Systems Biology in Industrial Biotechnology and Disease

Rasmussen, Simon; Brunak, Søren; Nielsen, Henrik Bjørn; Jarmer, Hanne Østergaard

Publication date:
2010

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

Citation (APA):
Rasmussen, S., Brunak, S., Nielsen, H. B., & Jarmer, H. Ø. (2010). Systems Biology in Industrial Biotechnology and Disease. Kgs. Lyngby, Denmark: Technical University of Denmark (DTU).

Systems Biology in Industrial Biotechnology and Disease

– PhD Thesis –

Simon Rasmussen

Center for Biological Sequence analysis
Department of Systems Biology
Technical University of Denmark

December 4, 2009

Cover illustration: *B. subtilis* genome tiling mRNA expression data in an artistic shape. Origin of replication is at 12 o'clock. Blue: Watson strand, Magenta: Crick strand, Green: genomic DNA hybridization. For all scales, darker is higher signal.

Preface

This PhD thesis was prepared at the Center for Biological Sequence analysis (CBS), Department of Systems Biology, Technical University of Denmark under the supervision of Hanne Jarmer, Henrik Bjørn Nielsen and Søren Brunak. The work was funded by the European Commission through the EU-IP-FP6-project ADIT (LSHB-2005-511977) and the Danish Strategic Research Council under the program "Associations between the diet, the composition of microbiota of the intestinal tract and human health". Additional funding for laboratory expenses came from EU-IP-FP6-project, BaSysBio (LSHG-CT2006-037469) and Augustinus Fonden.

Simon Rasmussen
Lyngby, December 2009

Contents

Abstract

Systems biology in industrial biotechnology and disease

Systems biology is a paradigm in biological science that has provided an alternative approach to the traditional reductionistic way of performing biological research. Rather than focusing on the individual parts of a biological system, the systems are recognized as inherently complex. Therefore the study of the entire system, such as with omics approaches, is more likely to be able to identify and explain the emergent properties of these systems. To power this, high-throughput technologies, allowing for data generation at a previously unparalleled scale, has been used in bottom-up approaches to perform data-driven research. Integrated and combined with top-down approaches, such as systems modeling, it is possible to investigate complex biological systems. In this thesis I provide examples of how systems biology, using DNA microarray based transcriptomic research, can be applied in biological research related to industrial biotechnology and disease.

With regard to industrial biotechnology, we have performed a tiling DNA microarray experiment identifying transcriptionally active regions in the genome of the gram-positive bacteria *Bacillus subtilis*. This organism is widely used in the industry for enzyme production and increasing systemic knowledge of *B. subtilis* is important for improving the efficiency in industrial applications. Here we extend an existing segmentation method to provide a thorough mapping of transcription and identify 125 novel putative antisense transcripts as well as 54 non-coding RNAs not previously described. Furthermore we identify conserved 3' UTRs that could be involved in protein-assembly or transport. Secondly we have performed an analysis of DNA microarray data on a *Saccharomyces cerevisiae* mutant which has been genetically engineered to have increased tolerance towards glucose and alcohol stress. Here we find that the increased fitness of the strain is only present in leucine poor media and that increased uptake or utilization of leucine is responsible for the phenotype.

In relation to disease research we have been a part of an Alzheimer's Disease drug discovery consortium. Alzheimer's Disease is the most common neurodegenerative disease leading to a progressive cognitive decline and threatens to become an even larger burden on society than it is today. Using gene expression analysis on a disease model we have identified gene targets for small molecule drug discovery of which small molecules targeting two genes are in lead development. Additionally by integrating expression data with protein interaction and phenotypic interaction data we have identified ADAM23 as associated with the disease in humans. Lastly gene expression profiling investigates the mechanism by which the probiotic bacterium *Lactobacillus acidophilus* NCFM induces a viral response from the immune system. This has been shown in clinical trials to reduce the risk of viral infections such as common cold and influenza and we find that the viral response is mediated by IFN-β in a TLR2 dependent mechanism.

Resumè

Systembiologi i industriel bioteknologi og sygdomme

Systembiologi er et paradigme indenfor biologisk videnskab der i forhold til traditionel reduktionisme har en alternativ tilgangsvinkel til biologisk forskning. I stedet for at fokusere pàde enkelte dele af et biologisk system skal systemerne forstås som komplekse i deres natur. Derfor er det med undersøgelser af hele systemet, hvilket er muligt med "omics"tilgange, mere sandssynligt at identificere og beskrive disses emergente egenskaber. En af drivkrafterne bag dette er high-throughput teknologier der muliggør at genere data i en hidtil uset skala og disse er blevet anvendt i bottom-up tilgange til at udføre data-drevet forskning. Via integrering med top-down tilgange, som systemmodellering, er det muligt at undersøge komplekse biologiske systemer. I denne afhandling giver jeg eksempler på hvordan systembiologi baseret på DNA mikroarray transkriptomics, kan anvendes i biologisk forskning relateret til industriel bioteknologi og sygdomme.

Med hensyn til industriel bioteknologi har vi udført et tiling DNA mikroarray forsøg hvor vi identificerer transkriptionelle aktive regioner i genomet af den gram-positive bakterie *Bacillus subtilis*. Denne organisme er meget anvendt i industrien til produktion af enzymer og det er vigtigt at øge den systemiske viden om *B. subtilis* for at øge effektiviteten i industrielle formål. Her udbygger vi en eksisterende segmenteringsmetode til at udføre en grundig kortlægning af transkriptionen og identificerer 125 nye formodede antisense transkripter såvel som 54 ikke tidligere beskrevede ikke-kodende RNA. Derudover identificerer vi en konserveret 3' UTR som kan være involveret i protein-samling og transport. For det andet har vi udført en analyse af DNA mikroarray data af en *Saccharomyces cerevisiae* mutant som er blevet genetisk konstrueret til at have øget tolerance overfor glukose og alkohol stress. Her finder vi at stammens øgede egenthed kun er tilstedet i leucin fattige medier og at øget optag eller udnyttelse af leucin er ansvarlig for fænotypen.

I relation til sygdomsforskning har vi været del af et konsortium med det formål at udvikle nye lægemidler mod Alzheimers sygdom. Alzheimers sygdom er den mest hyppigt forekommende nervedegenerative sygdom og medfører tiltagende kognitiv forfald, og sygdommen truer med at blive en endnu større byrde for samfundet end den er idag. Her har vi ud fra gen ekspressionsanalyse af en sygdomsmodel identificeret gener der kan være mål for små molekylære lægemidler, og lead molekyler er ved at blive udviklet for to af disse. Integration af ekspressionsdata med protein-protein interaktionsdata og fænotypedata har derudover ført til identifikationen af ADAM23 som associeret med sygdommen i mennesker. Som det sidste har gen ekspressionsanalyse givet indsigt i den mekanisme hvorved den probiotiske bakterie *Lactobacillus acidophilus* NCFM kan inducere et viralt respons fra immunsystemet. Dette har i kliniske studier vist at reducere risikoen for virus infektioner som forkølelse og influenza og her viser vi at det virale respons er medieret via produktion af IFN-β i en TLR2 afhængig mekanisme.

Acknowledgements

I am deeply grateful for the supervision I have received during my PhD from Associate Professor Hanne Jarmer, Associate Professor Henrik Bjørn Nielsen and Professor Søren Brunak. Especially the day-to-day supervision by Hanne and Bjørn has been fantastic – I have learned so much from you and I had a great time doing so.

Claus Andersen from Siena Biotech for letting me come to Siena and feel at home, and for all the tricks and guidance during the ADIT project. Additionally Alessia Tarditi, Guiseppe Pollio, Roberto Raggiaschi and Georg Terstappen also from SiBi.

Pia Friis for high quality technical assistance, for letting me use her reagents and pipets, and especially for running DNA microarrays en masse.

My collaborators, especially Jeroen Hoozemans, Richard Baerends, Emma Denham, Gudrun Weiss and Hanne Frøkiær.

My office-mates, Chris, Duygu, Peter Hagedorn, Greg and Kristoffer Kiil – and the neighboring office Aron and Nicolai. And thanks Aron for getting me started on and giving me a lot of tricks on LaTeX.

Flemming Hansen for learning me "real" microbiology and letting me use his lab for countless synchronization experiments. Susanne Koefoed for lots of reagents, LB plates, technical guidance and pleasant lab-talk.

Sheila, Tune, Niels, 2 × Daniel, Kasper, Bent, Qiyuan, Kirstine, Malene, Rikke, Thomas, Nico, Pernille, Rodrigo, Chico and probably a lot more for making CBS such as nice place.

Professor Stanley Brul, Luc Hornstra, Wouter Kallemeijn, Johan van Beilen and the rest of the dutch gang who helped me out with every possible thing on *B. subtilis* spores and the dutch language during my stay in Amsterdam.

Kasper Lage for being a protein-phenotype wizard and Niclas Tue Hansen for starting me up on disease gene finding scripts.

The administration, especially Marlene for mastering EU projects, Dorthe for knowing and fixing everything regarding my PhD study and paternity leave, and of course Lone for fixing just about everything else.

The system administration, Kristoffer, Peter Wad, Hans-Henrik and Olga for keep-

ing the system running and backing up my files.

And of course my wife, Marie-Louise, who has been even more patient than I could ever have imagined and for giving me a family of 3 (soon to be 4).

Papers included in the thesis

- **Paper I: Rasmussen S**, Nielsen HB, Jarmer H (2009) The transcriptionally active regions in the genome of *Bacillus subtilis*. Molecular Microbiology 73: 1043-1057 Cover illustration.

- **Paper II:** Baerends RJ, Qiu JL, **Rasmussen S**, Nielsen HB, Brandt A (2009) Impaired uptake and/or utilization of leucine by *Saccharamyces cerevisiae* is suppresed by the SPT15-300 allele of the TATA-binding protein gene. Applied Environmental Microbiology 75: 6055-6061.

- **Paper III: Rasmussen S**, Hondius D, Hoozemans JM, Nielsen HB, Brunak S. ADAM23 is associated with Alzheimer's Disease in the human brain. [manuscript in preparation].

- **Paper IV:** Weiss GM, **Rasmussen S**, Zeuthen LH, Nielsen BN, Jarmer H, Jespersen L, Frøkiær H *Lactobacillus acidophilus* induces virus immune defense genes in murine dendritic cells by a TLR-2 dependent mechanism. Immunology (Accepted).

Papers not included in the thesis

- Nicolas P, Leduc A, Robin S, **Rasmussen S**, Jarmer H, Bessières P (2009) Transcriptional landscape estimation from tiling array data using a model of signal shift and drift. Bioinformatics 25: 2341-234.

- Weiss GM, **Rasmussen S**, Nielsen LF, Jarmer H, Nielsen BN, Frøkiær H *Bifidobacterium bifidum* actively changes the gene expression profile induced by *Lactobacillus acidophilus* in murine dendritic cells. (Submitted to PLoS ONE).

Work in progress

- From the work performed on Alzheimer's Disease (Chapter 5) hopefully a drug molecule will be patented in the not so near future. Currently lead molecules have been identified for two of the gene targets that we identified from the DNA microarray analysis.

Part I

General Introduction

Chapter 1

Introduction

1.1 Systems biology

The field of systems biology is not uniformly defined and covers several aspects. One view is that it can be thought of as a paradigm of biological research where data integration and understanding emergent properties of complex biological systems is in focus. It has been well formulated using analogies of a radio or an airplane, where knowing the individual parts of these systems is not sufficient for describing how the machine functions. To understand the complex properties and functionalities of these, knowledge of the wiring and interplay of the individual parts is critical [1, 2]. Likewise, the genome sequence of an organism and identification of its genes is not sufficient for describing the complex properties of that biological system. Although reductionistic approaches has proven adequate when the number of parts is relatively low, such as modeling planetary movement in the solar system, biological systems are inherently complex. The change from reductionism to studying systems as a whole, is at the fundamental level for understanding complex processes and diseases.

One factor in the emergence of this holistic approach has been the development of experimental techniques that has allowed for massive parallel generation of data describing different aspects of biology. These include technologies for studying genomes, genetics, transcripts, proteins, metabolites, protein-protein interactions, protein-DNA interactions and more, leading to the generation of the “omics” concept. Generally termed “bottom-up”, these approaches are used to describe the system starting from the lowest state (e.g. genes, proteins, metabolites) and are at the base of highly data-driven research (Figure 1.1) [3]. These may be integrated

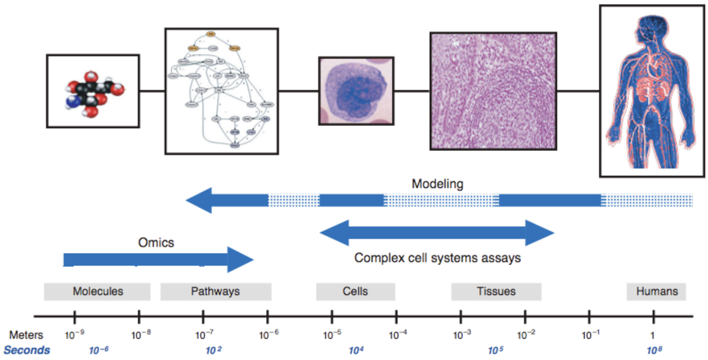


Figure 1.1 – Approaches to systems biology as omics-based (bottom-up) and model-based (top-down). By integrating omics approaches with modeling better understanding of biological systems can be achieved. Reprinted with permission from [7].

with "top-down approaches" such as system and network models, where models are build covering elements from a lower state. Using this combined approach it is possible to perform research from which induction of biological knowledge and novel hypotheses may be obtained [4–6]. In this view systems biology is perceived as an iterative approach generating biological insight by combining "traditional" life science experiments (mainly high-throughput) with model development and evaluation (Figure 1.2).

Data integration

By the generation of massive amounts of data from several aspects of biology, integration of different data types has become of increasing importance. One advantage of data integration is that the type of errors from orthogonal data sources are not likely to be the same, hereby increasing the power and value of the data. However one of the most obvious reasons for data integration is that the complex problems in systems biology are only solvable if multiple parts of the underlying biology can be studied and combined [7]. For example for common complex disorders such as cancers, diabetes, obesity and brain diseases, there are probably many different, and rare, genetic factors responsible for disease development. Additionally the majority of these may only have a weak impact on disease risk making the discovery of these not straightforward. The complexity of these diseases can be emphasized by the fact that at the time of writing, 35 genes have been associated with Late-Onset Alzheimer's Disease [8]. However, recent large scale Genome Wide Associations Studies (GWAS) have had success in identifying the genetics of some common and low-penetrance Single Nucleotide Polymorphisms (SNPs) as-

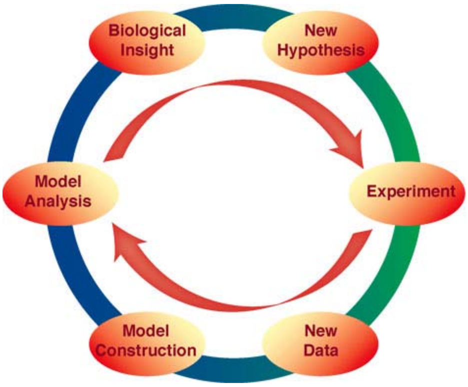


Figure 1.2 – Systems biology as an combined and iterative discipline. To be able to explain complex biological systems the combined effort of high-throughput experiments, data integration, model development and evaluation allow for the revision and development of new hypotheses.

sociated with common diseases [6, 9–11]. From a drug discovery perspective the nature of these rare disease alleles with low penetrance may increase the risk of drug development. Drugs targeting these may not have a huge clinical impact, because they will only be effective in a minority of the patients.

From protein sciences it has been established that proteins function together in complexes and knowledge of these interactions have become of vital importance for studying diseases. When a protein in a complex is dysfunctional it will often disrupt the functionality of the entire complex [12]. In this sense diseases are modular – similar disease phenotypes are seen when any of the proteins in the complex are disrupted [13–15]. Because of this protein-protein interaction networks have become important and powerful resources for studying cellular processes, and are at the core of Integrative Systems Biology. In our study of Alzheimer's Disease (AD) we integrate transcriptomic data with inferred protein-protein interaction networks and phenotypic data assembled from the Online Mendelian Inheritance in Man (OMIM) database. From this we identify a candidate gene (*ADAM23*) and show that the protein product is up regulated and associated with AD in human post-mortem brain tissue.

In our tiling DNA microarray study of *B. subtilis* (Paper I) we take a different integrative approach not utilizing protein-interaction networks. This is primarily because the goal of the study is to expand the understanding of the transcriptome including findings such as UnTranslated Regions (UTRs), operon structure and novel non-coding RNA (ncRNA). These findings achieve increased confidence by integration with known and predicted sigma factor binding sites and Rho-independent terminators.

1.2 Transcription

In the majority of all living organisms DeoxyriboNucleic Acid (DNA) contains the genetic information needed to develop and perform the functions of a particular organism. DNA contains the blueprint of how to build the components of the living organism such as proteins and RiboNucleic Acid (RNA) molecules. Following the central dogma information flows from the stretches of DNA, that are known as genes, via RNA to proteins, that are the primary active gene products. This occurs by transcription in which information is transferred from DNA to messenger RNA (mRNA), and then by translation that through mRNA instructs ribosomes in how to assemble a particular protein. Hence the mRNAs present at a given time will to a large extent represent the active genes in a cell at that moment (Figure 1.3). Even though almost all cells in multicellular organisms contain the same genetic information (DNA) they have very diverse roles and functionalities, because only subsets of genes are active at certain cell stages or developments. By identifying all the transcripts present and measuring their expression levels and changes, information about active and repressed processes can be obtained. And though RNA transcripts primarily are the information carriers and therefore only proxies leading to protein expression, the relatively ease of use, makes transcriptomics an excellent tool for studying biology.

Genes

When doing biological research, one of the concepts that is most often used is “gene” and most scientists within the field has a sense for what a gene is. However when writing this thesis I have realized that a gene might not be such an easy entity to define. Since the invention of the term “gene” by Wilhelm Johannsen in 1909 to describe inheritable biological traits, a concept developed by Gregor Mendel in 1866, the definition has changed several times [16]. When I first encountered biological chemistry during my studies a decade ago, a gene was defined as an inheritable discrete DNA (or RNA) element that contribute to the phenotype. Except for rRNA and tRNA they were generally defined as protein-coding and 98% of the DNA not coding for genes were thought of as primarily non functional and termed “junk DNA”[1]. However today tiling DNA microarrays, Next Generation Sequencing (NGS) and other transcriptomic technologies have shown that not only protein coding genes are transcribed. Generally the transcriptome of eukaryotic and especially mammalian organisms have been shown to comprise of a myriad of transcripts where NGS studies have shown that 30-40% of all sequence reads to map to unannotated regions [17–19]. A source of these transcripts have been

[1] Any discrepancies to actually knowledge at that time is entirely at my own responsibility.

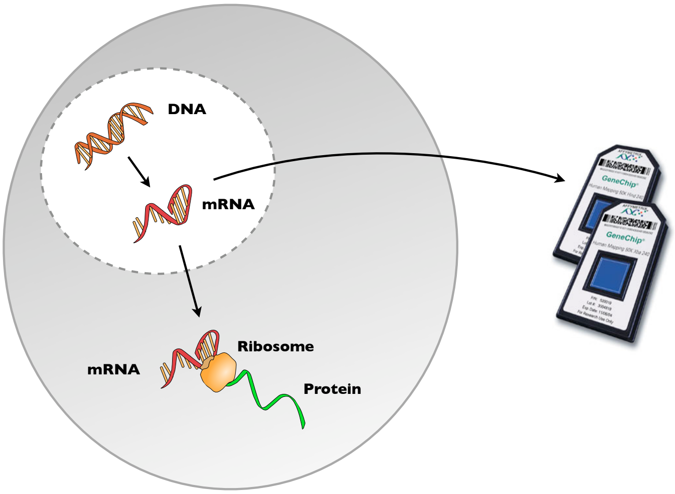


Figure 1.3 – The central dogma in molecular biology showing the general transfer of information within cells. Information flows from DNA (information storage) to mRNA (information carrier) via transcription and hereafter translation where ribosomes synthesize proteins (active cell machinery) using this information. Generally mRNA abundance is assayed using DNA microarrays.

shown to be bi-directional promoters from where there seem to be a tendency of the transcriptional machinery to start transcription from regions that are nucleosome free [20]. Natural antisense transcription has increased in abundance to such an extent that antisense transcription is thought of as a pervasive feature of eukaryotic and mammalian genomes [21]. These discoveries challenge the definition of a gene.

Determinants of transcription

To start transcription from a DNA region, the DNA must be accessible to the cellular machinery. There are some differences between prokaryotes and eukaryotes in how and where transcription occurs. For eukaryotes, DNA is packed as chromatin, stored in the cell nucleus and generally exists as closed or open forms known as hetero- and euchromatin. For DNA to be available for transcription it needs to be unwound to euchromatin before general transcription factors, RNA polymerase and mediators can bind [22]. Regarding prokaryotes, that does not contain a nucleus, the chromosome is present in the cytoplasm where it is generally less condensed. Here RNA polymerase and sigma factors initiate transcription by unwinding the coiled DNA. Generally transcription is more complex in eukaryotes compared to prokaryotes, one example being that eukaryotes has three different types of RNA polymerases whereas prokaryotes (bacteria) has only one [23, 24].

The promoter region, which is where transcription initiates, is subjected to several different regulatory mechanisms. Relatively close upstream and downstream regions can be the target of transcription factors leading to repression or activation of transcription from the promoter. Typically the transcription factors work in concert to mediate their regulation and prediction of this is not a straightforward task. Additionally enhancer and silencer regions, which does not necessarily have to be close in terms of DNA sequence to the transcription start site, influence eukaryotic transcription. There are even examples of enhancers found within introns and on different chromosomes [22, 25, 26]. However, similarly as with the understanding of genes, the abundant existence of bi-directional promoters challenge the consensus view of a promoter.

Regarding termination of transcription in eukaryotes, different mechanisms exists for the three RNA polymerases. In the case of RNA polymerase II, which is responsible for the production of protein-coding mRNA, termination generally occurs when a cleavage-specific RNA sequence is synthesized. This leads to cleavage of the 3' and polyadenylation, a signal involved in nuclear export, translation and stability. Alternative polyadenylation can occur which can influence exon conformation and include or exclude miRNA binding sites in the 3' of the transcript [24]. On the contrary polyadenylation of RNA in bacteria is needed for efficient degradation of RNA molecules [27].

The understanding of bacterial promoters as a concept is thought of as relatively well established, however mapping of Transcription Start Sites (TSSs) is far from accomplished. This is exemplified by the fact that for the model organism of gram-positive bacteria, *B. subtilis*, only 660 TSSs are known [28]. This should be seen in the perspective of our tiling array study of *B. subtilis* (Paper I) where more than 3.500 Transcriptionally Active Regions (TARs) are identified. Additionally polycistronic operons can be transcribed from more than one promoter and contain internal read-through terminators allowing for several different transcripts from the same region. Termination of transcription in bacteria generally occurs by Rho-independent or Rho-dependent termination. In the former transcription is terminated when a hairpin structure followed by uracils, known as a terminator structure, is formed, whereas the latter approach is dependent on the binding of the protein Rho [23]. Hopefully the findings of our study will attribute to improving the mapping of TSSs and transcripts, as well as the understanding of transcription in *B. subtilis*.

What is in a transcript?

When performing transcriptomics all identifiable transcripts and their abundance from a genome is studied. Simplified mRNA transcripts can be thought of as information carriers for protein-coding genes, leading to protein production from the information in the gene. However transcripts does not only contain the coding sequence that instructs how the protein has to be assembled. For eukaryotes, transcripts contain a mixture of introns and exons, where introns are spliced out of the transcript leading to a transcript consisting of exons. Alternative splicing of the transcript including or excluding different exons lead to different coding sequences and hence alternative proteins [24]. Transcripts have been identified that are composed of exons from different genes and even from different chromosomes [16, 19]. For prokaryotes genes are often structured into operons where several genes are present on the same transcript, but each of the coding sequences are translated into separate proteins.

Transcripts also contain UnTranslated Regions (UTRs), which in the 5' and 3' of the transcript is target of many regulatory events. The length and utilization of UTRs for regulatory mechanisms has been driven by evolution, and the average length of 3' UTRs correlates with organismal complexity [29]. Paper I shows that for the *B. subtilis* transcriptome the 5' UTRs are longer compared to 3' UTRs, a finding in good correlation with the above. Examples of transcriptional regulation in UTRs are transcriptional attenuation in bacteria where secondary structure in the 5' of the transcript can block transcription by forming hairpin structures. For eukaryotes microRNA is a widespread example of additional features within a transcript, as they can be encoded from within introns or 3' UTRs, from where they are spliced and processed to functional miRNAs. These can bind to complementary mRNA and induce transcript cleavage or translational inhibition. Interestingly 30% of the human mRNA pool is thought to be regulated by miRNAs [30].

Natural antisense transcription

The discovery that natural antisense transcription is a pervasive feature of eukaryotic genomes has introduced a complexity not previously recognized. Additionally one must question the ability of the traditional protein-coding gene centric transcriptomic approaches such as non-tiling DNA microarrays for gene expression profiling. Although the large amount of positive results already obtained from these technologies are probably still essentially true, many underlying changes and mechanisms must have escaped our attention. Especially in complex diseases such as cancers, diabetes and brain diseases where the exact underlying mechanisms have been hard to fully identify. An example directly relevant to the Alzheimer's Disease work in this thesis is that an antisense transcript *BACE1*-AS has been shown to mask the miRNA binding site in the 3' of the *BACE1* transcript. This leads to increased stability of the *BACE1* transcript, coding for a crucial protein in the production of the Aβ peptide [31]. This and other examples demonstrate that the traditional understanding of antisense transcription as being exclusive negative regulators of their sense transcript is not true. Both anti-correlated (discordant) and correlated (concordant) relationships between sense and antisense transcript pairs have been shown, however the mechanisms by which they occur are not all well understood. Several mechanisms of action have been proposed and to a greater or lesser extent validated, however the general consensus is that natural antisense transcripts are heterogeneous, both in term of composition and mode of action. Regarding their generation they can be produced head-to-head with overlapping 5', tail-to-tail with overlapping 3' or fully/partial overlapping inside the sense transcript. Functionalities such as epigenetic regulation, including DNA methylation and chromatin modifications, genomic imprinting, alternative splicing, transport modification, modification of mRNA stability and translation has been shown. Additionally endogenous small interfering RNA (siRNA) are generated from double-stranded RNA [17–21].

Generally antisense transcription has not been pronouncedly described for bacteria, with the example of *B. subtilis* having only two known antisense transcripts. In Paper I using tiling DNA microarrays we identify 125 putative novel antisense transcripts for *B. subtilis* verifying and expanding the phenomenon in bacteria. Among these we even describe an antisense transcript tail-to-tail with *sigA*, the major housekeeping sigma factor in *B. subtilis*. The antisense transcripts observed are generally expressed at relatively low levels and experiments using Next Generation Sequencing is likely to improve greatly on the resolution and mapping of these. Additionally the presence of several antisense transcripts has recently been shown for *Listeria monocytogenes* and the archaea *Sulfolobus solfataricus* [32,33].

1.3 Next-Generation RNA-sequencing

When writing a thesis largely based on analyzing transcriptomic data it is impossible not to mention the emergence of Next Generation Sequencing technology. For more than a decade DNA microarrays have been a key tool in biological research, allowing highly parallel studies of gene expression. However, just as DNA microarrays revolutionized biological research NGS is now introducing a new revolution. The technology enables high-throughput and deep sequencing of DNA, and has been used to sequence and re-sequence genomes at a fraction of the cost of traditional Sanger sequencing. Besides this, NGS can be used for gene expression analysis (RNA-seq) and show several advantages compared to DNA microarrays, primarily because NGS relies on direct measurement of sequence reads and not the inferring of nucleotide concentration from hybridization levels. As an analogy one can think of RNA-seq as providing "digital measurements" of transcript abundance, whereas DNA microarrays in some sense is an analog system. This allows for single base resolution, low background noise and a high dynamic range. Additionally novel transcriptomic features such as discovery of non-coding RNA, novel exons, poly-adenylation sites, rare transcripts and novel splice variants is greatly enhanced by the fact that the cDNA sequence is directly identified [19, 34, 35]. Furthermore it is possible to identify and monitor SNPs within RNA sequences and to perform allele-specific transcription.

Although NGS is a developing technology one may expect it to outperform

DNA microarray based methods for the majority, if not all, of the traditional DNA microarray applications. Although the process may be gradual, the emphasis and new developments in the near future are likely going to be based and focused on NGS technology. However, as RNA-seq is still based on investigating the transcriptome the majority of the data analysis methods and data integration approaches will be overlapping.

Chapter 2

The DNA microarray technology

The purpose of DNA microarray technology is to determine the abundance of specific RNA or DNA molecules in an experiment. The technology was developed from the classical molecular biology techniques Southern blotting and northern blotting that are used to detect and measure the relative abundance of DNA or RNA molecules, respectively [36,37]. The underlying principle in DNA microarrays is a massive parallelization of these blots, where the arrays of today incorporate the technological improvements from multiple fields. The use of DNA microarrays have additionally evolved from expression analysis to several other applications and is a multi-discipline field involving physics, chemistry, biology, statistics and informatics. Whether the goal of the DNA microarray experiment is to measure relative abundance of messenger RNA, identify alternative splicing events, novel transcripts, chromosomal DNA variations, transcription factor binding sites or SNPs associated with disease, the basic principle is always the same – hybridization of complementary nucleotide sequences.

A short story of DNA microarrays

A common conception is that DNA arrays was developed in the late 1980s and appeared in scientific publications in the beginning and mid 1990s. However, DNA arrays, in the sense of parallel blots on filter paper called *dot blots*, was introduced in the late 1970s for homology studies and for keeping libraries of complementary DNA (cDNA) clones. Similarly expression analysis, that has been the primary application of DNA microarrays for a more than a decade, began in the early

1980s using filter papers spotted with cDNA clones. This was achieved by blotting from cDNA libraries to two filter papers and then hybridizing radioactive labeled cDNA from two different samples to each filter. Further technological advances in e.g. the Polymerase Chain Reaction (PCR) automation lead to the adaption to membranes instead of filters and the emergence of a more robust and higher throughput system known as *macroarrays* in the early 1990s [38, 39].

As the name implies *microarrays* were developed as a result of miniaturization of the macroarray technology. At Stanford University, Schena et al. in 1995 published a study where 45 *Arabidopsis thanliana* cDNA clones were spotted on microscope glass slides [40]. This together with fluorescent labeling and the use of confocal laser scanners enabled mRNA abundance measurements of high sensitivity. For these cDNA arrays two-color labeling was used, meaning that samples were labeled with separate dyes and then mixed prior to hybridization on glass slides. By scanning for each dye, the relative abundance of mRNA species between two samples can be measured in just one microarray experiment [41].

In 1991 Fodor et al., later co-founder of the company Affymetrix, developed an array manufacturing approach using technology from the semiconductor industry, photolitography, and combinatorial chemistry. In this approach light is used to control the synthesis of nucleotides on a quartz wafer by applying pre-designed masks that will either block or allow light to hit selected areas. Each of these areas, also termed features, contains millions of copies of certain 25-mer oligonucleotides allowing the detection of its complementary oligonucleotide. In contrast to cDNA microarrays, manufacturing of these *oligonucleotide* arrays only requires knowledge about the genome sequence, no cDNA library is needed as the probes are manufactured *in situ* [42]. However due to the use of relatively short probes, several probes are needed for each gene in order to achieve reliable measurements. Additionally both perfect match (PM) and mismatch probes (MM) have traditionally been used, where the MM probes are similar to the PM probes except for the middle base substituted by the complementary base. Theoretically this would allow for probe background detection, however using these have not proven beneficial for the data quality. Affymetrix arrays today utilize background probes with similar GC content to estimate background signals.

In addition several other companies have ventured in to the DNA microarray industry. Similar to the Affymetrix approach, two other companies use *in situ* synthesis of oligonucleotides. Roche NimbleGen utilizes a photolitography method for the manufacturing of microarrays where micromirrors instead of pre-designed masks, guide the light to control the synthesis at each spot. Agilent uses ink-jet printing and phosphoramidite chemistry where the oligonucleotides are synthesized by dropping nucleotides in the desired order at the spots [43, 44].

Another strategy for microarray manufacturing, used by Illumina, is bead arrays, where probes are linked to beads. In contrast to the previous approaches, the beads containing the probes are randomly distributed on the array surface

Name	Technology	Features/Array	Probe length	Detection
cDNA arrays	Spotted	98k	NA[a]	Two-color
Affymetrix	Photolitography	6.8M	25	One-color
NimbleGen	Photolitography	2.1M	<85	One and Two-color
Agilent	Ink-jet	1M	60	One and Two-color
Illumina	Bead based	1.2M	>100	One and Two-color

Table 2.1 – Overview of the major DNA microarray technologies available. [a] Probe length is defined by the cDNA or PCR product spotted and can vary.

and then decoded prior to hybridization. This is possible since a DNA barcode is incorportated into each probe sequence. An overview of the major microarray technologies is shown in Table 2.1.

2.1 Applications

The popularity of DNA microarrays was mainly founded by the ability to fast profile thousands of genes as long as a cDNA library or genome sequence was available. Additionally sharing of data from microarray experiments in large databases such as NCBI Gene Expression Omnibus (GEO) and EBI Array express has proven valuable. This enables other researchers to re-analyse and combine data sets in novel ways, an approach we used for Paper II [45, 46]. Today, the applications of DNA microarrays have evolved into a variety of fields – here I briefly describe the major applications (see Figure 2.1).

Expression analysis

In 1997 the first gene expression profiling experiment covering all protein-coding genes of an organism was performed on *S. cerevisiae* investigating transition from fermentation to respiration [47]. Following this experiment, these microarrays were used to assay other cellular phenomenons in yeast such as the cell cycle and sporulation [48–50]. From these it was clear that whole-genome DNA microarray experiments made it possible to characterize the genetic wiring responsible for complex cellular processes.

A very common application of DNA microarrays is to compare gene expression between *Case* and *Control*, often being disease vs. healthy tissue or gene knockout/overexpressor vs. wild-type. In addition to providing information on which genes that may be responsible for a particular disease phenotype, it can also help identifying candidates for drug intervention. Knocking out a gene in an organism and measuring the corresponding changes in gene expression will provide information on processes this gene may be involved in or part of controlling. More advanced experimental designs such as a two-factor ANOVA design enables the identification of genes differentially regulated by two different factors and their interaction.

Another approach is to profile different tissues creating tissue expression maps. Knowing where a gene is expressed can provide clues of functionality but can also help to indicate whether a drug targeting this gene is likely to have unintended effects. Additionally approaches like this can identify chromosomal stretches or regions of co-expressed genes that associate to different tissues. One such gene atlas for human and mouse tissues is available from Su et al. (2004) [51]. Furthermore gene expression signatures have been widely used to classify cancers and to predict treatment outcomes based on selection of differentially regulated genes. From these types of experiments prognostic information can be achieved on tumor sensitivity and whether survival or relapse can be expected from treatment with a particular drug.

Tiling arrays

Rather than probing certain areas of genes or exons, another approach is to design tiling DNA microarrays that cover genomic regions or the entire genome independent of gene organization. Spacing of the probes can be overlapping, end-to-end or average spaced depending on application and array technology. In general oligonucleotide tiling arrays are preferable compared to PCR or Bacterial Artificial Chromosomes (BAC) arrays as the two latter will have both nucleotide strands of the probe present, meaning that they are not strand-specific [52]. Whole-genome tiling is possible on single arrays for organisms with small genomes, whereas current technology requires several arrays to tile the entire human genome. Such arrays allow the identification of novel features such as novel genes and exons, non-coding and small RNAs, novel splicing, antisense RNAs and elucidation of transcript organization such as bacterial operons. In addition to the above, tiling arrays are also at the core of Chromatin ImmunoPrecipitation on chip (ChIP-chip), DNA methylation on chip, array CGH and Copy Number Variation (CNV) studies, and re-sequencing using DNA microarrays.

Chromatin immunoprecipitation on chip

Chromatin is the complex between DNA and protein that form the chromosomes and this technique aims at identifying specific proteins binding to DNA. Originally published in 2000 and 2001 the authors investigated where several transcription factors involved in carbon source, mating pheromones and cell cycle control bound to the genome of *S. cerevisiae* [53, 54]. The approach is a merger of Chromatin ImmunoPrecipitation and DNA microarray technology (ChiP-chip). By crosslinking the chromatin using formaldehyde and thereafter shearing it by sonication, fragments of protein-DNA complexes can be extracted using antibodies for the protein of interest. By reversing the crosslink between the purified protein-DNA, the DNA can be isolated, labeled and hybridized to tiling arrays making it possi-

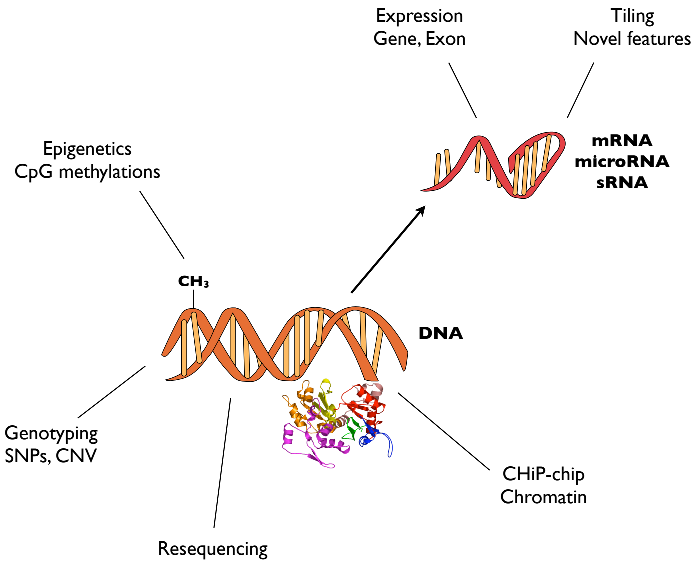


Figure 2.1 – Applications of DNA microarray technology in biology.

ble to identify the DNA regions that the particular protein of interest binds [55]. ChiP-chip has been used to decipher genetic regulatory networks at varying conditions by identifying targets of transcription factors, binding of histones and other DNA-binding proteins [56].

DNA methylation

The field of epigenetics revolves around the question of heritable changes in gene expression and phenotype that can not be explained by changes in the DNA sequence. One of the major mediators of this effect is DNA methylation of cytosines, being when a cytosine followed by a guanine (termed CpG) is methylated to 5-methylcytosine. These methylation patterns changes during cell differentiation and to external stimuli such as maternal care during early childhood and chemical compunds. Implications of DNA methylation have been shown on gene expression where methylation of promoters leads to gene silencing and demethylation of promoters to gene expression. This is often seen in cancers leading to silencing of tumor suppressor genes or over-expression of oncogenes. Additionally microRNA expression has been found to be controlled by DNA methylation [57–59].

Genotyping & Single Nucleotide Polymorphisms

The conceptually simplest form of genetic variation between genomes within the same species is when a base is substituted with another base, also known as Single Nucleotide Polymorphisms (SNPs). For more than 2.200 mendelian, and hence inherited, disorders known genes has been identified for which a SNP in a gene is the cause of the disorder [9]. Generally these have high penetrance and are relatively easy to identify compared to low penetrance SNPs that are thought to be involved in many common disorders. For the identification of these, Genome-Wide Association Studies (GWAS) using DNA microarray technology have been applied, where a large number of SNPs are screened for disease association. In the HapMap project SNPs have been collected and validated with currently more than 3.1 million SNPs identified and more than 10 million predicted to exist [6, 60–62].

Copy number variations

Chromosomal Copy Number Variations (CNVs) are defined as chromosomal areas that are either deleted or duplicated (or multiplicated) in the genome of a cell. They have been found to comprise a significant portion of genetic variation in the human genome – in terms of nucleotides covered it involves more bases than SNPs. The HapMap project have estimated CNVs to exist for ∼30% of the human reference genome and CNVs have also been reported to have higher "mutation" rates than SNPs. The functional role of a CNV can be through several mechanisms such as altered gene dosage, gene interruption and gene fusion – hence they have a large impact on the phenotype. In this respect, CNVs are associated with human diseases such as cancers, brain diseases and numerous other phenotypes. They are, however, also associated with evolution of the human genome by gene duplication and exon shuffling, and are hereby involved in the development of human beneficial traits such as cognition [63, 64].

Resequencing

Initially the inventors of the Affymetrix technology envisaged DNA microarrays to be used in re-sequencing and mutation detection [65]. The approach for sequencing using DNA microarrays is to compare the sequence in question against a reference genomic sequence and design tiling probes covering the sequence. This approach have been used for genotyping disease genes involved in breast cancer (BRCA1), cystic fibrosis (CFTR) and the HIV protease (HIV-1 PR) [66–68]. Additionally small genomes such as the mitochondrial genome and stretches of pathogenic genomes have been re-sequenced for forensics and detection of human pathogens [69]. Resequencing arrays have been directly developed for use in drug discovery such as Affymetrix Drug Metabolizing Enzymes and Transporters (DMET) and Roche AmpliChip CYP450, that resequences key drug metabolizing genes. Especially the CYP2D6 and CYP2C19 proteins are involved in the metabolism of ∼25% of all administered drugs with the implications that different genotypes will affect the metabolizing speed [70].

2.2 Experimental preparation

When designing a microarray experiment several aspects are important to ensure success. They may seem obvious, however keeping them in mind is important – and as always the “garbage in, garbage out” phrase applies. Some key points are (a) Construct a hypothesis and a subsequent experimental setup that allows for the hypothesis to be partially or fully proved true. (b) Keeping the goal of the experiment in mind and the experimental setup relatively simple is more likely to give interpretable results. (c) The importance of biological replicates allowing for statistical interpretation of the data. (d) Minimize and balance effects such as sampling time, growth conditions, handling procedures and dye swap. The procedures presented below is primarily described for gene expression profiling and RNA tiling experiments.

While performing the experiment handling of RNA samples must be performed rapidly. This is due to the fact that mRNA is being rapidly degraded by RNases, and cells to be extracted are therefore normally snap frozen in liquid nitrogen or immersed in RNA protecting solutions. Several methods are available for RNA extraction, where the most common ones are using guanidinium thiocyanate, phenol and chloroform extraction or solid-phase (column) based extraction [71, 72]. One obvious disadvantage of using solid-phase techniques is that RNA molecules shorter than 200 nucleotides are not extracted, which can have implications for experiments where small RNA, non-coding RNA and microRNA are studied. As mRNA only represents 1-5% of the total RNA an enrichment of mRNA can be performed in eukaryotes using techniques that target the polyA-tail [72]. Extracted mRNA is transcribed into double-stranded DNA using Reverse Transcriptase PCR (RT-PCR) and can hereafter be used as template for *in vitro* transcription (IVT) creating amplified RNA (aRNA) or be directly labeled. If aRNA is produced it is normally labeled during the IVT by incorporation of biotinylated nucleotides. Hereafter the cDNA or aRNA is fragmented and hybridized to DNA microarrays, washed and stained depending on the technology. Last, the DNA microarrays are scanned using a laser scanner and intensity readings for each feature is generated. The general approach for expression profiling using Affymetrix GeneChip hybridizations is shown in Figure 2.2.

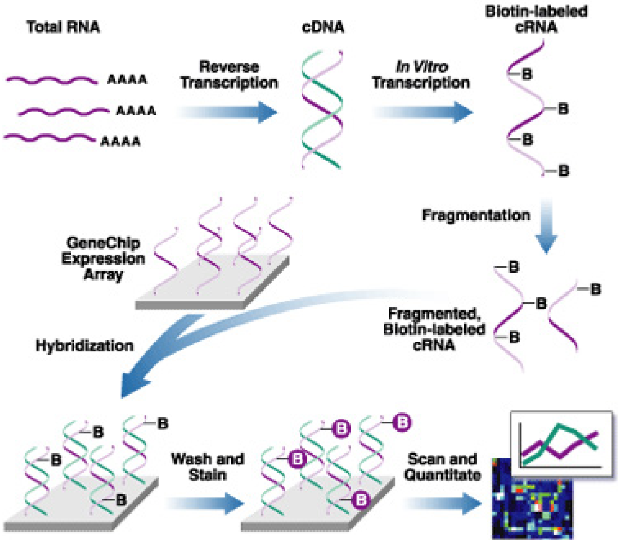


Figure 2.2 – Working scheme for Affymetrix GeneChips. From Affymetrix.

2.3 Data analysis

Several methods have been developed for analysis of DNA microarray generated data, some are commercial, whereas others are free. Of the free software, the statistical programming language *R* provides the base for the Bioconductor project which is an open source platform for computational biology and bioinformatics [73,74]. There are several packages containing software for data analysis and among the most used within DNA microarray analysis are the *affy* and *limma* packages [75, 76]. These provide basic tools for processing and analysis of Affymetrix and cDNA data, respectively. Many other packages exists for the other applications of DNA microarrays mentioned in section 2.1, see `www.bioconductor.org`. This section will focus on data analysis for gene expression analysis and tiling array data analysis, primarily using Affymetrix and NimbleGen technology, respectively.

Preprocessing

The intensity readings from the thousands to millions of features on the DNA microarrays have to be preprocessed before statistical testing and biological information can be extracted. This preprocessing is needed because raw intensity readings, in addition to gene specific signals, contain variance and noise that can lead to false biological conclusions. Uninteresting variation originates from sources such as unintended effects during sample preparation such as dye labeling and chemistries, hybridization conditions and photodetection during scanning [77]. These effects are primarily observed between different arrays, however within array variations may also occur. In general three preprocessing steps are needed, background correction, normalization and gene expression index calculations. Tools for all three steps are implemented in the R package *rma* [78].

Background signal, which is ambient non-specific signals, can arise from several sources such as non-specific binding to the surface of the array, effects from the washing stage and optical noise from the scanner [79]. This may have a pronounced effect when using two-colour microarrays where variance stabilizing methods have been shown to perform best in terms of precision and bias. Importantly the background correction method of subtracting signal from the background of a spot has been found to perform worse than using non-corrected values [79]. Regarding Affymetrix arrays background is estimated from the signal intensities themselves, where observed intensity is modeled as the sum of a signal from an exponential distribution and a background component from a normal distribution [78,80]. Additionally *gcrma* utilizes the GC content of each probe to estimate probe affinity. This type of probe affinity background correction may be useful for tiling array data (see section 2.5).

For normalization several methods exists. The simplest solution is scaling each array by a factor relative to the array with median of the median intensities – this was originally proposed by Affymetrix and generally performs poor [81]. One of the reasons for this is that the effect of global variations is primarily signal-dependent (see Figure 2.3 a and b) meaning that a non-linear normalization method is needed to normalize the data. The most commonly used non-linear normalization methods are the invariant set method, LOWESS regression, qspline and quantile normalization [81–84]. For quantile normalization (*rma*) the highest value from each array is replaced by their average, then the next-highest value from each array is replaced by their average and so forth. Hence quantile normalization forces the intensity distributions to be equal and can be thought of as a one of the most rigorous normalization methods.

For DNA microarrays that utilize several probes per gene or exon, the intensities are condensed into one value termed an expression index. As of today several methods have been proposed, with the *rma* approach shown to be the best performing. Here a robust fitting technique that protects against outliers, median polish, is used to condense probe level intensities to expression indexes [78].

Statistical testing

In addition to the variations described above biological systems contains stochastic noise, emphasizing the need for statistical testing to identify the true biological effects. For this parametric statistical tests such as the t-test and the ANalysis Of VAriance (ANOVA) are often used, testing either the means or variance to assess the null hypothesis. If the null hypothesis, that the means are equal, is rejected

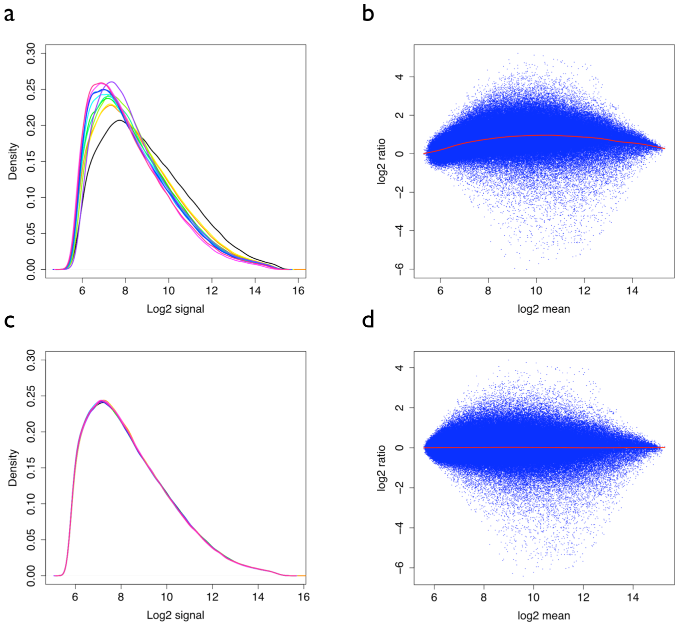


Figure 2.3 – Effect of normalization on 12 Affymetrix Gene Chip Mouse genome 430 2.0 arrays (Chapter 6). **a**. Density plot of raw intensities before normalization for each array, showing differences in the intensity distributions. **b**. MvA plot comparing two samples. The $\log_2$ mean for each probe is plotted vs. the $\log_2$ ratio of each probe. Deviations from 0 shows non-linear relationship between the samples. The red line is a fitted spline. **c**. Density plot of qspline normalized intensities where intensity distributions have been normalized and are now highly similar. **d** MvA plot as in b, but data normalized with qspline. The signal-dependant bias has clearly been removed by the normalization.

(p-value ∼0) the gene is found to be differentially regulated. One-way ANOVA can be applied when multiple levels of the same factor is tested, such as treatment of cells with three or more types of stimulants and time series. Factorial ANOVA can be used when two factors are analyzed, normally in a 2 by 2 factorial experiment. An example of this could be a design with a genetic factor and an environmental factor such as wild-type vs. gene knockout and medium vs. medium + alcohol and glucose, such as used in Paper II (Chapter 4).

The assumptions for the tests are that the populations compared follow the normal distribution, have equal variance (the Student's t-test) and are independent. Additionally large sample sizes are needed to generate enough statistical power, however DNA microarray experiments typically only have three or a few more biological replicates. As the assumptions are rarely fulfilled, the p-value is therefore not strictly trustable and false positives are an issue in DNA microarray experiments. A solution for the normality issue is to use non-parametric tests that does not assume normality, however these have lower statistical power. Instead the statistical tests are used for ranking the data in a meaningful way. Additionally tests based on Bayesian statistics that are more robust to small sample sizes have been developed for DNA microarrays [85].

As the statistical test is applied for each gene on the microarray this gives rise to the issue of multiple testing. When a hypothesis is tested multiple times, e.g. ~30.000 times for an array covering the human genome, 300 of the genes would be expected to have an p-value < 0.01. Therefore the significance values must be adjusted for multiple testing to control the number of False Positives (FP, type I errors). Two methods that has been suggested for this are by Bonferroni and Benjamini-Hochberg, however these are often too strict for DNA microarray data [86]. However through resampling the statistics by permuting the sample classification, it is possible to estimate the False Discovery Rate (FDR) from the data. The number of known true null hypotheses (from permutations) allows the estimation of the FDR at given number of accepted genes as,

$$P_i = \frac{FP_q}{i} \tag{2.1}$$

where i is the number of accepted genes, P_i is the FDR at i and FP_q is the number of true null hypotheses from the permutations at the observed p-value q or lower. Using this approach it is possible to control and accept a certain FDR within the accepted genes.

2.4 Detecting trends

The primary output of gene expression profiling experiments are long lists of tables with significant differentially expressed genes. The data, perhaps containing hundreds of differentially expressed genes, has to be interpreted in the biological context of the experiment to enable a refined or new hypothesis to be formulated. Several tools have been developed to enhance the biological interpretation of microarray data and here classical tools such as Principal Component Analysis (PCA), clustering and annotation enrichment will be discussed.

Principal Component Analysis

For identification of trends in DNA microarray experiments visualization of the data is useful. However an experiment covering m genes each measured using n arrays creates a $m \times n$ -dimensional space, which it is not possible to visualize directly. This may be solved by dimension-reduction techniques such as PCA that projects high-dimensional data to a low-dimensional space. This is performed by computing the first Principal Component (PC1) along the axis with the most variation within the data followed by the second Principal Component (PC2) as the axis containing the maximum variation orthogonal to PC1. For DNA microarray experiments PCA is implemented as Singular Value Decomposition (SVD),

$$X = USV^T \tag{2.2}$$

where X is the expression data matrix, U are the left eigenvectors (eigenassays), V are the right eigenvectors (eigengenes) and S are the singular values. The use of SVD is exemplified in an experiment where mouse dendritic cells are stimulated with two different bacteria in a 2x2 factorial ANOVA designed experiment (Chapter 6). Figure 2.4a shows the information content in the singular values, showing that the first dimension contains by far the most information. From Figure 2.4b it can be seen that PC1 captures the variation between untreated and treated cells (illustrated by the size of the circles) and that PC2 and PC3 together contain the difference in response between the treatments. This is observed by Z9 treated cells showing a similar response to NCFM + Z9 treated cells (yellow and green in lower right corner) whereas NCFM treated are in the diagonal corner from these.

Clustering

In addition to PCA, cluster analysis may be used to reduce the dimensionality of the data and to identify correlations within an experiment. Additionally clustering approaches can be used to classify samples based on gene expression profiles, an approach widely used for different cancer types or sub-types [87].

Clustering is based on calculating the distance between observations using a distance measure. Several distance metrics and inter-cluster distance methods are available, where the most commonly used distance metrics are euclidean, angle vector and pearson correlation distance. These metrics determine how distance is calculated, whereas inter-cluster linkage distance determine where in the clusters the distance is calculated between. For expression data complete and average linkage are preferred methods [88].

For this work Partitioning Around Mediods (PAM) clustering has been used. It is similar to k-means clustering, but is more robust as it minimizes dissimilarities instead of euclidean distances [89]. One object is selected as mediod for each of the k clusters and the objects are clustered based on minimizing the sum of dissimilarities to the mediods in iterative steps. The Pearson correlation coefficient (r_{xy}, eq. 2.3) is used to calculate the correlation between points $\mathbf{x}$ and $\mathbf{y}$, which will take a value between -1 and 1. This corresponds to genes anti-correlating or correlating, respectively. As distance measures are positive, the dissimilarity metric is calculated by transforming r_{xy} to between 0 and 1, where genes with similar profiles will have the distance of 0 and dissimilar genes 1 (eq. 2.4).

