Design, analysis and applications of custom high-density oligonucleotide microarrays

Gard Olav Sundby Thomassen

Thesis for the degree of Philosophiae Doctor (PhD)

Centre for Molecular Biology and Neuroscience (CMBN)
Institute of Medical Microbiology, Department of Molecular Biology
Oslo University Hospital
University of Oslo, Norway

2010

Series of dissertations submitted to the
Faculty of Medicine, University of Oslo
No. 965

ISBN 978-82-8072-495-3

Cover: Inger Sandved Anfinsen.
Printed in Norway: AiT e-dit AS.

Produced in co-operation with Unipub.
The thesis is produced by Unipub merely in connection with the
thesis defence. Kindly direct all inquiries regarding the thesis to the copyright
holder or the unit which grants the doctorate.

Table of Contents

Acknowledgements

The work included in this thesis was carried out at the Bioinformatics group at the Centre for Molecular Biology and Neuroscience (CMBN) at Rikshospitalet, Oslo University Hospital from 2004 to 2009, and at the Department of Cancer Prevention at Rikshospitalet, Oslo University Hospital from 2007 to 2009.

I would like to thank my supervisors and co-workers at CMBN and at the Department of Cancer Prevention.

First of all I would like to thank Assoc. Prof. Torbjørn Rognes for being my supervisor from the start of my Cand. scient. degree all the way through my Ph.D. thesis. I really do not know how he could bear with me for that many years!

I would also like to thank Rolf Skotheim for bringing me into the Fusion Gene project, and everyone at the Department of Cancer Prevention for their warm welcome.

My co-workers at the Bioinformatics group at CMBN have supported me for many years, and I would really like to express my thanks to everyone in the group. Specially Alexander Rowe for his great support while working on my data. I would also like to thank the microbiologists at CMBN. I would specially like to mention Professor Magnar Bjørås, Ragnhild Weel-Sneve, Knut-Ivan Kristiansen and James Booth for their great support with various aspects of the bacterial microarray experiments.

Additionally, I would like to thank Professor Eivind Hovig for giving invaluable advice, Sigve Nakken for being the best office-buddy in the world and Thomas Hansen for priceless Apple support and coffee supplies.

Finally I would like to thank my family, friends, co-workers and co-authors for help and support.

Oslo, May 2010
Gard Thomassen

Abbreviations

actD	Actinomycin D
DNA	DeoxyriboNucleic Acid
DMD	Digital Micromirror Device
DLP	Digital Light Processor
DSB	Double Stranded Break
gcRMA	GeneChip Multi-array Average (alias gc-adjusted Multi-array Average)
GLM	Generalised Linear Model
GSD	Global Standard Deviation
HDONA	High-density OligoNucleotide Arrays
HMM	Hidden Markov Model
ncRNA	non-coding RiboNucleic Acid
MM	MisMatch (probe)
MCTPP	Minimal-Cost Tiling Path Problem
PM	Perfect Match (probe)
RNA	RiboNucleic Acid
RMA	Robust Multi-arrray Average
RT-qPCR	Reverse Transcriptase quantitative Polymerase Chain Reaction
VSN	Variance stabilisation normalization
UTR	Untranslated Region

List of papers

Paper I.

Gard O.S. Thomassen, Alexander D. Rowe, Karin Lagesen, Jessica M. Lindvall, and Torbjørn Rognes. **Custom design and analysis of high-density oligonucleotide bacterial tiling microarrays.** *PLoS One* 2009 Jun 17;4(6):e5943.

Paper II.

Gard O.S. Thomassen, James A. Booth, Alexander D. Rowe, Ragnhild Weel-Sneve, Karin Lagesen, Knut I. Kristiansen, Magnar Bjørås, Torbjørn Rognes, and Jessica M. Lindvall. **Transcriptome analysis of MNNG treated *Escherichia coli* reveals a widespread transcription of coding and non-coding RNA.** Manuscript submitted.

Paper III.

Gard O.S. Thomassen, Ragnhild Weel-Sneve, Alexander D. Rowe, James A. Booth, Jessica M. Lindvall, Karin Lagesen, Knut I. Kristiansen, Magnar Bjørås, and Torbjørn Rognes. **Tiling array analysis of UV treated *Escherichia coli* predicts novel differentially expressed small peptides.** Manuscript submitted.

Paper IV.

Rolf I. Skotheim, Gard O. S. Thomassen, Marthe Eken, Guro E. Lind, Francesca Micci, Franclim R. Ribeiro, Nuno Cerveira, Manuel R. Teixeira, Sverre Heim, Torbjørn Rognes, and Ragnhild A. Lothe. **A universal assay for detection of oncogenic fusion transcripts by oligo microarray analysis.** *Molecular Cancer* 2009, 8:5.

Chapter 1 Introduction

The merging of computer science and molecular biology has become increasingly tighter and important as the design and analysis of biological experiments gains complexity and produces increasing amounts of data. Computers and computer science were developed with the main goal of solving problems involving too much data or too many operations for man to finish in a reasonable time. On the other hand, molecular biology emerged as a field of study when tools and curiosity combined enabled man to venture beyond biological studies in the more classical terms of Mendel and Darwin. As novel tools are developed and further understanding of the complexity of molecular biology is gained, the need for general and specific high-throughput experiments and analysis tools has become clearer than ever. One specific area of high throughput experiments and analysis is the field of microarrays. All microarray studies involve an array design, an experimental design, a hybridisation of some biological component to the array and a subsequent analysis of the hybridisation results.

In this work I have unleashed the power of microarrays for observation of DNA events in bacteria and humans. This has been achieved by going all the way from array design to result interpretation, and clearly demonstrates the power contained in high-throughput methodology in the combined field of molecular biology and computer science.

Since the molecular biology and bioinformatics of this work are somewhat separated, an introduction to the different aspects is presented in two parts below. As a bioinformatician one tries to be a bridge between computer science and biology, and as a trained computer scientist my short introduction to molecular biology would be basic to a biologist, but important to the computer scientist. I further describe common algorithms, tools, microarray platforms and computer programs that are relevant to the studies presented in the papers.

1.1 Molecular biology

The study has focused on custom high-density oligonucleotide microarray design and analysis for detection of DNA repair systems and DNA damage. As the microarray application areas differ, the introductory aspects of these are separated for the sake of clarity.

1.1.1 DNA and RNA basics

”DNA makes RNA, RNA makes protein, and proteins make us.” - *Francis Crick*

All organisms, except viruses, are made up of one or more cells. To construct and to maintain a living organism a recipe is needed, and this recipe is stored in the genome. In single cell organisms (prokaryotes), like bacteria, the genome is one circular piece of DeoxyriboNucleic Acid (DNA), while multi-cellular organisms (eukaryotes) have chromosomal DNA organisation and mitochondrial DNA. Normal chromosomal DNA is made up of two complimentary strands formed in a double-helix structure, each strand is a linear variation over the four nucleotides (bases): adenine (A), thymine (T), cytosine (C) and guanine (G). The double-helix structure is maintained by base pairing between the two linear DNA strands (Figure 1).

RiboNucleic Acid (RNA) is a less stable single stranded version of DNA, where the base T has been changed to uracil (U), (Figure 1). An RNA molecule is generated by “copying” one strand of the DNA “nucleotide by nucleotide” in a process called transcription. Dependent upon its sequence type, the RNA can now serve as either a functional component in the cell or be further translated into a protein (Figure 1). The total RNA product within in a cell is called the transcriptome.

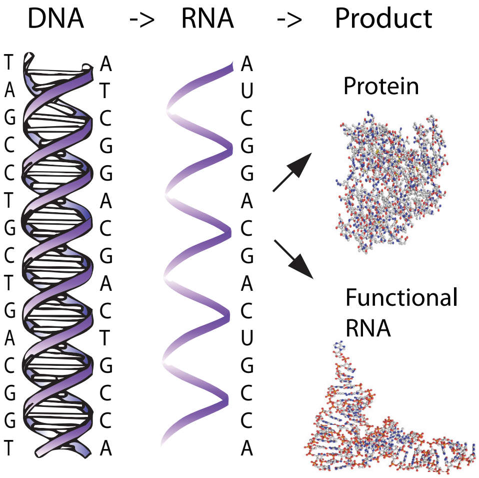
DNA -> RNA -> Product
T A G C C T G C T G A C G G T
A T C G G A C G A C T G C C A
A U C G G A C G A C U G C C A
Protein
Functional RNA

Figure 1. DNA, RNA, ncRNA and protein. The DNA is the storage facility for the hereditary material. During transcription RNA is created and subsequently either folded into a functional RNA or translated into amino acids and folded into a protein.

1.1.2 Genes

A gene is a piece of DNA which can be transcribed into some RNA product that is of use to the cell. In this thesis all DNA regions coding for some useful RNA product are referred to as genes, rather than limiting the description to those responsible for protein coding messenger RNAs (mRNAs). A gene is encoded on one of the DNA strands (coding strand), the opposite DNA strand is named the template strand. The section of an mRNA being translated into a protein is called the open reading frame (ORF), while the 5' start and the 3' end regions of mRNAs which are not part of the ORF are referred to as untranslated regions (UTR) (Figure 2).

In prokaryotes, a gene is organised as one continuous stretch of DNA, while in eukaryotes a gene is usually made up of at least two exons (coding DNA) separated by introns (non-coding DNA) (Figure 2). The introns are removed when the mRNA is assembled from the initial transcript, which includes both the exon and intron sequences. Hence, eukaryotes can compose several different versions (splice variants) of a single protein from a coding gene by including or excluding exons in the mRNA assembly before translating it into a protein.

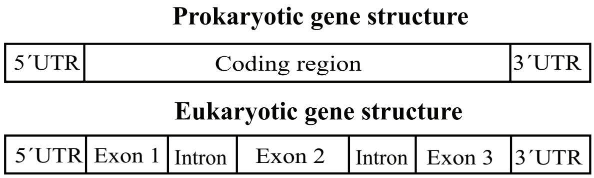

Figure 2. The structure of prokaryotic and eukaryotic genes. The prokaryotic genes are organised as continuous stretches, while the eukaryotic genes are usually made up of exons and introns.

Several genes (in prokaryotes) can be organised into operons; leading to a single transcript containing at least two protein coding sequences. Operons often consist of genes that are functionally closely related [1, 2]. One operon theory is that they are formed to ensure co-regulation [3], while another theory (the "selfish operon model") claim that genes form operons by horizontal gene transfer to prevent their removal from the genome [4].

1.1.3 Non-coding RNA

RNA molecules have for many years been viewed as the cornerstones of the protein synthesis (mRNA, transfer RNA (tRNA), ribosomal RNA (rRNA)). As more genomes were sequenced, the numbers of protein coding genes suggested that organism complexity could not be completely explained by the relatively limited number of protein coding genes. Already in 1993 it was shown that the *lin-4* gene in *Caenorhabditis elegans* was not coding for an mRNA, tRNA or rRNA, but for a functional RNA crucial to the development of the organism [5]. The detection of this functional RNA molecule was made in the "protein era", but today the revelation of *lin-4* and its function is perhaps viewed as more important than at the time of discovery. This novel RNA molecule fuelled the detection of a wide range of RNA molecules, for example; short-, micro-, and short-interfering –RNAs, generically named non-coding RNAs (ncRNAs) or small RNAs (sRNAs), as they do not code for proteins.

Several studies have tried to map the entire transcriptomes of bacteria [6, 7], yeast [8] or entire human chromosomes [9], and widespread transcription is detected in all cases. Novel transcripts are detected from intergenic regions and from the opposite strand of known coding regions. This abundance of novel transcripts is sometimes referred to as the "dark matter in the genome" since the transcripts await an explanation [10]. One possibility is that many of these transcripts are ncRNAs acting as a critical, but until now, hidden layer of gene regulation, with more complex organisms containing larger amounts of ncRNAs [11]. Other explanations propose that these results are mixtures of experimental artefacts [12], biological “artefacts”, true ncRNAs and transcripts that code for short peptides instead of ncRNAs [10].

Some RNAs have known functions; for example *lin-4* triggers a certain developmental stage, while the majority have unknown functions. One characteristic of these functional RNA molecules is that they are often short (< 100 nts) and they can behave in a regulatory fashion by binding to mRNAs with a complementary nucleotide sequence. Regulatory ncRNAs can regulate mRNA translation depending on the level of complementarity (ncRNAs have been reviewed and discussed in for instance Eddy [13], Szymanski *et al.* [14], Huttenhofer *et al.* [15], and Mattick & Makunin [16, 17]). Today many ncRNAs, from different species, have been detected with laboratory methods. See for instance miRBase [18-20] or Rfam [21] for all miRNAs and ncRNAs, and even more are *in silico* predicted [22-25]. As studies continue to unravel ncRNAs as important gene regulators, it follows that detection of ncRNAs and functional ncRNA studies are fundamental to a full understanding of functional genomics.

1.1.4 *Escherichia coli*

Escherichia coli is well known as a bacterium causing diarrhoea (e. g. strain 0157:H7). *E. coli* is a gram-negative bacterium and some of the harmless *E. coli* strains are part of the normal flora of the human gut. There are more than 60 sequenced strains of *E. coli* (NCBI 29th of June, 2009). The *E. coli* strain studied in this thesis is MG1655, which is non-pathogenic and has a circular genome of about 4.6 megabases [26]. *E. coli* is perhaps the best characterised organism in molecular biology, and was documented already in 1885 and in 1919 named after the German doctor Theodor Escherich who made the discovery.

The genome of *E. coli* MG1655 is annotated with 4131 protein coding genes and 172 non coding genes (NCBI, NC_000913, updated 27th of January, 2009) [21], and it was sequenced as early as in 1997 [26]. Although the *E. coli* genome is among the most thoroughly studied, the functions of the majority of the genes are unknown. In prokaryotes it is assumed that the non-coding strand does not hold any important coding information. Additionally, intergenic DNA (regions located between known genes) is by many considered to be “junk-DNA”. Today these non-coding regions are emerging as an interesting field of research, as experiments indicate widespread transcription from non-coding DNA in a variety of species [6, 8, 9], including *E. coli*.

Since 2001 several studies have been published on computational and/or experimental detection of *E. coli* ncRNAs [22-25, 27-29] (reviewed in [30]). The methods vary from whole-genome microarrays, via searches for conserved regions and structures, through searches for transcription signals. Currently (24th of March 2009) there are 57 annotated ncRNAs (not including tRNAs or rRNAs) in *E. coli* MG1655 (Rfam [21]), and more than 1000 ncRNA candidates from *in silico* studies [30]. The usual methods for verification of ncRNAs are northern blots and reverse transcriptase polymerase chain reaction (RT-PCR). All verified *E. coli* ncRNAs are located in DNA regions without previously known protein coding DNA on either strand (some of these have been mistakenly reported as short coding regions before), and some predicted ncRNAs reside inside annotated protein coding regions. Most known *E. coli* ncRNAs reside in intergenic regions with size [300-900] nts, and only rarely in longer regions, since those regions are dominated by repeats.

1.1.5 DNA repair

A remarkable feature of DNA is the way in which the structure, via the strict strand complementarity, has a redundancy which enables DNA damage repair. All organisms suffer from DNA-damage from for example chemical reagents or UV-irradiation. For example, about 18,000 purine residues are lost in every human cell every day because of hydrolysis. Although most DNA damage is non-lethal, all organisms are critically dependent upon DNA repair functionality. DNA repair is defined as "*a cellular response to DNA damage that results in the restoration of normal nucleotide sequence and DNA structure*" [31]. Hence, a DNA repair mechanism has to either reverse the damage or exchange the damaged part with a healthy part. The initial recognition of DNA repair was done in *E. coli* [32].

DNA repair mechanisms are often divided into three different groups [31]:

i) *Reversal of base damage.* Repair systems of this category reverse damage inflicted on single bases in the DNA.

ii) *Excision repair of damaged, mispaired, or incorrect bases*. Excision repair has several subgroups defined by the part of the DNA that is excised. The central function of these repair mechanisms cause damaged DNA base(s) to be cut away and replaced by a healthy part which is identical to the original.

iii) *Strand break repair.* The DNA strand can be broken either on one or both strands. Strand break repair has two subgroups; single strand break repair and double strand break repair.

DNA-damage can be inflicted by different sources, from chemical reagents to irradiation. Typical damage include loss of purines, damage inflicted to the bases, base-exchanges, single strand breaks and double stranded breaks of the DNA [31]. DNA damage mechanisms can cause mutations if the repair system is insufficient or the damage level too high, and although most mutations are benign some are known to cause diseases in humans like xeroderma pigmentosum, Cockayne's syndrome and different types of cancer (reviewed by Hoeijmakers [33]). Since the genome should be kept stable without introducing malign mutations, the DNA repair system has several backup strategies; if a primary repair pathway fails, some secondary system will hopefully repair the damage.

Not all repair processes are able to perfectly repair the damaged DNA, and instead might introduce errors (error prone repair) or might allow DNA replication to continue although there are errors present in the DNA (lesion bypass). Error prone / lesion bypass repair genes are described in SOS-response section (1.1.5.1) below.

DNA damage tolerance systems are mechanisms often combined with DNA repair. These are mechanisms enabling the cell to tolerate a certain level of DNA damage. It should be noted that favourable genetic changes are most often introduced by means that do not include the removal of DNA damage [31].

In multicellular organisms, a final answer to severe and unrecoverable damage is apoptosis *i.e.* programmed cell death. DNA damage and uncompleted repair can cause many diverse, but severe, diseases (including cancer). Therefore it is of utmost importance to fully understand all DNA repair mechanisms, and *E. coli* is perhaps the most widely used model organism for DNA repair research. A better understanding of the *E. coli* DNA repair systems, combined with the fact that DNA repair genes are remarkably well conserved between species [34], may make it easier to define therapeutic targets and design drugs for DNA damage related diseases in higher organisms.

An important DNA damage response in many organisms is the SOS response [32] (first formal paper, 1973), named after the international distress signal. The SOS response was first discovered in *E. coli* and is known to include more than 40 genes [35]. DNA damage induces the SOS response and the most important regulatory proteins involved are the LexA repressor and the RecA protein (Figure 3). The SOS response has been reviewed extensively [36-40] and is well understood.

Both *in vivo* and *in vitro* studies postulate that the induction signal for the SOS response is the interruption of normal replication, or the creation of regions of single stranded DNA (ssDNA) originating from replication attempts on damaged DNA templates [31]. The RecA protein binds to the ssDNA making RecA-ssDNA nucleoprotein filaments, and the coprotease activity of RecA is activated, enabling cleavage of LexA. The LexA repressor fuses with the nucleoprotein filaments causing cleavage of the Lex repressor and subsequently a rapid decrease in the amount LexA [41]. LexA functions as a repressor by binding to an operator sequence (SOS box) of the genes it represses. The decrease of LexA repressors induces expression of the normally repressed repair proteins at different times and levels. According to the time-point microarray gene expression study by Courcelle *et al.* [6] the different SOS repair systems are induced in an ordered fashion (commented in [37]). The first class of genes induced is nucleotide excision repair (NER) consisting of *uvrA, uvrB uvrD* genes and the endonuclease UvrC homologue *cho* which repairs damaged nucleotides in double stranded DNA. Secondly different homologous recombination repair functions with genes including *recA, recN, ruvA* and *ruvB* are induced. To allow time to complete the repair process the cell division inhibitor *sfiA* (alias *sulA*) is induced. If the genome is not fully repaired about 40 minutes after damage discovery, a final SOS response step is tried. This final repair process is carried out by the error prone DNA repair polymerase Pol V (encoded by the *umuC* and *umuD* genes). Pol V can repair double stranded DNA lesions, but Pol V might introduce errors to the genome during this process.

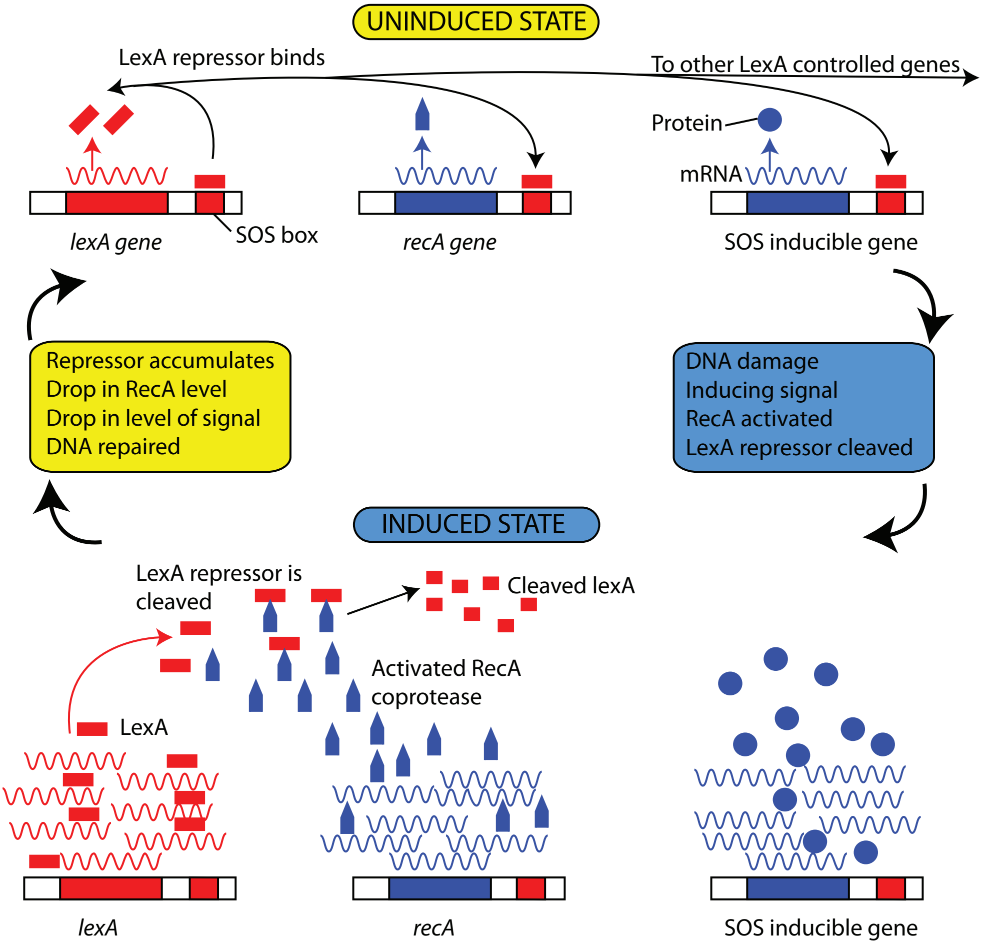

Figure 3. Induction of the LexA controlled SOS response genes. A schematic view of the regulation in the *lexA-recA* regulon. In the uninduced state the LexA repressor protein binds to the SOS box of the LexA regulated genes and represses these. DNA damage activates the coprotease activity of the RecA protein and the activated RecA cleaves the LexA repressor. In the induced state the de-repression of *recA* results in large amounts of RecA proteins. When the DNA damage induced signal decreases (due to DNA repair), the coprotease activity of RecA decreases, LexA repressor protein accumulates and the LexA controlled genes are again repressed by LexA. (Figure adapted from Friedberg *et al.* [42].)

The upstream region of SOS genes, which the LexA repressor binds to when residing its normal non-cleaved form, is named the SOS box [31]. These operator sequences are imperfect palindromes and the consensus sequence is described as 5´-TACTG(TA)$_5$CAGTA-3´ [43], and is usually located within -200 to +40 bases of putative translational start codons of LexA regulated genes [35]. When the *E. coli* genome sequence was published in 1997, whole-genome scanning for SOS-boxes could be applied to detect possible LexA binding sites and consequently possible SOS-response genes. A computational search by Fernándes de

Henestrosa *et al.* revealed 69 possible LexA regulated genes/operons, from which seven new LexA regulated genes were identified [35]. The findings of this *in silico* search also included all previously predicted LexA regulated genes from a previous study by Lewis *et al*. [44] from 1994. Of the 62 known SOS boxes most are found upstream of single genes or operons, but some are found between genes transcribed in opposite directions. Due to DNA strand complementarity and the palindromic form of the SOS-boxes the LexA binding sites may regulate 69 genes or operons in total [35].

Ultraviolet (UV) irradiation, mitomycin C and nalidixic acid are examples of well established SOS response inducers in *E. coli.* UV-irradiation causes DNA lesions that transiently block both replication and transcription in *E. coli*. The first high throughput examination of the *E. coli* SOS response was a time-point study done in 2000 with UV-stress and microarrays covering 95.5% of all known ORFs [6]. In this study Courcelle *et al.* show significant upregulation of several known LexA dependent and independent genes, in addition to a wide range of downregulated genes in response to UV-irradiation in the *E. coli* strain MG1655. As expected, known SOS inducible genes were among those upregulated, additionally there were many upregulated genes with unknown functions and/or that had not previously been described as SOS inducible. The measured decrease in expression levels in a wide range of genes is suggested to be a global stress response, but the experiment could not verify if this decrease is caused by a decrease of transcription or accelerated transcript degradation. Interestingly the normal *E. coli* transcriptome showed a significantly larger number of reduced transcript levels compared to the LexA deficient version. This indicates some LexA dependent transcript inhibition or degradation as another result of the LexA controlled machinery.

Mitomycin C is a DNA damaging reagent generating inter-strand cross-links [45, 46]. The crosslinks prevent separation of the DNA strands and can completely block replication and translation, which is the reason mitomycin C and other inter strand cross link inducers have been used in cancer chemotherapy [31, 45]. The *E. coli* transcriptional response to mitomycin C was measured in the Khil and Camerini-Otero time-point study from 2002 [47], and revealed regulation of more than 1000 genes, including upregulated genes from LexA independent repair responses. The study was performed using microarrays covering all known ORFs in the *E. coli* genome. Of the total of ~1200 (~30% of all ORFs) significantly regulated genes detected in this study two thirds were downregulated, which coincides with the general agreement of a general downregulation combined with a specific upregulation in response to DNA damage. The remarkable difference in the amount of differentially regulated genes compared to the UV-irradiation time point study by Courcelle *et al.* [6] can possibly be explained by a difference in growth medium, the *E. coli* strain used (*E. coli* W3100 versus MG1655), the dosage and type of stress and the microarrays used.

Notably, mitomycin C is an alkylating agent that do not invoke the adaptive response to alkylation damage (described below), since the Ada protein remains unmethylated by mitomycin C induced DNA damage. Interestingly, crosslink damage of DNA may introduce translocations and other DNA damage known to be present in cancer. Hence, mitomycin C may both induce cancer related damage and be applied chemotherapeutically to treat cancer. Usage of mitomycin C in cancer chemotherapy has been reviewed by Thomasz [45].

DNA damage induced by nalidixic acid is an example of damage severely compromising the normal DNA, without invoking the SOS response directly. Nalidixic acid results in double stranded breaks (DSB) in the DNA. The DSBs alone are not sufficient to induce the SOS response and recBCD processing of the DSBs is needed for SOS induction. The recBCD enzyme degrades and unwinds the DSBs and the unwinding results in SOS induction, supposedly due to detection of damaged ssDNA [31].

1.1.5.1 Adaptive response to alkylation damage

Alkylating base damage is a type of DNA damage where bases in the DNA are damaged by the addition of a methyl group or an alkyl group. The main DNA repair pathway for repair of alkylation damage in *E. coli* is called the adaptive response to alkylation damage [48]. This damage is typically induced by exposure to alkylating agents inside the cell or in the environment. Hence, this response is important to different species and it is highly conserved in many bacterial genomes, although sometimes differentially organised [49]. As with the SOS-response the genes involved in the adaptive response have a main regulator; the Ada protein. The Ada protein is both a repair protein and a regulator of the adaptive response pathway consisting of the *ada-alkB* operon and the *alkA* and the *aidB* genes (Figure 4). The adaptive response in bacteria has been reviewed by Teo *et al.* [50], Sakumi & Sekiguchi [51], Landini & Volkert [52] and Sedgwick & Lindahl [49].

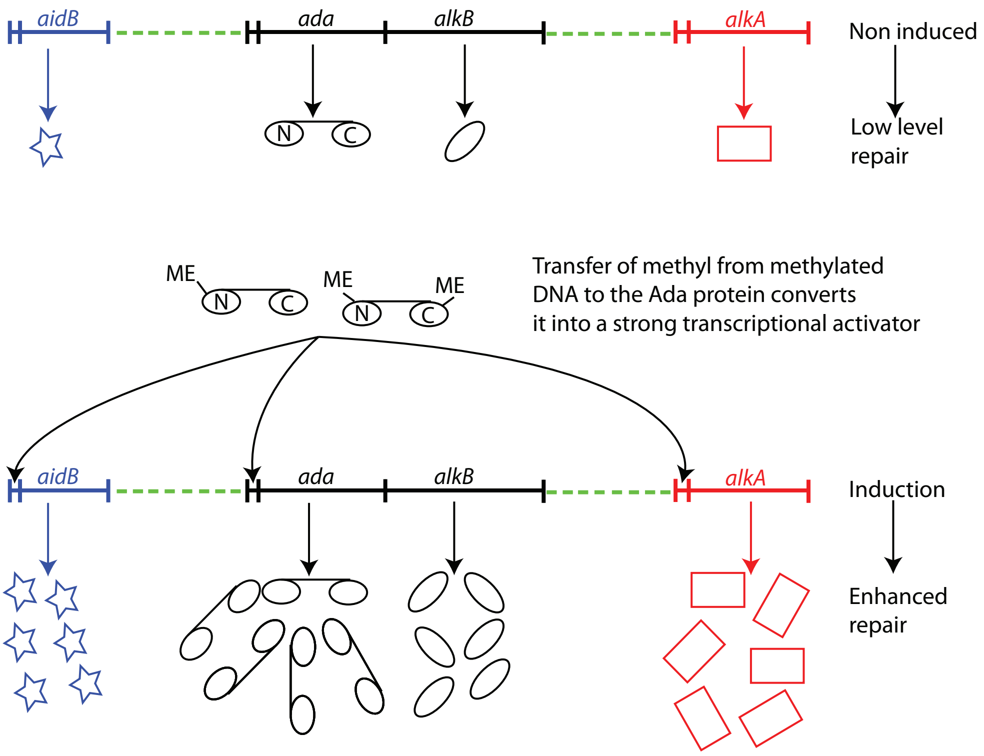
aidB
ada
alkB
alkA
Non induced
Low level repair
N
C
ME
ME
ME
Transfer of methyl from methylated DNA to the Ada protein converts it into a strong transcriptional activator
Induction
Enhanced repair

Figure 4. Induction of the adaptive response. Exposure of *E. coli* to methylating agents makes the cellular DNA alkylated at many cites. The Ada protein, only present at a low level in the cell (top), will be methylated during repair of methylated DNA (middle). This process converts Ada into a strong transcriptional activator of the *ada* inducible genes (middle). The induction of the *ada* inducible genes strongly enhances the repair of methylated DNA (bottom). (Figure adapted from Lindahl *et al.* [53].)

The *ada* gene product has at least four roles: repair of alkylated DNA lesions, sensor of alkylation damage, it is a transcriptional activator of the *ada* gene itself and other adaptive response genes, and it also terminates the adaptive response.

To be turned into a transcriptional regulator Ada needs to be activated. Methylated phosphates in DNA are the signals that transfer Ada to its transcriptionally active form. The Ada protein repairs such lesions by removal of the methyl groups from the damaged DNA, and during this process Ada becomes methylated (^{me}Ada). The methylation turns Ada into a strong transcriptional regulator of the Ada regulon genes, and is followed by a large increase of the number of Ada proteins and increased transcription of the *alkA, alkB* and *aidB* genes.

The *ada* and *alkB* genes are organised as an operon, and the alkB protein is known to repair 1-methyladenine and 3-methylcytosine in DNA by oxidative demethylation [54, 55]. The alkA protein is a glycosylase that catalyses removal of several methylated base lesions [49]. The final (known) player in the Ada DNA repair regulon is the *aidB* gene. The function of the *aidB* gene is somewhat uncertain, but it is believed to be important to the adaptive response. It has been shown to reduce the mutagenic effect of N-methyl-N'-nitro-N-nitrosoguanidine (MNNG) [56] and it is suggested that it directly destroys certain alkylating agents [49].

