Volume 2 | Issue 1 | 31 May 2012 | ISSN 2182-0287

# Journal of Integrated OMICS

a methodological journal

# JIOMICS

## Journal of Integrated OMICS

### Focus and Scope

Journal of Integrated OMICS, JIOMICS, provides a forum for the publication of original research papers, preliminary communications, technical notes and critical reviews in all branches of pure and applied "-omics", such as genomics, proteomics, lipidomics, metabolomics or metallomics. The manuscripts must address methodological development. Contributions are evaluated based on established guidelines, including the fundamental nature of the study, scientific novelty, and substantial improvement or advantage over existing technology or method. Original research papers on fundamental studies, and novel sensor and instrumentation development, are especially encouraged. It is expected that improvements will also be demonstrated within the context of (or with regard to) a specific biological question; ability to promote the analysis of molecular mechanisms is of particular interest. Novel or improved applications in areas such as clinical, medicinal and biological chemistry, environmental analysis, pharmacology and materials science and engineering are welcome.

## Associated editors

Departamento de Química y Ciencia de losMateriales, Facultad de Ciencias Experimentales, Universidad de Huelva, Spain

**Theodore Alexandrov**
University of Bremen, Center for Industrial Mathematics, Germany

**Thole Züchner**
Ultrasensitive Protein Detection Unit, Leipzig University, Center for Biotechnology and Biomedicine, Institute of Bioanalytical Chemistry, Germany

**Tiziana Bonaldi**
Department of Experimental Oncology, European Institute of Oncology, Via Adamello 16, 20139 Milan, Italy

**Tsangaris George**
Proteomics Research Unit, Center of Basic Research II Foundation of Biomedical Research of the Academy of Athens, Athens, Greece

**ÜnerKolukisaoglu**
Center for Plant Molecular Biology, EberhardKarls University Tübingen, Tübingen, Germany

**Valeria Bertagnolo**
Department of Morphology and Embryology Universityof Ferrara, Italy

**Vera Muccilli**
DipartimentodiScienzeChimiche, UniversitàdiCatania, V.le A. Doria 6, 95126 Catania, Italy

**Vicenta Martínez-Zorzano**
Department of Biochemistry, Genetics and Immunology, University of Vigo, Spain

**Virginie Brun**
French Atomic Energy Commission and *French National Institute* for *Health* and Medical Research, France

**Vladislav Khrustalev**
Department of General Chemistry, Belarussian, State Medical University, Dzerzinskogo, 83 220029, Communisticheskaya 7-24, Minsk, Belarus

## SOUTH AMERICA

**Deborah Schechtman**
Department of Biochemistry, Chemistry Institute, University of São Paulo, Brazil

**Fabio Ribeiro Cerqueira**
Department of Informatics and NuBio (Research Group for Bioinformatics), University of Vicosa, Brazil

**Luis Pacheco**
Institute of Health Sciences, Federal University of Bahia
Salvador, Brazil

**Mário HiroyukiHirata**
Laboratório de Biologia Molecular Aplicado ao Diagnóstico, Departamento de Análises Clínicas e Toxicológicas, Faculdade de Ciências Farmacêuticas, Universidade de São, Paulo ,Brazil

**Jorg Kobarg**
Centro Nacional de Pesquisa em Energia e Materiais, Laboratório Nacional de Biociências, Brazil

**Marcelo Bento Soares**
Cancer Biology and Epigenomics Program, Children's Memorial Research Center, Professor of Pediatrics, Northwestern University's Feinberg School of Medicine

**Rinaldo Wellerson Pereira**
Programa de Pós Graduação em Ciências Genômicas e Biotecnologia
Universidade Católica de Brasília, Brazil

**Roberto Bobadilla**
BioSigma S.A., Santiago de Chile, Chile

**Rossana Arroyo**
Department of Infectomic and Molecular Biology, Center of Research and Advanced Studies of the National, Polytechnical Institute (CINVESTAV-IPN), Mexico City, Mexico

**Rubem Menna Barreto**
Laboratorio de Biología Celular, Instituto Oswaldo Cruz, FundaçaoOswaldo Cruz, Rio de Janeiro, Brazil

**Vasco Azevedo**
BiologicalSciencesInstitute, Federal Universityof Minas Gerais, Brazil

## NORTH AMERICA

**Adam Vigil**
University of California, Irvine, USA

**Akeel Baig**
Hoffmann-La Roche Limited, Pharma Research Toronto, Toronto, Ontario, Canada

**Alexander Statnikov**
Center for Health Informatics and Bioinformatics, New York University School of Medicine, New York

**Anthony Gramolini**
Department of Physiology, Faculty of Medicine, University of Toronto, Canada

**Amrita Cheema**
Georgetown Lombardi Comprehensive Cancer Center, USA

**Anas Abdel Rahman**
Department of Chemistry, Memorial University of Newfoundland
Newfoundland and Labrador St. John's, Canada

**Christoph Borcher**
Biochemistry & Microbiolog , University of Victoria, UVic Genome British Columbia Proteomics Centre, Canada

**David Gibson**
University of Colorado Denver, Anschutz Medical Campus, Division of Endocrinology, Metabolism and Diabetes, Aurora, USA

**Eustache Paramithiotis**
CaprionProteomicsInc., Montreal, Canada

**Fouad Daayf**
Department of Plant Science, University of Manitoba, Winnipeg, Manitoba, Canada

**Haitao Lu**
Washington University School of Medicine, Saint Louis, USA

**Hsiao-Ching Liu**
232D Polk Hall , Department of Animal Science, North Carolina State University Raleigh, USA

**Hui Zhang**
Johns Hopkins University, MD, USA

**Ing-Feng Chang**
Institute of Plant Biology, National Taiwan University, Taipei, Taiwan

**Irwin Kurland**
Albert Einstein College of Medicine, Associate Professor, Dept of Medicine, USA

**Jagjit Yadav**
Microbial Pathogenesis and Toxicogenomics, Laboratory, Environmental Genetics and Molecular, Toxicology Division, Department of Environmental Health, University of Cincinnati College of Medicine, Cincinnati, Ohio, USA

**Jianbo Yao**
Division of Animal and Nutritional Sciences, USA

**Jiaxu Li**

JOURNAL OF INTEGRATED OMICS

*A methodological Journal*

# CONTENTS OF VOLUME 2 | ISSUE 1 | MAY 2012

## REVIEW ARTICLES



## ORIGINAL ARTICLES

JOURNAL OF INTEGRATED OMICS
A METHODOLOGICAL JOURNAL
HTTP://WWW.JIOMICS.COM

REVIEW ARTICLE | DOI: 10.5584/jiomics.v2i1.76

# Infant gut microbial colonization and health: recent findings from metagenomics studies

Adrien Fischer*[1], Katrine Whiteson[1], Vladimir Lazarevic[1], Jonathan Hibbs[1], Patrice Francois[1], Jacques Schrenzel[1].

[1]Genomic Research Laboratory, Department of Internal Medicine, Geneva University Hospitals, Rue Gabrielle-Perret-Gentil 4, 1211 Geneva 14, Switzerland.

**Received:** 24 October 2011 **Accepted:** 11 April 2012 **Available Online**: 13 April 2012

ABSTRACT

New DNA sequencing technologies have emerged in the last decade enabling in-depth study of human gut microbiota. The bacterial communities inhabiting the gut influence our immune development and maturation with consequences for general health. However, the balance between host and bacterial community is affected by changes in lifestyle. Increasing rates of caesarean delivery, formula-feeding, antibiotic treatments, high fat diet, urbanization and hygiene have led to important changes in the colonization of the gut microbiota. Emergent diseases and conditions including asthma, allergies, autoimmunity, necrotizing enterocolitis (NEC), obesity and type I diabetes may be related to modifications in the microbiota. In this review we focus on studies related to early bacterial colonization of the gut, and how the evolution of gut microbiota during the first years of life may lead to new perspectives on the treatment of these diseases. Diet complementation with pre- or probiotics in formula or replacement of a disease associated-microbiota with a healthy one are currently the most studied approaches in the treatment of microbiota-related disorders. Bacteriophages may provide an alternative means for manipulating gut bacterial communities. However, the question is whether we can alter infant gut microbiota without any risk to health. High-throughput sequencing (HTS) techniques give access to the composition of the gut microbiome, and its evolution over time or in response to different circumstances. This review discusses these techniques, evaluates the impact of microbiome composition on infant development and outlines possible improvements in health care based on this knowledge.

**Keywords:** Microbiota; Infant; Gut; Delivery; Feeding; Probiotic; Prebiotic.

## 1. Introduction

The last decades have given rise to new genomic approaches for the study of uncultivated cells and viruses. Micro-organism community structure and diversity from inert samples (soil or sea) [1-3] or associated with humans and other eukaryotic organisms can now be studied [4-6]. Jo Handelsman originally described this as metagenomics, the study of all the microbial genomes in a community (metagenome) including different species and kingdoms (i.e. bacteria and viruses).

This culture-independent approach is now widely used to study the composition of human microbiome and viriome and evaluate their contribution to health. Bacterial cells in the human gut are ten times more abundant than the host's somatic cells [7,8] and are in a confined space. Studying the evolution of bacterial populations in this context with high resolution methods will supplement classical techniques which are probably able to identify less than 50% of bacterial species present [9]. The greatest part of the human microbiome corresponds to the bacterial community in the distal part of the gut, while the bacterial load at other anatomical sites (i.e. skin and mucosae) is relatively low [10]. The microbiota can shape the development of human gut epithelial cells and contribute to our digestion by their ability to synthesize enzymes to hydrolyze otherwise indigestible oligosaccharides [11,12]. Gut microbiota play an important role in the maturation of the intestine [13-17] and act as a barrier against pathogens. They also influence host metabolism (drug, amino-acid, lipid and anti-oxidant),

***Corresponding author:** Adrien FISCHER. Genomic Research Laboratory, Department of Internal Medicine, Geneva University Hospitals, 1211 Geneva 14, Switzerland. Phone (+4122) 372 9818, Fax (+4122) 372 9830, Email Address: adrien.fischer@genomic.ch

nutrient absorption as well as immune development and response [18,19,15]. There is a complex relationship between gut microbiome, environment, diet and even genetic predisposition to specific diseases, especially auto-immune disorders [20].

In the 20th century, advances in agriculture, transportation, urbanization and medical care, among other factors, have changed our diet and lifestyle significantly. These modifications may have a deep impact on our gut microbiota composition [21] that may be related to the emergence of new diseases like asthma, obesity, autoimmune disorders or inflammatory bowel diseases in adults and children [22-27]. Changes in microbial community structure have been observed in many human diseases including type I diabetes and inflammatory bowel diseases [23,28-34].

Human gut colonization begins at birth with microbes and viruses found in the environment and present in maternal skin or vagina, depending on the delivery mode. During the first two years of life, our microbiome and viriome change and evolve to a community different from that found at birth [35]. A long list of important questions is being addressed in this quickly growing field by researchers around the world. Understanding bacterial colonization of the infant gut as a function of environment, genetic composition or phage community may help us avoid and manage diseases not only in infancy, but throughout life. Will this new genomic approach lead to new treatments? Are gut microbiota associated with disease as cause, consequence, or both? Does the virus population in an infant's gut play a role in the emergence of new diseases? Could manipulation of the gut microbiota prevent, treat, or worsen diseases of infants, older children or adults?

## 2. Tools for physical characterization of microbial diversity

Studying variation in the human gut microbiota is currently possible with a range of techniques which are decreasing in cost. Many of these techniques have been used by ecologists studying soil or ocean microbial communities before the explosion of interest in human microbial communities.

### 2.1 *Metabolites based method*

Most methods use nucleic acids as their substrates. It is possible, however, to sample microbial populations using other molecular substrates.

#### *Mass spectrometry*

Mass spectrometry (MS) can directly detect the chemical products of microbial metabolism. For example, MS shows that the gut microbiome has an impact on mammalian blood metabolites [36], highlighting the influence of the microbiome on the drug processing capacity of the host. This method is not yet widely used to study commensal bacteria. Introducing MS complements the phenotypic information derived from nucleic acid-based methods, and may help us to learn more about bacterial metabolism. Because it requires no prior assumptions about what molecular components will be present, mass spectrometry is an open-ended technique and allows potential detection of previously unknown organisms.

### 2.2 *Nucleic acid based methods*

Variations in the 16S rRNA sequence allow phylogenetic classification of bacterial and archeal species [37]. The 16S rRNA gene is ubiquitous among prokaryotes, and has become the main genetic and phylogenetic marker to characterize and compare bacterial communities using classic techniques or high-throughput sequencing (HTS).

Polymerase chain reaction (PCR)-based techniques have been used to characterize prokaryotic phylogeny since the 1980's [38,39]. More recently, PCR has been used to complement other methods like fluorescence in situ hybridization [40] for the study of microorganism variation. In addition to the 16S based approach, bacterial communities may be investigated analyzing random metagenomic fragments [41].

#### *Restriction Fragment Length Polymorphism*

In the 1990's PCR with fluorescent primers was used to amplify ribosomal DNA [42]. The resulting amplicons were digested with restriction enzymes, and the resulting fragments obtained were separated electrophoretically. This method, called Terminal Restriction Fragment Length Polymorphism (T-RFLP) has been used to determine the microbial composition of samples from ocean crust [43], cystic-fibrosis affected lungs [44], and many other sources. The resolution of T-RFLP is rather moderate, since the number of restriction sites in the 16S rRNA genes is limited.

#### *Automated ribosomal intergenic spacer analysis*

A PCR based method relying on the length heterogeneity in the intergenic region spacer of the 16S-23S ribosomal RNA was developed by Fisher and Triplett in 1999 [45]. This method, ARISA (Automated Ribosomal Intergenic Spacer Analysis), allows for the comparison of bacterial communities and provides an estimation of microbial richness and diversity from many kinds of samples including soil, marine (ocean) or human gut [46-50]. ARISA was used in concert with other classical analysis tools to assess a correlation between gut microbiome and diabetes in rats [51]. This method is cheaper than most of the others and allows a rapid estimation of bacterial composition, but has some limits because bacterial 16S and 23S rRNA genes are not always organized as an operon; *in silico* modeling shows a loss of linearity when the species richness increases [52].

The utilization of complementary methods, such as statistics [53] or sequencing [54] could help to correct this bias.

*Fluorescence in situ hybridization*

Fluorescence in situ hybridization (FISH) reveals the location of specific DNA sequences after hybridization with specific fluorescently labeled oligonucleotide probes [55,56]. It was used to compare different bacterial communities e.g. the gut microbiota of 6 month old formula-fed or breast-fed babies. This method allowed Rinne and colleagues to identify differences related to a decrease of *Bifidobacterium* in formula-fed infants gut microbiota [40]. FISH has the advantage of observing specific members of the bacterial community directly in the sample, for instance, within buccal epithelial cells [57,58]. The basic limitation of the method is that it only detects taxa covered by the probes. FISH is an example of a closed-ended diagnostic technique, like T-RFLP or ARISA, while metagenomics includes the study of unknown bacterial communities, which can only be revealed through more open-ended methods, assuming little prior knowledge of the bacteria under study.

*Microarray*

A microarray consists of a huge number of single-strand DNA molecules spotted on glass slides. Hybridization with labeled cDNA with known taxonomy allows for identification of bacterial species, differences between strains and the presence or absence of genes of interest. Microarray assays may be used in comparison with high-throughput sequencing. For example, microarray analysis revealed a dramatic shift in the gut viral population between 1 and 2 weeks of life [59], while high-throughput sequencing detected some stable members in viral community until 3 months of age [59]. Another phenotypic microarray analysis of bacterial communities in patients suffering noma disease revealed no increase in *Fusobacterium necrophorum* compared to controls, exonerating one suspected etiological agent [60,61].

Using microarrays alone may miss important unpredicted or unknown members of a bacterial community. This is because microarrays in general are another closed-ended technique, designed in advance with known sequences. This approach, which assumes knowledge of all sequences of interest before the experiment, is considered to be more efficient than using random sequences on the array. It introduces a bias against previously unknown organisms, but it does allow for screening of many known sequences, enough to identify most Operational Taxonomic Units (OTUs) known at the time the experiment is done. Samples obtained from two separate samples, processed similarly with identical microarrays, may provide information regarding the differences in their microbiota profiles.

*High-Throughput sequencing methods*

Recently, the Roche/454 and Illumina/Solexa HTS proved their potential in analyzing microbial communities [62,63,41]. Currently, 454 pyrosequencing produces reads close to $10^3$ bases, whereas the length of Illumina reads is limited to $10^2$ bases. However, the number of sequence reads generated by Illumina is, for the same cost, about 150x that produced using the Roche/454 platform. HTS-based study of microbiota may include random sequencing of metagenomic fragments or sequencing of 16S rDNA amplicon libraries [62,63,41]. 16S rRNA contains very stable as well as highly variable regions, allowing for phylogenetic comparison. The HTS of partial 16S rRNA genes of a microbial community generates a large amount of classifiable sequences. All target sequences, however, are not amplified with the same efficiency using "universal" 16S primers, the copy number is not stable across organisms (which complicates abundance analysis), and the classification of 16S rDNA sequences is typically limited to the genus level [64]. When genomic DNA from the community is randomly sequenced, a higher fraction of sequences remain taxonomically unassigned because of the lack of homology in sequence databases. Nevertheless, the taxonomy of sequences corresponding to different groups of organisms, including bacteria, archaea and fungi as well as viruses may be assessed simultaneously.

While 10 sequences per sample may be enough to identify differences between bacterial communities [28], it is necessary to increase the number of sequences per sample to identify underrepresented taxa. In some cases, even after obtaining one million sequences from a single sample, further sequencing will still lead to the discovery of new phylotypes [65]. HTS has aspects of a closed-ended system since it assumes all the target organisms have conserved elements (of the 16s gene, for example). It allows considerable variability within those limits, however (including the detection of new 16s variable sequences pertaining to a previously unknown species, such as the discovery of *Trophyrema whippelii* [66]). In this sense it is a partially open-ended system, allowing for some discovery of unexpected organisms that fit fairly broad pre-experimental assumptions. The possibility of sequencing the complete genome of bacteria without any prior knowledge of their sequences aside from the 16s primers confirmed the open-ended aspects of HTS.

Although the HTS culture-independent approach could out-compete other existing diagnostic and typing methods in microbiology, questions remain about the cost and best strategy for analysis of the massive datasets generated by this technology.

*Analytical strategies for HTS data*

Data analysis is one of the most labor intensive aspects of metagenomic projects. Computing advances do not keep pace with constant huge increases in the data generated by high-throughput sequencing. This leads to problems with storage space and computer memory as well as with the

selection of consistent and reproducible analytic algorithms [67].

A typical HTS data set must be cleaned following predefined parameters for the study. For example, sequences which are too short for a given technology contain ambiguous characters, or those that do not carry the primer sequence are removed. Sequences containing putative amplification and sequencing errors are also removed. Remaining sequences are then clustered into Operational Taxonomic Units (OTU or phylotypes) with varying percentages of sequence identity [68]. Currently 97% identity is often used to represent the species level cut-off. In many cases a representative sequence is chosen from each OTU, by largest abundance or even a randomly chosen sequence, and this is used to build a phylogenetic tree. The phylogenetic diversity in the tree can be confirmed by BLASTing all the reads in a parallel analysis. Finally, the taxonomy can be assigned to each OTU with publicly available programs such as the RDP classifier [69,70]. Much can be learned about the composition of the microbial community in a particular sample by simply looking at the abundance and identity of the top ten or more OTUs; doing this at different identity cut-offs is also revealing.

Within sample (alpha) and between sample (beta) diversity may be estimated from the abundance of 16S rDNA sequences [71]. Although different studies with different approaches will each have their own unique bias, comparisons of the communities found in different samples with the same methods are likely to reveal biologically relevant differences. Branch length-based phylogenetic diversity comparisons using UniFrac [72], or nearest taxon index (NTI) [73] are widely used. UniFrac metric, that calculates the fraction of the branch length shared by two communities in their common phylogenetic tree, can be used to estimate beta diversity. For more than two communities, a distance matrix that relates each pair of communities may be subject to hierarchical clustering and Principal Component Analysis [74,75]. ANOSIM (Analysis of Similarity) [76] is a powerful way of testing the similarity of community composition of samples with a shared condition against a null hypothesis that they are randomly distributed. PERMANOVA [77,76] determines which conditions are associated with community differences. Powerful software is emerging to accommodate these types of analysis, including MG-RAST [78], QIIME [79], mothur [80], RDP [69], CARMA [81], CAMERA [82], IMG/M [83], PRIMER [76] and PC-ORD (MjM Software Design, Gleneden Beach, OR).

## 3. Using metagenomic tools to study development of the human gut microbiome during infancy

Colonization of an infant's gut begins just after birth. This initial colonization may be important as it could establish long term microbial communities. Gut microbiota are highly variable and poorly diversified during the first two years of life before the more stable and diversified adult microbiota is established, often around the time of weaning [84,85]. Recent research suggests the importance of the first colonization (bacterial or viral), which is linked to the delivery mode, the environment, food source and genetic background in infant's health [86-95].

### 3.1 *Initial colonization*

During the first month of life, *Bifidobacterium* are predominant in the gut of almost all infants [96]. A recent study on gut microbiota from vaginally delivered and at least partially breast-fed infants that were never exposed to surgical intervention, antibiotics, pre- or probiotics, provided insight into "normal" bacterial colonization and evolution of the gut microbiota during the first 4 months of life [97]. The study, partially represented in table 1, identified a large number of bacteria and revealed the presence of *Staphylococcus*, *Bifidobacterium* and γ-Proteobacteria in most 4 days old infants' gut. The authors highlighted that the evolution of gut microbiota includes shifts in both composition and structure. From day 4 to 120, *Lactobacillus 2a*, *Veillonella*, *Lachnospiraceae2* and *Bifidobacterium 1* populations increased whereas *Staphylococcus*, *Streptococcus* and an uncultured bacterium population decreased. Some strains could cause severe diseases only in particular contexts. For example, if low-level toxin-producing staphylococcal strains are supplanted by a high-level toxin producer, the gut microbial community might not appear to change but the baby may become ill. Moreover, the use of antibiotics may have a drastic effect on development of gut microbiota and health especially if infants are not breast-fed [88,98-100].

### 3.2 *Delivery Mode*

#### *3.2.1 Vaginal vs. C-section delivery*

The most important first source of inocula after birth is the mother's vaginal and fecal microbiota [92,93] from the birth canal. This ecosystem contains a limited number of bacterial taxa [101,102]. Depending on the delivery mode, these first inocula could be completely different (Table 1). Indeed, if babies are delivered vaginally, their gut microbiota just after birth (before 24 hours of life) is similar to their mother's vaginal (and fecal) microbiota (sampled 1 hour before delivery) dominated by *Lactobacillus*, *Prevotella* or *Sneathia* spp [94]. In contrast, children delivered by C-section are colonized by species similar to those from their mother's skin microbiota such as *Staphylococcus*, *Corynebacterium* and *Propionibacterium* spp [94] or from the environment (equipment, air, other infants, nurses) [95]. *Staphylococcus aureus* (*S. aureus*) colonization is more common in the gut of infants delivered by C-section likely because of the presence of this species on their mothers' skin [103,104]. However, Eggesbo and colleagues [97] demonstrated

*Staphylococcus aureus* in 95% of healthy 4 days old vaginally delivered and breast-fed infants decreased after 120 days to a prevalence around 60% [97]. Further studies over longer time periods and larger samples are required to establish whether delivery mode affects *S. aureus* colonization.

Babies born by C-section have a microbiome similar to vaginally born infants whose mothers received antibiotics before normal delivery or during breast-feeding [95]. This may be a consequence of a shift in the mother's vaginal microbiota due to antibiotic treatment or a direct effect of antibiotics from mother's milk on the vaginally-delivered babies gut microbiota. Common use of antibiotics before C-section may also contribute to this observed similarity. The delivery mode has a major role in first colonization of the infant gut. The major question in this field is, however, whether delivery mode and its effect on microbial community composition affect the development of increasingly common disorders including asthma, allergy, auto-immune diseases and colitis?

### *3.2.2 Clostridium difficile and C-section*

*Clostridium difficile* (*C. difficile*) is an anaerobic spore-forming gram positive rod. Under some circumstances, it produces an exotoxin that kills epithelial cells. In extreme cases, enteric infection with *C. difficile* causes inflammatory diarrhea, toxic megacolon and death. The organism can be recovered from the stool of healthy individuals [106], however, exposure to antibiotics alters gut flora in a way that favors overgrowth of *C. difficile,* secretion of its toxin, and disease [107]. Clindamycin, an anti-anaerobic agent to which *Clostridium difficile* is resistant, is most strongly associated with *C. difficile* diarrhea, but risk also increases even with short courses of many other antibiotics, such as pre-surgical prophylaxis with cefazolin [108,109].

A correlation between *C. difficile* colonization and delivery mode was not found in a 1986 study in Italy [110]. Thirteen percent of the newborns were colonized with *C. difficile* and the colonization rate was higher but insignificant in caesarean than in vaginally delivered infants [110]. However, according to Penders *et al*, infants born by caesarean are more likely to be colonized with *C. difficile* [88]. One study involving 1032 one month old infants found that the prevalence of *C. difficile* in infants' gut was 42% for C-section delivered, 26% for vaginally-delivered in the hospital, and only 19% for home delivered babies [88]. A recent paper from the Chicago area highlighted the increased prevalence of *C. difficile* infections in C-section delivered infants with an incidence of 2.2 / 1000 births while only 0.2 / 1000 of

| | | Bacteroidetes | | Actinobacteria | | | | Firmicutes | | | | | | | Fusobacteria | Proteobacteria |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Microbiota | *Bacteroides* | *Prevotella* | *Bifidobacterium* | *Coriobacterineae / Atopobium* | *Propionibacterium* | *Corynebacterium* | *Lactobacillus* | *Streptococcus* | *Staphylococcus* | *Clostridium difficile* | *Clostridiumcoccoides* | *Veillonella* | *Lachnospiraceae* | *Sneathiaspp* | γ-proteobacteria |
| Mother | Vagina | + | + | | + | | | + | + | | | | | | + | |
| | Skin + Environment | | | | | + | + | | | + | | | | | | |
| Infants, birth | Caesarean | | | | | + | + | | + | + | + | + | | | | |
| | Vaginal delivery | + | + | + | + | | | + | | | | | | | + | |
| | Prematurity | | | | | | | | | +* | + | | | | | |
| | 4 days of life ** | + | | + | + | | | + | + | + | | | + | + | | + |

**Table 1.** ***Infant gut bacterial population depends on the delivery mode.*** *: After 8 days of life; **: Babies vaginally delivered, breast-fed and without any treatments (babies and mother) before the study; +: predominant. This Table is based on the results presented in the following studies: Dominguez-Bello *et al* [94]; Fallani *et al* [95]; Eggesbo *et al* [97]; Morowitz *et al* [105]; Hyman *et al* [101]; Penders *et al* [88].

vaginally delivered new born were infected [111]. It is possible that the different findings reflect relatively higher antibiotic use with caesarean delivery in the U.S. and Italy; higher antibiotic use would be expected to promote more *C. difficile*. However, 60% to 70% of infants guts (vaginally or C-section delivered in Boston) are asymptomatically colonized with *C. difficile* [112]. Asymptomatic *C. difficile* colonization correlates with high levels of Firmicutes (also seen in C-section born infants, except for *Staphylococcus*) or *Bacteroides* but it is clearly not correlated with the presence of *Bifidobacterium* [95,113,107]. *Clostridium difficile*-associated disease correlates with the absence of Bacteroidetes [114]. Usually, *C. difficile* disappears in children between 1 and 2 years old. The gut microbiota of infants, characterized by a low diversity, might not provide an efficient protection against *C. difficile* infection. The absence of some bacteria may allow unusual *C. difficile* colonization and overgrowth leading to disease. Table 2 and Khoruts' study (described below) support the idea that *Bacteroides* colonization decreases the chances of *C. difficile* pathogenicity. Therefore, decreased *Bacteroides* and *Bifidobacterium* populations in infant gut, due to the use of antibiotics [88] may increase the risk of *C. difficile* infection. In older patients *C. difficile*-associated disease is related to a decrease in overall bacterial populations [34]. Rousseau *et al* therefore suggested that *C. difficile* colonization results in a microbial community disorder of the gut. Alternatively, disturbed microbial populations may predispose to overgrowth of *C. difficile*. Varma and colleagues observed that the supernatant from *Lactobacillus fermentum* culture alone is enough to inhibit *S. aureus* and *Pseudomonas aeruginosa* growth *in vitro*, suggesting that compounds released by *Lactobacillus* inhibit pathogen growth and may be used to treat hospital-acquired infections [115]. Whether competition from particular bacteria or by inhibitory compounds released during their growth affects bacterial colonization remains to be clarified.

The association between disease-causing *C. difficile* and C-section, found in some studies [88,111] was not confirmed in the others [110,112]. One study pointed to a shift or a decrease in overall bacterial populations in *C. difficile*-associated disease. Therefore, it appears that *C. difficile* may be a member of the gut microbiota of healthy infants and that illness might be triggered by factor which causes a shift in the gut microbiome.

Initial microbial colonization in infants may influence health in later life. The gut microbiome undergoes several drastic shifts in composition during the first two years of life, and later becomes more stable [116,117,85]. Bacterial taxa differ in their ability to persist in the infant gut [97]. The forces that shape the formation of bacterial communities in the gut, including the immune system and the consequences of random environmental variables, are the subject of intense study.

### 3.3 *Feeding mode*

#### *3.3.1 Lifestyle and geography*

During the last decades, our life style has changed drastically. Nowadays, more people live in cities than in rural areas. We use antibiotics, eat heavily processed food and pay more attention to hygiene. All these changes have an impact on our microbiota and thus on our health. Carlotta de Filippo and colleagues have compared the gut microbiome of African children having a life style close to the time of the birth of agriculture to the gut microbiome from European children [22]. They highlighted a diet richer in fiber from complex grains for African children, and noted an enrichment of Bacteroidetes and a depletion of Firmicutes among African compared with European children. The effect of dietary fibers on the relative abundance of Bacteroidetes and Firmicutes was recently confirmed in the ob/ob mouse model [118]. Gut microbiota of African children are enriched in bacteria containing genes for cellulose or xylose hydrolysis, as well as for short-chain fatty acid metabolism.

#### *3.3.2 Breast-feeding*

Human milk and thus breast-feeding affects growth, immune system development, cognitive development, as well as susceptibility to toxins and immune diseases like asthma and allergy [119-125]. At least some of these effects may be mediated by the effect of breast-feeding on gut microbiome.

Differences in breast-feeding practice influence the observed microbiota differences in the gut between human subpopulations [22,95]. In the De Filippo's study [22], the microbial communities of breast-fed African and Italian children were quite similar, but diverged when breast-feeding was discontinued in the Italian infants. A study of infants in Granada and Stockholm found that the Granada infants had lower carriage of *Bifidobacteria* (19 vs. 60%) and higher carriage of *Bacteroides* (21% vs. 6%), which the

**Table 2. *C. difficile* role in infants' gut colonization.** -: unrelated; +: predominant; ND: no data. This table summarizes the results from the following studies: Rousseau *et al* [107]; Khoruts *et al* [114]; Fallani *et al* [95]; Collignon *et al* [113]; Penders *et al* [88].

| *Clostridium difficile* | Mode of delivery | NEC | Diarrhea | Firmicutes | Veillonella | Lachnospiraceae | *Bifidobacterium* | *Bacteroides* |
|---|---|---|---|---|---|---|---|---|
| Asymptomatic | ND | | | + | | | - | + |
| Associated disease | C-section | + | + | + | + | + | | - |

authors attributed to the lower rate of breast-feeding in Grenada (43% vs. 76%) [95]. On the other hand, limited evidence suggests that feeding mode is not required for *S. aureus* colonization. Sequencing the gut metagenome of an 8-day-old premature breast-fed neonate revealed a limited population of *S. aureus* [105]. *S. aureus* population grew significantly at 10 days of age, suggesting *S. aureus* reservoir was independent of feeding type [105]. In another case, contamination by a methicillin resistant *S. aureus* (MRSA) through mother's breast milk was reported [126], suggesting that the colonization by *S. aureus* may be in some cases related to the feeding mode. While term-delivered infant gut microbiota is less diverse than that of adults, it remains more diversified than premature infants gut microbiota [85,127,33]. Among low weight infants colonized by coagulase negative staphylococci (CoNS), breast-feeding decreases duration of persistent CoNS bacteremia [128]. Breast-feeding might be a good way to protect against *S. aureus* diseases. It must be conceded, however, that the apparent ecological relationship between breast-feeding and gut microbiota variations is highly susceptible to confounding factors, since children living in Western cities are not exposed to the same physical or social environment as children living in developing countries or in farms.

In addition to ecologic and demographic demonstrations, microbiologic studies of individuals also tend to support the influence of breast-feeding on the infant gut microbiota. Breast-feeding gives rise to a less diversified microbiota mainly composed of *Bifidobacterium* and *Lactobacillus* [40,129,130], while formula-feeding promotes *Bacteroides* and *Clostridium coccoides* gut colonization [95,131], resulting in a microbiota profile closer to that of adults. One obvious explanation for such differences, illustrated in the Figure 1, is the presence of abundant and complex oligosaccharides in human milk absent in formula. These oligosaccharides, indigestible by humans, are a major substrate for *Bifidobacterium* (*Bifidobacterium bifidum* and *Bifidobacterium longum* subsp. *infantis*) [132] and are necessary for the "normal" infant gut microbiota development [132]. They promote the growth of *Bifidobacterium* and in some cases of *Lactobacillus* [133]. As oligosaccharides from human milk are similar to some human cell surface carbohydrates, they play an important role in the protection against pathogenic bacteria by competing with their target ligands [134,135]. Oral supplementation with *Bifidobacteria* or *Lactobacillus* (probiotics) may have beneficial effects. For instance, the addition of *Lactobacillus* in the diet of infants, aged between 2 weeks and 13 years during the three months preceding antimicrobial therapy, decreased the incidence of diarrhea 2 weeks after the introduction of the antimicrobial treatment [136]. Administration of Lactobacillus in children from 2 months to 6 years old was associated with lower rotavirus diarrhea duration and, improved effect of parenteral rehydration [137]. Co-administration of *Lactobacillus* and *Bifidobacterium* was associated with reduced incidence of NEC in low-birth weight infants [138], and reduced diarrhea duration [139] as well as stool frequency [139]. However in some instances, the benefit from administration of *Lactobacillus* or *Bifidobacterium* was not observed, e.g. on the duration of non-rotaviral diarrhea [137], and the incidence of death, NEC or nosocomial infections in low birth-weight infants [140,141].

These observations raise the question of whether prebiotics should be added to formula. Prebiotics are selectively fermented ingredients that promote growth and/or activity of the gut bacteria and confer benefits to the host [142]. Several studies demonstrated that introducing prebiotics like galacto-oligosaccharides or fructo-oligosaccharides into formula shifted the gut microbiota to be more similar to breast-fed infants [129,130,40,143,131]. Moro and colleagues [143] observed a dose-related positive association between prebiotics in formula and lactobacilli number in infants gut but Fallani *et al* [95] found a similar abundance of lactobacilli in formula-fed without prebiotic supplementation and in breast-fed infants. The distribution of *Bifidobacterium* species in the fecal samples from standard and prebiotic-supplemented formula-fed infants resembled those of breast-fed infants and adults, respectively [129]. The infant's gut microbial community is thus highly malleable and sensitive to changes in the composition of the diet. The use of prebiotics should be approached with caution, because their long-term effects are currently unknown.

## 4. Role of intestinal flora in the emergence of new diseases

Intestinal colonization is the host's earliest contact with micro-organisms [147], with important consequences on maturation of the immune system and on metabolism [148]. In this section, we examine mechanisms by which intestinal flora might influence pathophysiology of diseases mediated by nutrient absorption and by the host immune system. This is a dynamic literature, and every week new studies demonstrate associations between altered microbiota and disease states. We note, however, that at this stage most published observations lack the resolution to disentangle cause and effect. In most cases we cannot yet determine whether altered microbiota cause disease or are actually a marker of underlying problems.

### 4.1 *Obesity and nutrition*

Gut microbiota may modulate the ability to access nutrients and therefore body weight and composition. Studies comparing germ free (GF) mice raised in sterile conditions with normal mice have found that they respond differently to the same high fat diet and suggested that GF mice can resist obesity. Compared to mice with normal gut flora, GF mice consumed fewer calories, increased lipid excretion and enhanced insulin sensitivity [149]. No differences were observed during a low fat diet between GF and conventional mice [150,151], suggesting a primary

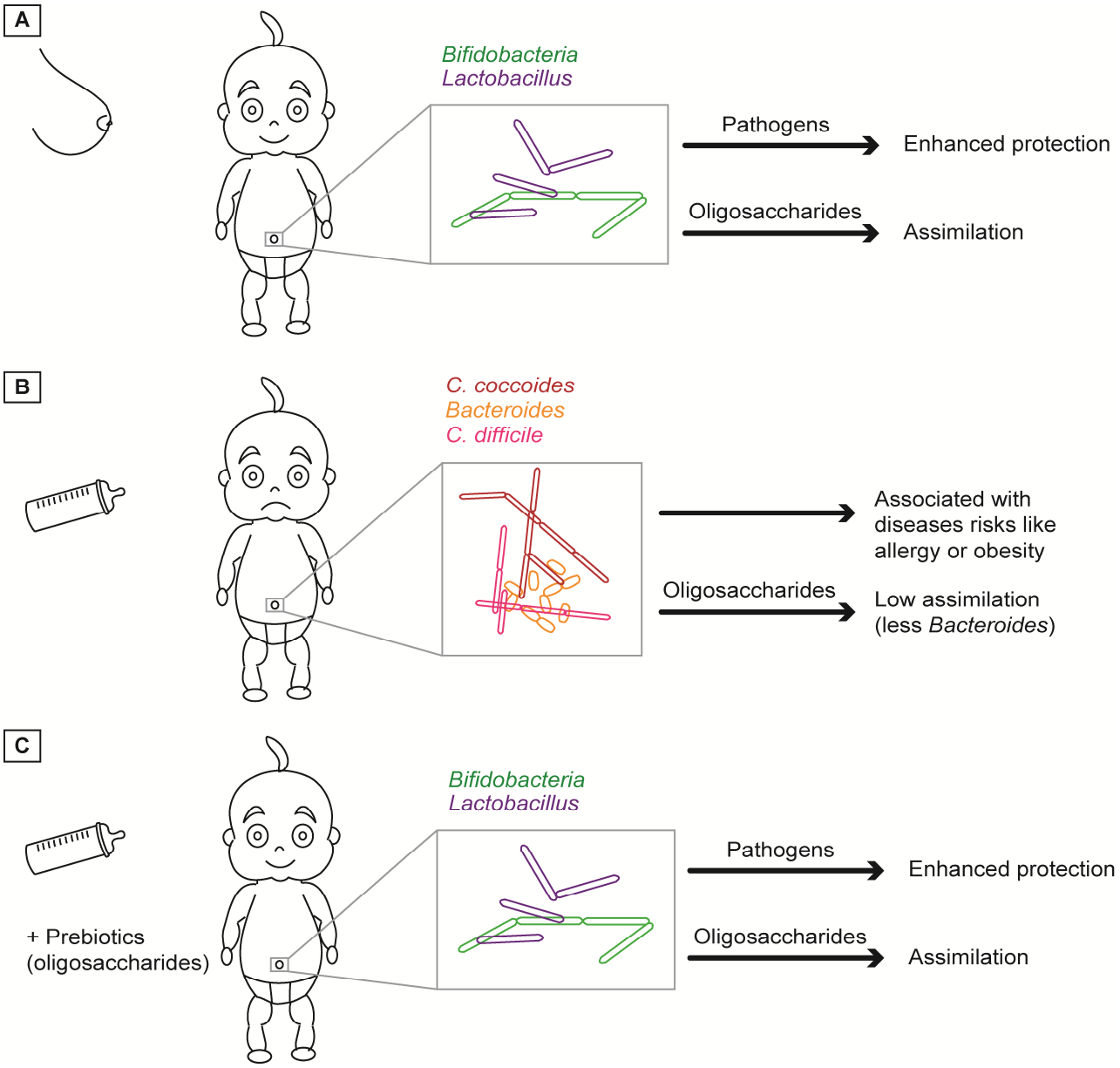


**Figure 1.** ***The importance of breast-feeding.*** Breast-fed children will develop a gut microbiota composed with *Bifidobacteria* and *Lactobacillus* among other species, enhancing protection against pathogens and the assimilation of oligosaccharides (A), while bottle-fed children will be mostly colonized with *Clostridium coccoides* or *difficile* and *Bacteroides* that does not allow oligosaccharide assimilation and are associated with the emergence of diseases (B). Complementation of formula-feeding with oligosaccharides, almost recovers the same microbiota as breast-fed children with its benefits (C). The species represented in this figure are predominant, not exclusive. Information to build figure 1 were found in the following studies: Harmsen *et al* [121]; Schack Nielsen *et al* [122,123]; Rinne *et al* [40]; Haarman *et al* [129,130]; Roberfroid *et al* [133]; Ninonuevo *et al* [144,145]; Zivkovic *et al* [132]; Moro *et al* [143]; Fallani *et al* [95]; Mira *et al* [21]; de Filippo *et al* [22]; Turnbaugh *et al* [146].

impact of the gut flora on absorption of fats.

Intestinal microbiota may also influence nutrition by their effect on gastrointestinal hormone production. *Helicobacter pylori* (*H. pylori*), an ancient and dominant commensal microbial inhabitant, regulates ghrelin and leptin production, two hormones involved in body weight regulation [152-155]. As *H. pylori* has become less prevalent in infants' commensal flora since the beginning of the early 1900s, these two hormones persist and may cause infants' obesity and type-2-diabetes.

Immunologically active receptors may be another means by which gut microbiota influence obesity. High levels of FIAF (fasting-induced adipose factor), AMPK (AMP-activated protein kinase regulating fatty acid oxidation) activation and a lack of ileal TNF-α induction have been observed in GF mice that remain lean despite a high fat diet

[156,150,151]. Lean GF mice treated with gut flora from obese mice become obese [157].

Whatever the mechanism, it seems clear that obesity is epidemiologically associated with the gut microbiome, and in particular with a predominance of Firmicutes over Bacteroidetes [23]. Among discordant twins, bacterial communities in the obese twins are less diverse [65].

### 4.2 *The enteric immune system*

Healthy development of the immune system has been associated with infants' gut microbiota colonization [158,159]. Complex interactions with commensal microbiota modulate the response of the immune system to self molecules, harmless and pathogenic microbes. Commensal microbiota is not essential for animal life [160,161], and its presence may even appear disadvantageous in some respects. For instance, GF mice have decreased pro-inflammatory T cell response (and therefore less pathology) in an induced encephalomyelitis model [162]. However, GF mice also show altered antibody production, lymph node structure and gut capillary and lymphoid tissue development [163-166]. Changes in interactions between the immune system and the microbiota may lead to the emergence of allergy, asthma, type I diabetes, obesity or gastric disorders [158,29,167,168,146]. Immune response is involved in this process as Crohn's disease is TNF-α and INF-γ dependent [20]. Pathology in apparently "autoimmune" diseases may require interaction between the host genome and specific microbiological exposure, as with the cytokine-mediated induction of a Crohn's disease-like illness in mice by a combination of a host mutation with a specific norovirus infection [20].

Migration of Mast cells from blood to the intestine might be promoted by the commensal microbiota as GF mice carry less Mast cells in their intestine [169]. Mast cells or Mastocytes are involved in allergic disease, but also in wound healing and protection against pathogens [170]. GF mice underexpress Keratinocyte-derived cytokine, lipopolysaccharide-induced CXC chemokine, and macrophage inflammatory protein 2, ligands for the receptor CXCR2 required for intestinal localization of Mast cells [169]. Oligosaccharides have a big impact on infant gut microbiota, and their similarity to pathogenic ligands allows them to compete with pathogenic bacteria to protect infants from infection [134,135]. Kunii *et al* [169] hypothesize that commensal bacteria in mice stimulate production of a ligand necessary for the intestinal recruitment of Mast cells.

### 4.3 *Are treatments available to regulate gut microbial community disorders?*

Even without a sophisticated understanding of how gut microbial community are assembled, acceptance of their importance allows for novel medical treatments of infectious diseases. *C. difficile* diarrhea is a paradigmatic example. Antibiotic treatments before surgery often lead to loss of "healthy bacteria" and susceptibility to opportunistic pathogens, including *C. difficile*. One such patient suffered chronic intestinal *C. difficile* associated disease, with diarrhea every 15 minutes, dramatic weight loss and confinement to a wheelchair [114,171]. The patient received a fecal transplant from her husband, and had a normal bowel movement the next day. Fecal transplants have been used successfully for this kind of infection in dozens of patients since the 1950s, but the treatment is far from accepted routinely. In this case, rapid and drastic changes in the patient's gut microbiota was followed using metagenomics, showing how the patient's gut bacteria came to resemble the healthy donor community, with a predominance of *Bacteroidetes* spp. strains, and the end of the patient's symptoms. These results showed that in the absence of healthy gut microbiota, pathogenic bacteria overgrew and caused life-threatening diarrhea. If the opportunistic strain is exterminated and its niches are quickly populated by "healthy microbiota", the pathogen will not be able to colonize the patient.

#### *4.3.1 Prebiotics/Probiotics as a solution?*

Many studies report the benefit of using pre or probiotics to treat infants' disease linked to gut microbiota modifications. Studies underway in Europe, the US (At the UC Davis Children's hospital http://ww.ucdmcucdavis.edu/children/pediatric_research) and South America (http:clinicaltrialsfeeds.org/clinical-trials/show/NCT00727363) all seek to understand initial infant colonization, and whether probiotics protect against NEC more powerfully than antibiotics as previously proposed [172-177]. In some hospitals, notably in Sweden, all premature babies are given probiotics to prevent gut disorders such as NEC, despite the lack of certainty described by Millar *et al* [178]. Figure 2 summarizes the involvement of some compounds in infants' intestinal diseases and makes a link with inflammatory immune reactions in the intestine. Although -biotics help to prevent or cure some diseases, like NEC or allergic rhinitis [179,180], their role remains to be clarified. Indeed, adding prebiotic to breast-fed infants lead to a different gut microbiota than those who were breast-fed exclusively [181].

##### *4.3.1.1 Necrotizing Enterocolitis*

It is not clear whether NEC has a specific agent or agents, or if so what those agents might be. *Enterobacter sakazakii*, found in infants' formula powder, is one candidate [193,194] while Morowitz and colleagues proposed *Citrobacter* strains [105]. However, they observed that healthy children also carry *Citrobacter* strains. In other studies NEC was related to *C. difficile*, γ-proteobacteria or decreased bacterial diversity in infants' gut [33,127]. This suggests that NEC is not caused by one particular bacterium. Moreover, Morowitz and colleagues observed increased *Citrobacter* phage counts, while others did not look at the phage population. The

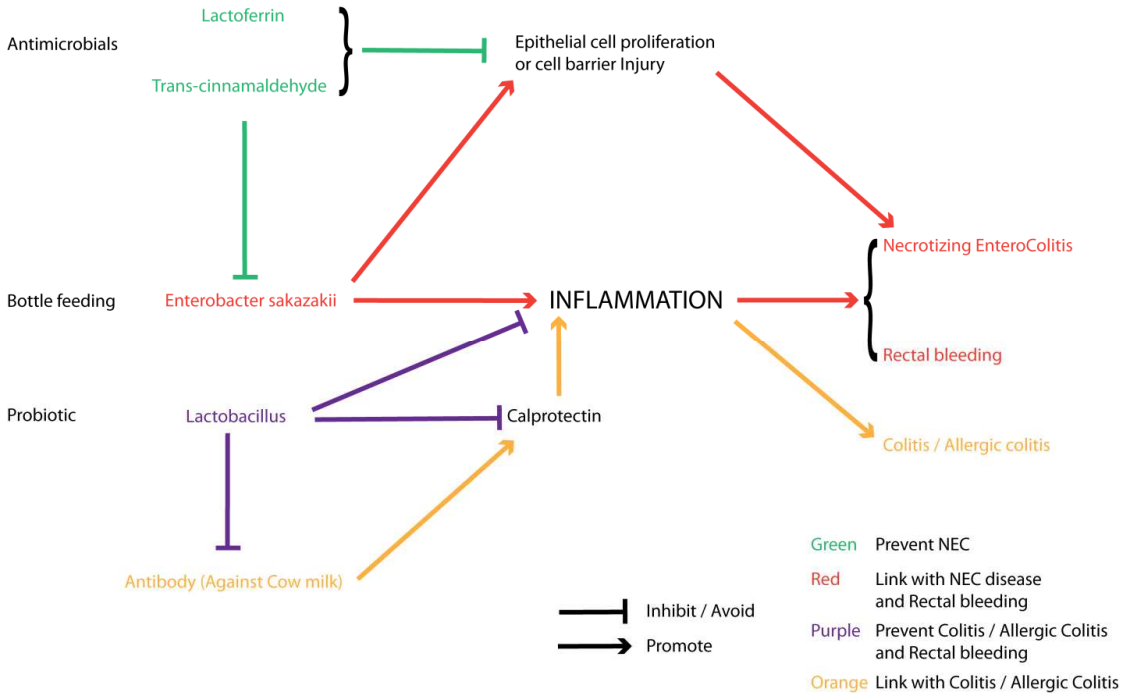


**Figure 2.** ***Prevention of enteric diseases in infants?*** Antimicrobials like Lactoferrin or Trans-cinnamaldehyde protects children from NEC or rectal bleeding, preventing infection with *Enterobacter sakazakii* found in formula that causes inflammation and NEC or rectal bleeding. Probiotics like *Lactobacillus* prevent inflammations as well as production of antibody responsible for the production of the Calprotectin enhancing inflammations responsible for allergic colitis. Results presented in figure 2 were described in the following studies: Venkatesh *et al* [182]; Oguchi *et al* [183]; Buccigrossi *et al* [184]; Amalaradjou *et al* [185]; Faber *et al* [186]; Abdelhamid *et al* [187]; Savino *et al* [181,188]; Baldassarre *et al* [189]; Nermes *et al* [190]; Voganatsi *et al* [191]; Arvola *et al* [192].

largest study of gut phage and viruses in humans did not find a tight link between viral population and regulation of bacterial population [195]. However, by analogy with aquatic environments [196,197] it is likely phages act as regulators of microbiota structure in the gut. Treatments may need to target not one, but several possible pathogens, and do so without adversely affecting beneficial microbiota.