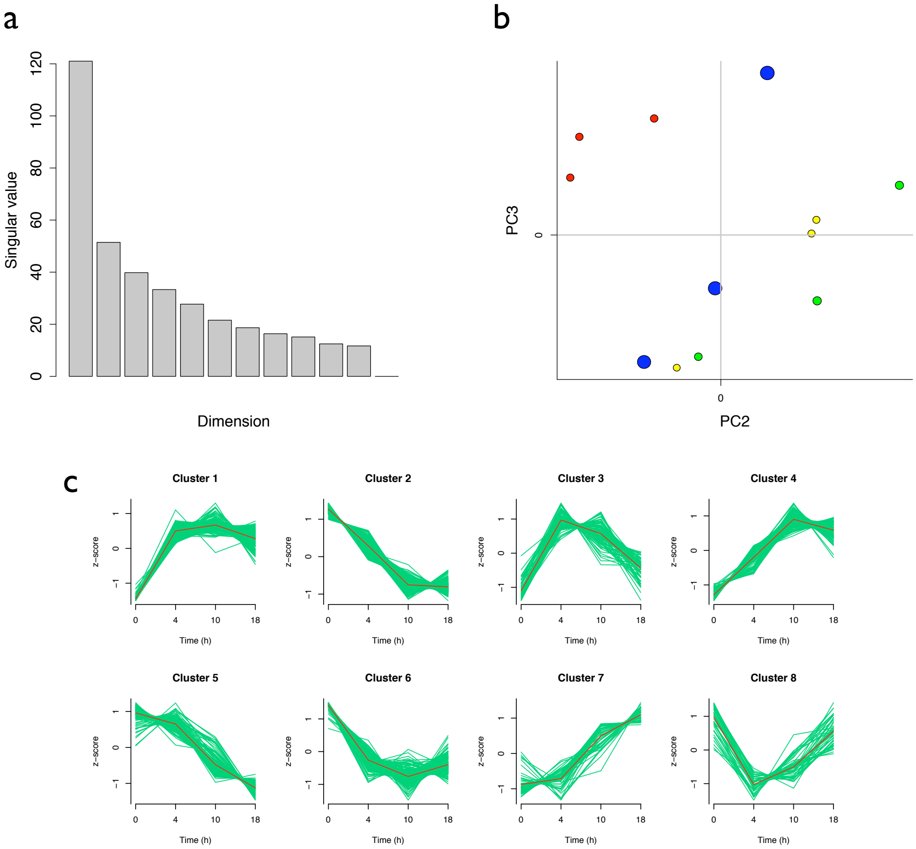


Figure 2.4 – Singular value decomposition and clustering of 12 mouse samples (Affymetrix). The experiment is designed as a 2x2 factorial ANOVA with *L. acidophilus NCFM* and *B. bifidum Z9* as the two factors. **a**. Individual components of SVD and the singular value (information content). **b**. SVD plot of the 12 samples of the first three components. x-axis: PC2, y-axis: PC3, size: PC1. Blue: untreated, red: NCFM, green: Z9, yellow: NCFM+Z9. **c**. PAM clustering of a time-course experiment where mouse DCs are stimulated with *L. acidophilus NCFM* for 0, 4, 10 and 18 hours.

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{2.3}$$

$$dis(x, y) = r_{xy} \times (-0.5) + 0.5 \tag{2.4}$$

An example of PAM clustering on the 1000 most significant genes in a time course experiment from Paper IV (Chapter 6) is shown in Figure 2.4c, where different time-resolved responses are captured.

Annotation enrichment analysis

Co-regulated genes are likely to be involved in similar biological processes. To systematically investigate this, ontologies such as Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) can be used [90, 91]. GO is a collection of controlled vocabularies describing the biology of a gene product in any organism. It is organized in three independent sets Molecular Function (MF), Biological Process (BP) and Cellular Component (CC), each in a tree structure of parent and child terms. Each ontology is structured as a directed acyclic graph where the specificity of the term increases as one moves down the network. KEGG is a database of protein networks and a chemical space providing a different angle than GO.

For DNA microarray analysis these ontologies are useful for identifying enriched terms in e.g. a cluster of co-expressed genes. This can be done by sampling without replacement using the hypergeometric distribution, also known as a hypergeometric test or a one-tailed Fisher's exact test. Other approaches are Gene Set Enrichment Analysis (GSEA) or Parametric GSEA (PGSEA) that aims at identifying changes in minimally changed gene expression experiments [92, 93]. In this approach pre-defined gene sets, such as GO terms or Molecular Signatures Databases (MSigDB) [92] are tested for induction or repression using *all* genes on the array. In Paper IV we use a defined gene set from GO to show that a viral response is significantly induced in the experiment. These approaches are useful as a parallel approach to GO term enrichment, but also for comparison of different data sets. Comparing data sets at a pathway level is more general than at gene-level and is more likely to yield interpretable results.

2.5 Tiling arrays

When designing arrays that tile regions or the entire genome of an organism, probe affinity for the target becomes increasingly important. The freedom in chromosomal location of a probe is limited, especially if the array is designed with overlapping probes. This makes it hard to avoid features such as cross-hybridization, secondary structure, areas of low complexity and irregular probe-affinities derived from base-content. Regarding base composition, especially guanine and cytosine (GC) bases have higher binding energies and contributes more to probe $\mathrm{T_m}$ compared to adenine and thymine. Solutions for this problem can be to normalize the signal from each base in an iterative quantile normalization procedure or to use genomic DNA hybridization as a reference [94, 95]. The design used in Paper I is based on NimbleGen 375k arrays with probe lengths varying between 45 and 65 nt, making it possible to design iso-thermal probes.

Segmentation

A critical step when analyzing expression data from tiling arrays is to identify transcript boundaries. This can be performed by segmentation methods where breakpoints (or changepoints) are identified dividing the data in segments. The problem is related to aCGH where areas of chromosomal copy number variations (CNVs) have to be identified. To assess which approach to use we have tested several existing segmentation methods: Circular Binary Segmentation (CBS) part of the *DNAcopy* package developed for aCGH, Structural Change Model (SCM) from the *tilingArray* package, and a Hidden Markov Model (HMM) based approach developed by our collaborators using the data generated in Paper I [96–98]. The benchmark of these are shown in Chapter 3. We decided to extend the SCM model in an approach, in the following termed *TAR* method, and use this for segmentation of our data. As it is based on SCM a short description of this is given below.

SCM models the data as a piecewise constant function of chromosomal coordinates, where z_{ki} is the signal from the k-th probe on the i-th replicate,

$$z_{ki} = \mu_s + \varepsilon_{ki} \qquad t_s \leq k < t_{s+1} \tag{2.5}$$

where μ_s is the mean of the s-th segment, ε_{ki} are the residuals and $t_2, \ldots, t_s$ are the breakpoints on the chromosomal strand. Using dynamic programming the model is fitted by minimizing the sum of squared residuals (from eq. 2.5: $\varepsilon_{ki} = z_{ki} - \mu_s$),

$$G(t_1, \ldots, t_S) = \sum_{s=1}^{S} \sum_{i=1}^{I} \sum_{k=t_s}^{t^{s+1}-1} (z_{ki} - \hat{\mu}_s)^2 \tag{2.6}$$

where S is the number of segments, I is the number of replicate arrays and $\hat{\mu}_s$ is the arithmetic mean of z_{ki} in segment s. That minimizing the sum of residuals is a good approach for defining breakpoints (segments) is exemplified in Figure 2.5. An example of the outcome of the segmentation methods is shown in Figure 2.6.

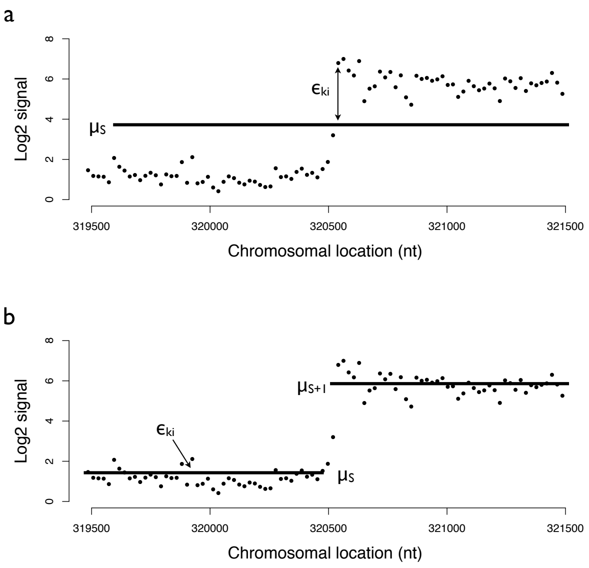


Figure 2.5 – Fitting the SCM model to *B. subtilis* tiling data from 320 to 321 kb with (**a**) one segment or (**b**) two segments. Segment mean is given by μ_S and residuals for individual probes per replicate array are given by ε_{ki}. By minimizing the sum of squared residuals (eq. 2.6) the breakpoints (segment boundaries) can be identified, in this case S = 2 (**b**).

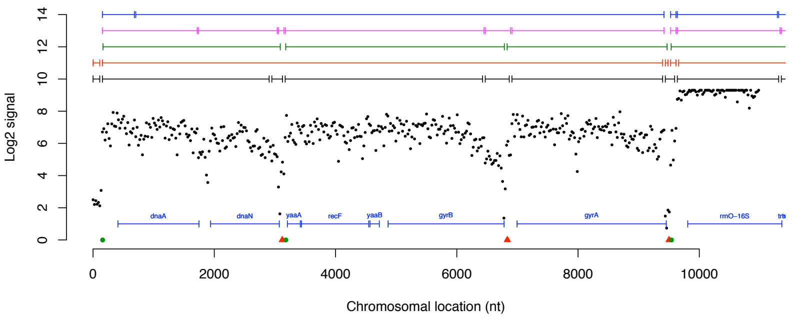


Figure 2.6 – Segmentation of the first 10kb of *B. subtilis* positive strand using different methods. Black: CBS, red: HMM, green: TAR, magenta: SCM, blue: SCM+gDNA norm. For the HMM and SCM+gDNA normalization methods, the underlying data is different from the data shown due to different normalization methods. Known transcripts are: *dnaA*-*dnaN*, *yaaA*-*recF*-*yaaB*-*gyrB* and *gyrA*. Red triangles: known Rho-independent terminators, green circles: known SigA binding sites.

Part II

Systems biology in industrial biotechnology

Chapter 3

Genome-wide tiling of *B. subtilis*

This chapter describes our efforts in reporting the first tiling array data from the Gram-positive model organism *Bacillus subtilis*. *B. subtilis* is a rod-shaped bacterium, able to form endospores and naturally found in soil and on plants. It is generally regarded as safe and is used in cooking for fermenting bean products around the world, such as *B. subtilis* natto used for the traditional Japanese dish, *natto*. It is widely used in the industry for enzyme production as it is safe, easy to transform with foreign DNA and has excellent protein export capabilities.

B. subtilis is closely related to the pathogen *Bacillus anthracis* of which the spores are the cause of anthrax. Additionally it is related to other pathogens, one such being *Staphylococcus aureus* that is becoming an increasing burden for the health care system due to the emergence of multidrug-resistant strains. This work was performed as a part of the BaSysBio EU-FP6 project.

Experimental considerations

Initially we wanted to perform the experiment using cell cycle synchronized *B. subtilis*. To date the cell cycle of living organisms have been studied using transcriptomics for the majority of the branches of life, such as human cells, *Saccharomyces cerevisiae*, *Schizosaccharomyces pombe* and *Arabidopsis thaliana* for eukaryotes, *Caulobacter crescentus* for gram-negative bacteria and *Sulfolobus acidocaldarius* for archaea [50, 99–103]. However no study has yet been performed on gram-positive bacteria making this an experiment worth striving for. By using tiling arrays we would, a part from the traditional analysis of protein-coding genes, have been able to identify small RNAs and antisense transcripts expressed in a cell cycle manner. The synchronization approaches that we attempted were using a temperature sensitive mutant in DNA replication initiation, Percoll gradient fractionation and centrifugational size selection [104–107]. Additionally I visited Prof. Stanley Brul's lab at Universiteit van Amsterdam in the hope of using germinating spores to establish cell cycle synchronized cultures [108], however again without satisfying results. The requirements for running the experiments were to achieve two consecutive cell cycles synchronized, however from the different approaches tested we were able to synchronize one cell cycle regularly with the cells going out of sync in the second cycle. However, while performing test experiments of the tiling arrays, which we designed for the BaSysBio project, we discovered that the data was of high quality. We therefore decided to thoroughly analyze these data assaying the *B. subtilis* transcriptome during exponential growth in rich (Luria-Bertani, LB) and poor media (M9).

Tiling arrays

Using OligoWiz 2.0 [109] we designed tiling DNA microarrays with overlapping isothermal probes covering the entire genome of *B. subtilis* with a spacing of 22 nt on each strand. In combination with a strand specific labeling protocol developed by NimbleGen, we are able to generate a comprehensive dataset of transcriptional signals from both chromosomal strands. As the arrays were designed and run before the re-sequencing of *B. subtilis* (AL009126.2) we re-annotated the probes to the updated sequence (AL009126.3) at the time of publication. The effects are fairly small with only ∼300 probes of 375.000 that could not be mapped due to low homology. As the re-sequencing also introduces ∼2.000 extra nucleotides in the genome seven small regions have gaps with the most affected being *trpF*. This is in good agreement with the fact that the laboratory strain used for more than 25 years is a tryptophan auxotroph, and that the *trp* region has not been under evolutionary pressure.

Segmentation methods

To perform the segmentation of our data we tested three different existing methods and the TAR method that we developed. The latter was an adaption of the SCM method addressing the issue, that the number of segments S has to be guessed or estimated using a penalized log-likelihood criteria, however this is not straightforward for real biological data [96]. Instead by deliberately overestimating the number of segments and introducing a joining step the issue is avoided. The algorithm for this is to accept all segments with mean signal (μ_s) above background and then join neighboring segments if, (a) five probes on each side of the breakpoint

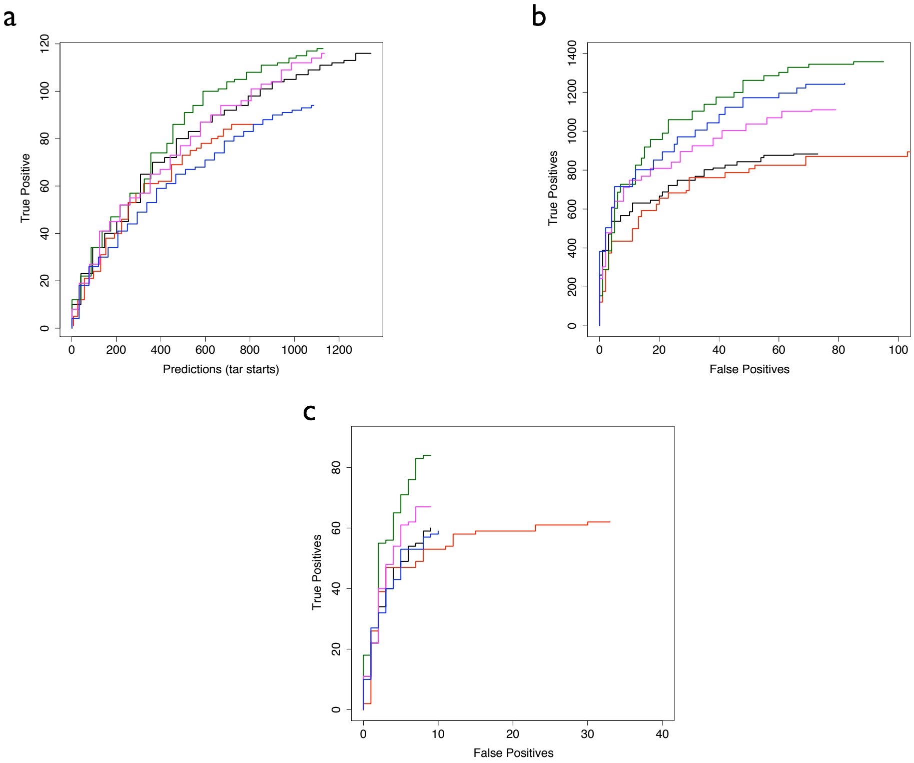


Figure 3.1 – Benchmark of segmentation methods using *B. subtilis* tiling data, positive strand. Black: CBS, red: HMM, green: TAR, magenta: SCM, blue: SCM+gDNA norm. For the SCM approach the number of segments used were the same as for the TAR approach. **a**. Known Transcription Start Sites (TSS) identified within 20 nt. **b**. Known genes as they are identified. True positives: known annotated genes identified fully inside a segment, false positives: known annotated genes but on the opposite strand inside a segment. **c**. Known transcripts from Hoon et al. [110]. True positives: Segment containing known transcript within 100 nt upstream and downstream of breakpoints. False positives: Segment containing known transcript on the opposite strand.

are above background, and (*b*) a Student's t-test does not reject the hypothesis that probes in each segment belong to the same signal-intensity distribution with a p-value $> 1 \times 10^{-10}$. To test the performance of this and the other methods we performed benchmarks against (*a*) known Transcription Start Sites (TSS), (*b*) annotated genes inside segments, and (*c*) known transcripts (Figure 3.1). The SCM methods were benchmarked using the number of segments identified from the TAR approach.

From this the best performing method was the found to be the TAR approach, whereas the HMM approach generally performed poor. The poor performance of the latter approach was probably due to estimating too long segments, as it shows similar performance in the TSS benchmark (Figure 3.1a), whereas benchmark against known genes (b) and known transcripts (c) had poor performance. Additionally we found that normalization with gDNA hybridization signal did not improve SCM predictions. This is possibly due to local structures such as Rho-independent terminators that can form hairpins and hereby avoid detection. The extent of this is likely to vary between organisms depending on the use of Rho-independent terminators – with regard to *B. subtilis* transcriptional control is to a large extent dependent on this phenomenon [111].

3.1 Paper I

From our experiments we are able to provide a genome-wide view of RNA expression combined with Rho-independent terminator and sigma factor binding site predictions and annotations. We identify putative novel non-coding RNAs and reveal that antisense transcription may be much more pronounced than previously thought – prior to this study only two antisense transcripts were known in *B. subtilis.* Supplementary material are found in Appendix A. Included are Fig S1, Fig S3-9 and 6 pages of Fig S2 (the pages mentioned in the paper). Full supplementary figures and tables are available at the publishers web-site as *Online Open* material: `http://www3.interscience.wiley.com/journal/122536032/suppinfo`.

Molecular Microbiology (2009)

doi:10.1111/j.1365-2958.2009.06830.x

The transcriptionally active regions in the genome of *Bacillus subtilis*

Simon Rasmussen, Henrik Bjørn Nielsen and Hanne Jarmer*
Center for Biological Sequence Analysis, Department of Systems Biology, Technical University of Denmark, 2800 Lyngby, Denmark.

Summary

The majority of all genes have so far been identified and annotated systematically through *in silico* gene finding. Here we report the finding of 3662 strand-specific transcriptionally active regions (TARs) in the genome of *Bacillus subtilis* by the use of tiling arrays. We have measured the genome-wide expression during mid-exponential growth on rich (LB) and minimal (M9) medium. The identified TARs account for 77.3% of the genes as they are currently annotated and additionally we find 84 putative non-coding RNAs (ncRNAs) and 127 antisense transcripts. One ncRNA, *ncr22*, is predicted to act as a translational control on *cstA* and an antisense transcript was observed opposite the housekeeping sigma factor *sigA*. Through this work we have discovered a long conserved 3′ untranslated region (UTR) in a group of membrane-associated genes that is predicted to fold into a large and highly stable secondary structure. One of the genes having this tail is *efeN*, which encodes a target of the twin-arginine translocase (Tat) protein translocation system.

Introduction

The bacterial genome is a highly compact structure. Both strands are densely covered by genes, of which a large part is organized into the even more gene-dense arrangements of operons. Recent technological advances have allowed for an empirical assessment of the prevalence of transcriptionally active regions (TARs) across an entire genome – by the use of either high-throughput sequencing of RNA-derived cDNA (Nagalakshmi *et al.*, 2008) or high-density oligo-nucleotide tiling arrays (Tjaden *et al.*, 2002; Bertone *et al.*, 2004; David *et al.*, 2006; Li *et al.*, 2006; Reppas *et al.*, 2006). Where the studies of Tjaden and co-workers and Reppas and co-workers investigated the transcriptional landscape of *Escherichia coli*, we report here the first findings of a high-density tiling-array study performed on the Gram-positive *Bacillus subtilis*. *B. subtilis* was first described in 1835 by the German scientist Christian Gottfried Ehrenberg as the hay/grass-associated bacterium, *Vibrio subtilis* (Ehrenberg, 1835). In 1872 another German scientist, Ferdinand Julius Cohn, renamed it *Bacillus subtilis* (Cohn, 1872). In 1876 Cohn showed for the first time that *B. subtilis* is capable of changing into an endospore state, and hereby surviving environmental changes not suitable for vegetative growth (Cohn, 1876). In 1930 the American bacteriologist, Harold Joel Conn, published a description of the Marburg strain of *B. subtilis* (American Type Culture Collection No. 6051) (Conn, 1930; Teas, 1949) and in 1947 this particular strain was subjected to both X-rays and UV light by Burkholder and Giles (Burkholder and Giles, 1947; Teas, 1949). Charles Yanofsky provided a number of stable auxotrophs, which had been isolated from these experiments, to John Spizizen (Spizizen, 1984), which studied their ability to develop natural competence (Spizizen, 1958; Zeigler *et al.*, 2008). Further investigations resulted in the development of a highly efficient two-step protocol for transformation of the #168 strain (Anagnostopoulos and Spizizen, 1961), a success drawing the attention of the research community to such an extent that this strain was selected as the *B. subtilis* model strain. Today *B. subtilis* is widely used as an industrial production strain, and has even been shown to possess probiotic properties (Huang *et al.*, 2008). And now, more than 10 years after fully sequencing and annotating the genome the first time (Kunst *et al.*, 1997) and only shortly after the recent re-sequencing (Barbe *et al.*, 2009), we experimentally validate and extend these efforts.

Results and discussion

Identification of transcriptionally active regions (TARs)

Hybridization of labelled RNA to densely tiled microarrays allows for a high-resolution mapping of genome-wide expression on both strands, and we have found that

Accepted 23 July, 2009. *For correspondence. E-mail hanne@cbs.dtu.dk; Tel. (+45) 45 25 61 48; Fax (+45) 45 93 15 85.

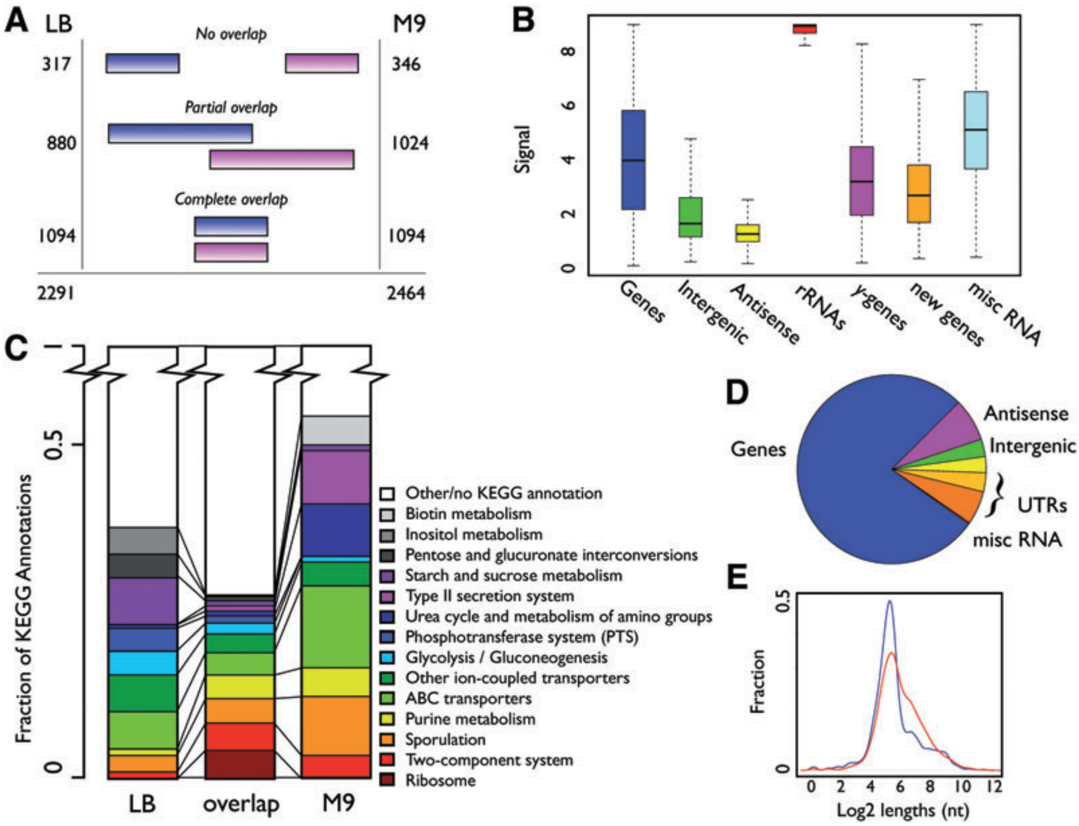


Fig. 1. Expression in LB and M9.
A. Diagram showing the overlap between TARs identified in the two media. No overlap: less than 5% overlap; Partial overlap: between 5% and 85% overlap (can overlap multiple TARs); Complete overlap: overlap of 85% or more.
B. Box plot showing the log2-transformed signal range of the probes within annotated genes (non-y-genes), the regions between genes (Intergenic), antisense to known annotation, rRNAs, y-genes, new genes and misc RNAs as by the re-annotation by Barbe *et al.*
C. Representation of the top 14 KEGG terms from genes uniquely expressed in LB and M9, and genes common to the two media.
D. Pie chart illustrating the physical position of the probes returning a signal above background. Blue: ORF/gene; dark orange: 5′ UTR; orange: intergenic UTR; yellow: 3′ UTR; green: intergenic region (IR); magenta: antisense; red: misc RNA (Barbe *et al.*, 2009).
E. Density plot showing the log2 lengths (nt) of 5′ UTRs (red) and 3′ UTRs (blue) as they are determined in the study.

during growth in rich medium (LB) *B. subtilis* expresses 2291 transcriptionally active regions (TARs), whereas the corresponding number using minimal medium (M9) is 2464 TARs (all listed in Table S1). To determine how many of these were unique we have calculated the TAR overlap between the two conditions (see Fig. S1). If less than 5% of the TAR overlapped we define it as a unique TAR and likewise we define a common TAR if more than 85% overlap. This leads to the identification of 1094 common TARs, whereas 317 and 346 TARs are unique for LB and M9 respectively (Fig. 1A). In total 3662 non-redundant (overlap < 85%) TARs have been identified. An overview of the results in terms of identified genes, gene-like features and TARs can be seen in Table 1. Additionally we have annotated the TARs with experimentally verified and HMM predicted sigma factor binding sites and experimentally verified and predicted Rho-independent terminators (see *Experimental procedures*). A total of 10.5% and 27.3% of the TARs have been annotated with at least one experimental or predicted sigma factor binding site, respectively, and similarly 6.2% and 54% with an experimental or predicted Rho-independent terminator. Together the identified TARs account for 77.3% of the genes as they are currently annotated, and the overlap between the two media is 2843 genes corresponding to 64% of the 4422 known genes (Table S2). The whole-genome expression data, along

Table 1. Overview of current annotation and the findings in this study.

Type	Current annotation[a]	LB	M9	Unique
Total CDS	4244	3189	3074	3420
Genes	1912	1514	1469	1627
y-genes	2332	1675	1605	1793
New genes	171[b]	106	103	119
rRNA	30	30	30	30
tRNA	86	82	83	83[c]
ncRNA	16[d]	50	68	84
Antisense	2[e]	60	99	127
TARs	–	2291	2464	3662[f]

a. GenBank (AL009126.3).
b. Genes annotated as new from Barbe *et al.* (2009).
c. The missing tRNAs are: *trnD-Leu2*, *trnSL-Arg1* and *trnSL-Arg2*.
d. Ando *et al.* (2002); Suzuma *et al.* (2002); Licht *et al.* (2005); Silvaggi *et al.* (2006); Gaballa *et al.* (2008); Saito *et al.* (2009).
e. Silvaggi *et al.* (2005); Eiamphungporn and Helmann (2009).
f. TARs were unique if the overlap was less than 85% between the two conditions.
Columns LB and M9 show number of occurrences in that particular category and Unique are unique occurrences in the two media combined.

with the predicted transcripts, sigma factor binding sites and Rho-independent terminators, are visualized in a figure spanning 48 pages (Fig. S2) and we encourage the reader to explore the findings.

Hybridization of labelled genomic DNA (gDNA) to the tiling array results in a uniform signal level throughout the genome (as can be seen in Fig. S2). However, we found that low gDNA signals coincide with experimentally verified and predicted Rho-independent terminators (Fig. 3A). This may be explained by the formation of stable structures possibly forming in both the probe and the target, which hereby prevents detection (Ratushna *et al.*, 2005). These findings may also explain why normalization using gDNA hybridizations, as performed by Huber *et al.*, did not improve the performance of our TAR findings (data not shown) (Huber *et al.*, 2006). This normalization are in areas with Rho-independent termination introducing significant noise and interferes with the determination of correct transcript boundary.

We have benchmarked the prediction of TARs against gene coverage, known transcription start sites (TSSs) and signal autocorrelation (see Fig. S3). From this we see that of 2500 genes predicted to be covered by TARs, only ~2.5% are estimated to be false positives, here defined as TARs covering genes expressed at the opposite strand (not taking possible antisense transcripts into account). Regarding TSS, our findings are in general within 20 nt from the experimentally verified starts. Additionally it is interesting to note that we do observe a spatial gene expression dependence – neighbouring genes tend to be coexpressed in operons. Experimentally we verify the TSSs of five of the determined transcripts using RNA ligase-mediated rapid amplification of cDNA ends (RLM-RACE) and the results are summarized in Table S3. The verified transcript start sites are within 30 nt of our findings.

Comparison in gene utilization using two different growth sources

The distribution of the most common KEGG annotations (Kanehisa *et al.*, 2008) are shown for genes expressed in both media (common) compared with genes expressed uniquely to either of the conditions. As expected, it becomes evident that *B. subtilis* utilizes different pathways when growing in the two different media. One example is the difference in the *Glycolysis/gluconeogenesis*, where a closer inspection reveals that the gluconeogenesis is inactive when the cell is growing in minimal medium, which is expected (Fillinger *et al.*, 2000). Likewise, a large portion of genes involved in the development of competence (with the KEGG annotation *Type II secretion*) is active when the cell is starving. Whereas, many of the genes exclusively expressed when the cell is growing in the rich medium include a large proportion of genes encoding products responsible for uptake and metabolism of various carbon sources, which is expected from growth in a complex medium (Deutscher *et al.*, 2002). A puzzling observation is that sporulation genes, based on KEGG annotation (Fig. 1C), are seen expressed at conditions when sporulation should not occur. However when investigated in detail it becomes clear that the majority of these are expressed at levels close to our detection limit and that the few highly expressed are sporulation initiation control genes such as response regulator aspartate phosphatase genes/operons (Auchtung *et al.*, 2006). To further ensure that sporulation is indeed not occurring we have analysed the expression of the sporulation regulons σ^F, σ^E, σ^G and σ^K (Steil *et al.*, 2005) and reassuringly we find that all of these are expressed below background (shown in Fig. S4). We therefore contribute the above phenomenon to genes involved in sporulation control and/or genes with divergent functionality.

Determination of untranslated regions (UTRs)

Forty per cent of the probes tiling the genome give a signal above background. As is shown in Fig. 1D the majority of these seemingly expressed elements are generally localized within regions expected to give a signal, either within an annotated gene or in the putative untranslated regions (UTRs) – as they have been determined in this study. The majority of the probes located in the intergenic regions (IRs) and the antisense regions (ARs) have signals below background level. Additionally 5.6% of the probes with signal above background fall within putative 5′ UTRs as they are determined in this study, whereas probes in the 3′ UTRs only comprise 2.7% of the expressed probes, which is even fewer than for intergenic UTRs (3.2%). This corresponds to 1648/1633 (LB/M9) transcripts with a defined 5′ UTR and 1371/1506 (LB/M9) with defined 3′ UTRs. The majority of this difference between 5′ and 3′ UTRs may be explained by their difference in length, as is shown in Fig. 1E. The median lengths are 47 and 36 nt for 5′ and 3′ UTRs respectively (significant in a two-sided Wilcoxon rank sum test with a *P*-value of 1×10^{-24}). This is opposite to what is observed in higher organisms, such as in the study of David *et al.* (2006) in *Saccharomyces cerevisiae*, where the 3′ UTRs are found to be longer than the 5′ UTRs. It does however correspond well to the previous discovery that the average length of the 3′ UTRs is increasing as a function of the organismal complexity (Mazumder *et al.*, 2003). These findings point at emphasis on 5′ UTRs or lack thereof on 3′ UTRs compared with higher organisms in transcriptional and post-transcriptional control in *B. subtilis.* Already well-studied examples of such 5′ UTR-mediated control in *B. subtilis* is the control of the tryptophan operon and *S*-adenosyl methionine (SAM) riboswitch (Grundy and Henkin, 1998; Gollnick *et al.*, 2005).

Novel protein-coding genes

The new annotation by Barbe *et al.* has identified 171 putative novel protein-coding genes increasing the amount of protein-coding genes in *B. subtilis* to 4244 and here we report the first expression data covering these. In general the novel protein-coding genes are expressed at lower signals compared with the remaining protein-coding genes with expression means of 3.0 and 3.9 respectively (Fig. 1B). Additionally only 70% are found to be expressed above background signal, which is less than the protein-coding genes in general (77%). This combined with the short lengths of the newly annotated genes (median 159 versus 258 aa for remaining) and the fact that some of these were found to have sequence errors explains why these have not been annotated before.

We have investigated whether the novel protein-coding genes are expressed mono- or polycistronic and we find that 26 of the 119 expressed genes are encoded monocistronic, which may provide experimental evidence for the existence of these genes. The novel protein-coding genes are listed in Table S4 together with their expression values and the genes annotated to the TAR they belong to.

An interesting monocistronic expressed new gene is *ybzH*, within the *pro1* prophage-like element (see Fig. 2A), positioned on a transcript with clearly defined boundaries. RLM-RACE mapped the TSS to 7 nt downstream of our observed boundary and exactly at a predicted +1 of SigA factor binding site. Additionally, a Rho-independent terminator is predicted at the transcript end. Regarding functionality, Barbe *et al.* reported BLAST hits with high similarity to proteins of the arsenic resistance transcriptional regulator family (ArsR) from different *Bacilli* and *Geobacilli* species. This is in agreement with findings that prophages have been shown to confer protective traits to heavy metals such as arsenic (Cervantes *et al.*, 1994).

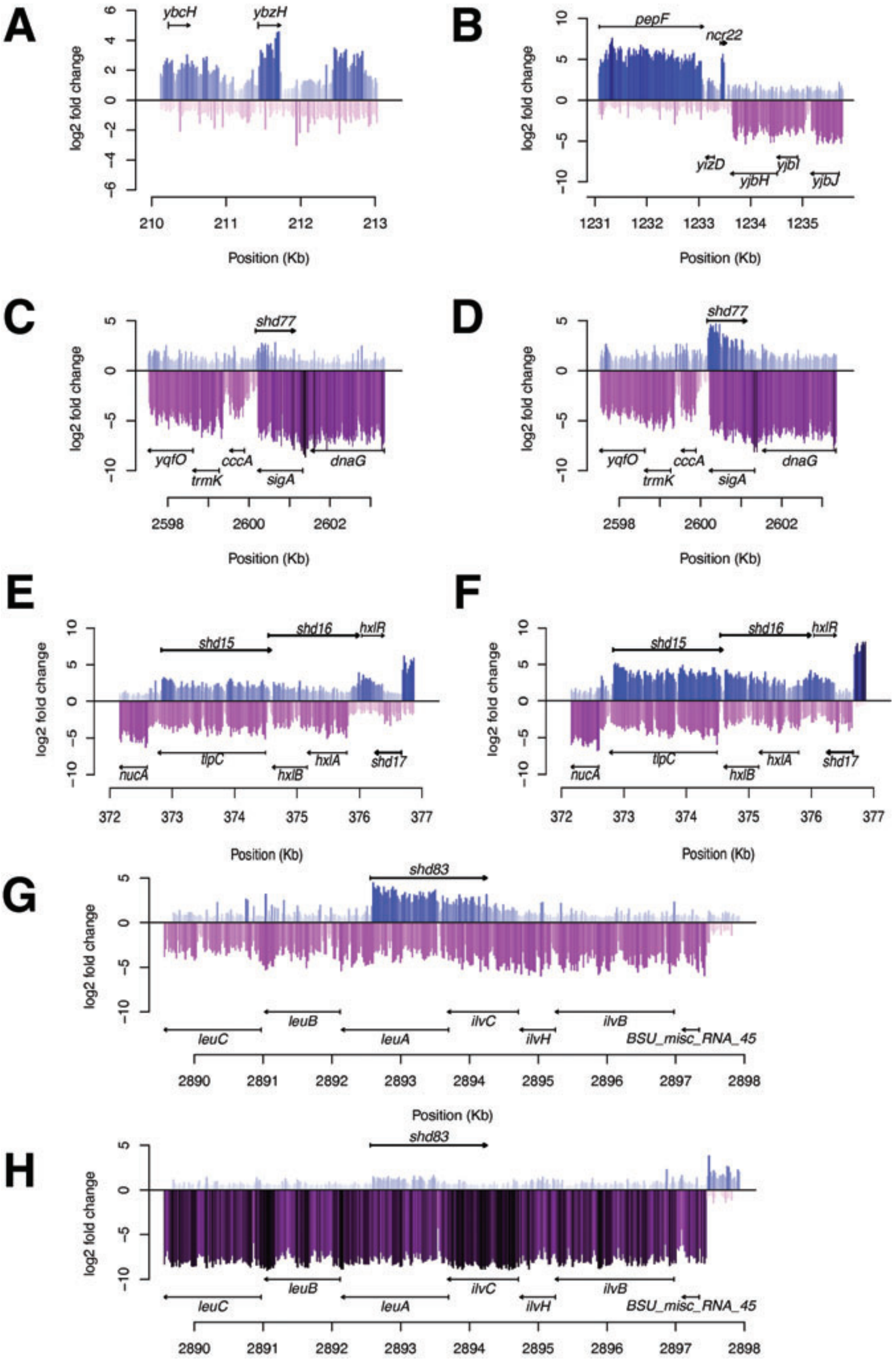
Alternative ORFs as the result of TAR identification

The transcriptional map also uncovers irregularities in the current annotation, e.g. the region containing the annotated translation start site or stop codon is not expressed. The discrepancy in the annotations of these genes might be due to sequence errors at the time of annotation; however, re-sequencing and re-annotation efforts of Barbe *et al.* seem to have corrected several of these irregularities. An example is the *ykvS* gene that was re-annotated from 143 to 62 aa and is now confined within the observed transcript (see Fig. S2, at 1447 kb). Examples of irregularities between gene annotation and TARs are *cgeD*, *ybcL*, *ybcM*, *ycgN*, *yxxF*, *yqjD* and *ydbO* (Fig. S2). However irregularities may also be explained by alternative internal promoters. The latter is the case of the *hisC*, *tyrA* and *aroE* operon which is transcribed from a promoter residing inside the *trpA* gene (see Fig. S2, at 2372 kb) (Gollnick *et al.*, 2005).

Expression of prophage elements

The data generated here are well suited for a systematic investigation of the prophage elements in *B. subtilis* and

Fig. 2. Expression of different regions of the *B. subtilis* genome during growth, where the position and direction of genes are indicated with arrows. Expression on the Watson strand is blue, Crick strand is magenta and the colour intensity also indicates signal strength.
A. Expression in the 210–212 kb region in LB showing new protein-coding gene *ybzH* expressed monocistronic.
B. Expression during growth in LB medium in the region 1231–1236 kb, showing expression of the novel non-coding RNA *ncr22*.
C. Antisense expression in the region 2598–2603 kb in LB (*shd77*) of *sigA*.
D. As (C), but expression in M9.
E. Antisense expression in the region 372–377 kb in LB (*shd15-shd17*) of *tlpC, hxlB, hxlA* and *hxlR*.
F. As (E), but expression in M9.
G. Antisense expression in the region 2890–2898 kb during growth in LB (*shd83*) of the operon *ilvBHC-leuABC*.
H. As (G), but expression in M9.


A
log2 fold change
ybcH
ybzH
210
211
212
213
Position (Kb)
B
log2 fold change
pepF
ncr22
yizD
yjbH
yjbI
yjbJ
1231
1232
1233
1234
1235
Position (Kb)
C
log2 fold change
shd77
yqfO
trmK
cccA
sigA
dnaG
2598
2600
2602
Position (Kb)
D
log2 fold change
shd77
yqfO
trmK
cccA
sigA
dnaG
2598
2600
2602
Position (Kb)
E
log2 fold change
shd15
shd16
hxlR
nucA
tlpC
hxlB
hxlA
shd17
372
373
374
375
376
377
Position (Kb)
F
log2 fold change
shd15
shd16
hxlR
nucA
tlpC
hxlB
hxlA
shd17
372
373
374
375
376
377
Position (Kb)
G
log2 fold change
shd83
leuC
leuB
leuA
ilvC
ilvH
ilvB
BSU_misc_RNA_45
2890
2891
2892
2893
2894
2895
2896
2897
2898
Position (Kb)
H
log2 fold change
shd83
leuC
leuB
leuA
ilvC
ilvH
ilvB
BSU_misc_RNA_45
2890
2891
2892
2893
2894
2895
2896
2897
2898

we have examined the expression of the prophage elements *PBSX*, *SPβ* and *skin*, and the prophage-like elements *pro1–7* (Zahler *et al*., 1977; Wood *et al*., 1990; Takemaru *et al*., 1995; Nicolas *et al*., 2002). The functionality of the genes expressed from the prophage elements during exponential growth would be expected to be involved in control of the bistable lysogenic equilibrium, conferring immunity to the phages or to be functional genes obtained via hitch-hiking. Genes with unknown functions that are expressed during exponential growth from within these elements are then likely not to be inducing the lytic cycle, but confer beneficial traits during growth in the natural habitat (Lazarevic *et al*., 1999).

In previous microarray studies large clusters of genes in prophage elements were found to be expressed at low levels (Helmann *et al*., 2001). Our analysis reveal that this is in particular true for the *skin* element and to some degree for *pro2* and *pro7* (Figs S5 and S6). Characteristic of these clusters is that their expression are at extremely low levels, indicating that even low expression of these genes is undesirable during exponential growth. This trend is not observed to the same extent within the *SPβ* prophage, where there are low, but not non-existing, basal gene expression levels. Additionally the sublancin genes and neighbouring area (*bdbB* to *sunI*) are highly expressed within *SPβ* and exemplify that prophage genes may confer beneficial traits that are not essential. *yolA* in *SPβ* is the highest expressed gene within the prophage elements and is one of the highest expressed in the entire genome (above the 98% quantile). The gene encodes a 155 aa protein predicted to contain a signal peptide and is hence a putatively exported protein.

The prophage and prophage-like elements *PBSX*, *pro3*, *pro4* and to some extent *pro5* show high levels of gene expression. For the *PBSX* element it is in agreement with previous observations (Krogh *et al*., 1996) and coincides with the fact that it has similar base composition to the native *B. subtilis* sequence in contrast to typical AT-rich prophage elements (Nicolas *et al*., 2002). These expression profiles indicate limited phage functionality or viability of *PBSX*, *pro3*, *pro4* and *pro5*, whereas *skin*, *pro1* and *pro7* may contain gene products undesirable during exponential growth. Expression of all genes, including prophage and prophage-like elements, are listed in Table S2.

Identification of novel non-coding RNAs

In order to extract high-confidence new non-coding RNAs we have set-up a list of criteria that should be fulfilled. In total we extract 84 non-coding RNAs from segments that fulfil the following criteria: (i) no annotated transcription according to the latest GenBank version (AL009126.3), (ii) higher signal level than neighbouring segments, (iii) higher signal than the corresponding antisense region, (iv) maximum 5% of the probes cross-hybridize to other regions of the genome (using a BLAT-scoring scheme; Kent, 2002), and (v) if shorter than five probes, the signal should be observed in both media. These putative noncoding RNAs (ncRNAs) (*ncr1–84*) are listed in Table S5. They have a median length of 197 nt and range from 55 to 571 nt. From *E. coli* it is known that the functions of ncRNAs cover a wide range (Kawano *et al*., 2005). Figure S7 shows how conserved the *ncr* genes are across species. We annotate 65% (55) of the *ncr*s with experimental or predicted sigma factor binding sites and 70% (59) with an experimental or predicted Rho-independent terminator.

Of the 16 already known ncRNAs in *B. subtilis*, other than rRNAs and tRNAs, we identify 10: *surA*, *ssrSB*, *ssrSA*, *bsrF*, *bsrG*, *bsrH*, *bsrI*, *fsrA*, *scr* and *ssrA* (Ando *et al*., 2002; Suzuma *et al*., 2002; Silvaggi *et al*., 2006; Gaballa *et al*., 2008; Saito *et al*., 2009). The remaining ncRNAs *bsrC*, *bsrD*, *bsrE*, *surC*, *SR1* and *polC-ylxS* are not identified in our study (Licht *et al*., 2005; Silvaggi *et al*., 2006; Saito *et al*., 2009). Even though we do not identify the *bsrE* transcript, we do find *ncr40* expressed from the opposite strand at the same genomic location (Saito *et al*., 2009). However as there are expression from both strands in this region *ncr40* may be an antisense transcript of *bsrE*. Reasons for the absence of the other RNAs might for *surC* and *polC-ylxS* be that they were identified as being expressed under sporulating conditions (Silvaggi *et al*., 2006). Regarding the *bsrC* and the *SR1* transcripts the regions are expressed (*ydaG-ydaH* and *slp-speA* respectively); however, segments are not identified. The *bsrC* segment is joined with the upstream gene and *SR1* is weakly expressed and is therefore not identified as a segment in our analysis; however, visual inspection reveals a possible transcript at the position (Licht *et al*., 2005). *bsrD* is actually well defined in M9; however, it fails to meet the criteria as it is only two probes and not present in LB (Saito *et al*., 2009). In addition to these non-coding RNAs, 22 riboswitches such as purine, SAM, TPP, FMN, glycine and lysine, and T-box elements are identified as *ncr* elements (Barbe *et al*., 2009). This leaves 54 non-coding RNA elements not previously described.

An example of a novel putative non-coding transcript is *ncr22*, which is located between *yizD* and *yjbH* and is a clearly defined transcript showing high expression in both media (Fig. 2B). Using 5′ RLM-RACE we map the TSS to 18 nt upstream of the observed boundary; however, we are not able to identify a probable sigma factor binding site (Fig. 3B). A Rho-independent terminator sequence is positioned in the 3′ of the transcript where the stem-loop is folded from the last 16 nt of the transcript and 5 nt outside the 3′, and the T-tail following these. This results in

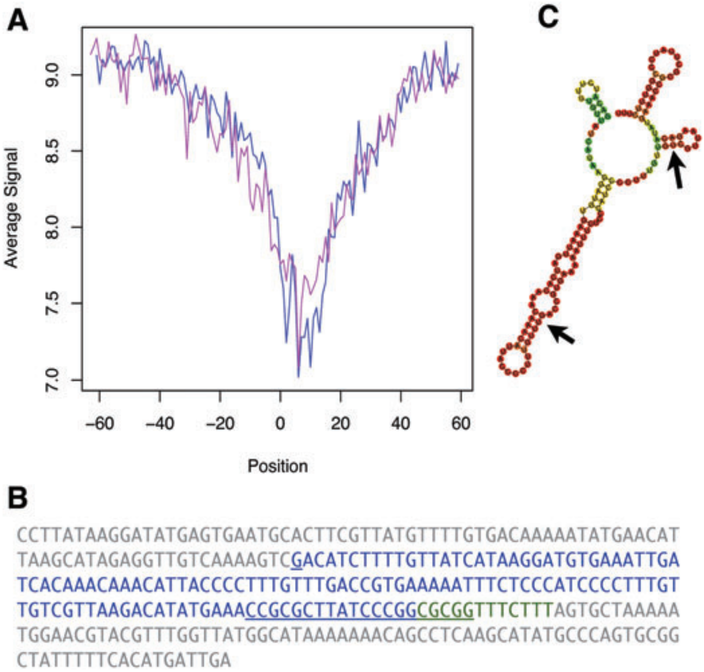


Fig. 3. A. Average genomic DNA signal intensity over Kingsford predicted terminators (Kingsford *et al*., 2007). Position 0 corresponds to the middle nucleotide of the predicted terminator. Blue: Watson strand; magenta: Crick strand.
B. The *ncr22* transcript and ~100 nt upstream and downstream. Grey: intergenic nt; blue: identified transcript; underlined: transcription start site (+1) determined by 5′ RLM-RACE and stem-loop of terminator sequence; green: last part of terminator stem-loop and T-tail not within the identified transcript.
C. Fold of *ncr22* transcript using RNAfold, coloured as base-pair probabilities. Blue equals zero and red equals 1. The two arrows indicate binding sequence to *cstA* transcript upstream of the start codon.

a transcript of 133 nt, which when folded using RNAfold seems to fold into a stable structure with a minimum free energy (MFE) of –33.8 kcal mol^{-1}. Additionally the *ncr22* transcript is highly conserved in the *Bacillus*, *Geobacillus* and *Staphyloccocus* genera providing supporting evidence for this transcript (Fig. S7). As most bacterial ncRNAs act in a *trans*-acting regulatory role, we have searched for mRNA targets for *ncr22* using targetRNA (Tjaden *et al*., 2006; Vogel and Wagner, 2007). Interestingly the best hit is in the 5′ of the carbon starvation-induced protein messenger (*cstA*). The interaction between the two RNAs occur from +13 to –22 (relative to the start codon) in the *cstA* transcript, covering the region containing the Shine–Dalgarno (SD) sequence, and nucleotides 61–95 in *ncr22* (Fig. 3B). In *E. coli cstA* is under translational control of the RNA-binding protein CsrA and the sRNAs CsrB and CsrC (Dubey *et al*., 2003); however, the CsrA homologue in *B. subtilis* does not seem to have binding affinity for the *cstA* transcript (similarity search using CsrA-binding domains; Yakhnin *et al*., 2007). The above suggests that *cstA* may be under translational control in *B. subtilis* not by CsrA but possibly by *ncr22*.

Identification of antisense RNAs

We identify 127 TARs fulfilling the same criteria as for non-coding RNAs, except that they are expressed antisense to already known genes with an overlap of more than 10%. We term these shadow genes and name the TARs *shd1–shd127*; details on these are listed in Table S6. The median length of shadow expressed transcripts is 681 nt and ranges from 197 to 3516 nt.

A possible function of these antisense transcripts is as *cis*-acting regulators, as described by Eiamphungporn and Helmann (2009) for the *yabE* gene and Silvaggi *et al*. (2005) for *yqdB*. In this study we detect antisense expression to *yabE* (*shd4*) during growth in both media and in addition an antisense signal (*shd3*) for the upstream gene *yabD* (Fig. S2, at 49 kb). Likewise we observe the other known *B. subtilis* antisense transcript *ratA* (*shd80*) expressed in both media as antisense to *yqdB* (Silvaggi *et al*., 2005).

As an example of a novel antisense transcript *shd77* should be mentioned since this could potentially be of significant importance as it is expressed antisense to *sigA*, the principal sigma factor during vegetative growth (Haldenwang, 1995) (Fig. 2C and D). Sigma A and E binding sites are predicted at –10 and +10, respectively, of the observed 5′-TAR boundary. This finding adds to the complexity of the regulation of the *yqxD-dnaG-sigA* operon, which is already known to be controlled via at least seven different promoters (Wang *et al*., 1999). Furthermore, we have experimentally verified the TSS of *shd15* (Fig. 2E and F and Table S3) and found it to correspond to the TAR TSS prediction and identify a putative

SigA site with –35: TTGATT and –10: TATGAT. This transcript appears to be one of three antisense transcripts (*shd15–17*) antisense to *tlpC*, a methyl-accepting chemotaxis protein, *hxlAB*, formaldehyde detoxification system and *hxlR*, which encodes a positive regulator of *hxlAB*.

If antisense transcripts are acting as negative *cis*-regulatory elements the signal levels of sense and antisense ratio would be expected to anticorrelate which should be possible to observe if there are differential regulation between the conditions tested. Generally, and in line with expectations, we do see anticorrelation (Pearson correlation –0.22) when comparing antisense and sense ratios (LB versus M9) (Fig. S8). When investigating this for *anti-yabE* (*shd4*) the antisense transcript level increases 1.1 log2-fold (LB to M9) with a concomitant 2.7 log2-fold decline in the sense *yabE* signal. However, this trend may not always be observed if multiple regulatory mechanisms control the sense expression or the area is not differentially expressed, as exemplified by the antitoxin *ratA* (*shd80*) with the log2 antisense and sense ratios of –0.9 and –1.1. The antisense–sense transcript pair with the strongest change observed when comparing LB with M9 medium is *shd83*, which partially overlaps *leuA* and *ilvC* in the *ilvBHC-leuABC* operon (Fig. 2G and H). The products of the operon are enzymes involved in branched chain amino acid synthesis and the full-length mRNA is subjected to transcriptional regulation by tRNA^{Leu} T-box in the 5′ UTR, CodY, CcpA, TnrA and processed into smaller units (Mäder *et al*., 2004; Shivers and Sonenshein, 2005). Due to the many regulatory modes of the *ilvB* operon further experiments are needed to understand whether the observed expression change can be explained by antisense RNA expression.

The fraction of sense coding sequence covered by antisense transcripts seems to be divided in two distributions, transcripts covering close to or full length of genes and another existing of transcripts only partially covering genes (Fig. S8). The groups are exemplified by the two already known antisense transcripts *anti-yabE* and *ratA*, which are predicted to cover 72–80% and ~35% of the coding sequences respectively (see Fig. S2, at 49 and 2678 kb, and Table S6; Silvaggi *et al*., 2005; Eiamphungporn and Helmann, 2009).

In addition to this some of the antisense transcripts seem to be UTRs that overlap genes on the opposing strand. We annotate eight of the transcripts as putative overlapping 5′ UTRs and 26 as putative 3′ UTRs. A closer inspection of the latter reveals that 35% of these have start sites in a 50 nt range of an experimental or predicted Rho-independent terminator (for such an example see *shd49*, Fig. S2 at 1261 kb). This suggests that some 3′ UTR antisense transcripts may arise from terminator read-through events.

We predict sigma factor binding sites for 42% (50% when leaving out putative overlapping 3′ UTRs) of the antisense transcripts near the observed 5′ TSSs. Furthermore, only 17% (22) of the antisense transcripts were predicted to have an Rho-independent terminator at the 3′, which is significantly lower than what is observed for the identified non-coding RNAs (70%).