The Ada-dependent promoters located upstream of *alkA*, *aidB* and the *ada-alkB* operon only allow a high level of transcription when methylated Ada (^{me}Ada) is present in the cell. The actual interaction of ^{me}Ada at the *ada*/*alkB* promoter has been presented in different versions [52]. Lately several studies indicate that a ^{me}Ada-σ^{70} (RNA polymerase subunit) interaction is what enables transcription activation [57, 58]. The Ada role at the *alkA* promoter is different from the *ada* promoter and transcription of *alkA* is allowed with ^{me}Ada and with normal Ada, but in the latter case at a far lower level [59, 60].

A recent study by He *et al.* [61] shows that the regulatory elements of *ada* regulon promoters consist of two boxes; A box (AAT) and B box (GCAA) separated by a six nucleotide spacer. This novel promoter region is similar for all four Ada regulon genes. Normal Ada proteins do not achieve a strong enough binding with the A box and B box to activate the promoter, while a ^{me}Ada protein will stably bind to the *ada* regulon promoters and enhance transcription. It is not known why the methyl group so strongly enhances the Ada binding ability, but the difference in the binding efficiency is significant.

Methylation of the Ada protein is irreversible and some functionality for turning off the adaptive response genes after completed DNA repair is needed, as high expression of the adaptive response genes is lethal in healthy cells [62]. In the final stages of the repair process unmethylated Ada will accumulate in the cell. High concentrations of normal Ada in the cell (>200 molecules pr cell) will deactivate *ada* transcription caused by ^{me}Ada, and the adaptive response will rapidly be switched off if no further damage occurs [63].

Methylating agents are by far the best inducers of the adaptive response, some induction can be seen during DNA damage from large alkyl groups, but large alkyl lesions are usually more efficiently repaired by the uvrABC-dependent nucleotide excision repair pathway [52].

1.1.6 Human fusion genes

In humans the genome is arranged in chromosomes, and the genes consist of one or more exons separated by introns. As described, genes can occur in many splice variants by including or excluding exons when processing the mRNA prior to translation. Splice-variants give the resulting protein different functions and are part of the normal variation in healthy human cells. The genomes of human cancer cells are often heavily damaged. The DNA damage in cancer cells often includes numerical gains or losses of whole chromosomes, structural changes such as translocations, intrachromosomal deletions, amplifications or inversions. Such abnormalities are often the pathogenically essential features of cancer genomes, and gene fusions caused by chromosomal rearrangements are the most widespread genomic alterations known in cancers [64]. A fusion gene is caused by inter or intra-chromosomal rearrangements, and arise when a gene *A*, often having a strong promoter, fuse with a gene *B* (Figure 5).

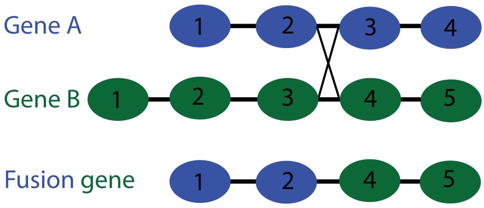

Figure 5. Fusion genes. Fusion genes occur when a chromosomal rearrangement creates a "fused" gene consisting of exons from two different genes. (Picture courtesy of Marthe Løvf).

The resulting fusion gene thus has exons from both gene *A* (the upstream gene) and gene *B* (the downstream gene). The fusion break point usually occurs in the intronic regions of the genes [65] and the expression level is controlled by the *A* gene's promoter [66]. Consequently, fusion gene products are often highly expressed and the protein functions diverge from their origins (often towards malignant behaviour). The chromosomal translocations usually create two fusion genes as they are reciprocal, but detection of both gene fusions are uncommon as only one is controlled by an active promoter [67]. One known rearrangement in leukemia results in a fusion gene with exons from the *TCF3* gene being followed by exons from the *PBX1* gene [68]. Particularly, fusion genes seem characteristic of haematological malignancies and sarcomas, and they can serve as potential drug targets [69, 70], diagnostic markers [71-73], and prognostic parameters.

The Philadelphia chromosome in chronic myeloid leukaemia discovered by Novell and Hungerford was the first cancer related chromosomal rearrangement detected [71, 72, 74]. Several reviews on fusion genes and cancer have been published since then; for example Mitelman *et al.* [75], Gasparini *et al.* [67], Morris et *al.* [76] and Kumar-Sinha *et al.* [77]. Fusion genes can be found in neoplastic cells from a wide range of cancers, and their presence may be used in differential diagnosis and/or therapeutic decision making. Hence there is a great interest in, and a need for, development of high throughput screening methods for known and novel chimeric fusion genes. Today, detection of fusion genes in cancer usually includes laborious and time-consuming methods like chromosome banding analysis (karyotyping), fluorescent *in-situ* hybridisation (FISH) and reverse transcriptase polymerase chain reaction (RT-PCR). But as none of these methods have any high-throughput screening potential they all require some educated guess about where to start the search. Without any clue as to which genes fuse, all genes and their possible exon-exon combinations must be tested, and the testing is unfeasible within a reasonable time.

There are two aspects of the detection of fusion genes; on the one hand cancer researchers want to detect all fusion genes in a wide range of cancer types so that further research can focus on the function of chimeric fusions and possible roles as drug targets. On the other hand a high throughput chimeric fusion gene screening method is very interesting as a tool in diagnostics, hopefully at a very early stage in the cancer development.

Oligonucleotide microarrays have been used in a few studies to detect known fusions in cancer cell-lines, but these arrays have only covered a small set of predefined fusion junction sequences [78-81]. Another option is to apply exon arrays, *i.e.* microarrays giving intensity measurements for all (or a selection of) human exons, to look for intragenic gene expression profiles that have significant internal edges between the up- and downstream parts. A major problem with this method is that even though possible junction points for fusion genes can be detected, one has no idea of which genes fuse, as no fusion sequences (sequences made up of part from both the *A* and the *B* gene) are probed.

Recent developments of ultra-high throughput sequencing platforms have enabled whole transcriptome sequencing, and consequently several novel fusion genes have been detected in cancer cell lines [82-85]. The limitations of these studies today are the number of genomes that can be investigated due to the high cost and computational challenges.

1.2 Microarrays

Different microarray types have a wide application range, and they all apply the basic principle of having known compounds (probes) organised in a grid-like fashion on some sort of slide (array), followed by a hybridisation to some labelled unknown biological compound (sample). By knowing that the sample compound will, according to biological rules, bind to the probes, one can use the labelling of the samples when reading a microarray (chip) to detect the concentration of compounds in the sample (target) bound (hybridised) to any of the known probes on the array. Subsequent data analysis will focus on investigating which probes that have hybridised with the target, and very often the interest lies in detecting difference between reference and treated sample hybridisations. Microarrays today have a wide range of applications from DNA expression studies [86] to protein interactions [87] (Table 1), but only oligonucleotide arrays will be discussed here (reviewed by Lockhart and Winzeler [88], Yazaki *et al.* [89] and Liu [90].

An oligonucleotide array is a microarray where the probes are short (25-80 nts) single stranded DNA sequences (oligonucleotides), and the sample is a collection of short, labelled DNA or RNA sequences (locked nucleic acids [91] and other synthetic variants are disregarded here). At the time of writing, high-density oligonucleotide arrays (HDONAs) have feature numbers ranging up to ~$6.1x10^6$ (Affymetrix GeneChip® exon arrays [92]). A feature is a unique microscopic region on the microarray (most often a glass or silicon slide) housing similar probe sequences measuring the presence of one unique complementary sample sequence. The actual feature number of an array varies with the price and the platform type.

Table 1 – Overview of a variety of microarray technologies and usage.

Application /technology	**Usage**
Custom arrays	*Custom designs to address needs not addressed by any other array*
Tiling arrays	*Arrays with overlapping probes to give a dense representation of the DNA/RNA investigated, these arrays may be used for transcriptome mapping and detection of alternative splice forms*
Resequencing arrays	*Arrays made to control and refine genome sequences derived from different sequencing methods*
SNP detection	*Arrays to identify Single Nucleotide Polymorphism within or between species or populations. Usage may be within genotyping, forensic analysis, drug candidate identification etc..*
Alternative splicing detection	*Also referred to as "exon junction arrays". These are made to detect alternative splicing of genes. Some of these arrays have probes consisting of sequences from two exons.*
Exon arrays	*Compared to gene expression arrays with 1-3 probes pr gene these arrays have several probes per exon within each gene. These arrays may also be used for alternative splicing detection.*
Chromatin immunoprecipitation on Chip	*These arrays are used to locate protein binding sites throughout the genome. One may use several different chip types for ChIP on Chip studies.*
GeneID	*Often relatively small chips (low feature number) applied to identify organisms in food and animal feed. This approach is often combined with PCR verification.*
Comparative genomic hybridisation	*This technique is applied to investigate the genomic content within different cells or between closely related organisms.*
Gene expression profiling	*Arrays used for measurements of the expression level of several thousand genes simultaneously, often applied to detect changes.*
Capture array	*Arrays used to capture oligonucleotides of special interest, these might subsequently be used in high throughput sequencing.*

A labelled sample (target) is hybridized to the microarray, and binds to complementary probes (Figure 6). When hybridising one sample per array, one colour labelling is used, and if using two samples per array the samples have one colour each. In the one colour case, comparisons are done between arrays; typically by hybridising a reference sample to one chip and a treated sample to another chip. The two colour arrays also enable intra chip comparison as probes with one dominating colour would be those differentially expressed in the reference and the control.

Most microarrays are designed for one specific task or organism, exceptions include microarrays made to discover expression on related unsequenced organisms. Today, a variety of arrays can be bought off the shelf for a wide range of organisms. Commercial arrays now come with a definition of each probe sequence, their exact position on the array and most often a mapping to the genomic location of each probe sequence.

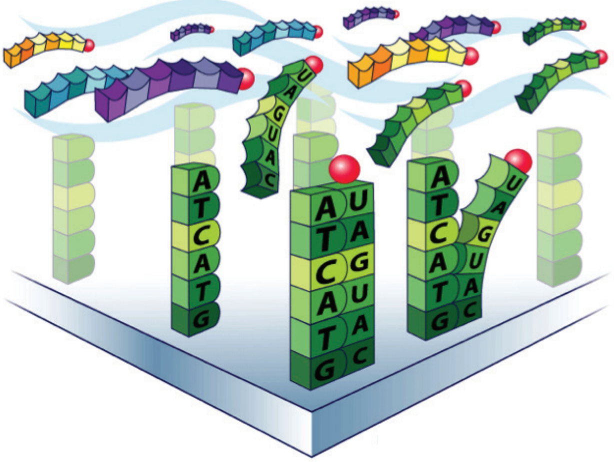

Figure 6. Hybridisation of labelled nucleotide-strings from the sample to the oligonucleotide probes. The labelled (with a red dot) nucleotide sequences from the sample hybridises with completely complementary probes located on the array (picture from the educational materials at *www.affymetrix.com*).

The microarray technology for the studies performed in this thesis is the NimbleGen platform. The NimbleGen platform occurred as a novel product after 2002 when the company managed to develop maskless *in situ* synthesizing of oligonucleotides, by modifying the standard photolithography process using a digital light processor (DLP) [93, 94]. This Maskless Array Synthesis (MAS) technique relies on the DLP projection technology where digital micromirror devices (DMDs) are programmed to create a "virtual" mask to control the oligonucleotide synthesis on the array. Applying this "programmable light" NimbleGen could eventually synthesise as many probe sequences simultaneously as there were micromirrors in the DMD connected to the DLP.

On the NimbleGen arrays each feature (often referred to as probe) is made up of approximately 1 x 10^6 similar oligonucleotide sequences (probes) connected to the array surface by linker sequences. The linker sequence is designed to obtain some distance from the slide to the actual probe sequence since probe sequences too close to the slide might behave suboptimally. The novel NimbleGen technology has opened up affordable custom-designed arrays with far more probes per array, when compared to any previous offers. The major advantage for the NimbleGen arrays compared to the Affymetrix GeneChip® arrays is the maskless production process, as production of the GeneChip® arrays is time consuming, expensive and has little flexibility due to the mask-dependence. Hence, programmable mirrors for maskless array production are both far more flexible and also less expensive.

A major and ever present challenge to microarray technology has been experimental noise. As the microarrays might have as many as 6 million features it is obvious that a great number of false positives will occur just by random chance. Highly increased production accuracy combined with strict lab procedures, randomized probe arrangement and introduction of control probes, assessing both the sample preparation and the hybridisation, have greatly improved array data quality, but normalization is still needed [95]. The role of the microarray, as the most popular and widely used high-throughput tool for expression studies, is now challenged by ultra-high throughput sequencing. How this novel technology will influence the future of microarrays will be discussed later.

1.2.1 High-density oligonucleotide arrays

High-density oligonucleotide arrays (HDONAs) are microarrays with a large amount of short-sequence DNA features on each array (typically > 100,000). HDONAs have for several years been applied for detection of (differential) expression by probing genomic regions known or expected to code for proteins. The regions probed were most often RefSeq sequences [96] and/or computationally annotated genes (homology or start/stop codon searches). Some of the first HDONAs were the Affymetrix GeneChip® arrays [97], designed with each gene covered by a set of probes (*probe set*), and each probe set consisted of about 11-20 *probe pairs*. Every probe pair consists of one perfect match (PM) and one mismatch (MM) probe (PM and MM probes will be discussed later). The major manufacturers of HDONAs are Roche NimbleGen, Agilent Technologies and Affymetrix.

1.2.2 High-density oligonucleotide tiling arrays

As high-density oligonucleotide arrays emerged (Affymetrix GeneChip®), the large probe numbers enabled not only several probes per coding region, but also alternative designs with probes placed in an overlapping manner (like roof-tiles) along all genes investigated (Figure 7). The resolution (sometimes coverage) of a *tiling array* design denotes the number of nucleotides between the start-point of two neighbouring probes, as one cannot pinpoint any transcript start or stop with a higher accuracy than the experiment resolution. In response to the revelation of fewer protein coding genes than previously believed (approximation in humans; 20-25.000 in 2004 [98] versus ~70-100.000 in 1993/94 [99, 100]), regions previously described as "junk" DNA are now more often probed. The resulting interest in probing all genomic regions, combined with the rapidly increasing feature number on arrays, has brought forward whole-genome tiling arrays. Whole-genome tiling arrays have been designed for organisms having relatively small genomes (10^6 nts) and for some human chromosomes. Currently there are a few whole-genome tiling arrays commercially available.

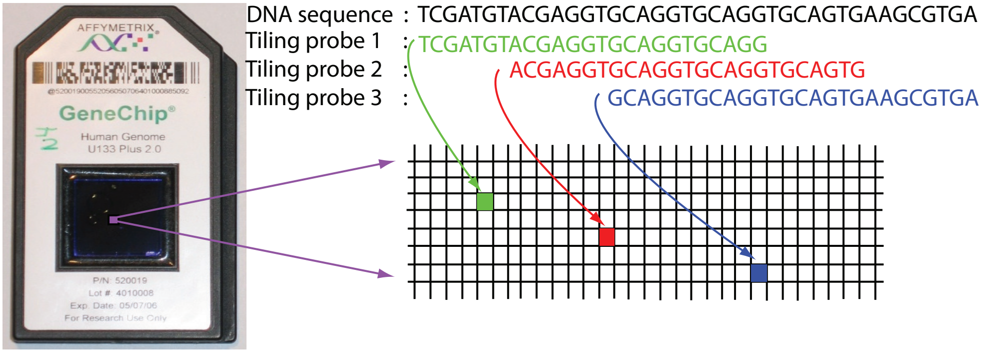
AFFYMETRIX
GeneChip®
Human Genome U133 Plus 2.0
P/N 520019
Lot # 4010008
Exp. Date 05/07/06
For Research Use Only
DNA sequence : TCGATGTACGAGGTGCAGGTGCAGGTGCAGTGAAGCGTGA
Tiling probe 1 : TCGATGTACGAGGTGCAGGTGCAGG
Tiling probe 2 : ACGAGGTGCAGGTGCAGGTGCAGTG
Tiling probe 3 : GCAGGTGCAGGTGCAGTGAAGCGTGA

Figure 7. Picture of an Affymetrix GeneChip® microarray, and a schematic view of a simple tiling design. On the left is a picture of the Affymetrix GeneChip® microarray (actual size 4x7cm). The GeneChip® is not a tiling array and is only used for the illustration. On the right a DNA sequence has been covered by three 25 nt tiling probes. Multiple copies of unique probe sequences are hybridised to unique spots defined by the mask on the Affymetrix array. Every defined spot makes up a feature on the array. The density of the design is the distance between the start of the probes (here 7nts).

1.2.3 Custom design of high-density olignucleotide microarrays

When your field of study falls outside the range of pre-designed arrays, the only other option, if microarrays should be applied, was, and still is, custom design. Custom designs of arrays with a low feature number (<< 100,000) have been used for many years, but only lately has high-density (> 100,000 probes per array) custom design been affordable and hence possible [101]. The upside of custom design is the complete and accurate control of each and every probe on the array, while the downside includes the sometimes laborious work of probe-design and the, often even more, complex task of analysing an array that no regular analysis-tool knows how to handle. The array design parts of this thesis will focus on issues related to design arrays such as those from NimbleGen arrays, but it can easily be applied to other HDONA platforms.

The basic tiling strategy is simple: i) remove (mask) all repetitive regions in the DNA investigated, ii) place a probe at each N^th^ nucleotide all along the DNA. The size of N is dependent on the array feature number and the length of the investigated genome. Experiments performed to understand the importance of uniform probe affinities (hybridisation capabilities) uncovered the need of different tiling designs [95, 102]. The idea of a more sophisticated design is to avoid probe quality bias before array production rather than trying to adjust for it by normalization after hybridisation. Several programs and algorithms have been developed for optimal probe selection [103-111], these are more or less useful for alternative tiling designs and will be discussed later. But since a simple tiling design of one probe at every N^th^ nucleotide is easily made, only a few alternative whole-genome tiling probe selection strategies have been reported (Bertone *et al.[103]*, Graf *et al.* [112] and Schliep *et al.* [109]).

The ultimate, but unachievable goal of a microarray design, is to have i) high coverage of all interesting sequences, ii) many good control probes, iii) uniform probe affinities of all probes, iv) all probes equidistantly spaced, and v) no genomically closely located probes placed adjacent on the array.

A design process should begin with a biological question: What is the hypothesis of the study, and what is the question asked? "People tend to go out blindly and do (microarray) experiments, then go back and try to analyse them and figure out what the question is afterwards" (citation J. Quackenbush, Nature 2004, [113]). In the case of oligonucleotide arrays the single most important part is that the probes on the array match the nucleotide sequences one intends to measure. Not being able to buy commercial arrays with this capability is what drives custom design. Hence, the designer must make an array design fulfilling this need in parallel with meeting the desired goals mentioned above.

1.2.4 Criteria of high quality probes and arrays

The oligonucleotide probes constituting each feature are exactly similar, but each feature should differ unless multiple copies are specifically wanted. The sequence of the different probes may also vary in length not only in nucleotide composition.

The nature of DNA-binding makes the design of each probe on the array far more difficult than the simple task of finding an oligonucleotide sequence complementary to each of the sequences you wish to probe for. As the binding between A-T is somewhat weaker than C-G the nucleotide distribution of the probe will affect the probe affinity [114, 115]. Another problem occurs when DNA sequences only partially complementary to a probe bind and cause unspecific hybridisation. Further considerations include the temperature at which the labelled single stranded sample will unwind from its secondary structure to enable hybridisation. This temperature should be equal for all DNA in the sample that is complementary to any probe on the array. Sequence complexity of the probes should also be accounted for, meaning that a probe should not consist of short repeats of nucleotides which might invite more unspecific hybridisation, and the probe sequences should also be unique not only to the chip, but should occur only once in the target genome. A probe that is non-unique to the target genome cannot be pinpointed back to one single occurrence on the genome. Hence, one would not know the origin of the measured RNA. The probe quality characteristics make up the probe quality (sometimes affinity) of every probe on an array. The number of features available on the chip is also important since high coverage is one main goal, but the biological question, genome sizes, number of genes and types of splice-variants will also play an important role when designing arrays.

The criteria defining a high quality microarray probe can (although they are very closely related) be divided into three: i) homogeneity, ii) sensitivity and iii) specificity. These criteria should be met as far as possible for each probe, and it is easily seen that the more probes on the array, or the stricter the placement boundaries for each probe are, the harder (sometimes impossible) it is to select high quality probes only. The basics of the three criteria are described here:

i) Homogeneity: To enable a fair analysis, where equal intensity measurements from different probes can be considered to be signals from comparable probe binding levels, it is important that the probes are homogenic, *i.e.* their abilities to bind to the sample are equal under the same hybridisation condition. The hybridisation ability of probes are sometimes computed by using the Nearest Neighbour Model [116], but more often predicted by other algorithms (also focusing on the melting temperature of each probe (T_m)) due to the demand for computational speed [105, 108, 110]. The T_m computations differ in complexity between the algorithms that are applied, and sometimes other thermodynamic properties (entropy, enthalpy and free energy) of the hybridisation are also included in the algorithms.

ii) Sensitivity: The idea behind the sensitivity criteria is to ensure no self-binding among the selected probes (increasingly important with probe length [108]), as self-binding makes the probe inaccessible for hybridisation to the sample cDNA. The usual way to meet the sensitivity criterion is to predict the secondary structure of the probe sequence; the more stable, the less sensitive. The most widely applied algorithms for secondary structure predictions used are MFOLD [117] and Vienna RNAfold [118]. These software packages are highly computational intensive, and other faster algorithms have been introduced by for instance Li & Stormo [119] and by Wernersson & Nielsen [110]. The Li & Stormo algorithm exploits the fact that the lowest energy structure can be fairly precisely computed using simple heuristics on the probe versus target alignment [119]. This alignment is already applied in the design process, thus overall computational speed is increased. Wernersson & Nielsen base their algorithm on a di-nucleotide dynamic programming algorithm, where each di-nucleotide has a predefined stacking energy.

iii) Specificity: The optimal specificity is when a probe only binds to the perfect complementary sequence. The basic criteria are that the probe should not in general sequence terms be too similar to any other part of the genome than the target, additionally no subsequence longer than some defined cutoff should be exactly similar to any other part of the genome. Specificity increases with probe length, as the free energy involved in hybridisation of the probe-target complex increases with length, but overly long probes could result in the hybridisation of complete matches of short cDNA samples to parts of the probes, hence producing false positives [120]. The specificity of probes is often computed by one of two methods; Hamming distance [121] or BLAST [122, 123] searches. The Hamming distance is computationally intensive and time consuming (runtime for probe-length m in a genome of length n is $O(mn^2)$), and hence often impractical to use. BLAST searches will be time-dependent upon the number of sequences in the BLAST database, hence if probes are designed to detect species specific genes in a pool of genes from a set of species, BLAST may also be very time expensive [108].

Another important factor is to have an error free array production making each probe represent exactly the sequence it has been designed to represent on all arrays of one design. The error rate of maskless *in-situ* hybridisation production of microarrays is not given from any manufacturer, but NimbleGen claims it is very low (a verification study of the low error rate has not been seen yet, and I assume it would be very difficult to make one).
Finally, genomically closely related probes should never occur closely located on the array to minimise the effect of systematic spatial errors.

The various experiments performed with microarrays to understand sequence dependent binding affinities, probe placement, usage of hybridization controls, spike-ins and probes for background signal estimation seem to have forced commercial array manufacturers focus more on production of arrays that produce high quality data than before.

1.2.5 Control probes

In addition to probes designed to investigate the labelled sample, a proper microarray should also include control probes. Control probes are designed solely to investigate the quality of the experiment, and control probe signal intensities can be used to adjust for systematic bias (discussed later). Today, control probes are mainly used to investigate the target preparation and the hybridisation process. The basic idea of control probes is that they probe for some predefined RNA or DNA sequence that, if possible, should be entirely foreign to the target genome. The hybridisation control probes are plainly named hybridisation controls (less commonly called spike-ins), and typically probe for a set of four different genes. In the hybridisation cocktail, prelabelled cDNAs (if cDNA is hybridised) of known concentrations are added for each gene (different concentration per gene). Consequently the researcher knows what to expect from the intensity readout of these four genes (when comparing them to each other) during analysis. The second type of widely applied control probes is process control probes (more commonly named spike-ins). These are made to investigate the entire target preparation process. Such a spike-in set also consists of a set of genes and corresponding probes, but in this case, unlabelled, unfragmented RNA is added at an early stage in the target preparation process (at different, but known, concentrations for each spiked-in gene). Usage of the spike-ins is similar to the hybridisation controls, and there is an expected relative probe intensity value for each gene, which can be used to verify the quality of the sample preparation. Often the spiked-in concentrations are designed to make a graph of the $\log_2$ intensity values, that can be described by a linear function. This is to enable easy quality control read-out. Inter-array differences in the intensity values can subsequently be used to normalize the arrays with respect to each other, and sometimes the values are also used for probe normalization within arrays.

1.2.6 Perfect match and mismatch probes

High-density oligonucleotide microarrays were originally introduced by Affymetrix. The first design covered each gene by a *probe set* consisting of *probe pairs*. Each probe pair consisted of a 25nt *perfect match* (PM) probe and a 25nt *mismatch* (MM) probe. The PM probe is a perfect complimentary sequence to the target sequence, while the MM probe is made with the middle nucleotide (13th) changed (Affymetrix exchanges it with the compliment) to create a mismatch (MM) probe. The idea behind this strategy was to use the MM probe to estimate the level of non-specific binding that introduces error to the PM probe.

Several different studies began suggesting MM-independent normalization and background signal estimations as it was shown that MM probes were not only measuring non-specific hybridisation [114, 124, 125]). It has also been shown that the inserted nucleotide strongly affects the probe affinity and that random probes for background signal estimates may be better [126].

Eventually Affymetrix recognised MM independent background estimates and at the same time made room for almost twice the amount of PM probes on their GeneChip® arrays covering all human exons [92]. These exon arrays have included a different type of background estimation probes. First all PM probes are separated into groups (bins) according to their nucleotide compositions. Secondly the background estimation probes are created in groups (bins), where the probes in each bin are designed to measure the background signal of the PM probe bin it corresponds to (in terms of nucleotide composition). By using this bin-based method, Affymetrix claims precise background estimation using far fewer probes than one MM per PM. As long as there is no clear argument that can be made for including MM probes, they will remain irrelevant. The most common problems associated with MM probes will be discussed further in the analysis chapter.

1.2.7 Tools and methods for high-density oligonucleotide (tiling) array design

Several algorithms and computer programs have been made for the purpose of optimal probe selection, of which some have been made specifically for the design of tiling arrays. Repeatmasker [127] and DUST (Tatusov & Lipman, unpublished) are typical representatives for programs applied in the early process of tiling array design. These programs are made to filter out repeat regions, regions with extreme A/T or G/C content, or stretches of polypurine/polypyrimidine bases. These are typical regions that are believed to carry little genomic information, and they are known to cause cross-hybridisation problems and non-specific binding. DUST can even be used for more complex filtering by applying an information entropy-based model of sequence analysis [103].

OligoArray 2.0 [108] from 2003 was the successor to OligoArray [107] and was designed to meet the needs for automated design of short oligonucleotide probe sequences (25mers). The OligoArray 2.0 program requires BLAST [122, 123] and applies a thermodynamic approach (using MFOLD [117]) to predict secondary structures and the probe specificity, and the user can adjust parameters such as the probe length, GC-content and the T_m range.

CommOligo (and its accompanying CommOligo Parameter Estimator) from 2005 is software made to either design whole-genome arrays (coding regions only) or probes for highly homologous sequences [105]. The probes are selected by applying multiple filters for all possible probe sequences; sequence identity, free energy, continuous stretch, GC-content, self-annealing, the distance to the 3'UTR and the melting temperature (T_m). The programs were tested on 50-mer designs and the performance was good compared to existing tools like OligoArray [107], OligoArray 2.0 [108] and OligoPicker [128]. The probes were score based on their similarity to non-targets and longest continuous stretch, but the program is relatively slow.

OligoWiz 2.0 from 2005 is a web-tool for probe design [106, 110]. The probe selection program presents a graphical interface showing each probe and several default and optional parameter scores. The default parameters are cross hybridization, T_m, low-complexity, position and folding. The cross-hybridization score is a BLAST based and defines how similar this probe sequence is to any other sequence in the entire input-string for this probe selection. The T_m score defines the optimal melting temperature for the chosen probe sequence, and the low-complexity score measures whether fragments of the chosen probe sequence are part of common species-specific repeat sequences. The position score is made to compensate for the fact that the reverse transcriptase is likely to fall off the transcript during cDNA reverse transcription, and hence, it is preferable to place the probe close to the reverse transcription start site. And the folding score procedure is described in section 1.2.4 point ii). The possibility of applying additional custom parameters, along with the default parameters that are fairly common to probe selections, makes OligoWiz 2.0 very flexible, and applicable to tiling design. Usage of OligoWiz 2.0 has been presented by Wernersson *et al.* [129], and the program is available at http://www.cbs.dtu.dk/services/OligoWiz2.

Bertone *et al.* [103] presented two algorithms, a technique for finding the optimal tiling path through longer sequences, and a method to determine the similarities between probe sequences covering large genomes. Their study from 2006 concentrated on tiling array application to eukaryotic genomes where a major goal of the design process is to remove repeat regions, as such regions often contribute to unspecific hybridisation, and removal gives better coverage in genomic regions viewed as more interesting. They emphasise the value of applying a tiling strategy that optimises probe affinities (thermodynamic properties) compared to having a uniform tiling resolution. The algorithms were embedded in a web tool available at http://tiling.gersteinlab.org.

MAMMOT (2006) is a web-tool made both for design of tiling arrays and visualisation and processing of experimental data [111]. MAMMOT is MySQL based, which enables usage on other genomes than those already installed and arrays other than the original PCR product based arrays MAMMOT was designed for. The probe-selection algorithm is based on matching probes towards PCR primers, and the visualisation requires the target genome to be uploaded into the MySQL database. MAMMOT is available at http://www.mammot.org.

A randomized probe selection algorithm for microarray design was presented by Gasieniec *et al.* in 2007 [104]. This algorithm is mainly focusing on selection of one or a few probes that are optimal for measuring genes, typically for detection of species specific genes in an unknown pool of RNA to detect the species present. The algorithm first filters the search space for probes using three probe selection criteria (quantitative (*i.e.* nucleotide composition), homogeneity and sensitivity), and then a "randomized" probe selection is made. The random algorithm is based upon the general idea that a probe similar to some random generated probe "seed" will in general yield a high probe quality. Hence, instead of searching for all good probes to pick from, the program suggests a presumed good random generated probe, searches the input sequence for a similar sequence and then runs quality checks on this probe candidate alone and not all possible candidates. This enables a significant speedup compared to previous probe selection algorithms aiming at the same problem. The three probe selection criteria are defined by the nucleotide composition, T_m values and Hamming distances [121], respectively. The name of the algorithm is RandPS and it is available at http://www.csc.liv.ac.uk/~cindy/RandPS/RandPS.htm.

From the total set of all possible probes covering a genomic sequence one can select "the globally optimal tiling path" through the sequence. One of the very few studies that has tried another approach (instead of enforcing a defined distance between probes) for whole-genome tiling, has applied this strategy to design microarrays with high coverage and high quality probes, by introducing non equidistantly distributed probes [109]. While previous design strategies have typically focused on obtaining a uniform T_m, and have either scored probes due to their cross-hybridisation abilities or the maximum contiguous match length (between probes) for probes targeting a filtered (by DUST or Repeatmasker) sequence, the algorithm presented solves the complete multicriterion optimization problem and works with unfiltered sequences. The algorithm can accomplish this by the usage of unequal probe spacing instead of the very strict probe placement-criteria implied by inter-probe distance parameters used in other designs. They address the probe design problem as a Minimal-Cost Tiling Path Problem (MCTPP). The approach applies a score to the T_m, the probe-quality (here cross-hybridisation and maximum contiguous match) and the inter-probe distance, subsequently the optimal tile path can be found by optimizing the total score based on score penalties for suboptimal individual scores. The feature number and genome length are important constraints in the algorithm. In short, all possible probe candidates scoring over some defined threshold are evaluated first and then the MCTPP algorithm finds the optimal tiling path, in a time scaling linear with the input size. The MCTPP algorithm is able to probe some of the regions that other algorithms would filter out, but some regions would still be left unprobed as no high quality probes can be found within the regions.

