makes amplification of the signal unnecessary, because the confocal microscope scanner is sensitive enough to directly detect the emission signal. The system was optimised using a set of strains that were previously tested by the membrane based method.[14] No false positive were detected and there were less than 1% false negative reactions and therefore the results of the two assays were in good agreement.

Multiplex PCR has previously been used for *E. coli* virulence gene testing.[10,24] After amplification the multiplex PCR products were analysed on gels or hybridised to microarrays. However, the number of genes that can be detected in one multiplex PCR is limited and very restrictive, thereby not allowing the detection of variants. In contrast, our method uses random amplification of genomic DNA and is a more rapid way of creating a broad-range hybridisation target. The results we obtained show that the use of genomic DNA and rapid random amplification does not affect the sensitivity of the method and that individual genes are still detectable. The positive signals were strong and there was no background hybridisation from DNA from control strains.

Investigation of the distribution of virulence genes in *E. coli* is important for public health and food microbiology. The slide based hybridisation system we describe can be used for typing a wide range of *E. coli* isolates from different sources. Strains carrying virulence factors are likely to cause health problems, but *E. coli* isolated from healthy individuals may also have a variety of virulence genes as is shown in the microarray analysis of the ECOR strains. This confirms previous findings from Mühldorfer and colleagues who found virulence genes in *E. coli* isolated from water and stool samples of healthy individuals.[11] A rapid screening system for many different genes in one array would complement current diagnostic tests. Furthermore, it would allow virulence testing of the 'non pathogenic' *E. coli* strains that are used in biotechnological applications, and would allow the testing of animal isolates for food safety purposes. Further developments of automated equipment for microarray technology should make this system suitable for high throughput testing in diagnostic laboratories.

## ACKNOWLEDGEMENTS

We thank Yvonne Schlatter and Hélène Berthoud for their help in probe preparation.

## REFERENCES

1. Lipshutz, R. J., Fodor, S. P., Gingeras, T. R. & Lockhart, D. J. (1999). High density synthetic oligonucleotide arrays. *Nature Genetics* **21**, 20–4.
2. Anthony, R. M., Brown, T. J. & French, G. L. (2000). Rapid diagnosis of bacteremia by universal amplification of 23S ribosomal DNA followed by hybridization to an oligonucleotide array. *Journal of Clinical Microbiology* **38**, 781–8.
3. Kane, M. D., Jatkoe, T. A., Stumpf, C. R., Lu, J., Thomas, J. D. & Madore, S. J. (2000). Assessment of the sensitivity and specificity of oligonucleotide (50mer) microarrays. *Nucleic Acids Research* **28**, 4552–7.
4. Duggan, D. J., Bittner, M., Chen, Y., Meltzer, P. & Trent, J. M. (1999). Expression profiling using cDNA microarrays. *Nature Genetics* **21**, 10–14.
5. Hacia, J. G. (1999). Resequencing and mutational analysis using oligonucleotide microarrays. *Nature Genetics* **21**, 42–7.
6. Gingeras, T. R., Ghandour, G., Wang, E. *et al.* (1998). Simultaneous genotyping and species identification using hybridization pattern recognition analysis of generic *Mycobacterium* DNA arrays. *Genome Research* **8**, 435–48.
7. Porwollik, S., Wong, R. M. & McClelland, M. (2002). Evolutionary genomics of Salmonella: Gene acquisitions revealed by microarray analysis. *Proceedings of National Academic Sciences of United States of America* **99**, 8956–61.
8. Dorrell, N., Mangan, J. A., Laing, K. G. *et al.* (2001). Whole genome comparison of *Campylobacter jejuni* human isolates using a low-cost microarray reveals extensive genetic diversity. *Genome Research* **11**, 1706–15.
9. Salama, N., Guillemin, K., McDaniel, T. K., Sherlock, G., Tompkins, L. & Falkow, S. (2000). A whole-genome microarray reveals genetic diversity among *Helicobacter pylori* strains. *Proceedings of National Academic Sciences of United States of America* **97**, 14668–73.
10. Chizhikov, V., Rasooly, A., Chumakov, K. & Levy, D. D. (2001). Microarray analysis of microbial virulence factors. *Applied Environmental Microbiology* **67**, 3258–63.
11. Muhldorfer, I., Blum, G., Donohue-Rolfe, A. *et al.* (1996). Characterization of *Escherichia coli* strains isolated from environmental water habitats and from stool samples of healthy volunteers. *Research in Microbiology* **147**, 625–35.
12. O'Brien, S. J., Murdoch, P. S., Riley, A. H. *et al.* (2001). A foodborne outbreak of Vero cytotoxin-producing *Escherichia coli* O157:H-phage type 8 in hospital. *Journal of Hospital Infection* **49**, 167–72.
13. Su, C. & Brandt, L. J. (1995). *Escherichia coli* O157:H7 infection in humans. *Annual Internal Medicine* **123**, 698–714.
14. Kuhnert, P., Hacker, J., Muhldorfer, I., Burnens, A. P., Nicolet, J. & Frey, J. (1997). Detection system for *Escherichia coli*-specific virulence genes: absence of virulence determinants in B and C strains. *Applied Environmental Microbiology* **63**, 703–9.
15. Kuhnert, P., Boerlin, P. & Frey, J. (2000). Target genes for virulence assessment of *Escherichia coli* isolates from water, food and the environment. *FEMS Microbiological Review* **24**, 107–17.
16. De Boer, E. & Heuvelink, A. E. (2000). Methods for the detection and isolation of Shiga toxin-producing *Escherichia coli. Symposium Series Society of Applied Microbiology* **29**, 133S-43S.
17. Murray, A. E., Lies, D., Li, G., Nealson, K., Zhou, J. & Tiedje, J. M. (2001). DNA/DNA hybridization to microarrays reveals gene-specific differences between closely related microbial genomes. *Proceedings of National Academic Sciences of United States of America* **98**, 9853–8.
18. Savarino, S. J., Fasano, A., Watson, J. *et al.* (1993). Enteroaggregative *Escherichia coli* heat-stable enterotoxin 1 represents another subfamily of *E. coli* heat-stable toxin. *Proceedings of National Academic Sciences of United States of America* **90**, 3093–7.
19. Taniguchi, T., Fujino, Y., Yamamoto, K., Miwatani, T. & Honda, T. (1995). Sequencing of the gene encoding the major pilin of pilus colonization factor antigen III (CFA/III) of human enterotoxigenic *Escherichia coli* and evidence that CFA/III is related to type IV pili. *Infection Immunology* **63**, 724–8.
20. Gomez-Duarte, O. G., Ruiz-Tagle, A., Gomez, D. C. *et al.* (1999). Identification of lngA, the structural gene of longus type IV pilus of enterotoxigenic *Escherichia coli. Microbiology* **145**, 1809–16.