As Xu *et al*. (2009) report bi-directional promoters as a source of antisense transcription in *S. cerevisiae*, we investigated whether such a phenomenon could also explain some of the antisense transcription in *B. subtilis*. We identified putative sigma factor binding sites on the opposite strand of the predicted antisense TSS and in the case of 16 (13%) antisense TSSs a predicted or experimental site was identified. These findings point at antisense transcription in *B. subtilis* as a 'directed' effort and perhaps to a lesser extent the result of bi-directional promoters.

Sequence and structurally conserved 3′ UTRs

During the extraction of non-coding RNAs 39 putative *ncr*s were excluded, due to cross-hybridizing probes within the transcripts. An investigation of these revealed that a group of genes have a long 3′ UTR (~220 nt) with high sequence similarity and according to RNAfold a highly stable secondary structure (see Fig. 4 and Fig. S9). The latter will in our data be apparent by a local decline in signal in both RNA and DNA hybridizations, hence causing the TAR to be split up into a gene containing TAR and a downstream ncr-like TAR. The nine genes having these conserved 3′ UTRs are listed in Table 2 along with their function (known/predicted). A closer inspection reveals that most of these are somehow membrane-associated, either physical sitting in the membrane, in complex with a membrane protein, or being exported. One possible exception is *ytvA*, which is a blue-light-sensing protein positively regulating the sigma-B pathway. Figure 5A shows RNA and DNA hybridization of a ncr-like TARs downstream of *efeN* (former *ywbN*) together with the predicted structure (Fig. 5B). Experimentally we mapped the *efeN* 3′ UTR using 3′ RLM-RACE to 225 nucleotides downstream of the *efeN* stop codon, hereby showing that the conserved sequence is indeed part of the transcript (Fig. 5C).

EfeN is a substrate of the twin-arginine translocase (Tat) protein translocation system and is expressed as a part of the *efeUMN* operon. The operon has been shown to be regulated by Fur (ferric uptake regulator) and EfeN is predicted to function as a Fe(III) permease of the dye-decolorizing Dyp-peroxidase family (Jongbloed *et al*., 2004; Ollinger *et al*., 2006). Interestingly the Tat system is able to transport folded proteins and proteins with bound cofactors and to some extent only correctly folded pro-

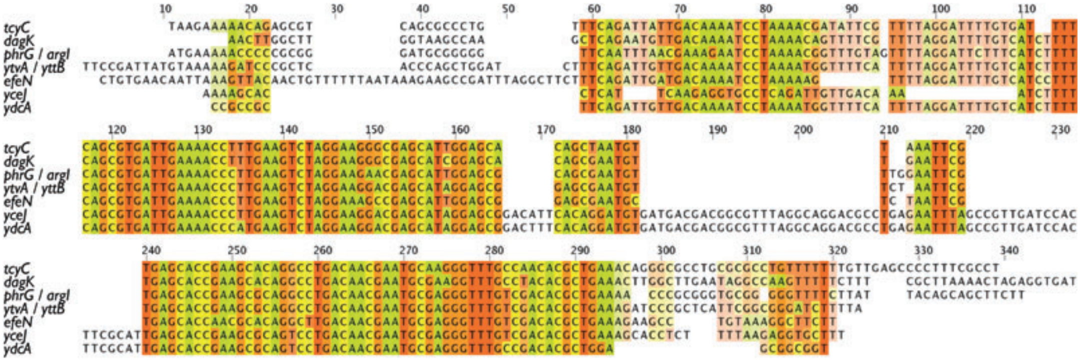


Fig. 4. Multiple alignment in CLUSTALW2 (Larkin *et al.*, 2007) of the 250 nucleotides downstream of *tcyC*, *dagK*, *phrG* (shared with *argI)*, *ytvA* (shared with *yttB)*, *efeN*, *yceJ* and *ydcA*. *efeN*, *yceJ* and *ydcA* have the reverse complement of the sequence and are here aligned using the reverse complement. Bases are coloured A = green, T = red, C = yellow, G = orange and the intensity at each position indicates base conservation. No conservation is uncoloured.

teins are transported (DeLisa *et al.*, 2003). As previous studies on EfeN in *B. subtilis* have focused on the Tat signal peptide the expression of *efeN* coding sequence has been performed via a *xylA-efeN-myc* cassette, without the conserved 3′ UTR (Jongbloed *et al.*, 2004). From this it has been observed that while EfeN expressed from the *xylA-efeN-myc* cassette has been identified in extracellular extracts, the wild-type EfeN has never been detected either inside or outside the cell (H. Antelmann and J.M. van Dijl, pers. comm.). To this respect we, in these expression data, see that during vegetative growth, *efeU*, *efeM* and *efeN* are expressed at high rates (Table S2). From this we speculate that the 3′ UTR of the *efeN* transcript may have a function in regulating the translation and/or the physical location of EfeN. Upon completion of folding or cofactor binding the protein would be available for translocation or insertion into the membrane. In *E. coli* Tat proofreading exists, where a protein binds and hereby blocks the Tat signal peptide, so that it is shielded from the translocase until proper assembly has been completed (DeLisa *et al.*, 2003). Examples of these Tat signal binding peptides in *E. coli* are TorA and NapD (Maillard *et al.*, 2007), of which no homologues are found in the *Bacilli*.

Another possible function of the 3′ UTRs could be to inhibit 3′ directed RNA degradation as double-stranded RNA and stable helical regions have been shown to block the activity of the major 3′ exoribonuclease in *B. subtilis* PNPase and RNase II. Additionally the 3′ exoribonuclease RNase R, which has been shown to be able to degrade double-stranded RNA and RNA with secondary structures, needs a single-stranded RNA tail to be active. It has been reported to be active with single-stranded tails of more than 40 nt, and was demonstrated not to be active on RNA with only a 12 nt single-stranded tail (Oussenko *et al.*, 2005). The single-stranded tail of the conserved 3′ UTRs ranges from 2 to 10 nt for *phrG* and *efeN*, respectively, meaning that they may be protected from 3′ exoribonuclease degradation (Fig. 5 and Fig. S9).

Conclusions

Since these findings are based on the first experimental attempt to map expression on a genome-wide scale in

Table 2. The nine genes with the conserved 3′ UTR and their function/predicted function.

Gene	Protein function/genetic organization	Reference
dagK	Essential diacylglycerol kinase, lipoteichoic acid (LTA) production	Jerga *et al.* (2007)
tcyC	Part of the *tcyABC* operon encoding an L-cysteine uptake system	Burguière *et al.* (2004)
ydcA	Rhomboid-like membrane protein[a]	–
yceJ	Similar to multidrug-efflux transporter[a]	–
phrG	Phosphatase (RapG) regulator, exported, divergent of *argI*	Ogura *et al.* (2003)
argI	Arginase, part of *rocDE-argI* operon (RocE: arginine permease)	Gardan *et al.* (1995)
yttB	Similar to multidrug resistance protein,[a] divergent of *ytvA*	–
ytvA	Blue-light sensor, positive regulator of the sigma-B pathway	Gaidenko *et al.* (2006)
efeN	Similar to Dyp-type peroxidases,[a] Tat-translocated protein	Jongbloed *et al.* (2004)

a. BLAST search result.

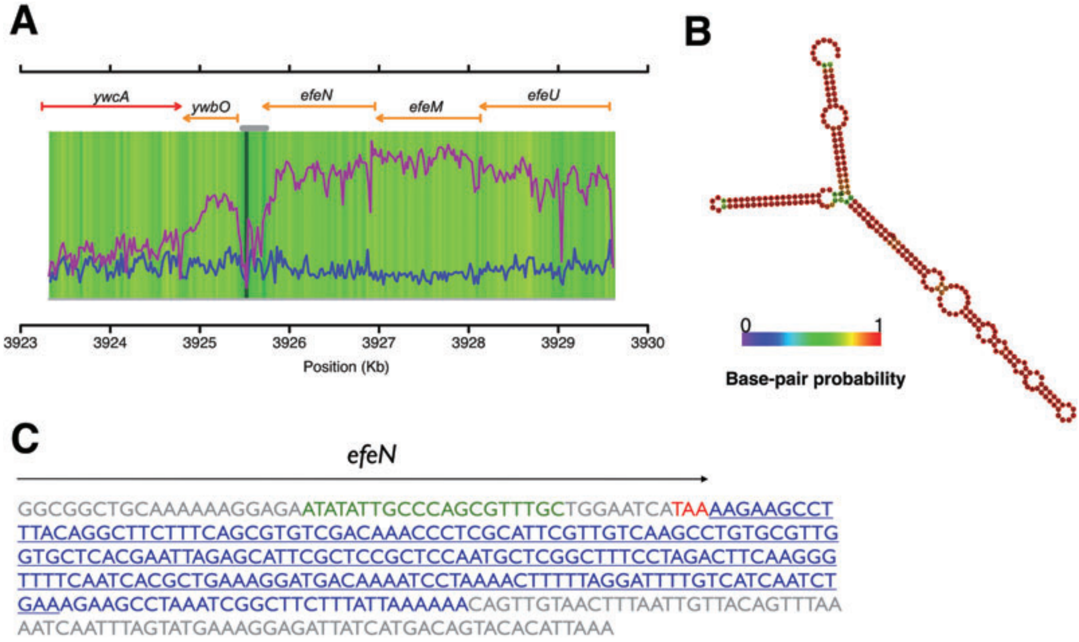


Fig. 5. A. Expression of the genomic area near *efeN* as an example of the identified conserved, stable structure forming 3′ UTRs. Watson strand is blue, Crick strand is magenta and the results from the DNA hybridization are shown in a colour gradient from dark green (low signal) to yellow (high signal). The grey bar indicated the location of the 3′ UTR transcript.
B. RNA structure, folded using RNAfold, of 220 nt downstream of stop codon of *efenN* coloured as base-pair probabilities. Blue equals zero and red equals 1.
C. 3′ sequence of *efeN* containing transcript. Grey: *efeN* CDS, and intergenic nt; green: RLM-RACE primer; red: *efeN* stop codon; blue: 3′ UTR identified by RLM-RACE and the sequence folded in (B); underlined: the conserved sequence.

B. subtilis, they are to a large extent allowing us to refine our knowledge about the *B. subtilis* transcriptome. But since almost 1:5 of the currently annotated protein-coding genes are not expressed in this study more studies using a large spectrum of different growth conditions and perturbations are needed in order to reveal the full transcriptional map of *B. subtilis*.

As this is, taken the above into consideration, a work in progress, we report the expression map of *B. subtilis* in two different media, LB and M9, as 2291 and 2464 TARs, respectively, which in total adds to 3662 non-redundant TARs. The predicted TARs clearly describe the spatial expression patterns expected from genes expressed from operons as is the case in *B. subtilis*. Additionally, and as expected, a significant difference has been observed between the length of 5′ and 3′ UTRs with medians of 47 and 36 nt respectively.

By the use of KEGG annotation we clearly see expected differences in gene expression when comparing the two growth sources. Regarding the novel protein-coding genes predicted in the re-sequencing project of Barbe and co-workers, we here report them to be expressed at low levels compared with the previously annotated protein-coding genes although 70% (119) of them are expressed above background levels. Additionally 26 of these are found to be expressed on monocistronic transcripts, providing experimental evidence for their existence. The TSS of *yzbH* has here been mapped using 5′ RLM-RACE. Furthermore the annotation of seven genes did not match well with the expression signals seen, suggesting re-annotation of these.

An analysis was also performed on prophage and prophage-like elements, revealing large clusters of genes from the *skin* element, *pro2* and *pro7*, that are not expressed. On the contrary *PBSX*, *pro3*, *pro4* and *pro5* show high expression and we identify an uncharacterized putative exported protein, *yolA*, within the *SPβ* prophage to be among the most abundantly expressed genes on the genome.

We discover a range of high-confidence novel features covering 84 non-coding RNAs and 127 antisense transcripts. We identify 10 out of the 16 known ncRNAs known in *B. subtilis* (excluding tRNAs and rRNAs), a putative ncRNA on the opposite strand of *bsrE* and 22 riboswitches. Additionally the 5′ of *ncr22* was mapped using RLM-RACE and it may act as a putative *trans*-acting inhibitor on translation of the carbon starvation protein gene *cstA*. Regarding the antisense transcripts, 27% of them could be overlapping 5′ or 3′ UTRs and 50% of the non-3′ UTR antisense transcripts have predicted sigma factor binding sites near the observed TSS. The TSSs of 16 antisense transcripts are opposing an experimental or predicted sigma factor binding site and may be products of bi-directional promoters. The expression of antisense transcripts was found to be anticorrelated to their sense counterparts.

In addition, the analysis of gDNA hybridization has led us to discover stable structures in the 3′ UTRs of several transcripts and one of these tails was experimentally verified for EfeN, a Tat-translocated protein.

Experimental procedures

Design of the tiling array BaSysBio Bsub T1

A total of 385 000 feature NimbleGen arrays have been designed, using OligoWiz 2.0 (Wernersson and Nielsen, 2005), with long iso-thermal probes (45–65 nt) covering the entire genome of *B. subtilis* #168 Trp$^+$ (AL009126.2) in 22 nt intervals on each strand and an 11 nt offset between the strands. The microarray design and data are available at the Gene Expression Omnibus (GEO) database at NIH as '*BaSysBio Bacillus subtilis T1385K array version 1*' with the records GPL8486 and GSE16086 respectively. The data were remapped to the re-sequenced genome (AL009126.3) using BLAT (BLAST-like alignment tool) and 383 probes were removed due to low match (Kent, 2002; Barbe *et al.*, 2009).

The bacterial strain, growth conditions and sample processing

Three *B. subtilis* #168 Trp$^-$ cultures were grown in LB medium and three in M9 medium at 37°C and 120 r.p.m. until the OD_{600} had reached a value of 0.5. Generation times for *B. subtilis* in the experiments were 26 and 78 min respectively. Media compositions were: LB (Sigma-Aldrich): 10 g l^{-1} Tryptone, 5 g l^{-1} yeast extract and 5 g l^{-1} NaCl; M9: 0.3% glucose, 0.1 mM $CaCl_2$, 1 mM $MgSO_4$, 0.05 mM $FeCl_3$, 8.5 g l^{-1} $Na_2HPO_4{\cdot}2H_2O$, 3 g l^{-1} KH_2PO_4, 1 g l^{-1} NH_4Cl, 0.5 g l^{-1} NaCl, 1 mg l^{-1} $MnCl_2$, 1.7 mg l^{-1} $ZnCl_2$, 0.43 mg l^{-1} $CuCl_2{\cdot}2H_2O$, 0.6 mg l^{-1} $CoCl_2{\cdot}6H_2O$ and 0.6 mg l^{-1} $Na_2MoO_4{\cdot}2H_2O$. A total of 25 ml from each culture was transferred to a 40 ml tube 1/3-filled with crushed ice and spun at 7000 r.p.m. for 5 min, after which the supernatant was discarded and the cell pellet frozen by dumping the closed tube into liquid nitrogen. Total RNA was extracted by the use of the *FastRNA PRO Blue Kit* from Qbiogene as recommended by the supplier, but with an additional shake in the *FastPrep* instrument and a 1 min incubation on ice between the two shakings. DNA was extracted (from four independent cultures grown in LB under the same conditions as described above) using the *DNeasy Blood tissue kit* from Qiagen as recommended by the supplier. Both RNA and DNA were send to NimbleGen labelled and hybridized to the BaSysBio Bsub T1 chip using a protocol for strand-specific hybridization developed during this work (the *BaSysBio* protocol), and the NimbleGen-standard protocol for double-stranded DNA respectively. All samples were labelled with Cy3 and in the case of RNA first-strand cDNA was produced by random priming and Actinomycin D inhibition of the reverse transcriptase polymerase effect (as suggested by Perocchi *et al.*, 2007). We found that the optimal enzyme concentration was 40 μg μl^{-1}.

RNA ligase-mediated rapid amplification of cDNA ends (RLM-RACE)

Transcription start sites were mapped for five transcripts using FirstChoice® RLM-RACE Kit (Ambion) following the manufacturer's protocol. DNase-treated RNA from an independent LB experiment was used as template and nested PCR was performed using primers listed in Table S3. Single-band PCRs were purified using Qiaquick PCR Purification Kit (Qiagen) and multiple bands were excised from gels and purified using Qiaquick Gel Extraction Kit (Qiagen) and sequenced. Transcript end mapping was performed for *efeN* by poly-adenylating DNase-treated RNA using Poly(A) Polymerase (Epicentre Biotechnologies) and Firstchoice® RLM-RACE kit (Ambion) following manufacturer's protocol. Only a single PCR was needed for the 3′ RLM-RACE and the primers are listed in Table S3.

Data preprocessing, segmentation and TAR creation

Segmentation was performed using the Structural Change Model (SCM) described by Huber *et al.* (2006), in the *Bioconducter* package *tilingArray*. We used default settings allowing 3000 segment to be created for each strand with a maximum length of 400 probes (~8800 bp). Normalization by reference (gDNA data) was not used as it according to our benchmarking decreased performance, and the optimal detection limit (background) was determined to the 60% quantile (2.2 log2 signal) of the signal intensities. Following the segmentation we created the resulting TARs by accepting all segments above background and joining neighbouring segments if the five probes on each side of a breakpoint were all above background, and when a Student's *t*-test did not rejected the hypothesis that these two sets of probes belonged to the same signal-intensity distribution (*P*-value > 1e-10). Finally short TARs (< 5 probes) between two highly expressed segments were removed. The resulting list of TARs is shown in Table S1.

Assessment of breakpoints

To determine the accuracy of the TAR breakpoint predictions these were benchmarked against the 654 experimentally verified TSS, which were extracted from the DataBase of Transcriptional regulation in *B. subtilis* (DBTBS, release 5) (Sierro *et al.*, 2008), and 425 experimentally verified Rho-independent terminators (Hoon *et al.*, 2005). Both the sigma factor binding sites and Rho-independent terminator annotation was transferred to the re-sequenced genome (AL000926.3) using BLAT (Kent, 2002). The TAR-signal ends were adjusted 9 and 51 nucleotides downstream to optimally predict TSS and TES. The belonging receiver operating characteristic (ROC) curves (Swets, 1988) are shown in Fig. S3.

Annotation of TARs and UTRs

Known genes were annotated to the TAR with the maximal overlap to it, and only if more than 50% of the gene was covered by the given TAR. The reported 5′ UTR lengths are the distances from the 5′ end of the given TAR to the start of the first ORF in the TAR (if any) and likewise the 3′ UTR lengths are the distances from the stop codon of the last ORF to the TAR 3′ end. Internal UTRs were calculated as the distance between stop and start for two neighbouring ORFs inside TARs.

Sigma factor and terminator predictions

All identified transcripts were annotated with experimentally verified sigma factor binding sites and Rho-independent terminator sequences (Hoon *et al*., 2005; Sierro *et al*., 2008). The co-ordinates of the above were transferred to the re-sequenced genome (AL009126.3) using BLAT. Additionally the transcripts were also annotated with predicted sigma factor binding sites from two sources, sigma A sites from Jarmer and co-workers and sigma A, B, E, D, G, F, K, H, X, W predicted by a HMMbuild from all known alignments from DBTBS (Release 5) (Jarmer *et al*., 2001; Sierro *et al*., 2008). The HMM was created using HMMbuild and HMMcalibrate and was used by HMMsearch to search in the sequences 100 nt upstream and 50 nt downstream the TSS. The sigma factors I, M, Y, Z and YlaC and YvrI had too few known sites to build HMMs. Terminators were predicted using TransTermHP 2.0 and in the case of more than one terminator within 50 nt of the TES the closest one was used for annotation (Kingsford *et al*., 2007).

KEGG analysis

KEGG annotations for *B. subtilis* were downloaded from the KEGG website (September 2008) (Kanehisa *et al*., 2008). KEGG annotations were counted for the genes present exclusively expressed in the LB medium, the M9 medium and genes expressed in both (common). From each of these three categories, the five most occurring annotations were selected and the occurrences were plotted as shown in Fig. 1.

Identification of novel ncRNAs

Segments of five or more probes without known annotation according to the latest GenBank annotation (AL009126.3) and no ORF predicted by EasyGene (Nielsen and Krogh, 2005) were accepted as putative novel ncRNAs if they were expressed above background and neighbouring segments, contained a maximum of 5% potentially cross-hybridizing probes and had higher signal level than same area on the opposite strand. Segments with less than five probes fulfilling the criteria and expressed in both media were also accepted as possible ncRNAs. In addition, all potentially novel ncRNAs were inspected visually. The ncRNAs were named *ncr1–ncr84* and are listed in Table S5. We also searched the first 100 nt of each ncRNA for ribosome binding sites with SD (AGGAGG) and 4–10 nt after that a start codon (ATG/CTG/GTG), resulting in five of these coding for small putative CDSs. The DNA sequences corresponding to the 84 segments that passed the above criteria were extracted from the genome sequence (AL009126.3), and were compared by BLAST to all available Firmicute genome or plasmid sequences [34 species within 42 strains resulting in 225 entries from the CBS Genome Atlas Database version 2.0 (Hallin and Ussery, 2004)] (Altschul *et al*., 1990). For each species the best hit was recorded as per cent identity over the entire ncRNA length. These results are shown in Fig. S7.

Identification of antisense RNAs

Segments were subjected to the same criteria as for identification of novel ncRNAs, except expression did not have to be higher than on the opposing strand. Additionally the transcripts are antisense to a known gene (GenBank: AL00926.3) with an overlap of more than 10%. The identified antisense transcripts were manually curated leading to 127 transcripts that were named *shd1–127* and are listed in Table S6.

3′ UTR identification

Transcripts with a conserved 3′ UTR structure were identified based on multiple alignment of the 220 nucleotides downstream of all genes, performed using CLUSTALX2 (Larkin *et al*., 2007). Structures were made using RNAfold v. 1.6 and the Vienna RNA web suite (Gruber *et al*., 2008).

Acknowledgements

This work was funded by the European Commission through the EU-IP-FP6-project, BaSysBio (LSHG-CT2006-037469). The sample preparation was funded by Augustinus Fonden. The authors wish to thank Pia Friis for her technical assistance in the lab and Peter Wad Sackett for his helping with the technical validation of the array design. Additionally the authors wish to thank Jan Marten van Dijl and Emma Denham for fruitful discussions and ideas on the 3′ UTRs and EfeN protein.

References

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990) Basic local alignment search tool. *J Mol Biol* **215:** 403–410.

Anagnostopoulos, C., and Spizizen, J. (1961) Requirements for transformation in *Bacillus subtilis*. *J Bacteriol* **81:** 741–746.

Ando, Y., Asari, S., Suzuma, S., Yamane, K., and Nakamura, K. (2002) Expression of a small RNA, BS203 RNA, from the *yocI–yocJ* intergenic region of the *Bacillus subtilis* genome. *FEMS Microbiol Lett* **207:** 29–33.

Auchtung, J.M., Lee, C.A., and Grossman, A.D. (2006) Modulation of the ComA-dependent quorum response in *Bacillus subtilis* by multiple Rap proteins and Phr peptides. *J Bacteriol* **188:** 5273–5285.

Barbe, V., Cruveiller, S., Kunst, F., Lenoble, P., Meurice, G., Sekowska, A., *et al.* (2009) From a consortium sequence to a unified sequence: the *Bacillus subtilis* 168 reference genome a decade later. *Microbiology* **155:** 1758–1775.

Bertone, P., Stolc, V., Royce, T.E., Rozowsky, J.S., Urban, A.E., Zhu, X., *et al.* (2004) Global identification of human transcribed sequences with genome tiling arrays. *Science* **306:** 2242–2246.

Burguière, P., Auger, S., Hullo, M.F., Danchin, A., and Martin-Verstraete, I. (2004) Three different systems participate in L-cystine uptake in *Bacillus subtilis*. *J Bacteriol* **186:** 4875–4884.

Burkholder, P.R., and Giles, N.H. (1947) Induced biochemical mutations in *Bacillus subtilis*. *Am J Bot* **34:** 345–348.

Cervantes, C., Ji, G., Ramirez, J., and Silver, S. (1994) Resistance to arsenic compounds in microorganisms. *FEMS Microbiol Rev* **15:** 355–367.

Cohn, F. (1872) Untersuchungen über Bakterien. *Beitr Biol Pflanzen* **1:** 127–224.

Cohn, F. (1876) Untersuchungen über Bakterien, IV. Beiträge zur Biologie der Bacillen. *Beitr Biol Pflanzen* **2:** 249–277.

Conn, H.J. (1930) The identity of *Bacillus subtilis*. *J Infect Dis* **46:** 341–350.

David, L., Huber, W., Granovskaia, M., Toedling, J., Palm, C.J., Bofkin, L., *et al.* (2006) A high-resolution map of transcription in the yeast genome. *Proc Natl Acad Sci USA* **103:** 5320–5325.

DeLisa, M.P., Tullman, D., and Georgiou, G. (2003) Folding quality control in the export of proteins by the bacterial twin-arginine translocation pathway. *Proc Natl Acad Sci USA* **100:** 6115–6120.

Deutscher, J., Galinier, A., and Martin-Verstraete, I. (2002) Carbohydrate uptake and metabolism. In *Bacillus subtilis and its closest relatives*. Sonenshein, A.L., Hoch, J.A., and Losick, R. (eds). Washington, DC: American Society for Microbiology, pp. 129–150.

Dubey, A.K., Baker, C.S., Suzuki, K., Jones, A.D., Pandit, P., Romeo, T., and Babitzke, P. (2003) CsrA regulates translation of the *Escherichia coli* carbon starvation gene, *cstA*, by blocking ribosome access to the *cstA* transcript. *J Bacteriol* **185:** 4450–4460.

Ehrenberg, C.G. (1835) Dritter Beitrag zur Erkenntniss grosser Organisation in der Richtung des kleinsten Raumes. In Physikalische Abhandlungen der Koeniglichen Akademie der Wissenschaften zu Berlin aus den Jahren 1833–1835, pp. 145–336.

Eiamphungporn, W., and Helmann, J.D. (2009) Extracytoplasmic function sigma factors regulate expression of the *Bacillus subtilis yabE* gene via a *cis*-acting antisense RNA. *J Bacteriol* **191:** 1101–1105.

Fillinger, S., Boschi-Muller, S., Azza, S., Dervyn, E., Branlant, G., and Aymerich, S. (2000) Two glyceraldehyde-3-phosphate dehydrogenases with opposite physiological roles in a nonphotosynthetic bacterium. *J Biol Chem* **275:** 14031–14037.

Gaballa, A., Antelmann, H., Aguilar, C., Khakh, S.K., Kyung-Bok, S., Smaldone, G.T., and Helmann, J.D. (2008) The *Bacillus subtilis* iron-sparing response is mediated by a Fur-regulated small RNA and three small basic proteins. *Proc Natl Acad Sci USA* **105:** 11927–11932.

Gaidenko, T.A., Kim, T.J., Weigel, A.L., Brody, M.S., and Price, C.W. (2006) The blue-light receptor YtvA acts in the environmental stress signaling pathway of *Bacillus subtilis*. *J Bacteriol* **188:** 6387–6395.

Gardan, R., Rapoport, G., and Débarbouille, M. (1995) Expression of the *rocDEF* operon involved in arginine catabolism in *Bacillus subtilis*. *J Mol Biol* **249:** 843–856.

Gollnick, P., Babitzke, P., Antson, A., and Yanofsky, C. (2005) Complexity in regulation of tryptophan biosynthesis in *Bacillus subtilis*. *Annu Rev Genet* **39:** 47–68.

Gruber, A.R., Lorenz, R., Bernhart, S.H., Neuböck, R., and Hofacker, I.L. (2008) The Vienna RNA Web suite. *Nucleic Acids Res* **36:** W70–W74.

Grundy, F.J., and Henkin, T.M. (1998) The S box regulon: a new global transcription termination control system for methionine and cysteine biosynthesis genes in Gram-positive bacteria. *Mol Microbiol* **30:** 737–774.

Haldenwang, W.G. (1995) The sigma factors of *Bacillus subtilis*. *Microbiol Mol Biol Rev* **59:** 1–30.

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

Helmann, J.D., Wu, M.F., Kobel, P.A., Gamo, F.J., Wilson, M., Morshedi, M.M., *et al.* (2001) Global transcriptional response of *Bacillus subtilis* to heat shock. *J Bacteriol* **183:** 7318–7328.

Hoon, M.J.L., Makita, Y., Nakai, K., and Miyano, S. (2005) Prediction of transcriptional terminators in *Bacillus subtilis* and related species. *PLoS Comput Biol* **1:** 212–222.

Huang, J.M., La Ragione, R.M., Nunez, A., and Cutting, S.M. (2008) Immunostimulatory activity of *Bacillus* spores. *FEMS Immunol Med Microbiol* **53:** 195–203.

Huber, W., Toedling, J., and Steinmetz, L.M. (2006) Transcript mapping with high-density oligonucleotide tiling arrays. *Bioinformatics* **22:** 1963–1970.

Jarmer, H., Larsen, T.S., Krogh, A., Saxild, H.H., Brunak, S., and Knudsen, S. (2001) Sigma A recognition sites in the *Bacillus subtilis* genome. *Microbiology* **147:** 2417–2424.

Jerga, A., Lu, Y.J., Schujman, G.E., de Mendoza, D., and Rock, C.O. (2007) Identification of a soluble diacylglycerol kinase required for lipoteichoic acid production in *Bacillus subtilis*. *J Biol Chem* **282:** 21738–21745.

Jongbloed, J.D.H., Grieger, U., Antelmann, H., Hecker, M., Nijland, R., Bron, S., and van Dijl, J.M. (2004) Two minimal Tat translocases in *Bacillus*. *Mol Microbiol* **54:** 1319–1325.

Kanehisa, M., Araki, M., Goto, S., Hattori, M., Hirakawa, M., Itoh, M., *et al.* (2008) KEGG for linking genomes to life and the environment. *Nucleic Acids Res* **36:** D480–D484.

Kawano, M., Reynolds, A.A., Miranda-Rios, J., and Storz, G. (2005) Detection of 5′- and 3′-UTR-derived small RNAs and *cis*-encoded antisense RNAs in *Escherichia coli*. *Nucleic Acids Res* **33:** 1040–1050.

Kent, W.J. (2002) BLAT – the BLAST-like alignment tool. *Genome Res* **12:** 656–664.

Kingsford, C.L., Ayabule, K., and Salzberg, S.L. (2007) Rapid, accurate, computational discovery of Rho-independent transcription terminators illuminates their relationship to DNA uptake. *Genome Biol* **8:** R22.

Krogh, S., O'Reailly, M., Nolan, N., and Devine, K.M. (1996) The phage-like element PBSX and part of the *skin* element, which are resident at different locations on the *Bacillus subtilis* chromosome are highly homologous. *Microbiology* **142:** 2031–2040.

Kunst, F., Ogasawara, N., Moszer, I., Albertini, A.M., Alloni, G., Azevedo, V., *et al.* (1997) The complete genome sequence of the Gram-positive bacterium *Bacillus subtilis*. *Nature* **390:** 249–256.

Larkin, M.A., Blackshields, G., Brown, N.P., Chenna, R., McGettigan, P.A., McWilliam, H., *et al.* (2007) ClustalW and ClustalX version 2. *Bioinformatics* **23:** 2647–2648.

Lazarevic, V., Düsterhöft, A., Soldo, B., Hilbert, H., Maeuël, C., and Karamata, D. (1999) Nucleotide sequence of the *Bacillus subtilis* temperate bacteriophage *SPβc2*. *Microbiology* **145:** 1055–1067.

Li, L., Wang, X., Stolc, V., Li, X., Zhang, D., Su, N., *et al.* (2006) Genome-wide transcription analyses in rice using tiling microarrays. *Nat Genet* **38:** 124–129.

Licht, A., Preis, S., and Brantl, S. (2005) Implication of CcpN in the regulation of a novel untranslated RNA (SR1) in *B. subtilis*. *Mol Microbiol* **58:** 189–206.

Mäder, U., Henning, S., Hecker, M., and Homuth, G. (2004) Transcriptional organization and posttranscriptional regulation of the *Bacillus subtilis* branched-chain amino acid biosynthesis genes. *J Bacteriol* **186:** 2240–2252.

Maillard, J., Spronk, C.A.E.M., Buchanan, G., Lyall, V., Richardson, D.J., Palmer, T., *et al.* (2007) Structural diversity in twin-arginine signal peptide-binding proteins. *Proc Natl Acad Sci USA* **104:** 15641–15646.

Mazumder, B., Seshadri, V., and Fox, P.L. (2003) Translational control by the 3′-UTR: the ends specify the means. *Trends Biochem Sci* **28:** 91–98.

Nagalakshmi, U., Wang, Z., Waern, K., Shou, C., Raha, D., Gerstein, M., and Snyder, M. (2008) The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science* **320:** 1344–1349.

Nicolas, P., Bize, L., Muri, F., Hoebeke, M., Rodolphe, F., Ehrlic, S.D., *et al.* (2002) Mining *Bacillus subtilis* chromosome heterogeneities using hidden Markov models. *Nucleic Acids Res* **30:** 1418–1426.

Nielsen, P., and Krogh, A. (2005) Large-scale prokaryotic gene prediction and comparison to genome annotation. *Bioinformatics* **21:** 4322–4329.

Ogura, M., Shimane, K., Asai, K., Ogasawara, N., and Tanaka, T. (2003) Binding of response regulator DegU to the *aprE* promoter is inhibited by RapG, which is counteracted by extracellular PhrG in *Bacillus subtilis*. *Mol Microbiol* **49:** 1685–1697.

Ollinger, J., Song, K.B., Antelmann, H., Hecker, M., and Helmann, J.D. (2006) Role of the Fur regulon in iron transport in *Bacillus subtilis*. *J Bacteriol* **188:** 3664–3673.

Oussenko, I.A., Abe, T., Ujiie, H., Muto, A., and Bechhofer, D.H. (2005) Participation of 3′- to 5′-exoribonucleases in the turnover of *Bacillus subtilis* mRNA. *J Bacteriol* **187:** 2758–2767.

Perocchi, F., Xu, Z., Clauder-Münster, S., and Steinmetz, L.M. (2007) Antisense artifacts in transcriptome microarray experiments are resolved by actinomycin D. *Nucleic Acids Res* **35:** 1–7.

Ratushna, V.G., Weller, J.W., and Gibas, C.J. (2005) Secondary structure in the target as a confounding factor in synthetic oligomer microarray design. *BMC Genomics* **6:** 31.

Reppas, N.B., Wade, J.T., Church, G.M., and Struhl, K. (2006) The transition between transcriptional initiation and elongation in *E. coli* is highly variable and often rate limiting. *Mol Cell* **24:** 747–757.

Saito, S., Kakeshita, H., and Nakamura, K. (2009) Novel small RNA-encoding genes in the intergenic regions of *Bacillus subtilis*. *Gene* **428:** 2–8.

Shivers, R.P., and Sonenshein, A.L. (2005) *Bacillus subtilis ilvB* operon: an intersection of global regulons. *Mol Microbiol* **56:** 1549–1559.

Sierro, N., Makita, Y., de Hoon, M., and Nakai, K. (2008) DBTBS: a database of transcriptional regulation in *Bacillus subtilis* containing upstream intergenic conservation information. *Nucleic Acids Res* **36:** D93–D96.

Silvaggi, J.M., Perkins, J.B., and Losick, R. (2005) Small untranslated RNA antitoxin in *Bacillus subtilis*. *J Bacteriol* **187:** 6641–6650.

Silvaggi, J.M., Perkins, J.B., and Losick, R. (2006) Genes for small, noncoding RNAs under sporulation control in *Bacillus subtilis*. *J Bacteriol* **188:** 532–541.

Spizizen, J. (1958) Transformation of biochemically deficient strains of *Bacillus subtilis* by deoxyribonucleate. *Proc Natl Acad Sci USA* **44:** 1072–1078.

Spizizen, J. (1984) Citation classic – Transformation of biochemically deficient strains of *Bacillus subtilis* by deoxyribonucleate. *Curr Contents Life Sci* **19:** 15.

Steil, L., Serrano, M., Henriques, A.O., and Völker, U. (2005) Genome-wide analysis of temporally regulated and compartment-specific gene expression in sporulating cells of *Bacillus subtilis*. *Microbiology* **151:** 339–420.

Suzuma, S., Asari, S., Bunai, K., Yoshino, K., Ando, Y., Kakeshita, H., *et al.* (2002) Identification and characterization of novel small RNAs in the *aspS–yrvM* intergenic region of the *Bacillus subtilis* genome. *Microbiology* **148:** 2591–2598.

Swets, J.A. (1988) Measuring the accuracy of diagnostic systems. *Science* **240:** 1285–1293.

Takemaru, K., Mizuno, M., Sato, T., Takeuchi, M., and Kobayashi, Y. (1995) Complete nucleotide sequence of a *skin* element excised by DNA rearrangement during sporulation in *Bacillus subtilis*. *Microbiology* **141:** 323–327.

Teas, H.J. (1949) Mutants of *Bacillus subtilis* that require threonine or threonine plus methionine. *J Bacteriol* **59:** 93–104.

Tjaden, B., Goodwin, S.S., Opdyke, J.A., Guillier, M., Fu, D.X., Gottesman, S., and Storz, G. (2006) Target prediction for small, noncoding RNAs in bacteria. *Nucleic Acids Res* **34:** 2791–2802.

Tjaden, B., Saxena, R.M., Stolyar, S., Haynor, D.R., Kolker, E., and Rosenow, C. (2002) Transcriptome analysis of *Escherichia coli* using high-density oligonucleotide probe arrays. *Nucleic Acids Res* **30:** 3732–3738.

Vogel, J., and Wagner, E.G.H. (2007) Target identification of small noncoding RNAs in bacteria. *Curr Opin Microbiol* **10:** 262–270.

Wang, L., Park, S., and Doi, R.H. (1999) A novel *Bacillus subtilis* gene, *antE*, temporally regulated and convergent to and overlapping *dnaE*. *J Bacteriol* **181:** 353–356.

Wernersson, R., and Nielsen, H.B. (2005) OligoWiz 2.0 – integrating sequence feature annotation into the design of microarray probes. *Nucleic Acids Res* **33:** W611–W615.

Wood, H.E., Dawson, M.T., Devine, K.M., and McConnell, D.J. (1990) Characterization of *PBSX*, a defective prophage of *Bacillus subtilis*. *J Bacteriol* **172:** 2667–2674.

Xu, Z., Wei, W., Gagneur, J., Perocchi, F., Clauder-Münster, S., Camblon, J., *et al.* (2009) Bidirectional promoters generate pervasive transcription in yeast. *Nature* **457:** 1033–1037.

Yakhnin, H., Pandi, P., Petty, T.J., Baker, C.S., Romeo, T., and Babitzke, P. (2007) CrsA of *Bacillus subtilis* regulates translation initiation of the gene encoding the flagellin protein (*hag*) by blocking ribosome binding. *Mol Microbiol* **64:** 1605–1620.

Zahler, S.A., Korman, R.Z., Rosenthal, R., and Hemphill, H.E. (1977) *Bacillus subtilis* bacteriophage *SP*β: localization of the prophage attachment site and specialized transduction. *J Bacteriol* **129:** 556–558.

Zeigler, D.R., Prágai, Z., Rodriguez, S., Chevreux, B., Muffler, A., Albert, T., *et al.* (2008) The origins of 168, W23, and other *Bacillus subtilis* legacy strains. *J Bacteriol* **190:** 6983–6995.

Supporting information

Additional supporting information may be found in the online version of this article.

Please note: Wiley-Blackwell are not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.

3.2 Perspectives

We hope that the *B. subtilis* community will investigate and use our findings by integrating the data into their own research. An example is that we were contacted by the editor of Molecular Microbiology, Dr. John Helmann, informing us that when observing *shd83* (shown in Figure 2G and H) he fifteen years ago identified a SigA binding site that matches the TSS of that antisense transcript.

Regarding the conserved 3' UTRs that we identified in the *efeN* transcript, we hypothesized that this might be a part of a translational control mechanism similar to the Tat-proofreading in *E. coli.* A recent study in *E. coli* [112] showed that a DNase (TatD) was crucial for the Tat-proofreading, which could add more to the hypothesis of the 3' UTR tail. Interestingly, we identified an antisense transcript (*shd3*) for *yabD*, a gene coding for a protein that contains a TatD DNase domain.

With collaborators at University Medical Center Groningen (UMCG) we are currently investigating the effects of the non-coding RNA *ncr26*. *ncr26* is a 351 nt putative non-coding RNA expressed from a predicted SigW/SigX promoter for both conditions tested. No protein has been identified from the sequence and from *ncr26*$^{-}$ experiments it seems to be involved in protein secretion. At the moment we are analyzing tiling array data from this and other related mutants. Regarding *ncr22*, the non-coding transcript hypothesized to have translational regulatory control on *cstA*, we are hoping to start similar experiments.

Chapter 4

Characterizing a *Saccharomyces cerevisiae* mutant

Saccharomyces cerevisiae is an eukaryotic organism widely used in the industry for fermentation purposes, for baking (bakers yeast) and for the production of recombinant proteins and bioethanol. In 1996 it became the first eukaryotic genome to be sequenced and it is one of the best studied model organisms for eukaryotic life with more than 40.000 research publications [113]. There are an abundance of data available from almost any aspect of biological research making *S. cerevisiae* an obvious model for systems biology [114].

Together with Carlsberg Research Center (CRC) we engaged in characterizing a *S. cerevisiae* BY4741 mutant, a strain that was created and published by Alper et al. in Science, 2006 [115]. The purpose of that study was to develop a strain with improved tolerance towards glucose and ethanol as these characteristics are important in very high gravity fermentations in the industry. This form of fermentation is characterized by high sugar concentration in the beginning and high ethanol concentration in the end of the batch run. As it has previously been found that tolerance to ethanol and glucose mixtures is not controlled at monogenic level, the authors employed a global Transcription Machinery Engineering (gTME) approach. Here random mutagenesis is applied to key proteins in the transcription with the target of the particular study, *SPT15*, being a TATA-binding protein. Modulation of TATA-binding proteins has previously been shown to induce changes in the specificity of RNA polymerase II towards promoters and may therefore be used to induce changes in gene expression of multiple genes. A mutant with three mutations in *SPT15* (termed *spt15-300*) was found to display

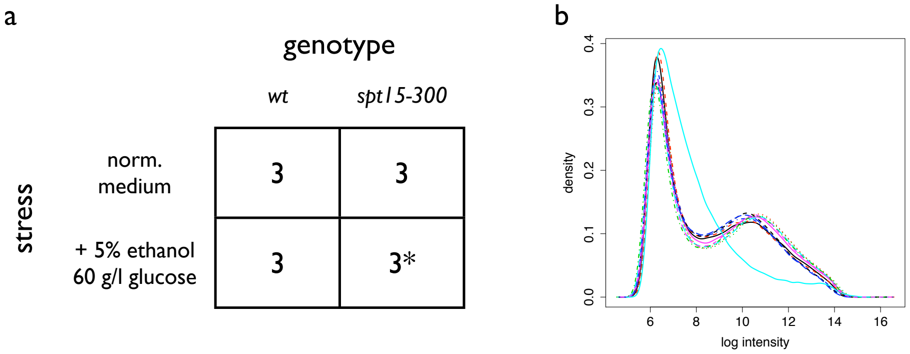


Figure 4.1 – Experimental setup and outlier removal. **a**. two factor ANOVA setup, the factors "genotype" and "medium" are shown and the number of biological replicates indicated inside the box. *: One array was removed as an outlier from this group. **b**. Density plot of probe intensities. The light blue line represents the outlier.

increased glucose and ethanol tolerance.

In the original publication DNA microarray analysis in combination with gene-knockout and overexpression analysis of SPT15 targeted genes were applied, however they were not able to identify the genetic network responsible for the phenotype. Our collaborators at CRC attempted to use the *spt15-300* allele in industrial important strains, however they found that the increased tolerance phenotype could only be reproduced in media with small amounts of the amino acid leucine. Interestingly the BY4741 strain is deleted for the *LEU2* gene involved in the biosynthesis of leucine and rescuing this deletion abolished the differences. We therefore re-investigated the DNA microarray data from the original publication with regard to leucine uptake, synthesis and utilization.

4.1 Paper II

The DNA microarray experiment originally performed was designed as a two-factor ANOVA (Figure 4.1a). For the analysis the authors only used the unstressed conditions – hence they only assessed the genotype effect and ignored samples from the glucose/ethanol conditions. On the contrary we analyzed the experiment as a two-factor ANOVA approach, removing one sample from the *spt15-300* and 5% ethanol, 60 g/l glucose group because we identified it as an outlier (Figure 4.1b).

From the analysis we found that the stress induced by glucose and ethanol significantly regulates expression of roughly half of the known genes (3100 genes at FDR = 0) in *S. cerevisiae*. Likewise the genotype effect (*spt15-300* mutation) significantly regulates 700 genes (FDR = 0), indicating that both pertubations have large impact on the cells. In a situation like this, being able to focus on a few biological pathways can greatly enhance the chance of success.

APPLIED AND ENVIRONMENTAL MICROBIOLOGY, Oct. 2009, p. 6055–6061 Vol. 75, No. 19
0099-2240/09/$08.00+0 doi:10.1128/AEM.00989-09

Impaired Uptake and/or Utilization of Leucine by *Saccharomyces cerevisiae* Is Suppressed by the *SPT15-300* Allele of the TATA-Binding Protein Gene[▽]

Richard J. S. Baerends,[1] Jin-Long Qiu,[1] Simon Rasmussen,[2] Henrik Bjørn Nielsen,[2] and Anders Brandt[1]*

Carlsberg Laboratory, Gamle Carlsberg Vej 10, DK-2500 Copenhagen Valby, Denmark,[1] and Center for Biological Sequence Analysis, Department of Systems Biology, Technical University of Denmark, DK-2800 Kongens Lyngby, Denmark[2]

Received 30 April 2009/Accepted 31 July 2009

Successful fermentations to produce ethanol require microbial strains that have a high tolerance to glucose and ethanol. Enhanced glucose/ethanol tolerance of the laboratory yeast *Saccharomyces cerevisiae* strain BY4741 under certain growth conditions as a consequence of the expression of a dominant mutant allele of the *SPT15* gene (*SPT15-300*) corresponding to the three amino acid changes F177S, Y195H, and K218R has been reported (H. Alper, J. Moxley, E. Nevoigt, G. R. Fink, and G. Stephanopoulos, Science 314:1565–1568, 2006). The *SPT15* gene codes for the TATA-binding protein. This finding prompted us to examine the effect of expression of the *SPT15-300* allele in various yeast species of industrial importance. Expression of *SPT15-300* in leucine-prototrophic strains of *S. cerevisiae*, *Saccharomyces bayanus*, or *Saccharomyces pastorianus* (lager brewing yeast), however, did not improve tolerance to ethanol on complex rich medium (yeast extract-peptone-dextrose). The enhanced growth of the laboratory yeast strain BY4741 expressing the *SPT15-300* mutant allele was seen only on defined media with low concentrations of leucine, indicating that the apparent improved growth in the presence of ethanol was indeed associated with enhanced uptake and/or utilization of leucine. Reexamination of the microarray data published by Alper and coworkers likewise suggested that expression of genes coding for the leucine permeases, Tat1p and Bap3p, were upregulated in the *SPT15-300* mutant, as was expression of the genes *ARO10*, *ADH3*, *ADH5*, and *SFA1*, involved in leucine degradation.

Improvement of stress tolerance in microorganisms applied in industrial fermentations for the production of ethanol is of major interest (26, 34). Based on screens for ethanol sensitivity/tolerance in *Saccharomyces cerevisiae* (12, 16–18, 35, 37, 40), it appears that this trait in yeast is possibly controlled by several genes acting in concert. Using global transcription machinery engineering (gTME), a tool to reprogram gene transcription for eliciting new phenotypes important for technological applications, Alper et al. (2) found mutants of *S. cerevisiae* with improved glucose/ethanol tolerance. In that work, mutated versions of the *SPT15* gene, which codes for the TATA-binding protein, were generated by random in vitro mutagenesis and expressed in the laboratory strain BY4741. The authors identified one dominant allele, *SPT15-300*, which corresponds to the three amino acid changes F177S, Y195H, and K218R, that conferred increased tolerance of the yeast to ethanol (2). Although extensive analyses, such as transcriptional profiling and deleting and overexpressing individual genes, were carried out, a particular pathway or a genetic network responsible for the observed growth gain of the *SPT15-300*-expressing strain could not be identified (2). During our attempts to analyze the effect of the mutant *SPT15-300* alleles in various yeast species of industrial importance, we discovered that the described improvement of growth in the presence of ethanol of the standard laboratory strain BY4741, the strain used by Alper et al. (2), is associated with improved uptake and/or utilization of leucine on media containing small amounts of leucine.

MATERIALS AND METHODS

Strains, media, and molecular procedures. The *Saccharomyces* strains investigated in this study were *S. cerevisiae* strains BY4741 (*MAT***a** *his3ΔD1 leu2Δ0 met15Δ0 ura3Δ0*) (5), in which the *LEU2* gene is completely deleted (obtained from Euroscarf, Frankfurt, Germany), and Y55 (23) derivative JT20150 (*MAT***α** *MAL1*) (obtained from J. M. Thevelein, Katholieke Universiteit, Leuven, Belgium); *S. bayanus* NRRL Y-11845 (MCYC 623) (7, 20) (provided by C. P. Kurtzman, Microbial Genomics and Bioprocessing Research Unit, Peoria, IL); and *S. pastorianus* W-34/70 (25) (obtained from Hefebank Weihenstephan, Freising, Germany).

Yeast cells were cultured aerobically in complex rich medium YPD (1% [wt/vol] yeast extract, 2% [wt/vol] peptone), 2% [wt/vol] glucose) (32) supplemented when necessary with G418 (final concentration of 100 or 300 μg/ml as indicated), synthetic complete minimal (SC) medium (6.7 g/liter yeast nitrogen base [without amino acids] supplemented with amino acids as specified in reference 32) without uracil (SC−Ura) and a modified composition containing five times the amount of leucine (i.e., 150 mg/liter instead of 30 mg/liter) (SC−Ura 5 × Leu), SC lacking leucine (SC−Leu), yeast synthetic complete (YSC) medium (6.7 g/liter yeast nitrogen base [without amino acids] supplemented with Qbiogene CSM-URA [a commercial amino acid mixture]) lacking uracil (containing 100 mg/liter leucine) as described by Alper et al. (2), and YSC lacking leucine (prepared as described for YSC−Ura, with Qbiogene CSM-LEU [2]). SC media were buffered (pH 5.5) with 1% (wt/vol) succinic acid and 0.6% (wt/vol) NaOH. SC and YSC media were supplemented with glucose and/or ethanol as indicated. *S. cerevisiae* strains were incubated at 20 or 30°C (as indicated), and *S. bayanus* and *S. pastorianus* were cultivated at 20°C. *Saccharomyces* species were transformed by use of the lithium acetate method (3).

Escherichia coli strain DH5α (Invitrogen A/S, Taastrup, Denmark) was used for plasmid selection/propagation and cultivated as described previously (31).

* Corresponding author. Mailing address: Carlsberg Laboratory, Gamle Carlsberg Vej 10, DK-2500 Copenhagen Valby, Denmark. Phone: 45 3327 5236. Fax: 45 3327 4765. E-mail: anders.brandt@crc.dk.

[▽] Published ahead of print on 7 August 2009.

TABLE 1. Constructed *SPT15* expression vectors[a]

Vector	Selection marker	Promoter	Insert
pCJR2	G418	P_{TEF1}	Multiple-cloning site
pCJR3	G418	P_{TEF1}	*S. cerevisiae*-type *SPT15-300*
pCJR4	G418	P_{TEF1}	*S. cerevisiae*-type *SPT15*
pCJR5	G418	P_{TEF1}	Non-*S. cerevisiae*-type *SPT15*
pCJR6	G418	P_{TEF1}	Non-*S. cerevisiae*-type *SPT15-300*
pCJR7	G418	$P_{TEF1mut2}$	*S. cerevisiae*-type *SPT15*
pCJR8	G418	$P_{TEF1mut2}$	*S. cerevisiae*-type *SPT15-300*
pCJR11	*URA3*	$P_{TEF1mut2}$	*S. cerevisiae*-type *SPT15*
pCJR12	*URA3*	$P_{TEF1mut2}$	*S. cerevisiae*-type *SPT15-300*
pCJR17	*LEU2*	$P_{TEF1mut2}$	*S. cerevisiae*-type *SPT15*
pCJR18	*LEU2*	$P_{TEF1mut2}$	*S. cerevisiae*-type *SPT15-300*

[a] See "Plasmid constructions" in Materials and Methods.

Plasmid constructions. Standard recombinant DNA manipulations were performed as described previously (31). DNA-modifying enzymes were obtained from Invitrogen (Invitrogen A/S, Taastrup, Denmark), New England Biolabs (Medinova Scientific A/S, Glostrup, Denmark), and Promega (Promega Biotech AB, Nacka, Sweden) and used as recommended by the suppliers. PCRs were carried out with Phusion high-fidelity DNA polymerase (Finnzymes, Medinova Scientific A/S, Glostrup, Denmark). DNA sequencing and oligonucleotide synthesis were performed by Eurofins MWG (Ebersberg, Germany); oligonucleotide sequences are available on request.

SPT15 expression vectors (Table 1) were constructed basically as described by Alper et al. (2). As displayed in Table 1, four vector sets were constructed.

(i) One set of *SPT15* variants (see below) was inserted into vector pCJR2 (a *CEN*-based vector with a native *S. cerevisiae TEF1* promoter and G418 selection), which was constructed by cloning a 852-bp SacI-PvuII-fragment of p416TEF (24) (obtained from ATCC, LGC Standards AB, Boras, Sweden) containing the wild-type *S. cerevisiae TEF1* promoter into a 4471-bp SacI-EcoRV-digested vector fragment of pCJR1. This plasmid was constructed by cloning a 1,447-bp BglII (blunt ended by DNA polymerase [Klenow fragment])-SacI fragment (KanMX4 cassette) of pUG6 (15) (obtained from Euroscarf, Frankfurt, Germany) into a 3,082-bp TthIII1 (blunt ended with Klenow fragment)-SacI vector fragment of pRS315 (33).

(ii) The second set of *SPT15* expression vectors was constructed identically to pCJR2, except that the wild-type *S. cerevisiae TEF1* promoter was exchanged with mutant version 2 as described previously (1, 27). To accomplish this, a 403-bp SacI-XbaI-digested synthetic DNA fragment (GenScript, Piscataway, NJ, USA) containing the mutant *TEF1* promoter (1) was used to replace the native SacI-SpeI-digested promoter.

