DNA Microarrays in Comparative Genomics and Transcriptomics

Willenbrock, Hanni; Wassenaar, Gertrude Maria; Ussery, David; Lund, Ole

Publication date:
2007

Document Version
Publisher's PDF, also known as Version of record

PhD Thesis:

DNA Microarrays in Comparative Genomics and Transcriptomics

Hanni Willenbrock

September 2006

Center for Biological Sequence Analysis
BioCentrum
Technical University of Denmark
DK-2800 Kgs. Lyngby
Denmark

Abstract

During the past few years, innovations in the DNA sequencing technology has led to an explosion in available DNA sequence information. This has revolutionized biological research and promoted the development of high throughput analysis methods that can take advantage of the vast amount of sequence data. For this, the DNA microarray technology has gained enormous popularity due to its ability to measure the presence or the activity of thousands of genes simultaneously.

Microarrays for high throughput data analyses are not limited to a few organisms but may be applied to everything from bacteria to higher Eukaryotes and new applications are constantly being reported. In this PhD thesis, various applications for DNA microarrays are explored. Consequently, research results are presented where the use of microarray data has been essential. The thesis comprises three main topics: gene expression analysis, analysis of chromosomal aberrations and DNA sequence dependent gene expression.

First, this thesis contains a description of how the gene expression profiles from children with acute lymphoblastic leukemia may be used to improve the diagnosis of these patients and potentially improve their treatment. Next, a new method is presented that utilizes a large repository of gene expression microarray data to derive functional associations between for instance a mutant and a compendium of gene expression responses. By this approach, an extensive functional characterization of a given mutant or experimental factor such as compound treatment may be obtained. The same characterization could otherwise be time consuming and require an extensive biological knowledge of the investigated biological system.

Often, solid tumors are characterized by a multitude of chromosomal aberrations where parts of the chromosomes have either been lost or additional copies might have been gained. By targeting microarrays at chromosomal DNA, it is possible to measure the so-called DNA copy number and thereby obtain a DNA copy number profile of each chromosome. Numerous analysis methods have been published that aims at identifying the exact breakpoints where DNA has been gained or lost. In this thesis, three popular methods are compared and a realistic simulation model is presented for generating artificial data with known breakpoints and known DNA copy number. By using simulated data, we obtain a realistic evaluation of each method's ability to analyze DNA copy number data. Moreover, our study shows that analysis methods developed for cancer research may also successfully be applied to DNA copy number profiles from bacterial genomes. However, here the purpose is to characterize variations in the gene content of various strains of the bacteria, e.g. *Escherichia coli*, with regard to genes involved in pathogenesis.

Finally, this thesis present results demonstrating that the gene expression level is sequence dependent, that is, it depends on both DNA structure and codon usage bias. Here, microarray data was used to verify predictions of highly expressed genes. Moreover, the codon bias of microbial genomes was found to constitute an environmental signature. For example, soil bacteria have very similar codon bias.

Resumé

Inden for de sidste få år har store fremskridt i udviklingen af DNA-sekventeringsteknologien medført en eksplosion i tilgængelig DNA-sekvensinformation. Dette har revolutioneret den biologiske forskning og fremmet udviklingen af 'high-throughput' analysemetoder, som kan drage fordel af disse svimlende mængder af sekvensdata. I denne forbindelse har DNA-microarray-teknologien vundet enorm popularitet pga. dens evne til simultant at måle tilstedeværelsen eller aktiviteten af tusindvis af gener.

Microarrays til brug for high-throughput dataanalyser er ikke begrænset til enkelte organismer, men kan bruges på alt fra bakterier til højere Eukaryoter og teknologien finder derfor hele tiden nye anvendelsesmuligheder. I denne Ph.d.-afhandling udforskes forskellige anvendelsesmuligheder for DNA microarrays. Således præsenteres forskningsresultater hvor microarraydata har spillet en væsentlig rolle. Afhandlingen omfatter tre hovedemner: genekspressions-analyse, analyse af kromosomale afvigelser og genekspressionens afhængighed af DNA-sekvensen.

Først vises hvorledes genekspressions-profilerne fra børn med akut lymfocytisk leukæmi kan benyttes til at forbedre diagnosen af disse patienter og på sigt potentielt forbedre deres behandling. Dernæst præsenteres en ny metode, der udnytter de store mængder af offentlig tilgængeligt genekspressions-microarray-data til at finde funktionelle associationer imellem f.eks. en mutant og et kompendium af genekspressionsresponser. Denne metode viser sin anvendelighed både for bagegær og planten Arabidopsis thaliana (gåsemad). Med denne fremgangsmåde opnås en omfattende funktionel karakterisering af en given mutant; en karakterisering som ellers kan være både tidskrævende og afhænge af en stor biologisk baggrundsviden af det pågældende system.

Mange kræftformer er karakteriseret ved forskellige kromosomale afvigelser, hvor dele af kromosomerne er enten gået tabt eller blevet multipliceret. Vha. microarrays rettet mod kromosomalt DNA er det muligt at måle det såkaldte DNA kopital og derved opnå en DNA-kopitals-profil af de enkelte kromosomer. Rigtig mange analysemetoder er blevet publiceret til at analysere denne type data og derved identificere de præcise brudpunkter, hvor DNA er enten gået tabt eller blevet duplikeret. Vi sammenligner 3 populære metoder og præsenterer samtidig en virkelighedstro simuleringsmodel til at generere data med kendte brudpunkter og kendte DNA kopital. Ved brug af simuleret data opnår vi en realistisk evaluering af de forskellige metoders evne til at analysere DNA-kopitals-data. Vores studier viser desuden at analysemetoder udviklet til kræftforskningen også kan bruges til at analysere DNA-kopitals-profiler fra bakterielle genomer. Her er målet dog at karakterisere variationer i tilstedeværende genmateriale for forskellige stammer af den samme bakterie, f.eks. *Escherichia coli*, bl.a. med hensyn til gener involveret i patogenese.

Endelig indeholder denne afhandling studier af hvorledes DNA-sekvensen i form af strukturelle egenskaber og foretrukne codons har indflydelse på genekspressionsniveauet. Her blev microarraydata benyttet til at verificere forudsigelser om højtudtrykte gener. Desuden viste mønsteret for mikroorganismers foretrukne codons at være en vigtig faktor for deres foretrukne miljø. Således har f.eks. tarmbakterier meget ens codon præferencer.

Preface

This Ph.D. thesis is written for BioCentrum, The Technical University of Denmark, under the Biotechnology program. The majority of the research has been done at the Center for Biological Sequence Analysis at BioCentrum supervised by associate professor Steen Knudsen (1st year) and associate professor David W. Ussery (2nd and 3rd year), and co-supervised by assistant professor Henrik Bjørn Nielsen. Part of the research was done at the University of California at San Francisco (UCSF) supervised by assistant professor Jane Fridlyand. The project has been finance by a PhD scholarship from the Technical University of Denmark.

This thesis will provide an introduction to the microarray technology and data analysis, followed by three parts describing different applications of the microarray technology and its usage. These three parts comprise a total of seven papers including the following topics: microarrays for gene expression analysis, microarrays for comparative genomics and use of microarray data for estimating sequence dependent gene expression.

Hanni Willenbrock, September 2006

Acknowledgements

I wish to thank everyone who has helped me during the past 3 years and made it such a pleasurable experience. I would also like to express my deepest gratitude to everybody at CBS for making it such a great place to work. In particular, I wish to thank:

- My first supervisor, Steen Knudsen, for getting me started.

- My last supervisor, David W. Ussery, for welcoming me into his group when my previous supervisor sought new challenges elsewhere, for the inspiring discussions, and the great taco events at his home.

- Jane Fridlyand for taking such good care of me while I was in San Francisco.

- My co-supervisor, Henrik Bjørn Nielsen, for always being very helpful and willing to discuss ideas and technical questions.

- Past and present members of the microarray group, Laurent Gautier, Carsten Friis, Hanne Jarmer and Chris Workman for always being willing to help and share their knowledge.

- Agnieszka S. Juncker for everything from general project discussions to advanced discussions about luck and the meaning of our lives, for someone to laugh and cry with, and for being such a good friend.

- Members of the Genome Atlas group, especially Peter Hallin for his persistent help with database issues, 'make rules' and general programming tasks.

- Past and present office mates for great conversation and for making me want to come to work even when things were not going that well project wise.

- All co-authors on the papers included in this thesis: Agnieszka S. Juncker, Kjeld Schmiegelow, Steen Knudsen, and Lars P. Ryder, Henrik Bjørn Nielsen, Jane Fridlyand, Anne Petersen, Camilla Sekse, Kristoffer Kiil, Yngvild Wasteson, David W. Ussery, Carsten Friis. Without you, this thesis would not have been possible.

- My beloved husband, Claus Thomsen, for always being there for me.

- Feedback and proof-reading team: Henrik Bjørn Nielsen, Claus Thomsen, Chris Workman, and David W. Ussery.

Publications

Paper I:

Hanni Willenbrock, Agnieszka S. Juncker, Kjeld Schmiegelow, Steen Knudsen, and Lars P. Ryder. Prediction of immunophenotype, treatment response, and relapse in childhood acute lymphoblastic leukemia using DNA microarrays. *Leukemia* 2004 18(7): 1270-7.

Paper II:

Henrik Bjørn Nielsen and **Hanni Willenbrock**. Functional Association by Response Overlap (FARO). *Manuscript in preparation.*

Paper III:

Hanni Willenbrock and Jane Fridlyand. A comparison study: applying segmentation to array CGH data for downstream analyses. *Bioinformatics* 2005; 21(22): 4084-91.

Paper IV:

Hanni Willenbrock, Anne Petersen, Camilla Sekse, Kristoffer Kiil, Yngvild Wasteson, and David W. Ussery. Design of a Seven-Genome *Escherichia coli* Microarray for Comparative Genomic Profiling. *Journal of Bacteriology*. Epub Sept 8, 2006. Doi:10.1128/JB.01043-06.

Paper V:

Hanni Willenbrock and David W. Ussery. Chromatin architecture and gene expression in Escherichia coli. *Genome Biology* 2004;5(12):252.

Paper VI:

Hanni Willenbrock, Carsten Friis, Agnieszka S. Juncker, and David W. Ussery. An Environmental Signature for 323 Microbial Genomes based on Codon Adaptation Indices. *Submitted to Genome Biology (July 2006).*

Paper VII:

Hanni Willenbrock and David W. Ussery. Chromatin dependent gene expression in microbes. Submitted to *Nucleic Acids Research (September 2006).*

Table of Contents

Part I

INTRODUCTION

The DNA Microarray Revolution

Over the past few years, innovations in DNA-sequencing technology has led to an explosion in DNA sequence information resulting in availability of sequences from more than 300 bacterial genomes and about 30 archaeal genomes, in addition to extensive nucleotide sequences information for higher eukaryotes, including human (International Human Genome Sequencing Consortium, 2004), mouse (Nadeau*, et al.*, 2001) and fly (Adams*, et al.*, 2000).

This has revolutionized biological research. The explosion in nucleotide sequence information allows for the comparison of genetic information between individuals or the analysis of gene expression including possible splice-variants to provide detailed disease patterns, and possibly tailoring treatment. Moreover, the vast amount of sequence data may be searched for industrial purposes, to identify enzymes able to, for example, break down oil pollutants or to synthesize chemicals with less stress on the environment.

This sudden explosion in available sequence information data has promoted the development of high-throughput genetic approaches for analyzing and utilizing vast amount of sequence data, including computational approaches to comparative genomics and experimental approaches such as the microarray technology.

During the past few years, the DNA microarray technology has become popular both among the scientific community and in the industry due to its ability to simultaneously measure the presence, the activities in term of gene expression and the interactions of thousands of genes, thus, providing new insights into the mechanisms of living systems. In fact, no other methodological approach has transformed biological research more in the recent years. With the microarray technology, researches are no longer restricted to studies of individual biological functions of a few related genes. Consequently, microarrays have been applied in a vast range of biological studies and have immediately yielded new and interesting biological insight.

Nonetheless, due to the large volumes of data generated, the analysis of microarray data is far from trivial and in many cases, advanced statistics are required. In the following three introductory chapters, microarray technology will be presented and relevant analysis approaches will be introduced. Examples of the application of the technology are then given in the subsequent chapters, including classification of childhood acute lymphoblastic leukemia in Chapter 4, prediction of functional associations by response overlap (FARO) in Chapter 5, analysis of genomic DNA variations in Chapter 6 and Chapter 7, and analysis of sequence dependent gene expression in Chapter 8, Chapter 9 and Chapter 10.

Chapter 1 DNA Microarrays

Microarray-based methods for high-throughput monitoring of gene expression was first described in 1995 (Schena, *et al.*, 1995), although the idea date all the way back to the discovery of DNA hybridization in the 1960s (Marmur and Doty, 1961), the invention of the blotting technology in the 1970s (Southern, 1975), and the suggested potential of array technology in genomics in the 1980s (Poustka, *et al.*, 1986). During the past decade, the technology has rapidly developed into a complex field comprising both genomics, transcriptomics, informatics and advanced statistics.

The microarray technique is used in a wide variety of applications like gene expression analysis (Schena, *et al.*, 1995; Schena, *et al.*, 1996), including analysis of gene expression profiles from cancer (DeRisi, *et al.*, 1996), single nucleotide polymorphism (Cutler, *et al.*, 2001), splice-variant analysis, identification of unknown exons (Hoheisel, 2006), and analysis of DNA-protein interactions (Bulyk, *et al.*, 1999).

Microarrays directed at the genome sequence have been widely used for comparative genomics, to identify differences in gene content such as changes in DNA copy number and chromosomal aberrations often found in cancer and developmental abnormalities including mental retardation (Albertson, *et al.*, 2003; Menten, *et al.*, 2006; Vissers, *et al.*, 2003) or for complex mutations in human disease genes (see for example Chapter 6). Recently, developments in the technology have allowed the analysis of chromosomal imbalances in a single cell (Le Caignec, *et al.*, 2006).

In particular, development of solid tumors is associated with acquisition of complex genetic alterations. Consequently, microarray methods may be employed to extensively map cancer genomes and detect chromosomal aberrations. Moreover, by using the same arrays for DNA copy number analysis and expression analysis, it is possible to assess the relationship of mRNA expression levels to DNA copy numbers broadly across the genome (Pollack, *et al.*, 2002).

Microarray approaches are also useful in microbial comparative genomics and have been used to detect variations in the baseline sequence, such as in emerging pathogenic strains (Anjum, *et al.*, 2003; Fukiya, *et al.*, 2004; Winterberg, *et al.*, 2005) and to detect horizontal gene transfer (Fitzgerald, *et al.*, 2001). Due to their much lower complexity than mammalian genomes, it is – in fact - easier to obtain copy number information from bacteria. Because the concentration of each portion of the genome in the hybridization mixture is relatively higher, the corresponding signals will also be higher and easier discernible (see for example Chapter 7).

1.1 The Technology

A DNA microarray is a high-density array of known single stranded DNA (ssDNA) sequences attached to a solid surface. These sequences are called probes and complementary single stranded sample sequences, so-called targets, can be hybridized to these probes (see Figure 1-1). By labeling the targets with fluorescence or radioactivity, the amount of hybridized target can be measured. Due to their small dimensions, a vast number of targets might be measured much faster using DNA microarrays than by traditional gel-based analysis methods. Moreover, only very small sample volumes are required which is an important feature when dealing with expensive or limited material.

Two general methods exist for manufacturing DNA microarrays based on two different strategies for immobilizing DNA onto a chip. The first one is spotted microarrays, where pre-synthesized DNA is immobilized onto a substrate surface, and the second method is 'in situ' synthesis of DNA directly on a substrate surface.

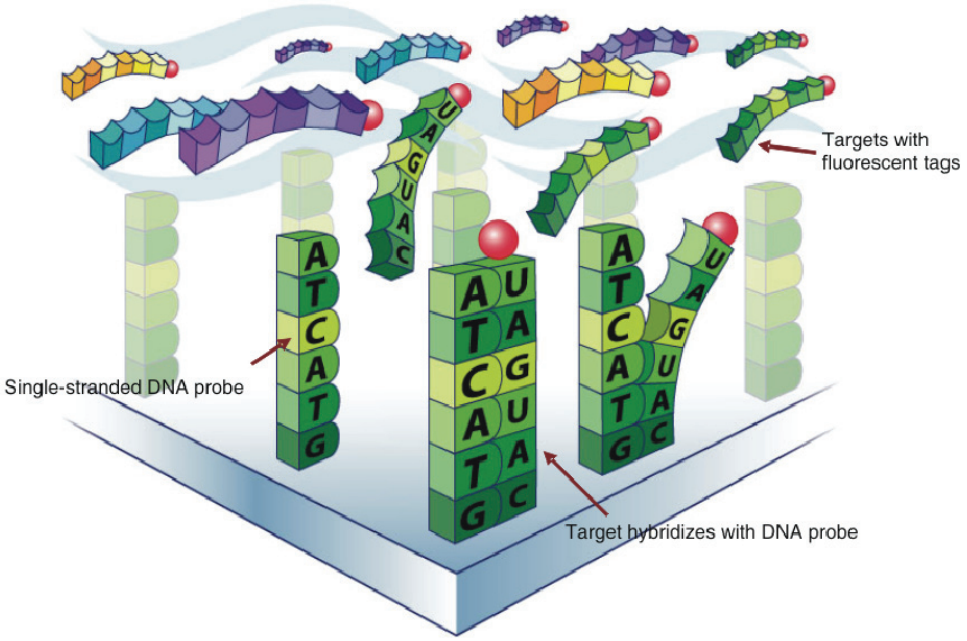

Figure 1-1. Cartoon illustration of DNA microarray probes and target hybridization (source: Wellcome Trust Centre for Human Genetics, http://www.well.ox.ac.uk/).

Owing to their flexibility and value, mechanically spotted microarrays have been the most popular platform. However, with the recent advances in flexibility for customizing 'in situ' synthesized DNA microarrays, this platform has become increasingly popular. An overview of both technologies is given in the following.

Spotted Microarrays are often known as Stanford cDNA arrays. However, the term 'Stanford array' refers specifically to the array being developed at Stanford University in the 90's (Schena*, et al.*, 1995; Schena*, et al.*, 1996), while 'spotted arrays' includes any array being fabricated using a spotter. In this type of arrays, a robot is used to move small quantities of probes in solution from a microtiter plate to the surface of a glass slide. Here, probes may be cDNA, PCR-products or synthetic oligonucleotides. The probes may be immobilized onto a substrate surface by several means, and the binding can be either covalent, including co-polymerisation or non-covalent, including non-covalent charge interactions (Auburn*, et al.*, 2005).

Besides their great flexibility in customization of the microarray content, another advantage of spotted arrays is that they usually allow for hybridization of two samples simultaneously to each slide by labeling targets from the first sample with green fluorescent dye (Cy3) and targets from the second sample with red fluorescent dye (Cy5). Consequently, resulting signal ratios are not dependent on the hybridization efficiency of individual probes but only on the relative amount present in the two samples, where an equal amount of targets in the two samples will result in a yellow fluorescence.

Because of the numerous technical challenges that robotic spotting poses, such as variable spot size and varying binding efficiency of spotted ssDNA, an experienced microarray facility

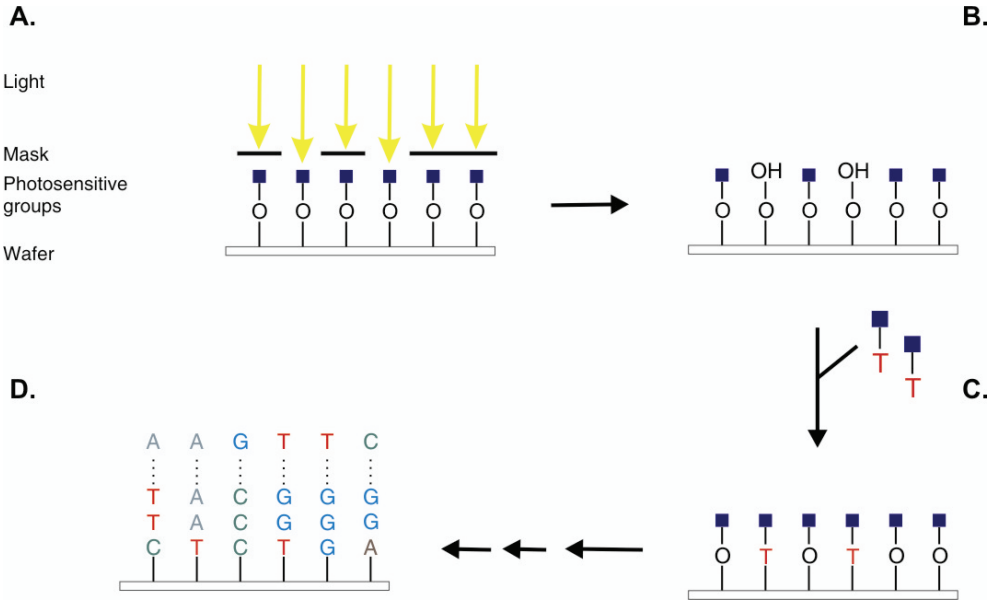

Figure 1-2. Principle of photolithography: light directed DNA synthesis on a wafer surface. (A) Photosensitive groups are exposed to UV light through a mask. (B) The exposed groups are converted to a hydroxy-group. (C) A specific photosensitive nucleotide, T, is attached to the hydroxy-groups. (D) After several cycles, where step (A), (B) and (C) have been repeated with different masks for the four nucleotides: A, T, G and C, an oligonucleotide is build.

is required to set up and manufacture this type of microarrays to reduce artefacts from the spotting process.

While oligonucleotides may be fabricated and spotted onto the microarray as described above, they may also be synthesized '*in situ'* directly onto a microarray slide by light directed synthesis, so-called photolithography. Photolithography combines the power of producing oligomer arrays of extremely high density and flexible patterns with a relatively simple procedure for independently directing the sequence of the molecules synthesized at the individual array positions. In addition, it facilitates large-scale chip production.

The principle of this method is illustrated in Figure 1-2 (Beier and Hoheisel, 2000; Maskos and Southern, 1993; Southern*, et al.*, 1999). Generally this method of DNA immobilisation has several advantages. The yield is high and the distribution of the DNA is consistent over the array. On the other hand, the *in situ* method is not optimal for immobilisation of longer oligonucleotides, compared to other methods (Southern*, et al.*, 1999). Another problem is that the necessary techniques are not easily available in individual laboratories. Because of the latter, the *in situ* technique has - until recently - been provided mainly through the chip manufacturing company, Affymetrix.

Recently, another provider, NimbleGen Systems Inc, has introduced an alternative approach comprising a maskless method of *in situ* DNA synthesis using a digital micromirror array (Singh-Gasson*, et al.*, 1999). This new approach has increased the flexibility of *in situ* synthesis tremendously, since it is now possible to order customized high-density oligonucleotide microarrays at a fraction of the cost of customized Affymetrix arrays manufactured by the use of masks.

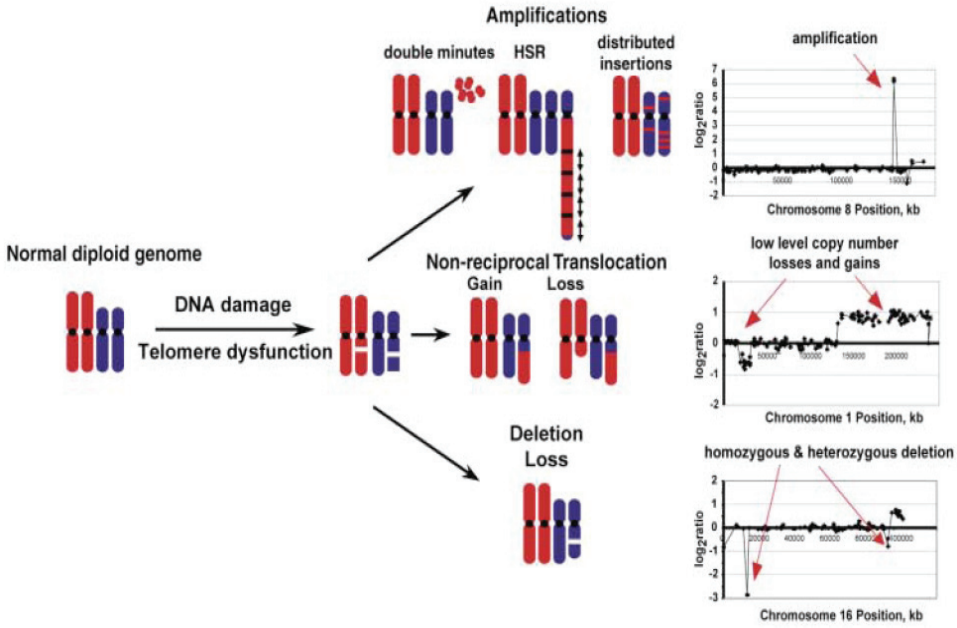

Figure 1-3. Illustration of the various copy number aberrations detectable by aCGH. HSR: homogeneously staining region (Albertson and Pinkel, 2003).

1.2 Array comparative genomic hybridization (aCGH)

Comparative genomic hybridization (CGH) is a technique by which it is possible to detect and map genetic changes such as chromosomal aberrations involved in gain or loss of genomic DNA. Figure 1-3 provides an overview of the various chromosomal aberrations that may be detected using CGH and the corresponding chromosomal profiles that may be obtained. While a normal diploid genome, such as the human, may experience a number of different amplifications, non-reciprocal translocations and deletions, it is not possible to distinguish between subtypes, e.g. double minutes (extra-chromosomal amplifications of specific DNA fragments) will result in the same profile as multiple distributed insertion amplifications. Nonetheless, microarray formats of CGH, array CGH (aCGH), provide a high throughput and relatively fast procedure for obtaining high-resolution copy number data. For this purpose, a variety of array platforms have been used, including large insert genomic clones, such as bacterial artificial chromosomes (BACs) (Snijders*, et al.*, 2003; Veltman*, et al.*, 2003), cDNA clones (Pollack*, et al.*, 1999) and oligonucleotides for array spots (Carvalho*, et al.*, 2004).

The experimental procedure may vary slightly depending on the array platform. Typically, a two-color scheme is used (Figure 1-4) where a test sample and a reference genomic sample from a healthy individual or healthy tissue from the same patient are co-hybridized to a representation of the genome. Then corresponding intensity ratios are measured for each clone. From this, copy number changes can be identified (Albertson*, et al.*, 2003). To block repetitive sequences in the genome, differentially labeled genomic DNA is combined with unlabeled Cot-1 DNA (Fridlyand*, et al.*, 2004; Pollack*, et al.*, 1999).

Another strategy to obtain high resolution copy number data is to reduce the complexity of human samples by using representations (e.g. small (<1.2 kb) *bgl*II restriction fragments), which has been found to improve signal-to-noise performance (Lucito*, et al.*, 2003). This is

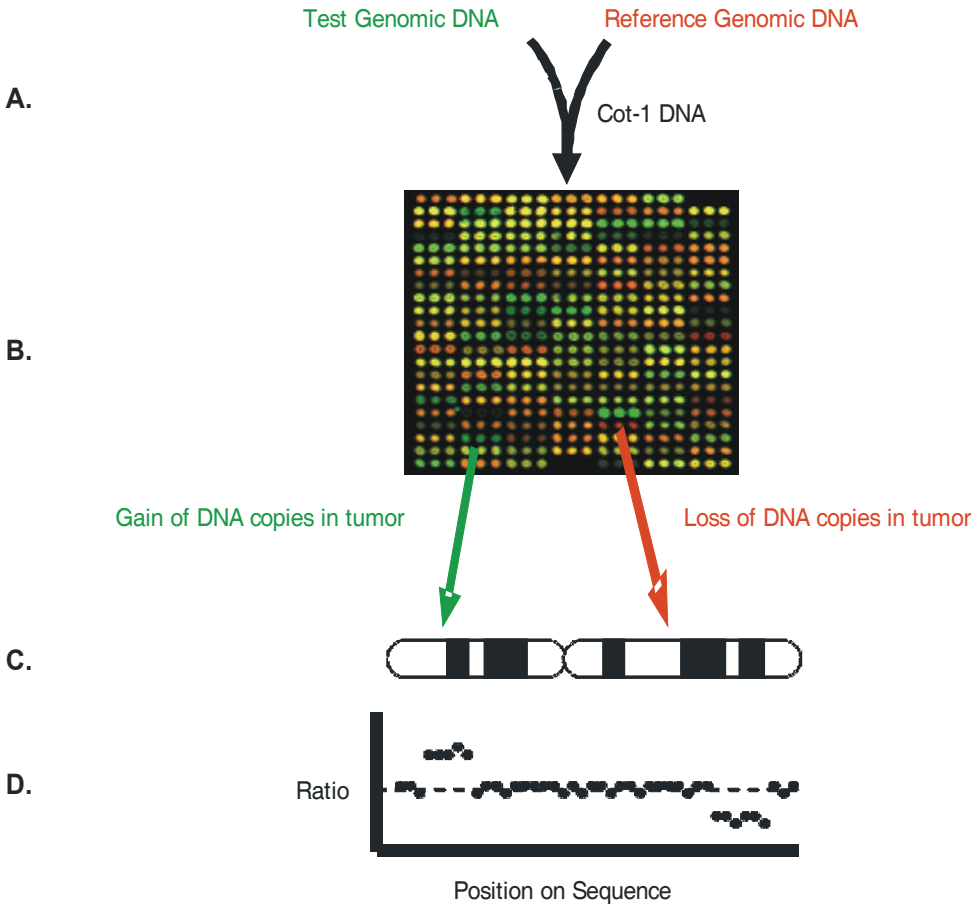

Figure 1-4. The experimental procedures of aCGH. **(A)** Labelled test and reference genomic DNA is pre-hybridized with cot-1 DNA and **(B)** co-hybridized to a microarray slide. Measured intensity ratios may be **(C)** mapped to a location on the chromosome **(D)** producing a 'copy number profile'.

the idea behind representational oligonucleotide microarray analysis (ROMA), which has evolved from an earlier method, representational difference analysis (RDA). Here, a microarray is constructed to consist of 70mer oligonucleotide probes designed to hybridize to representations of the human genome where the representations are characterized by reduced nucleotide complexity to increase the concentration of DNA complementary to the probes. Samples are prepared to consists of corresponding representations of the genome, e.g. by using PCR to select for small (<1.2 kb) *bgl*II restriction fragments. This results in roughly 200,000 fragments interrogating approximately 2.5 percent of the human genome (Lucito*, et al.*, 2003).

Chapter 2 Microarray Data Pre-processing

In the following, the most common pre-processing steps for microarray data will be introduced. While estimation of gene expression indices applies solely to gene expression data, use of spatial information in terms of segmentation approaches, may aid considerably in noise reduction for data from DNA copy number arrays.

2.1 Normalization

Microarray data from gene expression experiments are widely known for their high level of noise partly due to mRNA instability and sample size variations. Consequently, in order to make microarray data comparable, the intensity values must be normalized. Numerous statistical and physical models have been proposed to model these variations to normalize the data, i.e. to remove systematic sources of variation and make experiments comparable.

Simple linear scaling based on assumptions like constant total amount of RNA or constant expression of housekeeping genes may be applied for normalization (Knudsen, 2002). However, these assumptions are not always valid. The total amount of RNA is, for example, not constant when comparing starved cells with normal cells, and the expression of household genes have been shown to vary under different conditions (Schadt*, et al.*, 2001).

Generally, linear scaling is at best sub-optimal, since the distribution of gene expression data is rarely linear as illustrated in Figure 2-1A. This may be the effect of, for example, non-linear scaling of the fluorescence signal and probe saturation. Figure 2-1 also demonstrates an MA-plot, a popular way of illustrating probe level intensities from two gene expression microarray samples. It was originally suggested for comparing the red and green (R,G) – Cy3 and Cy5 cyanine dyes - intensities from two-color microarrays (Dudoit*, et al.*, 2002b), and the illustrated data transformation has been found particularly useful for normalization of

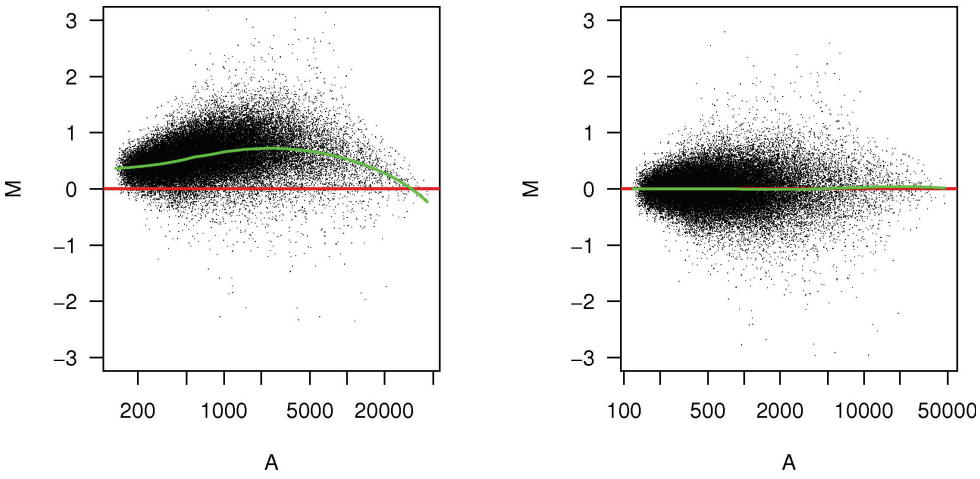

Figure 2-1. MA-plots for two arrays, before (I) and after (II) normalization. Red: line corresponding to the expected M value of all probes in case of equal gene expression for the two samples. Green: The lowess fit to the actual M values.

microarray data (Yang, *et al.*, 2002). For the MA-plot, the axes correspond to:

$M = \log_2 R/G$

$A = \log_2 \sqrt{(R \times G)}$.

Figure 2-1 illustrates how easy it is to perform non-linear, intensity-dependent normalization using the robust scatter plot smoother 'lowess' implemented in the statistical software package R (Yang, *et al.*, 2002). Numerous additional non-linear normalization methods have been developed, including 'qspline' developed here at CBS (Workman, *et al.*, 2002) and 'quantile' developed at Berkeley (Bolstad, *et al.*, 2003).

After normalization of the microarray data, the exact amount of antisense RNA (aRNA) applied to each chip is not so crucial anymore, since all chips are normalized against each other. For spotted microarrays, additional considerations might have to be taken into account when normalizing the data, e.g. considering print tip variation by performing a within-print-tip-group normalization where each print tip group is fitted individually (Yang, *et al.*, 2002).

All of the normalization procedures described above have been developed for pre-processing of gene expression data and as such, their assumptions regarding the distribution of the data, does not always apply to DNA copy number data (Snipen, *et al.*, 2006). Often median centering of DNA copy data may provide sufficient normalization for this type of data (Snijders, *et al.*, 2001; Snijders, *et al.*, 2003). However, recently, some methods have been proposed for normalization for experimental artifacts such as spatial bias (Neuvial, *et al.*, 2006) and other systematic biases (Khojasteh, *et al.*, 2005).

2.2 Expression Index

For oligonucleotide arrays (including both spotted and *in situ* synthesized arrays), measured intensities may be summarized for probes targeting the same gene. These so-called expression values must be calculated for each gene. On the other hand, cDNA arrays and arrays spotted with PCR products do not require probe level summary to extract gene expression levels since a single probe usually span the entire gene sequence.

On an Affymetrix chip there is a perfect match (PM) and a mismatch (MM) for each probe (probe pair). The purpose of the mismatch is to represent the background, and the expression index for a probe was originally calculated by Affymetrix as the average difference of probe pairs in a probe set. However, the importance of the mismatch has been discussed (Irizarry, *et al.*, 2003b; Li and Wong, 2001b). Often, an MM probe can exhibit higher intensity levels than the corresponding PM probe (Naef and Magnasco, 2003), resulting in many negative expression values when simply subtracting the intensity for the MM probe from the intensity of the PM probe. Since this makes little biological sense, Li & Wong suggested calculating the expression index using only perfect matches (Li, *et al.*, 2001b). Furthermore, their method for calculating the expression index is based on the fact that all probes are not equally good, but some tend to always have an intensity level lower or higher than the average. Therefore, a scaling factor for each probe is found based on empiric data and a multiplicative model fitted using least squares (Li, *et al.*, 2001b).

A similar model may be fitted using a more robust method than least squares, such as median polish (Holder, *et al.*, 2001). Also, the same robust linear fitting procedure may be used to fit a quite different log scale linear model to estimate log scale expression values from background-corrected, quantile normalized and $\log_2$-transformed probe intensities (Irizarry, *et al.*, 2003a; Irizarry, *et al.*, 2003b). This robust multi-array average (RMA) expression measure may further be adjusted for the GC content of the oligonucleotides (GCRMA) (Wu, *et al.*, 2004).

2.3 Choice of pre-processing method for gene expression data

The choice of preprocessing methods depend on the desired analysis and may impact the list of differentially expressed genes significantly (Shedden*, et al.*, 2005). For example, a recent publication found that for co-expression analyses, the Li-Wong summary method is the preferred method, while the RMA/GCRMA method is recommended for detection of differentially expressed genes (Harr and Schlotterer, 2006).

Affycomp II (http://affycomp.biostat.jhsph.edu/) is another resource that may help the researchers to choose between the vast range of pre-processing algorithms available. This web-based resource is set up to benchmark Affymetrix GeneChip expression measures (Cope*, et al.*, 2004; Irizarry*, et al.*, 2006). Here, authors of a pre-processing algorithm can benchmark their method on a number of spike-in datasets and compare the performance to previously submitted methods.

2.4 Segmentation

Finding a clear separation between DNA segments corresponding to different copy numbers of DNA is essential for the analysis of DNA copy number data. Consequently, the use of spatial information for noise reduction applies mainly to data from DNA copy number arrays such as array CGH. To reduce noise and increase the reliability of change point detection, anything from simple *smoothing* to advanced statistical *segmentation* algorithms has been proposed. The latter to automatically partitioning the probe measurements into sets corresponding to the same copy number by exploiting the physical dependency of the nearby probes.

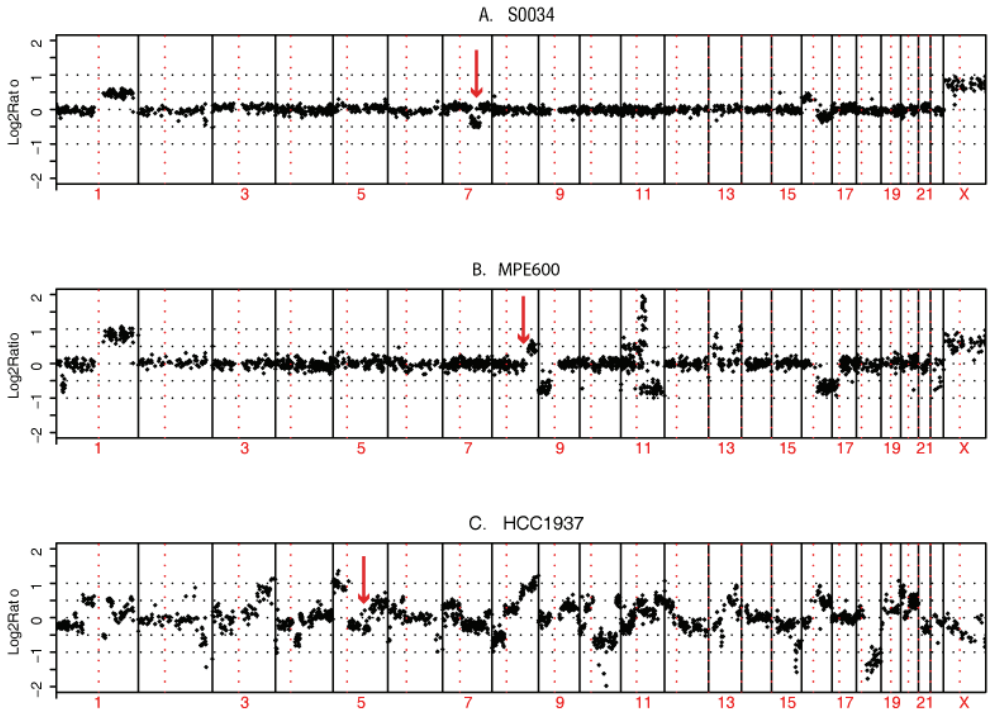

Figure 2-2. Illustration of copy number profiles obtained from three human cancer cell lines. Here, the $\log_2$-ratio between cancer sample and normal control is plotted as a function of chromosomal position for 22 chromosomes and the X chromosome. Gains are visible with $\log_2$-ratios above 0 and losses are visible with $\log_2$-ratios below 0. Examples of clearly visible breakpoints are indicated with red arrows.

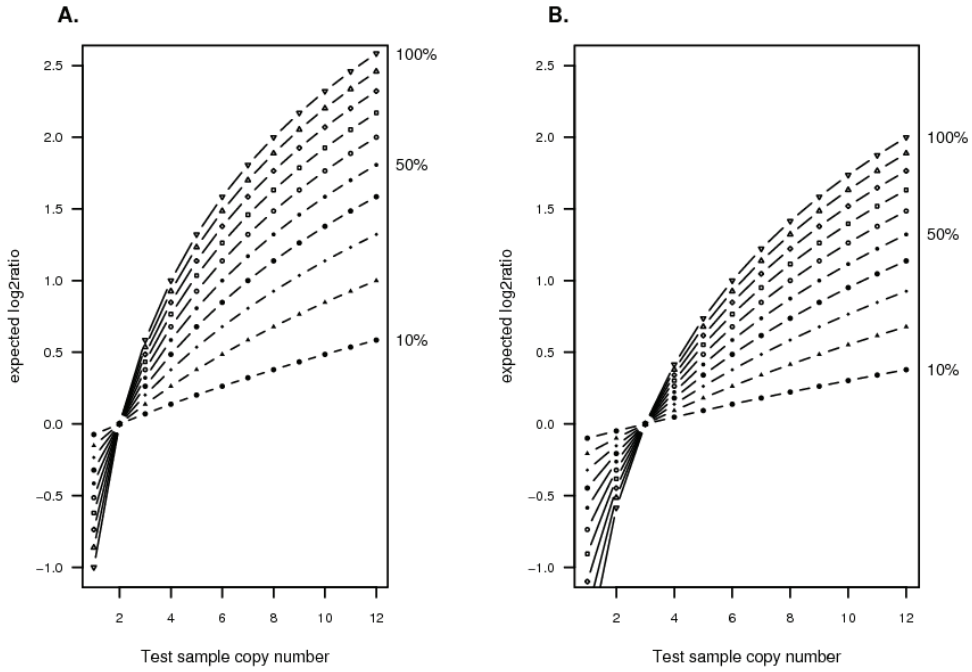

Figure 2-3. Illustration of the expected $\log_2$-ratio as a function of the true copy number in the abnormal cells and the proportion of abnormal cells in a sample. **(A)** Reference cell ploidy = 2. **(B)** Reference cell ploidy = 3.

Figure 2-2 provides a typical example of the differing complexity of copy number profiles from three human cancer cell lines. In the first example, the breakpoints are easy to spot (Figure 2-2A), but as the complexity increases, it becomes virtually impossible to manually spot the breakpoints (Figure 2-2C). The clear separation between segments corresponding to different copy numbers may be complicated by various sources of noise, including impurities in the sample (a mixture of tumor cells and surrounding cells), and unknown ploidy of the cells (Figure 2-3). Other sources of noise include random experimental noise and disease heterogeneity, e.g. certain aberrations occur only with a certain frequency for a given cancer subtype.

Chapter 3 Downstream Data Analysis

Microarray experiments produce enormous amounts of data. Therefore, the scientist faces a huge challenge, both in terms of selecting an appropriate statistical method to interpret the results with, and to exploit available computational power to process the data. While the number of samples is usually low due to cost or other limitations in sample availability, the number of genes probed for is usually very high. Consequently, use of advanced multivariate statistical methods and multiple testing procedures is necessary to obtain the correct interpretation of the data. Applying these methods, all sorts of analyses may be performed, including identification of differentially expressed genes, pathway analysis, classification of samples and genes and identification of copy number alterations.

3.1 Testing

Following preprocessing of the microarray data, intensity values may now be compared to identify up- or down-regulated genes or to identify gain or loss of genetic material. A number of different statistical tests may be applied to derive the significance, depending on the experimental design and the specific question in mind. The most popular tests include paired and un-paired two-sample T-test and the analysis of variance (ANOVA). The power of the t-test and the ANOVA is highly dependent on the number of replicates since the estimation of the P-value is based on variance. Thus, with a high number of replicates for each condition in the microarray, experimental variance estimates may be obtained with more confidence. While the t-test and ANOVA both are fairly robust to moderate departures from the underlying assumptions of normally-distributed data and equality of variance, the presence of very small or unequal sample sizes can decrease the statistical power considerably (Jafari and Azuaje, 2006).

The student's T-test is based on the assumptions that the data are normally distributed, and that the variances are the same for both groups. The T-test estimates the probability that the gene expressions for both groups come from the same T-distribution, from which it derives a probability, the so-called P-value. If the P-value is low (typically <0.05), there is a 5% chance of incorrectly rejecting H_0, the hypothesis that no true difference exists between expression levels in the two tested groups. The T-test is then said to be significant at a 5% significance level (Montgomery, 2000). When evaluating gene expressions, the variance for the two groups can differ significantly. In this case, Welsh T-test, which assumes unequal variances, may be performed. Here, testing of the hypothesis of equal means takes the number of degrees of freedom into account. The paired T-test may increase statistical power in cases where two conditions can be assumed to be dependent, such as patient samples before/after treatment. In this case, the hypothesis is that the difference of all pairs is zero, thereby reducing noise from between patient variance (Montgomery, 2000).

When experiments may be divided into more than two classes, an analysis of variance (ANOVA) may be applied. Using the ANOVA, it can be tested if one or more groups differ significantly from the others. The ANOVA uses the sum of squares as a measure of variance and compares the variance between groups to the variance within groups. The resulting F-statistic is compared to an F-distribution to determine significance. This test is based on the assumptions of normally-distributed data and equal within-group variance (Montgomery, 2000). The ANOVA may further be applied in cases where two or more conditions are varied simultaneously, e.g. a two-way ANOVA with two types of treatment and two types of disease subtypes.