21. Nagy, G., Dobrindt, U., Kupfer, M., Emody, L., Karch, H. & Hacker, J. (2001). Expression of hemin receptor molecule ChuA is influenced by RfaH in uropathogenic *Escherichia coli* strain 536. *Infection Immunology* **69**, 1924–8.
22. Torres, A. G. & Payne, S. M. (1997). Haem iron-transport system in enterohaemorrhagic *Escherichia coli* O157:H7. *Molecular Microbiology* **23**, 825–33.
23. Watson, A., Mazumder, A., Stewart, M. & Balasubramanian, S. (1998). Technology for micro-array analysis of gene expression. *Current Opinion in Biotechnology* **9**, 609–14.
24. Pass, M. A., Odedra, R. & Batt, R. M. (2000). Multiplex PCRs for identification of *Escherichia coli* virulence genes. *Journal of Clinical Microbiology* **38**, 2001–4.

10

# Microarrays for Bacterial Typing

## *Realistic Hope or Holy Grail?*

Carola Van Ijperen and Nicholas A. Saunders

### Abstract

Microbiology has entered the post-genomic era and it is clear that bacterial typing should aim to be based on analysis of complete genomes. Although complete genome sequencing for epidemiological typing remains unrealistic for the present, microarrays that provide information on gene content are now becoming available. Microarrays comprised of several thousand probes on glass slides can now be manufactured in the laboratory using robotic arrayers. The gene probes are either PCR products or synthetic oligonucleotides that can be irreversibly attached to a reactive glass surface. The target nucleic acids to be hybridized to the probe array are tagged with fluorescent dyes. Relative probe hybridization signals can be measured when two or more different preparations are labeled with distinguishable fluorophores. Microarrays that include probes for every gene within a genome provide excellent comparative data, although a focus on variable genes may be more useful for typing purposes. Composite arrays of variable genes are under development.

### Key Words

Bacterial typing, microarray, probe hybridization, comparative genomics, genetic diversity.

## 1. Introduction

### *1.1. Bacterial Typing*

The traditional methods of typing bacteria are based on phenotypic characteristics such as serotype and phage susceptibility. Although these methods are widely used and remain important, they are gradually being replaced by methods that differentiate strains on the basis of genotypic differences, such as

From: *Methods in Molecular Biology, vol. 266: Genomics, Proteomics, and Clinical Bacteriology*
Edited by: N. Woodford and A. Johnson © Humana Press Inc., Totowa, NJ

pulsed-field gel electrophoresis (PFGE), amplified fragment-length polymorphism (AFLP), and mulilocus sequence typing (MLST; *see* **Chapter 15**). Both genotypic and phenotypic methods can be tailored to give information about strains that may be useful for disease management (e.g., drug resistance markers), but current genotypic methods generally rely on the analysis of anonymous markers that have no known phenotypic correlates. To illustrate this point, one of the features of MLST is that the products of the genes analyzed have essential "housekeeping" functions, and therefore it is highly unlikely that polymorphisms in the open reading frames would have gross phenotypic consequences. A disadvantage of many of the early genotyping methods such as PFGE and restriction fragment-length polymorphisms (RFLP) was that laboratories found it difficult to standardize methods and compare data. Despite this, the genotypic methods are finding favor because they offer greater flexibility and discriminatory ability than the phenotypic methods.

### 1.2. Arrays

The term *array* is applied to a range of different technological platforms. A common feature of all arrays is that they comprise a set of defined nucleic acid probes, each placed at specific *X*, *Y* coordinates on a surface. The probes on the array can be exposed to labeled target nucleic acids, and hybridization occurs if complementary sequences are present. The term *microarray* is usually applied when probes are placed a very small distance apart (approx 200 μm). Robotic laboratory arrayers that exploit split-pin or ring-and-pin technology (**Fig. 1**; *see* **Note 1**) can produce microarrays comprised of >20,000 precisely positioned probes on standard low-fluorescence microscope slides. The construction, handling, and potential applications of this type of microarray are described and discussed in this chapter. High-density microarrays are available commercially and comprise up to 100 to 200 times more probes per unit area than can be achieved using a laboratory arrayer. These arrays are produced using photolithography to build the DNA oligonucleotide probes *in situ* (*see,* for example, http://www.affymetrix.com).