(iii) A third set of *SPT15* vectors (*CEN*-based vector, mutant *S. cerevisiae TEF1* promoter, *URA3* selection) was constructed by inserting SacI-EagI-digested fragments of pCJR7 (1,434-bp fragment with *S. cerevisiae*-type *SPT15*) or pCJR8 (1,430-bp fragment with *S. cerevisiae*-type *SPT15-300*) into a 4,805-bp SacI-EagI-digested vector fragment of p416TEF.

(iv) A fourth set of *SPT15* vectors (*CEN*-based vector, mutant *S. cerevisiae TEF1* promoter, *LEU2* selection) was constructed by inserting SacI-EagI-digested fragments of pCJR7 (1,434-bp fragment with *S. cerevisiae*-type *SPT15*) or pCJR8 (1,430-bp fragment with *S. cerevisiae*-type *SPT15-300*) into a 6,005-bp SacI-EagI-digested vector fragment of pRS315.

The lager brewing yeast, *Saccharomyces pastorianus*, is a hybrid of *S. cerevisiae* and a *Saccharomyces* species related to *S. bayanus* (21, 25). Genes in the genome of lager brewing yeast that have high identity with genes found in *S. cerevisiae* are called *S. cerevisiae* type, while genes more distantly related are called non-*S.*

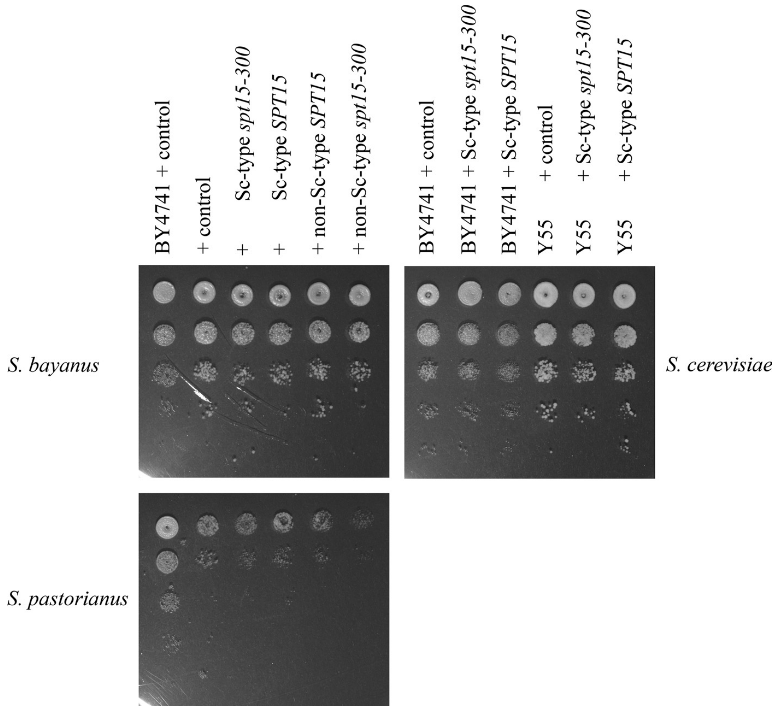


FIG. 1. Growth assays for ethanol tolerance of *SPT15* transformants. *S. cerevisiae* strains BY4741 and Y55, *S. bayanus* NRRL Y-11845 (MCYC 623), and *S. pastorianus* W-34/70 were used. Tenfold serial dilutions of cultures were spotted on YPD agar plates supplemented with 8% ethanol. Plates were photographed after 4 days of incubation at 20°C. Media were supplemented with G418 for plasmid selection. *SPT15* expression was under the control of the wild-type *S. cerevisiae TEF1* promoter. For comparison, *S. cerevisiae* strain BY4741 with the control vector pCJR2 was spotted on the first lane of each plate.

TABLE 2. Growth of *S. cerevisiae* BY4741 transformants in YSC medium

Plasmid selection	Avg OD_{600} ± SEM with:		
	Control plasmid	*SPT15*	*SPT15-300*
Uracil prototrophy	0.006 ± 0.001	0.009 ± 0.001	0.038 ± 0.002
G418 resistance	0.042 ± 0.005	0.028 ± 0.002	0.092 ± 0.005

[a] Transformants were incubated for 20 h at 30°C with shaking. Cultures were inoculated to an OD_{600} of 0.01 in YSC medium with 2% glucose and 6% ethanol. *SPT15* and *SPT15*-300 expression was under the control of the weaker *TEF1* promoter ($P_{TEF1mut2}$; *URA3* or G418 resistance selection). Cultivations were performed in triplicate.

cerevisiae type. *SPT15* gene variants (Table 1) were obtained as follows: (i) a 753-bp BamHI (blunt ended with Klenow fragment)-SpeI-digested *S. cerevisiae*-type *SPT15* fragment was amplified from *S. cerevisiae* BY4741 genomic DNA by PCR, (ii) a 749-bp SpeI-SmaI-digested *S. cerevisiae*-type *SPT15-300* fragment was obtained as a synthetic gene (GenScript, Piscataway, NJ) according to the mutant sequence as described by Alper et al. (2), (iii) a 775-bp SpeI-EcoRV-digested non-*S. cerevisiae*-type *SPT15* fragment was amplified from *S. pastorianus* W-34/70 genomic DNA by PCR, and (iv) a fragment containing non-*S. cerevisiae*-type *SPT15-300* was constructed by cloning a 259-bp SpeI-BglII-digested 5′ fragment of non-*S. cerevisiae*-type *SPT15* (in which the BglII site was introduced by silent mutation using PCR) to a 504-bp BglII-SmaI digested 3′ fragment of *S. cerevisiae*-type *SPT15-300* (exchange of gene fragments was possible, since the encoded *S. cerevisiae*-type and non-*S. cerevisiae*-type Spt15 proteins differ only at amino acids 31 and 36). The correct sequence of each vector was confirmed by DNA sequencing.

Growth and ethanol tolerance assays. The growth phenotypes of *SPT15* transformants were examined as described by Alper et al. (2). In short, yeast transformants were precultured in YSC (2) or SC (32) medium containing various amounts of glucose as indicated and diluted to an optical density at 600 nm (OD_{600}) of 0.01 in fresh medium supplemented with various amounts of ethanol as indicated. The OD_{600} was measured after 20 h of incubation at 30°C with shaking. In the case of G418 selection, YSC media were supplemented with 300 μg/ml G418. The ethanol tolerance of *SPT15* transformants was analyzed on plate assays in which solid medium (as indicated) was supplemented with 6% or 8% ethanol (as indicated). Tenfold serial dilutions of cell cultures, pregrown in appropriate media (as indicated), at an OD_{600} of 1.0 (initial dilution) were spotted on plates. Growth assays were performed in triplicate. Results of representative experiments are shown.

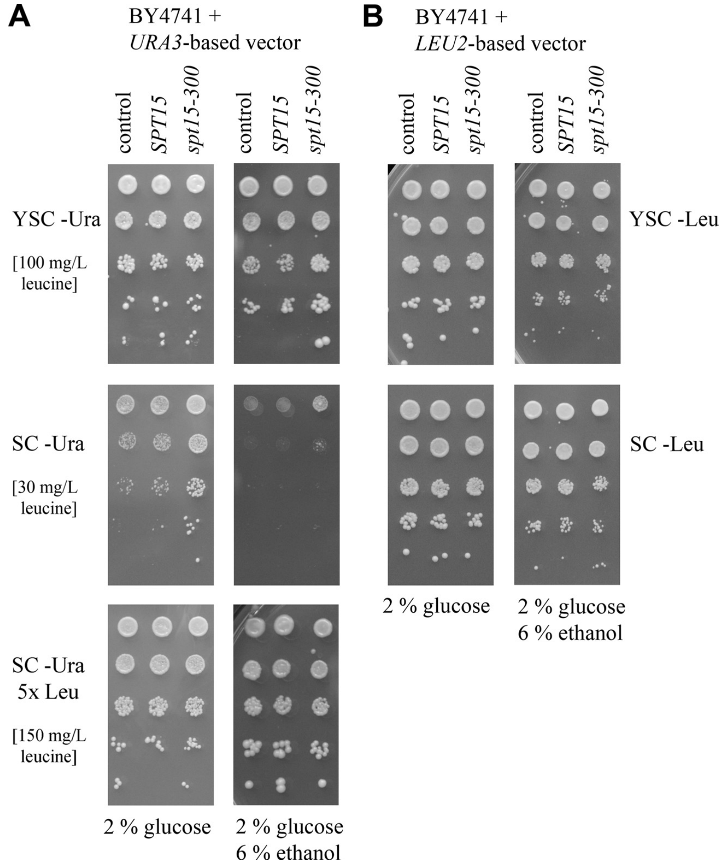


FIG. 2. Growth of transformants of *S. cerevisiae* strain BY4741 harboring *URA3*-based (A) or *LEU2*-based (B) vectors without insert (control) or with *SPT15* or *SPT15-300* on different defined media in the presence or absence of 6% ethanol. *SPT15* expression was under the control of the mutant *S. cerevisiae TEF1* promoter, i.e., $P_{TEF1mut2}$. Tenfold serial dilutions of cell cultures were spotted on the plates. Plates were photographed after 2 (A) (2% glucose), 3 (B) (2% glucose), 4 (B) (2% glucose plus 6% ethanol), and 7 (A) (2% glucose plus 6% ethanol) days of incubation at 30°C.

Microarray analysis. The microarray data (accession no. GSE5185) were downloaded from the Geo Expression Omnibus database (4) and analyzed using R and Bioconductor (13). Array "GSM116825" (*SPT15-300* plus 60 g/liter glucose and 5% ethanol) was identified as an outlier and removed, and only probes specific for *S. cerevisiae* were used in our analyses. *rma* was used for quantile normalization and probe index calculations, and these were subsequently normalized using Qspline (19, 39). For statistical testing, two-factor analysis of variance was used, with the factors "genotype" (i.e., wild type versus *SPT15-300*) and "medium" (i.e., 20 g/liter glucose [medium A] versus 60 g/liter glucose and 5% ethanol [medium B]). The false-discovery rate (FDR) was estimated using a Monte Carlo approach, and statistical significance was set at an FDR of 0.005.

RESULTS AND DISCUSSION

Encouraged by the report by Alper and coworkers (2), we were interested in applying gTME to yeasts in order to improve their ethanol tolerance and ultimately fermentation performance. As a first step, we decided to evaluate the effect of the *SPT15-300* mutant allele identified by Alper et al. (2) in various yeast species of industrial importance. The lager brewing yeast, *Saccharomyces pastorianus*, is a hybrid between *S. cerevisiae* and a *Saccharomyces* species related to *S. bayanus* (21, 25). Genes in the genome of lager brewing yeast that show a high percentage of sequence identity to genes found in *S. cerevisiae* are called *S. cerevisiae* type, while genes more distantly related are called non-*S. cerevisiae* type. We introduced the three point mutations identified in *SPT15-300* in both types of genes, and the wild-type and mutant *SPT15* genes were subsequently inserted into plasmids under the control of the wild-type *S. cerevisiae TEF1* promoter. Since ethanol sensitivity/tolerance screens are generally performed in rich complex media, i.e., YPD supplemented with various amounts of ethanol ranging from 6 to 12.5% (12, 16–18, 35, 37, 40), we analyzed the growth of *SPT15* transformants of *S. cerevisiae* BY4741 and Y55 (JT20150), *S. bayanus* NRRL Y-11845, and *S. pastorianus* W-34/70 on rich complex solid medium (i.e., YPD) supplemented with 8% ethanol. This ethanol percentage was arbitrarily chosen in order to analyze "ethanol-resistant" and "ethanol-sensitive" yeasts (such as *S. cerevisiae* Y55 and *S. pastorianus* W34/70, respectively) under one single condition. Unfortunately, none of the yeasts harboring the mutant *SPT15-300* gene displayed the expected improved ethanol tolerance (Fig. 1). As a control and in order to repeat the experiments described by Alper et al. (2), plasmids that carried the wild-type and mutant *SPT15* genes under the control of the weaker mutant version of the *S. cerevisiae TEF1* promoter were constructed (1, 2, 27). In agreement with the findings of Alper et al., we found that the *S. cerevisiae* laboratory strain BY4741 transformed with *SPT15-300* showed an apparent increase in ethanol tolerance in liquid YSC medium (2) regardless of whether the *SPT15* genes were expressed from a *URA3*- or a G418 selection-based plasmid (Table 2). Evaluation of the contribution of promoter strength to the appearance of ethanol tolerance in cells expressing *SPT15-300* (i.e., comparison of *SPT15* expression using the native and mutant *S. cerevisiae TEF1* promoters) demonstrated that the effect was stronger when the native promoter was used than when the weaker mutant version was used (data not shown). The enhanced ethanol tolerance of cells carrying the *SPT15-300* allele was also apparent on SC plates with 30 mg/liter leucine but not on SC plates containing 150 mg/liter leucine (Fig. 2A). When transformants were grown in YSC medium that contained 100 mg/liter leucine, prepared as described by Alper et al. (2), the enhanced ethanol tolerance was only marginally manifested (Fig. 2A). In the absence of ethanol, the enhanced growth of cells with the *SPT15-300* allele was also noticeable on SC−Ura medium with 30 mg/liter leucine (Fig. 2A). Thus, the apparent

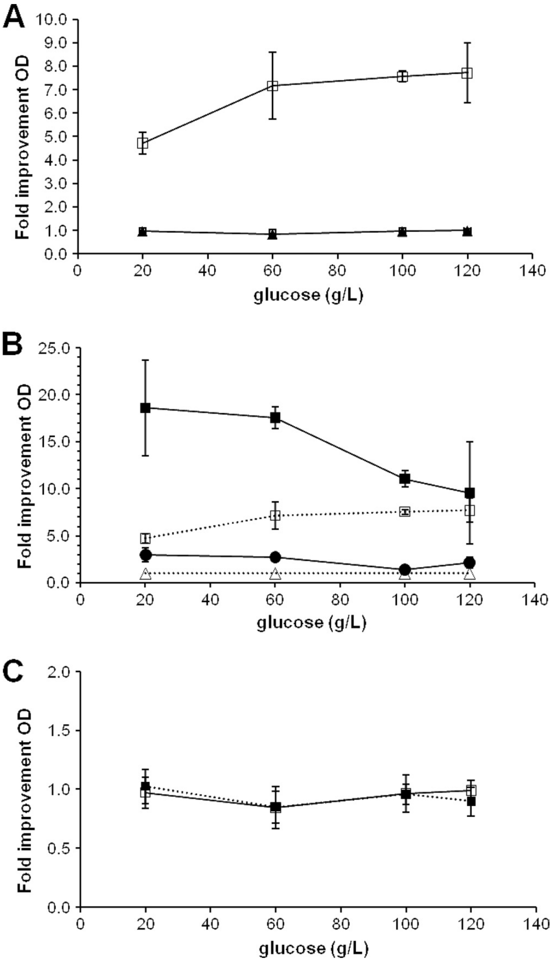


FIG. 3. Growth (OD_{600}) of *S. cerevisiae* strain BY4741 expressing the *SPT15-300* mutant gene relative to that of cells expressing the wild-type *SPT15* gene (expression was under the control of the mutant *S. cerevisiae TEF1* promoter, i.e., $P_{TEF1mut2}$). Cells were inoculated in SC media with 20, 60, 100, or 120 g/liter glucose. Cultivation analyses were performed in triplicate (error bars display standard deviations). (A) Cells expressing *SPT15-300* or *SPT15* from a *URA3* plasmid were inoculated into SC−Ura containing 30 mg/liter leucine (open squares), or cells expressing *SPT15-300* or *SPT15* from a *LEU2* plasmid were inoculated into SC−Leu (closed triangles). (B) Cells expressing *SPT15-300* or *SPT15* from a *URA3* plasmid were inoculated into SC−Ura containing 30 mg/liter leucine in the absence of ethanol (open squares) or in the presence of 4% (closed squares), 5% (closed circles), or 6% (open triangles) ethanol. (C) Cells expressing *SPT15-300* or *SPT15* from a *LEU2* plasmid were inoculated into SC−Leu in the absence of ethanol (open squares) or in the presence of 6% ethanol (closed squares).

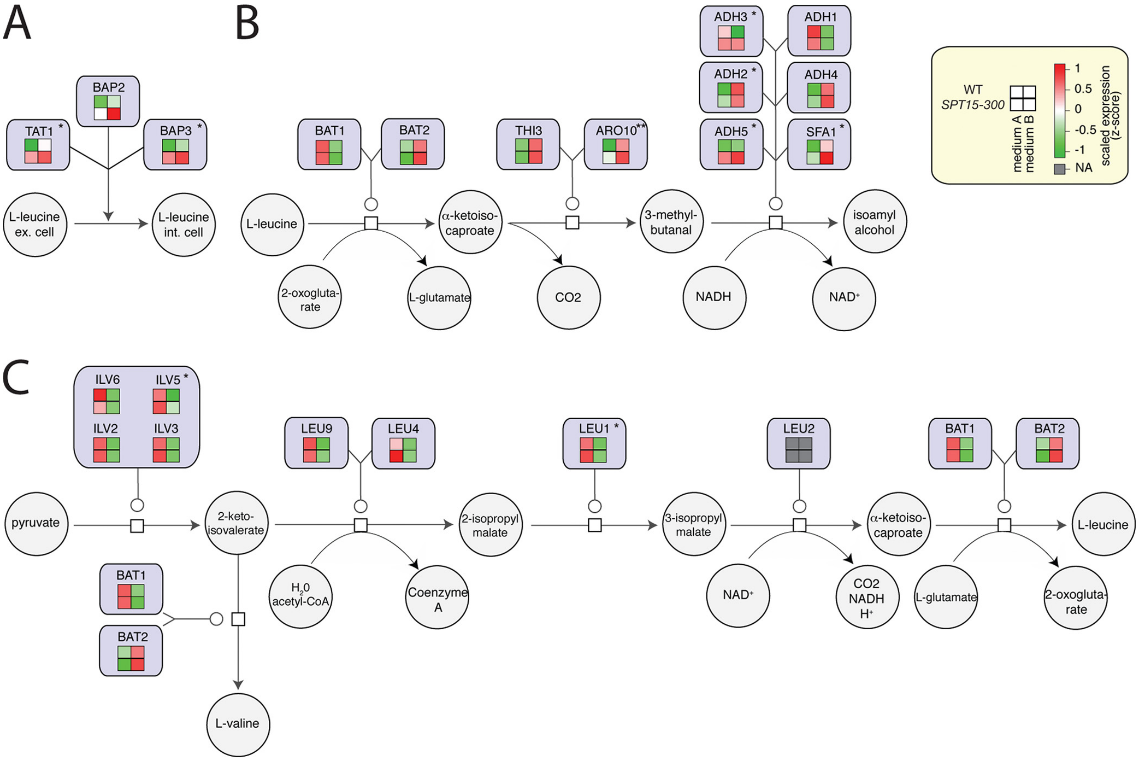


FIG. 4. Reevaluation of the expression data published by Alper et al. (2), specified to genes involved in leucine uptake (A), leucine degradation (B), and leucine biosynthetic (C) pathways in *S. cerevisiae* obtained from SGD (http://www.yeastgenome.org). The color map under each gene shows the scaled regulation of each gene (*z* score), with green being less expressed and red highly expressed. The color map is organized so that the rows represent the genotype (upper, wild-type *SPT15*; lower, mutant *SPT15-300*) and the columns represent the applied media (left, medium A [20 g/liter glucose]; right, medium B [60 g/liter glucose, 5% ethanol]) (see yellow box). All genes except *ADH2*, *ADH3*, *ADH5*, and *LEU4* are significant (FDR of 0.005) for medium B. Genes for which expression is significantly changed by the genotype (wild-type versus *SPT15-300*) are indicated with a single asterisk; those that are significantly altered by the genotype and medium are highlighted by a double asterisk. The *LEU2* gene is deleted in the BY4741 strain. Abbreviations: ex. cell., extracellular; int. cell., intracellular.

improved ethanol tolerance could be related to the improved growth of the *SPT15-300* mutant in media containing smaller amounts of leucine. The *S. cerevisiae* laboratory strain BY4741 is deficient in leucine biosynthesis due to the deletion of the *LEU2* gene, encoding β-isopropylmalate dehydrogenase, the third enzyme in leucine biosynthesis (5). Therefore, we examined the effect of the *SPT15* wild-type and mutant alleles inserted into a *LEU2*-containing plasmid in BY4741 cells transformed to leucine prototrophy. Cells containing either the wild-type or the mutant allele grew equally well on solid media lacking leucine (SC−Leu) whether or not ethanol was present (Fig. 2B). This indicated that the observed increased growth of cells harboring the *SPT15-300* allele indeed is related to improved uptake and/or utilization of leucine.

We also tested the growth of the transformed cells in liquid SC-based media containing different amounts of glucose (Fig. 3). In cells transformed with *URA3*-based plasmids, the *SPT15-300* mutant showed enhanced growth in SC−Ura medium containing 30 mg/liter leucine at all glucose concentrations tested, while cells transformed with *LEU2*-based plasmids did not show this effect of the *SPT15-300* allele (Fig. 3A). In SC−Ura media containing 30 mg/liter of leucine and 5% or 6% ethanol, growth was severely slowed and thus the apparent growth advantage of the *SPT15-300* mutant was reduced (Fig. 3B). However, at 4% ethanol the growth advantage of the mutant was still noticeable, in particular at lower glucose concentrations (Fig. 3B). When the growth experiments were performed with liquid SC medium containing 20 g/liter glucose and 150 mg/liter leucine, the growth advantage of the mutant *SPT15-300* allele was absent (1.1-fold growth improvement; standard deviation, 0.0 [data not shown]). In the presence of 4, 5, or 6% ethanol, the fold growth improvement was limited (1.7 to 1.8; standard deviation, 0.1 [data not shown]). As was the case on solid media, cells transformed with the wild-type and mutant *SPT15* alleles on a *LEU2* plasmid grew equally well in SC−Leu media with different glucose concentrations, without or with 6% ethanol (Fig. 3C).

These growth experiments illustrate that the enhanced growth of cells with the *SPT15-300* mutant allele could be distinguished only in media with limiting amounts of leucine and when expressed from plasmids that do not complement the *LEU2* mutation in BY4741 (i.e., *URA3*- or G418 selection-based plasmids). This implies that the beneficial growth advantage of cells expressing the *SPT15-300* mutation is the result of enhanced uptake and/or improved utilization of leucine.

The ethanol sensitivity of *S. cerevisiae* strains with single-gene deletions (commonly leucine auxotrophic strains) has been determined mainly in rich complex media (12, 16–18, 35, 37, 40). Therefore, a possible effect of ethanol on leucine uptake and/or utilization has not been reported in these global screens. However, impairment of amino acid transport and/or utilization in yeast by ethanol has been described (11). Recently, Hirasawa et al. reported that tryptophan uptake might be inhibited by high concentrations of ethanol (16). Overexpression of *TAT2*, encoding a high-affinity tryptophan and tyrosine permease (30), yielded yeast cells that acquired a higher tolerance toward ethanol (16). Likewise, the known growth defect of *S. cerevisiae leu2* strains (e.g., BY4741) on SC media (8) could be alleviated by overexpression of *TAT1* or *BAP2* (both encoding amino acid permeases that transport leucine [14, 30]) or by reintroducing *LEU2* (8). Several studies have demonstrated that the amount of leucine provided in commonly used synthetic media is limiting for growth of leucine-requiring strains, and authors therefore recommend supplementing synthetic media with at least 400 mg leucine per liter (6, 29).

Based on the growth phenotypes, we decided to reinvestigate the microarray data published by Alper et al. (2), now focusing on uptake and metabolism of leucine. Genes involved in the uptake and degradation of leucine showed differential expression due to the *SPT15-300* mutations but also in the presence of increased glucose and ethanol (medium B). The *TAT1* gene, which codes for a tyrosine and tryptophan amino acid permease, and the *BAP3* gene, which codes for a branched-chain amino acid permease, showed increased expression in cells with the mutant *SPT15-300* allele in both media (i.e., media A and B) compared to cells harboring the wild-type *SPT15* (Fig. 4A). *BAP2* (coding for another branched-chain amino acid permease) showed a similar expression profile but was not significant at a FDR of 0.005. These three genes code for permease proteins that are able to transport leucine across the plasma membrane (14, 30), and increased expression of *TAT1* and *BAP2* has been shown to alleviate reduced growth of BY4741 on SC media (8). A majority of the genes involved in leucine utilization and degradation show statistically significant differential expression for the *SPT15-300* mutations (a genotype effect, i.e., an effect on gene expression when comparing cells expressing the *SPT15-300* allele to those expressing the wild-type allele). *ARO10*, coding for one of the Ehrlich pathway decarboxylases involved in leucine degradation, is significantly upregulated both by the *SPT15-300* mutations and by the presence of increased glucose and ethanol (Fig. 4B) (38). Additionally *ADH3*, *ADH5*, and *SFA1*, coding for alcohol dehydrogenases, show upregulated expression in the *SPT15-300* mutant compared to wild-type *SPT15* cells, suggesting that higher rates of NADH reoxidation via 3-methylbutanal reduction in the *SPT15-300* mutant could account for the increased fitness of cells with the *SPT15-300* allele under leucine-limiting conditions. Finally, genes involved in leucine biosynthesis were downregulated by the presence of ethanol (Fig. 4C). Expression of the *ILV5* and *LEU1* genes was only slightly changed due to the presence of the *SPT15-300* mutations compared to wild-type *SPT15* (genotype effect), though the direction of the response was unchanged. As expected, the *LEU2* gene showed only background expression, while genes coding for branched-chain amino acid aminotransferases, i.e., *BAT1* and *BAT2*, showed inverse expression correlating with the cells transitioning from logarithmic to growth-arrested phase (10). It is therefore likely that the improved growth of the *SPT15-300* mutant under leucine-limiting conditions is due to increased uptake and utilization of leucine (9, 28, 36).

Alper et al. (2) unambiguously demonstrated that gTME is applicable to *S. cerevisiae* for altering its properties. Unfortunately, the properties of cells with the mutant *SPT15-300* allele did not result in increased ethanol-tolerant phenotypes of yeast in rich complex media, but the application of gTME has been reported to improve xylose fermentation in *S. cerevisiae* (22).

ACKNOWLEDGMENTS

We thank Lisbeth Faldborg for excellent technical assistance, J. M. Thevelein for providing the *S. cerevisiae* JT20150 strain, and C. P. Kurtzman for kindly donating *S. bayanus* strain NRRL Y-11845. We are grateful to M. C. Kielland-Brandt and J. Dietvorst for valuable discussions and critical reading of the manuscript.

REFERENCES

1. **Alper, H., C. Fischer, E. Nevoigt, and G. Stephanopoulos.** 2005. Tuning genetic control through promoter engineering. Proc. Natl. Acad. Sci. USA **102:**12678–12683.
2. **Alper, H., J. Moxley, E. Nevoigt, G. R. Fink, and G. Stephanopoulos.** 2006. Engineering yeast transcription machinery for improved ethanol tolerance and production. Science **314:**1565–1568.
3. **Ausubel, F. M., R. Brent, R. E. Kingston, D. D. Moore, J. G. Seidman, J. A. Smith, and K. Struhl.** 1996. Current protocols in molecular biology. John Wiley & Sons Inc., New York, NY.
4. **Barrett, T., D. B. Troup, S. E. Wilhite, P. Ledoux, D. Rudnev, C. Evangelista, I. F. Kim, A. Soboleva, M. Tomashevsky, and R. Edgar.** 2007. NCBI GEO: mining tens of millions of expression profiles—database and tools update. Nucleic Acids Res. **35:**D760–D765.
5. **Brachmann, C. B., A. Davies, G. J. Cost, E. Caputo, J. Li, P. Hieter, and J. D. Boeke.** 1998. Designer deletion strains derived from *Saccharomyces cerevisiae* S288C: a useful set of strains and plasmids for PCR-mediated gene disruption and other applications. Yeast **14:**115–132.
6. **Çakar, Z. P., U. Sauer, and J. E. Bailey.** 1999. Metabolic engineering of yeast: the perils of auxotrophic hosts. Biotechnol. Lett. **21:**611–616.
7. **Cliften, P., P. Sudarsanam, A. Desikan, L. Fulton, B. Fulton, J. Majors, R. Waterston, B. A. Cohen, and M. Johnston.** 2003. Finding functional features in *Saccharomyces* genomes by phylogenetic footprinting. Science **301:**71–76.
8. **Cohen, R., and D. Engelberg.** 2007. Commonly used *Saccharomyces cerevisiae* strains (e.g. BY4741, W303) are growth sensitive on synthetic complete medium due to poor leucine uptake. FEMS Microbiol. Lett. **273:**239–243.
9. **Derrick, S., and P. J. Large.** 1993. Activities of the enzymes of the Ehrlich pathway and formation of branched-chain alcohols in *Saccharomyces cerevisiae* and *Candida utilis* grown in continuous culture on valine or ammonium as sole nitrogen source. J. Gen. Microbiol. **139:**2783–2792.
10. **Eden, A., G. Simchen, and N. Benvenisty.** 1996. Two yeast homologs of ECA39, a target for C-myc regulation, code for cytosolic and mitochondrial branced-chain amino acid aminotransferases. J. Biol. Chem. **271:**20242–20245.
11. **Ferreras, J. M., R. Iglesias, and T. Girbés.** 1989. Effect of the chronic ethanol action on the activity of the general amino-acid permease from *Saccharomyces cerevisiae* var. *ellipssoideus*. Biochim. Biophys. Acta **979:**375–377.
12. **Fujita, K., A. Matsuyama, Y. Kobayashi, and H. Iwahashi.** 2006. The genome-wide screening of yeast deletion mutants to identify the genes required for tolerance to ethanol and other alcohols. FEMS Yeast Res. **6:**744–750.
13. **Gentleman, R. C., V. J. Carey, D. M. Bates, B. Bolstad, M. Dettling, S. Dudoit, B. Ellis, L. Gautier, Y. Ge, J. Gentry, K. Hornik, T. Hothorn, W. Huber, S. Iacus, R. Irizarry, F. Leisch, C. Li, M. Maechler, A. J. Rossini, G. Sawitzki, C. Smith, G. Smyth, L. Tierney, J. Y. H. Yang, and J. Zhang.** 2004. Bioconductor: open software development for computational biology and bioinformatics. Genome Biol. **5:**R80.
14. **Grauslund, M., T. Didion, M. C. Kielland-Brandt, and H. A. Andersen.** 1995.

BAP2, a gene encoding a permease for branched-chain amino acids in *Saccharomyces cerevisiae*. Biochim. Biophys. Acta **1269:**275–280.

15. **Güldener, U., S. Heck, T. Fiedler, J. Beinhauer, and J. H. Hegemann.** 1996. A new efficient gene disruption cassette for repeated use in budding yeast. Nucleic Acids Res. **24:**2519–2524.
16. **Hirasawa, T., K. Yoshikawa, Y. Nakakura, K. Nagahisa, C. Furusawa, Y. Katakura, H. Shimizu, and S. Shioya.** 2007. Identification of target genes conferring ethanol stress tolerance to *Saccharomyces cerevisiae* based on DNA microarray data analysis. J. Biotechnol. **131:**34–44.
17. **Hu, X. H., M. H. Wang, T. Tan, J. R. Li, H. Wang, L. Leach, R. M. Zhang, and Z. W. Luo.** 2007. Genetic dissection of ethanol tolerance in the budding yeast *Saccharomyces cerevisiae*. Genetics **175:**1479–1487.
18. **Inoue, T., H. Iefuji, T. Fujii, H. Soga, and K. Satoh.** 2000. Cloning and characterisation of a gene complementing the mutation of an ethanol-sensitive mutant of *sake* yeast. Biosci. Biotechnol. Biochem. **64:**229–236.
19. **Irizarry, R. A., B. Hobbs, F. Collin, Y. D. Beazer-Barclay, K. J. Antonellis, U. Scherf, and T. P. Speed.** 2003. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics **4:**249–264.
20. **Kellis, M., N. Patterson, M. Endrizzi, B. Birren, and E. S. Lander.** 2003. Sequencing and comparison of yeast species to identify genes and regulatory elements. Nature **423:**241–254.
21. **Kodama, Y., M. C. Kielland-Brandt, and J. Hansen.** 2005. Lager brewing yeast, p. 145–164. *In* P. Sunnerhagen and J. Piškur (ed.), Comparative genomics: using fungi as models. Springer-Verlag, Berlin, Germany.
22. **Liu, H., L. Xu, M. Yan, C. Lai, and P. Ouyang.** 2008. gTME for construction of recombinant yeast co-fermenting xylose and glucose. Chin. J. Biotech. **24:**1010–1015.
23. **McCusker, J. H., and J. E. Haber.** 1988. Cycloheximide-resistant temperature-sensitive lethal mutations of *Saccharomyces cerevisiae*. Genetics **119:** 303–315.
24. **Mumberg, D., R. Mailer, and M. Funk.** 1995. Yeast vectors for the controlled expression of heterologous proteins in different genetic backgrounds. Gene **156:**119–122.
25. **Nakao, Y., T. Kanamori, T. Itoh, Y. Kodama, S. Rainieri, N. Nakamura, T. Shimonaga, M. Hattori, and T. Ashikari.** 2009. Genome sequence of the lager brewing yeast, an interspecies hybrid. DNA Res. doi:10.1093/dnares/dsp003.
26. **Nevoigt, E.** 2008. Progress in metabolic engineering of *Saccharomyces cerevisiae*. Microbiol. Mol. Biol. Rev. **72:**379–412.
27. **Nevoigt, E., J. Kohnke, C. R. Fischer, H. Alper, U. Stahl, and G. Stephanopoulos.** 2006. Engineering of promoter replacement cassettes for fine-tuning of gene expression in *Saccharomyces cerevisiae*. Appl. Environ. Microbiol. **72:**5266–5273.
28. **Overkamp, K. M., B. M. Bakker, P. Kötter, A. van Tuijl, S. de Vries, J. P. van Dijken, and J. T. Pronk.** 2000. In vivo analysis of the mechanisms for oxidation of cytosolic NADH by *Saccharomyces cerevisiae* mitochondria. J. Bacteriol. **182:**2823–2830.
29. **Pronk, J. T.** 2002. Auxotrophic yeast strains in fundamental and applied research. Appl. Environ. Microbiol. **68:**2095–2100.
30. **Regenberg, B., L. Düring-Olsen, M. C. Kielland-Brandt, and S. Holmberg.** 1999. Substrate specificity and gene expression of the amino-acid permeases in *Saccharomyces cerevisiae*. Curr. Genet. **36:**317–328.
31. **Sambrook, J., E. F. Fritsch, and T. Maniatis.** 1989. Molecular cloning: a laboratory manual, 2nd ed. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.
32. **Sherman, F.** 1991. Getting started with yeast. Methods Enzymol. **194:**3–21.
33. **Sikorski, R. S., and P. Hieter.** 1989. A system of shuttle vectors and yeast host strains designed for efficient manipulation of DNA in *Saccharomyces cerevisiae*. Genetics **122:**19–27.
34. **Stephanopoulos, G.** 2007. Challenges in engineering microbes for biofuels production. Science **315:**801–804.
35. **Takahashi, T., H. Shimoi, and K. Ito.** 2001. Identification of genes required for growth under ethanol stress using transposon mutagenesis in *Saccharomyces cerevisiae*. Mol. Genet. Genomics **265:**1112–1119.
36. **van Dijken, J. P., E. van den Bosch, J. J. Hermans, L. R. de Miranda, and W. A. Scheffers.** 1986. Alcoholic fermentation by 'non-fermentative' yeasts. Yeast **2:**123–127.
37. **van Voorst, F., J. Houghton-Larsen, L. Jønson, M. C. Kielland-Brandt, and A. Brandt.** 2006. Genome-wide identification of genes required for growth of *Saccharomyces cerevisiae* under ethanol stress. Yeast **23:**351–359.
38. **Vuralhan, Z., M. A. Luttik, S. L. Tai, V. M. Boer, M. A. Morais, D. Schipper, M. J. H. Almering, P. Kötter, J. R. Dickinson, J. M. Daran, and J. T. Pronk.** 2005. Physiological characterization of the *ARO10*-dependent, broad-substrate-specificity 2-oxo acid decarboxylase activity of *Saccharomyces cerevisiae*. Appl. Environ. Microbiol. **71:**3276–3284.
39. **Workman, C., L. J. Jensen, H. Jarmer, R. Berka, L. Gautier, H. B. Nielsen, H. H. Saxild, C. Nielsen, S. Brunak, and S. Knudsen.** 2002. A new non-linear normalization method for reducing variability in DNA microarray experiments. Genome Biol. **3:**research0048.1–0048.16.
40. **Yoshikawa, K., T. Tanaka, C. Furusawa, K. Nagahisa, T. Hirasawa, and H. Shimizu.** 2009. Comprehensive phenotypic analysis for identification of genes affecting growth under ethanol stress in *Saccharomyces cerevisiae*. FEMS Yeast Res. **9:**32–44.

4.2 Perspectives

In the original publication of the *SPT15-300* mutant the authors attempted to identify the underlying genetic changes of the mutant using among others, a Systems Biology based approach including protein-protein interaction networks. *S. cerevisiae* is one of the best characterized model organisms, with unparalleled depth of protein-protein interaction knowledge, however the pertubations had very large impact on cellular transcription, probably concealing the true underlying biology. Here we with the added information of additional laboratory experiments were able to take a more directed approach analyzing the data at the level of biological pathways.

Part III

Systems biology in disease

Chapter 5

Alzheimer's Disease - the ADIT project

This chapter is about the work that we have performed as a part of the European Union (EU) Framework Programme 6 (FP6) project: "Design of Small Molecule Therapeutics for the Treatment of *A*lzheimer's *D*isease Based on the Discovery of *I*nnovative Drug *T*argets" (ADIT). It is one of three large integrated projects from FP6 that spans from basic molecular and cellular understanding of the disease to the identification, validation and development of drug targets. The project was launched in June 2005 with eight partners in several EU countries and Siena Biotech S.p.A. in Italy as the coordinating partner. The project aims at identifying novel drug targets for Alzheimer's Disease (AD) and to develop small molecule drugs targeting these. In this process we have been involved in identifying biological entities suitable for drug targeting part of creating the base for further progress in the project. To do this we have performed gene expression profiling using Affymetrix DNA microarrays on a rat AD model and used integrative approaches to understand the neuronal response.

The work presented in the following is unpublished but will, hopefully, be part of future patent applications. Currently two targets from the gene expression profiling have progressed to small molecule (hit and lead) identification and optimization, and additional three targets have progressed to immunohistochemistry in human post-mortem brain tissue. The chapter is organized with an introduction to AD, a short discussion of the drug discovery process taken and a manuscript in preparation covering the identification of ADAM23 as associated with AD.

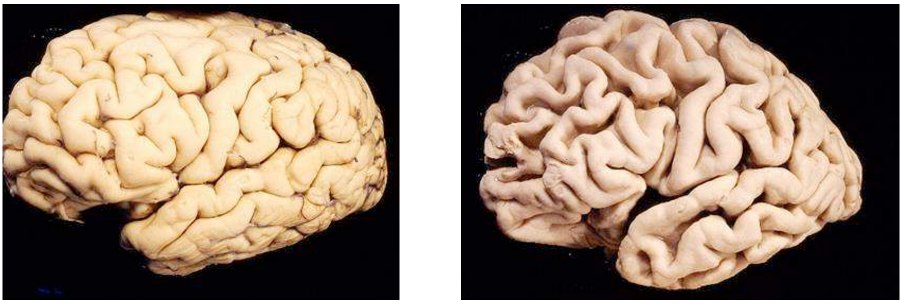


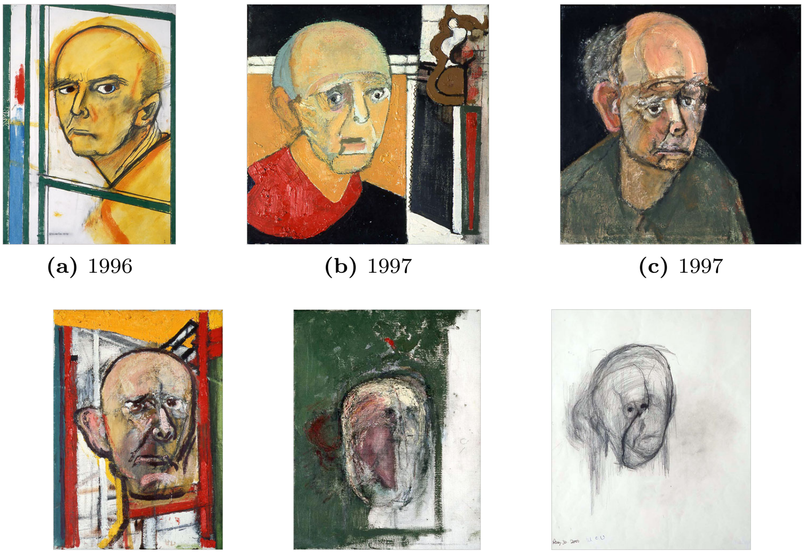
Figure 5.1 – Normal brain of an aged person (**Left**) and brain of a patient suffering from Alzheimer's Disease (**Right**).

5.1 Introduction to Alzheimer's Disease

Alzheimer's Disease (AD) is the most common neurodegenerative disease and causes dementia that leads to a progressive pattern of cognitive impairments [116]. It is named after the German researcher Dr. Alois Alzheimer that in 1906 presented the case of "*Frau Auguste D.*", a patient that had developed an unusual dementia at age 51. Here he described the main pathological hallmarks of AD, the marked neurofibrillary tangles, widespread presence of plaques and neuronal degeneration [117]. The severe degree of degeneration in AD is obvious from Figure 5.1 where a brain from an AD patient is compared to a brain from a non-demented person of similar age.

The disease consists of four clinical stages: Pre-dementia, early/mild dementia, moderate dementia and severe dementia with an average survival after clinical diagnosis of 5 to 8 years. Typically other diseases such as pneumonia followed by myocardial infarctions and sepsis (a whole-body inflammation) is the cause of death [116]. The clinical manifestations of AD begins from mild impairment in acquiring new information (pre-state) to a rapid decline in cognitive abilities such as memory, learning, language and reading resulting in changes of the personality such as depression, delusion, restlessness, aggression and disorientation. The cognitive decline has been described artistically by William Utermohlen (1933-2007) which was diagnosed with Alzheimer's Disease and decided to visualize the progression through self-portraits painted at different time points – gradually closer to his death (see Figure 5.2) [118]. The character of the disease naturally sets a high strain on the caretakers and family with high social and emotional impact [116]. The average annual costs of AD is estimated to US$80-100 billion and 55 billion euro in the US and EU, respectively, resulting in a very high economical impact on society. As AD occurs with an increasing rate in elderly people ($\sim$4.5 million total in the US and 42% in the population above 84 years of age) coupled with increasing life expectancy, AD threatens to become an even larger burden on society [119–121].

Even though AD has been known for more than a century and been studied intensively, e.g. a search for "Alzheimer's Disease" on Pubmed retrieves more than 48.000 abstracts, no curative therapy exist [122]. There are four symptomatic medications approved for AD that have two different modes of actions. Three of the medications inhibit the breakdown of acetylcholine by the enzyme acetylcholinesterase. This increases the concentration of acetylcholine in the brain which counteracts the loss of this neurotransmitter caused by failing and dying cholinergic neurons [123]. The other type of treatment is based on the neurotransmitter glutamate and is focused on the inhibition of NMDA receptors. Because glutamate is such as powerful neurotransmitter too high concentrations can, via the NMDA receptor, lead to excitotoxicity and neuronal death [124]. However the two approaches, and a combination hereof, have shown only little curative effect and in addition especially the acetylcholinesterase inhibitors are associated with side effects [125,126]. There is therefore a high unmeet medical demand for medications with curative effects on AD.



(a) 1996 **(b)** 1997 **(c)** 1997

(d) 1998 **(e)** 1999 **(f)** 2000

Figure 5.2 – Six self-portraits by the artist William Utermohlen 1933-2007 describing his gradual decent into the dementia of Alzheimer's Disease.

Genetics of Familial AD

In general AD can be categorized into two categories, one being Familial Alzheimer's Disease (FAD) and the other being Late-Onset Alzheimer's Disease (LOAD). The FAD form is caused by mutations in the amyloid beta (A4) precursor protein (APP), presenilin 1 (PSEN1) or presenilin 2 (PSEN2) and is therefore inherited - hence a *familial* form of AD. The disease phenotype is autosomal-dominant, meaning that a heterozygous state of a mutation in any of these genes is enough to cause disease. Additionally FAD is characterized by occurring before the age of 60, but is a rare form of AD responsible for only a fraction ($< 5\%$) of the total amount of cases [127, 128]. Understanding the genetics of FAD and AD in general sparked with the discovery that the Amyloid-β (Aβ) peptide, present in the brain plaques of AD patients, was also found in Down syndrome patients [129]. The Amyloid Precursor Protein (*APP*) was cloned and it was discovered to be located on chromosome 21, the same chromosomal area that is copied in Down syndrome (trisomy 21) [130,131]. From this several mutations in the APP protein was discovered in AD patients leading to the understanding that the Aβ peptide is naturally occurring in the brain, but also that perturbations in the processing of APP to Aβ may be at the core of the disease [132]. However this also added new questions regarding AD as mutations in *APP* could only explain FAD and not LOAD and additionally only 10% of the FAD cases had mutations in the *APP* gene. This began further investigations of disease causing alleles for FAD and lead to the identification of *PSEN1* and *PSEN2*. These proteins are located at the catalytic center of γ-secretase, one of the protein complexes involved in processing of APP to Aβ [133,134]. To date all cases of autosomal dominant AD cases (FAD) can be explained through Aβ [135].

Genetics of Late-Onset AD

On the contrary LOAD which comprises by far the majority of AD cases (>95%) does not show obvious familial aggregation and is hence also termed sporadic AD [136]. It occurs in patients older than 65 years of age and the risk of developing LOAD increases rapidly with age to almost 50% among those older than 85 [121]. The genetics of LOAD has proven more difficult to assess than FAD, however it was discovered in 1993 that the $\epsilon4$ allele of Apolipoprotein E (*APOE*) is a large risk factor for developing the disease [137–139]. In humans the *APOE* gene exists in three different polymorphic alleles, $\epsilon2$, $\epsilon3$ and $\epsilon4$ with the $\epsilon3$ allele being the most frequent one (77%). The $\epsilon2$ and $\epsilon3$ alleles are not associated with Alzheimer's disease, whereas the *APOE*-$\epsilon4$ allele has been shown to be implicated in more than 50% of LOAD cases [139]. Additionally gene dosage of this allele increases the risk of developing the disease and minimizes the age of onset. This was shown in a study by Corder et al. where 90% of the persons with this genotype developed LOAD [137]. The APOE protein is involved in brain-cholesterol transport where APOE delivers cholesterol and other lipids from astrocytes to neurons. Additionally it may be involved in signaling pathways with functionalities such as neuronal migration and synaptic plasticity, and binding and trafficking of Aβ. However the role of APOE in AD it is not completely understood [139].

Although the *APOE*-ϵ4 allele is a risk factor for developing LOAD, the identification of additional genetic risk factors have proven to be a challenging task. Where the *APOE* locus has a large impact on the risk of developing LOAD, other risk alleles are thought to have small penetrance. These alleles demand high statistical power to be identified, exemplified by the Genome-Wide Association Study (GWAS) of Coon *et al.* where 1086 brain donors (664 AD cases, 422 controls) were assayed for Single Nucleotide Polymorphisms (SNPs) associated with Alzheimer's Disease [136, 140]. χ^2-test for disease association revealed only one very significant SNP[1], which is in strong Linkage Disequilibrium (LD) with the *APOE* locus (Figure 5.3). However recently two very large studies of Harold et al. and Lambert et al. both identify another apolipoprotein, clusterin (CLU), also know as APOJ, as a risk factor for LOAD [10, 11]. Interestingly APOE and APOJ are the most abundant apolipoproteins in the central nervous systems [141, 142]. At the time of writing 35 genes in total have been associated with LOAD [8].

In addition to the genetic component of LOAD, which is thought to comprise 60% or more of the disease susceptibility, environmental risk factors are also involved [136, 143]. As previously mentioned, age is a large risk factor for AD, but other factors such as long-term hypertension, diabetes and obesity, viral infections and neuroinflammation, chronic stress and head injury have been associated with the disease. On the contrary protective factors have also been described, among them are intellectual stimulation and social interactions, regular physical activity as well as vitamin and omega-3 fatty acid rich diets [144].

The amyloid hypothesis

During the development of Alzheimer's Disease the pathological hallmarks neuron degeneration, plaques and neurofibrillary tangles appear at different stages. Extracellular plaques that consists of mostly of Aβ and also cellular material are the first to appear during the disease development with intracellular neurofibrillary tangles (NFTs) consiting of hyperphosphorylated microtubule-associated protein tau (MAPT) appearing downstream of these. The disease progression mainly develops from the transentorhinal and entorhinal cortex, and hippocampal area with plaques, NFTs and neuronal dysfunction and loss, spreading throughout the cerebral cortex [145–147]. There have been many discussions regarding which of these that are the primary disease causing effects [148, 149]. For example, it is not well established how Aβ exerts toxic effects on neurons and the amount of

[1] Although in the study GAB2 was identified having weak association to LOAD when samples were divided according to *APOE*-ϵ4 allele status

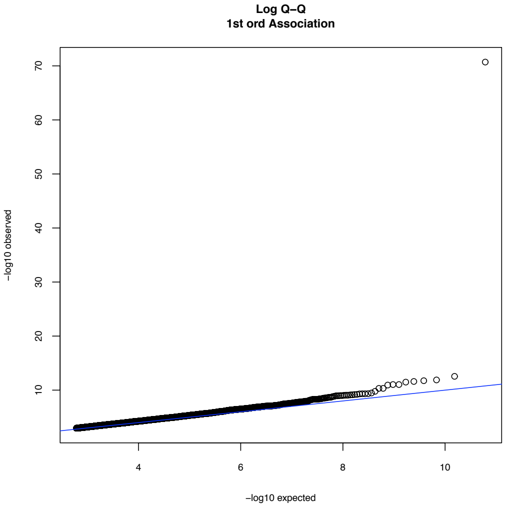


Figure 5.3 – GWAS using the Affymetrix 500K GeneChip by Coon *et al.* [140]. -log10 to expected vs. observed p-values is shown from χ^2 test of disease association between cases and controls. The very significant SNP is in strong LD with the *APOE* locus.

Aβ plaques does not correlate well with cognitive impairment. Although it has been suggested that the amount of soluble oligomeric Aβ peptides correlate with cognitive impairment this is still disputed [150, 151]. On the contrary NFTs correlate well with cognitive impairment, and studies of Frontotemporal Dementia and Parkinsonism linked to chromosome 17 (FTDP17) show that tau mutations are sufficient to trigger neuronal degeneration [152]. However tau aberrations and NFTs are thought to develop downstream in time of Aβ plaques as these are not observed in in FTDP17 [149]. Issues like these are plenty-fold within the AD field, underlining the complexity and the lack of thorough understanding of the disease. Still the amyloid hypothesis is the most supported hypothesis for AD [148].

Generation of Aβ

Amyloid-β is generated from the processing of APP. APP is a transmembrane protein expressed in a variety of cells in different isoforms, with the most abundant form in the brain of 695 aa (APP695). The protein undergoes proteolytic cleavage by either α- or β-secretases, and the resulting peptides from both processing pathways are cleaved by γ-secretase (Figure 5.4). If APP is first cleaved by α-secretase, the subsequent cleavage by γ-secretase results in the production of a

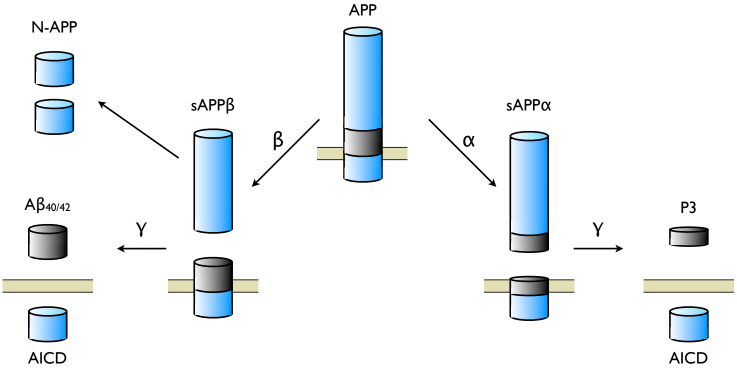


Figure 5.4 – Different fates of APP upon proteolytic cleavage by α, β and γ secretase. (**Left**) Intial cleavage by β secretase generates toxic Aβ species, whereas (**Right**) initial cleavage by α secretase generates the non-toxic fragment P3.

non-toxic secreted peptide (P3). However if β-secretase is the first to cleave APP, the additional cleavage by γ-secretase results in the generation of the toxic species, Aβ_{40} and Aβ_{42} [145]. In addition to the above peptides, the α- and β-secretase pathways also generates the sAPPα and sAPPβ peptides of which the latter can be cleaved to the N-terminal APP peptide (N-APP) [153, 154]. Furthermore from both proteolytic pathways the APP IntraCellular Domain (AICD) is produced.

Regarding the secretases, α-secretases are members of the A Disintegrin And Metalloprotease (ADAM) family where ADAM9, ADAM10, ADAM17 and ADAM19 have been associated to the cleavage of APP, whereas β-secretase has been defined as beta-site APP-cleaving enzyme 1 (BACE1) [155–158]. γ-secretase is thought to be composed of presenilins 1 or 2 (PSEN1/2), which as mentioned above are among the genes involved in the development of Familial AD, nicastrin, Aph-1 and Pen-2 [159].

Function of APP and derived peptides

The complete functionality of APP and its derived peptides is not completely understood, however it is suggested that APP processing controls signaling for multiple physiological functions. Processing via the α-secretase pathway the sAPPα peptide is involved in promoting neuronal survival and neurite outgrowth [154,160]. On the contrary processing via BACE1 and hence the β pathway generates products that impair and deactivate neuronal function and viability. Long-Term-Potentiation (LTP), which is the basis for inducing learning and memory is inhibited by nanomolar concentrations of oligomeric Aβ_{42} and it can trigger anti-synaptic function. Generation of N-APP have recently been shown to induce axon

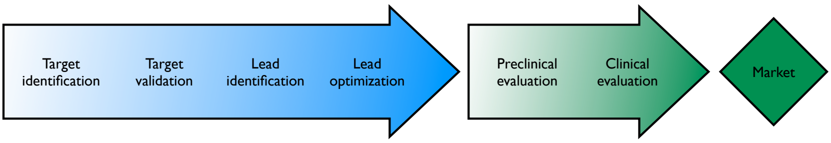


Figure 5.5 – Drug development pipeline. Cost of the steps increases dramatically as the candidates are progressed towards clinical evaluation and the market. Adapted from [162].

pruning and neuron death by binding to death receptor 6 (DR6). It is important to emphasize that even the "negative" functions of APP signaling are necessary physiological behavior, e.g. to control precise neuronal connectivity (N-APP) and to prevent over-excitation of neurons ($A\beta_{42}$) [154]. To add to the complexity of APP mediated signaling Puzzo et al. (2008) has shown that picomolar concentrations of $A\beta_{42}$ actually can enhance LTP, the opposite of nanomolar concentrations of $A\beta_{42}$ [161]. Additionally the AICD peptide, generated from both processing pathways, can form a transcriptionally active complex with Fe65 and Tip60 [160].

