

TKK Dissertations 40
Espoo 2006

**EXPRESSION MICROARRAY TECHNOLOGY AS A TOOL
IN CANCER RESEARCH**

Doctoral Dissertation

Antti Kokko



**Helsinki University of Technology
Department of Chemical Technology
Laboratory of Bioprocess Engineering**

TKK Dissertations 40
Espoo 2006

EXPRESSION MICROARRAY TECHNOLOGY AS A TOOL IN CANCER RESEARCH

Doctoral Dissertation

Antti Kokko

Dissertation for the degree of Doctor of Science in Technology to be presented with due permission of the Department of Chemical Technology for public examination and debate in Auditorium AS1 at Helsinki University of Technology (Espoo, Finland) on the 29th of September, 2006, at 12 noon.

**Helsinki University of Technology
Department of Chemical Technology
Laboratory of Bioprocess Engineering**

**Teknillinen korkeakoulu
Kemian tekniikan osasto
Bioprosessitekniiikan laboratorio**



**University of Helsinki
Haartman Institute and Biomedicum Helsinki
Molecular and Cancer Biology Research Program
Department of Medical Genetics**

**Helsingin yliopisto
Haartman-instituutti ja Biomedicum Helsinki
Molekyyli- ja syöpäbiologian tutkimusohjelma
Lääketieteellisen genetiikan osasto**



UNIVERSITY OF HELSINKI

Distribution:

Helsinki University of Technology
Department of Chemical Technology
Laboratory of Bioprocess Engineering
P.O. Box 6100 (Kemistintie 1)
FI - 02015 TKK
FINLAND

Tel. +358-9-451 2541

Fax +358-9-462 373

URL: <http://www.tkk.fi/Units/BioprocessEngineering/>

E-mail: bio-info@list.hut.fi

© 2006 Antti Kokko

ISBN 951-22-8337-9

ISBN 951-22-8338-7 (PDF)

ISSN 1795-2239

ISSN 1795-4584 (PDF)

URL: <http://lib.tkk.fi/Diss/2006/isbn9512283387/>

TKK-DISS-2168

Yliopistopaino

Helsinki 2006



HELSINKI UNIVERSITY OF TECHNOLOGY P. O. BOX 1000, FI-02015 TKK http://www.tkk.fi		ABSTRACT OF DOCTORAL DISSERTATION	
Author Antti Kokko			
Name of the dissertation Expression microarray technology as a tool in cancer research			
Date of manuscript 06.06.2006		Date of the dissertation 29.09.2006	
<input type="checkbox"/> Monograph		<input checked="" type="checkbox"/> Article dissertation (summary + original articles)	
Department	Department of Chemical Technology		
Laboratory	Laboratory of Bioprocess Engineering		
Field of research	Bioprocess Engineering		
Opponent(s)	Professor Riitta Lahesmaa, MD, PhD		
Supervisor	Professor Matti Leisola, PhD		
(Instructors)	Academy Professor Lauri A. Aaltonen, MD, PhD and Diego Arango, PhD		
Abstract DNA microarray technology has in a decade been rapidly adopted by biomedical researchers and emerged as a very prominent research tool. In this study, microarray technology, together with supporting methods, was utilized in studies of human cancer. The study focused on two types of cancer, a hereditary syndrome called Hereditary Leiomyomatosis and Renal Cell Cancer (HLRCC) and on colorectal cancer (CRC). HLRCC is a disease caused by mutations in the Krebs cycle gene fumarase, where some of the patients develop an aggressive and early-onset renal cancer or uterine leiomyosarcoma. CRC is one of the leading causes of death in the Western world. In the first study, yeast models with fumarase mutations were subjected to microarray profiling and functional experiments to reveal changes caused by two different fumarase mutations and to find potential candidate genes for the renal cancer observed in some of the HLRCC patients. No significant differences in fumarase gene or protein expressions or in enzyme activities were observed. This indicated that modifying genes, rather than genotype-phenotype effects, play a role in the formation of the malign tumors. In the second study, Dukes' C stage colorectal tumors with good and bad prognosis were studied using microarray profiling, and a molecular signature separating these two groups with differing prognoses was identified. The study showed that gene expression profiling of surgical samples can predict the recurrence of Dukes' C patients. In the third study, serrated colorectal carcinomas, which differ morphologically from conventional colorectal carcinomas, were distinguished from each other using expression microarrays. The separation by unsupervised clustering indicated that serrated tumors differ biologically from conventional ones. Statistical analyses were used to identify key genes with differential expression between these two tumor types and the results were further validated by immunohistochemical analyses. A key gene, <i>EPHB2</i> , revealed by the expression data analysis of serrated CRC, was further characterized in the last two studies to find out more about the relevance of this gene to colorectal tumorigenesis. Germline mutations in <i>EPHB2</i> were found in few CRC patients, but did not appear to be a major contributor in CRC susceptibility. Aberrant promoter hypermethylation and frameshift mutations in a repetitive track of the gene were, however, found to be frequent mechanisms of <i>EPHB2</i> inactivation in CRC. In general, it was observed that the use of combined research methods greatly enhance the power of microarray studies, and enable focusing of the analyses. Although the technology is presently used primarily in basic research, clinical applications are foreseeable and slowly emerging.			
Keywords DNA microarray technology, cancer, HLRCC, CRC, <i>EPHB2</i>			
ISBN (printed)	951-22-8337-9	ISSN (printed)	1795-2239
ISBN (pdf)	951-22-8338-7	ISSN (pdf)	1795-4584
ISBN (others)		Number of pages	73 p. + app. 35 p.
Publisher Yliopistopaino			
Print distribution Helsinki University of Technology, Laboratory of Bioprocess Engineering			
<input checked="" type="checkbox"/> The dissertation can be read at http://lib.tkk.fi/Diss/2006/isbn9512283387/			



TEKNILLINEN KORKEAKOULU PL 1000, 02015 TKK http://www.tkk.fi		VÄITÖSKIRJAN TIIVISTELMÄ	
Tekijä Antti Kokko			
Väitöskirjan nimi Expression microarray technology as a tool in cancer research			
Käsikirjoituksen jättämispäivämäärä 06.06.2006		Väitöstilaisuuden ajankohta 29.09.2006	
<input type="checkbox"/> Monografia		<input checked="" type="checkbox"/> Yhdistelmäväitöskirja (yhteenveto + erillisartikkelit)	
Osasto	Kemian tekniikan osasto		
Laboratorio	Bioprosessiteknikan laboratorio		
Tutkimusala	Bioprosessiteknikka		
Vastaväittäjä(t)	Professori Riitta Lahesmaa, LT		
Työn valvoja	Professori Matti Leisola, TkT		
(Työn ohjaajat)	Akatemiaprofessori Lauri A. Aaltonen, LT ja Diego Arango, PhD		
Tiivistelmä DNA-mikrosiruteknologia on vuosikymmenen kuluessa omaksuttu nopeasti osaksi biolääketieteellistä tutkimusta ja noussut lupaavaksi menetelmäksi syöpätutkimuksessa. Tässä työssä tutkittiin mikrosiruteknologian hyväksikäyttöä kahdella syöpätyypillä, periytyvällä HLRCC -syndroomalla sekä kolorektaalisyövällä (CRC). HLRCC aiheutuu muutoksista Krebsin syklin fumaraasi geenissä. Osalla potilaista ilmenee kohdun leiomyosarkooma tai aggressiivinen ja nuorella iällä todettu munuaissyöpä, jonka epäillään johtuvan vielä tuntemattoman geenin vaikutuksesta. CRC on yksi yleisimmistä syöpätyypeistä ja johtavista kuolinsyistä länsimaissa. Työn ensimmäisessä osassa tutkittiin kahden eri fumaraasimutaation vaikutusta hiivamallissa mikrosiruanalyysien ja funktionaalisten kokeiden avulla. Eri mutaatioiden todettiin olevan vaikutukseltaan keskenään samanlaisia, viitaten mahdollisten muiden geenien vaikutukseen munuaissyövän synnyssä. Toisessa osajulkaisussa vertailtiin mikrosiruteknologialla keskenään hyvän ja huonon ennusteen saaneita Dukes' C -luokituksen omaavia kolorektaalikasvaimia. Nämä kaksi ryhmää kyettiin molekyyliogeneettisten erojensa perusteella erottamaan toisistaan, mikä osoitti leikkausnäytteiden perusteella tehtävän geeniekspressio-profiloinnin olevan näissä syöväissä mahdollista ja kykenevän ennustamaan tämän syöpätyypin uusiutumista potilailla. Kolmannessa osajulkaisussa mikrosiruteknologiaa käytettiin hyväksi luokiteltaessa CRC:n sahalaitaista muotoa molekyyliogeneettisen profiilinsa perusteella tavanomaisesta CRC:stä. Geneettinen profilointi jakoi nämä kaksi ryhmää omiksi alatyypeikseen, viitaten näiden rakenteellisesti tavanomaisesta CRC:stä poikkeavien sahalaitaisten kasvainten erilaiseen biologiseen taustaan. Tilastollisten analyysien perusteella sahalaitaisissa CRC:ssä erilaisimmin ilmeneviä genejä valittiin immunohistokemiallisiin jatkotutkimuksiin, joiden avulla todennettiin mikrosiruanalyysin löydökset. Viimeisissä osajulkaisuissa tutkittiin mikrosiruanalyysien perusteella esiin tulleen <i>EPHB2</i> geenin vaientamisen mekanismeja sekä merkitystä CRC:n muodostumisen alttiudessa. Geenin ituratamutoksia löydettiin muutamasta tutkitusta CRC näytteestä, mutta niiden vaikutus CRC:n alttiuteen katsottiin vähäiseksi. Promootorialueen hypermetylaation sekä geenin lukualueella sijaitsevan toistojakson muutosten aiheuttaman geenin vaientamisen sen sijaan havaittiin olevan yleistä. Yleisellä tasolla mikrosiruanalyysien havaittiin hyötyvän samanaikaisesti tehtävistä muista tutkimusmenetelmistä, joiden avulla tutkimuksen kohteita kyettiin rajaamaan. Vaikka teknologiaa nykyisellään sovelletaan lähinnä perustutkimukseen, on lupauksia kliinisistä käyttökohteista nähtävissä.			
Asiasanat	DNA-mikrosiruteknologia, syöpä, HLRCC, CRC, EPHB2		
ISBN (painettu)	951-22-8337-9	ISSN (painettu)	1795-2239
ISBN (pdf)	951-22-8338-7	ISSN (pdf)	1795-4584
ISBN (muut)		Sivumäärä	73 s. + liit. 35 s.
Julkaisija	Yliopistopaino		
Painetun väitöskirjan jakelu	Teknillinen korkeakoulu, Bioprosessiteknikan laboratorio		
<input checked="" type="checkbox"/> Luettavissa verkossa osoitteessa http://lib.tkk.fi/Diss/2006/isbn9512283387/			

Preface

This work was carried out at the Department of Medical Genetics, at Haartman Institute and Biomedicum Helsinki, University of Helsinki during 2003-2006. The project was initially started in 2002, while I was doing my Master's thesis at the Governmental Research Center in Espoo, Otaniemi (VTT Biotekniikka) in collaboration with the Human Tumorigenesis Group at the University of Helsinki.

First, I wish to express my gratitude to my supervisor, professor Lauri Aaltonen, for providing me the opportunity to work and conduct my PhD studies in his research group. Lauri is a great person and an excellent leader; motivating, inspiring, and understanding, and possesses a good sense of humor that lightens up the day. I am deeply grateful for these years that I have been working with him.

I also wish to express my deep gratitude to Diego Arango, a dear friend and my other supervisor. In the early stages of my studies, Diego was always there for me when I needed advice – either scientific or personal. Muchas gracias, “*perro*”.

I would like to extend my thanks to professor Matti Leisola, who supervised my studies at the University of Technology, for the discussions and advice given during all these years.

I sincerely wish to thank Jaakko Hollmen and Eija Hyytinen for reviewing the thesis and providing excellent comments to improve it.

I am deeply grateful to Jussi Jääntti and Sirkka Keränen for the supervision given during my first article and for initiating me into the yeast studies.

I would like to express my gratitude to Päivi Laiho, a former member of the Aaltonen group, who initiated me into the field of microarray research and provided multivarious help during my projects.

I would like to thank all the collaborators and co-authors in my publications, Sampsa Hautaniemi, Jarno Tuimala, Paula Salmikangas, Maarit Takatalo, Johanna Schleutker, Sanna Korja, Jaakko Astola, Akseli Hemminki, Simo Schwartz Jr, Hafid Alazzouzi, Veronica Davalos, Charis Eng, Ian Tomlinson, Luis Carvajal-Carmona, Daniel Nicorici, Eloi Espín, John Mariadason, Manel Armengol, Lars Konrad, Andrew Wilson, Imai Kohzoh, Hiroyuki Yamamoto, Enric Domingo, Johannes Gebert and Stefan Woerner. I would especially like to thank Jukka-Pekka Mecklin, Heikki Järvinen, Markus Mäkinen, Tuomo Karttunen and Karoliina Tuppurainen for the invaluable help in providing clinical samples used in these works and for the substantial collaboration provided in the colon cancer projects.

The assistance and services of Outi Monni, Janna Saarela, and the rest of the Biomedicum Biochip Center people, and help given by the Biomedicum Bioinformatics

Unit in microarray related issues and sequencing services provided by the Molecular Medicine Sequencing Laboratory in Biomedicum are gratefully acknowledged.

I wish to thank all the people at the Aaltonen Group, both past and present. A special thanks goes to my room mates Sini, Matti and Sakari for the good atmosphere, lively discussions and all the assistance given in various projects. The evenings spent at work with Sakari, writing our theses and chatting, are still very clear in my mind. Many thanks belong also to the post doc room ladies and lad(s), Rainer for helping me out with dHPLC and data management, Pia V. for all the help in EPHB2 works and microarrays – and the encouragement and help given while writing this thesis ☺, Virpi and Auli for the theoretical and practical advice given, the other students Taru, Heli L, Heli S, Pia A, Marianna, Sanna, Silva, Laura, Edwin, Sari, Anniina, Iina, Mia and Marko for being good friends and helping me in my projects, the lab personell Inga-Lill, Mikko and Päivi for the technical assistance and the former members of the group Tuija, Susa, Maija, Reijo, Nina and Kirsi for all the help given. Thanks guys, for all the great time spent together, both in and out of office, and for NOT publishing all the pictures taken while partying!

I would like to express my deep gratitude to my parents, family (in the large sense, including Mikkeli and Ylöjärvi) and friends for their continuous support, encouragement and understanding over all these years. I love you all, very much. Special thanks goes to the *Wanha Jengi* (in alphabetic order): Hanna, Heikki, Jarno, Kisse, Lauri, Marko, Sanna H, Sanna J, Sari, Pasi, Paula, Peter, Petteri and Virva, for long-term friendship, for all the great time spent together, and especially for the unforgettable theme parties.

Finally, my deepest gratitude belongs to my beloved wife Mikaela, for making me understand what really is important in life. In addition to fishing.

The financial support of the Finnish Foundation of Technology (TES), Finnish Cultural Foundation, Finnish Cancer Society, Research and Science Foundation of Farnos, Maud Kuistila Foundation, Ida Montin Foundation and the University of Helsinki Research Funds for Medicine are gratefully acknowledged.

Contents

LIST OF PUBLICATIONS	11
THE AUTHOR'S CONTRIBUTION IN THE APPENDED PUBLICATIONS	12
LIST OF ABBREVIATIONS	13
1 REVIEW OF THE LITERATURE	15
1.1 INTRODUCTION	15
1.2 GENES AND GENE EXPRESSION	15
1.3 EXPRESSION MICROARRAY TECHNOLOGY	17
1.3.1 Introduction.....	17
1.3.2 History of microarrays.....	17
1.3.3 Microarray fabrication	18
1.3.4 Microarray sample preparation and hybridization.....	21
1.4 MICROARRAY DATA ANALYSIS	23
1.4.1 Introduction.....	23
1.4.2 Image analysis and signal adjustment	23
1.4.3 Data normalization	24
1.4.4 Statistical analysis.....	25
1.4.5 Sensitivity and reliability in microarray measurements.....	27
1.5 CANCER	28
1.5.1 Introduction.....	28
1.5.2 Cancer and genes.....	29
1.5.3 Hereditary Leiomyomatosis and Renal Cell Cancer.....	30
1.5.4 Colorectal cancer.....	30
1.6 DNA MICROARRAY TECHNOLOGY AND CANCER	32
2 AIMS OF THE STUDY	35
3 MATERIALS AND METHODS.....	36
3.1 YEAST STRAINS (I).....	36
3.2 CLINICAL SAMPLES, HEALTHY CONTROLS AND CELL LINES (II-V)	36
3.3 MICROARRAY PREPARATION AND HYBRIDIZATION	37
3.3.1 YG-S98 expression microarrays (I).....	37
3.3.2 HG-U133A expression microarrays (II, III)	37
3.4 EXPRESSION MICROARRAY DATA ANALYSIS	37
3.4.1 Statistical analysis of the YG-S98 arrays (I).....	37
3.4.2 Statistical analysis of the HG-U133A arrays (II, III).....	38
3.4.3 Functional gene enrichment analyses (I, II, III).....	38
3.4.4 Unsupervised hierarchical clustering (II, III).....	39
3.4.5 Class prediction analyses (II, III)	39
3.5 SUPPORTING METHODS	40
3.5.1 Functional studies of the yeast strains (I)	40
3.5.2 Immunohistochemistry and tissue microarray (II, III).....	40
3.5.3 Quantitative RT-PCR (II).....	40
3.5.4 Mutation analyses and allelic imbalance (II, III, IV, V)	41
3.5.5 Methylation analysis and western blotting (III, V).....	41
4 RESULTS AND DISCUSSION.....	42
4.1 MODELING TUMOR PREDISPOSING <i>FH</i> MUTATIONS IN YEAST (I).....	42
4.1.1 Expression microarray profiling of the yeast transcriptome.....	42
4.1.2 Effect of fumarase alterations to the Krebs cycle gene expression levels	42
4.1.3 Functional studies	43
4.1.4 Evidence of a modifier effect in HLRCC.....	43
4.2 PREDICTION OF RECURRENCE IN DUKES' C COLORECTAL CANCER (II).....	45
4.2.1 Differentially expressed genes and functional group enrichment analysis	46
4.2.2 Molecular signatures for good or bad prognosis.....	46

4.2.3	<i>RHOA as a prognostic marker in Dukes' C colorectal cancer</i>	47
4.2.4	<i>Microarray profiling of colorectal carcinomas</i>	47
4.3	MOLECULAR CLASSIFICATION OF SERRATED COLORECTAL CANCER (III)	48
4.3.1	<i>Gene expression profiling and class prediction of the serrated CRC</i>	49
4.3.2	<i>Validation of the expression array results</i>	49
4.3.3	<i>Serrated colorectal tumors differ from conventional adenocarcinomas</i>	50
4.4	EPHB2 AND COLORECTAL TUMORIGENESIS (IV, V)	52
4.4.1	<i>EPHB2 germline variants in colorectal tumorigenesis (IV)</i>	52
4.4.2	<i>Mechanisms of EPHB2 inactivation in colorectal tumors (V)</i>	53
4.4.3	<i>EPHB2 in colorectal tumorigenesis</i>	53
4.5	FUTURE PROSPECTS OF ARRAY TECHNOLOGIES	54
5	CONCLUSIONS	55
6	REFERENCES	57

List of Publications

This thesis is based on the following five publications, which are, throughout the text, referred to as their Roman numerals.

- I. Kokko A, Ylisaukko-oja SK, Kiuru M, Takatalo MS, Salmikangas P, Tuimala J, Arango D, Karhu A, Aaltonen LA, Jääntti J. Modeling tumor predisposing FH mutations in yeast; effects on fumarase activity, growth phenotype and gene expression profile. *International Journal of Cancer*. 2006;118:1340-1345.
- II. Arango D, Laiho P, Kokko A, Alhopuro P, Sammalkorpi H, Salovaara R, Nicorici D, Hautaniemi S, Alazzouzi H, Mecklin J-P, Järvinen H, Hemminki A, Astola J, Schwartz Jr. S, Aaltonen LA. Gene expression profiling predicts recurrence in Dukes' C colorectal cancer. *Gastroenterology*. 2005;129:874-884.
- III. Laiho P*, Kokko A*, Vanharanta S, Salovaara R, Sammalkorpi H, Järvinen H, Mecklin J-P, Karttunen TJ, Tuppurainen K, Davalos V, Schwartz Jr. S, Arango D, Mäkinen MJ, Aaltonen LA. Serrated carcinomas form a subclass of colorectal cancer with distinct molecular basis. *Oncogene*. 2006 Jul 3; [Epub ahead of print].
- IV. Kokko A, Laiho P, Lehtonen R, Korja S, Carvajal-Carmona LG, Järvinen H, Mecklin J-P, Eng C, Schleutker J, Tomlinson IPM, Vahteristo P, Aaltonen LA. EPHB2 germline variants in patients with colorectal cancer or hyperplastic polyposis. *BMC Cancer*. 2006; 6:145.
- V. Alazzouzi H, Davalos V, Kokko A, Domingo E, Woerner SM, Wilson AJ, Konrad L, Laiho P, Espín E, Armengol M, Kohzoh I, Yamamoto H, Mariadason JM, Gebert JF, Aaltonen LA, Schwartz Jr. S, Arango D. Mechanisms of inactivation of the receptor tyrosine kinase EPHB2 in colorectal tumors. *Cancer Research*. 2005;65:10170-10173.

* Equal contribution

The Author's contribution in the appended publications

Publication I: Most of the experimental work and all of the data analysis was done by Antti Kokko. The manuscript was prepared jointly by Auli Karhu, Lauri Aaltonen, Jussi Jääntti and Antti Kokko, who wrote the first version of the manuscript.

Publication II: Antti Kokko assisted in designing the experimental methods, carried out the experimental microarray work together with Päivi Laiho and Diego Arango, performed the tissue microarray analyses together with Diego Arango and assisted in the preparation of the manuscript.

Publication III: Antti Kokko was in charge of the data analysis of the expression microarray experiment and carried out the direct sequencing of the *EPHB2* gene. The manuscript was prepared jointly by Antti Kokko, Päivi Laiho, Markus Mäkinen, Sakari Vanharanta and Lauri Aaltonen. The authors have agreed that both Antti Kokko and Päivi Laiho will use this publication in their doctoral dissertations.

Publication IV: Majority of the experimental work and interpretation of the results was carried out by Antti Kokko, with other authors assisting as described in the publication. The manuscript was prepared by Antti Kokko, Pia Vahteristo and Lauri Aaltonen.

Publication V: Antti Kokko carried out the mutation analysis and interpretation of the results in the Finnish samples in the study and assisted in the preparation of the manuscript. Veronica Davalos conducted the methylation studies for the article. The authors have agreed that both Antti Kokko and Veronica Davalos may use this publication in their doctoral dissertations.