Microarrays represent a powerful new method for analyzing bacterial genotypes. The main advantage of arrays is their flexibility in both the format of the array and the versatility of the probes to test sample material. The great promise of microarray technology lies in the fact that the probes can be used to detect DNA sequences linked to specific phenotypic characteristics. Future arrays might be used to generate typing data for epidemiological purposes linking disease presentation, transmissibility, and changes in the bacterial and host populations.

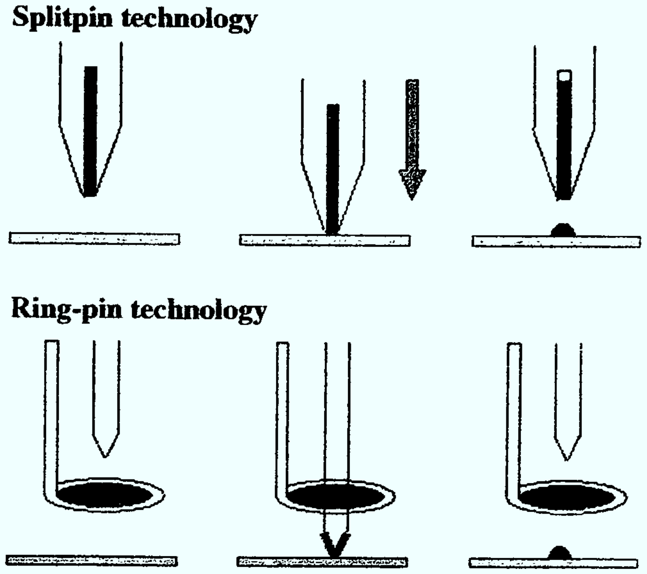

Fig. 1. Schematic diagram of probe deposition onto the slides using split-pin technology (top) and ring-pin technology (bottom).

### 1.3. Comparative Genomics

The number of complete bacterial genome sequences available is expanding rapidly. The technical and logistical problems presented by sequencing a few megabases of DNA appear to have been solved (*see* Chapter 3), principally owing to advances in robotics and the software for assembly of the data. Currently, the major bottleneck appears to be the full annotation of the available sequence data (**Fig. 2**). Many of the major bacterial pathogens that infect humans now have several fully sequenced strains, and precise comparison of these genomes is now possible (for example, *see* Baba et al. *(1)*; *see* Chapter 4). This work illustrates the significance of lateral gene transfer and gene deletions as mechanisms in the evolution of bacterial species. Although genera vary in the degree of genomic plasticity evident, a common feature is the presence of a core of genes that are always present, together with a category of "divergent" genes that are present in only some strains. These genes are usually clustered, indicating that they are lost or acquired as a unit, and are often found in the vicinity of chromosomal elements that are associated with gene translocation (e.g., transposons and bacteriophage components). Comparison of complete genome sequences is likely to remain the most informative and sensitive method for typing bacterial strains. Unfortunately, for the present, genome sequencing remains far from being a practical typing method for most

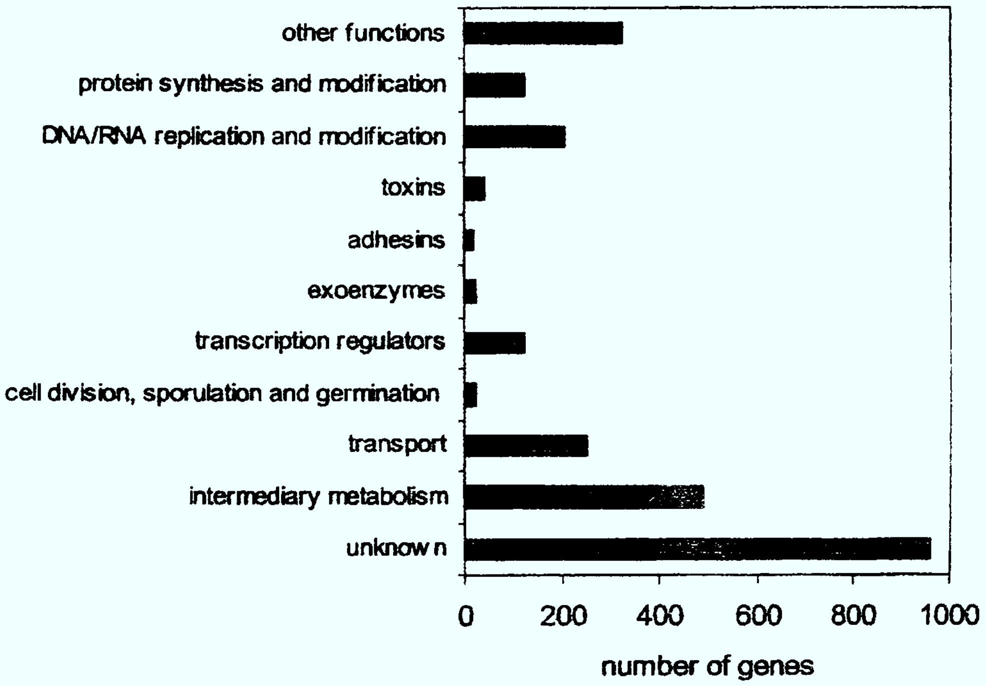

Fig. 2. The functions of the 2595 ORFs of *Staphylococcus aureus* strain N315 are shown. Approximately two-thirds of the ORFs encode proteins with function that are either known or suspected from homology. The data were extracted from Kuroda and colleagues *(18)*.

purposes, because of its high cost. However, the rapid decrease in the cost per base of sequencing shows no current sign of coming to an end. Furthermore, technologies for resequencing (i.e., finding variants in a known sequence) in particular, are very promising in terms of low cost and high capacity. For this reason it seems entirely plausible that in the not-too-distant future, comparison of complete genome sequences will be the primary method used for epidemiological typing. Bacterial gene microarrays use the newly acquired wealth of sequence information to provide a practical shortcut to this ultimate goal of whole-genome comparison. Using arrays, genomes are compared gene by gene instead of base by base. This approach provides data comparable with that of sequencing in terms of the presence and absence of genes, but at a fraction of the current cost.