5.2 Drug discovery in ADIT

The process of drug discovery and design can be described as beginning with identifying an unmeet need or by the discovery of a new disease, until the arrival of a new drug. In the case of AD, the unmeet need is huge as no curative therapies exist and as incidence correlates with increasing age. This is emphasized by the many ongoing clinical trials for AD where, by the time of writing this thesis, at least 11 drug candidates were in clinical phase II or III [163].

The initial step in drug development for a particular disease, is to establish the targets involved in the pathology of the disease. In this project focus has been on the traditional small molecule drug targets such as G-Protein-Coupled Receptors (GPCRs), nuclear receptors, ligand- and voltage-gated ion channels [164,165]. These protein families are generally relatively easy to drug due to their function and localization. Additionally proteins that require inhibition and not activation has been preferred as it is more likely to design a small molecule that can block the function of an enzyme or ion-channel, compared to increase the efficiency of a protein that, by nature, is already very efficient. The latter may in some cases be possible if an inhibitor of the protein can be targeted.

Identification of a gene target gives rise to target validation, which is the process of gathering scientific information through either literature mining or experiments. Several approaches have been applied for this, ranging from qPCR verifications to immunohistochemistry in human post-mortem AD brains and AD transgenic mice models. To functionally validate the role of a target, single-cell imaging in combi-

Step	**Technology**
Target identification	Transcriptomics, Proteomics
Confirmation of differential expression	qPCR
Non-experimental evaluation	IP, Literature, AD focused pathways
Experimental evaluation (mRNA)	Gene expression profile in human tissues panel and post-mortem AD brains
Experimental evaluation (protein)	Target analysis in AD transgenic animals and human post-mortem AD brains
Functional validation	Neuronal viability to Aβ by overexpression and silencing

Table 5.1 – Technologies used in target identification and validation in the ADIT project.

nation with protein over-expression and RNAi has been used – all the techniques that has been applied are summarized in Table 5.1.

When a target is validated as having a functional and drugable role screening for chemical compounds begins. The approach for this is to use High-Throughput Screening (HTS), where numerous chemical compounds from compound libraries are tested for activity against the target. The hits that are identified do usually only have some degree of activity and have to be further developed to lead candidates and lead series. For this several parameters has to be considered such as Absorption, Distribution, Metabolism, Excretion and Toxicity (ADMET) properties, bioavailability and the Lipinski rule-of-five. Additionally the activity against the drug target has to be optimized while reducing the activity against unrelated targets to prevent side effects [166]. This, and other experiments, progresses the lead candidates into one or few clinical candidates which are advanced through clinical phases leading to a fully developed drug (Figure 5.5). As the cost of developing novel drugs increases dramatically while progressing through the drug discovery pipeline, it is important to avoid continuation of non-optimal targets, hits and leads. The average spending per chemical entity developed as a drug is US$500-800 million [167].

5.3 Transcriptomics of Alzheimer's Disease

The system used to model AD was primary cultures of rat cortical neurons treated with Aβ peptides, which induces dendritic degeneration and apoptosis of the neurons [168]. Previous studies utilizing this model have mainly focused on the late events in the neuronal response whereas this approach aims at the early response. These effects may contain more disease-relevant mechanisms than later stages – as effects measured at late stages are likely to be secondary and tertiary effects.

The gene expression changes were investigated using the Affymetrix Gene Chip Rat genome 230 2.0 arrays in a total of 6 experimental batches (Table 5.2). Several forms of amyloid-β peptides exists and at least 18 different forms have been determined in AD patients [169]. The, believed, most pathological relevant species are A$\beta_{1\text{-}42}$ and to a lesser degree A$\beta_{1\text{-}40}$, however laboratory experiments using A$\beta_{1\text{-}42}$ peptide is not trivial. Issues of formulation of the peptide and reproducibility of the results have led to widespread use of the non-naturally occuring A$\beta_{25\text{-}35}$ peptide. Although an artificial truncated peptide, it has proven to give reproducible results and to elicit neuronal toxicity and biochemical changes observed in animal models and AD patients [170]. This peptide was therefore the primary peptide applied in the experiments. The use of A$\beta_{1\text{-}42}$ was attempted in batch 4, however formulation of this peptide was later shown to be inaccurate and a re-formulated version was assayed in batch 6.

5.4 Materials and methods

Primary neuronal cultures

Pure primary cortical cultures were prepared from E16 embryos of Sprague-Dawley rat neocortex by mechanical dissociation and cultured in Neurobasal medium supplemented with B27 (Invitrogen). All experiments were carried out according to the ECC guidelines for animal care (DL 166/92, application of the European Communities Council Directive 86/609/EEC). Cultures were maintained in a humidified incubator at 5% CO_2 at 37°C and grown in 96-well plates (200 μl medium/well). The stimulant (A$\beta_{25\text{-}35}$, A$\beta_{35\text{-}25}$, A$\beta_{1\text{-}42}$ or A$\beta_{42\text{-}1}$) was applied for the indicated times (0, 45, 120, 180 or 240 min) at concentrations at 1, 10 or 50 μM. The Aβ peptides were rehydrated in sterile water and incubated for 2 h at 37°C to allow fibril formation prior to treating cells. RNA was extracted using the RNAeasy Plus Mini Kit and RNA quality was verified by the use of Agilent Bioanalyzer.

RNA extraction, microarray data generation and analysis

1 μg RNA per stimulation condition was converted into cDNA, and biotin-labeled aRNA was synthesized using the MessageAmpTM II-Biotin Enhanced Kit (Ambion) according to the manufacturers instructions. The samples were hybridized to Gene Chip Rat genome 230 2.0 Array (Affymetrix), comprising 31.000 probe sets representing over 28.000 rat genes. The arrays were stained, washed and scanned according to the manufacturer's instructions. The data was analyzed using *R* and *Bioconductor* [171]. Raw probe intensities were normalized using *qspline* and expression index calculations were performed using *rma* [78, 172]. Statistical testing was performed using either t-test or anova and were performed using logit-transformation and at probe-level [173]. The false discovery rate (FDR) [86] was estimated using a Monte Carlo approach and used to determine statistical significance.

Batch	Inoculates	Concentration (μM)	Replicates	Timepoints (min)	Arrays
1	control, A$\beta_{25\text{-}35}$	50	3	0, 45, 240	9
2	control, A$\beta_{25\text{-}35}$	50	3	0, 45, 240	8
3	control, A$\beta_{25\text{-}35}$	50	3	0, 45, 120, 180, 240	13
4	control, A$\beta_{25\text{-}35}$, A$\beta_{1\text{-}42}$, A$\beta_{35\text{-}25}$	50	3	0, 45, 240	12
5	control, A$\beta_{25\text{-}35}$	1, 10	3	0, 45, 240	18
6	control, A$\beta_{25\text{-}35}$, A$\beta_{1\text{-}42}$, A$\beta_{35\text{-}25}$, A$\beta_{42\text{-}1}$	50	5	0, 240	30
All					90

Table 5.2 – Microarray experiments performed in the ADIT project

5.5 Results

Microarray data

The microarray experiments were performed from January 2006 (batch 1) to July 2008 (batch 6). Performing DNA microarray experiments over an extended period of time will generate data biases due to changes in the experimental environment. This is observed by Singular Value Decomposition of the data resulting in a clear separation of batches in four clusters (Figure 5.6). Prior to batch 4, the experiments had primarily investigated the reproducibility of batch 1 and attempted to expand the time resolution by two time points of 120 and 180 minutes of stimulation. For batch 4 a reversed peptide A$\beta_{35\text{-}25}$ was applied to test the specificity of the A$\beta_{25\text{-}35}$ response and to eliminate genes responding to peptides in general. Additionally, to compare the data generated from A$\beta_{25\text{-}35}$, the pathological A$\beta_{1\text{-}42}$ peptide was applied, however the peptide formulation was later demonstrated to be of unsure quality. Due to the inter-batch differences revealed from the SVD, a reference data-set for use in target identification analyses was chosen. For this benchmarking of each batch against four gene sets were performed, where the gene sets was composed by AD genes from Metacore, genes co-occurencing in PubMed abstract with "Alzheimer" or "Amyloid-β", AD neuronal inflammation literature and proteins identified from phospho-proteomic approaches within the ADIT project (Figure 5.7). As there were no major overall differences in the performances and that batch 1 had already been extensively used for target identification this batch was chosen as the reference set. By selecting the reference set we decided to investigate A$\beta_{25\text{-}35}$ concentration response in batch 5.

For the last approach, batch 6, a new formulation of A$\beta_{1\text{-}42}$ was used to attempt to finish what had begun in batch 4 – comparing the truncated peptide with the pathological one. A SVD of the data clearly reveal differences between the two peptides, however the reversed A$\beta_{42\text{-}1}$ show similar effects as the functional peptide (Figure 5.6b). Again this emphasizes the value of A$\beta_{25\text{-}35}$ in *in vitro* settings, where the reversed A$\beta_{35\text{-}25}$ elicits a response similar to saline control. Issues with non-reproducible results and non-functional control experiments would not be limited

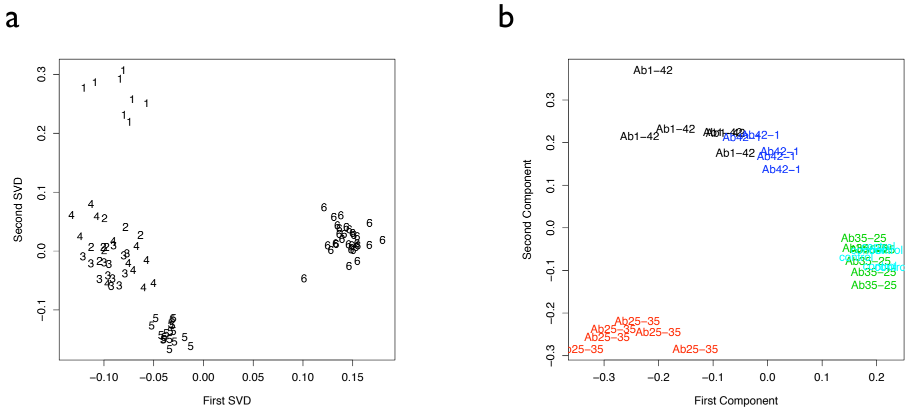


Figure 5.6 – Singuar Value Decomposition (SVD) of microarray data. **a**. All DNA microarrays experiments run in the ADIT project, each array is represented by the batch number. **b**. Top 1000 ranked genes from ANOVA of batch 6 data. Arrays are represented and colored by stimulant added. Control: cyan, A$\beta_{25\text{-}35}$: red, A$\beta_{35\text{-}25}$: green, A$\beta_{1\text{-}42}$: black, A$\beta_{42\text{-}1}$: blue.

to the target identification approaches, but would additionally be a problem for the functional validation assays.

5.6 Perspectives

Current status on drug discovery in the project is that lead molecules have been identified for two targets and these are currently being progressed to lead development. Additionally a third target is undergoing High-Throughput Screening to identify small molecule hits. Two of the targets originates from the transcriptomic approaches whereas a third target was identified from pre-existing knowledge. Additionally it has not been possible to progress targets from proteomic approaches.

At the moment we are in the initial phases of setting up experiments for profiling the transcriptional response of neurons when exposed to lead molecules and known target antagonists in combination with Aβ. From these experiments we hope to investigate the mode of action of the lead compounds and possibly identify side effects. Additionally a manuscript is in preparation for the discovery of ADAM23, which is presented below.

5.7 Paper III

In this manuscript we have used a method originally devised for prioritization of disease candidate genes within linkage intervals [15]. In this approach we prioritize

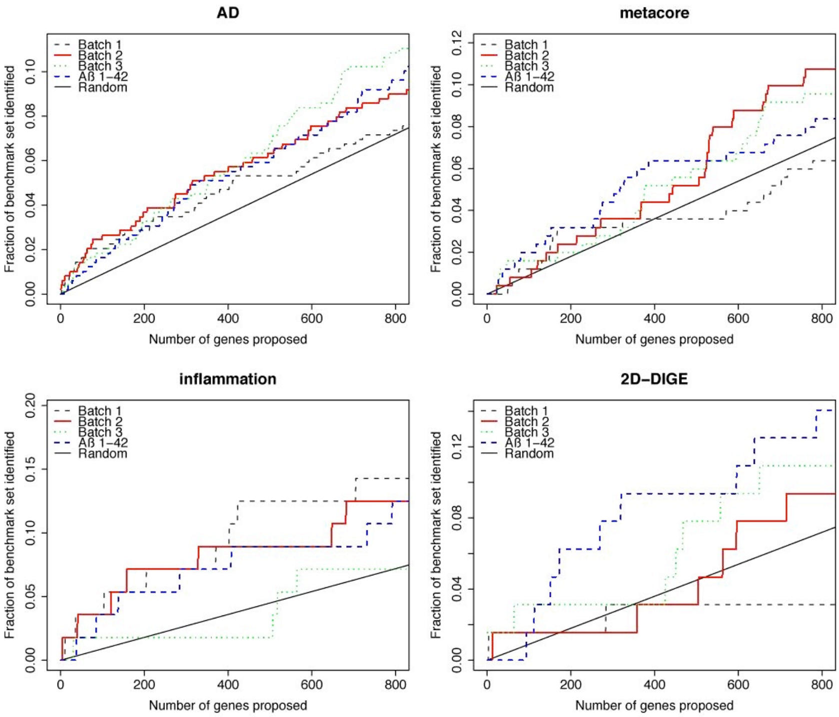


Figure 5.7 – Fraction of benchmark set identified as a function of genes proposed from batch 1–4 sorted by statistical significance. Batch 1–3 is based on A$\beta_{25\text{-}35}$ treatment whereas batch 4 is based on A$\beta_{1\text{-}42}$ treatment. Batch 1: black striped, batch 2: red full, batch 3: green dotted, batch 4: blue striped and random performance: black full.

the significant genes from the microarray study based on protein interactions and the association of each protein in the resulting networks to Alzheimer's Disease. From this we identified A Disintegrin And Metalloprotease 23 (ADAM23) as highly prioritized, and as co-localized and interacting with several core AD proteins. Investigation of ADAM23 protein expression in human post-mortem brain tissue of AD cases revealed it to be up regulated and localized in NeuroFibrillary Tangles (NFTs), one of the pathological hallmarks of AD. However some of the experiments in the project has not yet been performed and we are currently performing or planning the following,

- Immunostaining has only been performed using one antibody (Anti-ADAM23 propetide domain – Abcam 28304) and there could be possible cross-reaction with ADAM22 propeptide domain. We are in the process of reproducing the results with another antibody.

- Up regulation of ADAM23 in AD brain tissue is currently being tested using Western blotting analysis to allow for statistical evaluation of ADAM23 expression.
- We intend to investigate the expression of ADAM23 protein and its localization pattern at different stages of AD development.
- Additionally we are considering to perform qPCR of RNA extracted from post-mortem brain tissue to investigate if known splice variants are differentially expressed. At least three different transcript variants have been described for ADAM23, which differ in the C-terminal part [174].

ADAM23 is associated with Alzheimer's Disease in the human brain

Simon Rasmussen[1], David C. Hondius[2], Jeroen J.M. Hoozemans[2], Henrik B. Nielsen[1], and Søren Brunak[1]

[1] Center for Biological Sequence Analysis, Department of Systems Biology, Technical University of Denmark, DK-2800 Lyngby, Denmark
[2] Department of Pathology, VU University Medical Center, Amsterdam, The Netherlands

Abstract

Gene expression profiling, in combination with integrative protein-phenotype data, was used to investigate the early response of rat cortical neurons to amyloid-beta ($A\beta_{25\text{-}35}$). From this A Disintegrin And Metalloprotease 23 (ADAM23) was identified as the only differentially regulated gene in an interaction network of proteins associated with AD such as APOE, CLU, APLP1/2 and APBB1. Using immunohistochemical staining of ADAM23 protein levels in post-mortem Alzheimer's Disease brain tissue, ADAM23 was found to have increased expression in neurons of the temporal cortex. The expression was observed as tangle-like staining suggesting that the increased expression of ADAM23 protein is associated with neurofibrillary tangles (NFTs) in human AD. However, as ADAM23 is normally localized at the cell membrane in neurites, this suggests that the increased protein levels does not confer to an increase of ADAM23 functionality. ADAM23 functionality is thought to be through its integrin binding domain in cell-extracellular matrix signaling and to mediate axonal growth. Additionally ADAM23 has been found to bind the cellular prion protein, which has recently been shown to mediate $A\beta$ inhibition of Long-Term Potentiation. From this we suggest that ADAM23 is neuroprotective.

5.8 Introduction

Alzheimer's Disease (AD) is the most common neurodegenerative disease leading to progressive cognitive impairment and dementia. AD pathology is characterized by the major hallmarks, extracellular amyloid plaques and intracellular neurofibrillary tangles (NFTs). The most established hypothesis of AD pathogenesis is based on the amyloid-beta (Aβ) peptide where imbalance in Aβ production and clearance is hypothesized as the primary cause of AD [149, 175]. The second hallmark, NFTs primarily constitutes of aberrantly phosphorylated Microtubule-Associated Protein Tau (MAPT), an important protein in the dynamics of the microtubules. In AD, NFTs are primarily observed in neurons and their occurrence correlates well with the severity of cognitive impairment [176, 177]. For Late-Onset Alzheimer's Disease (LOAD), which comprises by far the majority of AD cases, the epsilon-4 (ϵ4) allele of apolipoprotein E (APOE) has been identified as carrying a significant part of the genetic heritability. Recently very large scale Genome-Wide Association (GWA) studies have associated clusterin (CLU), another apolipoprotein, with LOAD, however the majority of genes associated, except APOE, are characterized by low penetrance [10, 11, 178].

To study AD, we employed primary cultures of rat cortical neurons treated with Aβ peptides inducing dendritic degeneration and apoptosis of the neurons [168]. Using DNA microarrays and proteomic approaches this model has previously been employed to characterize the molecular response of neurons to Aβ [179–182]. These are mainly focused on late events whereas this approach aims at the early Aβ response increasing the likelihood of discovering primary events in the neuronal response [183].

A Disintegrin And Metalloprotease 23 (ADAM23) is a member of the ADAM family of transmembrane proteins of which ADAM11, ADAM22 and ADAM23 are predominantly expressed in the nervous system and conserved in mammals [184, 185]. Specifically ADAM23 is found highly expressed in the cerebral cortex pyramidal cells, in the CA1 and CA3 pyramidal cells of the hippocampus and cerebellar Purkinje cells [186–188]. The metalloprotease domain, which is a characteristic feature of ADAM proteins seem to be inactive in ADAM23 and functionality is thought to be mediated via the integrin binding domain. This domain has been found to bind $\alpha v\beta 3$ integrin of nervous cells promoting cell adhesion and signaling [189, 190]. Additionally ADAM23 has been found to mediate neurite outgrowth as binding of LGI1 to ADAM23 *in vitro* stimulates neurite outgrowth in hippocampal and cortical cultures [186]. Furthermore, a study by Costa et al. (2009) showed that ADAM23 can bind via its integrin domain to the cellular prion protein PrP$^{\mathrm{C}}$ [191]. Mice devoid of ADAM23, die by postnatal day 14, show less dendritic arborization *in vivo* and develop severe tremor and ataxia [186, 192]. Other members of the ADAM family have been related to AD, such as ADAM9, ADAM10, ADAM17 and ADAM19 which have been associated with α-secretase and cleavage of the amyloid precursor protein (APP), however ADAM23 has not been associated with the disease before [155, 157, 158].

Here we use gene expression profiling of an Aβ rat model and a protein-phenotype integrative approach to identify ADAM23 as differentially regulated and part of a disease candidate complex interacting with core AD proteins. When investigating ADAM23 expression in AD human post-mortem tissue we found increased ADAM23 protein levels and that this is localized with NFTs in neurons.

5.9 Materials and Methods

Primary neuronal cultures

Pure primary cortical cultures were prepared from E16 embryos of Sprague-Dawley rat neocortex by mechanical dissociation and cultured in Neurobasal medium supplemented with B27 (Invitrogen). All experiments were carried out according to the ECC guidelines for animal care (DL 166/92, application of the European Communities Council Directive 86/609/EEC). Cultures were maintained in a humidified incubator at 5% CO_2 at 37°C and grown in 96-well plates (200 μl medium/well). The stimulant (A$\beta_{25\text{-}35}$) was applied at a concentration of 50μM and the neurons were incubated for 45 mins or 4 hours. The Aβ peptides were rehydrated in sterile water and incubated for 2 h at 37°C to allow fibril formation prior to treating cells. RNA was extracted using the RNAeasy Plus Mini Kit and RNA quality was verified by the use of Agilent Bioanalyzer.

RNA extraction, microarray data generation and analysis

1 μg RNA per stimulation condition was converted into cDNA, and biotin-labeled aRNA was synthesized using the MessageAmpTM II-Biotin Enhanced Kit (Ambion) according to the manufacturers instructions. The samples were hybridized to Gene Chip Rat genome 230 2.0 Array (Affymetrix), comprising 31.000 probe sets representing over 28.000 rat transcripts. The arrays were stained, washed and scanned according to the manufacturer's instructions. The data was analyzed using *R* and *Bioconductor* [171]. Raw probe intensities were normalized using *qspline* and expression index calculations were performed using *rma* [78, 172]. The false discovery rate (FDR) [86] was estimated using a Monte Carlo approach, and statistical significance was set at an FDR < 0.01 and absolute fold change of $\log_2 > 0.5$. Only rat genes with human ortholog genes/proteins (Ensembl) were considered resulting in 121 human orthologs for further analysis. The microarray data is available at the Gene Expression Omnibus (GEO) database as GSE0000.

Phenotype association

Phenotype analysis was essentially performed as Lage et al. except human orthologs identified from the micrarray analysis were used as input rather than linkage intervals [15]. In short, for each input gene a virtual pull-down of protein-protein interactions is made. This protein interaction data consist of a pool from seven of the largest databases on human interactions and additionally inferred inter-species model organism data, resulting in a network of ∼62.000 high-confidence interactions. The virtual pull-down is performed by a Bayesian predictor ensuring only high-confidence and interactions supported by network topology, literature, reliable small-scale interaction experiments or a combination of these. Next, for each of the proteins in the network, termed candidate complexes, similarity to Alzheimer's Disease is calculated based on text-mining of Online Mendelian Inheritance in Man (OMIM) records. This score is a measure of phenotypic overlap between phenotypes associated with the proteins in the candidate complexes and Alzheimer's Disease. The OMIMs considered were AD1-14 and Amyloid beta A4 precursor protein, AD susceptibility to mitochondrial and AD familial early-onset, with coexisting amyloid and prion pathology (OMIMs: 104300, 104310, 607822, 606889, 602096, 605526, 606187, 607116, 608907, 609636, 609790, 611073, 611152, 611154, 104760, 502500, 605055). Lastly a second Bayesian predictor scores the candidate complexes by the phenotypes associated with the complex and they are ranked according to how likely the protein complex is involved in Alzheimer's Disease of all input genes considered. The sum of all posterior probabilities of adds to 1 [15]. Clustering of phenotype similarity (OMIMs) were done using the cosine similarity scores as a distance measure using hierarchical clustering in *R*.

Human post-mortem brain tissue

Human brain specimens of probable AD, other dementias and age-matched non-demented control cases were obtained at autopsy with a short post mortem interval (The Netherlands Brain Bank, Amsterdam, The Netherlands). Clinical diagnosis was defined according to DSM-III-R criteria and the severity of dementia was evaluated according to the Global Deterioration Scale of Reisberg (GDS) [193]. Neuropathological evaluation was performed on formalin fixed, paraffin embedded tissue from different sites, including the frontal cortex (F2), temporal pole cortex, parietal cortex (superior and inferior lobule), occipital pole cortex and the hippocampus (essentially CA1 and entorhinal area of the parahippocampal gyrus). The distribution and the density of neurofibrillary tangles was determined using Bodian staining and immunohistochemistry for hyperphosphorylated tau. Senile plaques were stained with the methenamine silver method [194]. Staging of AD was evaluated according to Braak and Braak [147, 195].

Immunohistochemistry of brain tissue

Sections from the temporal cortex (5 μm thick) were mounted on superfrost plus tissue slides (Menzel-Gläser, Germany) and dried overnight at 37°C. For all stainings sections were deparaffinised and subsequently immersed in 0.3% H_2O_2 in methanol for 30 min to quench endogenous peroxidase activity. Normal sera and antibodies were dissolved in phosphate-buffered saline (PBS) containing 1% (w/v) bovine serum albumin (BSA, Boehringer Mannheim, Germany). Sections were treated in 10 mM pH 6.0 sodium citrate buffer heated by autoclave during 10 minutes for antigen retrieval. For the detection of ADAM23 sections were incubated overnight at 4°C with rabbit anti-ADAM23 (1:1600 dilution, Abcam). After washing with PBS sections were incubated with EnVision solution (goat anti-mouse/rabbit HRP, undiluted, DAKO, Glostrup, Denmark). Color was developed using 3,3'-diaminobenzidine (DAB, EnVision Detection system/HRP, 1:50 dilution, DAKO, 10 minutes) as chromogen. Sections were counterstained with hematoxylin and mounted using Depex (BDH Laboratories Supplies, Poole, England).

5.10 Results

Phenotype association identifies ADAM23 candidate complex

Investigating the response of rat cortical neurons stimulated with A$\beta_{25\text{-}35}$ for 45 minutes or 4 hours revealed 121 significantly regulated genes that could be mapped to a human ortholog. These were used as input for the phenotype association study where candidate protein complexes were generated for 112 of the input genes. Each of the proteins in the complexes are linked to their corresponding phenotypes and assessed for similarity to Alzheimer's Disease phenotypes. This yields posterior probabilities of the input genes being involved in AD for the 112 candidate protein complexes times the 17 phenotypes (Figure 5.8a). For 12 of the 17 phenotypes, the protein complex generated by the ADAM23 input gene has high scoring posterior probabilities. As the AD OMIMs aim at describing the Alzheimer's Disease phenotypes – albeit slightly different phenotypes – they are not completely independent from each other. By clustering the pair-wise cosine similarity scores distinct clusters are formed where the candidate complex derived from ADAM23 is the top scoring candidate complex in all clusters except one (Figure 5.8b). Only two other candidate complexes, created from cAMP-dependent protein kinase catalytic subunit beta (PRKACB) and beta-sarcoglycan (SGCB), have posterior probabilities greater than 0.2.

The candidate complex generated from ADAM23 (Figure 5.9), contains proteins that are associated with the amyloid hypothesis such as Apolipoprotein E (APOE), amyloid precursor protein binding B1 (APBB1 / Fe65), APP-like protein 1 and 2 (APLP1/2) and the prion protein (PRNP). Additionally clusterin (CLU),

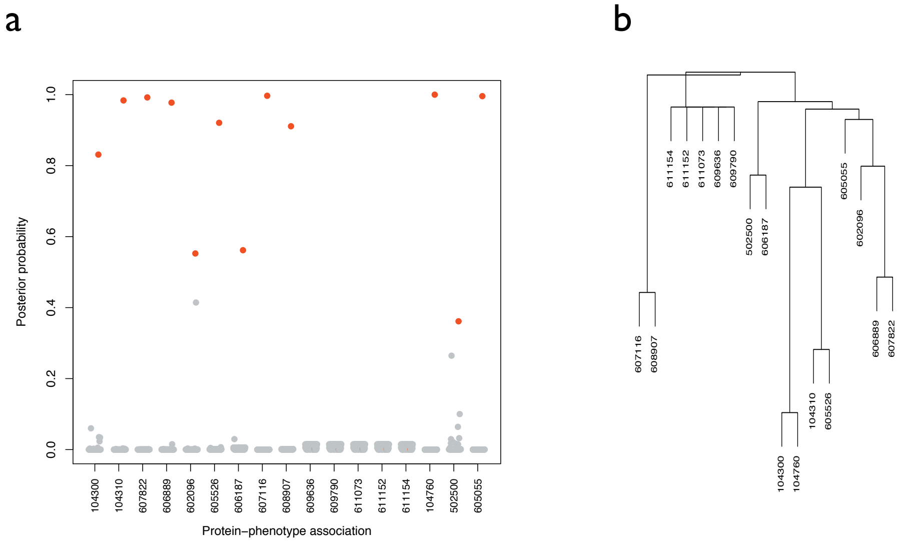


Figure 5.8 – **a**. Posterior probabilities of Bayesian predictor for phenotype association between significant genes from the microarray experiment and 17 Alzheimer's disease OMIM records. ADAM23 candidate complex posterior probabilities are shown in red. **b**. Hierarchical clustering of OMIM record cosine similarity scores showing that the OMIMs are not independent of each other.

which has been found in two recent genome-wide association studies to be associated with AD [10, 11], is also identified in the complex. The interactions in this part of the network are from a study of Schmitt-Ulms et al. (2004) investigating the lipid-raft membrane micro-enviroment of the prion protein in mice [196]. These can therefore not be strictly defined as observations of protein-protein binding, but rather co-localizations and putative protein-protein binding. However ADAM23 has recently been shown to physically bind the cellular prion protein (PrP^C) and to function as a receptor for LGI1, a secreted protein associated with forms of epilepsy [186, 191]. Additional proteins with phenotype similarities with Familial AD (AD1: 104300) in the protein network are L1CAM and RYR1. L1CAM is a neural recognition molecule involved in axonal growth and mutations in L1CAM are responsible for X-linked hydrocephalus, whereas the latter, RYR1 is a sarcoplasmatic/endoplasmatic reticulum Ca^{2+} release channel expressed in muscle and brain [197, 198].

ADAM23 was in the microarray experiment found to be down regulated by $\log_2$ -0.5, whereas none of the other genes coding for the proteins in the ADAM23 candidate complex were significantly regulated (data not shown).

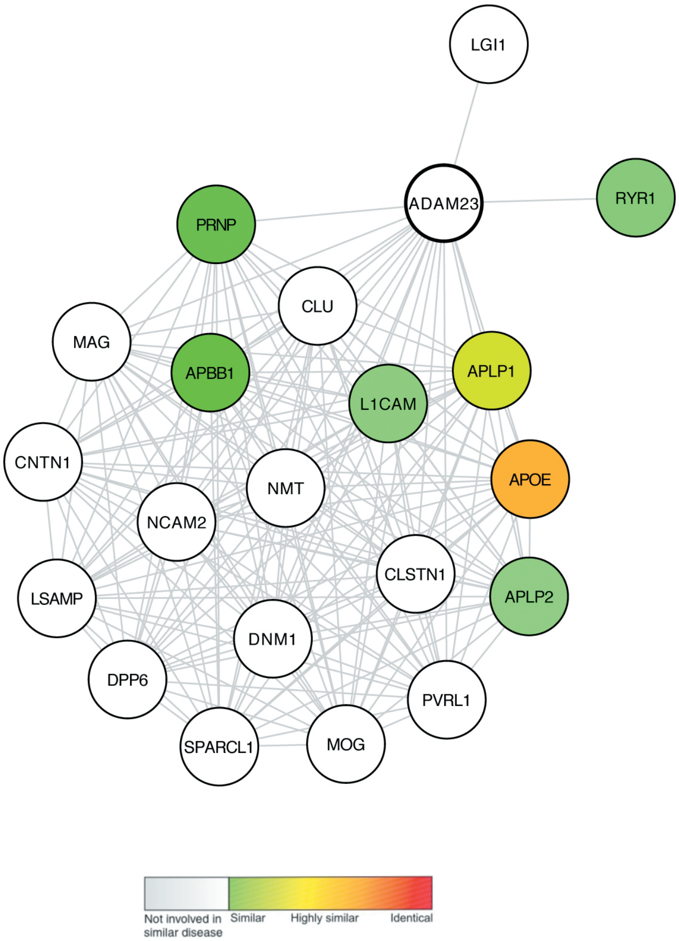


Figure 5.9 – ADAM23 candidate complex where each node is a protein and edge an interaction. The proteins are colored according to phenotype similarity to Familial AD (AD1: 104300), where white: no similarity, green: similar, yellow: highly similar and red: identical. An interaction between YWHAZ and ADAM23 was removed as it was not supported in the literature and the recent published interaction between ADAM23 and LGI1 was added [186].

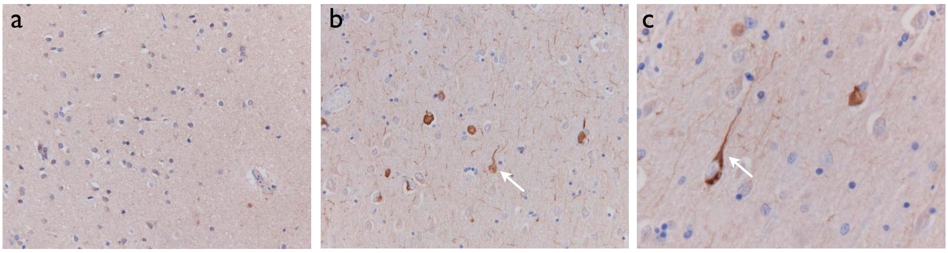


Figure 5.10 – Immunohistochemistry of ADAM23 in AD and control tissue. **a**. ADAM23 staining in the temporal cortex of a non-demented age-matched control case. **b**. The temporal cortex of an AD case, showing tangle-like (arrow) staining of ADAM23. **c**. As *b*, but at greater magnification.

Immunohistochemistry of ADAM23 in human AD brains

Investigation of ADAM23 protein levels in human post-mortem brain tissue revealed increased expression of ADAM23 in AD brain tissue compared to non-demented age-matched control cases (Figure 5.10). Staining of ADAM23 revealed intracellular tangle-like structures in neurons indicating that increased levels of ADAM23 protein is localized in neurofibrillary tangles. In addition, a diffuse cytoplasmic staining of ADAM23 was observed in both AD and control cases (not shown). This is in good agreement with the findings of Goldsmith et al. (2004) which identified a small cytoplasmic fraction of ADAM23 in healthy rat brain neurons [188]. (To add: We are performing Western Blotting experiments quantifying the differences statistically).

5.11 Discussion

Using a well-established *in vitro* rat model of Aβ induced neuro-toxicity we identified several differentially regulated genes. By applying a protein-phenotype integrative approach to these data ADAM23 was identified as the only differentially regulated gene in a protein interaction network based on membrane co-localization and putative interaction with AD-associated proteins APOE, APBB1, APLP1/2 and CLU. Additionally PrP^{C} and LGI1 are identified as part of the network. ADAM23 has not previously been associated with AD and we therefore investigated protein expression level in the temporal cortex of AD human post-mortem brain tissue. On the contrary of what could have been expected from the transcriptomics data, the immunohistochemistry demonstrated increased levels of ADAM23 staining of AD cases compared to non-demented control cases. The discrepancy could be explained by the different conditions of the two systems such as the organisms studied, Aβ species, concentration of Aβ and by applying an *in vitro* system for studying complex tissue and disease. One may speculate that the decrease in the *in vitro* system reflects an early/fast response. The increase observed in post mortem brain tissue could be due to post-transcriptional events such as alternative splicing, increased translation or decreased removal. However the important finding is that ADAM23 is the only differentially regulated transcript in an early response to Aβ of several core AD related genes.

The immunohistochemistry of ADAM23 revealed a tangle-like staining suggesting that the increased protein levels of ADAM23 in AD brains is localized in neurofibrillary tangles. This indicates that the increased expression of ADAM23 does not confer to an up regulation of ADAM23 functionality, rather retainment in the cytosol, compared to normal membrane bound localization in neurites, would correspond to non-functional ADAM23. An explanation for the association with NFTs, could be that the aggregation of tau and the dysfunction of the microtubule system might impair transport and correct docking of ADAM23 along the axons of the neurons [176]. As the functionalities of ADAM23 is as a regulator of neuronal growth, cell differentiation and integrin contact between cells and the extracellular matrix this can be expected to have detrimental effects on neuronal viability and signaling. This and the finding that ADAM23 transcript is down regulated as an early response of neurons to Aβ suggests that ADAM23 is neuroprotective. Increased expression of ADAM23 protein in AD could be a survival response of the neurons.

Interestingly ADAM23 has recently been shown to co-localize and bind to the cellular prion protein (PrP^{C}), which mediates brain derived Aβ inhibition of Long-Term Potentiation (LTP) in hippocampal CA3 and CA1 pyramidal cells [191,199]. Additionally PrP^{C} has been identified as an inhibitor of β-secretase cleavage of APP and shown to reduce Aβ formation, and other members of the ADAM protein family, ADAM10 and ADAM17, has been shown to process PrP^{C} [200,201]. Overlapping functionality between ADAM23 and PrP^{C} such as involvement in axonal growth, cell adhesion and cell-matrix adhesion and the identification of ADAM23 and PrP^{C} as molecular binders implicate ADAM23 as functionally linked to PrP^{C} [202–204]. Whether ADAM23 can influence the activity or processing of PrP^{C} remains to be seen, however the loss of function of ADAM23 could be important for the role of PrP^{C} in AD.

Interactions between ADAM23 and the other proteins identified in the candidate complex provide additional interesting links to AD. The ϵ4 allele of APOE compose the largest genetic risk factor for LOAD and recently CLU, another apolipoprotein has been associated with the disease. Interestingly APOE was originally associated with AD when it was identified as localized with NFTs before genetic studies identified the ϵ4 allele as a risk factor for LOAD [205]. However currently no physical interaction has been shown between ADAM23 and these.

5.12 Conclusion

Using gene expression profiling and integration with protein-phenotype interaction data ADAM23 is identified as member of a network consisting of core AD-associated proteins. Immunohistological staining of human post-mortem brain tissue revealed that ADAM23 protein levels are significantly increased in AD tissue compared to non-demented control. The functionality of ADAM23 as involved in signaling and neurite outgrowth suggests that stressed neurons could benefit from increased ADAM23 levels, however localization with NFTs implicate that ADAM23 functionality is not effectively increased. The physical interaction between ADAM23 and $\mathrm{PrP^C}$ may suggest a role of ADAM23 in regulating $\mathrm{PrP^C}$ activity and hereby $A\beta$ production or $A\beta$ toxicity. The co-localization in lipid rafts with other core AD proteins such as APOE and CLU may provide interesting alternative pathways to consider.

5.13 Acknowledgements

The work has been supported by a grant from the European Commision FP6 (ADIT, contract no. LSHB-CT-2005-511977). The authors wish to thank Pia Friis for excellent technical assistance, Claus Andersen for ideas, comments and reading the manuscript, and additionally Niclas Tue Hansen and Kasper Lage for valuable input on phenotypic data integration.

Chapter 6

Probiotic bacteria and human health

Microorganisms use many strategies for avoiding the immune system. Of particular interest in this study is the scheme of inducing an intracellular focused response (viral) as an extra-cellular organism. In the case of *Lactobacillus acidophilus*, which is considered beneficial to human health (probiotic), it has been shown in clinical trials to reduce the risk of viral infections [206–208]. In the light of the current (H1N1) and future influenza epidemics this may prove valuable. In the study we investigate the mechanism by which this can occur and show that it is primarily mediated by Toll-Like Receptor 2 (TLR2).

6.1 Introduction to the relevant immunology

The immune system is highly complex and this prompted my previous professor in immunology, Ib Søndergaard, to say,

> "You do not understand anything of immunology, until you understand all of immunology."

Here I will nonetheless try to give a short overview with relation to the topic studied in this paper. Immunology may be divided into two categories – the innate which is non-specific and the adaptive which is specific and can acquire memory. Additionally the adaptive immune system can be divided into two subsystems, the humoral and the cellular immunity. The former is mediated by B cells and plasma cells through antibodies, whereas the latter is mediated by T cells.

Innate Immunity

The first line of defence against infections is the innate immune system, which is comprised of a set of barriers that the pathogen must cross. First, the anatomic barriers (skin and mucous membranes) prevent entry of microbes into the host organism by acting as mechanical barriers as well as creating a hostile environment for the microbes. Next, physiologic barriers such as temperature and pH are an array of defences that the body employ. However once inside the host, pathogens have to deal with phagocytic barriers - cells such as monocytes, tissue macrophages, dendritic cell and neutrophils which ingest and kill foreign microorganisms. Finally there are inflammatory barriers where tissue damage and detection of pathogens results in an influx of phagocytic cells as well as vascular fluid containing antibacterial serum proteins into the affected area.

Innate and adaptive immunity are closely associated. An example of this is that phagocytic cells are vital for the initiation of an adaptive immune response. Additionally, many signaling cytokines are released during an adaptive response affecting the innate immune system. During antigenic challenge both the innate and the adaptive immune systems work together to eliminate the pathogen [209].

Adaptive Immunity

The adaptive immune response reacts to specific foreign microorganisms or molecules. This response is very specific, and can differentiate between very subtle differences in protein structure. Accordingly, the immune system is capable of generating an enormous amount of different recognition molecules to be able to recognize unique structures on antigens. When the immune system has responded to a pathogen, immunological memory is generated, resulting in a faster and more efficient response should the pathogen be re-encountered. This attribute of the adaptive immunity can confer life-long immunity against many diseases and is the reason for vaccination. The most important cells in the adaptive immune response are the B-lymphocytes, the T-lymphocytes and the antigen-presenting cells [209].

Dendritic Cells

Dendritic Cells (DCs) exist throughout the body as immature DCs in tissue where they may locate antigen. They are a part of the innate immune system but has an important role in triggering the adaptive immune system. Immature DCs specialize in using a variety of membrane receptors to capture protein antigens and process them to form MHC-peptide complexes. Various stimuli, such as tissue damage, inflammatory mediators, microorganisms, and chemokines, cause DCs to mature and migrate out of the nonlymphoid tissues into the blood and lymph system to present the antigen to T-cells and Natural Killer (NK) cells. Both macrophages and DCs express proteins that recognize particular microbial molecules, known as Pattern Recognition Receptors (PRR). An important group of PRRs is called Toll-Like Receptors (TLRs) which recognize lipoproteins, polysaccharides, peptides and nucleic-acids and are important for initiating an appropriate immune response [210].

6.2 Paper IV

For this manuscript a time series of microarray experiments were performed, in an exploratory attempt, to investigate DC maturation when stimulated with the probiotic bacteria *L. acidophilus* NCFM. DCs were stimulated for 4, 10 or 18 hours allowing time-resolved investigation corresponding to early, intermediate and late stages of maturation. However the transformation from immature DCs to mature DCs corresponds to a large change in cell phenotype, which is reflected by the number of genes differentially expressed. Using an ANOVA >3300 genes were significantly differentially expressed (at FDR = 0) compared to non-stimulated immature DCs, which is an overwhelming number to consider for analysis. To provide a starting point for downstream analysis we clustered gene expression data using PAM clustering and used these for GO term enrichment. This revealed that genes involved in generating a viral response were among the highest up regulated genes and our collaborators therefore initiated laboratory experiments investigating the mechanism by which this could occur. The work is an example of transcriptomics data being used for data-driven hypothesis generation.

Lactobacillus acidophilus induces virus immune defense genes in murine dendritic cells by a TLR-2 dependent mechanism

Gudrun Weiss[1], Simon Rasmussen[2], Louise Hjerrild Zeuthen[1], Birgit Nøhr Nielsen[1], Hanne Jarmer[2], Lene Jespersen[3], and Hanne Frøkiær[1*],

[1] University of Copenhagen, Faculty of Life Sciences, Department of Basic Sciences and Environment, Thorvaldsensvej 40, DK-1871 Frederiksberg C, Denmark
[2] Center for Biological Sequence Analysis, Department of Systems Biology, Technical University of Denmark, DK-2800 Kongens Lyngby, Denmark
[3] University of Copenhagen, Faculty of Life Sciences, Department of Food Microbiology, Rolighedsvej 30, DK-1958 Frederiksberg C, Denmark
[*] Corresponding author: Hanne Frøkiær, Phone: +45 35 33 27 14, Fax: +45 35 33 23 98, e-mail: hafr@life.ku.dk

Abstract

Lactobacilli are probiotics that, among other health promoting effects, have been ascribed immunostimulating and virus preventive properties. Certain *Lactobacillus* spp. have been shown to possess strong IL-12 inducing properties. As IL-12 production depends on the up-regulation of type I interferons (IFN), we hypothesized that the strong IL-12 inducing capacity of *Lactobacillus acidophilus* NCFM in murine bone marrow derived dendritic cells (DC) is caused by an up-regulation of IFN-β, which subsequently induces IL-12 and the dsRNA binding toll like receptor (TLR)-3. The expression of the genes encoding IFN-β, TLR-3, IL-12 and IL-10 in DC upon stimulation with *L. acidophilus* NCFM was determined. *L. acidophilus* NCFM induced a much stronger expression of *Ifn-β*, *Il-12* and *Il-10* compared to the synthetic dsRNA ligand Poly I:C, whereas the levels of expressed *Tlr-3* were similar. Whole genome microarray gene expression analysis revealed that other genes related to the viral defense were significantly up-regulated and among the strongest induced genes in DC stimulated with *L. acidophilus* NCFM. The ability to induce IFN-β was also detected in another *L. acidophilus* strain

(X37), but was not a property of other probiotic strains tested i.e. *Bifidobacterium bifidum* Z9 and *Escherichia coli* Nissle 1917. The IFN-β expression was markedly reduced in TLR-2 -/- DC, dependent on endocytosis, and the major cause of the induction of *Il-12* and *Tlr-3* in *L. acidophilus* NCFM stimulated DC. Collectively, our results reveal that certain lactobacilli trigger the expression of viral defense genes in DC in a TLR-2 manner depended on IFN-β.

6.3 Introduction

Lactic acid bacteria (LAB) are inhabitants of the gastrointestinal (GI) tract, and some species are considered to have probiotic properties offering a number of benefits to health and well being [211–213]. Some probiotics have been shown to reduce the risk of virus infections such as the common cold and influenza [206–208]. So far, the mechanisms causing the reductions in respiratory tract infections and other symptoms are unknown. It is likely that these positive effects are due to the ability of probiotics to modulate immune stimulatory responses upon interaction with dendritic cells (DC). DC are central gatekeepers and regulators of the immune response interacting with mucosally encountered antigens, including the gut microbiota and viruses. The innate immune cell activation occurs predominantly through the interaction of TLRs and other pathogen recognition receptors on the surfaces of antigen presenting cells [214]. Exposure to microorganisms induces up-regulation of surface markers and the production of several cytokines that modulate the function of DC [215]. Probiotics exert differential stimulatory effects on DC *in vitro*, giving rise to varying production of different cytokines and accordingly different effector functions [216, 217]. Members of the lactobacillus and bifidobacterium genera are well-recognized for their probiotic properties, but also certain other bacteria, including some *Escherichia coli* strains, have shown to exert probiotic features.

Upon viral infection, type I interferons (IFN), cytokines with anti-viral and immune-regulatory functions, are produced. Toll-like receptors (TLRs) of DC have emerged as key transducers of type I IFN during viral infections [218]. TLR-3, a receptor localized in the endosomal compartment, recognizes dsRNA motifs of viruses and Poly I:C (a synthetic dsRNA), and induces the transcription of type I IFNs (IFN-β and IFN-β) [219, 220]. Type I IFNs exert their antiviral function by binding specifically to a unique receptor (IFNAR), thereby initiating a signaling cascade that controls the expression of hundreds of interferon-stimulated genes (ISGs) and other genes involved in an innate host response against viruses [221]. Type I IFNs, although best known for their antiviral properties, are potent regulators of cell growth and can modulate both innate and adaptive immune responses. Synthesis of type I IFNs was originally associated with viral infections, however, many pathogenic bacteria are equally able to induce the up-regulation of type I IFN, leading to modulation of the innate antibacterial response [222]. Several Gram-negative bacteria, such as *Salmonella enterica* Serovar Typhimurium, *Shigella flexneri* and *Escherichia* spp., stimulate type I IFN synthesis in phagocytosing cells [222]. Recently, pathogenic Gram-positive bacteria, such as group A and B *Streptococcus* spp. [223–225], *Listeria monocytogenes* [226, 227], and the spirochetal bacterium *Borrelia burgdorferi* [228] were likewise reported to induce the production of high quantities of type I IFN during infection. Manusco et al. [225] reported that the production of type I IFNs was critical for the clearance of infection by the host. In relation to intracellular bacteria, in particular TLR-3, TLR-7 and TLR-9 are involved in the type I IFN induction [229], whereas in connection with other bacteria, TLRs and other pathogen recognition receptors on the cell surface are of particular importance. However, no clear picture of which receptors are involved exists or which role these receptors play in the bacterial induced of IFN-β production. For *Streptococcus* spp. and *Listeria* spp., the intracellular located TLR-9 was essential for the induction of IFN-β in in vitro stimulated monocytes or DC [223, 227]. In *Borrelia burgdorferi*, the induction of IFN-β was independent of TLR-2 [228]. Only for the Gram negative *Pseudomonas auroginosa*, a role of TLR-2 has been suggested in the induction of a pro-inflammatory response in human monocytes [230]. It was demonstrated that a TLR-2 and mannose receptor synergistically were involved in the induction of the cytokines TNF-α, IL-6 and IL-1β. However, IFN-β was not included in the study. To our knowledge, neither lactic acid bacteria nor commensal bacteria have the capability to induce IFN-β in DC upon stimulation.

TLR-mediated IL-12p70 synthesis has been reported to be strongly reduced in the absence of type I IFN [231], demonstrating a critical role of type I IFN in controlling the production of the pro-inflammatory cytokine IL-12p70. We have previously reported that certain members of the *Lactobacillus* genus, including *L. acidophilus*, demonstrated remarkable IL-12 inducing properties [232]. On account of these observations, we hypothesized that *L. acidophilus*, despite its non-pathogenic phenotype and health promoting properties, is able to induce IFN-β production in DC and consequently help mature DC into anti-virus phenotype cells.

The aim of this study was to investigate whether *L. acidophilus* has the ability to induce the anti-viral defense gene expression in DC. We analyzed the gene expression profile of TLR-3 and IFN-β, key players involved in viral defense, in murine bone marrow derived DC *in vitro* stimulated with *L. acidophilus* NCFM. Genome wide microarray analysis confirmed our hypothesis showing a general, significant up-regulation of anti-viral defense genes. The IFN-β inducing property was likewise detected in another *L. acidophilus* strain, but not in a probiotic bifido or *E. coli* strain. This ability to induce IFN-β was dependent on TLR-2 recognition and required phagocytic activity in the DC. Our results reveal that, in contrast to Poly I:C, the expression of *Tlr-3* in *L. acidophilus* stimulated DC was dependent on the production of IFN-β. This study is the first to report that *L.*

acidophilus NCFM, a widely used probiotic bacterium, is able to induce the viral defense in murine bone marrow derived DC.

6.4 Materials and Methods

Bacterial strains, growth conditions and preparation of UV-killed bacteria

Lactobacillus acidophilus NCFM (Danisco, Copenhagen, Denmark), *L. acidophilus* X37 (Copenhagen University, Department of Food Microbiology, Faculty of Life Sciences, Denmark), *Bifidobacterium bifidum* Z9 (Copenhagen University, Department of Food Microbiology, Faculty of Life Sciences, Denmark), which are all considered to have probiotic properties, were grown anaerobically overnight at 37 °C in de Man Rogosa Sharp (MRS) broth (Merck, Darmstadt, Germany) and subcultured twice. Cells were harvested by centrifugation at 2,000 x g for 15 min, washed twice in phosphate-buffered saline (PBS, Bio Whittaker, East Rutherford, NJ, USA) and resuspended in 1/10 the growth volume of PBS. The bacteria were killed by 20 min exposure to UV light. *Escherichia coli* Nissle 1917 O6:K5:H1 (Statens Serum Institut, Copenhagen, Denmark), a Gram negative probiotic bacterium, was grown aerobically overnight at 37 °C in Luria-Bertani (LB) broth (Merck) and killed by a 45 min exposure to UV-light. The bacteria were stored at -80 °C, the concentration was determined as the content of dry matter per ml upon lyophilisation, and the dry weight was corrected for buffer salt content. Absence of viable cells was verified by plating the UV-exposed bacteria on MRS and LB agar.

Generation of murine dendritic cells

Bone marrow-derived dendritic cells (DC) were prepared as previously described [216]. Briefly, bone marrow from wild type (WT) or TLR-2-/- knock out C57BL/6 mice was flushed out from the femur and tibia and washed twice in sterile PBS. 3×10^5 bone marrow cells were seeded into Petri dishes in 10 ml RPMI 1640 (Sigma-Aldrich, St. Louis, MO) containing 10% (v/v) heat-inactivated fetal calf serum supplemented with penicillin (100 U/ml), streptomycin (100 μg/ml), glutamine (4 mM), 50 μm 2-mercaptoethanol (all purchased from Cambrex Bio Whittaker) and 15 ng/ml murine GM-CSF (harvested from a GM-CSF transfected Ag8.653 myeloma cell line). The cells were incubated for 8 days at 37 °C in 5 % CO_2 humidified atmosphere. On day 3, 10 ml of complete medium containing 15 ng/ml GM-CSF was added. On day 6, 10 ml were removed and replaced by fresh medium. Non-adherent immature DC were harvested on day 8.

Stimulation of murine dendritic cells with bacteria

Immature DC (2×10^6 cells/ml) were resuspended in fresh medium supplemented with 10 ng/ml GM-CSF, and 500 μl/well were seeded in 48-well tissue culture plates (Nunc, Roskilde, Denmark). The stimuli were suspended in medium and added (100 μl/well) in a final concentration of 10 μg/ml (*L. acidophilus* NCFM, *L. acidophilus* X37 and *E. coli* Nissle 1917) and 40 μg/ml (B. bifidum Z9). Optimal bacterial concentrations were determined in a previous study [232]. As a positive control, Poly I:C (InvivoGen, San Diego, CA, USA), a synthetic analog of dsRNA, was added in a final concentration of 10 μg/ml. The cell cultures were incubated at 37 °C in 5 % CO_2.

Immunstaining and flow cytometry

DC were harvested and resuspended in cold PBS containing 1 % (v/v) fetal bovine serum and 0.15 % (w/v) sodium azide (PBS-Az) containing anti-mouse FcγRII/III (3 μg/ml, BD Biosciences, San Jose, CA, USA) to block non-specific binding of antibody reagents. The cells were stained with phycoerythrin (PE)-conjugated anti-mouse MHCII, allophycocyanin (APC)-conjugated anti-mouse CD86, PE-conjugated anti-mouse CD11c (Southern Biotech, Birmingham, AL). Analysis was performed using BD FACSarray flow cytometer (BD Biosciences) based on counting 10.000 cells. The geometric mean fluorescence intensity (MFI) was determined representing the level of expression.