Lactoferrin is an antimicrobial (reviewed by Jenssen *et al* [198]) produced by our immune system [199,200] that stimulates gut epithelial cell growth, proliferation and differentiation [183,184], and also displays antiviral properties [201]. The use of lactoferrin may prevent infant NEC [182] by inducing proliferation and differentiation of intestinal cells but cannot cure this illness [202]. A recent study of NEC showed a link between the use of the vegetal antimicrobial trans-cinnamaldehyde and the inactivation of the pathogen *Enterobacter sakazakii* [185]. This could prevent NEC in infants. However in preterm infants, NEC may be related to an allergic reaction to cow milk [186,187]. The question of whether this chain of events can be prevented or disrupted by appropriate microbiota remains unanswered, though several studies are in progress.

*4.3.1.2 Allergic Colitis*

Allergic colitis is often due to antigens against cow's milk [203,204]. *Lactobacillus* is thought to preserve the infant intestine by favoring antigen degradation. *Lactobacillus* preserves the intestinal mucosal barrier, competing with pathogenic bacteria. It also enhances the production of cytokines IL-10 and TGF-β [205,206], both involved in the process of immune tolerance during inflammation or allergy [207,208]. Savino and colleagues observed improved gut motility and decreased pain in 24 of 25 infants suffering colitis after 21 days of *Lactobacillus* diet complementation, but only 15 of 21 children treated with placebo [181,188,189]. Infants given probiotic containing *Lactobacillus* were better protected against rectal bleeding and allergic colitis. The importance of *Lactobacillus* in immune protection against inflammatory colitis was documented by Nermes and colleagues [190]. They added *Lactobacillus* to the diet of 19 infants between 3 and 13 months of age during three months exclusively formula-fed. Compared with 20 controls, the *Lactobacillus* infants had decreases in the number of cells secreting IgA and IgM (by 7% and 20% respectively) while the controls had increases in the same cell populations (by 22% and 31% respectively) between the beginning and the end of the study. At the same time the *Lactobacillus*-enriched diet correlated with an increase of CD19+ and CD27+ B memory cells compared with controls. The authors observed no difference in *Bifidobacterium* populations between the treated and the

control groups, both of which were breast-fed before the study. *Lactobacillus* supplementation reduced the level of calprotectin, a marker of cow's milk allergic colitis, and allowed a better recovery of the intestinal mucosa inflammation in infants suffering from hematochezia [189]. Arvola and colleagues reported a group of infants suffering rectal bleeding who were largely breast-fed [192]. The cow's milk elimination diet in a subset of infants did not affect the duration of rectal bleeding. Compared to the control group, infants suffering from rectal bleeding had lower bacteria counts and their populations of *Bifidobacterium* and *Lactobacillus* were around ten times lower than in healthy infants. The authors thus suggested the possibility of probiotic intervention aimed at normalizing the level of bifidobacteria and lactobacilli.

*4.3.2 Dangers of antibiotherapy*

The use of antibiotics continues to increase in human medicine [209]. In addition, low dose antibiotics are used routinely to increase the weight and growth rate of livestock [210]. The widespread use of antibiotics resulted in the emergence of multidrug resistant bacteria and possibly in changes in the persistent human-associated microbial communities. At present, immune development is influenced by a decrease in microbial diversity compared to previous generations [210]. Animal models suggest the microbiota shift associated with increased antibiotic use may increase the incidence of obesity [210] which has reached pandemic proportions among humans. The use of antibiotics affects the colonization of the infant gut by a potential pathogen *S. aureus* [103,211,212]. Lindberg and colleagues found a higher prevalence of *S. aureus* colonization in the gut of Swedish than in Italian infants until the age of one year. This difference was not related to the delivery mode or the diet and was attributed to a common use of antibiotics effective against *S. aureus* in Italy, while penicillin V, inefficient against *S. aureus*, is mostly used on Swedish infants. Not only the use of antibiotics may create new niches for resistant bacteria but may also eradicate useful bacteria like *Lactobacillus*, or favor the overgrowth of pathogenic bacteria like *C. difficile* [213,98,107]. Therefore, the use of antibiotics in preterm infants or babies delivered by C-section may bias studies on *C. difficile* and *S. aureus* carriage and their relationship with disease. While several studies suggest that changes in the overall bacterial community are associated with *S. aureus* pathogenicity, we do not know if the changes in bacterial community are causative, as they are thought to be with *C. difficile*.

Treating pneumonia with ceftriaxone for 5 days in babies < 6 months old results in extirpation of *Lactobacillus* from infants' gut microbiota, as well as a decreased population of other commensal bacteria like Enterobacteriaceae [98]. However, 20 days after the end of the treatment, the gut bacterial communities were greatly recovered and included *Lactobacillus* that had disappeared during the treatment. Recent research on adult gut microbiota modifications following 4 days of ciprofloxacin treatment repeated twice over 2 months reported a likely permanent shift in the rare gut microbiota species [214]. However, some of the newly acquired species were close to those present before the treatment. Recently, Hviid and colleagues pointed out the relationship between the number of courses of antibiotics and risk of inflammatory bowel disease in infants [215]. Difficulties with gut microbiota recovery after antibiotherapy in early life highlights the importance of the initial gut microbiota establishment and a potential danger of antibiotic (over)use [216].

**5. Viral gut community**

5.1 *Who are they?*

The human microbiota is associated with an abundant and diverse community of viruses, in particular phages [217-220,59]. These phage populations are likely to play an important role in fast developing infant gut microbiome diversity [59]. However, a relative stability of both phage and bacterial populations in adult feces samples led to the hypothesis that a predatory viral microbial dynamic does not exist in the adult human gut [195]. The interaction of phages and bacteria in the gut remains relatively poorly understood. One of the main issues in identification of phages, and viruses in general, is the lack of sequence similarity between metagenomic sequence reads and known viral genomes [195].

5.2 *Relationship between viruses and bacteria in the gut*

Phage communities in Western infants' gut present low diversity but they might influence the abundance of the microbial community [59]. Breitbart and colleagues suggested that fecal phages do not originate from a dietary source as they are different from the phages present in mothers' milk or in formulas [59]. They hypothesize that the first viruses in the gut originate from prophage induction from the first colonizing bacteria rather than from an environmental source. Persistence of the phage population in infants' gut over extended periods raised the hypothesis of a completely different mode of colonization than that of bacteria. Indeed no significant clustering of viral population could be observed between co-twins or between twins and their mothers [195] while their bacterial populations were closer compared to those of unrelated people [146]. However, with a different approach, using microarrays, Breitbart *et al* reported a dramatic shift in the gut viral population between 1 and 2 weeks of life [59], related to an important modification in bacterial population [221]. Therefore, the results obtained in different studies should be considered in the light of the methodology used. Combining the data generated using different methodologies will improve our understanding of the gut viral communities.

## 6. Outlook

Metagenomic studies of infants' gut microbiota provide important insights into understanding bacterial colonization of the gut and its role in human health. However, differences in sampling procedures and methodologies used as well as the lack of data on previous use of antibiotics can complicate comparisons of the results from different studies. While metagenomics reveals changes in bacterial communities as a function of geography, diet and medical treatments, in many instances the causal relation between altered microbiota and a disease remains unclear. To deal with this issue, long-term studies with large cohorts are needed. Given the complexity of the interactions between the host and environmental factors, and members of the gut bacterial community, one of the challenges is to define normal and abnormal gut microbiota. In that light, the exposure of germ-free mice to defined bacterial consortia under controlled conditions may provide insights into the impact of early bacterial colonization on the host physiology. As shown for prebiotics and probiotics, bacteriophage supplementation in infant's diet might have health-promoting effect. However, possible benefits and risk of phage-, prebiotic- and probiotic supplements on the health in later life remain to be determined.

The prevalence and incidence of allergies and autoimmune diseases have been increasing over recent decades. Metagenomics provides tools to relate these conditions to the gut microbiota composition and structure which may contribute to the development of novel prophylactic and therapeutic approaches.

## Acknowlegments

This work was supported by funding from the Swiss National Science Foundation (Grant 3100A0-112370/1 to J.S.) and from bioMérieux.

## References

1. G.W. Tyson, J. Chapman, P. Hugenholtz, E.E. Allen, R.J. Ram, P.M. Richardson, V.V. Solovyev, E.M. Rubin, D.S. Rokhsar, J.F. Banfield, Nature 428 (2004) 37-43.
2. H. Bertrand, F. Poly, V.T. Van, N. Lombard, R. Nalin, T.M. Vogel, P. Simonet, J.Microbiol.Methods 62 (2005) 1-11.
3. J.C. Venter, K. Remington, J.F. Heidelberg, A.L. Halpern, D. Rusch, J.A. Eisen, D. Wu, I. Paulsen, K.E. Nelson, W. Nelson, D.E. Fouts, S. Levy, A.H. Knap, M.W. Lomas, K. Nealson, O. White, J. Peterson, J. Hoffman, R. Parsons, H. Baden-Tillson, C. Pfannkoch, Y.H. Rogers, H.O. Smith, Science 304 (2004) 66 -74.
4. S.R. Gill, M. Pop, R.T. Deboy, P.B. Eckburg, P.J. Turnbaugh, B.S. Samuel, J.I. Gordon, D.A. Relman, C.M. Fraser-Liggett, K.E. Nelson, Science 312 (2006) 1355-1359.
5. P.J. Turnbaugh, R.E. Ley, M.A. Mahowald, V. Magrini, E.R. Mardis, J.I. Gordon, Nature 444 (2006) 1027-1031.
6. R.E. Ley, F. Backhed, P. Turnbaugh, C.A. Lozupone, R.D. Knight, J.I. Gordon, Proc.Natl.Acad.Sci.U.S.A 102 (2005) 11070-11075.
7. T.D. Luckey, Am.J.Clin.Nutr. 25 (1972) 1292-1294.
8. R.D. Berg, Trends Microbiol. 4 (1996) 430-435.
9. A.L. Goodman, G. Kallstrom, J.J. Faith, A. Reyes, A. Moore, G. Dantas, J.I. Gordon, Proc.Natl.Acad.Sci.U.S.A 108 (2011) 6252-6257.
10. V. Lazarevic, K. Whiteson, D. Hernandez, P. Francois, J. Schrenzel, BMC.Genomics 11 (2010) 523.
11. J.L. Sonnenburg, J. Xu, D.D. Leip, C.H. Chen, B.P. Westover, J. Weatherford, J.D. Buhler, J.I. Gordon, Science 307 (2005) 1955-1959.
12. H.J. Flint, E.A. Bayer, M.T. Rincon, R. Lamed, B.A. White, Nat.Rev.Microbiol. 6 (2008) 121-131.
13. E. Isolauri, Y. Sutas, P. Kankaanpaa, H. Arvilommi, S. Salminen, Am.J.Clin.Nutr. 73 (2001) 444S-450S.
14. L.A. Hanson, M. Korotkova, S. Lundin, L. Haversen, S.A. Silfverdal, I. Mattsby-Baltzer, B. Strandvik, E. Telemo, Ann.N.Y.Acad.Sci. 987 (2003) 199-206.
15. F. Backhed, R.E. Ley, J.L. Sonnenburg, D.A. Peterson, J.I. Gordon, Science 307 (2005) 1915-1920.
16. S.K. Mazmanian, C.H. Liu, A.O. Tzianabos, D.L. Kasper, Cell 122 (2005) 107-118.
17. A. Are, L. Aronsson, S. Wang, G. Greicius, Y.K. Lee, J.A. Gustafsson, S. Pettersson, V. Arulampalam, Proc.Natl.Acad.Sci.U.S.A 105 (2008) 1943-1948.
18. M.R. Carrera, G.F. Kaufmann, J.M. Mee, M.M. Meijler, G.F. Koob, K.D. Janda, Proc.Natl.Acad.Sci.U.S.A 101 (2004) 10416 -10421.
19. W.R. Wikoff, G. Pendyala, G. Siuzdak, H.S. Fox, J.Clin.Invest 118 (2008) 2661-2669.
20. K. Cadwell, K.K. Patel, N.S. Maloney, T.C. Liu, A.C. Ng, C.E. Storer, R.D. Head, R. Xavier, T.S. Stappenbeck, H.W. Virgin, Cell 141 (2010) 1135-1145.
21. A. Mira, R. Pushker, F. Rodriguez-Valera, Trends Microbiol. 14 (2006) 200-206.
22. F.C. De, D. Cavalieri, P.M. Di, M. Ramazzotti, J.B. Poullet, S. Massart, S. Collini, G. Pieraccini, P. Lionetti, Proc.Natl.Acad.Sci.U.S.A 107 (2010) 14691-14696.
23. R.E. Ley, P.J. Turnbaugh, S. Klein, J.I. Gordon, Nature 444 (2006) 1022-1023.
24. P.J. Turnbaugh and J.I. Gordon, J.Physiol 587 (2009) 4153-4158.
25. P.J. Turnbaugh, V.K. Ridaura, J.J. Faith, F.E. Rey, R. Knight, J.I. Gordon, Sci.Transl.Med. 1 (2009) 6ra14.
26. M.J. Blaser, Trans.Am.Clin.Climatol.Assoc. 116 (2005) 65-75.
27. D.P. Strachan and C.H. Sanders, J.Epidemiol.Community Health 43 (1989) 7-14.
28. J. Kuczynski, E.K. Costello, D.R. Nemergut, J. Zaneveld, C.L. Lauber, D. Knights, O. Koren, N. Fierer, S.T. Kelley, R.E. Ley, J.I. Gordon, R. Knight, Genome Biol. 11 (2010) 210.
29. N. Larsen, F.K. Vogensen, F.W. van den Berg, D.S. Nielsen, A.S. Andreasen, B.K. Pedersen, W.A. Al-Soud, S.J. Sorensen, L.H. Hansen, M. Jakobsen, PLoS.One. 5 (2010) e9085.

30. U. Gophna, K. Sommerfeld, S. Gophna, W.F. Doolittle, S.J. Veldhuyzen van Zanten, J.Clin.Microbiol. 44 (2006) 4136-4141.
31. R. Bibiloni, M. Mangold, K.L. Madsen, R.N. Fedorak, G.W. Tannock, J.Med.Microbiol. 55 (2006) 1141-1149.
32. D.N. Frank, A.L. St Amand, R.A. Feldman, E.C. Boedeker, N. Harpaz, N.R. Pace, Proc.Natl.Acad.Sci.U.S.A 104 (2007) 13780-13785.
33. Y. Wang, J.D. Hoenig, K.J. Malin, S. Qamar, E.O. Petrof, J. Sun, D.A. Antonopoulos, E.B. Chang, E.C. Claud, ISME.J. 3 (2009) 944-954.
34. J.Y. Chang, D.A. Antonopoulos, A. Kalra, A. Tonelli, W.T. Khalife, T.M. Schmidt, V.B. Young, J.Infect.Dis. 197 (2008) 435-438.
35. C.F. Favier, E.E. Vaughan, W.M. De Vos, A.D. Akkermans, Appl.Environ.Microbiol. 68 (2002) 219-226.
36. W.R. Wikoff, A.T. Anfora, J. Liu, P.G. Schultz, S.A. Lesley, E.C. Peters, G. Siuzdak, Proc.Natl.Acad.Sci.U.S.A 106 (2009) 3698-3703.
37. C.R. Woese, Proc.Natl.Acad.Sci.U.S.A 99 (2002) 8742-8747.
38. N.R. Pace, G.J. Olsen, C.R. Woese, Cell 45 (1986) 325-326.
39. N.R. Pace and T.L. Marsh, Orig.Life Evol.Biosph. 16 (1985) 97-116.
40. M.M. Rinne, M. Gueimonde, M. Kalliomaki, U. Hoppu, S.J. Salminen, E. Isolauri, FEMS Immunol.Med.Microbiol. 43 (2005) 59-65.
41. J. Qin, R. Li, J. Raes, M. Arumugam, K.S. Burgdorf, C. Manichanh, T. Nielsen, N. Pons, F. Levenez, T. Yamada, D.R. Mende, J. Li, J. Xu, S. Li, D. Li, J. Cao, B. Wang, H. Liang, H. Zheng, Y. Xie, J. Tap, P. Lepage, M. Bertalan, J.M. Batto, T. Hansen, P.D. Le, A. Linneberg, H.B. Nielsen, E. Pelletier, P. Renault, T. Sicheritz-Ponten, K. Turner, H. Zhu, C. Yu, S. Li, M. Jian, Y. Zhou, Y. Li, X. Zhang, S. Li, N. Qin, H. Yang, J. Wang, S. Brunak, J. Dore, F. Guarner, K. Kristiansen, O. Pedersen, J. Parkhill, J. Weissenbach, P. Bork, S.D. Ehrlich, J. Wang, Nature 464 (2010) 59-65.
42. W.T. Liu, T.L. Marsh, H. Cheng, L.J. Forney, Appl.Environ.Microbiol. 63 (1997) 4516-4522.
43. O.U. Mason, T. Nakagawa, M. Rosner, J.D. Van Nostrand, J. Zhou, A. Maruyama, M.R. Fisk, S.J. Giovannoni, PLoS.One. 5 (2010) e15399.
44. G.B. Rogers, S. Skelton, D.J. Serisier, C.J. van der Gast, K.D. Bruce, J.Clin.Microbiol. 48 (2010) 78-86.
45. M.M. Fisher and E.W. Triplett, Appl.Environ.Microbiol. 65 (1999) 4630-4636.
46. J.A. Fuhrman, I. Hewson, M.S. Schwalbach, J.A. Steele, M.V. Brown, S. Naeem, Proc.Natl.Acad.Sci.U.S.A 103 (2006) 13104-13109.
47. D.P. Lejon, R. Chaussod, J. Ranger, L. Ranjard, Microb.Ecol. 50 (2005) 614-625.
48. R. Danovaro, G.M. Luna, A. Dell'anno, B. Pietrangeli, Appl.Environ.Microbiol. 72 (2006) 5982-5989.
49. R.S. Esworthy, D.D. Smith, F.F. Chu, Int.J.Inflam. 2010 (2010) 986046.
50. S. Sepehri, R. Kotlowski, C.N. Bernstein, D.O. Krause, Inflamm.Bowel.Dis. 13 (2007) 675-683.
51. L.F. Roesch, G.L. Lorca, G. Casella, A. Giongo, A. Naranjo, A.M. Pionzio, N. Li, V. Mai, C.H. Wasserfall, D. Schatz, M.A. Atkinson, J. Neu, E.W. Triplett, ISME.J. 3 (2009) 536-548.
52. A. Kovacs, K. Yacoby, U. Gophna, Res.Microbiol. 161 (2010) 192-197.
53. L. Ranjard, F. Poly, J.C. Lata, C. Mougel, J. Thioulouse, S. Nazaret, Appl.Environ.Microbiol. 67 (2001) 4479-4487.
54. P. Offre, B. Pivato, S. Siblot, E. Gamalero, T. Corberand, P. Lemanceau, C. Mougel, Appl.Environ.Microbiol. 73 (2007) 913-921.
55. J.G. Bauman, J. Wiegant, P. Borst, D.P. van, Exp.Cell Res. 128 (1980) 485-490.
56. P.R. Langer, A.A. Waldrop, D.C. Ward, Proc.Natl.Acad.Sci.U.S.A 78 (1981) 6633-6637.
57. H.F. Jenkinson and R.J. Lamont, Trends Microbiol. 13 (2005) 589-595.
58. J.D. Rudney, R. Chen, G.J. Sedgewick, J.Dent.Res. 84 (2005) 59-63.
59. M. Breitbart, M. Haynes, S. Kelley, F. Angly, R.A. Edwards, B. Felts, J.M. Mahaffy, J. Mueller, J. Nulton, S. Rayhawk, B. Rodriguez-Brito, P. Salamon, F. Rohwer, Res.Microbiol. 159 (2008) 367-373.
60. A. Huyghe, P. Francois, Y. Charbonnier, M. Tangomo-Bento, E.J. Bonetti, B.J. Paster, I. Bolivar, D. Baratti-Mayer, D. Pittet, J. Schrenzel, Appl.Environ.Microbiol. 74 (2008) 1876-1885.
61. A. Huyghe, P. Francois, J. Schrenzel, Infect.Genet.Evol. 9 (2009) 987-995.
62. M.L. Sogin, H.G. Morrison, J.A. Huber, W.D. Mark, S.M. Huse, P.R. Neal, J.M. Arrieta, G.J. Herndl, Proc.Natl.Acad.Sci.U.S.A 103 (2006) 12115-12120.
63. V. Lazarevic, K. Whiteson, S. Huse, D. Hernandez, L. Farinelli, M. Osteras, J. Schrenzel, P. Francois, J.Microbiol.Methods 79 (2009) 266-271.
64. Q. Wang, G.M. Garrity, J.M. Tiedje, J.R. Cole, Appl.Environ.Microbiol. 73 (2007) 5261-5267.
65. P.J. Turnbaugh, C. Quince, J.J. Faith, A.C. McHardy, T. Yatsunenko, F. Niazi, J. Affourtit, M. Egholm, B. Henrissat, R. Knight, J.I. Gordon, Proc.Natl.Acad.Sci.U.S.A 107 (2010) 7503-7508.
66. D.A. Relman, T.M. Schmidt, R.P. MacDermott, S. Falkow, N.Engl.J.Med. 327 (1992) 293-301.
67. K. Chen and L. Pachter, PLoS.Comput.Biol. 1 (2005) 106-112.
68. P.D. Schloss and J. Handelsman, Appl.Environ.Microbiol. 71 (2005) 1501-1506.
69. J.R. Cole, Q. Wang, E. Cardenas, J. Fish, B. Chai, R.J. Farris, A.S. Kulam-Syed-Mohideen, D.M. McGarrell, T. Marsh, G.M. Garrity, J.M. Tiedje, Nucleic Acids Res. 37 (2009) D141-D145.
70. J.A. Aas, B.J. Paster, L.N. Stokes, I. Olsen, F.E. Dewhirst, J.Clin.Microbiol. 43 (2005) 5721-5732.
71. I. Nasidze, J. Li, D. Quinque, K. Tang, M. Stoneking, Genome Res. 19 (2009) 636-643.
72. C. Lozupone and R. Knight, Appl.Environ.Microbiol. 71 (2005) 8228-8235.
73. C.O. Webb, Am.Nat. 156 (2000) 145-155.
74. A. Ramette, FEMS Microbiol.Ecol. 62 (2007) 142-160.
75. G. Talbot, E. Topp, M.F. Palin, D.I. Masse, Water Res. 42

(2008) 513-537.
76. K.R. Clark, Australian Journal of Ecology 18 (1993) 117-143.
77. M.J. Anderson, Austral Ecology 26 (2001) 32-46.
78. F. Meyer, D. Paarmann, M. D'Souza, R. Olson, E.M. Glass, M. Kubal, T. Paczian, A. Rodriguez, R. Stevens, A. Wilke, J. Wilkening, R.A. Edwards, BMC.Bioinformatics. 9 (2008) 386.
79. J.G. Caporaso, J. Kuczynski, J. Stombaugh, K. Bittinger, F.D. Bushman, E.K. Costello, N. Fierer, A.G. Pena, J.K. Goodrich, J.I. Gordon, G.A. Huttley, S.T. Kelley, D. Knights, J.E. Koenig, R.E. Ley, C.A. Lozupone, D. McDonald, B.D. Muegge, M. Pirrung, J. Reeder, J.R. Sevinsky, P.J. Turnbaugh, W.A. Walters, J. Widmann, T. Yatsunenko, J. Zaneveld, R. Knight, Nat.Methods 7 (2010) 335-336.
80. P.D. Schloss, S.L. Westcott, T. Ryabin, J.R. Hall, M. Hartmann, E.B. Hollister, R.A. Lesniewski, B.B. Oakley, D.H. Parks, C.J. Robinson, J.W. Sahl, B. Stres, G.G. Thallinger, D.J. Van Horn, C.F. Weber, Appl.Environ.Microbiol. 75 (2009) 7537-7541.
81. W. Gerlach and J. Stoye, Nucleic Acids Res. 39 (2011) e91.
82. R. Seshadri, S.A. Kravitz, L. Smarr, P. Gilna, M. Frazier, PLoS.Biol. 5 (2007) e75.
83. V.M. Markowitz, N. Ivanova, K. Palaniappan, E. Szeto, F. Korzeniewski, A. Lykidis, I. Anderson, K. Mavromatis, V. Kunin, M.H. Garcia, I. Dubchak, P. Hugenholtz, N.C. Kyrpides, Bioinformatics. 22 (2006) e359-e367.
84. P. Trosvik, N.C. Stenseth, K. Rudi, ISME.J. 4 (2010) 151-158.
85. C. Palmer, E.M. Bik, D.B. DiGiulio, D.A. Relman, P.O. Brown, PLoS.Biol. 5 (2007) e177.
86. T.M. Marques, R. Wall, R.P. Ross, G.F. Fitzgerald, C.A. Ryan, C. Stanton, Curr.Opin.Biotechnol. 21 (2010) 149-156.
87. G. Biasucci, B. Benenati, L. Morelli, E. Bessi, G. Boehm, J.Nutr. 138 (2008) 1796S-1800S.
88. J. Penders, C. Thijs, C. Vink, F.F. Stelma, B. Snijders, I. Kummeling, P.A. van den Brandt, E.E. Stobberingh, Pediatrics 118 (2006) 511-521.
89. M.M. Gronlund, S. Salminen, H. Mykkanen, P. Kero, O.P. Lehtonen, APMIS 107 (1999) 655-660.
90. M.M. Gronlund, O.P. Lehtonen, E. Eerola, P. Kero, J.Pediatr.Gastroenterol.Nutr. 28 (1999) 19-25.
91. K. Orrhage and C.E. Nord, Acta Paediatr.Suppl 88 (1999) 47-57.
92. M. Gueimonde, S. Sakata, M. Kalliomaki, E. Isolauri, Y. Benno, S. Salminen, J.Pediatr.Gastroenterol.Nutr. 42 (2006) 166-170.
93. P.A. Vaishampayan, J.V. Kuehl, J.L. Froula, J.L. Morgan, H. Ochman, M.P. Francino, Genome Biol.Evol. 2 (2010) 53-66.
94. M.G. Dominguez-Bello, E.K. Costello, M. Contreras, M. Magris, G. Hidalgo, N. Fierer, R. Knight, Proc.Natl.Acad.Sci.U.S.A 107 (2010) 11971-11975.
95. M. Fallani, D. Young, J. Scott, E. Norin, S. Amarri, R. Adam, M. Aguilera, S. Khanna, A. Gil, C.A. Edwards, J. Dore, J.Pediatr.Gastroenterol.Nutr. 51 (2010) 77-84.
96. D.A. Sela, J. Chapman, A. Adeuya, J.H. Kim, F. Chen, T.R. Whitehead, A. Lapidus, D.S. Rokhsar, C.B. Lebrilla, J.B. German, N.P. Price, P.M. Richardson, D.A. Mills, Proc.Natl.Acad.Sci.U.S.A 105 (2008) 18964-18969.
97. M. Eggesbo, B. Moen, S. Peddada, D. Baird, J. Rugtveit, T. Midtvedt, P.R. Bushel, M. Sekelja, K. Rudi, APMIS 119 (2011) 17-35.
98. F. Savino, J. Roana, N. Mandras, V. Tarasco, E. Locatelli, V. Tullio, Acta Paediatr. 100 (2011) 75-78.
99. D.S. Newburg, Adv.Exp.Med.Biol. 501 (2001) 3-10.
100. M.A. Quigley, P. Cumberland, J.M. Cowden, L.C. Rodrigues, Arch.Dis.Child 91 (2006) 245-250.
101. R.W. Hyman, M. Fukushima, L. Diamond, J. Kumm, L.C. Giudice, R.W. Davis, Proc.Natl.Acad.Sci.U.S.A 102 (2005) 7952-7957.
102. X. Zhou, C.J. Brown, Z. Abdo, C.C. Davis, M.A. Hansmann, P. Joyce, J.A. Foster, L.J. Forney, ISME.J. 1 (2007) 121-133.
103. E. Lindberg, I. Adlerberth, B. Hesselmar, R. Saalman, I.L. Strannegard, N. Aberg, A.E. Wold, J.Clin.Microbiol. 42 (2004) 530-534.
104. G. Reid, J.A. Younes, H.C. Van der Mei, G.B. Gloor, R. Knight, H.J. Busscher, Nat.Rev.Microbiol. 9 (2011) 27-38.
105. M.J. Morowitz, V.J. Denef, E.K. Costello, B.C. Thomas, V. Poroyko, D.A. Relman, J.F. Banfield, Proc.Natl.Acad.Sci.U.S.A 108 (2011) 1128-1133.
106. E. Ozaki, H. Kato, H. Kita, T. Karasawa, T. Maegawa, Y. Koino, K. Matsumoto, T. Takada, K. Nomoto, R. Tanaka, S. Nakamura, J.Med.Microbiol. 53 (2004) 167-172.
107. C. Rousseau, F. Levenez, C. Fouqueray, J. Dore, A. Collignon, P. Lepage, J.Clin.Microbiol. 49 (2011) 858-865.
108. S.G. McNeeley, Jr., G.D. Anderson, B.M. Sibai, Obstet.Gynecol. 66 (1985) 737-738.
109. B.S. Block, L.J. Mercer, M.A. Ismail, A.H. Moawad, Am.J.Obstet.Gynecol. 153 (1985) 835-838.
110. E. Manso, P. Strusi, M.A. Stacchiotti, R. Vincenzi, M. Tiriduzzi, P.U. Del, Boll.Ist.Sieroter.Milan 65 (1986) 118-124.
111. A.A. Venugopal, D.N. Gerding, S. Johnson, Am.J.Infect.Control (2010).
112. S. Jangi and J.T. Lamont, J.Pediatr.Gastroenterol.Nutr. 51 (2010) 2-7.
113. A. Collignon, L. Ticchi, C. Depitre, J. Gaudelus, M. Delmee, G. Corthier, Eur.J.Pediatr. 152 (1993) 319-322.
114. A. Khoruts, J. Dicksved, J.K. Jansson, M.J. Sadowsky, J.Clin.Gastroenterol. 44 (2010) 354-360.
115. P. Varma, N. Nisha, K.R. Dinesh, A.V. Kumar, R. Biswas, J.Mol.Microbiol.Biotechnol. 20 (2011) 137-143.
116. E.K. Costello, C.L. Lauber, M. Hamady, N. Fierer, J.I. Gordon, R. Knight, Science 326 (2009) 1694-1697.
117. R.E. Ley, C.A. Lozupone, M. Hamady, R. Knight, J.I. Gordon, Nat.Rev.Microbiol. 6 (2008) 776-788.
118. A. Everard, V. Lazarevic, M. Derrien, M. Girard, G.G. Muccioli, A.M. Neyrinck, S. Possemiers, H.A. Van, P. Francois, W.M. De Vos, N.M. Delzenne, J. Schrenzel, P.D. Cani, Diabetes 60 (2011) 2775-2786.
119. 119. M.C. Daniels and L.S. Adair, J.Nutr. 135 (2005) 2589-2595.
120. J.B. German, C.J. Dillard, R.E. Ward, Curr. Opin. Clin. Nutr. Metab Care 5 (2002) 653-658.
121. H.J. Harmsen, A.C. Wildeboer-Veloo, G.C. Raangs, A.A. Wagendorp, N. Klijn, J.G. Bindels, G.W. Welling,

J.Pediatr.Gastroenterol.Nutr. 30 (2000) 61-67.
122. L. Schack-Nielsen and K.F. Michaelsen, Ugeskr.Laeger 169 (2007) 985-989.
123. L. Schack-Nielsen and K.F. Michaelsen, J.Nutr. 137 (2007) 503S-510S.
124. A. Bener, M.S. Ehlayel, S. Alsowaidi, A. Sabbah, Eur.Ann.Allergy Clin.Immunol. 39 (2007) 337-343.
125. W. Karmaus, A.L. Dobai, I. Ogbuanu, S.H. Arshard, S. Matthews, S. Ewart, J.Asthma 45 (2008) 688-695.
126. P. Behari, J. Englund, G. Alcasid, S. Garcia-Houchins, S.G. Weber, Infect.Control Hosp.Epidemiol. 25 (2004) 778-780.
127. M.F. de la Cochetiere, H. Piloquet, R.C. des, D. Darmaun, J.P. Galmiche, J.C. Roze, Pediatr.Res. 56 (2004) 366-370.
128. N. Linder, A. Hernandez, L. Amit, G. Klinger, S. Ashkenazi, I. Levy, Eur.J.Pediatr. (2011).
129. M. Haarman and J. Knol, Appl.Environ.Microbiol. 71 (2005) 2318-2324.
130. M. Haarman and J. Knol, Appl.Environ.Microbiol. 72 (2006) 2359-2365.
131. F.F. Rubaltelli, R. Biadaioli, P. Pecile, P. Nicoletti, J.Perinat.Med. 26 (1998) 186-191.
132. A.M. Zivkovic, J.B. German, C.B. Lebrilla, D.A. Mills, Proc.Natl.Acad.Sci.U.S.A (2010).
133. M. Roberfroid, G.R. Gibson, L. Hoyles, A.L. McCartney, R. Rastall, I. Rowland, D. Wolvers, B. Watzl, H. Szajewska, B. Stahl, F. Guarner, F. Respondek, K. Whelan, V. Coxam, M.J. Davicco, L. Leotoing, Y. Wittrant, N.M. Delzenne, P.D. Cani, A.M. Neyrinck, A. Meheust, Br.J.Nutr. 104 Suppl 2 (2010) S1-63.
134. D.S. Newburg, J.Anim Sci. 87 (2009) 26-34.
135. C. Kunz and S. Rudloff, Adv.Exp.Med.Biol. 606 (2008) 455-465.
136. T. Arvola, K. Laiho, S. Torkkeli, H. Mykkanen, S. Salminen, L. Maunula, E. Isolauri, Pediatrics 104 (1999) e64.
137. H. Szymanski, J. Pejcz, M. Jawien, A. Chmielarczyk, M. Strus, P.B. Heczko, Aliment.Pharmacol.Ther. 23 (2006) 247-253.
138. T.D. Braga, G.A. da Silva, P.I. de Lira, L.M. de Carvalho, Am.J.Clin.Nutr. 93 (2011) 81-86.
139. S. Rerksuppaphol and L. Rerksuppaphol, Ann.Trop.Paediatr. 30 (2010) 299-304.
140. W.A. Mihatsch, S. Vossbeck, B. Eikmanns, J. Hoegel, F. Pohlandt, Neonatology. 98 (2010) 156-163.
141. F.N. Sari, E.A. Dizdar, S. Oguz, O. Erdeve, N. Uras, U. Dilmen, Eur.J.Clin.Nutr. 65 (2011) 434-439.
142. G.R. Gibson, H.M. Probert, J.V. Loo, R.A. Rastall, M.B. Roberfroid, Nutr.Res.Rev. 17 (2004) 259-275.
143. G.E. Moro and S. Arslanoglu, Acta Paediatr.Suppl 94 (2005) 14-17.
144. M.R. Ninonuevo, P.D. Perkins, J. Francis, L.M. Lamotte, R.G. LoCascio, S.L. Freeman, D.A. Mills, J.B. German, R. Grimm, C.B. Lebrilla, J.Agric.Food Chem. 56 (2008) 618-626.
145. M.R. Ninonuevo and L. Bode, Pediatr.Res. 64 (2008) 8-10.
146. P.J. Turnbaugh, M. Hamady, T. Yatsunenko, B.L. Cantarel, A. Duncan, R.E. Ley, M.L. Sogin, W.J. Jones, B.A. Roe, J.P. Affourtit, M. Egholm, B. Henrissat, A.C. Heath, R. Knight, J.I. Gordon, Nature 457 (2009) 480-484.
147. R.I. Mackie, A. Sghir, H.R. Gaskins, Am.J.Clin.Nutr. 69 (1999) 1035S-1045S.
148. J.J. Cebra, Am.J.Clin.Nutr. 69 (1999) 1046S-1051S.
149. S. Rabot, M. Membrez, A. Bruneau, P. Gerard, T. Harach, M. Moser, F. Raymond, R. Mansourian, C.J. Chou, FASEB J. 24 (2010) 4948-4959.
150. C.K. Fleissner, N. Huebel, M.M. Abd El-Bary, G. Loh, S. Klaus, M. Blaut, Br.J.Nutr. 104 (2010) 919-929.
151. S. Ding, M.M. Chi, B.P. Scull, R. Rigby, N.M. Schwerbrock, S. Magness, C. Jobin, P.K. Lund, PLoS.One. 5 (2010) e12191.
152. N. Salles, A. Menard, A. Georges, M. Salzmann, L. de, V, M.A. de, J.P. Emeriau, H. Lamouliatte, F. Megraud, J.Gerontol.A Biol.Sci.Med.Sci. 61 (2006) 1144-1150.
153. D.W. Jun, O.Y. Lee, Y.Y. Lee, H.S. Choi, T.H. Kim, B.C. Yoon, Dig.Dis.Sci. 52 (2007) 2866-2872.
154. J. Roper, F. Francois, P.L. Shue, M.S. Mourad, Z. Pei, A.Z. Olivares de Perez, G.I. Perez-Perez, C.H. Tseng, M.J. Blaser, J.Clin.Endocrinol.Metab 93 (2008) 2350-2357.
155. J. Weigt and P. Malfertheiner, Curr.Opin.Clin.Nutr.Metab Care 12 (2009) 522-525.
156. F. Backhed, J.K. Manchester, C.F. Semenkovich, J.I. Gordon, Proc.Natl.Acad.Sci.U.S.A 104 (2007) 979-984.
157. H. Tilg, J.Clin.Gastroenterol. 44 Suppl 1 (2010) S16-S18.
158. D.P. Strachan, BMJ 299 (1989) 1259-1260.
159. R. Martin, A.J. Nauta, A.K. Ben, L.M. Knippels, J. Knol, J. Garssen, Benef.Microbes. 1 (2010) 367-382.
160. H.A. Gordon and L. Pesti, Bacteriol.Rev. 35 (1971) 390-429.
161. K. Smith, K.D. McCoy, A.J. Macpherson, Semin.Immunol. 19 (2007) 59-69.
162. Y.K. Lee, J.S. Menezes, Y. Umesaki, S.K. Mazmanian, Proc.Natl.Acad.Sci.U.S.A (2010).
163. T.S. Stappenbeck, L.V. Hooper, J.I. Gordon, Proc.Natl.Acad.Sci.U.S.A 99 (2002) 15451-15455.
164. H. Hooijkaas, R. Benner, J.R. Pleasants, B.S. Wostmann, Eur.J.Immunol. 14 (1984) 1127-1130.
165. D. Bouskra, C. Brezillon, M. Berard, C. Werts, R. Varona, I.G. Boneca, G. Eberl, Nature 456 (2008) 507-510.
166. T. Yamanaka, L. Helgeland, I.N. Farstad, H. Fukushima, T. Midtvedt, P. Brandtzaeg, J.Immunol. 170 (2003) 816-822.
167. P.G. De, I. Nadal, M. Medina, E. Donat, C. Ribes-Koninckx, M. Calabuig, Y. Sanz, BMC.Microbiol. 10 (2010) 63.
168. T.R. Murgas and J. Neu, J.Perinatol. 31 Suppl 1 (2011) S29-S34.
169. J. Kunii, K. Takahashi, K. Kasakura, M. Tsuda, K. Nakano, A. Hosono, S. Kaminogawa, Immunobiology (2010).
170. C. Prussin and D.D. Metcalfe, J.Allergy Clin.Immunol. 111 (2003) S486-S494.
171. A. Khoruts and M.J. Sadowsky, Mucosal.Immunol. 4 (2011) 4-7.
172. M.P. Venkatesh and S.A. Abrams, Cochrane.Database.Syst.Rev. (2010) CD007137.
173. M. Mshvildadze and J. Neu, Early Hum.Dev. 85 (2009) S71-S74.
174. M.S. Caplan, J.Perinatol. 29 Suppl 2 (2009) S2-S6.
175. N. Ladd and T. Ngo, Proc.(Bayl.Univ Med.Cent.) 22 (2009) 287-291.
176. M. Millar, M. Wilks, K. Costeloe, Arch.Dis.Child Fetal Neonatal Ed 88 (2003) F354-F358.
177. K. Alfaleh, J. Anabrees, D. Bassler, Neonatology. 97 (2010) 93-99.
178. M. Millar, M. Wilks, P. Fleming, K. Costeloe, Arch.Dis.Child Fetal Neonatal Ed (2010).
179. E.C. Claud, Anaerobe. (2011).
180. J.C. Nogueira and M.C. Goncalves, Braz.J.Otorhinolaryngol. 77 (2011) 129-134.
181. F. Savino, L. Cordisco, V. Tarasco, E. Palumeri, R. Calabrese,

R. Oggero, S. Roos, D. Matteuzzi, Pediatrics 126 (2010) e526-e533.

182. M. Venkatesh and S. Abrams, Expert.Rev.Anti.Infect.Ther. 7 (2009) 515-525.
183. S. Oguchi, W.A. Walker, I.R. Sanderson, Biol.Neonate 67 (1995) 330-339.
184. V. Buccigrossi, M.G. de, E. Bruzzese, L. Ombrato, I. Bracale, G. Polito, A. Guarino, Pediatr.Res. 61 (2007) 410-414.
185. M.A. Amalaradjou, T.A. Hoagland, K. Venkitanarayanan, Int.J.Food Microbiol. 129 (2009) 146-149.
186. M.R. Faber, P. Rieu, B.A. Semmekrot, J.H. Van Krieken, J.J. Tolboom, J.M. Draaisma, Acta Paediatr. 94 (2005) 1514-1515.
187. A.E. Abdelhamid, S.L. Chuang, P. Hayes, J.M. Fell, Pediatr.Res. 69 (2011) 165-169.
188. F. Savino, E. Castagno, S. Viola, J.Pediatr. 157 (2010) 174-175.
189. M.E. Baldassarre, N. Laforgia, M. Fanelli, A. Laneve, R. Grosso, C. Lifschitz, J.Pediatr. 156 (2010) 397-401.
190. M. Nermes, J.M. Kantele, T.J. Atosuo, S. Salminen, E. Isolauri, Clin.Exp.Allergy 41 (2011) 370-377.
191. A. Voganatsi, A. Panyutich, K.T. Miyasaki, R.K. Murthy, J.Leukoc.Biol. 70 (2001) 130-134.
192. T. Arvola, T. Ruuska, J. Keranen, H. Hyoty, S. Salminen, E. Isolauri, Pediatrics 117 (2006) e760-e768.
193. FAO/WHO, FAO/WHO Regional Conference on Food Safety for Asia and the Pacific, Seremban, Malaysia (2004) 24-27.
194. A.J. van, S.F. de, G. Muyldermans, A. Bougatef, A. Naessens, S. Lauwers, J.Clin.Microbiol. 39 (2001) 293-297.
195. A. Reyes, M. Haynes, N. Hanson, F.E. Angly, A.C. Heath, F. Rohwer, J.I. Gordon, Nature 466 (2010) 334-338.
196. K.E. Wommack and R.R. Colwell, Microbiol.Mol.Biol.Rev. 64 (2000) 69-114.
197. T. Bouvier and P.A. del Giorgio, Environ.Microbiol. 9 (2007) 287-297.
198. H. Jenssen and R.E. Hancock, Biochimie 91 (2009) 19-29.
199. P.L. Masson, J.F. Heremans, E. Schonne, J.Exp.Med. 130 (1969) 643-658.
200. K.B. Pryzwansky, E.K. MacRae, J.K. Spitznagel, M.H. Cooney, Cell 18 (1979) 1025-1033.
201. B.W. van der Strate, L. Beljaars, G. Molema, M.C. Harmsen, D.K. Meijer, Antiviral Res. 52 (2001) 225-239.
202. D. Vieten, A. Corfield, P. Ramani, R. Spicer, Pediatr.Surg.Int. 22 (2006) 50-56.
203. H.R. Jenkins, J.R. Pincott, J.F. Soothill, P.J. Milla, J.T. Harries, Arch.Dis.Child 59 (1984) 326-329.
204. S. Mehr, A. Kakakios, K. Frith, A.S. Kemp, Pediatrics 123 (2009) e459-e464.
205. S. Sierra, F. Lara-Villoslada, L. Sempere, M. Olivares, J. Boza, J. Xaus, Anaerobe. 16 (2010) 195-200.
206. L.H. Zeuthen, L.N. Fink, S.B. Metzdorff, M.B. Kristensen, T.R. Licht, C. Nellemann, H. Frokiaer, BMC.Immunol. 11 (2010) 2.
207. M. Gavin and A. Rudensky, Curr.Opin.Immunol. 15 (2003) 690-696.
208. M.A. Perez-Machado, P. Ashwood, M.A. Thomson, F. Latcham, R. Sim, J.A. Walker-Smith, S.H. Murch, Eur.J.Immunol. 33 (2003) 2307-2315.
209. J. Carlet, P. Collignon, D. Goldmann, H. Goossens, I.C. Gyssens, S. Harbarth, V. Jarlier, S.B. Levy, B. N'Doye, D. Pittet, R. Richtmann, W.H. Seto, J.W. van der Meer, A. Voss, Lancet 378 (2011) 369-371.
210. M.J. Blaser,(2011) International Human Microbiome Congress, Conference presentation, Vancouver, 2011.
211. E. Lindberg, F. Nowrouzian, I. Adlerberth, A.E. Wold, Pediatr.Res. 48 (2000) 741-747.
212. E. Lindberg, I. Adlerberth, P. Matricardi, C. Bonanno, S. Tripodi, V. Panetta, B. Hesselmar, R. Saalman, N. Aberg, A.E. Wold, Clin.Microbiol.Infect. (2010).
213. M.J. Blaser and S. Falkow, Nat.Rev.Microbiol. 7 (2009) 887-894.
214. L. Dethlefsen and D.A. Relman, Proc.Natl.Acad.Sci.U.S.A 108 Suppl 1 (2011) 4554-4561.
215. A. Hviid, H. Svanstrom, M. Frisch, Gut 60 (2011) 49-54.
216. M. Blaser, Nature 476 (2011) 393-394.
217. M. Breitbart, P. Salamon, B. Andresen, J.M. Mahaffy, A.M. Segall, D. Mead, F. Azam, F. Rohwer, Proc.Natl.Acad.Sci.U.S.A 99 (2002) 14250-14255.
218. R. Cornax, M.A. Morinigo, F. Gonzalez-Jaen, M.C. Alonso, J.J. Borrego, Zentralbl.Bakteriol. 281 (1994) 214-224.
219. T. Zhang, M. Breitbart, W.H. Lee, J.Q. Run, C.L. Wei, S.W. Soh, M.L. Hibberd, E.T. Liu, F. Rohwer, Y. Ruan, PLoS.Biol. 4 (2006) e3.
220. M. Breitbart, I. Hewson, B. Felts, J.M. Mahaffy, J. Nulton, P. Salamon, F. Rohwer, J.Bacteriol. 185 (2003) 6220-6223.
221. T. Tapiainen, S. Ylitalo, E. Eerola, M. Uhari, APMIS 114 (2006) 812-817.

JOURNAL OF INTEGRATED OMICS
A METHODOLOGICAL JOURNAL
HTTP://WWW.JIOMICS.COM

ORIGINAL ARTICLE | DOI: 10.5584/jiomics.v2i1.83

# An Improved Isotope Coded Affinity Tag Technology for Thiol Redox Proteomics

Ning Zhu[#1], Mengmeng Zhu[#1], Shaojun Dai[1,2], Ran Zheng[3], Sixue Chen[*1,3].