### *1.4. Arrays for Genomotyping*

*Genomotyping* is a new term *(2)* applied to describe the analysis of bacteria by comparison of their genomes using microarrays. For example, Dorrell and colleagues *(3)* have described the use of arrays comprised of probes derived by amplification of clones produced for sequencing the genome of *Campylobacter jejuni*. Each probe was thus specific for a part of the genome of the strain used to generate the clones. When DNA from different strains was hybridized to the array, it was possible to identify sequences that were present only in the strain used to construct the array, from the pattern of probe hybridization. The advantage of this approach is that it is relatively inexpensive, because a single primer pair (with annealing sites in the cloning vector) can be used to amplify all of the necessary probes. Unfortunately this method has the significant disadvantage that a single probe may include sequences derived from more than one gene and, consequently, the data may be misleading. An alternative method of array construction, which avoids this drawback, is to design probes specific for each gene in a bacterial genome sequence *(4,5)*. Probes for individual genes may be produced by amplification using specific PCR primers, or may be oligonucleotides *(6)*. Computer software for the design of large numbers of PCR primer pairs *(7)* and oligonucleotides *(8)* is available. An array of probes for every open reading frame (ORF) in a bacterial genome will include only a strain-specific subset of the "divergent" genes maintained by the species. Since these are, in some ways, the most interesting genes, most array designs now include additional probes for the "divergent" genes present in other key strains (e.g., *see* Smoot et al.; *5*). These composite arrays give a more sensitive and complete genomotype than arrays based on the genes of a single strain.

### 1.5. Applications in Clinical Bacteriology

The analysis of bacterial genomes using genome-scale microarrays has already contributed to our understanding of the molecular evolution and diversity of human pathogens including *C. jejuni (3), Vibrio cholerae (4), Streptococcus pyogenes (5), Helicobacter pylori (9), Mycobacterium bovis (10),* and *Staphylococcus aureus (11)*. These data are a valuable contribution to our effort to understand the pathogenic mechanisms of these bacteria. It is already clear that human pathogens rely on the appropriately timed expression of a large number of different genes, including those for survival in the host and for transmission between hosts. Diversity in these genes between different strains of a species appears to have developed in response to host defense mechanisms. Typing of certain human pathogens *(11,12)* by analysis of their virulence-associated gene content is likely to be highly discriminatory, but the most significant advantage of this approach may be that it should be possible to correlate gene content with disease transmission or presentation. From a typing viewpoint, there is likely to be a great deal of redundancy in the data provided by genome-scale microarrays; therefore, it should be advantageous to select subsets of informative genes for the arrays. This approach, which has been demonstrated by van Ijperen et al. *(12)*, reduces costs and the time required for data analysis.

### *1.6. Future Developments*

The current microarray format, in which each glass slide may carry several thousand probes, is well suited to hybridizations between target material and genome-scale probe arrays. However, although several arrays of less than 1000 probes can easily be accommodated on a single slide, it is not then a simple matter to hybridize each array to a different target. Systems that allow hybridization of multiple targets to a single slide would be very convenient for typing applications. This equipment is likely to become available as probe sets for typing are developed further.

Genomotyping has been the main focus of bacterial typing work using microarrays. However, an alternative would be to develop arrays of probes each complementary to a different amplified fragment made by the AFLP method. Hybridization of the array to targets composed of AFLP fragments from an individual strain would result in a strain-specific pattern of reactions *(13)*. The discriminatory ability of arrays of AFLPs would depend upon the choice of probes used, and could be very high. In contrast to the genomotyping approach, the probes used would be selected on the basis of their contribution to the discriminatory ability of the typing method. In AFLP, the absence of a fragment in the amplified material shows only that one of the restriction sites is absent. This may indicate that a single base mutation has occurred in one of the sites or that the sequence is entirely absent.

## 2. Materials

### *2.1. Slides*

Microarrays are usually printed onto glass slides coated with reactive groups. Slides with different coatings are available and are used in variations of the microarray method. There are two classes of slides in common use, and these are most suitable for spotting either PCR products or oligonucleotides. Slides for spotting PCR products are usually coated with polylysine and aminosilane; these surfaces bind unmodified DNA covalently via negatively charged phosphate groups (**Fig. 3A**). The second class of slides, which are modified with aldehyde (**Fig. 3B**) and epoxy (**Fig. 3C**) groups, are usually recommended for printing oligonucleotides that have a free aminolinker. The amino group of the modified oligonucleotide binds to the active groups on the slide. This has the advantage that all of the molecules are bound in the same place and have the same orientation on the slide. This is important for relatively short oligonucleotide sequences, since steric hindrance can significantly affect the specificity