Effect of endocytic activity during stimulation

DC were pre-treated with cytochalasin D (0.5 μg/ml), chlorpromazine (25 μg/ml), methyl-β-cyclodextrin (1 mM) (Sigma-Aldrich), or medium alone for 1 h at 37 °C in 5 % CO_2 prior to addition of *L. acidophilus* NCFM (10 μg/ml) or Poly I:C (10 μg/ml) as previously described [233]. The cells were harvested after 3 h incubation at 37 °C in 5 % CO_2, and RNA was extracted.

IFN-β inhibition assay

Mouse IFN-β polyclonal antibody (R&D Systems, Minneapolis, USA) was added to the DC in different concentrations (0.01 μg/ml, 1 μg/ml, 10 μg/ml and 50 μg/ml) immediately after addition of *L. acidophilus* NCFM (10 μg/ml). The cells were harvested after 10 h of stimulation at 37 °C in 5 % CO2, and RNA was extracted.

RNA extraction

Murine DC were harvested at various stimulation time points, homogenised by QIAshredder (Qiagen, Ballerup, Denmark), and RNA was extracted using the

RNeasy Plus Mini Kit (Qiagen). RNA quality was verified by Bioanalyzer (Agilent, Santa Clara, USA), and the concentration was determined by Nanodrop (Thermo, Wilmington, USA).

Microarray analysis

Immature DC from three C57BL/6 mice were stimulated with L. acidophilus NCFM, and DC cells were harvested after 4 h, 10 h and 18 h. RNA was extracted, 1 μg RNA per stimulation condition was converted into cDNA, and biotin-labeled aRNA was synthesized using the MessageAmpTM II-Biotin Enhanced Kit (Ambion, Austin, TX, USA) according to the manufacturers instructions. The aRNA samples were hybridized to Gene Chip Mouse genome 430 2.0 Array (Affymetrix, Santa Clara, CA, USA), comprising 45.000 probe sets representing over 34.000 mouse genes. The arrays were stained, washed and scanned according to the manufacturer's instructions. The microarray data was analyzed using R and Bioconductor [73]. Raw probe intensities were normalized using *qspline* and expression index calculations were performed using *rma* [78, 83]. For statistical testing, ANOVA was performed using stimulation time as factor where all untreated samples were treated as one group. The false discovery rate (FDR) was estimated using a Monte Carlo approach, and statistical significance was set at an FDR of 0 yielding 4947 highly significant probe sets corresponding to 3319 unique genes annotated by Mouse Genome Informatics (MGI) [234].

Quantitative Real Time PCR analysis

DCs were harvested after 2 h, 4 h and 10 h of stimulation. RNA was extracted, and 1μg of total RNA was reverse transcribed by the TaqMan Reverse Transcription Reagent kit (Applied Biosystems, Foster City, USA) using random hexamer primers according to the manufacturer's instructions. The obtained cDNA was stored in aliquots at -80 °C. For the selection of primer and probe sequences, the regions coding for the genes investigated were retrieved from the GenBank EMBL databases. Following gene sequences were applied: TLR-3 (NM_126166), IFN-β (NM_010510), IL-12 p40 (NM_008352), IL-10 (NM_010548) and beta actin (NM_007393). Primers and probes were designed using the software Primer Express 3.0 (Applied Biosystems) and tested for specificity by the basic alignment search tool BLAST. HPLC purified forward and reverse primers were manufactured by DNA Technology (Aarhus, Denmark). The probes were labelled with the 5' reporter dye 6-carboxy-fluorescein (FAM) and the 3' quencher dye NFQ-MGB (Applied Biosystems). Sequences of primers and probes are listed in Table B.1. Primer and probe concentrations were optimized and to determine the efficiency of the amplifications dilution, standard curves were made for each set of primers and probe (data not shown). The amplifications were carried out in a total volume of 20 μl containing 1×TaqMan Universal PCR Master Mix (Applied Biosystems), forward and reverse primer (concentration 900 nM each), 200 nM TaqMan MGB probe, and purified target cDNA. The cycling parameters were initiated by 20 sec at 95 °C, followed by 40 cycles of 3 sec at 95 °C and 30 sec at 60 °C using the ABI Prism 7500 (Applied Biosystems). Amplification reactions were performed in triplicate, and DNA contamination controls were included. The amplifications were normalised to the expression of beta actin. Relative transcript levels were calculated applying the equation described by Pfaffl [235].

Cytokine quantification by ELISA

After 24 h of stimulation, culture supernatants were collected and stored at − 80 °C for later cytokine analysis. The production of murine IL-12(p70), IL-10, IL-6, TNF-α and IFN-β was analysed using commercially available enzyme-linked immunosorbent assay (ELISA) kits (R&D Systems, Minneapolis, USA).

Statistical analysis

Statistical calculations were performed using the software program GraphPad Prism 5 (San Diego, CA, USA). For each experiment, results were analysed by ANOVA with Bonferroni as post test, and p-values of < 0.05 were considered significant.

6.5 Results

Lactobacillus acidophilus NCFM induces IFN-β and TLR-3 up-regulation in murine dendritic cells

The expression of the genes encoding IFN-β and TLR-3 was determined after 2 h, 4 h and 10 h of stimulation with *L. acidophilus* NCFM and Poly I:C (Figure 6.1A). The two stimulators gave rise to highly distinct expression patterns. The strongest up-regulation of *Ifn-β* was detected after stimulation with *L. acidophilus* NCFM, as it was only slightly up-regulated after 2 h (38-fold) but reached a significant maximum after 4h (589-fold) that declined to 100-fold after 10 h. Contrary to *L. acidophilus* NCFM, the synthetic dsRNA analogue Poly I:C induced a strong expression of *Ifn-β* after 2h (180-fold). However, this induction decreased to 20-fold after 4 h and was raised to 220-fold again after 10 h. The highly potent *Ifn-β* inducing property of *L. acidophilus* NCFM was additionally confirmed by ELISA (Figure 6.1B), as the protein production of IFN-β was more than 5-times higher in DC stimulated with *L. acidophilus* NCFM compared to Poly I:C (1639 pg/ml versus 318 pg/ml, respectively). Both *L. acidophilus* NCFM and Poly I:C strongly sustained *Tlr-3* expression after 4 h and 10 h stimulation, indicating that L. acidophilus NCFM is capable of triggering up-regulation of TLR-3 to the same extent as the synthetic ligand Poly I:C. Stimulation with *L. acidophilus* NCFM resulted

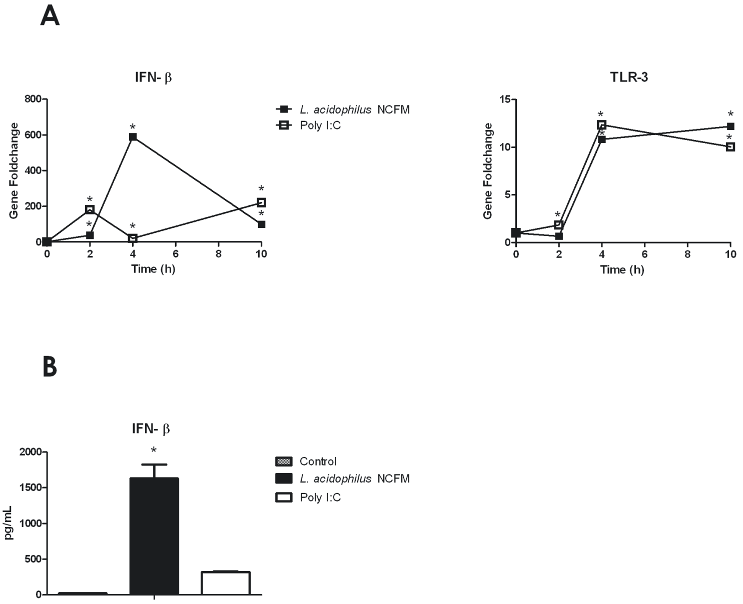


Figure 6.1 – *L. .acidophilus* NCFM induces gene expression of IFN-β and TLR-3. Bone marrow derived DC were stimulated with *L. acidophilus* NCFM and Poly I:C (10 μg/ml) for 2 h, 4 h and 10 h. RNA was extracted, and the induction of the genes encoding IFN-μ and TLR-3 was determined by RT-PCR analysis. The mRNA levels were normalised to the relative expression of beta-actin. The error bars depict the mean value +/- standard error of three individual measurements from one experiment. The data represent one of at least 7 independent experiments, * P<0.05.

in upregulation of the surface markers CD40 and CD86 (Figure 6.2), showing that the DC mature upon stimulation with *L. acidophilus* NCFM.

L. acidophilus NCFM was also observed to induce a much stronger expression of the pro-inflammatory cytokine IL-12 and the regulatory cytokine IL-10 (Figure 6.3A). IL-12 and IL-10, as well as the pro-inflammatory cytokines IL-6 and TNF-α, were detected in high concentrations measured by ELISA in the supernatants of DC stimulated with *L. acidophilus* NCFM, whereas the concentration of these cytokines upon stimulation with Poly I:C was just slightly increased as compared to non-stimulated cells (Figure 6.3B).

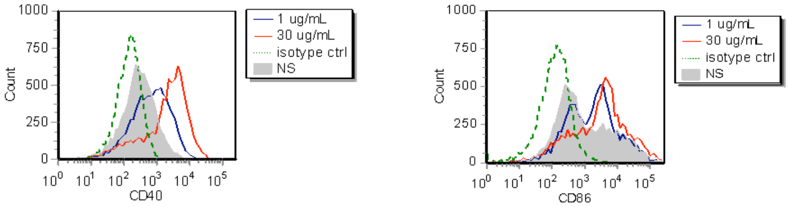


Figure 6.2 – Maturation profile of murine DC upon incubation with *L. acidophilus* NCFM. Flow cytometric analysis of expression of the surface markers CD40 and CD86 in DC after stimulation with *L. acidophilus* NCFM.

Identification of multiple virus-defence related genes by genome wide expression analysis in dendritic cells stimulated with *Lactobacillus acidophilus* NCFM

Genome wide microarray analysis was performed to further investigate the up-regulation of virus-related genes during stimulation of DC with *L. acidophilus* NCFM. To generate a comprehensive view of the expression profile, samples were harvested at different time points (4 h, 10 h and 18 h). Differential expression was assessed using ANOVA resulting in 3319 significant regulated genes at a false discovery rate (FDR) of 0 (p-value $<$ 1e-4). These findings point to a very strong response of DC upon stimulation, which is in good agreement with the phenotypical changes (e.g. production of cytokines and up-regulation of various surface markers) observed. The data generated were deposited in NCBI's Gene Expression Omnibus [236] and are accessible through GEO Series accession number GSE18460.

Focusing on genes that are virus-defense related, we used the Gene Ontology (GO) term GO:0009615 'Response to virus' to test whether the distribution of their expressions was different from the entire distribution. The Wilcoxon rank sum test (Mann-Whitney test) with a p-value of 3×10^{-11} revealed a strong, significant up-regulation of these genes (Figure 6.4). The induction of virus related genes was most prominent for the gene encoding Rsad2 (700-fold). Rsad2 (Radical S-adenosyl methionine domain containing 2), also known as viperin, encodes a cytoplasmic antiviral protein induced by interferons. This protein impairs virus budding by disrupting lipid rafts at the plasma membrane, a feature which is essential for the budding process of many viruses [237]. The genes encoding TGTP2 (interferon-induced T-cell specific GTPase), ISG15 (interferon-stimulated gene 15), interferon-regulatory factor (IRF-7) and toll-like receptor 3 (TLR-3), all involved in the viral immune defence and induced by IFN-β, were likewise among the highest significantly up-regulated.

In addition to the genes in the 'Response to virus' GO term, microarray data analysis revealed a significant induction of numerous genes related to viral in-

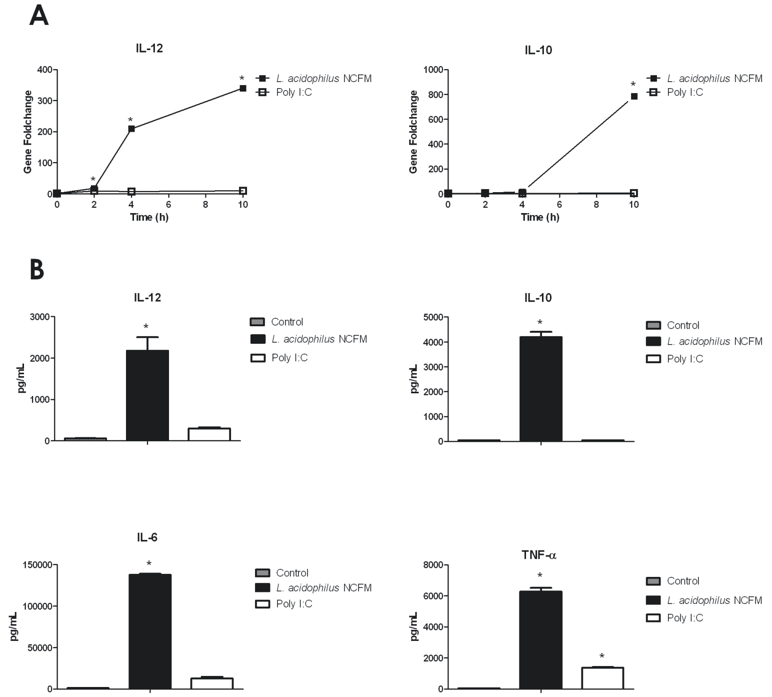


Figure 6.3 – *L. acidophilus* NCFM induces significant higher expression of the cytokines IL-12, IL-10, IL-6 and TNF-α compared to Poly I:C. **A**. Bone marrow derived DC were stimulated with L. acidophilus NCFM and Poly I:C (10ng/ml) for 2 h, 4 h and 10 h. RNA was extracted, and the induction of the genes coding for IL-12 and IL-10 was determined by RT-PCR analysis. The mRNA levels were normalised to the relative expression of beta-actin. **B**. Protein concentration of IL-12, IL-10, IL-6 and TNF-α was measured by ELISA in the supernatant 24 h after stimulation of DC with *L. acidophilus* NCFM and Poly I:C (10ng/ml). The error bars depict the mean value +/- standard error of three individual measurements from one experiment. The data represent one of at least 7 independent experiments. * P<0.05.

fection (Table B.2). The majority of these genes are classical Interferon Sensitive Genes (ISG) induced upon stimulation with IFN-β, e.g. members of the interferon-stimulated gene 56 family (ISG56), which are known to be strongly induced in response to virus infection, type I IFNs and dsRNA. In mouse, this family comprises three members (ISG56, ISG54, and ISG49) which associate with large protein complexes and block the translation pathway at different steps [238, 239]. Another strongly induced gene belonging to the classical family of ISGs codes for the well studied antiviral enzyme double-stranded RNA-dependent protein kinase (EIF2AK2, also termed PKR), which phosphorylates various substrates including the protein synthesis initiation factor eIF2α and acts by blocking the translation of viral RNA [240]. The 2'-5'oligoadenylate synthetases (OAS), a family of enzymes activated by dsRNA, were likewise strongly induced. These enzymes produce 2'-5'linked oligoadenylates activating the latent ribonuclease RNase L, which degrades viral mRNA [241]. The Myxovirus-resistance (Mx) proteins, IFN-inducible GTPases, were up-regulated in a similar manner. These proteins have a wide antiviral spectrum against different types of viruses and form complexes with dynamin, which disrupts intracellular transport or interferes with the activity of viral polymerases [242].

Induction of anti-viral mechanisms in dendritic cells is confined to certain probiotic strains

To elucidate whether the induction of the antiviral response is unique for *L. acidophilus* NCFM, universal for *L .acidophilus* strains, or a common property of probiotics, we further stimulated DC with another *L. acidophilus* (X37), a *B. bifidum strain* (Z9), and the Gram negative probiotic *E. coli* Nissle 1917. Gene expression analysis by RT-PCR revealed that *L. acidophilus* X37 was similarly able to trigger a virus response, as the genes encoding IFN-β and TLR-3 were significantly induced (Figure 6.5A). In contrast, neither *B. bifidum* Z9 nor *E. coli* Nissle 1917 gave rise to a strong up-regulation. Both strains resulted in a small peak of *Ifn-β* expression after 2 h of stimulation, followed by a rapid decrease to almost background level and a lower and less sustained up-regulation of *Tlr-3* transcription compared to the *L. acidophilus* strains. The rapid but low up-regulation of *Ifn-β* transcription upon stimulation with bifidobacteria and *E. coli* Nissle 1917 corresponded to the peak observed upon stimulation with Poly I:C. However, stimulation with Poly I:C showed a reemerging rise in the transcription after 10 h. The results obtained for *Ifn-β* were verified on a protein level by ELISA (Figure 6.5B). The highest production of IFN-β was measured upon stimulation of DC with *L. acidophilus* X37, which induced 18 times more IFN-β compared to *E. coli* Nissle 1917. *L. acidophilus* NCFM induced the production of IFN-β more than 14 times compared to *E. coli* Nissle 1917, whereas DC stimulated with *B. bifidum* Z9 did not produce detectable levels of IFN-β.

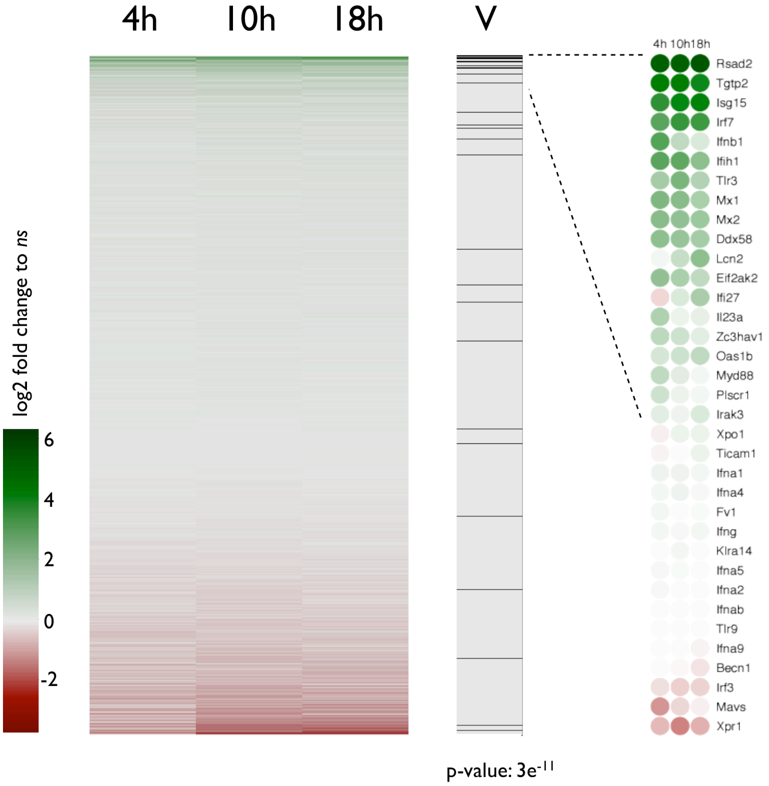


Figure 6.4 – *L. acidophilus* NCFM induces expression of multiple genes related to viral immune defence. Bone marrow derived DC from three mice were individually stimulated with *L. acidophilus* NCFM for 4 h, 10 h and 18 h, RNA was extracted, and microarray analysis was performed. Heatmap of $\log_2$ fold changes versus no stimulation for all probes on the array. Probe sets are sorted according to maximal $\log_2$ fold change at any time point. Green and red colours represent up and down regulations, respectively. In column "V" the position of genes in the gene ontology term GO:0009615 "Response to virus" is presented as black lines together with the significance of this distribution in a two-sided Wilcoxon Rank sum test (Mann-Whitney). Detailed expression of the genes is shown rightmost.

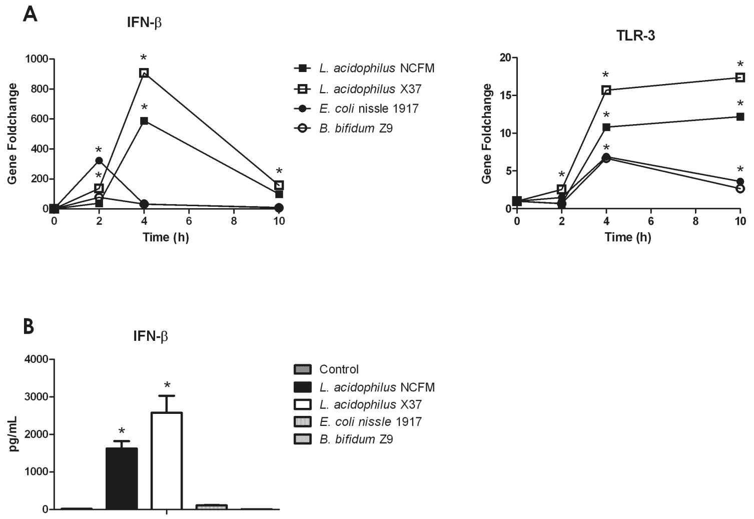


Figure 6.5 – *L. acidophilus* strains, but not *B. bifidum* and *E. coli*, induce IFN-β expression in DC. **A**. Bone marrow derived DC were stimulated with *L. acidophilus* NCFM, *L. acidophilus* X37, *E. coli* nissle 1917 (10 μg/ml) and *B. bifidum* Z9 (40 μg/ml) for 2 h, 4 h and 10 h. RNA was extracted, and the induction of the gene encoding IFN-β and TLR-3 was determined by RT-PCR analysis. The mRNA levels were normalised to the relative expression of beta-actin. **B**. Bone marrow derived DC were stimulated with *L. acidophilus* NCFM, *L. acidophilus* X37, *E. coli* nissle 1917 (10 μg/ml) and *B. bifidum* Z9 (40 μg/ml) for 24 h. The supernatant was harvested and protein concentrations were measured by ELISA. The error bars depict the mean value +/- standard error of three individual measurements from one experiment. The data represent one of at least three independent experiments,* P<0.05.

Induction of IFN-β and TLR-3 is dependent on TLR-2

The bacterial strains investigated in this study, capable of inducing strong *Ifn-β* and *Tlr-3* expression levels, were also the strains that gave rise to a high IL-12 production. As we previously have found that the IL-12 production is to a great extent dependent on TLR-2 stimulation [232], we hypothesized that TLR-2 might likewise be involved in the stimulation of DC with *L. acidophilus*, leading to the transcription of *Ifn-β* and *Tlr-3* (along with other virus related genes).

To investigate whether TLR-2 is required for the induction of IFN-β (or whether the induction of IFN-β requires TLR-2 recognition), we generated bone marrow derived DC from WT and TLR-2 -/- deficient mice. The expression of the gene encoding *IFN-β* was determined in DC upon stimulation with *L. acidophilus* NCFM, Poly I:C, *E. coli Nissle* 1917, and *B. bifidum* Z9 after 2 h, 4 h and 10 h. As depicted in Figure 6.6A, the lack of TLR-2 resulted in a dramatic decrease in the *Ifn-β* expression peak after 4 h induced by *L. acidophilus* NCFM. The *Ifn-β* expression profile was only moderately affected upon Poly I:C stimulation, with a slight increase in *Ifn-β* expression after 2 h and a decrease after 10 h. In contrast, when the DC were stimulated with either *B. bifidum* Z9 or *E. coli* Nissle 1917, the weak expression peaks observed after 2 h in wild type DC were markedly increased in TLR-2 -/- cells. Thus, whereas the absence of TLR-2 was central for the IFN-β production upon stimulation with *L. acidophilus* NCFM, TLR-2 seemingly exhibited the opposite role upon stimulation with *E. coli* Nissle 1917 and *B. bifidum* Z9, as the *Ifn-β* induction was higher in TLR-2-/- DC. Our gene expression results were confirmed by the presence of IFN-β in culture supernatants measured by ELISA after 24 h of stimulation Figure 6.6B).

Figure 6.7 illustrates the expression of *Tlr-3* upon stimulation of WT DC and TLR-2 -/- DC with *L. acidophilus* NCFM, Poly I:C, *E. coli* Nissle 1917, and *B. bifidum* Z9 for 2 h, 4 h and 10 h . In case of *L. acidophilus* NCFM and Poly I:C, the expression of *Tlr-3* was not affected by the absence of TLR-2 after 2 h and 4 h. However, after 10 h *Tlr-3* was significantly reduced in TLR-2 -/- DC compared to WT DC. In TLR-2-/- DC stimulated with *E. coli* Nissle 1917, the expression of *Tlr-3* was, in contrast to WT DC, only slightly lower (1-fold after 2 h, 4 h and 10 h). Upon incubation of DC with *B. bifidum* Z9, the up-regulation of *Tlr-3* was increased in TLR-2 -/- DC compared to WT DC at all time points.

To further investigate the dependency of IL-12 on IFN- β, and hence indirectly on TLR-2, we followed the Il-12 gene expression profile (Figure 6.8) and measured the protein production of IL-12 and three other cytokines (IL-10, IL-6 and TNF-α) in WT and TLR-2 -/- DC upon stimulation with *L. acidophilus* NCFM, Poly I:C, *B. bifidum* Z9, and *E. coli* Nissle 1917. The expression profiles of *Il-12* corresponded to the expression of *Ifn-β*, which were also reflected in the protein concentration of IL-12 and IFN-β measured in the supernatants by ELISA after 24 h of stimulation. In contrast, Il-10 induction was largely unaffected upon stimulation with *L. acidophilus* NCFM, Poly I:C, and *E.coli* Nissle 1917. However, in TLR-2-/- DC stimulated with *B. bifidum* Z9 the *Il-10* induction was significantly reduced (Figure 6.9). For all for stimulation regimes, the TNF-α protein concentration was slightly reduced in the supernatants of TLR-2 -/- DC, whereas IL-6 concentration was increased upon *E. coli* stimulation and decreased upon *L. acidophilus* stimulation (data not shown).

Taken together, these results show that TLR-2 plays an important role in the strong induction of IFN-β in DC upon stimulation with *L. acidophilus* NCFM. This observation is also reflected in the expression of the genes encoding IL-12 and TLR-3. In contrast, the same genes were largely unaffected when DC were stimulated with Poly I:C. In case of *E. coli* Nissle 1917 and *B. bifidum* Z9, TLR-2

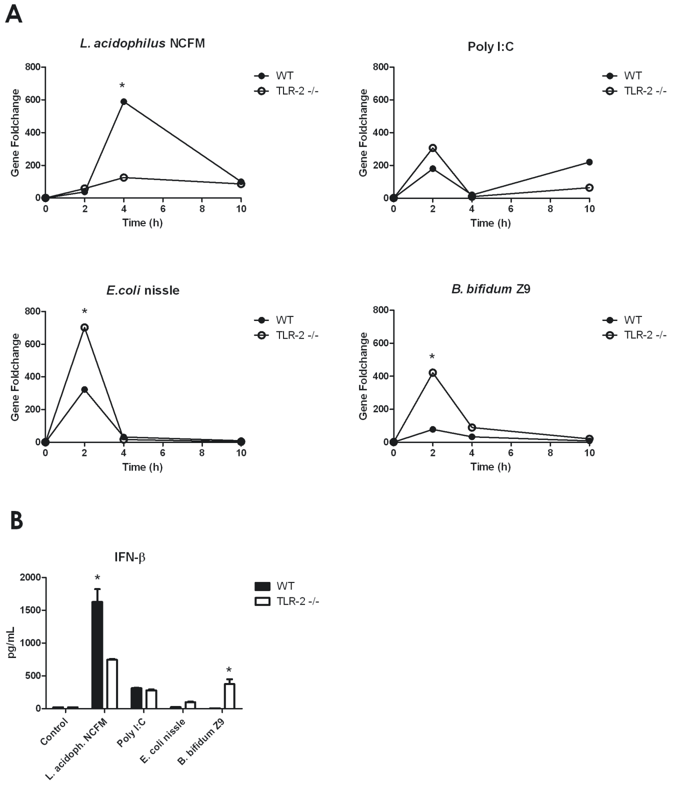


Figure 6.6 – IFN-β stimulating activity of L. acidophilus NCFM is dependent on TLR-2. **A**. Bone marrow derived DC from both WT and TLR-2 -/- were stimulated with *L. acidophilus* NCFM, Poly I:C, *E. coli* nissle 1917 (10 μg/ml), and B. bifidum Z9 (40 μg/ml) for 2 h, 4 h and 10 h. RNA was extracted, and the induction of the gene coding for IFN-β was determined by RT-PCR analysis. The mRNA levels were normalised to the relative expression of beta-actin. B. Bone marrow derived DC from both WT and TLR-2 -/- were stimulated with *L. acidophilus* NCFM, Poly I:C, *E. coli* nissle 1917 (10 μg/ml), and B. bifidum Z9 (40 μg/ml) for 24 h. The supernatant was harvested and protein concentrations of IFN-β were measured by ELISA. The error bars depict the mean value +/- standard error of three individual measurements from one experiment. The data represent one of at least two independent experiments. * $P<0.05$ values (WT versus TLR-2 -/-) are indicated.

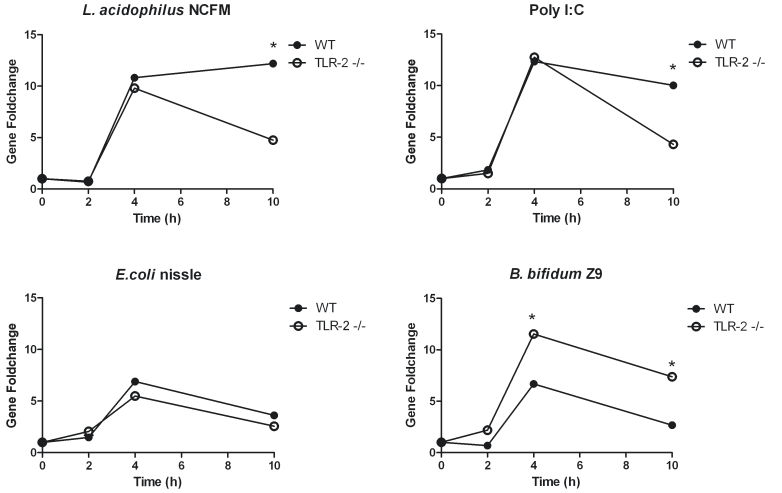


Figure 6.7 – The TLR-3 stimulating activity of *L. acidophilus* NCFM is dependent on TLR-2. Bone marrow derived DC from both WT and TLR-2 -/- were stimulated with *L. acidophilus* NCFM, Poly I:C, *E. coli* nissle 1917 (10 μg/ml), and *B. bifidum* Z9 (40 μg/ml) for 2 h, 4 h and 10 h. RNA was extracted, and the induction of the gene coding for TLR-3 was determined by RT-PCR analysis. The mRNA levels were normalised to the relative expression of beta-actin. The data represent one of at least two independent experiments. * P<0.05 values (WT versus TLR-2 -/-) are indicated.

seems to hold a suppressive role.

The clathrin-mediated endocytic pathway is required for the induction of IFN-β and TLR-3 upon stimulation with *L. acidophilus*

Poly I:C stimulated IFN-β induction in DC has recently been shown to depend on clathrin mediated endocytosis [233]. We have observed in previous studies that a prerequisite for a strong IL-12 response upon stimulation with *L. acidophilus* is that the bacterium is intact [232]. As a consequence, we speculated that the IFN-β and IL-12 inducing mechanism could involve phagocytosis or endocytosis triggering events. Accordingly, we used pharmacological inhibitors to investigate whether bacterial uptake of *L. acidophilus* NCFM is required for the induction of IFN-β, and, in turn, IL-12 and TLR-3. The effect of cytochalasin D (phagocytosis inhibitor), methyl-β-cyclodextrin (calveolae-mediated endocytosis inhibitor) and chlorpromazine (clathrin-mediated endocytosis inhibitor) on the stimulation

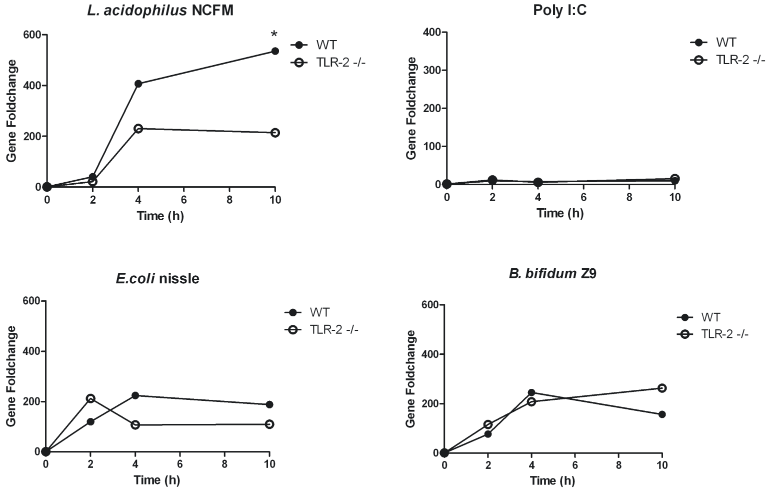


Figure 6.8 – Induction of Il-12 in DC stimulated with various probiotic bacteria and Poly I:C is dependent on TLR-2. Bone marrow derived DC from both WT and TLR-2 -/- were stimulated with *L. acidophilus* NCFM, Poly I:C, *E. coli* nissle 1917 (10 μg/ml), and *B. bifidum* (40 μg/ml) for 2 h, 4 h and 10 h. RNA was extracted, and the induction of the gene coding for IL-12 was determined by RT-PCR analysis. The mRNA levels were normalised to the relative expression of beta-actin. The data represent one of at least two independent experiments. * $P<0.05$ values (WT versus TLR-2 -/-) are indicated.

profile of DC after incubation with either *L. acidophilus* NCFM or Poly I:C was investigated Figure 6.10). Upon stimulation with *L. acidophilus* NCFM, the expression of the genes encoding IFN-β, TLR-3 and IL-12 was significantly inhibited when the DC were pre-treated with cytochalasin D and chlorpromazine. This inhibition was absent when the DC were pre-treated with methyl-β-cyclodextrin. The pharmacological inhibitors did not have the same impact on the expression of the gene encoding IL-10, as only a slight reduction was observed. We obtained similar results when DC were stimulated with Poly I:C. Pre-treatment with cytochalasin D and chlorpromazine of DC had a significant inhibitory effect on the expression of the genes coding for IFN-β and TLR-3, whereas pre-treatment with methyl-β-cyclodextrin did not have an impact. Our results indicate that the clathrin-mediated endocytic pathway participates in uptake of *L. acidophilus* NCFM as an important step in the stimulation of the transcription of IFN-β and TLR-3 and, in turn, IL-12.

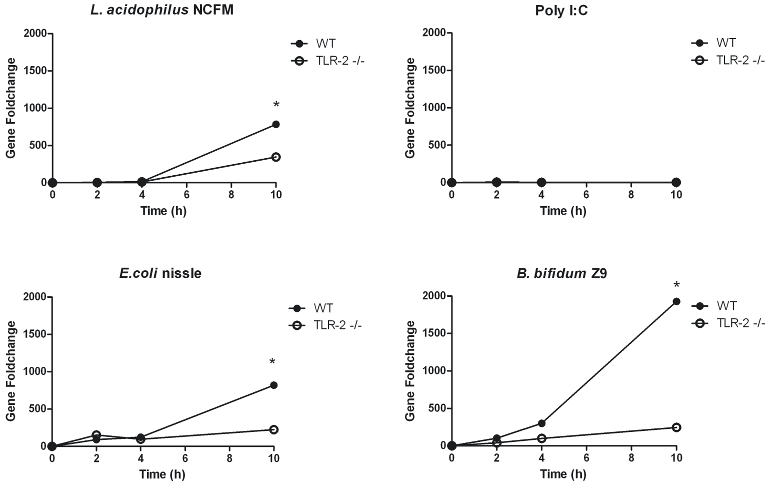


Figure 6.9 – Induction of *Il-10* in DC stimulated with various probiotic bacteria and Poly I:C is dependent on TLR-2. Bone marrow derived DC from both WT and TLR-2 -/- were stimulated with *L. acidophilus* NCFM, Poly I:C, *E. coli* nissle 1917 (10 μg/ml), and *B. bifidum* (40 μg/ml) for 2 h, 4 h and 10 h. RNA was extracted, and the induction of the gene coding for IL-10 was determined by RT-PCR analysis. The mRNA levels were normalised to the relative expression of beta-actin. The data represent one of at least two independent experiments. * P<0.05 values (WT versus TLR-2 -/-) are indicated.

Induction of IL-12 and TLR-3 by *L. acidophilus* NCFM is dependent on IFN-β

Despite the vast difference in the *Ifn-β* expression profiles of DC stimulated with *L. acidophilus NCFM* and Poly I:C, the *Tlr-3* expression profiles obtained were highly similar. We therefore speculated that the *Tlr-3* expression was caused by distinct mechanisms, i.e. that *L. acidophilus* NCFM *Tlr-3* expression was induced through the action of IFN-β and that the Poly I:C induced *Tlr-3* expression was primarily due to a direct binding of Poly I:C to TLR-3. To test the role of IFN-β in expressing *Tlr-3*, we added polyclonal anti-IFN-β antibodies to the cell cultures prior to stimulation with *L. acidophilus* NCFM, and measured the expression profiles of *Tlr-3* and *Il-12* (Figure 6.11). We observed a dose-dependent inhibitory effect of polyclonal IFN-β antibodies on the expression of *Tlr-3* and *Il-12.* This effect was strongest on *Tlr-3*, as the expression was almost completely inhibited when high antibody concentration was applied. By contrast, the same antibody concentrations reduced the expression of *Il-12* by approximately 50 %, which was

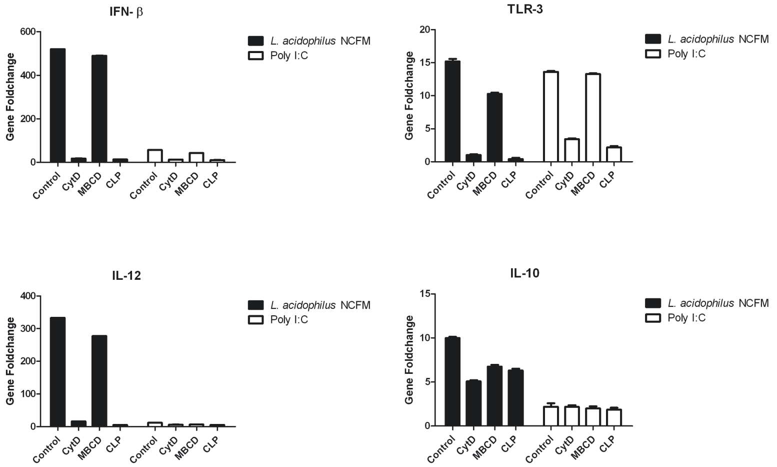


Figure 6.10 – A clathrin-dependent endocytic pathway participates in *L. acidophilus* NCFM-induced IFN-β production. Bone marrow derived DC were pretreated with cytochalasin D (CytD, 0.5 μg/ml), chlorpromazine (CLP, 25 μg/ml), methyl-β-cyclodextrin (MBCD, 1 mM) or medium alone for 1 h. Subsequently, the cells were stimulated with *L. acidophilus* NCFM (10 ng/ml) and Poly I:C, (10 ng/ml). RNA was extracted, and the induction of the gene coding for IFN-β, TLR-3, IL-12 and IL-10 was determined by RT-PCR analysis. The mRNA levels were normalised to the relative expression of beta-actin. The data represent one of at least four independent experiments.

confirmed by ELISA in the supernatant harvested after 24 h (data not shown).

6.6 Discussion

In this study we have shown that the probiotic bacterium *L. acidophilus* possesses the capability to induce a viral defense phenotype in bone marrow derived murine dendritic cells. Such properties have been demonstrated earlier by pathogenic bacteria [223, 243], but to our knowledge this has not been demonstrated for bacteria regarded as non-pathogenic or even beneficial for the immune system. The induction of viral defense mechanisms may explain the ability of some probiotic bacteria to stimulate the immune system as demonstrated in a number of clinical trials, including their ability to protect against viral infection.

The up-regulation of viral response genes seems to a great extent to be caused by a rapid and strong transient up-regulation of *Ifn-β*, which in turn stimulates transcription of a high number of other genes involved in viral defense. This was demonstrated in our microarray based kinetics study, as the gene encoding IFN-β appeared to belong to a minor group of genes with a rapid transient profile. By this approach we showed that virtually all genes related to viral defense were among the most up-regulated genes and that a high number of these genes is known to be directly regulated through the action of type I IFNs [220, 221].

The up-regulation of *Ifn-β* in DC was much stronger upon stimulation with *L. acidophilus* compared to cells stimulated with Poly I:C, *E. coli* Nissle 1917 and *B. bifidum* Z9. The up-regulation of *Ifn-β* correlated with an increased expression of *Tlr-3* as well as *Il-12*, thus supporting the connection between IFN-β and IL-12 as found by others [231]. In contrast, the up-regulation of *Tlr-3* was similar after stimulation of DC with *L. acidophilus* NCFM and Poly I:C. This indicates that up-regulation of *Tlr-3* does not exclusively depend on IFN-β, but may be affected by other mechanisms. Poly I:C has recently been shown to up-regulate IFN-β in a TLR-3 dependent manner in HEK293 cells and DC [244], hence there is evidence that ligand binding to TLR-3 induces IFN-β and, conversely, that IFN-β is able to induce *Tlr-3* expression. This may explain our observation that *Tlr-3* is up-regulated to the same extent upon Poly I:C stimulation as upon *L. acidophilus* NCFM stimulation despite the considerable difference in the produced IFN-β.

Not all probiotic bacteria were able to induce an up-regulation of *Ifn-β* and *Tlr-3* in DC, as demonstrated here with *B. bifidum* Z9, whereas another *Lactobacillus* strain, *L. acidophilus* X37, induced an IFN-β and TLR-3 response in a similar manner as *L. acidophilus* NCFM. Likewise, the Gram negative *E. coli* Nissle 1917, also considered probiotic, was not capable of inducing a significant expression of the genes coding for IFN-β or TLR-3. This is in accordance with the lack of capability of these bacteria to induce an extensive IL-12 production in the DC [232]. To which extent other probiotic bacteria are capable of inducing IFN-β and viral defense genes is currently under investigation.

As the IL-12 response was shown to be dependent on TLR-2 in a previous study [232], we investigated the IFN-β response in DC from TLR-2-/- mice and found that TLR-2, as for IL-12 expression, is mandatory for an induction of IFN-β upon *L. acidophilus* NCFM stimulation. In contrast, lack of TLR-2 resulted in a marked increase of IFN-β upon stimulation with *B. bifidum* Z9 and *E. coli* Nissle 1917. Hence, TLR-2 is not only playing a major role in the strong IL-12 and IFN-β response induced by *L. acidophilus*, it simultaneously plays a role in suppression of the same response upon stimulation of DC with other bacteria, such as *B. bifidum* Z9 and *E. coli* Nissle 1917 investigated in the present study. This dualism in TLR-2s role is not well described. Whereas the response to the TLR-2 ligand Pam3Cys generally is reported to be weak [232, 245], stimulation with whole bacteria through TLR-2 is, in a few cases, reported to give rise to a strong pro-inflammatory response [228, 232].

In human DC, IFN-β was found to be induced through a clathrin dependent en-

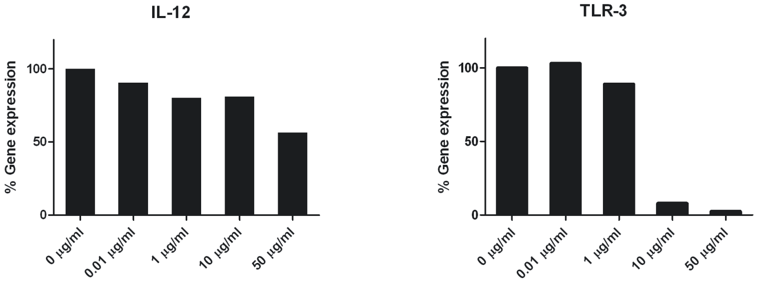


Figure 6.11 – Induction of *Il-12* and *Tlr-3* in *L. acidophilus* stimulated DC is dependent on IFN-β. Simultaneously with addition of *L. acidophilus* NCFM to DC, polyclonal IFN-β antibody was added in various concentrations (0 ng/ml, 10 ng/ml, 1000 ng/ml, 10000 ng/ml and 50000 ng/ml). Cells were harvested after 10 h, RNA was extracted and the gene expression of *Il-12* and *Tlr-3* was analysed by RT-PCR. The data represent one of at least two independent experiments.

docytotic mechanism [233]. We also found that the induction of *Ifn-β* was related to phagocytosis, possibly through a clathrin mediated mechanism, as addition of both the actin inhibitor cytochalasin and the clathrin inhibitor chlorpromazine abolished the induction of the gene coding for IFN-β. We have previously shown that UV-killed, but intact, bacteria, in particular *L. acidophilus*, induce a response corresponding to live bacteria which leads to much stronger IL-12 and TNF-β production compared to fragments or isolated cell walls of the bacteria. Taken together, this indicates that active uptake of the bacteria by endocytosis is important for the IFN-β induction. As TLR-2 was shown to be involved, our study suggests that TLR-2 plays an active role in the endocytosis dependent IFN-β upregulation, a phenomenon that to our knowledge has not been described before. However, whether there is a connection between the dependency of TLR-2 and endocytosis cannot be firmly established from the presented results. Maturation of DC is generally considered to abolish endocytosis in these cells, but a number of studies reports that some degree of maturation may take place in DC without abolishment of endocytosis. Weck et al. [245] found that in contrast to activation through TLR-3 and TLR-4, activation through TLR-2 with the synthetic TLR-2 agonist Pam3Cys did not abolish endocytosis. However, in contrast to Pam3Cys, ligands like peptidoglycan and lipopeptides present in close proximity in high number in an intact microorganism may stimulate several TLRs - or other receptors - simultaneously and hence work through a completely distinct mechanism. Such kind of receptor collaboration is well established for TLR-2 together with TLR-1 or TLR-6 [246]. Moreover, *Pseudomonas aeruginosa* was shown to induce a strong pro-inflammatory response by a TLR-2 and mannose receptor dependent mechanism [230]. The mannose receptor and TLR-2 form complexes on the cell surface during early phagocytosis and are found co-localized in endosomes for up to one hour after addition of the bacteria to the cells. We did not investigate the involvement of the mannose receptor in the present study but it is by all means conceivable that mannose receptor or another receptor collaborates with TLR-2 in the activation of a pro-inflammatory response.

Charrel-Dennis and colleagues [223] found that only live bacteria (streptococci) stimulated a strong induction of IFN-β, however, they compared with heat killed bacteria while we stimulated with UV-killed bacteria. This indicates that some protein-containing or heat vulnerable compound may be involved. Our previous studies showed that LTA, but not Pam2Cys or Pam3Cys, was involved in the TLR-2 dependent stimulation of IL-12 production in DC [232], but proteins or other molecules of importance for the intact bacterium may be involved, perhaps in collaboration with a TLR-2 ligand. Salazar and colleagues [228] stimulated monocytes with *Borrelia burgdorferi* which responded through both a TLR-2 dependent and TLR-2 independent pathway, but only the TLR-2 independent response lead to an induction of IFN-β. This is in contrast to our finding, as we observed a dramatic effect in TLR-2-/- DC. Thus, it is indicated that different microorganisms stimulate antigen presenting cells by distinct mechanisms giving rise to various cellular phenotypes.

Taken together, these results add to the picture of TLR-2 as an important receptor for both pro-inflammatory and regulatory responses in antigen presenting cells. Our study reveals that *L. acidophilus* is capable of stimulating a pro-inflammatory and antiviral response by a TLR-2 dependent mechanism, thus pointing towards TLR-2 as a receptor playing a central role in endocytosis dependent stimulation of a pro-inflammatory response in DC.

6.7 Acknowledgements

This study was supported by the Danish Strategic Research Council under the program “Associations between the diet, the composition of microbiota of the intestinal tract and human health”. The skilled technical support by Marianne K. Pedersen, Anni Mehlsen and Pia Friis was highly appreciated.

6.8 Perspectives

We are currently performing follow-up experiments aiming at identification of genes central for driving the viral response when DCs are stimulated with *L. acidophilus* NCFM. We have performed a microarray experiment investigating the effect of *Bifidobacterium bifido* Z9 on *L. acidophilus* NCFM induced activation of DCs. *B. bifido* is known to inhibit Lactobacilli response and a Singular Value Decomposition of the data in Figure 2.4 clearly reveals this effect.

Interestingly a Parametric Gene Set Enrichment of GO terms on the data, testing for differences between only stimulating using *L. acidophilus* NCFM versus simulation by both *L. acidophilus* and *B. bifido* Z9, the only significantly term is "Response to virus". By clustering the gene expression data and focusing on genes that are inhibited by adding *B. bifido* we hope to identify central genes in driving the viral response induced by *L. acidophilus.*

As of printing the thesis, we have submitted a manuscript to PLoS ONE showing that *B. bifido* actively inhibits *L. acidophilus* presumably via the JNK1/2 pathway. The gene encoding Jun dimerization protein 2 (JDP2) was only up-regulated in cells stimulated with *B. bifidum* and is a candidate gene for the inhibitory regulation.

Part IV

Conclusions

Chapter 7

Perspectives

Systems biology and integrative approaches are effective tools for increasing the power of data analysis in biological science. In this thesis I have given examples of transcriptomic data that have been integrated to achieve added biological value. The power of the approach is best observed in our study of Alzheimer's Disease where protein interaction data in combination with phenotypic data identifies a disease relevant gene. Additionally the protein-protein interaction network provides a starting point for hypothesis-generation due to the disease relevant protein interactions. If we had not taken this approach it is not likely that ADAM23 would have been identified as a gene associated with AD. In the case of the tiling array experiment of *B. subtilis* the aim was not to identify a particular gene or protein, but rather to provide a comprehensive map of the transcriptionally active regions and to identify novel features such as non-coding RNAs. In this case integration with e.g. protein-protein interaction data, which is sparse for *B. subtilis*, is not likely to improve the results. On the contrary we integrated known and predicted sigma factor binding sites and Rho-independent terminator sites to refine, benchmark and add value to the findings. Additionally from this we were able to expand an existing segmentation method hereby increasing performance in identifying known transcripts. The identification of 125 putative novel antisense transcripts suggests that natural antisense transcription is also a pervasive feature of prokaryotic transcriptomes.

With regard to the *S. cerevisiae* and the *L. acidophilus* work we took a more directed approach analyzing the data at the level of biological pathways in close contact with the underlying biology. The original authors of the *S. cerevisiae* data were not successful in using systems biology approaches, such as protein interaction data, to identify the underlying genetic network of the mutant strain. This illustrates that even though transcriptomics are very suitable for integrative approaches, the circumstances of an experiment determines to which extend success may be expected. In this particular case we were able to use added information from growth experiments showing that the uptake and/or utilization of leucine could describe the increased fitness of the mutant. In the profiling experiment using probiotic *L. acidophilus* NCFM to stimulate dendritic cells, we established that the bacterium can induce a Toll-like receptor 2 dependent viral response. Here pathway analysis revealed that virus response genes were among the strongest up regulated genes.

Taken together this show that systems biology and integrative approaches can improve the analysis and outcome of transcriptomics for both disease and industry focused applications. The outcome of such approaches are, naturally, dependent on the experimental conditions, available data and most importantly the objective of the study.

Where are we going next?

Systems biology and transcriptomics are likely to continue to be important aspects of biological research. However the form in which it is performed today is gradually changing, with Next Generation Sequencing probably replacing DNA microarray technology due to a higher throughput, accuracy and dynamic range. This, and a general increase in throughput of other research fields, allow data generation and experimental designs which are not confined to traditional studies of model organisms and univariate data analysis. Instead large multivariate experimental designs, such as cohort studies, can be used for studying complex biological diseases and systems. The increase in depth of the biological layers that can be analyzed and the increase in sheer data output will continue to emphasize the importance of bioinformatics and systems biology in biological research. Issues such as how to store, process and analyze these data are highly relevant and seem similar to the questions asked a decade ago, when increasing amounts of genomic and transcriptomic data began to appear. Certainly computational approaches and data integration will be important parts of the solution and the approaches developed for DNA microarray based methods are likely to be applicable to future experiments.

Regarding transcriptomics, especially eukaryotic organisms will benefit greatly from RNA-seq, as especially tiling DNA microarrays are not easily applied to large genomes [19, 34]. Detailed investigation on the complexity of the transcriptome, where mapping of non-coding RNAs such as natural antisense transcripts and identification of novel splice variants are achieved, could advance interactomics to a new level. Integration of RNA and protein interaction data may lead to new insights into complex biological processes and diseases that has not yet been recognized. Knowledge such as allele-specific expression and simultaneous investigation of coding SNPs will increase the resolution of the transcriptome. Additionally, NGS technology has already revolutionized metagenomics, where deep sequencing of genomic DNA is used to sample biodiversity in the human microbiome or environmental samples such as different water sources and soil [35, 247, 248]. This provides a huge potential for discovery of novel genes and investigation of the effect of different bacterial species on human diseases. From this sequencing of clinical pathogens directly from patients, may allow determination of individual treatment strategies. Likewise the opportunity to study not only the human reference genome, but thousands of human genomes, can facilitate the development of personalized medicine, where medication can be customized to the individual patient based on their genotypes [6].

Appendix A

Paper I

This section contains supplementary material for Paper I. First are supplementary Fig S1 and S3-9, followed by 6 pages of Fig S2. These are the pages of Fig S2 that are directly referenced to in the manuscript. Additionally the supplementary tables are not included. All of the supplementary tables are available at the publisher web-site as *Online Open* material: `http://www3.interscience.wiley.com/journal/122536032/suppinfo`.

Figure S1

Histogram of degree of overlap between TARs in the two medias. An overlap of 0 means that the TAR is unique to that media, whereas an overlap of 1 means that the TAR is identical to a tar in the other media.