1.2.8 Target preparation, labeling, hybridisation and scanning

The design and analysis process for HDONA studies is typical bioinformatics business, while the actual biological parts of microarray studies is taken care of by trained lab personnel. In a typical study the lab work can be divided into four parts:

i) Extraction of total RNA from the cells of the organisms investigated. Sometimes t- and r-RNA are removed due to their abundance in the target. Then the single stranded (ss) RNA is reversely transcribed into complementary ssDNA since ssDNA is more stable than ssRNA. The cDNA is amplified if necessary.

ii) The resulting cDNA is now either already labelled and then cut into smaller pieces by using restriction enzymes, or cut first and subsequently labelled with some material (fluorescent) that is detectable by the scanner.

iii) Hybridisation is the process by which the hybridisation cocktail (target cDNA plus any control cDNA) is applied onto the array and the complementary target sequences in the sample bind to the probe sequences. The hybridisation process most often takes place inside specially designed instruments made for the optimisation and control of all hybridisation parameters.

iv) The final step of scanning includes washing the arrays of all non-hybridised material. Then the arrays are placed in a special scanner that first aligns to the array by chip-specific alignment probes. Subsequently the array is scanned and one picture file and one raw intensity file is produced per array for typical high-density chips.

During these four steps it is very important to avoid introducing any systematic or random errors into the experiments. Such errors might occur when different personnel perform the same task but for different arrays, or if randomization is lacking in the experimental setup (one should for instance not systematically do all reference samples first, followed by the treated samples). Other error sources include procedures carried out with different equipment

(weights, pipettes, scanners etc), experiments carried out on different days with significant changes of temperature and humidity or any other possible way of introducing error.
One error source in many experiments is for example dust or fat on the array surface before scanning. There are lab protocols and commercially available kits for all procedures performed in the lab, but the methods and the results may vary. If one wishes to perform only a minimum of laboratory work (only production of target cDNA), companies like NimbleGen can perform the hybridisation and initial analysis. A set of files with pictures, raw intensity files and initial analysis is returned to the researcher. Some manufacturers might even assist with custom designs of tiling arrays in specific cases.

1.2.9 Analysis of high-density oligonucleotide microarrays

1.2.9.1 Overview

A microarray analysis consists of several steps. The first step is to read the chip and to define the signal intensity of each feature on the array (most often done automatically), followed by a second step of normalization. Normalization is a process whereby one tries to make comparable scales of the intensity data, array-wise and internally on the arrays. During this process one often has to discard some data-points. The intensity values for all features are generally $\log_2$ transformed during normalization. Either when normalising data, or as a separate step, one has to measure the background noise present on the array, meaning that one has to estimate the part of the signal intensity of each feature which is present due to non-specific binding and other signal errors. The final step is now to analyse the intensity values either according to some genomic annotation sharing what the probes are supposed to represent, or totally independently of any "template-like" information of the genome probes (a typical case for whole-genome tiling arrays). This final step creates an expression profile for the investigated transcriptome (when hybridising total RNA).

1.2.9.2 Reading the microarray / image analysis

When the chip hybridisation is finished, the chip is washed and is read by a laser scanner. The scanner aligns a grid system to the chip by using special alignment probes or spots on the chip. If the automatic alignment fails it must be performed manually. When the alignment is deemed optimal, the scanner measures the intensity of the light emitted from the labels bound to sample sequences still present on the array (as they have hybridised to probes). Different array platforms have different scanners and have different algorithms for defining the exact area over which the emitted light should be measured. The most important experimental consideration during scanning is to have clean chips, high resolution imaging and repeatable scanning (preferably using the same scanner to ensure similar reads).

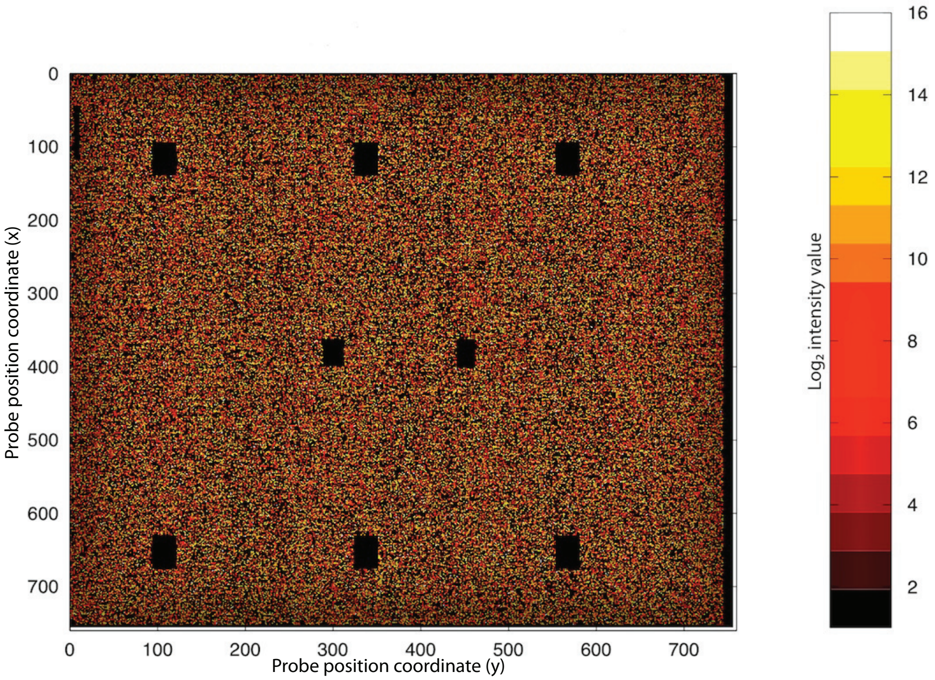

Figure 8. Raw intensity heatmap. Heatmap of raw intensity data for all positive strand probes from one reference sample from the MNNG study. The eight black rectangles are the masked out control probes. Minor vertical spatial patterns can be seen**.**

A useful assessment for detection of obvious spatial errors is to plot the resulting data readout as one heatmap per chip (Figure 8). Since probe placement on the arrays has been randomized regarding their genomic location one should not see any significant patterns of light or dark areas. Such areas would be representing systematic bias if present on the array.

1.2.9.3 Background estimates, normalization and expression measurements

When running a microarray experiment in a lab there will always be some random errors and/or systematic bias introduced, both internally on the arrays and between arrays. Causes might be as diverse as chip production batch, RNA extraction, humidity or a change of the lab-personnel. Additionally, the actual hybridisation on the chip will introduce noise. This noise is caused by random hybridisation, probe-specific effect, cross-hybridisation, half binder effects and spatial errors caused by saturation, dirt or fat. In addition, also the scanner may introduce errors, especially if comparing chips imaged by different scanners. Because of the (assumed) very high accuracy of the high-density array production methods (maskless or by mask) and their lack of robot related problems, as print-batch and print-tip errors, adjustments for manufacturing errors are not applicable here (even though no study on the high-density chip manufacturing error-rate has been seen yet).

The measured signal of each probe hence consists of a true signal and a background signal, where the background signal represents the sum of all signals not representing the true signal, *i.e.* the signal that would be present without any target specific cDNA in the hybridisation cocktail. Hence, it is crucial to form a proper estimate of the background to enable a more correct estimate of the true signal intensity. Normalization is the process where the intensity measurements of one or more probes or microarrays are adjusted to a common scale and/or distribution. Due to more or less systematic bias, this may be necessary both between arrays and internally within the arrays.

The normalization process manipulates the actual data and one should do as little normalization as possible [102], but until further development of the microarray technology and background adjustments matures into a standard approach, normalization should be performed in a manner suitable for the study performed. Many array manufacturers supply their own software packages for background adjustment, normalization and further data analysis; these programs (for instance Expression Console™ from Affymetrix) often have implementations of selected normalization and scaling algorithms. A very popular freeware for microarray analysis is the Bioconductor framework [130] running under R (http://www.r-project.org/), and most of the non-commercial implementations of background adjustment and normalization algorithms are implemented in Bioconductor. Presented here is a selected set of the most used and most relevant background adjustment and normalization methods applied on HDONAs. Many of these methods define an overall expression of the gene probed by a probe set and a measurement of the differential expression of genes between a treated and a reference sample.

The first analysis by Lockhart *et al.* [131] used the average difference (PM-MM) over probe sets (probes probing the same gene) to calculate whether an mRNA was present in the sample, and the log ratio (PM/MM) to compare data (typically between a treated and a reference sample hybridised on two different chips). These two approaches were implemented as AvgDiff and Average Log Ratio respectively in the Affymetrix Microarray Suite 4.0 (MAS 4.0). A major problem arising with the PM-MM or log(PM/MM) methods was that as many as 1/3 of all MM probes may have higher signals then their related PM probes, of which many will have MM >> PM [132]. The MM >> PM problem indicates this method as very inexact.

Model Based Expression Index (MBEI) [133] from 2001 was one of the first analysis methods that did not use MM probes [125]. Li and Wong show in these two papers that the variation of intensities is larger between different probes measuring the same gene (also shown by Naef *et al*. [124]), than between the same probe on different arrays. Following this, the main goal of their MBEI method becomes to adjust for probe-specific effects. Their method also automates detection and handling of contaminated array regions and other image artefacts. Their method is based on the assumption that the intensity measurement of a probe will increase linearly with the actual expression of the corresponding gene, but the rate of the increase will differ between probes. Additionally they assume that the PM intensity will increase at a greater rate than the MM intensity. Each probe (PM and MM) is also given a background noise intensity value and a random error, and probes (sometimes entire arrays) detected as outliers will be removed from the dataset. In their second paper on MBEI [125] the algorithm is made independent of MM probes, as their first MBEI paper indicated that some MM probes show little response to changes in the target gene expression level [133]. They also introduce a baseline array normalization method based on detecting non-differentially expressed genes (between samples) using internal rankings within the arrays. Subsequently they adjust all other arrays to the chosen baseline array using the expression levels of the non-differentially expressed genes. MBEI was implemented in the DNA-Chip Analyzer (http://www.dchip.org).

Micro Array Suite 5.0 (MAS 5.0) is a commercial microarray analysis software package provided by Affymetrix [134]. One major focus of MAS 5.0 was to remove the problems of negative expression values in previously applied algorithms, including the predecessor MAS 4.0. In MAS 5.0 robust estimators are used to avoid negative expression values, *i.e.* values outside the physical range. This algorithm utilizes both the PM and the MM values in the computation and assumes that the absolute signal is the sum of the *true* signal and a *stray* signal (elsewhere referred to as background) and the *stray* signal varies with the probe sequence. Special cases of PM < MM are treated by computing robustly estimated values for the MM intensities (robustly meaning a reasonable MM value giving a true signal inside the physical range). The problems faced due to the observation of varying stray-signals due to differential probe affinities are treated by always performing comparisons (reference versus treated). Since probe affinities remain relatively stable between experiments, the variation of the expression differences between treated and reference samples is much less than the variation between probes investigating the same gene. A problem with this approach is the probe affinity bias calculations of single array experiments (multiple arrays, but no comparison), but the authors suggest a solution by noise estimation through replicate experiments.

Quantile normalization tries to make all probe intensity distributions in a set of arrays the same [135]. By using the mean quantile of all datasets as the data value in the original dataset, equal intensity distributions in all datasets are achieved. In their paper, Bolstad *et al.* [135] showed the quantile normalization to be superior to previous methods. The methods proposed in their paper are implemented in Bioconductor.

Variance stabilization normalization (VSN) was introduced as a microarray pre-processing method in 2002 by Hüber *et al.* [136], and comprises data normalization and quantification of both the differential expression and the measurement error. VSN computes sample-to-sample variations and subsequently the intensities are transformed to a scale that has a variation approximately independent of the mean intensity.

Robust Multi-array Average (RMA) was suggested by Irizarry *et al.* [132] as a novel expression measurement that includes three steps: *i*) Background correction, *ii*) Quantile normalization and *iii*) Fitting a linear model to the background corrected, normalized and $\log_2$ transformed probe intensities. RMA was shown to perform better in terms of sensitivity and specificity than AvDiff, MBEI, quantile normalization and MAS 5.0, particularly for measurements of low intensity signals and small changes in expression levels. The RMA normalization procedure is implemented as a Bioconductor package.

GeneChip Robust Multi-array Average (gcRMA, also referred to as gc-adjusted RMA) was designed by Wu *et al.* [114]. The established RMA method was extended with a probe sequence specific normalization based on estimated probe affinities. The binding affinities were computed using hybridisation theories from molecular biology combined with data from studies designed to investigate probe affinity. Hence the probe affinity of any given probe can be computed by a formula considering the contribution of each base at each position. The authors finally conclude that the gcRMA usage of MM probes can be avoided by using an empirical approach to the background estimates at a low cost compared to the gain of more PM probes. The gcRMA method improves the fold-change accuracy of the original RMA method compared to MAS 5.0 while giving almost as precise expression measurements as RMA. gcRMA is implemented in Bioconductor.

A useful tool for quality assessment of microarray data (raw or normalized) is the arrayQualityMetrics [137] package for Bioconductor, which can easily provide a variety of quality assessment plots for different expression array platforms. The real advantage of this package is that it enables detection of experimental artefacts, and thereby outlier removal, and comparison of raw and normalized data quality and intensity distribution. The catch is that this package utilizes different older Bioconductor packages that depend on using MM probes, which narrows the applicability, since MM probes are now used less.

1.2.9.4 Whole-genome tiling array specific methods

The main difference between typical high-density oligonucleotide arrays, such as the Affymetrix GeneChip®, and tiling arrays is that the latter probes not only genomic regions known or believed to encode genes, but the entire genome (or parts of it). To apply a computationally derived annotation to learn which regions to measure intensity from would be an abuse of the data-material, except for using it as a control procedure to check the expression of genes with well known boundaries. One of the main motivations of the tiling arrays was to be able to perform an unbiased (bias from annotations) transcriptome mapping of entire genomic regions. Hence, analysis should be as independent of previous annotations as possible. Below I present a selection of the analysis methods that have been designed for tiling arrays.

DNA reference normalization applied with tiling arrays was presented by Huber *et al.* [138] in 2006. This approach was presented in combination with a segmentation method for detection of transcript boundaries, and was one of the first published methods designed particularly for whole-genome tiling arrays [138]. The DNA reference normalization method is based on computation of the intensities of all unique (*i.e.* only one genomic complementary sequence) perfect match probes when hybridised to total genomic DNA. In a perfect world, all probes would give the same intensity measurement, but on account of different probe affinities caused by the oligonucleotide composition of each probe, this would not be the case. The reference normalization uses the measurements of the total DNA hybridisation to adjust for probe-specific effects in the actual study. Transcription boundaries are detected by a structural change model (SCM), previously applied in array-CGH studies [139], that models the expression data as a piecewise constant function along the chromosome. The SCM will detect significant jumps in the expression pattern, and thus indicate transcriptional boundaries. The SCM and the DNA reference normalization are implemented in the *tilingArray* package in Bioconductor.

The Model-Based Analysis of Tiling-arrays (MAT) algorithm for detection of transcript boundaries from tiling arrays was presented by Johnson *et al.* in 2006 [140]. This method was designed for chromatin immunoprecipitation (ChIP) studies, but is flexible and can be applied to other tiling array studies. The normalization part of MAT makes two assumptions of the data; i) the majority of tiling array probes (in ChIP studies) measure non-specific binding, i.e. no complementary mRNA is present in the sample for these probes, ii) the large amount of probes per chip ($3x10^5$-$6x10^6$) gives accurate and precise predictions of probe specific effects. A probe behaviour model is derived from the estimated probe behaviour from each single array and used to calculate a standardized probe intensity value for each probe on every array. The actual detection of transcript boundaries is based on a sliding window (600bp window applied, and minimum 8 probes) and a scoring scheme based on a trimmed mean of all the standardized expression values inside the sliding window. Each region (window) is assigned a MATscore assumed to follow a normal distribution for the total of all windows. This can be used to define a P value for regions thought to be enriched in the ChIP study, or as expressed in a transcriptome mapping study. MAT is available at http://chip.dcfi.harvard.edu/~wli/MAT.

The Affymetrix tiling array analysis applied in at least three Affymetrix studies [9, 141, 142] uses quantile-normalization [135], background and probe-specific normalization algorithms as implemented in MAS 5.0 [134] and a smoothing sliding window to divide the genome into transcribed and non-transcribed fragments. Fragments are detected as transcribed if the average score is above a certain threshold set based on spike-ins.

The ExpressHMM method from 2006 focuses on removing previous *ad-hoc* solutions to the question of whether some given genomic region is transcribed or not [143]. The hidden Markov model (HMM) applied is trained on the correspondence between intensity signals and genomic annotation. To avoid bias from erroneous annotations, down-weighting is applied within regions of dubious annotation. A novel normalization method based on previous work [114, 115, 144] is also presented along with this HMM method. ExpressHMM is available at http://www.binf.ku.dk/~kasper/expresshmm.

1.2.9.5 Microarray data storage

Undoubtedly microarray experiments, and specially those of HDONAs, create a vast amount of data. To enable quality control of studies, and availability of this data for non-commercial research all array experiments published today must, in parallel, publish the raw data and preferably also the normalized data in open data repositories. To make the data accessible and understandable for other researchers, the microarray community have agreed upon how to represent the data, and public data repositories have been organised. The two main data repositories are the Gene Expression Omnibus (GEO), hosted by the National Centre of Biotechnology Information (NCBI), USA [145], and ArrayExpress at the European Bioinformatics Institute (EBI), UK [146]. To make the data accessible to others the MGED (Microarray Gene Expression Data) society, founded in 1999, has agreed upon the MIAME (Minimum Information About a Microarray Experiment) standards and the MAGE-ML (Microarray gene expression markup language) format and MGED Ontology with which to represent microarray data. Both the GEO and the ArrayExpress databases are curated, and all data submitted has to be approved before publication. Not only do these databases serve as storage for raw and normalized microarray data, some curated datasets also serve as the basis of a datawarehouse containing gene-expression measurements and other results.

Chapter 2 Aim of the study

The aim of the study was to design high-density oligonucleotide microarrays and custom analysis for detection of DNA repair systems and DNA damage.

The DNA repair systems chosen for this study were the SOS response system and the adaptive response to alkylation damage in *E. coli*, while the DNA damage detection was aimed at human cancer related fusion genes. In *E. coli* we emphasized detection of novel short differentially regulated transcripts. The DNA damage detection in human cancer cells included a completely novel oligonucleotide design, and was therefore aimed at establishing custom designed oligonucleotide microarrays as a means for fusion gene detection. The two applications of custom design and analysis for DNA repair and DNA damage detection are presented in more detail in the following sections.

2.1 Detection of transcriptional changes in *Escherichia coli* using high-density custom tiling microarrays

Although *E. coli* is one of the most studied organisms today, fairly little is known about the overall transcriptome, and how the entire transcriptome changes during exposure to different types of DNA-damaging agents. We wanted to map the *E. coli* reference transcriptome and compare it to the transcriptome when exposed to the methylating agent MNNG and to UV-irradiation. To address this problem we designed an oligonucleotide microarray with dense coverage of the entire *E. coli* genome including all regions; genes, regions opposite of genes (antisense) and regions considered as intergenic. As analysis of tiling arrays was a fairly new field when starting this project (2005), and custom design was in its early phase we also wanted to develop our own design and analysis methods optimized for bacterial tiling arrays.

Our main biological focus in the study of the transcriptional changes in *E. coli* was to detect novel transcripts, protein coding and non-coding, that might be of importance to the complicated systems of DNA repair and damage response.

2.2 Detection of fusion genes using custom designed microarrays

Detection of fusion genes have generally been inefficient and extremely time-consuming, but the fact that fusion genes are considered a major player in several types of cancer, reviewed by [77] and [147], has established a severe need for a high throughput method of fusion-gene detection. Not only is there a need for identification of novel fusion genes, but also the ability to detect known fusion genes present in cancer samples as both a diagnostic and research tool. The main focus has been to design, test and establish custom made high-density oligonucleotide microarrays as a method for a sensitive and specific high throughput detection of fusion genes (Figure 9).

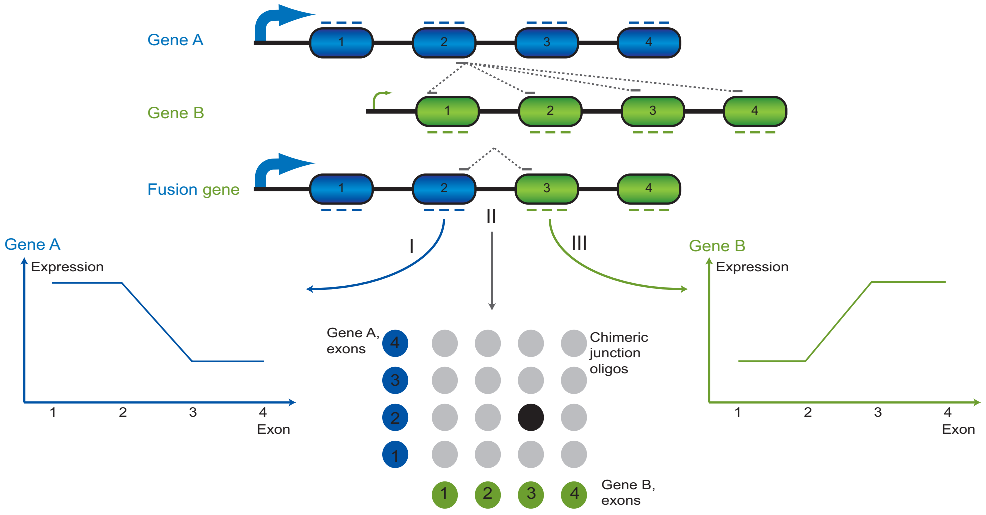

Figure 9. A schematic view from probe design to array based detection of fusion genes. First all possible fusion-junction probes covering how all exons (here exon 2) from the upstream (A) gene might fuse with any exon from the downstream (B) gene) are made (arrow II). Then a set of three inter-exonic probes is made for all exons of all possible fusion genes (arrows I and III). The inter-exonic probes are used to create expression profiles and the chimeric probes are used to detect fusion sites. (Picture courtesy of Marthe Løvf).

Chapter 3 Summary of papers

3.1 Paper I

In this methodology paper, we present our probe design approach and the analysis algorithms applied in paper II and III. The design process utilizes a program named OligoWiz 2.0 [110] for initial probe selection, while we apply custom Python scripts for final probe selection. The major novelty of the design methodology is the selective tiling approach, usage of several copies of each spike-in probe set and the usage of random negative control probes for background signal estimation. Selective tiling means that all probes included in the design have probe qualities that satisfy a minimum cut-off value.

The major part of the paper discusses our minimal normalization approach, and we show that it decreases data variation while maintaining the dynamic range of the raw data intensities. By comparisons to other methods RMA, gcRMA and VSN we show our method as a better alternative for bacterial tiling arrays.

Furthermore, we present an annotation-independent segmentation algorithm. This unguided approach divides the genome into transcribed and untranscribed regions, while simultaneously detecting transcripts differentially expressed between reference and treated bacteria. The algorithm is based on a sliding window of varying sizes and statistical t-tests for detection of significant changes. In the discussion we argue why not to use the established gcRMA method [114], except for the purpose of additional outlier detection. Further, we suggest improvements to our algorithm by introducing probe sequence specific weighting derived by a generalised linear model created on the basis of the array data.

3.2 Paper II

Here we present a high-density whole-genome tiling project to map the *E. coli* transcriptome and to detect transcriptional changes between MNNG treated and reference bacteria. MNNG is known to induce the adaptive response, hence upregulation of *ada, alkB* and *alkA* is expected [31]. We easily detect upregulation of the known adaptive response genes, except the *aidB* gene, which is a gene previously reported as an adaptive response gene. In total we detect 53 upregulated annotated genes, 171 downregulated annotated genes and 17 downregulated annotated ncRNAs. The majority of the differentially regulated genes are previously not described as MNNG responsive. The numbers of differentially regulated genes suggest gene specific upregulation of repair genes, as the specific damage response, and an overall downregulation of many other genes as a general defence mechanism. The most interesting upregulated genes are *recN* and *tisA/B*, which are known SOS response genes, not previously known to be MNNG responsive. The differential regulation of recN was verified by Northern analysis, and we suggest that *recN* might also be involved in base lesion repair, and that it might be regulated by an additional mechanism to the SOS response. We also investigated the regulation of genes found to have the *ada* A box and B box in their upstream regions and found both upregulated, downregulated and unmodulated genes. These findings suggest that the regulation of the *ada* regulon must involve more factors than simply the upstream boxes.

This work also describes 249 novel differentially expressed transcripts that are located ≥ 100 nts upstream and downstream of any annotation. Many of these transcripts correspond well with previous computational predictions of ncRNAs, actually fourteen differentially expressed transcripts overlap with at least two separate previous ncRNA predictions. Finally, we present reverse transcriptase quantitative polymerase chain reaction (RT-qPCR) verification of a set of differentially expressed known genes, ncRNAs and a set of novel MNNG responsive ncRNA candidates. We also experimentally verify the lack of differential expression of the previously annotated adaptive response gene *aidB*. The correspondence between array and RT-qPCR data is high and adds confidence to the array data.

3.3 Paper III

UV-irradiation is known to induce the SOS response in *E. coli.* The SOS response upregulates genes such as *lexA, sulA, recA, uvrA, uvrB, umuC* and *umuD* [31]. In Paper III we present a study using high-density whole-genome tiling arrays to measure a reference and a UV treated *E. coli* transcriptome. Along with the expected upregulation of almost all known SOS-genes we present a variety of other upregulated genes. We also detect 291 novel differentially expressed transcripts, of which many correspond to previous *in-silico* predictions of ncRNAs. The UV induced transcript modulation indicates a possible role in the stress response for these transcripts. Several of the differentially expressed known and novel transcripts were verified by RT-qPCR, showing high correlation to the array data.

To further investigate possible functions of the novel transcripts detected we searched all similarly and differentially non-gene overlapping transcripts for ORFs using GeneMark [148]. The ORF search resulted in six differentially and 17 similarly expressed ORFs, of which 17 are short peptides ([16-60] amino acids) and six are somewhat longer. Furthermore, we applied protein-BLAST to search for possible homologous amino acid sequences, and Jpred [149] and TMHMM [150, 151] to make secondary structure predictions. 16 of the 23 novel peptides had been annotated elsewhere, either as hypothetical proteins or proteins with some suggested function. Seven of the peptides did not have any obvious protein-BLAST hits. One of the upregulated small peptides forms a hydrophobic single transmembrane domain, indicating a possible regulatory function in the inner-membrane. In addition to the detection of this novel stress induced short inner-membrane peptide, we also confirmed transcription of 12 out of 18 small peptides detected in a study by Hemm *et al*. [152]. Two of the 12 are shown as UV-responsive in the present study, in total this adds three small peptides to the list of novel small stress induced peptides presented in their subsequent study on possible functions of small peptides [153].

Finally, we investigated the upstream regions of all significantly modulated genes, and as expected we detected the SOS box promoter sequence as the top motif. Interestingly some of the LexA independent upregulated genes were found to have a consensus sequence in their upstream regions, and we also saw a consensus sequence (different from the SOS box) for a set of upregulated genes, including several known LexA regulated genes.

3.4 Paper IV

In paper IV we present a novel high-density microarray strategy for high-throughput screening for all oncogenic fusion transcripts described in the literature with one single array. The microarray was designed solely for this purpose and gives measurements of chimeric junctions and exon-wise measurements of the individual fusion genes. The array has 68,861 probes and covers 275 pairs of fusion genes. Every fusion gene exon has three probes for detection of broken transcript profiles, and every possible chimeric exon-exon fusion is covered by probes that measure the presence of that specific nucleotide sequence in the sample mRNA. The array design algorithms have been implemented in Python. The analysis of the arrays is made up of three parts and the analysis algorithms have been implemented in Python. Every possible fusion has an upstream gene transcript profile, a downstream gene transcript profile and a set of chimeric exon-exon fusion intensities (Figure 9). The probabilities of broken upstream and downstream gene transcripts are computed (t-test based) and these scores are combined with the chimeric exon-exon intensity measurements. The three scores are differentially weighted based on the fusion gene literature, and the sum of the weighted scores is used to rank all possible fusion gene candidates (Figure 10).

Proof of principle was demonstrated by unguided detection of known fusion genes (such as *TCF3:PBX1*, *ETV6:RUNX1*, and *TMPRSS2:ERG*) from all six positive controls consisting of leukaemia cell lines and prostate cancer biopsies.

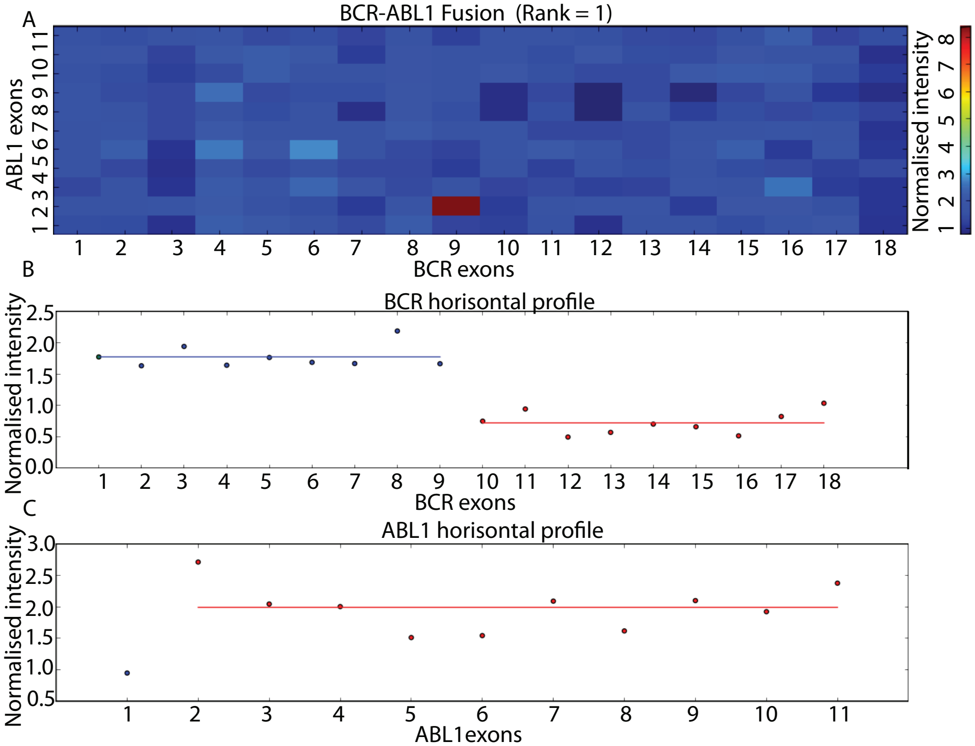

Figure 10. Fusion gene detection. Detection plot for the BCR-ABL1 (here exon 9 to exon 2) fusion found using the 2.1 version array on the K562 cell-line. The combined score of the chimeric probe intensity (red square, part A), the upstream gene (BCR) expression profile (part B) and the downstream gene (ABL1) expression profile (part C) gave this true fusion the top score of about 1x10^{6} possible fusions scored. (Intensity levels are not comparable between part A and parts B and C).

Chapter 4 Discussion

In this study, we have developed custom designs and custom analysis methods for high-density oligonucleotide microarrays to investigate DNA repair systems and to detect DNA damage. The discussion is, for the sake of clarity, divided into three major parts that are followed by a general discussion of future prospects including possible improvements.

4.1 Array design, normalization and analysis

As shown in this thesis, the availability of custom designed HDONAs now enables researchers to design arrays and experiments that suit their specific target optimally. The number of biological/medical questions that may be asked is infinite, the answers are only limited by money, time, availability, sensitivity and specificity. When considering a custom design for a microarray study one must first of all have a defined question or hypothesis that should be answered or tested. As discussed in chapter 1.2.3, a defined aim of for each study is imperative for its success.

The array platform will often be dictated by availability, in-house resources and money, but the choice must not compromise the project. The probe length and probe selection procedure will then again be directed by the actual biological question. If only certain genes are interesting to the study there is no need for a whole genome study, and if introns are interesting, an exon array will be useless. Additionally, a reasonable amount of control probes should be included. Experiment specific custom made control probes are, as shown in Paper I, a very useful quality assessment tool in addition to standard spike-sets.