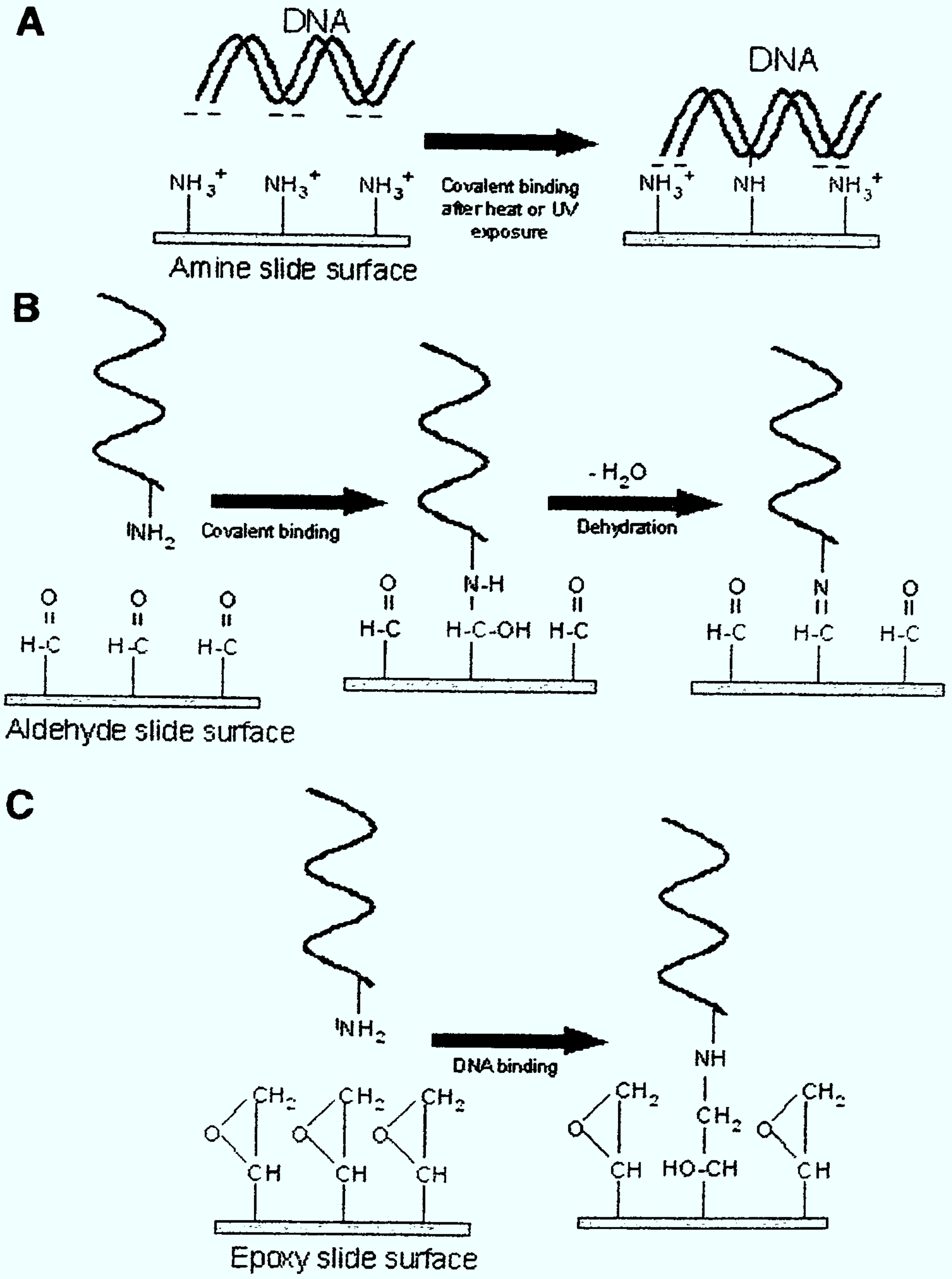

Fig. 3. Binding of DNA onto different slide surfaces. (**A**) Binding of DNA to amine slide in which positive amine groups covalently bind negatively charged DNA phosphate backbone. Binding of amino-linked oligonucleotides to (**B**) aldehyde and (**C**) epoxy slide surfaces by an active binding of the free amino group.

and stability of probe/target hybridization. Although nonmodified oligonucleotides can be covalently linked to epoxy slides, longer probes (>50-mer) are recommended *(14)*, and care should be taken to obtain good crosslinking by UV treatment or baking *(15)*.

### 2.2. Probes for Printing

In microarray technology, the PCR products or oligonucleotides spotted onto the slide are designated *probes*. Amplified probes are usually produced by PCR using gene-specific primers, with the genomic DNA of the organism of interest as substrate. Each PCR product represents a specified ORF and may either be an ORF-specific fragment or the complete ORF. Following amplification, the probes are checked for purity and identity. The most common approach is to analyze the products of each PCR by agarose-gel electrophoresis to check for the presence of a single amplicon of the expected molecular size. A proportion of products that varies depending on the resources available, are then sequenced to provide a positive identity check prior to arraying. When oligonucleotides are used as probes, most of these steps are unnecessary. Oligonucleotides ordered from commercial companies are purified and ready to print as soon as they arrive in the laboratory. This saves the time and resources required to run PCR reactions, analyze the products, and redesign any PCRs that do not give the expected product. However, short oligonucleotides are best bound to the glass surface via an aminolink group, and this increases the cost of synthesis. Depending on the length, oligonucleotides have the advantage of giving more specific hybridization signals. The longer PCR-amplified products, however, might detect similar genes with partial homology to the gene of interest, which may be useful for the detection of gene homologs.