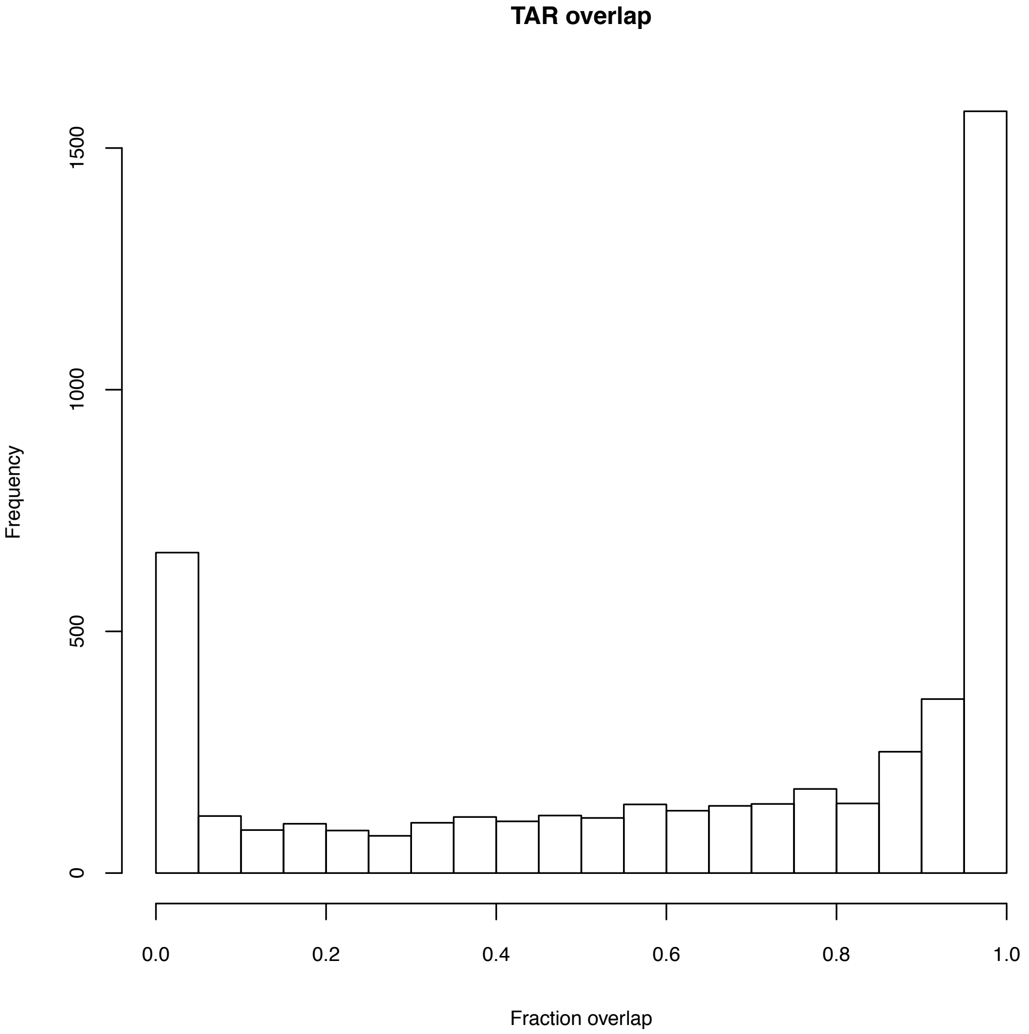


Figure S3

Benchmarking of TARs. (**a**) Shows the ROC-like curve of found genes. The True Positives (TP) are the genes as they are currently annotated and the False Positives (FP) are the same regions but on the opposite strand. (**b**) shows how many of the know Transcription Start Sites (TSSs) that are found as a function of the distance between this and the observed breakpoint. (**c**) Autocorrelation of expression signal as a function of spatial organization of genes (red) or TARs (blue).

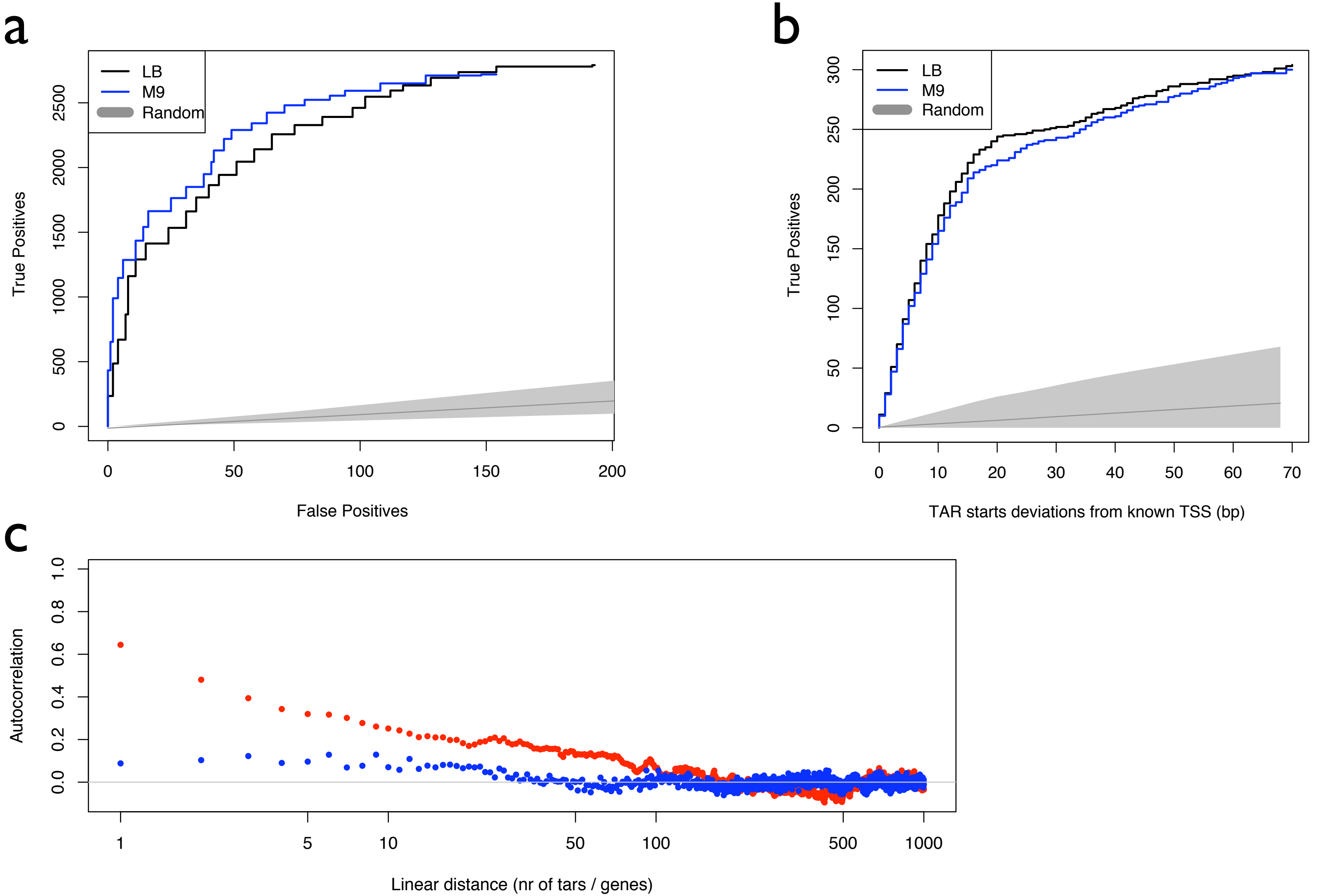
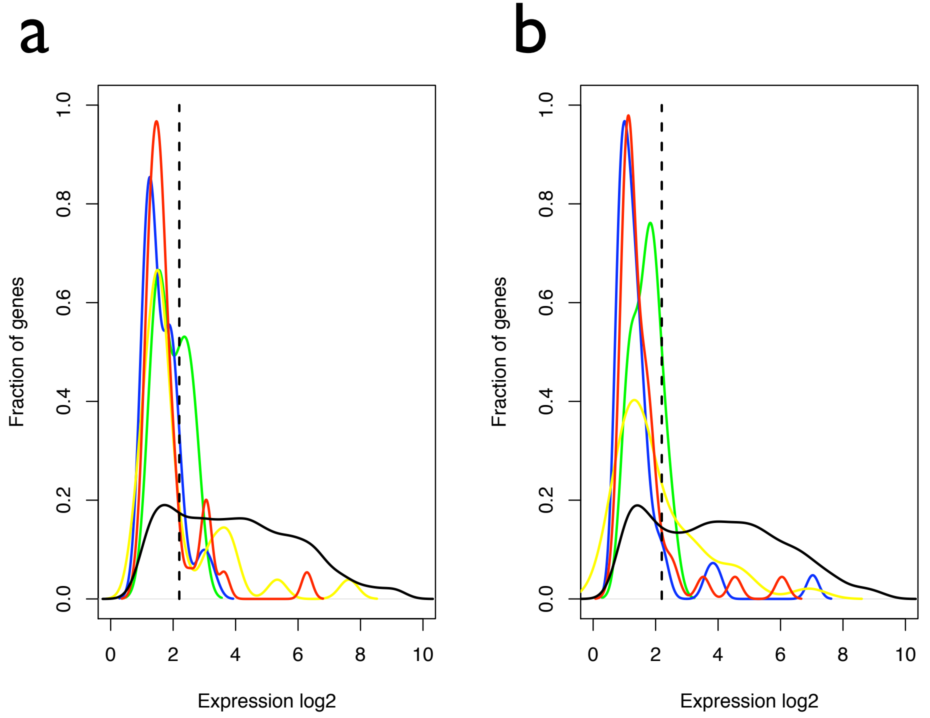


Figure S4

Density of gene expression of sporulation regulons, LB (a) and M9 (b). The regulons are color coded as: *sigF* (green), *sigE* (blue), *sigG* (yellow), *sigK* (red) and all genes (black). The vertical dotted line shows background signal. The composition of each regulon was taken from Steil *et al.*, 2005 and is shown in (c).

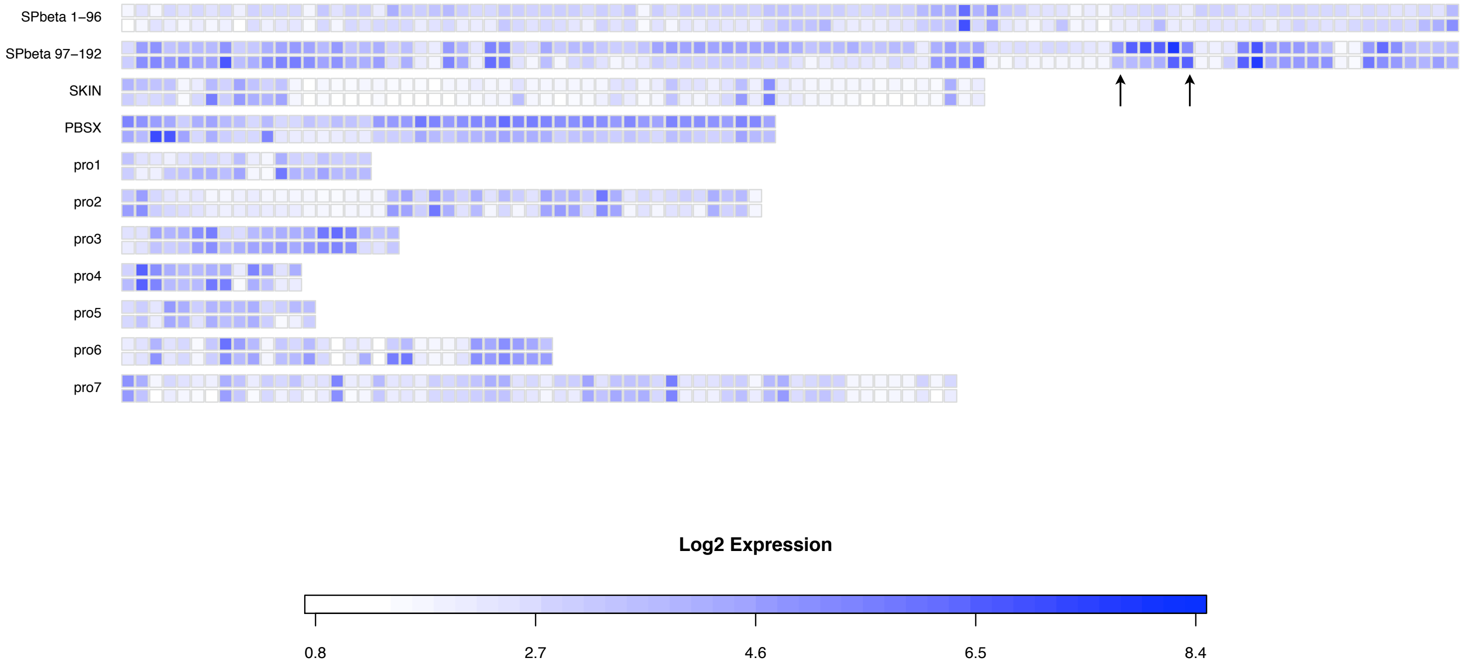


c

Regulon	Genes in regulon
sigF	*bofC, dacF, gpr, lonB, rsfA, spoIIQ, spoIIR, spoIVB, sspN, tlp, yphA, seaA*
sigE	*cotE, cotJA, cotJB, cotJC, cwlD, cwlJ, dacB, mmgA, mmgB, mmgC, mmgD, phoB, safA, spoIID, spoIIIAA, spoIIIAB, spoIIIAC, spoIIIAD, spoIIIAE, spoIIIAF, spoIIIAG, spoIIIAH, spoIIID, spoIVA, spoIVFA, spoIVFB, spoVD, spoVID, ysxE, spoVK, spoVR, usd, yaaH, ydhD, yjmC, exuR, exuT, uxuA, yjmD, uxuB, yknT, spoVM*
sigG	*coxA, csgA, gdh, gerAC, gerBA, gerBB, gerBC, gerD, sigG, sleB, splA, splB, spoIVB, spoVAA, spoVAB, spoVAC, spoVAD, spoVAEA, spoVAEB, spoVT, sspA, sspB, sspC, sspD, sspE, sspF, sspH, sspI, sspJ, sspK, sspL, sspN, tlp, ybaK, yhcN*
sigK	*cgeA, cgeB, cgeC, cgeE, cotA, cotB, cotD, cotF, cotG, cotH, cotS, cotV, cotW, cotX, cotY, cotZ, gerE, gerPA, gerPB, gerPC, gerPE, gerPF, spoIIIC, spoVFA, spoVFB, spsA, spsC, spsD, spsE, spsF, spsG, spsI, spsJ, spsK, sspG, tgl, yabG, ydgB, cotP, ydgA, ykvP, cotR, yvdP*

Figure S5

Gene expression of prophage and prophage-like elements. Upper row show phage element expression in LB media and lower row gene expression in M9 media. Squares represents genes in the phage elements. The color scale range from white (low expression) to blue (high expression). Arrows indicate the sublancin area (*bdbB* to *sunI*) in the *SPβ* prophage.



SPβ:2152-2286 kb, *PBSX*: 1316-1347 kb, *SKIN*: 2653-2700 kb, *pro1*: 202-220 kb, *pro2*:529-570 kb, *pro3*: 652-665 kb, *pro4*:1262-1270 kb, *pro5*:1879-1891 kb, *pro6*: 2046-2073 kb, *pro7*: 2707-2756 kb.

Figure S6

Density plots of gene expression of prophage and prophage-like elements. Expression for LB and M9 is shown blue and red, respectively. Prophage coordinates are shown in Fig. S5.

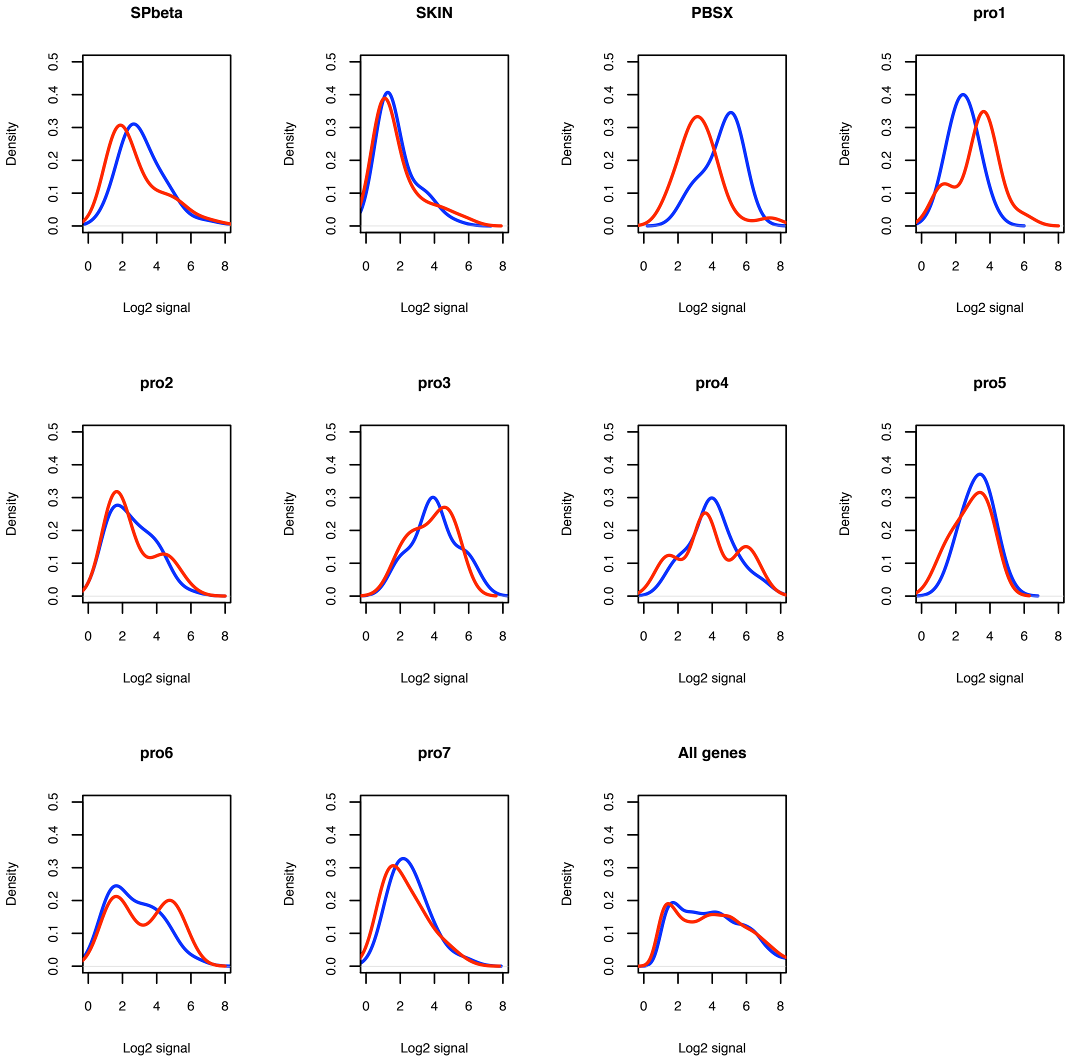


Figure S7

Conservation plot of non coding RNAs (*ncr*) identified. Nucleotide sequence of identified *ncr*s was compared against all available Firmicute sequences (genome and plasmid) and the maximum hit is plotted. The color scale ranges from black (0% identity) to white (100% identity).

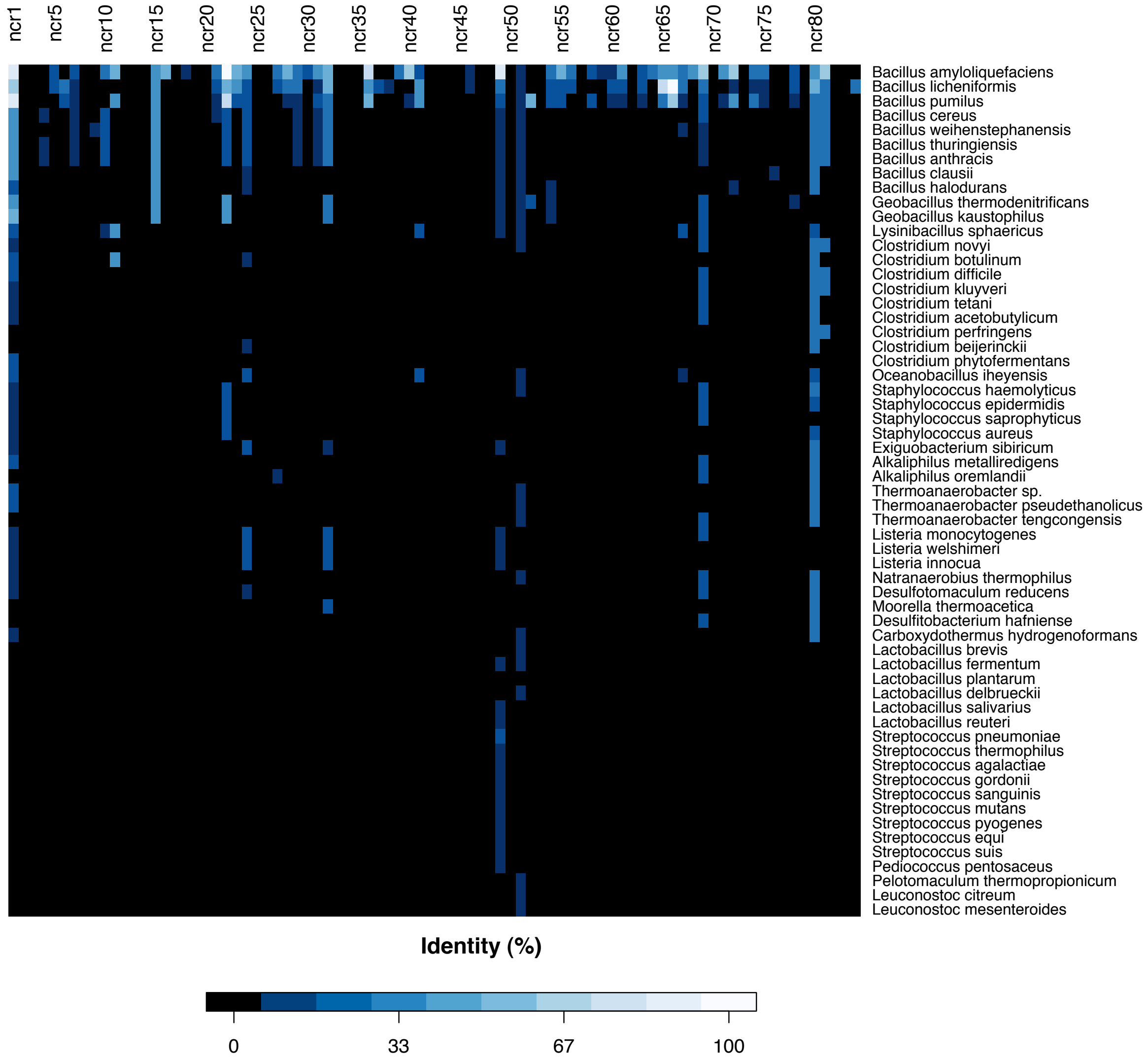


Figure S8

Sense overlap and antisense/ratio of antisense transcripts. (a) Fraction of sense genes (cds) overlapped by antisense transcript. Two distributions (< 0.7 and > 0.7) of overlaps can be seen. (b) For the 127 antisense transcripts, log2 sense (LB/M9) and log2 antisense (LB/M9) ratio is plotted. If antisense transcript is the primary regulator of the sense area at the conditions tested, the antisense and sense ratios would expect to be anti-correlated (upper left and lower right quarters). The number plotted corresponds to the shd nomenclature (e.g. 4 = *shd4*).

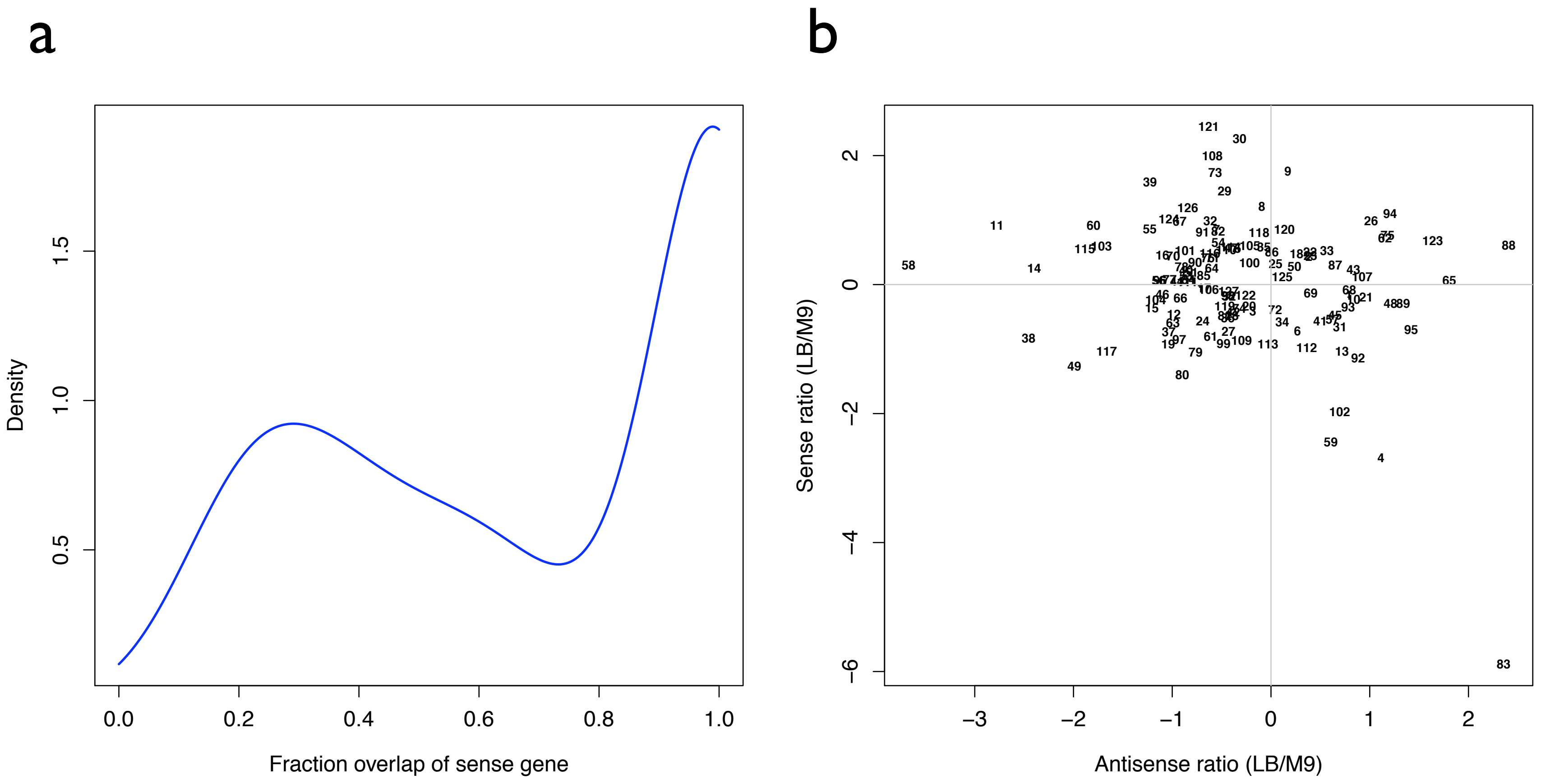


Figure S9

Folding of conserved 3'UTR RNA elements. All elements were folded using RNAfold v1.6 and bases are coloured according to base-pair probabilities, from 0 to 1 (purple, blue, green, yellow to red).

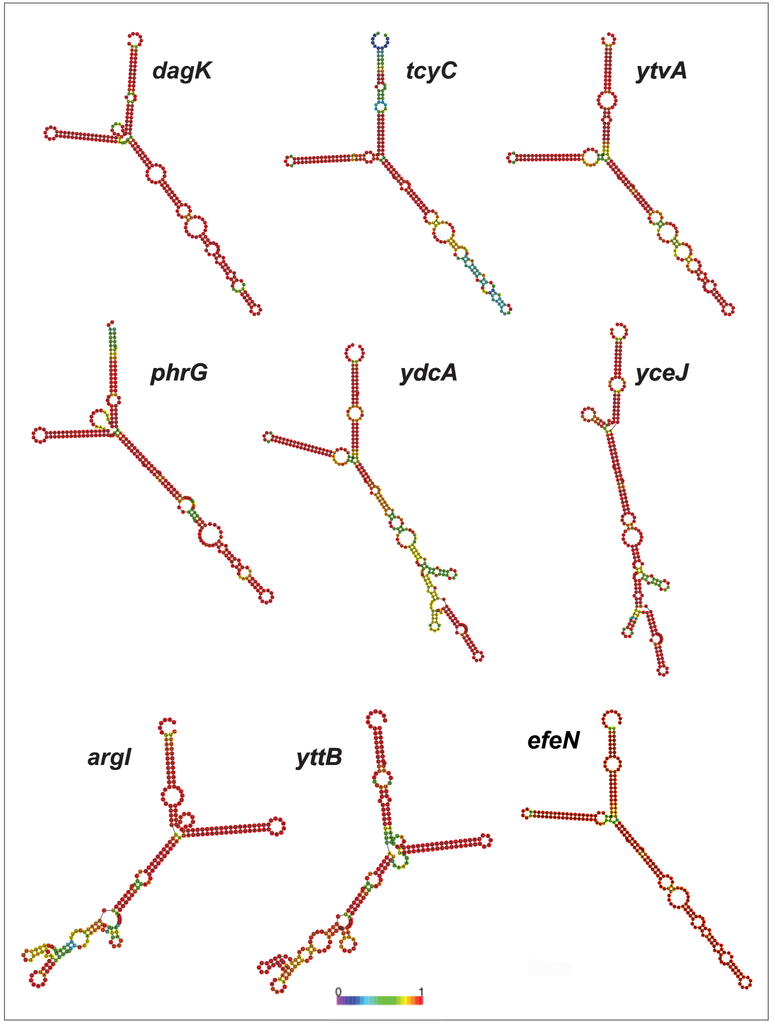
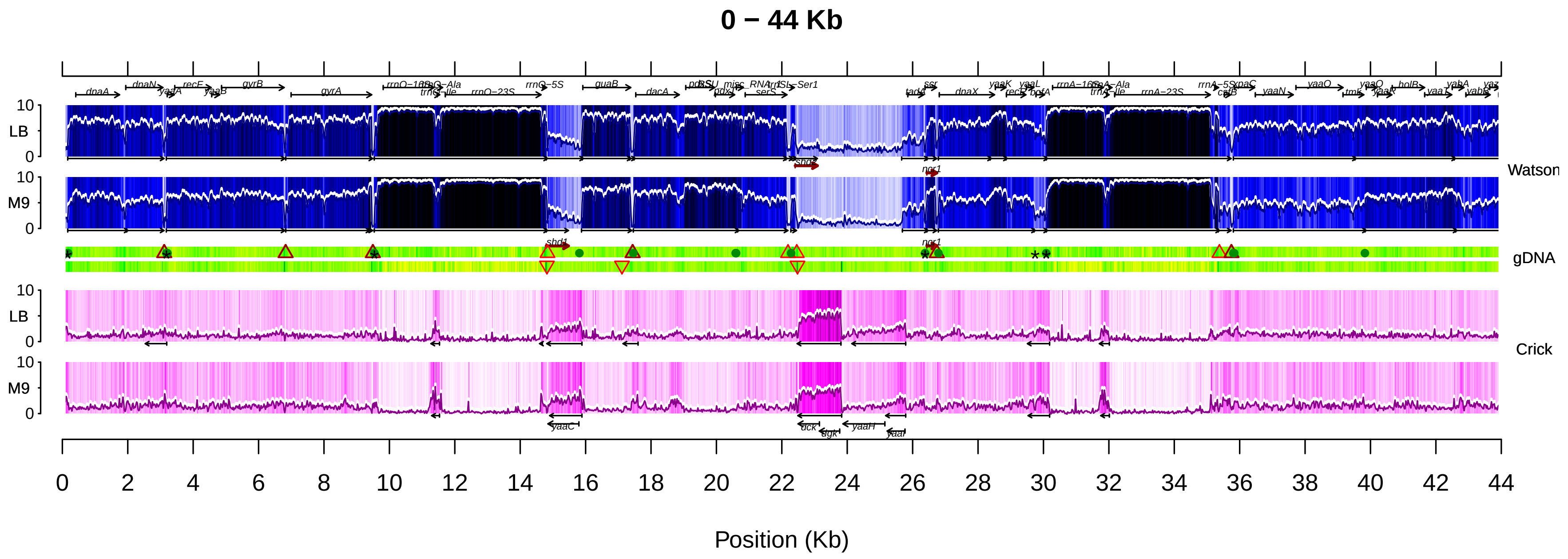
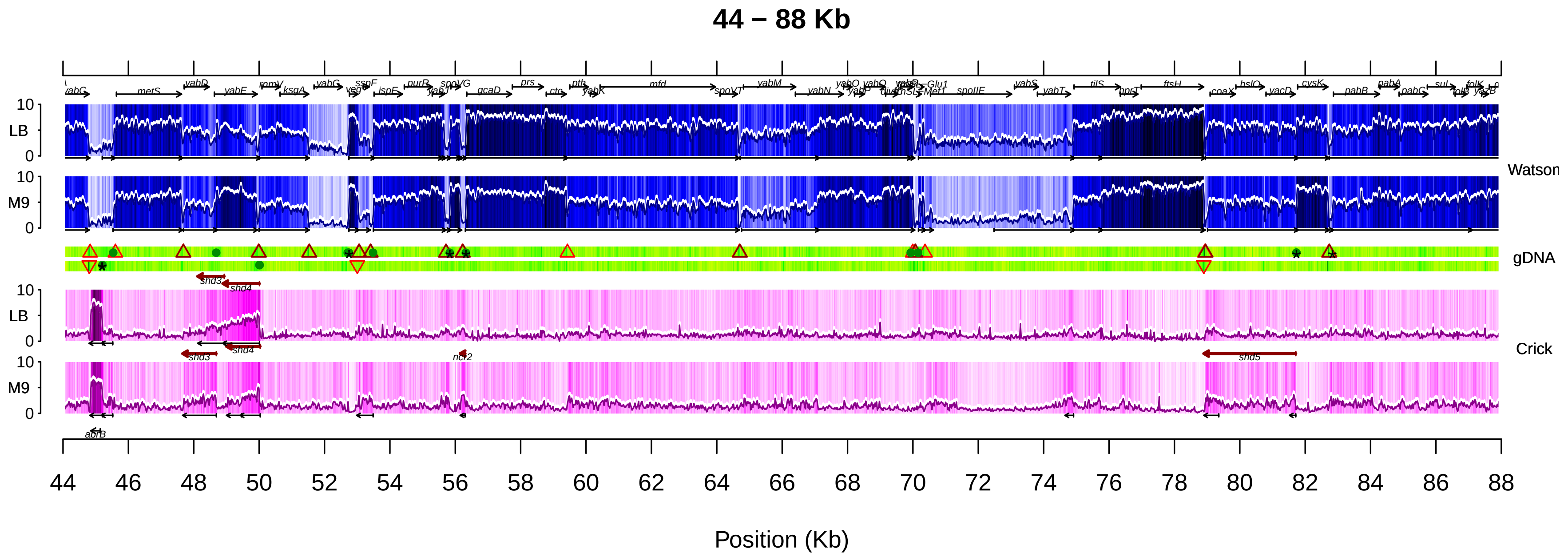


Figure S2

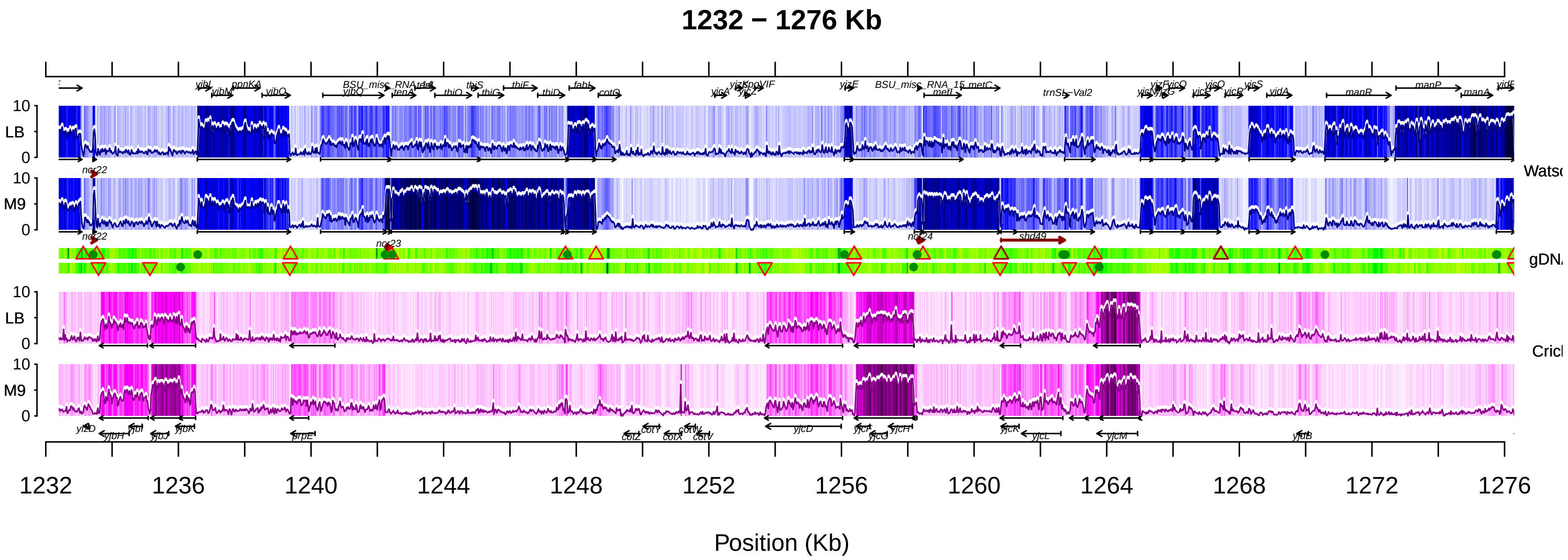
The following pages show the level of expression along the genome of Bacillus subtilis when growing on rich (LB) and minimal (M9) medium. The signal from genomic DNA hybridized to the same tiling chip (Rasmussen *et al.*, 2009) is also shown. Each page contain two similar plots each presenting ~44 Kb of the genome. The two blue bands show the expression in LB and M9 respectively on the positive strand, the two magenta bands show the minus strand. The darker the color the higher the expression in the given position. On top of the bands the expression level is plotted as a scaled version (0-10) of the log2 to the fold change (signal/background signal). In the middle of each plot the gDNA signal is shown for both strands, here the color scale is yellow-green-black, where black is low and yellow is high. Along the genome, in all 6 bands in the same positions, white regions indicate regions were the recent re-sequencing of the genome (AL009126.3) have shown either inserts or regions with very low similarity to the last version (see Rasmussen *et al.*, 2009). Annotated genes are shown as black arrows in the top and the bottom of each plot (annotation from AL009126.3), indicating genes on the positive and negative strand respectively. New genes are shown as red arrows below (+ strand) and above (- strand) each band. Both predicted and experimentally verified terminators and sigma-factor binding sites are shown on top of the gDNA bands. Dark red triangles: Experimentally verified terminators, red triangles: Predicted terminators, stars: Experimentally verified sigma-factor binding sites, green dots: Predicted sigma-factor binding sites.



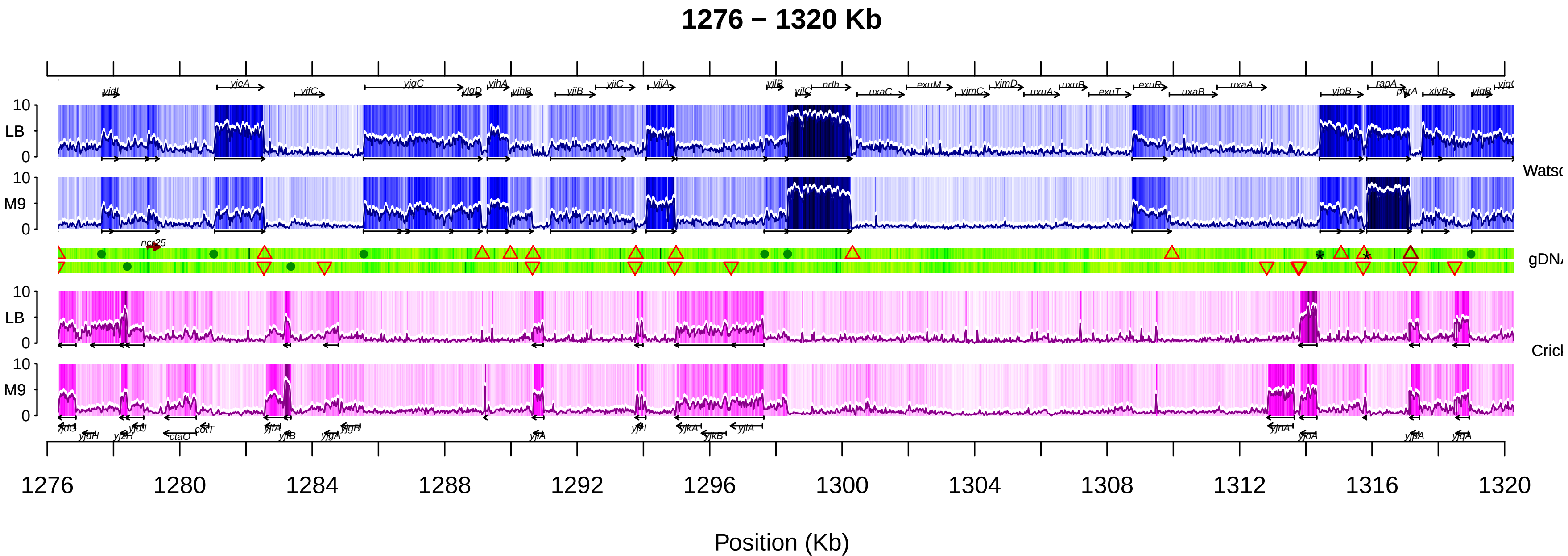
0 − 44 Kb



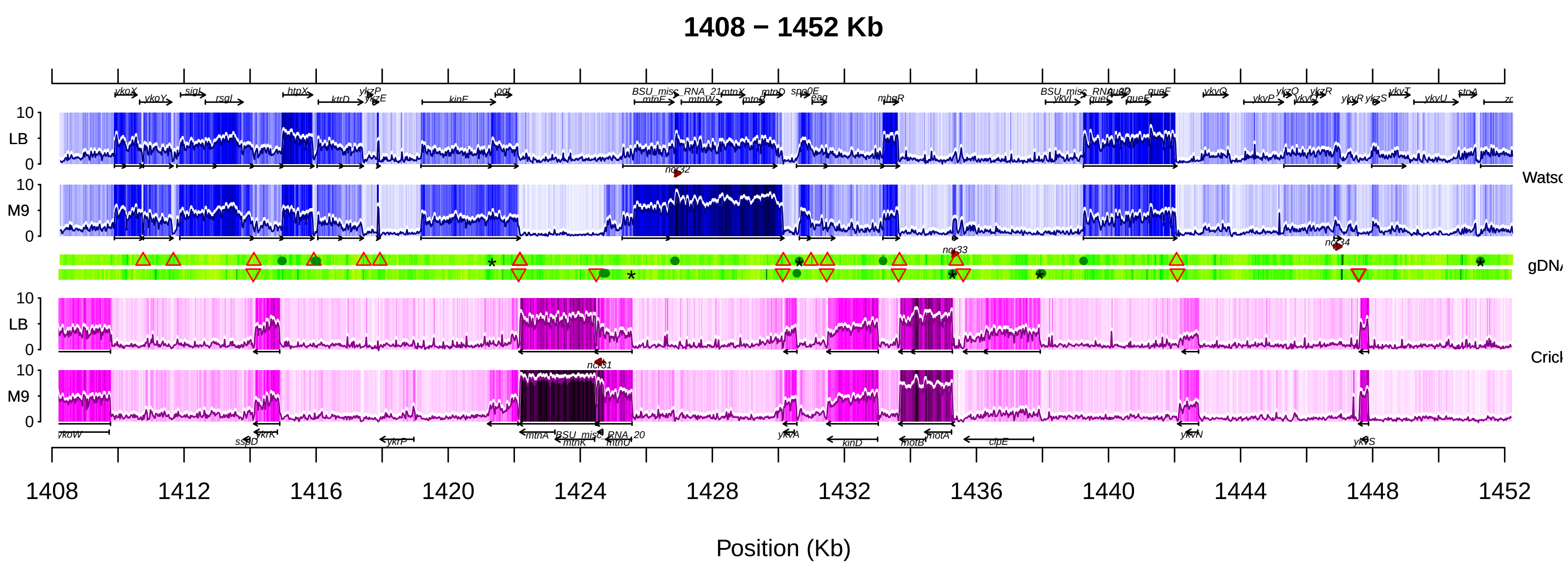
44 − 88 Kb

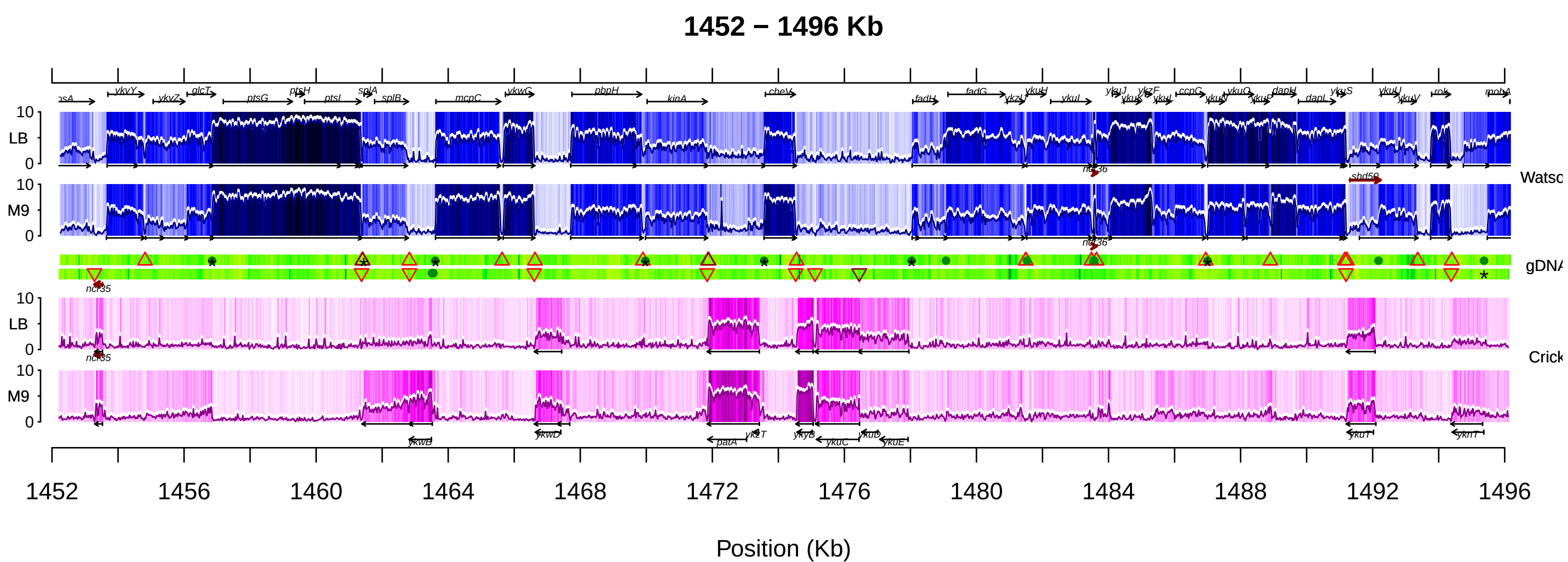


1232 − 1276 Kb
LB
M9
10
0
Watsc
gDNA
Cricl
1232
1236
1240
1244
1248
1252
1256
1260
1264
1268
1272
1276
Position (Kb)
ncr22
ncr23
ncr24
shd49
yjbL
ppnKA
yjbM
yjbO
BSU_misc_RNA_14
tenA
thiO
thiS
thiG
thiF
thiD
fabL
cotQ
yjcA
yjzKppVIF
yjzE
metI
metC
trnSL-Val2
yjzF
yjcO
yjcP
yjcS
yjdA
manR
manP
manA
yjzD
yjbH
yjbI
yjbJ
yjbK
prpE
cotZ
cotY
cotX
cotW
cotV
yjcD
yjcF
yjcG
yjcH
yjcK
yjcL
yjcM
yjdB

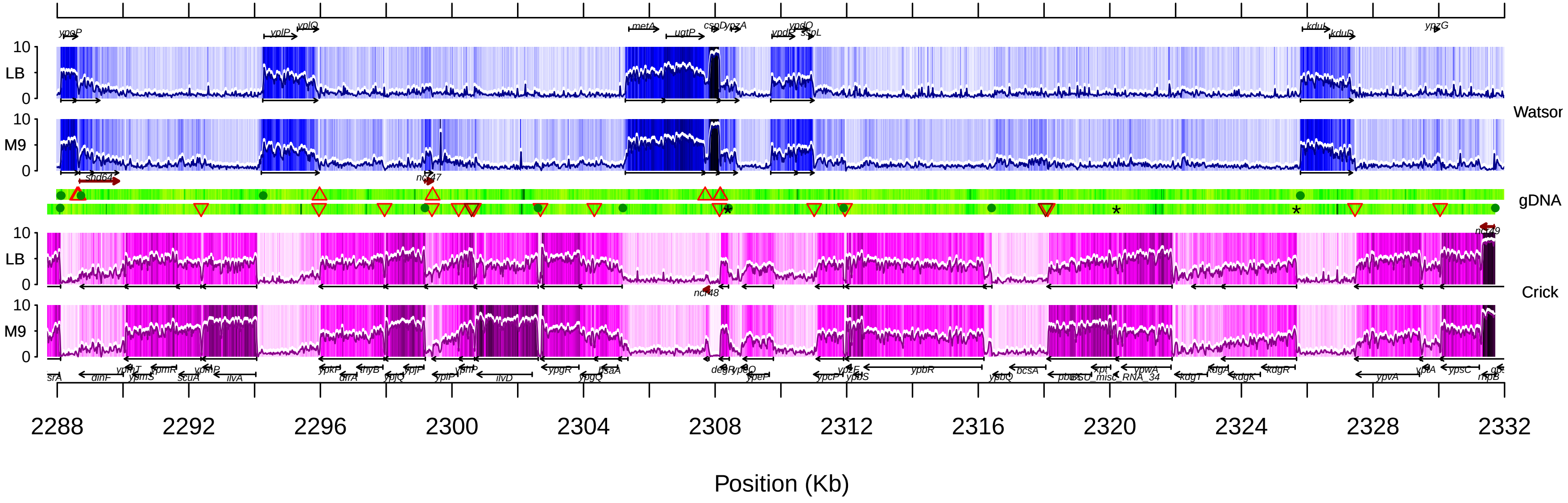


1276 − 1320 Kb
LB
M9
10
0
Watsc
gDNA
Cricl
1276
1280
1284
1288
1292
1296
1300
1304
1308
1312
1316
1320
Position (Kb)
ncr25
yjdI
yjeA
yjfC
yjgC
yjgD
yjhA
yjhB
yjiB
yjiC
yjjA
yjlB
yjlC
ndh
uxaC
exuM
yjmC
yjmD
uxuA
uxuB
exuT
exuR
uxaB
uxaA
yjoB
rapA
phrA
xlyB
yjqB
yjdG
yjdH
yjzH
yjdJ
ctaO
cotT
yjfA
yjfB
yjgA
yjgB
yjiA
yjzI
yjkA
yjkB
yjlA
yjnA
yjoA
yjpA
yjqA

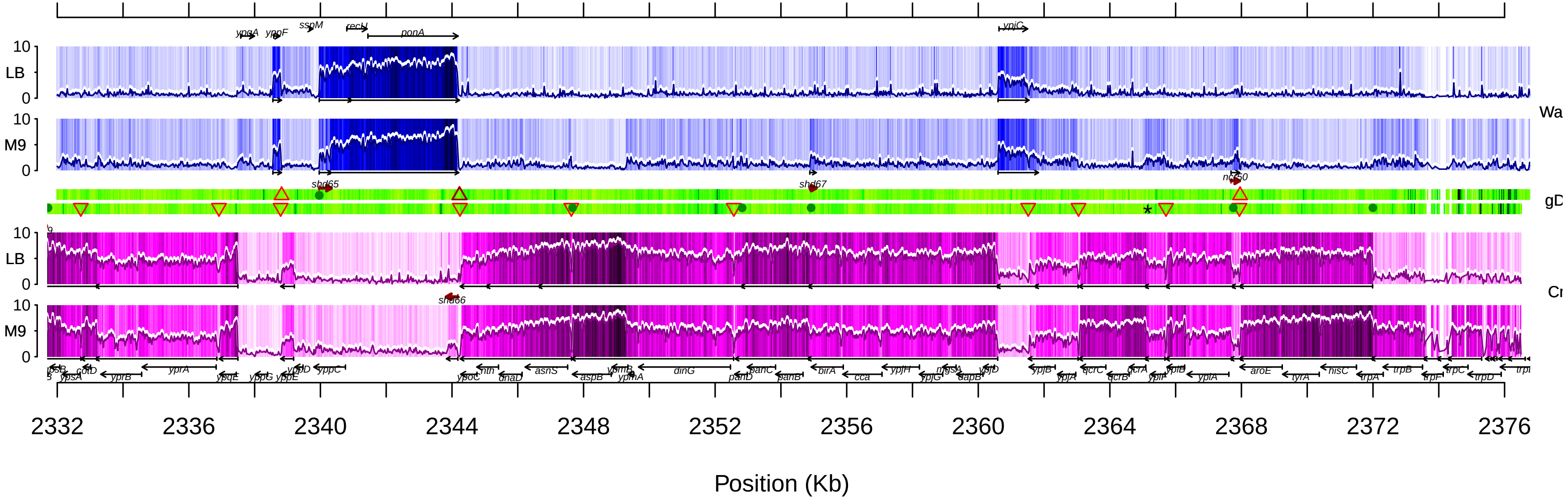

1408 − 1452 Kb
ykoX ykoY sigI rsgI htpX ktrD ykzP ykzE kinE ogt BSU_misc_RNA_21 mtnX mtnE mtnW mtnB mtnD spo0E eag mhqR BSU_misc_RNA ykvI queC queD queE queF ykvQ ykvP ykzQ ykvQ ykzR ykvR ykzS ykvT ykvU stoA
ncr32 ncr33 ncr34 ncr31
LB
M9
10
0
Watsc
gDNA
Crick
ykoW sspD ykrK ykrP mtnA BSU_misc_RNA_20 mtnK mtnU ykvA kinD motB motA clpE ykvN ykvS
1408 1412 1416 1420 1424 1428 1432 1436 1440 1444 1448 1452
Position (Kb)


1452 − 1496 Kb
ykvY ykvZ glcT ptsG ptsH ptsI splA splB mcpC ykwC pbpH kinA cheV fadH fadG ykzU ykuH ykuI ykuJ ykzF ykuK ykuL ccpC ykuO ykuN ykuP dapH dapL ykuS ykuU ykuV rok mobA
ncr36 shd50 ncr35
LB
M9
10
0
Watsc
gDNA
Crick
ykwB ykwD patA ykzT ykyB ykuC ykuD ykuE ykuT ykrT
1452 1456 1460 1464 1468 1472 1476 1480 1484 1488 1492 1496
Position (Kb)

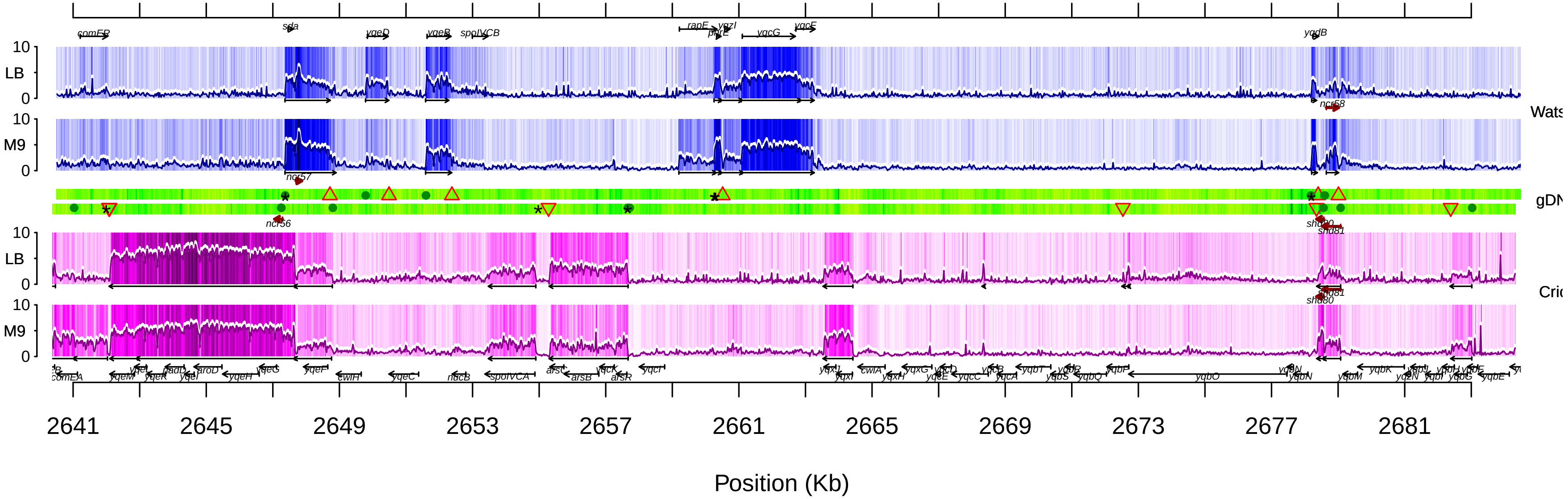
2288 − 2332 Kb



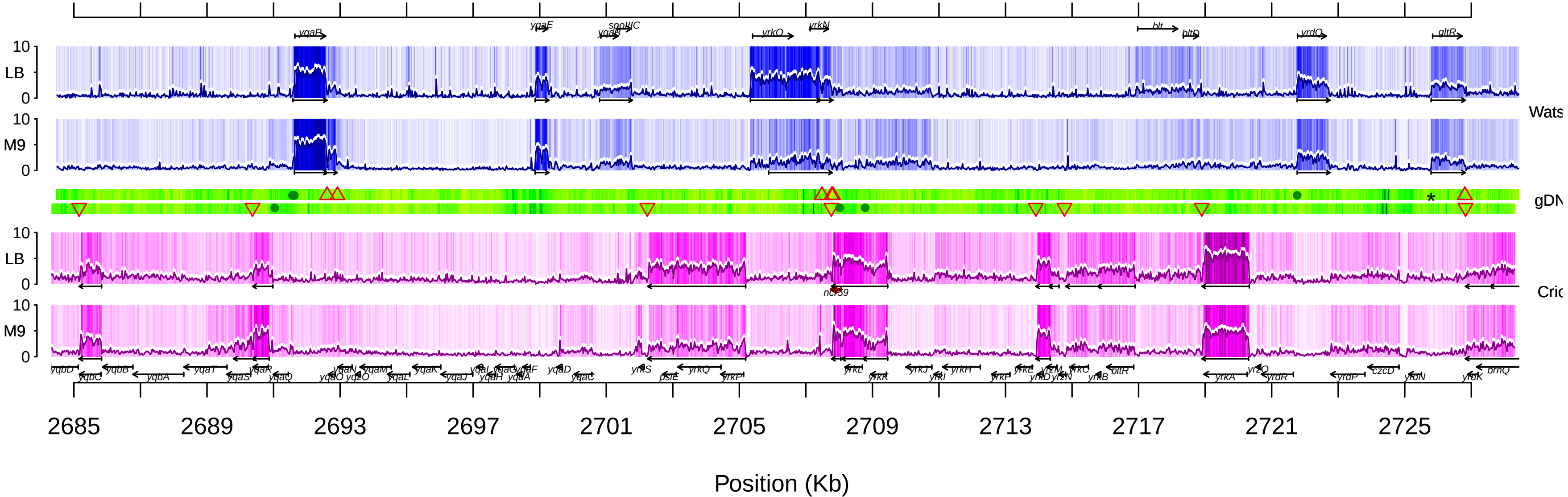
2332 − 2376 Kb



2641 − 2684 Kb



2685 − 2728 Kb



Appendix B

Paper IV

Target	Primers and probes	Sequence (5' – 3')
IFN-β (NM_010510)	Forward	CGGACTTCAAGATCCCTATGGA
	Reverse	TGGCAAAGGCAGTGTAACTCTTC
	Probe	ATGACGGAGAAGATGC
TLR-3 (NM_126166)	Forward	GATTCTTCTGGTGTCTTCCACAAA
	Reverse	AATGGCTGCAGTCAGCTACGT
	Probe	CAATGCACTGTGAGATAC
IL-12 p40 (NM_008352)	Forward	TGGAGCACTCCCCATTCCT
	Reverse	TGCGCTGGATTCGAACAA
	Probe	CTTCTCCCTCAAGTTC
IL-10 (NM_010548)	Forward	GATGCCCCAGGCAGAGAA
	Reverse	CACCCAGGGAATTCAAATGC
	Probe	CATGGCCCAGAAAT
Beta actin (NM_007393)	Forward	CGATGCCCTGAGGCTCTTT
	Reverse	TGGATGCCACAGGATTCCA
	Probe	CCAGCCTTCCTTCTT

Table B.1 – Primers and probes used for Real-Time PCR analysis

Table B.2 – Significant up-regulation of interferon-induced genes in murine dendritic cells stimulated with *Lactobacillus acidophilus* NCFM.