Alternatively, non-parametric tests may be used to avoid making any assumption as to the specific underlying distribution model, that is, non-parametric tests are distribution free. Therefore, they may be less sensitive to outliers and deviations from e.g. normality. The two-sample Kolmogorov-Smirnov test is one of the most useful and general non-parametric

procedures for two-sample comparisons. It may be used to determine whether two underlying probability distributions differ, that is, if data from x and y were drawn from the same or differing continuous distributions. Thus, it is sensitive to differences in both location and shape of the distributions (Conover, 1971).

Rank-based procedures, such as the Wilcoxon rank sum test, are based on a comparison of ranks. Any variable that can be ordered can be assigned ranks based on this ordering from smallest to largest. The Wilcoxon rank sum test compares the central location of two independent unpaired populations and is the non-parametric, rank-based analogue to the two sample t-test. Furthermore, the Ansari-Bradley two-sample test may be applied to test for differences in scale parameters, that is, if two distributions differ in variance (Bauer, 1972).

Finally, exact tests may be used for finding over- or under represented features in a list. Consider the case where a list with genes of interest has been identified by one of the above described procedures. Instead of just skimming the list and manually attempting to identify prominent traits, one wants to determine with a high statistical certainty, if a particular feature such as 'cancer oncogene' or 'ribosomal protein' is present in the list at a higher rate than expected by chance. For this, one may use an exact test such as Fisher's exact test or other tests in the hypergeometric or binomial distributions. The main difference between tests in these two distributions is that the binomial models sampling with replacement while the hypergeometric models sampling without replacement (Draghici and Krawetz, 2003). Exact tests are very popular for analysis of overrepresentation of genes within certain pathways, in particular, it has often been used for analysis of over representation of given gene ontology (GO) terms (Ben-Shaul*, et al.*, 2005; Young*, et al.*, 2005).

3.2 Correction for Multiple Testing

Since microarray data comprises thousands of genes, the same test is applied thousands of times to the same microarrays posing a multiple testing problem that has to be taken into account when determining the significance of an identified difference. In any testing situation, we may commit one of two types of errors: a type I error (a false positive) by falsely rejecting the null hypothesis - no true difference in means exists - or a type II error (false negative), when failing to reject the null hypothesis - true difference in means exists. While it is not feasible to simultaneously minimize the chance of committing either error type given the data, one usually seeks a trade-off between the two types of errors. Consequently, the type I error rate is usually controlled at an acceptable level, alpha, while selecting testing procedures that aim at minimizing the type II error rate, that is, maximize power, while retaining the type I error at level alpha.

Especially, when working with thousands of genes, the chances of a false positive among numerous tests will increase enormously if using standard significance thresholds for each individual test. For example, with a standard significance threshold, alpha = 0.05, one would expect 0.05 false positives when testing one gene, we may expect 500 false positives when testing 10 000 genes. Therefore, it is necessary to make an adjustment - multiple testing correction - to avoid a large number of false positive conclusions. This may be obtained by tightly controlling the type I error rate.

In multiple testing cases, the number of false positives (type I errors) may be controlled either by the family-wise error rate (FWER) - probability of at least one false positive, or by the false discovery rate (FDR) - the maximum expected proportion of false positives among predicted positives (Benjamini and Hochberg, 1995). The FWER is controlled, for example, by the classical Bonferroni procedure, where the P-values that are accepted at a significance level, alpha, must be below alpha divided by the number of tests (Bonferroni, 1936). The extent to which a Bonferroni correction is necessary has been discussed, since it is very conservative and results in only very few genes are being rejected below the

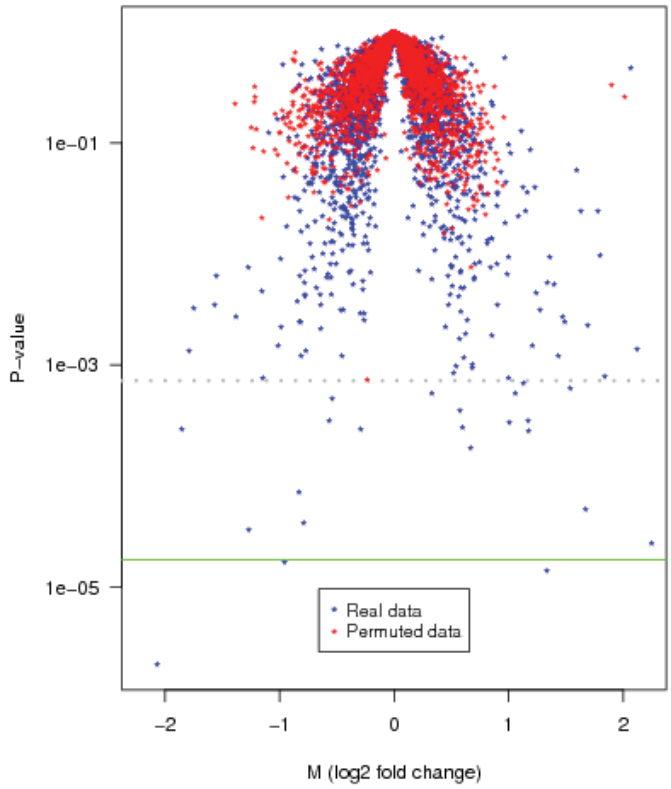

Figure 3-1. Volcano Plot. The P-value as a function of the $\log_2$-foldchange for (blue) real data example and (red) permuted data where the class labels have been shuffled before estimating P-values. The (green) horizontal line corresponds to the Bonferroni cut-off at alpha=0.05 for the 2839 genes in the example. Few genes are significant at this cut-off, while many genes have P-values below that expected by random (minimum random P-value is indicated with a grey dotted line).

corrected significance threshold when testing several thousands of genes. This may be exemplified in a volcano plot (Figure 3-1). However, the assumption that gene expression measures are independent is clearly not true since genes clusters, i.e. they are connected in networks and pathways. Consequently, other procedures for controlling the FWER have been suggested, for example a permutation based single-step maxT procedure (Westfall and Young, 1993) or a bootstrap re-sampling procedure to obtain consistent estimators of the null distribution for defining test-statistic cut-offs and derive adjusted P-values. One may also decide to accept a certain number of false positives, using the generalized family-wise error rate (gFWER), where the probability of k+1 false positives is maintained at level alpha (Dudoit*, et al.*, 2004). Here, k is an arbitrary number of additional false positives one is willing to accept compared to the standard FWER.

3.3 Cluster Analysis

Cluster analysis is a method for reducing the dimension of multivariate data in order to visualize the results of, e.g. a microarray experiment or to discover meaningful patterns. Consequently, the method is useful for class discovery purposes (Xing, 2003). For example by grouping genes or experiments in clusters with similar expression patterns, one may gain

$$X_{n\,x\,k} = \begin{bmatrix} X_{(1,1)} & X_{(1,2)} & \cdots & \cdots & X_{(1,k)} \\ X_{(2,1)} & X_{(2,2)} & \cdots & \cdots & X_{(2,k)} \\ \vdots & \vdots & \ddots & & \vdots \\ \vdots & \vdots & & \ddots & \vdots \\ X_{(n,1)} & X_{(n,2)} & \cdots & \cdots & X_{(n,k)} \end{bmatrix} \begin{matrix} \\ \\ \\ \\ n \end{matrix} \quad \Rightarrow \quad D = \begin{bmatrix} 0 & & & & \\ D(k_2,k_1) & 0 & & & \\ D(k_3,k_1) & D(k_3,k_2) & 0 & & \\ \vdots & \vdots & \ddots & 0 & \\ D(k_k,k_1) & D(k_k,k_2) & \cdots & D(k_k,k_{k-1}) & 0 \end{bmatrix}$$

Figure 3-2. Example of a n x k data matrix being transformed into a k x k distance matrix of pairwise distances.

an overview of which genes or samples that may be related somehow. The problem is formalized in the following.

Associated with each sample in the population of interest, is a vector of variables (e.g. gene expressions) describing each sample (Figure 3-2). Given these vectors of explanatory variables, $x_1, \ldots, x_n$, recorded for k samples, the task is to identify clusters, that is, groups or sets of similar samples and/or genes.

The methodology is not restricted to microarray data, and consequently, the variables x(n) can be a mix of quantitative numerical data such as gene expression indices (continous) or the number of genes (discrete); or qualitative such as tumor grade or pathogenecity.

Distance Measures and Linkage

Inherent in cluster analysis is a notion of distance or similarity between observations or features to be clustered. Consequently, the matrix of $n \times k$ values may be transformed into a $k \times k$ distance matrix of pairwise distances between the k samples to be clustered (or an $n \times n$ distance matrix of pairwise distances between the n observations) (Figure 3-2).

Before transforming observed values into a distance matrix, scaling and centering of the data might be necessary. Centering is performed by subtracting the sample means. Scaling is performed by dividing by the sample standard deviations. By centering and scaling, the effect of one observation or one variable is not considered more important than the next one by the clustering algorithm.

Several choices of distance metric exist: Euclidian distances, vector angle distance, '1 minus Pearson correlation' distances, etc. The applied distance measure may have a large impact on the results of a cluster analysis and it should therefore be carefully considered depending on the specific research question. For example, if one wants to identify genes in the same pathway, i.e. genes responding similarly to a treatment, '1-Pearson correlation' would be a good choice. Thereby genes that respond similarly will cluster together independently of their absolute gene expression values. On the other hand, if one wants to compare the codon preference of different organisms, Euclidian distances would give a better picture at how they are related in codon space (see for example Chapter 9).

A number of different ways exists for defining a distance between two clusters or between a single observation and a cluster of observations, e.g. *single linkage* which is the minimum distance between any two observations from each cluster; *average linkage* which is the average of all pairwise distances between the members of both clusters; *complete linkage* which is the maximum distance between any two observations from each cluster; and *centroid distance* which is the distance between their centroids, where a centroid is the cluster mean with regard to all variables. The shape of the resulting clusters depends on the

choice of linkage. For example, single linkage tends to lead to long, thin clusters, while average linkage tends to result in more round clusters.

Clustering Procedures

Several methods exist for the actual clustering. One of the most frequently used when working with microarray data analysis is hierarchical clustering and it may either be agglomerative (bottom up) or divisive (top down). Both approaches provide a hierarchy of clusters, from the smallest set, where all observations are in one cluster, through to the largest set, where each observation is in its own cluster. Hierarchical clustering results in one large cluster tree (dendrogram), which may be cut at any chosen level to give the desired number of individual clusters. The advantage with hierarchical cluster analysis is that it is deterministic; however, since hierarchical clustering methods usually make use of a distance matrix of dimension N^2, N being the number of data points, the size of the distance matrix becomes prohibitory for large N's in terms of memory and computational load.

Partitioning methods are usually more computationally efficient, although many are too complex to have exact solutions. Often, only approximate solutions are available and reproducibility may become an issue. Moreover, since these methods partition the observations into disjoint clusters, they usually require specification of the number of clusters. However, the number of reasonable clusters may be estimated by optimizing the Silhouette widths (Rousseeuw, 1987). Some examples of popular partitioning methods are K-means (Hartigan and Wong, 1979) and partioning around medoids (PAM) (Kaufman and Rousseeuw, 1990), where PAM is a robust version of the K-means and has been shown to work well for gene expression data (van der Laan*, et al.*, 2003)

3.4 Classification

When the classes are known *a priori*, supervised learning algorithms may be applied - commonly referred to as classification. Supervised classification methods try to predict/rediscover the known classes by various algorithms. Several supervised classification schemes have been devised. Some of the algorithms that have been used for classification and characterization of microarray data are simple algorithms such as the *k*-nearest neighbor classifier, and some are more complex and involves machine learning.

Feature Selection

Feature selection is the most important step in classification since all classification algorithms will perform well if a set of genes could be found that were entirely differentially expressed between the classes. Unfortunately, it is rarely possible to find such genes, and the challenge is then to search for a subset of genes that together might be able to distinguish the classes.

Most gene expressions in microarray data will not contribute to this class distinction (they have no discriminatory power) and may actually constitute an unwanted noise that many classification methods such as linear discriminant analysis cannot overcome. Therefore, the most important task of feature selection is to filter or remove these genes which often comprise all but a very small fraction of the entire feature set (Li and Weinberg, 2003).

Another important reason to do feature selection is to reduce the dimensionality of the data, as microarray experiments generally suffer from the problem that the number of samples, n, is relatively small compared to the number of genes, p. This may be a problem in statistical methods that uses the within-class covariance matrix which is singular if $n < p + 1$ (Antoniadis*, et al.*, 2003).

Feature selection may be performed either explicitly, before building the classifier, or implicitly as a part of the classifier training procedure, e.g. by a Bayesian approach (Krishnapuram*, et al.*, 2004). However, most common classification schemes do not employ

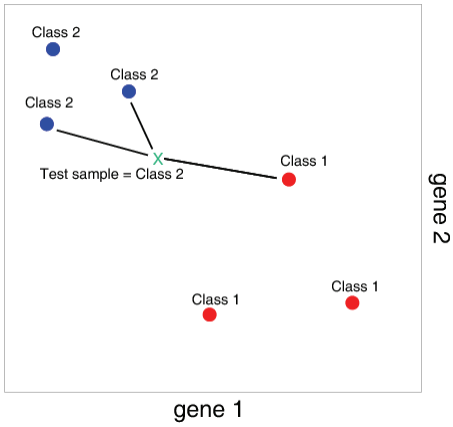

Figure 3-3. Selection algorithm for *k* nearest neighbour classification with k = 3. The 3 closest neighbours are determined by this distance function and indicated with lines. The class of the test sample is predicted as class 2, which is the class of the majority of the nearest neighbours.

any feature selection, and some sort of prior feature selection is usually required. The most simple feature selection methods are one-gene-at-a-time approaches. Here, genes are ranked according to a univariate test statistic, e.g. t-statistics (t-test) and F-statistics (ANOVA).

Classification Schemes

K-nearest Neighbors (KNN) and Nearest Centroid (NC) classification are both based on simple distance functions such as the Euclidian distance between pairs of samples in a g-dimensional space (g is the number of genes). The simple *k*-nearest neighbor (KNN) classification rule finds the *k* nearest samples by the distance function and predicts the class by a majority vote. The principle of KNN is illustrated in Figure 3-3. By always using an odd number of *k*, the situation of a vote tie is avoided (Dudoit and Fridlyand, 2003b). Nearest centroid classification is based on estimates of the class average of real samples of known class for each gene considered by the classifier (class centroid). By comparing the squared Euclidian distance of a test sample to these class centroids, it may be classified as the class whose centroid it is closest to (Dudoit and Fridlyand, 2003a).

Statistical approaches aim at finding a mathematical rule, a so-called discriminant function that can separate known classes. Some examples are Quadratic Discriminant Analysis (QDA), Linear Discriminant Analysis (LDA) and Diagonal Linear Discriminant Analysis (DLDA). DLDA is a simplification of the maximum likelihood estimator for linear discriminant analysis by using a diagonal covariance matrix (Dudoit*, et al.*, 2003b). Such maximum likelihood estimations has been used for classification by e.g. (Dyrskjot*, et al.*, 2003). Here, a test sample is classified according to its proximity to the centroid of a number of classes in much the same way as described in the above method of nearest centroid classification. However, the squared distances between a sample and class centroid for each gene are standardized by the estimated variance for each gene. Thereby, more weight is given to genes whose expression is more stable.

Artificial Neural Networks (ANN) and Support Vector Machines (SVMs) are among the most popular machine learning approaches in classification. SVMs represent a powerful technique for general linear and nonlinear classification and compared to many of the previously described classification methods, SVMs are better at handling data where the number of features (e.g. genes) exceeds the number of samples (Li*, et al.*, 2003).

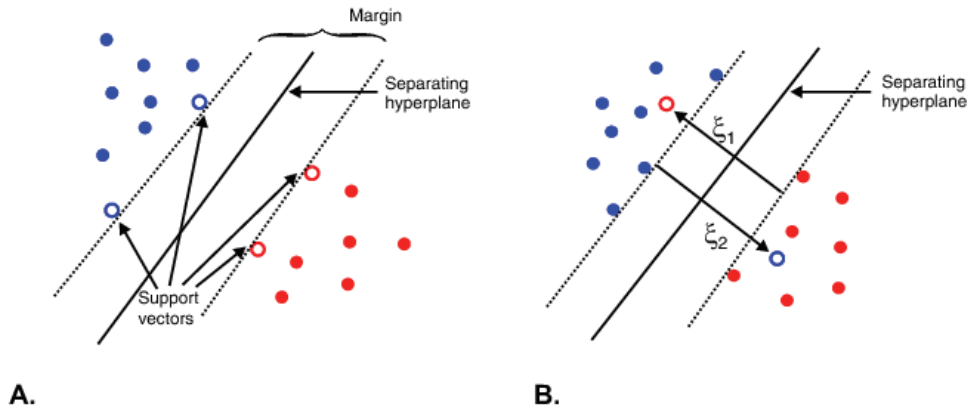

Figure 3-4. Illustration of the idea of Support Vectors. **(**A) The margin between the separating hyperplane and the closest points from the two classes (the support vectors) should be maximized and points from either class must fall on opposite sides of the separating hyperplane. The points lying on the boundaries, the open circles, are referred to as support vectors. The larger the margin, the larger is the margin for error when later classifying test samples. (B) Training samples may not be linearly separable. The data points on the "wrong" side of the discriminating margin may then be weighted down to reduce their influence by the use of "soft margins". Here, the distance of the i'th "misplaced" sample from its own class margin, ξ_i, is minimized as best as possible. The distances, ξ_1 and ξ_2 should be minimized simultaneously with the maximization of the margin.

The aim of SVMs is to maximize the margin between a separating hyperplane and the closest points from the two classes as sketched in Figure 3-4A. However, often samples are not linearly separable as illustrated in Figure 3-4B. Consequently, there is usually a trade-off between finding a hyperplane with a large margin and finding a hyperplane that separates the data well (minimizing ξ_i).

3.5 Performance Evaluation

Often, in the case of microarray data, the sample pool is not as large as desired and one cannot afford to leave out a large part of the data set for validation. However, the error rate may be severely under estimated when estimating the error rate of a classifier on the same data set as was used to build the classifier. Instead, cross validation is commonly used to provide a more accurate estimate of classification error rates. When using cross-validation, the training set is split into *n* smaller parts that in turn are used as test samples while the classifier is built using the remaining samples. When working with very small data sets, it is common practice to use leave-one-out cross-validation (LOOCV), where the data is trained on all but one sample and then tested on the last. On the downside, this training procedure carries a high computational burden as it requires the training procedure to be repeated *n* times. Since feature selection if often a major part of a classification scheme involving microarray data, when using cross validation to estimate the performance of the classifier, the feature selection step should be incorporated as the important features usually are unknown (Dudoit*, et al.*, 2003b).

Matthew's correlation coefficient may be used as a measure of classification performance. It considers the number of true and false classifications in each class as well as the number of samples belonging to each class:

$$CC_X = \frac{(TP \cdot TN) - (FP \cdot FN)}{\sqrt{(TN + FN)(TN + FP)(TP + FN)(TP + FP)}}$$

TP and FN are true positives and negatives, respectively, and FP and FN are false positives and negatives, respectively (Baldi and Brunak, 1998).

Pearson's product-moment correlation and Spearman's rho (often referred to as Pearson's and Spearman's correlation coefficients) may be estimated to determine the degree of association between two variables. Pearson's correlation coefficient is the most commonly used. It measures the strength of a linear relationship between two variables, whereas Spearman's correlation coefficient is a rank based measure of association between two variables. It is more resistant to outliers than Pearson's correlation coefficient since it relies on ranks. It measures the degree of monotonic relationship and is 1 for perfect monotonic increase (where for two variables, x and y, x increases as y increases) and -1 for perfect monotonic decreasing, while 0 indicates no monotonic relationship.

Sensitivity and specificity are both popular measures of performance. While sensitivity is the proportion of positive test examples that are correctly classified as positive, specificity is the corresponding proportion of negative test examples that are correctly classified as negative (Lazarus, 1999). When evaluating classification performance, these measures may be formalized as:

Sensitivity = TP / (TP + FN)
Specificity = TN / (TN + FP)

Note that another definition of *specificity* is also used frequently in the bioinformatics literature, although this measure is more correctly referred to as positive predictive value (PPV).

PPV = TP / (TP + FP).

A detailed performance analysis might be obtained by looking at the receiver-operating-characteristics (ROC) in a so-called ROC-curve. Here, sensitivity is plotted as a function of the false positive rate - corresponding to '1 minus specificity' - and the larger the area under the curve, the better the performance. Thus, areas approaching 1 corresponds to a near perfect performance.

Part II

GENE EXPRESSION ANALYSIS

Chapter 4 Paper I

Prediction of immunophenotype, treatment response, and relapse in childhood acute lymphoblastic leukemia using DNA microarrays

Hanni Willenbrock[1,4], Agnieszka Sierakowska Juncker[1,4], Kjeld Schmiegelow[2], Steen Knudsen[1] and Lars Peter Ryder[3]

[1]Center for Biological Sequence Analysis, Technical University of Denmark, DK-2800 Lyngby. Denmark.

[2]The Pediatric Clinic II, The University Hospital, Rigshospitalet, Copenhagen, Denmark.

[3]Department of Clinical Immunology, Tissue Typing Laboratory, The University Hospital, Rigshospitalet, Copenhagen, Denmark

[4]These two authors contributed equally to this work

ABSTRACT

Gene expression profiling is a promising tool for classification of pediatric acute lymphoblastic leukemia (ALL). We analyzed the gene expression at the time of diagnosis for 45 Danish children with ALL. The prediction of 5-years event-free survival or relapse after treatment by NOPHO-ALL92 or 2000 protocols resulted in a classification accuracy of 78% and a Matthew's correlation coefficient of 0.59 independently of immunophenotypes. The sensitivity and specificity for prediction of relapse was 87% and 69% respectively.

Prediction of high vs low levels of the minimal residual disease (MRD) on day 29 (≥0.1% or ≤0.01%) resulted in an accuracy of 100% for precursor-B samples. The classification accuracy of precursor-B- vs T-lineage immunophenotypes was 100% even in samples with as little as 10% leukemic blast cells, and the immunophenotype classifier constructed in this study was able to classify 131 of 132 samples from a previous study correctly. Our study indicates that the Affymetrix Focus Array GeneChip may be used without loss of classification performance compared to previous studies using the far more extensive U133A+B GeneChip set. Further studies should focus on prediction of MRD as this prediction would relate strongly to long-term outcome and could thus determine the intensity of induction therapy.

INTRODUCTION

In the Nordic countries acute lymphoblastic leukemia (ALL) has an annual incidence of approximately 3.9 per 100,000 children (Hjalgrim, *et al.*, 2003). The diagnosis of ALL is currently based on morphological, immunophenotypic, and cytogenetic analysis of a bone marrow sample, as well as clinical examinations. Based on biological and clinical features, the patients are assigned to risk group adapted therapy, and the 5 year event free survival (EFS) rate has increased to more than 75% within the last decade (Gustafsson, *et al.*, 2000; Pui, *et al.*, 2002).

Gene expression profiling of childhood ALL cases has previously shown great promise in diagnosing and risk classification of ALL. Several studies have shown that it is possible to distinguish between the two major immunophenotypes, precursor-B- (preB-) and T-lineage ALL, when applying classification methods based on the genetic profile (Golub, *et al.*, 1999; Ross, *et al.*, 2003; Yeoh, *et al.*, 2002). Furthermore, it has been reported possible to predict relapse as well as development of secondary acute myeloid leukemia within certain subgroups of ALL with 97% - 100% accuracy (Yeoh, *et al.*, 2002).

The aim of this study was to further explore the potential for prediction of relapse and classification of ALL subtypes. We present an attempt for prediction of long term outcome and treatment response using microarray analysis of diagnostic bone marrow samples from children with ALL who have been treated according to the NOPHO-ALL92 protocol (Gustafsson, *et al.*, 2000). The long-term outcome was here assessed by either continuous complete remission (CCR), as judged by 5-year EFS, or relapse in the same period. Treatment response was predicted according to either low or high (≤0.01% or ≥0.1%) minimal residual disease (MRD) measured on day 29 of treatment (Nyvold, *et al.*, 2002). Furthermore, we attempted the classification of the two major prognostically relevant immunophenotypes, preB- and T-lineage ALL, also for samples with less than 75% leukemic blasts. In this study, we based our gene expression analysis on the Affymetrix Focus Array GeneChip consisting of 8763 well-characterized human genes from the Affymetrix U133A GeneChip. Further material and raw data may be found at http://www.cbs.dtu.dk/~hanni/ALL.

METHODS

Patients and material

The study material included children with ALL for whom cryopreserved mononuclear bone-marrow cells had been stored at the time of diagnosis, and who were diagnosed between January 1st, 1992 and April 1st, 2003 and treated at the University Hospital, Rigshospitalet, in Denmark according to the NOPHO ALL-92 protocol or the NOPHO ALL-2000 protocol. Patients who failed to achieve remission, died during induction therapy or in remission, or who developed a second neoplasm were excluded from the study. Criteria for classification as standard (SR), intermediate (IR), high (HR), and very high risk (VHR) have been published previously (Gustafsson, *et al.*, 2000).

The material included 45 ALL patients, 15 girls and 30 boys with a median age of 8.3 years (range 1-15) at the time of diagnosis. Written consent was obtained for all included patients. The 38 patients diagnosed between January 1992 and June 2001 and treated according to the NOPHO ALL-92 included six cases of SR-ALL, 10 cases of IR-ALL, and 22 cases of HR/VHR-ALL. For these patients, induction, consolidation and maintenance therapy has been detailed elsewhere (Gustafsson, *et al.*, 2000). Of the 38 patients, 13 developed a relapse, while 21 patients had a 5-year EFS on April 1st, 2003 and are referred to as patients with continuous complete remission (CCR) (Table 4-1). The remaining four patients as well as the seven patients treated according to NOPHO ALL-2000 protocol have been followed less than 5 years from the date of diagnosis (Table 4-1). Among the patients for whom the

Table 4-1. ALL patients selected for our study according to 5-year outcome and immunophenotype.

	Immunophenotype		
Outcome	preB-lineage	T-lineage	Total
Relapse within 5 years	8	5	13
5-year event-free survival	13	8	21
Unknown 5-year outcome	5	6	11
Total	26	19	45

day-29 MRD level measurements were available, 24 patients had MRD levels ≥ 0.1% (high MRD) and 11 patients had MRD levels ≤ 0.01% (low MRD). During the first 29 days, all these 35 patients received identical therapy except an extra pulse of doxorubicin on day 8 for HR and VHR patients treated according to the NOPHO ALL-92 protocol. For more details about the patients see http://www.cbs.dtu.dk/~hanni/ALL.

Percentage of leukemic cells in samples

The percentage of leukemic cells present in each patient sample was estimated from data from immunophenotyping of the samples, based on expression of lineage-specific surface antigens like CD19, CD20 and CD3. Out of all 45 samples, 11 had a leukemic blast percentage of <75% (eight preB- and three T-lineage patients). Among these samples, one had 10%, four had 30-55% while the remaining six samples had 60-70% leukemic blasts.

RNA amplification and application to microarrays

Total RNA was purified from cryopreserved mononuclear cells using the ToTALLY RNA™ Kit (Ambion). For mRNA amplification, the MessageAmp™ aRNA Kit (Ambion) was applied, with the exception of the cDNA purification, which was made according to the Affymetrix clean-up protocol. All remaining steps were made according to the Affymetrix protocol. The final concentration was adjusted for starting amount of total RNA, and 10 µg of aRNA (or less if 10 µg of aRNA had not been obtained from the amplification step) was fragmented. Hybridization cocktails for Midi array format were prepared and samples were hybridized to Affymetrix Focus Array GeneChips for 15-17 hours and subsequently washed and stained with R-Phycoerythrin-streptavidin using the Midi_euk2v3 fluidics protocol. Finally, the GeneChips were scanned using the Agilent GeneArray® Scanner to determine the fluorescence intensity for each probe on the chip. Intensities for all probes were saved in a 'CEL file' for subsequent analysis.

Initial data treatment and statistical analysis

The R statistical software (Ihaka and Gentleman, 1996) was used for the initial data treatment, statistical analysis, and for classification.

Raw probe intensities were normalized using qspline, a nonlinear normalization method (Workman*, et al.*, 2002). Gene expression indices were calculated using the method of Li & Wong (Li and Wong, 2001a; Li and Wong, 2001c) with outlier detection using only perfect matches and background correction using a method implemented in the Robust Multichip Analysis method for calculating expression indices (Irizarry*, et al.*, 2003b). Unsupervised analysis was performed by hierarchical cluster analysis of patients using Euclidian or vector angle distances.

Feature (gene) selection for the classification was carried out by ranking genes according to their P-value in Welsh t-test. Further dimension reduction was done by principal component analysis on a number of selected genes (Knudsen, 2002). The classification was carried out on a training set consisting of 2/3 of the data samples randomly selected. Various classification methods were applied and evaluated: k-nearest neighbor (KNN) (Dudoit*, et al.*, 2003b; Knudsen, 2002), nearest centroid (NC) (Dudoit*, et al.*, 2003a), maximum likelihood

(ML) (Dyrskjot, *et al.*, 2003), nearest shrunken centroid (NSC) (Tibshirani, *et al.*, 2002), linear discriminant analysis (LDA) (Conradsen, 2002; Dudoit, *et al.*, 2003a) and support vector machines (SVM) (Cortes and V., 1995). Classifier performance was evaluated by leave-one-out-cross-validation (LOOCV) and classification accuracy as well as Matthews correlation coefficient (CC) (Matthews, 1975). For further information on the applied statistical data treatment, see http://www.cbs.dtu.dk/~hanni/ALL.

Prediction of immunophenotype

Classification of the preB- and T-lineage immunophenotypes was based on the 34 patients with ≥75% leukemic cells in the samples (18 had preB- and 16 T-lineage). For training, 2/3 of the data set (23 samples) was randomly chosen to comprise a training set. Only three simple classification methods were applied, KNN, NC and ML algorithms. For the choice of the number of 'general class discriminatory genes' the 50 top ranking genes were evaluated with regard to their appearance in each of the 23 LOOCV t-test to determine the genes present in all top 50 LOOCV t-tests. These genes were used for training and testing of one optimal classifier for each method. These optimal classifiers were also used for testing of the data set consisting of the 11 patient samples with less than 75% leukemic blast cells.

Prediction of immunophenotype for samples from a previous published study

The raw microarray data (CEL files) for the 132 samples from the study of Ross et al. (Ross, *et al.*, 2003) were obtained as test samples for our immunophenotype classifier. First, the CEL files were normalized against each other and expression indices were calculated by the same procedure as used for our own chip data. Each U133 GeneChip sample was reduced to probe sets included on the Focus GeneChips and each sample was subsequently normalized - one at a time - against all of the Focus GeneChips from our study, using the qspline normalization method (Workman, *et al.*, 2002). These data were applied as a test set for our immunophenotype classifier.

Prediction of relapse

Patients with either relapse or CCR as well as ≥75% of leukemic cells in the sample were used for classification of relapse. Here, 10 relapsed patients (six preB- and four T-lineage) and 18 CCR patients (10 preB- and eight T-lineage) were included. The number of input genes was varied from 2 to 150, and the number of principal components on selected genes was varied from 2 to 12. Several classification methods were applied: KNN, ML, NC, NSC, LDA and SVM. Random sampling, training on 2/3 of the data set (19 samples) and testing on the independent samples (nine samples) were performed a total of 10 times to ensure that the obtained classification performance was not due to sampling effects. The 30 top ranking genes were evaluated by their presence in at least nine of the 19 LOOCV t-tests in at least four of the 10 random samplings to retrieve the ‘general class discriminatory genes’.

Prediction of day-29 MRD levels

Patients with available day-29 MRD level measurements as well as ≥75% of leukemic cells in the sample were used for classification of MRD levels. Here, 18 patients with high day-29 MRD levels (nine preB- and nine T-lineage) and eight patients with low day-29 MRD levels (six preB- and two T-lineage) were included. The same classification and optimization approach was used as for the prediction of relapse.

RESULTS AND DISCUSSION

Unsupervised analysis of all samples

A hierarchical cluster analysis of all 45 ALL samples based on all 8763 gene expressions (Figure 4-1) showed that patients with the preB or T immunophenotype generally grouped separately, although a complete separation of the two subtypes was not observed. In three

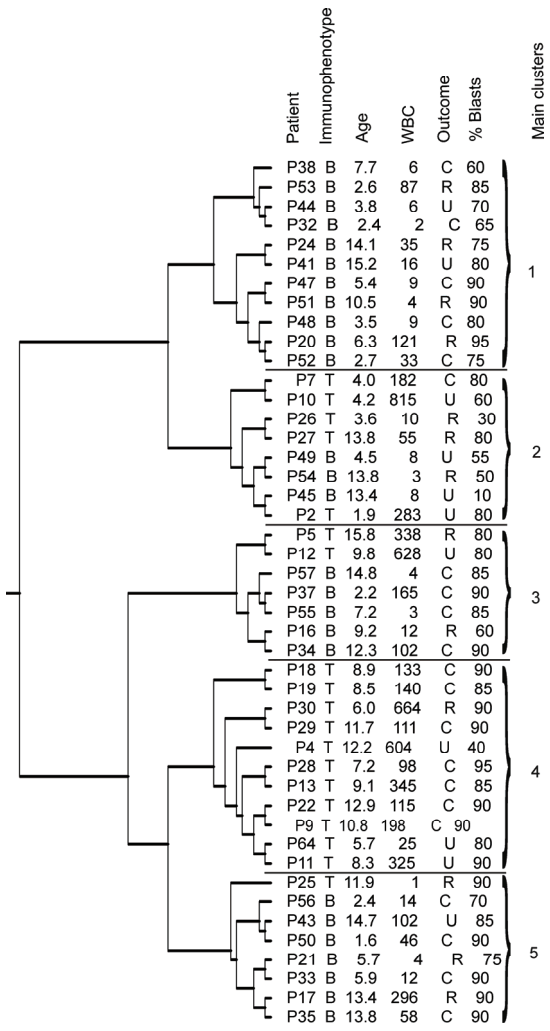

Figure 4-1. Hierarchical clustering of 45 ALL samples included in the study. From left: patient number, immunophenotype (B: preB-lineage, T: T-lineage), age, WBC (in 10^9/l), 5-year outcome (R: relapse, C: CCR, U: unknown), percentage of leukemic blasts in the sample. The five main clusters are marked.

of the five main clusters, preB and T samples were grouped together, except for one T sample (P25). Based on this unsupervised analysis it seemed as if the most apparent differences in gene expressions between the ALL patients in our study were those determined by the immunophenotype. The same results have previously been obtained by application of unsupervised analysis to ALL microarray data (Golub*, et al.*, 1999), where preB and T immunophenotypes were identified as the two major subclasses of ALL with almost complete separation. Regarding the prognostic factors: WBC and age at the time of diagnosis as well as relapse or CCR, no sub-clustering within the preB or T clusters was observed for the 45 patients.

Classification of immunophenotype

It has previously been reported that the gene expression profiles of preB and T-ALL were easily separable by means of classification (Golub*, et al.*, 1999; Yeoh*, et al.*, 2002), and classification of immunophenotypes was also attempted in our study. For all the applied classification methods (KNN, NC and ML) an accuracy of 100% was obtained both for training (23 samples) and testing the independent test set (11 samples) on the optimal classifier based on the 29 'general class discriminatory genes'. Among these 29 genes (see http://www.cbs.dtu.dk/~hanni/ALL), several were encoding well-known immunophenotype specific proteins, i.e. CD19 and CD3. Moreover, as little as one single gene, CD74, appeared to be enough to distinguish between the two immunophenotypes of ALL.

Prediction of immunophenotype for samples with <75% leukemic cells

In two previous extensive studies, all classifications were based on samples with ≥75% leukemic cells and it had been questioned if their subtype classifier might perform as well on samples with lower levels of leukemic blasts (Ross*, et al.*, 2003; Yeoh*, et al.*, 2002). For classification of the 11 samples with less than 75% leukemic blast cells as either preB or T-ALL by our simplified classifier using the 29 'general class discriminatory genes', all samples were classified correctly for the KNN (k = 1 and 3) and NC classification methods. Thus, our study indicates that the subtype-specific gene expression profile measured in ALL samples was characteristic enough even in samples with less than 75% leukemic cells, and that samples with as little as 10% leukemic blast cells may be classified correctly with respect to immunophenotype. Gene expression analysis might therefore be an improvement of the immunophenotype identification for patients with a small fraction of leukemic cells, where immunophenotyping might be difficult using the current flowcytometric methods.

Prediction of immunophenotype for samples from a previous published study

The results from the classification of preB and T immunophenotypes obtained in our study were evaluated by testing the 132 samples applied onto Affymetrix U133 GeneChips from the study of Ross et al. (Ross*, et al.*, 2003). Prediction of immunophenotype using the optimal classifiers designed in our study resulted in very good performance, where the NC method appeared to be superior (Table 4-2). Only one sample with MLL rearrangements was misclassified of all 132 samples. This subtype of preB ALL was, however, not represented in our data set.

Generally, our results confirmed, that the selected 29 'general class discriminatory genes' were not only applicable for prediction of our particular samples, but were general for prediction of immunophenotype. Moreover, nine out of the 29 'general class discriminatory genes' were identical to the 100 genes found to be characteristic for distinguishing between preB and T reported by Ross et al. (Ross*, et al.*, 2003).

Prediction of relapse

Hierarchical clustering of preB- and T-lineage patients separately as well as for the pooled

Table 4-2. Results from testing of Ross et al's data on the optimal classifier built based on the 29 'general class discriminatory genes', trained on 2/3 of the data set with ≥75% leukemic cells.

	T ALL (%)	preB ALL (%)					
		E2A-PBX1	Hyperdiploidy	MLL	BCR-ABL	TEL-AML1	Other
K-nearest neighbor, k=1	100	100	100	90	100	100	96.40
K-nearest neighbor, k=3	100	100	100	85	100	100	89.30
Nearest Centroid	100	100	100	95	100	100	100

Table 4-3. Prediction of relapse independently of immunophenotype (pooled preB and T ALL samples).

	Correlation coefficient[a]	Accuracy[a]	Optimal parameters	p-value[a,b]
K-nearest neighbor	0.51	0.77	45 genes, k^c=3	0.064
Maximum likelihood	0.44	0.72	4 genes	0.074
Nearest centroid	**0.59**	**0.78**	**30 genes**	**0.021**
Nearest shrunken centroid	0.47	0.72	3 genes	0.054
Linear discriminant analysis	0.41	0.71	2 PC (based on 4 genes)	0.082
Support vector machine	0.33	0.69	2 PC (based on 30 genes), c^d=3	0.074

[a]The values are the average of 10 random samples from LOOCV training of classifiers trained on 2/3 of the data set.
[b]As determined by a permutation test.
[c]Specific parameter for k-nearest neighbor.
[d]Specific parameter for support vector machines.

preB- and T-lineage patients based on all gene expressions did not reveal any clustering into groups with the same outcome when only the 28 relapse or CCR patients with ≥75% of leukemic cells in the sample were included in the analysis (see http://www.cbs.dtu.dk/~hanni/ALL).

When various classification methods were applied to predict either relapse or CCR of ALL patients, mean CCs in range of 0.33-0.59 and corresponding accuracies of 0.69-0.78 were obtained for the 19 random samples used for LOOCV training (Table 4-3). While most methods had an optimal performance using 3-45 gene expressions, LDA and SVM seemed to perform best when using dimension reduced data in the form of two principal components. The nearest centroid classifier had the best performance, CC=0.59±0.18, with an optimal number of 30 genes (as can be seen in Figure 4-2), and this method is also the only one that showed a significant classification performance (with an estimated P-value of

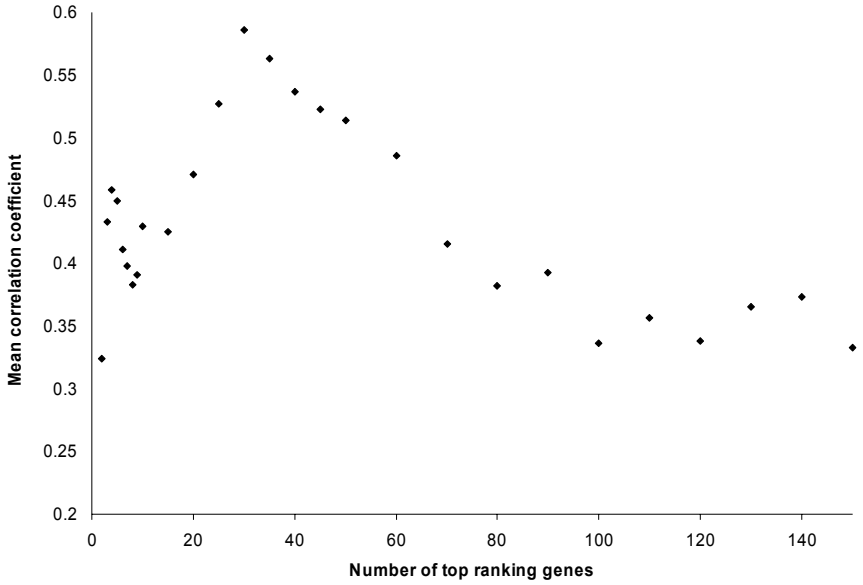

Figure 4-2. Prediction of relapse. The correlation coefficient (average value for the 10 random samplings) as a function of number of input genes for prediction of outcome (relapse or CCR) for the nearest centroid method.

0.021). Thus, for all other classifiers than NC, there is a statistically significant chance that classification performances matching the obtained CCs might have been obtained with random prior class assignments with a significance level of 5%. Therefore, the NC method seems to be most suitable for the outcome classification problem, and it is also reasonable that a simple method like NC is optimal when only a limited number of samples is available.

Testing of the nine samples in each of the 10 random independent test sets on the LOOCV nearest centroid classifiers resulted in a CC of 0.56 ± 0.2 and a corresponding accuracy of 74% ± 0.11. This CC and accuracy are thought to be a truer estimate of the classification performance than obtained for LOOCV during training and it is noteworthy that they do not differ significantly.

Interestingly, while the overall accuracy for prediction of outcome was found to be only 74%, the prediction accuracy for the relapse was 87%. The fact that almost all patients with relapse were found among the patients predicted as relapsed based on the diagnostic samples might be important for clinical application of the prediction method, since these patients could have been given an alternative or more intensive treatment. On the other hand, among the patients predicted to be CCR patients, a high percentage is in fact CCR patients (specificity of 92%). Future studies are needed to explore whether this subset can be cured with less intensive therapy. The specificity for patients predicted as relapse patients was only 69%. However, this is a far better overall specificity of relapse prediction than that obtained presently by conventional risk classification criteria such as age, white cell counts, immunophenotype, and cytogenetics.

By evaluation of the 30 top ranking genes, 19 'general class discriminatory genes' were retrieved (see http://www.cbs.dtu.dk/~hanni/ALL). A hierarchical cluster analysis of the 28 patients based on the gene expression of these 19 genes (Figure 4-3) illustrates that the relapsed patients group together with 4 CCR patients, while only one single relapse patient cluster with the CCR patients. The cluster analysis pattern thus supports the fact that several of the CCR patients are predicted as relapsed patients, while only few relapse patients are predicted as CCR patients.

It has previously been reported possible to predict relapse in certain subgroups of ALL by use of gene expression data (Yeoh*, et al.*, 2002) with a prediction accuracy of 97% for T-lineage ALL. However, it has later been discovered that this prediction accuracy was overestimated since it was based on LOOCV only during classifier training while the feature selection step had not been included in the LOOCV procedure. The performance was subsequently re-estimated (James R. Downing, December 2003) and resulted in a much lower classification accuracy of 73.5% using the top 50 ranked genes in a t-test and data pre-treatment by Affymetrix MAS 5.0. However, the specificity for relapse cases was only 25% giving a CC of 0.16.

The CCs for prediction of relapse independently of immunophenotype found in the present study (0.59 and 0.56 for the LOOCV training and the independent test sets, respectively) were significantly higher than the CC obtained for T-lineage samples for the re-evaluated data set from Yeoh et al. (Yeoh*, et al.*, 2002), while the prediction accuracy obtained in the present study (78%) was only slightly higher than the re-estimated accuracy obtained by Yeoh et al. (Yeoh*, et al.*, 2002). The better results obtained in our study might partly be due to the different treatments that patients had received in these two studies as well as differences in the period for EFS applied to define patients with CCR, where we defined the minimum period of EFS to be 5 years, while patients with shorter EFS period were included as CCR patients in the study of Yeoh et al. (Yeoh*, et al.*, 2002).

Moreover, it was reported by Yeoh et al. (Yeoh*, et al.*, 2002) that it was not possible to predict relapse across subtypes of ALL. However, the results from our study indicated that an at least as good classification performance could be obtained when predicting relapse

independently of ALL immunophenotype compared to prediction of relapse for preB and T patients separately (data not shown). However, this may partly be attributed to the fact that a limited number of patients were available for each of these subtype-specific classifiers. Especially, when taking into consideration that there are many subtypes of ALL, we cannot expect to find a common expression profile for relapse for all subtypes. Thus, the low prediction accuracy of clinical outcome is not surprising. The chances for cure for individual patients will reflect the leukemic clone, the host and the treatment. A number of different leukemia-related biological features such as chromosomal translocations, multiple drug resistance gene activity, and deregulated apoptotic pathways may influence clinical outcome, and their impact may differ between different subsets of ALL. In addition, the strongest prognostic factor is treatment itself. Thus, patients are assigned to different risk groups that are offered different treatment protocols, the bioavailability and disposition of the anticancer agents may differ among patients, and both physician and patient compliance to the treatment protocols may significantly influence the chances for cure. Further improvement of the outcome prediction using DNA microarrays may necessitate analysis of both tumor samples and patient germline samples that allow identification of genetic polymorphism that influence drug disposition. Such data should be analyzed within biological well-defined subsets of leukemias treated by similar therapeutic strategies.