Regarding normalization there are a variety of widely applied and accepted methods available in freeware and commercial software packages. A widely used measurement of probe quality in oligonucleotide arrays is the level of variation between probes covering the same known transcripts, as in theory they should yield the same intensity level [95]. If normalized data show higher variation over such probes than in the raw data, it follows that the normalization is no improvement to the data. Probe-sequence based normalization may indeed reduce the noise measured along a single transcriptome, but when comparing a treated and a reference transcriptome, this normalization has little or no value. Except if it is of any interest to perform a t-test between all reference and all treated sample probes for example. As described in Paper I; Supplementary File 2, we have tested, investigated and experimented with different normalization procedures before settling for the minimal normalization procedure presented in Paper I. I will here briefly discuss the reasons why established methods turned out to be suboptimal in the bacterial part of this study, and thus present the need for a thorough consideration of which normalization method to choose.

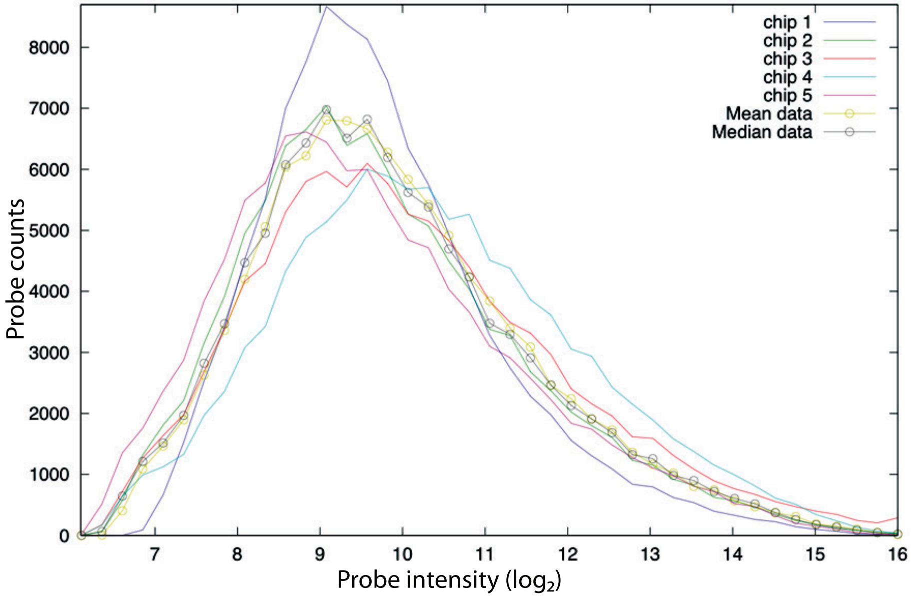

Figure 11. Histogram of raw probe intensities. The figure shows the raw intensity value histograms for all five arrays probing the positive strand of the reference sample in the UV-irradiation study. Deviations from the mean and median intensity lines are clearly seen, and the level of deviation varies between arrays and along the $\log_2$ intensity scale.

The need of a normalization procedure in the first place is easily seen in Figure 11, showing the DMSO treated positive strand data values and the mean and median values. Ideally the histograms of the intensity data from the 5 chips measuring the same data should overlap closely in Figure 11 (unless some noise has been introduced), but the variation appears quite large. Some noise could be explained by the weak patterns seen in the region-averaging heatmaps (Figure 12) (optimally no variation in intensity across the whole chip), but the patterns were far too weak to explain all variations. Thus, indicating additional sources of noise.

The purpose of any normalization has to be to improve signal-to-noise ration and this has to be quantified before any method can be accepted as making a useful contribution to the data. If raw data is better than the normalized, the raw data should be used.
To be able to compare normalization methods we computed the global standard deviation (GSD) of all probes based on the mean level of standard deviation present in all sets of five identical probes across the arrays. Further we assumed that all probes belonging to any well-known annotated gene should return approximately the same intensity level. Based on this assumption we calculated the average transcription level measured by all probes for each gene and subsequently calculated the deviation of all coding probes from their calculated gene-wide transcription mean. The distribution of these deviations would then be expected to form a single normal distribution. Our criteria defined that only methods that reduce the GSD and standard deviation within genes below the level seen in raw data should be candidate normalization procedures. With the microarray techniques of today one should furthermore assume that the raw data is of "good" quality and hence that a normalization method should not drastically affect the data, but rather give a light "tuning".

This allowed us to test and compare the established methods RMA, gcRMA and VSN on the data from our tiling microarrays. RMA and gcRMA both struggle with the probe distribution (dominantly from non-coding rather than coding regions) in our chip design, which was based on our specific interest in short transcripts located in non-coding regions, and imposed by the size of the available chip. RMA gave a small decrease in the GSD, but increased the variation around known coding regions. gcRMA on the other hand gave a large increase of both values, this indicated that both methods were performing worse than raw data. VSN on the other hand gave a fair decrease in variation both on the global level and around coding regions, suggesting nice improvement to the raw data. Closer inspection of the VSN data however revealed a remarkable decrease of the dynamic range of the data resulting in no detection of the response from well-known and tested stress-inducible genes, which is clearly visible in the raw data.

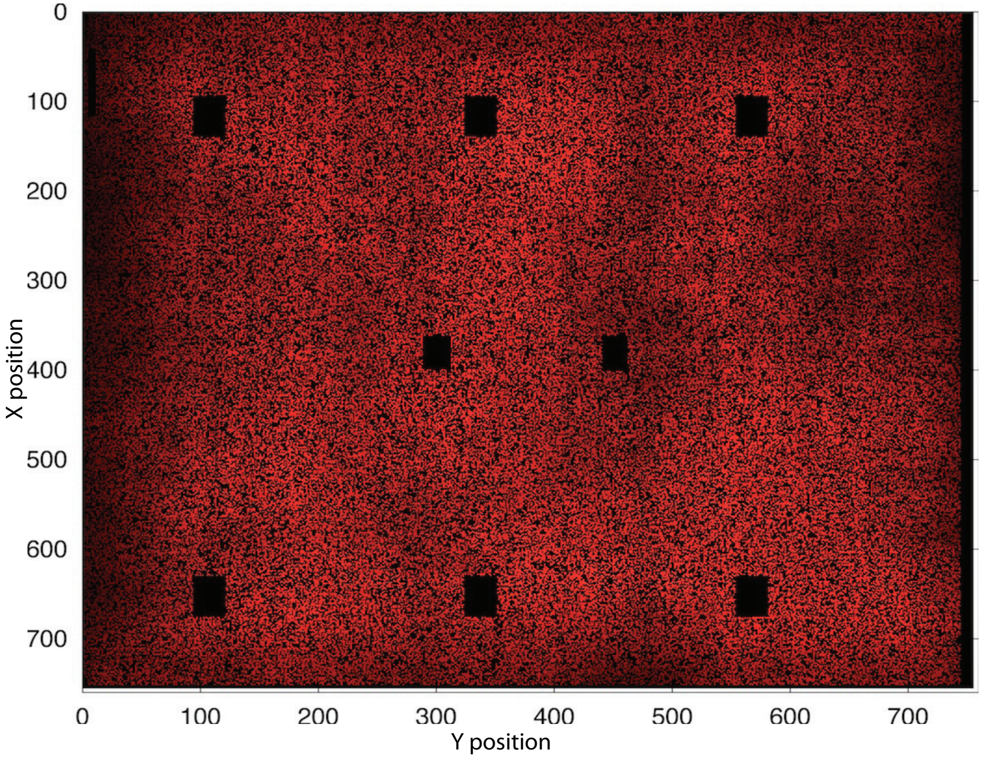

Figure 12. Region averaging heatmap. Heatmap averaged over a 50 point half window width, the colour-scale is from 9 (dark) to 10.5 (white). Averages are over a minimum of 609 local probes (average 2239). The nine black squares are the masked control probes. This heatmap shows the negative strand probes from one of the arrays of MNNG treated E. coli.

In short, we concluded that the probe design, probing pattern (number of probes probing coding versus non-coding regions) and the probe selection procedure have influence on the performance of the normalization procedures. We reason that RMA and gcRMA, both of which were more or less designed for the Affymetrix GeneChip® (probing only DNA assumed to be coding) are confused by the large amount of probes here that probe true intergenic regions and hence result in low signal levels. gcRMA is further tricked in some way by the selection of probes having fairly uniform affinities, and we see that the nucleotide-position weight graphs we presented in Paper I (Supplementary material) deviate from the same plots shown by the gcRMA designers [114]. This might not only originate in the methods applied and the transcriptome investigated, but also from the differences between cDNA and cRNA [102], as the GeneChip® uses complementary RNA, while this study used cDNA in the hybridisation process. Additionally we think that gcRMA is confused by the small number of probes given as negative controls compared to the usual case of PM-MM

probe pairs. On the other hand VSN manipulates the array data too heavily, resulting in overly similar signal levels on a global scale, making only a few very significant transcription changes detectable, and destroying the sensitivity of the technique.
It is, as shown in Paper I, of great importance that the chosen normalization method improves the raw data and does not severely narrow the signal range of the study. Hence, one should not settle too easily for the "usual" normalization method, or the one used in the last project when working with custom designed arrays. One rule of thumb is to control the normalized data against the raw data in terms of global variation and variation within genes, while keeping the sensitivity of the resulting signal range in mind.

Analysis of custom designed microarrays will always be heavily dependent upon the question asked, which again dictates the design process. There will be no single golden path through the analysis when the number of different questions that might be asked is infinite, but some rules must be kept in mind. One cannot say anything certain about any data that show signal intensity levels below background, hence it is very important to have a good background signal estimate. Furthermore, one must not be afraid of throwing away array data with systematic or non-systematic experimental errors that are detected, but cannot be accounted for in the normalization procedure. Another important issue is to ensure statistical significance (and preferably biological significance) for all findings through sufficient biological/technical replicates and probe coverage. In addition, a chosen set of the findings should be experimentally tested to verify the array data, this test set should ideally include results that represent the entire result range and not only the most significant findings.

Regular gene probing microarrays are made to investigate gene expression and disregard all genomic regions not annotated as potential genes. Analysis of these arrays is bound by the annotation that has been used to define the genomic regions that are probed. Until the emergence of the Affymetrix exon arrays (HuEx) the analysis of high-density gene probing arrays focused on the gene-wise expression, with the HuEx arrays one can detect splice variants as well, although with some limitations. The major difference between gene arrays and whole genome tiling arrays is that a large part (design and genome dependent) of the probes are now expected to yield signal intensities at the background noise level, as many probes investigate true non-coding regions (this discussion only considers transcriptome analysis). In the analysis the major differences are the density level of the data, and the ability to detect novel transcripts and define transcript boundaries. This calls for an analysis that is annotation independent, as annotation dependency would not take full account of the available data and possibilities. The most significant difference from an annotation dependent analysis is that the tiling analysis must be able to define transcriptional start and stop sites. There have been a couple of different approaches to solve this problem, and the ultimate goal is to be able to define these sites with one nucleotide resolution, which again demands single nucleotide resolution of the probe tiling.

The two major approaches have been HMMs [143, 154] and sliding windows [9, 140]. An HMM has to be trained, and the training set is based upon some annotated data. This disables identification of novel transcripts with properties that do not correspond well with the training data. On the other hand the sliding window approach decreases the resolution of the tiling array by investigating several probes at a time, i.e. the probes inside the window.

In this study we selected a sliding window method where we also included varying window sizes. Analysis of the data indicated that single probes with consistent significant differential expression (comparing reference and treated data) could be trusted (25 nts), and that the minimum length of a similarly expressed transcript should be 32 nts (i.e. covered by a minimum of 2 probes). Trusted here means that there is sufficient evidence to believe that it is an actual transcript, and not noise. These minimum lengths were based on different observations described in detail in Paper I. In short, one must first assume that probe specific effects behave similarly on the different arrays, and previous work has uncovered that the variation between the intensities of probes investigating the same gene on the same array is larger than between intensities from identical probes on different arrays [124, 125, 133]. This implies that if a short region (one probe) shows consistent significantly differential expression between scaled reference and treated data, there must be a biological difference in the samples that cause this systematic signal intensity difference. In the other case if one single probe yields similar intensity levels in the reference and the stressed dataset, and this signal is above the estimated average background intensity level, it may be simply due to a higher binding affinity for this single probe compared to the average non-expressed probe. Therefore, the expression of this genomic region is likely to become a false positive if not removed. Differentially transcripts can therefore be defined with fewer probes than similarly expressed transcripts. To enable single nucleotide resolution on tiling data, further knowledge must be gained on probe-specific effects so that the analysis algorithms can better adjust for this. Based on the findings by Naef and Magnasco [115] we investigated the possibilities of using a generalised linear model (GLM) to define a probe-specific score system, but we observed only an insignificant increase of the data quality.

In the two *E. coli* studies presented in this thesis, the main goal was detection of novel small differentially regulated transcripts. We claim that the varying window sizes applied in the sliding window method add strength to previous sliding window methods and are preferable to the HMM, although the sliding window disregards all information regarding typical transcript properties. As seen in paper III, we detected a short UV induced peptide with a possible transmembrane function. This peptide had no significant protein-BLAST hits. The ORF of this short peptide would probably not fit into the HMM, as no homologues are known and thereby most likely not present a training set. Additionally, the RT-qPCR verifications and the comparisons of detected transcript boundaries to annotations of well characterised genes also show that the sliding window algorithm has performed well.

4.2 Bacterial DNA repair study

These studies of treated and reference *E. coli* transcriptomes (Paper II and III) have shown a much higher transcriptional activity than previously annotated. Lately several studies on various genomes (including *E. coli*) have revealed similar high transcription levels [7-9]. This brings forth the question of the function of all these novel transcripts. The detected transcripts are likely to have some kind of function in the cell physiology, as transcription is not likely to occur by random. Not only do we detect high levels of transcription, several of the novel transcripts are detected as differentially expressed in the treated bacteria. In Paper II and III we presented two different lists of fourteen differentially regulated ncRNA candidates which also had been previously *in silico* predicted by Sætrom *et al.* [25] and listed as candidates by Hershberg *et al.* [30]. Of these candidates only three are found in both studies (Table 2). This indicates that these three ncRNAs are possible regulators of gene expression in general during stress, while the remainder are damage specific gene regulator candidates. Experimental work including over- and under-expression of these ncRNA candidates must be performed to further unveil their role in the damage response.

Table 2 – Differentially expressed ncRNA candidates reported in both paper II and III.

Start, end and strand as detected in Paper II	Sætrom ID [25]	Previous predictions	Left gene	Right gene	UV fold change ($\log_2$)	MNNG fold change ($\log_2$)
4532242-4532327,+	I253	[28]	*yjhX*	*yjhS*	- 1.5	- 1.0 *
4532456-4532559.+	I179	[23]	*yjhX*	*yjhS*	- 1.5	- 1.0 *
3645892-3645984,-	I238	[22]	*dinQ*	*arsR*	+ 1.1	+ 0.6

Differentially expressed ncRNA candidates reported in both paper II and III, which had also been previously *in silico* predicted by Sætrom *et al.* [25] and also found in the Hershberg *et al.* [30] compilation of *E. coli* ncRNA candidates. The left and right genes are given regardless of strand. * These candidates have also been verified by RT-qPCR in paper II.

In the early analysis of the *E. coli* tiling arrays widespread antisense transcription (transcription on the opposite strand of known genes) was detected. These findings were similar to reports from other tiling array studies from *E. coli* [7] and yeast [8, 155]. However a report from Perocchi *et al.* [12] concludes that the majority of such transcripts are artefacts due to second strand cDNA synthesis during reverse transcription from mRNA to cDNA. They show that most antisense transcripts correlate with the sense strand (2/3 in our studies; data not shown), and that these transcripts are removed by adding actinomycin D (ActD) during the reverse transcription. To further investigate this we performed strand specific RT-qPCR and Northern analysis (data not shown) of the most prominently differentially expressed antisense transcripts detected in the MNNG study; the *ada* transcript, and the antisense *recN* transcript from the UV study. These antisense transcripts were invalidated in both cases, and we selected to disregard all detected antisense transcripts as artefacts. However, some of these transcripts might be true transcripts. These would be the transcripts that are non-correlated to the sense strand.

The majority of the differentially regulated transcripts are downregulated in the treated bacteria. In total we see a ~3% general decrease of the total number of nucleotides detected as transcribed when comparing the two references with the two treated transcriptomes. A general decrease in transcriptional activity is a reasonable defence against stress, as the organism would then be able to concentrate more energy on maintenance and repair. Cell division, for instance, would seem to be a hazardous activity while knowing parts of the genome are damaged, therefore we believe that focus on repair is the reason behind the general down-regulation of the wide range of annotated and un-annotated transcripts we detected.

On the other hand, significant upregulation is seen in both of the treated samples (MNNG and UV). The top list of upregulated transcripts are dominated by known and important response genes (MNNG: *alkA*, *ada* and *alkB,* UV: *sulA, recN, umuC, umuD* and more) in the two different repair pathways invoked by the treatments given. Novel upregulated transcripts are clearly interesting candidates for being novel players within the respective repair pathways, and further experiments must be made to validate them. We detected 132 and 4 upregulated transcripts in the UV and MNNG stressed transcriptomes respectively, in addition to many upregulated genes and predicted ncRNAs, not previously reported as stress induced, in both studies. A reasonable approach to gain more knowledge of the function of these transcripts would be to investigate the phenotypes of knock-outs of the most prominently modulated candidates when inducing DNA damage. And of course, a selection of the most prominently modulated genes, previously not described as MNNG or UV responsive, should be included in the knock out studies. Among others, a very interesting candidate is the *recN* gene, a known SOS inducible gene, which was strongly upregulated in response to MNNG.

One of the most interesting findings made is the high number of differentially expressed transcripts that overlap with *in silico* predicted ncRNAs. In combination with the RT-qPCR verification of a novel short differentially expressed ncRNA candidates in each of the two studies this indicates that several of the small novel and previously in-silico predicted ncRNA candidates are important in DNA damage responses. The *in-silico* predicted ncRNAs detected as induced by UV and/or MNNG also adds to the list of transcripts that should be tested to unveil their function in the respective responses.

Several small peptides have been detected in E. coli[152], and some of them have lately been shown to be stress inducible [153]. These short peptides are often missed by normal peptide searches as they are very short (< 60 amino acids), and may thus be an overlooked part of the bacterial stress response, as argued by Hemm *et al*. [153]. We detected one novel UV-induced transmembrane protein and exhibited two previously detected small inner-membrane as UV-responsive. These findings definitively strengthen the view of many overlooked important small stress responsive peptides, and it shows that it is important to grasp the function of all these small peptides to obtain the complete picture of stress responses.

Upstream regulatory sequences are heavily involved in the gene expression during both adaptive response and SOS response. In Paper II a computational search for the A box and B box (spacer 5-7 nts) was performed on the upstream sequence of all MNNG responsive genes detected by either analysis method. Interestingly several genes, even some downregulated, had perfect A and B box sequences with a 6 nt spacer in their upstream regions. Our findings indicate that the distance from the transcriptional start site to the upstream regulatory sequences is important, and that the upregulated genes have a shorter distance than the downregulated. The unexpectedly upregulated *recN* gene (known SOS response gene, but not adaptive response gene) was found to have a perfect A and B box sequences only one nucleotide upstream of the transcriptional start site detected by the sliding window, which is 115 nts downstream of the annotated start site. As this unexpected upregulation of *recN* was also seen in an MNNG treated *E. coli ada* mutant (data not shown) we find the A and B box sequence unimportant to the regulation of *recN*. Anyhow, a perfect A and B box sequence was detected upstream of genes that did not follow the regulation of the known adaptive response genes, this indicates that the A box and B box alone is insufficient to control the regulation of the downstream genes alone.

The SOS box is central in the regulation of the SOS response genes, and the quality of the SOS box, the position of the SOS box and the number of SOS boxes upstream of the gene influence the induction of the various SOS genes [31]. In Paper III we performed a promoter search of the upstream regions of all differentially regulated transcripts using the MEME tool [156]. The top consensus sequence for upregulated transcripts was as expected the SOS box consensus sequence, but additional interesting motifs were also predicted. No consensus promoter sequence was found for the downregulated transcript. In short the promoter study indicates novel promoter sequences for LexA independent UV responsive genes, and it indicates alternative promoter sequences for some genes already known to have a functional SOS box. The lack of a consensus sequence for the downregulated genes strengthens the idea of a general unspecific downregulation of a variety of genes as part of different stress responses. This also follows from the belief that it is easier to control a general downregulation by lowering the general transcriptional activity, than by activating a sequence specific inhibitor.

4.3 Human fusion gene microarrays

The study presented in this thesis shows that the custom designed fusion gene microarray enables high throughput detection of chimeric fusion genes. In Paper IV array analysis clearly points out six out of six known positive controls as the top ranking fusion gene candidates in the tested cancer samples. As fusion genes are known in several cancer types, while their presence remains unknown in other cancers, high throughput detection of chimeric fusions is needed. In cancer types having known fusion genes, high-throughput detection might become an important means for early and/or differential diagnostics.

Once the pilot study established our high throughput method as successful, thoughts about the next generation array began. Multiplex microarrays with even more features (3 x 720,000 features per chip) were available and knowledge about the probe behaviour on the pilot chip guided an improved design. As we found no special benefit in the small nucleotide up- and down-shift of some of the chimeric fusion probe sequences we designed in the pilot study, we chose to exclude them in the second generation. To gain statistical strength, all chimeric fusion probes covering an exon-junction are present in three copies on the next generation array. Further, all intronic, intron-exon and exon-intron probes were removed from the chip. To gain a more uniform probe affinity distribution of the exon probes, the script constraints placing probes in the start, end and middle of each exon were changed. The updated design algorithm now selects the three best non-overlapping (if the exon length allows) probes per exon, regardless of position within the exon. Since the probe selection for the critical chimeric junction probes is very restricted and T_m is the only score applied, we use only this score in all other probe selections on the array as well. The reasoning behind this is that we prefer a uniform distribution of the probe affinities. A good T_m score will imply a certain GC-content (according to the T_m-calculation procedure), and all probes are designed to be unique, but we have not optimised the exon probes according to all probe-selection factors mentioned in this thesis. By analysis of the pilot array we know that our design is sufficient to detect fusion genes in cancer samples, hence a suboptimal selection of the least critical probes on the array can be justified. If a uniform and optimal probe selection should be made, many of the chimeric junctions would be left unprobed. This would disagree with the basic idea of the fusion gene array to provide a universal tool for detection of oncogenic fusion genes.

In addition to probing known or expected fusion genes, the second generation of the fusion gene microarray contains one single probe for each gene in the entire human genome. The probe selected is a unique sequence with optimal T_m that is present in one of the exons that is part of all, or as many as possible, transcripts annotated by Ensembl (www.ensembl.org). In addition to the fusion gene detection, the inclusion of these additional probes enables a general expression profiling, and the probes also enable investigation and comparisons of how our probes behave compared to measurements of the same samples with commercial exon or gene arrays. Future chip versions will continuously cover more fusion gene candidates and will be used to detect fusion genes in cancer types known to have such genes, and applied to cancer types with no previously known fusion genes.

Lately it has been shown that precursor mRNAs (mRNAs not processed and made ready for translation) from normal un-fused genes can be processed so that the mRNA product becomes similar to mRNA originating from chromosomal rearrangements. Certain types of such trans-spliced mRNAs can exist at a fairly low level in some healthy cells [157]. The revelation of such mRNA products may complicate mRNA screening as a diagnostic tool in certain cancers, but their existence are believed to be anecdotal more than being a common obstacle to fusion gene based diagnostics.

Another technology that was recently applied for fusion gene detection is ultra-high throughput sequencing methods, such as the Illumina Genome analyser II (www.illumina.com), Solid [158] and 454 [159] technologies. This technology has now been successfully applied in fusion gene detection studies [84, 160, 161], and one major upside of this technology is that the great dynamic range helps detecting the true “driver” fusions from the many less interesting fusion candidates detected [160]. Ultra-high throughput sequencing enables fast sequencing of enormous amounts of short nucleotide sequences (Illumina GX IIx: ~ 15 gigabases per run), which can subsequently be assembled into longer pieces. The strategy for fusion gene detection would be to sequence samples from neoplastic tissue to look for possible fusion genes, and the technology may also be combined with microarrays as a first step to narrow the search field. The narrowing function would imply a hybridisation to an array to look for possible candidates, followed by sequencing of all oligonucleotides that hybridized with the chimeric probes on the array. A problem with this strategy today is first of all the cost of running the sequencing and also the assembly of the sequenced oligonucleotides is a very computer intensive problem even though a reference genome is present. Ultra-high sequencing (sometimes combined with microarrays) will probably be the dominating technology detection of novel fusion genes in the near future. Nevertheless, if proven successful as a diagnostic means the fusion-gene microarrays will play an important role in diagnostics and fusion gene screening for many years to come.

4.4 Future perspectives and possible improvements

If this project was to be remade there should be some minor changes made based on the knowledge gained during the project, and based on work by other research groups.
When designing the bacterial tiling arrays in 2005 little information was academically available on the topic of custom design, except for the equally spaced designs. The changes that should be made to the tiling array design include: i) Equal coverage of regions defined as coding (the number of probes would be a function of the length of the gene investigated), ii) At least two different probe selection software packages should be applied to design high quality probes, iii) Random distribution of the control probes instead of placing them in defined rectangles on the array to measure spatial error more accurately and iv) Usage of non-uniform probe lengths will enable more flexibility when trying to obtain uniform binding affinities for all probes.
As the array technology has improved the entire *E. coli* genome can now be tiled with a very high-density on one single array. Anyhow, based on the observations of high variance of probe qualities, even though uniform probe qualities have been strived for, a selective tiling strategy seems preferable. Excepted is the case where total DNA could be hybridised to a set of equal spaced tiling arrays to generate a probe quality reference for data normalization.

As shown in Paper I, and in the discussion, our minimal normalization method performed better than gcRMA, RMA and VSN, which were not originally designed for analysis of tiling arrays. Therefore the novel normalization algorithm should definitely be applied in similar future experiments, along with possible upcoming alternatives. We have shown the sliding window method for segmentation of the expression data into similarly and differentially regulated known and novel transcripts, successful in paper II and III. This was verified by comparison to expected gene modulations, known annotations and RT-qPCR verifications of selected modulated and unmodulated transcripts. The sliding window method is thereby well suited to be applied on similar data in the future. Additionally, other tiling array analysis methods as (MAT [140] and ExpressHMM [143]) should also be taken into consideration.

The UV and MNNG studies would gain even more biological insight if performed as a time point study. Transcriptome mapping at different time points will, as shown by Courcelle *et al.*[6], provide a picture of the timing of transcript modulation in the two bacterial stress responses. Additionally, as discussed, ActD should definitely be included in the sample preparation control to be able to separate true antisense transcripts from reverse transcriptase artefacts. Another, now available, approach is to apply next-generation sequencing on the *E. coli* transcriptome at different time points after stress induction, and to compare this to a sequenced reference transcriptome.

Fusion gene arrays have now been designed in two refined versions, and the design changes have been described in the discussion. The pilot fusion array was proven capable of detecting fusion genes in Leukemia cell-lines with an unguided computational analysis, as shown in Paper IV. The newest version of the arrays has been tested in a blinded experiment with six sarcomas, and we detected four out of six fusion genes correctly (data not shown). There is an ongoing project with an iterative process of refining the design and analysis algorithms, and if sufficient specificity and sensitivity are proven the fusion gene microarrays might become a diagnostic tool in the near future.

This study has utilized only a fraction of the spectre of possible custom designed high-density oligonucleotide microarrays applications. The real challenge is to fully recognise how your project might benefit from custom high-density microarray experiments, and to develop the design and analysis accordingly. Today, next generation sequencing is the big star, but in many cases one should settle for microarrays. The microarray technology is fully capable of solving a variety of questions, of which many have not been stated yet, at a lower cost and without creating a vast overhead of perhaps unnecessary data.

Chapter 5 References

1. Romero PR, Karp PD: **Using functional and organizational information to improve genome-wide computational prediction of transcription units on pathway-genome databases.** *Bioinformatics* 2004, **20:**709-717.
2. Zheng Y, Szustakowski JD, Fortnow L, Roberts RJ, Kasif S: **Computational identification of operons in microbial genomes.** *Genome Res* 2002, **12:**1221-1230.
3. Price MN, Huang KH, Arkin AP, Alm EJ: **Operon formation is driven by co-regulation and not by horizontal gene transfer.** *Genome Res* 2005, **15:**809-819.
4. Lawrence JG, Roth JR: **Selfish operons: horizontal transfer may drive the evolution of gene clusters.** *Genetics* 1996, **143:**1843-1860.
5. Lee RC, Feinbaum RL, Ambros V: **The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14.** *Cell* 1993, **75:**843-854.
6. Courcelle J, Khodursky A, Peter B, Brown PO, Hanawalt PC: **Comparative gene expression profiles following UV exposure in wild-type and SOS-deficient Escherichia coli.** *Genetics* 2001, **158:**41-64.
7. Selinger DW, Cheung KJ, Mei R, Johansson EM, Richmond CS, Blattner FR, Lockhart DJ, Church GM: **RNA expression analysis using a 30 base pair resolution Escherichia coli genome array.** *Nat Biotechnol* 2000, **18:**1262-1268.
8. David L, Huber W, Granovskaia M, Toedling J, Palm CJ, Bofkin L, Jones T, Davis RW, Steinmetz LM: **A high-resolution map of transcription in the yeast genome.** *Proc Natl Acad Sci U S A* 2006, **103:**5320-5325.
9. Kapranov P, Cawley SE, Drenkow J, Bekiranov S, Strausberg RL, Fodor SP, Gingeras TR: **Large-scale transcriptional activity in chromosomes 21 and 22.** *Science* 2002, **296:**916-919.
10. Johnson JM, Edwards S, Shoemaker D, Schadt EE: **Dark matter in the genome: evidence of widespread transcription detected by microarray tiling experiments.** *Trends Genet* 2005, **21:**93-102.
11. Mattick JS: **The functional genomics of noncoding RNA.** *Science* 2005, **309:**1527-1528.
12. Perocchi F, Xu Z, Clauder-Munster S, Steinmetz LM: **Antisense artifacts in transcriptome microarray experiments are resolved by actinomycin D.** *Nucleic Acids Res* 2007, **35:**e128.
13. Eddy SR: **Non-coding RNA genes and the modern RNA world.** *Nat Rev Genet* 2001, **2:**919-929.
14. Szymanski M, Barciszewska MZ, Zywicki M, Barciszewski J: **Noncoding RNA transcripts.** *J Appl Genet* 2003, **44:**1-19.
15. Huttenhofer A, Schattner P, Polacek N: **Non-coding RNAs: hope or hype?** *Trends Genet* 2005, **21:**289-297.
16. Mattick JS, Makunin IV: **Small regulatory RNAs in mammals.** *Hum Mol Genet* 2005, **14 Spec No 1:**R121-132.
17. Mattick JS, Makunin IV: **Non-coding RNA.** *Hum Mol Genet* 2006, **15 Spec No 1:**R17-29.
18. Griffiths-Jones S: **The microRNA Registry.** *Nucleic Acids Res* 2004, **32:**D109-111.
19. Griffiths-Jones S, Grocock RJ, van Dongen S, Bateman A, Enright AJ: **miRBase: microRNA sequences, targets and gene nomenclature.** *Nucleic Acids Res* 2006, **34:**D140-144.