The probes are dissolved in solutions that interact physically with the slide surface via the depositing device to give aliquots of consistent volume that dry as round spots of equal size and shape. The spotting solution must also be of suitable volatility to ensure rapid drying of the deposited spots while preventing excessive evaporation of the sample during the printing process. This is also influenced by the local temperature and humidity. A common spotting solution is 50% dimethyl sulfoxide (DMSO), which is normally used on nonactivated surfaces. However, DMSO can interfere with some active groups on the slide surface, and it is less suitable for the printing oligonucleotides than alternatives such as saline–sodium citrate (SSC) solution. A wide range of commercial spotting buffers are also available. Changes in the volumes of the probe solutions dispensed into microtiter plate wells for spotting inevitably occur over time, even when care is taken to avoid the evaporation of solvent. To preserve valuable probes, plates affected by evaporation may be dried down fully and the probes redissolved in the original volume of solvent. It is usually unnecessary to adjust the volume for probe used in a previous print run.

### 2.3. Target

In the usual nomenclature, the "target" is the nucleic acid that is hybridized to the array. A wide range of target formulations are used. The well-known application of microarrays to gene-expression studies uses labeled cDNA made from mRNA extracts by reverse transcription *(16)*. Multiplex PCR products may be used *(17)*, and genomic DNA may be most appropriate for typing or screening for the presence or absence of genes.

Target nucleic acids are labeled with fluorescent dyes by either direct or indirect incorporation methods. The most common dyes used are cyanine three (Cy3, green) and cyanine five (Cy5, red). These two dyes have different absorption and emission spectra and can therefore be used in the same hybridization experiment. The control and a test sample are each labeled with a different dye, then mixed and hybridized to the array. The hybridization signal is detected with a high resolution scanner. The result at each probe is either a green, red, or yellow (mixed) signal. Green or red signals indicate that only one of the targets has hybridized to the probe, and yellow shows that both control and test sample have hybridized.

Direct incorporation of the Cy dyes by some polymerases is not always efficient due to steric hindrance by the large Cy groups. The degree of inhibition depends upon the structure of the Cy dye, and this may cause inaccuracies in experiments in which the ratio of two dyes is measured. Alternative indirect labeling methods have been developed using amino-allyl-modified dNTPs. The small amino-allyl dNTPs are incorporated into the target nucleic acid at a rate similar to that of unmodified nucleotides, and Cy dye is then bound to the target DNA by ester bonding. This approach gives good yields of highly labeled target, and thus minimizes experimental variation. Furthermore, as all Cy dyes can be incorporated equally, the normalization of signal intensities is less problematic.

### 2.4. Buffers

During the three stages of a microarray experiment, various buffers are used. The first step is to treat the slide with a blocking or prehybridization buffer before addition of the target, the purpose being to prevent nonspecific binding by blocking remaining sites on the glass surface that could interact with nucleic acids. Blocking buffers usually contain bovine serum albumin (BSA), sodium dodecyl sulphate (SDS), and SSC, but, depending on the application, competitor DNA may also be added. For example, addition of $C_0t$-1 DNA to the buffer reduces the hybridization of repetitive elements by binding to the labeled repetitive target sequences, while tRNA acts as a blocker of nonspecific hybridization. The next stage is the hybridization of target to the array. Various commercial hybridization buffers are available, but in-house buffers are also widely used. The use of formamide in the hybridization buffer has the advantage that the DNA remains denatured at lower temperatures. Hybridization at a lower temperature results in reduced evaporation from the target solution during incubation; evaporation leads to high background signals in locations where target has dried. Following hybridization, several washes are carried out to remove nonspecifically bound target. The stringency of the post-hybridization washes affects the strength of the hydrogen bonding between the probe and target. Washing stringency can be adjusted by changing the temperature and SSC concentration (stringency is raised at higher temperatures and lower salt concentrations). SDS is autofluorescent and should be omitted from the buffer used for the final post-hybridization wash.

### *2.5. Hybridization*

Water-tight chambers (Corning, Schiphol-Rijk, The Netherlands; and Genetix Limited, New Milton, UK) that fit the microarray slide exactly are used for hybridization. The chambers prevent the slide from drying out when used within a hybridization oven, but they can also be used in a water bath. Hybridization takes place under a coverslip so that a small amount of target is evenly spread over the array. Automatic hybridization stations are also commercially available. In these systems, hybridization takes place in a low-volume chamber with access ports for the addition of buffers. An advantage of this equipment is that the target solution can mix more freely than under a coverslip so the target has a greater opportunity to interact with the probe array; this increases signal intensities and ensures that all positions on the array are equally exposed.

### *2.6. Scanners and Software*

A variety of fluorescent scanners are now available for microarray work. The machines generally include a high-resolution confocal microscope scanner (10 to 4 μm), and have lasers with the appropriate wavelengths as light sources. The standard lasers excite Cy3- (532 nm) and Cy5- (635 nm) labeled targets, but up to six different lasers are now included with recent designs. Scanners generally come with user-friendly operation software and are therefore simple to use. The Affymetrix 428™, the Axon Genepix 4000®, and the PerkinElmer ScanArray™ are most commonly used.