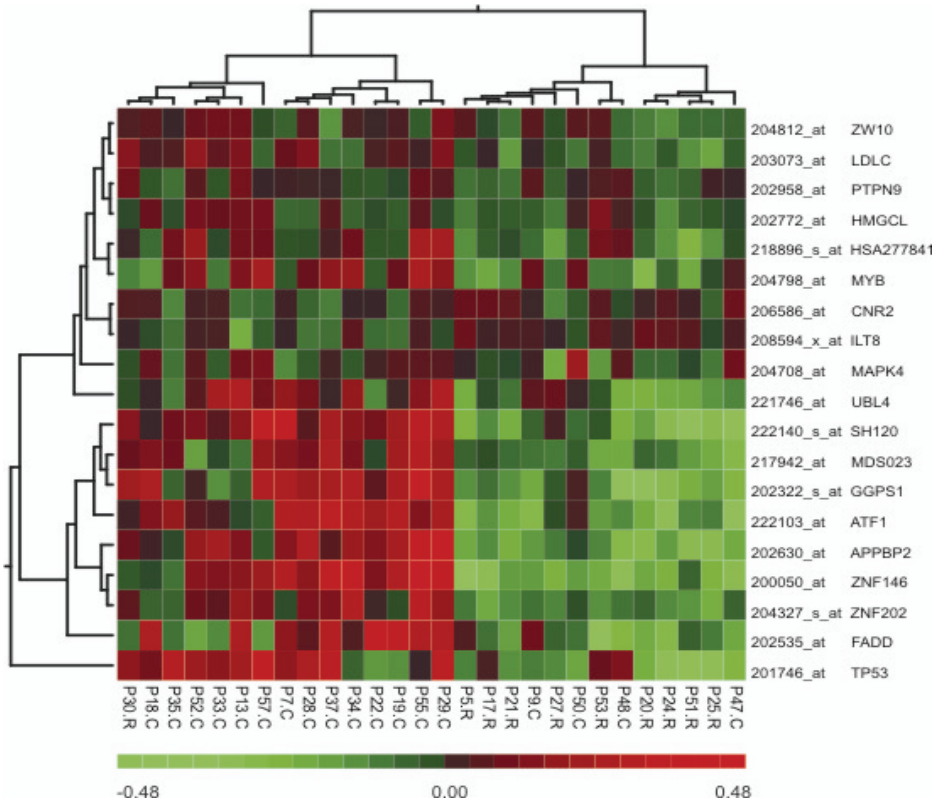

Figure 4-3. Horizontal: Hierarchical clustering of the 28 ALL patients with known 5-year outcome based on the gene expression levels of the 19 genes. For each patient the number and outcome, relapse (R) or CCR (C), are given. Vertical: Hierarchical clustering of the 19 genes found to be predictive for long-term outcome based on gene expression levels. The Affymetrix id and gene symbol is given for each gene. The color scale shows the logarithm of the gene expression value relative to the mean logarithmic gene expression for each gene.

Prediction of MRD level

Finally, prediction of high and low MRD level on day 29 after treatment initiation was attempted. Since the number of patients with available MRD data was limited, this prediction was only possible for pooled preB and T-ALL patient samples (26 patients) or for preB samples only (15 patients), where only samples with ≥75% leukemic blasts were included in both cases.

For the pooled preB and T patients a very low CC was obtained at LOOCV training on 2/3 of the data set (-0.05 to 0.23). On the contrary, for the preB samples only, a classification accuracy of 100% could be obtained during LOOCV training for the LDA and SVM methods on six PC based on the 120 top ranking genes. However, these promising results could not be tested on an independent test set due to the limited number of samples and when testing on the samples with <75% leukemic blasts only four out of the six preB samples with available MRD data were predicted correctly (66.7%).

If the MRD classifier was in fact as good as the results from the LOOCV training indicate, it would be highly useful in clinical settings for choice of induction therapy. Thereby, it would be possible to predict the treatment response on day 29 already at the time of treatment initiation, and for the patients with predicted high MRD, an alternative or more intensive therapy could subsequently be given. Another advantage of the MRD prediction is that all patients have received almost identical treatment during the first 29 days, which makes the classification results more easily interpretable and more general compared to the prediction of relapse where treatment during the first 5 years from diagnosis varied among the patients.

Microarray platform

All the analyses performed and the results obtained indicate that the use of the limited Focus Array platform does not result in a loss in classification performance compared to previous studies using the far more extensive U133A+B GeneChip set from Affymetrix for immunophenotype classification. On the contrary, by using the Focus Array, only the most validated genes are included in the analysis and much of the potential noise from probes against possible nonexisting human genes is therefore avoided. While this chip is less complex and easier to interpret than other chips on the market, it has other advantages, too. It is cheaper, requires smaller amounts of sample and is faster to run. Thus, the use of these chips for clinical gene expression profiling seems promising.

CONCLUSION

Altogether, our results indicate that gene expression analysis using DNA microarrays is a promising tool for prediction of relapse or treatment response in childhood ALL patients. Moreover, our immunophenotype classifier was able to classify correctly all but one sample from a previous independent study and thus, this technology shows potential for future clinical multicenter studies.

ACKNOWLEDGEMENTS

We thank staff members from the Section of Clinical Hematology and Oncology, Rigshospitalet, laboratory technicians from the Department of Clinical Immunology, Rigshospitalet, and people at Center for Biological Sequence Analysis, Technical University of Denmark, for their assistance. This work was supported financially by Knud Veilskov's Foundation, Ellen and Aage Fausbølls Health Foundation of 1975, Holger and Inez Petersens Foundation, Gangsted Foundation, Vilhelm Pedersens Foundation, Danish National Research Foundation, Danish Biotechnology Instrument Centre, Danish Center for Scientific Computing, Novo Nordisk, Novozymes, Carlsberg Foundation, The Danish Cancer Society (grant no.: 99 144 10 9132, 94-100-28, and 96-100-07), The Danish Cancer League, The Edith & Søren Kiilerich Hansen Family Foundation, The Emil and Inger Hertz Foundation, The Kornerup Foundation, The Lundbeck Foundation, The Medical Research Council in Denmark (grant no.: 9401011) and The Queen Louise´s Children´s Hospital Foundation.

Chapter 5 Paper II

Functional Associations by Response Overlap (FARO)

Henrik Bjørn Nielsen and Hanni Willenbrock

Center for Biological Sequence Analysis, Technical University of Denmark, DK-2800 Lyngby. Denmark.

ABSTRACT

Extensive utilization of gene expression data repositories is often restricted by limited comparability between experiments. Here, we present a novel and conceptually simple approach that overcome this restriction by deriving associations from overlaps in differentially expressed genes. This approach demonstrates an excellent capability to find biologically meaningful association between experimental factors even from independent studies. By this approach, published evidence of the roles of the *Arabidopsis* MAP kinase 4 could be confirmed and extended. Further results demonstrated that the approach is more powerful than existing methods such as co-expression analysis and has potential for cross-platform applications.

INTRODUCTION

During the past few years, the number of available transcriptomes has exploded due to recent advances in high-throughput techniques such as microarrays. However, only limited comparability between independent studies has been reported due to high variability between studies. Therefore, novel approaches for effective exploitation of the fast growing microarray data repositories are much needed.

In the areas of functional genomics, analyses of mutants are among the most successful approaches for deciphering the genetic makeup and the molecular functions of genes. However, the function of a mutant gene is rarely immediately obvious from the phenotype of the mutant. Even mutants brought to light under carefully designed screenings may be very difficult to dissect and may possess molecular functions not obvious at first. Global transcription analyses of mutants have, in several cases, proven valuable in determining or narrowing down the molecular function and affected pathways through careful expert examination of differentially expressed genes or statistical analysis of systematic gene annotations such as gene ontology (GO) or KEGG. Both of these approaches rely on high-quality annotations and expert knowledge of the biological system. Alternatively, use of gene expression microarray data for deriving associations between gene functions has previously been limited to analysis of co-expression and cluster analysis (Carlson*, et al.*, 2006; Hughes*, et al.*, 2000; Lee*, et al.*, 2004; Zhang*, et al.*, 2004). In addition, some advanced clustering approaches have been suggested, for example the utility of transcriptional consensus clusters derived from multiple cluster algorithms (Wu*, et al.*, 2002) or incorporation of prior gene function knowledge into the clustering procedure (Huang and Pan, 2006). However, a transcriptional response is typically restricted to a smaller subset of genes differentially expressed between the experimental conditions addressed. Therefore, inclusion of non-responding genes in such analyses is likely to introduce a significant amount of noise.

The underlying assumption in previous studies, as well as in the present, is that; if associated biological functions are affected, the response to these tends to be similar. Furthermore, we consider that the amplitudes of the responses may vary or be reversed, even when closely associated functions are affected. For example, if we have two proteins close to each other in a pathway, network or interaction complex, we expect overlapping sets of genes to respond when compromising either of them. However, If one protein is a repressor and the other an activator, the resulting responses are likely to affect overlapping gene sets in opposite directions.

Here, we show that response overlaps in terms of overlapping differentially expressed genes between gene expression microarrays experiments can be used for deriving associations between the factors (mutants, treatments, experimental conditions, etc.) of the experiments in question. We designate such associations 'Functional Associations by Response Overlap' (FARO). Applying this approach, we demonstrate that an *Arabidopsis* mutant, MAP kinase 4 (*mpk4*), may be functionally characterized in agreement with previously documented characteristics by assigning FARO associations to a compendium of *Arabidopsis* gene expression responses derived from a series of experiments originating from various laboratories. Furthermore, we demonstrate that the FARO approach is superior to co-expression analysis in associating genes accordingly to KEEG and MIPS annotations in the Rosetta Yeast compendium (Hughes*, et al.*, 2000). Finally, the approach demonstrates potential for cross-platform applications.

RESULTS

The FARO approach

To assign Functional Associations by Response Overlap (FARO) between an experimental factor and the experimental factors of a compendium of gene expression responses, a query

response, in terms of a list of differentially expressed genes, was compared to the responses of the compendium (Figure 5-1). The statistical significance of the overlap was estimated using Fishers exact test (Fisher, 1922) and overlaps were ranked thereby. The

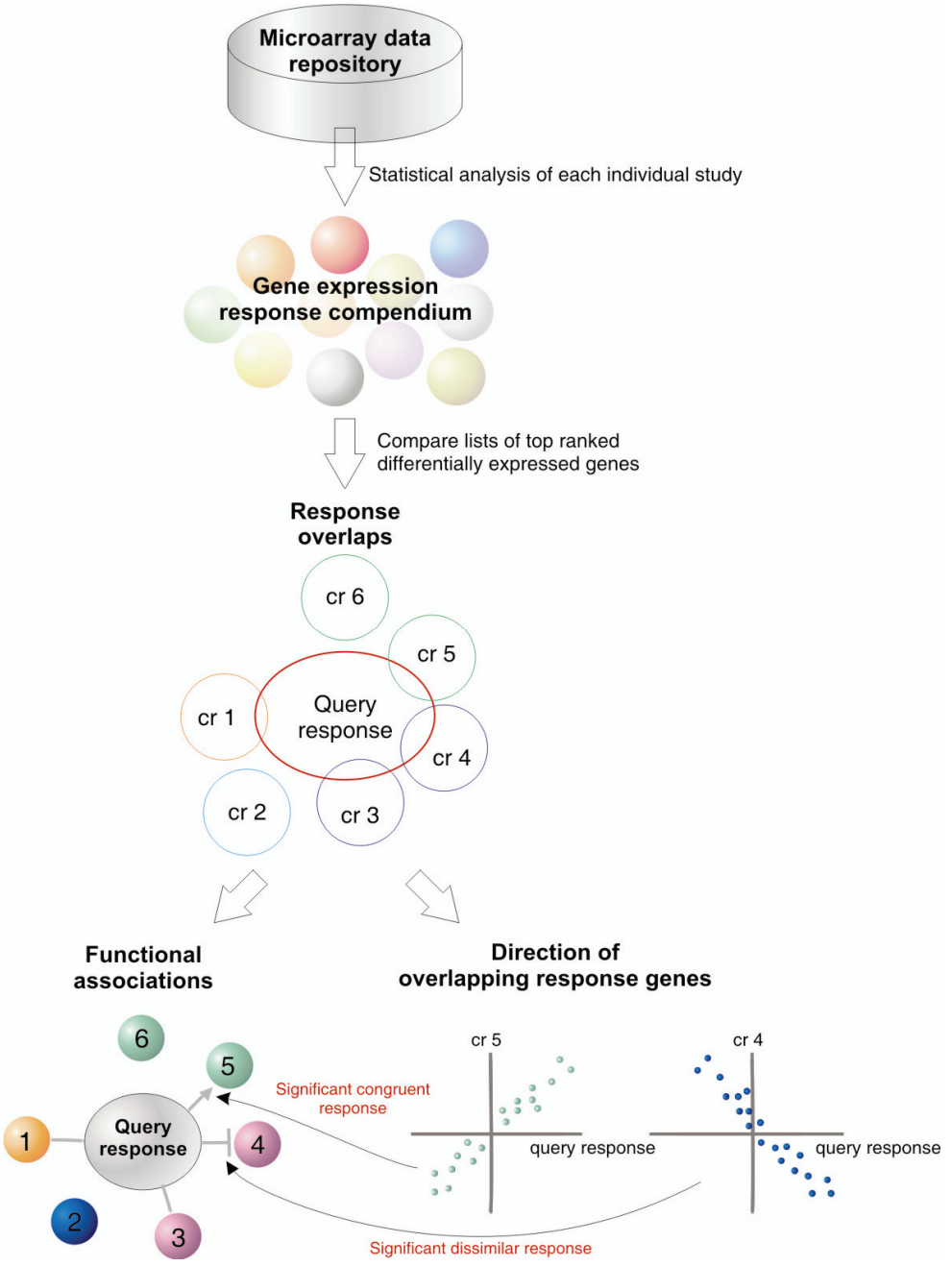

Figure 5-1. Overview of the FARO approach. Here, the query response demonstrates significant associations to compendium factor 1, 3, 4, and 5.

compendium of gene expression responses was constructed by first analyzing the individual studies in a collection of microarray studies in order to rank the genes by their significance of differential expression. Each individual experiment was analyzed separately; hence, individual samples were only compared directly within a study. Consequently, variations in experimental procedures between different experiments have no direct influence on the estimated responses. Assuming that the individual experimental designs were executed carefully, differentially expressed genes represent the response to the addressed factor and hence – provides an expression phenotype.

Functional characterization of the *Arabidopsis* knockout mutant *mpk4*

Figure 5-2 illustrates a FARO analysis of the *Arabidopsis* knockout mutant *mpk4* against a compendium of *Arabidopsis* gene expression responses. In Figure 5-2, the gene expression response of 245 experimental factors derived from this compendium, are represented as nodes. Edges from the central *mpk4* node to the compendium nodes represent the most significant response overlaps. In short, the associations above the threshold used in Figure 5-2 are almost all in agreement with previous knowledge and few if any of the factors below the threshold are known to be relate to *mpk4* (for details, see methods and supplementary Table 1).

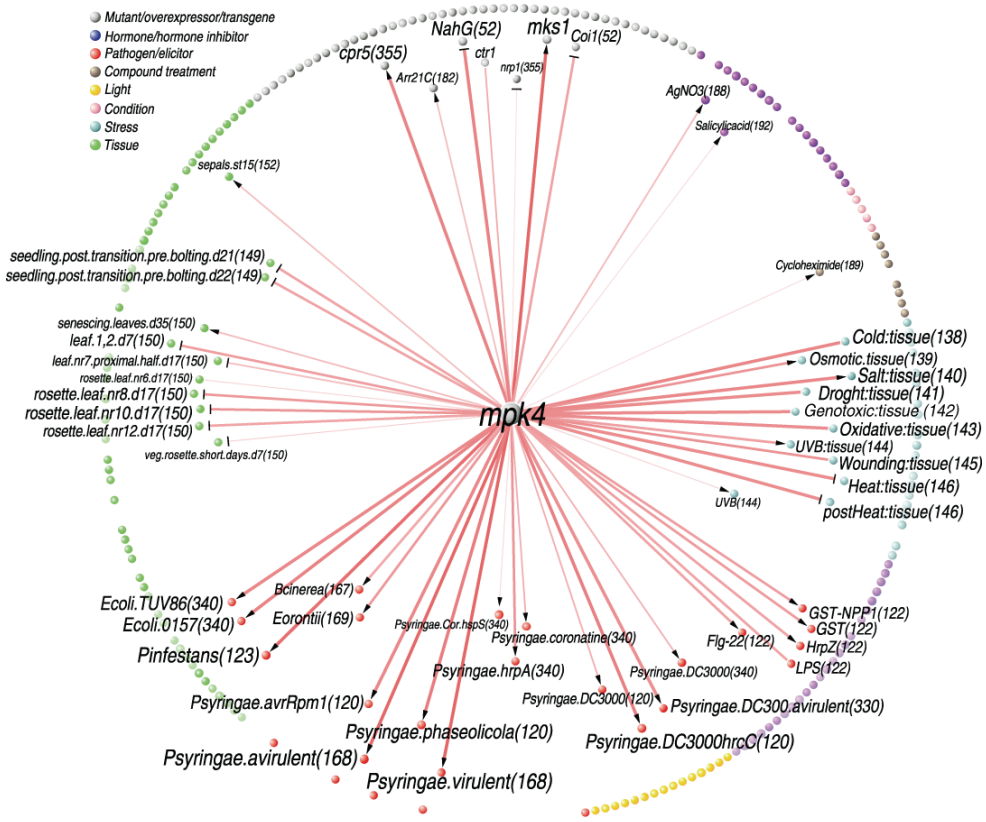

Figure 5-2. Graph illustrating a FARO analysis of *mpk4*. The thickness of the edges and the size of the node font indicate increasingly significant response overlaps. Only edges and names for factors with strong associations to *mpk4* are included. The edge arrows or bars indicate significant congruent or significantly opposite response direction of overlapping genes, respectively.

The analysis indicates a series of very strong associations between *mpk4* and plants subjected to various infections, including both virulent and avirulent pathogens as well as non-pathogens. In fact, 16 of the 18 infection studies in the compendium were among the significant associations. In extension to this, strong associations to the classical defense mutants NPR1 (non-expresser of pathogenesis-related genes 1 (Cao*, et al.*, 1994)) and CPR5 (constitutive expresser of pathogenesis-related genes 5 (Bowling*, et al.*, 1997)) were found. These findings are consistent with the previous observation that *mpk4* exhibit constitutive systemic acquired resistance (SAR) (Andreasson*, et al.*, 2005; Brodersen*, et al.*, 2006; Petersen*, et al.*, 2000).

MAP kinase substrate 1 (MKS1) interacts physically with MPK4 *in vivo* and is phosphorylated by MPK4 *in vitro (Andreasson, et al., 2005)*. In perfect agreement with this, one of the strongest associations to the *mpk4* knockout mutant is to the *mks1* overexpressor. Likewise, the strong association to coronatine-insensitive 1 (*coi1*), constitutive triple response 1 (*ctr1*) and the ethylene response-inhibiting agent, $AgNO_3$, all affecting the perception of the plant hormones jasmonic acid (JA) or ethylene (ET), also agrees with previously reported findings. Namely, that MPK4 plays a central role in plant's antagonistic mechanism between SA and both ET and JA (Brodersen*, et al.*, 2006; Petersen*, et al.*, 2000). Particularly, the epistatic relationship between *ctr1* and *mpk4* can be determined (Van Driessche*, et al.*, 2005) since global expression data is available for both mutants and the *ctr1/mpk4* double mutant (Brodersen*, et al.*, 2006). From this, we conclude that *mpk4* in some respect is epistatic to *ctr1* (for details, refer to supplementary Note 1).

In addition, the compendium contained 33 studies of *Arabidopsis* responses to various plant hormone treatments (AtGenExpress, 2006). Of these, only the response to SA treatment associates to *mpk4* despite the fact that SA is among the hormone studies with least statistical power due to only four samples in the study. While this is expected due to the elevated levels of SA previously observed in the *mpk4* mutant (Petersen*, et al.*, 2000), it illustrates that the approach, to some extend, can overcome limited statistical power in weak experimental designs of the underlying studies. Also, supporting the SA like expression phenotype of *mpk4* is the observed strong association to the *NahG* transgene plant (Gaffney*, et al.*, 1993), which expresses a SA hydroxylase that degrades SA (Buchanan-Wollaston*, et al.*, 2005).

The elevated SA levels in *mpk4* may also explain the significant response overlap between *mpk4* and various leaf types. Confounding with this, MPK4 is primarily expressed in leafs (Petersen*, et al.*, 2000). Here, the FARO analysis demonstrated remarkable consistency. Hence, of the 58 plant tissue specific experimental factors, the 16 addressing different leaf sections, stages or types all have associations to *mpk4* that rank 22 or better in respect to other tissues. The only other tissue with significant associations illustrated in Figure 5-2 is sepals (stage 15), which in many aspects resemble leafs. It is further noticed, that *mpk4* associates to seedlings post transition and prior to bolting (day 21, 22 and 23), a stage where SA plays a critical role and where the majority of the biomass is leafs.

Moreover, the congruence or dissimilarity in the direction of the observed responses also supports the association found by the response overlap. That is, several strong associations have very significant congruence or dissimilarity in terms of the direction of the gene expression response of the overlapping genes (Figure 5-3). Between *mpk4* and pathogen or elicitor treated plants, the congruence is close to 100%. The same is true for the overlap between *mpk4* and the *mks1* overexpressor (95%), whereas, not surprisingly, plants transgene for NahG have an inverted response (85%).

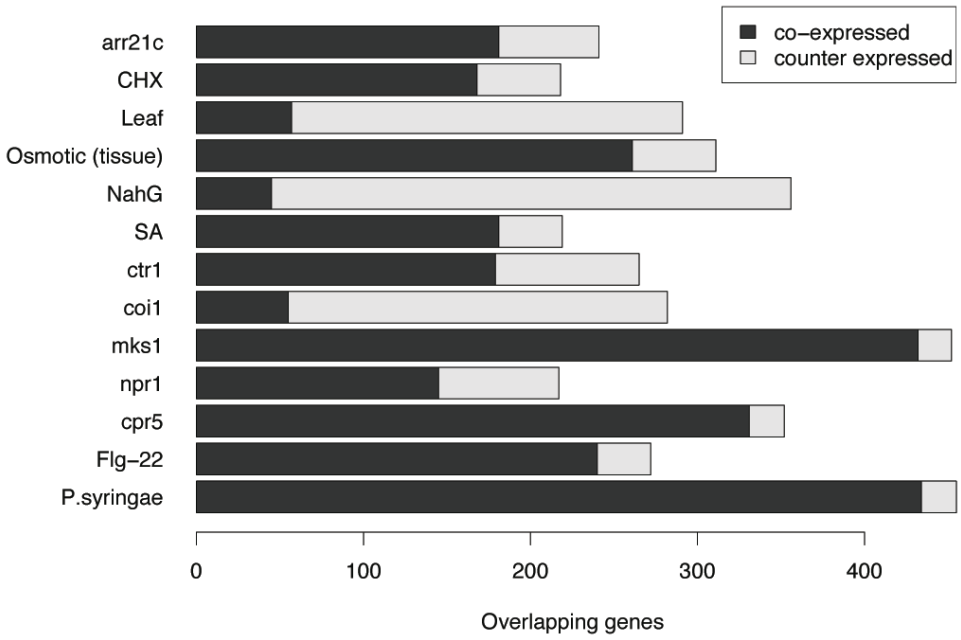

Figure 5-3. Bar plot of gene expression congruence.

Novel associations to *mpk4*

The *mpk4* mutant was first described by (Mizoguchi*, et al.*, 1993) in 1993 and has, since then, been subjected to intense investigations. It is, therefore, not surprising that most of the identified associations agrees with previous characterizations or derivatives of these. However, *mpk4* has not previously been associated with the *Arabidopsis* response regulator 21 (arr21) or the protein synthesis inhibitor cycloheximide (CHX) treatment. Transcriptional response to CHX typically indicates that there is a feed back loop from the protein to the mRNA stability or the transcription of the gene. The mRNA steady state level of *mpk4* itself does not change in response to cycloheximide. However, transcripts of the closely associated *MKS1* accumulate strongly as a result of CHX treatment (NASCArray 183).

The experimental factor Arr21C (Kiba*, et al.*, 2005) (NASCArray 183) corresponds to plants overexpressing the C-terminal DNA binding domain of ARR21 driven by the cauliflower mosaic virus 35S promoter. For a review on *Arabidopsis* response regulators, see (Mizuno, 2004). In contrast to the ARR21 knockout mutant, for which no phenotype is detected (Horák*, et al.*, 2003), the constitutive overexpressor demonstrates an extremely abnormal development with tissue resembling *in vitro* callus formations (Tajima*, et al.*, 2004). A second order FARO analysis, i.e. an analysis for overlap between the compendium and the *mpk4-arr21* overlap, characterized the *mpk4-arr21* association as a tissue specific stress and/or pathogen response with *Phytophthora infestans* being the predominant factor. A FARO analysis of ARR21C itself indicates strong associations to zeatin treatments, circadian rhythm, ARR22 over expression - in line with (Kiba*, et al.*, 2005; Mizuno, 2004), *P. infestans* as well as tissue specific stress response.

Multifactor FARO applications

The FARO analysis further indicates that MPK4 may be involved in or exhibiting a stress response. This is evident from strong associations to a series of stress responses, where tissue specificity could be derived (NASCArray 137-146). Interestingly, the overlapping genes demonstrate a strong tendency to respond to stress predominantly in the shoot (Figure 5-4). The 'single factor against all' FARO analysis, here, fails to distinguish clearly between the different tissue specific stress responses. The FARO approach, however, also allows us to investigate the relationships between multiple factors in one combined schema and thereby gain an overview. Doing so for all factors in the *Arabidopsis* compendium, a very tight cluster is revealed between tissue-specific responses to various stress conditions similar to what have also been reported for yeast (Gasch*, et al.*, 2000). Hence, comparison of top 1209 most significantly differentially expressed genes from each of the nine stress treatments (cold, drought, genotoxic, heat, osmotic, oxidative, salt, UV-B radiation and wounding) resulted in a collection of only 1858 different genes. Of these, 657 respond to all nine stress conditions. Surprisingly, the response direction is not conserved between the stress forms (Figure 5-5, average congruence 61%). This observation predicts that plants are unable to provide an adequate response to certain combinations of stress. This is interesting because it may contribute to understand what farmers and breeders already recognize, namely that combinations of abiotic stresses in the field (e.g. drought together with cold) cause the greatest losses to crop productivity worldwide (Mittler, 2006).

The multi factor FARO analysis further explains why *mpk4* associates to all tissue specific stress treatments rather than a subset. An analysis of congruence points toward *mpk4* as exhibiting an osmotic stress (Droillard*, et al.*, 2004), but also to some extend UV-B, salt or cold stress (Teige*, et al.*, 2004).

FARO has Cross-platform potential

Exploiting the vast amount of gene expression data available from central repositories may often be complicated by low cross-platform comparability. To investigate whether the FARO

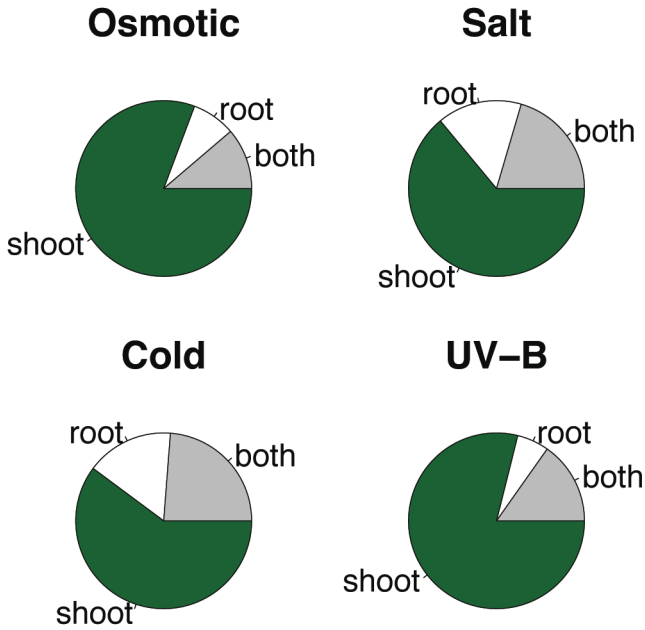

Figure 5-4. Pie chart, showing the fractions of mpk4 responding genes that are differentially expressed in shoot, root or both in response to the indicated stress types.

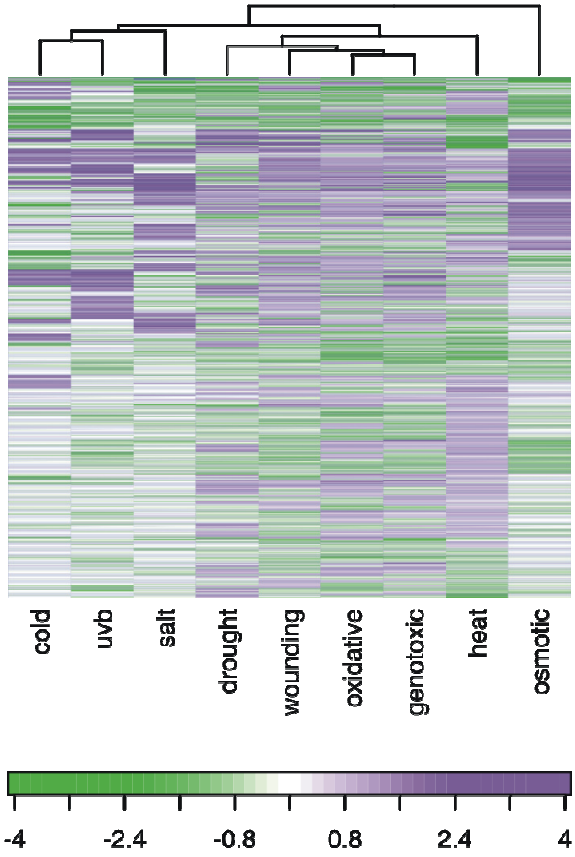

Figure 5-5. Expression profiles of the 657 tissue specific stress responding genes, through nine different stress conditions. The response is clearly different between these.

approach potentially may improve this comparability, gene expression responses were extracted from a number of cDNA studies (AFGC data) and compared to our Affymetrix ATH1 GeneChip® based *Arabidopsis* compendium. A predominant number of these comparisons demonstrated good compatibility. Most convincing was the cDNA gene expression response of 'white light treated' Colombia and Landsberg wild type *Arabidopsis* plants (NASCArray 250) that were highly associated (rank 3 and 4, respectively) to the '4 hours white light' compendium response phenotype (NASCArray 124) (for details, refer to supplementary Note 2).

Benchmarking on the Rosetta Yeast compendium

To validate the performance of the FARO approach in a more quantitative fashion, two benchmarking datasets were created from the Rosetta compendium of yeast gene expression profiles (Hughes*, et al.*, 2000). The Rosetta dataset consists of microarray gene expression data for a large number of yeast deletion mutants and a few chemical treatment experiments. The mutants within the Rosetta compendium may be associated by common KEGG category (71 mutant experiments) or by protein-protein interactions as annotated in MIPS PPI (30 mutant experiments), respectively.

Within each set, the strength of all associations was estimated by response overlaps. For the KEGG set, 39 correct associations were found stronger than the strongest false

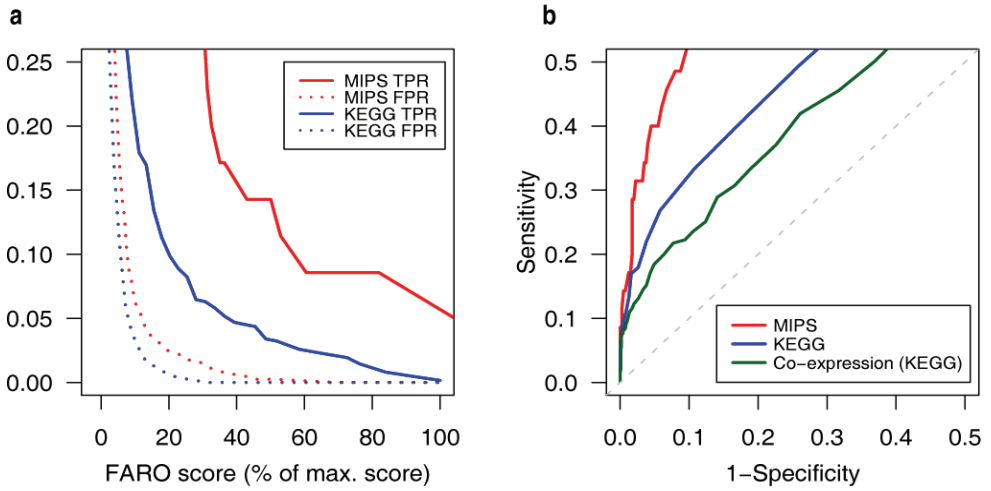

Figure 5-6. Benchmarking of the FARO approach. (a) True positive rate (TPR) and false positive rate (FPR) as a function of the relative FARO score for response overlap. (b) ROC curves of performance.

association and no false positive associations were found at FARO scores (minus $\log_{10}$ *p*-value) above 1/3 of that of the strongest association (Figure 5-6a).

Associations evaluated by use of the manually curated MIPS protein interaction annotations illustrates that the performance on this dataset, in fact, was even better than for the KEGG dataset (Figure 5-6b). Here, an extremely high initial true positive to false positive rate was observed in spite of the relatively low number of true associations in the MIPS set (MIPS: 35 true associations out of 436 possible vs. 619 true KEGG associations out of 2485 possible). Moreover, the eight chemical treatment experiments included in the Rosetta compendium consistently associated strongest to mutants in the pathway they affect (supplementary Note 3). From this, it is evident that the FARO approach certainly enrich for true associations. Furthermore, a comparison demonstrated that the performance for the FARO approach is clearly superior to a conventional co-expression analysis evaluated against corresponding associations in KEGG (Figure 5-6b).

DISCUSSION

Functional Association by Response Overlap (FARO) is a robust and conceptually straightforward approach for extracting information regarding the relatedness of experimental factors (knockout mutant, treatment, experimental condition, etc.) of microarray gene expression experiments, even when the experiments originate from independent studies from different laboratories. This allows for novel employments of available microarray data repositories and offers an advantage over existing analysis methods due to its robustness, simplicity and direct interpretability. A detailed characterization of the plant mutant *mpk4* is an example hereof. By comparing the result of a mutant/wild type gene expression study to a compendium of *Arabidopsis* gene expression responses, associations were derived to a meaningful subset of experimental factors within the compendium. The subset of *mpk4* associated factors together suggests that the mutant is involved in interactions with both virulent and avirulent pathogens, and that the mutant has a salicylic acid like expression profile. Furthermore, the mutant exhibits a gene expression response that resembles a shoot specific stress response. The subset also contained a series of

mutants involved in plant defense and/or perception of plant hormones (ET, JA) that are important for defense regulation. In short, this characterization of *mpk4* was consistent with previously reported characteristics and general understanding of plant biology. Moreover, two novel associations in relation to *mpk4* were identified: *Arabidopsis* response regulators 21 (ARR21C) and cycloheximide. Although, this analysis could not establish the exact relationship between ARR21C and MPK4, the association appears similar to a tissue specific stress response or pathogen response.

The analysis of the Rosetta Yeast compendium (Hughes*, et al.*, 2000) enabled a more quantitative benchmarking of the FARO approach. Held up against both KEGG and MIPS, the FARO approach demonstrated an astonishing ability to re-extract the groupings and protein interactions specified in these annotations. Furthermore, the FARO approach was clearly superior to the commonly applied method of co-expression analysis for deriving functional associations. Moreover, we show that the FARO approach is also applicable for cross-platform analyses.

For both analyses described here, the FARO approach demonstrated extremely high robustness toward experimental noise. Much of this robustness may be due to the indirect comparison of individual experimental results. That is, direct comparisons between microarrays are restricted to within experiment comparisons and only the outcomes of the statistical analyses in the form of differentially expressed genes are compared between experiments. Hence, the FARO approach benefits from the great care with which experimentalists have ensured comparability within their individual experimental designs. In addition, the extraction of differentially expressed genes serves as a feature selection step, enriching for genes that are characteristic for the given experimental factor. This reduces the amount of noise in the between factor comparison and consequently contributes significantly to the robustness of the analysis. Moreover, it renders the result more transparent.

Weak experimental designs or noisy experiments result in a less well-defined list of responding genes and tend to result in a smaller overlap than otherwise expected for truly associated factors. Such, weak experiments may result in false negatives, but is unlikely to result in false positive associations. Nonetheless, in the FARO analysis of the *Arabidopsis mpk4* knockout mutant presented here, cases of highly significant response overlaps were evident even to factors supported by weak data (e.g. salicylic acid).

While application of various clustering schemes also may provide a network of functional predictions for individual genes (Tavazoie*, et al.*, 1999; Wu*, et al.*, 2002), none of these measures are as easy interpretable as the FARO approach. Although, the interpretation of the FARO associations to some extend is up to the scientist, this approach offers an advantage over more abstract methods since the result may be further dissected into the actual genes that constitute the overlap. In fact, the interpretations of the FARO associations can be further investigated by any systematic analysis that may be applied to the list of overlapping response genes. Examples are GO-term overrepresentation analysis, chromosomal location bias analysis or even a second order FARO analysis. Consequently, the annotation of the overlapping genes may directly facilitate an interpretation of the functional association. Moreover, the response directions of the overlapping genes may add to the understanding of the relationship indicated by the association.

An essential advantage of the FARO approach over existing approaches utilizing co-expression measurements - apart from being more powerful - is its inherent ability to associate not only genes or proteins, but all kinds of factors that may be experimentally addressed, e.g. drug treatment and disease stages. Moreover, associations between analyzed experimental factors may be used to reveal clusters of factors in a functional association network that may be integrated with other data sources. Consequently, a FARO analysis enables exogenous factors to be associated directly to genotypes and as such unites bottom-up and top-down analysis approaches into a single association scheme.

METHODS

The *Arabidopsis* Compendium of Gene Expression responses

The Nottingham *Arabidopsis* Stock Center (NASC) compendium of global expression data (http://affymetrix.arabidopsis.info/) is a repository of microarray gene expression data from numerous *Arabidopsis* studies (Craigon*, et al.*, 2004). From this repository, we selected the AffyWatch II and III collection including the data from the AtGenExpress consortium: 21 studies of hormone treatments, 6 studies of pathogen infections, 3 studies of growth conditions, 9 studies of stress treatment, and 7 studies of different developmental stages. The set further include focused studies by 29 different authors from various experimental plant labs and two genotype studies of our own addressing the effect of the *Arabidopsis* knockout mutant of the constitutive triple response 1 (*ctr1*) gene (Brodersen*, et al.*, 2006; Kieber*, et al.*, 1993) and the MAP kinase 4 substrate 1 (*mks1*) overexpressor (Andreasson*, et al.*, 2005). From NASC, we further selected 6 cDNA studies from the AFGC data collection for cross-platform compatibility benchmarking.

Experimental factors were manually extracted from the description files, and each individual study was analyzed as a separate case with regard to the experimental factors in its design. Microarray data was pre-processed by RMA (Irizarry*, et al.*, 2003a; Irizarry*, et al.*, 2003b). Appropriate statistical tests (T-test, ANOVA) were used to obtain list of genes ranked by their significance of differential expression for the 245 different experimental factors. For a comprehensive list of included studies and their experimental factors, refer to supplementary Table 1.

KEGG and MIPS

By extracting mutants experiments that can be associated to other mutant experiments, within the Rosetta Yeast Expression Profile Compendium (Hughes*, et al.*, 2000), by common annotation in the Kyoto Encyclopedia of Genes and Genomes (KEGG: http://www.genome.jp/kegg/) or by protein-protein interactions as annotated in MIPS PPI (from the manually curated comprehensive *Saccharomyces cerevisiae* protein-protein interaction database at MIPS: http://mips.gsf.de/), two benchmarking sets was created. These sets comprised 71 and 30 mutant experiments, respectively. The KEGG category cell cycle was assigned to six additional genes, recently found to be involved in yeast cell cycle (de Lichtenberg*, et al.*, 2005). For the KEGG dataset, 619 proteins were associated by common KEGG category, among a total of 2485 possible between-mutant associations. For the MIPS dataset, 35 associations by MIPS interactions were present among a total of 435 possible between-mutant associations.

Statistical Significance

The statistical significance of the response overlap, in terms of overlap in differentially expressed genes, was estimated using Fisher's exact test (Fisher, 1922). Overlaps were ranked by minus $\log_{10}$ *p*-values (FARO score). The statistical significance of congruence in up/down regulation of overlapping genes was determined using an exact test in the binomial distribution (Conover, 1971; Hollander and Wolfe, 1973), where the hypothesized probability of success was fixed at 0.5.

With regard to an optimal number of top ranking genes to include in a comparison between experimental factors, we found it optimal to include genes that ranked higher than the median number of significant genes at a significance level lower than 0.05. While the inclusion of an increasing number of response specific genes will strengthen a true response overlap signature, including too many excessive genes may disturb the expression associations.

ACKNOWLEDGEMENTS

The authors wish to thank Lars Juhl Jensen, Anders Gorm Pedersen and Søren Brunak for critical reading of the manuscript. This study was supported financially by The Danish Center for Scientific Computing and The Danish Technical Research Council.

Part III

COMPARATIVE GENOMICS

Chapter 6 Paper III

A comparison study: applying segmentation to array CGH data for downstream analyses

Hanni Willenbrock[1] and J. Fridlyand[2]

[1]Center for Biological Sequence Analysis, Technical University of Denmark, DK-2800 Lyngby. Denmark.

[2]Department of Epidemiology and Biostatistics, University of California at San Francisco, San Francisco, CA 94143, USA

ABSTRACT

Motivation: Array comparative genomic hybridization (CGH) allows detection and mapping of copy number of DNA segments. A challenge is to make inferences about the copy number structure of the genome. Several statistical methods have been proposed to determine genomic segments with different copy number levels. However, to date, no comprehensive comparison of various characteristics of these methods exists. Moreover, the segmentation results have not been utilized in downstream analyses.

Results: We describe a comparison of three popular and publicly available methods for the analysis of array CGH data and we demonstrate how segmentation results may be utilized in the downstream analyses such as testing and classification, yielding higher power and prediction accuracy. Since the methods operate on individual chromosomes, we also propose a novel procedure for merging segments across the genome, which results in an interpretable set of copy number levels, and thus facilitating identification of copy number alterations in each genome.

Availability: http://www.bioconductor.org

Contact: Hanni@cbs.dtu.dk or jfridlyand@cc.ucsf.edu

Supplementary Information: http://www.cbs.dtu.dk/~hanni/aCGH/

INTRODUCTION

Development of solid tumors is associated with acquisition of complex genetic alterations. The particular types of genomic derangement seen in tumors reflect underlying failures in maintenance of genetic stability, as well as selection for changes that provide growth advantages. Comparative genomic hybridization (CGH) is a technique by which it is possible to detect and map genetic changes that involve gain or loss of segments of genomic DNA. Microarray formats of CGH provide copy number information at thousands of locations distributed throughout the genome. For a review of existing array platforms see (Pinkel and Albertson, 2005).

Genomic profiles greatly vary in their complexity. Depending on the instability present in the tumor and the selection environment, tumor cells may acquire alterations ranging from large segments with single copy number alterations to narrow homozygous deletions or high level amplifications. In many tumors the magnitude of measurable changes is reduced because the cell population is heterogeneous, thus frequently containing a significant proportion of normal cells. For a given genomic profile, the initial computational step is commonly referred to as *segmentation* and it involves reliable identification of locations with copy number transitions, or *breakpoints*. An example of how a genomic profile may look is illustrated in Figure 6-1 (A and B). Downstream analyses involve classifying the samples and finding copy number alterations that are associated with known biological markers. Thus, additional opportunities arise in the analysis of array CGH data compared to the established analyses of gene expression microarrays. In particular, one wants to make efficient use of the physical dependency of nearby clones.

Several segmentation methods have been proposed for partitioning clones into sets with the same copy number. Performances of a Hidden Markov Models (HMM) approach (Fridlyand et al., 2004), a non-parametric change-point method (DNAcopy) (Olshen et al., 2004) and a Gaussian model-based approach (GLAD) (Hupe et al., 2004) are compared in this article and these approaches are described in the *Method* section in detail. Additional segmentation methods involve building hierarchical Clustering-style trees along each chromosome (CLAC) (Wang*, et al.*, 2005), using a penalized likelihood criterion to estimate

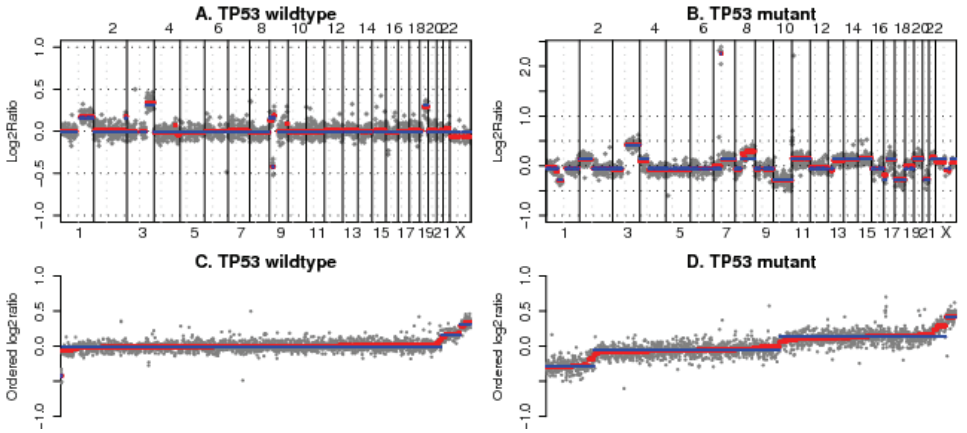

Figure 6-1. A+B: Genomic profiles for oral cancer samples segmented by DNAcopy and merged by MergeLevels. The observed log2-ratios are ordered according to their physical mapping along the genome C+D: Discretized log2-ratios by segmentation and merging. Log2-ratios are sorted according to predicted log2-ratios. Observed log2-ratios are shown in grey. Log2-ratios predicted by DNAcopy are shown in red and log2-ratios following the application of MergeLevels are shown in blue. The merged profiles yield better interpretability.

breakpoints (Picard*, et al.*, 2005); or applying an expectation–maximization-based method (Myers*, et al.*, 2004). Other proposals include a Bayesian model that uses parameterized prior distributions and a prior-less maximum a posteriori (MAP) technique to estimate the underlying model (Daruwala*, et al.*, 2004), a wavelet approach (Hsu*, et al.*, 2005) and use of a genetic local search algorithm to identify potential breakpoints and perform data smoothing (Jong, et al., 2004).