20. Griffiths-Jones S, Saini HK, van Dongen S, Enright AJ: **miRBase: tools for microRNA genomics.** *Nucleic Acids Res* 2008, **36:**D154-158.
21. Gardner PP, Daub J, Tate JG, Nawrocki EP, Kolbe DL, Lindgreen S, Wilkinson AC, Finn RD, Griffiths-Jones S, Eddy SR, Bateman A: **Rfam: updates to the RNA families database.** *Nucleic Acids Res* 2009, **37:**D136-140.
22. Carter RJ, Dubchak I, Holbrook SR: **A computational approach to identify genes for functional RNAs in genomic sequences.** *Nucleic Acids Res* 2001, **29:**3928-3938.
23. Chen S, Lesnik EA, Hall TA, Sampath R, Griffey RH, Ecker DJ, Blyn LB: **A bioinformatics based approach to discover small RNA genes in the Escherichia coli genome.** *Biosystems* 2002, **65:**157-177.
24. Rivas E, Klein RJ, Jones TA, Eddy SR: **Computational identification of noncoding RNAs in E. coli by comparative genomics.** *Curr Biol* 2001, **11:**1369-1373.
25. Saetrom P, Sneve R, Kristiansen KI, Snove O, Jr., Grunfeld T, Rognes T, Seeberg E: **Predicting non-coding RNA genes in Escherichia coli with boosted genetic programming.** *Nucleic Acids Res* 2005, **33:**3263-3270.
26. Blattner FR, Plunkett G, 3rd, Bloch CA, Perna NT, Burland V, Riley M, Collado-Vides J, Glasner JD, Rode CK, Mayhew GF, et al: **The complete genome sequence of Escherichia coli K-12.** *Science* 1997, **277:**1453-1474.
27. Argaman L, Hershberg R, Vogel J, Bejerano G, Wagner EG, Margalit H, Altuvia S: **Novel small RNA-encoding genes in the intergenic regions of Escherichia coli.** *Curr Biol* 2001, **11:**941-950.
28. Tjaden B, Saxena RM, Stolyar S, Haynor DR, Kolker E, Rosenow C: **Transcriptome analysis of Escherichia coli using high-density oligonucleotide probe arrays.** *Nucleic Acids Res* 2002, **30:**3732-3738.
29. Wassarman KM, Repoila F, Rosenow C, Storz G, Gottesman S: **Identification of novel small RNAs using comparative genomics and microarrays.** *Genes Dev* 2001, **15:**1637-1651.
30. Hershberg R, Altuvia S, Margalit H: **A survey of small RNA-encoding genes in Escherichia coli.** *Nucleic Acids Res* 2003, **31:**1813-1820.
31. Friedberg EC, Walker,G.C., Siede,W., Wood,R.D., Schultz,R.A. and Ellenberger,T.: **DNA Repair and Mutagenesis.** *Am Soc Microbiol* 2006:463-612.
32. Radman M: **SOS repair hypothesis: phenomenology of an inducible DNA repair which is accompanied by mutagenesis.** *Basic Life Sci* 1975, **5A:**355-367.
33. Hoeijmakers JH: **DNA damage, aging, and cancer.** *N Engl J Med* 2009, **361:**1475-1485.
34. Eisen JA, Hanawalt PC: **A phylogenomic study of DNA repair genes, proteins, and processes.** *Mutat Res* 1999, **435:**171-213.
35. Fernandez De Henestrosa AR, Ogi T, Aoyagi S, Chafin D, Hayes JJ, Ohmori H, Woodgate R: **Identification of additional genes belonging to the LexA regulon in Escherichia coli.** *Mol Microbiol* 2000, **35:**1560-1572.
36. Little JW, Mount DW: **The SOS regulatory system of Escherichia coli.** *Cell* 1982, **29:**11-22.
37. Michel B: **After 30 years of study, the bacterial SOS response still surprises us.** *PLoS Biol* 2005, **3:**e255.
38. Peterson KR, Ossanna N, Thliveris AT, Ennis DG, Mount DW: **Derepression of specific genes promotes DNA repair and mutagenesis in Escherichia coli.** *J Bacteriol* 1988, **170:**1-4.
39. Sutton MD, Smith BT, Godoy VG, Walker GC: **The SOS response: recent insights into umuDC-dependent mutagenesis and DNA damage tolerance.** *Annu Rev Genet* 2000, **34:**479-497.

40. Walker GC: **Inducible DNA repair systems.** *Annu Rev Biochem* 1985, **54:**425-457.

41. Sassanfar M, Roberts JW: **Nature of the SOS-inducing signal in Escherichia coli. The involvement of DNA replication.** *J Mol Biol* 1990, **212:**79-96.

42. Friedberg EC, Walker,G.C., Siede,W., Wood,R.D., Schultz,R.A. and Ellenberger,T.: *DNA repair and Mutagenesis.* 2005.

43. Walker GC: **Mutagenesis and inducible responses to deoxyribonucleic acid damage in Escherichia coli.** *Microbiol Rev* 1984, **48:**60-93.

44. Lewis LK, Harlow GR, Gregg-Jolly LA, Mount DW: **Identification of high affinity binding sites for LexA which define new DNA damage-inducible genes in Escherichia coli.** *J Mol Biol* 1994, **241:**507-523.

45. Tomasz M: **Mitomycin C: small, fast and deadly (but very selective).** *Chem Biol* 1995, **2:**575-579.

46. Tomasz M, Palom Y: **The mitomycin bioreductive antitumor agents: cross-linking and alkylation of DNA as the molecular basis of their activity.** *Pharmacol Ther* 1997, **76:**73-87.

47. Khil PP, Camerini-Otero RD: **Over 1000 genes are involved in the DNA damage response of Escherichia coli.** *Mol Microbiol* 2002, **44:**89-105.

48. Samson L, Cairns J: **A new pathway for DNA repair in Escherichia coli.** *Nature* 1977, **267:**281-283.

49. Sedgwick B, Lindahl T: **Recent progress on the Ada response for inducible repair of DNA alkylation damage.** *Oncogene* 2002, **21:**8886-8894.

50. Teo I, Sedgwick B, Kilpatrick MW, McCarthy TV, Lindahl T: **The intracellular signal for induction of resistance to alkylating agents in E. coli.** *Cell* 1986, **45:**315-324.

51. Sakumi K, Sekiguchi M: **Regulation of expression of the ada gene controlling the adaptive response. Interactions with the ada promoter of the Ada protein and RNA polymerase.** *J Mol Biol* 1989, **205:**373-385.

52. Landini P, Volkert MR: **Regulatory responses of the adaptive response to alkylation damage: a simple regulon with complex regulatory features.** *J Bacteriol* 2000, **182:**6543-6549.

53. Lindahl T, Sedgwick B, Sekiguchi M, Nakabeppu Y: **Regulation and expression of the adaptive response to alkylating agents.** *Annu Rev Biochem* 1988, **57:**133-157.

54. Falnes PO, Johansen RF, Seeberg E: **AlkB-mediated oxidative demethylation reverses DNA damage in Escherichia coli.** *Nature* 2002, **419:**178-182.

55. Trewick SC, Henshaw TF, Hausinger RP, Lindahl T, Sedgwick B: **Oxidative demethylation by Escherichia coli AlkB directly reverts DNA base damage.** *Nature* 2002, **419:**174-178.

56. Landini P, Hajec LI, Volkert MR: **Structure and transcriptional regulation of the Escherichia coli adaptive response gene aidB.** *J Bacteriol* 1994, **176:**6583-6589.

57. Landini P, Bown JA, Volkert MR, Busby SJ: **Ada protein-RNA polymerase sigma subunit interaction and alpha subunit-promoter DNA interaction are necessary at different steps in transcription initiation at the Escherichia coli Ada and aidB promoters.** *J Biol Chem* 1998, **273:**13307-13312.

58. Landini P, Busby SJ: **The Escherichia coli Ada protein can interact with two distinct determinants in the sigma70 subunit of RNA polymerase according to promoter architecture: identification of the target of Ada activation at the alkA promoter.** *J Bacteriol* 1999, **181:**1524-1529.

59. Akimaru H, Sakumi K, Yoshikai T, Anai M, Sekiguchi M: **Positive and negative regulation of transcription by a cleavage product of Ada protein.** *J Mol Biol* 1990, **216:**261-273.

60. Nakabeppu Y, Sekiguchi M: **Regulatory mechanisms for induction of synthesis of repair enzymes in response to alkylating agents: ada protein acts as a transcriptional regulator.** *Proc Natl Acad Sci U S A* 1986, **83:**6297-6301.

61. He C, Hus JC, Sun LJ, Zhou P, Norman DP, Dotsch V, Wei H, Gross JD, Lane WS, Wagner G, Verdine GL: **A methylation-dependent electrostatic switch controls DNA repair and transcriptional activation by E. coli ada.** *Mol Cell* 2005, **20:**117-129.

62. Berdal KG, Johansen RF, Seeberg E: **Release of normal bases from intact DNA by a native DNA repair enzyme.** *EMBO J* 1998, **17:**363-367.

63. Saget BM, Walker GC: **The Ada protein acts as both a positive and a negative modulator of Escherichia coli's response to methylating agents.** *Proc Natl Acad Sci U S A* 1994, **91:**9730-9734.

64. Futreal PA, Coin L, Marshall M, Down T, Hubbard T, Wooster R, Rahman N, Stratton MR: **A census of human cancer genes.** *Nat Rev Cancer* 2004, **4:**177-183.

65. Novo FJ, de Mendibil IO, Vizmanos JL: **TICdb: a collection of gene-mapped translocation breakpoints in cancer.** *BMC Genomics* 2007, **8:**33.

66. Aman P: **Fusion genes in solid tumors.** *Semin Cancer Biol* 1999, **9:**303-318.

67. Gasparini P, Sozzi G, Pierotti MA: **The role of chromosomal alterations in human cancer development.** *J Cell Biochem* 2007, **102:**320-331.

68. Jack I, Seshadri R, Garson M, Michael P, Callen D, Zola H, Morley A: **RCH-ACV: a lymphoblastic leukemia cell line with chromosome translocation 1;19 and trisomy 8.** *Cancer Genet Cytogenet* 1986, **19:**261-269.

69. Druker BJ, Sawyers CL, Kantarjian H, Resta DJ, Reese SF, Ford JM, Capdeville R, Talpaz M: **Activity of a specific inhibitor of the BCR-ABL tyrosine kinase in the blast crisis of chronic myeloid leukemia and acute lymphoblastic leukemia with the Philadelphia chromosome.** *N Engl J Med* 2001, **344:**1038-1042.

70. Druker BJ, Talpaz M, Resta DJ, Peng B, Buchdunger E, Ford JM, Lydon NB, Kantarjian H, Capdeville R, Ohno-Jones S, Sawyers CL: **Efficacy and safety of a specific inhibitor of the BCR-ABL tyrosine kinase in chronic myeloid leukemia.** *N Engl J Med* 2001, **344:**1031-1037.

71. Nowell PC, Hungerford DA: **Chromosome studies on normal and leukemic human leukocytes.** *J Natl Cancer Inst* 1960, **25:**85-109.

72. Nowell PC, Hungerford DA: **Chromosome studies in human leukemia. II. Chronic granulocytic leukemia.** *J Natl Cancer Inst* 1961, **27:**1013-1035.

73. Rowley JD: **Letter: A new consistent chromosomal abnormality in chronic myelogenous leukaemia identified by quinacrine fluorescence and Giemsa staining.** *Nature* 1973, **243:**290-293.

74. Nowell PC, Hungerford DA: **A minute chromosome in human chronic granulocytic leukemia.** *Science* 1960, **132:**1497.

75. Mitelman F, Johansson B, Mertens F: **The impact of translocations and gene fusions on cancer causation.** *Nat Rev Cancer* 2007, **7:**233-245.

76. Morris DS, Tomlins SA, Montie JE, Chinnaiyan AM: **The discovery and application of gene fusions in prostate cancer.** *BJU Int* 2008, **102:**276-282.

77. Kumar-Sinha C, Tomlins SA, Chinnaiyan AM: **Recurrent gene fusions in prostate cancer.** *Nat Rev Cancer* 2008, **8:**497-511.

78. Lu Q, Nunez E, Lin C, Christensen K, Downs T, Carson DA, Wang-Rodriguez J, Liu YT: **A sensitive array-based assay for identifying multiple TMPRSS2:ERG fusion gene variants.** *Nucleic Acids Res* 2008, **36:**e130.

79. Nasedkina T, Domer P, Zharinov V, Hoberg J, Lysov Y, Mirzabekov A: **Identification of chromosomal translocations in leukemias by hybridization with oligonucleotide microarrays.** *Haematologica* 2002, **87:**363-372.

80. Nasedkina TV, Zharinov VS, Isaeva EA, Mityaeva ON, Yurasov RN, Surzhikov SA, Turigin AY, Rubina AY, Karachunskii AI, Gartenhaus RB, Mirzabekov AD: **Clinical screening of gene rearrangements in childhood leukemia by using a multiplex polymerase chain reaction-microarray approach.** *Clin Cancer Res* 2003, **9:**5620-5629.

81. Shi RZ, Morrissey JM, Rowley JD: **Screening and quantification of multiple chromosome translocations in human leukemia.** *Clin Chem* 2003, **49:**1066-1073.

82. Campbell PJ, Stephens PJ, Pleasance ED, O'Meara S, Li H, Santarius T, Stebbings LA, Leroy C, Edkins S, Hardy C, et al: **Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing.** *Nat Genet* 2008, **40:**722-729.

83. Chen W, Kalscheuer V, Tzschach A, Menzel C, Ullmann R, Schulz MH, Erdogan F, Li N, Kijas Z, Arkesteijn G, et al: **Mapping translocation breakpoints by next-generation sequencing.** *Genome Res* 2008, **18:**1143-1149.

84. Maher CA, Kumar-Sinha C, Cao X, Kalyana-Sundaram S, Han B, Jing X, Sam L, Barrette T, Palanisamy N, Chinnaiyan AM: **Transcriptome sequencing to detect gene fusions in cancer.** *Nature* 2009, **458:**97-101.

85. Ruan Y, Ooi HS, Choo SW, Chiu KP, Zhao XD, Srinivasan KG, Yao F, Choo CY, Liu J, Ariyaratne P, et al: **Fusion transcripts and transcribed retrotransposed loci discovered through comprehensive transcriptome analysis using Paired-End diTags (PETs).** *Genome Res* 2007, **17:**828-838.

86. Stoughton RB: **Applications of DNA microarrays in biology.** *Annu Rev Biochem* 2005, **74:**53-82.

87. MacBeath G, Schreiber SL: **Printing proteins as microarrays for high-throughput function determination.** *Science* 2000, **289:**1760-1763.

88. Lockhart DJ, Winzeler EA: **Genomics, gene expression and DNA arrays.** *Nature* 2000, **405:**827-836.

89. Yazaki J, Gregory BD, Ecker JR: **Mapping the genome landscape using tiling array technology.** *Curr Opin Plant Biol* 2007, **10:**534-542.

90. Liu XS: **Getting started in tiling microarray analysis.** *PLoS Comput Biol* 2007, **3:**1842-1844.

91. You Y, Moreira BG, Behlke MA, Owczarzy R: **Design of LNA probes that improve mismatch discrimination.** *Nucleic Acids Res* 2006, **34:**e60.

92. Affymetrix: **Technical note: GeneChip ® Exon Array Design.**Available: http://www.affymetrix.com/support/technical/technotes/exon_array_design_technote.pdf.

93. Nuwaysir EF, Huang W, Albert TJ, Singh J, Nuwaysir K, Pitas A, Richmond T, Gorski T, Berg JP, Ballin J, et al: **Gene expression analysis using oligonucleotide arrays produced by maskless photolithography.** *Genome Res* 2002, **12:**1749-1755.

94. Singh-Gasson S, Green RD, Yue Y, Nelson C, Blattner F, Sussman MR, Cerrina F: **Maskless fabrication of light-directed oligonucleotide microarrays using a digital micromirror array.** *Nat Biotechnol* 1999, **17:**974-978.

95. Royce TE, Rozowsky JS, Bertone P, Samanta M, Stolc V, Weissman S, Snyder M, Gerstein M: **Issues in the analysis of oligonucleotide tiling microarrays for transcript mapping.** *Trends Genet* 2005, **21:**466-475.

96. Pruitt KD, Tatusova T, Maglott DR: **NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins.** *Nucleic Acids Res* 2007, **35:**D61-65.

97. Affymetrix: **Array Design for the GeneChip® Human Genome U133** Set.Available: http://www.affymetrix.com/support/technical/datasheets/ human_datasheet.pdf.

98. Consortium IHGS: **Finishing the euchromatic sequence of the human genome.** *Nature* 2004, **431:**931-945.

99. Antequera F, Bird A: **Number of CpG islands and genes in human and mouse.** *Proc Natl Acad Sci U S A* 1993, **90:**11995-11999.

100. Fields C, Adams MD, White O, Venter JC: **How many genes in the human genome?** *Nat Genet* 1994, **7:**345-346.

101. Lemoine S, Combes F, Le Crom S: **An evaluation of custom microarray applications: the oligonucleotide design challenge.** *Nucleic Acids Res* 2009, **37:**1726-1739.

102. Royce TE, Rozowsky JS, Gerstein MB: **Assessing the need for sequence-based normalization in tiling microarray experiments.** *Bioinformatics* 2007, **23:**988-997.

103. Bertone P, Trifonov V, Rozowsky JS, Schubert F, Emanuelsson O, Karro J, Kao MY, Snyder M, Gerstein M: **Design optimization methods for genomic DNA tiling arrays.** *Genome Res* 2006, **16:**271-281.

104. Gasieniec L, Li CY, Sant P, Wong PW: **Randomized probe selection algorithm for microarray design.** *J Theor Biol* 2007, **248:**512-521.

105. Li X, He Z, Zhou J: **Selection of optimal oligonucleotide probes for microarrays using multiple criteria, global alignment and parameter estimation.** *Nucleic Acids Res* 2005, **33:**6114-6123.

106. Nielsen HB, Wernersson R, Knudsen S: **Design of oligonucleotides for microarrays and perspectives for design of multi-transcriptome arrays.** *Nucleic Acids Res* 2003, **31:**3491-3496.

107. Rouillard JM, Herbert CJ, Zuker M: **OligoArray: genome-scale oligonucleotide design for microarrays.** *Bioinformatics* 2002, **18:**486-487.

108. Rouillard JM, Zuker M, Gulari E: **OligoArray 2.0: design of oligonucleotide probes for DNA microarrays using a thermodynamic approach.** *Nucleic Acids Res* 2003, **31:**3057-3062.

109. Schliep A, Krause R: **Efficient algorithms for the computational design of optimal tiling arrays.** *IEEE/ACM Trans Comput Biol Bioinform* 2008, **5:**557-567.

110. Wernersson R, Nielsen HB: **OligoWiz 2.0--integrating sequence feature annotation into the design of microarray probes.** *Nucleic Acids Res* 2005, **33:**W611-615.

111. Ryder E, Jackson R, Ferguson-Smith A, Russell S: **MAMMOT--a set of tools for the design, management and visualization of genomic tiling arrays.** *Bioinformatics* 2006, **22:**883-884.

112. Graf S, Nielsen FG, Kurtz S, Huynen MA, Birney E, Stunnenberg H, Flicek P: **Optimized design and assessment of whole genome tiling arrays.** *Bioinformatics* 2007, **23:**i195-204.

113. Eisenstein M: **Microarrays: quality control.** *Nature* 2006, **442:**1067-1070.

114. Wu Z, Irizarry R, Gentlemen R, Martinez-Murillo F, Spencer F: **A model-based background adjustment for oligonucleotide expression arrays.** *J Am Statist Assoc* 2004, **99:**909-917.

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

116. Kaderali L, Schliep A: **Selecting signature oligonucleotides to identify organisms using DNA arrays.** *Bioinformatics* 2002, **18:**1340-1349.
117. Zuker M: **Mfold web server for nucleic acid folding and hybridization prediction.** *Nucleic Acids Res* 2003, **31:**3406-3415.
118. Hofacker IL: **Vienna RNA secondary structure server.** *Nucleic Acids Res* 2003, **31:**3429-3431.
119. Li F, Stormo GD: **Selection of optimal DNA oligos for gene expression arrays.** *Bioinformatics* 2001, **17:**1067-1076.
120. Kreil DP, Russell RR, Russell S: **Microarray oligonucleotide probes.** *Methods Enzymol* 2006, **410:**73-98.
121. Hamming RW: **Error-detecting and error-correcting codes.** *Bell System Technical Journal* 1950, **26:**147-160.
122. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool.** *J Mol Biol* 1990, **215:**403-410.
123. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nucleic Acids Res* 1997, **25:**3389-3402.
124. Naef F, Lim DA, Patil N, Magnasco MO: **From feature to expression: High-density oligonucleotide array analysis revisited.** *Tech Report* 2001, **1:**1-9.
125. Li C, Hung Wong W: **Model-based analysis of oligonucleotide arrays: model validation, design issues and standard error application.** *Genome Biol* 2001, **2:**RESEARCH0032.
126. Seringhaus M, Rozowsky J, Royce T, Nagalakshmi U, Jee J, Snyder M, Gerstein M: **Mismatch oligonucleotides in human and yeast: guidelines for probe design on tiling microarrays.** *BMC Genomics* 2008, **9:**635.
127. Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, Walichiewicz J: **Repbase Update, a database of eukaryotic repetitive elements.** *Cytogenet Genome Res* 2005, **110:**462-467.
128. Wang X, Seed B: **Selection of oligonucleotide probes for protein coding sequences.** *Bioinformatics* 2003, **19:**796-802.
129. Wernersson R, Juncker AS, Nielsen HB: **Probe selection for DNA microarrays using OligoWiz.** *Nat Protoc* 2007, **2:**2677-2691.
130. Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J, et al: **Bioconductor: open software development for computational biology and bioinformatics.** *Genome Biol* 2004, **5:**R80.
131. Lockhart DJ, Dong H, Byrne MC, Follettie MT, Gallo MV, Chee MS, Mittmann M, Wang C, Kobayashi M, Horton H, Brown EL: **Expression monitoring by hybridization to high-density oligonucleotide arrays.** *Nat Biotechnol* 1996, **14:**1675-1680.
132. Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, Speed TP: **Exploration, normalization, and summaries of high density oligonucleotide array probe level data.** *Biostatistics* 2003, **4:**249-264.
133. Li C, Wong WH: **Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection.** *Proc Natl Acad Sci U S A* 2001, **98:**31-36.
134. Hubbell E, Liu WM, Mei R: **Robust estimators for expression analysis.** *Bioinformatics* 2002, **18:**1585-1592.
135. Bolstad BM, Irizarry RA, Astrand M, Speed TP: **A comparison of normalization methods for high density oligonucleotide array data based on variance and bias.** *Bioinformatics* 2003, **19:**185-193.

136. Huber W, von Heydebreck A, Sultmann H, Poustka A, Vingron M: **Variance stabilization applied to microarray data calibration and to the quantification of differential expression.** *Bioinformatics* 2002, **18 Suppl 1:**S96-104.

137. Kauffmann A, Gentleman R, Huber W: **arrayQualityMetrics--a bioconductor package for quality assessment of microarray data.** *Bioinformatics* 2009, **25:**415-416.

138. Huber W, Toedling J, Steinmetz LM: **Transcript mapping with high-density oligonucleotide tiling arrays.** *Bioinformatics* 2006.

139. Picard F, Robin S, Lavielle M, Vaisse C, Daudin JJ: **A statistical approach for array CGH data analysis.** *BMC Bioinformatics* 2005, **6:**27.

140. Johnson WE, Li W, Meyer CA, Gottardo R, Carroll JS, Brown M, Liu XS: **Model-based analysis of tiling-arrays for ChIP-chip.** *Proc Natl Acad Sci U S A* 2006, **103:**12457-12462.

141. Cheng J, Kapranov P, Drenkow J, Dike S, Brubaker S, Patel S, Long J, Stern D, Tammana H, Helt G, et al: **Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution.** *Science* 2005, **308:**1149-1154.

142. Kampa D, Cheng J, Kapranov P, Yamanaka M, Brubaker S, Cawley S, Drenkow J, Piccolboni A, Bekiranov S, Helt G, et al: **Novel RNAs identified from an in-depth analysis of the transcriptome of human chromosomes 21 and 22.** *Genome Res* 2004, **14:**331-342.

143. Munch K, Gardner PP, Arctander P, Krogh A: **A hidden Markov model approach for determining expression from genomic tiling micro arrays.** *BMC Bioinformatics* 2006, **7:**239.

144. Wu Z, Irizarry RA: **Stochastic models inspired by hybridization theory for short oligonucleotide arrays.** *J Comput Biol* 2005, **12:**882-893.

145. Edgar R, Domrachev M, Lash AE: **Gene Expression Omnibus: NCBI gene expression and hybridization array data repository.** *Nucleic Acids Res* 2002, **30:**207-210.

146. Parkinson H, Kapushesky M, Kolesnikov N, Rustici G, Shojatalab M, Abeygunawardena N, Berube H, Dylag M, Emam I, Farne A, et al: **ArrayExpress update--from an archive of functional genomics experiments to the atlas of gene expression.** *Nucleic Acids Res* 2009, **37:**D868-872.

147. Teixeira MR: **Recurrent fusion oncogenes in carcinomas.** *Crit Rev Oncog* 2006, **12:**257-271.

148. Besemer J, Borodovsky M: **GeneMark: web software for gene finding in prokaryotes, eukaryotes and viruses.** *Nucleic Acids Res* 2005, **33:**W451-454.

149. Cole C, Barber JD, Barton GJ: **The Jpred 3 secondary structure prediction server.** *Nucleic Acids Res* 2008, **36:**W197-201.

150. Krogh A, Larsson B, von Heijne G, Sonnhammer EL: **Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes.** *J Mol Biol* 2001, **305:**567-580.

151. Sonnhammer EL, von Heijne G, Krogh A: **A hidden Markov model for predicting transmembrane helices in protein sequences.** *Proc Int Conf Intell Syst Mol Biol* 1998, **6:**175-182.

152. Hemm MR, Paul BJ, Schneider TD, Storz G, Rudd KE: **Small membrane proteins found by comparative genomics and ribosome binding site models.** *Mol Microbiol* 2008, **70:**1487-1501.

153. Hemm MR, Paul BJ, Miranda-Rios J, Zhang A, Soltanzad N, Storz G: **Small stress response proteins in Escherichia coli: proteins missed by classical proteomic studies.** *J Bacteriol* 2010, **192:**46-58.

154. Du J, Rozowsky JS, Korbel JO, Zhang ZD, Royce TE, Schultz MH, Snyder M, Gerstein M: **A supervised hidden markov model framework for efficiently segmenting tiling array data in transcriptional and chIP-chip experiments: systematically incorporating validated biological knowledge.** *Bioinformatics* 2006, **22:**3016-3024.

155. Wilhelm BT, Marguerat S, Watt S, Schubert F, Wood V, Goodhead I, Penkett CJ, Rogers J, Bahler J: **Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution.** *Nature* 2008.

156. Bailey TL, Elkan C: **Fitting a mixture model by expectation maximization to discover motifs in biopolymers.** *Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology* 1994**:**pp.28-36.

157. Li H, Wang J, Mor G, Sklar J: **A neoplastic gene fusion mimics trans-splicing of RNAs in normal human cells.** *Science* 2008, **321:**1357-1361.

158. Shendure J, Porreca GJ, Reppas NB, Lin X, McCutcheon JP, Rosenbaum AM, Wang MD, Zhang K, Mitra RD, Church GM: **Accurate multiplex polony sequencing of an evolved bacterial genome.** *Science* 2005, **309:**1728-1732.

159. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen YJ, Chen Z, et al: **Genome sequencing in microfabricated high-density picolitre reactors.** *Nature* 2005, **437:**376-380.

160. Maher CA, Palanisamy N, Brenner JC, Cao X, Kalyana-Sundaram S, Luo S, Khrebtukova I, Barrette TR, Grasso C, Yu J, et al: **Chimeric transcript discovery by paired-end transcriptome sequencing.** *Proc Natl Acad Sci U S A* 2009, **106:**12353-12358.

161. Wang XS, Prensner JR, Chen G, Cao Q, Han B, Dhanasekaran SM, Ponnala R, Cao X, Varambally S, Thomas DG, et al: **An integrative approach to reveal driver gene fusions from paired-end sequencing data in cancer.** *Nat Biotechnol* 2009, **27:**1005-1011.

Chapter 6 Original papers

Paper I

Custom Design and Analysis of High-Density Oligonucleotide Bacterial Tiling Microarrays

Gard O. S. Thomassen[1,2], **Alexander D. Rowe**[2], **Karin Lagesen**[2], **Jessica M. Lindvall**[3], **Torbjørn Rognes**[2,3]*

1 Centre for Molecular Biology and Neuroscience (CMBN), Institute of Medical Microbiology, University of Oslo, Oslo, Norway, **2** Centre for Molecular Biology and Neuroscience (CMBN), Institute of Medical Microbiology, Oslo University Hospital, Rikshospitalet, Oslo, Norway, **3** Department of Informatics, University of Oslo, Oslo, Norway

Abstract

Background: High-density tiling microarrays are a powerful tool for the characterization of complete genomes. The two major computational challenges associated with custom-made arrays are design and analysis. Firstly, several genome dependent variables, such as the genome's complexity and sequence composition, need to be considered in the design to ensure a high quality microarray. Secondly, since tiling projects today very often exceed the limits of conventional array-experiments, researchers cannot use established computer tools designed for commercial arrays, and instead have to redesign previous methods or create novel tools.

Principal Findings: Here we describe the multiple aspects involved in the design of tiling arrays for transcriptome analysis and detail the normalisation and analysis procedures for such microarrays. We introduce a novel design method to make two 280,000 feature microarrays covering the entire genome of the bacterial species *Escherichia coli* and *Neisseria meningitidis*, respectively, as well as the use of multiple copies of control probe-sets on tiling microarrays. Furthermore, a novel normalisation and background estimation procedure for tiling arrays is presented along with a method for array analysis focused on detection of short transcripts. The design, normalisation and analysis methods have been applied in various experiments and several of the detected novel short transcripts have been biologically confirmed by Northern blot tests.

Conclusions: Tiling-arrays are becoming increasingly applicable in genomic research, but researchers still lack both the tools for custom design of arrays, as well as the systems and procedures for analysis of the vast amount of data resulting from such experiments. We believe that the methods described herein will be a useful contribution and resource for researchers designing and analysing custom tiling arrays for both bacteria and higher organisms.

Citation: Thomassen GO, Rowe AD, Lagesen K, Lindvall JM, Rognes T (2009) Custom Design and Analysis of High-Density Oligonucleotide Bacterial Tiling Microarrays. PLoS ONE 4(6): e5943. doi:10.1371/journal.pone.0005943

Editor: Janet Kelso, Max Planck Institute for Evolutionary Anthropology, Germany

Received February 12, 2009; **Accepted** May 18, 2009; **Published** June 17, 2009

Funding: The project was funded by the Research Council of Norway (FUGE and CoE grants) (www.rcn.no) and the University of Oslo (EMBIO) (www.embio.uio.no). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: torognes@ifi.uio.no

Introduction

The availability of affordable custom-made expression arrays is increasing, and the feature number on oligonucleotide microarrays has increased remarkably during the last few years. Traditional Affymetrix GeneChip arrays focus on probing the coding sequences of known genes, and the probes usually only cover the annotated transcripts' 3′ end, hence much information regarding new transcripts (e.g. microRNAs, anti-sense transcripts and new genes), as well as splice variants of both known and unknown transcripts, are never found [1,2]. Also, recent reports show that annotated genes tend to contain methylation sites with biased distribution towards the 3′ end. This bias in the expressed gene indicate that methylation might interfere with transcription initiation and termination [3,4]. To address this problem, new microarray approaches that enable mapping of the total genome have emerged [5]. Tiling probes on the microarrays is one strategy that has been developed to completely cover areas of the genome [6]. For the majority of completely sequenced genomes no such arrays are currently on the market. Researchers therefore need to design the tiling array themselves. One great advantage of custom made arrays is that they enable total control over chip content with regard to probes for expression measurements, control probes and the distribution of probes over the array.

There are many aspects that have to be taken into consideration in order to achieve high quality data when designing microarrays; including probe density, probe-length, melting temperature, probe placement, strand coverage, cross-hybridization/probe-sequence complexity, probe uniqueness and control probes. The probe-specific aspects mentioned above make up a set of probe-properties. All probes on an array should ideally have approximately the same properties to ensure a constant probability of hybridization [7], the mean value of all these properties can be referred to as the consensus property. The ultimate, but impossible achievement, is to obtain dense coverage of an entire genome by probes with high consensus properties.