[1]Department of Biology, Genetics Institute, University of Florida, Gainesville, FL 32610, USA; [2]Alkali Soil Natural Environmental Science Center, Northeast Forestry University, Key Laboratory of Saline-alkali Vegetation Ecology Restoration in Oil Field, Ministry of Education, Harbin 150040, China; [3]Proteomics Facility, Interdisciplinary Center for Biotechnology Research, University of Florida, Gainesville, FL 32610, USA. [#]These Authors contributed equally.

**Received:** 02 December 2011 **Accepted:** 24 January 2012 **Available Online**: 26 January 2012

ABSTRACT

Isotope Coded Affinity Tag (ICAT) is a gel-free technology for quantitative proteomics. In ICAT procedure, strong cation exchange chromatography (SCX) using increased potassium chloride gradient is recommended for peptide fractionation. Here we report optimization of hydrophilic interaction chromatography (HILIC) as an alternative strategy for peptide fractionation of ICAT samples. HILIC exhibits high separation efficiency and does not require any downstream desalting steps. Compared to SCX based ICAT, integration of HILIC into the ICAT technology has resulted in high rates of protein identification, cysteine mapping, and quantification of cysteine-containing peptides. The improved technology has shown utility in thiol redox proteomics. Interestingly, results from HILIC ICAT and SCX ICAT are complementary. Implementation of both provides high coverage analysis of a complex proteome.

**Keywords:** ICAT; HILIC; SCX; Proteomics; Mass spectrometry.

## 1. Introduction

Although two-dimensional gel electrophoresis (2-DE) has been used for decades to separate and quantify proteins on a large scale, its limitations in reproducibility and separation of membrane proteins, acidic, basic, very small and large proteins have promoted the invention of a gel-free isotope-coded affinity tag (ICAT) technology in 1999 [1, 2]. The ICAT reagent consists of a cysteine-reactive group, a linker containing stable isotope signatures and a biotin affinity tag enabling the isolation of cysteine-containing peptides. In the original ICAT reagent, the linker region of heavy form contains eight deuteriums and the light form contains no deuteriums. It was reported later that the light and heavy ICAT tagged peptides exhibited different retention on reverse phase HPLC columns [3]. In addition, the retention of the biotin group complicates tandem mass spectrometry (MS/MS) spectrum interpretation. To overcome the problems, a cleavable ICAT reagent (cICAT) was developed. Carbon-13 was used instead of deuterium in the linker and the mass difference between the heavy and light forms is 9 Dalton. Additionally, an acid cleavable moiety was introduced between the biotin group and the rest of the molecule [4, 5]. Because the ICAT technology overcomes the 2-DE limitations, it has been applied as an alternative to address a variety of biological questions including whole-cell protein expression changes [6-8], protein subcellular localization [9, 10], dynamics of protein complexes [11-13], and identification of redox sensitive proteins [14].

In the ICAT method, two protein samples are labeled with isotopically light and heavy ICAT reagents respectively, which covalently attach to cysteine residues of the samples. The two samples are combined, proteolyzed with trypsin, and the peptides are fractionated by strong cation exchange (SCX) chromatography using potassium chloride gradient. After desalting, ICAT labeled peptides are purified using

***Corresponding author:** Sixue Chen, Ph.D. Department of Biology, Genetics Institute, University of Florida, Gainesville, FL 32610, USA; Tel: (352) 273-8330; Fax: (352) 273-8284; E-mail Address: schen@ufl.edu

avidin affinity chromatography, followed by acid cleavage of the biotin group. The purified peptides are then subjected to mass spectrometry (MS) analysis for relative quantification and peptide sequencing [4, 7]. Obviously, the ICAT method has the advantage of simplifying sample complexity using SCX fractionation and biotin affinity purification of cysteine-containing peptides. However, SCX separation is solely based on the peptide charge state and often limited by poor retention of similarly charged peptides. Recent studies have shown that hydrophilic interaction chromatography (HILIC) offers a superior separation mechanism that is based on retention by hydrophilicity and electrostatic interaction [15, 16]. The stationary phase of HILIC column (e.g., Luna® HILIC column, Phenomenex Inc., USA) has a silica surface covered with cross-linked diol groups, which has been reported to have broad range of applications with high recovery and reproducibility [17]. Direct comparison of SCX and HILIC in peptide separation has demonstrated the enhanced separation power of HILIC [15, 16, 18, 19]. Since HILIC is a type of normal phase liquid chromatography, volatile solvent can be used and desalting steps for downstream applications are not necessary. HILIC may provide an excellent alternative to SCX in ICAT peptide fractionation.

Since the development of isobaric tag for relative and absolute quantification (iTRAQ) in 2004 [20], it has replaced ICAT in many applications because of the high efficiency of iTRAQ multiplexing and the labeling of all peptides [20-22]. One application that retains the strength of the ICAT technology is thiol redox protein analysis [14, 23]. Here we investigated the use of HILIC to replace SCX recommended in current ICAT method with the focus on the identification of redox sensitive proteins. Control and abscisic acid (ABA) treated guard cell proteins were used as starting materials. After ICAT labeling and trypsin digestion, optimized HILIC conditions were used to fractionate the peptides. The fractions were loaded onto avidin columns without desalting. In parallel, regular SCX with potassium chloride gradient was carried out for comparison. HILIC showed better retention and recovery of ICAT labeled peptides. MS analysis revealed more protein identification and cysteine mapping data of the HILIC samples than the SCX samples. Interestingly, the two methods turned out to be complementary. The improved HILIC ICAT together with the SCX ICAT can provide greater in-depth coverage analysis of quantitative changes of reactive cysteines in complex protein samples.

## 2. Material and Methods

### 2.1 *Plant growth, guard cell isolation and ABA treatment.*

*Brassica napus* var. Global seeds were kindly provided by Svalöv Weibull AB (Svalöv, Sweden). Plant growth and guard cell isolation were carried out as previously described [21]. ABA was incubated with the guard cells at 100 μM for 3 hours during the isolation [22].

### 2.2 *Protein preparation and ICAT labeling.*

A solution of 10% trichloroacetic acid (TCA) in acetone was used to precipitate protein on ice for 2 hours. Samples were washed with 80% acetone once followed by washing with 100% acetone twice. Pellets were dissolved in the ReadyPrep™ Sequential Extraction Reagent 3 (Bio-Rad Inc., USA) and quantified by a CB-X™ protein assay kit (G Biosciences Inc., USA). Protein aliquots of 100 μg were alkylated with 100 mM iodoacetamide (IAM) at 75°C for 5 min, followed by 1 hour incubation at 37°C [24]. Protein samples were precipitated in 100% cold acetone over night. The pellets were dissolved in 80 μL denaturing buffer (pH 8.5) provided in the ICAT kit (AB Sciex Inc., USA). Reduction, ICAT labeling and trypsin digestion were performed according to the manufacturer's manual. The tryptic peptides were fractionated using an Agilent 1100 HPLC with a Luna® HILIC column (150 × 2.0 mm, 3 μm, 200 Å, Phenomenex Inc., USA) or with a PolySULFOETHYL A™ SCX column (150 × 2.1 mm, 5 μm, 300 Å, Poly LC Inc., USA). Mobile phases for HILIC were 5 mM ammonium acetate in 90% acetonitrile, pH 5.8 as solvent A and 5 mM ammonium acetate in water, pH 5.8 as solvent B. Peptides were eluted at a flow rate of 200 μL/min with a linear gradient of 0-50% solvent B over 50 min, followed by ramping up to 100% solvent B in 5 min and holding for 5 min before equilibrating in 0% solvent B. Mobile phases for SCX were 10 mM $KH_2PO_4$, 25% acetonitrile, pH 3 as solvent A and 10 mM $KH_2PO_4$, 350 mM KCl, 25% acetonitrile, pH 3 as solvent B. Peptides were eluted at a flow rate of 200 μL/min with a linear gradient of 0-100% solvent B over 60 min, followed by holding at 100% solvent B for 5 min before equilibrating in 100% solvent A. The absorbance at 214 nm was monitored and a total of 10 fractions were collected. The SCX fractions were desalted by solid phase extraction (SPE) using Vydac® silica C18 MacroSpin™ column (The Nest Group Inc., USA). The peptides from each fraction were affinity purified using an avidin affinity cartridge provided in the ICAT kit. ICAT labeled peptides were released by incubating with the cleavage reagent at 37°C for two hours, followed by lyophilization to dryness (Figure 1).

### 2.3 *Protein identification using LC-MS/MS, database searching and data analysis.*

ICAT labeled peptides were dissolved in 10 μL solvent A (0.1% v/v acetic acid, 3% v/v acetonitrile) and loaded onto a C18 PepMap™ nanoflow column (75 μm id, 3 μm, 100 Å, Dionex, USA). The elution gradient of the column started at 3% solvent A, 97% solvent B and finished at 60% solvent A, 40% solvent B for 60 min. Solvent A consisted of 0.1% v/v acetic acid, 3% v/v acetonitrile, and 96.9% v/v water. Solvent B consisted of 0.1% v/v acetic acid, 96.9% v/v acetonitrile, and 3% v/water. Tandem MS analysis was carried out on a

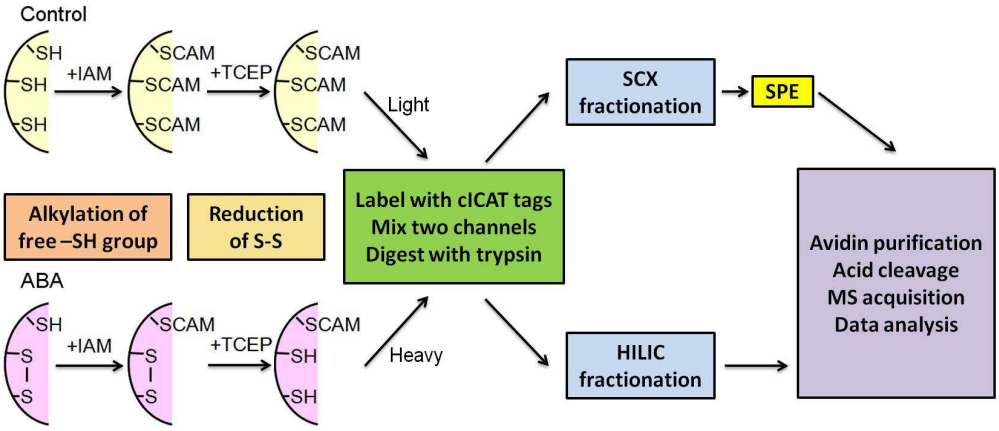


**Figure 1**. ICAT workflow with SCX and HILIC for thiol-based redox sensitive protein identification.

hybrid quadrupole time-of-flight mass spectrometer (QSTAR® XL, AB Sciex Inc., USA) as previously described [24]. Proteins were identified by searching the MS/MS data against a custom database containing *Arabidopsis thaliana* and *B. napus* protein sequences (downloaded from NCBI with a total of 33,365 entries) using ProteinPilot™ 4.0 software (AB Sciex Inc., USA) because the complete *B. napus* genome is not available and *A. thaliana* is a close relative sharing up to 87% protein sequence identity [25]. The following criteria were used to identify redox sensitive cysteines and proteins: i) at least 20% change of ICAT peptide ion intensity under ABA treatment [14] (Supplemental Figure 1), ii) peptide confidence over 95%, iii) peptide present in at least two replicates, and iv) each peptide assigned to only one protein. The MS data reported in this paper are available in the PRIDE database (www.ebi.ac.uk/prid) [26, 27] under accession number 16864-16867.

## 3. Results and Discussion

### 3.1 *Optimization of HILIC conditions for peptide fractionation.*

The factors that affect HILIC chromatography include type and content of organic solvent, salt concentration and pH, in addition to stationary phase properties. A HILIC buffer typically contains more than 70% acetonitrile and uses ammonium acetate or formate [15, 19, 28, 29]. This is because these reagents are volatile and compatible with mass spectrometry. In addition, among the organic solvents tested, acetonitrile was found to exhibit superior chromatography and analyte retention [29]. Salt was shown to increase the hydrophilicity of the liquid layer around the stationary phase and to facilitate analyte retention [18, 30]. Here we focus on optimizing a critical factor of HILIC chromatography, the pH, using bovine serum albumin (BSA) tryptic digest (100 pmol on column). As shown in Figure 2, it is clear that pH 5.8 (Figure 2A) gave much better retention and resolution of the peptides than pH 3.2 (Figure 2B) and pH 6.8 (Figure 2C). The buffer pH directly affects the charge state and hydrophilicity of peptides, and thus the interaction with the HILIC stationary phase. Under pH 3.2 and 6.8 conditions, the peptide hydrophilicity and electrostatic interaction with the stationary phase were not optimal for retention and separation on HILIC. Therefore, pH 5.8 solvents were chosen for HILIC separation of complex protein digests. Although pH 5.8 gave satisfactory results, it should be noted that the pH conditions can be further optimized by testing more pH units. An optimal pH depends on the pKa values of the peptides and should enable favorable retention and fractionation of most, if not all of the peptides in a sample.

### 3.2 *ICAT peptide fractionation by HILIC and SCX chromatography.*

Currently, many on-line and off-line 2D-LC MS experiments utilize SCX with reverse phase chromatography [22, 30-32]. In standard ICAT protocol, SCX is a critical step before avidin affinity purification of cysteine containing peptides from complex samples (Figure 1). Recent studies have shown that HILIC offers a superior separation mechanism based on retention by hydrophilicity and electrostatic interaction [15, 16, 18, 19]. Here we conducted a direct comparison of SCX and HILIC in the fractionation of peptides derived from guard cells ICAT samples. Control and ABA treated guard cell proteins were combined, digested with trypsin and the resulting peptides were aliquoted for SCX and HILIC separation. As shown in Figure 2, HILIC (Figure 2D) resembled SCX (Figure 2E), but showed enhanced peptide retention and higher peak intensity compared to SCX. Based on the LC chromatograms, the two separation methods exhibited high reproducibility in an independent experiment (Supplemental Figure 2). Since volatile solvent was used in HILIC, the fractions collected can be lyophilized and used directly for further separation, e.g., reverse phase chromatography without the desalting step that could lead to peptide loss (Figure 1). It becomes evident that HILIC provides an

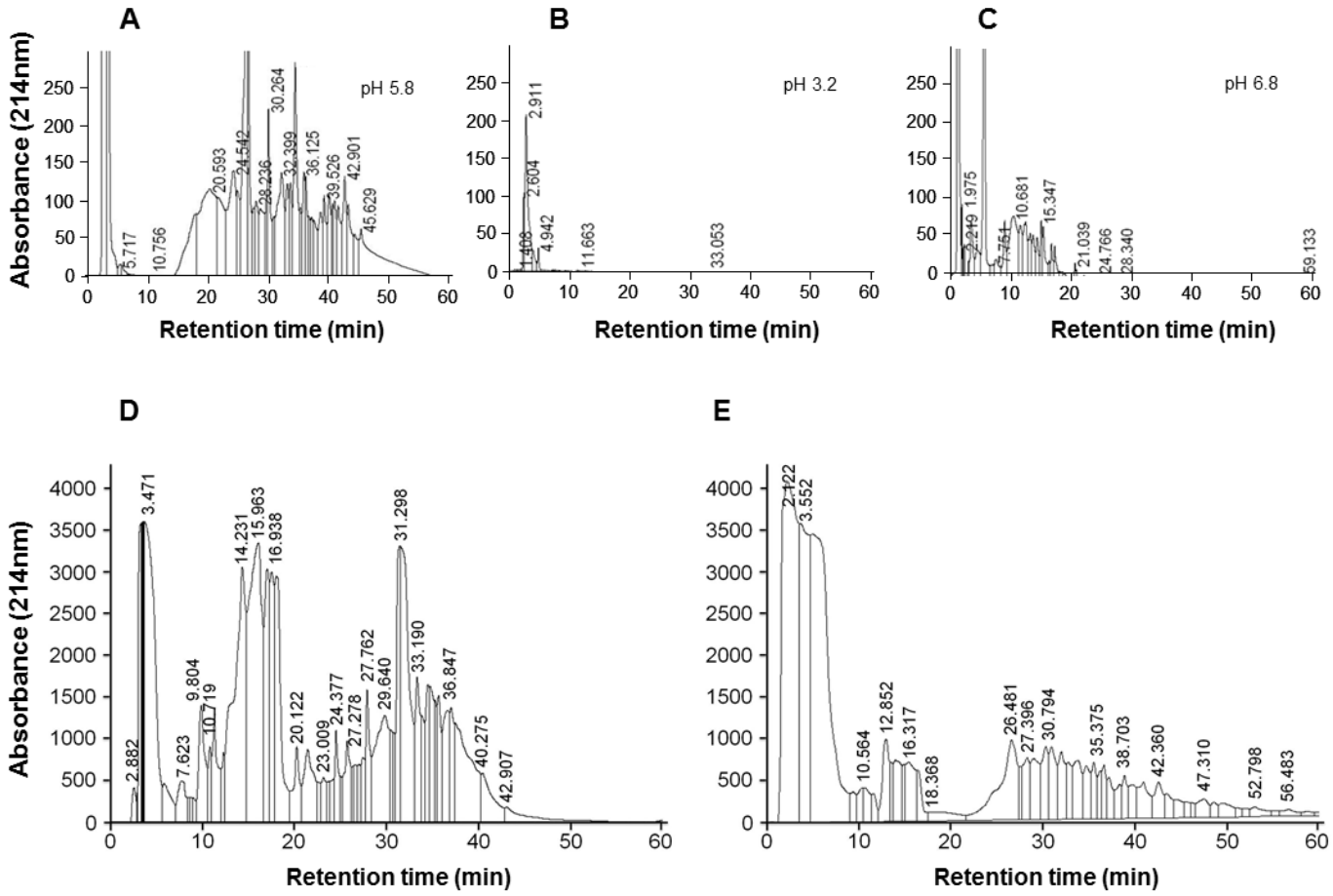


**Figure 2**. Representative HILIC and SCX chromatograms. (A-C) HILIC chromatograms of 100 pmol BSA digest separated using solvents of different pH values. (D) HILIC chromatogram of guard cell protein digest. (E) SCX chromatogram of guard cell protein digest.

excellent alternative to SCX in peptide fractionation of ICAT experiments.

### 3.3 *Protein identification and quantification using HILIC ICAT and SCX ICAT.*

In ICAT experiments, proteins can be identified and relatively quantified between two different samples. In addition, cysteine residues labeled by ICAT reagents can be mapped using the acquired MS/MS spectra (Figure 3). To obtain confident results, two independent experiments were conducted. Each experiment included a HILIC ICAT and a SCX ICAT of two guard cell samples, one control and the other ABA treated. It is interesting to note that in the two replicates of SCX ICAT, 59 and 53 proteins were identified, with only 16 proteins overlapping between the two replicates (Figure 4, Supplemental Table 1). In the two replicates of HILIC ICAT experiments, significantly more proteins, 91 and 77 were identified, with over a half of proteins (54 IDs) overlapping (Figure 4, Supplemental Table 1). These results showed the advantage of conducting replicate experiments [33] and the high reproducibility of the HILIC ICAT workflow. Comparison of both HILIC ICAT and SCX ICAT methods revealed that they are highly complementary. HILIC ICAT and SCX ICAT identified a total of 114 and 96 proteins, respectively, with 41 overlapping and the rest unique to each method. In addition to protein identification and quantification, many ICAT labeled cysteine residues were mapped (Figures 4, Supplemental Table 2). Here we analyze redox responsive cysteines after ABA treatment. Among the 44 peptides containing redox sensitive cysteines, HILIC ICAT identified 40 and SCX ICAT identified 30 peptides. Overall, the changes in ICAT ratios were consistent across replicates (Supplemental Table 2). The two methods produced overlapping qualitative as well as quantitative results (Figure 4, Supplemental Table 2). It is interesting to note that in the two replicates of SCX ICAT, only half of the 30 peptides were reproducibly identified. In the HILIC ICAT experiments, more peptides (38 out of 40) were identified in the both replicates (Supplemental Table 2). These results indicate a higher reproducibility of the HILIC ICAT workflow. Overall, the cysteine specificity of the ICAT technology is an inherent advantage for experiments focused on investigating cysteine modifications [14]. However, it


**Figure 3**. Example of protein identification and quantification result including summary of the protein (A), peptide quantification in the control sample (light) and ABA treated sample (heavy) (B), and peptide MS/MS spectrum indicating heavy ICAT labeled cysteine (C).

may compromise overall protein identification and quantification because non-cysteine containing peptides were excluded from downstream MS analysis. Other technologies such as iTRAQ [21, 22] are an alternative to overcome this limitation.

## 4. Conclusion

HILIC conditions have been optimized and successfully incorporated into the popular ICAT workflow to replace SCX. Considering the LC chromatograms and the number of proteins characterized, HILIC ICAT has clearly exhibited superior performance to SCX ICAT. The improvement can be attributed to HILIC separation based on retention by hydrophilicity and electrostatic interaction, and the unnecessary desalting steps. Interestingly, for either SCX ICAT or HILIC ICAT, replicate experiments increased the number of proteins identified and cysteines mapped. In addition to some common proteins detected by both SCX and HILIC ICAT methods, each identified a unique set of

A

| Experiment | Proteins | Cysteine peptides | Cysteines mapped | Redox peptides |
|---|---|---|---|---|
| SCX01 | 59 | 114 | 117 | 21 |
| SCX02 | 53 | 141 | 181 | 24 |
| HILIC01 | 91 | 386 | 451 | 40 |
| HILIC02 | 77 | 206 | 228 | 39 |

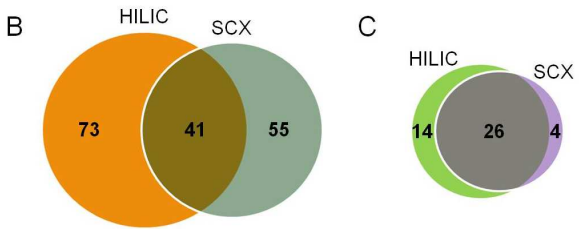


**Figure 4**. Comparison of protein identification and cysteine mapping results using the SCX ICAT and HILIC ICAT methods. (A) Summary of SCX ICAT and HILIC ICAT results. (B) Proteins identified by HILIC ICAT and SCX ICAT. (C) Significant redox sensitive cysteine peptides quantified by HILIC ICAT and SCX ICAT. (Detailed results are listed in Supplemental Tables 1 and 2).

proteins. Therefore, the novel combination of HILIC ICAT with SCX ICAT can significantly enhance the qualitative and quantitative analysis of proteins, especially thiol redox proteins. The usefulness of HILIC has also been found in shotgun proteomics and phosphoproteomics [16, 19] and can be extended to complex sample fractionation for iTRAQ analysis and the newly released six-plex cysteine reactive TMT tags (Thermo Scientific Inc., USA) in due course.

## 5. Supplementary material

Supplementary data and information is available at: http://www.jiomics.com/index.php/jio/rt/suppFiles/83/0

*Supplemental Table 1* - List of proteins and peptides identified in *Brassica napus* guard cells using SCX ICAT and HILIC ICAT methods;

*Supplemental Table 2* - Redox-sensitive proteins and peptides in ABA treated guard cells identified by SCX ICAT and HILIC ICAT. Increased ratios are highlighted in red and decreased ratios in green. Unused score (for protein) is the summed score for peptides that are not claimed by another protein; the % Error represents the error in the calculated ratio, calculated from the error for each of the peaks in the ratio; R in the redox switch stands for reduction while O stands for oxidation;

*Supplemental Figure 1* - Histogram of ICAT ratio distribution of each experiment. (A) SCX-ICAT replicate 1, with average ICAT ratio of 0.87 and % Error 5.48%; (B) SCX-ICAT replicate 2, with average ICAT ratio of 1.08 and % Error 4.42%; (C) HILIC-ICAT replicate 1, with average ICAT ratio of 1.19 and % Error 3.94%; (D) HILIC-ICAT replicate 2, with average ICAT ratio of 1.04 and % Error 4.19%. In all four replicates, the average ICAT ratio is 1.09 with % Error 4.24%.

*Supplemental Figure 2* - HILIC chromatogram (left) and SCX chromatogram (right) of guard cell protein digest after labeling the control sample with ICAT light reagent and ABA treated sample with ICAT heavy reagent. The result was obtained from an experiment independent of that shown in Figure 2D and 2E.

## Acknowledgements

We thank Dr. Cecilia Silva-Sanchez and Carolyn Diaz for critical reading of the manuscript. This work was supported by grants from the National Science Foundation (MCB 0818051) and the National Institute of Health (1S10RR025418-01) to S Chen, and grants from National Natural Science Foundation of China (No. 31071194) and Funds for Distinguished Young Scientists of Heilongjiang Province, China (No. JC201011) to S. Dai.

## References

1. S.P. Gygi, B. Rist, S.A. Gerber, F. Turecek, M.H. Gelb, R. Aebersold, Nat. Biotechnol. 17 (1999) 994-999. DOI:10.1038/1369.
2. Y. Wang, H. Li, S. Chen, Frontiers in Biology 5 (2010) 195-203. DOI:10.1007/s11515-010-0049-y.
3. R. Zhang, F. E. Regnier, J. Proteome Res. 1 (2002) 139-147. DOI: 10.1021/pr015516.
4. H. Zhou, J. A. Ranish, J. D. Watts, R. Aebersold, Nat. Biotechnol. 20 (2002) 512-515. DOI:10.1038/nbt0502-51.
5. K. C. Hansen, G. Schmitt-Ulms, R. J. Chalkley, J. Hirsch, M. A. Baldwin, A. L. Burlingame, Mol. Cell. Proteomics, 2 (2003) 299-314. DOI: 10.1074/mcp.M300021-MCP20.
6. T. Ideker, V. Thorsson, J.A. Ranish, R. Christmas, J. Buhler, J.K. Eng, R. Bumgarner, D.R. Goodlett, R. Aebersold, L. Hood, Science 292 (2001) 929-934. DOI: 10.1126/science.292.5518.92.
7. J. Li, H. Steen, S.P. Gygi, Mol. Cell. Proteomics 11 (2003) 1198-1204. DOI: 10.1074/mcp.M300070-MCP20.
8. U. B. Kang, Y. Ahn, J. W. Lee, Y. H. Kim, J. Kim, M. H. Yu, D. Y. Noh, C. Lee, BMC Cancer 10 (2010) 114. DOI: 10.1186/1471-2407-10-11.
9. J. Jin, G. E. Meredith, L. Chen, Y. Zhou, J. Xu, F. S. Shie, P. Lockhart, J. Zhang, Brain Res. Mol. Brain Res. 134 (2005) 119-138. DOI: 10.1016/j.molbrainres.2004.10.00.
10. T. P. Dunkley, R. Watson, J. L. Griffin, P. Dupree, K. S. Lilley, Mol. Cell. Proteomics 3 (2004) 1128-1134. DOI: 10.1074/mcp.T400009-MCP20.
11. J. A. Ranish, E. C. Yi, D. M. Leslie, S. O. Purvine, D. R. Goodlett, J. Eng, R. Aebersold, Nat. Genet. 33 (2003) 349-355. DOI: 10.1038/ng110.
12. M. Brand, J. A. Ranish, N. T. Kummer, J. Hamilton, K. Igarashi, C. Francastel, T. H. Chi, G. R. Crabtree, R. Aebersold, M. Groudine, Nat. Struct. Mol. Biol. 11 (2004) 73-80. DOI: 10.1038/nsmb71.
13. Y. Shiio, D. W. Rose, R. Aur, S. Donohoe, R. Aebersold, R. N. Eisenman, Mol. Cell. Biol. 26 (2006) 1386-1397. DOI: 10.1128/MCB.26.4.1386-1397.200.
14. C. Fu, J. Hu, T. Liu, T. Ago, J. Sadoshima, H. Li, J. Proteome Res. 7 (2008) 3789-3802. DOI: 10.1021/pr800233.
15. M. Gilar, P. Olivova, A. E. Daly, J. C. Gebler, Anal. Chem. 77 (2005) 6426-6434. DOI: 10.1021/ac050923.
16. P. J. Boersema, N. Divecha, A. J. R. Heck, S. Mohammed, J. Proteome Res. 6 (2007) 937-946. DOI: 10.1021/pr060589.
17. H. Tanak, X. Zho, O. Masayoshi J. Chromatogr. 987 (2003) 119-125. DOI:10.1016/S0021-9673(02)01949-.
18. P. J. Boersema, S. Mohammed, A. J. Heck, Anal. Bioanal. Chem. 391 (2008) 151-159. DOI: 10.1007/s00216-008-1865-.
19. C. P. Albuquerque, M. B. Smolka, S. H., Payne, V. Bafna, J. Eng, H. Zhou, Mol. Cell. Proteomics 7 (2008) 1389-1396. DOI: 10.1074/mcp.M700468-MCP20.
20. P. L. Ross, Y. N. Huang, J. N. Marchese, B. Williamson, K. Parker, S. Hattan, N. Khainovski, S. Pillai, S. Dey, S. Daniels, S. Purkayastha, P. Juhasz, S. Martin, M. Bartlet-Jones, F. He, A. Jacobson, D. J. Pappin, Mol. Cell. Proteomics 3 (2004) 1154-1169. DOI: 10.1074/mcp.M400129-MCP20.
21. M. Zhu, S. Dai, S. McClung, X. Yan, S. Chen, Mol. Cell. Proteomics 8 (2009) 752-766. DOI: 10.1074/mcp.M800343-MCP20.
22. M. Zhu, B. Simons, N. Zhu, D. G. Oppenheimer, S. Chen, J. Proteomics 73 (2010) 790-805. DOI:10.1016/j.jprot.2009.11.00.
23. L. I. Leichert, F. Gehrke, H. V. Gudiseva, T. Blackwell, M. Ilbert, A. K. Walker, J. R. Strahler, P. C. Andrews, U. Jakob, Proc. Natl. Acad. Sci. USA 105 (2008) 8197-8202. DOI: 10.1073/pnas.070772310.
24. S. Alvarez, M. Zhu, S. Chen, J. Proteomics 73 (2009) 30-40. DOI:10.1016/j.jprot.2009.07.00.

25. C. G. Love, A. J. Robinson, G. A. Lim, C. J. Hopkins, J. Batley, G. Barker, G. C. Spangenberg, D. Edwards, Nucleic Acids Res. 33 (2005) D656-659. DOI: 10.1093/nar/gki03.

26. L. Martens, H. Hermjakob, P. Jones, M. Adamski, C. Taylor, D. States, K. Gevaert, J. Vandekerckhove, R. Apweiler, Proteomics 5 (2005) 3537-3545. DOI:10.1002/pmic.20040130.

27. J. A. Vizcaíno, R. Côté, F. Reisinger, H. Barsnes, J. M. Foster, J. Rameseder, H. Hermjakob, L. Martens, Nucleic Acids Res. 38 (2010) D736-742. DOI: 10.1093/nar/gkp96.

28. J. Alpert, J. Chromatogr. 499 (1990) 177-196. DOI:10.1016/S0021-9673(00)96972-.

29. P. Hemstrom, K. Irgum, J. Sep. Sci. 29 (2006) 1784-1821. DOI:10.1002/jssc.20060019.

30. M. P. Washburn, D. Wolters, J. R. Yates, Nat. Biotechnol. 19 (2001) 242-247. DOI:10.1038/8568.

31. M. Q. Dong, J. D. Venable, N. Au, T. Xu, S. K. Park, D. Cociorva, J. R. Johnson, A. Dillin, J. R. Yates, Science 317 (2007) 660-663. DOI: 10.1126/science.113995.

32. M. Gilar, P. Olivova, A. B. Chakraborty, A. Jaworski, S. J. Geromanos, J. C. Gebler, Electrophoresis 30 (2009) 1157-1167. DOI: 10.1002/elps.20080063.

33. P. Chong, C. Gan, T. Pham, P. Wright, J. Proteome Res. 5 (2006) 1232-1240. DOI: 10.1021/pr060018.

JOURNAL OF INTEGRATED OMICS
A METHODOLOGICAL JOURNAL
HTTP://WWW.JIOMICS.COM

ORIGINAL ARTICLE | DOI: 10.5584/jiomics.v2i1.79

# Multivariate methods aid in pinpointing promising tumor marker candidates from colorectal biopsies

Ana M. Rodríguez-Piñeiro[#,*], Paula Álvarez-Chaver[#], Francisco J. Rodríguez-Berrocal, María Páez de la Cadena, Vicenta S. Martínez-Zorzano.

Departamento de Bioquímica, Genética e Inmunología, Facultad de Biología, Universidad de Vigo. As Lagoas-Marcosende s/n, 36310. Vigo, Spain. # These authors contributed equally.

**Received:** 22 November 2012 **Accepted:** 20 January 2012 **Available Online**: 04 February 2012

ABSTRACT

The application of proteomic techniques to the search for disease markers is widely reported nowadays. However, the data rendered by these methods is highly complex and requires mining through statistical methods. Since univariate tests are prone to false positives and require post-test correction, multivariate methods seem more suitable for the task. Here we show an example of their utility, applying both principal component analysis (PCA) and linear discriminant analysis (LDA) to the hydrophobic subproteome of the colorectal mucosa. In order to find proteins specifically altered by colorectal cancer, we compared both the tumor and the adjacent healthy mucosa. PCA followed by variable selection, and corroboration by LDA, pointed out the proteins vimentin and prohibitin as promising candidates for the diagnosis of colorectal tumors.

**Keywords:** Principal component analysis; discriminant analysis; multivariate statistics, colorectal cancer; hydrophobic proteins.

## Abbreviations

**2-DE:** two-dimensional electrophoresis; **FDR:** false discovery rate; **SGoF:** sequential goodness of fit; **PCA:** principal component analysis; **PC:** principal component; **LDA:** linear discriminant analysis; **CRC:** colorectal cancer; **MS:** mass spectrometry.

## 1. Introduction

Proteomics has become a very popular field in the aid of biomarker discovery, mostly because it allows the detection of many marker candidates at a time. This is an important advantage, as until present no single marker has been demonstrated to differentially diagnose one disease. Most of the markers employed nowadays lack either sensitivity or specificity, and thus several markers or other complementary methods (imaging techniques, exploration, etc.) have to be used for the final diagnosis [1].

The way proteomic results have been analyzed, especially those derived from two-dimensional electrophoresis (2-DE) or related methods, has led to an ever-increasing number of potential biomarkers, many of which eventually fail upon validation. Most of the literature available reports these candidates after independent analyses of each variable (for example, after comparison of the mean level of a spot in two study groups by the Student's *t* test). The major hindrance on these comparisons is that they usually consist on a high number of variables (e.g. spots in the case of a 2-DE experiment) and hence require multiple testing. However, the methods commonly employed, called univariate as they test one variable at a time (as the Student's *t* test mentioned), lack power when applied for multiple testing. Thus, there is a need for more efficient methods to analyze these types of datasets. Several alternatives have been proposed, being the simplest to implement some type of post-test correction after

***Corresponding author:** Ana M. Rodríguez-Piñeiro. Present address: Department of Medical Biochemistry and Cell Biology, University of Gothenburg. Box 440, 405 30 Gothenburg, Sweden. E-mail Address: ana.rodriguez@medkem.gu.se.

multiple testing of the independent variables, as the traditional Bonferroni method, the false discovery rate (FDR)-based alternatives [2], or other more recent and powerful methods as the sequential goodness of fit (SGoF) [3]. However, all the variables can be considered at once when using multivariate methods. These tests obviate the need for a post -test correction, thus there is an agreement they are an interesting path for further exploration in proteomic data analysis. To date, many groups have applied these methods with just a profiling or descriptive purpose, trying to find out if the 2-DE maps as a whole contain enough information to distinguish groups of samples [4-7]. Nonetheless these methods can be used to pinpoint one or several specific spots differentiating the sample groups. In this light, there are different and opposing proposals as whether to try to minimize or maximize the number of protein candidates eventually selected.

Recently Marengo *et al*. [8] developed a modification of the principal component analysis (PCA), which they called "Ranking PCA", coupling PCA to a variable selection algorithm that incorporates in each cycle the variable that gives more differential information between the groups. This is based on the "principle of exhaustiveness", aiming to find all the molecules with relevance on the disease studied, even if the information they provide is redundant. On the opposite side, we had published before an application of the PCA [9], which was also based in variable selection, but aiming at finding the minimum number of non-redundant potential markers to differentiate our groups. PCA is a multivariate method for dimension reduction, i.e., it reduces a high number of starting variables (spots in our case) to a new reference system containing a smaller set of variables called principal components (PCs). All the initial variables have to be quantitative and independent. The method calculates the *eigenvalues* and *eigenvectors* of a correlation matrix derived from the original matrix of data. Each PC is the linear combination of the original variables; they are built orthogonal to each other, and they hierarchically explain the maximum possible variance contained in the starting dataset. The final aim of the PCA is to find the smallest group of PCs able to explain the maximum variance from the original data. In that study we also explored the utility of the linear discriminant analysis (LDA) to further investigate the potential of the multivariate methods in highlighting relevant discriminant proteins. The LDA is defined as a classification method, as it allows “classification” or “discrimination” of the sample groups. It also defines which variables are necessary to reach the best classification possible. Thus from the original data it produces a function ("diagnostic function") formed by the linear combination of the variables providing the best (more accurate) classification of the samples into their original groups. These "discriminant variables" or the diagnostic function can then be used to classify new unknown samples. In our group, we use the LDA as a tool to assess the correct classification of samples given by the variables selected by PCA.

In our first description of this approach [9], we applied PCA and variable selection methods to a set of 2-DE maps were *N*-glycoproteins from serum had been resolved. To prove the utility of the method on maps obtained from other types of samples, we now applied the PCA-variable selection strategy to a 2-DE dataset obtained from the comparison of hydrophobic proteins from normal tissues and tumors from colorectal cancer (CRC) patients [10]. From all the potential tissue biomarkers highlighted by the 2-DE comparative strategy, multivariate methods allowed us to narrow down the candidates to two proteins, which were further validated with promising preliminary results.

## 2. Material and Methods

### 2.1. *Comparison of 2-DE maps from healthy and tumoral colorectal mucosa*

The procedure to obtain 2-DE maps from tumoral colorectal mucosa and its neighboring healthy counterpart is described in [10]. Briefly, pairs of samples from 5 patients with CRC were collected. Hydrophobic fractions were extracted by temperature-dependent phase partitioning using the ReadyPrep Protein Extraction Kit (Membrane I) (Bio-Rad), based on the use of Triton X-114. The procedure was repeated to ensure purity of the hydrophobic proteins. These were then separated by 2D-PAGE in 17-cm polyacrylamide gels, along a 4-7 pH gradient. Protein maps were visualized by a silver staining protocol compatible with mass spectrometry (MS) analysis [11]. The reproducibility of the 2D-PAGE for the same sample run in different days was calculated as 85% [10]. Map images were acquired with a GS-800 (Bio-Rad) calibrated densitometer, and compared with the PDQuest 7.1.1 software (Bio-Rad). Protein spots were detected, background was subtracted and images were filtered. The intensity level of a spot in a gel was normalized to the total protein intensity detected for the entire gel. Therefore the spots were relatively quantified and the protein amount was expressed as a relative volume (relative intensity x area of the spot).

The study was approved by the Galician Ethical Committee for Clinical Research (code 2006/326), and complied with the tenets of the Helsinki Declaration, the Oviedo Agreement, the Organic Law for Data Protection 15/1999 and the Royal Decree 1720/2007. Informed consent was obtained from patients or guardians, and anonymity was warranted through the use of clinical history numbers.

### 2.2. *Multivariate analyses*

Levels (relative volumes) of the spots compared between healthy and tumoral mucosa were analyzed by PCA and LDA as previously described [9]. Notice the use of paired samples minimizes the chances of detecting protein differences due to normal individual variability. Furthermore, only those spots present in all the 10 maps obtained from the 5 patients were tested, avoiding the inclusion of null values

and the application of replacement (inference) methods.

For dimension reduction, PCA was applied, and the PCs with *eigenvalues* higher than 1 were selected and considered statistically significant if $p < 0.05$ by the Mann-Whitney *U* test. Next, we examined the correlation matrix (component matrix), which contains the correlation between the original variables and each of the PCs extracted in the analysis. For the significant PCs, we selected the variables (spots) with a correlation higher than 0.8 (that is, where more than 80% of the "spot information" was contributing only to that PC).

In order to evaluate the discriminant function by LDA, we employed a chi-square transformed from Wilks' lambda, so that the classification function was significant (hence "useful") when $p < 0.05$. Results from the LDA were corroborated by leave-one-out cross-validation, removing in turn each one of the 10 cases and classifying them on the basis of the nine remaining ones.

### 2.3. *Identification by mass spectrometry*

Proteins were identified following the MS protocols described before [10]. First, protein digests were submitted to MALDI-TOF/TOF. Those not successfully identified were further processed through Cap-LC-nESI-Q-TOF. All the parameters for database search were kept as described in reference [10].

### 2.4. *Detection of caveolin-1, prohibitin and vimentin*

Total colorectal tissue and both the hydrophobic and the soluble fractions derived from it were resolved in 10% SDS-PAGE gels. Then they were transferred to PVDF membranes by Western blot. The primary antibodies employed were goat anti-human caveolin-1 (1 µg/mL; AbD Serotec); mouse anti-human vimentin Ab-2 (2 µg/mL; clone V9; Neomarkers); and mouse anti-human prohibitin (1 µg/mL; clone II-14-10; Neomarkers). Secondary antibodies were rabbit anti-goat or goat anti-mouse IgG (H+L), conjugated with alkaline phosphatase (1/2,000; Bio-Rad). Gel images were acquired with a GS-800 calibrated densitometer (Bio-Rad) and analyzed with Quantity One (v. 4.4.1, Bio-Rad).

## 3. Results and Discussion

### 3.1. *Origin of the dataset*

In a previous work [10], we extracted hydrophobic proteins from healthy and tumoral colorectal mucosa samples from 5 CRC patients. We then separated these proteins by 2-DE, obtaining 10 maps that were matched and analyzed by the univariate Student's *t*-test for paired samples. The 41 proteins that showed significantly altered levels were submitted to MALDI-TOF/TOF and Cap-LC-nESI-Q-TOF, and 23 of them were identified. From them we chose the protein vimentin for validation, on the basis of its levels, the available literature, and relevance for CRC, showing it was indeed altered in this pathology [10].

### 3.2. *Detection of caveolin-1 corroborates the enrichment of the sample*

To assess the enrichment in hydrophobic proteins in the samples used for the present study, we specifically detected caveolin-1, a plasma membrane protein. Results from the starting tissue were compared against the hydrophobic and hydrophilic fractions (figure 1). Caveolin-1 was detected as a band of approximately 21 kDa, almost 5 times more abundant in the hydrophobic fraction than in the starting tissue homogenate. In the hydrophilic sample only residual contamination was observed.

### 3.3. *Application of PCA and LDA*

The dataset with the relative volume of the 41 spots significantly differing in healthy and tumoral mucosa was analyzed first by PCA. This method allowed a reduction of the original data to 9 PCs, explaining 100% of the total variability (table 1). Applying the Mann-Whitney *U* test to these components, we found PC1 was significant with more than 99% confidence ($p < 0.01$). The plot in figure 2A shows how the PC1 effectively separates the 10 samples in their true original group (healthy mucosa or tumor). Comparison of this result with our previous application of PCA to a proteomic dataset [9] shows that with a complex starting sample (as the serum in that study) the information given by the first principal component is not so relevant (i.e. not statistically significant). However, in the present study the highly-enriched and

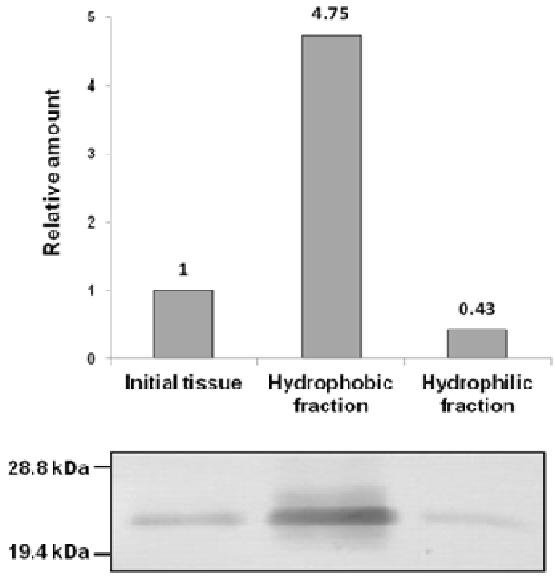


**Figure 1.** Detection of caveolin-1 in the initial tissue, and in the hydrophobic and hydrophilic fractions derived from it. Bars show the relative amount of caveolin-1, considering the starting tissue has a value of 1.

**Table 1.** Principal components (PCs) calculated from the 41 spots altered in colorectal tumors.

| Component | *Eigenvalues* | | |
|---|---|---|---|
| | Total | % Variance | Cumulative % |
| PC1 | 17.184 | 41.913 | 41.913 |
| PC2 | 6.601 | 16.101 | 58.014 |
| PC3 | 4.899 | 11.948 | 69.963 |
| PC4 | 3.977 | 9.701 | 79.663 |
| PC5 | 2.926 | 7.135 | 86.799 |
| PC6 | 2.056 | 5.015 | 91.814 |
| PC7 | 1.454 | 3.545 | 95.359 |
| PC8 | 1.116 | 2.721 | 98.080 |
| PC9 | 0.787 | 1.920 | 100.00 |

less-complex initial sample (hydrophobic proteins, estimated to be 10% of the total starting tissue [10]) surely contributed to our ability to detect the spots providing significant "differential information" through the first principal component. Even samples that have been enriched but are still complex (as in [12]) fail usually to provide such a neat separation through the first principal component.

As shown in table 1 and figure 2A, the separation of the groups given by PC1 is 100% effective. However, the individual contribution of each PC (percentage of variance explained, table 1) is small. Therefore we selected the variables with higher correlation with PC1 (the most informative and significant). A subsequent PCA using only the values of the spots with more than 80% correlation to PC1 (8 variables) resulted in a new set of PCs detailed in table 2. The new PC1, named as PC1', was again significantly different between healthy mucosa and tumor tissue ($p < 0.01$), and explained 72.6% of the variance, reaching with the first three components a 90% cumulative variance. As shown in figure 2B, it also allowed graphical separation of the two groups of tissues. More interestingly, the tumor tissues showed a closer distribution than the normal tissues, indicating a higher homogeneity. This effect has been noticed before by us [9, 12] and other authors [4, 13] in different types of samples and cancers, and highlights the good performance of this set of 8 spots for the diagnosis of the tumor tissue.

From this set of 8 spots selected by the new PCA we could identify the proteins vimentin, alpha-1B-glycoprotein, and prohibitin (table 3). The other 5 spots included in the set could not be identified. Since the alpha-1B-glycoprotein is involved in acute phase and inflammatory processes [14], which are not specific for CRC but appear also in benign pathologies of the colon and rectum, we discarded this protein and repeated the PCA with only the other two identified proteins (vimentin and prohibitin). We obtained again a similar explanation of variability (73% versus the previous 72.6%) by the first PC (named as PC1'' in this case). Again, we found a neat graphical separation of the cases by group (figure 2C). Therefore vimentin and prohibitin by themselves could be as informative as the 8 spots together.

To corroborate this, we applied LDA to the values obtained for vimentin and prohibitin in all our healthy and

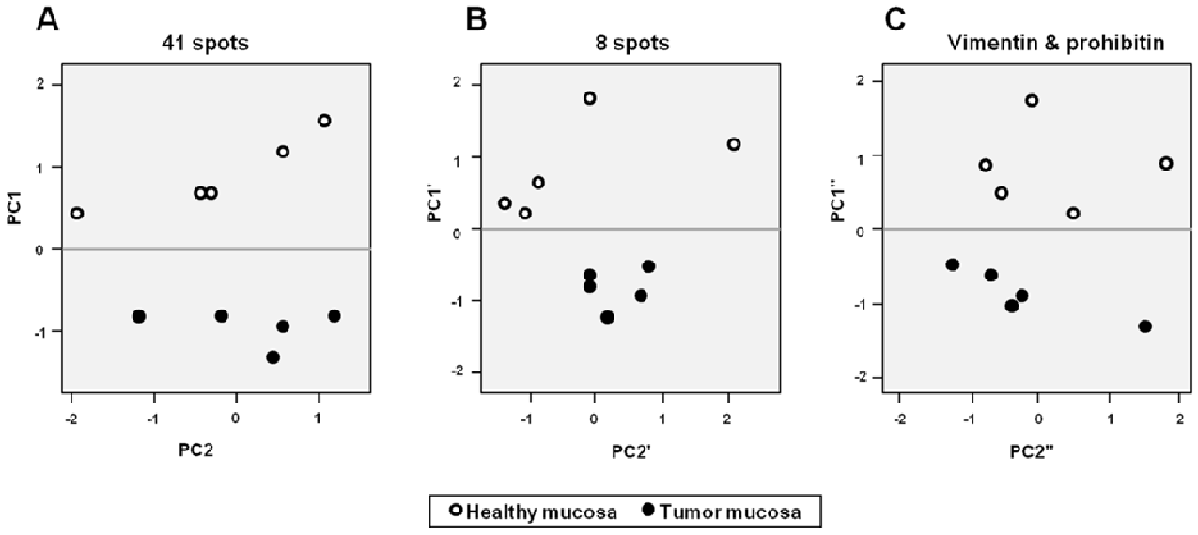


**Figure 2.** Separation of the healthy (open circles) and tumor (filled circles) mucosa samples on the basis of the first (and significant) principal component of each analysis. A) PCA on the 41 spots with different levels in healthy and tumor colonic mucosa; B) PCA on the 8 spots with higher contribution to the PC1 in the previous analysis; C) PCA performed just on the 2-DE spot values for vimentin and prohibitin. In all the analyses, healthy samples obtained positive values for the first (significant) principal component, whereas tumor samples showed negative values.