List of abbreviations

3'	3' end (of a DNA or RNA chain)
5'	5' end (of a DNA or RNA chain)
5-FU	5-Fluorouracil
ALL	acute lymphoid leukaemia
AML	acute myeloid leukemia
ANOVA	analysis of variance
APC	adenomatosis polyposis coli
BIOB	E. coli BioB gene biotin synthetase
BIOC	E. coli bioC gene
BIODN	E. coli bioD gene dethiobiotin synthetase
BRAF	v-raf murine sarcoma viral oncogene homolog B1
CCNT2	cyclin T2
CGH	comparative genomic hybridization
CRC	colorectal cancer
CREX	bacteriophage P1 cre recombinase
dCHIP	DNA chip (software)
DHPLC	denaturing high-performance liquid chromatography
DNA	deoxyribonucleic acid
DTT	dithiotreitol
EPHB2	ephrin receptor B2
FAP	familial adenomatous polyposis
FDR	false discovery rate
FGEA	functional gene enrichment analysis
FH	fumarate hydratase (fumarase)
FUM1	yeast fumarase
GO	Gene Ontology
HIF1A	hypoxia-inducible factor 1, alpha subunit
HLRCC	hereditary leiomyomatosis and renal cell cancer
HNPCC	hereditary nonpolyposis colorectal cancer
HPP	hyperplastic polyposis
HUGO	the human genome project
IDH1	isocitrate dehydrogenase 1
IHC	immunohistochemistry
KGD	alpha-ketoglutarate dehydrogenase
KNN	K nearest neighbors
K-RAS	Kirsten rat sarcoma 2 viral oncogene homolog
LOH	loss of heterozygosity
LOI	loss of imprinting
LPD	lipoamide dehydrogenase
MAS	micro array suite (software)
MBD4	methyl-CpG binding domain protein 4
MGED	microarray gene expression data society
MIAME	minimal information about microarrays
MM	mismatch (probe)
MMR	mismatch repair

mRNA	messenger RNA
MSI	microsatellite instability
MSS	microsatellite stable
MTA1	metastasis associated 1
NAPD	nicotinamide adenine dinucleotide phosphate
NMD	nonsense-mediated mRNA decay
OCT	optimal cutting temperature compound
PC	prostate cancer
PCA	principal component analysis
PCR	polymerase chain reaction
PJS	Peutz-Jeghers syndrome
PM	perfect match (probe)
PTCH	patched homolog (Drosophila)
RHOA	ras homolog gene family, member A
RMA	robust multichip average (software)
RNA	ribonucleic acid
rRNA	ribosomal RNA
RT-PCR	reverse transcriptase PCR
SAM	significance analysis of microarrays (software)
SAPE	streptavidin-phycoerythrin
SDH	succinate dehydrogenase complex
SNP	single nucleotide polymorphism
SOM	self-organizing map
SVD	singular value decomposition
TCF4	transcription factor 4
TNM	tumor, lymph node, metastasis
TP53	tumor protein p53
tRNA	transfer RNA
TS	training set
VS	validation set
WNT	Wingless-type
WT	wild type
XP	Xeroderma pigmentosum
YPD	yeast peptone dextrose medium

1 Review of the Literature

1.1 Introduction

Biomedical sciences have in the recent years taken great leaps forward in the understanding of human genetics, gene expression regulation and studies of human cancer. The development is in many ways linked to the advances made in the electronic industry and information technology, which have been successfully applied to the biological sciences. This has created new technologies and automated methods, such as microarray technology. Microarrays cover a broad range of applications ranging from genetic research and molecular biology related analysis to pharmacogenomic and forensic purposes. The focus of this dissertation is the expression microarray technology and its use in cancer research, and the following chapters will hence be limited to these subjects.

1.2 Genes and gene expression

The genetic material of most living organisms is known as deoxyribonucleic acid (DNA). DNA is composed of nucleotides containing a sugar-phosphate backbone and attached bases. There are four different bases: adenine (A), cytosine (C), guanine (G) and thymine (T). The bases bind to each other in a complementary manner; A to T and C to G, respectively, using hydrogen bonding. The sugar phosphate backbones of the nucleotides bind covalently to each other and form long chains of nucleotides. A DNA molecule consists of two complementary polynucleotide chains held together by the base-pair binding, hydrophobic effects and pi-stacking and form a right-handed double helix structure, which is typically millions of nucleotides in length. As the nucleotides are joined in an ordered fashion with the 3' hydroxyl group binding to the 5' hydroxyl group of the previous nucleotide, the two strands of the helix become polarized and are in an antiparallel orientation to each other with opposite 5' - 3' directions. The main purpose of DNA in a nucleus is to store information needed by the cell to function.[1-3]

Another information-containing material that the cells utilize is ribonucleic acid (RNA), which is analogous to DNA but differs from it in several important ways. In RNA, the backbone sugar is ribose and uracil (U) has replaced T in the bases, and is now complementary to A. Unlike DNA, most RNA molecules are single-stranded and only 75-50000 nucleotides in length. Cells contain several types of RNA, of which messenger RNA (mRNA), transfer RNA (tRNA) and ribosomal RNA (rRNA) are involved in the protein synthesis.

Functional and physical units of heredity passed from parent to offspring are called genes. They are fragments of DNA, encompassing coding DNA sequences and introns, which in most cases contain the information for making a specific protein. Proteins are biologically active molecules produced by the cell and are responsible for the organized functioning of the cell. In essence, they control the physical traits that an organism has, such as eye, skin or hair color. Proteins are composed of sequences of twenty different

types of amino acids that, when linked together, are called polypeptides. Each amino acid is encoded by a sequence of three bases in the DNA called codons. Although there are 64 possible triplets, some redundancy exists and multiple codons code for the same amino acid. In addition, three codons have not been assigned to any amino acids but serve as termination signals for the synthesis.

The process of synthesizing proteins is a two-step procedure that consists of transcription and translation, two processes collectively known as the central dogma of molecular biology. In the transcription step the information of the double-stranded DNA sequence is transferred into a single-stranded mRNA sequence. This is done by RNA polymerase, an enzyme that moves along one strand on DNA in 5'-3' direction and produces a complementary sequence of mRNA. The DNA which is used to produce the RNA is called the antisense strand, while the complementary DNA is named the sense strand. The produced immature mRNA strand is complementary to the antisense strand, and identical to the sense strand, with the exception of base T being changed to U. The transcription of a gene begins at regions of the DNA sequence known as promoter sites and ends at regions known as terminator sites.

Translation, the second stage of the protein synthesis, is the process of constructing an amino acid sequence from a mRNA molecule. This is done by using ribosomes in the cytoplasm consisting of rRNA and proteins. In translation, the ribosomes use tRNA to attach to the mRNA and translate the bases into amino acids. tRNA molecules bring the specified amino acids to the translation site, where the ribosome links them together, forming an elongating chain of peptides. The translation starts from an AUG codon that marks the start of the peptide chain and terminates when a codon representing a stop signal is recognized. The formed polypeptide chain is then released from the translation site and modified into an active protein to serve its purpose in the cell. This process of converting DNA sequence into a protein is called gene expression.

While the central dogma of molecular biology in principle still exists, the genetic and molecular interactions between nucleic acids and proteins have since then been discovered to compose of complex networks with multilateral signalling. Transcriptional and (post-)translational modifications, such as gene splicing, small RNAs and epigenetic methods such as methylation are examples of these processes.[4]

The term genome is used to describe the total genetic information of an organism. Each cell of an organism contain identical genetic information, that is, DNA for every gene is present in all the cells of that organism. The human genome consists of roughly 3 billion basepairs, and the current estimate of the total number of genes varies between 20000 – 25000. While the genetic information of every cell is the same, the mRNA and protein levels vary between the cells. The number of genes expressed in a cell depends on the various environmental conditions the cells possess. In general, only a fraction of the genes (on average several thousands) are expressed simultaneously in any given cell type or tissue. Thus, gene expression studies are used to describe the levels of mRNA molecules produced in a collection of cells at a given time, which in general is proportional to the amount of protein and reflect the ongoing biological and functional processes of the cell. Traditionally gene expression studies have been done one gene at a time using technologies such as reverse-transcription polymerase chain reaction (RT-

PCR) and northern blotting. Recently, the development of microarray technology has revolutionized this concept.

1.3 Expression microarray technology

1.3.1 Introduction

A DNA microarray is a miniature device containing cDNA fragments that are either synthesized directly or spotted onto glass or other matrix. Thousands of genes are represented in a single array, which is designed to simultaneously measure the expression levels of all these genes in a particular tissue or cell type. The basic concept of microarray technology is to hybridize preprocessed sequences of mRNA called *targets* to the complementary sequences called *probes* bound to a solid surface and to quantify the amount of specifically hybridized target, typically by fluorescence detection. Numerous technologies for this purpose have been developed, and these can be categorized based on the chip design, chemistry, manufacturing or signal detection. In general, microarrays are divided into complementary DNA (cDNA) arrays or oligonucleotide arrays based on the used probe material, *in-situ* synthesized or spotted (contact or non-contact) arrays based on the manufacturing method and one-color or two-color arrays based on the staining and detection technique. Other kinds of microarray systems, such as microelectronic arrays which use electrical fields to bind DNA to probes have also been developed,[5,6] but due to the limited scope of this dissertation these are not discussed further. Microarray technologies, applications and analysis methods have been extensively described in many reviews and books.[7-10]

1.3.2 History of microarrays

The roots of microarray technology can be found in the early biochemical “dot blot” experiments done in the 1970s, where DNA was immobilized on membranes and usually probed with radioactively labeled DNA or colorimetric assays.[11,12] The development of fluorescence assay methods in the late 1980s [13] and solid glass surfaces in the early 1990s [14,15] were significant improvements that laid the ground for modern microarrays and enabled faster and more precise analysis of signals. However, the major break-through that changed the whole field of genetic research came when photolithography [16] and printing [17], methods already used in the semiconductor industry, together with miniaturization and automation of processes, were applied to the manufacturing of microarrays. Miniaturization of the spots allowed better sensitivity and more genes to be analyzed from smaller amounts of samples [18]. The miniaturization of the array design also brought a massive parallelism, allowing the analysis of vast amount of genes in a single assay, which was a revolutionary innovation and enabled the analysis of the whole transcriptome (i.e. all the expressed sequences in the human genome) in a single experiment.

The history of modern solid-surface microarrays can be tracked down to 1995, when the first DNA microarrays with 45 cDNA probes were introduced by Schena *et al.*[17] The technological progress of the cDNA microarrays was extremely rapid; in 1996

publications with 1000 probes arrayed were already presented.[19-21] While impressive at the time, the pioneering company in the field has been Affymetrix (Santa Clara, CA, USA), which developed a novel technology of *in situ* hybridized oligonucleotide arrays based on photolithography combined with DNA-synthetic chemistry. [16,22,23] The use of light-directed synthesis to bind modified nucleotides to the chip surface enabled the manufacture of high-density oligonucleotide microarrays, which in 1996 contained already 135000 probes. Currently the GeneChip® (Affymetrix) expression microarrays are capable of containing millions of probes on a 1.28 cm² surface of quartz, and mass production of whole genome chips is available for 16 organisms, including human, mouse, rat, dog, yeast, *E. coli*, *C. elegans* and *A. thaliana*, among others. The rapid development of various chip platforms is connected to the general development in bioinformatics, as automated sequencing methods and public databanks have made the annotated total genome sequences available for microarray probe designing and selection.

1.3.3 Microarray fabrication

The two most commonly used microarray systems have been the cDNA and oligonucleotide arrays, which differ in the used probe materials. The cDNA array probes are usually products of the polymerase chain reaction (PCR), generated from cDNA libraries or clone collections. The probes are printed on glass slides or nylon membranes as spots at defined locations, typically 100-300 µm in size and roughly equal distance apart from each other. However, due to the rapid completing of the human genome (HUGO) project, sequence information alone has become sufficient to generate the DNA to be arrayed and subsequently oligonucleotide arrays have replaced the cDNA arrays to a great extent. While the short oligonucleotide probes are prone to less specific hybridization and reduced sensitivity, they offer the design of probes that represent the most unique part of a given transcript, thus making the detection of closely related genes or splice variants possible. Affymetrix oligonucleotide microarrays, sometimes others as well, are referred to as high-density microarrays, reflecting the massive amount of probes they can contain in comparison to traditional cDNA arrays. The following chapters focus on the widely-used oligonucleotide arrays that can be manufactured by various methods, of which the most important are the *in situ* synthesis method for high-density oligonucleotide arrays (*Figure 1*) used by Affymetrix and Agilent Technologies (Palo Alto, CA, USA) and the contact and non-contact (ink-jet) printing methods of presynthesized oligonucleotide probes used by many academic groups and commercial vendors.

The multiple microarray systems available have been reported to show divergence in measured gene expression levels across the different platforms.[24,25] While varying degrees of correlation have been measured, the commercial platforms that have been compared appear to be moderately standardized and have showed reasonable correlations, at least when noise reducing strategies, such as filtering of low-level expressing genes, have been used.[24,26] The gene expression data correlation between custom-made and commercial platforms has been reported to be slightly lower, though.[24]

1.3.3.1 *In situ* synthesized high-density oligonucleotide arrays

A photolithographic procedure to fabricate microarrays is used by Affymetrix for the GeneChip® microarrays, where a series of masks are used to synthesize 25-mer oligonucleotide probes onto a silicon wafer in such a manner that a large number of different sequences can be produced in parallel in a small number of steps.[16,22,23,27] The chip surface is first covalently modified with a silane reagent to provide hydroxyalkyl groups, which serve as the initial synthesis sites. The surface is then coated with a linker, which is a light-sensitive chemical compound that prevents coupling between the wafer and the first nucleotide of the DNA probe being created. In the next step, lithographic masks are used to direct light onto specific locations at the wafer surface to remove the protecting group from the exposed locations. The surface is subsequently flooded with a solution containing a modified base (A, T, C, or G) and coupling occurs only in those regions on the glass that have been deprotected through the illumination. The monomers are also protected at their 5' positions with a photolabile group, so the cycle can be repeated. In this way, by using repeated cycles of photodeprotection and nucleotide coupling, a desired length of any given oligonucleotide sequences can be built on the microarray. In practise, the probe lengths are limited to 25 bases, as the yield of full length probes drops rapidly as the sequence is extended.

The probes used for microarray synthesis are examined for specificity, potential for cross hybridization and predicted binding properties. To match the properties of the sample amplification procedure, the probes are 3' biased, but typically widely spaced along the sequence. The GeneChip® contains two types of probes; *Perfect Match* (PM) and *Mismatch* (MM) probes. The PM is exactly complementary to the sequence of interest. The MM is identical to the PM except at the central 13th base position, which differs from the PM probe. In theory, the MM probe can be used to quantify and remove non-specific hybridization. The GeneChip® typically uses 11 to 20 probes to interrogate a given gene. This collection of probes is called a *probeset*, and is used in downstream analyses by Affymetrix to give a representative value for a gene expression. The *in situ* synthesis is a very powerful method; the process can achieve extremely high spot densities (spot size of 5 µm in 2005) and the probe sequence can be chosen more or less randomly for each synthesis.

A competing method for the photolithography is the *in situ* synthesis of DNA microarrays by industrial inkjet printing process from digital sequence files. This technique has been adopted by Agilent Technologies (Palo Alto, CA, USA), and has the advantage of much longer probe sequences (60-100 bp) compared to the GeneChip®. The Agilent SurePrint Technology [28] is based on solid-phase phosphoramidite chemistry,[29] where the reactive sites on the nucleotides are blocked with chemical groups (dimethoxytrityl) that can be removed selectively. After the first base is printed, the trityl group that protects the 5' hydroxyl group on the nucleotide is removed and oxidized to activate it, enabling it to react with the 3' group on the next nucleotide. In between each step, the excess reagents are washed away to prevent random reactions later in the synthesis. The process of printing a nucleotide followed by de-tritylation, oxidation and washing is repeated 60 times. While the longer probe sequences guarantee better accuracy and precision for target hybridization, the spot

densities, however, are much lower (100-150 μm in size, at 100-200 μm center) than in light directed synthesis, mainly because of the natural limitation of liquid handling.

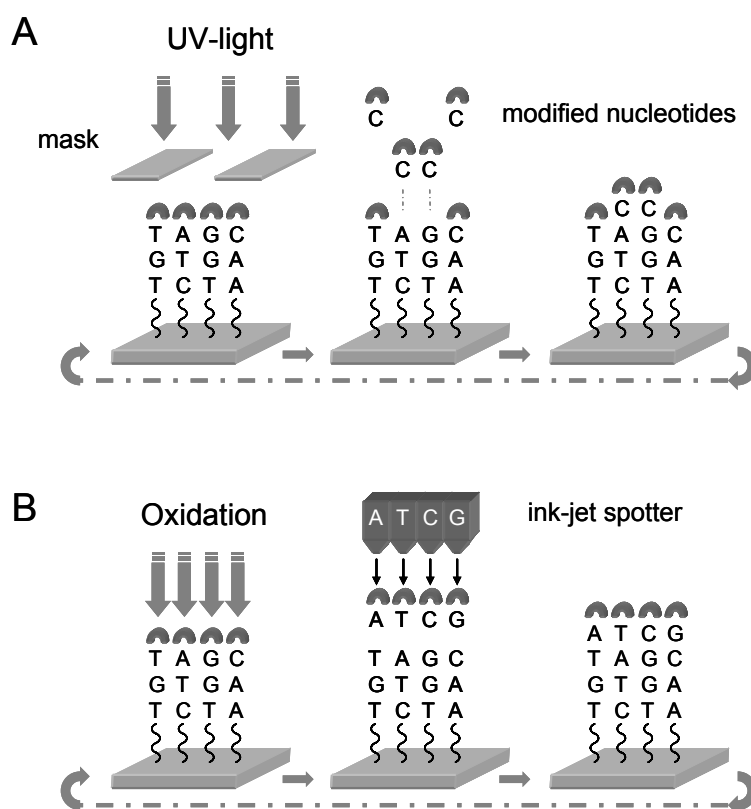


Figure 1. Schematic view of the typical “in situ” methods for microarray manufacturing. In the method used by Affymetrix (A) the probes are synthesized in repeated cycles using modified nucleotides and ultraviolet light. Photolithographic masks are used to selectively de-protect probes at given locations. In the method used by Agilent Technologies (B) the probes are synthesized using oxidative chemistry and ink-jet printing, where the nucleotides are fired from chambers containing a given nucleotide.

1.3.3.2 Spotted microarrays

The alternative for *in situ* synthesis is the use of presynthesized oligonucleotides that can be printed onto coated glass slides either with contact pins or by non-contact method using ink-jet technology, as mentioned earlier (Figure 2). These spotted microarrays can be custom made but are nowadays in increasing amounts commercially manufactured by a number of vendors. Microarray fabrication using contact printing is based on computer controlled robotic arms linked with a head of high definition pins or capillary devices.[17] The pins pick up small drops of the probe solution from multiwell plates and carry them to the microarray surface. Upon contact with the substrate the small amounts of probe solution are released and deposited to the surface. In non-contact printing, small dispensing devices mounted on robotic arms use ink-jet, bubble-jet or piezo-electric propulsion [29-33] to transfer the oligonucleotides to the microarray surface. The ink-jetting technique does not require direct surface contact,

but operates on a “drop-on-demand” basis, firing small droplets of probe solution from miniature nozzles to precise locations on the substrate surface.

Many factors, such as immobilization chemistry, spotting buffer, probe concentration and physical factors like spotter type, environmental conditions and the utilized pins influence the fabrication of DNA microarrays and have to be accounted for in the array design. Thus, whether using *in situ* synthesized or spotted arrays, the choice between platforms is an important decision, as this will affect the downstream operations of microarray sample processing and data analysis.

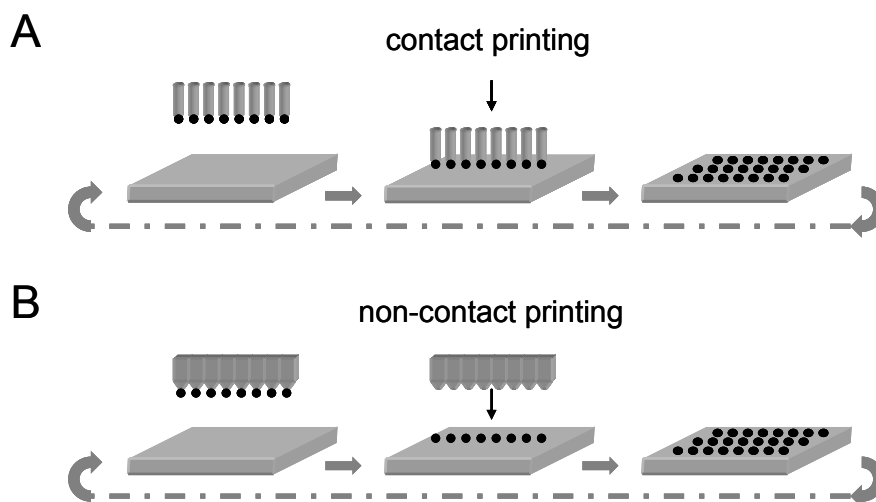


Figure 2. Schematic view of the fabrication of spotted microarrays, where presynthesized oligonucleotides are used. In contact printing (A), high-definition pins deposit small volumes of probe solution directly on the array surface. In non-contact printing (B), the probe is delivered by propelling a small droplet of the solution from a miniature nozzle into the array surface. By repeating the cycles, large amounts of oligonucleotides can be printed with high definition.

1.3.4 Microarray sample preparation and hybridization

In microarray analysis, the mRNA expression levels in a sample are typically detected using a fluorescence detection system. This method relies on prepared RNA that has been tagged with a fluorescence label and can be detected with a scanning device. An important difference lies in the target preparation between the *in situ* synthesized oligonucleotide arrays (Affymetrix) and the spotted microarrays. In both cases, total RNA (or mRNA) is isolated from the source tissue or cells and converted to cDNA, labeled with a fluorescent dye, hybridized to the probes on the microarray and detected by phospho-imaging or fluorescent scanning. However, the high reproducibility of Affymetrix system allows accurate comparison of signals generated by samples hybridized to separate arrays using only one fluorescent dye (one-color array), while in the case of spotted arrays two different fluorescent dyes (such as Cy3 and Cy5) are used (two-color array). This procedure is schematically presented in Figure 3.

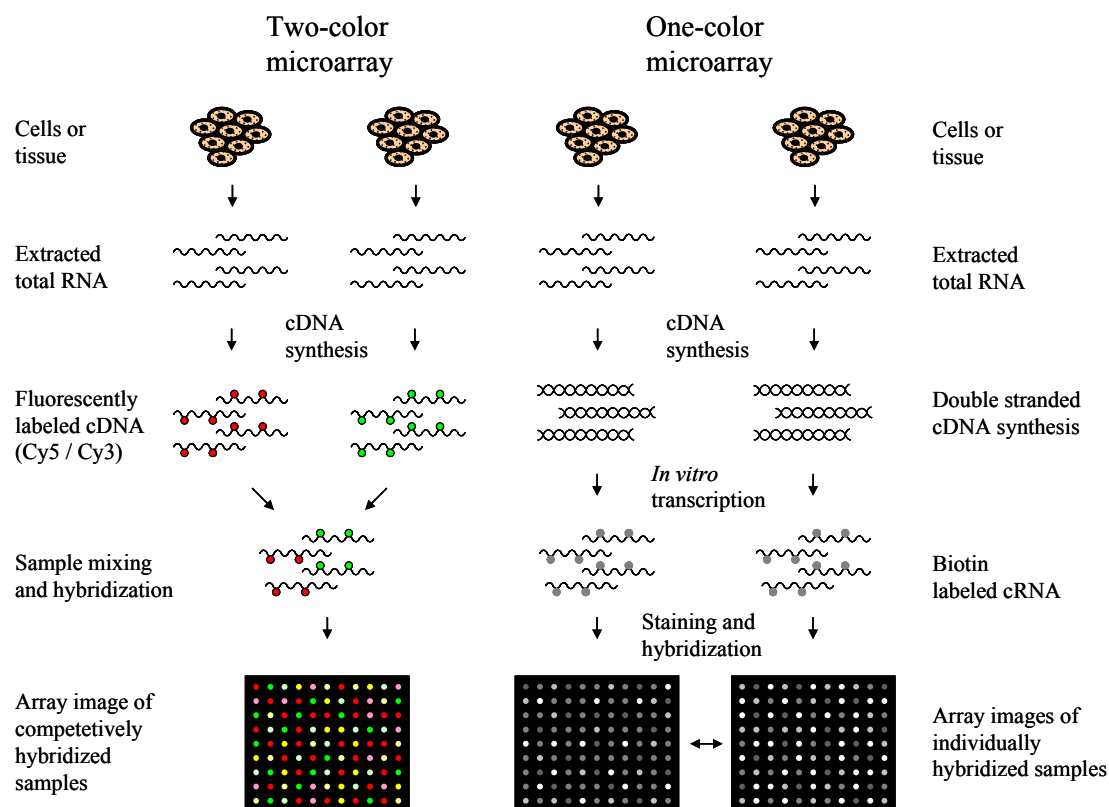


Figure 3. Overview of target preparation for spotted two-color arrays (cDNA and presynthesized oligonucleotide microarrays) and in-situ hybridized one-color arrays (Affymetrix).