Today, several methods for the estimation of background signal level (sum of noise and non-specific hybridization) and data normalisation exist, but these are designed to work with commercial arrays (MAS 5.0, RMA, MBEI, and gcRMA) [8–

11]. Such methods might rely on mismatch-probes [12] or assume that the majority of probes target coding regions, and are therefore often sub-optimal for non-standard custom arrays. Meanwhile, the more generally applicable analysis algorithm MAT (Model-based Analysis of Tiling-arrays) [13], originally designed for ChIP tiling arrays, would be sub-optimal for this study as it applies a 600 bp window which is far larger than the short transcripts targeted here (<60 nts). Other methods for dividing a transcriptome into discrete transcription segments involve different applications of hidden Markov models (HMMs), for instance the supervised Markov model framework of Du *et al.* [14]. One downside of HMM based methods is the need for a training set (generally originating from annotated regions of the genome) which necessarily guides the method towards the recognition of regions which are characteristically similar to the training set. Since a major goal of the approach presented here is to locate novel, short, differentially expressed transcripts in unannotated regions, a standard training set is not optimal. Finally, an HMM method which may successfully work on a single stressed or unstressed dataset will not simultaneously be applicable to data from a direct reference vs stress transcription comparison.

Present analysis methods for microarrays are mainly focused on known coding regions [8,10], and researchers soon run into problems when trying to analyse signals from intergenic regions or un-annotated genomes, because of the difficulty in defining consistently expressed segments of the genome without the aid of an annotation. These problems can be addressed by applying the methods presented here, and the annotation-independent analysis method can be applied to any tiling array project, regardless of whether the investigated regions are coding or non-coding, and without the need of any genomic annotation or training set.

In this manuscript we present a novel design method for tiling arrays, here targeting prokaryotic genomes, but easily applicable to eukaryotic genomes as well. We present a novel normalization method suited to equidistantly or un-equidistantly distributed probes on tiling arrays. Additionally, we show how increased numbers of control probes, including random controls, can be used to assess the levels of non-specific binding and noise, which is always more or less of a problem with microarrays. Finally, we present two different analysis methods for genome-wide tiling array data, of which the latter is independent of annotations and training-sets.

Methods

There are several important considerations regarding microarray design and analysis. Here we present a method for designing tiling arrays and methods for normalisation, background estimates/adjustments and data analysis of tiling experiments. As an initial project, two different prokaryotic genomes are used, the *E. coli* K12 MG1655 genome and *N. meningitidis* MC58 genome, respectively.

Microarray design

Genomic coverage will always be a trade-off between probe-length, genome size and array feature number. The choices made here ensure coverage comparable to regular gene chips of all genes with a known function, as well as a very high coverage of the remaining genome. The arrays used in this project are the 280,000 feature NimbleExpress [15–17] custom arrays provided by Affymetrix, as this was the most reasonable choice when considering the feature number versus production cost. The oligo length was set to 25 nucleotides. The bacterial genomes and annotations of *E. coli* K12 MG1655 [GenBank:NC000913] and *N. meningitidis* MC58 [GenBank:NC003112] used for the probe design were downloaded from the NCBI ftp-site (24th of May 2005). A basic tiling strategy places a probe at every Nth nucleotide (for some N where N<probe-length). Such an approach does not apply any probe-quality measures except for the widely used exclusion of repeat-elements from the target sequences (by using programs such as RepeatMasker [18] or Dust (Tatusov RL and Lipman DJ, unpublished). Use of probes covering repeat elements in the genome should be avoided because of the high risk of cross-hybridisation by similar probes with plural origin, generating meaningless data within these regions. If a more selective tiling approach is used, as described in this paper, it should be possible to choose a set of probes that are more homogeneous, reducing the noise that is otherwise introduced by significant probe-affinity differences.

A limited number of features on the arrays often prohibits a high density tiling strategy from covering the entire genome evenly. As these chips have a 280,000 feature size limit, the decision to split the genomes into two categories was taken; coding and non-coding. All regions annotated with an Open Reading Frame (ORF) having a known function on either strand were defined as coding regions, ORFs separated by less than 25 nucleotides were concatenated. The remaining regions were defined as intergenic (Figure 1). This process of dividing the genome into two categories does not introduce any bias to the applied analysis method, and is solely used for the purpose of probe design as the feature number is limited. For the genomes used in this design, the intergenic regions make up about 10 percent of the *E.coli* and 20 percent of the *N. meningitidis* genome. The terms "coding" and "non-coding" are used here *only* to describe the two categories defined during the design phase.

As *E. coli* and *N. meningitidis* differ in genome-sizes as well as the percentages of non-coding versus protein-coding regions, the probe densities in the coding and non-coding parts in the two genomes were set independently. This density trade-off was dictated by the percentage of coding and non-coding regions along with the total feature number available. The coding regions were covered by 19 and 32 probes per gene in *E. coli* and *N. meningitidis*, respectively. The probe density parameter details can be found in Table S1.

Several probe selection programs are available today, such as OligoArray 2.0 [19], CommOligo [20], OligoWiz 2.0 [21,22] and a web tool from the Gerstein lab (http://tiling.gerstein.org) [23]. OligoArray 2.0 from 2005 was designed for automated selection of short oligonucleotide probe sequences, it requires BLAST and uses MFOLD [24] for thermodynamic secondary structure and probe specificity predictions. CommOligo, accompanied by the Comm Oligo Parameter Estimator, on the other hand addresses whole genome array design or probe design from highly homologous sequences. OligoWiz 2.0, which is applied here, is an oligonucleotide selection software with several user defined parameters; ΔT_m, homology, low-complexity, position and "GATC" only, probe spacing and a maximum and a minimum probe number per sequence. The two algorithms from Bertone *et al.* [23] that form the Gerstein lab web tool concentrate on eukaryotic genome tiling, hence detection of similar probes or sub-sequences between probes is their main focus. Their work emphasise the value of a tiling strategy which optimises the probe affinities rather than a uniform tiling solution, as long as the obtained coverage is sufficient to answer the biological question asked.

As the target organisms here are bacteria, the large-scale eukaryotic similarity problems are excluded (i.e. the Gerstein lab web tool solution) and since the homology problems in bacteria are relatively small, the need for the CommOligo special functionality

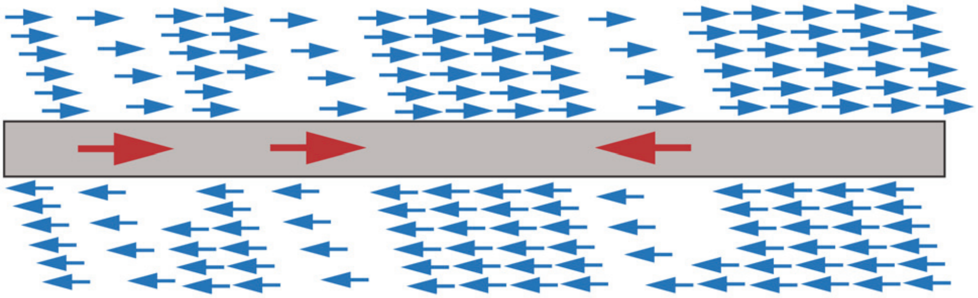

Figure 1. Tiling strategy. The genome was divided into coding and non-coding regions, and the two region types were probed with different densities. The grey bar represent the genome, red arrows represent genes and the blue arrows represent probes. The numbers of probes are not realistic here (see Table S1 for density details).
doi:10.1371/journal.pone.0005943.g001

relating to probe designs for highly homologous sequences is not as critical as for higher species/organisms. To make the initial oligo selection, OligoWiz 2.0 was chosen on the basis of functionality, and the implemented selection algorithms were well suited to the tiling design in these specific projects. Major factors contributing to the selection of OligoWiz 2.0 were the ability to adjust the score parameters to fit the selective tiling design and to apply different probe densities for known ORFs and intergenic regions. In addition, OligoWiz 2.0 is more compatible, since it can be run without the position score-filter, since every part of each probed region is equally important in terms of the detection of novel transcripts. Some recent methods for probe selection are discussed in the "Conclusion and method remarks" section at the end.

After the divison into coding and non-coding regions, the initial selection of probes was made using OligoWiz 2.0 [21,22]. From the resulting set of all possible probes, a subset was chosen by setting the selection parameters in OligoWiz 2.0 (see Tables S2 and S3). When choosing a small minimum inter-probe distance (≪probe-length) for the intergenic regions a "selective tiling" is achieved, i.e. high density, but with high quality probes only (see Table S1 for maximum probe density.) Repeat regions were not removed prior to the probe selection, but were avoided by the combination of OligoWiz 2.0 criteria followed by subsequent probe selection scripts. The main function of these scripts was to remove duplicates, see "probe-uniqueness" below. On the actual array no genomically adjacent probes were closely located on the chip, in order to minimize errors from spatial effects.

To ensure sufficient coverage of both strands, every probe on the array has a complementary probe (if unique) covering the opposite strand. This complementary design also enables all probes to be hybridized with DNA or RNA from both strands. One should keep in mind that hybridization to total DNA can give good probe-quality measurements, which is a useful mean for experimental probe-quality assessment [10]. To achieve this design, OligoWiz 2.0 was applied on one strand and then all probes were complemented to cover the reverse strand. Each complement probe was assigned the same score as its origin. Test-runs with OligoWiz 2.0 proved this approach reliable compared to applying OligoWiz 2.0 on both strands. The complementary probes were then checked for uniqueness (see below), and removed if non-unique (exemplified by the removal of 166 out of 273.414 probes from the original *E. coli* design).

The optimal melting temperature was estimated by OligoWiz 2.0. All regions were considered equally important, as the goal was to map the entire transcriptome. Therefore, the OligoWiz 2.0 position score was left unused. For future designs, variable probe length design (24–26 mers) might be considered in order to achieve a more uniform melting temperature distribution for all probes [25].

Cross-hybridization occurs when a piece of cDNA in the sample binds with, and hence add signal to, a probe that is not 100% complementary. This results in false positives that are almost impossible to identify and remove. This is considered to be a critical problem in array designs [26]. Therefore, the cross-hybridization threshold was the most heavily weighted score. The related sequence–complexity score was also set reasonably high to further decrease the risk of cross-hybridization, see Table S2. One major drawback regarding the probes selected by OligoWiz 2.0 is that the program is able to select identical probes from two different input sequences. The program can thus report two good probes while actually choosing two identical probe sequences. Similar probes on the chip therefore make it impossible to map the actual transcript back to the genome. To avoid this problem of non-unique probes, a computer program removing duplicates from the OligoWiz 2.0 output-files was written and applied (available upon request). The script uses a hash-table with all 13 nucleotide sub-sequences of all probes as keys, if similar keys are detected, all non-overlapping probes with this sub-sequence are removed. This allows a maximum similar continuous stretch of 12 nts. The removal is followed by a control of the regions from which the probes have been taken away. If the removal strongly affects the coverage, another probe with a lower OligoWiz 2.0 score is selected, from the set of all possible probe-sequences generated, to ensure sufficient probe coverage.

A quality assessment of the sample preparation, the hybridization-process and the intensity measurements can be obtained by using control-probes [27]. Control probes are sequences foreign to the target genome designed to assess cross-hybridization and background noise. There are several commercial sets of control probes made to measure the hybridization quality, as well as the RNA sample preparation, labelling and fragmentation process [28]. An improvement of the data quality measurement is sought here by the inclusion of multiple control sets in combination with multiple copies of each control probe. By distributing six copies of these control probes (seven including the hybridization controls, see Figure S1) around the arrays, more measurements can be taken to improve the quality control process. This control probe distribution is used particularly to assess chip-area specific hybridization artefacts. In total there are 4566 control probes distributed over seven separate patches on the chip, see Figure S1 and Table S4 for details. The standard controls used on these arrays are the Affymetrix hybridization control-set, the Affymetrix prokaryotic spike-in set (poly-A) for assessment of the sample preparation and labelling process and the HXB2-yeast spike-set (all three sets described in [28]). Additionally there is a custom made control probe-set consisting of 50 probes having a di-nucleotide composition similar to the *E.coli* specific probes. These custom probes were generated by computing all di-nucleotide frequencies for the target genome probe sequences. Then a probabilistic algorithm producing 25-mers with similar di-nucleotide composition to the target specific probes was implemented. The algorithm outputs the N first probes that differ on at least seven out of 25 nucleotide positions when compared to every *E. coli* specific probe.

The design method presented here was originally made for relatively small genomes (4×10^6). However, the design is easily adapted and scaled up to larger genomes. The target genome size and the feature number available, combined with the biological question asked, will decide whether a tiling approach with equidistantly distributed probes of the entire genome is possible or not. If this approach is considered, Gräf *et al.* [29] as well as Schliep *et al.* [30] recently presented more suitable methods for equidistant probing. The method presented here is on the other hand an elegant alternative for non-equidistant tiling designs. We believe that the division of the target genome into "high" and "less high" interest regions is trivial after the biological question has been stated. OligoWiz 2.0, or another well suited oligo selection tool depending on the biological question (see "Method remarks" section and [31]), should then be applied to design probes suitable for the feature number available and the resolution needed in the genomic region of interest. A probe selection as described here will then select the set of best unique probes for the final design. The control of uniqueness described here can be exchanged for a suffix array approach [32], if the hash-based method raises memory-limit problems. Also, if splice-variant related questions are raised, probes must be designed with probe sequences that represent both the end of exon$_A$ and the start of exon$_B$, as used by Skotheim *et al.* [33]. The control probe design method, including the random negative controls, is well suited to any genome or array size.

Data normalisation

There are a number of accepted normalization techniques that can be applied to microarray data, with varying levels of complexity and transparency. In many experiments, normalisation procedures have proved extremely advantageous; but, as discussed elsewhere [34], in the cases of relatively small genomes such as that of *E. coli* (~4.6 Mbp) and *N. meningitidis* (~2.3 Mbp) the benefits are usually minimal and the application of complex sequence based normalisation routines can in fact confound otherwise clean data (See File S1 for full discussion). It follows, therefore, that it is preferable to minimise normalisation solely to the removal of significant outliers from the data. Ideally, data from multiple arrays show a variance between the log$_2$ intensities of a single probe-set, which is independent of the mean log$_2$ intensity for the given probes for all but the extremes of the data. Plotting the standard deviation versus the intensity for all probe-sets after aligning the data by the mean values of all chips (red circles in Figure 2) allowed a mean level to be calculated for the standard deviation. This was considered as a global measure of the standard deviation (σ_g) between probes in the set of 5 chips (see Figure 2). The global standard deviation was then used to process the data set, by removing the worst-case outliers from the data sets. Here, exactly 46,321 out of 2,733,980 data points were removed from the MNNG experiment. Outlier detection was performed by sorting the five different array signal values from each probe into ascending order and taking the mean of the middle three points as the central value. If either of the remaining probes was found to be more than three global standard deviations ($3\sigma_g$) from the central mean value it was considered to be an outlier with >99% certainty and was therefore discarded. In all other cases, the probe values

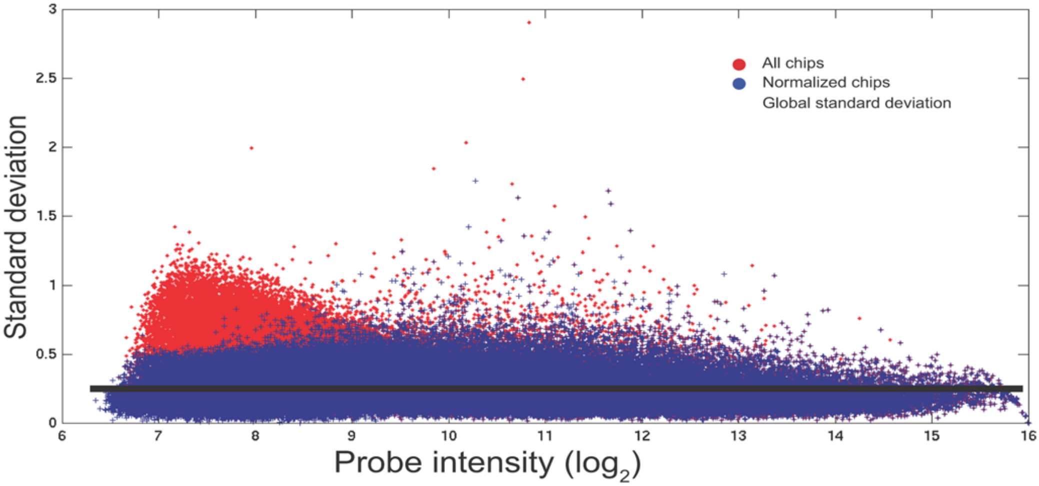

Figure 2. Standard deviation versus intensity for all probe sets. Plotting standard deviation versus intensity for all probes across the 5 arrays (red circles) allowed a mean level of interest to be calculated for the standard deviation. This was considered as a global measure of the standard deviation (σ_g) between probes in the set of 5 arrays. All extreme outliers were removed (see text for details) and the result from this filtering is shown by blue circles.
doi:10.1371/journal.pone.0005943.g002

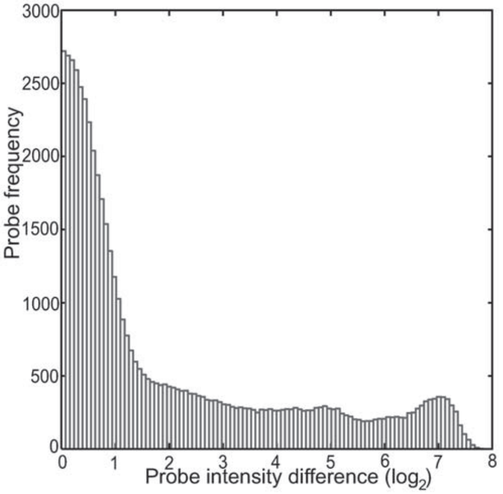

Figure 3. Probewise difference distribution between normalisation methods. Distribution of differences between our normalised data and the gcRMA normalised data is shown. Y-axis represent probe frequencies and the X-axis the absolute value of the difference (log_2).
doi:10.1371/journal.pone.0005943.g003

were retained. The result of this probe outlier filtering is shown as blue circles (Figure 2). This was done before a comparison of relative expression levels was performed on the data.

Given that adjacent probes within a single gene may differ in signal with a standard deviation >1 (on a log2 scale) [35] we have the option to create a very conservative dataset by selectively removing probes using the results of the gcRMA algorithm [11] run on the original raw dataset, in comparison to the dataset returned by the normalization procedure described above. As the original gcRMA algorithm (version 1.0) uses mismatch (MM) probes we applied gcRMA 2.0 (http://rss.acs.unt.edu/Rdoc/library/gcrma/doc/gcrma2.0.pdf). Our custom designed random negative control probes where used in the "bg.adjust.gcrma()" method call, that adjusts for background signals, instead of MM probes. Approximately 10% of all probes (28.594 out of 273.398 in the referred MNNG experiment) can subsequently be discarded where the difference between the gcRMA results and normalized data exceeded the set threshold. The threshold difference level was defined on the basis of the distribution of mean differences between the control and stressed data sets (Figure 3). At extreme difference values, >6 (log_2), there is clearly a secondary peak in the distribution, contributed by data points, which are in strong disagreement with the gcRMA algorithm. In order to minimize data adjustment, while removing the points with strongest disagreement, the threshold difference was set in the minimum region of the distribution between primary and secondary peak.

As previously stated, the large number of control probes assured good assessment of the labelling and hybridization process, respectively. The average signal intensity values of all the spike-in probes for two experiments with a reference and a treated dataset are shown in Figures 4 and 5. The intensities of non-specific probes (HXB2-yeast-, random- and *trpnX*-probes) give an estimate of the level of cross-hybridisation and background noise. An interesting observation is that HXB2-yeast spike set has slightly lower average signals than the custom-made experiment specific control probes, indicating that custom-made, genome specific, negative controls might be better for background signal estimation than these standard spike-sets. The, custom controls show a higher and probably more correct background signal intensity level than the standard sets. The background level was defined as the level at which low level transcription becomes indistinguishable from other background signals. Since low-level transcription predominates along the total length of the genome, this low-level intensity is defined by the peak of a histogram of probe intensities (Figure 6). Below this level it is impossible to separate error from transcription levels. Therefore the background level was set to a log_2 intensity level of 9.0 for the *E. coli* arrays, which is a slightly higher level than the intensities of the custom negative control probes (Figure 4 and 5). All signals below the background noise level are considered as uncertain since they might be a result of noise and/or cross-hybridisation.

Scaling of experimental data should be performed when comparing two datasets where a consistent difference can be detected between control probes designed to give equal signals at a range of different intensities. Here, the average difference showed little variation between probes at differing intensities and therefore the difference was applied as the baseline shift of the reference dataset (Figure 4 and 5).

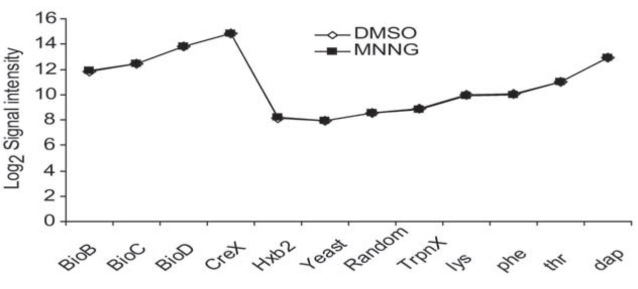

Figure 4. Reference and MNNG treated *E. coli* control probeset average intensities. Average signal intensities for all control probes in reference (Dimethyl Sulfoxide Reductase (DMSO) added only) and treated (N-methyl-N′-nitro-N-nitrosoguanidine (MNNG)) *E. coli*. It is easily seen that the lines overlap very well (sometimes one is hidden by the other), and hence the two experiments can easily be compared with only a minor baseline shift.
doi:10.1371/journal.pone.0005943.g004

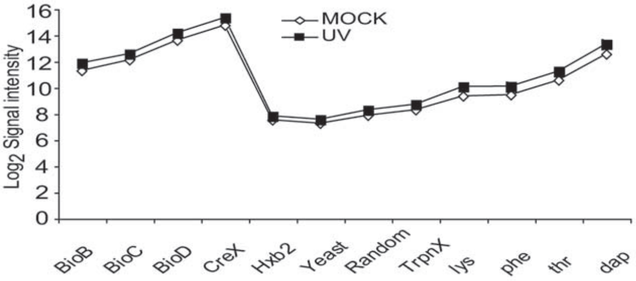

Figure 5. Reference and UV treated control probeset average intensities. Average intensity for all control probes in reference (Mock) and treated (UV irradiated) *E. coli*. Note the consistent difference on all spiked in genes.
doi:10.1371/journal.pone.0005943.g005

Probe-specific effects and estimation of the minimum length of a trustworthy signal

One important question regarding tiling arrays is how long a region is needed to be for its signal to be considered a true signal? A short stretch of the genome with unusual base-composition might result in probes with a very high or very low binding affinity [11]. Probes having low binding affinity might give rise to false negatives, while the ones with high affinity can produce false positives *only* when looking at the expression levels, and false negatives *only* when considering differentially expressed regions. These possible high or low affinity probes could be removed by the application of the gcRMA [11] based method described previously to the raw data. This decreased the number of probes that potentially have biased signal intensities due to highly diverging probe-affinities, although the design process tries to avoid such differences. In the case of differentially expressed regions, probe-artefacts should be equivalent in both conditions and hence regions detected as differentially expressed should be trusted, although there might be uncertainties connected to the absolute signal intensity values, due to the probe affinity problem.

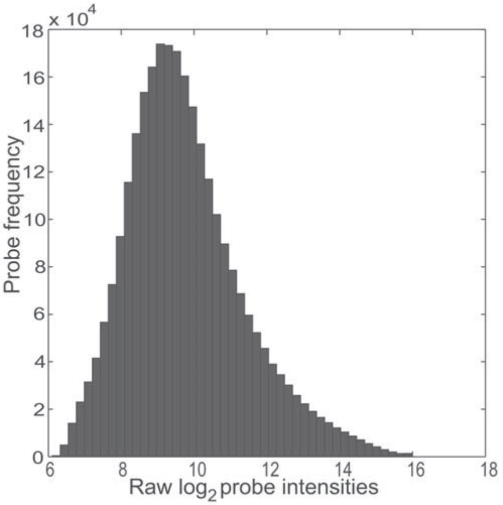

Figure 6. Raw data signal intensity distribution. Signal intensity distribution of all probes for reference (DMSO) and treated (MNNG treated) *E. coli* before data processing. Log_2 signal intensities on the X-axis and probe frequencies on the Y-axis.
doi:10.1371/journal.pone.0005943.g006

Similarly expressed regions and regions detected as present are a different matter. First one must consider the very high probe density, which inevitably will give rise to probes with diverging affinities, even though this has been striven against in the design and normalization process. Thus, with strict selection criteria, regions that are transcribed *in vivo* as long stretches of RNA might appear to be divided into several shorter stretches by the presence of low-affinity probes. On the other hand, short stretches appearing to be expressed in both conditions might be a result of probe artefacts, indicating that they might represent false positives. In addition to this, cDNA production and RNA degradation may, to some degree, represent certain sources of errors. This would likely be due to shortened or missing cDNA pieces from the sample, generating false negatives. Bearing in mind the above observations, differentially expressed regions with a length of only one probe (25 nts) will be considered significant in this study. To define a minimum length threshold for regions detected as present, or similarly expressed, the length distribution of the expressed regions ($\leq$50 nts) with a signal above the background level were plotted in a histogram (Figure 7). A cut-off of minimum 36 nts was set based on this distribution plot combined with the criterion of a separation of the two adjacent probes by at least 10 nts to ensure specific binding of the cDNA to both probes. The minimum spacing criteria of 10 nts is based on the Roche NimbleGen design guide [36]). This exclusion will inevitably exclude true positives, but still it will remove far more false positives and in the end increase the overall data-quality.

Analysis methods

The era of tiling arrays is fairly new [6] and there is not yet one preferred, established and thoroughly tested data analysis method. One problem is that most commercial and free-ware analysis tools are made solely for traditional gene arrays and are therefore not designed to handle the tiling strategy. Therefore the researcher has to create new functions to sub-optimal programs already available, or develop new data analysis tools to fit their specific need.

The percentage of transcribed DNA compared to total DNA is unknown with regards to the bacterial genomes considered in this paper, but is believed to be significantly higher than the percentage annotated today (based on previous tiling projects [37–42]). Nonetheless, tiling arrays are supposed to show far fewer high-intensity signals than normal for gene-targeting arrays

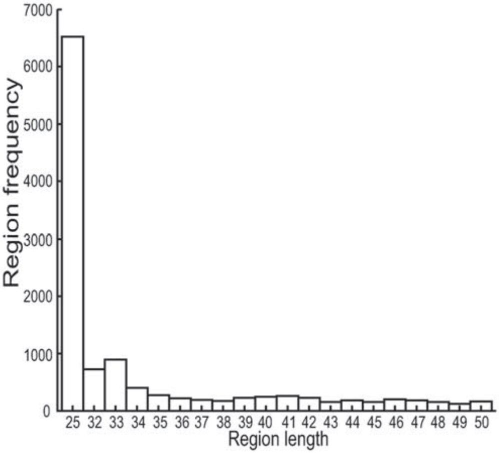

Figure 7. Distribution of short similarly expressed regions. Distribution plot of all similarly expressed regions (< = 50 nts in length) in the DMSO and the MNNG dataset
doi:10.1371/journal.pone.0005943.g007

probing only coding regions. When intergenic regions are probed there are no defined areas in which to look for signals, hence new considerations and adjustments have to be made.

As the goal of this project was divided into transcriptome mapping and detection of differentially as well as similarly expressed genes and transcripts, including novel short transcripts, different analysis methods needed to be developed. First, an annotation guided approach was applied in order to investigate similarly and differentially expressed annotated genes between reference and treated cells. Then, a novel and more complex sliding/expanding window approach, independent of previous annotations, was developed to segment the data and give a comparative analysis of the tiling-results. This approach also allowed transcriptome mapping independent of the comparison between reference and stress datasets.

Annotation based method. In the annotation guided approach all probe signals for each condition of an annotated gene were collected into two groups X_n and Y_m. X_i is probe *i* of a total of *n* probes probing the reference sample, while Y_j is probe *j* of a total of *m* probes probing the treated sample. As a result of the probe-by-probe normalization method, *n* and *m* are not necessarily equal. A two-tailed unpaired t-test was applied to compare the means of the signal values X_n and Y_m. A p-value of 0.05 was chosen as the threshold for rejection of the null hypothesis that the mean values of the two probe sets originate from the same distribution. This threshold equates to a 0.95 confidence of a differential expression between reference and treated data sets. Probe sets conforming to this condition were logged as candidates for differentially expressed genes. Subsequently the absolute average signal intensity difference (fold-change) between all X_n and Y_m probes was calculated. Genes having a probability $> = 0.95$ for differential regulation combined with an absolute fold-change $> = 0.5$ were finally considered as differentially expressed. In cases where the average of X_n or Y_m was below background signal, this average was adjusted to be equal to the background signal before the fold-change calculation was made. This excluded the possibility of false positives in difference calculations occurring due to the presence of erroneous low signals. Although it may be argued that the use of a t-test is suboptimal in cases where many probes are present in an annotated region, the subsequent application of the fold change rule ensures that regions defined as differentially expressed are valid. Meanwhile, when attempting to distinguish differential expression in the shortest fragments, which is our primary interest, application of the t-test as the first rule is the optimal solution.

The p-value returned by each t-test was recorded and subjected to a Bonferroni multiple-testing correction. In practice, these p-values were so small ($\ll 0.05$) that the entire genelist measured as differentially expressed all pass the Bonferroni test. Similar results were shown for the t-tests applied to the top two-hundred regions identified by the sliding window method (below).

Genes where the average of $X_n > =$ background and the average of $Y_m > =$ background and the probability of differential expression or the fold change was below either threshold value were considered similarly expressed. We are aware that a more correct term would be non-significantly differentially expressed but for simplicity similarly expressed is used. Genes having either the average X_n or Y_m below the background level were excluded, as the true signal value is uncertain. Inclusion could lead to false positives, while exclusion gives possible false negatives. The false negatives might be further investigated by looking at the dataset from the plain transcriptome-mapping data (see present/absent regions further below). The background adjustment is, as for the differentially expressed genes, adjusted for the "worst-case" scenario.

Sliding and expanding window method. The normalized data, i.e. after removal of datapoints defined as outliers compared to the gcRMA-normalized data, was sorted according to strand and genomic position.

A sliding and expanding window algorithm was then applied to run along the probes in order to perform calculations on window-sizes of one, three and five probes, for each consecutive probe. For every probe along the genome, a score (0 or 1) was computed for each of the three window sizes. First, an unpaired t-test was applied to calculate the probability of differential regulation between the reference and the stressed samples within the window. Second, the absolute difference of the average signal intensities (fold-change) of all the signals inside the window was computed. Third, the probability and the fold-change were used to define a boolean set of zeroes or ones for differential expression in each window at each probe-position where a 1 indicate that the window has a probability $> = 0.95$ for being differentially regulated, combined with a fold-change $> = 0.5$ ($\log_2$ value). On the other hand a 0 indicates that the probability and/or the fold-change criteria of differential expression are not met. Furthermore, no window could include regulation in both directions, if the window received a score of 1. This sliding and expanding window algorithm resulted in two large score matrices, one for each strand (example in Table 1). A selection algorithm was then applied on these score-matrices. This algorithm searches through the matrices sequentially and selects regions that are differentially regulated. Differentially regulated regions are identified by locating rows in the matrices where all window sizes (1 through 5) had a score of 1 and continues if the next row in the matrix is equal to one of the following [1 X X] or [0 1 1], where X can be either 0 or 1. If a single matrix row of [0 0 1] is located between two rows fulfilling either of the mentioned criteria, this row is also included in the differentially expressed region. In addition, the regulation has to be uniform (either up or down) on all the probes inside a detected region. For all regions detected, the overall t-test score and fold-change value was computed. The final step of the region-selecting

Table 1. Strand-wise score matrix from the Sliding window algorithm.