To date, most proposed segmentation methods have been evaluated on a simple simulation model and/or a small set of karyotyped Coriell cell lines containing a limited spectrum of one-copy number alterations. Some approaches to simulate array CGH data were to randomly and uniformly select breakpoints throughout the genome (Daruwala*, et al.*, 2004); assign loss, normal or gain according to a fixed probability transition matrix (Hupe*, et al.*, 2004); or to draw lengths of segments from a theoretical distribution and then assign either normal or one-copy gain (Hsu*, et al.*, 2005). Some additional variations have been used to make the simulation resemble real data, e.g. adding a trend parameter (Olshen*, et al.*, 2004) or simply adding random Gaussian noise to karyotyped Coriell cell lines (Fridlyand*, et al.*, 2004). However, many of these simulations produce unrealistically simple array CGH data involving few copy number changes. Moreover, until recently, no formal comparisons had been made among proposed algorithms except for (Hsu*, et al.*, 2005) who compare their method with a previous method in terms of its breakpoint detection ability. A very recent paper (Lai*, et al.*, 2005) describes an extensive study that compares the ability of a large number of methods to assign copy number alterations. However, they did not specifically examine the behavior of aberrations at the boundaries and their simulation model does not lead to sufficiently complicated genomic profiles. With the explosion of interest in copy number microarrays and of published computational approaches, there is a need for establishing a standard for systematic comparison of computational segmentation approaches. Here, we create a simulation schema that generates genomic profiles of comparable complexity to real life data. This is achieved by re-sampling segments from a large set of primary tumors. We use the simulated data to compare three original published segmentation methods that were chosen on the basis of free access and ability to output appropriate and comparable segmentation information.

All available methods operate on individual chromosomes. Thus, as a result of segmentation, profiles are partitioned into numerous copy number levels with varying means. This presents a problem when identifying regions of gain or loss. It is currently done on a clone-by-clone basis either by considering normal range using normal/normal hybridizations (Veltman*, et al.*, 2003; Wang*, et al.*, 2005) or by estimating the level of experimental noise for a given profile and considering all clones with values outside x times standard deviation range to be altered (Hodgson*, et al.*, 2001; Nakao*, et al.*, 2004) where *x* is frequently set to 3. In this paper, we present a novel level-merging algorithm. The merging step does not compromise detection accuracy of the breakpoints and is indispensable as it allows us to identify a genomic base level, if present, and thereby easily assign regions of copy number gain and loss to characterize individual genomes in terms of the number of copy number levels and to describe regions with respect to their relative copy number level.

Similarly, the physical positions of clones are ignored when identifying regions where the copy number is significantly associated with a phenotype of interest, e.g. a cancer subtype. A standard approach to the problem is to individually test each clone for the association on a “clone-by-clone” basis. In this paper, we evaluate the benefits of segmenting data before performing downstream analyses and introduce a novel idea of segmenting test statistics to identify entire genomic regions of interest, facilitating the interpretation of results. Thus we compare the downstream analyses such as testing and classification using simulated and real datasets by applying *clone-by-clone* and *region-based* approaches.

This paper is organized as follows: in the Methods section, we provide details on the three methods under comparison and on a novel level-merging algorithm. We also present novel approaches to incorporate segmentation into downstream analyses such as genome-wide testing and gain/loss detection. The simulation model and the primary tumor dataset are described in the Study Design section. In the Simulation Results section, we compare the ability of the three segmentation methods to detect breakpoints, identify altered regions and detect copy number associations with a phenotype of interest. In the Real Data Example section, we show a case study using a primary tumor array CGH dataset. Finally, in the Discussion and Conclusions, we discuss the limitations of the study and future work.

METHODS

The methods to be compared are available for the R statistical language from Bioconductor (http://www.bioconductor.org/) and have copy number level assignments as their main output.

aCGH: This package contains a HMM-based method that assigns clones to underlying states with constant copy number, thus allowing for determination of breakpoints. It fits an unsupervised HMM in which any state is reachable from any other state.

The state emission distributions are Gaussian with state-specific means and fixed variance. The re-estimation is done with a backward-forward algorithm. For a given number of states (k), the initialization is performed using k-means partitioning and transition probabilities are set to be proportional to the copy number distance between the pair of states. The number of states, k, is selected using a model selection criterion, e.g. Akaike Information Criterion (AIC) (Fridlyand, *et al.*, 2004). The HMM outputs two types of segmented values: predicted and smoothed $\log_2$-ratios, where the predicted values are state medians and smoothed values are state medians weighted by the estimated probability of being in each state. Here, we use aCGH version 1.1.4 and refer to the method as "HMM".

DNAcopy: This entirely non-parametric method is based on Circular Binary Segmentation (CBS), which is a modification of a change-point approach allowing for tertiary splits by connecting the two chromosomal ends. It splits the chromosomes into contiguous regions of equal copy number by modeling discrete copy number gains and losses. Using a permutation reference distribution, it bypasses parametric modeling of the data for assessing significance of the proposed splits (Olshen, *et al.*, 2004). The model selection is done in the forward way by repeatedly splitting each contiguous segment until no significant splits are found. As predicted values, DNAcopy outputs mean $\log_2$-ratios of each predicted segment. Here, we use DNAcopy version 1.1.0 and we refer to the method as "DNAcopy".

GLAD: This Gaussian-based approach detects chromosomal breakpoints by estimating a piecewise constant function that is based on adaptive weights smoothing (AWS). A local constant Gaussian regression model $Y_i = \theta(X_i)+\varepsilon_i$ is considered where the ε_i are independently and identically distributed as $N(0,\sigma^2)$, and $\theta(X_i)$ is a piecewise constant function, where the disjoint regions and the total number of regions are unknown. AWS is based on local-likelihood modeling and is an iterative algorithm that, around every location X_i, finds the maximal possible neighborhood in which the θ parameter is constant (Hupe, *et al.*, 2004). GLAD contains a procedure for merging segmented levels by iteratively removing excessive breakpoints and subsequently cluster segments across chromosomes to assign levels of copy number gain and loss (Hupe, *et al.*, 2004). We use the median of the original $\log_2$-ratios of each initial predicted level as unmerged GLAD data; and the median of the original $\log_2$-ratios for each predicted cluster as the GLADmerge values. Since we used GLAD version 1.0.1, it was modified slightly to optimize its performance in our comparison study (see supplementary material for details). We refer to this method as "GLAD", and "GLADmerge" for its level-merging procedure.

Merging of the copy number levels

As an alternative to model-based GLADmerge, which is not easily combined with other segmentation methods, we propose the following novel method (referred to as "MergeLevels") for merging copy number levels across the genome. The method merges two segmented levels if the distributions of the $\log_2$-ratios of the clones mapped to those segments are not significantly different or if the predicted level values are closer than a dynamically determined threshold. The algorithm performs the following steps: (1) Order distances between predicted levels using copy number scale ($2^{\text{level value}}$), where level value is the predicted value of the segment. (2) Starting from the smallest distance, test if two levels should be merged according to either of two criteria: (a) Wilcoxon rank sum test *P*-value > 1e-4 between observed values for two states or (b) distance less than a given threshold. States with <3 clones in each may only be merged based on the threshold criterion (b). (3) After a successful merge, step 1 and 2 are repeated until no two adjacent levels can be merged. (4) Step 1-3 is repeated for increasing thresholds. (5) For each threshold, we use Ansari-Bradley 2-sample test (Bauer, 1972) to determine whether the distribution of the current residuals (current merged values minus observed $\log_2$-ratios) is significantly different from the distribution of the original residuals (original segmented values minus observed $\log_2$-ratios). (6) Optimal threshold is chosen as the largest threshold where the Ansari-Bradley *P*-value > 0.05, i.e. where two types of residuals do not differ significantly. The Ansari-Bradley and Wilcoxon rank sum test significance thresholds were chosen based on an independent simulation data set. See supplementary information for details.

Using segmentation results for identifying regions of gain and loss, testing and classification

We test the application of segmentation followed by merging for identification of copy number alterations by defining the level of no alteration as the level with predicted $\log_2$-ratio closest to 0. Thus, all clones belonging to the remaining segments are either gained or lost. For comparison, we estimate experimental variability as the median absolute deviation (MAD) of difference between the observed and predicted $\log_2$-ratios and define threshold for determining gain and loss as 3 times MAD (factor of 2.5 is used in real data example).

We also introduce a region-based method for copy number association studies, which allows us to compute test statistics for entire regions rather than for individual clones. Student's t-test (equal variance) was used as a test statistic. For multiple testing corrections, we use a permutation based single-step maxT procedure to control the family-wise error rate (FWER) (Westfall*, et al.*, 1993). Thus, the reference distribution was estimated by repeatedly permuting a phenotype with respect to the copy number data, re-computing relevant statistics and recording a permutation absolute maximum. A total of 100 permutations were used for simulation data and 1000 for primary tumor data. Adjusted *P*-values were derived by comparing an observed statistic with the distribution of the permutation maxima. The significance was declared at maxT adjusted *P*-value < 0.05. Finally, we investigated whether using segmented values improved prediction accuracy for a phenotype predictor (e.g. *TP53* mutational status). For simplicity we used a linear discriminant analysis classifier with diagonal covariance matrix (DLDA) which has previously demonstrated very good performance in microarray studies (Dudoit*, et al.*, 2002a). Performance was assessed using leave-one-out cross validation for a varying number of input variables which were ranked by their F-statistic within each cross-validation.

STUDY DESIGN

Simulation model

The ratio profiles for array CGH data were simulated to emulate the complexity of real tumor profiles. To accomplish that, we segmented a primary breast tumor dataset of 145 samples

(Chin et al., unpublished data) using DNAcopy and randomly sampled copy number levels from the empirical distribution of segment mean values, where mean values were binned into the intervals less than -.4 (0 copies), between -.2 and -.4 (1 copy), between -.2 and .2 (2 copies), >.2 but < .4 (3 copies), between .4 and .6 (4 copies), and >.6 (5 copies). Note that defined intervals enrich for more extreme copy number changes and are not intended to present a realistic $\log_2$-ratio-copy number relationship but rather were constructed to increase complexity of the simulated genomes allowing for higher copy number diversity.

The lengths for normal levels (copy number 2) were assigned by sampling from the empirical length distribution of levels falling into the [-.2, .2] bin. Similarly, we assigned lengths to the altered segments by sampling from the length distribution for segments with levels outside that bin, i.e. altered segments, without distinguishing among length distributions with different copy numbers. Thus, the "true" breakpoints were derived and recorded. Each sample was assumed to be diploid and was assigned a proportion of tumor cells (P_t), which was drawn from a uniform distribution between 0.3 and 0.7 to resemble the proportion of tumor cells often seen in tumor biopsies and to incorporate this into our simulation model in a controlled way. Consequently, the expected $\log_2$-ratio for each clone was computed as $\log_2[(c \times P_t + 2\times(1- P_t))/2]$ where c was the assigned copy number.

Finally, Gaussian noise of mean 0 and varying variance were added to each sample. Appropriateness of the Gaussian distribution has previously been demonstrated using samples with limited number of alterations (Hodgson*, et al.*, 2001). Since hybridization quality and thus experimental variability of the samples may vary greatly, a sample-specific variance was added to each profile by drawing a standard deviation from a uniform distribution with range between 0.1 and 0.2. This variability reflects what is typically observed in the lower quality examples of UCSF BAC array hybridizations (data not shown). A total of 500 samples with 20 chromosomes containing 100 clones each were simulated with lengths of the edge segments truncated. This simulation was used to compare sensitivity and specificity of the three segmentation methods with regards to the breakpoint detection, to compare the two level-merging algorithms and to evaluate merging-based approach for identification of copy number alterations.

We created a different set of simulations to emulate real data sets with samples from two tumor classes. These datasets were used to specifically test whether the segmentation approach was more powerful than a univariate clone-by-clone approach for testing of copy number associations with a phenotype. For this simulation, we created 500 data sets each consisting of 20 samples drawn at random from either of two genome templates constructed as described above with a few exceptions: Without loss of generality the length of each genomic profile was reduced to 500 clones placed on just one chromosome and each sample was only assigned a probability of 0.7 of having a given aberration (i.e. in all samples approximately 30% of segments with copy number gains or losses were re-assigned a normal copy number of 2). Because the proportion of segments with copy number changes in each sample was decreased thereby, we doubled the probability of drawing altered segments from the copy number/segment length distribution. Segments with differences in copy number between the two classes were recorded. On average, each data set had 211 clones in such segments.

Breakpoint detection and merging

We compared the sensitivity and false discovery rate (FDR) of HMM, GLAD and DNAcopy to detect and correctly locate breakpoints for originally predicted segments as well as merged segments. Here, the sensitivity is the proportion of true breakpoints that were identified, whereas the FDR is the proportion of falsely predicted breakpoints among the predicted ones. Additionally, MergeLevels and GLADmerge were compared based on the precision of their predicted values relative to expected $\log_2$-ratios and the accuracy of identifying altered clones. We also considered all pair-wise combinations of the clones and

determined the proportion of clone pairs that were incorrectly assigned to the same or different states, referred to as discordant pairs.

Copy number association study: testing for differential copy number

A standard approach to identifying genomic regions associated with a particular phenotype, e.g. a cancer subtype, is to individually test each clone for an association, i.e. on a "clone-by-clone" basis. Here, comparisons were done between the standard approach and "region-based" approaches which included either performing *t*-tests on segmented $\log_2$-ratios or on the observed $\log_2$-ratios followed by segmenting the resulting statistic. Here, for HMM the segmented values corresponded to the HMM-smoothed values (weighted means of the state means). The performance of the methods was evaluated by sensitivity and specificity using a multiple testing corrected significance threshold and by comparing ROC curves.

Application to primary tumor data

Real array CGH data from BAC arrays with formalin-fixed primary oral squamous cell carcinomas (SCCs) (Snijders*, et al.*, 2005) were re-analyzed using the approaches introduced in this manuscript. The dataset consisted of 14 *TP53* mutant samples and 61 wildtype samples. The scientific question of interest was the comparison of genomic features between *TP53* mutant and *TP53* wildtype tumor samples. *TP53* status was determined by sequencing. Based on the methods' comparative performance assessment on simulated data, we chose to apply DNAcopy to the tumor data followed by merging with MergeLevels. The two tumor types were compared in terms of their overall genomic instability measured using the total number of breakpoints in each genome. We also assigned gain and loss status to each clone using threshold and segmentation based methods; and displayed an example of a typical disagreement between the two approaches. Furthermore, we tested for copy number associations with phenotypes using clone-by-clone and region-based approaches. Finally, we build a predictor of the *TP53* phenotype and demonstrate that providing segmented data as an input to a classifier greatly improves prediction accuracy estimated using leave-one-out cross-validation error rate.

SIMULATION RESULTS

Breakpoint identification and merging

From the output of each method, it is possible to infer predicted breakpoints. These were compared to the location of known breakpoints for the simulated data (15 breakpoints per sample on average). Figure 6-2 illustrates how the methods perform with regard to breakpoint detection at the correct position (w=0) or with an offset (localization error) of one or two clones, w=[1,2], within which a predicted breakpoint was assigned as correctly identified. As expected, the sensitivity increased while FDR decreases with larger accepted offsets. By merging, some true breakpoints were removed and consequently, sensitivity decreased slightly. Since many excessive breakpoints were removed as well, the FDR greatly decreased, especially for HMM and GLAD.

Of the compared methods, DNAcopy was most sensitive while having the lowest FDR (*P*-value < 2.2e-16, paired Wilcoxon rank-sum test). GLAD was least sensitive and HMM had the highest FDR. Not surprisingly, both merging procedures decreased FDR while reducing sensitivity for DNAcopy and GLAD (*P*-value < 2.2e-16, paired Wilcoxon rank-sum test). MergeLevels was less aggressive than GLADmerge in removing breakpoints resulting not only in higher sensitivity but also in higher FDR. Notice that DNAcopy is very sensitive and has a low FDR when applied alone, and thus benefits least from merging with regard to breakpoints. As an example, when accepting an offset of two clones, DNAcopy has a median sensitivity of 88% while having a median FDR of 6%. This corresponds to 1.8 missed breakpoints on average and 0.8 false breakpoints. Both HMM and GLAD had significantly more trouble identifying precise breakpoint locations than DNAcopy based on

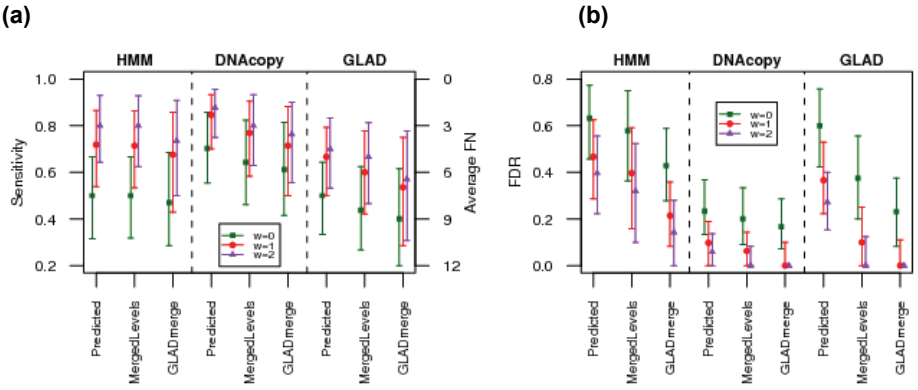

Figure 6-2. Results from simulation identifying breakpoints using either HMM, DNAcopy or GLAD or after removal of excessive breakpoints by MergeLevels or GLADmerge following segmentation. (a) It shows the median sensitivity and corresponding average number of false negatives (FN). (b) FDR for breakpoint detection with error bars depicting the interquartile range. Breakpoints were classified as correctly identified at its exact location (w=0) or if within an offset of 1 – 2 clones (w=1-2) of a correct breakpoint.

examination of the offset for predicted breakpoints. The comparative performance between methods was independent of the magnitude of the signal/noise ratio defined as the ratio of the proportion of the tumor cells to the variability of noise (P_t/sd), i.e. DNAcopy consistently performed the best while GLAD was least sensitive and HMM had the highest FDR (see supplementary material).

Additional studies indicated that the comparative performance did not change when introducing a larger proportion of smaller segments in the simulated data using empirical length distributions generated by either HMM or GLAD using the same primary breast tumor data set as for DNAcopy. However, further examination of the spatial resolution of the three segmentation methods revealed that HMM had the greatest power to detect the shortest segments with DNAcopy surpassing HMM for longer segments. However, DNAcopy had by far the lowest FDR for all segment lengths. GLAD consistently performed worse than the other two methods except for the detection of the longest segments (see supplementary information for details).

The merging step allows us to identify segments on different chromosomes corresponding to the same copy number. As an example, Figure 6-3 shows simulated data overlaid with known $\log_2$-ratios and with either HMM-segmented $\log_2$-ratios before merging (A) or after application of MergeLevels (B). For this example, merging clarifies the genomic profile and is able to correctly identify the base (no change) level as well as other copy number levels. This is also true for most other samples (see supplementary Figure S1). Note that for highly aberrant genomes, such a base level does not exist and it is not possible to infer gain and loss reliably.

To verify that merging performed reasonably, 4 different measures were considered: (1) sum of squared (SSQ) distance; (2) MAD between predicted $\log_2$-ratios and known $\log_2$-ratios; (3) accuracy of assigning copy number gain and loss; and (4) the proportion of discordant pairs (Table 6-1). Here, the SSQ distance and MAD were calculated with respect to the residuals between the observed (predicted, merged) values and the expected $\log_2$-ratios computed as a function of the copy number and the proportion of the tumor cells. All the clones with the true copy number not equal to 2 were considered to be “altered” and the “accuracy” was calculated as the proportion of the clones correctly assigned to altered or unaltered states.

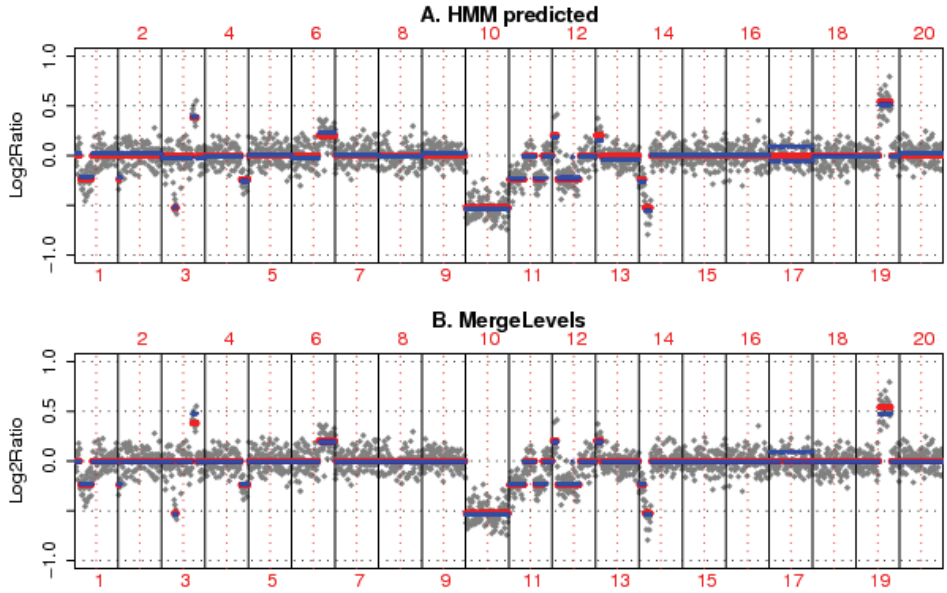

Figure 6-3. An example of simulated array CGH data with 100 clones on each of 20 chromosomes. The figure shows simulated $\log_2$-ratios in grey, ordered by position and chromosome. "True" $\log_2$-ratios were recorded from the simulations prior to the addition of Gaussian noise and are overlaid in red. (A) Predicted or merged $\log_2$-ratio levels are overlaid in blue for HMM predicted $\log_2$-ratios before merging and (B) after applying MergeLevels. Merging brings predicted values closer to their true copy numbers.

To calculate the proportion of the discordant pairs, all pair-wise combinations of the clones were considered and the proportion of clone pairs that were incorrectly assigned to the same or different copy number levels, referred to as discordant pairs was determined. Segmentation alone improved all 4 measures and both types of merging further decreased MAD, and as expected, further increased the accuracy of assigning copy number alterations

Table 6-1. Result using 4 difference performance measures for the array CGH analyses.

	Original	Predicted	*MergeLevels*	*GLADmerge*
SSQ distance	47.38	5.08	4.88	7.25
MAD	0.104	0.015	0.0044	0.0047
Accuracy	0.93	0.93	0.97	0.98
Proportion of discordant pairs	-	0.73	0.04	0.04

Median of each performance measure for original $\log_2$-ratios, HMM-predicted $\log_2$-ratios, and HMM-predicted $\log_2$-ratios merged by MergeLevels or by GLADmerge. Results are based on 500 simulated samples. SSQ and MAD are calculated with respect to the residuals between the observed (predicted, merged) values and the expected $\log_2$-ratios. Accuracy refers to the proportion of correctly assigned copy number alterations. The proportion of discordant pairs is the proportion of clone pairs that were incorrectly assigned to the same or different states relative to their true state.

The same four measures were used to assess the benefits from merging DNAcopy and GLAD segmented data and similar overall results were obtained. Moreover, to ascertain that our results and conclusions were not an artifact of our data simulation model or the DNAcopy segmentation results for determination of the empirical length distribution used in our simulation model, a second set of simulated data was generated using the model for high-rearrangement profiles as described by (Hupe, *et al.*, 2004) without their outlier addition. Their model led to much simpler datasets than the data simulated using our model, and consequently improved results for all methods. However, the comparative performance of the three methods was similar (see supplementary material for details).

and dramatically decreased the proportion of discordant clone pairs. No significant difference was observed between the performance of MergeLevels and GLADmerge except for the SSQ distance where the application of GLADmerge resulted in significantly larger squared error compared with those obtained when only applying segmentation. Thus, while some merging is beneficial – 'over-merging' may occur, which is also reflected in the sensitivity/specificity trade-off in Figure 6-2.

Copy number association power study: testing

We tested copy number associations of individual clones and of the genomic segments with the simulated binary phenotype, by testing whether a clone had a significantly different $\log_2$-ratio in samples from one subgroup (class 1 template) as compared with the $\log_2$-ratio for the same clone in samples from the other subgroup (class 2 template). Thus, we assessed the sensitivity and specificity of the clone-by-clone approach and the region-based approaches. The latter used segmented $\log_2$-ratios or segmented test statistics as described in Methods. For segmented test statistics, all clones assigned to the same segment would have the same test statistics. Here, the sensitivity is the proportion of known differential clones that were identified, while the specificity is the proportion of known non-differential clones identified as such.

ROC curves were used to evaluate the power to detect associations of the genomic alterations with a phenotype. Thus, in Figure 6-4, we plotted the median sensitivity over all datasets for small binned intervals of '1-specificity' corresponding to a sequence of different significance thresholds. It shows a combined ROC curve based on results from all 500 simulations. Application of any of the three methods resulted in greatly improved performance, which is evident by a higher sensitivity for any given specificity. Both region-based approaches are superior to the clone-by-clone (original) approach for all three segmentation methods with DNAcopy performing significantly better than HMM and GLAD (see also supplementary Figure S8). For HMM and GLAD, the family wise multiple testing cutoff was often driven by single extreme values. The levels were predicted correctly in most cases, but the cutoff derived from the maxT reference permutation distribution was too conservative, resulting in many distinct segments being classified as non-differential. We refer to (Westfall*, et al.*, 1993) and Supplementary material for details on the permutation-

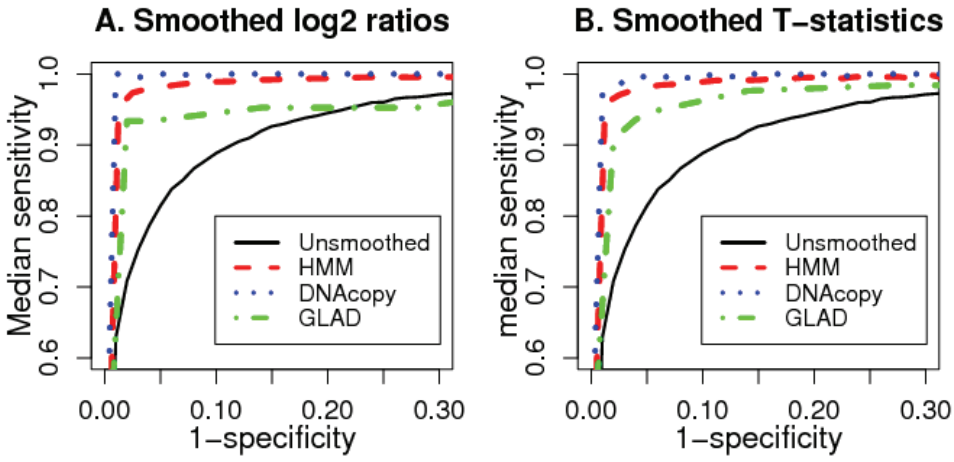

Figure 6-4. ROC curve illustrating the results from the copy number association power study. For varying thresholds, it shows the sensitivities versus "1-specificity" (false positive rate). Results are based on 500 simulations and binned median sensitivities are depicted. A: T-statistics based on segmented log2-ratios. B: Segmented T-statistics based on raw log2-ratios.

based single-step maxT procedure to control the FWER. Alternatively, when applying a gFWER(k)-controlling single-step common-cutoff augmentation procedure to define significance thresholds, the sensitivity increased greatly, especially for HMM and GLAD, whereas the specificity only decreased slightly (see supplementary information for details).

REAL DATA EXAMPLE: ORAL SQUAMOUS CELL CARCINOMA

Experimental data are inherently variable and segmentation involves bias/variance trade-off. We used DNAcopy and MergeLevels to re-analyze 75 oral SCC samples from a recently published study (Snijders*, et al.*, 2005) and demonstrate how segmentation may improve the analysis. The aim was to quantitatively compare the *TP53* mutant and wildtype tumor samples in terms of their genomic instability as measured by the number of breakpoints, to identify specific genomic regions associated with the *TP53* mutation and to use copy number data to predict mutation status of tumor samples.

Figure 6-1A and B illustrates profiles of a wildtype and a mutant sample showing original $\log_2$-ratios overlaid by segmented and merged $\log_2$-ratios. Figure 6-1C and D shows the effect of segmenting and merging, with merged and original $\log_2$-ratios sorted according to the values of predicted levels. A median of 17 and 28 breakpoints were identified in *TP53* wildtype and mutant samples, respectively. Thus, *TP53* mutant tumors were significantly more unstable genomically (*P*-value < 0.03). Similarly to simulations, merging only removed a small number of breakpoints for DNAcopy (final median of breakpoints: 16 and 24, respectively).

Now, recall that assigning alterations could be done either on a clone-by-clone basis by drawing a genome-specific threshold or by using merged segments. The difference between the proportions of autosomal clones declared to be altered was dramatic between these two approaches: median value of 5 versus 33%, respectively (see supplementary Figure S11). The large difference arose partly because of significant heterogeneity of the SCC samples combined with high experimental noise for paraffin-embedded tumors such as the samples

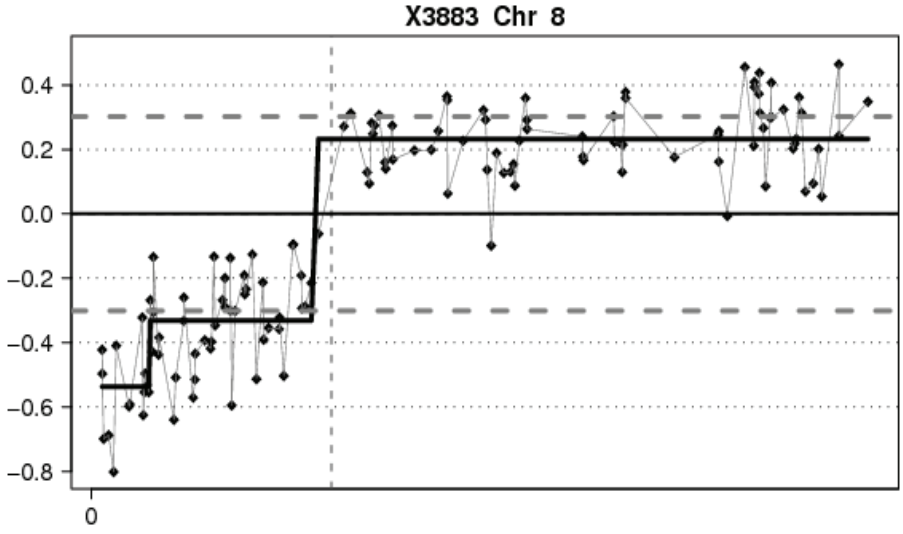

Figure 6-5. Identification of gained and lost clones using threshold-based and region-based approaches. A threshold for calling aberrations is indicated by a dashed horizontal line at -0.31 and 0.31. The solid curves indicate the segmented values. The baseline is at 0, thus all clones are altered according to the region-based approach with only a small proportion of clones altered with the threshold-based method.

in the SCC study. For these, a threshold-based approach is likely to miss many clones within real alterations. For instance, if the threshold is near a true copy number level, half of the clones with that copy number will be incorrectly declared unaltered. Figure 6-5 demonstrates the threshold-based and segmentation/merging-based methods for calling alterations. The dashed horizontal lines indicate the tumor-specific threshold. Thus, only clones above and below this threshold would be assigned an altered state. However, following segmentation and merging, assignment of the breakpoints agreed with the segmentation done using visual inspection and all clones on this chromosome can be assigned to an altered state. Of course, the threshold for the first method may be decreased; however, this would occur at the expense of a higher false positive rate as illustrated in Figure 6-6 (Note, the figure is based on results from breakpoint simulation study). This figure shows an ROC curve for assigning alterations on a clone-by-clone basis using original $\log_2$-ratios or those from a DNAcopy segmentation, and compares it with the results obtained by applying each of the level-merging algorithms. Segmentation by *DNAcopy* alone improves the results significantly; however, the merging approach is far superior to any threshold for the clone-by-clone approach illustrated by points to the left of both ROC curves.

Next, clones with significant differences in copy numbers between *TP53* mutants and wildtype samples were identified (see supplementary Figure S12 for resulting *t*-statistics). Only 29 clones were significantly differential for original $\log_2$-ratios. Using segmented $\log_2$-ratios for testing, 66 clones were found to be significant, and when using segmented T-statistics a total of 139 clones were identified as differential. These 139 clones were concentrated in segments on chromosome 8p, 8q, 11q and 18q. Compared to the 29 clones originally identified, only 4 were missing. They corresponded to a single clone on chromosome 2, and a small cluster of 3 clones separated by single non-significant clones on chromosome 10. Thus, the segments picked by region-based approaches produced more biologically meaningful results than traditional univariate testing method. Note that segmentation of the test statistic outputs entire regions of interest and thus eases the interpretation of the results.

Finally, to investigate whether noise reduction via segmentation would allow for more accurate classification, we constructed a predictor for *TP53* mutants versus wildtypes based

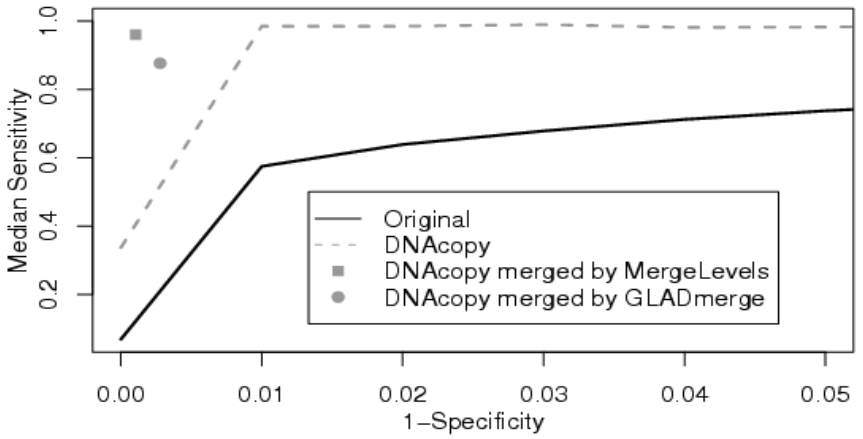

Figure 6-6. Simulation study results: ROC curve of calling gains and losses are shown for DNAcopy for varying $\log_2$-ratio thresholds. Median sensitivity based on 500 simulated samples is shown for bins of "1 minus specificity". Dots for merged results are shown for median sensitivity and median "1 minus specificity" for MergeLevels and GLADmerge.

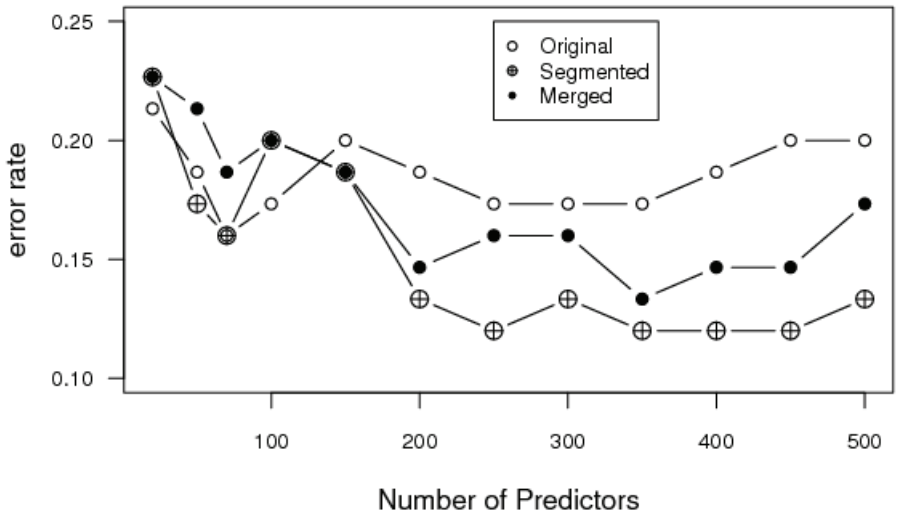

Figure 6-7. Misclassification error rate for DLDA classifier using original or segmented and merged data with an increasing number of variables re-selected at each leave-one-out cross-validation step.

on observed $\log_2$-ratios, predicted segmented $\log_2$-ratios, or segmented and merged $\log_2$-ratios as input to the classifier. Figure 6-7 illustrates the resulting error rate curve and demonstrates that segmentation decreases prediction error rate, while use of merged data result in inferior results compared with use of segmented data alone. However, this was to be expected as we have observed that merging occasionally removed a true breakpoint. It is also possible that the DLDA classifier is a suboptimal choice for the merged data which is discretized.

Discussion and Conclusion

Numerous methods have been proposed for segmentation of array CGH data, thus allowing for identification of copy number transitions. However, no comprehensive comparison or even basic evaluation of the performance of the proposed methods in terms of their breakpoint detection ability has been attempted; nor have the segmentation results been utilized in downstream analyses. Here, we have presented a realistic simulation study comparing three popular algorithms designed to segment array CGH data. Moreover, we have evaluated a novel merging algorithm linking segmentation output to downstream analyses. Finally, we have proposed a region-based testing algorithm and demonstrated its superior performance.

Our results have indicated that segmentation by any of the three methods aids downstream analyses of array CGH data. Of the methods under comparison, DNAcopy has the best operational characteristics in terms of its sensitivity and FDR for breakpoint detection. However, it should be noted that it is not able to identify single clone aberrations. While our comparison was limited to only three methods, albeit widely used, our study sets an example as a reference point for evaluating future algorithms. Also, our simulation model successfully emulates the complexity of real array CGH data. Moreover, our results agree well with the recently published results by (Lai*, et al.*, 2005), where they used a limited number of simple data simulations to demonstrate that DNAcopy generally performed better than GLAD and HMM with regard to detection of copy number alterations. Their results also indicated that HMM performed the best for small aberrations given a sufficient signal/noise ratio and GLAD did better than HMM for wider aberrations.

Merging of the resulting segments is of paramount importance in downstream use of the segmentation results. This aspect of the analysis has been largely ignored up to now except for a post-processing procedure in GLAD. We have introduced a novel merging algorithm and evaluated its performance against the existing one obtaining comparable results. We have also demonstrated that level-merging improves gain/loss detection, quantification of genomic instability for a tumor, and assignment of clones to the same copy number classes. However, small reductions in sensitivity brought on by merging may hurt some downstream analyses such as testing and classification since these analyses are very sensitive to the removal of even a few true breakpoints. Ideally, a merging step could be incorporated into the initial segmentation.

Currently, identifying regions with differential copy number is done using the same approaches as in transcriptional microarray studies without special consideration for known physical dependence. We have introduced a novel method for identifying such regions which explicitly uses segmentation results. The new approach delivers great improvements in detection power as demonstrated by our analysis.

In this paper we have demonstrated the superior performance of DNAcopy. However, an HMM approach is adaptable to perform a whole genome fit by doing constrained optimization of the segment means and variances across the entire genome, and thus consistently improving its performance with more observations. Moreover, in problems where simultaneous inferences need to be made, e.g. copy number and methylation, it may be of an advantage to use more model-based approaches such as an HMM and its extensions. Several papers on this have already been published, e.g. see (Zhao*, et al.*, 2004) and we are continuing working on evaluating and extending exciting methods to such problems.

Acknowledgements

The authors wish to thank Donna Albertson, Adam Olshen and E. S. Venkatraman for many useful discussions and Dan Pinkel for his ideas and critical reading of this manuscript. The authors would also like to acknowledge Peter Dimitrov for his assistance with implementation of the aCGH package. Finally, thanks to the anonymous referees for their useful comments. This work was partially supported by the grant NCI P50 CA58207 (JF) and The Danish Center for Scientific Computing and The Danish Technical Research Council (HW).

Chapter 7 Paper IV

Design of a Seven-Genome *Escherichia coli* Microarray for Comparative Genomic Profiling

Hanni Willenbrock[1], Anne Petersen[3], Camilla Sekse[2], Kristoffer Kiil[1], Yngvild Wasteson[2], and David W. Ussery[1]

[1]Center for Biological Sequence Analysis, Technical University of Denmark, DK-2800 Lyngby. Denmark.

[2]Norwegian School of Veterinary Science, P.O. Box 8146 Dep., N-0033 Oslo, Norway. [3]Diseased.

ABSTRACT

We describe the design and evaluate the use of a high density oligonucleotide microarray covering seven sequenced *E. coli* genomes in addition to several sequenced *E. coli* plasmids, bacteriophages, pathogenicity islands and virulence genes. Its utility is demonstrated for comparative genomic profiling of two unsequenced strains, O175:H16 D1 and O157:H7 3538 (Δ*stx*$_2$::*cat*) as well as two well-known control strains, K-12 W3110 and O157:H7 EDL933. By using fluorescently labelled genomic DNA to query the microarrays and subsequently analyse common virulence genes and phage elements, and perform whole genome comparisons, we observed that O175:H16 D1 is a K-12 like strain and confirmed that its ϕ3538 (Δ*stx*$_2$::*cat*) phage element originated from the *E. coli* 3538 (Δ*stx*$_2$::*cat*) strain with which it shares a substantial proportion of phage elements. Moreover, a number of genes involved in DNA transfer and recombination was identified in both new strains providing a likely explanation for their capability to transfer ϕ3538 (Δ*stx*$_2$::*cat*) between them. Analyses of control samples demonstrated that results using our custom designed microarray were representative of the true biology, e.g. by confirming the presence of all known chromosomal phage elements as well as 98.8 and 97.7 percent of queried chromosomal genes for the two control strains. Finally, we demonstrate that use of spatial information, in terms of the physical chromosomal locations of probes, improves the analysis.

INTRODUCTION

Escherichia coli is a complex group of bacteria comprising several intestinal and extra-intestinal pathogroups as well as commensal bacteria that are normal inhabitants of the intestinal tract of all warm-blooded animals and humans. Shiga toxin-producing *E. coli* (STEC) have emerged as important food borne pathogens causing diarrhea, hemorrhagic colitis and hemolytic uremic syndrome. Healthy ruminants such as cattle and sheep are regarded as the primary reservoir of STEC, which may be pathogenic to humans depending on their genomic content and combination of pathogenicity factors.

The Shiga toxins (Stx) are the main pathogenicity factors of STEC. Stx encoding genes (*stx*) are located on lamboid bacteriophages known as *stx* phages (Miao and Miller, 1999; Shaikh and Tarr, 2003). The *stx* phages are not only passive vectors for the dissemination of *stx*, but genetic entities where the characteristics of the phage itself may influence toxin production and thus, virulence of the host bacteria (Wagner*, et al.*, 1999; Wagner*, et al.*, 2001). Dissemination of *stx* genes by transduction is the most likely mechanism for intra- and intergenic spread of *stx* and subsequent development of new STEC. The host range of *stx* phages is highly variable, and phage transduction into *E. coli* and *Shigella* strains has been shown in different laboratory and animal experiments (Acheson*, et al.*, 1998; Gamage*, et al.*, 2004; James*, et al.*, 2001; Schmidt*, et al.*, 1999; Toth*, et al.*, 2003). Evidence for transduction of the bacteriophage ϕ3538 (*stx*$_2$::*cat*) from *E. coli* O157:H7 3538 (Δ*stx*$_2$::*cat*) (Schmidt*, et al.*, 1999) has been shown in porcine loops (Toth*, et al.*, 2003) and recently by feeding sheep with *E. coli* O157:H7 3538 (Δ*stx*$_2$::*cat*) (C. Sekse, H. Solheim, A. M. Urdahl, and Y. Wasteson et al., unpublished data). This latter experiment resulted in the isolation of a transductant, *E. coli* O175:H16 D1 from sheep feces. Consequently, *E. coli* O157:H7 3538 (Δ*stx*$_2$::*cat*) and *E. coli* O175:H16 D1 both contain Φ3538 (Δ*stx*$_2$::*cat*), a detoxified derivative of an *stx*$_2$ phage from a human *E. coli* O157:H7 type strain, in which most of the *stx*$_2$ is replaced by a chloramphenicol acetyltransferase gene, *cat* (Schmidt*, et al.*, 1999). However, little is known about host specificity of the *stx* phages, and similarities and differences of *E. coli* donor and recipient strains taking part in the transduction event.

Genome sizes among natural isolates of *E. coli* varies considerably, ranging by more than a million bp (Bergthorsson and Ochman, 1998). Furthermore, substantial diversity and genetic polymorphism exists even within the set of “core genes” found in most *E. coli* genomes (Anjum*, et al.*, 2003; Dobrindt*, et al.*, 2003; Fukiya*, et al.*, 2004; Ochman and Jones, 2000). Comparative genomic profiling using microarray chips designed to cover entire genomes is one strategy to obtain information about the variability between different strains of the same species and indications of horizontal gene transfer (Anjum*, et al.*, 2003; Fukiya*, et al.*, 2004; Ogura*, et al.*, 2006). Many commercial chips contain oligonucleotides from only one genome, such as the *C. jejuni* and *S. pneumoniae* chips and the *E. coli* K-12 chip (Ocimum Biosolutions, Affymetrix). The new *E. coli* Genome 2.0 array from Affymetrix covers four genomes; K-12 MG1655 and three pathogenic *E. coli* strains (CFT073, and two O157:H7 type strains). With at least seven *E. coli* genome sequences now publicly available, it is possible to design high density microarrays covering all seven of the fully sequenced genomes, in addition to selected genes for virulence factors, plasmids, phages, and mobile elements.