**Table 2.** Principal components (PCs) calculated from the 8 spots with higher correlation (≥ 80%) with PC1.

| Component | *Eigenvalues* | | |
|---|---|---|---|
| | Total | % Variance | Cumulative % |
| PC1' | 5.805 | 72.568 | 72.568 |
| PC2' | 0.860 | 10.745 | 83.313 |
| PC3' | 0.540 | 6.750 | 90.063 |
| PC4' | 0.426 | 5.328 | 95.391 |
| PC5' | 0.227 | 2.843 | 98.234 |
| PC6' | 0.117 | 1.459 | 99.693 |
| PC7' | 0.022 | 0.279 | 99.972 |
| PC8' | 0.002 | 0.028 | 100.00 |

tumor samples. We found thus a discriminant function explaining 100% of the tissue variability, i.e. the difference between healthy and tumoral mucosa. This function classified each sample in its correct group with a high confidence level (99.8%; $p = 0.002$), even higher than the one obtained for the initial set of 41 proteins (96.6%; $p = 0.034$). In both cases, 100% of the samples were correctly classified after leave-one-out cross-validation.

### *3.4. Validation of vimentin and prohibitin*

Since the multivariate analyses pointed out at vimentin and prohibitin as potential tissue biomarkers to distinguish the healthy colorectal mucosa from its tumoral counterpart, we aimed next to corroborate these results and verify the changes in these two proteins by specific immunodetection.

Vimentin is a type-III intermediate filament ubiquitously expressed by cells of mesenchymal origin, as fibroblasts, chondrocytes, macrophages, and endothelial cells [15]. It seems to act as a scaffolding protein to stabilize connective tissues and cells, or in signal transduction [16-17]. We quantified the relative levels of the spot identified as vimentin in all the maps from healthy and tumor tissues obtained from the 5 CRC patients. As shown in figure 3 (panels A through C), either the vimentin levels were decreased in tumors, or it was altogether absent. Representation of the levels of vimentin in each group of samples allowed us to find a cut-off point, which can distinguish the healthy tissues from the tumors (figure 3B). As mentioned above, the protein vimentin had been studied before by our group [10]. In that occasion, slot and Western blot analyses showed there was a decrease of vimentin in tumor tissues, with at least 3 isoforms of the protein showing different variations in amount. These specific changes were again corroborated in other patients (figure 3D). In certain carcinomas, such as breast cancer or melanoma, vimentin is up-regulated during epithelial-mesenchymal transition [18]. However, this phenomenon has not been observed in CRC; in fact, a reduced expression as the one found by us was observed. These results can be explained by the fact that the vimentin gene has been found methylated and epigenetically silenced in colorectal tumors and adenomas [19-20].

**Table 3.** Characteristics of the 8 spots with high correlation (≥ 80%) with the significant PC1.

| Spot no. (as in [10]) | Fold change in tumor tissue | Protein | Correlation with PC1 (%) |
|---|---|---|---|
| 3 | -7.6 | Vimentin | 80.5% |
| 19 | -5.8 | not identified | 82.8% |
| 20 | -2.5 | Alfa-1B-glycoprotein | 86.6% |
| 21 | -2.4 | not identified | 89.9% |
| 27 | +3.0 | Prohibitin | 80.2% |
| 31 | -10.7 | not identified | 84.2% |
| 35 | +4.9 | not identified | 80.7% |
| 39 | -8.7 | not identified | 81.2% |

Prohibitin is a highly conserved and widely expressed protein. At the subcellular level it has been localized to the cell membrane, mitochondrial inner membrane and cytoplasm, as well as to the nucleus, depending on the cell type and situation. Its subcellular localization influences its multiple roles within the cell. Although the best characterized function of prohibitin is as a chaperone involved in the stabilization of mitochondrial proteins, it has also been implicated in the regulation of proliferation, apoptosis, transcription, and as a cell-surface receptor. Recent data indicated a role of prohibitin in pathogenesis, including its potential involvement in cancer (reviewed in [21]). When we examined the levels of prohibitin in our samples, we found an increment in the amount of protein in 4 of the 5 patients, while the remaining one showed a slight decrease (figure 4). That made it impossible to establish an effective experimental cut-off for the relative levels of the protein in the tumor tissue regarding its healthy counterpart (as shown in figure 4B). Our finding that prohibitin is up-regulated in most of the colorectal tumors is in agreement with previous reports describing an up-regulation in CRC [4, 22]. Recently, Chen *et al.* [23] confirmed the increased expression of prohibitin in the adenoma-carcinoma sequence, only in those adenomas further developed into CRC and not in the non-malignant ones. These results are in line with our observations.

Prohibitin is a good example of the superiority of the multivariate methods over the univariate ones. As mentioned above, it does not vary in the same direction in all the patients analyzed. However, the multivariate approach highlights it not as a marker by itself, but as part of a group (in this case together with vimentin) that provides the best dis-

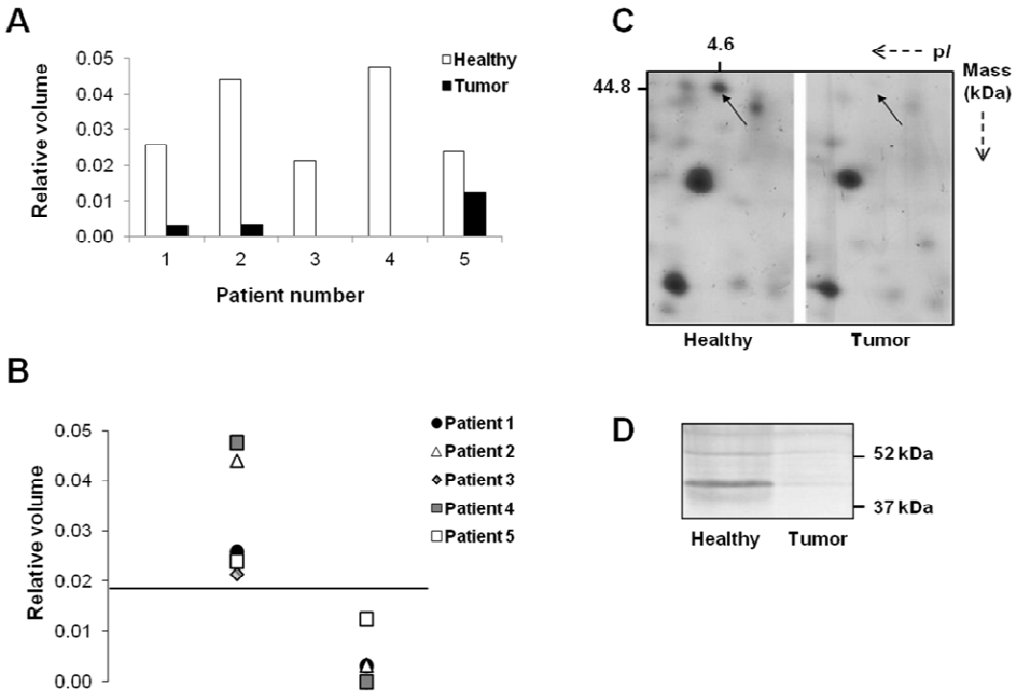


**Figure 3.** Validation of the differential expression of vimentin. A) Vimentin levels in the healthy and tumor tissues corresponding to each CRC patient. B) The difference between the vimentin levels in the healthy and the tumor mucosas can be visualized by a cut-off. C) Areas of 2 -DE maps from a healthy tissue and a tumor where the vimentin spot (pointed by arrows) is located (p*I*: isoelectric point). A decrease in the relative levels is clearly seen in this example. D) Immunodetection of vimentin isoforms in a healthy colorectal tissue and a tumor, corroborating the decrease of the protein levels in the latter.

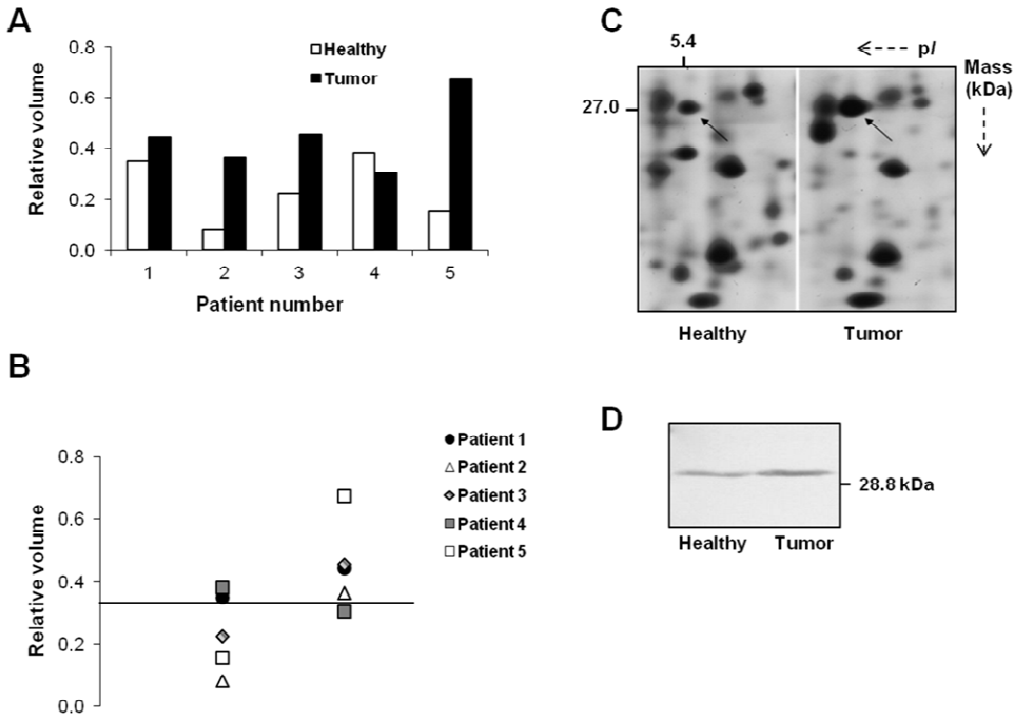


**Figure 4.** Analysis of the prohibitin levels. A) The relative amount of prohibitin increased in 4 of the 5 patients studied. B) A plot of the individual sample values showed it was not possible to establish an experimental cut-off to specifically distinguish healthy tissues from tumors. C) Prohibitin spot (arrow) shown in representative 2-DE maps from a healthy tissue and a colorectal tumor (p*I*: isoelectric point). An increase in the spot levels can be observed in the tumor. D) Specific immunodetection of prohibitin in a healthy colorectal tissue and a tumor.

crimination between the two types of samples studied (healthy *vs* tumoral). If we had done a simple univariate test, we would have found prohibitin was differentially expressed between the two tissues (for instance, $p < 0.05$ by a paired *t*-test), but we would have probably discarded it since it does show the opposite variation in 1 out of 5 cases. Vimentin would have stood successfully the univariate testing ($p < 0.05$); however a detailed examination of figure 3A reveals that patient 5 does not show such a high fold-variation (< 2 fold) as the other sample pairs, though the difference between the tissues is clearly seen in the prohibitin level (figure 4A). This example intuitively shows the major advantage of the multivariate approach, by selecting more than one protein to construct a panel that is expected to be more sensitive than a biomarker alone.

## 4. Concluding Remarks

This study has shown the utility of multivariate methods, especially PCA, for the evaluation of a 2-DE dataset, as well as for the selection of a reduced set of potential markers for the disease. In this case, the proteins vimentin and prohibitin were selected through this approach and showed significant changes in their levels between healthy and tumoral colorectal mucosa. Vimentin was able to correctly classify all the samples studied, whereas prohibitin could not do this alone. However, PCA and LDA showed that both proteins together were able to discriminate correctly 100% of the samples. Further studies will clarify if it is vimentin alone or the combination of both proteins that gives the best results for diagnosis of colorectal tissues.

## Acknowledgements

The authors thank A. Carvajal-Rodríguez for useful discussions, and C. Fiaño at "Complejo Hospitalario-Universitario de Vigo" for providing samples. This work was supported by Xunta de Galicia (grant 10PXIB310215PR) and FEDER funding (grant CN2011/024). AMRP was supported by the Angeles Alvariño program and PAC by a predoctoral fellowship, both from Xunta de Galicia (Spain).

## References

1. M. Sikaroodi, Y. Galachiantz, A. Baranova, Curr. Mol. Med. 10 (2010) 249-257.
2. Y. Benjamini, Y. Hochberg, J. Roy, Stat. Soc. Ser. B 57 (1995) 289-300.
3. A. Carvajal-Rodríguez, J. de Uña-Álvarez, E. Rolán-Álvarez, BMC Bioinformatics 10 (2009) 209.
4. U.J. Roblick, D. Hirschberg, J.K. Habermann, C. Palmberg, S. Becker, S. Krüger, M. Gustafsson, H.P. Bruch, B. Franzén, T. Ried, T. Bergmann, G. Auer, H. Jörnvall, Cell. Mol. Life Sci. 61 (2004) 1246-1255.
5. M. Ciaparrone, M. Quirino, G. Schinzari, G. Zannoni, D.C. Corsi, F.M. Vecchio, A. Cassano, G. La Torre, C. Barone, Oncology 70 (2006) 366-377.
6. E. Ragazzi, S. Pucciarelli, R. Seraglia, L. Molin, M. Agostini, M. Lise, P. Traldi, D. Nitti, J. Mass Spectrom. 41 (2006) 1546-1553.
7. U.J. Roblick, F.G. Bader, C. Lenander, U. Hellman, K. Zimmermann, S. Becker, A. Ost, A. Alaiya, H.P. Bruch, R. Keller, L. Mirow, B. Franzén, T. Ried, G. Auer, J.K. Habermann, Int. J. Colorectal Dis. 23 (2008) 483-491.
8. E. Marengo, E. Robotti, M. Bobba, F. Gosetti, Anal. Bioanal. Chem. 397 (2010) 25-41.
9. A.M. Rodríguez-Piñeiro, F.J. Rodríguez-Berrocal, M. Páez de la Cadena, J. Chrom. B 849 (2007) 251-260.
10. P. Álvarez-Chaver, A.M. Rodríguez-Piñeiro, F.J. Rodríguez-Berrocal, V.S. Martínez-Zorzano, M. Páez de la Cadena, Int. J. Biochem. Cell Biol. 39 (2007) 529-540.
11. A. Shevchenko, M. Wilm, O. Vorm, M. Mann, Anal. Chem. 68 (1996) 850-858.
12. P. Álvarez-Chaver, A.M. Rodríguez-Piñeiro, F.J. Rodríguez-Berrocal, A. García-Lorenzo, M. Páez de la Cadena, V.S. Martínez-Zorzano, J. Proteomics 74 (2011) 874-886.
13. B. Franzén, S. Linder, A.A. Alaiya, E. Eriksson, K. Uruy, T. Hirano, K. Okuzawa, G. Auer, Br. J. Cancer 74 (1996) 1632-1638.
14. S. Golbabapour, W.W. Pang, J. George, T. Pasupati, P.S. Abdul -Rahman, O.H. Hashim, Int. J. Mol. Sci. 12 (2011) 1030-1040.
15. M. Osborn, J. Invest. Dermatol. 81 (1983) 104s-109s.
16. H. Herrmann, M. Hesse, M. Reichenzeller, U. Aebi, T.M. Magin, Int. Rev. Cytol. 223 (2003) 83-175.
17. J.E. Eriksson, A.V. He, A.V. Trejo-Skalli, A.S. Harmala-Brasken, J. Hellman, Y.H. Chou, R.D. Goldman, J. Cell Sci. 117 (2004) 919-932.
18. T. Brabletz, F. Hlubek, S. Spadena, O. Schmalhofer, E. Hiendlmeyer, A. Jung, T. Kirchner, Cells Tissues Organs 179 (2005) 56-65.
19. H. Zou, J.J. Harrington, A.M. Shire, R.L. Rego, L.Wang, M.E.Campbell, A.L. Oberg, D.A. Ahlquist, Cancer Epidemiol. Biomarkers Prev. 16 (2007) 2686-2696.
20. A. Shirahata, M. Sakata, K. Sakuraba, T. Goto, H. Mizukami, M. Saito, K. Ishibashi, G. Kigawa, H. Ne moto, Y. Sanada, K. Hibi, Anticancer Res. 29 (2009) 279-281.
21. A.L. Theiss, S.V. Sitaraman, Biochim. Biophys. Acta 1813 (2011) 1137-1143.
22. J. Stulik, L. Hernychova, S. Porkertova, J. Knízek, A. Macela, J. Bures, P. Jandis, J.I. Langridgem, P.R. Jungblut, Electrophoresis 22 (2001) 3019-3025.
23. D. Chen, F. Chen, X. Lu, X. Yang, Z. Xu, J. Pan, Y. Huang, H. Lin, P. Chi, Int. J. Oncol. 37 (2010) 355-365.

JOURNAL OF INTEGRATED OMICS
A METHODOLOGICAL JOURNAL
HTTP://WWW.JIOMICS.COM

ORIGINAL ARTICLE | DOI: 10.5584/jiomics.v2i1.80

# *In silico* directed mutagenesis using software for glycosylation sites prediction as a new step in antigen design

Vladislav Victorovich Khrustalev*[1] and Eugene Victorovich Barkovsky[1].

[1]Department of General Chemistry, Belarussian State Medical University, Belarus, Minsk, Dzerzinskogo, 83

**Received:** 29 September 2011 **Accepted:** 09 February 2012 **Available Online**: 22 February 2012

ABSTRACT

*In silico* directed mutagenesis with the aim to estimate consequences of mutational pressure on number of N- and O-glycosylation sites has been proposed as an important step in antigen design. Using this kind of methodology one is able to estimate probabilities at which N- and O -glycosylation sites can be destroyed and created due to one-step missense mutations caused by mutational pressure in the subsequent gene and to select a region of a protein with a lowest probability of new glycosylation site occurrence. Mutational AT-pressure has been simulated in the region of *env* gene coding for HIV1 gp120. Consequences of 741 amino acid substitutions caused by missense GC to AT mutations have been predicted with the help of NetNGlyc 1.0 and NetOGlyc 3.1 algorithms. The probability of O-glycosylation site destruction (2.16%) in HIV1 gp120 protein due to a single missense GC to AT mutation in *env* gene is higher than the probability of a new site creation (0.40%). The probability of N-glycosylation site destruction in HIV1 gp120 protein is equal to the probability of its creation (5.53%), while the number of N-glycosylation sites which can be created due to a single missense GC to AT mutation in *env* gene is 1.27 times higher than the number of N-glycosylation sites which can be destroyed

**Keywords:** HIV1; N-glycosylation; O-glycosylation; gp120; mutational pressure; B-cell epitopes .

## 1. Introduction

Practical approach of the mutational pressure theory created by Noboru Sueoka [1] has been developed in our recent work [2]. Levels of mutability for different conformational B -cell epitopes of the same protein can be compared based on knowledge of mutational pressure directions in the subsequent gene [2]. The less mutable epitope is encoded by a part of a gene which is protected from missense nucleotide mutations occurrence better than others due to certain features of its nucleotide content and composition [2].

According to our results, there is AT-pressure in the *env* gene of Human immunodeficiency virus type 1 (HIV1), while the rates of G to A transitions in that gene are higher than rates of C to T(U) transitions, and the rates of C to A transversions are higher than rates of G to T(U) transversions [2].

Symmetric mutational AT-pressure leads to the decrease of the quantity and length of linear B-cell epitopes [3, 4]. However, antigenic properties of glycoproteins depend both on existence of immunogenic amino acid stretches forming conformational epitopes and on existence of glycans connected with them [5]. Glycan can be connected with protein via "-OH" group of serine or threonine side chain (this type of glycosylation is known as O-glycosylation) and via "-$NH_2$" group of asparagine side chain (this type of glycosylation is known as N-glycosylation) [5, 6]. Side chains of some other natural and modified amino acids can also be glycosylated. It is known that many hydroxylisine residues of collagen are O -glycosylated [5]. Rare cases of O-glycosylation via tyrosine and hydroxyproline side chains have been described, as well as a single (to this date) case of N-glycosylation via arginine

***Corresponding author:** Vladislav Victorovich Khrustalev; address: Belarus, Minsk, 220029, Communisticheskaya 7-24; telephone: 80172845957; E-mail Address: vvkhrustalev@mail.ru

side chain [5, 6]. Software for prediction of sites for those rare types of O- and N-glycosylation has not been created yet due to the insufficient volume of data.

The purpose of this study was to simulate AT-pressure in the region of *env* gene coding for HIV1 gp120, to estimate its influence on frequencies of N-glycosylation (via asparagine residues) and O-glycosylation (via serine and threonine residues) sites and to select B-cell epitope of gp120 with lowest probability of a new glycosylation site occurence.

Sites for O-glycosylation have been predicted by NetOGlyc 3.1 software [7] (http://www.cbs.dtu.dk/services/NetOGlyc). This software producing neural network predictions based on 299 known and verified mucin-type O-glycosylation sites has already been used in theoretical study [8].Sites for N-glycosylation have been predicted by NetNGlyc 1.0 software (http://www.cbs.dtu.dk/services/NetNGlyc), which has also showed good performance in several bioinformatical works [9, 10, 11].

Peptides corresponding to those B-cell epitopes of viral or bacterial proteins which are not glycosylated and have low probability to acquire a new site for glycosylation due to mutational pressure should show better performance as antigens for ELISA test systems and as components of synthetic vaccines. Immunization to that kind of peptides should decrease the probability of immune escaping by the way of mutation creating a new glycosylation site. Those B-cell epitopes which can acquire N- or O-glycosylation site due to a single amino acid substitution caused by mutational pressure should be excluded from antigen design study on its early *in silico* step. It will help future investigators to save time and funds for their *in vitro* and *in vivo* experiments (peptide synthesis, testing antigenic properties of peptides in ELISA, immunization of laboratory animals, affinity purification of antibodies and, finally, clinical trials).

## 2. Material and Methods

As a material we used nucleotide sequence of the HIV1 *env* gene region coding for gp120 protein from the reference strain of that virus [NC_001802]. We introduced all possible point missense nucleotide mutations of GC to AT direction in that coding region. Then consequences of each of those possible 741 amino acid substitutions have been predicted with the help of NetOGlyc 3.1 [7] and NetNGlyc 1.0 algorithms.

Introduction of point missense nucleotide mutations of GC to AT direction has been performed with a help of simple but useful original MS Excel algorithm entitled "Mutational Pressure Simulator". To use this software available via our webpage (www.barkovsky.hotmail.r) one should enter nucleotide sequence of a protein coding region in a special cell on its "nucleotide sequence" list. Then one should enter certain codon in which mutation should be introduced as well as resulting codon in cells on the same list. The set of amino acid sequences with introduced mutations can be found in a column of the "nucleotide sequence results" list. Each of those sequences possesses amino acid substitution resulting from single codon mutation. The algorithm uses universal genetic code to translate nucleotide sequences into amino acid sequences. In case if there are some deviations from universal genetic code in the genome of given specie, the code used by the algorithm may be changed manually. To make this operation one should introduce corrections into the genetic code table on the "genetic code" list. The output of the "Mutational Pressure Simulator" algorithm (amino acid sequences) is in FASTA format. It means that all the resulting sequences may be copied from the "nucleotide sequence results" list and pasted into the special field of NetNGlyc 1.0 or NetOGlyc 3.1.

"Mutational Pressure Simulator" is also able to introduce single amino acid mutations in the amino acid sequence entered in the cell on the "amino acid sequence" list. The set of sequences with introduced mutations can be found in a column of the "AA results" list in FASTA format.

Information from the output of NetNGlyc 1.0 and NetOGlyc 3.1 used in this study includes: 1) number of sites for N - and O-glycosylation in each mutated sequence; 2) location of a new N-glycosylation or O-glycosylation site; 3) location of a site for N-glycosylation or O-glycosylation which has been destroyed due to a single amino acid substitution.

To calculate total probability of N-glycosylation site creation due to a single amino acid substitution caused by AT-pressure we divided the number of amino acid substitutions creating new sites for N-glycosylation by the total number of possible amino acid substitutions produced by missense GC to AT mutations. Other probabilities, including those for each glycosylation site destruction and creation, have been calculated in a similar way.

## 3. Results

### 3.1 *Consequences of GC to AT missense mutations in env gene for number of O-glycosylation sites in HIV1 gp120 protein*

According to the results of NetOGlyc 3.1 prediction, there is a single site for O-glycosylation in HIV1 gp120 protein from reference strain. This site is quite unstable under the pressure of GC to AT nucleotide mutations (see Figure 1). Sixteen amino acid substitutions in the area near that site lead to its disappearance (the probability of its destruction due to a single missense GC to AT mutation is equal to 2.16%). Just two of those mutations lead to the replacement of threonine residue itself (See Supplementary material, Tables 1-4). Six of those mutations lead to the replacement of alanine residue near threonine, and four – to the replacement of proline residue. This data is in consistence with the known fact that threonine and serine residues are O-glycosylated in case if they are surrounded by proline and alanine residues [7]. Both proline and alanine are encoded by GC-rich codons (CCX and GCX, respectively).

Interestingly, only a single O-glycosylation site can be created in gp120 protein due to one-step missense GC to AT

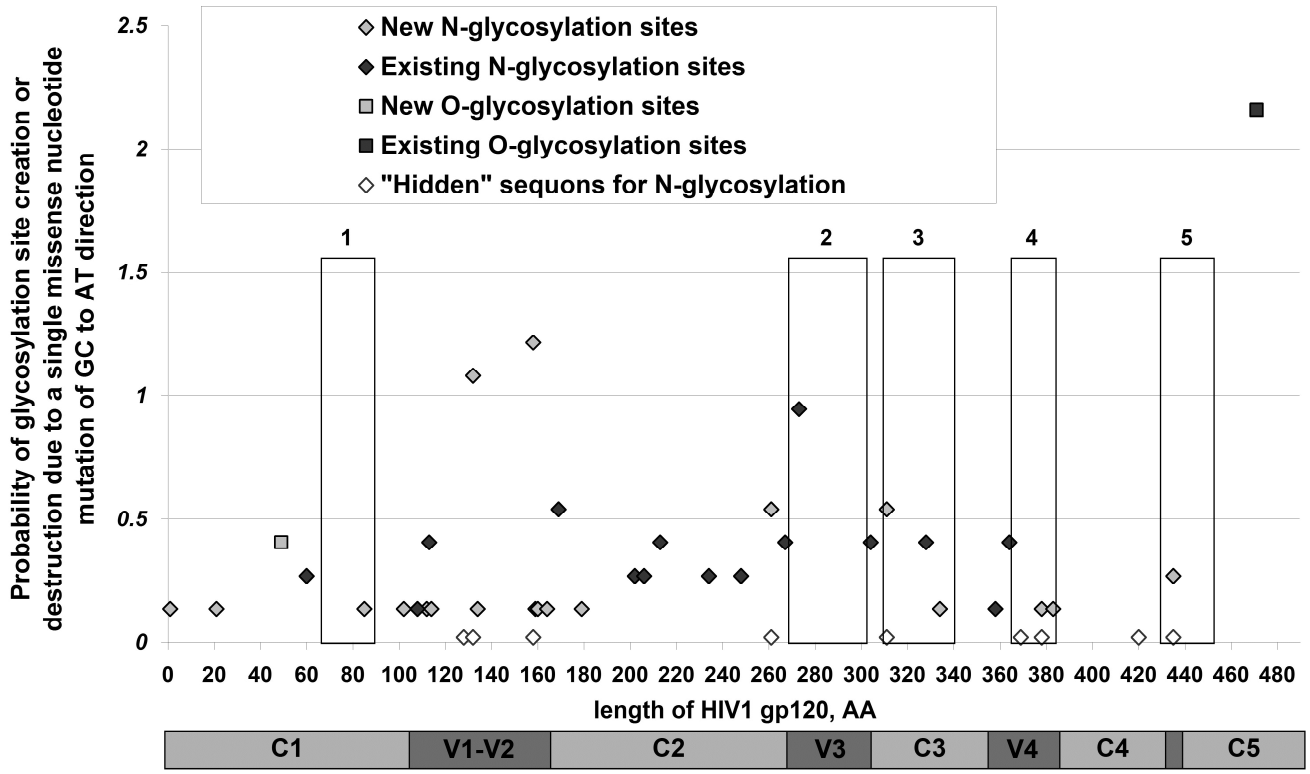


**Figure 1.** Probabilities of creation and destruction for N- and O-glycosylation sites along the length of gp120 protein from HIV1 reference strain. Five comformational epitopes [2] are designated by bars. Boarders of canonical conserved (C) and variable (V) regions of gp120 are provided. "Hidden sequons" for N-glycosylation are shown

mutations in the region of *env* gene coding for it. The probability of creation for that O-glycosylation site (0.40%) is 5.4 times lower than the probability of destruction for existing O -glycosylation site. New O-glycosylation site can appear in the sequence "P**T**DPNP" enriched by proline residues situated in C1 region of gp120.

### 3.2 *Consequences of GC to AT missense mutations in env gene for number of N-glycosylation sites in HIV1 gp120 protein*

According to the results of NetNGlyc 1.0 predictions, the probability of N-glycosylation site creation in HIV1 gp120 protein due to a single missense GC to AT mutation in *env* gene (5.53%) is exactly the same as the probability of N-glycosylation site destruction. Since asparagine is encoded by GC-poor codons (AAT/C), its level of usage increases due to GC to AT mutations. However, consensus sequence for N-glycosylation site (also known as "sequon") is "Asn-Xaa-Thr/Ser" (where Xaa is not Pro). It means that serine and threonine replacements due to AT-pressure lead to the destruction of N-glycosylation sequons.

There are fifteen sites for N-glycosylation in gp120 protein (according to NetNGlyc 1.0 prediction), each of which can be destroyed by AT-pressure due to replacement of serine or threonine residue in its sequon. However, some of those sites can be destroyed due to many other amino acid replacements (See Supplementary material, Tables 1-4). The most unstable N-glycosylation site is situated in the V3-loop of gp120: it can be destroyed due to seven different amino acid replacements caused by AT-pressure both in the sequon itself and in the area near it (see Figure 1).

The number of N-glycosylation sites which can appear due to AT-pressure is 1.27 times higher than the number of sites which can be destroyed. Indeed, since serine and threonine are encoded by codons of average GC-content, they can not only disappear, but also appear due to GC to AT missense mutations. So, in general, AT-pressure leads to the increase of the number of N-glycosylation sites in gp120 protein.

### 3.3 *Consequences of different types of GC to AT missense mutations in env gene for number of N-glycosylation sites in HIV1 gp120 protein*

For 215 possible missense G to A transitions (see Supplementary material, Table 1) the probability of N-glycosylation site creation is higher than that for N-glycosylation site destruction (4.19% versus 2.33%). Those one-step transitions can create nine and destroy four N-glycosylatyion sites.

Since G to A transitions occur in HIV1 genes more frequently than other types of GC to AT mutations, in general, this specific mutational A-pressure [12] should lead to the growth of the number of N-glycosylation sites in gp120 protein. Obviously, amino acid substitutions of Ser2 to Asn direction can not only create new sites for N-glycosylation due to Asn appearance, but also destroy previously existed ones due to Ser disappearance (See Supplementary material, Table 1).

For 128 possible missense C to T(U) transitions (see Supplementary material, Table 2) the probability of N-glycosylation site creation is lower than that for N-glycosylation site destruction (5.47% versus 11.72%). Those one-step transitions can create four and destroy ten N-glycosylatyion sites. Two thirds of destructive missence C to T(U) mutations lead to Thr to Ile replacement (see Supplementary material, Table 2). Sequons for N-glycosylation containing threonine in their third positions can be destroyed by this kind of amino acid substitution, as well as by Thr to Met substitution which may also be caused by missence C to T(U) mutation.

For 187 possible missense C to A transversions (see Supplementary material, Table 3) the probability of N-glycosylation site creation is also lower than that for N-glycosylation site destruction (7.49% versus 8.56%). Those one-step transversions can create nine and destroy twelve N-glycosylatyion sites. The main cause of the destructive effect of missence C to A transversions is in their ability to cause Thr to Lys amino acid substitutions which can destroy sequons for N-glycosylation (see Supplementary material, Table 3).

For 220 possible missense G to T(U) transversions (see Supplementary material, Table 4) the probability of N-glycosylation site creation is higher than that for N-glycosylation site destruction (5.00% versus 2.28%) mostly due to their ability to cause Lys to Asn substitutions. Those one-step transversions can create eight and destroy four N-glycosylatyion sites.

The probability to be N-glycosylated for a given site may become low not only due to direct destruction of the sequon (due to Asn, Ser or Thr disappearance), but also due to other types of amino acid substitutions. Theoretically, some of those amino acid substitutions should be able to make the region containing sequon less hydrophilic. For example, Pro to Leu4 substitution caused by C to T(U) mutation drastically decreased the score for N-glycosylation of a certain sequon (see Supplementary material, Table 2). As we have found out, C to T(U) mutations are prone to decrease the score of linear B-cell epitopes predicted by BepiPred 1.0 [13]. In general, missense C to T(U) mutations should have a destructive effect on antigenic determinants (both glycosylated and not glycosylated) helping the virus to escape humoral immune answer [3]. However, the percent of destructive amino acid replacements which are unable to destroy a sequon for N-glycosylation is relatively low (7.3%). In contrast, the percent of amino acid substitutions which are unable to create a new sequon but have made the score of the sequon higher than the threshold is equal to 53.7%. In other words, mutational AT-pressure destroys sequons for N-glycosylation mostly in direct manner (causing replacements of Asn, Ser and Thr), while it can 1) create new sequons for N-glycosylation and 2) make "hidden sequons" more suitable for that kind of posttranslational modification. "Hidden sequons" themselves may also be consequences of AT-pressure which increases the level of asparagine usage in proteins. They should be abundant in proteins encoded by GC-poor genes, such as HIV1 gp120 protein. Indeed, there are 24 sequons for N-glycosylation in the gp120 protein. All of those sequons may be glycosylated *in vitro* [14]. However, according to NetNGlyc 1.0 predictions, only 15 of them have a high probability to be glycosylated.

## 4. Discussion

### 4.1 *The role of in silico directed mutagenesis using software for glycosylation sites prediction in antigen design studies*

The final aim of antigen design study is in the development of new components for vaccines. Synthetic and recombinant vaccines are usually based on peptides corresponding to short fragments of proteins exposed on a surface of virions or bacterial cells. The main idea of synthetic vaccine creation is in the possibility to immunize person against the conserved antigenic determinant shared by the most of the strains of the given pathogen. To make this idea work well one should carefully select antigenic determinant of the protein of interest. The first step of that selection usually includes mapping of B-cell epitopes. There are numerous bioinformatical methods able to predict B-cell epitopes (either liner [13, 15, 16] or conformational [17, 18]) or regions of a protein exposed to a solvent [19]. However synthetic peptides corresponding to strongest B-cell epitopes often can not be recognized by antibodies against fragments of the native molecule and *vice versa*. One of the causes of the lack of immunological cross-reactivity is in the fact that many strong B-cell epitopes are mapped in regions containing proline and glycine residues [4]. Those residues usually form beta-turns which are situated on a surface of a molecule. However those beta-turns are formed mostly during the folding of the whole molecule. Short peptides corresponding to them usually have quite different conformations. So, one should select those fragments of a protein which conformation and secondary structure should not depend on long-distance interactions with other parts of a native full-length molecule, even though their score of immunogenicity is some lower than for other fragments [2].

The second step of antigen design study usually includes searching for conserved B-cell epitopes in a protein of interest. This step is usually based on alignment of amino acid sequences of that protein from different strains of the same pathogen [2]. As a result, B-cell epitope the structure of which is under the influence of the stronger negative selec-

tion should be determined.

In our works [2, 20] we introduced another step of antigen design studies based on mutational pressure theory. The aim of this step is in the selection of the less mutable B-cell epitope. To select the less mutable B-cell epitope one should estimate the most frequent types of nucleotide mutations in a gene coding for a protein of interest (main directions of the mutational pressure) and compare levels of mutability for regions coding for B-cell epitopes. The less mutable B-cell epitope should be encoded by a region of a gene with the lowest level of missense sites for the most frequent nucleotide mutations and the highest level of synonymous sites for them [2]. The probability to be missense for the most frequent types of nucleotide mutations should be low in a region of a gene coding for the less mutable B-cell epitope [2].

In the current article we introduce yet another important step of antigen design study which is also based on mutational pressure theory [1]. It is known that appearance of new sites for glycosylation in viral epitopes leads to the loss of protective effect in case of immunization against recombinant or synthetic peptides [21]. That is why it is important to estimate a risk of a new site for glycosylation appearance due to mutations of a preferable direction in each of the conformational epitopes of a protein of interest before the start of the vaccine design project. Synthetic peptides corresponding to B-cell epitopes which are glycosylated or have a high probability to be glycosylated after a single amino acid substitution caused by mutational pressure should not be used as antigens for vaccine development.

In case if all the conformational epitopes already possess sites for glycosylation, an epitope with the most stable site for glycosylation should be chosen for production of recombinant peptide in eukaryotic cells. Synthetic peptide can also be conjugated with a certain glycan. This kind of synthetic vaccine may be developed in case if the structure of a glycan is constant. It has been shown that different glycans are attached to different N-glycosylation sites of HIV1 gp120 protein in different cell lines [21, 22].

### 4.2 *Selection of the best antigen for vaccine design study using results of in silico directed mutagenesis with software for glycosylation sites prediction*

Five conformational B-cell epitopes have been mapped by us on HIV1 gp120 protein in the previous study [20]. For this purpose DiscoTope 1.2 [17], Epces [18] and Epitopia [19] algorithms have been used. In the current work we compared their suitability for inclusion in antigen design study based on data received from *in silico* directed mutagenesis session with software for glycosylation sites prediction.

In Figure 1 probabilities of destruction and creation are given for each N-glycosylation site along the length of gp120 protein from reference HIV1 strain. The highest probability of new N-glycosylation site appearance due to AT-pressure is characteristic to V1-V2 region of gp120.

There are no N-glycosylation sites in two from five conformational B-cell epitopes of gp120 predicted by us [2] (in epitope 1 from C1 region and in epitope 5 from C4-V5-C5 region). AT-pressure can create a single N-glycosylation site in epitope 1 at a probability which is two times lower than that for creation of a single N-glycosylation site in epitope 5. Two N-glycosylation sites can appear due to AT-pressure in epitope 3 (this region plays important role in CD4 receptor binding) and epitope 4 (in highly variable V4-loop).

Five sites for N-glycosylation which can appear due to several single amino acid substitutions are represented by "hidden sequons". Different types of amino acid substitutions near those sequons may drastically increase their probabilities to be N-glycosylated. In contrast, probabilities to be N-glycosylated for three other "hidden sequons" did not become higher than the threshold during the current session of *in silico* mutagenesis.

"Hidden sequons" which may become N-glycosylated due to a single amino acid substitution caused by mutational AT-pressure can be found in epitope 3, epitope 4 and epitope 5 (see Figure 1). Moreover, epitope 4 possesses yet another "hidden sequon" which cannot become suitable for N-glycosylation at least due to a single amino acid substitution caused by mutational AT-pressure.

Results showed that epitope 1 is more suitable for antigen design than four other epitopes predicted by us [2]. That region is not glycosylated, it has no "hidden sequons" for N-glycosylation and the probability of N-glycosylation site creation due to AT-pressure is lower for it than for other conformational B-cell epitopes. Moreover, this epitope is the less mutable one [2].

Results of our *in vitro* experiments showed that there are antibodies able to cross-react with the peptide NQ21 corresponding to the consensus sequence of that epitope in blood of 80.22% of persons with currently diagnosed HIV1-infection [2]. This high level of sensitivity for an ELISA test system based on the short NQ21 peptide conjugated with biotin could not be reached in case of a high probability of new N-glycosylation site appearance in the epitope 1 of gp120.

## 5. Concluding Remarks

*In silico* directed mutagenesis with estimation of the mutational pressure consequences on number of N- and O-glycosylation sites is an important step in the process of antigen design. Short peptides corresponding to those epitopes which are not glycosylated and have a lowest probability of new N-glycosylation site appearance due to mutational pressure are recommended for usage as new vaccine components and antigens for ELISA test systems.

## 6. Supplementary material

Supplementary data and information is available at: http://www.jiomics.com/index.php/jio/rt/suppFiles/80/0

Table 1.Consequences of G to A missense mutations.
Table 2. Consequences of C to T(U) missense mutations.
Table 3. Consequences of C to A missense mutations.
Table 4. Consequences of G to T(U) missense mutations.

## References


1. N. Sueoka, Proc. Natl. Acad. Sci. USA 85 (1988) 2653–2657.
2. V. V. Khrustalev, E.V. Barkovsky, A.E. Vasilevskaya, S.M. Skripko, V.L. Kolodkina, G.M. Ignatyev, P.A. Semizon, JIOMICS 1 (2011) DOI: 10.5584/jiomics.v2011i2011.64
3. V. V. Khrustalev, Molecular Immunology 47 (2010) 1635-1639. doi:10.1016/j.molimm.2010.01.006
4. V. V. Khrustalev, E.V. Barkovsky, J. Theor. Biol., 282 (2011) 71-79. doi:10.1016/j.jtbi.2011.05.018
5. K. Drickamer, M.E. Taylor, Introduction to Glycobiology, second ed., Oxford University Press, USA, 2006.
6. M. Lommel, S. Strahl, Glycobiology, 19 (2009) 816–828.
7. K. Julenius, A. Mølgaard, R. Gupta and S. Brunak, Glycobiology, 15 (2005) 153-164.
8. J. J Calvete, L. Sanz, Methods Mol. Biol. 446 (2008) 281-292.
9. L. Wang, F. Li, W. Sun, S. Wu, X. Wang, L. Zhang, D. Zheng, J. Wang and Y. Gao, Mol. and Cell. Proteomics 5 (2006) 560-562. doi: 10.1074/mcp.D500013-MCP200
10. S. K. Saxena, N. Mishra, R. Saxena, M. L. A. Swamy, P. Sahgal, S. Saxena, S. Tiwari, A. Mathur, M. P. Nair, J. Infect. Dev. Ctries 4 (2010) 1-6.
11. A. K. Tomar, B. S. Sooch, S. Yadav, Bioinformation 7 (2011) 69-75.
12. B. Berkhout, F.J. van Hemert, Nucleic Acids Res., 22 (1994) 1705–1711.
13. J. E. P. Larsen, O. Lund, M. Nielsen, Immunome Res. 2 (2006) 2.
14. C. K. Leonard, M. W. Spellman, L. Riddle, R. J. Harris, J. N. Thomas and T. J. Gregory, J. Biol. Chem. 265 (1990) 10373-10382.
15. T. P. Hopp, K. R. Woods, Mol. Immunol. 20 (1983) 483–489.
16. J. Kyte, R. Doolittle, J. Mol. Biol. 157 (1982) 105-132.
17. P. H. Andersen, M. Nielsen, O. Lund, Protein Science 15 (2006) 2558–2567.
18. S. Liang, D. Zheng, C. Zhang, M. Zacharias, BMC Bioinformatics 10 (2009) 302.
19. N. D. Rubinstein, I. Mayrose, E. Martz, T. Pupko, BMC Bioinformatics 10 (2009) 287.
20. V. V. Khrustalev, Immunological Investigations 39 (2010) 551-569.
21. M. Raska, K. Takahashi, L. Czernekova, K. Zachova, S. Hall, Z. Moldoveanu, M. C. Elliott, L. Wilson, R. Brown, D. Jancova, S. Barnes, J. Vrbkova, M. Tomana, P. D. Smith, J. Mestecky, M. B. Renfrow and J. Novak, J. Biol. Chem. 285 (2010) 20860-20869. doi: 10.1074/jbc.M109.085472.
22. T. Mizuochi, T. J. Matthews, M. Kato, J. Hamako, K. Titani, J. Solomon and T. Feizi, J. Biol. Chem. 265 (1990) 8519-8524.

JOURNAL OF INTEGRATED OMICS
A METHODOLOGICAL JOURNAL
HTTP://WWW.JIOMICS.COM

ORIGINAL ARTICLE | DOI: 10.5584/jiomics.v2i1.85

# Elucidation of carbon transfer in a mixed culture of *Acidiphilium cryptum* and *Acidithiobacillus ferrooxidans* using protein-based stable isotope probing

René Kermer[1], Sabrina Hedrich[2,3], Martin Taubert[1], Sven Baumann[4], Michael Schlömann[2], D. Barrie Johnson[3], Martin von Bergen[1,4] and Jana Seifert*[1].

[1]Department of Proteomics, UFZ – Helmholtz Centre for Environmental Research, Permoserstr. 15, D-04318 Leipzig, Germany; [2]Interdisciplinary Ecological Center, TU Bergakademie Freiberg, Leipziger Strasse 29, 09599 Freiberg, Germany; [3]School of Biological Sciences, College of Natural Sciences, Bangor University, Deiniol Road, Bangor LL57 2UW, U.K.; [4]Department of Metabolomics, UFZ - Helmholtz Centre for Environmental Research, Permoserstr. 15, D-04318 Leipzig, Germany.

**Received:** 16 December 2011 **Accepted:** 21 February 2012 **Available Online**: 08 March 2012

ABSTRACT

Although many examples of syntrophic cultures are known, details of carbon utilization and carbon transfers within them often remain elusive due to limitations in methods used to detect carbon flux. We have applied the recently developed method of protein-based stable isotope probing (protein-SIP) to track carbon flow in a mixed culture of acidophilic bacteria. The heterotroph *Acidiphilium cryptum* was grown in the presence of $^{13}$C-labeled galactose, together with the iron- and sulfur-oxidizing autotroph *Acidithiobacillus ferrooxidans*. Cultures were harvested at five time points, proteins extracted and separated by 1-dimensional SDS gel electrophoresis, peptides obtained by tryptic digest of gel slices and analyzed by UPLC Orbitrap MS/MS measurements. Syntrophic interactions were confirmed by analysis of the time-dependent incorporation of $^{13}$C into peptides, and quantified by calculation of relative isotope abundance (RIA) and labeling ratio (lr) from mass spectral isotope patterns. $^{13}CO_2$ formed by catabolism of galactose by *A. cryptum*, was found to be assimilated by *At. ferrooxidans* which used tetrathionate as electron donor. Mass spectral data indicated that $^{13}$C-labeled organic substances, mostly peptides, secreted by the chemoautotroph were assimilated by the heterotroph. The data provided unequivocal evidence for two-way transfer of carbon in mixed cultures of autotrophic and heterotrophic acidophilic bacteria.

**Keywords:** Mass spectrometry; protein-SIP; acidophiles; carbon flux.

## 1. Introduction

Stable isotope probing (SIP) is an emerging tool for investigating metabolic key players in microbial communities and their interaction mechanisms [1]. In brief, substrates are labeled with heavy isotopes, such as $^{13}$C, $^{15}$N or $^{36}$S, which are then incorporated into biomolecules by the metabolic activity of the cells [2-4]. In addition to the application of SIP with nucleic acids, the use of proteins in SIP has recently been established [2, 3, 5, 6]. As proteins closely reflect the cell activity at any time point, this approach enables concomitant functional analysis of microbial communities.

In protein-SIP, the incorporation of heavy isotopes can be detected as a mass shift of peptides derived from the analyzed proteins which can be measured by mass spectrometric analysis. From a peptide mass spectrum, the quantification of the fraction of heavy isotopes of a specific element, e.g. $^{13}$C, referred to as relative isotope abundance (RIA), is possible either in a sequence-dependent or sequence-independent approach, with the latter using the half decimal place rule [7-9]. For the sequence-dependent approach, an accuracy of ± 0.1% $^{13}$C RIA using at least 20 peptides has been claimed [10] .

Besides the RIA data, the intensity ratio between labeled and unlabeled peptide forms is also of interest, especially in time-series experiments. This 'labeling ratio' (lr) was recently

***Corresponding author:** Jana Seifert, E-mail Address: jana.seifert@ufz.de; Phone Number: +49 341 235 1352; Fax Number: +49 341 235 1786 .

used to detect induction of specific proteins after a substrate shift based on differences in protein synthesis rates [8]. The potential of protein-SIP to detect metabolic activities in defined communities as well as in enrichment cultures was also proven in recent studies [2, 11, 12], though a demonstration of the applicability of protein-SIP to detect the carbon flux between different species has not previously been reported.