In the case of spotted microarrays, mRNA from two different cell populations or tissues (such as normal and tumor tissue) is converted to cDNA, which is labeled with different dyes in both samples. The labeled samples are then mixed and hybridized to the same array, which results in competitive binding of the target to the arrayed sequences. After hybridization, the slide is scanned using two different wavelengths, corresponding to the dyes used. Typically microarray scanners are scanning confocal microscopes, where the wavelength emitted by the fluorescent dye is captured in a photomultiplier tube. Hence, for cDNA microarrays, two fluorescence images of the chip are captured, which are merged to produce a composite image of the microarray. The image contains the measurements of the transcript levels ratios for each gene represented on the array.

In the one-color target preparation process used by Affymetrix,[34] a reverse transcription procedure is used to produce double-stranded cDNA, which is *in vitro* transcribed and amplified to biotin-labeled cRNA. The biotinylated cRNA is then fragmented and hybridized on the chip. Following the hybridization, non-hybridized cRNA is removed from the array, and the chip is subjected to a series of washing and staining steps, where the fluorescent streptavidin-phycoerythrin (SAPE) dye binds with the biotin labeling of the cRNA. Finally, the array is scanned using a laser light, which excites the fluorescent staining agent. The result of a GeneChip® microarray scanning

is an image file that contains the recorded intensity values of each spot on the entire array, representing the emitted signal intensities that are relative to the amount of hybridized cRNA on the chip.

1.4 Microarray data analysis

1.4.1 Introduction

The data analysis of a microarray experiment is a multi-step process and always dependent on the chosen experimental design. The complex process of sample preparation and hybridization in microarray analysis produces image files containing probe signal intensities, which require image analysis, signal adjustment and data normalization to correct for the non-biological variability (or noise) inherent in the system. These procedures are often joined together and referred to as “data normalization”. Due to the multiple microarray platforms with different target preparation and hybridization methods, numerous alternatives to implement these corrections have been developed. After preprocessing, the normalized gene expression data can be analyzed using statistical tools and exploratory methods to extract genes or patterns with biological significance. The analysis of one-color and two color microarray data is more or less similar, the main difference between these methods lies in the normalization of the data.

1.4.2 Image analysis and signal adjustment

The image analysis of microarrays consists of a semi-automated procedure, where a computer-aligned grid is placed over the hybridized surface area. An image analysis software is then used to calculate the intensity of each spot or probe on the array and to store these measurements as numeric values into a text file. In the image analysis, the chip surface is also inspected for spatial hybridization biases and poorly hybridized spots to evaluate the hybridization quality and to eliminate these “bad spots” from further analyses.

The term signal adjustment, also referred to as background correction, describes a wide variety of methods. This step is performed for mainly three reasons. First, to correct for the background noise and processing effect induced during the array hybridization. Secondly, to adjust for cross hybridization caused by the binding of non-specific target (DNA or RNA) to the array. Thirdly, to adjust for expression estimates so that they fall on the proper scale, or are linearly related to concentration. On spotted arrays the pixels surrounding the spot can be used to compute the background adjustment, whereas for Affymetrix GeneChips® the probe intensities are very densely spaced on the array and the probes themselves must be used to determine any adjustment required. For this purpose, Affymetrix uses a statistical algorithm called the Ideal Mismatch procedure, where the adjusted signal is calculated using the PM and MM probe intensities.[35]

1.4.3 Data normalization

When dealing with experiments involving multiple arrays, inter-chip variation due to non-biological factors such as dye effects, sample or scanner differences, unequal quantities of starting RNA, etc. always exists. Normalization is a method that attempts to remove some of this variation. The idea behind normalization methods is that when comparing two or more samples on a genome-wide level, a vast majority of the gene expressions should remain unchanged and the data follow normal distribution. Thus, the expected mean intensity ratio between two channels (two-color data) or chips (one-color data) should be one. If deviation is observed, the data is mathematically processed to adjust this ratio to one. Numerous algorithms and applications have been developed for data normalization, but a standard method for microarray data normalization has not been defined. Typical steps in microarray data normalization include a data transformation procedure to stabilize the variance of the intensity values across the dataset and to make the distribution more symmetric, a procedure to remove non-biological variation within a single array, and a procedure to correct for non-biological variation between different arrays in a dataset. In addition, genes that are believed to be constantly expressed across a variety of conditions, often called house-keeping genes, or external control sequences from another organism, referred to as spike-in controls, can be used for normalization purposes. Biological and technical sample replicates are also utilized, and have been found useful in diminishing the effect of outlier samples and in enhancing the confidence of the data.[36] Robustness to the data is also achieved by utilizing summarized probe intensity values of multiple probes (one-color arrays) or replicate spots (two-color arrays) for expression measurements. Normalization is usually performed on the whole dataset, but can be based on either behavior of the whole data or only a subset of it. The selected method should depend on the prior knowledge or assumptions that are known regarding the behaviour of a particular dataset.

For two-color microarrays, a constant adjustment, such as scaling or log centering, is often used to force the distribution of the intensity log ratios to have a median of zero for each slide. However, such global normalization approaches are not adequate in situations where dye biases can depend on spot overall intensity or spatial location within the array. In these situations dye-swap or locally weighted linear regression (LOWESS) smoothing are examples of typically used normalization methods. As each experiment has its own unique features and needs for normalization, many different methods for two-color arrays have been developed and are well documented in the literature.[37-42]

For Affymetrix GeneChip® arrays, as mentioned earlier, probesets consisting of summarized probe intensity values are used to define a measure of expression representing the amount of the corresponding mRNA species. Several approaches to normalize expression data created with the GeneChip® system have been proposed, such as the model-based expression index (MBEI) used in dCHIP software,[43,44] the Microarray Suite (MAS) 5.0 statistical algorithm from Affymetrix [35] and the robust multi chip average (RMA) method by Irizarry *et al.* (2003).[45,46] These methods differ from each other considerably, and may lead to different results in data analysis. The MAS 5.0 algorithm uses a linear regression approach, where adjusted PM values

are log transformed and a robust mean is calculated on the resulting values. After obtaining a signal for each probeset as the antilog of the resulting value, the data is scaled using a trimmed mean.[47] In the method used by dCHIP, a chip with median intensity is selected as a baseline array for normalization. The algorithm uses an invariant set method, where a large number of probes are selected *ad-hoc* as references for comparison of two samples and a non-parametric curve (running median) is fitted through the datapoints. The most recently introduced normalization methods are the RMA and GCRMA, a modification of this method,[48] which use quantile normalization to give the same empirical distribution of intensities for each array in the experiment. In the quantile normalization method, the highest background corrected and log-transformed PM intensity on each GeneChip® is determined. These values are averaged, and the individual values replaced by the average. This process is then repeated for all intensities in descending order. Following the quantile normalization, an additive linear model is fit to the normalized data to obtain expression measure for each probe on each GeneChip®. The modified GCRMA method takes also advantage of sequence information to appropriately describe nonspecific background variation.

Due to the wide range of methods and applications available for data production and normalization, quality and comparability of the analysis results of microarray experiments has become a major challenge. For this purpose, guidelines for microarray data reporting and standards for minimum information about microarray experiments (MIAME) [49] have been developed by the Microarray Gene Expression Data Society (MGED). The MIAME standards and deposition of the original datasets to a public database are nowadays by many journals considered as a prerequisite for publication of microarray data.

1.4.4 Statistical analysis

Microarrays can be used to investigate problems in cell biology in various ways, with a range of differential approaches. At the other end, the main interest lies in finding a single change in gene expression that might be a key to a given alteration in phenotype. At the other extreme, the aim is to look at overall patterns of gene expression in order to understand the architecture of genetic regulatory networks. The basic idea behind the statistical analysis is to characterize the structure of the experimental data and extract statistically significant patterns from it. Because of the complexity of the problems and datasets generated by microarray experiments, the use of data analysis software is essential. To date, a large number of commercial and non-commercial software tools for statistical analysis and visualization of gene expression data have been developed, which all offer their own solutions to the problem at hand. GeneSpring (Agilent Technologies), Cluster and Treeview,[50] GeneCluster,[51,52] SAM [53] and dCHIP [43] are examples of these software tools, to name a few.

Methods utilized in the data analysis vary considerably. The analysis of microarray data is explorative by nature, and the components of the analysis depend upon the purpose of the experiment. Tools that are generally used include filters to remove redundant genes from the experiment, statistical tests to find differentially expressed genes and classification methods to discover pathway level expression patterns and find

distinguished expression profile signatures. The composition of a typical microarray data analysis is presented in Figure 4.

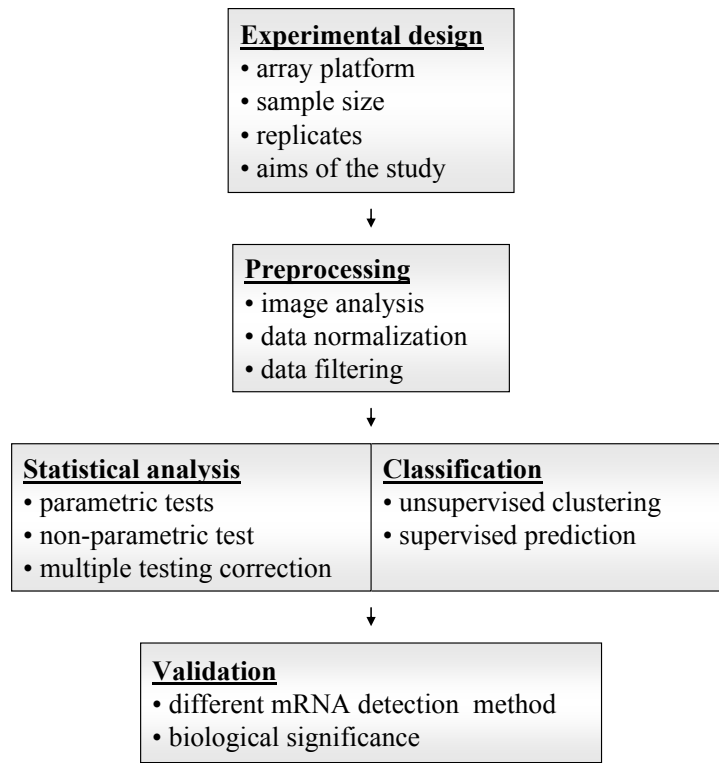


Figure 4. Overview of microarray data analysis for one-color and two-color microarrays. Due to the multiple different microarray platforms and experimental designs no standard protocol exists, and the composition of individual steps vary in each experiment.

The expression of a large number of genes that are irrelevant or unchanged add a high level of noise and uncertainty to the data, which makes the use of classical statistical tests problematic. Reducing the number of genes in the analyses benefits the power of statistical testing. Gene filtering and differentiation approaches can efficiently reduce the dimensionality of the data and help remove redundant genes. This helps to highlight the genes that are truly differential for the investigated trait. Typical tests used in microarray analysis are parametric tests such as Student's t-test or analysis of variance (ANOVA), which assume normal distribution of the data and try to estimate whether the variance in the data comes from the normal distribution or not. In addition, microarray experiments typically have large number of observations but only few samples that leads to testing of multiple hypotheses. As some of the observed differences are expected to happen by chance alone, correction for multiple testing is desired. These adjustments to the statistical tests include corrections such as the Bonferroni method and the false discovery rate (FDR) suggested by Benjamini and Hochberg (1995).[54] Permutation-based models are another approach to validate the results. Other methods for analysis include data transformations such as principal component analysis (PCA) and singular value decomposition (SVD), which reduce the dimensionality of the data and aim to find the major components explaining the

variance in the data. Fold change was among the first methods used to evaluate whether genes were differentially expressed, but is nowadays considered an inadequate test statistic when used alone, as it does not account for the variance and offers no associated level of confidence. The choice of an appropriate correction can be challenging, as many of the popular correction methods, such as the Bonferroni method, have not been designed for microarray data, where there are few cases but many observations per sample. This may lead to very stringent correction and loss of data, with no false positive findings, but also very few true positive findings. Therefore, permutations and FDR based methods with adjustable threshold levels have gained popularity in validation of microarray analyses.

Classification is a widely used analysis method for gene expression data, used either to discover new categories within a dataset or to assign cases to a given category, and is often referred to as clustering. Two principal categories of clustering exist: the unsupervised and supervised methods. In the unsupervised clustering (or class discovery), objects such as genes or samples are grouped into classes based on some sort of similarity metric that is computed for one or more variables. Typically, genes are grouped into classes on the basis of the similarity in their expression profiles across cases, tissues or conditions. Unsupervised clustering can further be split to hierarchical clustering methods, which produce a tree diagram (dendrogram) and non-hierarchical clustering methods such as self-organizing maps (SOM) or K-means clustering,[55,56] which typically divide the cases into a predetermined number of groups in a manner that maximizes a specific function. In the supervised clustering (or class prediction) methods, algorithms are developed to assign objects to predetermined categories. The supervised methods generally involve the use of a training data set and an independent validation data set, and aim to obtain a function or rule that uses expression data to predict whether a case is of one type or another. In cases where the dataset is too small to be effectively split, a cross-validation method such as leave-one-out or class permutation procedure is often used.

1.4.5 Sensitivity and reliability in microarray measurements

While the existence and direction of gene expression changes can be reliably detected for majority of genes in appropriate sensitivity ranges, accurate measurements of absolute expression levels and the reliable detection of low-level abundance genes are presently beyond the reach of microarray technology. [26] The detection limit of current microarray technology is estimated to be between one and ten copies of mRNA per cell, depending on the used microarray platform and target material. [57-59]

Sources of inaccuracy and inconsistencies in microarray measurements and analysis include the probe sequence design, redundant annotation, splice variant effects, folding of the target transcript and cross-hybridization.[26,60] Recently it has been pointed out, that Affymetrix GeneChip® –based expression analyses suffer from probe and probeset annotation problems due to utilization of genomic information, which since the design of the chips has gone through a tremendous progress.[60] Due to the high variability in analysis methods and increased probability of false positive findings in microarray data analysis, data validation is generally regarded as an integral part of the analysis process.

Data validation can be either internal or external. In internal validation, the existing data is typically re-sampled, and probabilities are calculated on permuted sets of the data to evaluate the original findings. In external validation, the results are typically validated using another method on a new sample (or data) set, such as mRNA measurement of a number of selected genes by quantitative reverse-transcriptase PCR (RT-PCR), which is believed to protect against erroneous inferences due to measurement quality problems. Recently, though, this type of validation procedure has been questioned,[36] and it remains to be seen what the requirements for acceptable microarray results evolve to in the future.

1.5 Cancer

1.5.1 Introduction

Tumor is a mass of abnormally proliferating (neoplastic) cells, which have no known purpose in the physiological function of the body. Two general types of tumors exist: benign and malignant. A benign tumor is composed of cells that may continue to grow in number abnormally but, unlike a malignant tumor, will not invade other unrelated tissues or organs of the body. A disease caused by a malignant tumor is referred to as cancer.

Cancer is a disease of genes, where both external and internal factors, such as chemicals, radiation, viruses, hormones, immune conditions, lifestyle and inherited mutations, play a role in the formation of tumorigenic mutations. The causal factors may act together, or in sequence, to initiate or promote carcinogenesis. It is estimated that worldwide there are 24.6 million people alive who have received a diagnosis of cancer in the last five years. Around half of these people live in Europe and North America. Each year over ten million people worldwide are diagnosed with cancer.[61] Cancer is a serious health problem, with annually increasing incidences. In Finland, prostate cancer is the most common cancer by incidence (5252 cases), with breast (3909), colorectal (2486), and lung (2157) cancers close behind (Finnish Cancer Registry, 2004). While the causes of cancer are gradually being understood, the mechanisms of carcinogenesis are still largely a mystery.

Cancer is a group of diseases characterized by uncontrolled growth and spread of abnormal cells. Cancer can arise in different parts of the body. Cancers associated with the digestive system include cancers of the colon and rectum (colorectal cancer, CRC) and cancers of the pancreas, stomach and esophagus. Haematological malignancies (cancers of the blood, plasma and marrow) include leukemia and myelomas and cancers of the urinary system are bladder and kidney cancers. Lymphomas are cancers of the lymphoid system. Thyroid cancer is the most common type of tumour in the endocrine system (hormones), while brain and spinal tumours are found in the central nervous system. Sarcomas are related to the musculoskeletal system and named after the tissue of origin (bone, soft tissue, or connective tissue tumours). Skin cancers include melanoma and basal cell carcinoma, and head and neck cancers include those of the oral and nasal cavities. Some cancers are particular to men and women. Breast,

cervical and ovarian cancers are gynecological cancers found in women, whereas in men the reproductive neoplasms include prostate, testis and penile cancers.

While the vast majority of the genetic alterations leading to cancer are somatic and found only in the tumors of the affected individuals, gene defects in the germline predisposing to cancer exist as well. Germline mutations are inherited from the parents and are present in every cell of the body. In the first case, where no predisposing inherited genetic factor is identifiable, the cancer is called sporadic. In the latter case, where an inherited mutation predisposing to cancer exists, the phrase hereditary cancer is used. Individuals with such mutations are at a higher risk at developing cancer, and are likely to do so at a younger age than in the general population, with an increased risk for multiple primary tumors.[62] In some cancers, as in CRC, patients with a familial risk can reach 20% of all cases.[63]

1.5.2 Cancer and genes

A cancer cell is a cell that has acquired mutations in critical genes, which allows a cell to escape the normal growth signals and proliferate in an uncontrollable manner. Cancer cells have defects in normal cellular functions that allow them to become malignant. Malignant tumors can be characterized by self-sufficiency to growth signals, insensitivity to growth-inhibitory signals, evasion of programmed cell death (apoptosis), limitless replicative potential, sustained angiogenesis, tissue invasion and metastasis. In addition, the cells gain capabilities to invasion of other surrounding tissues and metastasis through the vascular or lymphatic systems.[64] Thus cancer cells may spread to new areas and organs of the body.

Mutations in several genes are required for a cell to become cancerous. Mutation patterns at the cellular level can be either dominant or recessive. In the dominant case, an abnormality exists in only one of the two alleles of a gene and is sufficient to contribute to oncogenesis. Genes with this type of mutations are referred to as oncogenes, which are altered forms of normal cellular components called proto-oncogenes. These genes control cell proliferation rates by their expression. Oncogenes acquire their tumorigenic properties normally through "gain of function" –mutations such as activating point mutation, amplification or chromosomal translocation.[65]

In a recessive mutation pattern, both alleles of a gene are inactivated. In this case, the genes are called tumor suppressor genes. Tumor suppressor genes can be divided into gatekeepers, caretakers and landscapers according to their function.[66,67] Gatekeepers are thought to directly regulate tumor growth by inhibiting proliferation or promoting cell death, whereas caretakers inhibit tumor growth indirectly by maintaining genomic integrity through DNA repair and replication. While the gatekeepers and caretakers are suggested to act at the intracellular level, the landscaper genes are speculated to be involved in tumorigenesis via intercellular signalling by generating an abnormal environment in the adjacent stromal cells. According to the two-hit hypothesis suggested by Knudson (1971) [68] both copies of the tumor suppressor gene have to be inactivated in order for the cell to turn malignant. The inactivation can be caused by mutations, chromosomal alterations, such as large deletions causing loss of

heterozygosity (LOH) of the gene locus, or epigenetic modifications. Epigenetic modifications are alterations in the genome that do not involve the DNA sequence itself.[69] Examples of epigenetic modifications include hypermethylation and loss of imprinting (LOI). Hypermethylation refers to the aberrant methylation of CpG dinucleotide bases of the promoter regions of the genes and has been associated to the transcriptional silencing of tumor suppressor genes.[70] LOI refers to gene silencing, where genes that show preferential expression of either maternal or paternal allele through a specific methylation pattern of the other allele lose this normal genetic imprinting.

In this dissertation, two types of cancer were studied: a hereditary form of kidney cancer called Hereditary Leiomyomatosis and Renal Cell Cancer (HLRCC), and two classes of CRC: a locally advanced sporadic CRC and serrated colorectal carcinoma. These will be briefly introduced in the following sections.

1.5.3 Hereditary Leiomyomatosis and Renal Cell Cancer

HLRCC is an inherited cancer syndrome with predisposition to leiomyomas of the uterus and skin (benign tumors), uterine leiomyosarcoma and distinct papillary type 2 renal cell carcinoma. The tumor predisposition is caused by heterozygous mutations in the Krebs cycle fumarase gene (*FH*).[71-73] Since the discovery of loss of wild type (WT) allele in tumors, the *FH* gene has been proposed to act as a tumor suppressor gene. However, the cellular and molecular mechanisms leading to tumor development remain unclear. Interestingly, families from Finland have displayed a phenotype with high risk of early-onset renal cell carcinoma and uterine leiomyosarcoma,[74] while this has not been the case in other populations.[75-77] The data from all examined populations are compatible with the notion that some families are very prone to malignant tumors, whereas others are not. [71,75-78] The most simple explanation for this is genotype–phenotype association, where the severity of the functional defect might relate to the occurrence of malignant tumors. However, modifier gene effects have been thought of as an alternative explanation.

1.5.4 Colorectal cancer

CRC is currently the only form of cancer where a model for tumor formation and progression has been described.[79,80] CRC has long been the most prospective cancer type for tumorigenic studies, as many discoveries behind hereditary forms of the cancer, such as *APC* in familial adenomatous polyposis (FAP),[81,82] *MLH1*, *MSH2*, *PMS2* and *MSH6* in hereditary nonpolyposis colorectal cancer (HNPCC) [83-88] and *LKB1* in Peutz-Jeghers syndrome (PJS),[89,90] to name a few, have provided clues to the formation of sporadic CRCs as well. Other genes relevant to the tumorigenesis of CRC are *TP53*, *K-RAS* and *SMAD4*. The *TP53* tumor suppressor gene is important in maintaining DNA integrity and is thought to play an important role in the progression of CRC. Up to half of the colorectal cancers show mutations of *TP53*. [91,92] *K-RAS* is a mitogen activated protein kinase that transduces signals from receptor tyrosine kinases. Activation of the *K-RAS* oncogene has been implicated in colorectal carcinogenesis, and it is mutated in 30–60% of the adenocarcinomas. [93,94] The

chromosome arm 18q is deleted in approximately 50% of colorectal adenomas and 70% of carcinomas. Initial target in the region was *deleted in colorectal cancer (DCC)*, but later studies have implicated *SMAD4* to be perhaps more important reason for these deletions.[95,96] In addition, germline mutations in *SMAD4* predispose to familial colon cancer syndrome juvenile polyposis.[97] The development of cancer is a multistep process which may take years or even decades. Two well characterized pathways leading to CRC are the suppressor pathway and the mutator pathway. In the suppressor pathway, the tumorigenesis is thought to follow an adenoma-carcinoma sequence, where mutations in the wingless-type (Wnt) signaling pathway initiate a neoplastic process, and tumor progression through mutations in other genes leads to the development of an adenoma and finally a carcinoma.[79,80] Colorectal tumors evolving through the suppressor pathway often show large chromosomal instability (CIN), such as losses and amplifications of whole chromosomes.

In the mutator pathway, the progression of colorectal tumors from adenoma to carcinoma is somewhat different. In majority of the cases, the chromosomal material is near diploid, and underlying causes for tumorigenesis are defects in the mismatch repair genes (MMR), which are responsible for maintaining the integrity of DNA in our cells. The MMR deficiency usually leads to microsatellite instability (MSI), which is observed as frequent insertions and deletions within short repetitive sequences (microsatellites) in the genome, where they are most apparent. Mutations in the coding region of a gene may result to altered reading frame and lead to disrupted protein function, which in turn may confer a growth advantage to the cell. Such mutations are typically desirable and selected, and can be found in a significant percentage of colorectal tumors with an MSI phenotype. Approximately 15% of the tumors of the colon and the rectum display MSI. Tumors where MSI is not observed are called microsatellite stable (MSS).