High-density oligonucleotide arrays provide large amounts of data. Consequently, automated analysis tools are necessary to identify probes corresponding to the presence or absence of specific genomic segments. Comparative genomic DNA hybridization experiments of bacterial genomes typically use either simple cut-off values to partition data points into present and absent DNA sequence segments, e.g. based on estimates from known reference hybridizations (Anjum*, et al.*, 2003) or based on standard deviation estimates (Gagne*, et al.*, 2005). However, the physical chromosomal position (mapping) of a probe is often ignored when analyzing this type of data. Statistical approaches for this

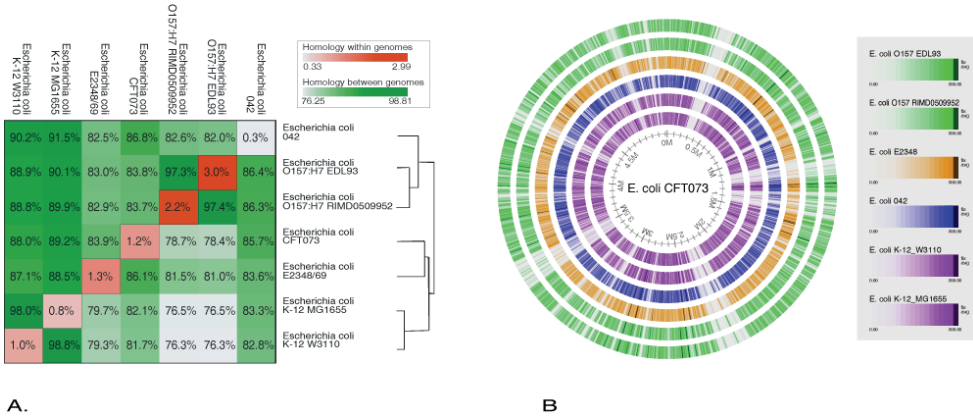

Figure 7-1. Comparison of sequenced E. coli genomes. A. Blast matrix comparing the 7 known genomes; the diagonal (red) represents internal homologues, and the other boxes (green) show the number and percentage of homologues for E. coli genomes in columns found in E. coli genomes in rows. On the right side: phylogenetic tree of the strains based on alignment of 16S rRNAs. B. Blast Atlas comparing the seven sequenced E. coli genomes. Here, the CFT073 genome is used as a reference and for each gene in this genome, the best match in the other genomes is plotted in the various circles.

purpose have been widely developed for copy number analyses in cancer research. These methods use statistics for partitioning probes into sets with the same copy number (corresponding to the same level of DNA). Recent advances and evaluation of their performance, demonstrate their usefulness and superiority compared to one-probe-at-a-time approaches (Willenbrock and Fridlyand, 2005).

In early 2005, when this study began, seven completely sequenced *E. coli* genomes were publicly available, including both pathogenic and non-pathogenic strains. These genomes vary in size from approximately 4.6 Mbp to 5.5 Mbp, and among these, there is a considerable amount of diversity as illustrated by the matrix shown in Figure 7-1A, which compares the coding sequence overlap between the seven different *E. coli* genomes. Moreover, next to the matrix, their relatedness is illustrated by a phylogenetic tree, based on their 16S rRNAs. The low relatedness of CFT073 to the other strains may also be illustrated by several large distinct chromosomal regions that contain genes unique to the CFT073 genome compared to other *E. coli* genomes (Figure 7-1B).

Here we describe the design and use of a high density oligonuclotide microarray covering seven sequenced *E. coli* genomes as well as several sequenced *E. coli* plasmids, bacteriophages, pathogenicity islands and virulence genes. The performance of this microarray is evaluated and its utility is illustrated for the hybridization of genomic DNA in order to compare two uncharacterised *E. coli* strains which have not been sequenced, with the seven known, sequenced *E. coli* strains. Recent advances in analysis of genomic DNA hybridization data were exploited. In particular, the physical mapping information was used to classify genes detected in the hybridization data into present and absent chromosomal segments.

MATERIALS AND METHODS

In this paper, we distinguish between the sequenced *E. coli* strains for which probes were designed on our custom made microarray chip and the genomic DNA from *E. coli* experimental strains that were actually hybridized to the custom designed microarrays.

Table 7-1. Overview of known sequenced *E. coli* genomes considered for the microarray probe design.

Strain	Isolate	Size bp	# easygene genes	# probes	Reference
K-12	MG1655	4,639,675	4,122	141,483	(Blattner, *et al.*, 1997)
	W3110	4,641,433	4,153	141,285	(Hayashi, *et al.*, 2006)
O157:H7	EDL933	5,528,445	4,990	139,445	(Perna, *et al.*, 2001)
	RIMD0509952	5,498,450	4,986	141,691	(Hayashi, *et al.*, 2001)
CFT073	-	5,231,428	4,653	127,261	(Welch, *et al.*, 2002)
O42	-	5,241,977	4,607	130,869	Sanger Institute, unpublished
E2348/69	-	5,074,835	4,599	124,103	Sanger Institute, unpublished

Microarray Probe Design

For probe design, the following sequences were considered (Table 7-1): whole genome sequences of seven *E. coli* strains - K-12 MG1655 (Blattner, *et al.*, 1997), K-12 W3110 (Hayashi, *et al.*, 2006), O157:H7 EDL933 (Perna, *et al.*, 2001), O157:H7 RIMD0509952 (Hayashi, *et al.*, 2001), CFT073 (Welch, *et al.*, 2002), 042 (Sanger Institute, unpublished), and E2348/69 (Sanger Institute, unpublished). These strains will be referred to as MG1655, W3110, EDL933, RIMD0509952, CFT073, 042 and E2348, respectively. Additionally, 104 *E. coli* genes involved in virulence (Dobrindt, *et al.*, 2003), 39 *E. coli* bacteriophages, 29 *E. coli* plasmids, 3 genomic islands from *E. coli* strain Nissle 1917 (Grozdanov, *et al.*, 2004) and 4 pathogenic islands from *E. coli* strain 536 (Dobrindt, *et al.*, 2002; Schneider, *et al.*, 2004) were extracted from Genbank release 146 (see supplementary material for a detailed list).

The probe design software, OligoWiz (Nielsen, *et al.*, 2003), was used to place probes both within unique areas and conserved areas of sequences shared by two or more open reading frames predicted by EasyGene (Larsen and Krogh, 2003). Conservation scores for aligned sequences were used by OligoWiz to place probes in the most conserved areas. Additional probes were placed in the 200 bp upstream regions of *E. coli* MG1655. A total of 271,693 *E. coli* specific probes were designed based on these sequences.

E. coli Experimental Strains and Culture Conditions

Experimentally, we examined the four *E. coli* strains - W3110 (Hayashi, *et al.*, 2006), EDL933 (O'Brien, *et al.*, 1984), O157:H7 3538 (Δ*stx*$_2$::*cat*) (referred to in the following as strain 3538) (Schmidt, *et al.*, 1999), O175:H16 D1 (referred to in the following as strain D1) (C. Sekse, H. Solheim, A. M. Urdahl, and Y. Wasteson et al., unpublished data), and bacteriophage Φ3538 (Δ*stx*$_2$::*cat*). The strains were grown overnight in Luria-Bertani (LB) broth with continuous agitation (Sambrook, *et al.*, 1989), and DNA was isolated using the Qiagen Genomic Tip 500/G (Qiagen, Hilden, Germany) and the Genomic DNA Buffer set (Qiagen). Independent triplicates of genomic DNA from each strain were prepared according to the manufacturer's protocol. The Φ3538 (Δ*stx*$_2$::*cat*) were induced from *E. coli* 3538 (Δ*stx*$_2$::*cat*) with mitomycin C, and DNA was extracted and purified as described by Muniesa et al. (Muniesa, *et al.*, 2003). Independent duplicates of the phage DNA were prepared.

Microarray Labelling and Hybridization

Seven micrograms of genomic DNA were fragmented with 0.7 Units of DNAseI (Amersham Biosciences) for 10-12 minutes at 37^{o}C in 1 x One-Phor All Plus buffer (Amersham Biosciences) to obtain fragments of 50-200 bp. Fragmented DNA was labeled according to the manufacturer's instructions (Affymetrix Inc.) for terminal labeling fragmented cDNA derived from mRNA for prokaryotic arrays. The labeled DNA was hybridized to custom-made NimbleExpress arrays (Affymetrix) for 15-17 hours at 45^{o}C. Standard protocols from Affymetrix for hybridization, washing and staining were followed using a hybridization oven, a Fluidics Station 450 and a GeneChip® Scanner 3000 (Affymetrix).

Data Analysis

Exact sequence matching was used to map each probe to specific chromosomal locations in the 7 *E. coli* design genomes and to specific locations within the 39 bacteriophage elements, 26 EDL933 or MG1655 genomic phage elements, 4 pathogenicity islands and 104 virulence genes. In the subsequent data analysis, a position dependent segmentation algorithm was employed to partition data points into present and absent sequence segments. For this, we used circular binary segmentation (Olshen*, et al.*, 2004) as implemented in DNAcopy developmental version 1.2.1 available for the R statistical language (http://bioconductor.org/). As recommended by the authors, the data was first smoothed and subsequently segmented. Segmentation was followed by merging the output with MergeLevels (Willenbrock*, et al.*, 2005). In cases where the algorithm was not able to find an optimal threshold, the threshold was fixed at the median absolute difference between segmented values assigned by DNAcopy and observed $\log_2$-intensities.

For the analysis of specific chromosomal genes, phage elements and virulence genes, only genes or phage elements to which at least 5 probes mapped were considered. Log_2-intensities were analysed using the above described segmentation approach. For chromosomal genes, it was safe to assume that a majority of them were present. Thereby, the present level was determined as the median value of merged segment means. For the analyses, segments with mean values at or above the level closest to the median for experimental strains and to the median of probes located in the known BP-933W phage sequence for ϕ3538 (Δ*stx*$_2$::*cat*) experiments were classified as present (BP-933W is the known sequenced equivalent of ϕ3538 (Δ*stx*$_2$::*cat*)). Chromosomal genes were considered present if at least two of the three replicate experiments had present probes spanning at least 90 % of the covered gene sequence. Virulence genes and phage elements were inspected visually if they met one of the following three criteria in at least one analyzed sample: (1) at least 10 percent of sequence in present segments, (2) a continuous segment spanning at least 100 bps (3) at least 5 percent of present probes in the largest segment.

Hierarchical cluster analysis was based on measurements for all probes using Pearson correlation distances and complete linkage. To reduce experimental data for replicate experiments into one set of probe values for each experimental strain, a one sided Student's T-test was used to estimate a P-value between 0 and 1 for each probe, where a P-value close to 0 corresponded to a probe being significantly below the median intensity for the 3 replicate experiments for a given experimental strain, and consequently, significantly absent. Corresponding sets of theoretical binary P-values of either 0 or 1 were constructed for each of the 7 known *E. coli* strains, where 0 corresponded to no match anywhere in the sequenced genome, and 1 corresponded to at least one match.

Atlases were created using the Genewiz software (Pedersen*, et al.*, 2000). The blast atlases were constructed as described previously (Skovgaard*, et al.*, 2002). Common *E. coli* genes as well as strain specific genes were identified by BLASTP version 2.2.11 (Altschul*, et al.*, 1997), using 1e-10 as E-value cut-off and minimum alignment ratio of 0.75 (ALR: the alignment length divided by the length of the longest compared gene).

Data Availability

The microarray data have been deposited in the Gene Expression Omnibus database (GEO: http://www.ncbi.nlm.nih.gov/geo/) with the series accession number GSE4690. Supplementary information and figures may be found at http://www.cbs.dtu.dk/~hanni/Ecolichip1.

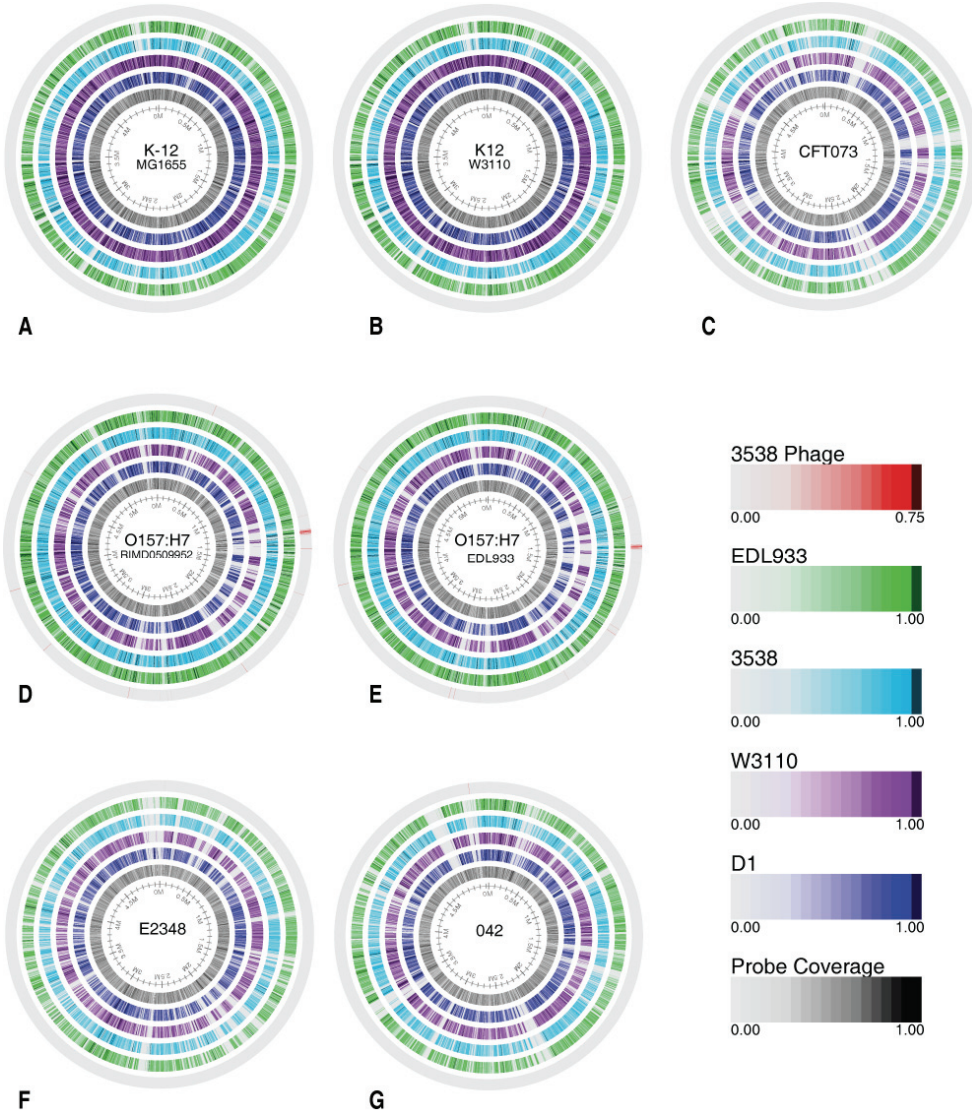

Figure 7-2. Hybridization Atlases, visualizing median probe intensities for the 4 experimental strains and the phage Φ3538 (Δ*stx*$_2$::*cat*), mapped to the 7 known *E. coli* genomes. Log intensities were

RESULTS AND DISCUSSION

Visualization

Probe intensities were visualized in whole genome hybridization atlases, as shown in Figure 7-2, for each of the seven known *E. coli* genomes considered in this study. Probes were mapped to each of the seven fully sequenced *E. coli* strains by sequence similarity to the known sequence and the resulting probe coverage patterns are visible in the innermost circle (grey). The probes appeared well distributed for all strains while several distinct regions existed for individual strains. Corresponding median intensities were visualized for each experimental *E. coli* strain (2nd to 5th circles) as well as ϕ3538 (Δ*stx*$_2$::*cat*) phage

experiments (outermost circle). It was possible to identify true distinct regions while neglecting gaps in the outer circles that were due to poor probe coverage. This allowed us to identify areas unique to each experimental strain. For instance, all experimental strains were missing large portions of the CFT073 genome at various sites (Figure 7-2C). As expected from the comparison to known sequenced genomes in Figure 7-1B, these regions were also missing in the two known *E. coli* strains included as control experiments (W3110 and EDL933).

Many probes were unique for individual strains, as evident by several gaps in the measured intensities. For example, EDL933 mapped to W3110 has gaps, whilst W3110 probe intensities covered the entire W3110 genome (Figure 7-2B). Moreover, both intensity patterns for the two control strains, EDL933 and W3110, closely resembled their corresponding probe coverage patterns, as expected (Figure 7-2B and Figure 7-2E).

Experimental strains D1 and 3538 (Δ*stx*$_2$::*cat*) possesses the same bacteriophage, ϕ3538 (Δ*stx*$_2$::*cat*), which has been transferred from strain 3538 (Δ*stx*$_2$::*cat*) to strain D1 (C. Sekse, H. Solheim, A. M. Urdahl, and Y. Wasteson et al., unpublished data). This phage is very similar to the BP-933W phage element in *E. coli* EDL933 located at ~1.33 to ~1.39 Mbp, and a region of extremely high similarity is clearly visible in the atlases for both *E. coli* O157:H7 type strains (red outermost circles in Figure 7-2D and E). A zoom of the BP-933W phage area on EDL933 clarifies the closer resemblance of phage ϕ3538 (Δ*stx*$_2$::*cat*) with the corresponding phage element from 3538 (Δ*stx*$_2$::*cat*) and D1, rather than with the BP-933W phage element from *E. coli* EDL933 (Figure 7-3).

Strain Comparison

To investigate how the *E. coli* D1, W3110, 3538 (Δ*stx*$_2$::*cat*) and EDL933 strains were related to each other, an unsupervised cluster analysis of all 3 replicate experiments for

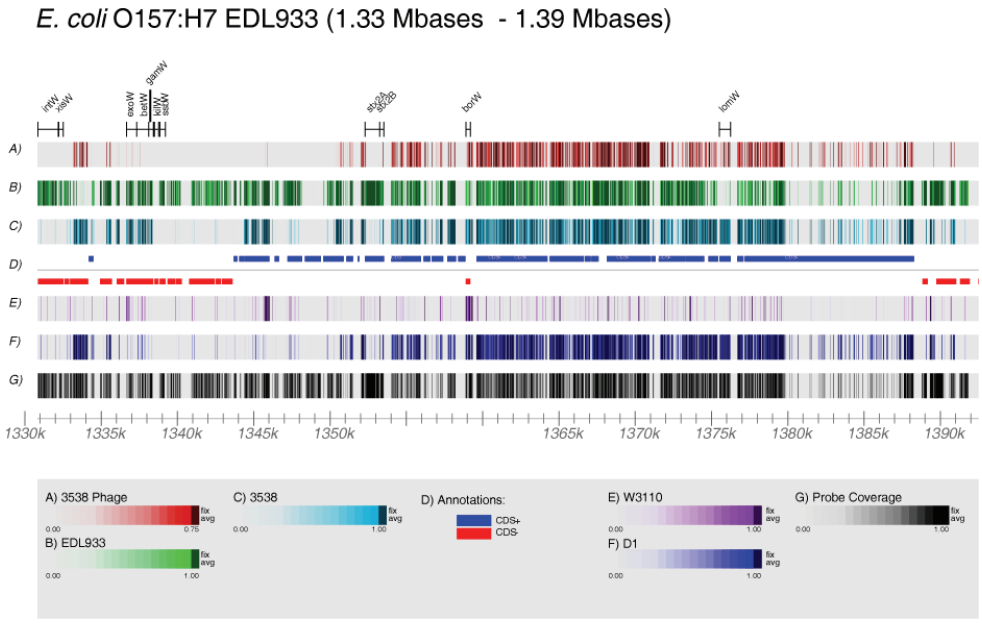

Figure 7-3. Zoom of the BP-933W phage area on EDL933 with known genes indicated. Intensity measurements for EDL933 experiments (C) are clearly as expected from the probe coverage pattern (G) and both the experimental strains 3538 (Δ*stx*$_2$::*cat*) (C) and D1 (F) has intensity patterns clearly similar to that of ϕ3538 (Δ*stx*$_2$::*cat*) (A).

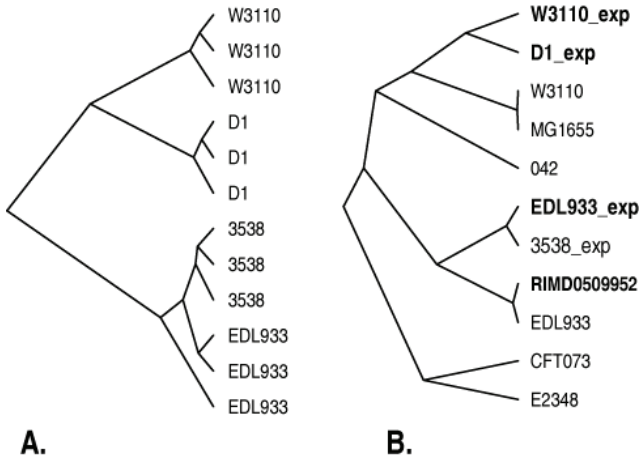

Figure 7-4. (A) Hierarchical Cluster Analysis based on all probe intensities for the 3 replicate experiments for each of the 4 experimental *E. coli* strains. (B) Cluster analysis based on continuous p-values between 0 and 1 (T-test) for the 4 experimental *E. coli* strains (indicated in bold and with postfix: "exp") and binary values of either 0 or 1 based on theoretical probe absence or presence for all 7 sequenced *E. coli* strains considered in this study. Both cluster analyses are based on Pearson correlation distances and complete linkage.

each of the four experimental strains was performed, based on intensities for all probes on the microarray, as shown in Figure 7-4A. Here, D1 appears closer related to W3110 than to the other experimental strains while 3538 clusters with EDL933. Since the 3538 (Δstx_2::*cat*) strain has been serotyped as 0157:H7 (same as EDL933), we expected these two to be more closely related to each other than to the other experimental strains.

Because experimental data to characterize the strains further with regard to their resemblance to other *E. coli* strains were not available, we attempted to construct theoretical data, based on all seven known *E. coli* strains considered in this study (see methods for details). In this cluster analysis, control strains clustered as expected from their phylogenetic tree based on their 16S rRNAs (Figure 7-1A), although experimental noise was significant. Thus, based on experimentally determined probe values, the K-12 type strains (and O157:H7 type strains) were more closely clustered than with corresponding theoretical strains (Figure 7-4B). However, since the two K-12 strains, MG1655 and W3110 are almost indistinguishable in terms of their genomic sequence, this result was expected. Moreover, this analysis confirms that D1 is much more related to K-12 strains than to other known strains such as E2348 and CFT073.

Analysis of Strain D1 and Strain 3538 Genes

Among the 7 *E. coli* strains used for chip design, 3475 genes were found to be in common by Blast analysis (the complete list may be found in the supplementary material). Of these *E. coli* 'core' genes, ~3100 were identified in D1 and 3538 samples, indicating that the D1 and 3538 strains have slightly different subsets of *E. coli* core genes than the 7 *E. coli* design strains. This is consistent with the observation that the number of *E. coli* core genes tend to decrease as the full genomic sequences of new *E. coli* strains continue to become available (Tipmann and Ussery, unpublished results). A thorough discussion of *E. coli* core genes will be presented elsewhere since there is now at least 20 sequenced *E. coli* genomes available for such an analysis (Binnewies*, et al.*, 2006).

Among non-core genes, we searched for genes specific to each of the 7 *E. coli* design strains, where genes specific to either the K-12 or the 0157:H7 type strains were combined

into lists of K-12 and 0157 specific genes (i.e. present in either W3110, MG1655 or both, but differing from non K-12 strains; or present in either EDL933, RIMD0509952 or both, but differing from non-0157:H7 strains).

D1 genes specific to the 7 *E. coli* design strains were analyzed further. A total of 150 K-12 specific genes were found, supporting the previous finding that D1 resembles the K-12 strains. Furthermore, the finding of 210 genes specific to 0157:H7, indicates that D1 has acquired many 0157:H7 specific genes in accordance with the known transfer of the 3538 phage. Although D1 has many 0157:H7 specific genes, it has much less than the known 0157:H7 type strain *E. coli* 3538, for which a total of 543 genes specific to 0157:H7 were identified.

Identified D1 and 3538 genes specific to the 7 *E. coli* design strains were annotated by Blastp comparison to the NCBI's non-redundant database (nr) (http://www.ncbi.nlm.nih.gov). Predicted genes, for which a reliable match was found, were examined closer (refer to supplementary for a detailed list). For the D1 genes specific to the 7 *E. coli* design strains, we identified a large number of 0157:H7 chromosomal phage genes (discussed further in the 'Benchmarking' section), while the majority of K-12 specific genes were unrelated to pathogenicity, e.g. genes in the phenylacetic acid degradation operon, genes involved in energy/metabolism, and membrane proteins. The CFT073 specific genes mainly consisted of genes involved in metabolism (e.g. pyruvate dehydrogenase) or translation/transcription (e.g. rpoC, RpoD, DNA polymerase 1). Among the 042 specific genes were 8 putative phage elements, a putative IS element, two putative transposases and the *cat* gene. The latter was expected since a similar *cat* gene is present in Φ3538 (Δ*stx*$_2$::*cat*). Finally, a whole series of conjugal transfer proteins (7 of TrbA - TrbJ, 17 of TraB - TraQ) were identified among E2348 specific genes. These genes comprise a large section of the E2348 chromosome and are clearly visible for D1 samples in the 5 Mb region of Figure 7-2F. This demonstrates that D1 is susceptible to foreign DNA, and might have facilitated the uptake of the *E. coli* 3538 genomic phage, Φ3538 (Δ*stx*$_2$::*cat*). Moreover, we found that for strain 3538, all but two of its 30 genes specific to *E. coli* E2348 were transposases, indicative of elevated levels of recombination in *E. coli* 3538 compared to other 0157:H7 type strains. This further provides a likely explanation for the observed transfer of Φ3538 (Δ*stx*$_2$::*cat*) from strain 3538 to strain D1.

D1 Pathogenicity

To further characterize the experimental strains - D1 and 3538 (Δ*stx*$_2$::*cat*), the data for probes covering known virulence genes and phage elements were analyzed. A minimum of 5 probes mapped within 96 of the 104 known virulence genes; and within all 39 non-MG1655 and non-EDL933 bacteriophages and all 4 pathogenicity islands.

After removal of sequences absent in all samples (see methods for details on filtering criteria), the numbers of sequences were further decreased to 21 (of 96) virulence genes, 14 (of 39) bacteriophages, and 2 (of 4) pathogenicity islands.

Results were illustrated for these remaining virulence genes (Figure 7-5) and for phage sequences + pathogenicity islands (supplementary Figure S1), by which present and missing fragments were clearly visible. While W3110 had few virulence factors, EDL933 had many, including the *stx* genes. Based on virulence genes, D1 clustered with the K-12 type strain (W3110) as when clustering based on all probe data Figure 7-4A), indicating that D1 and W3110 have more virulence genes in common with each other than with the other strains. Furthermore, as expected EDL933 and 3538 (Δ*stx*$_2$::*cat*) have more common virulence gene segments than with the other strains.

By further analyzing the virulence genes present in strain D1, we found that it did not have any hemolysin genes (*ehxA*), or type III secretion genes, which are located at the locus of enterocyte effacement in *E. coli* O157:H7 (*espA, B, D* and *tir*), and the *eae* gene which were

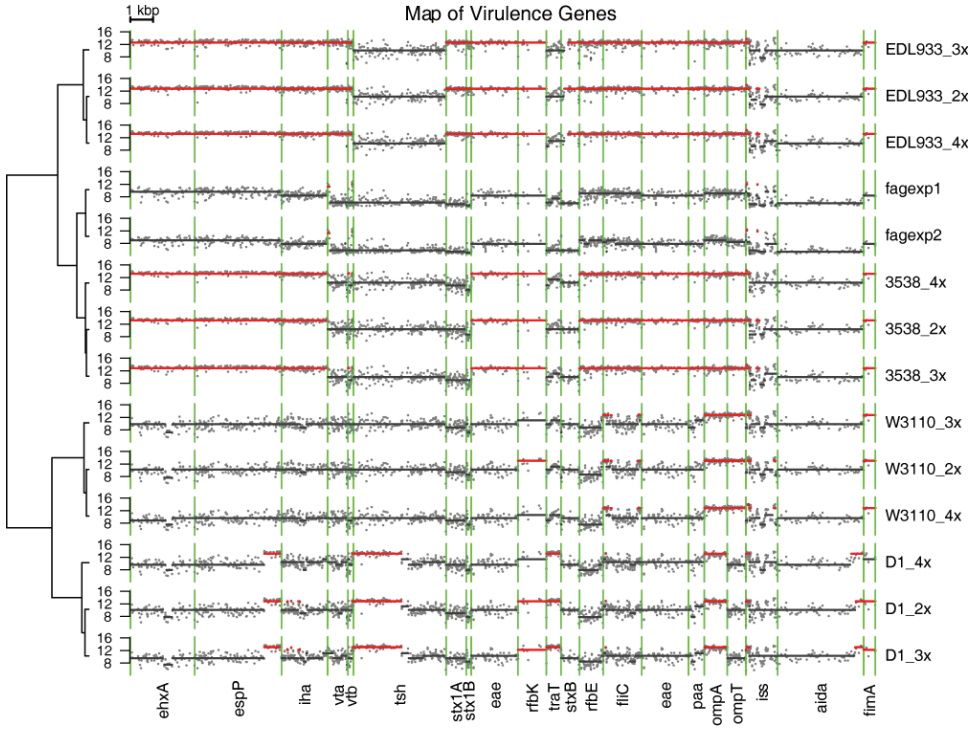

Figure 7-5. E. coli virulence genes. Illustration of log2-probe intensities (grey dots) overlaid with segmentation/merging results. Red lines correspond to segments identified as present in the experiment. Dark grey lines are those segments identified as not present. Only virulence genes with a segment present in at least one sample are included. Experiments are clustered according to segmentation and merging results (left). The sample is indicated to the right. Note: the order in which the virulence genes are concatenated does not signify importance.

present in EDL933 and 3538 (Δ*stx*$_2$::*cat*), as expected since they are human pathogens, EHECs (Caprioli*, et al.*, 2005; O'Brien*, et al.*, 1984). Almost the complete sequence of the bacteriophage V from *Shigella flexneri* was found in the genome of D1 and may be responsible for transferring the ϕ3538 (Δ*stx*$_2$::*cat*) to D1 as the V bacteriophage plays an important role in serotype conversion, and is associated with antigenic variation.

Although D1 has acquired genes often found in emerging pathogens, it is evident from the analysis of virulence genes that D1 is probably still a commensal *E. coli* and not yet a pathogen due to its relatedness to K-12 strains. While the K-12 strain is a commensal bacterium originally found in a stool sample from a diptheria patient in 1922, it has later developed into different sub-strains, none of which have been reported to cause illnesses. D1 is from a stool sample from a sheep, and its serotype O175:H16 has only been reported in the literature on a few occasions. While it can belong to a Shiga toxin producing *E. coli*, no illness have been related to this serotype (Pradel*, et al.*, 2000; Scheutz*, et al.*, 2004; Stephan and Hoelzle, 2000a; Stephan*, et al.*, 2000b), consistent with our findings.

Interestingly, based on the phage analysis (supplementary Figure S1), the D1 genome clusters with the phage ϕ3538 (Δ*stx*$_2$::*cat*) and 3538 (Δ*stx*$_2$::*cat*) samples rather than with W3110 samples, in this case disregarding EDL933 and MG1655 specific genomic phage elements. This indicates that D1 shares a significant proportion of phage elements with 3538

(Δ*stx*$_2$::*cat*). Supporting this, is the pattern of present segments in common for the Shiga toxin related phage elements, Stx1, Stx2-I, Stx2-II and VT2-Sa. Especially noticeable is the fact that although the same phage elements are present for EDL933 samples, the exact pattern differs, indicating obvious divergence in the phage sequence and confirming that a transfer of the phage Φ3538 (Δ*stx*$_2$::*cat*) element has taken place from 3538 (Δ*stx*$_2$::*cat*) to D1.

Hence, based on the above results, we can conclude that D1 is a non-pathogenic K-12 like strain with an increased ability to obtain foreign DNA, of which it has acquired a significant amount of 3538 (Δ*stx*$_2$::*cat*) phage elements. Nonetheless, the present analysis does not include potential virulence genes encoded on plasmids but only chromosomal genes. Therefore, D1 plasmids have to be purified and analyzed in a similar fashion with regard to potential virulence genes in order to say more about their possible role in pathogenicity.

Benchmarking

To estimate whether the above results reflected actual true biology, a number of quality issues were explored, including variability between replicate experiments and comparisons of results from control experiments to their known sequence.

First, the performance of the custom designed DNA microarray was evaluated further by analyzing the control strains, W3110 and EDL933, after mapping them to each other. By varying the threshold cutoff for calling absence/presence on raw data, a detailed performance analysis on the probe level could be achieved (Figure 7-6). In this way, it was possible to view how a gain in sensitivity (fraction of present probes that were identified) would concurrently increase the false positives rate (solid lines). The performance when using segmentation and merging (solid circles) was clearly above the ROC-curve for the simple threshold approach, indicating a superior performance and confirming that segmentation approaches improved the analysis.

Next, the 25 control EDL933 and MG1655 genomic phage elements were analyzed in the same way that virulence genes and non-EDL933 and non-MG1655 genomic phages were analyzed (supplementary Figure S2). Analysis of these genomic phage elements confirmed the reliability of our analysis approach, as they were all identified as present in the expected experimental strain. Thus, all K-12 isolate MG1655 phage elements where identified in their full length in the K-12 isolate W3110 samples, and also all O157:H7 isolate EDL933 phage

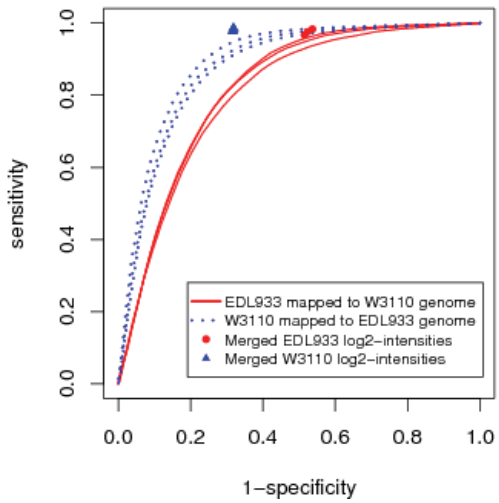

Figure 7-6. ROC curve showing the performance for different analysis approaches. Blue dotted lines and triangles: W3110 samples mapped to the EDL933 genome. Red solid line and dots: EDL933 samples mapped to the W3110 genome. The plot shows the performance when applying a threshold to $\log_2$ intensities (solid and dotted lines) and when segmenting and merging $\log_2$ intensities (solid dots and triangles).

elements were identified for the EDL933 samples. Only the end fragment of CP-933R was missing both in EDL933 and 3538 (Δ*stx*$_2$::*cat*) samples. However, since this fragment was part of an unstable cryptic prophage, it may easily have been lost since it is not useful to the bacteria.

Generally, the variability between replicate DNA samples was low, i.e. only a small fraction of genes were not found to be present or absent consistently across all replicates. For example, between replicate W3100 samples mapped to the EDL933 genome, the number of genes identified only differed by 0.9 percent. For the control strains, W3110 and EDL933, 98.8 and 97.7 percent of all genes were identified as present in their corresponding samples, respectively. Moreover, sensitivities of 0.92 and 0.94 were obtained when confirming the presence of W3110 genes in EDL933 samples and EDL933 genes in W3110 samples, respectively, while maintaining a false discovery rate (FDR) at 0.05. A closer examination of the false positives revealed that a majority of these corresponded to genes which might have been misclassified as negative by Blastp (e.g. an E-value close to the cut-off) while false negatives were most likely falsely predicted genes. Consequently, results obtained using our 7 *E. coli* genomes microarray platform are highly accurate and reflect true biology. Moreover, if repeating with high quality gene annotations when they become available, the sensitivity and FDR may even prove better than initially anticipated.

ACKNOWLEDGEMENTS

The authors would like to thank Peter Hallin for assistance with probe design and the Blast matrix; and the Sanger Institute for providing sequence data for *E. coli* strains produced by the Microbial Sequencing Group at the Sanger Institute (http://www.sanger.ac.uk/Projects/ Escherichia_Shigella/).

This study was supported partly by grant no. 147145 from the Research Council of Norway (AP, CS, YW); and The Danish Center for Scientific Computing and The Danish Technical Research Council (HW, KK, DWU).

Part IV

SEQUENCE DEPENDENT

GENE EXPRESSION

Chapter 8 Paper V

Minireview:

Chromatin architecture and gene expression in Escherichia coli

Hanni Willenbrock and David W. Ussery

Center for Biological Sequence Analysis, Technical University of Denmark, DK-2800 Lyngby. Denmark.

ABSTRACT

Two recent genome-scale analyses underscore the importance of DNA topology and chromatin structure in regulating transcription in *Escherichia coli*.

LOCATION, LOCATION, LOCATION

Expression of a gene is in a sense a bit like purchasing a new home – the value is strongly dependent on location. This value is context-dependent: it depends on who your neighbors are and also on the larger geographical picture. Two recent studies have analyzed DNA topology and chromatin structure on a genome-wide scale in *Escherichia coli* (Jeong*, et al.*, 2004; Peter*, et al.*, 2004). Both show that an important factor in determining transcription profiles – when and to what extent a gene is expressed - is the location of the gene within the context of the *E. coli* K-12 chromosome. Whereas this is old news for those who are interested mainly in eukaryotic chromosomes, it is an important concept that has often been overlooked (in our opinion) in bacterial transcriptomics. In eukaryotes, it is well known that there are two types of chromatin: heterochromatin, which remains condensed for the most part throughout the cell cycle and contains few genes, and euchromatin, which, on the other hand, contains gene-rich regions, and in some cases clusters of highly expressed genes.

Jeong *et al.* (Jeong*, et al.*, 2004) analyzed similarities in transcriptional activities of *E. coli* genes as a function of their position on the chromosome. An autocorrelation function identified three levels of spatial correlations of expressed genes: short-range (7-16 kilobase-pairs, kb), medium-range (approximately 100 kb) and long-range (over 700 kb). Figure 8-1 shows the gene expression data obtained by Jeong *et al.* (Jeong*, et al.*, 2004) together with that of Peter *et al.* (Peter*, et al.*, 2004), mapped onto the circular *E. coli* chromosome, with four circles (circles 3-6) corresponding to values obtained from the four experiments of Jeong *et al.* (Jeong*, et al.*, 2004). They took into account the transcription levels of nearly all genes, although only the more highly expressed genes are visible in Figure 8-1. Most of the genes in *E. coli* are transcribed around the time of replication (Dworkin and Losick, 2002), and only a small fraction (typically around 10%) of the genes are highly transcribed. These "clumps" or regions of highly expressed genes can be seen as dark bands in Figure 8-1, and some of these regions differ in the various experiments. The shortest level of spatial correlation found by Jeong *et al.* (Jeong*, et al.*, 2004) corresponds to between 7 and 15 genes that exhibit an apparently coherent transcriptional activity. These groups are larger than operons, and are likely to reflect small clusters of co-regulated genes, of between roughly three and five operons (assuming about three genes per operon), including the clusters of highly expressed genes mentioned above. This is the first level of the 'bigger picture' of spatial correlations, and is also the most clearly affected by DNA supercoiling, given that correlations at this level are significantly reduced by the addition of norfloxacin, a gyrase and topoisomerase IV inhibitor (data shown in circle 5 in Figure 8-1). Having said that, it should also be pointed out that all the correlations, including the longer range ones, were affected by gyrase mutations (circle 6 in Figure 8-1).

The results reported by Jeong et al. (Jeong*, et al.*, 2004) are slightly different from previous findings by Sousa *et al.* (Sousa*, et al.*, 1997), who looked at the expression of a reporter gene when it was inserted at different positions around the chromosome. Sousa *et al.* (Sousa*, et al.*, 1997) found that gene expression varies along the chromosome in a somewhat linear manner, forming a gradient in which the more highly expressed genes are localized near the replication origins and the region around the replication terminus contains few highly expressed genes. This was thought to be the result of gene dosage associated with the distance to the origin of replication: during the replication of the chromosome, there are more likely to be multiple copies of genes that are close to the replication origin. As can be seen in Figure 8-1, regions with highly expressed genes are not limited to the area close to the origin but are distributed in clumps throughout the chromosome, although there are few highly expressed regions around the replication terminus. Thus, in contrast to the predictions of Sousa *et al.* (Sousa*, et al.*, 1997), the experimental results of Jeong *et al.* (Jeong*, et al.*, 2004) show that a gene does not necessarily have to be located close to the

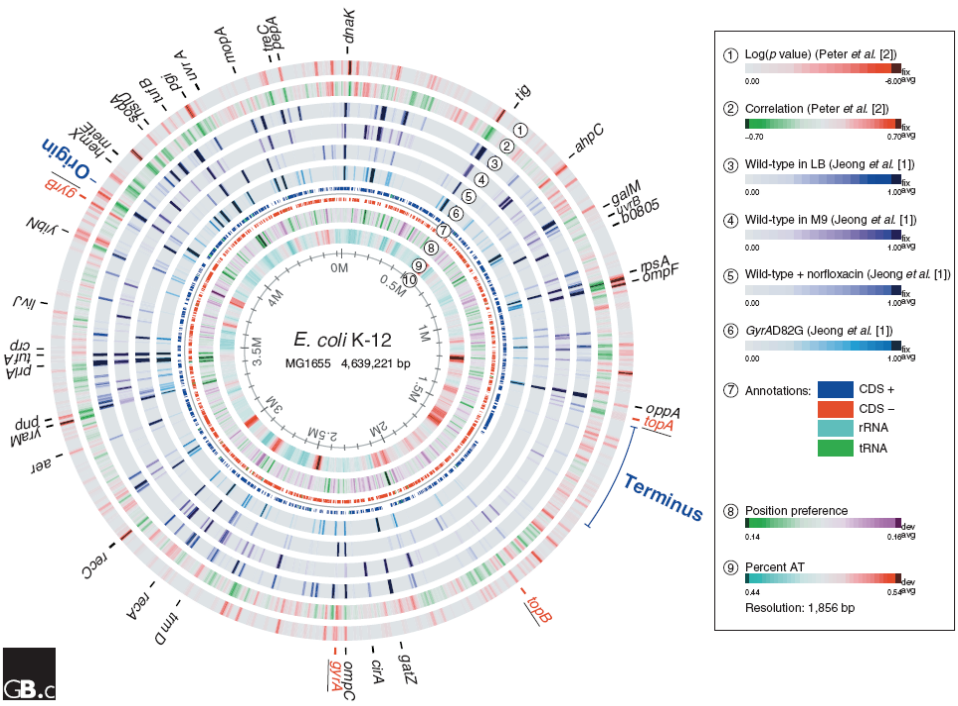

Figure 8-1. Expression atlas for the *E. coli* experimental data of Jeong *et al.* (Jeong, *et al.*, 2004) and Peter *et al.* (Peter, *et al.*, 2004). The atlas was constructed using the Genewiz software (Pedersen, *et al.*, 2000). DNA topoisomerase genes are marked in red, and the replication origin and terminus are marked in blue. The outer circle (1) shows the change in expression of genes in response to supercoiling (log *P*-values), where more negative values correspond to genes that are more significantly influenced by DNA relaxation; and circle (2) shows the correlation of these expression values with DNA supercoiling, where high absolute values correspond to gene expression levels that show most correlation or anti-correlation with measured levels of DNA relaxation; both sets of data are from Peter *et al.* (Peter, *et al.*, 2004). Shown in the next four circles (3-6) are the expression values of chosen experimental conditions from Jeong *et al.* (Jeong, *et al.*, 2004): (3) wild-type cells in rich medium (LB), (4) minimal medium (M9), (5) following 30 minutes of treatment with the gyrase inhibitor norfloxacin, and (6) cells carrying a mutation (*GyrA*D82G) in a gyrase gene, respectively. Circle (7) shows the location of protein coding sequences on the positive strand (CDS+), on the negative strand (CDS-), and the rRNA and tRNA genes. Circle (8) shows a running average of the absolute value of the nucleosomal position preference (Satchwell, *et al.*, 1986), and circle (9) the AT content (±3 standard deviations from chromosomal average). Expression data from Jeong *et al.* (Jeong, *et al.*, 2004) were centered and scaled. Circle (10) shows distance along the chromosome, in megabases (M), counting from the beginning of the GenBank sequence.

origin of replication to be highly expressed but its expression level is rather dependent on its location within a smaller confined sub-domain.

The long range correlations (several hundred thousand bp) found by Jeong *et al.* (Jeong, *et al.*, 2004) are more interesting than the short-range correlations and also have precedents in eukaryotic systems, where such clustering of highly expressed genes was postulated a very long time ago for the Drosophila polytene chromosomes a very long time ago (Ananiev and Gvozdev, 1974). More recently, there have been two studies on gene expression in human chromosomes that showed clustering of highly expressed genes (Gilbert, *et al.*, 2004; Versteeg, *et al.*, 2003). The topic of chromatin structure and gene expression in eukaryotes has generated considerably more interest (and publications) than for bacteria. In fact, at the

time of writing this article, a paper was recently published showing that the 'upstream binding factor' for RNA polymerase I causes the chromatin to form a more decondensed, open structure, allowing access to the polymerase enzyme for transcription (Chen*, et al.*, 2004). Although most animals have on the order of a thousand times as much DNA as bacteria, the level of compaction by chromatin is similar in both (about 7000-fold). But it is likely that the DNA compaction is more dynamic in bacteria, because of the higher coding densities. of the chromosome Furthermore, transcription and translation are coupled in bacteria, most likely for topological reasons (Gowrishankar and Harinarayanan, 2004). The long-range correlations found by Jeong *et al.* (Jeong*, et al.*, 2004) are consistent with a role for chromatin structure in gene expression in bacteria, showing once again that what is true for elephants can also apply to *E. coli*.