*Acidithiobacillus (At.) ferrooxidans* was the first acidophilic bacterium shown to oxidize ferrous iron as well as reduced sulfur and has subsequently been shown to oxidize hydrogen and formic acid. The pH range of the type strain is 1.5 to >4.0 with an optimum for growth of about 2.0 [13-16]. The genus *Acidithiobacillus* currently comprises four recognized species, two of which (*At. ferrooxidans* and the psychrotolerant species *At. ferrivorans*) are iron oxidizers [17, 18] and all of which are autotrophs. The physiological diversity and versatility of iron-oxidizing metabolic activity may be the reason for the wide distribution of this genus in a variety of acid mine drainage sites [19].

*Acidiphilium* spp. are obligately heterotrophic acidophiles (with the exception of *Acidiphilum (A.) acidophilum*, which is a facultative autotroph) that are widely distributed in metal-rich, acidic environments [20]. They can grow on a wide range of monosaccharides, dicarboxylic acids and tricarboxylic acids [21]. The type species of the genus *Acidiphilium cryptum* is often found in close association with acidophilic chemoautotrophs such as *At. ferrooxidans*. The syntrophic relationship of the acidophilic species involves organic substances that derive from the autotroph (as exudates or cell lysates) being used as carbon and energy sources by *Acidiphilium* spp., and other heterotrophic acidophiles [21, 22]. Ferric iron generated by iron-oxidizing acidithiobacilli can be used by *Acidiphilium* as an alternative electron acceptor to oxygen in oxygen-limited environments [19]. In addition, carbon dioxide generated by heterotrophic bacteria has been postulated to be used by autotrophic iron- and sulfur-oxidizing acidophiles [23], though this has not hitherto been proven.

Commonly carbon uptake and exchange within microbial consortia is detected by nucleotide-, fatty acid-SIP or by tracing $^{14}$C-carbon [24, 25]. In addition, methods such as Raman spectroscopy, microautoradiography, and nanoSIMS have been used to analyze the $^{12}$C/$^{13}$C carbon composition of cell compartments revealing a high resolution of single cells within complex communities [26, 27]. Besides the utmost level of sensitivity, a direct combination of isotope detection and phylogenetic origin is only possible by the concomitant use of phylogenetic probes and microscopy (FISH).

In this study, we used protein-SIP for tracking the carbon flux in a mixed culture of *At. ferrooxidans* and *A. cryptum*. A time-series experiment was set up using $^{13}$C-labeled galactose as substrate for *A. cryptum*. *At. ferrooxidans* is a strict chemoautotroph [21] and is not able to catabolize galactose or other monosaccharides. In a hypothetical model, a release of $^{13}CO_2$ by the heterotrophic species is assumed, which then is assimilated by the autotrophic species. In the course of the metabolic activity, $^{13}$C-labeled organic exudates are formed which might be used again by the heterotroph. The quantification of time-dependent $^{13}$C incorporation into the proteins of *At. ferrooxidans* and *A. cryptum* was measured by mass spectrometric analysis. The results from this study of a simple microbial community can be used to estimate the generic applicability of the protein-SIP method for carbon flux analysis.

## 2. Materials and methods

### 2.1 *Bacterial strains and growth conditions*

*A. cryptum*$^{T}$ (DSM-2389) and *At. ferrooxidans*$^{T}$ (DSM-14882) obtained from the German Collection of Microorganisms and Cell Cultures (DSMZ; Braunschweig, Germany) were used in the experimental work. *A. cryptum* was grown in a medium containing basal salts and trace elements (pH 3.0) [28] supplemented with 10 mM galactose, at 30°C and with continuous shaking (130 rpm). *At. ferrooxidans* was pre-cultured under the same growth conditions but the growth medium was supplemented with 2.5 mM potassium tetrathionate instead of galactose.

### 2.2 *Experimental set up and sampling*

Pure cultures of *At. ferrooxidans* and *A. cryptum*, grown as described above, were harvested in early stationary growth phase by centrifugation (11,000 x g, 30 min, 4°C) in an AvantiTM J-30I centrifuge (Beckman Coulter, Fullerton / USA). Cell pellets were washed several times with basal salts (pH 3.0) to remove any remaining traces of galactose or tetrathionate. Pellets were resuspended in basal salts pH 3.0 and cell numbers were determined using a Thoma counting chamber.

For the main experiment, 15 identical cultures were set up in 250 ml shake flasks, each containing 60 ml of growth medium (basal salts/trace elements, supplemented with 2.5 mM potassium tetrathionate and 1 mM $^{13}$C galactose (Campro Scientific, The Netherlands) at pH 3.0). Five additional shake flasks containing unlabeled galactose (and tetrathionate) were set up as controls. Cells from the pre-cultures of *At. ferrooxidans* and *A. cryptum* were inoculated into fresh media, giving a final cell number of $2*10^3$ cells/ml of each bacterium. The shake flasks were closed air-tight using screw-tops with central rubber septa, to allow sampling of the gas and liquid phases. The flasks were incubated in a rotary shaker (130 rpm) at 30°C. Control flasks were sampled on a daily basis to determine concentrations of galactose and tetrathionate and to measure pH. At different points in the culture growth cycles, triplicate flasks containing $^{13}$C galactose and one control flask were removed. Eight milliliters of the gas phase were removed from each flask using a syringe, and stored in 10 ml headspace vials (previously flushed with nitrogen). Next, the flasks were opened and 5 ml aliquots of the liquid phase were withdrawn for FISH analysis

(described below) and the cells in the remaining liquid phase were concentrated by centrifugation (11,000 x g, 30 mins, 4° C). The supernatant and the cell pellet were stored separately at -20°C for further analysis. A small volume of the supernatant was also used for final determination of pH and concentrations of galactose and tetrathionate.

2.3 *Physicochemical analysis*

The pH in the liquid phase was determined using a combined Sentix20 pH electrode (WTW, Germany) coupled to a Calimatic 766 pH meter (Knick, Germany). Tetrathionate was assayed by cyanolysis as described by Kelly et al. [29]. Galactose was determined as previously described [10].

2.4 *Fluorescent in situ hybridization (FISH)*

Five milliliters of each sample were concentrated by centrifugation for 15 min at 11,000 × g and 4°C. Cell pellets were suspended in 100 µl of the residual fluid and fixed by adding three volumes of a paraformaldehyde solution (4% (w/v) in phosphate-buffered saline (PBS)). After fixation for 3 h at 4°C, cells were washed in 1×PBS and stored in ethanol:1×PBS (1:1) at −20°C until analyzed [30, 31]. To determine total cell numbers, cells present in the samples were stained with 4-,6-diamidino-2-phenylindole (DAPI) and visualized by epifluorescence microscopy. Enumeration of the DAPI-stained cells was performed on Teflon-coated glass by using 2 µl of the fixed cell suspension. For hybridization of *At. ferrooxidans* cells the oligonucleotide probe used was TF539 (5'- CAG ACC TAA CGT ACC GCC -3') labeled with fluorescein at the 5'-end, according to [32], while for *A. cryptum* the probe used was Acdp821 (5'- AGC ACC CCA ACA TCC AGC ACA CAT -3') Cy3-labeled at the 5' end, as described in [33].

To adjust the probe specificity, the hybridization buffer for TF539 was supplemented with 20% (v/v) formamide and for Acdp821 with 25% (v/v) formamide. Hybridization was performed at 46°C and washing at 48°C for both probes. Washing solutions were prepared as described previously [34].

2.5 *Protein extraction*

Cell pellets were suspended in 200 µl of urea buffer (8 M urea, 40 mM Tris-HCl pH 7.5, 4 mM DTT, 1 mM PMSF, 1 µl/ml benzonase) and subsequently disrupted by three cycles of ultrasonic treatment under continuous cooling on ice (ultrasonic processor UP50H equipped with ultrasonic probe MS1, Hielscher ultrasonics, Teltow, Germany). After incubation for 1.5 h at room temperature, cell debris were separated by centrifugation at 16,000 x g for 20 min and 4°C (Laboratory Centrifuge 3K30, Sigma, Osterode, Germany). The supernatant was kept at 4°C and immediately subjected to further processing steps.

2.6 *Protein separation and mass spectrometric analysis*

Protein concentration was determined using the Bradford assay [35].{Marion M, 1976 #972} For 1-dimensional gel electrophoresis (1-DE), 50 µg of protein were precipitated with five-fold volume of ice-cold acetone and separated using a 12% acrylamide separating gel and the Laemmli-buffer system. After electrophoresis, protein bands were stained by colloidal Coomassie Brilliant Blue G-250 (Roth, Kassel, Germany). One protein band in the mass range of 55 – 65 kDa was cut and subsequently in-gel tryptic digestion was performed [3]. Tryptic peptides of each band were desalted by $C_{18}$ ZipTip columns before MS-analysis.

Peptides were analyzed by UPLC-LTQ Orbitrap-MS/MS as described in Bastida *et al.* [11]. The peptides of the $^{12}C$ sample and the first $^{13}C$ replicate were eluted over 16 min with 2–80% solvent B (acetonitrile + 0.1% formic acid) gradient using a nanoAcquity UPLC column (C18, 75 mmx15 cm, 1.75 mm, Waters). Continuous scanning of eluted peptide ions was carried out between 400-2,000 *m/z*, automatically switching to MS/MS CID mode on ions exceeding an intensity of 3000. Raw data were processed for database search using Thermo® Proteome Discoverer software (v1.0 build 43, Thermo Fisher Scientific). Search was performed by tandem mass spectrometry ion search algorithms from the Mascot house server (v2.2.1). The following parameters were selected: TaxID 524 (*A. cryptum*) and 920 (*At. ferrooxidans*) of NCBInr (National Center for Biotechnology Information, Rockville Pike, USA, version August 2010 and later) as criterion for taxonomy, tryptic cleavage, maximal two missed cleavage sites. A peptide tolerance threshold of ±10 ppm and an MS/MS tolerance threshold of ±0.2 Da were chosen. Carbamidomethylation at cysteines was given as static and oxidation of methionines as variable modification. Only proteins with at least 2 identified peptides with a false-positive probability less than 0.05 (ion score threshold >40) were considered for further analysis.

2.7 *Calculation of relative isotope abundance (RIA) and labeling ratio (lr)*

Mass spectra were analyzed as described previously [8, 10]. In brief, the spectra were analyzed manually using QualBrowser v2.0.7 (Xcalibur®, Thermo Fisher Scientific Inc., Waltham MA, USA). The percentage of incorporated $^{13}C$ atoms in relation to the total number of carbon atoms in a peptide (RIA) was calculated by a widely applied method based on comparison of theoretical and experimental spectral data. For semi-automatic calculation of the RIA, an Excel spreadsheet was used as described in Taubert et al. [8] (downloadable from the Helmholtz Centre for Environmental Research – UFZ, Department of Proteomics: http://www.ufz.de/index.php?en=20647).

The labeling ratio defined by Taubert et al. [8] as the proportion of a peptide's heavy isotope-labeled isotopologues of all isotopologues of that peptide, thereby having a range between 0 (only unlabeled peptide) and 1 (only labeled peptide), was calculated to compare the labeled and unlabeled

peptide abundances in any one sample.

The workflow for the calculation was similar to that described by Taubert et al. [8]: *(i)* a mass range covering the masses of heavy and light isotopologues of a peptide was selected and (*ii*) the peak intensities of both light and heavy isotopologues were compared within a sample using a timeframe adjusted to the elution time of the peptide to calculate the labeling ratio.

$$\text{labeling ratio} = \frac{\sum(\text{intensities of isotope peaks of specific incorporation pattern})}{\sum(\text{intensities of all isotope peaks of the peptide})}$$

If more than one incorporation pattern was detected, labeling ratios had to be calculated for all patterns individually. For calculation of the labeling ratios on protein level, data of at least three peptides were averaged.

## 3. Results and discussion

### 3.1 *Cultivation*

*At. ferrooxidans* and *A. cryptum* were grown as a mixed culture containing both $^{13}$C-labeled galactose and non-labeled tetrathionate, together with control cultures containing non-labeled galactose and tetrathionate. Microbial growth was monitored by measurement of pH, tetrathionate and galactose (Figure 1A). Tetrathionate originally present in the culture medium (2.3 mM) was completely oxidized (by *At. ferrooxidans*) after 10 days of incubation. Additional tetrathionate (2.3 mM) was added on day 16, which was oxidized within two days due to greater numbers of *At. ferrooxidans* being present by this time. The metabolic activity of *A. cryptum* was confirmed by the fact that all of the galactose in the medium had been removed by day 8. Although recent genome analysis of the type strain of *At. ferrooxidans* has identified a galactose proton symporter-like gene, this has been suggested to be involved in molecular signaling [36]. Cultivation experiments of *At. ferrooxidans* showed that the species does not catabolize galactose. The pH of the medium changed from 3.0 to 1.87 during the incubation period, due to oxidation of tetrathionate to sulfuric acid.

Growth of the two acidophiles was confirmed by cell counts using FISH probes specific for each of the two species. Cell numbers of *A. cryptum* (Figure 1B) increased from ~ $2.0*10^3$ to $1.0*10^7$ cells/ml, and those of *At. ferrooxidans* (Figure 1C) from ~ $1.5*10^3$ to $2.5*10^7$ cells/ml over the time course (18 days) of the experiment. Time-related changes in cell numbers were uneven, though this was a consequence of the experimental design which did not track individual cultures but instead involved the harvesting of entire liquid cultures at each time point. The apparent discrepancies were due to these cultures not being perfectly synchronized.

### 3.2 *Protein identification*

Protein extracts from each time point and replicate were separated via 1-DE, and a gel band covering the protein mass of 55 – 65 kDa was typically digested and further analyzed by mass spectrometry. A total of 104 proteins were identified in the combined searches from all non-labeled substrate cultures (see Table S1). This relatively low number can be explained by the measurement of the defined mass range and

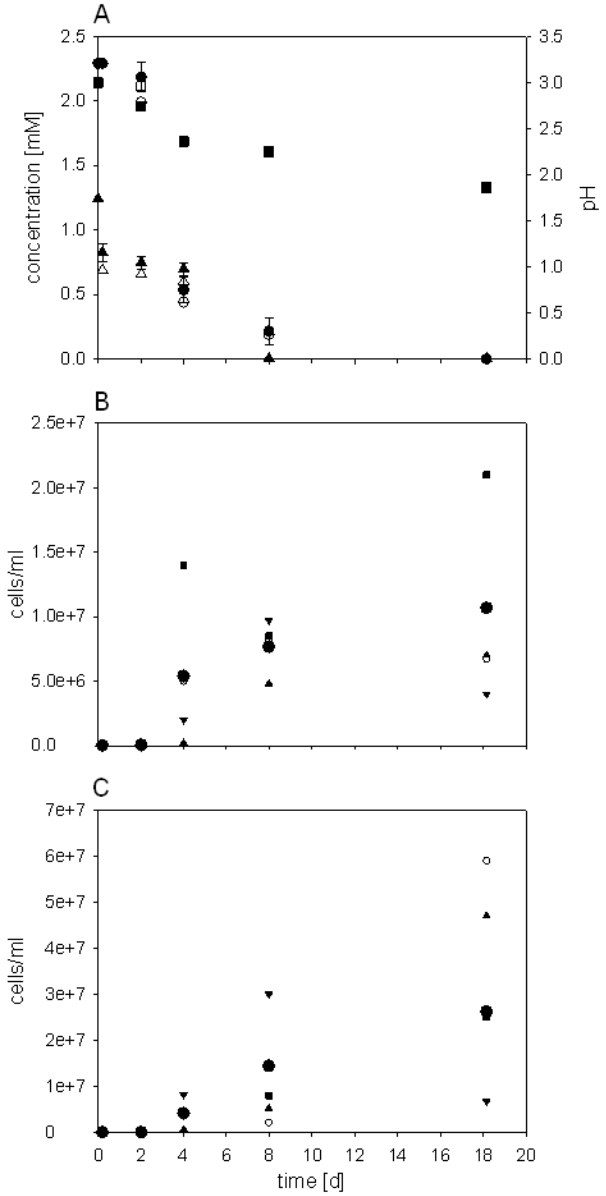


**Figure 1.** Time course of growth parameters during $^{12}$C and $^{13}$C cultivations. The upper plot (A) shows galactose concentration of $^{12}$C (Δ) and $^{13}$C cultures (▲), tetrathionate concentration of $^{12}$C (○) and $^{13}$C cultures (●), pH of $^{12}$C (□) and $^{13}$C cultures (■). Values for $^{13}$C cultures are given as means of three replicates and corresponding standard deviations. Middle and bottom plots display changes in cell numbers of (B) *A. cryptum* and (C) *At. ferrooxidans* during the experiment. Each plot shows cell numbers in the $^{12}$C culture (○) and in three replicates of $^{13}$C cultures (replicate one (▲), two (▼), three (■) and average over all replicates (●). Additional tetrathionate (2.3 mM) was added on day 16.

cell numbers being quite low, especially for the first time points.

### 3.3 *Incorporation of $^{13}C$ in proteins of A. cryptum*

The RIAs and the lr of peptides from three proteins of the heterotroph *A. cryptum* (chaperonin GroEL, dihydroxy-acid dehydratase and 30S ribosomal protein S1; Table 1) were calculated using the semi-automatic script described by Taubert et al. [8]. The peptides were found to show two distinct incorporation patterns caused by differential incorporation of $^{13}C$. As an example, the mass spectral incorporation patterns for peptide AGGLPAVIGELIR of dihydroxy-acid dehydratase at all time points are shown in Figure 2A. An overview of the averaged values from all peptides is given in Figure 3. The pattern showed a rapid appearance of the $^{13}C$

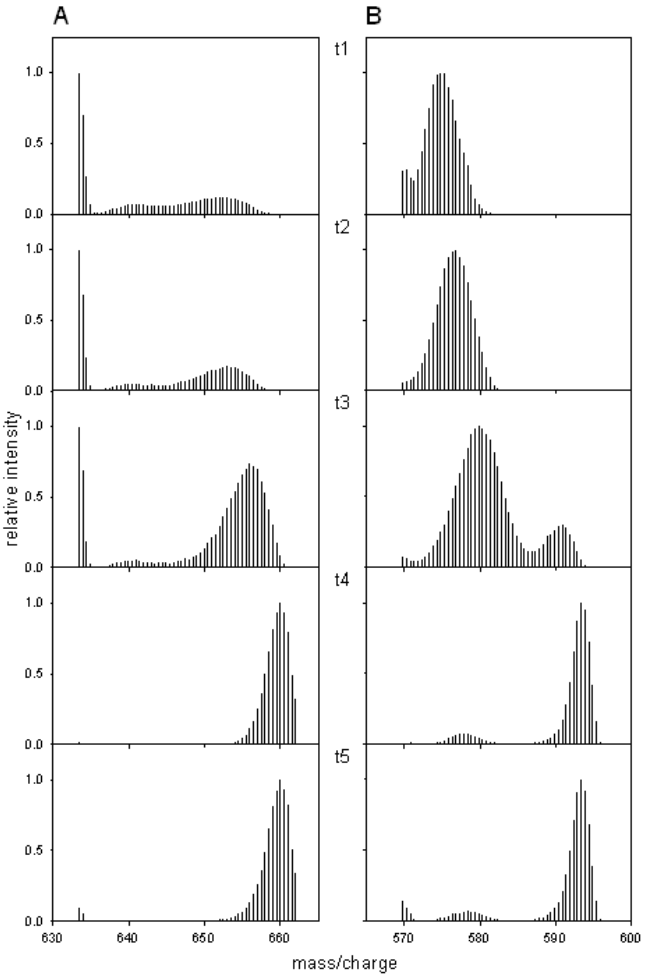


**Figure 2.** Changes of peptide ion mass distributions during cultivation of *A. cryptum* and *At. ferrooxidans* in mixed culture. Representative mass spectra showing the peptide ion mass distribution of (A) peptide AGGLPAVIGELIR of dihydroxy-acid dehydratase (*A. cryptum*, gi148259108) and (B) peptide AFDGSSIAGWK of glutamine synthetase, type I (*At. ferrooxidans*, gi198282766). Intensity values are given as relative intensity of highest peak in the section depicted.

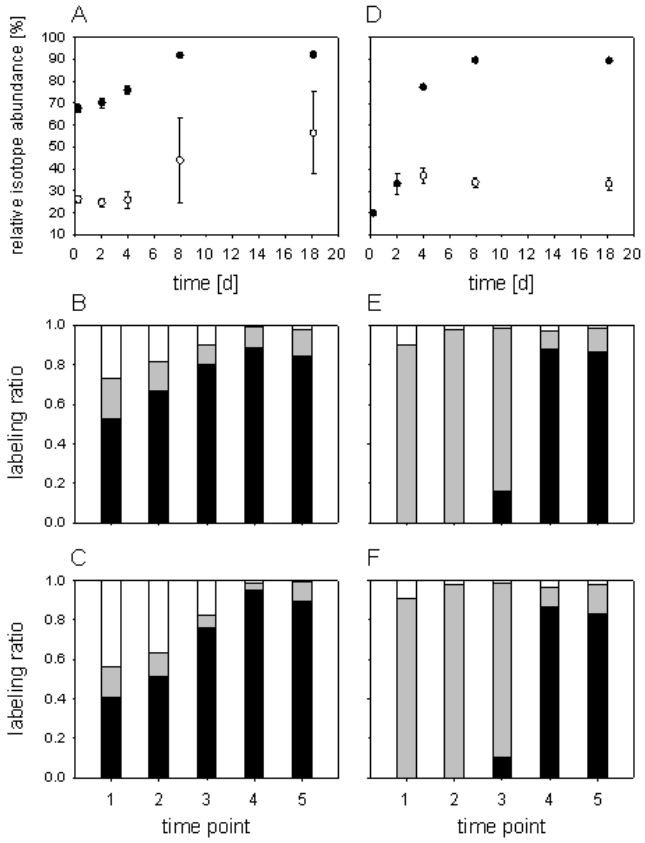


**Figure 3.** Development of relative isotope abundance (RIA) and labeling ratio (lr) over time in proteins of *A. cryptum* (left side) and *At. ferrooxidans* (right side). In the upper section the time courses of the RIA for (A) *A. cryptum* and (D) *At. ferrooxidans* are shown. In both cases a higher (●) and a lower RIA (○) was detected. Plots are shown as mean values of three replicates, and are based on three proteins in the case of *A. cryptum* (chaperonin GroEL, gi148261486, dihydroxy-acid dehydratase, gi148259108, 30S ribosomal protein, gi148259766) and two proteins in the case of *At. ferrooxidans* (glutamine synthetase type I, gi198282766, chaperonin GroEL, gi198282835). RIA calculations are based on at least three peptides per protein, time point and replicate.
The middle and lower section shows the development of the lr over time in proteins of *A. cryptum* (B and C) and *At. ferrooxidans* (E and F). Bars represent the distribution of the labeling ratios of $^{12}C$ peak (white), lower $^{13}C$ RIA peak (grey) and higher $^{13}C$ RIA peak (black) at each time point of cultivation. *A. cryptum* proteins shown are (B) 30S ribosomal protein S1, gi148259766 and (C) dihydroxy-acid dehydratase, gi148259108. *At. ferrooxidans* proteins shown are (E) glutamine synthetase type I, gi198282766 and (F) chaperonin GroEL, gi198282835. Calculations of lr are based on at least three peptides per protein, time point and replicate. Bars are averaged over three replicates.

isotopologues at time point one (within 5 hours of incubation), exceeding the intensity of the $^{12}C$ isotopologues at day 8. The first pattern, with an average RIA of around 25% $^{13}C$, decreased in relative intensity over time, from a lr of 0.19 ± 0.03 after 5 h of incubation to a final lr of 0.07 ± 0.06 at day 18. Due to the low intensity and the associated difficulties in detection of the pattern in the spectra, both RIA and lr val-

**Table 1.** List of proteins (gi numbers, organisms, analysed peptides) and the respective calculated RIA and lr values for cultivation on $^{13}$C-galactose for the time range 0.2 d to 18 d. Calculations for each protein were averaged for three replicate cultures.

| GI no. | Protein name (peptides) | Time [d] | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 0.2 | | 2 | | 4 | | 8 | | 18 | |
| | | RIA [%] | lr | RIA [%] | lr | RIA [%] | lr | RIA [%] | lr | RIA [%] | lr |
| *Acidithiobacillus ferrooxidans* ATCC 53993 | | | | | | | | | | | |
| gi198282766 | glutamine synthetase, type I | 19.8 ± 0.2 | 0.90 ± 0.02 | 32.7 ± 4.8 | 0.98 ± 0.01 | 36.4 ± 3.8 | 0.82 ± 0.04 | 89.6 ± 0.1 | 0.88 ± 0.02 | 89.4 ± 0.4 | 0.87 ± 0.02 |
| | (AFDGSSIAGWK, LVPHFEAPVLLAYSAK, | | | | | 77.5 ± 2.0 | 0.16 ± 0.05 | 33.0 ± 2.0 | 0.12 ± 0.02 | 32.8 ± 3.6 | 0.12 ± 0.02 |
| | GGYFPVPPVDSAQDLR, AINALTNPSTNSYK) | | | | | | | | | | |
| gi198282835 | chaperonin GroEL | 19.8 ± 0.3 | 0.91 ± 0.02 | 33.6 ± 4.7 | 0.98 ± 0.01 | 37.3 ± 3.7 | 0.88 ± 0.04 | 89.5 ± 0.2 | 0.84 ± 0.03 | 89.2 ± 0.5 | 0.84 ± 0.03 |
| | (EVAQVGTISANSDDSIGK, | | | | | 77.1 ± 0.9 | 0.11 ± 0.04 | 34.5 ± 2.5 | 0.15 ± 0.04 | 33.5 ± 2.7 | 0.14 ± 0.04 |
| | EVASQASDEAGDGTTTATVLAQAIIR, LESTTLADLGQAK, TANHDQDmGVAIIR, VVSEEIGmKLESTTLADLGQAK) | | | | | | | | | | |
| Average *Acidithiobacillus ferrooxidans* proteins | | 19.8 ± 0.2 | 0.90 ± 0.02 | 33.2 ± 4.6 | 0.98 ± 0.01 | 36.9 ± 3.7 | 0.85 ± 0.05 | 89.5 ± 0.1 | 0.86 ± 0.03 | 89.3 ± 0.4 | 0.85 ± 0.03 |
| | | | | | | 77.3 ± 1.5 | 0.14 ± 0.05 | 33.7 ± 2.3 | 0.13 ± 0.03 | 33.2 ± 3.1 | 0.13 ± 0.03 |
| *Acidiphilium cryptum* JF-5 | | | | | | | | | | | |
| gi148261486 | chaperonin GroEL | 68.3 ± 0.5 | 0.41 ± 0.02 | 70.6 ± 1.0 | 0.57 ± 0.03 | 74.3 ± 1.7 | 0.71 ± 0.04 | 91.3 ± 0.7 | 0.94 ± 0.09 | 91.5 ± 1.2 | 0.94 ± 0.04 |
| | (ITTPSETAQVGTISANGEAEIGK, TNDLAGDGTTTATVLAQAIVR, QIAENAGEDGAVISGK, VGGASETEVKER, GIQTELDVVEGmQFDR, LAGGVAVIR) | 25.6 ± 0.8 | 0.21 ± 0.03 | 24.0 ± 1.2 | 0.16 ± 0.03 | 24.4 ± 2.1 | 0.10 ± 0.02 | 43.6 ± 20.2 | 0.06 ± 0.08 | 49.4 ± 20.7 | 0.04 ± 0.04 |
| gi148259108 | dihydroxy-acid dehydratase | 65.8 ± 2.2 | 0.41 ± 0.03 | 69.2 ± 1.9 | 0.51 ± 0.04 | 76.3 ± 1.9 | 0.77 ± 0.05 | 91.5 ± 0.5 | 0.97 ± 0.06 | 91.7 ± 0.1 | 0.96 ± 0.05 |
| | (QSGTSGSPSILNASPESAVGGGLALLK, FTGQLETGAVLENAPQYQR, STQWFDNPDNPGmTALYLER, IGYEIPLLVNmmPAGK, SANILIDEADLAAR, AG-GLPAVIGELIR) | 26.7 ± 1.7 | 0.16 ± 0.02 | 24.9 ± 2.1 | 0.12 ± 0.03 | 27.1 ± 5.7 | 0.06 ± 0.02 | 47.2 ± 21.7 | 0.04 ± 0.03 | 69.1 ± 2.9 | 0.11 ± 0.01 |
| gi148259766 | 30S ribosomal protein S1 | 68.1 ± 2.3 | 0.53 ± 0.05 | 70.1 ± 3.2 | 0.67 ± 0.04 | 77.0 ± 1.5 | 0.80 ± 0.04 | 91.6 ± 0.6 | 0.97 ± 0.06 | 92.2 ± 1.2 | 0.93 ± 0.07 |
| | (VLDVDVEKER, DVTPLmGVQQPFQILK, QLETD-PWEGVDVK, SGDTFDLYIER, | 25.7 ± 1.3 | 0.20 ± 0.02 | 25.0 ± 2.2 | 0.15 ± 0.02 | 26.4 ± 4.1 | 0.10 ± 0.04 | 34.2 ± 0.1 | 0.12 ± 0.10 | 67.1 ± 3.0 | 0.15 ± 0.04 |
| | SQGADHTTGTEDFAALLDSTLGSNTGFEGSVVSGR, VDDVLTGFIR) | | | | | | | | | | |
| Average *Acidiphilium cryptum* proteins | | 67.4 ± 2.0 | 0.45 ± 0.07 | 70.0 ± 2.2 | 0.58 ± 0.07 | 75.8 ± 2.0 | 0.76 ± 0.06 | 91.5 ± 0.6 | 0.96 ± 0.07 | 91.8 ± 1.0 | 0.94 ± 0.06 |
| | | 26.0 ± 1.3 | 0.19 ± 0.03 | 24.6 ± 1.9 | 0.14 ± 0.03 | 25.8 ± 4.0 | 0.09 ± 0.03 | 43.8 ± 19.1 | 0.06 ± 0.07 | 56.2 ± 18.7 | 0.07 ± 0.06 |

ues for the later time points showed considerable variations.

The second pattern, with a RIA of 67.4 ± 2.0% $^{13}C$ after 5 hours and 91.8 ± 1.0 % $^{13}C$ at day 18, showed a strong increase of the lr from 0.45 ± 0.07 to 0.94 ± 0.06, with the highest increase in intensity between day 4 and 8 (Figure 2A). This represents a high synthesis rate of proteins from a highly $^{13}C$-labeled carbon source, and thus a rapid increase in microbial growth [8], which was also apparent from increasing cell numbers (Figure 1B). Within the present experimental setup, this was clearly a result of the metabolism of $^{13}C$-galactose. The deviation between the galactose RIA of >=98% and the peptide RIA of 67 to 92% can be explained by the assimilation of unlabeled carbon sources, internal carbon reserves or secreted metabolites of *At. ferrooxidans*. As the first pattern at lower RIA did not show any increase over time, it can be assumed that it was a result of metabolic processes proceeding prior to the first sampling time.

### 3.4 *Incorporation of $^{13}C$ in proteins of At. ferrooxidans*

For the autotroph *At. ferrooxidans*, values for RIA and lr of peptides from two proteins (glutamine synthetase, type I and chaperonin GroEL; table 1) were calculated. As an example, the mass spectral incorporation patterns for peptide AFDGSSIAGWK of glutamine synthetase, type I at all time points are shown in Figure 2B. An overview of the averaged values from all peptides is given in Figure 3. As with *A. cryptum,* the $^{13}C$ isotopologues pattern appeared immediately but the $^{12}C$ isotopologues showed a low intensity compared to the $^{13}C$ pattern. An incorporation pattern of around 20% RIA was visible at early time points (Figure 2B). This pattern is increasing in intensity up to day 4, and also a slight increase of the RIA to around 37% can be seen. From this time point on, no further increase in intensity was detected. This development of the pattern can be explained by an oxidation of galactose by *A. cryptum*, and the resulting $^{13}CO_2$ being assimilated by the autotroph *At. ferrooxidans*. Until day 4 the predominant RIA pattern showed rather low $^{13}C$-incorporation at a simultaneously high lr indicating the use of $^{13}C$-labeled and non-labeled $CO_2$ (Figure 2B). In addition, another pattern at higher RIA of ~77% appeared at the same time (day 4), which increased to ~ 89% on day 18, also strongly increasing in intensity. This increase of the RIA correlated with the highest metabolic activity of *A. cryptum* as illustrated by the strongly increasing lr between day 4 and 8 (Figure 2). The RIA in the $CO_2$ fraction would have been expected to increase significantly between day 4 and day 8, due to the oxidation of $^{13}C$ galactose to $^{13}CO_2$ by *A. cryptum*. Consequently, the shift of the incorporation pattern in *At. ferrooxidans* peptides can be explained by autotrophic $CO_2$ fixation and the increase of $^{13}CO_2$ produced by *A. cryptum*.

### 3.5 *Carbon flux analysis*

As the syntrophic interaction between *A. cryptum* and *At. ferrooxidans* under the present experimental conditions was based on the oxidation of $^{13}C$-galactose to $^{13}CO_2$ and the subsequent uptake of labeled $CO_2$, an increase of the labeling ratio of peptides from *A. cryptum* should correlate with an increase of the RIA of peptides from *At. ferrooxidans*. After a certain time, while a sufficient amount of $^{13}CO_2$ is produced by the heterotroph and assimilated by the autotroph, a similar RIA value should also have been detected in peptides of the autotroph. With the data obtained, both effects were observed. After an initial lag phase, *A. cryptum* displayed maximal protein biosynthesis rate (growth) between day 4 and day 8, as illustrated by both significantly increasing lr and cell numbers. Besides the use of $^{13}C$-labeled galactose, the RIA development of *A. cryptum* can be caused by the additional uptake of organic exudates and lysates from *At. ferrooxidans*, and the use of internal $^{12}C$ carbon reserves which are depleted after day 4. In addition, the increased value of the RIA at the last two time points could be explained by increased $^{13}C$-incorporation in organic exudates and lysates as a result of increasing uptake of $^{13}CO_2$ by *At. ferrooxidans*. $^{13}CO_2$ formation was detected by GC-C-IRMS measurements (data not shown). The overall syntrophic interaction between both species is shown in Figure 4.

Heterotrophic acidophilic microorganisms such as *Acidiphilium* spp. have often been found living in association with acidophilic chemolithoautotrophs such as *At. ferrooxidans* [37]. This partnership has been reported for several microbial communities in natural acidic habitats (e.g. the Tinto river), acid mine drainage and "biomining" environments [22, 38]. The heterotrophs are capable of using at least some of the organic compounds secreted by the autotrophs and, by removing these, help to maintain conditions that are conducive to the growth of the autotrophs. This is particularly important in closed environments (e.g. stirred tanks used for mineral bioprocessing) where concentrations of some soluble carbon exudates might otherwise inhibit the growth and activities of chemoautotrophic bacteria [23]. Liu et al. recently found that in an artificial mixed culture of *A. acidoph-*

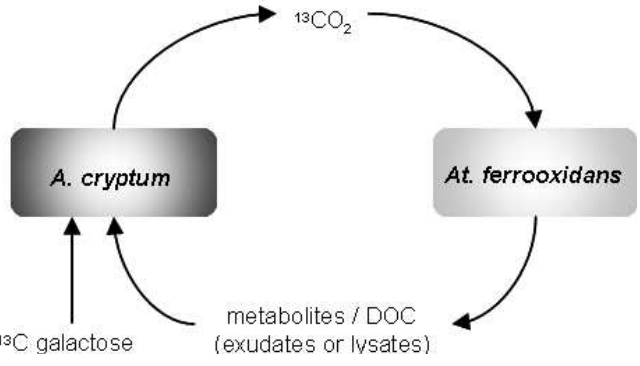


**Figure 4.** Carbon flow within the syntrophic interaction of *A. cryptum* and *At. ferrooxidans*. The heterotroph catabolized $^{13}C$ labeled galactose and released $^{13}CO_2$ which was taken up and assimilated by the autotroph. In course of its metabolism the autotroph secreted dissolved organic compounds (DOC) which in turn were metabolized by the heterotroph.

*ilum* and *At. ferrooxidans* growth of the autotroph was promoted by the heterotrophic partner by activating genes related to iron oxidation and $CO_2$ fixation [39]. The ability of heterotrophic acidophiles to remove inhibitory organic compounds and thereby facilitate the growth of iron- and sulfur-oxidizing acidophiles is also the basis of the "overlay" plate technique for cultivating the latter on solid media, though here the organic materials are thought mostly to derive from the acid hydrolysis of the gelling agent (agar and derivatives) [40]. The "feedback" mechanism where carbon dioxide released by the heterotrophs is used by the autotrophs has previously been alluded to [23] but had not been proven. Data from the current work has confirmed that this does indeed occur.

## 4. Concluding remarks

The applicability of protein-SIP to track carbon transfers in a defined mixed culture of autotrophic and heterotrophic bacteria was successfully demonstrated. Two-way transfer of carbon was confirmed by changes of mass spectral patterns. Protein-SIP is therefore a potential technique for examining more complex microbial communities where metabolic activities of microbial key players can be detected and quantitatively described. In addition, secondary consumers using metabolites from the primary species and putative scavengers are detectable by protein-SIP, allowing the construction of a carbon-based food web of a microbial community. Besides the relatively high amount of biomass needed and the dependency of appropriate genomic data, Protein-SIP has some major advantages compared to other SIP approaches: *i)* DNA/RNA-SIP has a 100 fold lower accuracy and sensitivity, and no conclusion of a direct or indirect metabolisation of the carbon source can be achieved; *ii)* PLFA-SIP has a low phylogenetic resolution and a concomitant detection of differentially labeled phospholipid fatty acids is not possible. The progress in bioinformatics tools allowing an automatic analysis will increase the potential of protein-SIP for complex studies.

## 5. Supplementary material

Supplementary data and information is available at: http://www.jiomics.com/index.php/jio/rt/suppFiles/85/0

Supplementary material includes Fig. S1 showing the pictures of the 1-dimensional (1-DE) SDS gels of the samples. Table S1 includes all proteins identified in the $^{12}C$ samples.

## Acknowledgements

S. Hedrich was funded by a Ph.D. scholarship of the German Federal Environmental Foundation (DBU) and M. Taubert was funded by the Deutsche Forschungsgemeinschaft (SPP 1319). We thank U. Günther for her support with GC-C-IRMS measurements.

## References

1. M. G. Dumont, J. C. Murrell, Nat. Rev. Microbiol. 3 (2005) 499-504.
2. N. Jehmlich, F. Schmidt, M. von Bergen, H. H. Richnow, C. Vogt, ISME J. 2 (2008) 1122-1133.
3. N. Jehmlich, F. Schmidt, M. Hartwich, M. von Bergen, H. H. Richnow, C. Vogt, Rapid Commun. Mass Spectrom. 22 (2008) 2889-2897.
4. N. Jehmlich, F. D. Kopinke, S. Lenhard, F. A. Herbst, J. Seifert, U. Lissner, U. Völker, F. Schmidt, M. von Bergen, Proteomics 12 (2012) 37-42.
5. N. Jehmlich, F. Schmidt, M. Taubert, J. Seifert, F. Bastida, M. von Bergen, H. H. Richnow, C. Vogt, Nat. Protoc. 5 (2010) 1957-1966.
6. N. Jehmlich, F. Schmidt, M. Taubert, J. Seifert, M. Von Bergen, H. H. Richnow, C. Vogt, Rapid Commun. Mass Spectrom. 23 (2009) 1871-1878.
7. N. Jehmlich, I. Fetzer, J. Seifert, J. Mattow, C. Vogt, H. Harms, B. Thiede, H. H. Richnow, M. von Bergen, F. Schmidt, Mol. Cell. Proteomics 9 (2010) 1221-1227.
8. M. Taubert, N. Jehmlich, C. Vogt, H. H. Richnow, F. Schmidt, M. von Bergen, J. Seifert, Proteomics 11 (2011) 2265-2274.
9. I. Fetzer, N. Jehmlich, C. Vogt, H. H. Richnow, J. Seifert, H. Harms, M. von Bergen, F. Schmidt, BMC Res. Notes 3 (2010) 178.
10. M. Taubert, S. Baumann, M. von Bergen, J. Seifert, Anal. Bioanal. Chem. 401 (2011) 1975-1982.
11. F. Bastida, M. Rosell, A. G. Franchini, J. Seifert, S. Finsterbusch, N. Jehmlich, S. Jechalke, M. von Bergen, H. H. Richnow, FEMS Microbiol. Ecol. 73 (2010) 370-384.
12. F. Bastida, S. Jechalke, P. Bombach, A. G. Franchini, J. Seifert, M. von Bergen, C. Vogt, H. H. Richnow, FEMS Microbiol. Ecol. 77 (2011) 357-369.
13. A. R. Colmer, K. L. Temple, M. E. Hinkle, J. Bacteriol. 59 (1950) 317-328.
14. E. Drobner, H. Huber, K. O. Stetter, Appl. Environ. Microb. 56 (1990) 2922-2923.
15. A. B. Jensen, C. Webb, Process Biochem. 30 (1995) 225-236.
16. K. B. Hallberg, D. B. Johnson, Adv. Appl. Microbiol. 49 (2001) 37-84.
17. A. Amouric, C. Brochier-Armanet, D. B. Johnson, V. Bonnefoy, K. B. Hallberg, Microbiology 157 (2011) 111-122.
18. S. Hedrich, M. Schlömann, D. B. Johnson, Microbiology 157 (2011) 1551-1564.
19. D. B. Johnson, FEMS Microbiol. Ecol. 27 (1998) 307-317.
20. N. Kishimoto, Y. Kosako, N. Wakao, T. Tano, A. Hiraishi, Syst. Appl. Microbiol. 18 (1995) 85-91.
21. D. B. Johnson, K. B. Hallberg, in: Poole, R. K. (Ed.), Advances in microbial physiology, Academic Press 2009, pp. 201-255.
22. E. Gonzalez-Toril, E. Llobet-Brossa, E. O. Casamayor, R. Amann, R. Amils, Appl. Environ. Microbiol. 69 (2003) 4853-4865.
23. D. E. Rawlings, D. B. Johnson, Microbiology 153 (2007) 315-324.
24. J. D. Neufeld, M. Wagner, J. C. Murrell, ISME J 1 (2007) 103-110.

25. N. Finke, T. M. Hoehler, B. B. Jorgensen, Environ. Microbiol. 9 (2007) 1060-1071.
26. W. E. Huang, K. Stoecker, R. Griffiths, L. Newbold, H. Daims, A. S. Whiteley, M. Wagner, Environ. Microbiol. 9 (2007) 1878-1889.
27. N. Musat, R. Foster, T. Vagner, B. Adam, M. M. M. Kuypers, FEMS Microbiol. Rev. (2011)
28. K. Wakeman, H. Auvinen, D. B. Johnson, Biotechnol. Bioeng. 101 (2008) 739-750.
29. D. P. Kelly, L. A. Chambers, P. A. Trudinger, Anal. Chem. 41 (1969) 898-902.
30. R. I. Amann, W. Ludwig, K. H. Schleifer, Microbiol. Rev. 59 (1995) 143-169.
31. S. J. Giovannoni, E. F. DeLong, G. J. Olsen, N. R. Pace, J. Bacteriol. 170 (1988) 720-726.
32. M. O. Schrenk, K. J. Edwards, R. M. Goodman, R. J. Hamers, J. F. Banfield, Science 279 (1998) 1519-1522.
33. J. Peccia, E. A. Marchand, J. Silverstein, M. Hernandez, Appl. Environ. Microbiol. 66 (2000) 3065-3072.
34. J. Pernthaler, F. O. Glöckner, W. A. Schönhuber, R. Amann, Meth. Microbiol. 30 (2001) 207-226.
35. M. M. Bradford, Anal. Biochem. 72 (1976) 248-254.
36. J. Valdes, I. Pedroso, R. Quatrini, R. J. Dodson, H. Tettelin, R. Blake, 2nd, J. A. Eisen, D. S. Holmes, BMC Genom. 9 (2008) 597.
37. D. B. Johnson, F. F. Roberto, in: Rawlings, D. E. (Ed.), Biomining: theory, microbes and industrial processes., Springer-Verlag/Landes Bioscience, Georgetown, Texas 1997.
38. A. P. Harrison, Jr., Annu. Rev. Microbiol. 38 (1984) 265-292.
39. H. Liu, H. Yin, Y. Dai, Z. Dai, Y. Liu, Q. Li, H. Jiang, X. Liu, Arch. Microbiol. 193 (2011) 857-866.
40. D. B. Johnson, J. Microbiol. Meth. 23 (1995) 205-218.

JOURNAL OF INTEGRATED OMICS
A METHODOLOGICAL JOURNAL
HTTP://WWW.JIOMICS.COM

ORIGINAL ARTICLE | DOI: 10.5584/jiomics.v2i1.86

# Comparative proteomic map among *van*A-containing *Enterococcus* isolated from yellow-legged gulls

Hajer Radhouani[1-4], Patrícia Poeta[3,4], Luís Pinto[1-4], Ricardo Monteiro[1-4], Júlio Nunes-Miranda[1-2], Susana Correia[1-4], Selma Vieira[1-4], Carlos Carvalho[1,2], Jorge Rodrigues[3,4], María López[5], Carmen Torres[5], Rui Vitorino[6], Pedro Domingues[6], Gilberto Igrejas[1,2,*].

[1]Institute for Biotechnology and Bioengineering, Center of Genomics and Biotechnology, University of Trás-os-Montes and Alto Douro, Vila Real, Portugal; [2]Department of Genetics and Biotechnology, University of Trás-os-Montes and Alto Douro; Vila Real, Portugal; [3]Center of Studies of Animal and Veterinary Sciences, Vila Real, Portugal; [4]Veterinary Science Department, University of Trás-os-Montes and Alto Douro, Vila Real, Portugal; [5]Biochemistry and Molecular Biology Area, University of La Rioja, Logroño, Spain; [6]Chemistry Department, University of Aveiro, Aveiro, Portugal.

**Received:** 21 December 2011 **Accepted:** 03 March 2012 **Available Online**: 08 March 2012

ABSTRACT


The increase of vancomycin resistance enterococci (VRE) represents a threat for patient care and creates a reservoir of mobile resistance genes for other, more virulent pathogens. The existence of VRE in different ecologic niches complicates the understanding of its epidemiology. The aim of the present study was to analyse the proteome of 2 *van*A strains recovered from seagull faecal samples. The *van*A *E. durans* and *van*A *E. faecium* strains presented different genomic patterns: *tet*(M)-*tet*(L)-*erm* (B) and *tet*(M)-*tet*(L)-*erm*(B)-*hyl*, respectively. A total of 123 spots were excised from two-dimensional gel electrophoresis (2-DE) gel of *van*A *E. durans* SG 2 strain, and 16 were successfully identified by using MS, representing 42 different proteins. For the *van*A *E. faecium* SG 50 strain, 93 spots were excised from the 2-DE gel and 23 were identified, representing 47 different proteins. The vancomycin/teicoplanin A-type resistance protein VanA in *E. durans* SG 2 strain was present in two different spots. The identified proteins have shown diverse functional activities including glycolysis, conjugation, translation, protein biosynthesis, among others. This work reports the impact of proteomics on knowledge of *van*A-containing enterococci and will be helpful to further understand the expression of proteins in antibiotic-resistant bacteria.

**Keywords:** Antimicrobial resistance; Bioinformatics; Databases; Proteomics; *Van*A-containing *Enterococcus*; Virulence factors.

## Abbreviations

**2-DE:** two-dimensional polyacrylamide gel electrophoresis; **PCR**: Polymerase Chain Reaction; **MALDI-TOF MS**: matrix-assisted laser desorption ionization time-of-flight mass spectrometry; **spp.**: subspecies; **MIC**: minimal inhibitory concentration

## 1. Introduction

Proteome analysis is a tool that can be used to visualize and, also, to compare complex mixtures of proteins, and to acquire important information about the individual proteins concerned in specific biological responses [1]. Proteomic is defined as the expressed complement of a genome. In comparison to the genome, the proteome is used to define a series of proteins expressed by a certain organism, under certain conditions, serving as unique and informative infor-

***Corresponding author:** Gilberto Igrejas, Institute for Biotechnology and Bioengineering, Centre of Genomics and Biotechnology, Department of Genetics and Biotechnology, University of Trás-os-Montes and Alto Douro, 5001-801 Vila Real, Portugal, Tel: + 351 259 350 530; Fax: + 351 259 350 572; E-mail: gigrejas@utad.pt.

mation of both its phenotypic state. This results in cell responses to physiological and environmental perturbations, and genomic information reflected in the amino acid sequences of expressed proteins. Therefore, the principal concern of proteomic resides, in the identification of proteins related to, in particular, cellular processes or presenting altered expression profiles as a consequence of different physiological conditions [2].