1.5.4.1 Dukes' C colorectal cancer

CRCs are clinically diagnosed using either Dukes' staging (A, B, C, D) or TNM classification (I-IV). These classifications define the stage of the tumor in terms of progression through the intestinal wall, spread to lymph nodes and distant metastasis, and are the basis of treatment selection. A large proportion of CRC patients is diagnosed with a disease with a regional lymph node metastasis (Dukes' C stage) and routinely receives 5-FU-based therapy in combination with surgical resection due to a high risk of recurrence. However, the chemotherapy only benefits 10-20% of the patients.[98,99] Thus far it has not been possible to accurately separate the patients at high risk of recurrence. The use of adjuvant treatment has also complicated the study of prognostic factors that predict recurrence after surgery, because chemotherapy confuses the interpretation of results obtained with recent tumor collections.

1.5.4.2 Serrated colorectal cancer

The serrated colorectal tumors are a heterogeneous group of lesions that combine the architectural features of hyperplastic polyps and the cytological features of conventional adenocarcinomas. Hyperplastic polyps are common lesions found in around 12% of individuals over the age of 50 years [100,101] that have traditionally

been considered non-neoplastic and without malignant potential. While at least some serrated tumors appear to evolve through the classical Wnt signaling pathway,[102] a recently introduced serrated neoplasia pathway has been proposed to act as a previously underrecognized route leading to CRC.[103-105] According to this concept, neoplastic progression originates in the hyperplastic colorectal polyp and drives the accumulation of genetic changes within the lesion, which at the tissue level is thought to result in the development of a serrated adenoma and/or the emergence of colorectal carcinoma.[104] A number of studies describing events such as inhibition of apoptosis by decreased expression of CD95 (Fas),[106] MSI, [107-109] CpG island methylator phenotype,[110,111] allelic imbalance of chromosomal regions 18q and 5q,[112] and *BRAF* mutations [113] have been proposed to explain the characteristic serration of the lesions, but these studies have examined only few molecular features at a time. The serrated tumors may differ from conventional CRCs,[104,109] but the biological background of these tumors is still largely unknown.

1.6 DNA microarray technology and cancer

The new bioinformatics tools and development of genome-wide microarray analyses both in human, mice and other model organisms have opened new windows in cancer research. The field of gene expression studies has greatly advanced the identification of novel tumor susceptibility genes, classification of tumors, prediction of outcome, treatment response, discovery of potential markers and targets for diagnosis of this malignant disease. Microarray technology and the statistical tools developed for it are an excellent option to study mRNA expression differences of normal and tumor tissues or various tumors and model organisms on a global scale. Class discovery methods such as hierarchical or K-means clustering or self organizing maps [55,56] provide a global overview of the cell transcript levels and can be very useful in identifying novel markers for cancer or to identify important genes or pathways for tumorigenesis. Class prediction methods, on the other hand, provide detailed information of specific genetic signatures in various tumor subtypes, which may previously have been very difficult to characterize by conventional methods. Additionally, expression array technology can provide a tool to diagnose clinical cases which may have been difficult to identify otherwise.

DNA microarray technology has expanded rapidly and has been applied to study several different types of human cancer, such as breast,[114-118] prostate,[119-121] colorectal [122] and ovarian cancer [123] as well as hematological malignancies.[52,124-126] A PubMed database search with keywords “microarray AND cancer” returned almost 4900 publications by August 2006, indicating the expansive use of microarray technology. (<http://www.ncbi.nlm.nih.gov/entrez/>)

Landmark studies in the field of class discovery and prediction are the studies by Golub *et al.* (1999),[52] where they showed that two types of leukemias, acute myeloid leukaemia (AML) and acute lymphoid leukaemia (ALL) could be distinguished from each other based on gene expression profiling and a later study (2001) where they demonstrated that pediatric ALL can be subclassified into different groups with very different gene expression profiles.[126] In 1999 and 2000, Alizadeh *et al.* [124,125]

identified a diffused large B-cell lymphoma using “LymphoChip”, a chip designed specifically to profile a hematologic disease. In 2000 Perou *et al.* [114] published an article about molecular portraits of breast tumors where they classified tumours from 42 individuals into distinguished subtypes by their gene expression patterns. The following year (2001) Sorlie *et al.* published a study about breast carcinomas and correlated tumor characteristics to clinical outcome using gene expression profiling.[115] Yet another study of importance is the one by van’t Veer *et al.* (2002),[116] where they were able to classify lymph node negative breast tumors into those with poor or good prognosis using a signature of 70 genes, with power that outstripped the available clinical prognostic markers. The impact of classification can be clinically very significant, as in prostate cancer (PC) where the surgical removal and risk that the surgery poses to these often older men has to be assessed. With molecular signatures being able to determine the possible outcome of the cancer these decisions can be greatly facilitated.

The use of microarray technology together with supporting methods such as linkage analysis increases the power of the study, as this enables the researchers to define specific genomic area and focus the analysis on a certain part of the expression data. Examples of this include the nonsense-mediated mRNA decay (NMD) microarray strategy and combinatorial use of SNP microarrays and linkage analysis together with expression microarrays.[127,128] In eukaryotes, mRNA containing nonsense mutations are selectively and rapidly degraded by the NMD pathway during translation.[129] By pharmacological blocking of the NMD pathway, mutated transcripts containing premature termination codons are stabilized and accumulate in the cells, and the mRNA levels before and after treatment can be monitored using expression microarrays. In a study by Huusko *et al.* (2004) [130] NMD microarrays were used in combination with comparative genomic hybridization (CGH) microarray data to identify truncating mutations in the receptor tyrosine kinase gene *EPHB2* in prostate cancer cell lines. Supporting analyses revealed frameshift, splice site, missense and nonsense mutations in clinical prostate cancer samples. Later publications have revealed this gene to be a putative tumor suppressor gene and inactivated also in CRC.[131,132] In a recent study by Vierimaa *et al.* (2006) [128] expression array data derived from peripheral blood extracted RNA was combined to linkage and SNP array analyses to identify a low-penetrance tumor susceptibility gene *AIP* predisposing to familial pituitary adenomas. This was a new approach, and demonstrated that peripheral blood can be effectively used to identify predisposition to hereditary cancer of various origin. This has also been observed by us in microarray data analyses of HNPCC patients with *MLH1* mutation and HLRCC patients with *FH* mutation (unpublished results). In principle, the use of blood as a source of RNA for expression arrays is in many ways desirable, as it is easily available and fresh, compared to tissue samples which can lead to tedious processing and result to poor quality arrays. Recently, blood derived RNA was also used in our expression array study where a genetic defect predisposing to a Xeroderma pigmentosum (XP) skin carcinoma disorder was discovered in a patient using only one affected and two non-affected chips (unpublished results). The disease in question could not be diagnosed using conventional methods, and demonstrated the diagnostic capabilities of expression arrays in cancer research.

Collectively, these studies serve to demonstrate the tremendous potential microarray technology has in cancer classification and clinical care. Numerous publications have appeared where class discovery has been utilised to distinguish different subtypes of various cancers, patient prognosis or response to treatment. While the microarray technology is most useful in providing transcriptional level information and genetic profiles, this is also the limitation of the technology. In general, other complementary techniques such as functional studies and/or sequencing of genomic DNA are often necessary to understand the molecular events involved and to gain mechanistic insight into the development or progression of cancer. The complexity and variability in data analysis and interpretation of the results, as well as the availability of sufficient quality RNA also remain obstacles in a wider clinical use of the technology. The possibility of using blood derived RNA, reducing the costs of the technology, and standardizing the array preparation and data analysis may provide an opportunity for a wider use of this approach in clinical practice and offer an additional tool in the search for new cancer susceptibility genes.

2 Aims of the Study

Expression microarray technology, together with the recent development in bioinformatics and other biomedical research methods, is emerging as a promising tool in studies of human cancer. The aim of this work was to use expression microarray technology together with supporting methods in cancer research in studies of HLRCC and CRC.

HLRCC is a syndrome caused by mutations in the gene encoding for fumarase (*FH*). While most families with *FH* mutations segregate a benign phenotype of multiple leiomyomas, others display an early-onset renal cancer and leiomyosarcoma, which may be due to modifier gene effects. In the first study (I) the aim was to perform a genome-wide gene expression profiling assay together with functional analyses to evaluate the effect of the *FH* mutation in yeast, and thus use it as a model organism for studies of HLRCC.

A large proportion of CRC patients is diagnosed with a Dukes' C stage disease and is treated with chemotherapy, which only benefits a fraction of the patients. Thus far, an accurate method to distinguish patients at high risk of recurrence has not been available. The second study (II) aimed to use gene expression profiling to identify molecular signatures that characterize tumors from Dukes' C patients with good and bad prognosis.

Serrated colorectal tumors differ from conventional adenocarcinomas morphologically, but not necessarily biologically. The third study (III) aimed to find out whether these two tumor types could be distinguished from each other based on their gene expression profiles and to subsequently identify key genes responsible for these differences. One of the interesting genes observed was *ephrin receptor B2* (*EPHB2*). The role of *EPHB2* in CRC predisposition and the mechanisms of inactivation in colorectal tumors was further analyzed in the last two studies (IV and V).

3 Materials and Methods

The materials and methods used in this work are presented in detail in the original publications (I–V). Here only essential procedures are described.

3.1 Yeast strains (I)

Six yeast strains, consisting of fumarase WT (*FUM1*, *FUM1^U*), mutant (*fum1^{H153R}*, *fum1^{K187R}*) and knockout (*fum1Δ*, *fum1^{vec1}*) strains, were created to study the effects of fumarase mutations. All strains were derived from a WT fumarase-containing *Saccharomyces cerevisiae* strain (*FUM1*) congenic to W303-1A. Fumarase gene knockout strain (*fum1Δ*) was constructed from the WT strain using a PCR-targeted gene disruption method utilizing the *kanMX4* module.[133] This knockout strain was used as a template for the creation of the 4 *FUM1* mutated yeast strains by site-directed mutagenesis. Integrative yeast expression vector (pRS406) containing different versions of the *FUM1* gene were integrated to the yeast genome by transformation that was carried out essentially as described previously.[134] The integration and copy number of the integrated genes were verified by Southern blotting.

3.2 Clinical samples, healthy controls and cell lines (II-V)

Majority of CRC cases used (II-V) came from a well characterized, previously described, consecutively collected and population-based series of 1042 Finnish fresh-frozen colorectal tumor samples.[135,136] In Publication II, 281 Dukes' C patients from this series were selected, of which 25 were used in the microarray analysis. In Publication III, 45 serrated and 115 conventional CRCs from the above mentioned series and from another well characterized and population-based Finnish collection of 466 cases [109] were utilized. Of these, 37 were used on the microarrays.

In Publication IV, 101 normal tissue samples from medical institutions in Finland, UK and USA from patients with 1) CRC and a personal or family history of PC, or 2) intestinal hyperplastic polyposis (HPP) were utilized in the initial mutation screening. The observed alterations were further examined in either Finnish familial prostate (n=164) and colorectal (n=159) cancer patients, or in additional UK HPP patients (n=40), respectively. The Finnish CRC patients belong to the aforementioned population based series of CRC patients.[135,136] The familial PC cohort has also been previously well characterized and described.[137,138] In addition, 282 samples from anonymous Finnish blood donors and 200 healthy UK individuals served as population-matched controls, respectively.

In Publication V, 246 clinical samples collected at medical institutions in Finland, Spain, Germany and Japan were utilized. The MSI status of these tumors was characterized as previously described.[135,136,139,140] The cell lines used in this study were obtained and maintained as previously described.[141]

3.3 Microarray preparation and hybridization

3.3.1 *YG-S98 expression microarrays (I)*

RNA extraction was done from 30 ml YPD cultures harvested at A_{600} 3.0 by homogenizing the samples with glass beads in a breaking buffer (20 mM Tris-HCl pH 7.5, 0.1 M KCl, 2 mM MgCl₂, 2 mM DTT). Total RNA was isolated from the lysate with RNEasy spin columns (Qiagen, Valencia, CA). The quality of RNA was analyzed using Agilent 2100 Bioanalyzer (Agilent Technologies, Palo Alto, CA). Hybridizations to the GeneChip YG-S98 arrays were done according to the manufacturer's instructions using 8 µg of total RNA as starting material (Affymetrix, Santa Clara, CA). Hybridization quality was controlled by inspecting that on each array the number of probes present on the chip, the 3'-5' ratios of the hybridization controls (BIOB, BIOC, BIODN, CREX), average background intensity and noise were on similar levels. Probe sets with unsuccessful hybridization were left out of analyses by masking them out using the MAS 5.0 software (Affymetrix).

3.3.2 *HG-U133A expression microarrays (II, III)*

Frozen tissue samples were macrodissected from the selected areas of the frozen OCT blocks and homogenized in 1 ml of Trizol reagent (Invitrogen, Carlsbad, CA) with a tissue homogenizer (Ultra-Turrax T8; IKA Labortechnik, Staufen, Germany). The total RNA was further purified using RNEasy spin columns (Qiagen). The RNA quality was analyzed using Agilent 2100 Bioanalyzer (Agilent Technologies). Biotin labeled and fragmented cRNA was prepared from 8 µg of total RNA with procedures recommended for the HG-U133 GeneChip expression analysis (Affymetrix). The HG-U133A chips were hybridized, scanned with GeneChip Scanner GA2500, and analyzed with MAS 5.0 software according to the manufacturer's instructions (Affymetrix). The hybridization quality of the chips was inspected by looking at the chip control values, which met the thresholds set by the Tumor Analysis Best Practices Working Group.[142]

3.4 Expression microarray data analysis

3.4.1 *Statistical analysis of the YG-S98 arrays (I)*

The expression data were normalized by centering array intensities on the median. Gene expression values were then scaled by dividing the expression measurement values on each chip with the corresponding control sample value (*FUMI*). The normalized expression data were filtered to remove expression values below detectable levels using the Affymetrix Detection Algorithm assigned flag calls. Subsequent analyses were restricted to genes with detectable expression in all 6 strains. The effect of biological variation resulting from strain processing and transformation methods such as heat shock treatment and insertion of bacterial DNA-containing vector was accounted for by filtering out genes with over 20% deviation in the expression values

between the created WT strain ($FUMI^U$) and the original WT strain ($FUMI$), resulting to a list containing 2113 genes. The threshold for differentially expressed genes was set to genes from this list with at least 2-fold change in the expression value of mutant or deletion strain in comparison to the $FUMI^U$ strain. The data was processed with GeneSpring 7.0 software (Silicon Genetics, Redwood City, CA).

3.4.2 Statistical analysis of the HG-UI33A arrays (II, III)

The expression data were normalized by centering first the array and then the gene intensities to their corresponding medians. The normalized data was filtered to remove expression values below detectable levels using the Affymetrix Detection Algorithm assigned flag calls.

In Publication II, the data analyses were restricted to genes with detectable expression levels in at least 50% of the samples (10035 genes). In samples with absent calls the gene expression was substituted with the average gene expression across all samples to allow class permutation analysis as previously described.[53] Genes differentially expressed in patients with good and bad prognosis were identified with a Student's t-test. Results validation used random permutation of the class labels, where the frequency of events when a p-value was lower than the initial p-value was calculated in one million permutations. Genes that were found to have a p-value lower than the initial p-value in <1% of the permutations were selected as differentially expressed.

In Publication III, the data analyses were restricted to genes with detectable expression in at least 29 of 37 samples (78%), which allowed inclusion of genes absent in the serrated samples while removing noise from the data. A total of 7928 genes fulfilled these criteria. Genes differentially expressed between the serrated and non-serrated CRC groups were identified by performing a Student's t-test and a Benjamini and Hochberg multiple test correction.[54] The data was processed with GeneSpring 6.2 software.

3.4.3 Functional gene enrichment analyses (I, II, III)

In functional group enrichment analysis (FGEA), Gene Ontology terms [143] were used to annotate genes based on their biological processes, cellular components, and molecular functions. Fisher's exact test was then performed to find out whether any of these categories were enriched in a list of differentially expressed genes, when compared to a list of all the annotated genes in that experiment. The analyses were implemented in the GoMiner software.[144]

In Publication I, the analysis was done in each of the mutant and deletion strains by comparing a list of 270 differentially expressed genes to a starting list of 2113 genes that were comparable in all the arrays. Genes that belonged to a category that had a Fisher's exact test p-value less than 0.05 and were over 2 fold over- or underexpressed in comparison to the WT strain $FUMI^U$ were considered enriched.

In Publication II, all genes with present calls in at least 50% of samples (10035 genes) were used for the annotation. This list was compared to a list of 218 unique genes differentially expressed in tumors from patients with good and bad prognosis. Categories with <5 genes or >500 genes were removed from the analysis, as they were considered to be too specific or too general to be useful. Categories with a p-value less than 0.01 were considered enriched.

In Publication III, genes with detectable expression in at least 29 of 37 samples (7928 genes) were used for the annotation and the resulting list compared to a list of 226 differentially expressed genes in serrated versus conventional CRCs. Categories with a p-value less than 0.01 were considered enriched.

3.4.4 Unsupervised hierarchical clustering (II, III)

In Publication II, 5102 genes that were expressed at detectable levels in all 25 samples were selected. The raw expression data for these genes was imported to the Cluster software, log₂-transformed and the intensity of both the arrays and the genes were centered to the median. Subsequently the transformed data was hierarchically clustered there and visualized with TreeView software.[50]

In Publication III, a total of 7928 genes detected in 78% of the samples were used in the unsupervised hierarchical clustering that used Spearman's rank correlation as the similarity metric. Both the clustering and the visualization of the normalized data were implemented in GeneSpring 6.2. software.

3.4.5 Class prediction analyses (II, III)

The class prediction analyses were used to test whether a small set of genes could be used to distinguish the tumor subtype of an unknown sample based on the expression profiles of a training set of samples. The analyses were implemented by using the GeneCluster II software.[51,52] The predictors were generated using a K-nearest neighbors (KNN) classifier and validated with a leave-one-out cross-validation procedure as previously described.[52,141,145]

In the prediction of Dukes' C recurrence (II), the list of 5102 genes expressed at detectable levels in all 25 samples were used for these analyses. Each gene profile was discretized independently to binary values across the samples by applying the Lloyd algorithm.[146] The Lloyd algorithm minimizes the average discretization error, which represents the distance between the data and the discrete representation. This algorithm can be understood as a particular case for 1-dimensional data of the K-means clustering method. The number of genes used in this classifier varied from 1 to 100, and the number of neighbors varied from 1 to 10. The classifier was further validated by using a random permutation of the class labels, which gave an estimate of the average prediction accuracy of the classifier.

In the molecular classification of serrated CRC (III), the predictor was built in using a list of 4413 genes detected in all 37 samples on the microarrays, 3 nearest neighbors

with signal-to noise feature selection (mean class estimate) and a $1/k$ weighting method, which weighs neighbors by the reciprocal of the rank of the neighbor's distance.

3.5 Supporting methods

3.5.1 Functional studies of the yeast strains (I)

The yeast strains were routinely grown on YPD media at 30°C. For plate growth tests yeast cultures were grown on fermentable, partially fermentable and non-fermentable carbon sources. The fumarase enzyme activities were measured from cell lysates by calculating the amount of NAPD consumed in the fumarate-to-malate reaction as previously described.[147] The fumarase protein production was determined by western blotting. The yeast cell cultures were grown on YPD media and collected at A_{600} 2.0. The methods are described in detail in the original publication (I).

3.5.2 Immunohistochemistry and tissue microarray (II, III)

For the immunohistochemical (IHC) assessment of RHOA expression levels in Dukes' C colorectal tumors (Publication II), an independent set of formalin-fixed, paraffin-embedded samples from 137 Dukes' C colorectal tumors and 16 lymph node metastases was used. The experiment was done on a tissue microarray that contained triplicate sections of each tumor sample in a fresh paraffin block. The correlation of RHOA staining level to the survival data was calculated using the mean survival, the hazard ratio and the log-rank p-values in the data set. Details of these analyses are explained in the original publication (II).

The IHC staining of ephrin receptor B2 (EPHB2), patched (PTCH), hypoxia-inducible factor 1-alpha (HIF1 α), cyclin T2 (CCNT2), metastasis associated 1 (MTA1), and methyl-CpG binding domain protein 4 (MBD4) were done in a training set (TS) and a validation set (VS) of samples utilizing paraffin embedded tissue blocks (Publication III). The TS consisted of the 37 (8 serrated and 29 conventional) CRC samples that entered expression array analysis while the VS included a separate set of 37 serrated and 86 conventional CRCs from the same collections. These antibodies were chosen based on observed differential expression at the RNA level, relevance of their biological function, and the commercial availability of the antibodies. Detailed procedures are explained in the original publication (III).

3.5.3 Quantitative RT-PCR (II)

Real-time reverse transcription polymerase chain reactions (RT-PCR) were done from aliquots of RNA (100 ng) using SuperScript II enzyme according to the manufacturer's recommendations (Invitrogen). An aliquot of undiluted RT-PCR product (5 μ l) was used to PCR-amplify the RNA with Assays-on-Demand TaqMan primers and probes and TaqMan Universal PCR Master Mix in a GeneAmp 5700 Sequence Detection

System (all from Applied Biosystems, Branchburg, NJ). Relative gene-expression levels were quantified with the $\Delta\Delta C_T$ method with β -actin as a housekeeping standard control, as previously described.[141,148,149]

3.5.4 Mutation analyses and allelic imbalance (II, III, IV, V)

In the Dukes' C recurrence study (II), an extended set of 46 fresh-frozen CRC samples were screened for *K-RAS* and *TP53* mutations and were assessed for allelic imbalance in chromosome 18q by direct sequencing. Two polymorphic microsatellite markers in 18q21 (D18S1110 and D18S1156) were used to assess the allelic imbalance in this region, as previously reported.[149,150] Allelic imbalance was scored if there was a difference >40% in the abundance of an allele between normal and tumor samples.[150,151] The mutation hotspots of *K-RAS* (codons 12, 13, and 61) were PCR-amplified as described previously.[152,153] Exons 2–11 of *TP53* were PCR-amplified as previously described.[145,154] The complete coding sequences of *RHOA* (exons 2–5) were sequenced in 5 of the tumor samples that showed the highest RHOA protein levels and in 5 of the tumors with the lowest expression.

The mutation analysis of *EPHB2* gene was done by direct screening of the genomic DNA as described in the respective publications (III, IV, V). Loss of heterozygosity (LOH) was scored in cases displaying single nucleotide polymorphisms (SNPs) by comparing sequences from the normal and the corresponding tumor DNA (III, IV). Frequencies of the observed alterations in familial CRC, PC and HPP patient samples, as well as in healthy population-matched controls, were determined by either direct sequencing or DHPLC (IV).

3.5.5 Methylation analysis and western blotting (III, V)

The DNA methylation status of the *EPHB2* promoter-associated CpG islands was determined by chemical conversion of unmethylated cytosines to uracil with bisulfite as previously described.[155] *EPHB2* promoter region was defined and primers created for it with Promoter Scan and MethPrimer software.[156,157] In vitro methylated DNA (CpG Genome Universal Methylated DNA; Chemicon International, Temecula, CA) was used as a positive control for methylated alleles, whereas DNA from normal lymphocytes and normal colon tissues were used as negative controls. The detailed procedures are presented in the relevant publications (III, V).

The *EPHB2* western blotting was done on SW620 cultures treated with varying amounts of the DNA methyltransferase inhibitor 5-aza-2'-deoxycytidine for 72 hours and probed with an anti-*EPHB2* antibody (Stratagene, La Jolla, CA). The membranes were then stripped and reprobed with an anti- β -actin antibody (clone AC74; Sigma, St. Louis, MO), which was used as a loading control. The analysis is described in detail in the original publication (V).

4 Results and Discussion

4.1 Modeling tumor predisposing *FH* mutations in yeast (I)

In this study, genetically modified yeast strains were used to study the effect of two conserved fumarase mutations; H153R, which has been described in 3 cancer predisposition families and K187R, which has displayed the benign phenotype in 3 reported families,[72,74] to assess whether cancer-related fumarase mutations differ from their benign phenotype-associated counterparts.