Probe start	Probe end	Window-size 1	Window-size 3	Window-size 5
49	74	0	1	1
57	*82*	*1*	*1*	*1*
64	*89*	*1*	*1*	*1*
72	*97*	*0*	*0*	*1*
79	*104*	*1*	*1*	*1*
86	*111*	*1*	*0*	*1*
94	*119*	*0*	*0*	*1*
102	*127*	*1*	*1*	*0*
110	*135*	*1*	*0*	*1*
119	*144*	*1*	*1*	*1*
129	*154*	*0*	*1*	*1*
137	*162*	*1*	*1*	*0*
145	170	0	1	0
152	177	1	0	0
165	190	1	1	0
195	220	0	1	0
229	254	1	1	0

A strand-wise score matrix generated by the sliding window algorithm. The example is fictional and illustrates different examples of how the algorithm expands a differentially expressed region. The region from 57 to 162 (*italics*) will be detected as differentially expressed, while the rest are non-differentially expressed regions.
doi:10.1371/journal.pone.0005943.t001

algorithm was to annotate all the detected regions. This was performed by searching for genes overlapping on the same or the opposite strand. If no such overlap was found, the distance to the closest upstream and downstream genes were calculated. For all regions not detected as differentially regulated, another algorithm was applied that located all similarly expressed regions, i.e. regions where both datasets had a signal average >background level but with t-test probability and/or the fold-change level below the threshold of a differentially regulated region, (0.95 and 0.5 respectively). Finally, all the similarly expressed regions were annotated as described earlier. As this method is independent of previous annotations, genes might be reported as partly similarly and partly differentially expressed. Also, there might be some overlap (<25 nt) between regions being differentially and similarly expressed due to the algorithm selection criteria and the overlapping probes (Figure 8).

Transcriptome mapping. An expressed region is a continuous stretch of probes that on average show a signal intensity value above the background noise level. All regions not detected as expressed (scored present) were reported as absent, i.e. missing. This present and absent calculation was done for the samples independently prior to the annotation procedure. Regions excluded by the applied algorithms for the selection of differentially and similarly expressed regions within the confines of the methods described above, can be investigated by comparing the present and absent data for the samples.

Normalisation method comparisons

The issue of normalization is critical in microarray experiments, since the data quality can be highly dependent upon the chosen algorithm. In the case of these custom arrays designed using the OligoWiz 2.0 probe selection program, a visual inspection of the data after application of the gcRMA normalization method [11] indicated data quality degradation. In order to quantify this impression we extracted the 87637 probe values from regions that are annotated and therefore expected to be consistently expressed. The strategy chosen was to use the mean value of all probes within a single similarly-expressed region in order to define the transcription level within this region. This led to the possibility to calculate the deviation – or sequence-dependent bias – of each individual probe from the mean transcription level. The measured biases were, as would reasonably be expected, normally distributed around zero. The quality of any normalizing algorithm was then easily defined by its influence on the normal distribution. A worthwhile normalization method would result in a reduction of the observed variance, while any increase in the variance would imply no improvement to the data quality, thus telling us that the chosen method is wrong for the dataset. Comparison of the variance between probes normalized by our method and the equivalent gcRMA normalized probes showed a variance of 1.17 and 6.84, respectively Therefore, in this case, application of the gcRMA method *severely* degrades the data quality. This in itself is intriguing and leads us to conclude that the design setup and the application of OligoWiz 2.0 (choosing uniform T_m values and GC-content) for probe selection defines a probe set which is incompatible with the gcRMA algorithm. The relative concentration of non-coding compared to coding region probes on our

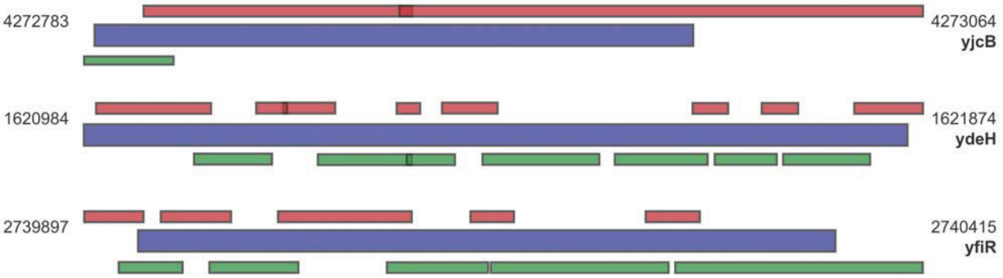

Figure 8. Genes reported as differentially and similarly expressed. A visualisation graph of how several regions can cover one single gene. The blue bars represent genes, differentially expressed regions are represented by the brown horizontal bars above the genes and similarly expressed regions are represented by the green bars below. The numbers indicate genomic start and stop coordinates.
doi:10.1371/journal.pone.0005943.g008

chips will also work against the gcRMA algorithm. Additionally the substitution of MM probes with random control probes, presumably having higher intensities than regular MM probes, will confuse the gcRMA algorithm. The decision was therefore taken not to apply any further normalisation to the data. (See discussion in File S1)

As a further exercise in understanding the sequence dependence of the bias, we compiled our data into histograms of the bias for each nucleotide type at each position along the probe (see Figure S3) and used this to generate a graph of the mean bias for each nucleotide at each position along the probe (Figure 9), which would act as the basis for any sequence dependent bias estimate. This is markedly different to the curve shown by Wu *et al.* and in their discussion of gcRMA [11], further confirming the incompatibility of our probe set with the gcRMA normalisation. Taking this even one step further and applying a generalized linear model (GLM), incorporating single nucleotide positions to the measured biases (using the SAS statistical package) we subsequently produced a set of additive coefficients for individual nucleotide positions (see Figure S2 and Table S5) with which sequence specific probe bias corrections could be made to the data set. Application of this sequence based correction show that a reduction in bias variance from 1.17 to 0.95 was attainable; thus implying that some sequence based normalization is achievable. Due to the time constraints imposed by related biological experiments that were necessary in order to confirm stress responses measured using this microarray data, this fine-tuning normalisation was not applied to the published data sets; however we include the outline of what is possible for the sake of completeness.

To investigate whether our probesets are compatible with standard normalization methods, gcRMA regular RMA and VSN [43] were applied to the data, and a variation comparison study was conducted. Details of these tests are in File S1, but the conclusion showed quite clearly that all three methods made the signal-to-noise ratio worse than unnormalized data. Thus we are vindicated in our choice not to apply standard methods.

Results and Discussion

Different genomes have different nucleotide-compositions, and one should always ensure that regions of special interest on the target genome have a sufficient coverage of probes. This is to ensure that no important genomic region goes un-probed due to some nucleotide composition abnormality.

Here we present a novel method that enables detection novel short (<60 nts) intergenic transcripts by custom made tiling arrays. To ensure sufficient intergenic coverage, overlapping tiling of probes was used in all intergenic regions (as far as the probe quality thresholds allowed). For the *E. coli* genome, a feature number of 386,000 is needed for a complete non-overlapping tiling. Since the array feature number (~280,000) was below 386,000 non-equidistant probing was applied. This probing strategy, which is considered dense, gives a very high intergenic coverage (up to 7 nt resolution), On the other hand, it gives sufficient coverage within regions of known genes. This probe density trade-off is balanced between the feature number and the biological questions asked. With our strict definition of coding and non-coding regions (see above) the applied design solution was considered optimal in terms of the biological aims. During the analysis of the arrays we have reconsidered this and would recommend equidistant coverage of coding regions combined with overlapping tiling of regions of special interest, if the total feature number does not allow dense coverage of the entire genome. In our existing case, the equal probe coverage of each known ORF implies equal data material for each gene to base the statistical analysis on and potentially enables the discovery of more individual gene features [31]. In the suggested case, probes should be tiled as densely as the feature number and the probe quality prediction allows.

Furthermore, by randomly distributing the control probes rather than grouping them in blocks as done here, one might obtain even better assessments of spatial bias. In the end it is the biological question underlying the design that decides where probes are of most efficient use. We still consider "selective tiling" better than a plain

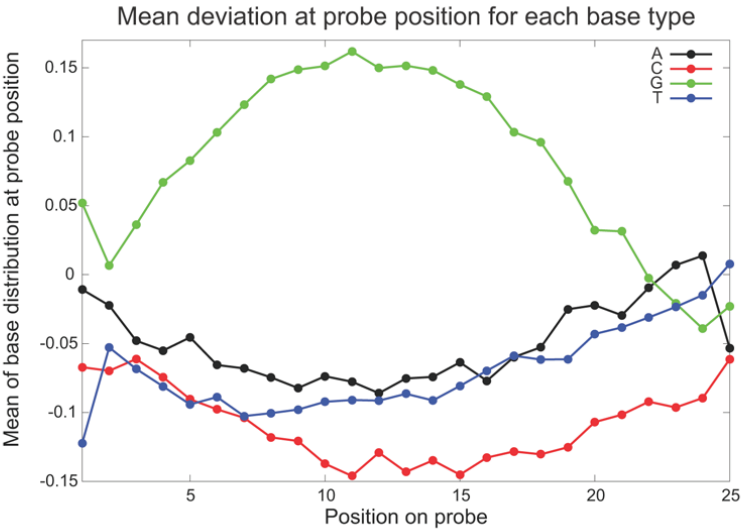

Figure 9. Probe nucleotide composition bias. Mean bias for each nucleotide type at each position along the probe for all probes within known annotated regions of the genome, illustrating the basis of the sequence dependence of individual probe biases.
doi:10.1371/journal.pone.0005943.g009

equidistant tiling approach, as high or low affinity probes would have to be heavily adjusted or thrown away during background predictions or normalisation procedures anyway. Additionally, a somewhat surprising increased transcription detected, and biologically validated, in regions opposite to some known genes indicate that, if the feature number allows, such regions should be prioritized with denser coverage

One may also think of experimenting with even more similar custom made control probes to find the "optimal similarity" when assessing background noise.

It should be noted that although OligoWiz 2.0 strives to obtain uniform probe affinities. Therefore, probe designers should be observant when designing probes for genomes with GC-content far higher or lower than 50%, as OligoWiz 2.0 has no GC-specific scoring filter. The GC-content is closely related to the T_m score and OligoWiz 2.0 would still select probes with uniform binding affinities but the optimal hybridisation temperature would be different and there are possibilities of a decrease or increase of cross-hybridisation due to the GC-content.

Since the actual array design several novel design algorithms and software have been introduced to the research community and are elegantly reviewed and compared in a recent study by Lemoine *et al.* [31]. Lemoine *et al.* show that OligoWiz 2.0 stand out as one of the best choices, as long as the studied organism is found in the OligoWiz 2.0 database. Of the competitors, CommOligo [20] could be considered if the target organism has a non-regular GC content or higher organisms with low-complexity regions. And ArrayOligoSelector [44] or OligoTiler (http://tiling.gersteinlab.org) should be considered when designing tiling arrays with feature numbers sufficient to provide equidistantly spacing of probes combined with sufficient coverage to answer the biological question asked.

Even though the tiling array technology has been around for several years now there is still no "all-in-one" programs and little "how-to" information are available. A few programs/algorithms have been developed for creating oligonucleotide tiling arrays [23,45,46] but none of these have the multi functionality that a chip-designer ideally would hope for. Also, as the interest in specific bacteria differs, one design algorithm might not give good results for two different species without modification.

The annotation based analysis method is a simple and straightforward method for the analysis of the coding parts of tiling experiments. But one should be aware that this method relies on *known* annotations. The sliding window approach, on the other hand, is novel *but* independent of previous annotations. This method is somewhat more complicated and time consuming. The array design, normalisation and data analysis methods presented here have produced a mass of biologically relevant results (manuscript in progress). This shows that the strategy from this work can be implemented on bacterial genomes, and on eukaryotic genomes after applying the minor changes suggested.

Additional information

The array definition and the datasets from the *E. coli* study has been submitted to the Gene Expression Omnibus [47] with accession number GSE 13829 and 13830 (data) and GPL 7714 (array). All computer programs made by the authors have been written in Python and MATLAB and can be obtained on request.

Supporting Information

Figure S1 Control probe distribution
Found at: doi:10.1371/journal.pone.0005943.s001 (0.09 MB PDF)

Figure S2 Nucleotide position bias
Found at: doi:10.1371/journal.pone.0005943.s002 (0.39 MB PDF)

Figure S3 Nucleotide bias histograms
Found at: doi:10.1371/journal.pone.0005943.s003 (0.87 MB PDF)

Table S1 Probe density parameter overview
Found at: doi:10.1371/journal.pone.0005943.s004 (0.05 MB PDF)

Table S2 Initial OligoWiz 2.0 parameter settings
Found at: doi:10.1371/journal.pone.0005943.s005 (0.05 MB PDF)

Table S3 Final OligoWiz 2.0 score-weight parameters
Found at: doi:10.1371/journal.pone.0005943.s006 (0.05 MB PDF)

Table S4 Overview of the control probes
Found at: doi:10.1371/journal.pone.0005943.s007 (0.07 MB PDF)

Table S5 Additive coefficients for sequence specific bias adjustments
Found at: doi:10.1371/journal.pone.0005943.s008 (0.07 MB PDF)

File S1 Supplementary file with a thorough discussion about the quality assessments done on the presented normalisation method and comparisons to gcRMA, RMA and VSN.
Found at: doi:10.1371/journal.pone.0005943.s009 (3.30 MB PDF)

Acknowledgments

We would like to thank our Affymetrix design team for great help. We would also like to thank the microbiologists at CMBN, Rikshospitalet (Oslo, Norway) for supplying information about the bacteria and for helping with all the lab work, Einar Rødland for rigorous mathematical insights and Ophelie Aussedat for helping with the figures.

Author Contributions

Conceived and designed the experiments: GOST JML TR. Performed the experiments: GOST. Analyzed the data: GOST ADR. Contributed reagents/materials/analysis tools: GOST ADR KL. Wrote the paper: GOST ADR JML TR.

References

1. Mockler TC, Chan S, Sundaresan A, Chen H, Jacobsen SE, et al. (2005) Applications of DNA tiling arrays for whole-genome analysis. Genomics 85: 1–15.
2. Bertone P, Gerstein M, Snyder M (2005) Applications of DNA tiling arrays to experimental genome annotation and regulatory pathway discovery. Chromosome Res 13: 259–274.
3. Carninci P, Kasukawa T, Katayama S, Gough J, Frith MC, et al. (2005) The transcriptional landscape of the mammalian genome. Science 309: 1559–1563.
4. Carninci P, Sandelin A, Lenhard B, Katayama S, Shimokawa K, et al. (2006) Genome-wide analysis of mammalian promoter architecture and evolution. Nat Genet 38: 626–635.
5. Johnson JM, Edwards S, Shoemaker D, Schadt EE (2005) Dark matter in the genome: evidence of widespread transcription detected by microarray tiling experiments. Trends Genet 21: 93–102.
6. Selinger DW, Cheung KJ, Mei R, Johansson EM, Richmond CS, et al. (2000) RNA expression analysis using a 30 base pair resolution Escherichia coli genome array. Nat Biotechnol 18: 1262–1268.

7. Mei R, Hubbell E, Bekiranov S, Mittmann M, Christians FC, et al. (2003) Probe selection for high-density oligonucleotide arrays. Proc Natl Acad Sci U S A 100: 11237–11242.
8. Affymetrix (2002) Statistical Algorithms Description Document. Technical report, Affymetrix Inc Discontinued support by Affymetrix: Available: http://www.lrgc.ca/Reading/Affymetrix%20Statistical%20Algorithm%20Description.pdf.
9. Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, et al. (2003) Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics 4: 249–264.
10. Li C, Wong WH (2001) Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. Proc Natl Acad Sci U S A 98: 31–36.
11. Wu Z, Irizarry R, Gentlemen R, Martinez-Murillo F, Spencer F (2004) A model-based background adjustment for oligonucleotide expression arrays. J Am Statist Assoc 99: 909–917.
12. Huber W, Toedling J, Steinmetz LM (2006) Transcript mapping with high-density oligonucleotide tiling arrays. Bioinformatics 22: 1963–1970.
13. Johnson WE, Li W, Meyer CA, Gottardo R, Carroll JS, et al. (2006) Model-based analysis of tiling-arrays for ChIP-chip. Proc Natl Acad Sci U S A 103: 12457–12462.
14. Du J, Rozowsky JS, Korbel JO, Zhang ZD, Royce TE, et al. (2006) A supervised hidden markov model framework for efficiently segmenting tiling array data in transcriptional and chIP-chip experiments: systematically incorporating validated biological knowledge. Bioinformatics 22: 3016–3024.
15. Affymetrix NimbleExpress Array Program. Available: http://www.affymetrix.com/products/arrays/specific/nimble.affx.
16. Albert TJ, Norton J, Ott M, Richmond T, Nuwaysir K, et al. (2003) Light-directed 5′→3′ synthesis of complex oligonucleotide microarrays. Nucleic Acids Res 31: e35.
17. Nuwaysir EF, Huang W, Albert TJ, Singh J, Nuwaysir K, et al. (2002) Gene expression analysis using oligonucleotide arrays produced by maskless photolithography. Genome Res 12: 1749–1755.
18. Smit AFA, Hubley R, Green P (1996–2004) RepeatMasker Open-3.0.
19. Rouillard JM, Zuker M, Gulari E (2003) OligoArray 2.0: design of oligonucleotide probes for DNA microarrays using a thermodynamic approach. Nucleic Acids Res 31: 3057–3062.
20. Li X, He Z, Zhou J (2005) Selection of optimal oligonucleotide probes for microarrays using multiple criteria, global alignment and parameter estimation. Nucleic Acids Res 33: 6114–6123.
21. Nielsen HB, Wernersson R, Knudsen S (2003) Design of oligonucleotides for microarrays and perspectives for design of multi-transcriptome arrays. Nucleic Acids Res 31: 3491–3496.
22. Wernersson R, Nielsen HB (2005) OligoWiz 2.0—integrating sequence feature annotation into the design of microarray probes. Nucleic Acids Res 33: W611–615.
23. Bertone P, Trifonov V, Rozowsky JS, Schubert F, Emanuelsson O, et al. (2006) Design optimization methods for genomic DNA tiling arrays. Genome Res 16: 271–281.
24. Zuker M (2003) Mfold web server for nucleic acid folding and hybridization prediction. Nucleic Acids Res 31: 3406–3415.
25. Wernersson R, Juncker AS, Nielsen HB (2007) Probe selection for DNA microarrays using OligoWiz. Nat Protoc 2: 2677–2691.
26. Wu C, Carta R, Zhang L (2005) Sequence dependence of cross-hybridization on short oligo microarrays. Nucleic Acids Res 33: e84.
27. van Bakel H, Holstege FC (2004) In control: systematic assessment of microarray performance. EMBO Rep 5: 964–969.
28. Affymetrix (2003) GeneChip® CustomExpress^{TM} Array Design Guide. Available: www.affymetrix.com/support/technical/other/custom_design_manual.pdf.
29. Graf S, Nielsen FG, Kurtz S, Huynen MA, Birney E, et al. (2007) Optimized design and assessment of whole genome tiling arrays. Bioinformatics 23: i195–204.
30. Schliep A, Krause R (2008) Efficient algorithms for the computational design of optimal tiling arrays. IEEE/ACM Trans Comput Biol Bioinform 5: 557–567.
31. Lemoine S, Combes F, Le Crom S (2009) An evaluation of custom microarray applications: the oligonucleotide design challenge. Nucleic Acids Res 37: 1726–1739.
32. Manber U, Myers G (1993) Suffix Arrays - a New Method for Online String Searches. SIAM J Comput 22: 935–938.
33. Skotheim RI, Thomassen GO, Eken M, Lind GE, Micci F, et al. (2009) A universal assay for detection of oncogenic fusion transcripts by oligo microarray analysis. Mol Cancer 8: 5.
34. Royce TE, Rozowsky JS, Gerstein MB (2007) Assessing the need for sequence-based normalization in tiling microarray experiments. Bioinformatics 23: 988–997.
35. Royce TE, Rozowsky JS, Bertone P, Samanta M, Stolc V, et al. (2005) Issues in the analysis of oligonucleotide tiling microarrays for transcript mapping. Trends Genet 21: 466–475.
36. Roche NimbleGen Probe Design Fundamentals, Technical Note TN-ARAY0100. Available: http://www.nimblegen.com/products/lit/probe_design_2007_2011_2013.pdf.
37. Bertone P, Stolc V, Royce TE, Rozowsky JS, Urban AE, et al. (2004) Global identification of human transcribed sequences with genome tiling arrays. Science 306: 2242–2246.
38. Tjaden B, Saxena RM, Stolyar S, Haynor DR, Kolker E, et al. (2002) Transcriptome analysis of Escherichia coli using high-density oligonucleotide probe arrays. Nucleic Acids Res 30: 3732–3738.
39. David L, Huber W, Granovskaia M, Toedling J, Palm CJ, et al. (2006) A high-resolution map of transcription in the yeast genome. Proc Natl Acad Sci U S A 103: 5320–5325.
40. Li L, Wang X, Stolc V, Li X, Zhang D, et al. (2006) Genome-wide transcription analyses in rice using tiling microarrays. Nat Genet 38: 124–129.
41. Jiao Y, Jia P, Wang X, Su N, Yu S, et al. (2005) A tiling microarray expression analysis of rice chromosome 4 suggests a chromosome-level regulation of transcription. Plant Cell 17: 1641–1657.
42. Stolc V, Samanta MP, Tongprasit W, Sethi H, Liang S, et al. (2005) Identification of transcribed sequences in Arabidopsis thaliana by using high-resolution genome tiling arrays. Proc Natl Acad Sci U S A 102: 4453–4458.
43. Huber W, von Heydebreck A, Sultmann H, Poustka A, Vingron M (2002) Variance stabilization applied to microarray data calibration and to the quantification of differential expression. Bioinformatics 18 Suppl 1: S96–104.
44. Bozdech Z, Zhu J, Joachimiak MP, Cohen FE, Pulliam B, et al. (2003) Expression profiling of the schizont and trophozoite stages of Plasmodium falciparum with a long-oligonucleotide microarray. Genome Biol 4: R9.
45. Berman P, Bertone P, Dasgupta B, Gerstein M, Kao MY, et al. (2004) Fast optimal genome tiling with applications to microarray design and homology search. J Comput Biol 11: 766–785.
46. Herold KE, Rasooly A (2003) Oligo Design: a computer program for development of probes for oligonucleotide microarrays. Biotechniques 35: 1216–1221.
47. Gene Expression Omnibus GSE13829. Available: http://www.ncbi.nlm.nih.gov/geo/.

6.2 Paper II (Manuscript submitted)

This article is removed.

6.3 Paper III (Manuscript submitted)

This article is removed.

6.3 Paper IV

Molecular Cancer

BioMed Central

Research

Open Access

A universal assay for detection of oncogenic fusion transcripts by oligo microarray analysis

Rolf I Skotheim*[1,2], Gard OS Thomassen[1,2,3], Marthe Eken[1,2,4], Guro E Lind[1,2], Francesca Micci[5], Franclim R Ribeiro[1,2,6], Nuno Cerveira[6], Manuel R Teixeira[2,6], Sverre Heim[5,7], Torbjørn Rognes[3,8] and Ragnhild A Lothe[1,2,4]

Address: [1]Department of Cancer Prevention, Institute for Cancer Research, Norwegian Radium Hospital, Rikshospitalet University Hospital, Oslo, Norway, [2]Centre for Cancer Biomedicine, University of Oslo, Oslo, Norway, [3]Centre for Molecular Biology and Neuroscience, Institute of Medical Microbiology, Rikshospitalet University Hospital, Oslo, Norway, [4]Department of Molecular Biosciences, University of Oslo, Oslo, Norway, [5]Department of Cancer Genetics, Norwegian Radium Hospital, Rikshospitalet University Hospital, Oslo, Norway, [6]Department of Genetics, Portuguese Oncology Institute, Porto, Portugal, [7]Medical Faculty, University of Oslo, Oslo, Norway and [8]Department of Informatics, University of Oslo, Oslo, Norway

Email: Rolf I Skotheim* - rolf.i.skotheim@rr-research.no; Gard OS Thomassen - g.o.s.thomassen@medisin.uio.no; Marthe Eken - marthe.eken@rr-research.no; Guro E Lind - guro.elisabeth.lind@rr-research.no; Francesca Micci - francesm@extern.uio.no; Franclim R Ribeiro - frsr@netcabo.pt; Nuno Cerveira - nscerveira@gmail.com; Manuel R Teixeira - mteixeir@ipoporto.min-saude.pt; Sverre Heim - sverre.heim@medisin.uio.no; Torbjørn Rognes - torbjorn.rognes@medisin.uio.no; Ragnhild A Lothe - rlothe@rr-research.no

* Corresponding author

Published: 19 January 2009

Molecular Cancer 2009, **8**:5 doi:10.1186/1476-4598-8-5

Received: 20 December 2008
Accepted: 19 January 2009

This article is available from: http://www.molecular-cancer.com/content/8/1/5

Abstract

Background: The ability to detect neoplasia-specific fusion genes is important not only in cancer research, but also increasingly in clinical settings to ensure that correct diagnosis is made and the optimal treatment is chosen. However, the available methodologies to detect such fusions all have their distinct short-comings.

Results: We describe a novel oligonucleotide microarray strategy whereby one can screen for all known oncogenic fusion transcripts in a single experiment. To accomplish this, we combine measurements of chimeric transcript junctions with exon-wise measurements of individual fusion partners. To demonstrate the usefulness of the approach, we designed a DNA microarray containing 68,861 oligonucleotide probes that includes oligos covering all combinations of chimeric exon-exon junctions from 275 pairs of fusion genes, as well as sets of oligos internal to all the exons of the fusion partners. Using this array, proof of principle was demonstrated by detection of known fusion genes (such as *TCF3:PBX1*, *ETV6:RUNX1*, and *TMPRSS2:ERG*) from all six positive controls consisting of leukemia cell lines and prostate cancer biopsies.

Conclusion: This new method bears promise of an important complement to currently used diagnostic and research tools for the detection of fusion genes in neoplastic diseases.

Background

Fusion genes created by structural chromosomal rearrangements such as translocations, deletions, and inversions are often the pathogenetically essential feature of cancer genomes. They seem to be particularly characteristic of hematological malignancies and sarcomas, where their identification may be crucial for differential diagnosis and therapeutic decision-making. Fusion genes have so far been found less frequently in the common solid cancers, but recent reports on prostate and lung carcinomas show that fusion transcripts may contribute significantly also to the development of these malignancies [refs. [1-3]; reviewed in [4,5]].

The detection of fusion genes in cancer is laborious and time-consuming and usually includes chromosome banding analysis (karyotyping) followed by fluorescence *in situ* hybridization (FISH) studies and molecular analyses based on the reverse transcriptase polymerase chain reaction (RT-PCR). Karyotyping requires the availability of fresh, vital cells for short-term culturing to obtain metaphase chromosomes, and the success rate of this approach may be particularly low for solid tumors. In addition to taking a lot of time, the method also requires highly trained and experienced personnel to interpret the karyotypes correctly and identify whatever rearrangements may exist. The main advantage of the approach is that it is global in nature; it screens without prejudice for all rearrangements at the chromosomal resolution level. FISH with locus-specific probes and RT-PCR, on the other hand, are precise and highly specific methods used for the analysis of one or a few candidate fusion genes at predefined breakpoints; the approach is therefore dependent on prior knowledge of the suspected diagnosis. The specificity of these methods at the same time highlights their main limitation; they have no screening ability.

Recent developments of high-throughput sequencing technologies enable genome-wide identification of novel fusion transcripts at an unprecedented level of resolution [6-9], but these technologies are as yet limited by the number of samples that can be analyzed within a reasonable time-frame and at an acceptable cost. A few studies have utilized oligo microarrays targeting junction sequences to detect fusion transcripts [10-13]. They have then relied on preceding amplification of a small selection of fusion transcripts by RT-PCR, thus limiting the coverage offered by these approaches to a predefined set of fusion junction sequences.

In this report, we present a new oligo microarray-based approach for simultaneous analysis of all known or predicted fusion gene variants, with all possible chimeric exon-exon junction combinations. The analysis can be performed in a single experiment and does not include prior sequence-specific amplification.

Methods

Cell lines and biopsies

To test our novel method for fusion gene detection, we selected four prostate cancer samples (fresh frozen tissue obtained from prostatectomy specimens of four independent patients) and two leukemic cell lines, all known to harbor a specific fusion gene. The cell lines, RCH-ACV [14] and REH [15,16], are of human B-cell precursor leukemia origin and were provided by Dr. Edith Rian.

Preparation of cDNA for microarray analysis and RT-PCR

Total RNA was isolated using the Trizol reagent (Life Technologies, Rockville, MD, USA), and the RNA quality was evaluated by use of the Agilent 2100 Bioanalyzer (Agilent Technologies, Palo Alto, CA, USA). To enrich for messenger RNA, we used the RiboMinus kit (Invitrogen, Carlsbad, CA, USA) which subtracts ribosomal RNA from total RNA. To ensure detection of fusion junctions far away from the poly-A tail, the first strand cDNA was prepared by random priming to avoid the 3' end bias introduced by oligo-dT labeling. Double stranded cDNA was labeled and hybridized onto the oligo microarrays.

Microarray design

We set up a database with a broad coverage of the reported fusion genes in cancer (351 to date), including information on which of the fusion partners are up- and downstream in the majority of the resulting fusion transcripts. See Additional file 1 for the identities and orientation of the 275 fusion genes included in the pilot microarray design. We used public genome sequence information from Biomart to extract the exon sequences of all listed transcript variants [17].

A script was written in the programming language Python for design of the oligos. For genes that constitute the 5' portion of fusion genes, we used the 3' end-sequences of the exons when constructing chimeric fusion junction oligos. For genes that are the 3' portion of fusion genes, we used the 5' start-sequences of the exons. Thus, for each fusion gene, we joined and listed all combinations of end-sequences and start-sequences. These chimeric sequences served as input for the design of chimeric fusion junction oligos, enabling detection of any breakpoint combination in the fusion genes. Chimeric oligos were constructed targeting all possible combinations of chimeric exon junctions between the up- and downstream partners of 275 known fusion genes. For a set of fusion genes, including the ones known to be present in the control samples, we extended the design to include four replicates of each of the exon-exon junctions, as well as altogether four extra control oligos for each exon-exon junction (oligos up- and down-shifted by two nucleotides as compared to the standard ones). Furthermore, a series of intragenic oligos were designed for measurements of longitudinal profiles of each of the fusion gene partners of altogether 115 genes, including all the positive control fusion genes. These were oligos targeting the start, mid, and end part of all exons and all introns, as well as oligos targeting the exon-exon, exon-intron, and intron-exon junctions. The exon-intron junctions and intron-exon junctions are also included among the single-gene oligos, as the pre-mRNA processing machinery may alter the splicing pattern following removal or introduction of cis-acting splicing regulatory sequences.

The constructed microarray included a design with 68,861 oligos, including 59,381 chimeric oligos (of which 55,482 were unique), which were synthesized onto custom-produced NimbleGen microarray slides (Roche NimbleGen, Inc., Madison, WI, USA). The chimeric oligos were designed to optimize for similar melting temperatures on each side of the junctions, thus reducing half-binder effects.

Two versions of the microarray were designed, differing as to the probe lengths. The set of shorter oligos, with lengths ranging from 34 to 40-mers, had a Tm optimum of 72°C. The set of longer oligos, with lengths ranging from 44 to 50-mers, had a Tm optimum of 75°C. All samples, except the REH cell line, were hybridized onto the short-oligo microarray, whereas the RCH-ACV and REH cell lines were hybridized onto the long-oligo microarray. The cell line RCH-ACV was analyzed by both microarray designs, and data from its positive control gene, *TCF3:PBX1*, demonstrated best performance of the short oligos due to substantial half-binder signals with the longer oligos (data not shown).

Because of the relatively short length of the sequences on each side of the junction, the binding may be sensitive to single nucleotide polymorphisms (SNPs). Thus, at known SNP positions, we created extra sets of probes, accounting for each of the SNP variants.

Data preprocessing and annotation

Data preprocessed by NimbleGen were further normalized by dividing all individual probe intensity values for each of the samples by the median of the three leukemia cell lines. We normalized based on these three samples (instead of all samples) because when the majority of the samples contain the same fusion gene and breakpoint (*TMPRSS2:ERG*, *e1:e4*), normalizing on all samples would level out the appearance of this fusion event in the dataset.

All oligonucleotide probes were mapped to their one or two respective genomic loci. For each locus, the Ensembl identifiers for exon (ENSE), transcript (ENST), and gene (ENSG) identities were used.

Raw and processed data were deposited to the Gene Expression Omnibus public repository for microarray data [accession number GSE14435] according to the MIAME, minimum information about a microarray experiment, recommendations for recording and reporting microarray-based gene expression data [18].