Various analytical tools such as two dimensional gel electrophoresis, mass spectrometry and searches in generalist and Expressed Sequence Tags (EST) databases, modified the protein identification process [3]. Although not usually used, proteomic approaches participate in determining antimicrobial resistance mechanism(s) and other cell metabolic alterations through the ability to study overall changes in bacteria [4]. The model of protein species linked to the antimicrobial resistance has been investigated in a diversity of microorganisms and with different antimicrobial agents [4-6]. The evaluation of changes in protein profiles in response to various mechanisms of stress, such as the susceptibility to antimicrobial agents or the modifications related to antibiotic resistance could represent an integrating method for the development of new therapeutic treatment and antimicrobial agents. Bacterial surface proteins are important for the host-pathogen interaction and they are commonly implicated in disease pathogenesis [4].

*Enterococcus* spp. live as commensals of the gastrointestinal tract of warm-blooded animals, being the most abundant Gram-positive cocci in humans and in animals [7]. Recently, the incidence of nosocomial enterococal infections has increased distinctly [8]. This genus is also recognized as important opportunistic pathogens, and reveals intrinsic resistance to a number of antimicrobial agents, in addition to the acquired multidrug resistance [9].

The increase of vancomycin-resistant enterococci (VRE) causes several challenges. Firstly, most of VRE are frequently also resistant to other available drugs e.g., aminoglycosides or ampicillin. Secondly, there is the possibility that the vancomycin resistance genes present in VRE could be transferred to other gram-positive microorganisms. In addition to the currently common detection of multiresistant bacteria in areas with high human density [10] their emergence in more remote areas like high mountain regions or natural reserve is even more alarming [11-13]. Monitoring the prevalence of resistance in indicator bacteria such as vancomycin-resistant enterococci in different populations, animals, patients and healthy humans, makes it feasible to compare the prevalence of resistance and to detect the transfer of resistant bacteria or resistance genes from animals to humans and vice versa [14]. Microbial resistance to antibiotics is a worldwide problem in human and veterinary medicine. Commonly, it is usual that the main risk factor for the increase of this situation is an extensive use of antibiotics that leads to the dissemination of resistant bacteria and resistance genes in animals and humans [15].

The wild birds seem to represent a significant reservoir, or at least source of *van*A enterococcal strains. Consequently, this may represent a significant hazard to human and animal health by transmitting these strains into waterways and other environmental sources via their faecal deposits. Although wild birds rarely come into contact with antimicrobial agents, arguing against the existence of direct selective pressure on birds, nevertheless they can be infected or colonized by resistant bacteria. Water contact and acquisition via food seem to be major aspects of transmission of the resistant bacteria of human or veterinary origin to wild animals [10,16]. Wild birds or wild animals in general could, therefore, serve as reservoirs of resistant bacteria and genetic determinants of the antimicrobial resistance [11].

In the present study, we examined the proteome of 2 *van*A strains recovered from seagull faecal samples. This evaluation was carried out in order to compare the proteins obtained from these strains with the results obtained in a previous published study using a different group of strains [13]. The combination of two high-resolution methods, isoelectric focusing and SDS polyacrylamide gel electrophoresis, permitted the separation of numerous proteins of *van*A-containing *Enterococcus* isolates and highlight the presence of vancomycin/teicoplanin A-type resistance protein *van*A.

## 2. Materials and Methods

### 2.1 *Samples and bacteria*

The phenotypic and genetic profiles of two *van*A-containing *Enterococcus* strains (*van*A *E. durans* SG2 and *van*A *E. faecium* SG50) as well as the SDS-PAGE of whole-cell extracts of them were studied in a previous report [13]. The complete proteomic analysis of these two *van*A-containing strains SG2 and SG50 has been the objective of the present study. These strains were previously obtained from faecal samples of yellow-legged seagulls, randomly recovered in the beaches of Berlengas Islands National Reserve of Portugal [13].

### 2.2 *Virulence factor genes*

The presence of genes encoding different virulence factors (*gelE*, *fsr*, *ace, cpd*, *agg and cyl*$L_LL_S$ABM) was verified by PCR using primers and conditions previously described [17-19]. Positive and negative controls obtained from the collection of the University of Trás-os-Montes and Alto Douro (Portugal) were included in all assays. The presence of *hyl* or *esp* gene in these strains was previously reported [13].

### 2.3 *Assay of gelatinase activity*

Gelatinase production was detected by inoculating the enterococci onto freshly prepared tryptic soy agar plates (Difco; 236950, Le Pont de Claix, France) containing 1.5% of skim milk (Difco; 232100). Plates were incubated overnight at 37° C and then cooled to ambient temperature for 2 h. The ap-

pearance of a transparent halo around the colonies was considered to be a positive indication of gelatinase production [20].

2.4 *Assay of hemolytic activity*

The production of hemolysin was determined by streaking bacterial cultures, grown overnight at 37°C in brain heart infusion agar (Difco; 241830), on columbia agar plates supplemented with 5% of horse blood (BioMérieux; 43050, La Balme, Les Grottes, France). Plates were incubated at 37°C for 72h in aerobic conditions and after that the plates were examined for haemolysis. The haemolytic reaction was recorded by the observation of a clear zone of hydrolysis around the colonies (β-haemolysis), a partial hydrolysis (α-haemolysis) and a non reaction (γ-haemolysis). When observed, greenish zones around the colonies were interpreted as α-haemolysis and taken as negative for the assessment of β-haemolytic activity [19].

2.5 *PCR amplification of pbp5 gene*

Total DNA was extracted from *van*A *E. faecium* SG 50 isolate by the InstaGene Matrix (Bio-Rad; 732-6030, Hercules, United States), and 10µl of DNA was used for the PCR reaction, the $MgCl_2$ concentration being of 3.5 mM. The primers used for PCR amplification of *pbp*5 gene were the following ones: F 5´-AACAAAATGACAAACGGG-3´; R 5´-TATCCTTGGTTATCAGGG-3`. PCR conditions were as follows: 95°C initially for 15 min; 94°C for 30 s, 54°C for 30 s, 72°C for 2 min over 30 cycles followed by a final 7 min extension period at 72°C [21].

2.6 *DNA sequence analysis*

The *pbp5* PCR products were purified with the QiaQuick PCR purification kit (Qiagen Inc.) according to the instructions of the manufacturer. The purified products were sequenced in both strands on the ABI Prism 3700 DNA sequencer (Perkin-Elmer). The obtained sequences were compared through bioinformatics tools to that of *pbp*5 included in GenBank accession no. X84860.

2.7 *Protein extraction*

Frozen *van*A-containing *Enterococcus* cell stocks were streaked onto Luria-Bertani (LB) plates and grown at 37ºC. Single colonies of *van*A-containing *Enterococcus* strains were conducted in 250 mL of M9 minimal medium supplemented with 4 $gL^{-1}$ glucose in covered 1 L Erlenmeyer flasks at 37 ºC. Cells were harvested from the exponential phase in all experiments. The cells were pelleted down at 10,000 rpm at 4ºC for 3 min. The pellet was supposed to be visible after spinning and resuspended in an equal volume of pre-warmed phosphate-buffered saline (PBS) pH 7.4 [22]. After new centrifugation pellet was suspended in 0.2 ml of SDS sample solubilization buffer. The sample was sonicated with an ultrasonic homogenizer (6 × 10 s, 4 °C at 100 W). The disrupted cells were centrifuged in an Eppendorf microfuge at maximum speed (14,000 rpm) for 30 minutes at 4°C. For SDS-PAGE experiment the supernatant was collected and resuspended in an equal volume of buffer containing 0.5 M Tris HCl pH 8.0, glycerol, SDS and bromophenol blue.

2.8 *Two-dimensional electrophoresis and proteomics*

The 2-DE was performed according to the principles of O'Farrell [23] but with IPG (Immobiline™ pH Gradient) technology [24]. Protein samples of *van*A *E. durans* isolate (SG 2) were used in parallel with those of *van*A *E. faecium* isolate (SG 50) proteins. For IEF, precast IPG strips with linear gradient of pH 4-7 were passively rehydrated overnight (12 to 16 hours) in a reswelling tray with rehydration buffer (8M urea, 1% CHAPS, 0.4% DTT, 0.5% carrier ampholyte IPG buffer pH 3-10) at room temperature and IPG strips were covered with DryStrip Cover Fluid (Plus One, Amersham Biosciences). Lyses buffer [9.5M urea, 1% (w/v) DTT, 2% (w/v) CHAPS, 2% (v/v) carrier ampholytes (pH 3-10) and 10 mM Pefabloc® proteinase inhibitor] was added to the two *van*A-positive enterococci (1:1). Samples containing a total of 73.5 µg of protein were loaded into 13 cm IPG strips (pH 4–7 NL, Amersham Biosciences, UK) [25]. The sample solution was then applied to the previously rehydrated IPG strips pH 4-7 by cup loading and then proteins were focused sequentially at 500 V for 1 h, 1000 V for 2 h, 8000 V for 2 h, 1000 V for 4h and, finally, 1000 V for 55min incremented to 23208 V/h on an EttanTM IPGPhor IITM (Amersham Biosciences, Uppsala, Sweden). Seven IEF replicate runs were performed according to Görg [24] and the GE Healthcare protocol for IPG strips pH 4-7 of 13 cm, in order to obtain the optimized running conditions, resulting in a final around 10 hour run. Focused IPG strips were then stored at -80ºC in plastic bags. Before running the second dimension, strips were equilibrated twice for 15 minutes in an equilibration buffer (6 M urea, 30% (w/v) glycerol, 2% (w/v) SDS in 0.05 M Tris-HCl buffer (pH 8.8)). In the first equilibration, 1% DTT was added to the original equilibration buffer and 4 % iodoacetamide to the second one, and, also, bromophenol blue was added to both solutions. The equilibrated IPG strips were gently rinsed with SDS electrophoresis buffer, blotted to remove excessive buffer, and then applied onto a 12.52% polyacrylamide gels in a Hoefer™ SE 600 Ruby® (Amersham Biosciences) unit. Some modifications were introduced in the SDS-PAGE technique previously reported by Laemmli [26], that allowed its resolution to be increased with a proper insertion of the IPG strips in the stacking gel [26,27]. After SDS-PAGE, the 2-DE gels were fixated on 40% methanol / 10% acetic acid for one hour and, afterwards, stained overnight in Coomassie Brilliant Blue G-250 [22]. Coomassie-stained gels were scanned with a flatbed scanner (Umax PowerLook 1100; Fremont, CA, USA), and the resulting digitized images were analyzed using Image

Master 5.0 software (Amersham Biosciences; GE Healthcare).

2.9 *Protein identification by MALDI-TOF/TOF*

Spots of expression in all gels were manually excised from the gels and analyzed using Matrix-Assisted Laser Desorption/Ionization-Time of Flight Mass Spectrometry (MALDI-TOF). The gel pieces were washed three times with 25mM ammonium bicarbornate/50 % ACN, one time with ACN and dried in a SpeedVac (Thermo Savant). 25 mL of 10mg/mL sequence grade modified porcine trypsin (Promega) in 25mM ammonium bicabornate was added to the dried gel pieces and the samples were incubated overnight at 37°C. The extraction of tryptic peptides was performed by adding 10% of formic acid (FA)/50% ACN three times and being lyophilised in a SpeedVac (Thermo Savant). Tryptic peptides were ressuspended in 10 mL of a 50% acetonitrile/0.1% formic acid solution. The samples were mixed (1:1) with a matrix consisting of a saturated solution of a-α–cyano-4-hydroxycinnamic acid prepared in 50% acetonitrile/ 0.1% formic acid. Aliquots of samples (0.5µL) were spotted onto the MALDI sample target plate. Peptide mass spectra were obtained through or from a MALDI-TOF/TOF mass spectrometer (4800 Proteomics Analyzer, Applied Biosystems, Europe) in the positive ion reflector mode.

Spectra were obtained in the mass range between 800 and 4500 Da with ca. 1500 laser shots. For each sample spot, a data dependent acquisition method was created to select the six most intense peaks, excluding those from the matrix, trypsin autolysis, or acrylamide peaks, for subsequent MS/MS data acquisition. Mass spectra were internally calibrated with autodigest peaks of trypsin (MH+: 842.5, 2211.42 Da) allowing a mass accuracy of better than 25 ppm.

2.10 *Database search*

Spectra were processed and analyzed by the Global Protein Server Workstation (Applied Biosystems), which uses internal MASCOT software (v 2.1.04, Matrix Science, London, UK) on searching the peptide mass fingerprints and MS/MS data. Swiss-Prot nonredundant protein sequence database was used for all searches under *Enterococcus*. Database search parameters are as follows: carbamidomethylation and propionamide of cysteine (+71Da) as a variable modification as well as oxidation of methionine (+16Da), and the allowance for up to two missed tryptic cleavages. The peptide mass tolerance was 25 ppm and the fragment ion mass tolerance was 0.3 Da. Protein identifications were considered as reliable when the MASCOT score was > 70 (MASCOT score was calculated as - 10 × log P, where *P* is the probability that the observed match is a random event.). This is the lowest score indicated by the program as significant ($P < 0.05$) and indicated by the probability of incorrect protein identification. All spectra were also processed in a reversed decoy database created for Swiss-Prot (consisting of normal and reverse sequences) to allow estimation of the false discovery rate (false positive peptides/(false positive peptides + total peptides))*100) which is by routine below than 1%.

2.11 *Sequence alignments and construction of the phylogenetic tree*

The analysis was performed on the Phylogeny.fr platform and comprised the following steps. Sequences were aligned with MUSCLE (v3.7) configured for the highest accuracy (MUSCLE with default settings). After the alignment, the positions with gap were removed from the alignment. The phylogenetic tree was reconstructed using the maximum likelihood method implemented in the PhyML program (v3.0 aLRT). The default substitution model was selected assuming an estimated proportion of invariant sites (of 0.000) and 4 gamma-distributed rate categories to account for the rate heterogeneity across sites. The gamma shape parameter was fixed (alpha=12.4). The reliability for internal branch was assessed using the aLRT test (SH-Like). A graphical representation of the phylogenetic tree (phenogram) was performed with Drawgram from the PHYLIP package (v3.66) [26].

## 3. Results

3.1 *Characteristics of the two van*A *strains included in the study*

The phenotype and genotype of antibiotic resistance of the two *van*A *Enterococcus* strains included in this study were previously reported [13]. Both strains showed high level of vancomycin and teicoplanin resistance, as well as resistance to tetracycline and erythromycin. *E. faecium* strain SG 50 showed also resistance to ciprofloxacin and ampicillin. The *van*A *E. durans* and *van*A *E. faecium* strains presented different genomic patterns: *tet*(M)-*tet*(L)-*erm*(B) and *tet*(M)-*tet*(L)-*erm*(B)-*hyl*, respectively.

Genes encoding virulence factors were studied in the two *van*A-containing *Enterococcus* isolates (SG2 and SG50). The *cpd* gene was detected in both *van*A enterococci. The $cylL_L$ gene was identified in the *van*A *E. durans* SG 2 strain . The two *vanA* strains expressed gelatinase activity and carried the *gel*E gene. No beta-haemolytic activity was identified in both strains, although they showed alfa-hemolysis.

The sequence of the C-terminal region of *pbp5* was analyzed in our *van*A *E. faecium* SG 50 strain that showed ampicillin-resistance. This strain presented 11 amino acid substitutions in PBP5 protein (408Q → H, 427I → M, 470H → Q, 485M → A, 496N → K, 497F → I, 499A → T, 525E → D, 586V → L, 629E → V, 634N → Q), in relation with the reference one.

3. 2 *Two-dimensional Electrophoresis*

The SDS-PAGE of whole-cell extracts of the 2 *van*A-

containing enterococci strains are shown in a previous study [13]. In the present study, a comparative analysis among the strains has been carried out. The protein expressions of the two vancomycin-containing enterococci (*E. durans* SG 2 and *E. faecium* SG 50) strains were analysed. *E. faecium* SG 50 was presented on 2-DE gel (Figure 1). The use of pH 4-7 IPG strips resulted in a well spread protein spots map which contributed to accurate image identification and the safe excision of the spots. For each sample of SG 2 and SG 50 strains, a total of 123 and 93 relevant protein spots, respectively, were collected for analysis using MALDI-TOF mass spectrometry (Figure 2). These spots are consistently and systematically present in all replicates. The peptide mass peaks were compared with those in the NCBI database (http://www.ncbi.nlm.nih.gov/). The protein identification data including Genebank ID, MW, PI value, mascot score, number of matched peptides and sequence coverage ratio (%) are listed in Table 1 for and *E. durans* SG 2 and Table 2 for the *E. faecium* SG 50 proteins. The identified proteins were showing diverse functional activities, including glycolysis, conjugation, translation, protein biosynthesis, among others Replicate sequences, truncated sequences, and sequences with partial alignments were removed from the BLAST results (not shown). From the collected sequences were selected to represent the initial tree. These sequences were aligned and a phylogenetic tree was constructed by using the mini-

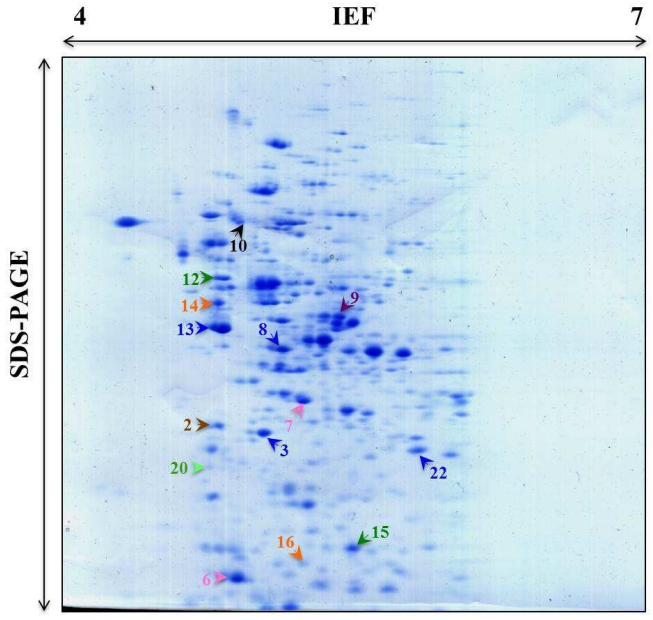


**Figure 1.** 2-DE gel image of SG 50 VRE with IPG strips pH4-7. Arrows indicates peptides that are part of one or more proteins whose have a role in such pathways: Blue: Glycolysis, Orange: Stress response, Pink: Protein folding/ Protein biosynthesis, Brown: Pyrimidine biosynthesis, Bright green: Transcription, Light green: ATP synthesis, Violet: Arginine metabolism, Black: Phosphotransferase system .

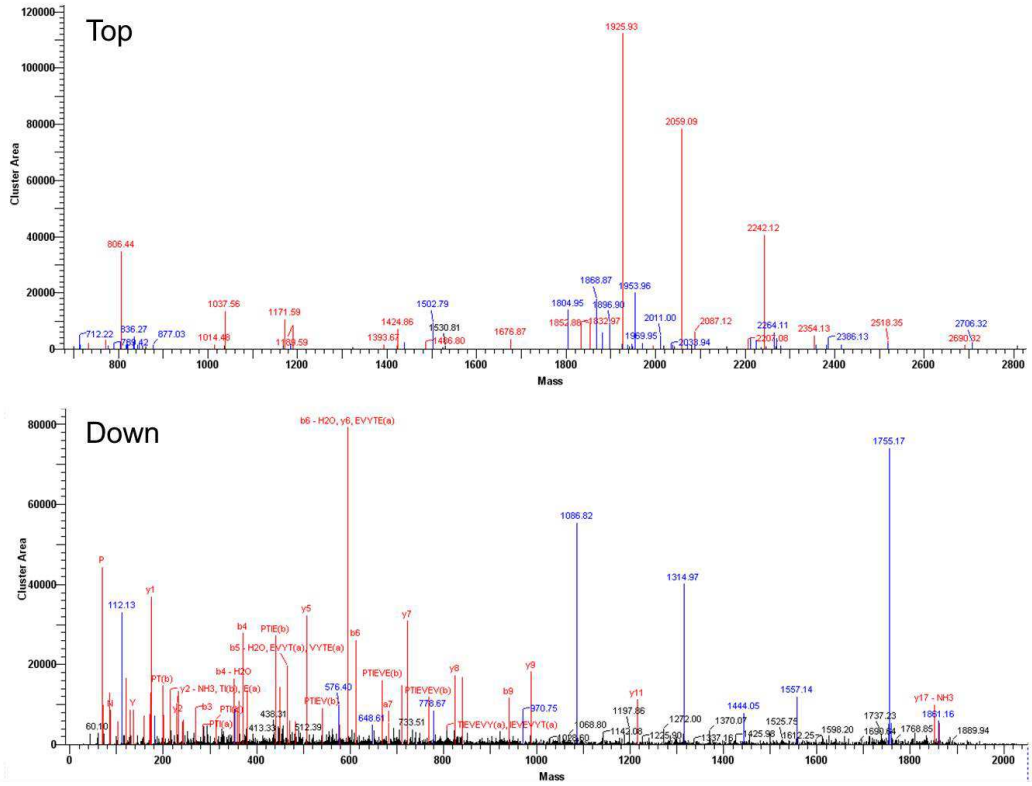


**Figure 2.** MALDI/MS spectra obtained for Enolase from *van*A *E. durans* SG 2 strain. Top - MS analysis of tryptic peptides. Peaks in red matching enolase. Down - MALDI-TOF-TOF MS tandem spectrum of tryptic peptide [M+H]+; m/z 1925.93 identified as enolase. The identified sequence is GNPTIEVEVYTESGAFGR. On spectrum, peaks in red correspond to y and b series .

mum-evolution method to root the tree. The clustering of the initial phylogenetic tree indicated that all of the proteins included in the data set diverged from a common ancestor (Figure 3).

## 4. Discussion

The worldwide appearance of antibiotic-resistant bacteria causes a severe threat to human health. Furthermore, it can add to the difficulties in controlling infectious diseases, the phenotype of resistance can generate metabolic changes that, in turn, can interfere with host-pathogen interactions. The commensal bacteria such as enterococci can carry virulence factors, and in that case, the pathogenic capacity can increase [17]. The two *van*A-containing *Enterococcus* strains included in the study were previously characterized for antibiotic resistance.

In our study, the *cpd* (a sex pheromone determinant) gene was detected in both *van*A strains (*E. durans* and *E. faecium*), which is similar to the results reported by others [28]. The isolates with sex-pheromone determinants have the potential to acquire the respective sex-pheromone plasmids and, hence, the associated virulence and resistance determinants. So, the sex pheromone production may promote the acquisition of vancomycin resistance and other linked traits from *E. faecium* strains and lead to an increased virulence [17].

Gelatinase, encoded by the *gel*E gene, is an extracellular zinc endopeptidase that hydrolyses collagen, gelatin, hemoglobin and other bioactive compounds, and it has been shown to exacerbate endocarditis in an animal model, although this activity is not required for pathogenesis [29]. Usually, the *gel*E gene has been detected with higher frequency suggesting that this virulence determinant is a common trait in the genus *Enterococcus*. In our study, the *gel*E was identified in both *van*A strains. Same results were observed in other reports [28,30]. The α-haemolysis could be a result of oxidation and a consequent lysis of the erythrocytes due to factors other than the production of enterococcal hemolysin [19]; therefore, the incubation under anaerobic conditions seems to be more reliable.

In this work, we performed a proteome analysis of two *van*A strains (*E. durans* SG 2 and *E. faecium* SG 50). A total of 123 spots were excised from 2-DE gel of SG 2 and 16 were successfully identified by MS, representing 42 different proteins. For the SG 50 strain, 93 spots were excised from the 2-DE gel and 23 were identified, representing 47 different proteins. It is important to highlight the presence of vancomycin/teicoplanin A-type resistance protein VanA in *E. durans* SG 2 strain, in two different spots (12 and 14). VanA is a protein capable of utilizing both hydroxyl acids and D-Ala as substrates with a concomitant switch from ester to peptide bond formation dependent on pH [28]. Vancomycin inhibits the extracellular steps of peptidoglycan synthesis by binding it to the C-terminal D-alanyl-D-alanine (DAla- D-Ala) residues of cell wall precursors in enterococci [32]. The D-Ala-D-Ala target residues are synthesized intracellularly as a dipeptide by a D-Ala:D-Ala ligase. Then, they are added to UDP-Nacetylmuramyl- L-Ala-g-D-Glu-L-Lys (UDP-MurNac-tripeptide) by using an adding enzyme [33]. The vancomycin/teicoplanin A-type resistance protein was also identified in *van*A-*E. durans* SG 3 isolate [13,34].

The presence of vancomycin/teicoplanin A-type resistance protein in vancomycin-resistant enterococci in natural environments could have impact on animal and human health. The emergence of vancomycin-resistant enterococci of human and veterinary origin [35], presumably leading to a great health concern [36].

Several proteins were found in multiple spots on the two gels. From the 42 different proteins identified in *van*A *E. durans* SG 2 isolate, five of them were involved in stress response. Usually, stress response requires heat molecular chaperones or shock proteins that preserve protein function or repair damage after cell injury. As such, the integrity of chaperone systems can seriously modify the progression of diseases associated with ageing, DNA damage and chronic injury [37]. While the molecular chaperone proteins are among the most evolutionarily preserved proteins and have a ubiquitous function in all repair processes, there is a high degree of tissue specificity in chaperone induction [38,39] showing that some cells have developed unique stress responses due to unique micro-environmental pressures. DnaK are abundant heat shock proteins that function as chaperones inside the bacterial cytoplasm [40]. The bacteria that overproduce the DnaK protein at all temperatures undergo a considerably reduced heat-shock response at high temperature. The DnaK protein is identified as an inhibitor of the heat-shock response [41] a very important reaction for the survival of bacteria such as enterococci, and that contributes to the antibiotic resistance. The protein DnaK was detected in spot 3 as linked to the *Enterococcus faecalis* strain (accession number Q835R7), the *Streptococcus agalactiae* serotype III strain (accession number P0A3J2), the *Streptococcus thermophilus* strain (Q5M6D1), the *Streptococcus pyogenes* serotype M12 strain (Q1JKD6) and the *Streptococcus mutans* (O06942).

It is important to underline the presence of both DnaK and DnaJ proteins (spot 14 and spot 16) in *E. faecium* SG 50 strain as related to the *Yersinia pseudotuberculosis* serotype O:1b and to the *Enterococcus faecalis* (*Streptococcus faecalis*), respectively. DnaJ proteins participate actively in the response to hyperosmotic and heat shock by preventing the aggregation of the stress-denatured proteins and by disaggregating proteins, also in an autonomous, DnaK-independent fashion. DnaJ proteins are identified as co-chaperones because they help another family of chaperones (DnaKs) with protein folding. The DnaJ and DnaK proteins must act together to facilitate protein folding [42].

*In vitro* the GroEL proteins raise the yield of functional protein during refolding by suppressing aggregation as a side-reaction and, probably, by shifting the substrate protein from off-pathway reactions back to the productive folding

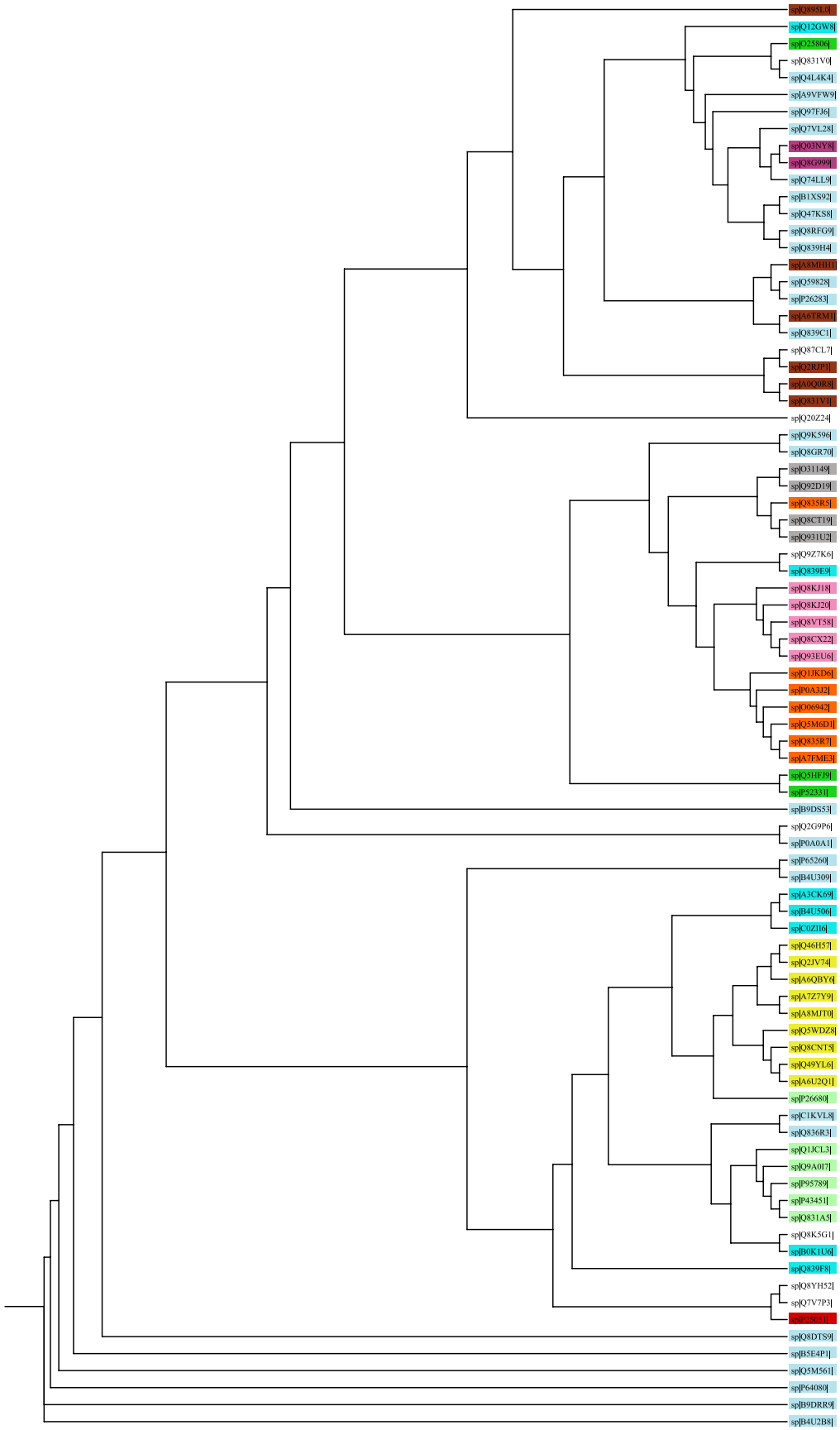


**Figure 3.** Phylogenetic tree of FASTA protein sequences of all proteins identified. The full alignment of these sequences were done with MUSCLE (v3.7) configured for highest accuracy.

pathway [43]. The GroEL protein is the major heat shock protein of a large number of bacteria and belongs to the chaperonin family. This protein avoids the misfolding of proteins and promotes the refolding and the proper assembly of unfolded polypeptiedes caused under stress condition [42]. In the *van*A *E. faecium* SG 2 isolate, the spot 4 showed the presence of the GroL protein (Q93EU6, Q8KJ20, Q8KJ18, Q8VT58 and Q8CX22) associated to the *Enterococcus faecalis*, the *Streptococcus anginosus*, the *Streptococcus constellatus*, the *Streptococcus gordonii*, the S*treptococcus agalactiae* serotype III, respectively. The dnaK and groL proteins were also detected in the enterococci strains from seagulls [13] and in the *Salmonella* strains from wild boars [44].

In the analysis of the obtained proteomes, the most abundant proteins identified in our *van*A SG 2 and *van*A SG 50 strains were those involved in glycolysis. In our study, the high number of proteins linked to the ATP synthesis, transferase, translation and protein folding is emphasized.

It is highly important to point out that proteins from *van*A enterococcal strains were identified in the obtained proteomes, some of which are involved in the antibiotic resistance. These proteins were controlled by vancomycin which, also, triggered innate signal regulators, adhesion factors, and metabolic gene expression in *E. faecalis*. Therefore, these responses may enable *Enterococcus* spp. to adapt, survive and remain pathogenic even under the pressure of the vancomycin treatment. Our results are in accordance with those observed in another report [45].

## 5. Concluding remarks

Our report showed that it became possible through a detailed proteomic approach and a 2-DE combined with mass spectrometry (MALDI/TOF-TOF) to obtain important information for further understanding of the antibiotic-resistant mechanism(s), but also for the evaluation of protein profiles in response to various stress mechanisms, such as sensitivity to antibiotics or modifications related to antibiotic resistance. All of these could represent a valid and integrating approach to the development of new therapeutic strategies [46]. In fact, the elaboration of a 2-D electrophoresis gel of the 2 *van*A enterococcal strains with phenotypic and genotypic profiles, indicating antimicrobial resistance, permitted us to identify and characterize the present proteins.

## 6. Supplementary material

Supplementary data and information is available at: http://www.jiomics.com/index.php/jio/rt/suppFiles/86/0

Table 1. Protein spots identification of 2-DE gels and MALDI-TOF sequencing results from *van*A *E. durans* SG 2 isolate .

Table 2. Protein spots identification of 2-DE gels and MALDI-TOF sequencing results from *van*A *E. faecium* SG 50 isolate .

## Funding

Hajer Radhouani was supported by a grant with reference SFRH/BD/60846/2009 of the Fundação para a Ciência e a Tecnologia (FCT) from Portugal.

## Competing interests

The authors declare that they have no competing interests.

## References

1. Patterson, S.D., Aebersold, R.H., 2003, Proteomics: the first decade and beyond. Nature genetics 33 Suppl, 311-323.
2. Fish, D.N., Ohlinger, M.J., 2006, Antimicrobial resistance: factors and outcomes. Crit Care Clin 22, 291-311, vii.
3. Jungblut, P.R., Holzhutter, H.G., Apweiler, R., Schluter, H., 2008, The speciation of the proteome. Chem Cent J 2, 16.
4. Cash, P., Argo, E., Ford, L., Lawrie, L., McKenzie, H., 1999, A proteomic analysis of erythromycin resistance in *Streptococcus pneumoniae*. Electrophoresis 20, 2259-2268.
5. Andrade, H.M., Murta, S.M.F., Chapeaurouge, A., Perales, J., Nirde, P., Romanha, A.J., 2008, Proteomic analysis of *Trypanosoma cruzi* resistance to benznidazole. Journal of Proteome Research 7, 2357-2367.
6. Yoo, J.S., Seong, W.K., Kim, T.S., Park, Y.K., Oh, H.B., Yoo, C.K., 2007, Comparative proteome analysis of the outer membrane proteins of in vitro-induced multi-drug resistant *Neisseria gonorrhoeae*. Microbiol Immunol 51, 1171-1177.
7. Tannock, G.W., Cook, G., 2002, Enterococci as members of the intestinal microflora of humans. Enterococci: Pathogenesis, Molecular Biology, and Antibiotic Resistance, 101-132 439.
8. Udo, E.E., Al-Sweih, N., Phillips, O.A., Chugh, T.D., 2003, Species prevalence and antibacterial resistance of enterococci isolated in Kuwait hospitals. J Med Microbiol 52, 163-168.
9. Coque, T.M., Arduino, R.C., Murray, B.E., 1995, High-level resistance to aminoglycosides: comparison of community and nosocomial fecal isolates of enterococci. Clin Infect Dis 20, 1048-1051.
10. Cole, D., Drum, D.J., Stalknecht, D.E., White, D.G., Lee, M.D., Ayers, S., Sobsey, M., Maurer, J.J., 2005, Free-living Canada geese and antimicrobial resistance. Emerg Infect Dis 11, 935-938.
11. Dolejska, M., Cizek, A., Literak, I., 2007, High prevalence of antimicrobial-resistant genes and integrons in *Escherichia coli* isolates from Black-headed Gulls in the Czech Republic. J Appl Microbiol 103, 11-19.
12. Pinto, L., Poeta, P., Radhouani, H., Coelho, A.C., Carvalho, C., Rodrigues, J., Torres, C., Vitorino, R., Domingues, P., Igrejas, G., 2011, Proteomic study in an *Escherichia coli* strain from seagulls of the Berlengas Natural Reserve of Portugal. Journal of Integrated OMICS 1.
13. Radhouani, H., Poeta, P., Pinto, L., Miranda, J., Coelho, C., Carvalho, C., Rodrigues, J., Lopez, M., Torres, C., Vitorino, R., Domingues, P., Igrejas, G., 2010b, Proteomic characterization of *van*A-containing *Enterococcus* recovered from Seagulls at the Berlengas Natural Reserve, W Portugal. Proteome Sci 8, 48.
14. Martel, J.L., Tardy, F., Sanders, P., Boisseau, J., 2001, New trends in regulatory rules and surveillance of antimicrobial resistance in bacteria of animal origin. Vet Res 32, 381-392.
15. van den Bogaard, A.E., Stobberingh, E.E., 2000, Epidemiology

of resistance to antibiotics. Links between animals and humans. Int J Antimicrob Agents 14, 327-335.

16. Kozak, G.K., Boerlin, P., Janecko, N., Reid-Smith, R.J., Jardine, C., 2009, Antimicrobial resistance in *Escherichia coli* isolates from swine and wild small mammals in the proximity of swine farms and in natural environments in Ontario, Canada. Appl Environ Microbiol 75, 559-566.
17. Eaton, T.J., Gasson, M.J., 2001, Molecular screening of *Enterococcus* virulence determinants and potential for genetic exchange between food and medical isolates. Appl Environ Microbiol 67, 1628-1635.
18. Pillai, S.K., Sakoulas, G., Gold, H.S., Wennersten, C., Eliopoulos, G.M., Moellering, R.C., Jr., Inouye, R.T., 2002, Prevalence of the fsr locus in *Enterococcus faecalis* infections. J Clin Microbiol 40, 2651-2652.
19. Semedo, T., Almeida Santos, M., Martins, P., Silva Lopes, M.F., Figueiredo Marques, J.J., Tenreiro, R., Barreto Crespo, M.T., 2003, Comparative study using type strains and clinical and food isolates to examine hemolytic activity and occurrence of the *cyl* operon in enterococci. J Clin Microbiol 41, 2569-2576.
20. Jett, B.D., Huycke, M.M., Gilmore, M.S., 1994, Virulence of enterococci. Clin Microbiol Rev 7, 462-478.
21. Jureen, R., Top, J., Mohn, S.C., Harthug, S., Langeland, N., Willems, R.J., 2003, Molecular characterization of ampicillin-resistant *Enterococcus faecium* isolates from hospitalized patients in Norway. Journal of clinical microbiology 41, 2330-2336.
22. Görg, A., Weiss, W., Dunn, M.J., 2004, Current two-dimensional electrophoresis technology for proteomics. Proteomics 4, 3665-3685.
23. Görg, A., Klaus, A., Luck, C., Weiland, F., Weiss, W., 2007, Two-Dimensional Electrophoresis with Immobilized pH Gradients for Proteome Analysis. Technische Universität Munchen.
24. Görg, A., Obermaier, C., Boguth, G., Harder, A., Scheibe, B., Wildgruber, R., Weiss, W., 2000, The current state of two-dimensional electrophoresis with immobilized pH gradients. Electrophoresis 21, 1037-1053.
25. Dereeper, A., Guignon, V., Blanc, G., Audic, S., Buffet, S., Chevenet, F., Dufayard, J.F., Guindon, S., Lefort, V., Lescot, M., Claverie, J.M., Gascuel, O., 2008, Phylogeny.fr: robust phylogenetic analysis for the non-specialist. Nucleic Acids Res 36, W465-469.
26. Laemmli, U.K., 1970, Cleavage of structural proteins during the assembly of the head of bacteriophage T4. Nature 227, 680-685.
27. Igrejas, G., 2000. Genetic, biochemical and technological factors associated to the utilization of common wheat (*Triticum aestivum* L.). Univesity of Trás-os-Montes and Alto Douro, Vila Real.
28. Silva, N., Igrejas, G., Vaz, J., Araujo, C., Cardoso, L., Rodrigues, J., Torres, C., Poeta, P., 2011, Virulence Factors in Enterococci from Partridges (*Alectoris rufa*) Representing a Food Safety Problem. Foodborne Pathog Dis. 8: 831-833.
29. Jones, R.N., Deshpande, L.M., 2003, Distribution of fsr among *Enterococcus faecalis* isolates from the SENTRY antimicrobial surveillance program. J Clin Microbiol 41, 4004-4005.
30. Radhouani, H., Pinto, L., Coelho, C., Sargo, R., Araujo, C., Lopez, M., Torres, C., Igrejas, G., Poeta, P., 2010a, MLST and a genetic study of antibiotic resistance and virulence factors in *van*A-containing *Enterococcus* from buzzards (*Buteo buteo*). Lett Appl Microbiol 50, 537-541.
31. Park, I.S., Lin, C.H., Walsh, C.T., 1996, Gain of D-alanyl-D-lactate or D-lactyl-D-alanine synthetase activities in three active-site mutants of the *Escherichia coli* D-alanyl-D-alanine ligase B. Biochemistry 35, 10464-10471.
32. Roper, D.I., Huyton, T., Vagin, A., Dodson, G., 2000, The molecular basis of vancomycin resistance in clinically relevant Enterococci: crystal structure of D-alanyl-D-lactate ligase (VanA). Proc Natl Acad Sci U S A 97, 8921-8925.
33. Evers, S., Courvalin, P., 1996, Regulation of VanB-type vancomycin resistance gene expression by the VanS(B)-VanR (B) two-component regulatory system in *Enterococcus faecalis* V583. J Bacteriol 178, 1302-1309.
34. Kak, V., Chow, J.W., 2002, Acquired antibiotic resistances in enterococci. Enterococci: Pathogenesis, Molecular Biology, and Antibiotic Resistance, 355-383.
35. Grobbel, M., Lubke-Becker, A., Alesik, E., Schwarz, S., Wallmann, J., Werckenthin, C., Wieler, L.H., 2007, Antimicrobial susceptibility of *Escherichia coli* from swine, horses, dogs and cats as determined in the BfT-GermVet monitoring program 2004-2006. Berl Munch Tierarztl Wochenschr 120, 391-401.
36. Martinez, J.L., 2009, The role of natural environments in the evolution of resistance traits in pathogenic bacteria. Proc Biol Sci 276, 2521-2530.
37. Mosser, D.D., Morimoto, R.I., 2004, Molecular chaperones and the stress of oncogenesis. Oncogene 23, 2907-2918.
38. Blake, M.J., Fargnoli, J., Gershon, D., Holbrook, N.J., 1991, Concomitant decline in heat-induced hyperthermia and HSP70 mRNA expression in aged rats. Am J Physiol 260, R663 -667.
39. Blake, M.J., Gershon, D., Fargnoli, J., Holbrook, N.J., 1990, Discordant expression of heat shock protein mRNAs in tissues of heat-stressed rats. J Biol Chem 265, 15275-15279.
40. Gauthier, A., Robertson, M.L., Lowden, M., Ibarra, J.A., Puente, J.L., Finlay, B.B., 2005, Transcriptional inhibitor of virulence factors in enteropathogenic *Escherichia coli*. Antimicrob Agents Chemother 49, 4101-4109.
41. Tilly, K., Mckittrick, N., Zylicz, M., Georgopoulos, C., 1983, The Dnak Protein Modulates the Heat-Shock Response of *Escherichia-Coli*. Cell 34, 641-646.
42. Paulsen, I.T., Banerjei, L., Myers, G.S., Nelson, K.E., Seshadri, R., Read, T.D., Fouts, D.E., Eisen, J.A., Gill, S.R., Heidelberg, J.F., Tettelin, H., Dodson, R.J., Umayam, L., Brinkac, L., Beanan, M., Daugherty, S., DeBoy, R.T., Durkin, S., Kolonay, J., Madupu, R., Nelson, W., Vamathevan, J., Tran, B., Upton, J., Hansen, T., Shetty, J., Khouri, H., Utterback, T., Radune, D., Ketchum, K.A., Dougherty, B.A., Fraser, C.M., 2003, Role of mobile DNA in the evolution of vancomycin-resistant *Enterococcus faecalis*. Science 299, 2071-2074.
43. Sparrer, H., Lilie, H., Buchner, J., 1996, Dynamics of the GroEL-protein complex: effects of nucleotides and folding mutants. J Mol Biol 258, 74-87.
44. Pinto, L., Poeta, P., Vieira, S., Caleja, C., Radhouani, H., Carvalho, C., Vieira-Pinto, M., Themudo, P., Torres, C., Vitorino, R., Domingues, P., Igrejas, G., 2010, Genomic and proteomic evaluation of antibiotic resistance in *Salmonella* strains. J Proteomics 73, 1535-1541.
45. Wang, X., He, X., Jiang, Z., Wang, J., Chen, X., Liu, D., Wang, F., Guo, Y., Zhao, J., Liu, F., Huang, L., Yuan, J., 2010, Proteomic analysis of the *Enterococcus faecalis* V583 strain and clinical isolate V309 under vancomycin treatment. J Proteome Res 9, 1772-1785.
46. Hiscock, D., Upton, C., 2000, Viral Genome DataBase: storing and analyzing genes and proteins from complete viral genomes. Bioinformatics 16, 484-485.

JOURNAL OF INTEGRATED OMICS
A METHODOLOGICAL JOURNAL
HTTP://WWW.JIOMICS.COM

ORIGINAL ARTICLE | DOI: 10.5584/jiomics.v2i1.81

# Optimisation of Downscaled Tandem Affinity Purifications to Identify Core Protein Complexes

Eric B. Haura*[1], Roberto Sacco [2], Jiannong Li[1], André C. Müller[2], Florian Grebien[2], Giulio Superti-Furga[2], Keiryn L. Bennett*[2].

[1] Department of Thoracic Oncology Program, H. Lee Moffitt Cancer Center and Research Institute, Tampa, Florida, USA; [2] CeMM Research Center for Molecular Medicine of the Austrian Academy of Sciences, Vienna, Austria.

**Received:** 01 December 2011 **Accepted:** 11 April 2012 **Available Online**: 18 April 2012

ABSTRACT

In this study we show that via stable, retroviral-expression of tagged EGFR del (L747-S752 deletion mutant) in the PC9 lung cancer cell line and stable doxycycline-inducible expression of tagged Grb2 using a Flp-mediated recombination HEK293 cell system, the SH-TAP can be downscaled to 5 to 12.5 mg total protein input (equivalent to 0.5 - 1 × 15 cm culture plate or 4 - 8 × $10^6$ cells). The major constituents of the EGFR del complex (USB3B, GRB2, ERRFI, HSP7C, GRP78, HSP71) and the Grb2 complex (ARHG5, SOS1, ARG35, CBL, CBLB, PTPRA, SOS2, DYN2, WIPF2, IRS4) were identified. Adjustment of the quantity of digested protein injected into the mass spectrometer reveals that optimisation is required as high quantities of material led to a decrease in protein sequence coverage and the loss of some interacting proteins. This investigation should aid other researchers in performing tandem affinity purifications in general, and in particular, from low quantities of input material.

**Keywords:** EGFR; Grb2; Orbitrap; TAP; downscale.

## Abbreviations

**1D-SG-CID,** one-dimensional shotgun collision-induced dissociation; **AGC,** automatic gain control; **BPI**, base peak intensity; **DMEM**, Dulbecco's modified eagle medium; **cDNA**, complementary deoxyribonucleic acid; **EGFR**, epidermal growth factor receptor; **ESI-MS,** electrospray ionisation mass spectrometry; **FBS**, foetal bovine serum; **GFP**, green fluorescent protein; **LC-MS**, liquid chromatography mass spectrometry; **PBS**, phosphate-buffered saline; **PCR**, polymerase chain reaction. **RIC**, reconstructed ion chromatogram; **PCT**, unique peptide counts; **SCT**, spectral counts. **SCV**, percent sequence coverage; **SH-TAP,** streptavidin-binding peptide haemagglutinin tandem affinity purificationl.; **TAP-MS**, tandem affinity purification mass spectrometry; **VSV-G**, vesicular stomatitis virus G.

## 1. Introduction

Tandem affinity purification (TAP) is a generic two-step purification protocol that enables the enrichment and isolation of non-covalent protein complexes under near-physiological conditions. Coupling of TAP to mass spectrometry (TAP-MS) has proven to be a powerful approach to generate large-scale protein-protein interaction networks in yeast [1-4] and mammalian [5, 6] systems. Although TAP-MS in yeast was relatively straightforward, similar studies were hampered in human cells by: (i) the labour-intensive generation of large collections of cell lines expressing epitope -tagged bait proteins; (ii) the low yield of the protein complexes isolated from the cell lines; and (iii) the limited sensitivity of MS-based protein identification. The recent introduction of the Flp-mediated recombination system in HEK-

***Corresponding authors:** *Keiryn L. Bennett, Ph.D.* CeMM Research Center for Molecular Medicine of the Austrian Academy of Sciences. Lazarettgasse 14, AKH, BT 25.3. Vienna, 1090, Austria. Phone: +43-1-40160-70010. Fax: +43-1-40160-970000. Email Address: mail: kbennett@cemm.oeaw.ac.at; *Eric B. Haura, M.D.* Department of Thoracic Oncology Program. H. Lee Moffitt Cancer Center and Research Institute. 12902 Magnolia Drive. Tampa, FL, 33612, USA. Phone: +1-(0)813-903-6287. Fax: +1-(0)813-903-6817. Email Address: eric.haura@moffitt.org.