4.1.1 Expression microarray profiling of the yeast transcriptome

In total, the microarray analysis of the 6 strains identified 2113 genes with comparable expression data between the yeast strains. Of these, 270 probes representing 173 genes and 97 poorly characterized sequences were found to be differentially expressed in the mutant and knockout strains in the yeast transcriptome. The differentially expressed genes belonged to categories related to cell respiratory functions and cell cycle regulation. Genes responsive to fumarase mutations or deletion were composed of multiple categories with functions integral to cell regulation, respiration, GTPase activity and mitochondrial trafficking. In analysis of genes regulating the cell cycle, 108 genes were found from the list of 2113 genes. Of these, 11 genes were found to have over 2 fold changes in at least 1 of the mutant strains (*fumI*^{H153R}, *fumI*^{K187R} or *fumI*^{vect}) in comparison to the WT strain (*FUMI*^U). However, on a pathway level, no significant differences could be distinguished.

A functional gene enrichment analysis performed on strains *fumI*^{H153R}, *fumI*^{K187R} and *fumI*^{vect} found 116 functional groups enriched in the 270 differentially expressed genes ($p < 0.05$). Although only one group was common to all strains, multiple genes inside the groups were overlapping with significantly enriched groups in the other strains. The analysis indicated that the mutant strain *fumI*^{K187R} and the knockout *fumI*^{vect} resembled each other more closely than the other mutant strain *fumI*^{H153R}, which had fewer groups in common with the other strains.

4.1.2 Effect of fumarase alterations to the Krebs cycle gene expression levels

The 15 yeast Krebs cycle genes were in general expressed at a decreased level in the knockout (*fumI*^{vect}) and mutant (*fumI*^{H153R} and *fumI*^{K187R}) strains with the mutant strains showing greater decrease. The only exception to this was the *IDH1* gene, where levels from both of the mutant strains *fumI*^{H153R} and *fumI*^{K187R} as well as the knockout strain *fumI*^{vect} showed increase in expression relative to the WT fumarase-containing strain *FUMI*^U. The mutant strains also showed decrease in expression levels in all of the 4 SDH subunit encoding genes, *KGD1* and *KGD2* as well as *LPD1*. As expected, there was no detectable fumarase expression in the knockout strain *fumI*^{vect}. After adjustment for copy number effects, no significant differences in fumarase expression

were seen in the mutant strains *fum1*^{H153R} and *fum1*^{K187R} in comparison to the WT strain *FUM1*^U.

4.1.3 Functional studies

In the literature, both heterozygous and homozygous mutations in the human fumarase gene (*FH*) causing both modifying and truncating forms of protein with different levels of enzyme activities have been described.[71,158] Previously, the K187R mutation has been one of the few mutations examined in homozygous state in human cells and shown to cause decreased activity.[158] Consequently, the effect of these mutations on fumarase activity in yeast carried out in this study, was of great interest.

The *in vitro* fumarase enzyme activities indicated that the fumarase activity levels differed considerably between the strains. The activity was highest in the WT strain *FUM1*^U and had dropped to a 1/11 fraction in the mutant strains *fum1*^{H153R} and *fum1*^{K187R}. Both mutants had similar fumarase activity levels. The deletion strains *fum1* Δ and *fum1*^{vect} showed no detectable activity. The western blotting results, however, showed fumarase protein in the WT strains *FUM1* and *FUM1*^U as well as in the mutant strains *fum1*^{H153R} and *fum1*^{K187R}. In agreement with the lack of fumarase mRNA observed, in the deletion strains *fum1* Δ and *fum1*^{vect}, no fumarase protein was observed.

To examine whether the detected residual fumarase activity was physiologically meaningful, the created yeast strains were tested for growth phenotype using combinations of carbon sources that were made gradually less fermentative. On fermentative carbon sources all strains had similar appearances, but when grown on less fermentative media the colonies formed more slowly and different phenotypes in appearances were observed in the knockout and mutant strains. On a nonfermentative carbon source the fumarase knockout strains *fum1* Δ and *fum1*^{vect} could not grow due to the loss of mitochondrial respiratory chain function, while the fumarase mutant and WT strains showed normal growth.

Due to the observed differences in growth abilities on nonfermentative plates, the yeast growth rates were measured also from liquid cultures for the WT and deletion strains. While the results indicated a trend towards knockout strains dividing faster than the WT strains, no statistical significance could be demonstrated by t-test.

4.1.4 Evidence of a modifier effect in HLRCC

Fumarase defects have been found to predispose to an autosomal dominant tumor predisposition syndrome HLRCC,[71-73] but the exact molecular mechanisms linking defective *FH* to malignant tumors are not known. The question on relation between *FH* defects and cancer is further complicated by the uneven occurrence of renal cell carcinoma and uterine leiomyosarcoma in families segregating *FH* mutations.[74] This could be explained by phenotype-genotype correlations, e.g. only severe mutations predisposing to cancer. However, as the same mutation can occur both in cancer predisposed and apparently low-risk families, modifying genes could equally well

explain the cancer phenotype. We addressed this issue by comparing the effects of 2 fumarase mutations, H153R with high-penetrance cancer predisposition and K187R with benign phenotype association, in a yeast background.

Microarray studies of the yeast strains consisted of a global analysis of genes most differentially expressed in the mutant and the knockout strains. Functional gene enrichment analysis was further used to look for pathway level differences. The array results reflected observations seen in activity assays and growth tests. When expression levels of cell cycle regulating genes were analyzed in different fumarase mutants, no significant differences were observed on a pathway level between the mutant strains. This correlates well with the finding that no growth rate differences were observed between the mutants. The expression level of most Krebs cycle genes was decreased and in general comparable to previously published results of fumarase knockout yeast.[159] The expression level decrease was more severe in the mutants, which may result from the fumarase point mutations. It was of interest to note that while fumarase mRNA and protein expressions were on comparable levels in both the mutants and the WT strain, a considerable difference was seen in the enzyme activities. Thus, although being partially defective, fumarase could still display interactions with other cellular components possibly even unrelated to respiration, such as cytosolic forms of the Krebs cycle enzymes or other yet unknown interaction partners. Interestingly, the array results on Krebs cycle *IDH1* gene expression levels were similar to those reported in human FH deficient myomas,[160] indicating that the mutations in the yeast fumarase may have similar effects to those observed in patients with HLRCC. This was true for the functional gene enrichment analysis results as well, where overlapping categories, such as oxidoreductase activity and carbohydrate metabolism were similarly affected in the yeast model system and in the human myomas with *FH* mutations.

The key observation of the study was the residual activity of the missense mutated fumarase proteins, which was physiologically relevant and similar in both the renal cancer and uterine leiomyosarcoma phenotype-associated H153R mutant and in the K187R mutant associated with the benign form of HLRCC (multiple cutaneous and uterine leiomyomatosis). This finding supports the hypothesis that HLRCC is not a classical one gene cancer predisposition syndrome with genotype-phenotype variations in severity, but that modifier genes appear to play a role in genesis of malignant tumors in the context of FH germline mutations.

On a more general level, this study demonstrated the applicability of a yeast model in an expression array study of the HLRCC syndrome. The *S. cerevisiae* as a model organism was found to be an adequate system; as an extensively studied eukaryote with a known genome,[161,162] yeast was a very approachable tool to study the molecular and cellular mechanisms of HLRCC. While the statistical analysis of expression data did not reveal candidate genes for the modifier hypothesis, it still supported the current view of Krebs cycle defects playing a major role in tumor predisposition. Retrospectively, the lack of biological replicates was a drawback in the array experiment, as it diminished the power of the statistical analysis. However, the resources available for these studies did not allow an experimental design with multiple replicates. By using more replicates per strain, more sophisticated analyses and better estimations of population mean expression levels could have been done, which would

have increased the confidence of the findings. Additionally, the study demonstrated that while the gene expression levels provided a global and accurate report of the transcriptional level events, the expression array technology was incapable to capture the translational level aspect of protein inactivity; fumarase mRNA and protein was expressed at comparable levels, but showed only minimal residual activity on enzyme assays. Thus, finding a gene which might be related to the HLRCC phenotype in humans by microarray studies can be challenging, as no prior assumptions of function or expected expressional differences are available. The usefulness of the expression arrays was linked to the utilization of other supporting methods, which together provide very useful expressional level information that can be used in further studies of HLRCC. In the future, by combining additional methods and information from other studies, the yeast array data could further benefit the search of a HLRCC modifier gene. A prospective study would be an experiment with yeast strains with both homo- and heterozygous mutations (in diploid yeast strains) or strains with the yeast fumarase replaced by the human fumarase gene.

In the literature, microarray technology has been successfully used in kidney cancer studies to profile different kidney cancer types and normal tissue,[163] and distinguish prognosis for papillary type 1 and 2 tumors,[164] but few studies exist of HLRCC. In a study by Vanharanta *et al.* (2006),[160] sporadic and familial *FH* mutation harbouring leiomyomas were compared to tumors with no *FH* mutations by microarray profiling. In this study, gene expression profiling clustered the tumors with sporadic and familial *FH* mutations together in a same cluster, while the tumors without *FH* mutations clustered separately on their own branch, demonstrating that HLRCC tumors were different from other leiomyomas. Interestingly, one heterozygous sample where one allele had a mutation while the other was intact clustered together with the *FH* mutation negative samples, indicating that two hits could be required to the formation of a HLRCC expression profile. In another study by Yang *et al.* (2005) [164] papillary type 1 and 2 renal cell cancers were distinguished from each other using supervised clustering. Based on the classifier used in that study, a renal cancer cell line with fumarase mutation used in our further studies was by the other group identified as a papillary type 2 tumor (B. Teh, personal communications), which is also the renal tumor type typically seen in HLRCC. However, relatively little is known of the mechanisms of HLRCC tumorigenesis, and thus far classification of the HLRCC syndrome relies on conventional genetic and functional studies to characterize the typical features of the syndrome.[74]

4.2 Prediction of recurrence in Dukes' C colorectal cancer (II)

In this study, fresh-frozen tumor samples from Dukes' C stage CRC patients with surgery as the only form of treatment were subjected to microarray analysis to identify patterns of gene expression that characterize tumors from patients with good and bad prognosis.

4.2.1 *Differentially expressed genes and functional group enrichment analysis*

Altogether 10035 probes were selected for statistical analysis, of which 236 sequences were identified as differentially expressed between patients with good and poor prognosis. Of these, 160 had lower expression and 58 had higher expression in tumors from patients with poor prognosis compared with tumors from patients with good prognosis. Gene ontology (GO) terms were used to classify all these 10035 sequences into 1385 GO categories. Functional group enrichment analysis of the differentially expressed genes identified 24 partially overlapping subcategories with a p-value < 0.01 among the 1385 gene categories tested. These overlapping subcategories could be grouped into 6 main functional groups: protein transport, protein folding, transfer RNA (tRNA) ligase activity, chemotaxis, muscle contraction, and negative regulation of enzyme activity. Most of the genes in protein transport category were involved in vesicle trafficking, a process that regulates multiple signaling mechanisms and has previously been linked to tumorigenesis.[165,166] The remaining genes in this category were involved in nuclear transport, mitochondrial transport, or endoplasmic reticulum trafficking, which are processes previously shown to be deregulated in CRC.[167-169] Protein folding category genes ensure that proteins are correctly folded to be functionally active and play an important role in tumorigenesis.[170-172] The tRNA aminoacylation synthetases are essential in protein synthesis but has recently been realized to have important additional functions in key processes such as tRNA processing, RNA splicing and trafficking, ribosomal RNA synthesis, apoptosis, angiogenesis, and inflammation.[173,174]

4.2.2 *Molecular signatures for good or bad prognosis*

The class discovery by unsupervised hierarchical clustering of the 25 samples divided the dataset into three sub-branches. Using 5102 genes, most of the patients with bad prognosis (7 of 10) clustered together in one of the main sub-branches while a great majority of the patients with good prognosis (12 of 15) clustered in the other 2 sub-branches. Moreover, a significant difference (p-value = 0.019) was observed in the disease-free survival of patients with good and bad prognosis.

In class prediction, a KNN-based classifier using a combination of 17 genes from a set of 72 genes was able to predict the prognosis of 22 of the 25 samples (88%) correctly in our data set. The classifier was built from genes on basis of differential expression levels between the two groups. The sensitivity and specificity of this classifier to identify patients with poor prognosis were 80% and 93.3%, respectively. Importantly, this classifier was able to identify 2 groups of patients with significantly different (p-value = 0.0001) disease-free survival after surgery.

The microarray-based expression profiling outperformed other genetic markers previously investigated, such as p53 and K-RAS status or allelic imbalance in chromosome 18q,[175-184] which in this study showed limited prognostic power in an extended sample set.

4.2.3 *RHOA as a prognostic marker in Dukes' C colorectal cancer*

In this study, statistical analysis of the microarray data revealed the RAS homologue *RHOA* as one of the genes with the most significant difference in expression between the groups with good and bad prognosis. The existing literature of RAS signaling and *RHOA*, [93,185-187] led us to investigate the potential of *RHOA* as a prognostic marker in Dukes' C CRC. The significance of the expressional level differences was assessed using IHC on a tissue microarray, where the level of *RHOA* expression was measured in an independent set of 137 formalin-fixed paraffin-embedded tumor samples from Dukes' C patients. These IHC results indicated that patients with low levels of *RHOA* protein in their tumors had a significantly worse overall (p-value = 0.03) and disease-free (p-value = 0.01) survival compared with patients whose tumors had high levels of *RHOA*. Shorter survivals could also be observed in patients who had surgery as the only form of treatment and in those who, in addition, received 5-FU-based adjuvant chemotherapy. In agreement with these results, lymph node metastases showed significantly (p-value = 0.017) decreased levels of *RHOA* protein than the primary tumors, which further suggests that reduced *RHOA* expression could favor tumor spread.

In the literature, RAS signaling deregulation is a common early event in CRC progression [93,185] and *RHOA* has been shown to regulate a signal transduction pathway that links plasma membrane receptors to the assembly of focal adhesions and actin stress fibers.[188] Recently it has been reported that high levels of *RHOA* can inhibit cell motility.[186,187] These results, together with the findings of this study, indicate that *RHOA* tumor expression levels can be a useful marker for predicting the probability of recurrence of Dukes' C patients treated with or without 5-FU-based chemotherapy and could be used to identify a subset of patients with poor prognosis who could benefit from more aggressive treatment.

4.2.4 *Microarray profiling of colorectal carcinomas*

Collectively, the results of this study show for the first time the potential of gene-expression profiling to predict the probability of recurrence of Dukes' C CRC after surgery. These primary tumors are locally advanced and likely to form distant metastases. The tumors were shown to display a distinctive expression signature; high-density oligonucleotide microarray analysis accurately distinguished patients with good and bad prognosis after surgery. Moreover, we showed that microarray-based genome-wide screening for genes with different levels of expression at the messenger RNA level can be used to identify single prognostic markers for these patients.

In the literature, previously performed studies have assessed the use of individual prognostic markers for CRC treated with 5-FU, such as LOH in chromosome 18q, tumor suppressor gene *p53* and *K-RAS*. [175-184] While the studies of *K-RAS* have suggested a prognostic role for *K-RAS* with cancer progression, the status of *p53* as a prognostic factor and a predictor of response to 5-FU chemotherapy is controversial: a number of studies have demonstrated that overexpression of *p53* correlates with poor survival in IHC analyses in Dukes' B stage CRC, but other studies have found that *p53*

does not display any independent prognostic role in early stage CRC.[189-192] As the currently used predictive markers have only limited value due to acquired or inherent drug resistance, gene expression profiling has been an active area of research that can potentially have major clinical implications to tailored therapies of CRC.

Several groups have used gene expression profiling to study the prognosis in various tumor types such as leukemia,[52] breast [193] and lung cancer.[194] Studies using DNA microarray profiling to classify CRC patients according to prognosis have also appeared in the literature, and have addressed, for example, response to 5-FU,[141] preoperative chemoradiotherapy [195] and recurrence of Dukes' B CRC.[196] Recently, expression profiling has also been used to separate colorectal tumors harboring *BRAF* and *K-RAS* mutations from each other.[197] In the study by Mariadason *et al.* (2003) [141] gene expression profiling combined with leave-one-out cross validation was used in 30 CRC cell lines to identify panels of genes that correlated with sensitivity to 5-FU and camptothecin treatment. The analysis predicted the response more effectively than the 4 previously established determinants of the 5-FU response: *thymidylate synthase*, *thymidine phosphorylase*, *p53* and MSI status. *Thymidylate synthase* and *thymidine phosphorylase* are genes related to 5-FU metabolism, contributing to the apoptotic effect of the active compound. A study by Bertucci *et al.* (2003) demonstrated that gene expression profiling was able to separate Dukes' stage D from stages A-C disease by hierarchical clustering and predict the likelihood of metastasis. Wang *et al.* (2004) used class prediction approaches in Dukes' B colon cancer to identify a 23-gene signature that predicts recurrence in these patients. The Dukes' B stage is a localized disease, where currently no adjuvant therapy is recommended. The study implicated that already at this stage, a gene expression signature could be identified to guide patients with a high risk of recurrence to a more aggressive therapy.

The use of class prediction to find predictive markers for recurrence of different stages of CRC is a demanding task, as the expression array results have thus far been performed on relatively small datasets and may thus suffer from overfitting models. Validation of the obtained prognostic markers in independent experiments is a critical issue, as the variation caused by sample heterogeneity and processing needs to be filtered from results before clinical applications can be designed. However, as multiple studies have shown, prediction of prognostic markers and therapeutic targets from gene expression data is feasible, and the microarray technology offers a good chance to enhance the treatment of cancer patients. In the near future, extended studies, meta-analyses of multiple datasets and further validation of currently predicted prognostic markers are likely to provide clinically usable markers for the prediction of CRC recurrence as well as other cancers.

4.3 Molecular classification of serrated colorectal cancer (III)

In this study, a genome-wide expression microarray profiling and IHC analyses of serrated and conventional colorectal tumors were performed to investigate the molecular basis of serrated CRCs.

4.3.1 Gene expression profiling and class prediction of the serrated CRC

The expression data analysis of the 37 colorectal tumor samples, of which 8 were serrated and 29 conventional CRC, consisted of a class discovery approach and a functional gene enrichment analysis of 7928 genes. In the unsupervised hierarchical clustering the samples formed two distinct branches, with all but one of the 8 serrated samples clustering to one branch, and all 29 conventional CRC cases to the other, with a highly significant p-value of 7.8×10^{-7} (Fisher's exact test for the distribution). None of the other parameters, such as sex, site, grade or stage of the tumors clustered together. Subsequently, a statistical analysis of the 7928 genes used in the clustering detected 226 differentially expressed genes between the serrated and the conventional CRCs (adjusted p-value < 0.05).

Gene Ontology terms were used to classify the probes used in the unsupervised clustering into 685 biological processes, cellular components, and molecular functions. Functional group enrichment analysis was used to identify categories with a significant enrichment in the number of genes differentially expressed in tumors with serrated and non-serrated histology. Nine categories with a p-value < 0.01 were found, of which 5 belonged to categories linked to morphogenesis (morphogenesis and organogenesis) and membrane associated genes (membrane, integral to membrane and integral to plasma membrane).

The class prediction method described by Golub *et al.* [52] was applied to find a set of differentially expressed genes between the serrated and conventional CRCs to find whether the dataset could be used to create an expression-based classifier for unknown samples. The predictor was generated using a KNN-based method and 4413 probes detected in all 37 samples with a leave-one-out cross-validation procedure. This classifier, using 3 neighbors and the expression data of 10 genes from a set of 27 genes, correctly categorized all 37 tumors as serrated or non-serrated. Because of the cross-validation procedure, a slightly different group of 10 genes was used in each prediction.

4.3.2 Validation of the expression array results

The expression microarray results were validated by IHC staining utilizing a training set (TS) and a validation set (VS) of paraffin embedded tissue blocks. The TS consisted of the 37 (8 serrated and 29 conventional) CRC samples that entered expression array analysis while the VS included a separate set of 37 serrated and 86 conventional CRCs from the same collections. Ephrin receptor B2 (EPHB2), patched (PTCH), hypoxia-inducible factor 1-alpha (HIF1 α), cyclin T2 (CCNT2), metastasis associated 1 (MTA1), and methyl-CpG binding domain protein 4 (MBD4), were chosen for further validation by IHC, based on microarray results, relevance of their biological function, and the commercial availability of the antibodies. The IHC results of EPHB2, PTCH and HIF1 α demonstrated statistically significant associations with serrated morphology of the CRC on both TS and VS, with EPHB2 and PTCH showing reduced and HIF1 α elevated levels of expression in the serrated tumors. The results of CCNT2 confirmed the array results on the TS, but failed to reach a significant level on the extended set of

samples. Both MTA and MBD4 showed differences on expression array analysis, but did not show correlation between cancer types in IHC.

To examine possible causes of reduced EPHB2 expression the genomic DNA of the available 24 serrated tumors was sequenced for somatic mutations in the *EPHB2* gene. Successful sequencing of 98% of the coding region yielded negative results. LOH was observed in 25% (3/12) of the serrated tumors displaying informative SNPs. Moreover, the *EPHB2* promoter was hypermethylated in 63% (5/8) of the tumors, suggesting that these mechanisms could contribute to the reduced levels of EPHB2 in serrated tumors.

Many clinical and pathological features in this study suggested that serrated CRCs may be more aggressive than conventional CRCs. Metastatic and poorly differentiated carcinomas were more frequent among the serrated CRCs. Patients with serrated CRC also tended to have worse survival curves than patients with conventional CRC with multiple studied parameters, as in male patients, MSS cancers and in moderately differentiated cancers. This trend has also been previously observed in a smaller set of serrated CRCs for distal and MSS serrated cancers.[109] The serrated CRCs also showed more variability in the WHO grade, tended to be more advanced, more frequently located in the proximal colon and represented a mucinous component more often than the conventional CRCs.

4.3.3 *Serrated colorectal tumors differ from conventional adenocarcinomas*

In this study, class discovery approach was used to investigate whether the serrated colorectal carcinoma is distinguishable from conventional CRC on gene expression level. Previously, serrated adenomas and hyperplastic polyps have been thought of as non-neoplastic lesions with no malignant potential.[198-200] Recently, though, evidence has been presented that certain types of hyperplastic polyps are linked with colorectal neoplasia.[104,108,109,112] The formation of serrated colorectal carcinoma is thought to differ from the conventional adenoma-carcinoma sequence and mutator pathway models, and instead follow an alternative route of colorectal carcinogenesis from hyperplastic polyps to serrated adenomas and adenocarcinomas,[104,201] where hypermethylation may play an essential role.[110,111] Other molecular features, such as frequent mutations in *BRAF* [113] and MSI,[107-109] have also been associated with the serrated phenotype. Currently, the serrated CRCs are diagnosed based on morphological differences. The recognition of serrated carcinomas as biologically different tumors would be of high significance, as identification of a genetic signature in the serrated lesions would allow a better understanding of the disease mechanisms and enable the development of optimal treatment methods for patients according to the specific tumorigenic signatures.

Gene expression profiling proved very useful in distinguishing serrated tumors as a distinctive subgroup of CRC, which was consistent with the hypothesis of a serrated neoplasia pathway proposed in the literature.[103-105] The unsupervised clustering together with the IHC analyses indicated that serrated and conventional CRCs display significant molecular differences, observed both on molecular and histopathological level. Interestingly, our expression microarray and IHC studies indicated that

upregulation of *HIF1 α* and suppression of *PTCH* and *EPHB2*, genes involved in biological processes closely related to the colon cell morphology, differentiation, gastrointestinal tract patterning and proliferation, demonstrated statistically significant association to the serrated phenotype. The IHC analysis also indicated that the immunohistochemical staining of these proteins were useful in distinguishing serrated cancers from conventional cancers, and could thus be potential markers for serrated CRC.