DNA SUPERCOILING AND GENE EXPRESSION

More than 20 years ago, it was postulated that supercoiling could be used to regulate gene expression in *E. coli* (Smith, 1981), and about a decade later (before microarray technology was readily available) the influence of supercoiling on the concentration of 88 proteins in *E. coli* was demonstrated (Steck*, et al.*, 1993). In the recently article by Peter *et al.* (Peter*, et al.*, 2004), the influence of DNA supercoiling on transcription was studied using DNA microarrays to systematically probe expression profiles of all *E. coli* genes. The authors (Peter*, et al.*, 2004) demonstrated that supercoiling may act as a' transcription factor' and that it can have either a negative or a positive effect on transcription of a specific gene. They identified 306 'supercoiling-sensitive genes' and the expression of most of these genes correlates very well with the amount of chromosomal relaxation in each experiment. The fact that most of these supercoiling-sensitive genes were localized in regions of high density 'clumps' that were affected by DNA relaxation agrees well with the findings by Jeong et al. (Jeong*, et al.*, 2004) that short-range correlations are dependent on negative supercoiling.

The outermost two circles in Figure 8-1 are based on data from the paper by Peter et al. (Peter*, et al.*, 2004) and show the locations of supercoiling-sensitive genes (log *P*-values, circle 1) and the correlation with chromosomal relaxation (circle 2). Anti-correlations corresponding to regions where expression decreases upon DNA relaxation were also found. As reported by Peter *et al.* (Peter*, et al.*, 2004), chromosomal regions with significant numbers of supercoiling-sensitive genes generally overlap with regions that are more correlated or anti-correlated with the level of chromosomal relaxation than regions with no supercoiling-sensitive genes.

Some of the chromosomal regions that are mostly correlated with supercoiling overlap with regions showing differential expression patterns among the experimental conditions used by Jeong *et al.* (Jeong*, et al.*, 2004). For example, *gyrA* and *gyrB* at 2.33 megabases and 3.88 megabases on the chromosome, respectively, are highly expressed in DNA-relaxed cells (wild-type cells grown with norfloxacin; circle 6 in Figure 8-1) but hardly expressed in wild-type cells grown in rich (LB;circle 3) or minimal (M9;circle 4) media. Because of the experimental conditions used in both studies, however, this picture is expected for the gyrase genes. These genes are known to be sensitive to supercoiling and are involved in maintaining a precise level of supercoiling in the cell. Thus the inhibition of these proteins is very likely to increase their mRNA expression. Surprisingly, a substantial number of additional genes were also affected by gyrase inhibition, indicating that this change in expression has to be due to the effect that gyrase inhibition has on DNA supercoiling - that is, chromosomal relaxation.

Peter et al. (Peter*, et al.*, 2004) also found that supercoiling-sensitive genes whose expression increased upon DNA relaxation were significantly more AT-rich in their upstream and coding regions compared to corresponding regions of genes not sensitive to supercoiling; the opposite was true for supercoiling-sensitive genes whose expression

decreased upon DNA relaxation. This may, however be due to the fact that AT-rich regions tend to be more curved than AT-poor regions. Supercoiling-sensitive genes may, therefore, be expected to be more AT-rich in upstream regions than genes that are regulated by means other than supercoiling. Nonetheless, these small local variations in upstream regions are not visible on the genome-scale atlas plot (Figure 8-1, circle 9). Because these supercoiling-sensitive genes are localized to specific regions, one would expect that in some cases a region would appear AT-rich if all of its supercoiling-sensitive genes were significantly AT-rich in their upstream regions.

A bit more context is needed here - at the risk of complicating the picture, there are two additional pieces of information which can help build a clearer picture of what is going on in terms of chromatin structure. The first is DNA curvature and the second is a bit more detail about DNA supercoiling. DNA has sequence-dependent structures, just like proteins, and certain sequences tend to coil in three-dimensional space. These 'DNA curves' are correlated with phased tracts of A residues, and have been found to be localized at the tips of supercoils (Pavlicek*, et al.*, 2004). The DNA in *E. coli* is known to be supercoiled, and curved DNA (which tends to be AT-rich) can result in the placement of certain DNA sequences at the apical tips of supercoils, as shown in Figure 8-2. The supercoils can be divided into two types: plectonemic and toroidal, depending on the shape (Figure 8-2). Roughly half of the supercoils in *E. coli* are toroidal – the DNA is wrapped around proteins and it is 'restrained', although this is transient in bacteria (but permanent in the form of

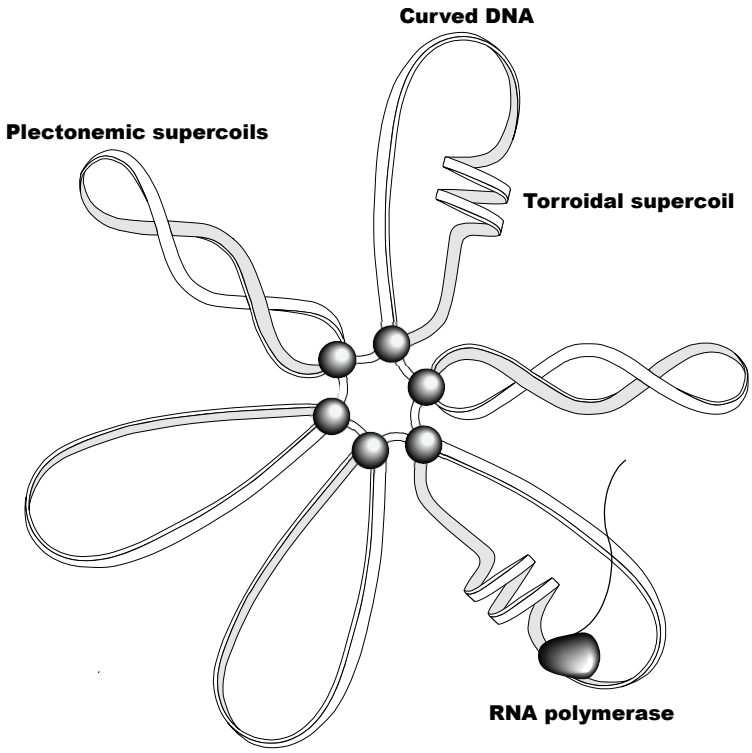

Figure 8-2. An illustration of DNA supercoiling domains in the *E. coli* chromosome. This is a cartoon of the chromosome; in real life there are perhaps as many as 400 different domains. Plectonemic (unrestrained) and toroidal (restrained, for example by wrapping around a protein) supercoiling is indicated. Curved DNA tends to be localized at the tips of supercoils. The illustration is modified with permission from (Sinden, 1994).

stable nucleosomes in eukaryotes). The other half of supercoils is plectonemic (unrestrained) and is under torsional stress, which can be relieved by formation of a bubble in the DNA helix. The ratio between plectonemic and toroidal supercoiling might vary along the chromosome and also with time, for example, an RNA polymerase can wrap DNA around it (a restrained toroidal supercoil) and then release the DNA later, creating an unrestrained supercoil. Furthermore, a region that in one set of experimental conditions contains mainly restrained supercoils can suddenly have most of the supercoils become "free" (plectonemic) in the absence of chromatin proteins.

From a DNA topology perspective, the plectonemic supercoils contain more potential energy, in terms of driving superhelical dependent transitions (such as melting the DNA helix). Thus, if there were regions along the chromosome that contained lots of binding sites for proteins involved in chromatin structure, most of the supercoiling would be transiently restrained, and hence less free energy would be available for transcription. In addition, the chromatin proteins can physically block the RNA polymerase from binding to the DNA. Because the *E. coli* chromatin proteins IHF and FIS show some sequence specificity, it is possible to predict binding sites throughout the chromosome. On a global scale, there tends to be an anti-correlation between these chromatin-binding sites and regions of highly expressed genes (Ussery*, et al.*, 2001). Finally, on the more local level of a few kilobases (for example, an operon), it is possible to predict regions that tend to exclude chromatin proteins, and hence might potentially be highly expressed (Dlakic*, et al.*, 2004). In Figure 8-1, this "nucleosomal position preference" measure is plotted in circle 8. As expected, regions of low position preference tend to correspond to the regions with highly expressed genes found by Jeong *et al.* (Jeong*, et al.*, 2004). However, the majority of cellular DNA is compacted transiently by chromatin proteins, and there are many regions that are not highly expressed but are nonetheless regulated, with their relative expression levels dependent on supercoiling.

Originally, it was postulated that the chromosome was divided into 12-80 topologically isolated loops, so-called domains, in which chromatin could be relaxed independently of supercoiling in nearby domains (Worcel and Burgi, 1972). Later this number was estimated more exactly at around 50 domains corresponding to a domain size of approximately 100 kb (Sinden and Pettijohn, 1981). Recently, Postow *et al.* (Postow*, et al.*, 2004) presented evidence of an even smaller domain size of approximately 10 kb on average, corresponding to as many as 400 distinct topologic domains in *E. coli*. This result corresponds very well with the finding by Jeong *et al.* (Jeong*, et al.*, 2004) that up to 16 genes exhibited apparent coherent transcriptional activity and the idea that genes may be organized into confined supercoiled domains with a size of up to 16 kb.

The fact that the genes identified as sensitive to supercoiling have a variety of functions, supports the hypothesis that supercoiling may act as a global transcriptional regulation mechanism and that the cell may use this mechanism as an environmental sensor because the topology of the chromosome may be affected by the surrounding environment. The chromatin protein H-NS regulates many environmental genes, probably through DNA topological changes (Rimsky, 2004).

One final aspect of this global view of regulation of transcription at the level of chromatin structure is that some of these environmentally regulated and supercoiling-sensitive genes are involved in bacterial pathogenesis. For example, in *Salmonella*, it has been shown that regulation of genes involved in invasion is regulated by DNA supercoiling (Leclerc*, et al.*, 1998). Thus, the global regulation of gene expression by DNA topology could prove to be an important aspect of understanding the mechanisms of bacterial virulence (Dorman, 1991).

ACKNOWLEDGEMENTS

This work was supported by a grant from the Danish Center for Scientific Computing.

Chapter 9 Paper VI

An Environmental Signature for 323 Microbial Genomes based on Codon Adaptation Indices

Hanni Willenbrock, Carsten Friis, Agnieszka S. Juncker, David W. Ussery

Center for Biological Sequence Analysis, Technical University of Denmark, DK-2800 Lyngby. Denmark.

ABSTRACT

Background: Codon adaptation indices (CAI) represent an evolutionary strategy to modulate gene expression and have been widely used to predict potentially highly expressed genes within microbial genomes. Here we evaluate and compare two very different methods for estimating CAI values, one corresponding to translational codon usage bias and the second obtained mathematically by searching for the most dominant codon bias.

Results: The level of correlation between these two CAI measures is a simple and intuitive measure of the degree of translational bias in an organism, and from this, we confirm that fast replicating bacteria are more likely to have a dominant translational codon usage bias than slow replicating bacteria and that this translational codon usage bias may be used for prediction of highly expressed genes. By analyzing more than 300 bacterial genomes as well as 5 fungal genomes, we are able to show that codon usage preference provides an environmental signature by which it is possible to group bacteria according to their lifestyle, e.g. soil bacteria and soil symbionts, sporeformers, enteric bacteria, aquatic bacteria and intercellular and extra-cellular pathogens.

Conclusions: The results and the approach may be used to acquire new knowledge regarding species lifestyle as well as deducing relationships between organisms originally thought to be evolutionary far apart.

BACKGROUND

Differential codon usage represents an evolutionary strategy to modulate gene expression and hence mathematical formulations of the codon usage bias have been widely used to predict gene expression on a genomic scale. This is based on the assumption that codon usage bias is correlated with protein levels. Indeed, highly expressed genes have been found to almost exclusively use those codons translated by abundant tRNAs in *Escherichia coli* and budding yeast, while non-highly expressed genes appear to be less biased in their codon usage that may be more strongly influenced by mutations than by selection during the course of evolution (Sharp and Li, 1987).

Based on these observations, several approaches to measure codon usage have been proposed in order to predict the level of protein expression, such as the frequency of optimal codons (Ikemura, 1981), the codon preference statistic (Gribskov*, et al.*, 1984), the codon adaptation index (Sharp*, et al.*, 1987), the 'effective number of codons' used in a gene (Wright, 1990), and predicted highly expressed genes (Karlin*, et al.*, 2003). Of these, the codon adaptation index (CAI), has survived the test of time and has now been cited more than 700 times with 58 citations just in 2005. This method is based on a known set of 27 very highly expressed *E. coli* genes (Sharp and Li, 1986a), from which a codon bias signature was deduced that was most likely to be efficient for translation. This bias was then used to derive codon adaptation indices for all genes in *E. coli*.

While the first species examined – *E. coli* and *S. cerevisiae* - provided strong evidence of high translational codon usage bias, recent studies report of bacterial species with little codon usage bias (Carbone*, et al.*, 2005; Carbone*, et al.*, 2003), *often* species with extreme AT or GC content. In these studies, whole genome information was used to obtain a universal CAI, applying a mathematical measure to derive the most dominant codon bias based on the codons from all potential open reading frames from a genome. This CAI, which ignored the codon usage of experimentally determined highly expressed genes, demonstrated that codon bias as such is not necessarily translational and correlating with gene expression, especially in slow growing bacteria (Carbone*, et al.*, 2003). Consequently, it is not trivial to deduce and compare codon usage biases across a vast range of bacterial species available in sequence databases, including AT or GC rich species, and to the best of our knowledge, this type of large-scale comparison has not been done previously.

Although an early paper found little correlation between mRNA and protein concentration, the correlation was considerably higher for highly expressed genes (Gygi*, et al.*, 1999) and a recent study found a significant relationship between protein levels and mRNA levels in yeast (Ghaemmaghami*, et al.*, 2003). Consequently, microarray gene expression data are useful for confirming predicted highly expressed genes - as a substitution for protein levels.

Here, we calculate and compare a translational codon adaptation index (tCAI) based on that proposed by Sharp and Li (Sharp*, et al.*, 1987) with a purely mathematical dominant codon adaptation index (dCAI) (Carbone*, et al.*, 2003) for 318 bacterial and 5 fungal genomes with their full sequence deposited in Genbank and available from the Genome Atlas Database (http://www.cbs.dtu.dk/services/GenomeAtlas/) version 19.1. We compare the ability for both types of CAI to estimate the translational codon bias of an organism and show that codon usage preferences provides an environmental signature by which it is possible to group bacteria according to their lifestyle. Furthermore, we examine how well each CAI measure correlates with microarray gene expression data for six selected organisms and show that the tCAI measure is generally better for predicting highly expressed genes than dCAI.

RESULTS AND DISCUSSION

The two types of codon adaptation indices were calculated for all genes in 318 bacterial strains and 5 fungal genomes, and the correlations between the derived tCAI and dCAI

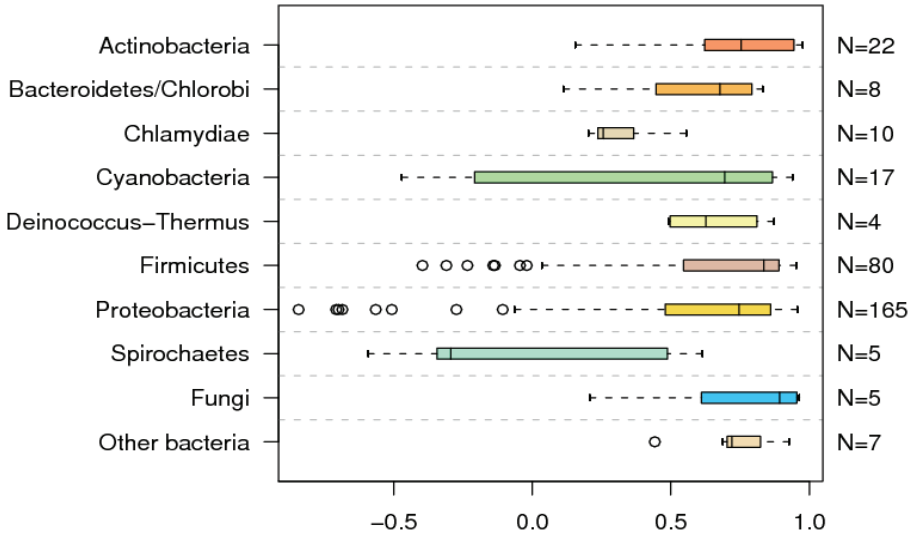

Figure 9-1. Boxplot summarizing correlations between tCAI and dCAI for 8 major bacterial phyla and fungi. The group "Other bacteria" comprises a number of minor bacterial phyla (Aquificae, Chloroflexi, Fusobacteria, Planctomycetes, Acidobacteria and Thermotogae) which could not meaningfully be included in any of the other categories.

values are illustrated for 8 different bacterial phyla, with any remaining bacterial species grouped into "Other bacteria", and fungi depicted separately (Figure 9-1). For most groups, the correlation between the two CAI measures is high (median above 0.5). Only for Chlamydiae and Spirochaetes are the median correlations below 0.5, indicating that the dominating codon biases are not translational for most of the species included in these groups. However, it is not surprising that there appears to be little selection for strong tCAI bias in these genomes since most of the bacteria in both of these phyla have slow replication times. Presumably, fast-replicating bacteria have optimized their replication machinery as opposed to slow replicating bacteria where other factors might be more important (Carbone*, et al.*, 2005; Carbone*, et al.*, 2003; Rocha, 2004). Consequently, we were able to confirm a significant relationship between the level of translational codon adaptation and replication time across the entire range of genomes (Spearman's rank correlation, rho ~ 0.46) – using the number of 16S rRNAs as an indirect measure of doubling time, as previously suggested (Sharp*, et al.*, 2005), since the number of 16S rRNAs indirectly influence replication times (Ussery*, et al.*, 2004).

Next, the codon preferences, which are measurable by the relative adaptiveness of each codon (w_{ij}), were compared between tCAI and dCAI and the difference (w_{ij} for tCAI minus w_{ij} for dCAI) was used for cluster analysis of all 318 bacterial strains and the 5 fungal genomes (See Figure 9-2A and supplementary Figure S1). The Figure shows a clear separation into several clusters with AT-rich bacteria towards the left and GC-rich bacteria towards the right, while bacteria with intermediate base composition are in the middle. This is also reflected in the clustering of codons which are separated into two distinct clusters, where either a codon preference for A/T (lower half) or G/C (upper half) in the third position for dCAI is evident, i.e. GC3/AT3 skew dominates over translational bias. However, although the AT content appears to be a significant factor in the clustering, merely ordering by AT content does not give the same highly distinguishable clusters. Consequently, the correlation between the level of translational codon adaptation (measured by the correlation between tCAI and dCAI) and the genomic AT content was indeed very low, but still significant (rho ~ -0.14, P-value ~

1.5e-2), supporting the minor although unmistakable correlation between AT content and clustering order visible in Figure 9-2A.

The middle area of Figure 9-2A appears most diverse and can be divided into three distinct regions (ignoring a few smaller clusters on its left side). This division results in a total of five distinct regions as illustrated in Figure 9-2A. Figure 9-2B provides a zoom of the third and fourth region from the left. The 3rd region consists mainly of 'enterics' (intestinal bacteria) living in the human intestine (e.g. *Escherichia*, *Shigella*, *Salmonella*, *Bacteroides*), the fly intestine (*Yersinia pestis*), and the animal intestine (*Yersinia pseudotuberculosis*). The yeast genome, *S. cerevisiae*, clusters with the enterics. Although fungi are obviously quite distant from bacteria phylogenetically, both can be relatively fast replicating and hence would face the same selective pressure on codon usage. Moreover, *Kluyveromyces lactis* also groups with the enterics, including *E. coli* K-12, with whom it is often grown together in fermentors to produce chymosin (rennet) on a commercial scale, reflecting similar preferences on growth environment.

The 4th region mostly consists of bacteria living in aquatic environments such as marine waters (*Thermotoga maritima*, *Prochlorococcus marinus*, *Desulfotalea psychrophila*, *Synechococcus species*), groundwater (*Dehalococcoides*), freshwater (*Synechococcus elongatus*), and hot springs (*Thermosynechococcus elongatus*). While other *P. marinus* strains cluster in the 1st region, strain MIT9313 is low-light-adapted and has almost as many strain specific genes as it has genes in common with its high-light-adapted relative, strain MED4 (Rocap*, et al.*, 2003), reflective of differing environmental preferences.

Looking at the remaining regions in Figure 9-2A, we observe that the 1st (leftmost) region consists of slow-growing intracellular pathogens (*Mycoplasma*, *Rickettsia*, *Chlamydia*, etc), and other small pathogens (*Bartonella*, *Helicobacter*, *Ehrlichia*, *Campylobacter*) mostly with genome sizes less than or close to 1 Mbp. The content of this region reflects the observation that many organisms with reduced genomes have very low GC content and supports the speculations that there is a selective pressure in this group of bacteria to lower the nitrogen requirement for DNA synthesis (Giovannoni*, et al.*, 2005), by adapting the codon usage to favor codons with more A's and U's.

The 2nd region mainly consists of sporeformers, including Gram positive bacteria (e.g. *Streptococcus*, *Lactococcus*, *Lactobacillus*, *Staphylococcus*, *Enterococcus*, *Bacillus*, *Oceanobacillus*, *Listeria*, *and Clostridium*). *Schizosaccharomyces pombe* (fission yeast) is found in this region and resembles the other microbes in that it can also reproduce by sporulation.

Many of the bacteria in this region can replicate quite fast, and exhibit other evidence of selective pressure for optimization of the genome for quick replication on demand. For example, the *Vibrio* (a Gram negative, non-spore former) and *Bacillus* (a Gram positive sporeformer) cluster close together; and they have the largest number of rRNAs and tRNAs out of several hundred bacterial genomes sequenced so far. Of the remaining fungal genomes, the plant infecting *Ashbya gossypii* and the diarrhea causing parasite *Encephalitozoon cuniculi* are found in a somewhat remote cluster in the 2nd region together with the soil bacterium, *Bacillus licheniformis*, which is mainly associated with plant materials and is toxinogenic, i.e. causing food poisoning in humans.

Finally, the 5th (rightmost) region mainly consists of soil bacteria (e.g. *Pseudomonas*, *Nocardia*, *Streptomyces*, *Desulfovibrio*, *Burkholderia*), and soil symbionts (e.g. *Xanthomonas*, *Agrobacterium*, *Rhizobium*) and plant pathogens (*Xylella fastidiosa*, *Pseudomonas*) as well as a few mammalian pathogens. Among additional bacteria in this region, we found an intercellular pathogen, *B. melitensis* that may have evolved from soil and plant associated bacteria (Paulsen*, et al.*, 2002) and a pathogen, *Wolinella succinogenes*, in which several soil related genes have been identified (Baar*, et al.*, 2003).

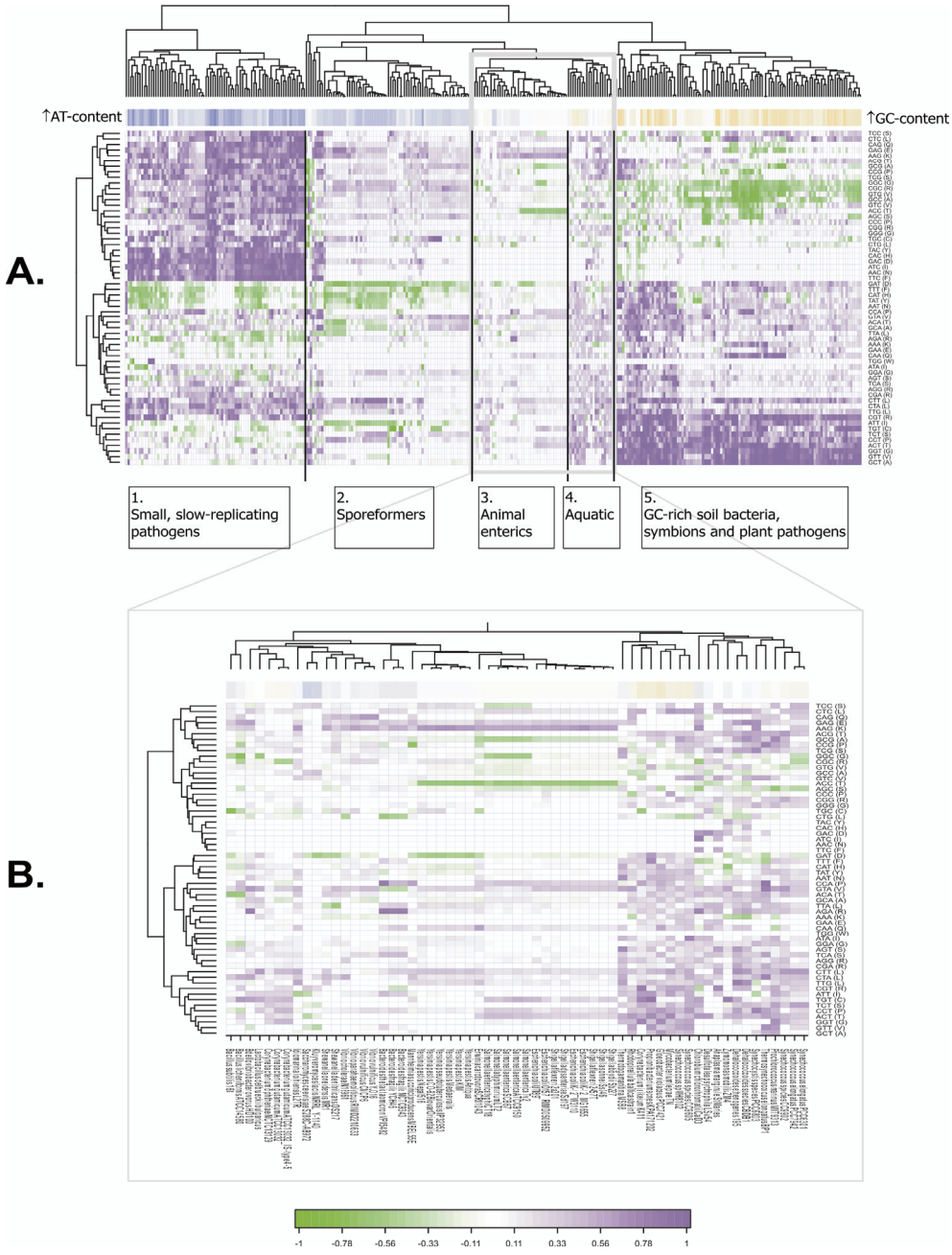

Figure 9-2. 2D cluster analysis of differential codon preferences for tCAI and dCAI. The differences in relative adaptiveness of each codon (w_{ij} for tCAI minus w_{ij} for dCAI) for each Genbank entry were clustered in two dimensions, one clustering codons and the other clustering Genbank entries. The clustering was performed as a hierarchical cluster analysis using Euclidian distances and complete linkage. Codons preferred relatively more by dCAI are red, while codons preferred relatively more by tCAI are green. Equal preference is indicated by white. A. Entire dendrogram. B. Zoom of the 3rd and the 4th region. Weights not considered: start codon 'ATG' and stop codons 'TGA', 'TAG' and 'TAA'.

Thus we find that, upon closer inspection, apparently misplaced genomes in a region may reflect similar shared ecological niches in the past.

By the above described approach, we were able to divide the organisms into three overall groups reflective of the genomic AT/GC content as previously demonstrated, based on distances between binarized codon weights from dCAI (Carbone*, et al.*, 2005). However, rather than merely discriminating between classes of lifestyle in terms of mesophily, thermophily and hyperthermophily - as previously shown based on either amino acid composition (Kreil and Ouzounis, 2001; Tekaia*, et al.*, 2002) or by codon usage (Carbone*, et al.*, 2005) - we obtained an environmental signature based on differences in codon weights between evolutionary more dominant codons and codons preferred by the translational machinery. Consequently, we demonstrate that differences in codon usage bias by tCAI and dCAI provide an environmental signature by which it is possible to group bacteria into environmental groups, such as soil bacteria, enterics, sporeformer and intracellular pathogens. That is, a clear environmental signature is evident in the composition of the clusters based on differences in relative adaptiveness of each codon as identified by either of the two CAI measures. These results build on a previous finding that GC content of microbial communities is influenced by the environment (Foerstner*, et al.*, 2005).

Prediction of highly expressed genes

Since tCAI is a “forced” measure of translational bias, while dCAI is a measure of the most dominating bias for an organism independently of the type of bias (i.e. GC skew bias, strand bias, *etc.*), the correlation between these two measures is a simple and intuitive, yet strong indication of whether the most dominating bias is translational or not, and consequently, how well the dCAI values explain gene expression. In this sense, it is not surprising that the correlation between the two CAI measures also gives an indication of how well tCAI explain the gene expression levels. This trend holds true at least for the six organisms for which we compared CAI values to microarray data, where the correlations between the two CAI measures are significantly correlated with the degree of how well tCAI correlates with gene expression (rho=0.6).

To further analyze and compare genes predicted as highly expressed by tCAI with genes having extreme codon bias according to dCAI values and with the highly expressed genes estimated by microarray analysis, the overlap between the top 10 percent genes was found and visualized in a Venn diagram (Figure 9-3). For both *S. cerevisiae* and *E. coli*, genes with high tCAI and dCAI values overlap significantly with each other as well as with genes identified as highly expressed in microarray experiments. For *B. subtilis*, a smaller but similar trend is evident. For the remaining bacteria, a significantly higher number of genes with high expression values (microarray data) overlap with genes with high tCAI values than with genes having high dCAI values. An investigation of the functional categories to which the dCAI reference genes (top 1% genes) belonged to, revealed that for *S. cerevisiae*, *E. coli* and *B. subtilis*, a significant fraction of ribosomal proteins were included, while for *P. aeruginosa*, *C. jejuni* and *G. sulfurreducens*, no ribosomal proteins where found among dCAI reference genes. This is in agreement with the ribosomal criterion defined by Carbone et al. (Carbone*, et al.*, 2005) saying that ribosomal proteins have significantly higher dCAI values than other protein encoding genes in translationally biased organisms. Thus, organisms having few or no ribosomal proteins among dCAI reference genes show little translational codon usage bias as compared to organisms having many ribosomal proteins among dCAI reference genes.

The above comparison of microarray data with tCAI values demonstrates that even for organisms evolutionary far apart from *E. coli*, it is possible to predict highly expressed genes by estimating the translational codon usage adaptation even when the most dominating bias in an organism is not translational, by comparing codon usage for each gene to that of genes involved in translation using tCAI. The level of confidence, however, decreases with

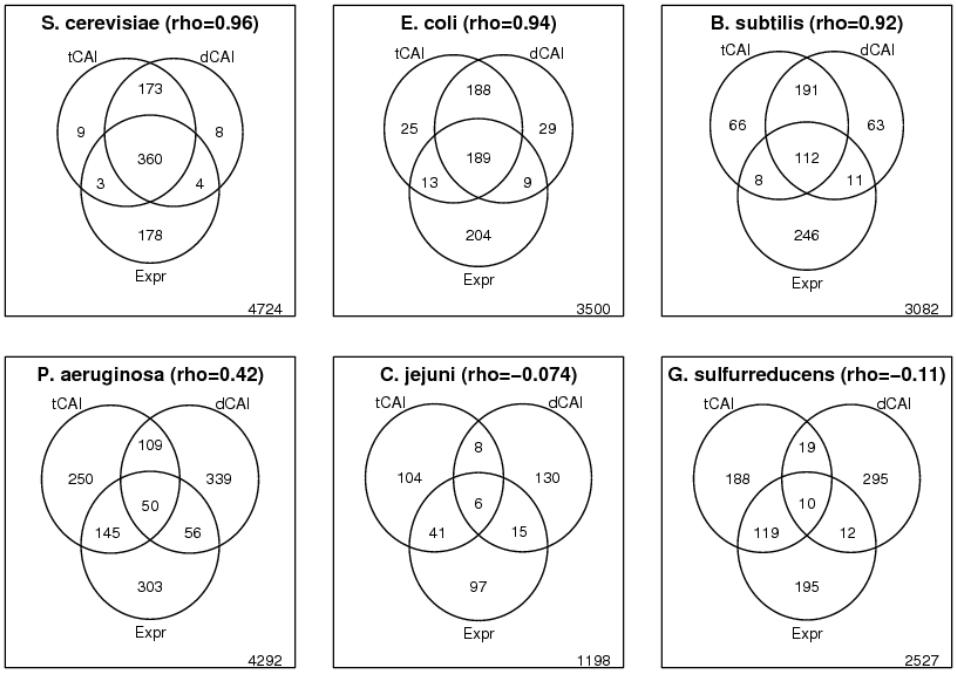

Figure 9-3. Overlap between genes with top 10% tCAI, dCAI and microarray gene expression values.

decreasing levels of translational codon adaptation in the dominating codon usage biases (as estimated from the correlation between tCAI and dCAI). On the other hand, as also observed by (Carbone*, et al.*, 2005), in cases where translational bias is not dominant, dCAI would not be useful for predicting gene expression.

CONCLUSION

Previously, it has been postulated that fast-growing bacteria share codon usage preferences since they have more abundant and similar tRNAs (Rocha, 2004). Here, we offer a biological explanation by showing a clear relationship between environment and similarities in codon usage biases, i.e. differences in codon preferences of translational codon adaptation and dominant codon adaptation provide an environmental signature by which it is possible to divide bacteria into groups representing different lifestyles, such as soil bacteria and symbionts, enterics, aquatic bacteria, sporeformers and small intercellular and extracellular pathogens.

Moreover, our study confirm across a wide range of bacteria and fungi that the observed variations in correlation between codon adaptation and gene expression are related to differences in replication times. For organisms with low correlations between tCAI and dCAI, the dominant codon bias is not translational, and consequently, the dCAI values do not reflect translational bias. Nonetheless, comparisons of microarray data with tCAI values indicate that this codon adaptation index is still useful for predicting a set of highly expressed genes although the level of confidence decreases along with the magnitude of the translational bias.

METHODS

All Genbank entries of completely sequenced genomes were taken from version 19.1 (May 26th, 2006) of the Genome Atlas Database (Hallin and Ussery, 2004).

Gene Expression Data

Gene expression data for *E. coli* was downloaded from Gene Expression Omnibus database (GEO, http://www.ncbi.nlm.nih.gov/projects/geo/): GSM18261 (Covert, *et al.*, 2004), and gene expression data for *C. jejuni*, 42°C reference experiments (Stintzi and Whitworth, 2003), and *P. aeruginosa*, MHH0122 (Salunkhe, *et al.*, 2005) were provided by the authors. For *S. cerevisiae*, preprocessed expression data were downloaded from GEO for two yeast strains, BY4741 (samples GSM6711, GSM6712 and GSM6713) (Bulik, *et al.*, 2003) and BY4716 (samples GSM35294, GSM35295 and GSM35296) (Ronald, *et al.*, 2005), both strains derived from the S288C strain.

All raw data were normalized with qspline (Workman, *et al.*, 2002) and expression indices were estimated (Li, *et al.*, 2001b). BY4741 expression data were log-transformed and all preprocessed S. cerevisiae data were re-normalized by qspline together with 179 additional expression profiles for the same Affymetrix YG-S98 chip downloaded from GEO. For *C. jejuni*, the median of normalized data was used and for *S. cerevisiae*, the mean of the two strain medians was used.

Additional processed expression data were downloaded from ArrayExpress for *G. sulfurreducens* ATCC 51573: GGS23_BR2_2S_12679025 (Methe, *et al.*, 2005), and *B. subtilis*: 25866GENEPIX25866 (Helmann, *et al.*, 2003). No further treatment of this data was carried out.

Translational Codon Adaptation Index (tCAI)

The CAI measure of translational adaptation was inspired by the original codon adaptation index from (Sharp, *et al.*, 1987) and in the following, we will refer to this CAI measure as the "translational codon adaptation index" (tCAI). However, although Sharp & Li in their original work from 1987 were forced by lack of data to assume a background codon usage corresponding to equal usage of the synonymous codons for any given amino acid, we now have vast libraries of complete genomic sequences available. Consequently, we calculate the relative synonymous codon usage (RSCU) for each organism by comparing the codon distribution from a set of highly expressed genes to a background distribution estimated from the codon usage of all coding regions in the genome as annotated in the Genbank entries:

$$RSCU_{ij} = \frac{X_{ij}}{\sum_{j=1}^{n_i} X_{ij}} \times \frac{\sum_{j=1}^{n_i} Y_{ij}}{Y_{ij}}$$

Here, X_{ij} represents the number of observations of the j'th codon for the i'th amino acid in the set of highly expressed genes, whereas Y_{ij} is the corresponding number of observations in the background set. Furthermore, n_i is the number of codons for the i'th amino acid, with $RSCU_{i,max}$ being the highest number from the vector of $RSCU_i = (RSCU_{ij=1},..., RSCU_{ij=n_i})$.

The relative adaptiveness of a codon (w_{ij}) is calculated as:

$$w_{ij} = RSCU_{ij} / RSCU_{i,\max}$$

Subsequently, codon adaptation indices for individual coding regions were obtained as follows:

$$CAI = \exp \frac{1}{L} \sum_{k=1}^{L} \ln w_k$$

Here, L is the number of codons in a given gene.

In order to identify a set of constitutively highly expressed genes for each of the 318 bacterial genomes analyzed in this work, the reference set of 27 very highly expressed *E. coli* genes originally compiled by (Sharp*, et al.*, 1986a) was aligned at the protein level against all genes annotated in the Genbank entry for each genome using BLASTP version 2.2.9 (Altschul*, et al.*, 1997). For each of these very highly expressed genes, the gene with the best alignment was added to a set of very highly expressed genes if it had an E-value below 10^{-6}, and these were used as a reference set for the given organism. Similarly for the 5 fungal genomes, we used the reference set of very highly expressed *S. cerevisiae* genes identified by (Sharp*, et al.*, 1986b), removing the second ribosomal protein 51 gene (rbs51B), resulting in a list of 37 genes.

By this procedure, we were able to construct reference sets containing a minimum of 15 genes for the Firmicute *Clostridium tetani* E88, and a maximum of 27 highly expressed *E. coli* reference genes for 26 Proteobacteria strains. Consequently, bacteria more related to *E. coli* showed a higher level of conservation. Thus, the number of identified reference genes ranged from a median of 24 for Proteobacteria to a median of 21 for Actinobacteria. For the fungal genomes, a median of 36 genes was found in the reference sets.

Dominating Codon Bias Index (dCAI)

A purely mathematical CAI measure was proposed by Carbone et al. (2003) and in this paper we refer to this CAI measure as “dominant codon adaptation index” (dCAI). It detects the most dominant codon bias in the genome, regardless of whether this bias is translational or not. The algorithm screens a genome for genes that score the highest values on the CAI scale and selects these as its reference set. For dCAI values, we have used the tool CAIJava available from the authors (http://www.ihes.fr/~materials/description.html).

Data treatment

All DNA and protein sequence information was extracted from each Genbank entry. For correlation estimates, we used Spearman’s rank correlation (Best and Roberts, 1975) to avoid any problems with possible deviations from normality in compared data (*e.g.* log-normal distribution for microarray data). Cluster analysis was based on hierarchical clustering of Euclidian distances by complete linkage.

ADDITIONAL DATA FILES

The following additional data are available at http://www.cbs.dtu.dk/~hanni/CAI/. Overview of the microbial genomes included in this study linked to estimated tCAI and dCAI values. Supplementary Figure S1 is a detailed version of the cluster analysis in Figure 9-2, providing the full organism names.

ABBREVIATIONS

tCAI: translational codon adaptation index

dCAI: dominant codon adaptation index

GEO: Gene Expression Omnibus

ACKNOWLEDGEMENT

This study was supported financially by The Danish Center for Scientific Computing (HW, DWU, ASJ, CF) and the Danish Research Agency (CF).

Chapter 10 Paper VII

Chromatin dependent gene expression in microbes

Hanni Willenbrock and David W. Ussery

Center for Biological Sequence Analysis, Technical University of Denmark, DK-2800 Lyngby. Denmark.

ABSTRACT

It is well known that gene expression is dependent on chromatin structure in Eukaryotes and it is likely that chromatin can play a role in bacterial gene expression as well. Here, we use a nucleosomal position preference measure of DNA flexibility to predict highly expressed genes in microbial genomes, and compare this to genes predicted to be highly expressed based on their codon adaptation index (CAI) values. The predictions are compared to experimental data for 6 different microbial genomes, with a particular interest in experimental data from *Escherichia coli*. We find that gene expression is not only regulated by DNA structural elements such as DNA flexibility in terms of nucleosomal position preference, but that absolute gene expression levels are highly correlated with their individual level of DNA flexibility in multiple microbial genomes. For these, flexible DNA may be more accessible to the transcriptional machinery. This newly gained insight into DNA structure dependent gene expression in microbes may be exploited for predicting the expression of non-translated genes such as non-coding RNAs that may not be predicted by any of the conventional codon usage bias approaches.

INTRODUCTION

The transcription of DNA is highly influenced by the bending and flexibility of the DNA double helix. These structural properties are dependant on the base sequence (Baldi*, et al.*, 1996), which in turn, is reflective of, or may influence the codon usage - also important in determining the relative expression of a given gene. Prediction of highly expressed genes and elucidation of the physical and biological properties of highly expressed genes has recently been addressed by a number of studies (Karlin*, et al.*, 2003; Raghava and Han, 2005; Sharp*, et al.*, 1987).

The translational 'codon adaptation index' (CAI) is highly correlated with the expression level in fast growing bacteria (Carbone*, et al.*, 2005). It is based on the finding that highly expressed genes almost exclusively use those codons translated by abundant tRNAs in *Escherichia coli* and budding yeast (Sharp & Li, 1987). Consequently, a codon bias signature was deduced that was most likely to be efficient for translation. This bias was then used to derive codon adaptation indices for all genes for a given organism, where high CAI values correspond to genes most likely to be highly expressed.

However, using CAI, one is only able to predict highly expressed proteins (translated genes) since this measure is based on codon usage bias. Unfortunately, this method cannot consider tRNAs, ribosomal RNAs, and other non-coding RNAs. Furthermore, when it comes to organisms with low translational bias – typically slow growing organisms - CAI is a less effective predictor of highly expressed genes.

On a more global scale, gene expression may be regulated from specific promoters that are sensitive to DNA superhelicity. That is, supercoiling may regulate gene expression at a genome-wide level (Peter*, et al.*, 2004; Willenbrock and Ussery, 2004). In this way, an organism may react rapidly to changes in growth and nutritional states as well as environmental conditions since DNA superhelicity varies with the cellular energy charge, which, for example, differs in log phase versus stationary phase or is influenced by environmental factors such as temperature or osmotic stress (Hatfield and Benham, 2002). Moreover, it is well known that DNA supercoiling can affect gene expression at the level of promoter activity by changing the shape of DNA. While negative supercoiling may facilitate promoter melting and consequently transcription initiation, it may also repress it (Drolet 2006). Some of the DNA supercoiling inside cells is restrained by wrapping around proteins in torroidal supercoils, and the remaining supercoils are "unrestrained", in the form of plectonemic supercoils. In eukaryotes, most of the supercoiling is thought to be restrained around nucleosomes, with little free supercoiling that can drive transitions such as opening of the double helix. However, in bacteria, which contain a much higher coding fraction of DNA, the ratio between restrained and unrestrained supercoiling is about equal.

The 'position preference' measure was originally derived for Eukaryotes using chicken DNA and is a trinucleotide model of nucleosome positioning patterns. It reflects the preference of a given trinucleotide for being found in a region where the DNA minor groove faces either towards or away from the nucleosome histone core (Satchwell*, et al.*, 1986). Here, we use a minor modification of the original nucleosomal positioning trinucleotide scale where absolute values reflect the magnitude of position preference (Pedersen*, et al.*, 1998). Thus, high absolute position preference reflects a high preference for nucleosomes; while low absolute position preferences reflect trinucleotides which tend to exclude nucleosomes. While this only makes sense in Eukaryotes since Prokaryotes do not have nucleosomes, these preference values are also a measure of DNA flexibility since flexible sequences can occupy any rotational position on nucleosomal DNA, while rigid sequences are restricted in their rotational location. Consequently, the 'position preference' measure may also describe a structural property of prokaryotic DNA. As a result, it has been used previously to show

structural characteristics in prokaryotic genomes (Pedersen*, et al.*, 1998; Pedersen*, et al.*, 2000).

By a cluster analysis of various structural properties including position preference, groups of genes were identified that contained all the ribosomal RNAs and a majority of the ribosomal proteins from *Escherichia coli* (Pedersen*, et al.*, 2000). These genes were characterized by higher than average DNaseI sensitivity (Brukner*, et al.*, 1995) and low position preference, indicating flexible DNA. Since the ribosomal genes are among the most highly expressed in actively dividing *E. coli* cells, it was hypothesized that their common structural features may play a role in regulating expression and that there exists a correlation between low position preference values and highly expressed genes (Dlakic*, et al.*, 2004). This makes sense because regions of DNA that are not condensed by chromatin are more accessible to the RNA polymerase. Consequently, transcription is thought to be governed by 'effective' superhelicity, where topoisomerases, the transcription machinery and chromatin proteins compete for available supercoils (Blot*, et al.*, 2006).

Here, we use the position preference measure to predict highly expressed genes and compare it to the CAI measure while evaluating the functional categories of genes with low position preference. The predictions are compared to experimental data for 6 different microbial genomes, including data from *E. coli* samples taken at various stages during growth, at which varying levels of global supercoiling is expected.

MATERIALS AND METHODS

Translational Codon Adaptation Index (CAI)

The codon adaptation index describes a codon usage bias in an organism (Sharp*, et al.*, 1987). Here, we use a translational codon adaptation index (CAI), in which a codon bias signature is deduced that is most likely to be efficient for translation (Willenbrock et al., submitted Genome Biology). In short, this method is based on a known set of 27 very highly expressed *E. coli* genes for bacterial genomes (Sharp*, et al.*, 1986a), and a set of 39 very highly expressed Yeast genes for Eukaryotes (Sharp*, et al.*, 1986b). In order to identify a set of constitutively highly expressed genes for each of the bacterial genomes analyzed in this work, the reference set of very highly expressed *E. coli* or Yeast genes is aligned at the protein level against all genes annotated in the Genbank entry for each genome using BLASTP version 2.2.9 (Altschul*, et al.*, 1997). For each of these very highly expressed genes, the gene with the best alignment was added to a set of very highly expressed genes if it had an E-value below 10^{-6}, and these were used as a reference set for the given organism. Using each genome specific reference set, a weight table including all codons is derived indicating the most translationally efficient codons. In turn, these weights are used for calculating a CAI value for each gene. The higher the CAI score, the more likely a gene is to be highly expressed.