Similarly, a variety of software packages are available for microarray image and data analysis. Although there are differences, most programs have basic functions to allow measurement of signal intensities and to present the data in convenient formats (i.e., histograms, scatterplots, and so on). The complexity of microarray data presents great challenges to the analysis software. The available programs are under continuous development, and new upgrades contain more and better tools for normalization, analysis, and statistical analysis. Any microarray software listing will be rapidly outdated due to the frequent introduction of new programs. Software should be selected to fulfil the needs of the individual user. Bacterial comparative genomics is a relatively simple application, and consequently the major software consideration is likely to be ease of use.

## 3. Methods

### *3.1. Probe Preparation and Microarray Printing*

Probes often comprise specific parts of ORFs that are amplified by PCR and then purified before printing. In the case of oligonucleotide microarrays, the probes are obtained from custom synthesis houses and may be aminolinker-modified depending on the application and slide coating employed. The probes spotted onto the slides must be very specific for the detection of the target gene or genes, since cross-hybridizations lead to array results that are very difficult to interpret. The steps required to prepare oligonucleotide or amplified probes are summarized in **Note 2**.

Microarrays are usually spotted using automated robotic arrayers, although manual instruments are available that may be suitable for producing arrays comprised of relatively few probes. Optimal environmental conditions for printing are between 40% and 50% humidity at around 20°C. These conditions prevent spots from drying too quickly. Handling of the slides after printing depends on the slide coating. Crosslinking by UV illumination is recommended for some slides, but quoted energy values vary and convincing optimization data are not available. Baking at 80°C for up to 2 h with occasional reintroduction of water by "fogging" the slide over warm water is recommended for other slides. Some protocols call for both UV and heat treatments, while others recommend curing at room temperature. Currently, most authors and slide manufacturers take an empirical approach to probe attachment, and a wide variety of conditions are satisfactory.

### *3.2. Hybridization and Posthybridization Processing of Slides*

For genomotyping experiments where the aim is to show the presence or absence of genes within the genome of a particular bacterial strain, the target is comprised of labeled total DNA. Cy dyes can be incorporated directly into representative DNA by primer extension using a random priming kit (Invitrogen. Paisley, UK). Purification of the target prior to hybridization is very important. For each hybridization on a 22 mm × 22 mm surface, 15 μL of probe is needed. The procedure is summarized in **Note 3**. Labeling of nucleic acid using direct methods is inefficient for some combinations of enzyme and modified nucleotide. Cy dyes in particular have large 3D structure and are therefore relatively inefficiently incorporated into DNA. Indirect labeling methods have been developed that use relatively small amino-allyl dNTPs as substrate for the enzymatic reaction (*see* **Subheading 2.3.**). These reactive primary amines react with NHS ester-modified Cy dyes in a nonenzymatic second step.

### 3.3. Analysis

Currently, the most time-consuming part of microarray experiments is data analysis. The reason for this is that many factors must be considered before any conclusions can be drawn from the raw data. The first steps in data analysis are required for normalization of the images. Standard methods of normalization calculate either the mean or median intensity for a microarray image. This value is then used to correct the intensity of each spot. The mean or median of all corrected spot intensities is finally adjusted to center around one, and then converted to $\log_{10}$ form so that the points follow a normal distribution centered around zero. Using logarithmic transformed data will help to normalize the distribution of the data and improves the applicability of statistical tests. Most statistical tests work with differences, although the scientists are more interested in a fold difference. This problem is solved by working with a log ratio, as $\log a - \log b = \log a/b$. The next step in analysis is to detect outliers. If the intensity of a spot is influenced by the strong fluorescence of a dust particle on the slide, this spot should be excluded from further analysis. In this way the data are reduced to a series of values that are in the form of ratios when different targets, each labeled with a distinct fluor, have been hybridized to the array. At this point the question to be considered is whether any differences observed are reproducible or significant. To answer this question, duplicate experiments are required, which in turn generate yet more data to analyze.

## 4. Bacterial Typing by Microarrays: Realistic Goal or Holy Grail?

Microarrays for typing is clearly not a holy grail, i.e., the objective of a quest that is unlikely to be realized without divine intervention. However, current microarray technology is not well-suited to high-throughput typing of many different strains. Considerable hands-on time is required for each sample, and the analysis can be time-consuming, especially while setting up a new application. At present, genomotyping using microarrays should be viewed as a complementary method that can provide large quantities of gene-content data and is appropriate in that context. These comparative genomic data allow the construction of detailed overviews of the relationships between strains and are extremely valuable. We anticipate rapid progress in the development of protocols and automation for microarray hybridization. These developments are likely to lead to the use of genomotyping as a routine tool for bacterial typing. This is certainly a realistic goal.

## 5. Notes

1. Array spotting technology. Various techniques for spotting small volumes of probe solutions have been described. The most common method of array printing currently is the use of capillary reservoir pins (split pins), which deposit picoliter volumes of probe onto the glass surface. The ring-pin system is similar in that pins are used to deposit the probe solution, but a loop filled with probe solution acts as the reservoir with the pin passing through the probe meniscus each time a spot is printed. Another method uses an inkjet-type spotter that fires the probe onto the surface without touching the surface.
2. Probe preparation. Probe preparation by PCR amplification includes the following steps: First, specific oligonucleotide primers have to be designed using the appropriate software, followed by the PCR amplification of the product. After removing the enzyme and unincorporated nucleotides, the identity of the products can be verified by size and/or sequencing. Finally, the concentrations are adjusted where necessary and probes are diluted in the appropriate spotting solution.