PTCH is a member of the Hedgehog (HH) signaling pathway, and essential for the maintenance of stem cells and for the patterning of the gastrointestinal tract during development.[202] Germline mutations in *PTCH* underlie the Gorlin syndrome, which is characterized by the early onset multiple basal cell carcinomas and increased rate of some other tumors such as medulloblastomas.[203,204]. This gene has also been found in many different sporadic tumors and thought to act as a tumor suppressor.[205] *HIF1 α* activates the transcription of genes that are involved in crucial aspects of cancer biology, including angiogenesis, cell survival, glucose metabolism and invasion, and has been associated with cell proliferation in colon carcinoma.[206-208]. *EPHB2* is a member of a receptor tyrosine kinase family that regulates several signaling pathways involved e.g. in cell growth and migration.[209] The gene maps to chromosome 1p36, which is a chromosomal region reported to be lost in 13% of hyperplastic polyps [210] and is also mutated in PC.[130] Recently, it has also been demonstrated to regulate the development and positioning of cells in the small intestine [211] and to play a critical role in CRC progression, [212] making it a putative tumor suppressor gene. Based on the study results and existing literature, these could be candidate genes involved in the serrated neoplasia pathway.

To study the solid tumors is challenging, as the tumor biopsies generally used in the studies may, in addition to the tumor cells, contain cells from the surrounding tissues, such as endothelia, connective and muscle tissue. To minimize these problems in this study, the array material was screened from fresh-frozen blocks and macrodissected from histologically confirmed locations. Additionally, selection bias was diminished by using population-based cohorts. However, the likely heterogeneous nature of serrated adenomas and the relatively small number of biological replicates in the class discovery and prediction analyses of this study are prone to cause sample-wise variation and model overfitting, as reported in the literature by Ambroise *et al.* (2002) and Allison *et al.* (2006).[36,213] In this study, the class prediction results were validated using a “leave-one-out” method, which does enhance the confidence of the classifier, but is not as good as an independent validation set. In these studies (II, III) however, the accuracy of the expression data was independently confirmed by IHC using an independent sample set. This greatly increases the level of confidence in these experiments, and suggests that overfitting was less of a concern in these studies.

Identification of sub-groups of tumors with different expression profiles is a classical example of the use of microarray technology in cancer research. This approach has been extensively demonstrated in a number of studies including colon cancer,[214] prostate cancer,[121] head and neck squamous cell carcinoma,[215] breast cancer [114] and leukemia,[52,126] to name a few. The earliest microarray-based classifications in CRC were published by Alon *et al.* (1999) and Notterman *et al.* (2001) [122,216] who

used clustering to distinguish CRC and normal tissue. Later publications by Bertucci *et al.* (2003) used unsupervised and supervised clustering to distinguish normal vs cancer tissues and metastatic vs non-metastatic tumors in colon cancer, which were further validated by IHC analyses using tissue microarray. A study by Frederiksen *et al.* (2003) [217] used KNN-based classifier to classify normal, and Dukes' B and C samples with less than 20% error. Despite the obvious challenges inherent in microarray classification of tumors, in the literature – as in this study - it has been possible to define expression profiles of subset of tumors, which is an important step towards individual tumor profiling and tailor-made tumor therapy.

4.4 *EPHB2* and colorectal tumorigenesis (IV, V)

In the previous study (III), expression microarray analyses indicated a decreased level of *EPHB2* gene expression in serrated colorectal tumors, which was validated on protein level experiments by IHC. Our findings, together with the concurrently growing evidence of the role of *EPHB2* in CRC tumorigenesis,[211,212] and PC [130] prompted us to further investigate the role of *EPHB2* in colorectal tumor predisposition. This was done by studying molecular mechanisms responsible for the inactivation of the gene.

4.4.1 EPHB2 germline variants in colorectal tumorigenesis (IV)

The *EPHB2* germline alterations were screened in a set of 101 samples consisting of CRC patients with either personal or family history of PC or intestinal hyperplastic polyposis (HPP). Overall, variants that may be associated with the disease phenotype were seen in 3/101 (3.0%) patients analyzed in the initial mutation screening. Two of the changes, I361V and R568W, were identified in Finnish CRC patients, but not in over 300 Finnish familial CRC or PC patients or more than 200 population-matched healthy controls. The third change, D861N, was observed in a UK HPP patient, but not in additional 40 UK HPP patients or in 200 UK healthy controls. A fourth change R80H, originally identified in a Finnish CRC patient, was also found in 1/106 familial CRC patients and in 9/281 healthy controls and was considered as a likely neutral polymorphism. The data from this study is compatible with the results by Oba *et al.* (2001), who found no *EPHB2* germline mutations among 50 CRC patients.[218]

The rarity of germline *EPHB2* alterations in this and in a previous study,[218] together with the sequential loss of *EPHB2* expression during colorectal carcinogenesis, suggests a limited role for *EPHB2* in CRC predisposition and speaks for the more pronounced role for *EPHB2* in tumor progression than in tumor initiation. Therefore, although some possibly disease associated germline *EPHB2* variants do exist and may play a role in colorectal tumor predisposition, the observed *EPHB2* inactivation in CRCs appears to be largely due to other mechanisms, such as promoter hypermethylation, LOH, and somatic mutations.[132,219] Notwithstanding, germline *EPHB2* variants do exist and may be associated with colon tumor predisposition, but further studies including functional analyses are needed to confirm this.

4.4.2 Mechanisms of *EPHB2* inactivation in colorectal tumors (V)

EPHB2 gene contains an A9 track in exon 17 that could be a target for mutation in tumors with MSI. To investigate the role of the A9 repeat of *EPHB2* in CRC tumorigenesis, we screened 24 MSI colorectal cell lines and 246 primary MSI colorectal tumors by direct sequencing and found that 9 of the cell lines (37.5%) and 101 of the tumors (41%) had mutations in the A9 repeat. All of the found mutations changed the translational reading frame, and either added a 26–amino acid tail or prematurely truncated the protein. In all cases, two serine residues (S1048 and S1052) were lost in the mutant *EPHB2*. These residues have been predicted to be phosphorylated in the wild-type protein [220] and might thus play a role in the protein kinase activity. In a screened set of 29 MSI adenomas [139] we found that 20.7% of these adenomas (6 of 29) had a mutation in the A9 repeat of *EPHB2*, which was significantly lower than in MSI carcinomas (41%, 101 of 246; χ^2 test p-value = 0.03). This observation was in good agreement with earlier reports showing that *EPHB2* expression was reduced or lost in colorectal carcinomas, but not in adenomas,[212] and further suggests that *EPHB2* inactivation may be important for the transition from adenoma to carcinoma.

Hypermethylation of cytosines located within CpG islands in the promoter of tumor suppressor genes is emerging as an important mechanism of gene silencing.[221] The proximal promoter of *EPHB2* contains a CpG island spanning 1400bp and provides a site for aberrant methylation. The DNA methylation status of this region was determined in 60 MSI and 41 MSS colorectal tumor samples and in 20 MSS CRC cell lines with methylation specific PCR. The results showed 53.4% (54/101) of the tumors and 25% of the cell lines (5/20) to be hypermethylated. No differences were seen between the MSI and MSS tumors. The treatment of a cell line showing *EPHB2* promoter methylation with a demethylating reagent resulted in a substantial up-regulation of *EPHB2* protein levels and demonstrated that aberrant methylation can regulate *EPHB2* expression.

This study described for the first time the mechanisms of *EPHB2* inactivation in colorectal tumors, caused by frequent mutations in repetitive sequences in exon 17 in MSI adenomas and carcinomas and hypermethylation of the *EPHB2* promoter in the majority of the tumors.

4.4.3 *EPHB2* in colorectal tumorigenesis

The Wnt signaling pathway plays a central role in the development of CRC. In the majority of cases the early events in tumorigenesis involve inactivation of the tumor suppressor gene APC and stabilization of β -catenin.[222,223] The constitutive activity of β -catenin/transcription factor 4 (*TCF4*) –complex leads to transcription of growth promoting genes, which together with subsequent inactivation of tumor suppressor genes drive the tumorigenesis further and enable the formation of abnormal growth patterns. Recently, *EPHB2* was demonstrated to be one of the direct transcriptional targets of β -catenin, and participate in the correct positioning of the cells along the crypt axis in the intestinal epithelium.[211] Furthermore, *EPHB2* maps to a

chromosomal region (1p36) frequently altered in colorectal tumors [218,224,225] as well as in many types of cancer,[226] and was recently shown to function as a putative tumor suppressor gene in CRC.[212] Recent findings have also demonstrated that silencing of *EPHB2* correlates inversely, and is a strong indicator of poor overall survival in CRC.[227,228] The evidence present thus far suggests that *EPHB2* is a new tumor suppressor gene in CRC with a role in tumor progression.

These studies demonstrate the benefits of integrated use of microarray technology and data mining together with other research methods. When connected to the existing literature and concurrently published studies, individual findings such as *EPHB2* in this study begin to have functional significance and thus become interesting targets for further studies. In general, microarray data analysis often requires extensive data mining, and conclusive evidence of the involvement of a gene in carcinogenesis is achieved only after validation of the results by functional testing.

4.5 Future prospects of array technologies

The future of microarray research in cancer is going towards a systems biology view, where the aim is to change focus from individual genes to pathways and more complex signalling networks in the cells. These kinds of integrative methods, such as meta-analyses of multiple datasets and studies of transcription factor networks have already been published and reviewed in the literature,[229] but require input from other high-throughput molecular approaches such as SNP arrays, array comparative genomic hybridization and proteomic solutions to become useful and more informative. However, retrospectively, it is astounding to note that microarray technology has been around for only a decade. When looking at the annually increasing rate of publications related to microarrays, it can be stated that biomedical sciences are evolving at an unprecedented pace.

Array technologies, together with bioinformatics methods related to it, have founded a novel platform for cancer research. Chip-based experiments are becoming a standard practise in basic research, both in studies of tumor biology and prognostic markers for the disease. Applications based on targeted therapies are also making their way to the mainstream clinical practise, as the expression of specific target genes in tumors and means to interfere with them are discovered and understood. When the prices of microarrays become comparable to currently utilized diagnostic tests and the target preparation and data analysis of the arrays have been standardized to meet the quality expected from medical treatment methods, gene expression profiling may eventually allow a personalized treatment of cancer by tailoring therapies to the biologically active pathways in each patient's tumor.

5 Conclusions

Molecular biology and the technology that supports it have lately advanced in ways that were inconceivable a decade earlier. The development of DNA microarray technology together with the mapping of the genomic sequence has made it possible to analyze the expression of the total human genome in a single experiment and perform extensive data mining at the subcellular level. This has given the scientists valuable insight into various diseases, such as human cancer, and increasing interest exists in changing the emphasis of tumor classification from morphologic to molecular.

In Publication I, a genome-wide expression profiling was performed on yeast strains with fumarase mutations that in human predispose to the HLRCC syndrome. The results showed that the mutant and knockout strains have similar expression profiles. While the fumarase mRNA and protein expression was at comparable levels between the mutant and WT strains, functional studies showed that the enzyme activity in the mutant strains was only a fraction of the activity seen in the WT strains. This residual activity was nonetheless crucial and sufficient to support normal growth phenotype. This supports the hypothesis that modifier gene(s) display a major role in determining tumor types in families segregating *FH* mutations. In addition, the study demonstrated that while microarray experiments on model organisms were useful in providing data of gene expression changes for HLRCC studies, functional analyses were essential to complement and understand the significance of the results at the cellular level.

In Publication II, expression microarrays were used to distinguish surgically cured Dukes' C CRC patients from those at high risk of recurrence. Unsupervised clustering and class prediction methods were able to identify groups of patients with significantly different survival and outperformed previously reported genetic markers of prognosis. One of the most differentially expressed genes in the experiment, *RHOA*, was further analyzed as a potential prognostic marker by using IHC and tissue microarray. The study showed that gene expression profiling of surgical samples can predict the recurrence of Dukes' C patients and could perhaps be used to guide decisions concerning the clinical management of these patients.

In Publication III, gene expression profiling was used for molecular classification of the serrated CRCs. Unsupervised clustering showed that serrated carcinomas appear to be a subclass of CRC with distinct molecular basis. The study provides a platform to understand the molecular basis of serrated CRC and in long term may contribute to the development of specific treatment options for this tumor type.

In Publications IV and V, *EPHB2* - a key target gene revealed by the expression data analysis in Publication III - was further studied in colon tumors and normal tissue to find out about its relevance to CRC tumorigenesis. These studies showed that *EPHB2* germline variations exist and may be associated with colon tumor predisposition. However, rarity of these events suggests a limited role for them and the mechanisms of inactivation appear to be largely due to frequent frameshift mutations as observed in MSI colorectal tumors and aberrant methylation of the regulatory sequences of *EPHB2*.

In conclusion, in this dissertation expression microarray technology and supporting technologies were used in studies of hereditary renal cell cancer and CRC. While the expression arrays were limited to detecting aberrant gene expression levels, and did not capture the changes at translational level, they still offered a very powerful platform to study genetic interactions and molecular mechanisms on a genome-wide level. Careful designing, sample preparation and use of sufficient amount of biological replicates in the experiments were found to be crucial elements. The usefulness of the technology was greatly enhanced by combined use of other research methods, such as functional studies, which enabled focusing of the analyses and provided better insight into the complex regulation of gene and protein networks involved in cancer. Although presently the expression array technology is primarily used in basic research, previous studies as well as results obtained from this dissertation indicate that clinical applications are foreseeable and slowly emerging, and can benefit the treatment of cancer in form of individually tailored therapies based on the molecular signature of a patient's tumor.

6 References

1. Watson JD, Crick FH. Genetical implications of the structure of deoxyribonucleic acid. *Nature* (1953), 171, 964-967.
2. Watson JD, Crick FH. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature* (1953), 171, 737-738.
3. Watson JD, Crick FH. The structure of DNA. *Cold Spring Harb. Symp. Quant. Biol.* (1953), 18, 123-131.
4. Mendes Soares LM, Valcarcel J. The expanding transcriptome: the genome as the 'Book of Sand'. *EMBO J.* (2006), 25, 923-931.
5. Heller MJ. An active microelectronics device for multiplex DNA analysis. *IEEE in Medicine and Biology* (1996), 15, 100-104.
6. Sosnowski RG, Tu E, Butler WF, O'Connell JP, Heller MJ. Rapid determination of single base mismatch mutations in DNA hybrids by direct electric field control. *Proc. Natl. Acad. Sci. U. S. A.* (1997), 94, 1119-1123.
7. Baldi P, Hatfield GW. DNA microarrays and gene expression: from experiments to data analysis and modeling. Cambridge, Cambridge University Press, 2002.
8. Kohane IS, Kho A, Butte AJ. Microarrays for an integrative genomics. Cambridge, Mass., MIT Press, London, 2003.
9. Heller MJ. DNA microarray technology: devices, systems, and applications. *Annu. Rev. Biomed. Eng.* (2002), 4, 129-153.
10. Dufva M. Fabrication of high quality microarrays. *Biomol. Eng.* (2005), 22, 173-184.
11. Kafatos FC, Jones CW, Efstratiadis A. Determination of nucleic acid sequence homologies and relative concentrations by a dot hybridization procedure. *Nucleic Acids Res.* (1979), 7, 1541-1552.
12. Saiki RK, Walsh PS, Levenson CH, Erlich HA. Genetic analysis of amplified DNA with immobilized sequence-specific oligonucleotide probes. *Proc. Natl. Acad. Sci. U. S. A.* (1989), 86, 6230-6234.
13. Kaiser RJ, MacKellar SL, Vinayak RS, Sanders JZ, Saavedra RA, Hood LE. Specific-primer-directed DNA sequencing using automated fluorescence detection. *Nucleic Acids Res.* (1989), 17, 6087-6102.
14. Maskos U, Southern EM. Parallel analysis of oligodeoxyribonucleotide (oligonucleotide) interactions. I. Analysis of factors influencing oligonucleotide duplex formation. *Nucleic Acids Res.* (1992), 20, 1675-1678.
15. Maskos U, Southern EM. Oligonucleotide hybridizations on glass supports: a novel linker for oligonucleotide synthesis and hybridization properties of oligonucleotides synthesised in situ. *Nucleic Acids Res.* (1992), 20, 1679-1684.
16. Fodor SP, Read JL, Pirrung MC, Stryer L, Lu AT, Solas D. Light-directed, spatially addressable parallel chemical synthesis. *Science* (1991), 251, 767-773.

17. Schena M, Shalon D, Davis RW, Brown PO. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* (1995), 270, 467-470.
18. Ekins R, Chu F, Biggart E. Multispot, multianalyte, immunoassay. *Ann. Biol. Clin. (Paris)* (1990), 48, 655-666.
19. Schena M, Shalon D, Heller R, Chai A, Brown PO, Davis RW. Parallel human genome analysis: microarray-based expression monitoring of 1000 genes. *Proc. Natl. Acad. Sci. U. S. A.* (1996), 93, 10614-10619.
20. Shalon D, Smith SJ, Brown PO. A DNA microarray system for analyzing complex DNA samples using two-color fluorescent probe hybridization. *Genome Res.* (1996), 6, 639-645.
21. DeRisi J, Penland L, Brown PO, Bittner ML, Meltzer PS, Ray M, Chen Y, Su YA, Trent JM. Use of a cDNA microarray to analyse gene expression patterns in human cancer. *Nat. Genet.* (1996), 14, 457-460.
22. Fodor SP, Rava RP, Huang XC, Pease AC, Holmes CP, Adams CL. Multiplexed biochemical assays with biological chips. *Nature* (1993), 364, 555-556.
23. Pease AC, Solas D, Sullivan EJ, Cronin MT, Holmes CP, Fodor SP. Light-generated oligonucleotide arrays for rapid DNA sequence analysis. *Proc. Natl. Acad. Sci. U. S. A.* (1994), 91, 5022-5026.
24. Jarvinen AK, Hautaniemi S, Edgren H, Auvinen P, Saarela J, Kallioniemi OP, Monni O. Are data from different gene expression microarray platforms comparable? *Genomics* (2004), 83, 1164-1168.
25. Tan PK, Downey TJ, Spitznagel EL, Jr, Xu P, Fu D, Dimitrov DS, Lempicki RA, Raaka BM, Cam MC. Evaluation of gene expression measurements from commercial microarray platforms. *Nucleic Acids Res.* (2003), 31, 5676-5684.
26. Draghici S, Khatri P, Eklund AC, Szallasi Z. Reliability and reproducibility issues in DNA microarray measurements. *Trends Genet.* (2006), 22, 101-109.
27. Lipshutz RJ, Fodor SP, Gingeras TR, Lockhart DJ. High density synthetic oligonucleotide arrays. *Nat. Genet.* (1999), 21, 20-24.
28. Agilent Technologies. Agilent SurePrint Technology. (2003), Technical Report 5988-8171EN.
29. Blanchard AP, Kaiser RJ, Hood LE. High-density oligonucleotide arrays. *Biosensors and Bioelectronics* (1996), 11, 687-690.
30. Baldeschwieler JD, Gamble RC, Theriault TP, inventors. Method and apparatus for performing multiple sequential reactions on a matrix. (1995), PTC patent WO/1995/025116.
31. Brennan TM, inventor. Method and apparatus for conducting an array of chemical reactions on a support surface. (1995), US patent 5474796.
32. Gutmann O, Niekrawietz R, Kuehlewein R, Steinert CP, Reinbold S, De Heij B, Daub M, Zengerle R. Non-contact production of oligonucleotide microarrays using the highly integrated TopSpot nanoliter dispenser. *Analyst* (2004), 129, 835-840.

33. Okamoto T, Suzuki T, Yamamoto N. Microarray fabrication with covalent attachment of DNA using bubble jet technology. *Nat. Biotechnol.* (2000), 18, 438-441.
34. Affymetrix.GeneChip® Expression Analysis Technical Manual. (2001), Technical Report 701029 Rev. 4.
35. Affymetrix.GeneChip® Expression Analysis Data Analysis Fundamentals. (2004), Technical Report 701190 Rev. 4.
36. Allison DB, Cui X, Page GP, Sabripour M. Microarray data analysis: from disarray to consolidation and consensus. *Nat Rev Genet* (2006), 7, 55-65.
37. Ekstrom CT, Bak S, Kristensen C, Rudemo M. Spot shape modelling and data transformations for microarrays. *Bioinformatics* (2004), 20, 2270-2278.
38. Quackenbush J. Microarray data normalization and transformation. *Nat. Genet.* (2002), 32 Suppl, 496-501.
39. Smyth GK, Speed T. Normalization of cDNA microarray data. *Methods* (2003), 31, 265-273.
40. Qin LX, Kerr KF, Contributing Members of the Toxicogenomics Research Consortium. Empirical evaluation of data transformations and ranking statistics for microarray analysis. *Nucleic Acids Res.* (2004), 32, 5471-5479.
41. Yang YH, Buckley MJ, Speed TP. Analysis of cDNA microarray images. *Brief Bioinform* (2001), 2, 341-349.
42. Yang YH, Dudoit S, Luu P, Lin DM, Peng V, Ngai J, Speed TP. Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Res.* (2002), 30, e15.
43. Li C, Wong WH. Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. *Proc. Natl. Acad. Sci. U. S. A.* (2001), 98, 31-36.
44. Schadt EE, Li C, Ellis B, Wong WH. Feature extraction and normalization algorithms for high-density oligonucleotide gene expression array data. *J. Cell. Biochem. Suppl.* (2001), Suppl 37, 120-125.
45. Irizarry RA, Bolstad BM, Collin F, Cope LM, Hobbs B, Speed TP. Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Res.* (2003), 31, e15.
46. Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, Speed TP. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* (2003), 4, 249-264.
47. Affymetrix.Statistical Algorithms Description Document. (2002), Technical Report 701137 Rev. 3.
48. Wu Z., Irizarry R. A., Gentleman R., Martinez-Murillo F. and Spencer F. A Model Based Background Adjustment for Oligonucleotide Expression Arrays. (2004), Johns Hopkins University, Dept. of Biostatistical Working Papers, Working Paper 1.
49. Brazma A, Hingamp P, Quackenbush J, Sherlock G, Spellman P, Stoeckert C, Aach J, Ansoorge W, Ball CA, Causton HC, Gaasterland T, Glenisson P, Holstege FC, Kim IF, Markowitz V, Matese JC, Parkinson H, Robinson A, Sarkans U, Schulze-Kremer S, Stewart J, Taylor R, Vilo J, Vingron M. Minimum information about a microarray

experiment (MIAME)-toward standards for microarray data. *Nat. Genet.* (2001), 29, 365-371.

50. Eisen MB, Spellman PT, Brown PO, Botstein D. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. U. S. A.* (1998), 95, 14863-14868.

51. Tamayo P, Slonim D, Mesirov J, Zhu Q, Kitareewan S, Dmitrovsky E, Lander ES, Golub TR. Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proc. Natl. Acad. Sci. U. S. A.* (1999), 96, 2907-2912.

52. Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, Bloomfield CD, Lander ES. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* (1999), 286, 531-537.

53. Tusher VG, Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl. Acad. Sci. U. S. A.* (2001), 98, 5116-5121.

54. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J.Roy.Stat.Soc.B Met.* (1995), 57, 289-300.

55. MacQueen J. Some Methods for classification and Analysis of Multivariate Observations. *Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability* (1967), 1, 281-297.

56. Kohonen T. Self-Organizing Maps. 3rd extended ed. New York, Springer, 2001.

57. Kane MD, Jatkoe TA, Stumpf CR, Lu J, Thomas JD, Madore SJ. Assessment of the sensitivity and specificity of oligonucleotide (50mer) microarrays. *Nucleic Acids Res.* (2000), 28, 4552-4557.

58. Holland MJ. Transcript abundance in yeast varies over six orders of magnitude. *J. Biol. Chem.* (2002), 277, 14363-14366.

59. Czechowski T, Bari RP, Stitt M, Scheible WR, Udvardi MK. Real-time RT-PCR profiling of over 1400 Arabidopsis transcription factors: unprecedented sensitivity reveals novel root- and shoot-specific genes. *Plant J.* (2004), 38, 366-379.