Position Preference

This is a model of DNA flexibility, which is derived experimentally from the preference demonstrated by individual trinucleotides to be positioned in a specific orientation in nucleosomal DNA (Satchwell*, et al.*, 1986). The values indicate the preference of triplets for being specifically positioned in nucleosomal DNA. High absolute values correspond to triplets with a strong preference for having minor grooves facing either towards or away from the nucleosome core, while triplets with close-to-zero preference can occupy any rotational position on the nucleosomal DNA, and are thus assumed to be flexible. Since the 'position preference' measure is based on a simple trinucleotide model, values are assigned to every nucleotide in the DNA sequence simply by looking up the values for the corresponding triplet, in which the nucleotide is centered (Baldi*, et al.*, 1996; Pedersen*, et al.*, 1998; Pedersen*, et al.*, 2000).

Assigning Cluster of Orthologous Genes (COGs)

The system for delineation of Clusters of Orthologous Groups of proteins (COGs) is based on orthologous relationships between genes and is useful for comparative genomics and facilitates the functional annotation of genomes. Here, genes were assigned a COG category by AutoFACTS, an automatic functional annotation tool (Koski*, et al.*, 2005) utilizing Blastx version 2.2.9 (Altschul*, et al.*, 1997) to blast open reading frames to a database of sequences with assigned cog categories available from NCBI (ftp://ftp.ncbi.nih.gov/pub/COG/COG/).

Prediction of ribosomal proteins

Ribosomal proteins for each Genbank entry were predicted using profile HMMs from Pfam (http://www.sanger.ac.uk/Software/Pfam/) since the quality of the annotations available from the Genbank entries varies tremendously. Pfam_ls profile HMMs for all ribosomal proteins were extracted (94 as per July 24th 2006). Pfam_ls files contain all the Pfam models for finding global or complete matches to a domain or family.

Topological domains

Starting from the beginning of the genomic sequence, position preference values for segments with an initial size of 55000 bp were compared to position preference values for the following 1000 bp's. These bp's were added to the initial segment unless it was significantly different (Kolmogorov-Smirnov P-value < 1e-6). Thereby, segments were 'grown' until no additional bp's could be added, following which, a new segment was started. The parameters (initial size and increment) were chosen by lowest possible SSQ (summed squared residuals of differences between original position preference values and region mean position preference values) among a number of test runs with varying initial size and varying increment size.

Gene Expression Data

Microarray based gene expression data were taken from (Willenbrock et al, submitted Genome Biology). Briefly, the dataset comprised pre-processed gene expression data for *E. coli* (Covert*, et al.*, 2004), *C. jejuni* (Stintzi*, et al.*, 2003), *P. aeruginosa*, *S. cerevisiae* (Bulik*, et al.*, 2003; Ronald*, et al.*, 2005), *G. sulfurreducens* (Methe*, et al.*, 2005), and *B. subtilis* (Helmann*, et al.*, 2003). Additional microarray gene expression data for *E. coli* at different growth stages were taken from (Tjaden*, et al.*, 2002), where raw data were normalized with qspline (Workman*, et al.*, 2002) and expression indices were estimated (Li*, et al.*, 2001a).

Data Treatment

All DNA and protein sequence information was extracted from each Genbank entry. For correlation estimates, we used Spearman's rank correlation (Best*, et al.*, 1975) to avoid any problems with possible deviations from normality in compared data (e.g. log-normal distribution for microarray data). Cluster analysis was based on hierarchical clustering of Euclidian distances using complete linkage.

Supplemental information

Additional data are available at http://www.cbs.dtu.dk/~hanni/Chromatin/. This website contains an overview of the microbial genomes included in this study linked to estimated position preference values. Supplementary Figure S1 is a detailed version of the heatmap sketched in Figure 10-1, providing the full organism names of all included microbial genomes. Supplementary table S1 and S2 provides some statistics for the comparison of expression values and CAI and position preference.

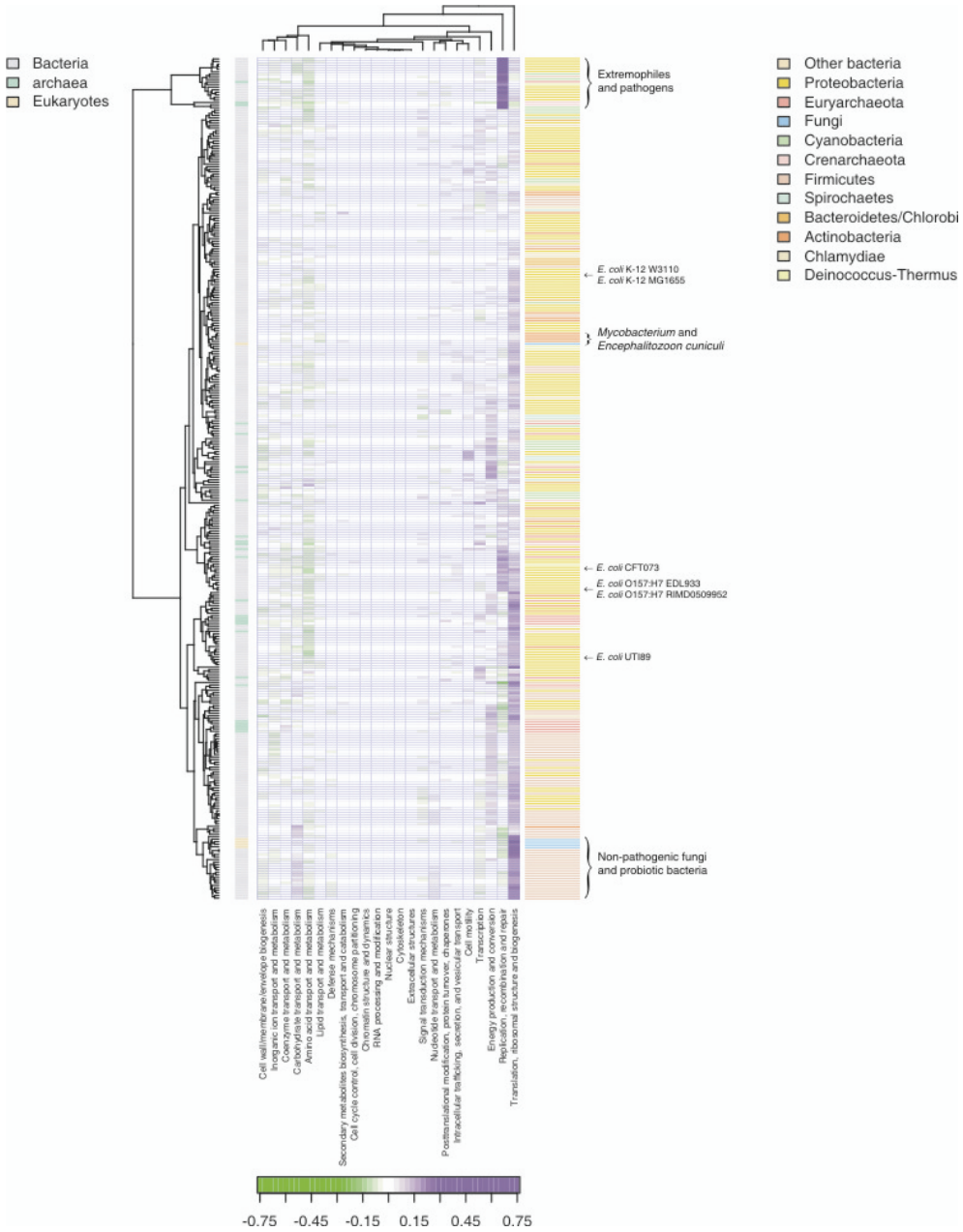

Figure 10-1. Heatmap of COG categories for genes with low position preference (10% lowest). Over-represented categories among genes with lowest position preference compared to the genomic background is indicated with purple, while green indicates under representation. The kingdom is indicated as a vertical color bar to the left of the heatmap and the phyla as a vertical color bar to the right.

RESULTS AND DISCUSSION

Functional categories of genes with low position preference

In fast growing organisms, ribosomal proteins and other proteins involved in translation and transcription are often highly expressed and are extremely biased in their codon usage preferences, that is, they have high CAI values (Carbone*, et al.*, 2005). Genes involved in translation, transcription, replication, and energy production are often encoded by flexible DNA in terms of low position preference values which is thought to be correlated with high gene expression (Figure 10-1). Figure 10-1 (and supplementary Figure S1) illustrates over-represented (purple) and under-represented (green) COG categories among genes with low position preference relative to the genomic background. The COG categories and the microbes are clustered in two dimensions by hierarchical clustering and the microbes do not cluster according to AT content (data not shown) as we found when clustering based on codon usage bias (Willenbrock *et al.,* submitted Genome Biology). Instead, it is possible to see the COG categories of genes encoded by DNA with low position preference. For most microbes, DNA with low position preference encodes genes involved in 'translation, ribosomal structure and biogenesis', 'energy production and conversion', 'transcription', and various types of metabolism.

It is clear that the clustering brings together organisms which are relatively distant phylogenetically (Figure 10-1, right side color bar representing the taxonomic phylum of each genome). One possible explanation for the clustering is similar environments as found based on CAI (Willenbrock *et al.,* submitted Genome Biology). However, in the present analysis, the ordering may also be related to the functionality of the microbe, i.e. pathogen versus non pathogen. For example, the COG category 'replication, recombination and repair' is particularly over represented amongst genes with low position preference for a distinct cluster at the top of Figure 10-1, consisting of extremophilic Archaea and Bacteria as well as pathogenic bacteria (mainly *Yersinia pestis* strains). The common feature of these organisms is that genes involved in replication, recombination and repair have very low position preference (and consequently are potentially highly expressed). Particularly genes involved in recombination and repair are essential for pathogens and microbes living under extreme conditions making it reasonable for them to be highly expressed. Supporting this observation, we find that the same COG category is over represented for pathogenic *E. coli* strains, O157:H7 EDL933, O157:H7 RIMD0509952, CFT073 and UTI89 as well as the *E. coli* like pathogen, *Shigella*, whereas, the same COG category is not dominating for the non pathogenic *E. coli* strains K-12 W3110 and K-12 MG1655. This provides us with a possible means for distinguishing pathogenic strains from non pathogenic strains.

Next, we observe that the eukaryotic intracellular parasite, *Encephalitozoon cuniculi,* clusters very close to the *Mycobacterium* species, also intracellular pathogens and similar in that they all have reduced genomes. Moreover, three non-pathogenic fungi are closely clustered with certain probiotic bacteria (*Lactobacillus*); it is interesting to note that these organisms can live in a similar ecological niche. Also, a few microbes contain genes with low position preference that are involved in carbohydrate transport and metabolism, especially the *Streptococcus* genomes found in the bottom cluster of Figure 10-1. Again, this might be reflective of their ecological niche.

Ribosomal proteins and non-translated RNA

Examining the ribosomal proteins for *E. coli*, we confirm that the average position preference is lower (mean=0.1399) than for other protein encoding genes (mean=0.1465) (Wilcoxon P-value 4e-11), and it is even more extreme for non-translated genes. For example, rRNAs, tRNAs, and miscellaneous RNAs have significantly lower position preference values than translated genes (P-value = 6e-34). This is true especially for rRNAs and some tRNAs, with 16S rRNAs having values of 0.132 or less and asparagine tRNA genes having values below

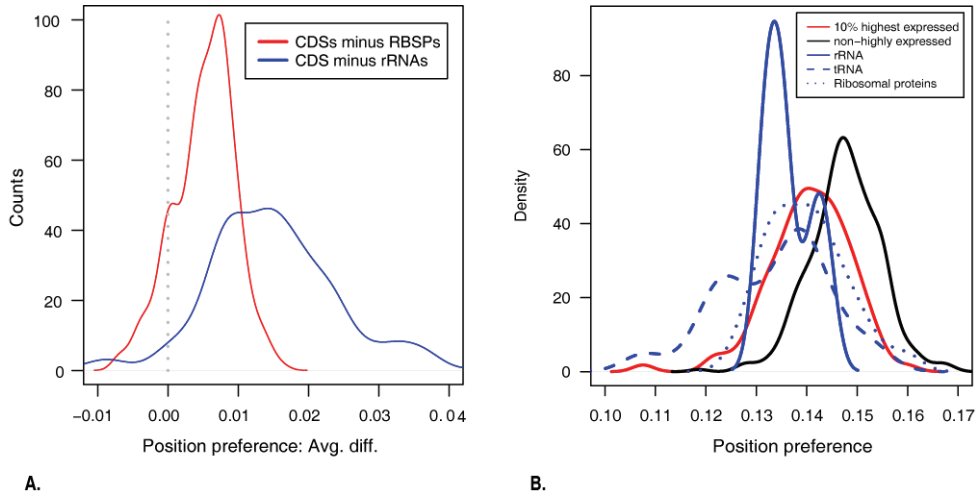

Figure 10-2. Position preference differences for 360 microbes. (A) Density plot for differences between translated coding sequences (CDSs) and ribosomal proteins or versus ribosomal RNA. (B) Densities of the 10% most highly expressed genes, non-highly expressed genes, rRNAs, tRNAs and ribosomal proteins in *E. coli*.

0.11. Although this difference was observed for a majority of microbial genomes, across a vast range of microbial genomes, the ribosomal proteins are not always encoded by flexible DNA (Figure 10-2A). However, the difference in position preference of ribosomal proteins and non-ribosomal proteins correlated very well with the replication times of the cells using the number of 16S rRNAs as an indirect measure of doubling time, as previously suggested (Sharp*, et al.*, 2005), since the number of 16S rRNAs indirectly influence replication times (Ussery*, et al.*, 2004). Consequently, as for CAI (Willenbrock *et al.,* submitted Genome Biology), fast replicating microbes have optimized their translational machinery by increasing the expression of proteins such as ribosomal proteins. As a result, their expression is optimized both by codon usage and by placing them on easy assessable flexible DNA (Wilcoxon P-value ~ 0, rho=0.42). The above results are somewhat in contrast to the findings by Segal and coworkers (Segal*, et al.*, 2006). They recently published a more refined model for nucleosomal positioning based on a combined experimental and computational approach. Although this model predicts a nucleosome pattern strikingly similar to that of the model used in our study (Satchwell*, et al.*, 1986; Segal*, et al.*, 2006), at least for Eukaryotes, they did not find nucleosome depletion at ribosomal proteins sites in Yeast, i.e. they predicted high nucleosome occupancy encoded over these genes, reasoning that the expression of these genes must be governed by other factors. However, although we only predict a slightly lower than average position preference for yeast ribosomal proteins, we find that the general trend observed across a large range of microbial genomes is that both DNA encoding ribosomal proteins and non-coding genes have lower position preference than the genomic average (Figure 10-2). This indicates a possible regulation of ribosomal proteins by DNA structural properties.

Prediction of highly expressed genes

The above analysis demonstrate that the functional categories of the genes with low position preference often resemble the functional categories of genes with high codon usage bias in terms of high CAI values (i.e. highly expressed ribosomal proteins). Consequently, we would expect a similar correlation between low position preference and high gene expression level. Nonetheless, a complete separation of highly expressed genes from the other genes was

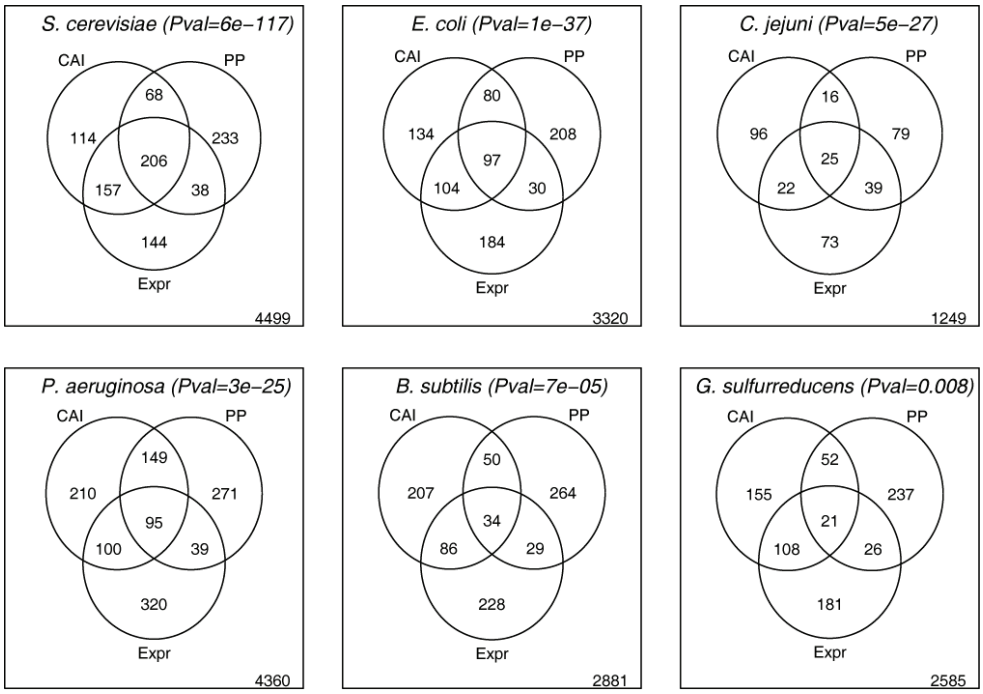

Figure 10-3. Venn diagram of overlap between genes with high CAI values (top 10%), low position preference (10%) and high microarray gene expression values (top 10%). The organisms are ordered by the significance of the overlap between position preference (PP) and highly expressed genes (Fisher's exact test).

not possible using the position preference measure (for example, see Figure 10-2B). However, this is hardly surprising since no structural or coding property singularly determines the level of gene expression, for which a large number of regulatory steps are involved. Consequently, the level of separation may reflect the influence of each measure on gene expression. For the five additional microbial genomes where we have examined expression levels based on microarrays, a clear difference was also observed between the distributions of CAI or position preference values for highly expressed genes and low expressed genes (supplementary Table S1).

As expected from the above analyses, we observe a significant enrichment in highly expressed genes among genes with low position preference (Figure 10-3) for all 6 organisms for which we have microarray gene expression data available. Moreover, the correlation between position preference values and microarray gene expression values is highly significant (supplementary Table S2). However, the overlap between genes with high CAI values and highly expressed genes is even more significant (Figure 10-3). While this is expected since codon usage is known to have a strong influence on protein expression, the DNA structural properties also influence gene expression, and it seems reasonable that DNA which cannot be condensed into tightly wrapped chromatin structures is more accessible to RNA polymerase, which is about the same size as a nucleosome. One likely explanation is that position preference, as a measure of chromatin structure, might not be the most optimal – particularly for bacterial genomes. This might also explain the considerably higher enrichment in highly expressed genes among *S. cerevisiae* genes with low position preference than observed for the bacterial genomes (Figure 10-3).

While CAI values are better predictors of high expression of proteins, DNA structural properties may be used for prediction of gene expression for non-translated genes such as transfer RNAs and micro RNAs. For example, for *E. coli*, gene expression levels were further available for some non-translated genes. Including these in the comparison, the correlation between gene expression and position preference was even more significant (P-value= 1.2e-52 compared to P-value = 1.7e-39 for translated genes only) and the overlap between genes with low position preference values and genes with high expression values were also more significant (p-value: 9.8e-47 compared to 1.3e-37 for translated genes only). This demonstrates that not only may the position preference measure be used for predicting the gene expression level for this type of coding regions, but since these regions are even more correlated with DNA flexibility than translated genes, they may consequently be under even more strict regulation by DNA structural properties. This makes sense since regulation by codon usage is not an option for these transcripts.

Topological domains

The *E. coli* chromosome is thought to consist of a number of fluid short-range distinct topological domains (Postow*, et al.*, 2004; Willenbrock*, et al.*, 2004). The lack of a single key component of bacterial chromatin can result in reorganization of these domains and affect supercoiling sensitivity dependent genomic transcription (Blot*, et al.*, 2006). On a chromosomal level, we observe that, for the *E. coli* genome, both CAI values and position preference values predict the same general regions of highly expressed genes, and indeed these regions correspond well with the experimental expression values (Figure 10-4). Consequently, our data show that not only is the gene expression regulated by DNA structural elements as demonstrated previously (Blot*, et al.*, 2006; Peter*, et al.*, 2004), but

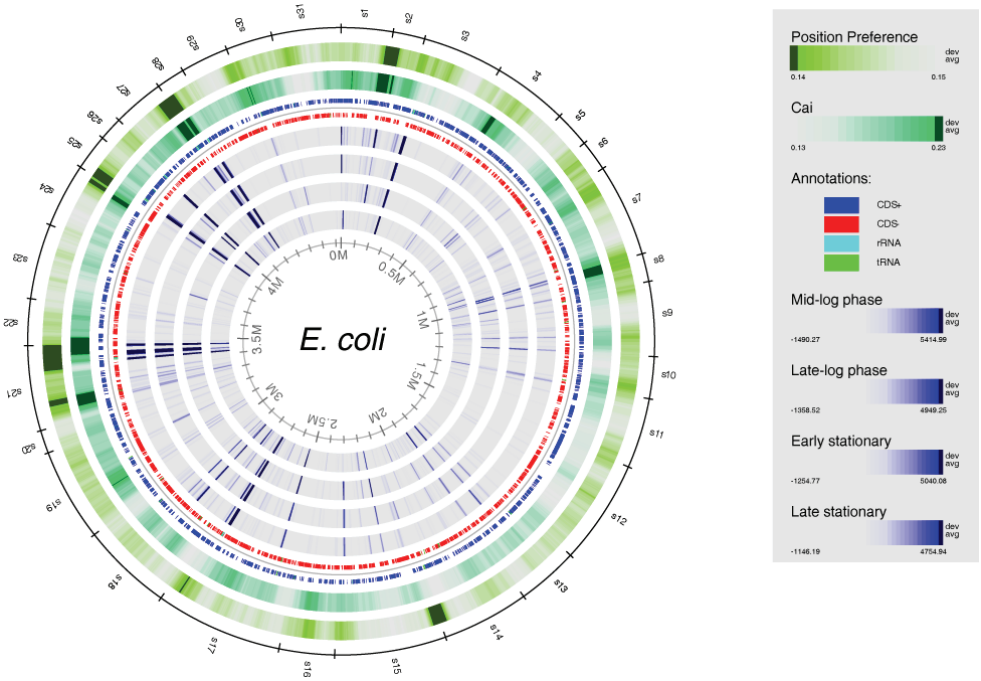

Figure 10-4. Atlas illustration of the *E. coli* genome. The atlas illustrates CAI values, position preference values and gene expression values at various growth stages based on data from (Tjaden*, et al.*, 2002).

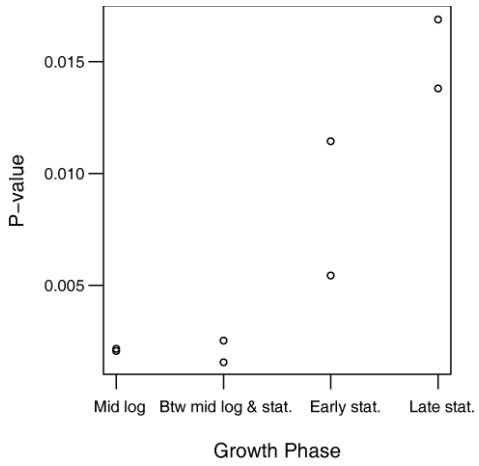

Figure 10-5. Illustration of the significance P-value of the correlation between mean position preference of 31 *E. coli* topological domains and the mean gene expression values of genes within each of these domains. The correlation is most significant in log-phase and less significant in stationary phase of growth. Gene expression data from (Tjaden*, et al.*, 2002).

the highly transcribed regions are also correlated with regions of the chromosome having low position preference. However, there are two regions where CAI and position preference differ significantly – the first is around 0.5 Mbp, where a cluster of highly expressed genes is predicted by CAI but not by position preference. This region contains the cyoA-cyoE genes involved in aerobic energy metabolism and they predominate during growth at high aeration. The second different region is towards the bottom of the atlas, around 2 Mbp, where there is a region with low position preference, but close to average CAI values. This region contains genes from the *flu* loci, which can be highly expressed under the appropriate environmental conditions (Schembri*, et al.*, 2002).

As can be seen in Figure 10-4, the gene expression levels vary between the different growth stages. To further investigate this apparent correspondence, we used a simple scanning statistical approach to divide the genome into distinct topological domains according to position preference. By this approach, the *E. coli* main chromosome was divided into 31 distinct regions (Figure 10-4, s1-s31) with varying mean position preference.

Within these topological domains, mean gene expression correlated well with mean position preferences of domains during log-phase growth, but were less significantly correlated for *E. coli* cells in stationary phase (Figure 10-5). This is consistent with the fact that genomic supercoiling is reduced in stationary phase as compared to the exponential growth phase, and that the coordination of growth-phase-dependent transcription involves a mechanism for reorganizing the supercoiling sensitivity (Blot*, et al.*, 2006) and regulating the overall transcription level.

CONCLUSION

We use a nucleosome position preference measure of DNA flexibility to predict highly expressed genes in microbial genomes, and compare it to a translational codon adaptation index for synonymous codon usage bias of potentially highly expressed genes. We hereby demonstrate that absolute gene expression levels are highly correlated with their individual level of DNA flexibility in multiple microbial genomes. This newly gained insight into DNA structure dependent gene expression may be exploited for predicting the expression of non-translated genes such as non-coding RNAs that may not be predicted by any of the conventional codon usage bias approaches. Genes often encoded by DNA with low position

preference values where mostly involved in ‘translation, ribosomal structure and biogenesis’, ‘energy production and conversion’, and transcription. For pathogens and microbes living in extreme environments, the predominant functional category was ‘replication, recombination and repair’. In particular, for *E. coli* pathogenic strains demonstrated this trait while non pathogenic strains did not. This provides a likely signature for distinguishing some pathogenic strains from non pathogens.

ACKNOWLEDGEMENTS

This study was supported financially by The Danish Center for Scientific Computing.

Perspectives

DNA microarrays provide for a much-needed high-throughput genetic methodology in biological research by facilitating the exploitation of the vast amount of DNA sequence information that is quickly becoming available. Thus, DNA microarrays will continue to play an increasing role in both genomic research and in transcriptomics. Nonetheless, the technology can still be improved significantly, since microarray data are still very noisy and poor reproducibility has been observed. However, the latter may be attributed the attempt to compare results from severely under-powered studies (too few replicates) or studies that are otherwise flawed in their experimental design or execution. Consequently, the experimental design and analysis of this data type requires at least basic statistical knowledge. Even so, within the data mining field, increasingly sophisticated algorithms are being published and semi-automated analysis tools are being developed both for gene expression analysis (Saeed*, et al.*, 2003; Saal*, et al.*, 2002) and comparative genomics (Liva*, et al.*, 2006; Menten*, et al.*, 2005; Myers*, et al.*, 2005). In the future, together with improvements in the technology to decrease noise levels, this might facilitate dissemination of the technology even further and lead to the discovery of even more remarkable biology than is currently being reported.

Gene Expression Microarrays

Due to the considerable noise levels associated with this kind of data, it has not proven very reliable, nor interesting for that matter, to analyze lists of differentially expressed genes by looking for 'interesting' genes. Consequently, the more successful studies have utilized alternative approaches for exploiting the advantages of such high-throughput data. For example, in Chapter 4, a classifier is build that utilizes a combination of predictive genes from samples from children with acute lymphoblastic leukemia. Hereby, it is possible to classify their subtype and to predict how well they respond to treatment. Especially the prediction of response to early treatment appears promising and may have clinical applications for treatment stratification in the early course of childhood leukemia. The latter has recently been supported by similar findings by Cario and coworkers (Cario*, et al.*, 2005). In the future, studies should aim at finding diagnostic traits for early and late treatment response both in terms of gene expression profiles as well as cytogenetic traits that may, for example, be addressed by microarrays directed at the genomic sequence as is the case for array CGH. With Affymetrix and FDA recently teaming up with several microarray manufacturers to advance the use of microarrays in clinical studies, we are now one step closer to actually using the DNA microarray technology clinically and not just in basic research. In the future, this could improve the diagnosis and prognosis of individual patients and promote the tailoring of individual treatments.

Currently, the field of gene expression microarray data analyses is moving towards more integrative approaches trying to fuse the microarray analysis results with various other data types or integrating microarray data from several studies. However, for the latter, only limited success has been reported due to high variability between studies as well as poor experimental designs. Even with these restrictions, in a novel study (Chapter 5), we present a method that successfully exploits the existing repositories of microarray data from multiple studies and various laboratories to derive functional associations between gene expression responses from, for instance, a given plant mutant compared to a compendium of gene expression responses from plant mutants and plants subjected to various treatments. By this approach, an extensive functional characterization of a given mutant may be obtained. The same characterization could otherwise be both time consuming and depend on extensive background knowledge of the investigated biological system. By limiting the direct comparison of samples to within an individual experiment or study, the method benefit from the great care with which each experimentalist has ensured comparability within their own experimental designs. Thus, this approach demonstrates an excellent capability for

promoting between study comparability and consequently, for deriving biologically meaningful functional association between experimental factors derived from microarray studies of independent laboratories.

In the future, existing or emerging techniques for gene silencing (Hilson*, et al.*, 2004) may provide a much needed short cut for quickly expanding the existing repositories of knockout gene expression responses, allowing even more detailed microarray based functional studies. Since promising results were obtained for the model plant, *Arabidopsis thaliana* and bakers' yeast, *Saccharomyces cerevisiae*, the same approach would be useful for deriving functional associations and characterization of other species, for which extensive microarray data is available, including both simpler organisms such as *E. coli* and more complex organisms such as humans. For example, the same approach might be useful for associating cancer gene expression response phenotypes to a compendium of cancer responses for diagnostic purposes. It would also be interesting and highly relevant to explore cross-species applications to functionally characterize experimental factors for organisms where extensive microarray data repositories are not available. Here, promising results have already been obtained in and attempt to characterize Barley experiments with regard to functional associations to our Arabidopsis compendium of gene expression responses.

Comparative Genomics

Several segmentation algorithms for the analysis of array CGH data have been proposed and it has been found that the application of many of these improve downstream analysis of DNA copy number data. Of the segmentation methods compared in Chapter 6, a non-parametric change-point method (DNAcopy) (Olshen et al., 2004) was found to have the best operational characteristics in terms of its sensitivity and false discovery rate for breakpoint detection. Furthermore, by applying the additional merging procedure, MergeLevel, copy numbers can be estimated directly across chromosomes.

Nonetheless, we also speculated that an HMM approach might be more naturally adaptable to perform such a whole genome fit and could perform constrained optimization of the segment means across the entire genome. Supporting this, a hidden markov model was recently proposed that incorporates relevant biological factors such as the distance between adjacent probes (Marioni*, et al.*, 2006). Also, a pseudolikelihood approach with a hidden Markov dependency structure has been suggested. It borrows strength across chromosomes and hybridizations and demonstrated a superior performance to the DNAcopy + MergeLevel approach (Engler*, et al.*, 2006).

To obtain a realistic and comparable measure of performance for different algorithms, a simulation model was developed for generating artificial data with known breakpoints and known DNA copy number. The simulation model provides a test data set that may be used repeatedly to benchmark the performance of new methods. If every new method was to be tested on this simulated data, it would provide an accurate and comparable evaluation of new method's ability to analyze DNA copy number data. Consequently, the literature would avoid being swamped by suggestions of new data analysis methods that do not perform better than the previous. This would facilitate the choosing of analysis tools for the non-statistical researcher.

Our next study shows that analysis methods developed for cancer research may also successfully be applied to analyze DNA copy number profiles from bacterial genomes. However, here the purpose was to characterize variations in the gene content of various strains of the bacteria, *Escherichia coli*, with regard to genes involved in pathogenesis, and to study horizontal gene transfer. Further developments would aim at optimizing the design for making high-throughput characterization of multiple strains feasible. For example, currently, the complete sequences of more than 20 *E. coli* strains are available. The challenge is then to optimize the microarray design for simultaneously targeting of all these

strains without redundancy. That is, to be able to distinguish between all existing strains as well as characterize new strains by optimizing consensus sequence design of backbone genes and at the same time targeting the between-strain sequence variations that are optimal for distinguishing between individual strains of the same species.

Nonetheless, with the prices on sequencing dropping continuously and the technology fast improving both in speed and quality, the array CGH technology may soon be rendered obsolete. Much more detailed information could quickly be obtained by merely sequencing the full sequence of an organism or even a cancer genome in the far future. However, use of super computers and advanced statistics to process this data and interpret the results would then become even more imperative.

Sequence Dependent Gene Expression

Gene expression is regulated by a large number of regulatory elements both at the transcriptional and translational level, and may consequently be influenced by, for example, DNA/RNA structural properties and codon usage. Gene expression may be measured both at the transcriptional level in form of RNA abundance and at the translational level in form of protein abundance. Often, we are most interested in the protein abundance since this is where a difference in abundance will usually have the largest impact on phenotype. However, the use of high quality experimental protein abundance measurements has been hampered by the fact that such data are not readily available. Consequently, we found that microarray gene expression data can be useful for confirming predicted highly expressed genes - as a substitution for protein levels, although they were – at best – very rough estimates of the true protein concentrations.

The newly gained insight into sequence dependent gene expression may be explored further for developing a reliable predictor of gene expression levels. This has already been attempted with promising results (Raghava*, et al.*, 2005). However, if available, high-throughput protein abundance data would assist in developing such a predictor even further, since a prediction method can never be better than the data. That is, if microarray gene expression data is only 60% correlated with the actual protein abundance levels, then the predictions will at the most also be 60% correlated with the true protein abundance. Recent developments in technologies measuring protein abundance, e.g. mass spectrometry, may result in an upcoming explosion in available proteomics data as seen for transcriptomics and genomics data during the past few years. This may facilitate the development of the above described predictor and improve our understanding of sequence dependent gene expression.

The prediction of DNA structure dependent gene expression is less optimal than predictions based on codon usage bias, especially for prokaryotic genomes. However, the used position preference measure of DNA flexibility was based on measurements in eukaryotes (chicken). In the future, development of a corresponding measure of DNA flexibility or chromatin structure specifically for prokaryotes, may improve the prediction of highly expressed genes in these microbes based on DNA structural properties.

A Look into the Future

In the future, DNA microarrays directed at the genomic sequence may become obsolete, at least for small microbial genomes due to reduced costs of sequencing. However, new applications of microarrays are constantly being sought after. For example, the new field of ‘metagenomics’ is currently expanding to ‘meta-transcriptomics’ for which, a high-throughput analysis methods such as DNA microarrays would be highly useful. Although data from gene expression microarrays are still considered quite noisy and while this would restrict such analyses, the technology constantly goes through major refinements and new sources of variation and how to deal with them are continuously being discovery. Consequently, in the future, microarray techniques with minimal experimental noise might evolve. On the

other hand, with the increasing amount of sequence data being published, microarray designs may become even more sophisticated in the attempt to cover multiple strains of the same species or even several different species. Discovery of whole pathogenicity networks may lead the way for designing a microarray aimed at identifying emerging pathogenic strains even before they become dangerous or for identification of pathogens in environmental samples. Multi-genome microarrays may also become efficient for environmental purposes, for example to determine degrees of pollution or certain environmental changes, since the specific composition of a microbial community largely depends on the surrounding environment.

New applications also include microarrays for proteomics. Exploitation of this technique in combination with existing techniques for transcriptomics and genomics, might aid in bridging the gap from transcription to translation and aid in discovering new regulatory elements in these mechanisms as well as improve our understanding of gene expression as a whole. Also, further advancements in this technology may allow protein science to enter a true ‘omics’ age as we have seen first for genomics and later for transcriptomics. In the future, this may result in extensive the analyses of the whole human proteome.

References

Acheson, D.W., Reidl, J., Zhang, X., Keusch, G.T., Mekalanos, J.J. and Waldor, M.K. (1998) In vivo transduction with shiga toxin 1-encoding phage, *Infection and Immunity*, **66**, 4496-4498.

Adams, M.D., Celniker, S.E., Holt, R.A.*, et al.* (2000) The genome sequence of Drosophila melanogaster, *Science*, **287**, 2185-2195.

Albertson, D.G. and Pinkel, D. (2003) Genomic microarrays in human genetic disease and cancer, *Human Molecular Genetics*, **12 Spec No 2**, R145-152.

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, *Nucleic Acids Research*, **25**, 3389-3402.

Ananiev, E.V. and Gvozdev, V.A. (1974) Changed pattern of transcription and replication in polytene chromosomes of Drosophila melanogaster resulting from eu-heterochromatin rearrangement, *Chromosoma*, **45**, 173-191.

Andreasson, E., Jenkins, T., Brodersen, P.*, et al.* (2005) The MAP kinase substrate MKS1 is a regulator of plant defense responses, *EMBO Journal*, **24**, 2579-2589.

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, *Infection and Immunity*, **71**, 4674-4683.

Antoniadis, A., Lambert-Lacroix, S. and Leblanc, F. (2003) Effective dimension reduction methods for tumor classification using gene expression data, *Bioinformatics*, **19**, 563-570.

AtGenExpress (2006) AtGenExpress: a multinational coordinated effort to uncover the transcriptome of *Arabidopsis*.

Auburn, R.P., Kreil, D.P., Meadows, L.A., Fischer, B., Matilla, S.S. and Russell, S. (2005) Robotic spotting of cDNA and oligonucleotide microarrays, *Trends in Biotechnology*, **23**, 374-379.

Baldi, P. and Brunak, S. (1998) *Bioinformatics – The Machine Learning Approach*. The MIT Press, Cambridge, Massachusetts, London, England.

Baldi, P., Brunak, S., Chauvin, Y. and Krogh, A. (1996) Naturally occurring nucleosome positioning signals in human exons and introns, *Journal of Molecular Biology*, **263**, 503-510.

Bauer, D.F. (1972) Constructing confidence sets using rank statistics, *Journal of the American Statistical Association*, **67**, 687–690.

Beier, M. and Hoheisel, J.D. (2000) Production by quantitative photolithographic synthesis of individually quality checked DNA microarrays, *Nucleic Acids Research*, **28**, E11.

Ben-Shaul, Y., Bergman, H. and Soreq, H. (2005) Identifying subtle interrelated changes in functional gene categories using continuous measures of gene expression, *Bioinformatics*, **21**, 1129-1137.

Benjamini, Y. and Hochberg, Y. (1995) Controlling the false discovery rate:
A practical and powerful approach to multiple testing., *Journal of the Royal Statistical Society Series*, **B 57**, 289–300.

Bergthorsson, U. and Ochman, H. (1998) Distribution of chromosome length variation in natural isolates of *Escherichia coli*, *Molecular Biology and Evolution*, **15**, 6-16.

Best, D.J. and Roberts, D.E. (1975) Algorithm AS 89: The Upper Tail Probabilities of Spearman's rho, *Applied Statistics*, **24**, 377-379.

Binnewies, T.T., Motro, Y., Hallin, P.F., Lund, O., Dunn, D., La, T., Hampson, D.J., Bellgard, M., Wassenaar, T.M. and Ussery, D.W. (2006) Ten years of bacterial genome sequencing: comparative-genomics-based discoveries, *Funct Integr Genomics*, **6**, 165-185.

Blattner, F.R., Plunkett, G., 3rd, Bloch, C.A.*, et al.* (1997) The complete genome sequence of *Escherichia coli* K-12, *Science*, **277**, 1453-1474.

Blot, N., Mavathur, R., Geertz, M., Travers, A. and Muskhelishvili, G. (2006) Homeostatic regulation of supercoiling sensitivity coordinates transcription of the bacterial genome, *EMBO Rep*, **7**, 710-715.

Bolstad, B.M., Irizarry, R.A., Astrand, M. and Speed, T.P. (2003) A comparison of normalization methods for high density oligonucleotide array data based on variance and bias, *Bioinformatics*, **19**, 185-193.

Bonferroni, C.E. (1936) CE Teoria statistica delle classi e calcolo delle probilitá., *Pubblicazioni del R Istituto Superiore de Scienze Economiche e Commerciali di Firenze*, **8**, 3-62.

Bowling, S.A., Clarke, J.D., Liu, Y., Klessig, D.F. and Dong, X. (1997) The cpr5 mutant of *Arabidopsis* expresses both NPR1-dependent and NPR1-independent resistance, *Plant Cell*, **9**, 1573-1584.

Brodersen, P., Petersen, M., Bjorn Nielsen, H., Zhu, S., Newman, M.A., Shokat, K.M., Rietz, S., Parker, J. and Mundy, J. (2006) *Arabidopsis* MAP kinase 4 regulates salicylic acid- and jasmonic acid/ethylene-dependent responses via EDS1 and PAD4, *Plant Journal*.

Brukner, I., Sanchez, R., Suck, D. and Pongor, S. (1995) Sequence-dependent bending propensity of DNA as revealed by DNase I: parameters for trinucleotides, *EMBO Journal*, **14**, 1812-1818.

Buchanan-Wollaston, V., Page, T., Harrison, E., Breeze, E., Lim, P.O., Nam, H.G., Lin, J.F., Wu, S.H., Swidzinski, J., Ishizaki, K. and Leaver, C.J. (2005) Comparative transcriptome analysis reveals significant differences in gene expression and signalling pathways between developmental and dark/starvation-induced senescence in *Arabidopsis*, *Plant Journal*, **42**, 567-585.

Bulik, D.A., Olczak, M., Lucero, H.A., Osmond, B.C., Robbins, P.W. and Specht, C.A. (2003) Chitin synthesis in Saccharomyces cerevisiae in response to supplementation of growth medium with glucosamine and cell wall stress, *Eukaryot Cell*, **2**, 886-900.

Bulyk, M.L., Gentalen, E., Lockhart, D.J. and Church, G.M. (1999) Quantifying DNA-protein interactions by double-stranded DNA arrays, *Nature Biotechnology*, **17**, 573-577.

Baar, C., Eppinger, M., Raddatz, G.*, et al.* (2003) Complete genome sequence and analysis of Wolinella succinogenes, *Proceedings of the National Academy of Sciences of the United States of America*, **100**, 11690-11695.

Cao, H., Bowling, S.A., Gordon, A.S. and Dong, X. (1994) Characterization of an *Arabidopsis* Mutant That Is Nonresponsive to Inducers of Systemic Acquired Resistance, *Plant Cell*, **6**, 1583-1592.

Caprioli, A., Morabito, S., Brugere, H. and Oswald, E. (2005) Enterohaemorrhagic *Escherichia coli*: emerging issues on virulence and modes of transmission, *Veterinary Research*, **36**, 289-311.

Carbone, A., Kepes, F. and Zinovyev, A. (2005) Codon bias signatures, organization of microorganisms in codon space, and lifestyle, *Molecular Biology and Evolution*, **22**, 547-561.

Carbone, A., Zinovyev, A. and Kepes, F. (2003) Codon adaptation index as a measure of dominating codon bias, *Bioinformatics*, **19**, 2005-2015.

Cario, G., Stanulla, M., Fine, B.M., Teuffel, O., Neuhoff, N.V., Schrauder, A., Flohr, T., Schafer, B.W., Bartram, C.R., Welte, K., Schlegelberger, B. and Schrappe, M. (2005) Distinct gene expression profiles determine molecular treatment response in childhood acute lymphoblastic leukemia, *Blood*, **105**, 821-826.

Carlson, M.R., Zhang, B., Fang, Z., Mischel, P.S., Horvath, S. and Nelson, S.F. (2006) Gene connectivity, function, and sequence conservation: predictions from modular yeast co-expression networks, *BMC Genomics*, **7**, 40.

Carvalho, B., Ouwerkerk, E., Meijer, G.A. and Ylstra, B. (2004) High resolution microarray comparative genomic hybridisation analysis using spotted oligonucleotides, *Journal of Clinical Pathology*, **57**, 644-646.

Chen, D., Belmont, A.S. and Huang, S. (2004) Upstream binding factor association induces large-scale chromatin decondensation, *Proceedings of the National Academy of Sciences of the United States of America*.

Conover, W.J. (1971) *Practical nonparametric statistics*. John Wiley & Sons, New York.

Conradsen, K. (ed) (2002) *Diskriminantanalyse*. IMM, Lyngby.

Cope, L.M., Irizarry, R.A., Jaffee, H.A., Wu, Z. and Speed, T.P. (2004) A benchmark for Affymetrix GeneChip expression measures, *Bioinformatics*, **20**, 323-331.

Cortes, C. and V., V. (1995) Support-vector network, *Machine Learning*, **20**, 1-25.

Covert, M.W., Knight, E.M., Reed, J.L., Herrgard, M.J. and Palsson, B.O. (2004) Integrating high-throughput and computational data elucidates bacterial networks, *Nature*, **429**, 92-96.

Craigon, D.J., James, N., Okyere, J., Higgins, J., Jotham, J. and May, S. (2004) NASCArrays: a repository for microarray data generated by NASC's transcriptomics service, *Nucleic Acids Research*, **32**, D575-577.

Cutler, D.J., Zwick, M.E., Carrasquillo, M.M., Yohn, C.T., Tobin, K.P., Kashuk, C., Mathews, D.J., Shah, N.A., Eichler, E.E., Warrington, J.A. and Chakravarti, A. (2001) High-throughput variation detection and genotyping using microarrays, *Genome Research*, **11**, 1913-1925.

Daruwala, R.S., Rudra, A., Ostrer, H., Lucito, R., Wigler, M. and Mishra, B. (2004) A versatile statistical analysis algorithm to detect genome copy number variation, *Proc Natl Acad Sci U S A*, **101**, 16292-16297.

de Lichtenberg, U., Jensen, L.J., Fausboll, A., Jensen, T.S., Bork, P. and Brunak, S. (2005) Comparison of computational methods for the identification of cell cycle-regulated genes, *Bioinformatics*, **21**, 1164-1171.