293 cells combined with the streptavidin-binding peptide haemagglutinin (SH)-tag and gel-free one-dimensional liquid chromatography mass spectrometry (LC-MS) analysis of the purified protein complexes [7] overcame most of the shortcomings proffered by the traditional [8] and improved TAP-tag protocols [9].

To yield sufficient material for multiple LC-MS experiments, the current recommendation is to perform an SH-TAP from approximately $3 \times 10^7$ HEK293 Flp-In cells (around 50 mg total protein from 5 × 15 cm culture plates) [7]. Successful identification of abundant protein complexes (*e.g.*, the PPP2R2B complex) has been achieved from $4 \times 10^6$ HEK293 cells (25% eluate volume per LC-MS injection) [7]. Although considerably lower quantities of cells are required compared to the previous recommendations of $5 \times 10^8 - 1 \times 10^9$ cells [10, 11] and $5 \times 10^7$ cells [9], the design of certain experiments and choice of bait protein (*e.g.*, innate immune sensors requiring cell lines that express the necessary modulating proteins) may prohibit the use of HEK293 Flp-In cell lines. Additionally, the Flp-In system is neither amenable to characterisation of protein complexes from cells that are difficult to cultivate (neuronal, immune) nor from primary cells.

Diseases such as cancer are increasingly recognised as the result of aberrantly-functioning protein networks that drive cells toward a diseased state. Affinity purification and mass spectrometry has the ability to map such disease networks and enable the realisation of so-called 'network medicine' [12-14]. Using this methodology, we are currently undertaking a systematic approach to map the protein-protein interaction network driven by somatic and activating mutations of the epidermal growth factor receptor (EGFR) in lung cancer. Activated EGFR can drive numerous downstream signalling pathways responsible for cancer growth and survival [15]. PC9 cells are widely-used as a model system to study EGFR mutations in lung cancer. These cells are robust and easily-cultured, however, the generation of large amounts of cells is associated with a number of costs and increased time requirements. The use of high quantities of cellular material may be restrictive when specialised cells are required. Such cells may take longer periods to grow in culture and/or require specific media and/or growth factor/cytokine supplementation.

As EGFR biology has been extensively investigated, there is a wealth of information available to assist in supporting experimentally-derived data. Thus, the protein is ideal as a case study to evaluate a downscaling of the SH-TAP methodology and examine the lower limit of cells from which the core complex can still be successfully purified and identified by mass spectrometry. C-terminally-tagged SH-EGFR del (L747-S752 deletion mutant) [16] was expressed in PC9 lung cancer cells via retroviral transduction, and the EGFR del core complex enriched via a modified two-step affinity purification procedure [17] (Figure 1). The proteins from the complex were digested with trypsin and the resultant peptides identified by LC-MS. Different percentages of the eluate volume were analysed by LC-MS to ascertain the outcome on the number and nature of the identified EGFR-interacting proteins. Ultimately, this information can assist researchers performing generic TAP experiments from low quantities of protein input material and in particular, TAP from more difficult cell types.

A common misconception with mass spectrometric analyses of biological samples is that the higher the quantity of digested material injected onto an LC-MS system, the more peptides (and thus proteins) will be identified. This is not necessarily the case and there are numerous publications and reviews that support this [18-20] and references therein. Increasing the quantity of injected material does not always provide more information. There are a number of important factors that can influence peptide identification including: (i) ion suppression and reduced ionisation efficiency [21]; (ii) resolving capacity of the chromatographic system [22]; (iii) composition of HPLC mobile phase system [23]; (iv) loading capacity of the pre-column; (v) mass spectrometer automatic gain control (AGC) settings; and (vi) TAP purification background (protein and other molecular groups). These points are discussed in the context of downscaled tandem affinity

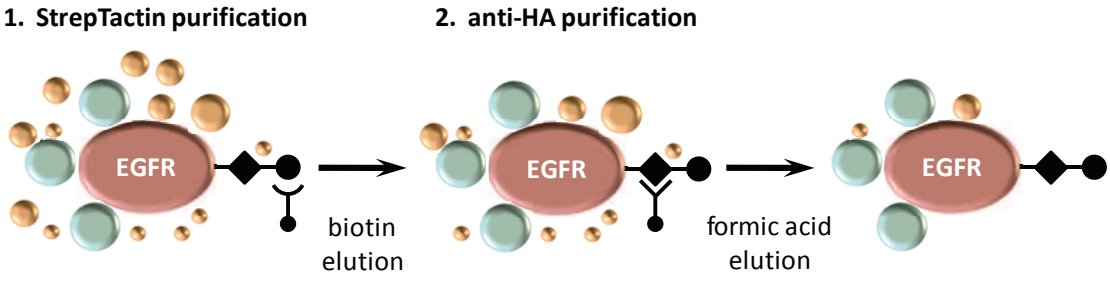


**Figure 1.** Schematic representation of the tandem affinity purification (TAP) procedure (adapted from Glatter *et al.* [7]). PC9 or HEK293 cells expressing SH-tagged EGFR del or Grb2, respectively, were lysed and purified from total protein extracts using streptavidin sepharose (streptactin beads). After several wash steps, tagged proteins were eluted with 2.5 mM D-biotin for subsequent immunoaffinity purification using anti-HA agarose. After further wash steps, protein complexes were eluted with 100 mM formic acid (pH 2.5) and processed for mass spectrometric analysis.

purification coupled to LC-MS.

The aims of this study were to determine the: (i) optimal quantity of eluate to inject onto the LC-MS system to maximise the number of peptides and proteins identified from the SH-EGFR del purified complex; and (ii) lowest amount of starting input protein required to perform a successful tandem affinity purification. Success is defined as identifying the relevant interacting proteins, as observed from a large-scale purification conducted in our laboratory combined with information from the literature. To confirm that our approach was not restricted to TAP of SH-EGFR complexes in PC9 cells, the study was extended to a second core complex (Grb2) using an alternative system (doxycycline-inducible Flp-In HEK293 cells). The ultimate future prospective of the optimised downscale tandem affinity purification methodology is to discover novel binding partners in specialised cell types. By assessing two well-known and characterised protein complexes at low levels of starting input material, we demonstrate that the discovery of new interactors is attainable in systems where the quantity of available material is limited.

## 2. Material and Methods

### 2.1 *Materials*

Iodoacetamide, dithiothreitol, 1 M triethylammonium bicarbonate (TEAB), protease inhibitor cocktail, anti-HA agarose, polybrene, doxycycline, HA7 antibody, 2-propanol, LC-MS grade (SIGMA-Aldrich, St. Louis, MO); trypsin (Promega Corp., Madison, WI); formic acid (HCOOH) (MERCK, Darmstadt, Germany); Strep-Tactin sepharose (IBA TAGnologies, Göttingen, Germany); D-biotin (Alfa Aesar, Karlsruhe, Germany); micro Bio-Spin chromatography columns (Bio-Rad, Hercules, CA); Gateway LR Clonase™ II Enzyme Mix Kit, foetal bovine serum, lipofectamine 2000, pDONR201, pcDNA6/TR, pOG44 (Invitrogen, Carlsbad, CA); methanol, LC-MS grade (Fisher Scientific, Schwerte, Germany).

### 2.2 *Generation of Stable Cell Lines: Retroviral Infection*

cDNA for EGFR del (L747-S752 deletion mutant) [16] was provided by Dr. William Pao (Vanderbilt University, Nashville, TN). Design of the PCR primers, amplification and pENTR™ TOPO® cloning of EGFR del was performed as previously described [24]. EGFR del was inserted into a gateway-compatible version of pMSCV-C-SH IRES GFP from pENTR™ TOPO® vector by Gateway LR Clonase™ II Enzyme Mix Kit. The retroviral expression clone was verified by DNA sequencing using an Applied Biosystems 3130X1 Genetic analyser (HITACHI) with data analysis performed using Lasergene software V7.2. Phoenix HEK293 cells were obtained from ATCC (catalogue № SD3514, Manassas, VA) and grown in DMEM (4.5 g/L glucose, 2 mM L-glutamine, 50 mg/mL penicillin, 50 mg/mL streptomycin) containing 10% FBS. On day one, $8 \times 10^5$ Phoenix cells per well were seeded in a 6-well plate. On day two, cells were transfected with 3 μg VSV-G and 5 μg retroviral plasmids using lipofectamine 2000. Six hours after transfection, the supernatant was replaced with 2 mL DMEM containing 20% FBS and the cells incubated in a 5% $CO_2$ incubator at 32°C for 48 h. The supernatant (viruses) was collected by centrifugation at 4°C; either stored at -80°C or used immediately to infect the target cells. PC9 cells were maintained in RPMI-1640 medium supplemented with 10% FBS. For retroviral transduction, $2 \times 10^5$ cells per well were seeded in a 6-well plate. After overnight incubation, cells were infected with 800 μL virus supernatant plus 6 μg/mL polybrene for 24 h and then supplemented with 4 mL of media per well. Cells were grown continuously until cell sorting. One week after infection, GFP-positive PC9 cells were sorted using a FACSVantage (BD Biosciences, San Diego, CA). GFP positivity and HA expression were assessed by flow cytometry and immunoblot, respectively, before expanding the cells to 10 × 15 cm dishes. When approximately 90% confluent, the EGFR del-tagged PC9 cells were washed with ice-cold PBS containing 1 mM sodium orthovanadate and scraped with a cell lifter on ice. Two cell pellets (each consisting of 5 × 15 cm dishes) were collected in 15 mL conical tubes by centrifugation at 129 × *g* at 4°C and stored at -80°C until required.

### 2.3 *Generation of Stable Cell Lines: Doxycycline-inducible Expression in HEK293 Flp-In Cells*

To obtain an entry vector, the human full length Grb2 cDNA was cloned by BP Clonase™ recombination into the Gateway®-compatible vector pDONR201. An LR Clonase™ recombination was performed between the entry and destination vector pcDNA5/FRT/TO/SH/GW [7] to generate an expression vector for tetracycline-controlled expression of an N-terminally SH-tagged version of human Grb2. To generate a stable cell line expressing the SH-tagged version of Grb2 in an inducible manner, a Flp-In T-Rex HEK293 cell line stably-transfected with the pcDNA6/TR regulatory vector was cultured in DMEM (containing 10% FBS) supplemented with 100 μg/mL Zeocin™ and 15 μg/mL blasticidin and co-transfected with the expression vector and the pOG44 vector using the FuGENE transfection reagent. Two days after transfection, cells were selected in DMEM supplemented with hygromycin (100 μg/mL) for 2-3 weeks and positive clones were pooled and amplified. After induction with 1 μg/mL doxycycline for 24 h, the inducible expression of the construct was verified by immunoblot.

### 2.4 *SH-tandem-affinity Purification (adapted from Glatter et al.) [7]*

To minimise experimental and daily variation, *e.g.*, different lot numbers of the streptactin and anti-HA beads, different stock solutions of buffers *etc.*; the 50 mg full pulldowns (n = 2 biochemical replicates) were performed simultaneous-

ly. Note that a biochemical replicate refers to a separate affinity-purification experiment that was performed from the same cell lysate. The series of half pulldowns (20, 12.5, 5 and 2.5 mg) were also performed together (n = 2 biochemical replicates, total 8 purifications) under identical experimental conditions. EGFR del-expressing PC9 cells or Grb2-expressing HEK293 cells were lysed in TNN-HS buffer (50 mM HEPES pH 8.0, 150 mM NaCl, 5 mM EDTA, 0.5% NP-40, 50 mM NaF, 1.5 mM $Na_3VO_4$, 1.0 mM PMSF and protease inhibitor cocktail). Insoluble material was removed by centrifugation at 39,443 × *g* for 15 min at 4°C. 200 µL streptactin sepharose (400 µL slurry/pulldown) was transferred to a 14 mL dust-free Falcon tube and washed with 2 × 1 mL TNN-HS buffer. The lysates (50, 20, 12.5, 5 or 2.5 mg) were added to the washed streptactin sepharose and rotated for 20 min at 4°C. The sepharose beads and supernatant were transferred to a micro Bio-Spin chromatography column (Bio-Rad, Hercules, CA) and gravity drained. The sepharose was washed with 4 × 1 mL TNN-HS buffer, and bound proteins eluted from the sepharose with 3 × 300 µL freshly-prepared 2.5 mM D-biotin in TNN-HS buffer into a fresh dust-free 1.5 mL eppendorf tube. 100 µL anti-HA agarose beads (200 µL slurry/pulldown) were transferred into a 1.5 mL eppendorf tube, washed with 1 × 1 mL TNN-HS buffer, centrifuged at 200 × *g* for 1 min at 4°C. The supernatant was removed and the agarose resuspended in 100 µL TNN-HS buffer. The anti-HA agarose beads were added to the biotin eluate and rotated for 1 h at 4°C. The samples were centrifuged at 200 × *g* for 1 min at 4°C and the supernatant removed. The agarose beads were resuspended in 1 mL TNN-HS buffer, the washed beads and buffer loaded into a fresh dust-free micro Bio-Spin column and gravity drained. The anti-HA agarose was washed with 3 × 1 mL TNN-HS buffer and then with 2 × 1 mL TNN-HS buffer consisting of only HEPES, NaCl and EDTA. Retained proteins were eluted from the column with 500 µL 100 mM HCOOH directly into a glass HPLC vial and immediately neutralised with 125 µL 1 M TEABC [17, 25]. Except for the Grb2 half pulldowns, 200 µL (*i.e.*, 33% of the final eluate volume) were routinely removed for silver-stain one-dimensional gel electrophoresis and/or immunoblot analysis as required. The remaining sample was frozen at -20°C until further processing. All volumes in the protocol were halved for the half pulldown series of experiments.

### 2.5 *Solution Tryptic Digestion*

Proteins from 66% (all EGFR experiments and Grb2 full pulldown) or 100% (Grb2 half pulldowns) of the TEAB neutralised acid eluate were reduced with dithiothreitol, alkylated with iodoacetamide and digested with 2 µg (full pulldown) or 1 µg (half pulldown) modified porcine trypsin. Multiples of 0.25 to 50% of the total eluate volume were desalted and concentrated with customised reversed-phase stage tips [26]. The volume of the eluted sample was reduced to approximately 2 µL in a vacuum centrifuge and reconstituted to 8 µL with 5% formic acid in water and multiples thereof.

### 2.6 *Anti-haemagglutinin Immunoblot*

The samples were denatured in 4× reducing SDS and boiled for 5 min, separated by 1D-SDS-PAGE and transferred onto a nitrocellulose membrane. The membrane was incubated in blocking solution (5% non-fat milk in PBS/0.05% Tween-20) for 1 h at room temperature, then with murine anti-HA-peroxidase antibody (diluted 1:5000 in blocking solution) for 1 h at room temperature. After thorough washing with PBS/0.05% Tween-20, the antibody signal was visualised by ECL™ Western Blotting Detection Reagents.

### 2.7 *Liquid Chromatography Mass Spectrometry*

Mass spectrometry was performed on a hybrid LTQ-Orbitrap XL mass spectrometer (ThermoFisher Scientific, Waltham, MA) using the Xcalibur version 2.0.7 coupled to an Agilent 1200 HPLC nanoflow system (dual pump system with one precolumn and one analytical column) (Agilent Biotechnologies, Palo Alto, CA) via a nanoelectrospray ion source using liquid junction (Proxeon, Odense, Denmark). Solvents for LC-MS separation of the digested samples were as follows: solvent A consisted of 0.4% formic acid in water and solvent B consisted of 0.4% formic acid in 70% methanol and 20% 2-propanol. From a thermostatted microautosampler, 8 µL of the tryptic peptide mixture were automatically loaded onto a trap column (Zorbax 300SB-C18 5µm, 5×0.3 mm, Agilent Biotechnologies, Palo Alto, CA) with a binary pump at a flow rate of 45 µL/min. 0.1% TFA was used for loading and washing the pre-column. After washing, the peptides were eluted by back-flushing onto a 16 cm fused silica analytical column with an inner diameter of 50 µm packed with C18 reversed phase material (ReproSil-Pur 120 C18-AQ, 3 µm, Dr. Maisch GmbH, Ammerbuch-Entringen, Germany). The peptides were eluted from the analytical column with a 27 minute gradient ranging from 3 to 30 percent solvent B, followed by a 25 minute gradient from 30 to 70 percent solvent B and, finally, a 7 minute gradient from 70 to 100 percent solvent B at a constant flow rate of 100 nL/min. The analyses were performed in a data-dependent acquisition mode using a top 6 collision-induced dissociation (CID) method. Dynamic exclusion for selected ions was 60 seconds. No lock masses were employed. Maximal ion accumulation time allowed on the LTQ Orbitrap in CID mode was 150 ms for $MS^n$ in the LTQ and 1,000 ms in the C-trap. Automatic gain control was used to prevent overfilling of the ion traps and were set to 5,000 (CID) in $MS^n$ mode for the LTQ, and $10^6$ ions for a full FTMS scan. Preview mode for FTMS master scans was enabled to determine the charge state of the intact peptides prior to a scan at 100,000 resolution. Monoisotopic precursor ion selection was enabled and unassigned charge states were not selected for fragmenta-

tion.

2.8 *Data Analysis*

The acquired data were processed with Bioworks V3.3.1 SP1 (ThermoFisher, Manchester, UK). The .dta files were extracted from the .raw files with the extract_msn.exe program. All .dta files were merged into a single peak list file (.mgf) with an internally-developed Perl script. The merged peak list was searched against the human SwissProt database version v57.4 (34,579 sequences, including isoforms as obtained from varsplic.pl) with the search engines MASCOT (v2.2.03, MatrixScience, London, UK) and Phenyx (v2.5.14, GeneBio, Geneva, Switzerland) [27]. Submission to the search engines was via a Perl script that performs an initial search with relatively broad mass tolerances (MASCOT only) on both the precursor and fragment ions ( ±10 ppm and ±0.6 Da, respectively). High-confidence peptide identifications are used to recalibrate all precursor and fragment ion masses prior to a second search with narrower mass tolerances (±4 ppm and ±0.3 Da). One missed tryptic cleavage site was allowed. Carbamidomethyl cysteine was set as a fixed modification, and oxidised methionine was set as a variable modification. To validate the proteins, MASCOT and Phenyx output files were processed by internally-developed parsers. For MASCOT, two unique peptides with an ion score >18 (plus additional peptides from proteins fulfilling the criteria with an ion score >10) are required. For Phenyx, two unique peptides with a z-score >4.5 and a P-value <0.001 are required (plus additional peptides from proteins fulfilling the criteria with a z-score >3.5 and a P-value <0.001). The validated proteins retrieved by the two algorithms are merged, any spectral conflicts discarded and grouped according to shared peptides. A false positive detection rate (FDR) of <0.25 % and <0.1 % (including the peptides with lower scores) was determined for proteins and peptides, respectively, by applying the same procedure against a reversed database. Comparisons between analytical methods involved comparisons between the corresponding sets of identified proteins. This was achieved by an internally -developed program that simultaneously computes the protein groups in all samples and extracts statistical data such as the number of distinct peptides (PCT), number of spectra (SCT), and sequence coverage (SCV).

## 3. Results and Discussion

The results reported in this study are divided into three areas. Firstly, using typical starting quantities of cellular lysate, the quantity of eluate required for injection onto the LC -MS system was adjusted to maximise the number of peptides and proteins identified for SH-tagged EGFR del. Secondly, the initial input material was reduced and also the injection quantities varied to determine the minimal protein input required to identify all core components of the EGFR del complex. Finally, the entire strategy was assessed by evaluating another bait in a different cell line to both confirm and generalise our findings.

3.1 *Optimisation of Injection Quantity*

The standard protocol established in our laboratory for the mass spectrometric analysis of streptavidin-binding peptide haemagglutinin (SH)-tagged proteins by one-dimensional gel-free or 'shotgun' collision-induced dissociation (1D-SG-CID), is to inject 3% [17] of the total eluate volume from tryptically-digested, acid-eluted and neutralised protein complexes (50 mg protein input). N-terminally-tagged green fluorescent protein (GFP) and C-terminally-tagged EGFR del were retrovirally-expressed in PC9 cells and analysed accordingly. The protein groups identified from the GFP negative control (TBA1B, TBB5, TBB2C, TBA4B, ADT2, ADT3, ACTB, ACTBL, RS27) were considered as non-specific proteins that were interacting with the SH-tag and/or GFP and were thus removed from the results obtained for EGFR del. The filtered data consisted of 8 proteins (n = 4). EGFR (bait) was the most abundant protein with a maximum sequence coverage (SCV) of 50% (52 unique peptide counts, PCT, from 936 spectral counts, SCT). The proteins identified as part of the core complex of EGFR were: UBS3B, GRB2, ERRFI, HS90A, HS90B, SHC1, and CDC37. A summary of the PCT, SCT and SCV are given in Table 1. UBS3B, GRB2, ERRF1, HSP90, and SHC1 are all *bona fide* interacting proteins with EGFR (www.hprd.org) while CDC37 is a co-chaperone of kinases along with HSP90 [28-30]. GRP78, HSP7C, HSP71 and HS71L were identified in both the EGFR del and GFP experiments, however, the spectral counts were markedly increased in the former thereby indicating that the interaction of EGFR del with some heat shock proteins is enriched and not entirely due to non-specific binding.

**Table 1.** The core complex proteins identified from the 50 mg protein input SH-EGFR del-TAP (n = 4) following subtraction of proteins identified in the green fluorescent protein (GFP) negative control. PCT (peptide counts); SCT (spectral counts); SCV (percent sequence coverage).

| ACCESSION | SWISSPROT | PCT | SCT | SCV |
|---|---|---|---|---|
| P00533-1 | EGFR | 52 | 936 | 50 |
| Q8TF42 | UBS3B | 14 | 77 | 29 |
| P62993-1 | GRB2 | 8 | 66 | 46 |
| Q9UJM3 | ERRFI | 6 | 20 | 20 |
| P07900-1 | HS90A | 5 | 19 | 8 |
| P08238 | HS90B | 3 | 8 | 5 |
| P29353-1 | SHC1 | 3 | 5 | 7 |
| Q16543 | CDC37 | 2 | 3 | 7 |

Shown in Figure 2 is the effect on maximum unique PCT when the amount of injected material was increased from 3% of the eluate volume (n = 4) to 5% (n = 4), 10% (n = 4) and 20% (n = 2). The general tendency for the major proteins was a gradual increase in maximum PCT from 3 to 5% eluate injected. UBIQ was identified when a higher quantity of material was analysed, however, CDC37 disappeared with increasing quantity of injected sample. From 5% to 10% to 20%, the tendency was for the maximum PCT to gradually decrease. When 20% of the eluate was injected, however, an additional protein (H90B4) with low PCT was identified. Only two unique peptides were observed for this protein: VILHLKEDQTEYLEER, (shared with HS90A); and IMEESNVK (specific to H90B4). It would appear that although the PCT for all the identified proteins decreased at 20% of the eluate injected, sufficient intensity for the specific peptide from H90B4 was reached to be selected and fragmented in the mass spectrometer.

Unique peptide counts (PCT) and sequence coverage (SCV) (see below, section 3.2) were chosen in this study as the metric for monitoring the effect of increasing the quantity of digested protein that is injected onto the LC-MS system. Contrary to SCT [31], PCT and SCV are by no means quantitative in nature. Nonetheless, both PCT and SCV demonstrated a phenomena that was not observed with SCT. Namely, PCT and SCV have a tendency to decrease with increasing quantities of injected sample, whereas SCT continues to increase. A example of the effect on SCT with increasing quantities of injected material is included in the supplementary section (Supplementary Figure S2).

It has long been established that ESI-MS is a highly-sensitive detection method for polar molecules, however, at elevated concentrations (>$10^{-5}$ M), the approximate linearity of the ESI response is often lost and ion suppression becomes an increasing problem [21]. Shown in Figure 3 is the reconstructed ion chromatogram (RIC) for one of the two unique peptides (LQAEAQQLR, theoretical $[M+2H]^{2+}$ = 528.7935) that were identified from CDC37 across four LC-MS analyses (biochemical replicate 1: 3, 5, 10 and 20% eluate injected). Inset is the *m/z* and isotopic pattern of the ion. The peptide is present in the MS spectrum across all analyses, however, the MSMS spectrum was only matched to the peptide (Figure 3C) from the technical replicate with 3% of the eluate injected (Figure 3B). In all the other analyses, no MSMS spectrum was recorded. The general trend was that the maximal ion intensity of the $^{12}C$ isotope for LQAEAQQLR actually decreased from 3% to 20% of the eluate injected. More notable is that the isotopic pattern of the ion was also compromised with increasing quantities of injected material. For the spectra shown in Figure 3A and 3D-H, the ion representative of the peptide containing two $^{13}C$ atoms is not evident. As a consequence, the mass spectrometer did not switch to MSMS. Conversely, the population of ions representative of the peptide with two $^{13}C$ atoms is only evident in Figure 3B and thus for this analysis only, an MSMS spectrum was recorded (Figure 3C). Thus, it fol-

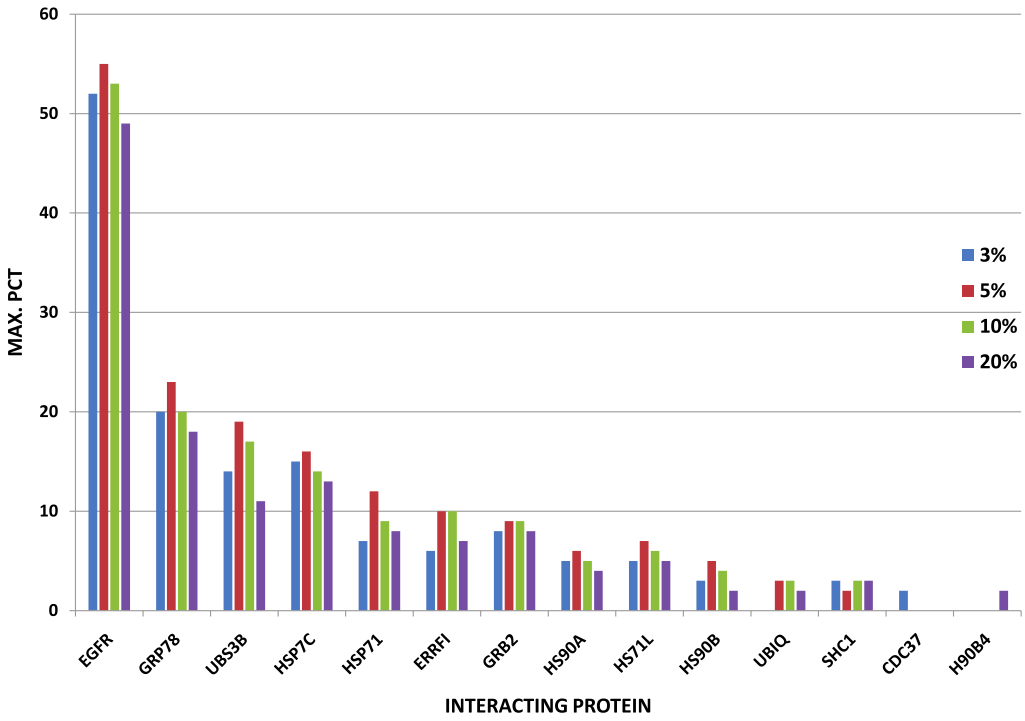


**Figure 2.** The effect on maximum unique PCT for 50 mg protein input with increasing quantities of injected digested SH-EGFR del-TAP eluate onto the LC-MS system. Blue (3%), red (5%), green (10%) and purple (20%). For the injected quantities of 3, 5 and 10%, n = 4 and for 20%, n = 2.

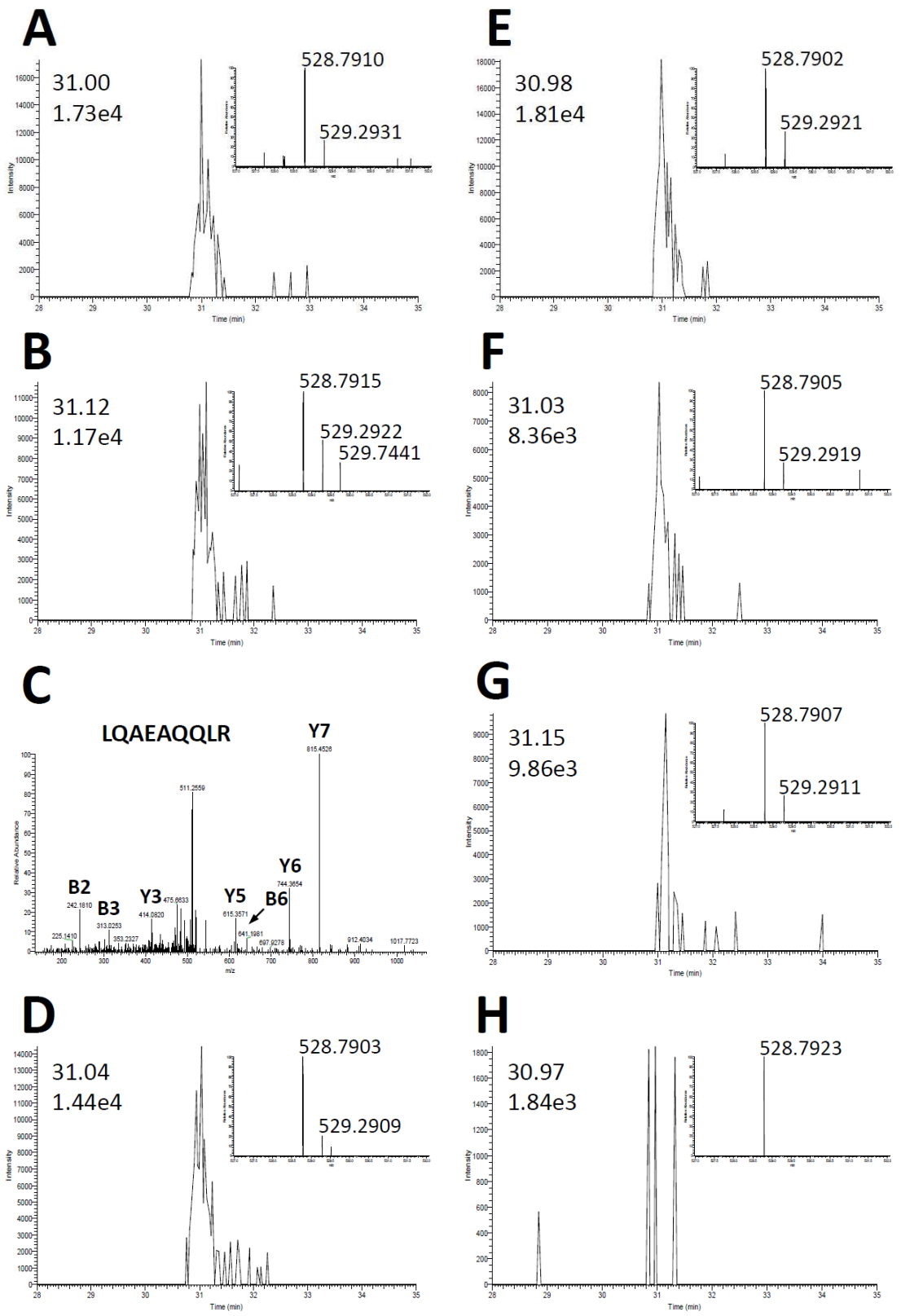


**Figure 3.** Reconstructed ion chromatogram (RIC) for the peptide LQAEAQQLR (theoretical $[M+2H]^{2+}$ = 528.7935) from CDC37. (A) and (B) 3%; (C) MSMS spectrum from (B); (D) and (E) 5%; (F) and (G) 10% and (H) 20% eluate volume injected. The observed retention time ($T_r$); and maximum ion intensity are given for each eluate injection percentage. Shown inset is the isotopic distribution and observed *m/z* for the peptide ion.

lows that at a lower concentration of analyte, the apparent ion suppression effect on CDC37 is reduced and the ions are available for selection in the mass spectrometer for fragmentation.

The resolving capacity of the chromatographic system is another factor that can influence the identification of peptides by mass spectrometry. The overall separation efficiency of an analytical column depends on such factors as column length, chosen solvent system, flow rate, gradient profile and type of packing material. Loss of resolution between consecutive chromatographic analyses can be caused by overloading effects or degradation of the reversed-phase material. This can be monitored by changes in peak shape, peak width and altered retention times. The effect of increasing quantities of protein injected on the chromatography is given in Figure 4. Overall the base peak intensity (BPI) chromatograms are relatively weak and the predominant peaks (T) are due to autodigestion products of trypsin. Shown inset is the reconstructed ion chromatogram (RIC) for one of the most intense doubly-charged peptide (NLQEILHGAVR, theoretical $[M+2H]^{2+}$ = 625.3542) from EGFR del (normalised to the maximal intensity of the ion across the four analyses). The peak full-width-at-half-maximum-height (FWHM) was 0.49, 0.42, 0.48 and 0.42 min for 3, 5, 10 and 20% eluate volume injected, respectively. Throughout all analyses in this study, a Zorbax 300SB-C18 5 µm, 5×0.3 mm precolumn was used (see Materials and Methods). As stated by the manufacturer, the loading capacity of this column is in the range of 1 to 10 µg. In chromatography, column overloading effects are observed if there is an increase in peak width of greater than 10%. Since this is not the case, the precolumn capacity is sufficient. The peak profile is unaffected by the increasing amounts of injected material and chromatographic resolution and separation are not compromised. Thus, the quantity of sample injected onto the precolumn does not appear to exceed the specifications.

The composition of HPLC mobile phase system can also play an important role in peptide identification. It is stand-

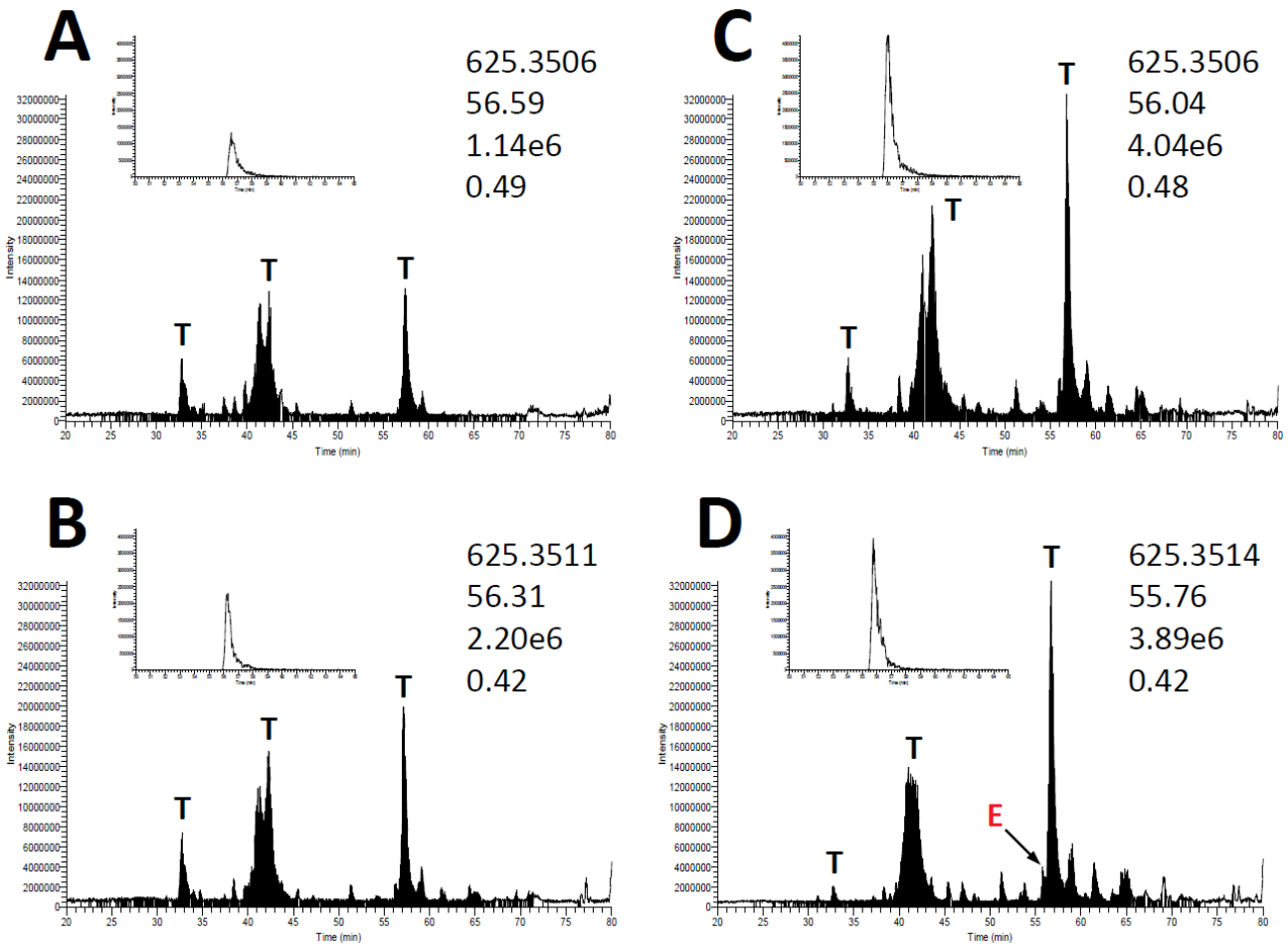


**Figure 4.** Chromatographic separation of the tryptic digest of the proteins from an SH-EGFR del TAP (50 mg protein input). Shown inset is the reconstructed ion chromatogram (RIC) for NLQEILHGAVR (theoretical $[M+2H]^{2+}$ = 625.3542) from EGFR (normalised to the maximal intensity of the ion across the four analyses). (A) 3; (B) 5; (C) 10 and (D) 20% eluate volume injected. The observed $[M+2H]^{2+}$ retention time ($T_r$); maximum ion intensity; and peak full-width-at-half-maximum-height (FWHM) are given for each eluate injection percentage. (T), autodigestion products of trypsin; (E) extracted EGFR peptide ion.

ard practice in our laboratory to use a non-traditional mobile phase B containing 70% methanol, 20% 2-propanol (see Materials and Methods). A systematic evaluation of alternative solvent systems compared to the more commonly-used acetonitrile has been performed by our group (unpublished data); however, the data is not included here as this is outside the scope of this manuscript. Nonetheless, in our hands we have found that a 70% methanol, 20% 2-propanol mixture provides superior data compared to other mobile phase combinations (including acetonitrile). Interestingly enough, it is known from the literature that the use of methanol in the mobile phase B actually augments and improves the limits of peptide detection compared to acetonitrile [23].

The AGC settings on an Orbitrap mass spectrometer are used to prevent overfilling of the ion traps. As indicated in the Material and Methods, the AGC was set to 5,000 (CID) in $MS^n$ mode for the LTQ, and $10^6$ ions for a full FTMS scan. Naturally, it would have been possible to increase the AGC settings to compensate for the lower quantities of injected material; however, the negative implication is that it takes longer to accumulate more ions. Any gain in adjusting the AGC are counteracted by longer filling times. To maintain consistency during the analyses of the different input quantities of starting material and allow direct comparison between the data sets, the choice was made to limit the number of variables and as such the same AGC settings were maintained throughout the analyses.

The TAP purification background (protein and other molecular groups) can also play a role in the observed decrease in identified peptides and proteins. For the non-specific proteins (TBA1B, TBB5, TBB2C, TBA4B, ADT2, ADT3, ACTB, ACTBL, RS27), the following observations were made (see Table 2). From 3% to 20% of the eluate injected (n = 4) RS27 was identified at the lower injection quantities but was not apparent when 20% of the eluate was analysed. TBB5 and ADT2 were identified across all injection quantities, however, the PCT decreased at 20% injected. ADT3 was only identified when 3% of the eluate was injected. At the higher injection quantities, no unique peptides were identified and the protein was grouped with ADT2. TBB2C, TBA4B, ACTBL and ACTB were only observed at 10% or higher injection percentages. TBA1B was the only non-specific protein that showed a gradual increase in PCT as the amount of eluate injected increased. Overall, the non-specific protein contaminants identified in the SH-EGFR del-TAP exhibited similar traits to the core complex proteins. See Supplementary Table S1 for an overview of the PCT, SCT and SCV for all proteins identified at the different eluate injection volumes. In addition to background non-specifically-interacting proteins, residual non-volatile substances (*e.g.*, salts, buffers, chaotropic agents, stabilisers and detergents) may play a role in the observed phenomena. For example, NP-40 is a pre-requisite component of the cell lysis buffer. The detergent is necessary to not only solubilise membrane and membrane-associated proteins, but also to maintain the structural integrity of the protein-protein complexes prior to

**Table 2.** The non-specific interacting proteins identified from the SH-EGFR del-TAP (50 mg protein input) with increasing quantities of injected eluate onto the LC-MS system. 3, 5 and 10%, n = 4 and 20%, n = 2. PCT (peptide counts); SCT (spectral counts); SCV (percent sequence coverage).

| ACCESSION CODE | SWISSPROT IDENTIFIER | 3% | | 5% | | 10% | | 20% | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | PCT | | SCT | | SCV | | PCT | | SCT | | SCV | | PCT | | SCT | | SCV |
| P68363 | TBA1B | 6 | | 18 | | 23 | | 6 | | 24 | | 20 | | 8 | | 37 | | 26 | | 8 | | 17 | | 27 | |
| P07437 | TBB5 | 3 | | 4 | | 9 | | 7 | | 23 | | 19 | | 7 | | 32 | | 21 | | 6 | | 21 | | 19 | |
| P68371 | TBB2C | 0 | | 0 | | 0 | | 0 | | 0 | | 0 | | 6 | | 28 | | 18 | | 5 | | 15 | | 16 | |
| Q9H853 | TBA4B | 0 | | 0 | | 0 | | 0 | | 0 | | 0 | | 0 | | 0 | | 0 | | 2 | | 3 | | 4 | |
| P05141 | ADT2 | 3 | | 16 | | 12 | | 4 | | 22 | | 15 | | 4 | | 37 | | 15 | | 5 | | 23 | | 19 | |
| P12236 | ADT3 | 3 | | 9 | | 12 | | 0 | | 0 | | 0 | | 0 | | 0 | | 0 | | 0 | | 0 | | 0 | |
| P60709 | ACTB | 0 | | 0 | | 0 | | 2 | | 3 | | 9 | | 2 | | 6 | | 8 | | 2 | | 2 | | 3 | |
| Q562R1 | ACTBL | 0 | | 0 | | 0 | | 0 | | 0 | | 0 | | 0 | | 0 | | 0 | | 2 | | 2 | | 9 | |
| P42677 | RS27 | 2 | | 7 | | 23 | | 2 | | 10 | | 23 | | 2 | | 5 | | 23 | | 0 | | 0 | | 0 | |

enrichment via TAP. Although the methodology is well-established in our laboratory and particular care is taken with the washing steps to remove detergent (and other non-volatile material) prior to elution with acid, it is feasible that residual detergent (and other components) remain and are concentrated as more of the eluate is injected onto the LC-MS. The negative repercussions are that a very small proportion of the detergent can also lead to reduced peptide intensity as detergents are a scavenger of charge [32]. Additionally, detergents also bind strongly to reversed-phase material which ultimately compromises peak shape and column performance.

When undertaking a TAP-MS experiment, it is important to bear in mind that there are a number of critical factors that can markedly influence the data obtained. These elements play a central role in the quality of the end result. Thus, for novice researchers embarking into protein-protein interactomics, this manuscript serves as a guide to performing TAP and offers plausible explanations for observed anomalies.

### 3.2 *Optimisation of Protein Input Quantity*

A series of half pulldowns was performed to determine the lowest quantity of input protein from which the major constituents of the EGFR del core complex could be identified. The total input protein quantities were: 20, 12.5, 5 and 2.5 mg and the inject quantities ranged between 0.25 to 20% (Figure 5).

At 20 mg protein input (Figure 5A) and 0.25% of the eluate injected (n = 4), only EGFR del was identified plus 2 peptides from UBS3B. Increasing the percentage of injected eluate to 0.5, 2, 3 and 6% led to the identification of GRP78, GRB2, HSP7C, HS71L (0.5%); HSP71, HS90A (2%); ERRFI (3%), and HS90B (6%), respectively. A further increase in injection quantity to 10% (n = 2) did not lead to the identification of any additional interacting proteins. As observed for the full pulldown experiments, increasing the quantity of injected material decreased the maximum sequence coverage (SCV) of all the identified proteins. When the protein input was 12.5 mg (Figure 5B) and 3% of the eluate volume was injected (n = 4), EGFR del was identified with a maximum SCV of 31% (31 PCT from 321 SCT). GRP78, GRB2, HSP7C, UBS3B and ERRFI were also observed. No specific peptides from HS71L and HSP71 were identified but the proteins are grouped with HSP7C. Increasing the percentage of injected eluate volume to 6% resulted in the identification of HS90A (plus HS90B grouped with HS90A). A further increase, however, in the injected eluate volume to 10 and 15% did not lead to the identification of additional interacting proteins. Conversely, HS90A was no longer apparent. Concurrent with this observation was the plateauing of the maximum SCV for EGFR del and the core complex proteins.

At 5 mg input material (equivalent to approximately 1 × 10 cm plate or 4 × $10^6$ cells, Figure 5C) and 15% of the eluate injected (n = 4), EGFR, GRP78, GRB2, HSP7C, UBS3B and

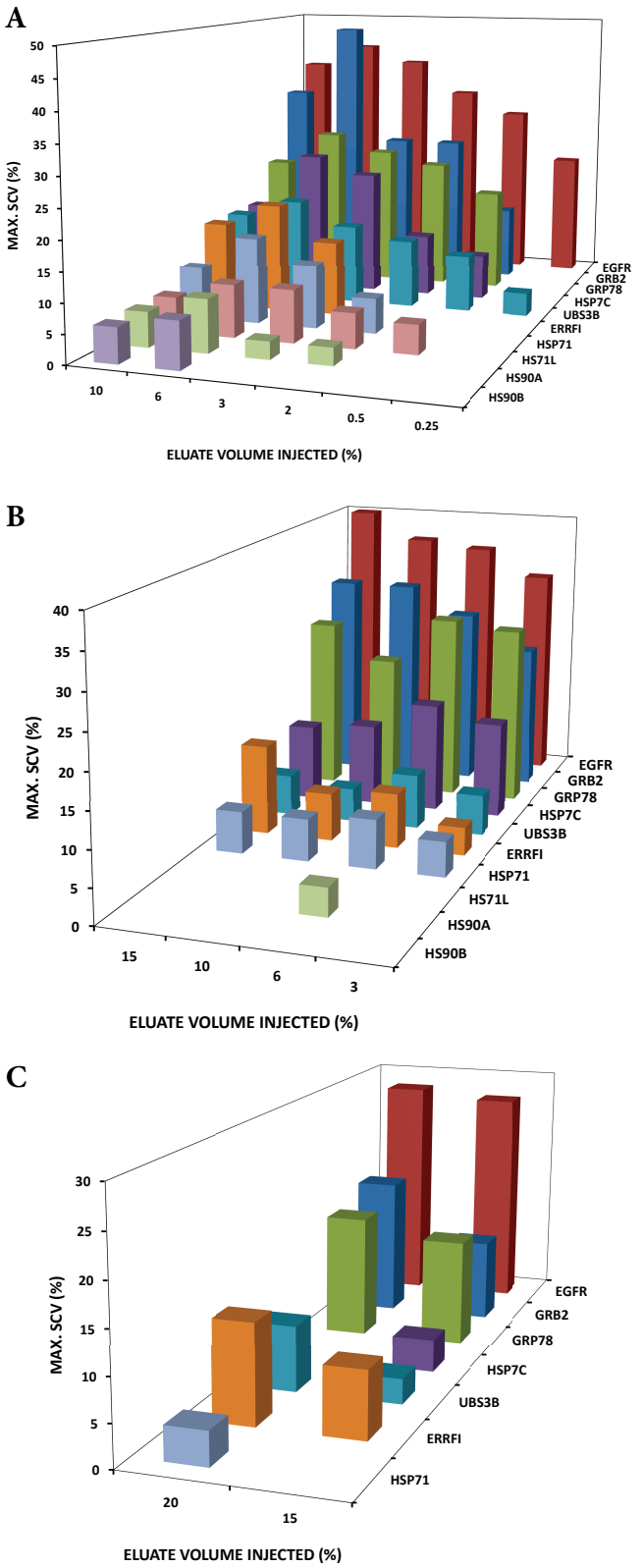


**Figure 5.** Maximum sequence coverage (SCV) obtained for proteins from the EGFR del core complex at variable total protein input quantities (A) 20 mg; (B) 12.5 mg; and (C) 5 mg. Quantity of injected material (% eluate volume) ranged from 0.25 to 20% per LC-MS analysis.