60. Dai M, Wang P, Boyd AD, Kostov G, Athey B, Jones EG, Bunney WE, Myers RM, Speed TP, Akil H, Watson SJ, Meng F. Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data. *Nucleic Acids Res.* (2005), 33, e175.

61. Bray F, Sankila R, Ferlay J, Parkin DM. Estimates of cancer incidence and mortality in Europe in 1995. *Eur. J. Cancer* (2002), 38, 99-166.

62. Fearon ER. Human cancer syndromes: clues to the origin and nature of cancer. *Science* (1997), 278, 1043-1050.

63. Lynch HT, de la Chapelle A. Hereditary colorectal cancer. *N. Engl. J. Med.* (2003), 348, 919-932.

64. Hanahan D, Weinberg RA. The hallmarks of cancer. *Cell* (2000), 100, 57-70.

65. Bishop JM. Molecular themes in oncogenesis. *Cell* (1991), 64, 235-248.

66. Kinzler KW, Vogelstein B. Cancer-susceptibility genes. Gatekeepers and caretakers. *Nature* (1997), 386, 761, 763.
67. Kinzler KW, Vogelstein B. Landscaping the cancer terrain. *Science* (1998), 280, 1036-1037.
68. Knudson AG, Jr. Mutation and cancer: statistical study of retinoblastoma. *Proc. Natl. Acad. Sci. U. S. A.* (1971), 68, 820-823.
69. Verma M, Srivastava S. Epigenetics in cancer: implications for early detection and prevention. *Lancet Oncol.* (2002), 3, 755-763.
70. Baylin SB, Herman JG. DNA hypermethylation in tumorigenesis: epigenetics joins genetics. *Trends Genet.* (2000), 16, 168-174.
71. Tomlinson IP, Alam NA, Rowan AJ, Barclay E, Jaeger EE, Kelsell D, Leigh I, Gorman P, Lamlum H, Rahman S, Roylance RR, Olpin S, Bevan S, Barker K, Hearle N, Houlston RS, Kiuru M, Lehtonen R, Karhu A, Vilkki S, Laiho P, Eklund C, Vierimaa O, Aittomaki K, Hietala M, Sistonen P, Paetau A, Salovaara R, Herva R, Launonen V, Aaltonen LA. Germline mutations in FH predispose to dominantly inherited uterine fibroids, skin leiomyomata and papillary renal cell cancer. *Nat. Genet.* (2002), 30, 406-410.
72. Kiuru M, Lehtonen R, Arola J, Salovaara R, Jarvinen H, Aittomaki K, Sjoberg J, Visakorpi T, Knuutila S, Isola J, Delahunt B, Herva R, Launonen V, Karhu A, Aaltonen LA. Few FH mutations in sporadic counterparts of tumor types observed in hereditary leiomyomatosis and renal cell cancer families. *Cancer Res.* (2002), 62, 4554-4557.
73. Launonen V, Vierimaa O, Kiuru M, Isola J, Roth S, Pukkala E, Sistonen P, Herva R, Aaltonen LA. Inherited susceptibility to uterine leiomyomas and renal cell cancer. *Proc. Natl. Acad. Sci. U. S. A.* (2001), 98, 3387-3392.
74. Kiuru M, Launonen V. Hereditary leiomyomatosis and renal cell cancer (HLRCC). *Curr. Mol. Med.* (2004), 4, 869-875.
75. Alam NA, Rowan AJ, Wortham NC, Pollard PJ, Mitchell M, Tyrer JP, Barclay E, Calonje E, Manek S, Adams SJ, Bowers PW, Burrows NP, Charles-Holmes R, Cook LJ, Daly BM, Ford GP, Fuller LC, Hadfield-Jones SE, Hardwick N, Highet AS, Keefe M, MacDonald-Hull SP, Potts ED, Crone M, Wilkinson S, Camacho-Martinez F, Jablonska S, Ratnavel R, MacDonald A, Mann RJ, Grice K, Guillet G, Lewis-Jones MS, McGrath H, Seukeran DC, Morrison PJ, Fleming S, Rahman S, Kelsell D, Leigh I, Olpin S, Tomlinson IP. Genetic and functional analyses of FH mutations in multiple cutaneous and uterine leiomyomatosis, hereditary leiomyomatosis and renal cancer, and fumarate hydratase deficiency. *Hum. Mol. Genet.* (2003), 12, 1241-1252.
76. Chan I, Wong T, Martinez-Mir A, Christiano AM, McGrath JA. Familial multiple cutaneous and uterine leiomyomas associated with papillary renal cell cancer. *Clin. Exp. Dermatol.* (2005), 30, 75-78.
77. Toro JR, Nickerson ML, Wei MH, Warren MB, Glenn GM, Turner ML, Stewart L, Duray P, Tourre O, Sharma N, Choyke P, Stratton P, Merino M, Walther MM, Linehan WM, Schmidt LS, Zbar B. Mutations in the fumarate hydratase gene cause hereditary

- leiomyomatosis and renal cell cancer in families in North America. *Am. J. Hum. Genet.* (2003), 73, 95-106.
78. Martinez-Mir A, Glaser B, Chuang GS, Horev L, Waldman A, Engler DE, Gordon D, Spelman LJ, Hatzibougias I, Green J, Christiano AM, Zlotogorski A. Germline fumarate hydratase mutations in families with multiple cutaneous and uterine leiomyomata. *J. Invest. Dermatol.* (2003), 121, 741-744.
79. Fearon ER, Vogelstein B. A genetic model for colorectal tumorigenesis. *Cell* (1990), 61, 759-767.
80. Kinzler KW, Vogelstein B. Lessons from hereditary colorectal cancer. *Cell* (1996), 87, 159-170.
81. Groden J, Thliveris A, Samowitz W, Carlson M, Gelbert L, Albertsen H, Joslyn G, Stevens J, Spirio L, Robertson M. Identification and characterization of the familial adenomatous polyposis coli gene. *Cell* (1991), 66, 589-600.
82. Kinzler KW, Nilbert MC, Su LK, Vogelstein B, Bryan TM, Levy DB, Smith KJ, Preisinger AC, Hedge P, McKechnie D. Identification of FAP locus genes from chromosome 5q21. *Science* (1991), 253, 661-665.
83. Fishel R, Lescoe MK, Rao MR, Copeland NG, Jenkins NA, Garber J, Kane M, Kolodner R. The human mutator gene homolog MSH2 and its association with hereditary nonpolyposis colon cancer. *Cell* (1993), 75, 1027-1038.
84. Leach FS, Nicolaides NC, Papadopoulos N, Liu B, Jen J, Parsons R, Peltomaki P, Sistonen P, Aaltonen LA, Nystrom-Lahti M. Mutations of a mutS homolog in hereditary nonpolyposis colorectal cancer. *Cell* (1993), 75, 1215-1225.
85. Bronner CE, Baker SM, Morrison PT, Warren G, Smith LG, Lescoe MK, Kane M, Earabino C, Lipford J, Lindblom A. Mutation in the DNA mismatch repair gene homologue hMLH1 is associated with hereditary non-polyposis colon cancer. *Nature* (1994), 368, 258-261.
86. Nicolaides NC, Papadopoulos N, Liu B, Wei YF, Carter KC, Ruben SM, Rosen CA, Haseltine WA, Fleischmann RD, Fraser CM. Mutations of two PMS homologues in hereditary nonpolyposis colon cancer. *Nature* (1994), 371, 75-80.
87. Papadopoulos N, Nicolaides NC, Wei YF, Ruben SM, Carter KC, Rosen CA, Haseltine WA, Fleischmann RD, Fraser CM, Adams MD. Mutation of a mutL homolog in hereditary colon cancer. *Science* (1994), 263, 1625-1629.
88. Miyaki M, Konishi M, Tanaka K, Kikuchi-Yanoshita R, Muraoka M, Yasuno M, Igari T, Koike M, Chiba M, Mori T. Germline mutation of MSH6 as the cause of hereditary nonpolyposis colorectal cancer. *Nat. Genet.* (1997), 17, 271-272.
89. Hemminki A, Tomlinson I, Markie D, Jarvinen H, Sistonen P, Bjorkqvist AM, Knuutila S, Salovaara R, Bodmer W, Shibata D, de la Chapelle A, Aaltonen LA. Localization of a susceptibility locus for Peutz-Jeghers syndrome to 19p using comparative genomic hybridization and targeted linkage analysis. *Nat. Genet.* (1997), 15, 87-90.
90. Hemminki A, Markie D, Tomlinson I, Avizienyte E, Roth S, Loukola A, Bignell G, Warren W, Aminoff M, Hoglund P, Jarvinen H, Kristo P, Pelin K, Ridanpaa M,

- Salovaara R, Toro T, Bodmer W, Olschwang S, Olsen AS, Stratton MR, de la Chapelle A, Aaltonen LA. A serine/threonine kinase gene defective in Peutz-Jeghers syndrome. *Nature* (1998), 391, 184-187.
91. Iacopetta B. TP53 mutation in colorectal cancer. *Hum. Mutat.* (2003), 21, 271-276.
92. Russo A, Bazan V, Iacopetta B, Kerr D, Soussi T, Gebbia N, TP53-CRC Collaborative Study Group. The TP53 colorectal cancer international collaborative study on the prognostic and predictive significance of p53 mutation: influence of tumor site, type of mutation, and adjuvant treatment. *J. Clin. Oncol.* (2005), 23, 7518-7528.
93. Bos JL. ras oncogenes in human cancer: a review. *Cancer Res.* (1989), 49, 4682-4689.
94. Kressner U, Bjorheim J, Westring S, Wahlberg SS, Pahlman L, Glimelius B, Lindmark G, Lindblom A, Borresen-Dale AL. Ki-ras mutations and prognosis in colorectal cancer. *Eur. J. Cancer* (1998), 34, 518-521.
95. Takagi Y, Kohmura H, Futamura M, Kida H, Tanemura H, Shimokawa K, Saji S. Somatic alterations of the DPC4 gene in human colorectal cancers in vivo. *Gastroenterology* (1996), 111, 1369-1372.
96. Miyaki M, Iijima T, Konishi M, Sakai K, Ishii A, Yasuno M, Hishima T, Koike M, Shitara N, Iwama T, Utsunomiya J, Kuroki T, Mori T. Higher frequency of Smad4 gene mutation in human colorectal cancer with distant metastasis. *Oncogene* (1999), 18, 3098-3103.
97. Howe JR, Roth S, Ringold JC, Summers RW, Jarvinen HJ, Sistonen P, Tomlinson IP, Houlston RS, Bevan S, Mitros FA, Stone EM, Aaltonen LA. Mutations in the SMAD4/DPC4 gene in juvenile polyposis. *Science* (1998), 280, 1086-1088.
98. Moertel CG, Fleming TR, Macdonald JS, Haller DG, Laurie JA, Goodman PJ, Ungerleider JS, Emerson WA, Tormey DC, Glick JH. Levamisole and fluorouracil for adjuvant therapy of resected colon carcinoma. *N. Engl. J. Med.* (1990), 322, 352-358.
99. Moertel CG, Fleming TR, Macdonald JS, Haller DG, Laurie JA, Tangen CM, Ungerleider JS, Emerson WA, Tormey DC, Glick JH, Veeder MH, Mailliard JA. Fluorouracil plus levamisole as effective adjuvant therapy after resection of stage III colon carcinoma: a final report. *Ann. Intern. Med.* (1995), 122, 321-326.
100. Imperiale TF, Wagner DR, Lin CY, Larkin GN, Rogge JD, Ransohoff DF. Risk of advanced proximal neoplasms in asymptomatic adults according to the distal colorectal findings. *N. Engl. J. Med.* (2000), 343, 169-174.
101. Lieberman DA, Weiss DG, Bond JH, Ahnen DJ, Garewal H, Chejfec G. Use of colonoscopy to screen asymptomatic adults for colorectal cancer. Veterans Affairs Cooperative Study Group 380. *N. Engl. J. Med.* (2000), 343, 162-168.
102. Sawyer EJ, Cerar A, Hanby AM, Gorman P, Arends M, Talbot IC, Tomlinson IP. Molecular characteristics of serrated adenomas of the colorectum. *Gut* (2002), 51, 200-206.
103. Goldstein NS. Serrated Pathway and APC (Conventional)-Type Colorectal Polyps. *Am J Clin Pathol* (2006), 125, 146-153.

104. Hawkins NJ, Bariol C, Ward RL. The serrated neoplasia pathway. *Pathology* (2002), 34, 548-555.
105. Higuchi T, Jass JR. My approach to serrated polyps of the colorectum. *J. Clin. Pathol.* (2004), 57, 682-686.
106. Tateyama H, Li W, Takahashi E, Miura Y, Sugiura H, Eimoto T. Apoptosis index and apoptosis-related antigen expression in serrated adenoma of the colorectum: the saw-toothed structure may be related to inhibition of apoptosis. *Am. J. Surg. Pathol.* (2002), 26, 249-256.
107. Iino H, Jass JR, Simms LA, Young J, Leggett B, Ajioka Y, Watanabe H. DNA microsatellite instability in hyperplastic polyps, serrated adenomas, and mixed polyps: a mild mutator pathway for colorectal cancer? *J. Clin. Pathol.* (1999), 52, 5-9.
108. Hawkins NJ, Ward RL. Sporadic colorectal cancers with microsatellite instability and their possible origin in hyperplastic polyps and serrated adenomas. *J. Natl. Cancer Inst.* (2001), 93, 1307-1313.
109. Makinen MJ, George SM, Jernvall P, Makela J, Vihko P, Karttunen TJ. Colorectal carcinoma associated with serrated adenoma--prevalence, histological features, and prognosis. *J. Pathol.* (2001), 193, 286-294.
110. Chan AO, Issa JP, Morris JS, Hamilton SR, Rashid A. Concordant CpG island methylation in hyperplastic polyposis. *Am. J. Pathol.* (2002), 160, 529-536.
111. Park SJ, Rashid A, Lee JH, Kim SG, Hamilton SR, Wu TT. Frequent CpG island methylation in serrated adenomas of the colorectum. *Am. J. Pathol.* (2003), 162, 815-822.
112. Yashiro M, Laghi L, Saito K, Carethers JM, Slezak P, Rubio C, Hirakawa K, Boland CR. Serrated adenomas have a pattern of genetic alterations that distinguishes them from other colorectal polyps. *Cancer Epidemiol. Biomarkers Prev.* (2005), 14, 2253-2256.
113. Kambara T, Simms LA, Whitehall VL, Spring KJ, Wynter CV, Walsh MD, Barker MA, Arnold S, McGivern A, Matsubara N, Tanaka N, Higuchi T, Young J, Jass JR, Leggett BA. BRAF mutation is associated with DNA methylation in serrated polyps and cancers of the colorectum. *Gut* (2004), 53, 1137-1144.
114. Perou CM, Jeffrey SS, van de Rijn M, Rees CA, Eisen MB, Ross DT, Pergamenschikov A, Williams CF, Zhu SX, Lee JC, Lashkari D, Shalon D, Brown PO, Botstein D. Distinctive gene expression patterns in human mammary epithelial cells and breast cancers. *Proc. Natl. Acad. Sci. U. S. A.* (1999), 96, 9212-9217.
115. Sorlie T, Perou CM, Tibshirani R, Aas T, Geisler S, Johnsen H, Hastie T, Eisen MB, van de Rijn M, Jeffrey SS, Thorsen T, Quist H, Matese JC, Brown PO, Botstein D, Eystein Lonning P, Borresen-Dale AL. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc. Natl. Acad. Sci. U. S. A.* (2001), 98, 10869-10874.
116. van't Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AAM, Mao M, Peterse HL, van der Kooy K, Marton MJ, Witteveen AT, Schreiber GJ, Kerkhoven RM, Roberts C, Linsley PS, Bernards R, Friend SH. Gene expression profiling predicts clinical outcome of breast cancer. *Nature* (2002), 415, 530-536.

117. Hedenfalk I, Duggan D, Chen Y, Radmacher M, Bittner M, Simon R, Meltzer P, Gusterson B, Esteller M, Kallioniemi OP, Wilfond B, Borg A, Trent J. Gene-expression profiles in hereditary breast cancer. *N. Engl. J. Med.* (2001), 344, 539-548.
118. Cunliffe HE, Ringner M, Bilke S, Walker RL, Cheung JM, Chen Y, Meltzer PS. The gene expression response of breast cancer to growth regulators: patterns and correlation with tumor expression profiles. *Cancer Res.* (2003), 63, 7158-7166.
119. Bubendorf L, Kolmer M, Kononen J, Koivisto P, Mousses S, Chen Y, Mahlamaki E, Schraml P, Moch H, Willi N, Elkahoul AG, Pretlow TG, Gasser TC, Mihatsch MJ, Sauter G, Kallioniemi OP. Hormone therapy failure in human prostate cancer: analysis by complementary DNA and tissue microarrays. *J. Natl. Cancer Inst.* (1999), 91, 1758-1764.
120. Dhanasekaran SM, Barrette TR, Ghosh D, Shah R, Varambally S, Kurachi K, Pienta KJ, Rubin MA, Chinnaiyan AM. Delineation of prognostic biomarkers in prostate cancer. *Nature* (2001), 412, 822-826.
121. Singh D, Febbo PG, Ross K, Jackson DG, Manola J, Ladd C, Tamayo P, Renshaw AA, D'Amico AV, Richie JP, Lander ES, Loda M, Kantoff PW, Golub TR, Sellers WR. Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell* (2002), 1, 203-209.
122. Alon U, Barkai N, Notterman DA, Gish K, Ybarra S, Mack D, Levine AJ. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc. Natl. Acad. Sci. U. S. A.* (1999), 96, 6745-6750.
123. Ono K, Tanaka T, Tsunoda T, Kitahara O, Kihara C, Okamoto A, Ochiai K, Takagi T, Nakamura Y. Identification by cDNA microarray of genes involved in ovarian carcinogenesis. *Cancer Res.* (2000), 60, 5007-5011.
124. Alizadeh A, Eisen M, Davis RE, Ma C, Sabet H, Tran T, Powell JI, Yang L, Marti GE, Moore DT, Hudson JR, Jr, Chan WC, Greiner T, Weisenburger D, Armitage JO, Lossos I, Levy R, Botstein D, Brown PO, Staudt LM. The lymphochip: a specialized cDNA microarray for the genomic-scale analysis of gene expression in normal and malignant lymphocytes. *Cold Spring Harb. Symp. Quant. Biol.* (1999), 64, 71-78.
125. Alizadeh AA, Eisen MB, Davis RE, Ma C, Lossos IS, Rosenwald A, Boldrick JC, Sabet H, Tran T, Yu X, Powell JI, Yang L, Marti GE, Moore T, Hudson J, Jr, Lu L, Lewis DB, Tibshirani R, Sherlock G, Chan WC, Greiner TC, Weisenburger DD, Armitage JO, Warnke R, Levy R, Wilson W, Grever MR, Byrd JC, Botstein D, Brown PO, Staudt LM. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* (2000), 403, 503-511.
126. Golub TR. Genomic approaches to the pathogenesis of hematologic malignancy. *Curr. Opin. Hematol.* (2001), 8, 252-261.
127. Wolf M, Edgren H, Muggerud A, Kilpinen S, Huusko P, Sorlie T, Mousses S, Kallioniemi O. NMD microarray analysis for rapid genome-wide screen of mutated genes in cancer. *Cell Oncol.* (2005), 27, 169-173.
128. Vierimaa O, Georgitsi M, Lehtonen R, Vahteristo P, Kokko A, Raitila A, Tuppurainen K, Ebeling TML, Salmela PI, Paschke R, Gündogdu S, De Menis E,

Mäkinen MJ, Launonen V, Karhu A, Aaltonen LA. Pituitary Adenoma Predisposition Caused by Germline Mutations in the AIP Gene. *Science* (2006), 312, 1228-1230.

129. Maquat LE. Nonsense-mediated mRNA decay in mammals. *J. Cell. Sci.* (2005), 118, 1773-1776.

130. Huusko P, Ponciano-Jackson D, Wolf M, Kiefer JA, Azorsa DO, Tuzmen S, Weaver D, Robbins C, Moses T, Allinen M, Hautaniemi S, Chen Y, Elkahloun A, Basik M, Bova GS, Bubendorf L, Lugli A, Sauter G, Schleutker J, Ozelik H, Elowe S, Pawson T, Trent JM, Carpten JD, Kallioniemi OP, Mousses S. Nonsense-mediated decay microarray analysis identifies mutations of EPHB2 in human prostate cancer. *Nat. Genet.* (2004), 36, 979-983.

131. Batlle E, Bacani J, Begthel H, Jonkheer S, Gregorieff A, van de Born M, Malats N, Sancho E, Boon E, Pawson T, Gallinger S, Pals S, Clevers H. EphB receptor activity suppresses colorectal cancer progression. *Nature* (2005), 435, 1126-1130.

132. Alazzouzi H, Davalos V, Kokko A, Domingo E, Woerner SM, Wilson AJ, Konrad L, Laiho P, Espin E, Armengol M, Imai K, Yamamoto H, Mariadason JM, Gebert JF, Aaltonen LA, Schwartz SJ, Arango D. Mechanisms of inactivation of the receptor tyrosine kinase EPHB2 in colorectal tumors. *Cancer Res.* (2005), 65, 10170-10173.

133. Wach A, Brachat A, Pohlmann R, Philippsen P. New heterologous modules for classical or PCR-based gene disruptions in *Saccharomyces cerevisiae*. *Yeast* (1994), 10, 1793-1808.

134. Gietz D, St Jean A, Woods RA, Schiestl RH. Improved method for high efficiency transformation of intact yeast cells. *Nucleic Acids Res.* (1992), 20, 1425.

135. Aaltonen LA, Salovaara R, Kristo P, Canzian F, Hemminki A, Peltomaki P, Chadwick RB, Kaariainen H, Eskelinen M, Jarvinen H, Mecklin JP, de la Chapelle A. Incidence of hereditary nonpolyposis colorectal cancer and the feasibility of molecular screening for the disease. *N. Engl. J. Med.* (1998), 338, 1481-1487.

136. Salovaara R, Loukola A, Kristo P, Kaariainen H, Ahtola H, Eskelinen M, Harkonen N, Julkunen R, Kangas E, Ojala S, Tulikoura J, Valkamo E, Jarvinen H, Mecklin JP, Aaltonen LA, de la Chapelle A. Population-based molecular detection of hereditary nonpolyposis colorectal cancer. *J. Clin. Oncol.* (2000), 18, 2193-2200.

137. Seppala EH, Ikonen T, Autio V, Rokman A, Mononen N, Matikainen MP, Tammela TL, Schleutker J. Germ-line alterations in MSR1 gene and prostate cancer risk. *Clin. Cancer Res.* (2003), 9, 5252-5256.

138. Schleutker J, Matikainen M, Smith J, Koivisto P, Baffoe-Bonnie A, Kainu T, Gillanders E, Sankila R, Pukkala E, Carpten J, Stephan D, Tammela T, Brownstein M, Bailey-Wilson J, Trent J, Kallioniemi OP. A genetic epidemiological study of hereditary prostate cancer (HPC) in Finland: frequent HPCX linkage in families with late-onset disease. *Clin. Cancer Res.* (2000), 6, 4810-4815.

139. Woerner SM, Kloor M, Mueller A, Rueschoff J, Friedrichs N, Buettner R, Buzello M, Kienle P, Knaebel HP, Kunstmann E, Pagenstecher C, Schackert HK, Moslein G, Vogelsang H, von Knebel Doeberitz M, Gebert JF. Microsatellite instability of selective target genes in HNPCC-associated colon adenomas. *Oncogene* (2005), 24, 2525-2535.