DeRisi, J., Penland, L., Brown, P.O., Bittner, M.L., Meltzer, P.S., Ray, M., Chen, Y., Su, Y.A., Trent, J.M., Bulyk, M.L., Gentalen, E., Lockhart, D.J. and Church, G.M. (1996) Use of a cDNA microarray to analyse gene expression patterns in human cancer
Quantifying DNA-protein interactions by double-stranded DNA arrays, *Nature Genetics*, **14**, 457-460.

Dlakic, M., Ussery, D. and Brunak, S. (2004) DNA bendability and nucleosome positioning in transcriptional regulation. In Ohyama, T. (ed), *DNA Conformation in Transcription*. Landes Bioscience.

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, *Journal of Bacteriology*, **185**, 1831-1840.

Dobrindt, U., Blum-Oehler, G., Nagy, G., Schneider, G., Johann, A., Gottschalk, G. and Hacker, J. (2002) Genetic structure and distribution of four pathogenicity islands (PAI I(536) to PAI IV(536)) of uropathogenic *Escherichia coli* strain 536, *Infection and Immunity*, **70**, 6365-6372.

Dorman, C.J. (1991) DNA supercoiling and environmental regulation of gene expression in pathogenic bacteria, *Infection and Immunity*, **59**, 745-749.

Draghici, S. and Krawetz, S.A. (2003) Global functional profiling of Gene Expression Data. In D.P., B., Dubitzky, W. and Granzow, M. (eds), *A Practical Approach to Microarray Data Analysis*. Kluwer Academic Publishers.

Droillard, M.J., Boudsocq, M., Barbier-Brygoo, H. and Lauriere, C. (2004) Involvement of MPK4 in osmotic stress response pathways in cell suspensions and plantlets of *Arabidopsis thaliana*: activation by hypoosmolarity and negative role in hyperosmolarity tolerance, *FEBS Letters*, **574**, 42-48.

Dudoit, S. and Fridlyand, J. (2003a) Classification in microarray experiments. In Speed, T.P. (ed), *Statistical Analysis of Gene Expression Microarray Data*. Interdisciplinary Statistics, CRC Press.

Dudoit, S. and Fridlyand, J. (2003b) Introduction to Classification in Microarrray Experiments. In D.P., B., Dubitzky, W. and Granzow, M. (eds), *A Practical Approach to Microarray Data Analysis*. Kluwer Academic Publishers.

Dudoit, S., Fridlyand, J. and Speed, T.P. (2002a) Comparison of discrimination methods for the classification of tumors using gene expression data, *Journal of the American Statistical Association*, **97**, 77-87.

Dudoit, S., van der Laan, M.J. and Birkner, M.D. (2004) Multiple testing procedures for controlling tail probability error rates, *Division of Biostatistics, University of California, Berkeley*, **Technical Report 166**.

Dudoit, S., Yang, Y.H., Callow, M.J. and Speed, T.P. (2002b) Statistical methods for identifying genes with differential expression in replicated cDNA microarray experiments, *Statistica Sinica*, 111-139.

Dworkin, J. and Losick, R. (2002) Does RNA polymerase help drive chromosome segregation in bacteria? *Proceedings of the National Academy of Sciences of the United States of America*, **99**, 14089-14094.

Dyrskjot, L., Thykjaer, T., Kruhoffer, M., Jensen, J.L., Marcussen, N., Hamilton-Dutoit, S., Wolf, H. and Orntoft, T.F. (2003) Identifying distinct classes of bladder carcinoma using microarrays, *Nature Genetics*, **33**, 90-96.

Engler, D.A., Mohapatra, G., Louis, D.N. and Betensky, R.A. (2006) A pseudolikelihood approach for simultaneous analysis of array comparative genomic hybridizations, *Biostatistics*, **7**, 399-421.

Fisher, R.A. (1922) On the interpretation of $\chi 2$ from contingency tables, and the calculation of P, *Journal of the Royal Statistical Society*, **85**, 87-94.

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, *Proceedings of the National Academy of Sciences of the United States of America*, **98**, 8821-8826.

Foerstner, K.U., von Mering, C., Hooper, S.D. and Bork, P. (2005) Environments shape the nucleotide composition of genomes, *EMBO Rep*, **6**, 1208-1213.

Fridlyand, J., Snijders, A.M., Pinkel, D., Albertson, D.G. and Jain, A.N.A.N. (2004) Hidden Markov models approach to the analysis of array CGH data, *Journal of Multivariate Analysis*, **90**, 132.

Fukiya, S., Mizoguchi, H., Tobe, T. and Mori, H. (2004) Extensive genomic diversity in pathogenic *Escherichia coli* and *Shigella* Strains revealed by comparative genomic hybridization microarray, *Journal of Bacteriology*, **186**, 3911-3921.

Gaffney, T., Friedrich, L., Vernooij, B., Negrotto, D., Nye, G., Uknes, S., Ward, E., Kessmann, H. and Ryals, J. (1993) Requirement of salicylic acid for the induction of systemic acquired resistance, *Science and Justice*, 754–756.

Gagne, S.E., Jensen, R., Polvi, A., Da Costa, M., Ginzinger, D., Efird, J.T., Holly, E.A., Darragh, T. and Palefsky, J.M. (2005) High-resolution analysis of genomic alterations and human papillomavirus

integration in anal intraepithelial neoplasia, *Journal of Acquired Immune Deficiency Syndromes*, **40**, 182-189.

Gamage, S.D., Patton, A.K., Hanson, J.F. and Weiss, A.A. (2004) Diversity and host range of Shiga toxin-encoding phage, *Infection and Immunity*, **72**, 7131-7139.

Gasch, A.P., Spellman, P.T., Kao, C.M., Carmel-Harel, O., Eisen, M.B., Storz, G., Botstein, D. and Brown, P.O. (2000) Genomic expression programs in the response of yeast cells to environmental changes, *Molecular Biology of the Cell*, **11**, 4241-4257.

Ghaemmaghami, S., Huh, W.K., Bower, K., Howson, R.W., Belle, A., Dephoure, N., O'Shea, E.K. and Weissman, J.S. (2003) Global analysis of protein expression in yeast, *Nature*, **425**, 737-741.

Gilbert, N., Boyle, S., Fiegler, H., Woodfine, K., Carter, N.P. and Bickmore, W.A. (2004) Chromatin architecture of the human genome: gene-rich domains are enriched in open chromatin fibers, *Cell*, **118**, 555-566.

Giovannoni, S.J., Tripp, H.J., Givan, S., Podar, M., Vergin, K.L., Baptista, D., Bibbs, L., Eads, J., Richardson, T.H., Noordewier, M., Rappe, M.S., Short, J.M., Carrington, J.C. and Mathur, E.J. (2005) Genome streamlining in a cosmopolitan oceanic bacterium, *Science*, **309**, 1242-1245.

Golub, T.R., Slonim, D.K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J.P., Coller, H., Loh, M.L., Downing, J.R., Caligiuri, M.A., Bloomfield, C.D. and Lander, E.S. (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring, *Science*, **286**, 531-537.

Gowrishankar, J. and Harinarayanan, R. (2004) Why is transcription coupled to translation in bacteria? *Molecular Microbiology*, **54**, 598-603.

Gribskov, M., Devereux, J. and Burgess, R.R. (1984) The codon preference plot: graphic analysis of protein coding sequences and prediction of gene expression, *Nucleic Acids Research*, **12**, 539-549.

Grozdanov, L., Raasch, C., Schulze, J., Sonnenborn, U., Gottschalk, G., Hacker, J. and Dobrindt, U. (2004) Analysis of the genome structure of the nonpathogenic probiotic *Escherichia coli* strain Nissle 1917, *Journal of Bacteriology*, **186**, 5432-5441.

Gustafsson, G., Schmiegelow, K., Forestier, E., Clausen, N., Glomstein, A., Jonmundsson, G., Mellander, L., Makipernaa, A., Nygaard, R. and Saarinen-Pihkala, U.M. (2000) Improving outcome through two decades in childhood ALL in the Nordic countries: the impact of high-dose methotrexate in the reduction of CNS irradiation. Nordic Society of Pediatric Haematology and Oncology (NOPHO), *Leukemia*, **14**, 2267-2275.

Gygi, S.P., Rochon, Y., Franza, B.R. and Aebersold, R. (1999) Correlation between protein and mRNA abundance in yeast, *Molecular and Cellular Biology*, **19**, 1720-1730.

Hallin, P.F. and Ussery, D.W. (2004) CBS Genome Atlas Database: a dynamic storage for bioinformatic results and sequence data, *Bioinformatics*, **20**, 3682-3686.

Harr, B. and Schlotterer, C. (2006) Comparison of algorithms for the analysis of Affymetrix microarray data as evaluated by co-expression of genes in known operons, *Nucleic Acids Research*, **34**, e8.

Hartigan, J.A. and Wong, M.A. (1979) A K-means clustering algorithm., *Applied Statistics*, 100-108.

Hatfield, G.W. and Benham, C.J. (2002) DNA topology-mediated control of global gene expression in *Escherichia coli*, *Annual Review of Genetics*, **36**, 175-203.

Hayashi, K., Morooka, N., Yamamoto, Y., Fujita, K., Isono, K., Choi, S., Ohtsubo, E., Baba, T., Wanner, B.L., Mori, H. and Horiuchi, T. (2006) Highly accurate genome sequences of *Escherichia coli* K-12 strains MG1655 and W3110, *Molecular Systems Biology*, **Epub Feb 21**.

Hayashi, T., Makino, K., Ohnishi, M.*, et al.* (2001) Complete genome sequence of enterohemorrhagic *Escherichia coli* O157:H7 and genomic comparison with a laboratory strain K-12, *DNA Research*, **8**, 11-22.

Helmann, J.D., Wu, M.F., Gaballa, A., Kobel, P.A., Morshedi, M.M., Fawcett, P. and Paddon, C. (2003) The global transcriptional response of Bacillus subtilis to peroxide stress is coordinated by three transcription factors, *Journal of Bacteriology*, **185**, 243-253.

Hilson, P., Allemeersch, J., Altmann, T.*, et al.* (2004) Versatile gene-specific sequence tags for Arabidopsis functional genomics: transcript profiling and reverse genetics applications, *Genome Research*, **14**, 2176-2189.

Hjalgrim, L.L., Rostgaard, K., Schmiegelow, K., Soderhall, S., Kolmannskog, S., Vettenranta, K., Kristinsson, J., Clausen, N., Melbye, M., Hjalgrim, H. and Gustafsson, G. (2003) Age- and sex-specific incidence of childhood leukemia by immunophenotype in the Nordic countries, *Journal of the National Cancer Institute*, **95**, 1539-1544.

Hodgson, G., Hager, J.H., Volik, S., Hariono, S., Wernick, M., Moore, D., Nowak, N., Albertson, D.G., Pinkel, D., Collins, C., Hanahan, D. and Gray, J.W. (2001) Genome scanning with array CGH delineates regional alterations in mouse islet carcinomas, *Nature Genetics*, **29**, 459-464.

Hoheisel, J.D. (2006) Microarray technology: beyond transcript profiling and genotype analysis, *Nat Rev Genet*, **7**, 200-210.

Holder, D., Raubertas, R.F., Pikounis, V.B., Svetnik, V. and Soper, K. (2001) Statistical Analysis of High Density Oligonucleotide Arrays: A Safer Approach, *Proceedings of the ASA Annual Meeting 2001*.

Hollander, M. and Wolfe, D.A. (1973) *Nonparametric statistical inference*. John Wiley & Sons, New York.

Horák, J., Brzobohatý, B. and Lexa, M. (2003) Molecular and Physiological Characterisation of an Insertion Mutant in the *ARR21* Putative Response Regulator Gene from *Arabidopsis thaliana*, *Plant biology*, **5**, 245-254.

Hsu, L., Self, S.G., Grove, D., Randolph, T., Wang, K., Delrow, J.J., Loo, L. and Porter, P. (2005) Denoising array-based comparative genomic hybridization data using wavelets, *Biostatistics*, **6**, 211-226.

Huang, D. and Pan, W. (2006) Incorporating biological knowledge into distance-based clustering analysis of microarray gene expression data, *Bioinformatics*, **22**, 1259-1268.

Hughes, T.R., Marton, M.J., Jones, A.R.*, et al.* (2000) Functional discovery via a compendium of expression profiles, *Cell*, **102**, 109-126.

Hupe, P., Stransky, N., Thiery, J.P., Radvanyi, F. and Barillot, E. (2004) Analysis of array CGH data: from signal ratio to gain and loss of DNA regions, *Bioinformatics*, **20**, 3413-3422.

Ihaka, R. and Gentleman, R. (1996) R: a language for data analysis and graphics, *Journal of Computational and Graphical Statistics*, **5**, 299-314.

Ikemura, T. (1981) Correlation between the abundance of *Escherichia coli* transfer RNAs and the occurrence of the respective codons in its protein genes: a proposal for a synonymous codon choice that is optimal for the *E. coli* translational system, *Journal of Molecular Biology*, **151**, 389-409.

International Human Genome Sequencing Consortium (2004) Finishing the euchromatic sequence of the human genome, *Nature*, **431**, 931-945.

Irizarry, R.A., Bolstad, B.M., Collin, F., Cope, L.M., Hobbs, B. and Speed, T.P. (2003a) Summaries of Affymetrix GeneChip probe level data, *Nucleic Acids Research*, **31**, e15.

Irizarry, R.A., Hobbs, B., Collin, F., Beazer-Barclay, Y.D., Antonellis, K.J., Scherf, U. and Speed, T.P. (2003b) Exploration, normalization, and summaries of high density oligonucleotide array probe level data, *Biostatistics*, **4**, 249-264.

Irizarry, R.A., Wu, Z. and Jaffee, H.A. (2006) Comparison of Affymetrix GeneChip expression measures, *Bioinformatics*, **22**, 789-794.

Jafari, P. and Azuaje, F. (2006) An assessment of recently published gene expression data analyses: Reporting experimental design and statistical factors, *BMC Med Inform Decis Mak*, **6**, 27.

James, C.E., Stanley, K.N., Allison, H.E., Flint, H.J., Stewart, C.S., Sharp, R.J., Saunders, J.R. and McCarthy, A.J. (2001) Lytic and lysogenic infection of diverse *Escherichia coli* and *Shigella* strains with a verocytotoxigenic bacteriophage, *Applied and Environmental Microbiology*, **67**, 4335-4337.

Jeong, K.S., Ahn, J. and Khodursky, A.B. (2004) Spatial patterns of transcriptional activity in the chromosome of *Escherichia coli*, *Genome Biology*, **5**, R86.

Karlin, S., Barnett, M.J., Campbell, A.M., Fisher, R.F. and Mrazek, J. (2003) Predicting gene expression levels from codon biases in alpha-proteobacterial genomes, *Proceedings of the National Academy of Sciences of the United States of America*, **100**, 7313-7318.

Kaufman, L. and Rousseeuw, P.J. (1990) *Finding Groups in Data: an Introduction to Cluster Analysis*. John Wiley & Sons.

Khojasteh, M., Lam, W.L., Ward, R.K. and MacAulay, C. (2005) A stepwise framework for the normalization of array CGH data, *BMC Bioinformatics*, **6**, 274.

Kiba, T., Naitou, T., Koizumi, N., Yamashino, T., Sakakibara, H. and Mizuno, T. (2005) Combinatorial microarray analysis revealing *arabidopsis* genes implicated in cytokinin responses through the His->Asp Phosphorelay circuitry, *Plant and Cell Physiology*, **46**, 339-355.

Kieber, J.J., Rothenberg, M., Roman, G., Feldmann, K.A. and Ecker, J.R. (1993) CTR1, a negative regulator of the ethylene response pathway in *Arabidopsis*, encodes a member of the raf family of protein kinases, *Cell*, **72**, 427-441.

Knudsen, S. (2002) *A Biologist's Guide to Analysis of DNA Microarray Data*. Wiley-Interscience, New York.

Koski, L.B., Gray, M.W., Lang, B.F. and Burger, G. (2005) AutoFACT: an automatic functional annotation and classification tool, *BMC Bioinformatics*, **6**, 151.

Kreil, D.P. and Ouzounis, C.A. (2001) Identification of thermophilic species by the amino acid compositions deduced from their genomes, *Nucleic Acids Research*, **29**, 1608-1615.

Krishnapuram, B., Hartemink, A.J., Carin, L. and Figueiredo, M.A. (2004) A bayesian approach to joint feature selection and classifier design, *IEEE Trans Pattern Anal Mach Intell*, **26**, 1105-1111.

Lai, W.R., Johnson, M.D., Kucherlapati, R. and Park, P.J. (2005) Comparative analysis of algorithms for identifying amplifications and deletions in array CGH data, *Bioinformatics*, **21**, 3763-3770.

Larsen, T.S. and Krogh, A. (2003) EasyGene--a prokaryotic gene finder that ranks ORFs by statistical significance, *BMC Bioinformatics*, **4**, 21.

Lazarus, R.M. (1999) Definition of sensitivity and specificity, *American Journal of Clinical Nutrition*, **69**, 158-.

Le Caignec, C., Spits, C., Sermon, K., De Rycke, M., Thienpont, B., Debrock, S., Staessen, C., Moreau, Y., Fryns, J.P., Van Steirteghem, A., Liebaers, I. and Vermeesch, J.R. (2006) Single-cell chromosomal imbalances detection by array CGH, *Nucleic Acids Research*, **34**, e68.

Leclerc, G.J., Tartera, C. and Metcalf, E.S. (1998) Environmental regulation of Salmonella typhi invasion-defective mutants, *Infection and Immunity*, **66**, 682-691.

Lee, H.K., Hsu, A.K., Sajdak, J., Qin, J. and Pavlidis, P. (2004) Coexpression analysis of human genes across many microarray data sets, *Genome Research*, **14**, 1085-1094.

Li, C. and Wong, W. (2001a) Model-based analysis of oligonucleotide arrays: model validation, design issues and standard error application, *Genome Biology*, **2**, 1-11.

Li, C. and Wong, W. (2001b) Model-based analysis of oligonucleotide arrays: model validation, design issues and standard error application, *Genome Biol*, **2**, RESEARCH0032.

Li, C. and Wong, W.H. (2001c) Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection, *Proceedings of the National Academy of Sciences of the United States of America*, **98**, 31-36.

Li, L. and Weinberg, C.R. (2003) Gene selection and sample classification using a genetic algorithm and k-nearest neighbor method. In Berrar, D.P., Dubitzky, W. and Granzow, M. (eds), *A Practical Approach to Microarray Data Analysis*. Kluwer Academic Publishers.

Liva, S., Hupe, P., Neuvial, P., Brito, I., Viara, E., La Rosa, P. and Barillot, E. (2006) CAPweb: a bioinformatics CGH array Analysis Platform, *Nucleic Acids Research*, **34**, W477-481.

Lucito, R., Healy, J., Alexander, J.*, et al.* (2003) Representational oligonucleotide microarray analysis: a high-resolution method to detect genome copy number variation, *Genome Research*, **13**, 2291-2305.

Marioni, J.C., Thorne, N.P. and Tavare, S. (2006) BioHMM: a heterogeneous hidden Markov model for segmenting array CGH data, *Bioinformatics*, **22**, 1144-1146.

Marmur, J. and Doty, P. (1961) Thermal renaturation of deoxyribonucleic acids, *Journal of Molecular Biology*, **3**, 585-594.

Maskos, U. and Southern, E.M. (1993) A novel method for the analysis of multiple sequence variants by hybridisation to oligonucleotides, *Nucleic Acids Research*, **21**, 2267-2268.

Matthews, B.W. (1975) Comparison of the predicted and observed secondary structure of T4 phage lysozyme, *Biochimica et Biophysica Acta*, **405**, 442-451.

Menten, B., Maas, N., Thienpont, B.*, et al.* (2006) Emerging patterns of cryptic chromosomal imbalance in patients with idiopathic mental retardation and multiple congenital anomalies: a new series of 140 patients and review of published reports, *Journal of Medical Genetics*, **43**, 625-633.

Menten, B., Pattyn, F., De Preter, K.*, et al.* (2005) arrayCGHbase: an analysis platform for comparative genomic hybridization microarrays, *BMC Bioinformatics*, **6**, 124.

Methe, B.A., Webster, J., Nevin, K., Butler, J. and Lovley, D.R. (2005) DNA microarray analysis of nitrogen fixation and Fe(III) reduction in Geobacter sulfurreducens, *Applied and Environmental Microbiology*, **71**, 2530-2538.

Miao, E.A. and Miller, S.I. (1999) Bacteriophages in the evolution of pathogen-host interactions, *Proceedings of the National Academy of Sciences of the United States of America*, **96**, 9452-9454.

Mittler, R. (2006) Abiotic stress, the field environment and stress combination, *Trends Plant Sci*, **11**, 15-19.

Mizoguchi, T., Hayashida, N., Yamaguchi-Shinozaki, K., Kamada, H. and Shinozaki, K. (1993) ATMPKs: a gene family of plant MAP kinases in *Arabidopsis thaliana*, *FEBS Letters*, **336**, 440-444.

Mizuno, T. (2004) Plant response regulators implicated in signal transduction and circadian rhythm, *Current Opinion in Plant Biology*, **7**, 499-505.

Montgomery, D.C. (2000) *Design and analysis of experiments*. Wiley, New York.

Muniesa, M., de Simon, M., Prats, G., Ferrer, D., Panella, H. and Jofre, J. (2003) Shiga toxin 2-converting bacteriophages associated with clonal variability in *Escherichia coli* O157:H7 strains of human origin isolated from a single outbreak, *Infection and Immunity*, **71**, 4554-4562.

Myers, C.L., Chen, X. and Troyanskaya, O.G. (2005) Visualization-based discovery and analysis of genomic aberrations in microarray data, *BMC Bioinformatics*, **6**, 146.

Myers, C.L., Dunham, M.J., Kung, S.Y. and Troyanskaya, O.G. (2004) Accurate detection of aneuploidies in array CGH and gene expression microarray data, *Bioinformatics*, **20**, 3533-3543.

Nadeau, J.H., Balling, R., Barsh, G.*, et al.* (2001) Sequence interpretation. Functional annotation of mouse genome sequences, *Science*, **291**, 1251-1255.

Naef, F. and Magnasco, M.O. (2003) Solving the riddle of the bright mismatches: labeling and effective binding in oligonucleotide arrays, *Phys Rev E Stat Nonlin Soft Matter Phys*, **68**, 011906.

Nakao, K., Mehta, K.R., Fridlyand, J., Moore, D.H., Jain, A.N., Lafuente, A., Wiencke, J.W., Terdiman, J.P. and Waldman, F.M. (2004) High-resolution analysis of DNA copy number alterations in colorectal cancer by array-based comparative genomic hybridization, *Carcinogenesis*, **25**, 1345-1357.

Neuvial, P., Hupe, P., Brito, I., Liva, S., Manie, E., Brennetot, C., Radvanyi, F., Aurias, A. and Barillot, E. (2006) Spatial normalization of array-CGH data, *BMC Bioinformatics*, **7**, 264.

Nielsen, H.B., Wernersson, R. and Knudsen, S. (2003) Design of oligonucleotides for microarrays and perspectives for design of multi-transcriptome arrays, *Nucleic Acids Research*, **31**, 3491-3496.

Nyvold, C., Madsen, H.O., Ryder, L.P., Seyfarth, J., Svejgaard, A., Clausen, N., Wesenberg, F., Jonsson, O.G., Forestier, E. and Schmiegelow, K. (2002) Precise quantification of minimal residual disease at day 29 allows identification of children with acute lymphoblastic leukemia and an excellent outcome, *Blood*, **99**, 1253-1258.

O'Brien, A.D., Newland, J.W., Miller, S.F., Holmes, R.K., Smith, H.W. and Formal, S.B. (1984) Shiga-like toxin-converting phages from *Escherichia coli* strains that cause hemorrhagic colitis or infantile diarrhea, *Science*, **226**, 694-696.

Ochman, H. and Jones, I.B. (2000) Evolutionary dynamics of full genome content in *Escherichia coli*, *EMBO Journal*, **19**, 6637-6643.

Ogura, Y., Kurokawa, K., Ooka, T., Tashiro, K., Tobe, T., Ohnishi, M., Nakayama, K., Morimoto, T., Terajima, J., Watanabe, H., Kuhara, S. and Hayashi, T. (2006) Complexity of the Genomic Diversity in Enterohemorrhagic *Escherichia coli* O157 Revealed by the Combinational Use of the O157 Sakai OligoDNA Microarray and the Whole Genome PCR scanning, *DNA Research*, **13**, 3-14.

Olshen, A.B., Venkatraman, E.S., Lucito, R. and Wigler, M. (2004) Circular binary segmentation for the analysis of array-based DNA copy number data, *Biostatistics*, **5**, 557-572.

Paulsen, I.T., Seshadri, R., Nelson, K.E.*, et al.* (2002) The Brucella suis genome reveals fundamental similarities between animal and plant pathogens and symbionts, *Proceedings of the National Academy of Sciences of the United States of America*, **99**, 13148-13153.

Pavlicek, J.W., Oussatcheva, E.A., Sinden, R.R., Potaman, V.N., Sankey, O.F. and Lyubchenko, Y.L. (2004) Supercoiling-induced DNA bending, *Biochemistry*, **43**, 10664-10668.

Pedersen, A.G., Baldi, P., Chauvin, Y. and Brunak, S. (1998) DNA structure in human RNA polymerase II promoters, *Journal of Molecular Biology*, **281**, 663-673.

Pedersen, A.G., Jensen, L.J., Brunak, S., Staerfeldt, H.H. and Ussery, D.W. (2000) A DNA structural atlas for *Escherichia coli*, *Journal of Molecular Biology*, **299**, 907-930.

Perna, N.T., Plunkett, G., 3rd, Burland, V.*, et al.* (2001) Genome sequence of enterohaemorrhagic *Escherichia coli* O157:H7, *Nature*, **409**, 529-533.

Peter, B.J., Arsuaga, J., Breier, A.M., Khodursky, A.B., Brown, P.O. and Cozzarelli, N.R. (2004) Genomic transcriptional response to loss of chromosomal supercoiling in *Escherichia coli*, *Genome Biololy*, **5**, R87.

Petersen, M., Brodersen, P., Naested, H.*, et al.* (2000) *Arabidopsis* map kinase 4 negatively regulates systemic acquired resistance, *Cell*, **103**, 1111-1120.

Picard, F., Robin, S., Lavielle, M., Vaisse, C. and Daudin, J.J. (2005) A statistical approach for array CGH data analysis, *BMC Bioinformatics*, **6**, 27.

Pinkel, D. and Albertson, D.G. (2005) Array comparative genomic hybridization and its applications in cancer, *Nature Genetics*, **37 Suppl**, S11-17.

Pollack, J.R., Perou, C.M., Alizadeh, A.A., Eisen, M.B., Pergamenschikov, A., Williams, C.F., Jeffrey, S.S., Botstein, D. and Brown, P.O. (1999) Genome-wide analysis of DNA copy-number changes using cDNA microarrays, *Nature Genetics*, **23**, 41-46.

Pollack, J.R., Sorlie, T., Perou, C.M., Rees, C.A., Jeffrey, S.S., Lonning, P.E., Tibshirani, R., Botstein, D., Borresen-Dale, A.L. and Brown, P.O. (2002) Microarray analysis reveals a major direct role of DNA copy number alteration in the transcriptional program of human breast tumors, *Proceedings of the National Academy of Sciences of the United States of America*, **99**, 12963-12968.

Postow, L., Hardy, C.D., Arsuaga, J. and Cozzarelli, N.R. (2004) Topological domain structure of the *Escherichia coli* chromosome, *Genes and Development*, **18**, 1766-1779.

Poustka, A., Pohl, T., Barlow, D.P., Zehetner, G., Craig, A., Michiels, F., Ehrich, E., Frischauf, A.M. and Lehrach, H. (1986) Molecular approaches to mammalian genetics, *Cold Spring Harbor Symposia on Quantitative Biology*, **51 Pt 1**, 131-139.

Pradel, N., Livrelli, V., De Champs, C., Palcoux, J.B., Reynaud, A., Scheutz, F., Sirot, J., Joly, B. and Forestier, C. (2000) Prevalence and characterization of Shiga toxin-producing *Escherichia coli* isolated from cattle, food, and children during a one-year prospective study in France, *Journal of Clinical Microbiology*, **38**, 1023-1031.

Pui, C.H., Relling, M.V., Campana, D. and Evans, W.E. (2002) Childhood acute lymphoblastic leukemia, *Rev Clin Exp Hematol*, **6**, 161-180; discussion 200-162.

Raghava, G.P. and Han, J.H. (2005) Correlation and prediction of gene expression level from amino acid and dipeptide composition of its protein, *BMC Bioinformatics*, **6**, 59.

Rimsky, S. (2004) Structure of the histone-like protein H-NS and its role in regulation and genome superstructure, *Current Opinion in Microbiology*, **7**, 109-114.

Rocap, G., Larimer, F.W., Lamerdin, J.*, et al.* (2003) Genome divergence in two Prochlorococcus ecotypes reflects oceanic niche differentiation, *Nature*, **424**, 1042-1047.

Rocha, E.P. (2004) Codon usage bias from tRNA's point of view: redundancy, specialization, and efficient decoding for translation optimization, *Genome Research*, **14**, 2279-2286.

Ronald, J., Akey, J.M., Whittle, J., Smith, E.N., Yvert, G. and Kruglyak, L. (2005) Simultaneous genotyping, gene-expression measurement, and detection of allele-specific expression with oligonucleotide arrays, *Genome Research*, **15**, 284-291.

Ross, M.E., Zhou, X., Song, G., Shurtleff, S.A., Girtman, K., Williams, W.K., Liu, H.C., Mahfouz, R., Raimondi, S.C., Lenny, N., Patel, A. and Downing, J.R. (2003) Classification of pediatric acute lymphoblastic leukemia by gene expression profiling, *Blood*, **102**, 2951-2959.

Rousseeuw, P.J. (1987) Silhouettes: A graphical aid to the interpretation and validation of cluster analysis., *Journal of Computational and Applied Mathematics*, **20**, 53–65.

Saeed, A.I., Sharov, V., White, J.*, et al.* (2003) TM4: a free, open-source system for microarray data management and analysis, *Biotechniques*, **34**, 374-378.

Salunkhe, P., Topfer, T., Buer, J. and Tummler, B. (2005) Genome-wide transcriptional profiling of the steady-state response of Pseudomonas aeruginosa to hydrogen peroxide, *Journal of Bacteriology*, **187**, 2565-2572.

Sambrook, J., Fritsch, E.F. and Maniatis, T. (1989) *Molecular cloning: a laboratory manual*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, N.Y.

Satchwell, S.C., Drew, H.R. and Travers, A.A. (1986) Sequence periodicities in chicken nucleosome core DNA, *Journal of Molecular Biology*, **191**, 659-675.

Schadt, E.E., Li, C., Ellis, B. and Wong, W.H. (2001) Feature extraction and normalization algorithms for high-density oligonucleotide gene expression array data, *Journal of Cellular Biochemistry. Supplement*, **Suppl 37**, 120-125.

Schembri, M.A., Ussery, D.W., Workman, C., Hasman, H. and Klemm, P. (2002) DNA microarray analysis of fim mutations in *Escherichia coli*, *Mol Genet Genomics*, **267**, 721-729.

Schena, M., Shalon, D., Davis, R.W.*, et al.* (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray
Parallel human genome analysis: microarray-based expression monitoring of 1000 genes
Use of a cDNA microarray to analyse gene expression patterns in human cancer
Quantifying DNA-protein interactions by double-stranded DNA arrays, *Science*, **270**, 467-470.

Schena, M., Shalon, D., Heller, R.*, et al.* (1996) Parallel human genome analysis: microarray-based expression monitoring of 1000 genes
Use of a cDNA microarray to analyse gene expression patterns in human cancer
Quantifying DNA-protein interactions by double-stranded DNA arrays, *Proceedings of the National Academy of Sciences of the United States of America*, **93**, 10614-10619.

Scheutz, F., Cheasty, T., Woodward, D. and Smith, H.R. (2004) Designation of O174 and O175 to temporary O groups OX3 and OX7, and six new *E. coli* O groups that include Verocytotoxin-producing *E. coli* (VTEC): O176, O177, O178, O179, O180 and O181, *APMIS*, **112**, 569-584.

Schmidt, H., Bielaszewska, M. and Karch, H. (1999) Transduction of enteric *Escherichia coli* isolates with a derivative of Shiga toxin 2-encoding bacteriophage phi3538 isolated from *Escherichia coli* O157:H7, *Applied and Environmental Microbiology*, **65**, 3855-3861.

Schneider, G., Dobrindt, U., Bruggemann, H., Nagy, G., Janke, B., Blum-Oehler, G., Buchrieser, C., Gottschalk, G., Emody, L. and Hacker, J. (2004) The pathogenicity island-associated K15 capsule determinant exhibits a novel genetic structure and correlates with virulence in uropathogenic *Escherichia coli* strain 536, *Infection and Immunity*, **72**, 5993-6001.

Segal, E., Fondufe-Mittendorf, Y., Chen, L., Thastrom, A., Field, Y., Moore, I.K., Wang, J.P. and Widom, J. (2006) A genomic code for nucleosome positioning, *Nature*, **442**, 772-778.

Shaikh, N. and Tarr, P.I. (2003) *Escherichia coli* O157:H7 Shiga toxin-encoding bacteriophages: integrations, excisions, truncations, and evolutionary implications, *Journal of Bacteriology*, **185**, 3596-3605.

Sharp, P.M., Bailes, E., Grocock, R.J., Peden, J.F. and Sockett, R.E. (2005) Variation in the strength of selected codon usage bias among bacteria, *Nucleic Acids Research*, **33**, 1141-1153.

Sharp, P.M. and Li, W.H. (1986a) Codon usage in regulatory genes in *Escherichia coli* does not reflect selection for 'rare' codons, *Nucleic Acids Research*, **14**, 7737-7749.

Sharp, P.M. and Li, W.H. (1987) The codon Adaptation Index--a measure of directional synonymous codon usage bias, and its potential applications, *Nucleic Acids Research*, **15**, 1281-1295.

Sharp, P.M., Tuohy, T.M. and Mosurski, K.R. (1986b) Codon usage in yeast: cluster analysis clearly differentiates highly and lowly expressed genes, *Nucleic Acids Research*, **14**, 5125-5143.

Shedden, K., Chen, W., Kuick, R., Ghosh, D., Macdonald, J., Cho, K.R., Giordano, T.J., Gruber, S.B., Fearon, E.R., Taylor, J.M. and Hanash, S. (2005) Comparison of seven methods for producing Affymetrix expression scores based on False Discovery Rates in disease profiling data, *BMC Bioinformatics*, **6**, 26.

Sinden, R.R. (1994) *DNA Structure and Function*. Academic Press.

Sinden, R.R. and Pettijohn, D.E. (1981) Chromosomes in living *Escherichia coli* cells are segregated into domains of supercoiling, *Proceedings of the National Academy of Sciences of the United States of America*, **78**, 224-228.

Singh-Gasson, S., Green, R.D., Yue, Y., Nelson, C., Blattner, F., Sussman, M.R. and Cerrina, F. (1999) Maskless fabrication of light-directed oligonucleotide microarrays using a digital micromirror array, *Nature Biotechnology*, **17**, 974-978.

Skovgaard, M., Jensen, L.J., Friis, C., Stærfeldt, H.H., Worning, P., Brunak, S. and Ussery, D.W. (2002) The atlas visualisation of genome-wide information. In Wren, B. and Dorrell, N. (eds), *Methods in Microbiology*. Academic Press, London, 49-63.

Smith, G.R. (1981) DNA supercoiling: another level for regulating gene expression, *Cell*, **24**, 599-600.

Snijders, A.M., Nowak, N., Segraves, R.*, et al.* (2001) Assembly of microarrays for genome-wide measurement of DNA copy number, *Nature Genetics*, **29**, 263-264.

Snijders, A.M., Nowee, M.E., Fridlyand, J., Piek, J.M., Dorsman, J.C., Jain, A.N., Pinkel, D., van Diest, P.J., Verheijen, R.H. and Albertson, D.G. (2003) Genome-wide-array-based comparative genomic hybridization reveals genetic homogeneity and frequent copy number increases encompassing CCNE1 in fallopian tube carcinoma, *Oncogene*, **22**, 4281-4286.

Snijders, A.M., Schmidt, B.L., Fridlyand, J., Dekker, N., Pinkel, D., Jordan, R.C. and Albertson, D.G. (2005) Rare amplicons implicate frequent deregulation of cell fate specification pathways in oral squamous cell carcinoma, *Oncogene*.

Snipen, L., Repsilber, D., Nyquist, L., Aakra, A., Ziegler, A. and Aastveit, A. (2006) Detection of divergent genes in microbial aCGH experiments, *BMC Bioinformatics*, **7**, 181.

Sousa, C., de Lorenzo, V. and Cebolla, A. (1997) Modulation of gene expression through chromosomal positioning in *Escherichia coli*, *Microbiology*, **143 (Pt 6)**, 2071-2078.

Southern, E., Mir, K. and Shchepinov, M. (1999) Molecular interactions on microarrays, *Nature Genetics*, **21**, 5-9.

Southern, E.M. (1975) Detection of specific sequences among DNA fragments separated by gel electrophoresis, *Journal of Molecular Biology*, **98**, 503-517.

Steck, T.R., Franco, R.J., Wang, J.Y. and Drlica, K. (1993) Topoisomerase mutations affect the relative abundance of many *Escherichia coli* proteins, *Molecular Microbiology*, **10**, 473-481.

Stephan, R. and Hoelzle, L.E. (2000a) Characterization of shiga toxin type 2 variant B-subunit in *Escherichia coli* strains from asymptomatic human carriers by PCR-RFLP, *Letters in Applied Microbiology*, **31**, 139-142.

Stephan, R., Ragettli, S. and Untermann, F. (2000b) Prevalence and characteristics of verotoxin-producing *Escherichia coli* (VTEC) in stool samples from asymptomatic human carriers working in the meat processing industry in Switzerland, *Journal of Applied Microbiology*, **88**, 335-341.

Stintzi, A. and Whitworth, L. (2003) Investigation of the Campylobacter jejuni Cold Shock response by global gene expression analysis, *Journal of Genome Science and Technology*, **2**, 18-27.

Saal, L.H., Troein, C., Vallon-Christersson, J., Gruvberger, S., Borg, A. and Peterson, C. (2002) BioArray Software Environment (BASE): a platform for comprehensive management and analysis of microarray data, *Genome Biology*, **3**, SOFTWARE0003.

Tajima, Y., Imamura, A., Kiba, T., Amano, Y., Yamashino, T. and Mizuno, T. (2004) Comparative studies on the type-B response regulators revealing their distinctive properties in the His-to-Asp phosphorelay signal transduction of *Arabidopsis thaliana*, *Plant and Cell Physiology*, **45**, 28-39.

Tavazoie, S., Hughes, J.D., Campbell, M.J., Cho, R.J. and Church, G.M. (1999) Systematic determination of genetic network architecture, *Nature Genetics*, **22**, 281-285.

Teige, M., Scheikl, E., Eulgem, T., Doczi, R., Ichimura, K., Shinozaki, K., Dangl, J.L. and Hirt, H. (2004) The MKK2 pathway mediates cold and salt stress signaling in *Arabidopsis*, *Molecular Cell*, **15**, 141-152.

Tekaia, F., Yeramian, E. and Dujon, B. (2002) Amino acid composition of genomes, lifestyles of organisms, and evolutionary trends: a global picture with correspondence analysis, *Gene*, **297**, 51-60.

Tibshirani, R., Hastie, T., Narasimhan, B. and Chu, G. (2002) Diagnosis of multiple cancer types by shrunken centroids of gene expression, *Proceedings of the National Academy of Sciences of the United States of America*, **99**, 6567-6572.

Tjaden, B., Haynor, D.R., Stolyar, S., Rosenow, C. and Kolker, E. (2002) Identifying operons and untranslated regions of transcripts using *Escherichia coli* RNA expression analysis, *Bioinformatics*, **18 Suppl 1**, S337-344.

Toth, I., Schmidt, H., Dow, M., Malik, A., Oswald, E. and Nagy, B. (2003) Transduction of porcine enteropathogenic *Escherichia coli* with a derivative of a shiga toxin 2-encoding bacteriophage in a porcine ligated ileal loop system, *Applied and Environmental Microbiology*, **69**, 7242-7247.

Ussery, D., Larsen, T.S., Wilkes, K.T., Friis, C., Worning, P., Krogh, A. and Brunak, S. (2001) Genome organisation and chromatin structure in *Escherichia coli*, *Biochimie*, **83**, 201-212.

Ussery, D.W., Hallin, P.F., Lagesen, K. and Coenye, T. (2004) Genome update: rRNAs in sequenced microbial genomes, *Microbiology*, **150**, 1113-1115.

van der Laan, M.J., Pollard, K.S. and Bryan, J. (2003) A New Partitioning Around Medoids Algorithm, *Journal of Statistical Computation and Simulation*, **73**, 575--584.

Van Driessche, N., Demsar, J., Booth, E.O., Hill, P., Juvan, P., Zupan, B., Kuspa, A. and Shaulsky, G. (2005) Epistasis analysis with global transcriptional phenotypes, *Nature Genetics*, **37**, 471-477.

Veltman, J.A., Fridlyand, J., Pejavar, S., Olshen, A.B., Korkola, J.E., DeVries, S., Carroll, P., Kuo, W.L., Pinkel, D., Albertson, D., Cordon-Cardo, C., Jain, A.N. and Waldman, F.M. (2003) Array-based comparative genomic hybridization for genome-wide screening of DNA copy number in bladder tumors, *Cancer Research*, **63**, 2872-2880.

Versteeg, R., van Schaik, B.D., van Batenburg, M.F., Roos, M., Monajemi, R., Caron, H., Bussemaker, H.J. and van Kampen, A.H. (2003) The human transcriptome map reveals extremes in gene density, intron length, GC content, and repeat pattern for domains of highly and weakly expressed genes, *Genome Research*, **13**, 1998-2004.

Vissers, L.E., de Vries, B.B., Osoegawa, K.*, et al.* (2003) Array-based comparative genomic hybridization for the genomewide detection of submicroscopic chromosomal abnormalities, *American Journal of Human Genetics*, **73**, 1261-1270.

Wagner, P.L., Acheson, D.W. and Waldor, M.K. (1999) Isogenic lysogens of diverse shiga toxin 2-encoding bacteriophages produce markedly different amounts of shiga toxin, *Infection and Immunity*, **67**, 6710-6714.

Wagner, P.L., Neely, M.N., Zhang, X., Acheson, D.W., Waldor, M.K. and Friedman, D.I. (2001) Role for a phage promoter in Shiga toxin 2 expression from a pathogenic *Escherichia coli* strain, *Journal of Bacteriology*, **183**, 2081-2085.

Wang, P., Kim, Y., Pollack, J., Narasimhan, B. and Tibshirani, R. (2005) A method for calling gains and losses in array CGH data, *Biostatistics*, **6**, 45-58.

Welch, R.A., Burland, V., Plunkett, G., 3rd*, et al.* (2002) Extensive mosaic structure revealed by the complete genome sequence of uropathogenic *Escherichia coli*, *Proceedings of the National Academy of Sciences of the United States of America*, **99**, 17020-17024.

Westfall, P.H. and Young, S.S. (1993) *Resampling-based Multiple testing: examples and methods for p-value adjustment*. Wiley, New York.

Willenbrock, H. and Fridlyand, J. (2005) A comparison study: applying segmentation to array CGH data for downstream analyses, *Bioinformatics*, **21**, 4084-4091.

Willenbrock, H. and Ussery, D.W. (2004) Chromatin architecture and gene expression in *Escherichia coli*, *Genome Biology*, **5**, 252.

Winterberg, K.M., Luecke, J., Bruegl, A.S. and Reznikoff, W.S. (2005) Phenotypic screening of *Escherichia coli* K-12 Tn5 insertion libraries, using whole-genome oligonucleotide microarrays, *Applied and Environmental Microbiology*, **71**, 451-459.

Worcel, A. and Burgi, E. (1972) On the structure of the folded chromosome of *Escherichia coli*, *Journal of Molecular Biology*, **71**, 127-147.

Workman, C., Jensen, L.J., Jarmer, H., Berka, R., Gautier, L., Nielser, H.B., Saxild, H.H., Nielsen, C., Brunak, S. and Knudsen, S. (2002) A new non-linear normalization method for reducing variability in DNA microarray experiments, *Genome Biology*, **3**, research0048.

Wright, F. (1990) The 'effective number of codons' used in a gene, *Gene*, **87**, 23-29.

Wu, L.F., Hughes, T.R., Davierwala, A.P., Robinson, M.D., Stoughton, R. and Altschuler, S.J. (2002) Large-scale prediction of *Saccharomyces cerevisiae* gene function using overlapping transcriptional clusters, *Nature Genetics*, **31**, 255-265.

Wu, Z., Irizarry, R.A., Gentleman, R., Murillo, F.M. and Spencer, F. (2004) A Model Based Background Adjustment for Oligonucleotide Expression Arrays, *John Hoostatistics Working Paperspkins Univerisyt, Department of B*.

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 Research*, **30**, e15.

Yeoh, E.J., Ross, M.E., Shurtleff, S.A.*, et al.* (2002) Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling, *Cancer Cell*, **1**, 133-143.

Young, A., Whitehouse, N., Cho, J. and Shaw, C. (2005) OntologyTraverser: an R package for GO analysis, *Bioinformatics*, **21**, 275-276.

Zhang, W., Morris, Q.D., Chang, R.*, et al.* (2004) The functional landscape of mouse gene expression, *J Biol*, **3**, 21.

Zhao, X., Li, C., Paez, J.G., Chin, K., Janne, P.A., Chen, T.H., Girard, L., Minna, J., Christiani, D., Leo, C., Gray, J.W., Sellers, W.R. and Meyerson, M. (2004) An integrated view of copy number and

allelic alterations in the cancer genome using single nucleotide polymorphism arrays, *Cancer Research*, **64**, 3060-3071.