Refseq	**Gene name**	**4h**	**10h**	**18h**	**Description**
NM_126166	TLR3	3.1	4.2	2.6	Toll-Like Receptor 3
NM_010510	IFNB1	4.1	1.9	1.2	Interferon-β
NM_021384	RSAD2	5.9	6	6.6	Interferon-induced protein Viperin
NM_020583	ISG20	3.9	5.5	5.4	Interferon-stimulated exonuclease
NM_011163	PKR	2.7	2.2	1.5	dsRNA-activated protein kinase
					P56 family
NM_008331	ISG56	5.8	5.1	5.2	Interferon-stimulated gene 56
NM_008332	ISG54	5.9	5.9	5.4	Interferon-stimulated gene 54
NM_010501	ISG49	5.3	5.7	4.9	Interferon-stimulated gene 49
					OAS family
NM_145209	OASL1	4.1	4.7	4.6	Oligoadenylate synthetase-like 1
NM_011854	OASL2	3.7	3.3	3	Oligoadenylate synthetase-like 2
NM_145227	OAS2	2.2	2.3	1.6	Oligoadenylate synthetase 2
NM_145226	OAS3	2.4	2.6	2.4	Oligoadenylate synthetase 3
NM_011852	OAS1G	1.9	2	1.7	Oligoadenylate synthetase 1G
NM_033541	OAS1C	1.1	1.2	1.2	Oligoadenylate synthetase 1C
					Mx proteins
NM_013606	MX2	3.3	3.1	2.6	Myxovirus resistance 2
NM_010846	MX1	2.7	2.7	1.9	Myxovirus resistance 1
					p200 gene family
NM_001045481	IFI203	3.4	2.9	2.5	Interferon activated gene 203
NM_008329	IFI204	2.5	3.4	2.7	Interferon activated gene 204
NM_172648	IFI205	3.1	3.2	3	Interferon activated gene 205
XM_001477431	LOC623121	4.4	5.1	4.6	Novel interferon-beta induced gene
NM_027320	IFI35	1.9	1.5	0.89	Interferon-induced protein 35
NM_133871	IFI44	3.5	4.8	4.6	Interferon-induced protein 44
					Interferon-induced GTPases
NM_021792	IIGP1	4.7	5.2	4.9	Interferon inducible GTPase 1
NM_019440	IIGP2	3.1	2.3	1.8	Interferon inducible GTPase 2
NM_001039160	GVIN1	1.8	1.4	1.3	Interferon inducible GTPase

Table B.2 – Continued

Refseq	**Gene name**	**4h**	**10h**	**18h**	**Description**
					Interferon-induced helicases
NM_172689	DDX58	2.9	2.7	2.3	RNA helicase DDX58
NM_027835	IFIH1	3.6	3.5	2.6	Interferon induced with helicase C1
					Protein Ubiquitination
NM_022329	ISG15	1.9	1.7	2	Interferon-stimulated gene 15
XM_001478484	HERC5	3	3.4	3.3	IFN-induced E3 protein ligase
NM_019949	UBE2L6	1.9	2.7	2.1	ISG-15-conjugating enzyme
NM_011909	USP18	3.7	3.6	3.3	Protease specifically removing ISG15
NM_023738	UBE1l	1.6	2.3	2	Ubiquitin-activating enzyme E1-like
NM_019949	UBE2l6	1.9	2.7	2.1	Ubiquitin-conjugating enzyme E2L6
XR_005074	LOC677168	4.2	4.8	4.8	Similar to ISG15 ubiquitin-like modifier
					Interferon regulatory factors (IRF)
NM_016850	IRF7	4.1	4.5	4.2	Interferon regulatory factor 7
					Misc
NM_028864	Zc3hav1	1.6	1.2	0.76	Antiviral zinc and RNA binding protein
NM_001038587	Adar	1.9	1.7	1.8	Adenosine deaminase (binds dsRNA)
NM_175397	Sp110	1.5	1.1	0.43	Sp110 nuclear body protein
NM_011636	Plscr1	1.1	0.35	0.2	Phospholipid scramblase 1

Bibliography

1. Lazebnik Y (2002) Can a biologist fix a radio?–or, what i learned while studying apoptosis. Cancer Cell 2: 179-82.

2. Kitano H (2002) Systems biology: a brief overview. Science 295: 1662-4.

3. Guido NJ, Wang X, Adalsteinsson D, McMillen D, Hasty J, et al. (2006) A bottom-up approach to gene regulation. Nature 439: 856-60.

4. Bray D (2003) Molecular networks: the top-down view. Science 301: 1864-5.

5. Barabási AL, Oltvai ZN (2004) Network biology: understanding the cell's functional organization. Nat Rev Genet 5: 101-13.

6. Auffray C, Chen Z, Hood L (2009) Systems medicine: the future of medical genomics and healthcare. Genome Med 1: 2.

7. Butcher EC, Berg EL, Kunkel EJ (2004) Systems biology in drug discovery. Nat Biotechnol 22: 1253-9.

8. http://www.alzgene.org/topresults.asp.

9. Altshuler D, Daly MJ, Lander ES (2008) Genetic mapping in human disease. Science 322: 881-8.

10. Harold D, Abraham R, Hollingworth P, Sims R, Gerrish A, et al. (2009) Genome-wide association study identifies variants at clu and picalm associated with alzheimer's disease. Nat Genet 41: 1088-1093.

11. Lambert JC, Heath S, Even G, Campion D, Sleegers K, et al. (2009) Genome-wide association study identifies variants at clu and cr1 associated with alzheimer's disease. Nat Genet 41: 1094-1099.

12. Gavin AC, Aloy P, Grandi P, Krause R, Boesche M, et al. (2006) Proteome survey reveals modularity of the yeast cell machinery. Nature 440: 631-6.

13. Wu X, Jiang R, Zhang MQ, Li S (2008) Network-based global inference of human disease genes. Mol Syst Biol 4: 189.

14. Oti M, Brunner HG (2007) The modular nature of genetic diseases. Clin Genet 71: 1-11.

15. Lage K, Karlberg EO, Størling ZM, Olason PI, Pedersen AG, et al. (2007) A human phenome-interactome network of protein complexes implicated in genetic disorders. Nat Biotechnol 25: 309-16.

16. Gerstein MB, Bruce C, Rozowsky JS, Zheng D, Du J, et al. (2007) What is a gene, post-encode? history and updated definition. Genome Res 17: 669-81.

17. 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.

18. Wilhelm BT, Marguerat S, Watt S, Schubert F, Wood V, et al. (2008) Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution. Nature 453: 1239-43.

19. Wang Z, Gerstein M, Snyder M (2009) Rna-seq: a revolutionary tool for transcriptomics. Nat Rev Genet 10: 57-63.

20. Xu Z, Wei W, Gagneur J, Perocchi F, Clauder-Münster S, et al. (2009) Bidirectional promoters generate pervasive transcription in yeast. Nature 457: 1033-7.

21. Faghihi MA, Wahlestedt C (2009) Regulatory roles of natural antisense transcripts. Nat Rev Mol Cell Biol 10: 637-43.

22. Kornberg RD (2007) The molecular basis of eukaryotic transcription. Proc Natl Acad Sci U S A 104: 12955-61.

23. Browning DF, Busby SJ (2004) The regulation of bacterial transcription initiation. Nat Rev Microbiol 2: 57-65.

24. Alberts B (2008) Molecular biology of the cell. Garland Science, 5 edition.

25. Szutorisz H, Dillon N, Tora L (2005) The role of enhancers as centres for general transcription factor recruitment. Trends Biochem Sci 30: 593-9.

26. Spilianakis CG, Lalioti MD, Town T, Lee GR, Flavell RA (2005) Interchromosomal associations between alternatively expressed loci. Nature 435: 637-45.

27. Deutscher MP (2006) Degradation of rna in bacteria: comparison of mrna and stable rna. Nucleic Acids Res 34: 659-66.

28. Sierro N, Makita Y, de Hoon M, Nakai K (2008) Dbtbs: a database of transcriptional regulation in bacillus subtilis containing upstream intergenic conservation information. Nucleic Acids Res 36: 93-96.

29. Mazumder B, Seshadri V, Fox PL (2003) Translational control by the 3'-utr: the ends specify the means. Trends Biochem Sci 28: 91-98.

30. Chang TC, Mendell JT (2007) micrornas in vertebrate physiology and human disease. Annu Rev Genomics Hum Genet 8: 215-39.

31. Faghihi MA, Modarresi F, Khalil AM, Wood DE, Sahagan BG, et al. (2008) Expression of a noncoding rna is elevated in alzheimer's disease and drives rapid feed-forward regulation of beta-secretase. Nat Med 14: 723-30.

32. Toledo-Arana A, Dussurget O, Nikitas G, Sesto N, Guet-Revillet H, et al. (2009) The listeria transcriptional landscape from saprophytism to virulence. Nature 459: 950-6.

33. Wurtzel O, Sapra R, Chen F, Zhu Y, Simmons BA, et al. (2009) A single-base resolution map of an archaeal transcriptome. Genome Res .

34. Morozova O, Hirst M, Marra MA (2009) Applications of new sequencing technologies for transcriptome analysis. Annu Rev Genomics Hum Genet 10: 135-51.

35. Mardis ER (2008) The impact of next-generation sequencing technology on genetics. Trends Genet 24: 133-41.

36. Southern EM (1975) Detection of specific sequences among dna fragments separated by gel electrophoresis. J Mol Biol 98: 503-17.

37. Alwine JC, Kemp DJ, Stark GR (1977) Method for detection of specific rnas in agarose gels by transfer to diazobenzyloxymethyl-paper and hybridization with dna probes. Proc Natl Acad Sci U S A 74: 5350-4.

38. Jordan B (2002) Historical background and anticipated developments. Ann N Y Acad Sci 975: 24-32.

39. Dufva M (2009) Introduction to microarray technology. Methods Mol Biol 529: 1-22.

40. Schena M, Shalon D, Davis RW, Brown PO (1995) Quantitative monitoring of gene expression patterns with a complementary dna microarray. Science 270: 467-70.

41. Brown PO, Botstein D (1999) Exploring the new world of the genome with dna microarrays. Nat Genet 21: 33-7.

42. Dalma-Weiszhausz DD, Warrington J, Tanimoto EY, Miyada CG (2006) The affymetrix genechip platform: an overview. Methods Enzymol 410: 3-28.

43. Wolber PK, Collins PJ, Lucas AB, De Witte A, Shannon KW (2006) The agilent in situ-synthesized microarray platform. Methods Enzymol 410: 28-57.

44. Dufva M (2009) Fabrication of dna microarray. Methods Mol Biol 529: 63-79.

45. Parkinson H, Kapushesky M, Kolesnikov N, Rustici G, Shojatalab M, et al. (2009) Arrayexpress update–from an archive of functional genomics experiments to the atlas of gene expression. Nucleic Acids Res 37: D868-72.

46. Barrett T, Troup DB, Wilhite SE, Ledoux P, Rudnev D, et al. (2009) Ncbi geo: archive for high-throughput functional genomic data. Nucleic Acids Res 37: D885-90.

47. DeRisi JL, Iyer VR, Brown PO (1997) Exploring the metabolic and genetic control of gene expression on a genomic scale. Science 278: 680-6.

48. Cho RJ, Campbell MJ, Winzeler EA, Steinmetz L, Conway A, et al. (1998) A genome-wide transcriptional analysis of the mitotic cell cycle. Mol Cell 2: 65-73.

49. Chu S, DeRisi J, Eisen M, Mulholland J, Botstein D, et al. (1998) The transcriptional program of sporulation in budding yeast. Science 282: 699-705.

50. Spellman PT, Sherlock G, Zhang MQ, Iyer VR, Anders K, et al. (1998) Comprehensive identification of cell cycle-regulated genes of the yeast saccharomyces cerevisiae by microarray hybridization. Mol Biol Cell 9: 3273-97.

51. Su AI, Wiltshire T, Batalov S, Lapp H, Ching KA, et al. (2004) A gene atlas of the mouse and human protein-encoding transcriptomes. Proc Natl Acad Sci U S A 101: 6062-7.

52. 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: 466-75.

53. Ren B, Robert F, Wyrick JJ, Aparicio O, Jennings EG, et al. (2000) Genome-wide location and function of dna binding proteins. Science 290: 2306-9.

54. Iyer VR, Horak CE, Scafe CS, Botstein D, Snyder M, et al. (2001) Genomic binding sites of the yeast cell-cycle transcription factors sbf and mbf. Nature 409: 533-8.

55. Buck MJ, Lieb JD (2004) Chip-chip: considerations for the design, analysis, and application of genome-wide chromatin immunoprecipitation experiments. Genomics 83: 349-60.

56. Park PJ (2009) Chip-seq: advantages and challenges of a maturing technology. Nat Rev Genet 10: 669-80.

57. Bird A (2007) Perceptions of epigenetics. Nature 447: 396-8.

58. Suzuki MM, Bird A (2008) Dna methylation landscapes: provocative insights from epigenomics. Nat Rev Genet 9: 465-76.

59. Esteller M (2008) Epigenetics in cancer. N Engl J Med 358: 1148-59.

60. International HapMap Consortium (2005) A haplotype map of the human genome. Nature 437: 1299-320.

61. International HapMap Consortium, Frazer KA, Ballinger DG, Cox DR, Hinds DA, et al. (2007) A second generation human haplotype map of over 3.1 million snps. Nature 449: 851-61.

62. McCarthy MI, Abecasis GR, Cardon LR, Goldstein DB, Little J, et al. (2008) Genome-wide association studies for complex traits: consensus, uncertainty and challenges. Nat Rev Genet 9: 356-69.

63. Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH, et al. (2006) Global variation in copy number in the human genome. Nature 444: 444-54.

64. Zhang F, Gu W, Hurles ME, Lupski JR (2009) Copy number variation in human health, disease, and evolution. Annu Rev Genomics Hum Genet 10: 451-81.

65. Fodor SP, Read JL, Pirrung MC, Stryer L, Lu AT, et al. (1991) Light-directed, spatially addressable parallel chemical synthesis. Science 251: 767-73.

66. Hacia JG, Brody LC, Chee MS, Fodor SP, Collins FS (1996) Detection of heterozygous mutations in brca1 using high density oligonucleotide arrays and two-colour fluorescence analysis. Nat Genet 14: 441-7.

67. Cronin MT, Fucini RV, Kim SM, Masino RS, Wespi RM, et al. (1996) Cystic fibrosis mutation detection by hybridization to light-generated dna probe arrays. Hum Mutat 7: 244-55.

68. Kozal MJ, Shah N, Shen N, Yang R, Fucini R, et al. (1996) Extensive polymorphisms observed in hiv-1 clade b protease gene using high-density oligonucleotide arrays. Nat Med 2: 753-9.

69. Lin B, Wang Z, Vora GJ, Thornton JA, Schnur JM, et al. (2006) Broad-spectrum respiratory tract pathogen identification using resequencing dna microarrays. Genome Res 16: 527-35.

70. Roche. Amplichip cyp450 test. URL `http://www.amplichip.us/documents/CYP450_P.I._US-IVD.pdf`.

71. Chomczynski P, Sacchi N (1987) Single-step method of rna isolation by acid guanidinium thiocyanate-phenol-chloroform extraction. Anal Biochem 162: 156-9.

72. Peirson SN, Butler JN (2007) Rna extraction from mammalian tissues. Methods Mol Biol 362: 315-27.

73. Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, et al. (2004) Bioconductor: open software development for computational biology and bioinformatics. Genome Biol 5: R80.

74. Team RDC (2008) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. URL `http://www.R-project.org`.

75. Gautier L, Cope L, Bolstad BM, Irizarry RA (2004) affy–analysis of affymetrix genechip data at the probe level. Bioinformatics 20: 307-15.

76. Smyth GK (2005) Limma: linear models for microarray data. In: Gentleman R, Carey V, Dudoit S, R Irizarry WH, editors, Bioinformatics and Computational Biology Solutions using R and Bioconductor, New York: Springer. pp. 397-420.

77. Hartemink AJ, Gifford DK, Jaakkola TS, Young RA (2001) Maximum likelihood estimation of optimal scaling factors for expression array normalization. SPIE BIOS .

78. Irizarry RA, Bolstad BM, Collin F, Cope LM, Hobbs B, et al. (2003) Summaries of affymetrix genechip probe level data. Nucleic Acids Res 31: e15.

79. Ritchie ME, Silver J, Oshlack A, Holmes M, Diyagama D, et al. (2007) A comparison of background correction methods for two-colour microarrays. Bioinformatics 23: 2700-7.

80. Wernisch L. Background correction in the rma algorithm. URL `http://www.biochem.ucl.ac.uk/~harry/MAD/rma_bg.pdf`.

81. Bolstad BM, Irizarry RA, Astrand M, Speed TP (2003) A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics 19: 185-93.

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

83. Workman C, Jensen LJ, Jarmer H, Berka R, Gautier L, et al. (2002) A new non-linear normalization method for reducing variability in dna microarray experiments. Genome Biol 3: research0048.

84. Yang YH, Dudoit S, Luu P, Lin DM, Peng V, et al. (2002) Normalization for cdna microarray data: a robust composite method addressing single and multiple slide systematic variation. Nucleic Acids Res 30: e15.

85. Baldi P, Long AD (2001) A bayesian framework for the analysis of microarray expression data: regularized t -test and statistical inferences of gene changes. Bioinformatics 17: 509-19.

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

87. Quackenbush J (2001) Computational analysis of microarray data. Nat Rev Genet 2: 418-27.

88. Shay E. Microarray cluster analysis and applications. URL `http://www.science.co.il/enuka/Essays/Microarray-Review.pdf`.

89. Maechler M, Rousseeuw P, Struyf A, Hubert M (2005) luster analysis basics and extensions. Rousseeuw et al provided the S original which has been ported to R by Kurt Hornik and has since been enhanced by Martin Maechler: speed improvements, silhouette() functionality, bug fixes, etc. See the 'Changelog' file (in the package source).

90. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, et al. (2000) Gene ontology: tool for the unification of biology. the gene ontology consortium. Nat Genet 25: 25-9.

91. Kanehisa M, Araki M, Goto S, Hattori M, Hirakawa M, et al. (2008) Kegg for linking genomes to life and the environment. Nucleic Acids Res 36: 480-484.

92. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, et al. (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A 102: 15545-50.

93. Kim SY, Volsky DJ (2005) Page: parametric analysis of gene set enrichment. BMC Bioinformatics 6: 144.

94. Huber W, Toedling J, Steinmetz LM (2006) Transcript mapping with high-density oligonucleotide tiling arrays. Bioinformatics 22: 1963-1970.

95. Royce TE, Rozowsky JS, Gerstein MB (2007) Assessing the need for sequence-based normalization in tiling microarray experiments. Bioinformatics 23: 988-97.

96. Huber W, Toedling J, Steinmetz LM (2006) Transcript mapping with high-density oligonucleotide tiling arrays. Bioinformatics 22: 1963-70.

97. Venkatraman ES, Olshen AB (2007) A faster circular binary segmentation algorithm for the analysis of array cgh data. Bioinformatics 23: 657-63.

98. Nicolas P, Leduc A, Robin S, Rasmussen S, Jarmer H, et al. (2009) Transcriptional landscape estimation from tiling array data using a model of signal shift and drift. Bioinformatics 25: 2341-2347.

99. Laub MT, McAdams HH, Feldblyum T, Fraser CM, Shapiro L (2000) Global analysis of the genetic network controlling a bacterial cell cycle. Science 290: 2144-8.

100. Whitfield ML, Sherlock G, Saldanha AJ, Murray JI, Ball CA, et al. (2002) Identification of genes periodically expressed in the human cell cycle and their expression in tumors. Mol Biol Cell 13: 1977-2000.

101. Menges M, Hennig L, Gruissem W, Murray JAH (2003) Genome-wide gene expression in an arabidopsis cell suspension. Plant Mol Biol 53: 423-42.

102. Lundgren M, Andersson A, Chen L, Nilsson P, Bernander R (2004) Three replication origins in sulfolobus species: synchronous initiation of chromosome replication and asynchronous termination. Proc Natl Acad Sci U S A 101: 7046-51.

103. Peng X, Karuturi RKM, Miller LD, Lin K, Jia Y, et al. (2005) Identification of cell cycle-regulated genes in fission yeast. Mol Biol Cell 16: 1026-42.

104. Laurent SJ, Vannier FS (1973) Temperature-sensitive initiation of chromosome replication in a mutant of bacillus subtilis. J Bacteriol 114: 474-84.

105. Dwek RD, Kobrin LH, Grossman N, Ron EZ (1980) Synchronization of cell division in microorganisms by percoll gradients. J Bacteriol 144: 17-21.

106. Hart A, Edwards C (1987) Buoyant density fluctuations during the cell cycle of bacillus subtilis. Arch Microbiol 147: 68-72.

107. Hassan AK, Moriya S, Ogura M, Tanaka T, Kawamura F, et al. (1997) Suppression of initiation defects of chromosome replication in bacillus subtilis dnaa and oric-deleted mutants by integration of a plasmid replicon into the chromosomes. J Bacteriol 179: 2494-502.

108. Keijser BJF, Ter Beek A, Rauwerda H, Schuren F, Montijn R, et al. (2007) Analysis of temporal gene expression during bacillus subtilis spore germination and outgrowth. J Bacteriol 189: 3624-34.

109. Wernersson R, Nielsen HB (2005) Oligowiz 2.0–integrating sequence feature annotation into the design of microarray probes. Nucleic Acids Res 33: W611-5.

110. de Hoon MJ, Makita Y, Nakai K, Miyano S (2005) Prediction of transcriptional terminators in bacillus subtilis and related species. PLoS Comput Biol 1.

111. Ingham CJ, Dennis J, Furneaux PA (1999) Autogenous regulation of transcription termination factor rho and the requirement for nus factors in bacillus subtilis. Mol Microbiol 31: 651-63.

112. Matos CFRO, Di Cola A, Robinson C (2009) Tatd is a central component of a tat translocon-initiated quality control system for exported fes proteins in escherichia coli. EMBO Rep 10: 474-9.

113. Peña-Castillo L, Hughes TR (2007) Why are there still over 1000 uncharacterized yeast genes? Genetics 176: 7-14.

114. Nielsen J, Jewett MC (2008) Impact of systems biology on metabolic engineering of saccharomyces cerevisiae. FEMS Yeast Res 8: 122-31.

115. Alper H, Moxley J, Nevoigt E, Fink GR, Stephanopoulos G (2006) Engineering yeast transcription machinery for improved ethanol tolerance and production. Science 314: 1565-8.

116. Förstl H, Kurz A (1999) Clinical features of alzheimer's disease. Eur Arch Psychiatry Clin Neurosci 249: 288-290.

117. Berchtold NC, Cotman CW (1998) Evolution in the conceptualization of dementia and alzheimer's disease: Greco-roman period to the 1960s. Neurobiol Aging 19: 173-189.

118. Crutch SJ, Isaacs R, Rossor MN (2001) Some workmen can blame their tools: artistic change in an individual with alzheimer's disease. Lancet 357: 2129-2133.

119. Hebert LE, Scherr PA, Bienias JL, Bennett DA, Evans DA (2003) Alzheimer disease in the us population: prevalence estimates using the 2000 census. Arch Neurol 60: 1119-1122.

120. Cupers P, Sautter J, Vanvossel A (2006) European union research policy and funding for alzheimer disease. Nat Med 12: 774-775.

121. Zhu CW, Sano M (2006) Economic considerations in the management of alzheimer's disease. Clin Interv Aging 1: 143-154.

122. http://www.ncbi.nlm.nih.gov/pubmed/. URL `http://www.ncbi.nlm.nih.gov/pubmed/`.

123. Kaduszkiewicz H, Zimmermann T, Beck-Bornholdt HP, van den Bussche H (2005) Cholinesterase inhibitors for patients with alzheimer's disease: systematic review of randomised clinical trials. BMJ 331: 321-327.

124. Lipton SA (2006) Paradigm shift in neuroprotection by nmda receptor blockade: memantine and beyond. Nat Rev Drug Discov 5: 160-170.

125. van Marum RJ (2009) Update on the use of memantine in alzheimer's disease. Neuropsychiatr Dis Treat 5: 237-247.

126. Francis PT, Palmer AM, Snape M, Wilcock GK (1999) The cholinergic hypothesis of alzheimer's disease: a review of progress. J Neurol Neurosurg Psychiatry 66: 137-147.

127. Raux G, Guyant-Maréchal L, Martin C, Bou J, Penet C, et al. (2005) Molecular diagnosis of autosomal dominant early onset alzheimer's disease: an update. J Med Genet 42: 793-795.

128. Janssen JC, Beck JA, Campbell TA, Dickinson A, Fox NC, et al. (2003) Early onset familial alzheimer's disease: Mutation frequency in 31 families. Neurology 60: 235-239.

129. Glenner GG, Wong CW (1984) Alzheimer's disease and down's syndrome: sharing of a unique cerebrovascular amyloid fibril protein. Biochem Biophys Res Commun 122: 1131-1135.

130. Goldgaber D, Lerman MI, McBride OW, Saffiotti U, Gajdusek DC (1987) Characterization and chromosomal localization of a cdna encoding brain amyloid of alzheimer's disease. Science 235: 877-880.

131. Kang J, Lemaire HG, Unterbeck A, Salbaum JM, Masters CL, et al. (1987) The precursor of alzheimer's disease amyloid a4 protein resembles a cell-surface receptor. Nature 325: 733-736.

132. Goate A, Chartier-Harlin MC, Mullan M, Brown J, Crawford F, et al. (1991) Segregation of a missense mutation in the amyloid precursor protein gene with familial alzheimer's disease. Nature 349: 704-706.

133. Levy-Lahad E, Wasco W, Poorkaj P, Romano DM, Oshima J, et al. (1995) Candidate gene for the chromosome 1 familial alzheimer's disease locus. Science 269: 973-977.

134. Sherrington R, Rogaev EI, Liang Y, Rogaeva EA, Levesque G, et al. (1995) Cloning of a gene bearing missense mutations in early-onset familial alzheimer's disease. Nature 375: 754-760.

135. Hardy J (2009) The amyloid hypothesis. The ADIT project.

136. Bertram L, Tanzi RE (2008) Thirty years of alzheimer's disease genetics: the implications of systematic meta-analyses. Nat Rev Neurosci 9: 768-778.

137. Corder EH, Saunders AM, Strittmatter WJ, Schmechel DE, Gaskell PC, et al. (1993) Gene dose of apolipoprotein e type 4 allele and the risk of alzheimer's disease in late onset families. Science 261: 921-923.

138. Strittmatter WJ, Saunders AM, Schmechel D, Pericak-Vance M, Enghild J, et al. (1993) Apolipoprotein e: high-avidity binding to beta-amyloid and increased frequency of type 4 allele in late-onset familial alzheimer disease. Proc Natl Acad Sci U S A 90: 1977-1981.

139. Bu G (2009) Apolipoprotein e and its receptors in alzheimer's disease: pathways, pathogenesis and therapy. Nat Rev Neurosci 10: 333-344.

140. Coon KD, Myers AJ, Craig DW, Webster JA, Pearson JV, et al. (2007) A high-density whole-genome association study reveals that apoe is the major susceptibility gene for sporadic late-onset alzheimer's disease. J Clin Psychiatry 68: 613-618.

141. May PC, Finch CE (1992) Sulfated glycoprotein 2: new relationships of this multifunctional protein to neurodegeneration. Trends Neurosci 15: 391-396.

142. Roheim PS, Carey M, Forte T, Vega GL (1979) Apolipoproteins in human cerebrospinal fluid. Proc Natl Acad Sci U S A 76: 4646-4649.

143. Gatz M, Reynolds CA, Fratiglioni L, Johansson B, Mortimer JA, et al. (2006) Role of genes and environments for explaining alzheimer disease. Arch Gen Psychiatry 63: 168-174.

144. Stozická Z, Zilka N, Novák M (2007) Risk and protective factors for sporadic alzheimer's disease. Acta Virol 51: 205-222.

145. Mattson MP (2004) Pathways towards and away from alzheimer's disease. Nature 430: 631-639.

146. Braak H, Braak E (1995) Staging of alzheimer's disease-related neurofibrillary changes. Neurobiol Aging 16: 271-8; discussion 278-84.

147. Braak H, Braak E (1991) Neuropathological stageing of alzheimer-related changes. Acta Neuropathol 82: 239-59.

148. Hardy J (2009) The amyloid hypothesis for alzheimer's disease: a critical reappraisal. J Neurochem 110: 1129-1134.

149. Hardy J, Selkoe DJ (2002) The amyloid hypothesis of alzheimer's disease: progress and problems on the road to therapeutics. Science 297: 353-356.

150. Kayed R, Head E, Thompson JL, McIntire TM, Milton SC, et al. (2003) Common structure of soluble amyloid oligomers implies common mechanism of pathogenesis. Science 300: 486-489.

151. Lambert MP, Barlow AK, Chromy BA, Edwards C, Freed R, et al. (1998) Diffusible, nonfibrillar ligands derived from abeta1-42 are potent central nervous system neurotoxins. Proc Natl Acad Sci U S A 95: 6448-6453.

152. Hutton M, Lendon CL, Rizzu P, Baker M, Froelich S, et al. (1998) Association of missense and 5'-splice-site mutations in tau with the inherited dementia ftdp-17. Nature 393: 702-5.

153. Nikolaev A, McLaughlin T, O'Leary DDM, Tessier-Lavigne M (2009) App binds dr6 to trigger axon pruning and neuron death via distinct caspases. Nature 457: 981-9.

154. Kim D, Tsai LH (2009) Bridging physiology and pathology in ad. Cell 137: 997-1000.

155. Lammich S, Kojro E, Postina R, Gilbert S, Pfeiffer R, et al. (1999) Constitutive and regulated alpha-secretase cleavage of alzheimer's amyloid precursor protein by a disintegrin metalloprotease. Proc Natl Acad Sci U S A 96: 3922-3927.

156. Cai H, Wang Y, McCarthy D, Wen H, Borchelt DR, et al. (2001) Bace1 is the major beta-secretase for generation of abeta peptides by neurons. Nat Neurosci 4: 233-234.

157. Asai M, Hattori C, Szabó B, Sasagawa N, Maruyama K, et al. (2003) Putative function of adam9, adam10, and adam17 as app alpha-secretase. Biochem Biophys Res Commun 301: 231-235.

158. Tanabe C, Hotoda N, Sasagawa N, Sehara-Fujisawa A, Maruyama K, et al. (2007) Adam19 is tightly associated with constitutive alzheimer's disease app alpha-secretase in a172 cells. Biochem Biophys Res Commun 352: 111-117.

159. Kaether C, Haass C, Steiner H (2006) Assembly, trafficking and function of gamma-secretase. Neurodegener Dis 3: 275-283.

160. Thinakaran G, Koo EH (2008) Amyloid precursor protein trafficking, processing, and function. J Biol Chem 283: 29615-9.

161. Puzzo D, Privitera L, Leznik E, Fà M, Staniszewski A, et al. (2008) Picomolar amyloid-beta positively modulates synaptic plasticity and memory in hippocampus. J Neurosci 28: 14537-45.

162. Terstappen GC, Reggiani A (2001) In silico research in drug discovery. Trends Pharmacol Sci 22: 23-26.

163. Mucke L (2009) Neuroscience: Alzheimer's disease. Nature 461: 895-7.

164. Overington JP, Al-Lazikani B, Hopkins AL (2006) How many drug targets are there? Nat Rev Drug Discov 5: 993-6.

165. Imming P, Sinning C, Meyer A (2006) Drugs, their targets and the nature and number of drug targets. Nat Rev Drug Discov 5: 821-34.

166. Bleicher KH, Böhm HJ, Müller K, Alanine AI (2003) Hit and lead generation: beyond high-throughput screening. Nat Rev Drug Discov 2: 369-78.

167. Rawlins MD (2004) Cutting the cost of drug development? Nat Rev Drug Discov 3: 360-4.

168. Loo DT, Copani A, Pike CJ, Whittemore ER, Walencewicz AJ, et al. (1993) Apoptosis is induced by beta-amyloid in cultured central nervous system neurons. Proc Natl Acad Sci U S A 90: 7951-7955.

169. Portelius E, Westman-Brinkmalm A, Zetterberg H, Blennow K (2006) Determination of beta-amyloid peptide signatures in cerebrospinal fluid using immunoprecipitation-mass spectrometry. J Proteome Res 5: 1010-6.

170. Kaminsky YG, Marlatt MW, Smith MA, Kosenko EA (2009) Subcellular and metabolic examination of amyloid-beta peptides in alzheimer disease pathogenesis: Evidence for abeta(25-35). Exp Neurol .

171. Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, et al. (2004) Bioconductor: open software development for computational biology and bioinformatics. Genome Biol 5.

172. Workman C, Jensen LJ, Jarmer H, Berka R, Gautier L, et al. (2002) A new non-linear normalization method for reducing variability in dna microarray experiments. Genome Biol 3.

173. Lemon WJ, Liyanarachchi S, You M (2003) A high performance test of differential gene expression for oligonucleotide arrays. Genome Biol 4: R67.

174. Sun YP, Deng KJ, Wang F, Zhang J, Huang X, et al. (2004) Two novel isoforms of adam23 expressed in the developmental process of mouse and human brains. Gene 325: 171-8.

175. Hardy J, Allsop D (1991) Amyloid deposition as the central event in the aetiology of alzheimer's disease. Trends Pharmacol Sci 12: 383-8.

176. Ballatore C, Lee VMY, Trojanowski JQ (2007) Tau-mediated neurodegeneration in alzheimer's disease and related disorders. Nat Rev Neurosci 8: 663-72.

177. Gendron TF, Petrucelli L (2009) The role of tau in neurodegeneration. Mol Neurodegener 4: 13.

178. Bertram L, Tanzi RE (2009) Genome-wide association studies in alzheimer's disease. Hum Mol Genet 18: 137-145.

179. Lovell MA, Xiong S, Markesbery WR, Lynn BC (2005) Quantitative proteomic analysis of mitochondria from primary neuron cultures treated with amyloid beta peptide. Neurochem Res 30: 113-22.

180. Paratore S, Parenti R, Torrisi A, Copani A, Cicirata F, et al. (2006) Genomic profiling of cortical neurons following exposure to beta-amyloid. Genomics 88: 468-79.

181. Sultana R, Newman SF, Abdul HM, Cai J, Pierce WM, et al. (2006) Protective effect of d609 against amyloid-beta1-42-induced oxidative modification of neuronal proteins: redox proteomics study. J Neurosci Res 84: 409-17.

182. Thomas SN, Soreghan BA, Nistor M, Sarsoza F, Head E, et al. (2005) Reduced neuronal expression of synaptic transmission modulator hnk-1/neural cell adhesion molecule as a potential consequence of amyloid beta-mediated oxidative stress: a proteomic approach. J Neurochem 92: 705-17.

183. Pollio G, Hoozemans JJM, Andersen CA, Roncarati R, Rosi MC, et al. (2008) Increased expression of the oligopeptidase thop1 is a neuroprotective response to abeta toxicity. Neurobiol Dis 31: 145-58.

184. Sagane K, Yamazaki K, Mizui Y, Tanaka I (1999) Cloning and chromosomal mapping of mouse adam11, adam22 and adam23. Gene 236: 79-86.

185. Yang P, Baker KA, Hagg T (2006) The adams family: coordinators of nervous system development, plasticity and repair. Prog Neurobiol 79: 73-94.

186. Owuor K, Harel NY, Englot DC, Hisama F, Blumenfeld H, et al. (2009) Lgi1-associated epilepsy through altered adam23-dependent neuronal morphology. Mol Cell Neurosci .

187. Lein ES, Hawrylycz MJ, Ao N, Ayres M, Bensinger A, et al. (2007) Genome-wide atlas of gene expression in the adult mouse brain. Nature 445: 168-76.

188. Goldsmith AP, Gossage SJ, ffrench Constant C (2004) Adam23 is a cell-surface glycoprotein expressed by central nervous system neurons. J Neurosci Res 78: 647-58.

189. Sagane K, Ohya Y, Hasegawa Y, Tanaka I (1998) Metalloproteinase-like, disintegrin-like, cysteine-rich proteins mdc2 and mdc3: novel human cellular disintegrins highly expressed in the brain. Biochem J 334 (Pt 1): 93-8.

190. Cal S, Freije JM, López JM, Takada Y, López-Otín C (2000) Adam 23/mdc3, a human disintegrin that promotes cell adhesion via interaction with the alphavbeta3 integrin through an rgd-independent mechanism. Mol Biol Cell 11: 1457-69.

191. Costa MDM, Paludo KS, Klassen G, Lopes MH, Mercadante AF, et al. (2009) Characterization of a specific interaction between adam23 and cellular prion protein. Neurosci Lett 461: 16-20.

192. Mitchell KJ, Pinson KI, Kelly OG, Brennan J, Zupicich J, et al. (2001) Functional analysis of secreted and transmembrane proteins critical to mouse development. Nat Genet 28: 241-9.

193. Reisberg B, Ferris SH, de Leon MJ, Crook T (1982) The global deterioration scale for assessment of primary degenerative dementia. Am J Psychiatry 139: 1136-9.

194. Yamaguchi H, Haga C, Hirai S, Nakazato Y, Kosaka K (1990) Distinctive, rapid, and easy labeling of diffuse plaques in the alzheimer brains by a new methenamine silver stain. Acta Neuropathol 79: 569-72.

195. Braak H, Braak E (1991) Demonstration of amyloid deposits and neurofibrillary changes in whole brain sections. Brain Pathol 1: 213-6.

196. Schmitt-Ulms G, Hansen K, Liu J, Cowdrey C, Yang J, et al. (2004) Time-controlled transcardiac perfusion cross-linking for the study of protein interactions in complex tissues. Nat Biotechnol 22: 724-31.

197. Kenwrick S, Watkins A, De Angelis E (2000) Neural cell recognition molecule l1: relating biological complexity to human disease mutations. Hum Mol Genet 9: 879-86.

198. Zalk R, Lehnart SE, Marks AR (2007) Modulation of the ryanodine receptor and intracellular calcium. Annu Rev Biochem 76: 367-85.

199. Laurén J, Gimbel DA, Nygaard HB, Gilbert JW, Strittmatter SM (2009) Cellular prion protein mediates impairment of synaptic plasticity by amyloid-beta oligomers. Nature 457: 1128-32.

200. Vincent B, Paitel E, Saftig P, Frobert Y, Hartmann D, et al. (2001) The disintegrins adam10 and tace contribute to the constitutive and phorbol ester-regulated normal cleavage of the cellular prion protein. J Biol Chem 276: 37743-6.

201. Parkin ET, Watt NT, Hussain I, Eckman EA, Eckman CB, et al. (2007) Cellular prion protein regulates beta-secretase cleavage of the alzheimer's amyloid precursor protein. Proc Natl Acad Sci U S A 104: 11062-7.

202. Graner E, Mercadante AF, Zanata SM, Forlenza OV, Cabral AL, et al. (2000) Cellular prion protein binds laminin and mediates neuritogenesis. Brain Res Mol Brain Res 76: 85-92.

203. Mangé A, Milhavet O, Umlauf D, Harris D, Lehmann S (2002) Prp-dependent cell adhesion in n2a neuroblastoma cells. FEBS Lett 514: 159-62.

204. Hajj GNM, Lopes MH, Mercadante AF, Veiga SS, da Silveira RB, et al. (2007) Cellular prion protein interaction with vitronectin supports axonal growth and is compensated by integrins. J Cell Sci 120: 1915-26.

205. Namba Y, Tomonaga M, Kawasaki H, Otomo E, Ikeda K (1991) Apolipoprotein e immunoreactivity in cerebral amyloid deposits and neurofibrillary tangles in alzheimer's disease and kuru plaque amyloid in creutzfeldt-jakob disease. Brain Res 541: 163-6.

206. Hatakka K, Savilahti E, Pönkä A, Meurman JH, Poussa T, et al. (2001) Effect of long term consumption of probiotic milk on infections in children attending day care centres: double blind, randomised trial. BMJ 322: 1327.

207. Leyer GJ, Li S, Mubasher ME, Reifer C, Ouwehand AC (2009) Probiotic effects on cold and influenza-like symptom incidence and duration in children. Pediatrics 124: e172-9.

208. Rautava S, Salminen S, Isolauri E (2009) Specific probiotics in reducing the risk of acute infections in infancy–a randomised, double-blind, placebo-controlled study. Br J Nutr 101: 1722-6.

209. Goldsby RA, Kindt TK, Osborne BA, Kuby J (2003) Immunology. New York: W.H. Freeman and Company, 5th edition.

210. Trinchieri G, Sher A (2007) Cooperation of toll-like receptor signals in innate immune defence. Nat Rev Immunol 7: 179-90.

211. Fuller R (1991) Probiotics in human medicine. Gut 32: 439-42.

212. Gill HS, Guarner F (2004) Probiotics and human health: a clinical perspective. Postgrad Med J 80: 516-26.

213. Parvez S, Malik KA, Ah Kang S, Kim HY (2006) Probiotics and their fermented food products are beneficial for health. J Appl Microbiol 100: 1171-85.

214. Akira S, Takeda K (2004) Toll-like receptor signalling. Nat Rev Immunol 4: 499-511.

215. Banchereau J, Steinman RM (1998) Dendritic cells and the control of immunity. Nature 392: 245-52.

216. Christensen HR, Frøkiaer H, Pestka JJ (2002) Lactobacilli differentially modulate expression of cytokines and maturation surface markers in murine dendritic cells. J Immunol 168: 171-8.

217. Zeuthen LH, Christensen HR, Frøkiaer H (2006) Lactic acid bacteria inducing a weak interleukin-12 and tumor necrosis factor alpha response in human dendritic cells inhibit strongly stimulating lactic acid bacteria but act synergistically with gram-negative bacteria. Clin Vaccine Immunol 13: 365-75.

218. Iwasaki A, Medzhitov R (2004) Toll-like receptor control of the adaptive immune responses. Nat Immunol 5: 987-95.

219. Alexopoulou L, Holt AC, Medzhitov R, Flavell RA (2001) Recognition of double-stranded rna and activation of nf-kappab by toll-like receptor 3. Nature 413: 732-8.

220. Stetson DB, Medzhitov R (2006) Type i interferons in host defense. Immunity 25: 373-81.

221. Katze MG, He Y, Gale M Jr (2002) Viruses and interferon: a fight for supremacy. Nat Rev Immunol 2: 675-87.

222. Bogdan C, Mattner J, Schleicher U (2004) The role of type i interferons in non-viral infections. Immunol Rev 202: 33-48.

223. Charrel-Dennis M, Latz E, Halmen KA, Trieu-Cuot P, Fitzgerald KA, et al. (2008) Tlr-independent type i interferon induction in response to an extracellular bacterial pathogen via intracellular recognition of its dna. Cell Host Microbe 4: 543-54.

224. Gratz N, Siller M, Schaljo B, Pirzada ZA, Gattermeier I, et al. (2008) Group a streptococcus activates type i interferon production and myd88-dependent signaling without involvement of tlr2, tlr4, and tlr9. J Biol Chem 283: 19879-87.

225. Mancuso G, Midiri A, Biondo C, Beninati C, Zummo S, et al. (2007) Type i ifn signaling is crucial for host resistance against different species of pathogenic bacteria. J Immunol 178: 3126-33.

226. O'Connell RM, Vaidya SA, Perry AK, Saha SK, Dempsey PW, et al. (2005) Immune activation of type i ifns by listeria monocytogenes occurs independently of tlr4, tlr2, and receptor interacting protein 2 but involves tnfr-associated nf kappa b kinase-binding kinase 1. J Immunol 174: 1602-7.

227. Stockinger S, Kastner R, Kernbauer E, Pilz A, Westermayer S, et al. (2009) Characterization of the interferon-producing cell in mice infected with listeria monocytogenes. PLoS Pathog 5: e1000355.

228. Salazar JC, Duhnam-Ems S, La Vake C, Cruz AR, Moore MW, et al. (2009) Activation of human monocytes by live borrelia burgdorferi generates tlr2-dependent and -independent responses which include induction of ifn-beta. PLoS Pathog 5: e1000444.

229. Kawai T, Akira S (2008) Toll-like Receptor and RIG-1-like Receptor Signaling. Blackwell Publishing.

230. Xaplanteri P, Lagoumintzis G, Dimitracopoulos G, Paliogianni F (2009) Synergistic regulation of pseudomonas aeruginosa-induced cytokine production in human monocytes by mannose receptor and tlr2. Eur J Immunol 39: 730-40.

231. Gautier G, Humbert M, Deauvieau F, Scuiller M, Hiscott J, et al. (2005) A type i interferon autocrine-paracrine loop is involved in toll-like receptor-induced interleukin-12p70 secretion by dendritic cells. J Exp Med 201: 1435-46.

232. Zeuthen LH, Fink LN, Frøkiaer H (2008) Toll-like receptor 2 and nucleotide-binding oligomerization domain-2 play divergent roles in the recognition of gut-derived lactobacilli and bifidobacteria in dendritic cells. Immunology 124: 489-502.

233. Itoh K, Watanabe A, Funami K, Seya T, Matsumoto M (2008) The clathrin-mediated endocytic pathway participates in dsrna-induced ifn-beta production. J Immunol 181: 5522-9.

234. Bult CJ, Eppig JT, Kadin JA, Richardson JE, Blake JA, et al. (2008) The mouse genome database (mgd): mouse biology and model systems. Nucleic Acids Res 36: D724-8.

235. Pfaffl MW (2001) A new mathematical model for relative quantification in real-time rt-pcr. Nucleic Acids Res 29: e45.

236. Edgar R, Domrachev M, Lash AE (2002) Gene expression omnibus: Ncbi gene expression and hybridization array data repository. Nucleic Acids Res 30: 207-10.

237. Suh HS, Zhao ML, Rivieccio M, Choi S, Connolly E, et al. (2007) Astrocyte indoleamine 2,3-dioxygenase is induced by the tlr3 ligand poly(i:c): mechanism of induction and role in antiviral response. J Virol 81: 9838-50.

238. Fensterl V, White CL, Yamashita M, Sen GC (2008) Novel characteristics of the function and induction of murine p56 family proteins. J Virol 82: 11045-53.

239. Sen GC, Lu L, Fensterl V, White C, Yamashita M, et al. (2008) Sy-1 induction, functions and viral evasion of the isg56 family of genes. Cytokine 43.

240. Lu J, O'Hara EB, Trieselmann BA, Romano PR, Dever TE (1999) The interferon-induced double-stranded rna-activated protein kinase pkr will phosphorylate serine, threonine, or tyrosine at residue 51 in eukaryotic initiation factor 2alpha. J Biol Chem 274: 32198-203.

241. Zhou A, Hassel BA, Silverman RH (1993) Expression cloning of 2-5a-dependent rnaase: a uniquely regulated mediator of interferon action. Cell 72: 753-65.

242. Stranden AM, Staeheli P, Pavlovic J (1993) Function of the mouse mx1 protein is inhibited by overexpression of the pb2 protein of influenza virus. Virology 197: 642-51.

243. Sing A, Merlin T, Knopf HP, Nielsen PJ, Loppnow H, et al. (2000) Bacterial induction of beta interferon in mice is a function of the lipopolysaccharide component. Infect Immun 68: 1600-7.

244. Trumpfheller C, Caskey M, Nchinda G, Longhi MP, Mizenina O, et al. (2008) The microbial mimic poly ic induces durable and protective cd4+ t cell immunity together with a dendritic cell targeted vaccine. Proc Natl Acad Sci U S A 105: 2574-9.

245. Weck MM, Grünebach F, Werth D, Sinzger C, Bringmann A, et al. (2007) Tlr ligands differentially affect uptake and presentation of cellular antigens. Blood 109: 3890-4.

246. Warshakoon HJ, Hood JD, Kimbrell MR, Malladi S, Wu WY, et al. (2009) Potential adjuvantic properties of innate immune stimuli. Hum Vaccin 5: 381-94.

247. Venter JC, Remington K, Heidelberg JF, Halpern AL, Rusch D, et al. (2004) Environmental genome shotgun sequencing of the sargasso sea. Science 304: 66-74.

248. MacLean D, Jones JDG, Studholme DJ (2009) Application of 'next-generation' sequencing technologies to microbial genetics. Nat Rev Microbiol 7: 287-96.