140. Boland CR, Thibodeau SN, Hamilton SR, Sidransky D, Eshleman JR, Burt RW, Meltzer SJ, Rodriguez-Bigas MA, Fodde R, Ranzani GN, Srivastava S. A National Cancer Institute Workshop on Microsatellite Instability for cancer detection and familial predisposition: development of international criteria for the determination of microsatellite instability in colorectal cancer. *Cancer Res.* (1998), 58, 5248-5257.
141. Mariadason JM, Arango D, Shi Q, Wilson AJ, Corner GA, Nicholas C, Aranes MJ, Lesser M, Schwartz EL, Augenlicht LH. Gene expression profiling-based prediction of response of colon carcinoma cells to 5-fluorouracil and camptothecin. *Cancer Res.* (2003), 63, 8791-8812.
142. Hoffman EP, Awad T, Palma J, Webster T, Hubbell E, Warrington JA, Spira A, Wright G, Buckley J, Triche T, Davis R, Tibshirani R, Wenzhong X, Jones W, Tompkins R, West M. Expression profiling - best practices for data generation and interpretation in clinical trials. *Nat. Rev. Genet.* (2004), 5, 229-237.
143. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* (2000), 25, 25-29.
144. Zeeberg BR, Feng W, Wang G, Wang MD, Fojo AT, Sunshine M, Narasimhan S, Kane DW, Reinhold WC, Lababidi S, Bussey KJ, Riss J, Barrett JC, Weinstein JN. GoMiner: a resource for biological interpretation of genomic and proteomic data. *Genome Biol.* (2003), 4, R28.
145. Arango D, Wilson AJ, Shi Q, Corner GA, Aranes MJ, Nicholas C, Lesser M, Mariadason JM, Augenlicht LH. Molecular mechanisms of action and prediction of response to oxaliplatin in colorectal cancer cells. *Br. J. Cancer* (2004), 91, 1931-1946.
146. Lloyd SP. Least Squares Quantization in PCM. *IEEE Trans Information Theory* (1982), IT-28 (2 March), 129-137.
147. Hatch MD. A simple spectrophotometric assay for fumarate hydratase in crude tissue extracts. *Anal. Biochem.* (1978), 85, 271-275.
148. Arango D, Mariadason JM, Wilson AJ, Yang W, Corner GA, Nicholas C, Aranes MJ, Augenlicht LH. c-Myc overexpression sensitises colon cancer cells to camptothecin-induced apoptosis. *Br. J. Cancer* (2003), 89, 1757-1765.
149. Alhopuro P, Alazzouzi H, Sammalkorpi H, Davalos V, Salovaara R, Hemminki A, Jarvinen H, Mecklin JP, Schwartz S, Jr, Aaltonen LA, Arango D. SMAD4 levels and response to 5-fluorouracil in colorectal cancer. *Clin. Cancer Res.* (2005), 11, 6311-6316.
150. Alazzouzi H, Alhopuro P, Salovaara R, Sammalkorpi H, Jarvinen H, Mecklin JP, Hemminki A, Schwartz SJ, Aaltonen LA, Arango D. SMAD4 as a prognostic marker in colorectal cancer. *Clin. Cancer Res.* (2005), 11, 2606-2611.
151. Canzian F, Salovaara R, Hemminki A, Kristo P, Chadwick RB, Aaltonen LA, de la Chapelle A. Semiautomated assessment of loss of heterozygosity and replication error in tumors. *Cancer Res.* (1996), 56, 3331-3337.

152. Servomaa K, Kiuru A, Kosma VM, Hirvikoski P, Rytomaa T. p53 and K-ras gene mutations in carcinoma of the rectum among Finnish women. *Mol. Pathol.* (2000), 53, 24-30.
153. Laiho P, Launonen V, Lahermo P, Esteller M, Guo M, Herman JG, Mecklin JP, Jarvinen H, Sistonen P, Kim KM, Shibata D, Houlston RS, Aaltonen LA. Low-level microsatellite instability in most colorectal carcinomas. *Cancer Res.* (2002), 62, 1166-1170.
154. Hsu IC, Metcalf RA, Sun T, Welsh JA, Wang NJ, Harris CC. Mutational hotspot in the p53 gene in human hepatocellular carcinomas. *Nature* (1991), 350, 427-428.
155. Herman JG, Graff JR, Myohanen S, Nelkin BD, Baylin SB. Methylation-specific PCR: a novel PCR assay for methylation status of CpG islands. *Proc. Natl. Acad. Sci. U. S. A.* (1996), 93, 9821-9826.
156. Prestridge DS. Predicting Pol II promoter sequences using transcription factor binding sites. *J. Mol. Biol.* (1995), 249, 923-932.
157. Li LC, Dahiya R. MethPrimer: designing primers for methylation PCRs. *Bioinformatics* (2002), 18, 1427-1431.
158. Coughlin EM, Christensen E, Kunz PL, Krishnamoorthy KS, Walker V, Dennis NR, Chalmers RA, Elpeleg ON, Whelan D, Pollitt RJ, Ramesh V, Mandell R, Shih VE. Molecular analysis and prenatal diagnosis of human fumarase deficiency. *Mol. Genet. Metab.* (1998), 63, 254-262.
159. McCammon MT, Epstein CB, Przybyla-Zawislak B, McAlister-Henn L, Butow RA. Global transcription analysis of Krebs tricarboxylic acid cycle mutants reveals an alternating pattern of gene expression and effects on hypoxic and oxidative genes. *Mol. Biol. Cell* (2003), 14, 958-972.
160. Vanharanta S, Pollard PJ, Lehtonen HJ, Laiho P, Sjoberg J, Leminen A, Aittomaki K, Arola J, Kruhoffer M, Orntoft TF, Tomlinson IP, Kiuru M, Arango D, Aaltonen LA. Distinct expression profile in fumarate-hydratase-deficient uterine fibroids. *Hum. Mol. Genet.* (2006), 15, 97-103.
161. Goffeau A, Barrell BG, Bussey H, Davis RW, Dujon B, Feldmann H, Galibert F, Hoheisel JD, Jacq C, Johnston M, Louis EJ, Mewes HW, Murakami Y, Philippsen P, Tettelin H, Oliver SG. Life with 6000 genes. *Science* (1996), 274, 546, 563-567.
162. Foury F, Roganti T, Lecrenier N, Purnelle B. The complete sequence of the mitochondrial genome of *Saccharomyces cerevisiae*. *FEBS Lett.* (1998), 440, 325-331.
163. Jones J, Otu H, Spentzos D, Kolia S, Inan M, Beecken WD, Fellbaum C, Gu X, Joseph M, Pantuck AJ, Jonas D, Libermann TA. Gene signatures of progression and metastasis in renal cell cancer. *Clin. Cancer Res.* (2005), 11, 5730-5739.
164. Yang XJ, Tan MH, Kim HL, Ditlev JA, Betten MW, Png CE, Kort EJ, Futami K, Furge KA, Takahashi M, Kanayama HO, Tan PH, Teh BS, Luan C, Wang K, Pins M, Tretiakova M, Anema J, Kahnoski R, Nicol T, Stadler W, Vogelzang NG, Amato R, Seligson D, Figlin R, Beldegrun A, Rogers CG, Teh BT. A molecular classification of papillary renal cell carcinoma. *Cancer Res.* (2005), 65, 5628-5637.


165. Cheng KW, Lahad JP, Kuo WL, Lapuk A, Yamada K, Auersperg N, Liu J, Smith-McCune K, Lu KH, Fishman D, Gray JW, Mills GB. The RAB25 small GTPase determines aggressiveness of ovarian and breast cancers. *Nat. Med.* (2004), 10, 1251-1256.
166. Fukata M, Kaibuchi K. Rho-family GTPases in cadherin-mediated cell-cell adhesion. *Nat. Rev. Mol. Cell Biol.* (2001), 2, 887-897.
167. Kau TR, Way JC, Silver PA. Nuclear transport and cancer: from mechanism to intervention. *Nat. Rev. Cancer.* (2004), 4, 106-117.
168. Garrido C, Gurbuxani S, Ravagnan L, Kroemer G. Heat shock proteins: endogenous modulators of apoptotic cell death. *Biochem. Biophys. Res. Commun.* (2001), 286, 433-442.
169. Rao RV, Ellerby HM, Bredesen DE. Coupling endoplasmic reticulum stress to the cell death program. *Cell Death Differ.* (2004), 11, 372-380.
170. Brunagel G, Shah U, Schoen RE, Getzenberg RH. Identification of calreticulin as a nuclear matrix protein associated with human colon cancer. *J. Cell. Biochem.* (2003), 89, 238-243.
171. Schubert A, Grimm S. Cyclophilin D, a component of the permeability transition-pore, is an apoptosis repressor. *Cancer Res.* (2004), 64, 85-93.
172. Xanthoudakis S, Roy S, Rasper D, Hennessey T, Aubin Y, Cassady R, Tawa P, Ruel R, Rosen A, Nicholson DW. Hsp60 accelerates the maturation of pro-caspase-3 by upstream activator proteases during apoptosis. *EMBO J.* (1999), 18, 2049-2056.
173. Han JM, Kim JY, Kim S. Molecular network and functional implications of macromolecular tRNA synthetase complex. *Biochem. Biophys. Res. Commun.* (2003), 303, 985-993.
174. Francklyn C, Perona JJ, Puetz J, Hou YM. Aminoacyl-tRNA synthetases: versatile players in the changing theater of translation. *RNA* (2002), 8, 1363-1372.
175. Bell SM, Scott N, Cross D, Sagar P, Lewis FA, Blair GE, Taylor GR, Dixon MF, Quirke P. Prognostic value of p53 overexpression and c-Ki-ras gene mutations in colorectal cancer. *Gastroenterology* (1993), 104, 57-64.
176. Benhattar J, Losi L, Chaubert P, Givel JC, Costa J. Prognostic significance of K-ras mutations in colorectal carcinoma. *Gastroenterology* (1993), 104, 1044-1048.
177. Goh HS, Yao J, Smith DR. p53 point mutation and survival in colorectal cancer patients. *Cancer Res.* (1995), 55, 5217-5221.
178. Ogunbiyi OA, Goodfellow PJ, Herfarth K, Gagliardi G, Swanson PE, Birnbaum EH, Read TE, Fleshman JW, Kodner IJ, Moley JF. Confirmation that chromosome 18q allelic loss in colon cancer is a prognostic indicator. *J. Clin. Oncol.* (1998), 16, 427-433.
179. Lanza G, Matteuzzi M, Gafa R, Orvieto E, Maestri I, Santini A, del Senno L. Chromosome 18q allelic loss and prognosis in stage II and III colon cancer. *Int. J. Cancer* (1998), 79, 390-395.

180. Watanabe T, Wu TT, Catalano PJ, Ueki T, Satriano R, Haller DG, Benson AB, 3rd, Hamilton SR. Molecular predictors of survival after adjuvant chemotherapy for colon cancer. *N. Engl. J. Med.* (2001), 344, 1196-1206.
181. Laurent-Puig P, Olschwang S, Delattre O, Remvikos Y, Asselain B, Melot T, Validire P, Muleris M, Girodet J, Salmon RJ. Survival and acquired genetic alterations in colorectal cancer. *Gastroenterology* (1992), 102, 1136-1141.
182. Jen J, Kim H, Piantadosi S, Liu ZF, Levitt RC, Sistonen P, Kinzler KW, Vogelstein B, Hamilton SR. Allelic loss of chromosome 18q and prognosis in colorectal cancer. *N. Engl. J. Med.* (1994), 331, 213-221.
183. Diep CB, Thorstensen L, Meling GI, Skovlund E, Rognum TO, Lothe RA. Genetic tumor markers with prognostic impact in Dukes' stages B and C colorectal cancer patients. *J. Clin. Oncol.* (2003), 21, 820-829.
184. Cohn KH, Ornstein DL, Wang F, LaPaix FD, Phipps K, Edelsberg C, Zuna R, Mott LA, Dunn JL. The significance of allelic deletions and aneuploidy in colorectal carcinoma. Results of a 5-year follow-up study. *Cancer* (1997), 79, 233-244.
185. Oliveira C, Westra JL, Arango D, Ollikainen M, Domingo E, Ferreira A, Velho S, Niessen R, Lagerstedt K, Alhopuro P, Laiho P, Veiga I, Teixeira MR, Ligtenberg M, Kleibeuker JH, Sijmons RH, Plukker JT, Imai K, Lage P, Hamelin R, Albuquerque C, Schwartz SJ, Lindblom A, Peltomaki P, Yamamoto H, Aaltonen LA, Seruca R, Hofstra RM. Distinct patterns of KRAS mutations in colorectal carcinomas according to germline mismatch repair defects and hMLH1 methylation status. *Hum. Mol. Genet.* (2004), 13, 2303-2311.
186. Vial E, Sahai E, Marshall CJ. ERK-MAPK signaling coordinately regulates activity of Rac1 and RhoA for tumor cell motility. *Cancer. Cell.* (2003), 4, 67-79.
187. Cox EA, Sastry SK, Huttenlocher A. Integrin-mediated adhesion regulates cell polarity and membrane protrusion through the Rho family of GTPases. *Mol. Biol. Cell* (2001), 12, 265-277.
188. Jaffe AB, Hall A. Rho GTPases in transformation and metastasis. *Adv. Cancer Res.* (2002), 84, 57-80.
189. Soong R, Grieco F, Robbins P, Dix B, Chen D, Parsons R, House A, Iacopetta B. p53 alterations are associated with improved prognosis in distal colonic carcinomas. *Clin. Cancer Res.* (1997), 3, 1405-1411.
190. Giatromanolaki A, Stathopoulos GP, Tsiobanou E, Papadimitriou C, Georgoulas V, Gatter KC, Harris AL, Koukourakis MI. Combined role of tumor angiogenesis, bcl-2, and p53 expression in the prognosis of patients with colorectal carcinoma. *Cancer* (1999), 86, 1421-1430.
191. Bhatavdekar JM, Patel DD, Chikhlikar PR, Shah NG, Vora HH, Ghosh N, Trivedi TI. Molecular markers are predictors of recurrence and survival in patients with Dukes B and Dukes C colorectal adenocarcinoma. *Dis. Colon Rectum* (2001), 44, 523-533.
192. Allegra CJ, Paik S, Colangelo LH, Parr AL, Kirsch I, Kim G, Klein P, Johnston PG, Wolmark N, Wieand HS. Prognostic value of thymidylate synthase, Ki-67, and p53 in patients with Dukes' B and C colon cancer: a National Cancer Institute-National

- Surgical Adjuvant Breast and Bowel Project collaborative study. *J. Clin. Oncol.* (2003), 21, 241-250.
193. van de Vijver MJ, He YD, van't Veer LJ, Dai H, Hart AA, Voskuil DW, Schreiber GJ, Peterse JL, Roberts C, Marton MJ, Parrish M, Atsma D, Witteveen A, Glas A, Delahaye L, van der Velde T, Bartelink H, Rodenhuis S, Rutgers ET, Friend SH, Bernards R. A gene-expression signature as a predictor of survival in breast cancer. *N. Engl. J. Med.* (2002), 347, 1999-2009.
194. Beer DG, Kardia SL, Huang CC, Giordano TJ, Levin AM, Misek DE, Lin L, Chen G, Gharib TG, Thomas DG, Lizyness ML, Kuick R, Hayasaka S, Taylor JM, Iannettoni MD, Orringer MB, Hanash S. Gene-expression profiles predict survival of patients with lung adenocarcinoma. *Nat. Med.* (2002), 8, 816-824.
195. Ghadimi BM, Grade M, Difilippantonio MJ, Varma S, Simon R, Montagna C, Fuzesi L, Langer C, Becker H, Liersch T, Ried T. Effectiveness of gene expression profiling for response prediction of rectal adenocarcinomas to preoperative chemoradiotherapy. *J. Clin. Oncol.* (2005), 23, 1826-1838.
196. Wang Y, Jatkoe T, Zhang Y, Mutch MG, Talantov D, Jiang J, McLeod HL, Atkins D. Gene expression profiles and molecular markers to predict recurrence of Dukes' B colon cancer. *J. Clin. Oncol.* (2004), 22, 1564-1571.
197. Kim IJ, Kang HC, Jang SG, Kim K, Ahn SA, Yoon HJ, Yoon SN, Park JG. Oligonucleotide microarray analysis of distinct gene expression patterns in colorectal cancer tissues harboring BRAF and K-ras mutations. *Carcinogenesis* (2006), 27, 392-404.
198. Morson BC. Precancerous lesions of the colon and rectum. Classification and controversial issues. *JAMA* (1962), 179, 316-321.
199. Lane N, Lev R. Observations on the origin of adenomatous epithelium of the colon. Serial section of minute polyps in familial polyposis. *Cancer* (1963), 16, 751-764.
200. Arthur JF. Structure and significance of metaplastic nodules in the rectal mucosa. *J. Clin. Pathol.* (1968), 21, 735-743.
201. Jass JR. Serrated route to colorectal cancer: back street or super highway? *J. Pathol.* (2001), 193, 283-285.
202. Oldak M, Grzela T, Lazarczyk M, Malejczyk J, Skopinski P. Clinical aspects of disrupted Hedgehog signaling (Review). *Int. J. Mol. Med.* (2001), 8, 445-452.
203. Hahn H, Wicking C, Zaphiropoulos PG, Gailani MR, Shanley S, Chidambaram A, Vorechovsky I, Holmberg E, Unden AB, Gillies S, Negus K, Smyth I, Pressman C, Leffell DJ, Gerrard B, Goldstein AM, Dean M, Toftgard R, Chenevix-Trench G, Wainwright B, Bale AE. Mutations of the human homolog of Drosophila patched in the nevoid basal cell carcinoma syndrome. *Cell* (1996), 85, 841-851.
204. Johnson RL, Rothman AL, Xie J, Goodrich LV, Bare JW, Bonifas JM, Quinn AG, Myers RM, Cox DR, Epstein EH, Jr, Scott MP. Human homolog of patched, a candidate gene for the basal cell nevus syndrome. *Science* (1996), 272, 1668-1671.

205. Lindstrom E, Shimokawa T, Toftgard R, Zaphiropoulos PG. PTCH mutations: distribution and analyses. *Hum. Mutat.* (2006), 27, 215-219.
206. Harris AL. Hypoxia--a key regulatory factor in tumour growth. *Nat.Rev.Cancer* (2002), 2, 38-47.
207. Semenza GL. Targeting HIF-1 for cancer therapy. *Nat.Rev.Cancer* (2003), 3, 721-732.
208. Zhong H, De Marzo AM, Laughner E, Lim M, Hilton DA, Zagzag D, Buechler P, Isaacs WB, Semenza GL, Simons JW. Overexpression of hypoxia-inducible factor 1alpha in common human cancers and their metastases. *Cancer Res.* (1999), 59, 5830-5835.
209. Kullander K, Klein R. Mechanisms and functions of Eph and ephrin signalling. *Nat. Rev. Mol. Cell Biol.* (2002), 3, 475-486.
210. Rashid A, Houlihan PS, Booker S, Petersen GM, Giardiello FM, Hamilton SR. Phenotypic and molecular characteristics of hyperplastic polyposis. *Gastroenterology* (2000), 119, 323-332.
211. Batlle E, Henderson JT, Beghtel H, van den Born MM, Sancho E, Huls G, Meeldijk J, Robertson J, van de Wetering M, Pawson T, Clevers H. Beta-catenin and TCF mediate cell positioning in the intestinal epithelium by controlling the expression of EphB/ephrinB. *Cell* (2002), 111, 251-263.
212. Batlle E, Bacani J, Begthel H, Jonkheer S, Gregorieff A, van de Born M, Malats N, Sancho E, Boon E, Pawson T, Gallinger S, Pals S, Clevers H. EphB receptor activity suppresses colorectal cancer progression. *Nature* (2005), 435, 1126-1130.
213. Ambroise C, McLachlan GJ. Selection bias in gene extraction on the basis of microarray gene-expression data. *Proc. Natl. Acad. Sci. U. S. A.* (2002), 99, 6562-6566.
214. Bertucci F, Salas S, Eysteris S, Nasser V, Finetti P, Ginestier C, Charafe-Jauffret E, Loriod B, Bachelart L, Montfort J, Victorero G, Viret F, Ollendorff V, Fert V, Giovaninni M, Delpero JR, Nguyen C, Viens P, Monges G, Birnbaum D, Houlgatte R. Gene expression profiling of colon cancer by DNA microarrays and correlation with histoclinical parameters. *Oncogene* (2004), 23, 1377-1391.
215. Chung CH, Parker JS, Karaca G, Wu J, Funkhouser WK, Moore D, Butterfoss D, Xiang D, Zanation A, Yin X, Shockley WW, Weissler MC, Dressler LG, Shores CG, Yarbrough WG, Perou CM. Molecular classification of head and neck squamous cell carcinomas using patterns of gene expression. *Cancer Cell* (2004), 5, 489-500.
216. Notterman DA, Alon U, Sierk AJ, Levine AJ. Transcriptional gene expression profiles of colorectal adenoma, adenocarcinoma, and normal tissue examined by oligonucleotide arrays. *Cancer Res.* (2001), 61, 3124-3130.
217. Frederiksen CM, Knudsen S, Laurberg S, Orntoft TF. Classification of Dukes' B and C colorectal cancers using expression arrays. *J. Cancer Res. Clin. Oncol.* (2003), 129, 263-271.
218. Oba SM, Wang YJ, Song JP, Li ZY, Kobayashi K, Tsugane S, Hamada GS, Tanaka M, Sugimura H. Genomic structure and loss of heterozygosity of EPHB2 in colorectal cancer. *Cancer Lett.* (2001), 164, 97-104.

219. Guo DL, Zhang J, Yuen ST, Tsui WY, Chan AS, Ho C, Ji J, Leung SY, Chen X. Reduced expression of EphB2 that parallels invasion and metastasis in colorectal tumours. *Carcinogenesis* (2006), 27, 454-464.
220. Blom N, Gammeltoft S, Brunak S. Sequence and structure-based prediction of eukaryotic protein phosphorylation sites. *J. Mol. Biol.* (1999), 294, 1351-1362.
221. Jones PA, Baylin SB. The fundamental role of epigenetic events in cancer. *Nat. Rev. Genet.* (2002), 3, 415-428.
222. Bienz M, Clevers H. Linking colorectal cancer to Wnt signaling. *Cell* (2000), 103, 311-320.
223. Powell SM, Zilz N, Beazer-Barclay Y, Bryan TM, Hamilton SR, Thibodeau SN, Vogelstein B, Kinzler KW. APC mutations occur early during colorectal tumorigenesis. *Nature* (1992), 359, 235-237.
224. Gerdes H, Chen Q, Elahi AH, Sircar A, Goldberg E, Winawer D, Urmacher C, Winawer SJ, Jhanwar SC. Recurrent deletions involving chromosomes 1, 5, 17, and 18 in colorectal carcinoma: possible role in biological and clinical behavior of tumors. *Anticancer Res.* (1995), 15, 13-24.
225. Praml C, Finke LH, Herfarth C, Schlag P, Schwab M, Amler L. Deletion mapping defines different regions in 1p34.2-pter that may harbor genetic information related to human colorectal cancer. *Oncogene* (1995), 11, 1357-1362.
226. Sulman EP, Tang XX, Allen C, Biegel JA, Pleasure DE, Brodeur GM, Ikegaki N. ECK, a human EPH-related gene, maps to 1p36.1, a common region of alteration in human cancers. *Genomics* (1997), 40, 371-374.
227. Jubb AM, Zhong F, Bheddah S, Grabsch HI, Frantz GD, Mueller W, Kavi V, Quirke P, Polakis P, Koeppen H. EphB2 is a prognostic factor in colorectal cancer. *Clin. Cancer Res.* (2005), 11, 5181-5187.
228. Lugli A, Spichtin H, Maurer R, Mirlacher M, Kiefer J, Huusko P, Azorsa D, Terracciano L, Sauter G, Kallioniemi OP, Mousses S, Tornillo L. EphB2 expression across 138 human tumor types in a tissue microarray: high levels of expression in gastrointestinal cancers. *Clin. Cancer Res.* (2005), 11, 6450-6458.
229. Rhodes DR, Chinnaiyan AM. Integrative analysis of the cancer transcriptome. *Nat. Genet.* (2005), 37 Suppl, S31-S37.



ISBN 951-22-8337-9
ISBN 951-22-8338-7 (PDF)
ISSN 1795-2239
ISSN 1795-4584 (PDF)