ERRFI were identified. Increasing the quantity of injected eluate volume to 20% resulted in the appearance of HSP71 but with the concomitant disappearance of HSP7C. Neither HS90A nor HS90B were identified. Overall, from the 5 mg protein input experiments, 6 interacting proteins (plus EGFR) were observed. When the total protein input was decreased to 2.5 mg and 33% of the eluate volume injected per technical replicate (n = 4), only EGFR (20% max. SCV, 17 PCT from 99 SCT) and two non-specifically-interacting proteins (GFAP, S10A9) were identified. Therefore this low quantity of input material is insufficient to purify/observe the EGFR del complex and thus provides a lower cut-off value for a successful protein complex purification from retrovirally-transduced cells using the instrumentation and workflow described in this study. See Supplementary Table S2 for an overview of the PCT, SCT and SCV for all proteins identified at the different protein input quantities and eluate injection volumes.

Across the entire half pulldown series of experiments, 9 EGFR del-interacting proteins were identified in common with the data obtained from the full pulldown series. Only the proteins identified with low PCT were absent (SHC1, CDC37, UBIQ, H90B4). Interestingly, when 6% of the eluate volume was injected from the 12.5 mg protein input sample, basically all the major EGFR-interacting proteins were identified. No specific peptides from HS71L and HS90B were selected for fragmentation by mass spectrometry but the proteins are grouped with HSP71 and HS90A, respectively. The data is comparable with the 3% of the eluate volume from the 50 mg protein input sample. Thus, from four times less input material and only double the injection volume, the same results were obtained.

As for the full pulldown series of experiments, there was also evidence of a decrease in PCT and SCV with a concomitant loss of certain proteins with increasing quantities of injected material. When 10% of the digested eluate from 20 mg protein input sample was analysed by LC-MS, there was a decrease in maximum SCV for all identified proteins compared to the data obtained when 6% of the same eluate was analysed (Figure 5A). The effects were not as pronounced for the 12.5 mg protein input (Figure 5B) and not apparent with the 5 mg protein input (Figure 5C).

For all of the purifications performed with EGFR del, the data was filtered to remove single peptide identifications (Material and Methods, section 2.8). Considering that the level of protein input in these experiments was very low, the discarded data was assessed for know interactors of EGFR del. Overall, the majority of the removed proteins were abundant, non-specifically interacting proteins (*e.g.,* tubulin, ribosomal *etc*), however CDC37 (50, 25 and 12.5 mg input material), HS90A (12.5 mg) and UBS3B (2.5 mg) were also present as single peptide identifications. The inclusion of such peptides could be beneficial, particularly in the discovery of novel interactors.

Comparison of the mass spectrometric data with the anti-HA immunoblot for SH-tagged EGFR del revealed some surprising findings. As observed in Supplementary Figure S1, the levels of the SH-tagged EGFR del following the second step of the purification are extremely low, and at 5 and 2.5 mg protein input (Figure S1G and I) the bait protein is below the detection limit of the anti-HA antibody. Although the immunoblot was very weak for the 12.5 mg protein input, the same amount of eluate volume (6%) analysed by LC-MS unequivocally identified all the core proteins of the EGFR del complex. The whole cell lysate (100 µg) and the biotin eluate from first step of the TAP are also given for comparison. As reported by Glatter *et al.* [7], elution from the streptavidin column with biotin is highly-efficient (>90% recovery) with negligible bait protein remaining on the streptavidin beads following elution with Laemmli buffer. From the signal intensity of the anti-HA immunoblot of the final eluate, these workers also estimated that the overall yield of the TAP was 30–40% of bait protein present in the cell lysate. Based on Supplementary Figure S1, the data presented here reflects some of the findings of Glatter *et al.* [7]. This is particularly evident for the one-step biotin elution for the 20 and 12.5 mg protein inputs. The yield of the second-step formic acid elution, however, appears to be slightly lower than that reported. Nonetheless, regardless of the yield determined by immunoblot techniques, the fact remains that it is feasible to purify and identify non-covalently-interacting protein complexes from low input quantities of protein. This alone is a substantial achievement and opens the possibility of exploring and charting molecular pathways from primary cells and/or cells that are particularly difficult to cultivate.

### 3.3 *Grb2 Core Complex*

To confirm the observations made with downscaling the EGFR del SH-TAP, a second protein (Grb2) was cloned with an N-terminal-HA tag via the alternative doxycycline-inducible Flp-In system in HEK293 cells [7]. Tandem affinity purifications were performed with 50, 12.5 and 5 mg protein input material. Following subtraction of the protein groups observed in the GFP negative control (HSP71, HSP7C), 12 proteins remained in the purification from 50 mg total protein input and 3% of the eluate injected (n = 2). Grb2 (bait) was the most abundant protein with a maximum sequence coverage (SCV) of 92% (24 unique peptide counts, PCT, from 295 spectral counts, SCT). The proteins identified as part of the core complex of Grb2 were: ARHG5, SOS1, DYN2, ARG35, SOS2, CBLB, CBL, DYN1, WIPF2, IRS4, and PTPRA. A summary of the PCT, SCT and SCV are given in Table 3A. Increasing the quantity of injected eluate to 5% (n = 2) resulted in the identification of the same proteins, plus one additional non-specific interactor (HNRPK). At 10% (n = 2) and 20% of the eluate injected (n = 1), a total of 6 further proteins were identified, albeit with low PCT, SCT and SCV. These were: WASL, ERRFI, DYN3 (interactors), ACACA (SH-TAP purification contaminant); SNX18 and FA59B (likely non-specific binders).

Figure 6 shows the effect on maximum unique PCT when

**Table 3.** The core complex proteins identified from the SH-Grb2-TAP. (A) 50 mg (n = 2); and (B) 12.5 mg (n = 2) protein input following subtraction of proteins identified in the green fluorescent protein (GFP) negative control. PCT (peptide counts); SCT (spectral counts); SCV (percent sequence coverage). n.i.; not identified.

| ACCESSION CODE | SWISSPROT IDENTIFIER | A | | | B | | |
|---|---|---|---|---|---|---|---|
| | | PCT | SCT | SCV | PCT | SCT | SCV |
| P62993-1 | GRB2 | 24 | 295 | 92 | 17 | 149 | 79 |
| Q12774-1 | ARHG5 | 45 | 95 | 32 | 35 | 66 | 27 |
| Q07889 | SOS1 | 39 | 77 | 29 | 24 | 42 | 19 |
| P50570-2 | DYN2 | 15 | 23 | 15 | 2 | 2 | 2 |
| A5YM69 | ARG35 | 9 | 20 | 20 | 8 | 15 | 20 |
| Q07890 | SOS2 | 11 | 18 | 9 | 5 | 7 | 5 |
| Q13191-1 | CBLB | 9 | 17 | 9 | 6 | 12 | 7 |
| P22681 | CBL | 11 | 18 | 15 | 8 | 13 | 13 |
| Q05193-2 | DYN1 | 8 | 11 | 7 | n.i. | n.i. | n.i. |
| Q8TF74-1/A6NGB9 | WIPF2/3 | 3 | 7 | 11 | 2 | 2 | 6 |
| O014654 | IRS4 | 4 | 6 | 4 | 2 | 2 | 2 |
| P18433-1 | PTPRA | 5 | 8 | 7 | 6 | 11 | 10 |

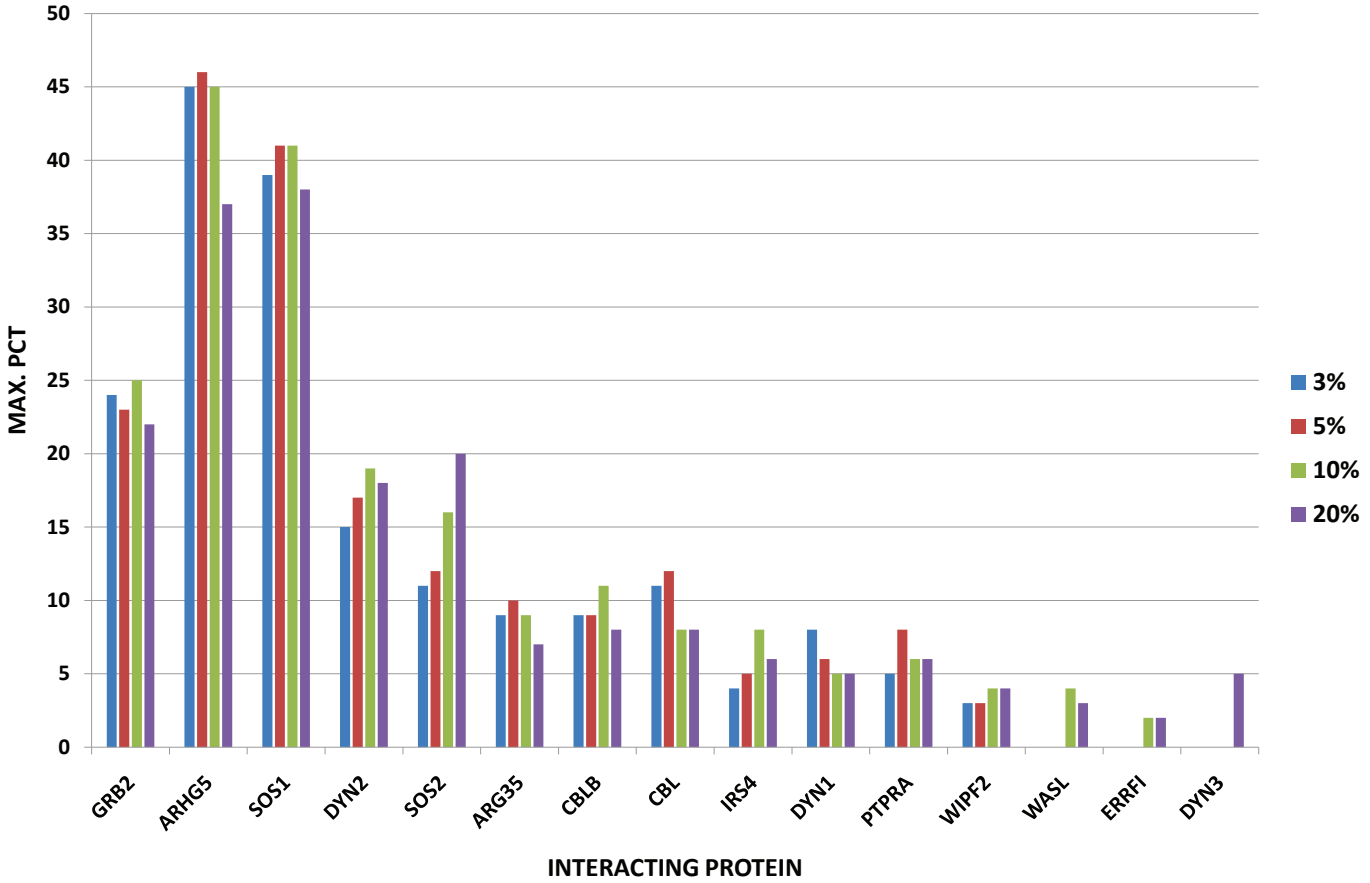


**Figure 6.** The effect on maximum unique PCT for 50 mg protein input with increasing quantities of injected digested SH-Grb2-TAP eluate onto the LC-MS system. Blue (3%), red (5%), green (10%) and purple (20%). For the injected quantities of 3, 5 and 10%, n = 2 and for 20%, n = 1.

the amount of injected material was increased from 3% of the eluate volume (n = 2) to 5% (n = 2), 10% (n = 2) and 20% (n = 1). A similar pattern to the EGFR experiments was obtained (Figure 2). That is, the general tendency for the major proteins was a gradual increase in maximum PCT from 3 to 5% eluate injected (ARHG5, SOS1, ARG35, CBL, PTPRA) or from 3 to 5 to 10% (DYN2, CBLB, IRS4, WIPF2) and then a decrease in PCT as more eluate was injected. DYN 1 had a maximum number of unique PCT at 3% injected, whilst SOS2 was the only protein that steadily increased in the number of unique PCT with increasing percentage of material injected. The low-abundance proteins WASL and ERRFI were only identified from 10% of the eluate injected and above. DYN3 was only observed at the maximum injected percentage.

Decreasing the protein input quantity to 12.5 mg and injection of 50% of the eluate, resulted in the identification of 17 proteins (n = 2). After removal of the proteins observed in the negative control (HSP71, HSP7C, GRP78) and apparent non-specific binders (H14, NFH, ACTA), 11 of the 15 proteins (ARHG5, SOS1, ARG35, CBL, CBLB, PTPRA, SOS2, WIPF3, DYN2, and IRS4) observed in the 50 mg full pull-down experiments remained. Grb2 had a maximum SCV of 79% (17 PCT from 149 SCT). The PCT, SCT and SCV for the identified proteins are summarised in Table 3B. When the protein input quantity was decreased further to 5 mg, only Grb2 (59% max. SCV, 15 PCT from 79 SCT), HSP71 and NFH were identified, thus providing the lower cut-off value for this protein in the HEK293 Flp-In cells. A summary of the proteins identified is given in Supplementary Table S3. As for EGFR del, the single peptide identifications were also assessed. Two known interactors of Grb2, WIPF3 [33] and M4K5 [34] also became evident. Overall, the data generated with the Grb2 experiments were reflective of the results obtained from the EGFR experiments.

## 4. Concluding Remarks

The ability to scale down the SH-TAP to low quantities of input material opens numerous possibilities and potential applications and shows that it is feasible to perform an SH-TAP experiment from a single 10 cm or 15 cm plate of cells (equivalent to 4 - 8 × $10^6$ cells). From this study, we believe that it is now possible to map protein-protein interaction networks from cells that are difficult to cultivate in large quantities (*i.e.*, primary cells) or cells from which it is not feasible to generate a stable cell line without using viral expression systems or the Flp-FRT recombination system. This may allow elucidation of biological networks in a larger panel of cell types and conditions, *e.g.*, detailed experiments with dynamic network perturbations with drugs or RNA interference molecules. Identification of these networks would be of considerable benefit in delineating disease-related signalling pathways in aberrant tissues obtained from mouse models and/or patient material.

Refinement of the liquid chromatography mass spectrometry system (*e.g.*, lower flow rates, smaller diameter analytical columns, submicron ESI needles) and sample preparation techniques (*e.g.*, fractionation at the peptide or protein level, removal of abundant bait protein) could further increase both sensitivity and dynamic range. This then has the potential to allow the identification of protein complexes from even lower quantities of protein material. To improve TAP-MS analyses based on the results from this study, our recommendation to other researchers that are embarking into the field of protein-protein interactomics is to find the balance between the (i) quantity of starting material for the tandem affinity purification; and (ii) the quantity of eluate injected onto the LC-MS system. As each protein is most certainly different, the results from this manuscript act as a guide for other researchers endeavouring to establish TAP in their laboratory or to perform TAP on proteins from unusual cell lines with low quantities of input material. Additionally, we put forward several plausible explanations to account for the phenomena observed in our study. Namely, that increasing quantities of injected material eventually leads to a decrease in the number of proteins identified. Researchers endeavouring to enter this field of research are thus aware of the anomalies that can occur when working at this low level of sensitivity and are therefore fully-prepared for interpretation of the data.

## 5. Supplementary material

Supplementary material regarding this manuscript is online available in the web page of JIOMICS.
http://www.jiomics.com/index.php/jio/rt/suppFiles/81

## Acknowledgements

The authors would like to thank William Pao for providing EGFR del cDNA, Matthias Gstaiger for supplying the original SH constructs and providing valuable advice in establishing SH-TAP protocols at CeMM, Fumi Kinose for assistance with cell culture, Christian Knoll and Elena Rudashevskaya for helpful discussion and input. Work in our laboratory is supported by the Austrian Academy of Sciences, the Austrian Federal Ministry for Science and Research (Gen-Au projects, APP and BIN) and by the Austrian Science Fund FWF (P22282-B11, FG). This study was also funded in part by NIH (5P50CA119997, EBH).

## Competing interests

The authors declare that there are no competing interests.

## References

1. Y. Ho, A. Gruhler, A. Heilbut, G. D. Bader, L. Moore, S. L. Adams, A. Millar, P. Taylor, K. Bennett, K. Boutilier, L. Yang,

C. Wolting, I. Donaldson, S. Schandorff, J. Shewnarane, M. Vo, J. Taggart, M. Goudreault, B. Muskat, C. Alfarano, D. Dewar, Z. Lin, K. Michalickova, A. R. Willems, H. Sassi, P. A. Nielsen, K. J. Rasmussen, J. R. Andersen, L. E. Johansen, L. H. Hansen, H. Jespersen, A. Podtelejnikov, E. Nielsen, J. Crawford, V. Poulsen, B. D. Sorensen, J. Matthiesen, R. C. Hendrickson, F. Gleeson, T. Pawson, M. F. Moran, D. Durocher, M. Mann, C. W. Hogue, D. Figeys, M. Tyers, Nature 415 (2002) 180-183.

2. A. C. Gavin, M. Bosche, R. Krause, P. Grandi, M. Marzioch, A. Bauer, J. Schultz, J. M. Rick, A. M. Michon, C. M. Cruciat, M. Remor, C. Hofert, M. Schelder, M. Brajenovic, H. Ruffner, A. Merino, K. Klein, M. Hudak, D. Dickson, T. Rudi, V. Gnau, A. Bauch, S. Bastuck, B. Huhse, C. Leutwein, M. A. Heurtier, R. R. Copley, A. Edelmann, E. Querfurth, V. Rybin, G. Drewes, M. Raida, T. Bouwmeester, P. Bork, B. Seraphin, B. Kuster, G. Neubauer, G. Superti-Furga, Nature 415 (2002) 141-147.
3. A. C. Gavin, P. Aloy, P. Grandi, R. Krause, M. Boesche, M. Marzioch, C. Rau, L. J. Jensen, S. Bastuck, B. Dumpelfeld, A. Edelmann, M. A. Heurtier, V. Hoffman, C. Hoefert, K. Klein, M. Hudak, A. M. Michon, M. Schelder, M. Schirle, M. Remor, T. Rudi, S. Hooper, A. Bauer, T. Bouwmeester, G. Casari, G. Drewes, G. Neubauer, J. M. Rick, B. Kuster, P. Bork, R. B. Russell, G. Superti-Furga, Nature 440 (2006) 631-636.
4. N. J. Krogan, G. Cagney, H. Yu, G. Zhong, X. Guo, A. Ignatchenko, J. Li, S. Pu, N. Datta, A. P. Tikuisis, T. Punna, J. M. Peregrin-Alvarez, M. Shales, X. Zhang, M. Davey, M. D. Robinson, A. Paccanaro, J. E. Bray, A. Sheung, B. Beattie, D. P. Richards, V. Canadien, A. Lalev, F. Mena, P. Wong, A. Starostine, M. M. Canete, J. Vlasblom, S. Wu, C. Orsi, S. R. Collins, S. Chandran, R. Haw, J. J. Rilstone, K. Gandi, N. J. Thompson, G. Musso, P. St Onge, S. Ghanny, M. H. Lam, G. Butland, A. M. Altaf-Ul, S. Kanaya, A. Shilatifard, E. O'Shea, J. S. Weissman, C. J. Ingles, T. R. Hughes, J. Parkinson, M. Gerstein, S. J. Wodak, A. Emili, J. F. Greenblatt, Nature 440 (2006) 637-643.
5. T. Bouwmeester, A. Bauch, H. Ruffner, P. O. Angrand, G. Bergamini, K. Croughton, C. Cruciat, D. Eberhard, J. Gagneur, S. Ghidelli, C. Hopf, B. Huhse, R. Mangano, A. M. Michon, M. Schirle, J. Schlegl, M. Schwab, M. A. Stein, A. Bauer, G. Casari, G. Drewes, A. C. Gavin, D. B. Jackson, G. Joberty, G. Neubauer, J. Rick, B. Kuster, G. Superti-Furga, Nat Cell Biol 6 (2004) 97-105.
6. C. Behrends, M. E. Sowa, S. P. Gygi, J. W. Harper, Nature 466 (2010) 68-76.
7. T. Glatter, A. Wepf, R. Aebersold, M. Gstaiger, Mol Syst Biol 5 (2009) 237.
8. G. Rigaut, A. Shevchenko, B. Rutz, M. Wilm, M. Mann, B. Seraphin, Nat Biotechnol 17 (1999) 1030-1032.
9. T. Burckstummer, K. L. Bennett, A. Preradovic, G. Schutze, O. Hantschel, G. Superti-Furga, A. Bauch, Nat Methods 3 (2006) 1013-1019.
10. A. K. Al-Hakim, O. Goransson, M. Deak, R. Toth, D. G. Campbell, N. A. Morrice, A. R. Prescott, D. R. Alessi, J Cell Sci 118 (2005) 5661-5673.
11. J. Gregan, C. G. Riedel, M. Petronczki, L. Cipak, C. Rumpf, I. Poser, F. Buchholz, K. Mechtler, K. Nasmyth, Nat Protoc 2 (2007) 1145-1151.
12. A. L. Barabasi, N Engl J Med 357 (2007) 404-407.
13. T. Pawson, R. Linding, FEBS Lett 582 (2008) 1266-1270.
14. A. L. Barabasi, N. Gulbahce, J. Loscalzo, Nat Rev Genet 12 (2011) 56-68.
15. T. Yoshida, G. Zhang, E. B. Haura, Biochem Pharmacol 80 (2010) 613-623.
16. W. Pao, V. Miller, M. Zakowski, J. Doherty, K. Politi, I. Sarkaria, B. Singh, R. Heelan, V. Rusch, L. Fulton, E. Mardis, D. Kupfer, R. Wilson, M. Kris, H. Varmus, Proc Natl Acad Sci U S A 101 (2004) 13306-13311.
17. E. B. Haura, A. Muller, F. P. Breitwieser, J. Li, F. Grebien, J. Colinge, K. L. Bennett, J Proteome Res 10 (2011) 182-190.
18. K. L. Bennett, L. A. Hick, S. V. Smith, R. J. W. Truscott, M. M. Sheil, J. Mass Spectrom. 30 (1995) 769-771.
19. T. M. Annesley, Clin Chem 49 (2003) 1041-1044.
20. J.-P. Antignac, K. d. Wasch, F. Monteau, H. D. Brabander, F. Andre, B. L. Bizec, Analytica Chimica Acta 529 (2005) 129-136.
21. P. Kebarle, L. Tang, Anal. Chem. 65 (1993) 972A.
22. E. Grushka, Analytical Chemistry 42 (1970) 1142-&.
23. F. Giorgianni, A. Cappiello, S. Beranova-Giorgianni, P. Palma, H. Trufelli, D. M. Desiderio, Anal Chem 76 (2004) 7028-7038.
24. J. Li, U. Rix, B. Fang, Y. Bai, A. Edwards, J. Colinge, K. L. Bennett, J. Gao, L. Song, S. Eschrich, G. Superti-Furga, J. Koomen, E. B. Haura, Nat Chem Biol 6 (2010) 291-299.
25. N. V. Fernbach, M. Planyavsky, A. Muller, F. P. Breitwieser, J. Colinge, U. Rix, K. L. Bennett, J Proteome Res 8 (2009) 4753-4765.
26. J. Rappsilber, Y. Ishihama, M. Mann, Anal Chem 75 (2003) 663-670.
27. J. Colinge, A. Masselot, M. Giron, T. Dessingy, J. Magnin, Proteomics 3 (2003) 1454-1463.
28. K. Kowanetz, N. Crosetto, K. Haglund, M. H. Schmidt, C. H. Heldin, I. Dikic, J Biol Chem 279 (2004) 32786-32795.
29. T. Shimamura, A. M. Lowell, J. A. Engelman, G. I. Shapiro, Cancer Res 65 (2005) 6401-6408.
30. P. J. Gray, Jr., T. Prince, J. Cheng, M. A. Stevenson, S. K. Calderwood, Nat Rev Cancer 8 (2008) 491-495.
31. D. H. Lundgren, S. I. Hwang, L. Wu, D. K. Han, Expert Rev Proteomics 7 (2010) 39-53.
32. O. N. Jensen, A. Shevchenko, M. Mann: Protein analysis by mass spectrometry. In *Book Protein analysis by mass spectrometry* (Editor Hames, B.D.). pp. 29-57. City: Oxford University Press Inc.; 1997:29-57.
33. N. Bisson, D. A. James, G. Ivosev, S. A. Tate, R. Bonner, L. Taylor, T. Pawson, Nat Biotechnol 29 (2011) 653-658.
34. C. Wu, M. H. Ma, K. R. Brown, M. Geisler, L. Li, E. Tzeng, C. Y. Jia, I. Jurisica, S. S. Li, Proteomics 7 (2007) 1775-1785.

JOURNAL OF INTEGRATED OMICS
A METHODOLOGICAL JOURNAL
HTTP://WWW.JIOMICS.COM

ORIGINAL ARTICLE | DOI: 10.5584/jiomics.v2i1.84

# Proteomic Response to Arsenic Stress in *Chromobacterium violaceum*

Alessandra Ciprandi[1,2], Rafael Azevedo Baraúna[2], Agenor Valadares Santos[2], Evonnildo Costa Gonçalves[2], Marta Sofia Peixe Carepo[3], Maria Paula Cruz Schneider[4], Artur Silva*[1,2]

[a]Núcleo de Medicina Tropical, Universidade Federal do Pará, Belém, Brasil. [b]Laboratório de Polimorfismo de DNA, Instituto de Ciências Biológicas, Universidade Federal do Pará, Belém, Brasil. [c]REQUIMTE/CQFB, Departamento de Química, Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa, Caparica, Portugal.

**Received:** 09 December 2011 **Accepted:** 11 April 2012 **Available Online**: 19 April 2012

ABSTRACT

Exposure to arsenic, whether acute or chronic, is a public health problem in many parts of the world and is associated with various types of deleterious effects on human health. One way that risk of contamination with this metalloid can be reduced is through bioremediation of areas contaminated with arsenic. Natural resistance mechanisms are widely distributed in microorganisms, meriting further study in an effort to improve their efficiency. Chromobacterium violaceum is a betaproteobacterium found in tropical and subtropical regions; its resistance to arsenic is controlled by an operon, arsRBC. The proteins expressed by the operon ars have been well studied; however, the overall cell response that determines resistance to this metalloid is little understood. We investigated changes in protein expression in response to arsenite. This was done through two-dimensional differential gel electrophoresis (2D-DIGE). Quantities of 26 proteins were altered after treatment with arsenite, 23 of which increased. The differential spots were analyzed with MS and MS/MS; eight proteins were identified that are involved in response to oxidative stress (SOD, GST, Grx), in DNA repair and in the metabolism of lipids, amino acids and coenzymes. We conclude that the response of C. violaceum to arsenite involves defense mechanisms against oxidative stress and alterations in cell metabolic cycles.

**Keywords:** *Chromobacterium violaceum*; Arsenic; Proteomics; Bioremediation.

## 1. Introduction

Arsenic is a metalloid that is widely distributed on the earth's surface. The forms that are most important biologically are arsenate ($As^{5+}$) and arsenite ($As^{3+}$) [1]. The arsenic found in water and soil is there principally due to natural processes, such as volcanic emissions and leaching; though it is intensified by human activities, including mining and arsenical-containing fungicides, pesticides, herbicides, and wood preservatives [2].

Exposure to arsenic in drinking water is a public health problem that affects various parts of world. Inorganic arsenic is a human carcinogen and is associated with chronic and cardiovascular diseases [3]. Arsenite bonds strongly to thiol groups of proteins, while arsenate is a structural analog of phosphate and competes with this element in many phosphorylation reactions [4].

Currently, remediation of arsenic in water and soil is done through physical-chemical interventions; however, these are high cost, energy-intensive measures, and the resulting waste still needs to be discarded or eliminated [5]. Microorganisms have developed various strategies to counter the toxicity of arsenic that can be used in bioremediation, through bioprocesses of adsorption, precipitation, complexation and bioaccumulation [6]. Improvement of the capacity of these bacteria requires a better understanding of the resistance mechanisms of these organisms.

In nature, microorganisms respond to arsenic in different ways, but the most common pathway in bacteria is mediated by the operon *ars* [7], which codes for three proteins: an arsenate reductase (ArsC), an efflux pump (ArsB) and a transcription regulator (ArsR). Additionally, in some bacte-

***Corresponding author:** A. Silva. Laboratório de Polimorfismo de DNA, Instituto de Ciências Biológicas. Universidade Federal do Pará. Rua Augusto Corrêa, s/n, Guamá. CEP 66075-900, Belém, PA, Brasil. Phone: +55-9132018426, Fax +55-9132017568, Email Address: asilva@ufpa.br

ria, the operon can include genes that code for an ATPase (ArsA) and a metallochaperone (ArsD) [8]. Arsenate reductase uses glutaredoxin or thioredoxin to reduce the arsenate to arsenite, which is expulsed from the cell by ArsB or by the ArsAB pump [1].

*Chromobacterium violaceum* is a Gram-negative, facultative anaerobic betaproteobacterium, found in the water and soil of tropical and subtropical regions [9]. It is an opportunistic pathogen and has a very versatile metabolism, making it capable of confronting diverse types of environmental conditions [10].

The most notable phenotypic characteristic of *C. violaceum* is violacein, a violet pigment that has antibacterial, antifungal, cytotoxic and antiviral activity [9]. This bacterium produces chitinases [11], polyhydroxyalkanoates [12], cellulase [13], and it is capable of solubilizing metals [14].

Genome analysis of *C. violaceum* demonstrated an operon *ars* of the type *ars*RBC [9, 15]. This operon was shown to be functional by Azevedo and colleagues [16], who demonstrated that expression of the gene *ars*R is increased in response to arsenite and expression of this gene is correlated with resistance to this metalloid. The arsenical resistance tests in *C. violaceum* indicated that this bacteria was able to tolerate micromolar concentrations of arsenite, and the IC50 value for arsenite was determined to be approximately 500 μM [Rocha et al., unpublished data]. However, the global effects of arsenite on metabolism were yet not reported. Here, we used a differential proteomic approach focused on metabolic cycles involved in the adaptation of *C. violaceum* to arsenite that are not directly correlated with this resistance.

## 2. Materials and Methods

### 2.1 *Bacterial growth and treatment with arsenite*

We used *C. violaceum* ATCC 12472 in our study. The bacteria were grown on Luria-Bertani (LB) medium containing 0.1 mg/mL ampicillin, at 28°C with 200 rpm agitation, according to [16]. One hundred microliters of pre-inoculum were diluted in 10 mL of fresh medium containing 10 μM sodium arsenite and incubated for 16 h. This inoculum was then diluted in 50 mL of fresh culture medium containing 100 μM sodium arsenite, order to obtain an optical density (OD) of 0.05 at 720 nm [17], and incubated until it reached an exponential growth phase ($DO_{720}$ = 1.00). As a control, the bacteria were grown under the same conditions, but without arsenite.

### 2.2 *Protein Extraction*

The bacteria were harvested by centrifuging the culture medium at 5,000 rpm, for 10 min, at 4°C. The liquid medium was discarded, and the bacterial pellet was washed three times with 50 mM Tris-HCl pH 7.5 and resuspended in 1 mL of lysis buffer (7 M urea, 2 M thiourea, 2% CHAPS, 75 mM DTT, and a cocktail of protease inhibitors). Cells were sonicated on ice for five cycles of 10s, with 10 s intervals. The lysate was centrifuged for 40 min, at 14,000 rpm, at 4°C. The soluble proteins were stored at -70°C until use. The protein samples were quantified using the 2D Quant kit (GE Healthcare).

### 2.3 *2D-DIGE*

The proteins in the extracts of three replicates from arsenic -treated and control bacteria were precipitated with methanol/chloroform and dissolved in sample buffer (7 M urea, 2 M thiourea, 4% CHAPS, 30 mM Tris-HCl pH 9). Analytical gels were prepared with 50 μg of protein labeled with fluorescent dyes Cy3 or Cy5 (400 pmol), according to manufacturer's instructions (GE Healthcare). An internal control was made with a mixture of treated and control group material, totaling 50 μg of protein, labeled with Cy2. The labeling reaction was stopped by adding 1 μL of 10 mM lysine for 10 min. The differentially labeled samples were combined and diluted with rehydration buffer (7 M urea, 2 M thiourea, 2% CHAPS, 1% ampholytes pH 4-7, 75 mM DTT and 0.002% bromophenol blue), were used to rehydrate immobilized pH gradient (IPG) gel strip for isoelectric focusing (IEF) (pH 4-7, 18 cm, GE Healthcare). The IEF was performed in *Ettan IPGphor* (GE-Healthcare) until 80,000 Vh. After equilibration for 15 min in a 50 mM Tris HCl (pH 8.8) buffer solution containing 6 M urea, 2% SDS, 30% glycerol, 0.001% bromophenol blue and 10 mg/mL DTT, the strips were equilibrated for 15 min in the same solution, except that the DTT was replaced by 25 mg/mL iodoacetamide. The SDS-PAGE was performed on a 15% polyacrylamide gel using an Ettan DALTsix device (GE-Healthcare). DIGE gels were scanned using a Ettan DIGE Imager (GE Healthcare), and image analysis was performed with ImageMaster 2D Platinum 7.0 software (GE Healthcare). Spots were considered to be differential when the mean ratio between volumes was ± 1.3 times and $p < 0.05$ in the ANOVA test.

For the preparative gels, 600 μg of proteins from extracts of bacteria that had been treated or not with arsenite were resolved by 2DE as previously described. The proteins were stained with colloidal Coomassie blue. The gels were digitized using ImageScanner (GE Healthcare), and image analysis was performed with ImageMaster 2D Platinum 7.0 software (GE Healthcare).

### 2.4 *Digestion in gel with trypsin*

The differential spots were excised from the preparative gels (three for each condition), using SpotPicking Ettan (GE Healthcare). The digestion was performed according to a previously described protocol [18], with some modifications. The spots were destained with 25 mM ammonium bicarbonate /50% acetonitrile (v/v), treated with acetonitrile for 5 min and completely dried. The gel pieces were digested with modified trypsin (Promega) 20 ng/μL in 25 mM ammonium bicarbonate, at 37°C for 16 h. The peptides were extracted

with 2.5% formic acid/50% acetonitrile (v/v). The peptide solution was concentrated in a SpeedVac (Savant, USA) to a volume of about 10 iL, and desalted using P10 ZipTip C18 (MilliPore).

2.5 *Mass Spectrometry and Protein Identification*

For identification of the proteins, 0.5 µL of the peptide solution and 0.5 µL of the alpha-cyano-4-hydroxycinnamic acid matrix 10 mg/mL were mixed on the target. Analysis through PMF was made in an AXIMA-CRF MALDI-TOF mass spectrometer (Shimadzu). The acquisition parameters were: positive reflector mode, laser repetition rate of 5 Hz, 20 kV acceleration voltage, m/z range of 500 - 4,000 Da; 200 shots were accumulated per spectrum and spectra were processed usisng the MALDI-MS Application software tool (Shimadzu). The MS/MS analysis was done using a AutoFlex III MALDI-TOF/TOF mass spectrometer (Bruker Daltonics). The acquisition parameters were: positive reflector mode, laser repetition rate of 50 Hz, 25 kV acceleration voltage, 500 to 4,000 Da m/z range; spectra were processed using FlexControl software tool (Bruker Daltonics). Protein identification was performed using the Mascot software (http://www.matrixscience.com/) to search the genome of *C. violaceum* deposited in the NCBI data bank (http://www.ncbi.nlm.nih.gov/NCBInr). The search parameters were: other proteobacteria as the taxonomic group, monoisotopic mass, tolerance of 0.5 Da, 1 missed cleavage, carbamidomethylation of cysteine as a fixed modification, and oxidation of methionine as a variable modification.

## 3. Results and Discussion

Exposure of *C. violaceum* to arsenite altered the expression of 26 spots on the DIGE, 23 of which were increased and 3 reduced (Fig. 1). The range of the different ratio of spot intensities of bacteria treated with arsenite in comparison to control levels is between 1.3 and 2.4. The differential spots were cut, digested with trypsin and submitted to analysis by MS and MS/MS. Eight proteins were identified in 10 spots, two of which were identified by PMF, four by MS/MS and four by PMF and MS/MS (Table 1).

The differential proteome of *C. violaceum* in response to arsenite revealed that the major alterations caused by this metalloid are associated with oxidative stress. Three proteins directly involved in the cellular response to oxidative stress had their expression increased in response to arsenite: superoxide dismutase (SOD), glutathione-S-transferase (GST) and glutaredoxin (Grx). Spots 907 and 909, identified as the enzyme SOD, had the same molecular weight and a small difference in pI, which could be due to posttranslational

modifications. The same was found for spots 891 and 901, identified as a protein of the GST family.

Oxidative stress occurs when reactive oxygen species (ROS) are elevated, which could cause lipid peroxidation, DNA damage and protein oxidation [19], as well as inactivation of iron-sulfur centers, which are essential for many electron transfer proteins.

The enzyme SOD converts the superoxide ion into hydrogen peroxide, which is then reduced to $H_2O$ by peroxidases, such as the peroxiredoxins (Prx), which are dependent on the thioredoxin system [20]. The GSTs are enzymes that catalyze the conjugation of glutathione with toxic agents and participate in the regeneration of S-thiolated proteins through oxidative stress [21]. The GST CV_0289 of *C. violaceum* belongs to the beta class, specific to bacteria, and can aid in the response to arsenic-induced oxidative stress. Induction of these enzymes observed in *C. violaceum* would be due to the generation of ROS within the cell, induced by arsenic, and has been reported from other bacteria, including *Comamonas sp.* [22], *Herminiimonas arsenicoxydans* [23] and *Leptospirillum ferriphilum* [24]. Arsenic disturbs the redox equilibrium through mechanisms such as activation of NADPH oxidase, inhibition of glutathione-peroxidase and binding of thiol groups of regulatory proteins [3, 25].

An increase in the expression of a single-strand DNA-binding protein (SSB), CV_1889, was also observed in the proteome of *C. violaceum*. The SSBs have various functions involved in the replication, repair and recombination of DNA [26]. CV_1889 is part of DNA repair pathways through recombination and the SOS response of *C. violaceum* [27]; and increased expression of this protein could have protective role against the DNA damage caused by oxidative stress due to arsenic.

Arsenic also affects metabolic pathways that are important for cell maintenance, increasing the expression of two enzymes involved in lipid metabolism, acyl-CoA dehydrogenase (ACAD) and enoil-CoA hidratase (ECAH), an enzyme involved in amino acid metabolism, 3-hydroxyisobutyrate

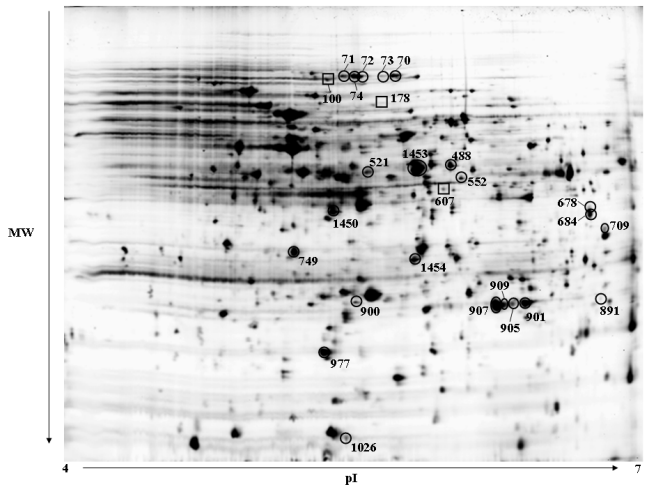


**Figure 1**. 2D-DIGE image of proteins of C. violaceum treated with arsenite. The circles show proteins whose expression increased and squares show proteins whose expression decreased with arsenite treatment.

**Table 1.** *C. violaceum* proteins that are differentially expressed in response to arsenite, identified by MS and MS/MS .

| Spot[a] | RefSeq[b] | Protein | Theoretical MW / pI | Score[c] | Ars/Ctr[d] |
|---|---|---|---|---|---|
| *Replication, recombination and repair*[h] | | | | | |
| 977[e] | NP_901559 | CV_1889 Single-strand DNA-binding protein | 17.11 / 5.25 | 94 | 1.73 |
| *Posttranslational modification, protein turnover, chaperones*[h] | | | | | |
| 749[f] | NP_901706 | CV_2036 Peroxiredoxin/glutaredoxin family protein | 26.97 / 5.09 | 68 | 1.35 |
| 891[f] | NP_899959 | CV_0289 Glutathione S-transferase family protein | 22.60 / 5.91 | 248 | 1.45 |
| 901[g] | NP_899959 | CV_0289 Glutathione S-transferase family protein | 22.60 / 5.91 | 89 | 1.71 |
| *Inorganic ion transport and metabolism*[h] | | | | | |
| 907[g] | NP_902174 | CV_2504 Superoxide dismutase | 21.63 / 5.87 | 77 | 1.59 |
| 909[f] | NP_902174 | CV_2504 Superoxide dismutase | 21.63 / 5.87 | 148 | 1.59 |
| *Coenzyme transport and metabolism*[h] | | | | | |
| 900[e] | NP_902057 | CV_2387 Riboflavin synthase subunit alpha | 22.28 / 5.34 | 107 | 1.31 |
| *Lipid transport and metabolism*[h] | | | | | |
| 488[g] | NP_901754 | CV_2084 Acyl-CoA dehydrogenase | 42.00 / 5.64 | 85 | 1.41 |
| 1454[g] | NP_901753 | CV_2083 Enoyl-CoA hydratase | 29.40 / 6.34 | 88 | 1.30 |
| *Amino Acid transport and metabolism*[h] | | | | | |
| 709[f] | NP_901751 | CV_2081 3-Hydroxyisobutyrate dehydrogenase | 30.10 / 6.24 | 68 | 1.34 |

[a.] Spot number refers to Fig. 1. [b.] RefSeq = NCBI access number. [c.] Score = MASCOT *Mowse score*. [d.] Ratio of spot intensities of extracts of *C. violaceum* treated with arsenite, compared to control levels. [e.] Spot identified by PMF; f. spot identified by MS/MS; [g.] spot identified by PMF and confirmed by MS/MS. [h.] Functional classification according to COG (Clusters of Orthologous Groups) .

dehydrogenase (3-HIBADH), and an enzyme involved in coenzyme metabolism, riboflavin synthase.

The two beta-oxidation enzymes that had increased expression, ACAD and ECAH, catalyze, respectively, the first and second steps of the reaction cycle that liberates acetyl-CoA, directed towards the citric acid or glyoxylate cycles, and electrons, carried by the coenzymes NADH and $FADH_2$ for the electron transport chain of energy generation. The activity of beta-oxidation enzymes is low when fatty acids are lacking; its genes are regulated by transcription factor FadR, cAMP, long-chain fatty acids and the life-cycle [28]. Expression of these enzymes can respond to stress conditions, such as salinity [29] and biosynthesis of polyhydroxyalkanoates [30].

Riboflavin synthase catalyzes the final step of riboflavin biosynthesis (vitamin B2), precursor of the coenzymes FMN and FAD, which are flavoproteins that are essential for energy metabolism, oxidation-reduction reactions and biosynthetic pathways, including the violacein biosynthetic pathway [31]. The riboflavin biosynthesis pathway responds to stress provoked by superoxides [32], metals and low phosphate levels [33]. In *C. violaceum*, expression of the riboflavin synthase alpha subunit was increased by arsenic or by arsenic-induced oxidative stress; this could be a response to a need for coenzymes for oxidation-reduction reactions. The riboflavin biosynthesis pathway also consumes ribulose-5-phosphate, produced by the pentose phosphate pathway, together with NADPH.

*C. violaceum* treated with arsenite had increased expression of 3-HIBADH, a key enzyme in the metabolism of valine, which catalyzes the reversible oxidation of 3- hidroxyisobutyrate to methylmalonate semialdehyde. This reaction produces NADPH and can have L-serine, D-threonine and other 3-hydroxy acid derivatives as substrates [34]. The function of 3-HIBADH can be understood together with that of other enzymes of the oxidative pathways that had increased expression, suggesting an overall effort by the bacteria to supply redox repair reactions, recruiting equivalent reducers normally involved in the oxidation of amino acids and fatty acids.

## 4. Conclusions

This is the first proteomic analysis of resistance of *C. violaceum* to arsenite, demonstrating the complexity of bacterial adaptation to a stress condition. The results suggest that this bacterium develops an ample response to oxidative stress provoked by arsenite, activating various enzymes involved in the elimination of oxygen radicals and repair of damage caused by these radicals, both at protein and DNA-repair levels. Arsenite in the culture medium also induces the expression of various proteins involved in cell processes, including energy metabolism and production of equivalent reducers.

## Acknowledgements

This work was supported by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Fundação de Amparo à Pesquisa do Estado do Pará (FAPESPA), and Centrais Elétricas do Norte do Brasil

(Eletronorte). MSPC would like to thank Programa Ciência 2007 and PEst-C/EQB/LA0006/2011 of Fundação para a Ciência e Tecnologia.

## References


1. BP Rosen, FEBS Lett 529 (2002) 86-92.
2. D Lièvremont, PN Bertin, MC Lett, Biochimie 91 (2009) 1229 -1237.
3. LC Platanias, J Biol Chem 284 (2009) 18583-18587.
4. LM Mateos, E Ordóñez, M Letek, JA Gil, Int Microbiol 9 (2006) 207-215.
5. S Wang, X Zhao, J Environ Manage 90 (2009) 2367-2376.
6. SL Tsai, S Singh, W Chen, Curr Opin Biotechnol 20 (2009) 659-667.
7. S Silver, LT Phung, Appl Environ Microbiol 71(2005) 599-608.
8. R Mukhopadhyay, BP Rosen, Environ Health Perspect 110 (2002) 745-748.
9. Brazilian National Genome Project Consortium, Proc Natl Acad Sci USA 100 (2003) 11660-11665.
10. CI Lima-Bittencourt, S Astolfi-Filho, E Chartone-Souza, FR Santos, AM Nascimento, BMC Microbiol 7 (2007) 58.
11. LS Chernin, MK Winson, JM Thompson, S Haran, BW Bycroft, I Chet, P Williams, GS Stewart, J Bacteriol 180 (1998) 4435-4441.
12. A Steinbuèchel, EM Debzi, RH Marchessault, A Timm, Appl. Microbiol. Biotechnol 39 (1993) 443-449.
13. DO Recouvreux, CA CarminattI, AK Pitlovanciv, CR Rambo, LM Porto, RV Antônio, Curr Microbiol 57 (2008) 469-476.
14. MA Faramarzi, M Stagars, E Pensini, W Krebs, H Brandl, J Biotechnol 113 (2004) 321-6.
15. MS Carepo, JS Azevedo, JI Porto, AR Bentes-Sousa, JDA Batista, AL Silva, MP Schneider, Genet Mol Res 3(2004) 181-194.
16. JS Azevedo, R Silva-Rocha, A Silva, MS Peixe Carepo, MP Cruz Schneider, Can J Microbiol 54 (2008) 137-142.
17. RD Demoss, ME Happel, J Bacteriol 77 (1959) 137-141.
18. A Shevchenko, H Tomas, J Havlis, JV Olsen, M Mann, Nat Protoc 1 (2006) 2856-2860.
19. M Kohárýová, M Kolárová, Gen Physiol Biophys 27 (2008) 71 -84.
20. T Rabilloud, M Heller, F Gasnier, S Luche, C Rey, R Aebersold, M Benahmed, P Louisot, J Lunardi, J Biol Chem 277 (2002) 19396-19401.
21. D Sheehan, G Meade, VM Foley, CA, Biochem J 360 (2001) 1 -16.
22. Y Zhang, YF Ma, SW Qi, B Meng, MT Chaudhry, SQ Liu, SJ Liu, Microbiology 153 (2007) 3713-3721.
23. J Cleiss-Arnold, S Koechler, C Proux, ML Fardeau, MA Dillies, JY Coppee, F Arsène-Ploetze, PN Bertin, BMC Genomics 11 (2010) 709.
24. B Li, J Lin, S Mi, J Lin, Bioresour Technol 101 (2010) 9811-9814.
25. J Dai, RS Weinberg, S Waxman, Y Jing, Blood 93 (1999) 268–277.
26. RR Meyer, PS Laine, Microbiol Rev 54 (1990) 342-380.
27. FT Duarte, FM Carvalho, U Bezerra e Silva, KC Scortecci, CA Blaha, LF Agnez-Lima, SR Batistuzzo de Medeiros, Genet Mol Res 3 (2004) 167-180.
28. CC Dirusso, PN Black, JD Weimar, Prog Lipid Res 38 (1999) 129-197.
29. RK Kapardar, R Ranjan, A Grover, M Puri, R Sharma, Bioresour Technol 101 (2010) 3917-3924.
30. K Peplinski, A Ehrenreich, C Döring, M Bömeke, A Steinbüchel, Appl Microbiol Biotechnol 88 (2010) 1145-1159.
31. CJ Balibar, CT Walsh, Biochemistry 45 (2006) 15444-15457.
32. AG Vitreschak, DA Rodionov, AA Mironov, MS Gelfand, Nucleic Acids Res 30 (2002) 3141-3151.
33. FH Knegt, LV Mello, FC Reis, MT Santos, R Vicentini, LF Ferraz, LM Ottoboni, Res Microbiol 159 (2008) 423-431.
34. T Yao, L Xu, H Ying, H Huang, M Yan, Appl Biochem Biotechnol 160 (2010) 694-703.