Automated scoring algorithm

Downstream fusion partners will generally have higher expression values for exons downstream of the fusion breakpoint. For each exon-exon junction of downstream fusion partner genes, two probabilities were calculated. One probability was based on a t-test for whether values from all upstream and all downstream exons are likely to belong to different populations. A second probability was based on a t-test for whether the values from the immediate up- and downstream exons are likely to belong to different populations.

A fusion score was calculated as the product of the normalized expression value for the chimeric oligo and the probabilities of the exon-exon junction of the corresponding position in the downstream fusion partner being a breakpoint in the longitudinal profile [Fusion score = Chimeric junction score * P(B-gene transcript) * P(B-gene exon)].

To keep the values within scale, the following thresholds were applied: when the normalized values for chimeric oligos were larger than 5, they were set to 5 (approximately 5 per 10,000 values). Similarly, when probabilities for a breakpoint in the longitudinal profiles were < 0.10, they were set to 0.10. When the values from the downstream exons were lower than the values from the upstream exons, the probability was set to 0.10 as well.

Experimental validation of fusion transcript breakpoints

We used RT-PCR followed by DNA sequencing to validate the actual fusion junctions in the positive control fusion genes. The following primers were applied: *TCF3:PBX1*: *TCF3*, exon 15, forward, 5'-CACCCTCCCTGACCTGTCT-3', and *PBX1*, exon 3, reverse, 5'-TGCTCCACTGAGTT-GTCTGAA-3'; yielding a chimeric fusion product of 218 basepairs. *ETV6:RUNX1*: *ETV6*, exon 5, forward, 5'-CACTCCGTGGATTTCAAACA-3', and *RUNX1*, exon 2, reverse, 5'-CGTGGACGTCTCTAGAAGGA-3'; yielding a chimeric fusion product of 204 basepairs. *TMPRSS2:ERG* [as published in ref. [19]]: *TMPRSS2*, exon 1, forward, 5'-TAGGCGCGAGCTAAGCAGGAG-3', and *ERG*, exon 6, reverse, 5'-CTGCCGCACATGGTCTGTAC-3'; yielding a chimeric fusion product of 597 basepairs. The PCR products were separated by gel electrophoresis in a 2% agarose gel. For all fusion genes, DNA was isolated from the appropriate PCR bands (MiniElute Gel Extraction kit, Qiagen Co., Valencia, CA, USA) and sequenced in both directions using the same primers as for the RT-PCR (ABI Prism 3730; Applied Biosystems, Foster City, CA, USA).

Cytogenetics

Cell cultures from the leukemia cell lines were harvested for chromosome banding analysis. Chromosome preparations were made and G-banded using trypsin (DIFCO Laboratories, Detroit, MI, USA) and Leishman staining (BDH, Poole, England). For metaphase FISH, commercially available probes for the *TCF3:PBX1* (*TCF3* FISH DNA probe, split signal, DAKO Denmark A/S, Glostrup, Denmark) and *ETV6:RUNX1* (dual color, Dual Fusion Translocation Probe Set; Vysis, Abbott Laboratories, Abbott Park, IL, USA) fusion genes were used. The denaturation and hybridization conditions as well as the subsequent detection procedures were in accordance with the manufacturers' protocols. Two hundred successive, whole, and single nuclei were examined through a Zeiss fluorescence microscope (Zeiss Axioplan, Oberkochen, Germany) for each FISH experiment.

Results

We have developed a novel strategy for the detection of oncogenic fusion transcripts enabling simultaneous analysis of all known or predicted fusion gene variants, with all possible chimeric exon-exon junction combinations targeting each possible fusion gene junction on the processed mRNA level (Figure 1). We combine the use of chimeric oligos, spanning the two potential fusion gene partners, with the use of single-gene oligos that provide measurements along the length of each individual partner.

We analyzed cDNA from a set of six positive control samples with known presence of one fusion gene in each. This included two leukemia cell lines, RCH-ACV and REH, known to carry the *TCF3:PBX1* and *ETV6:RUNX1* fusion genes, respectively, and four prostate cancer samples positive for the *TMPRSS2:ERG* fusion gene.

To combine the information from the chimeric junction measurements with that of the longitudinal intragenic profiles, a fusion score was calculated for all fusion transcripts and their respective breakpoints (details in Materials and Methods). This enabled an objective and automated evaluation of the presence of fusion genes, and the fusion score was calculated for 10,297 possible fusion events. The positive control fusion transcripts, with their correct breakpoint positions, was ranked as the number one hit in four out of the five samples run on the short-oligo microarray (Figures 2A and 3A), thus validating the concept. For prostate cancer sample P140, the expected *TMPRSS2* exon 1:*ERG* exon 4 fusion gene was assigned a fusion score rank of 95 within the 10,297 fusion breakpoints (and number one within the 154 measured junctions of *TMPRSS2:ERG*). When dissecting the values behind the fusion score for this positive control, we see that the intensity of the chimeric oligo was particularly low. This is also in compliance with RT-PCR results from the prostate cancer samples, demonstrating that this sample had a low expression level of the fusion gene as compared to the other samples (data not shown).

To evaluate the top fusion score hits and positive control fusion genes further, we visualized them via two independent paths, using either the chimeric probe set (Figures 2B and 3B) or the longitudinal intragenic probe set (Figures 2C and 3C). The positive control fusion genes were clearly visualized for all six analyzed samples.

Discussion

A novel microarray-based strategy is presented to screen for all known oncogenic fusion transcripts in a given sample, combining measurements of chimeric transcript junctions with exon-wise measurements of individual fusion partners. This provides a viable alternative to the existing cytogenetic and PCR-based methods for fusion gene detection, as it enables an objective and automated genome-wide analysis in which all known as well as predicted fusion genes are assessed without requiring any *a priori* knowledge as to the likelihood of the clinical or genetic diagnosis. Furthermore, the precise mapping information on the fusion breakpoint is given within every positive hit. Finally, the method is carried out in a single experiment and does not include prior sequence-specific amplification.

Because fusion breakpoints mainly map to intronic sequences, the resulting fusion transcripts will, after pre-mRNA processing, consist of whole exonic building blocks. In fact, more than 90% of the mapped fusion breakpoints are located in intronic sequences [20]. Thus, independently of the intra-intronic location of the breakpoints, a detection of all exon-exon junctions between two fusion gene partners would in principle provide specific detection of fusion transcripts.

To our knowledge, this is the first time chimeric oligos targeting fusion gene junctions have been used in combination with measurements of longitudinal profiles of the individual fusion partners. Furthermore, the earlier publications on fusion gene measurements by oligo microarrays have not attempted to be genome-wide, restricting their use to either a few pre-defined fusion junctions and fusion genes [10-13] or to the exclusive use of intragenic oligos [21]. Our pilot experiment alone included 68,861 oligos, and the current version of the NimbleGen micro-

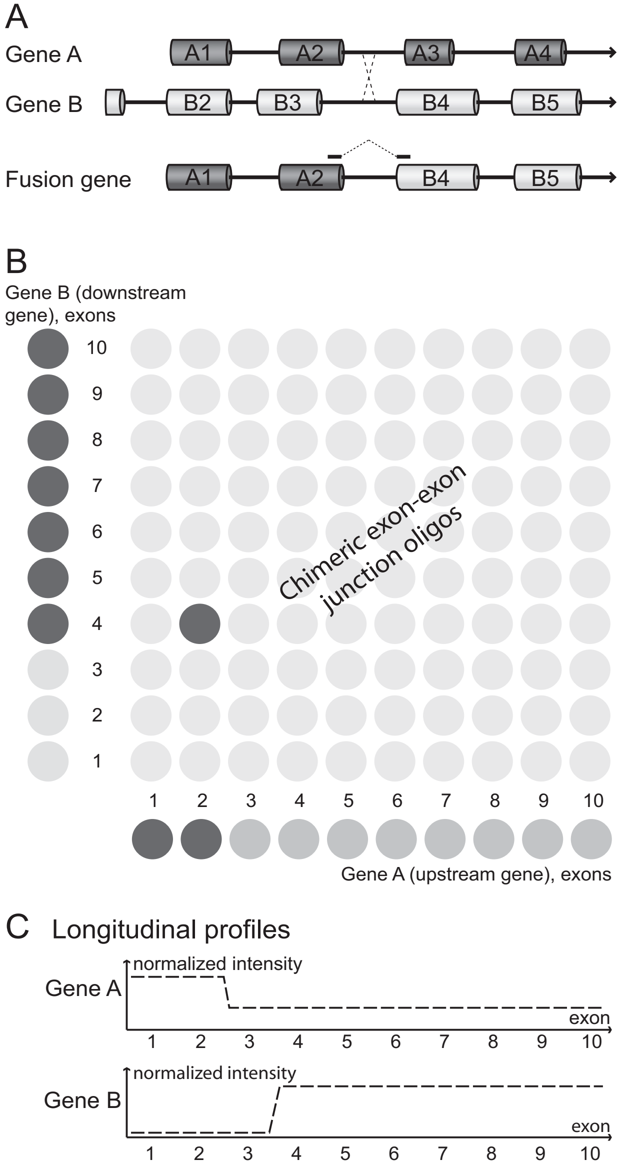

Figure 1

Microarray data for a positive fusion gene hit. (A) This theoretical example of a fusion gene has a crossing-over event between sequences in intron 2 of gene A and intron 3 of gene B. (B) If the genes A and B both have 10 exons, the microarray will contain 10 × 10 = 100 oligos to cover all chimeric exon-to-exon junction combinations for this particular fusion gene. The A2-B4 oligo detects the fusion transcript from part (A). (C) In true fusion events, the longitudinal profiles generated from intragenic oligos targeting each exon and exon-to-exon junction will provide independent confirmation.

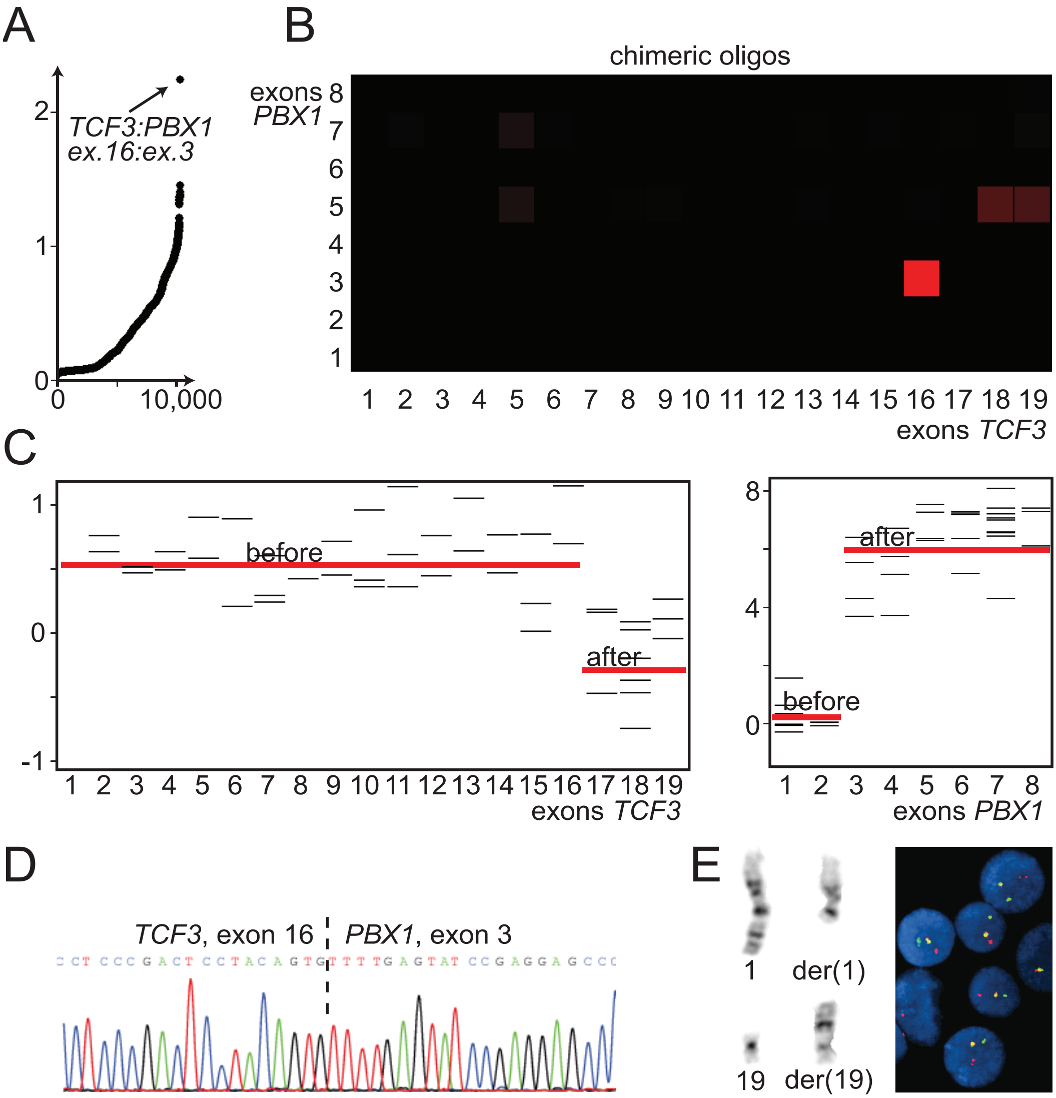

Figure 2
***TCF3:PBX1* in a leukemia cell line**. (A) The highest ranking fusion score (y-axis) among 10,297 chimeric combinations (ranked along the x-axis) indicated a fusion event between exons 16 and 3 of *TCF3* and *PBX1*, respectively. (B) Fusion map of each chimeric exon-exon junction of *TCF3* and *PBX1*. Intensities of red indicate the relative values of the medians for the four replicate oligos for each chimeric exon-exon junction, and the square with strongest intensity indicates the correct fusion breakpoint. (C) Measurements from intragenic oligos (intra-exon probes) for each of the two fusion partners are indicative of the same fusion breakpoints as seen from the chimeric oligos. (D) The exact fusion breakpoint between *TCF3* and *PBX1* was confirmed by cDNA sequencing. (E) Chromosome banding and fluorescence *in situ* hybridization analyses of the same cell line demonstrated rearranged chromosomes from the translocation t(1;19)(q23;p13), which implicates the loci of *TCF3* and *PBX1*.

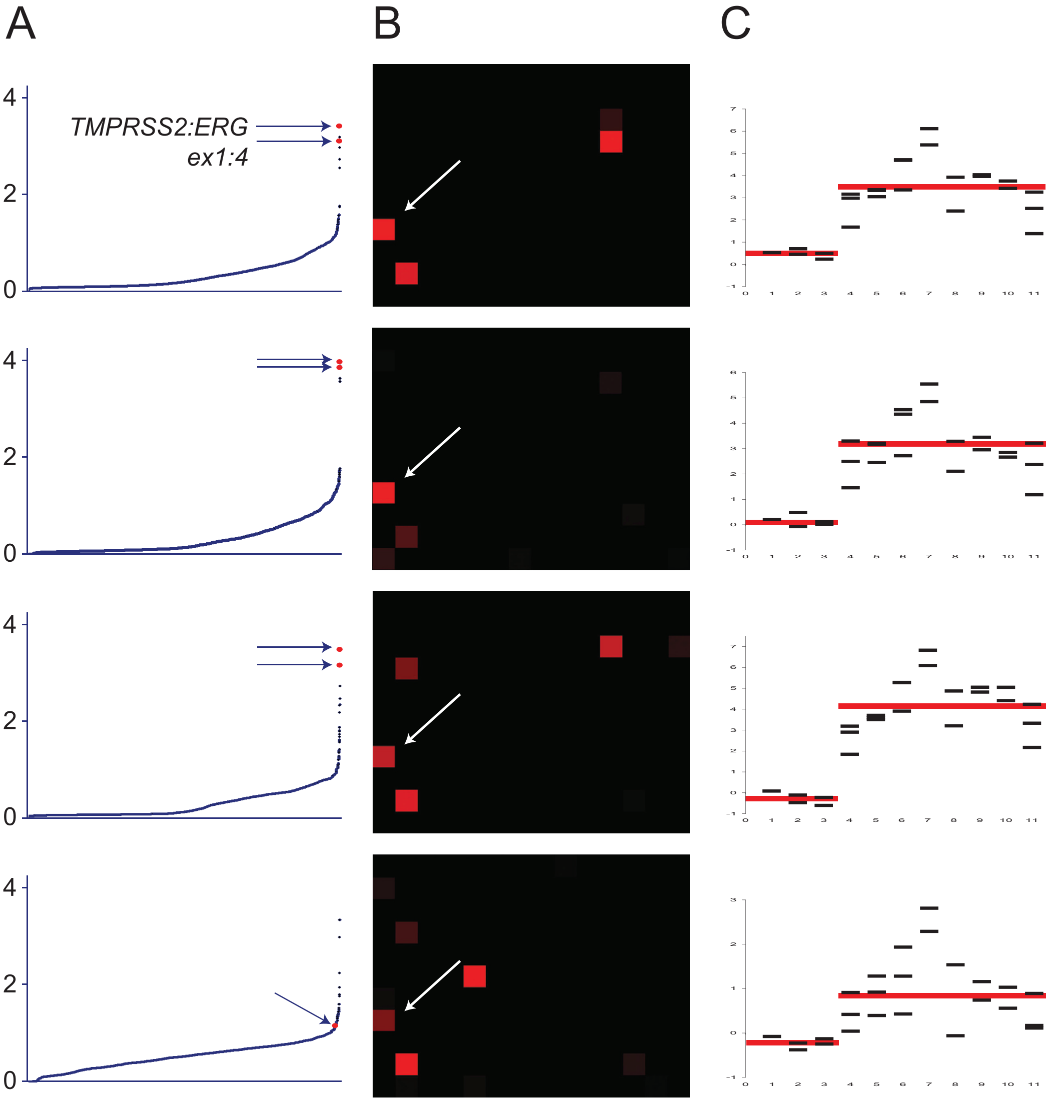

Figure 3

Fusion gene plots for four individual prostate cancer samples with the same fusion event. The four samples all had fusion transcripts with junctions between *TMPRSS2* exon 1 and *ERG* exon 4. (A) Chimeric sequences plotted with increasing fusion scores. For the three first samples, the chimera of *TMPRSS2:ERG* exon1:4 had the highest ranking out of the 10,297 tested combinations. (B) Fusion map of each chimeric exon-exon junction of *TMPRSS2* and *ERG*. Intensities of red indicate the relative values of the medians for the four replicate oligos for each chimeric exon-exon junction and the white arrows point to the correct fusion breakpoints. (C) Measurements from intragenic oligos (intra-exon probes) for *ERG* demonstrate a shift in intensities between exons three and four.

array platform enables analysis of up to 2.1 million oligos on a single microarray slide. Thus, scaling up to include all known fusion genes, as well as sets of novel candidate fusion genes detected by high-throughput sequencing strategies, can easily be achieved with the same resolution level as the genes included in our pilot run.

Next-generation sequencing approaches are beginning to provide numerous new pairs of fusion genes in individual biological samples [6-9]. However, this methodology is not feasible for screening purposes on large clinical sample series. The current microarray-based approach is suitable for assessing whether members of this growing set of novel fusion transcripts (alongside with the already known fusion genes) are indeed pathogenetic players in the various subgroups of cancer.

The reported fusion gene detection platform can be used irrespective of the tumor type in question. Detection of certain fusion genes has direct diagnostic implications in many leukemias and sarcomas, whereas other fusion genes are more promiscuous and can be found in several different cancer types. An example of the latter is the karyotypically cryptic translocation t(12;15)(p13;q25), resulting in the *ETV6:NTRK3* fusion gene, which occurs in histologically and developmentally completely disparate tumors such as kidney and breast tumors, infantile fibrosarcoma, and acute myeloid leukemia [22].

Conclusion

We have developed a novel high throughput method for detection of fusion genes with potentially significant applications in cancer diagnostics. Also, for research applications, there is a clear potential for detection of putative fusion genes and discovery of already known fusion genes in new cancer types.

List of abbreviations

FISH: fluorescence *in situ* hybridization; RT-PCR: reverse transcriptase polymerase chain reaction.

Competing interests

A patent application has been filed for the fusion gene microarray methodology.

Authors' contributions

RIS coordinated the project and drafted the manuscript. GOST designed and programmed algorithms for oligo design and data analysis. ME, GEL, FM, FRR, and NC performed laboratory analyses. MRT, SH, TR, and RAL supervised parts of the project, and contributed to design of the study and discussion of results. All authors contributed to the writing, and have read and approved the final manuscript.

Additional material

Additional File 1

Supplementary table. Fusion genes included in the microarray design.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1476-4598-8-5-S1.pdf]

Acknowledgements

We are grateful for the technical assistance from Stine Aske Danielsen, Mette Eknæs, and Lisbeth Haugom. This work was supported by grants from the Norwegian Cancer Society; the Legacy of Harald Andersen; and the Research Council of Norway.

References

1. Mitelman F, Johansson B, Mertens F: **Fusion genes and rearranged genes as a linear function of chromosome aberrations in cancer.** *Nat Genet* 2004, **36:**331-334.
2. Tomlins SA, Rhodes DR, Perner S, Dhanasekaran SM, Mehra R, Sun XW, Varambally S, Cao X, Tchinda J, Kuefer R, Lee C, Montie JE, Shah RB, Pienta KJ, Rubin MA, Chinnaiyan AM: **Recurrent fusion of *TMPRSS2* and *ETS* transcription factor genes in prostate cancer.** *Science* 2005, **310:**644-648.
3. Helgeson BE, Tomlins SA, Shah N, Laxman B, Cao Q, Prensner JR, Cao X, Singla N, Montie JE, Varambally S, Mehra R, Chinnaiyan AM: **Characterization of TMPRSS2:ETV5 and SLC45A3:ETV5 gene fusions in prostate cancer.** *Cancer Res* 2008, **68:**73-80.
4. Teixeira MR: **Recurrent fusion oncogenes in carcinomas.** *Crit Rev Oncog* 2006, **12:**257-271.
5. Kumar-Sinha C, Tomlins SA, Chinnaiyan AM: **Recurrent gene fusions in prostate cancer.** *Nat Rev Cancer* 2008, **8:**497-511.
6. Ruan Y, Ooi HS, Choo SW, Chiu KP, Zhao XD, Srinivasan KG, Yao F, Choo CY, Liu J, Ariyaratne P, Bin WG, Kuznetsov VA, Shahab A, Sung WK, Bourque G, Palanisamy N, Wei CL: **Fusion transcripts and transcribed retrotransposed loci discovered through comprehensive transcriptome analysis using Paired-End diTags (PETs).** *Genome Res* 2007, **17:**828-838.
7. Chen W, Kalscheuer V, Tzschach A, Menzel C, Ullmann R, Schulz MH, Erdogan F, Li N, Kijas Z, Arkesteijn G, Pajares IL, Goetz-Sothmann M, Heinrich U, Rost I, Dufke A, Grasshoff U, Glaeser B, Vingron M, Ropers HH: **Mapping translocation breakpoints by next-generation sequencing.** *Genome Res* 2008, **18:**1143-1149.
8. Campbell PJ, Stephens PJ, Pleasance ED, O'meara S, Li H, Santarius T, Stebbings LA, Leroy C, Edkins S, Hardy C, Teague JW, Menzies A, Goodhead I, Turner DJ, Clee CM, Quail MA, Cox A, Brown C, Durbin R, Hurles ME, Edwards PA, Bignell GR, Stratton MR, Futreal PA: **Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing.** *Nat Genet* 2008, **40:**722-729.
9. Maher CA, Kumar-Sinha C, Cao X, Kalyana-Sundaram S, Han B, Jing X, Sam L, Barrette T, Palanisamy N, Chinnaiyan AM: **Transcriptome sequencing to detect gene fusions in cancer.** *Nature* 2009 in press.
10. Nasedkina T, Domer P, Zharinov V, Hoberg J, Lysov Y, Mirzabekov A: **Identification of chromosomal translocations in leukemias by hybridization with oligonucleotide microarrays.** *Haematologica* 2002, **87:**363-372.
11. Nasedkina TV, Zharinov VS, Isaeva EA, Mityaeva ON, Yurasov RN, Surzhikov SA, Turigin AY, Rubina AY, Karachunskii AI, Gartenhaus RB, Mirzabekov AD: **Clinical screening of gene rearrangements in childhood leukemia by using a multiplex polymerase chain reaction-microarray approach.** *Clin Cancer Res* 2003, **9:**5620-5629.
12. Shi RZ, Morrissey JM, Rowley JD: **Screening and quantification of multiple chromosome translocations in human leukemia.** *Clin Chem* 2003, **49:**1066-1073.
13. Lu Q, Nunez E, Lin C, Christensen K, Downs T, Carson DA, Wang-Rodriguez J, Liu YT: **A sensitive array-based assay for identify-**

ing multiple TMPRSS2:ERG fusion gene variants. *Nucleic Acids Res* 2008.

14. Jack I, Seshadri R, Garson M, Michael P, Callen D, Zola H, Morley A: **RCH-ACV: a lymphoblastic leukemia cell line with chromosome translocation 1;19 and trisomy 8.** *Cancer Genet Cytogenet* 1986, **19:**261-269.
15. Rosenfeld C, Goutner A, Choquet C, Venuat AM, Kayibanda B, Pico JL, Greaves MF: **Phenotypic characterisation of a unique non-T, non-B acute lymphoblastic leukaemia cell line.** *Nature* 1977, **267:**841-843.
16. Matsuo Y, Drexler HG: **Establishment and characterization of human B cell precursor-leukemia cell lines.** *Leuk Res* 1998, **22:**567-579.
17. **BioMart, Martview** *Versions: Ensembl 48, release 43, Feb. 2007 (Sanger), NCBI36* [http://www.biomart.org/].
18. Brazma A, Hingamp P, Quackenbush J, Sherlock G, Spellman P, Stoeckert C, Aach J, Ansorge W, Ball CA, Causton HC, Gaasterland T, Glenisson P, Holstege FC, Kim IF, Markowitz V, Matese JC, Parkinson H, Robinson A, Sarkans U, Schulze-Kremer S, Stewart J, Taylor R, Vilo J, Vingron M: **Minimum information about a microarray experiment (MIAME)-toward standards for microarray data.** *Nat Genet* 2001, **29:**365-371.
19. Cerveira N, Ribeiro FR, Peixoto A, Costa V, Henrique R, Jerónimo C, Teixeira MR: ***TMPRSS2-ERG* gene fusion causing *ERG* overexpression precedes chromosome copy number changes in prostate carcinomas and paired HGPIN lesions.** *Neoplasia* 2006, **8:**826-832.
20. Novo FJ, de Mendíbil IO, Vizmanos JL: **TICdb: a collection of mapped translocation breakpoints in cancer.** *BMC Genomics* 2007, **8:**33.
21. Jhavar S, Reid A, Clark J, Kote-Jarai Z, Christmas T, Thompson A, Woodhouse C, Ogden C, Fisher C, Corbishley C, De-Bono J, Eeles R, Brewer D, Cooper C: **Detection of TMPRSS2-ERG Translocations in Human Prostate Cancer by Expression Profiling Using GeneChip Human Exon 1.0 ST Arrays.** *J Mol Diagn* 2008, **10:**50-57.
22. Lannon CL, Sorensen PH: **ETV6-NTRK3: a chimeric protein tyrosine kinase with transformation activity in multiple cell lineages.** *Semin Cancer Biol* 2005, **15:**215-223.

Appendix Errata

Errata list

The following corrections have been made to Paper III

1. Page 1, in the author list "Lagensen" was changed to "Lagesen".
2. Page 13, line 4 was changed to "Supplementary Files 3-12".
3. Page 15, the last full sentence starting with "Seven of these are …" was changed to "Six of these are detected within transcripts that also span other annotated genes and/or ncRNAs (examples include *dinQ* and *istR2*) as well (candidate ID R3, R7-R9, R13 and R14 in Table 3 and Supplementary File 3), while the rest are located inside transcripts that are located ≥ 100 nts from any upstream and downstream annotation (Table 3 and Supplementary File 7)."
4. Page 16, line 2 was changed to "Supplementary Files 13-16".
5. Page 16, Table 2, the text under Table 2 was changed to "Intersection represents the ncRNAs candidates present in both, while union represents the number of ncRNA candidates present in either, of the two studies (Saetrom and Hershberg)."
6. Page 16, Table 3, last sentence of the table caption was changed to "The ID column can identify the candidate transcripts in the Supplementary Files; Supplementary File 3: R3, R7-R9, R13, R14 and Supplementary File 7: R1-R2, R4-R6, R10-R12."
7. Page 18, just below the middle the sentence within parenthesis was changed to "(Figure 5, see Supplementary File 17 for primer sequences)".
8. Page 24, the text in parenthesis on the first line was changed to "(Supplementary File 18)".
9. Page 24, the text in parenthesis on the third line under Table 6 was changed to "(Supplementary File 19)".
10. Page 25, the text in parenthesis on line 5 was changed to: "(Supplementary File 18)".
11. The overview of Supplementary Files was changed to:

    Additional files

    Supplementary_file_1.xls – Differentially expressed genes, annotation method
    All genes detected as differentially expressed by the annotation method.

    Supplementary_file_2.xls – Similarly expressed genes, annotation method
    All genes detected as similarly expressed by the annotation method.

    Supplementary_file_3.xls – Differentially expressed transcripts touching genes, window method
    All differentially expressed transcripts touching annotated genes (detected by the sliding window method).

    Supplementary_file_4.xls – Differentially expressed possible operon elements, window method
    All differentially expressed possible operon elements (detected by the sliding window method).

Supplementary_file_5.xls – Differentially expressed possible 5'UTRs, window method
All differentially expressed possible 5'UTRs (detected by the sliding window method).

Supplementary_file_6.xls – Differentially expressed possible 3'UTRs, window method
All differentially expressed possible 5'UTRs (detected by the sliding window method).

Supplementary_file_7.xls – Differentially expressed novel transcripts, window method
All differentially expressed transcripts that could not be assigned any possible role (detected by the sliding window method).

Supplementary_file_8.xls – Similarly expressed transcripts touching genes, window method
All similarly expressed transcripts touching annotated genes (detected by the sliding window method).

Supplementary_file_9.xls – Similarly expressed possible operon elements, window method
All similarly expressed possible operon elements (detected by the sliding window method).

Supplementary_file_10.xls – Similarly expressed possible 5'UTRs, window method
All similarly expressed possible 5'UTRs (detected by the sliding window method).

Supplementary_file_11.xls – Similarly expressed possible 3'UTRs, window method
All similarly expressed possible 5'UTRs (detected by the sliding window method).

Supplementary_file_12.xls – Similarly expressed novel transcripts, window method
All similarly expressed transcripts that could not be assigned any possible role (detected by the sliding window method).

Supplementary_file_13.xls – Differentially expressed transcripts overlapping ncRNA predictions done by Saetrom *et al.*
All differentially expressed transcripts that overlap ncRNA predictions by Saetrom *et al.* [11], transcripts detected by the sliding window method.

Supplementary_file_14.xls – Differentially expressed transcripts overlapping previous ncRNA predictions from the Hershberg list.
All differentially expressed transcripts that overlap any previous ncRNA prediction found in the Hershberg list [28], transcripts detected by the sliding window method.

Supplementary_file_15.xls – Similarly expressed transcripts overlapping ncRNA predictions done by Saetrom *et al.*
All similarly expressed transcripts that overlap ncRNA predictions by Saetrom *et al.* [11], detected by the sliding window method.

Supplementary_file_16.xls – Similarly expressed transcripts overlapping previous ncRNA predictions from the Hershberg list.
All similarly expressed transcripts that overlap any previous ncRNA prediction found in the Hershberg list [28], transcripts detected by the sliding window method.

Supplementary_file_17.pdf – Primer sequences for RT-qPCR
All primer sequences used for the RT-qPCR verification

Supplementary_file_18.pdf – The 23 novel peptides
In this file all NT-sequences, AA sequences, BLAST search results and the Jpred secondary structure predictions can be found for the 23 novel peptides.

Supplementary_file_19.pdf – Overlap to previously predicted small peptides
All overlaps between similarly and differentially expressed transcripts from this study and the 18 small peptides predicted by Hemm *et al*. [32].