   If probes are prepared by oligonucleotide synthesis, fewer steps are involved. Oligonucleotides are designed using specialized software and bioinformatics knowledge, and users decide on the basis of the length of the oligonucleotide whether an aminolink is necessary. Oligonucleotides are produced synthetically by commercial companies. After adjustment of the concentration, probes are ready for printing.
3. Genomic DNA hybridization on microarrays. When genomic DNA is used for the preparation of target material, up to 2 μg can be labeled directly using Cy-labeled nucleotides and Klenow enzyme. Removing enzyme and unincorporated nucleotides is crucial for a low background signal after hybridization. The target is dried under vacuum and dissolved in 15 μL of hybridization buffer. After denaturing, the target can be applied to the slide, covered with a disposable coverslip, and incubated in a chamber at the appropriate temperature. After hybridization, slides are washed several times in 2X SSC buffer containing 0.1% SDS, and finally in 2X SSC buffer. Slides are dried by centrifugation in 50-mL tubes.

## References

1. Baba, T., Takeuchi, F., Kuroda, M., Yuzawa, H., Aoki, K., Oguchi, A., et al. (2002) Genome and virulence determinants of high virulence community-acquired MRSA. *Lancet* **359,** 1819–1827

2. Lucchini, S., Thompson, A., and Hinton, J. C. D. (2001) Microarrays for microbiologists. *Microbiol.* **147,** 1403–1414.
3. Dorrell, N., Mangan, J. A., Laing, K. G., Hinds, J., Linton, D., Al-Ghusein, H., et al. (2001) Whole genome comparison of *Campylobacter jejuni* human isolates using a low-cost microarray reveals extensive genetic diversity. *Genome Res.* **11,** 1706–1715.
4. Dziejman, M., Balon, E., Boyd, D., Fraser, C. M., Heidelberg, J. F., and Mekalanos J. J. (2002) Comparative genomic analysis of *Vibrio cholerae*: genes that correlate with cholera endemic and pandemic disease. *Proc. Natl. Acad. Sci. USA* **99,** 1556–1561.
5. Smoot, J. C., Barbian, K. D., Van Gompel, J. J., Smoot, L. M., Chaussee, M. S., Sylva, G. L., et al. (2002) Genome sequence and comparative microarray analysis of serotype M18 group A Streptococcus strains associated with acute rheumatic fever outbreaks. *Proc. Natl. Acad. Sci. USA* **99,** 4668–4673.
6. Relógio, A., Schwager, C., Richter, A., Ansorge, W., and Valcárcel, J. (2002) Optimization of oligonucleotide-based DNA microarrays. *Nucl. Acids Res.* **30,** e51.
7. Rozen, S. and Skaletsky, H. J. (1996–1998) Primer3. Code available at http://www-genome.wi.mit.edu/genome_software/other/primer3.html
8. Rouillard, J-M., Herbert, C. J., and Zuker, M. (2002) OligoArray: genome-scale oligonucleotide design for microarrays. *Bioinformatics* **18,** 486–487.
9. Salama, N., Guillemin, K., McDaniel, T. K., Sherlock, G., Tompkins, L., and Falkow, S. (2000) A whole-genome microarray reveals genetic diversity among *Helicobacter pylori* strains. *Proc. Natl. Acad. Sci. USA* **97,** 14,668–14,673.
10. Behr, M. A., Wilson, M. A., Gill, W. P., Salamon, H., Schoolnik, G. K., Rane, S., et al. (1999) Comparative genomics of BCG vaccines by whole-genome DNA microarray. *Science* **284,** 1520–1523.
11. Fitzgerald, J. R., Sturdevant, D. E., Mackie, S. M., Gill, S. R., and Musser, J. M. (2001) Evolutionary genomics of *Staphylococcus aureus:* insights into the origin of methicillin-resistant strains and the toxic shock syndrome epidemic. *Proc. Natl. Acad. Sci. USA* **98,** 8821–8826.
12. Van Ijperen, C., Kuhnert, P., Frey, J., and Clewley, J. P. (2002). Virulence typing of *Escherichia coli* using microarrays. *Mol. Cell Probes* **16,** 371–378.
13. Hu, H., Lan, R., and Reeves, P. R. (2002) Fluorescent amplified fragment length polymorphism analysis of *Salmonella enterica* serovar *Typhimurium* reveals phage-type-specific markers and potential for microarray typing. *J. Clin. Microbiol.* **40,** 3406–3415.
14. Kane, M. D., Jatkoe, T. A., Stumpf, C. R., Lu, J., Thomas, J. D., and Madore, S. J. (2000). Assessment of the sensitivity and specificity of oligonucleotide (50mer) microarrays. *Nucl. Acids Res.* **28,** 4552–4557.
15. Massimi, A., Harris, T., Childs, G., and Somerville, S. (2003) *DNA Microarray: A Molecular Cloning Manual.* (Bowtell, D. and Sambrook, J., eds.), Cold Spring Harbor Laboratory, Cold Spring Harbor, New York. p. 78.
16. Watson, A., Mazumder, A., Stewart, M., and Balasubramanian, S. (1998) Tech-

nology for microarray analysis of gene expression. *Curr. Opin. Biotechnol.* **9,** 609–614.

17. Chizhikov, V., Rasooly, A., Chumakov, K., and Levy, D. D. (2001) Microarray analysis of microbial virulence factors. *Appl. Environ. Microbiol.* **67,** 3258–3263.
18. Kuroda, M., Ohta, T., Uchiyama, I., Baba, T., Yuzawa, H., Kobayashi, I., et al. (2001) Whole genome sequencing of meticillin-resistant *Staphylococcus aureus*. *Lancet* **357,** 1225–1240.