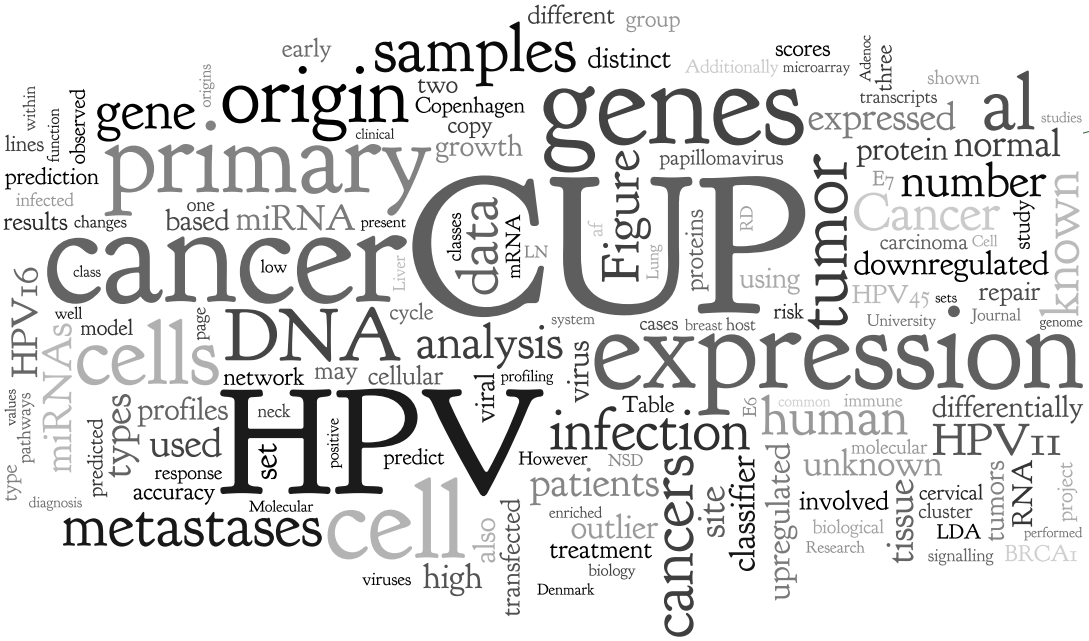
FACULTY OF SCIENCE
UNIVERSITY OF COPENHAGEN

PhD thesis

Bogumił Kaczkowski

Computational Cancer Biology: From Carcinogenesis to Metastasis



Academic advisors:
Prof. Anders Krogh
Asso. Prof. Ole Winther

Submitted: April, 2012

Computational Cancer Biology: From Carcinogenesis to Metastasis

by

Bogumił Kaczkowski

A dissertation submitted in partial satisfaction
of the requirements for the degree
of
Doctor of Philosophy in Bioinformatics
at
The PhD School of Science, Faculty of Science,
University of Copenhagen, Denmark.

April 2012

Supervisors
Prof. Anders Krogh
Asso. Prof. Ole Winther

The Bioinformatics Centre, and
Biotech Research and Innovation Centre
Department of Biology,
University of Copenhagen

To My Parents

Moim Rodzicom

Summary

Cancer biology is an exciting and dynamic field of research. Recently, it has become increasingly dependent on high throughput technologies to generate biological data. Due to massive amount of (often noisy) data, cancer biology research needs computational and machine learning methods to handle the data.

In the cell, the genomic information is stored in DNA in the nucleus. Messenger RNA (mRNA) is a working copy of DNA, generated by transcription; it passes the information to the cytoplasm, where it is transcribed to proteins. The process is called gene expression. MicroRNAs are small, non-protein coding RNAs, that regulate gene expression at post-transcriptional level. The high throughput profiling of DNA, mRNA and miRNAs provides valuable insight into different layers of biological activities within the cell.

In the first part of the thesis, I present the analysis of mRNA and miRNA expression in the cell model of Human Papilloma Virus (HPV) infection. HPV is responsible for 5% of cancers worldwide and better understanding of the molecular mechanisms of the infection can lead to improved treatment, diagnostics and prevention of these cancers. The generated mRNA expression profiles of the infected cells have been analyzed with integration of the available knowledge about cellular pathways and protein-protein interaction. The results show differential expression of many interesting genes, and deregulation of several vital signaling pathways, such as Interleukin-2, JAK-STAT, TGF-β, NOTCH and tyrosine kinase signaling. Profiling of microRNA expression within the model, showed differential expression of dozens of cellular microRNAs, and provided targets for further experimental research.

The second part of this thesis focuses on building a classifier that can predict the primary site of Cancers of Unknown Primary (CUP). CUP is generally a highly aggressive disease with poor prognosis and is forth most common cause of death by cancer in developed countries. The prediction of its primary site enables more targeted therapy, hopefully improving the response to the treatment. In the first project, mRNA expression profiles of more than 2400 tumor samples are used to train a classifier. The classifier is reasonably successful with predicting the origin of primary and metastatic samples However the expression profiles of 60 CUP patients appeared distinct from the primary tumor and metastases of known origin. Therefore CUP patients may require different diagnostic strategy and treatment. In the second project, DNA copy number data are used to build similar classifier. The results from primary tumors and cancer cell lines are promising

and open a way for development of a novel classifier of primary site.

Dansk resumé

Kræftbiologi er et spændende og dynamisk forskningsfelt. High throughput teknologier bliver i stigende grad vigtige til generering af biologisk data i forskningen. På grund af massive mængde af ofte støjfyldte data, er der inden for tumorbiologi behov for computerbaserede beregningsmodeller og automatiske læringsmetoder til at håndtere data.

I cellen er den genetiske information lagret i DNA i cellekernen. Messenger RNA (mRNA) er en arbejdskopi af DNA, der genereres af transkription. Således overleveres information til cytoplasmaet, hvor det transkriberes til proteiner. Processen kaldes genekspression. MicroRNA (miRNA) er små, ikke-proteinkodende RNA'er, der regulerer genekspression på post-transkriptionelt niveau. High thoughput profiler af DNA, mRNA og miRNA giver os værdifuld indsigt om forskellige niveauer af biologisk aktivitet i cellen.

I den første del af afhandlingen, præsenterer jeg en analyse af mRNA og miRNA ekspression i en cellemodel af Human Papilloma Virus (HPV) infektion. HPV er ansvarlig for 5% af kræfttilfælde i verden og bedre forståelse af infektionens molekylære mekanismer, kan føre til forbedret behandling. De genererede mRNA-ekspressionsprofiler af de inficerede celler er blevet analyseret med integration af tilgængelig viden om cellulære processer og protein-protein-interaktion. Resultaterne viser differentieret ekspression af mange interessante gener og ændret regulering af adskillige vigtige signaleringsveje, såsom interleukin-2, JAK-STAT, TGF-β, NOTCH og tyrosin kinase signalering. Profilering af microRNA ekspression i modellen viser differentieret ekspression af dusinvis af cellulære microRNA'er, og resulterer i opdagelser som kan målrette yderligere eksperimentel forskning.

Den anden del af afhandlingen fokuserer på at opbygge en klassificeringsmodel, der kan forudsige oprindelsesstedet for kræft med ukendt primær tumor (UPT). UPT er generelt en meget aggressiv sygdom med dårlig prognose og er den fjerde mest almindelige årsag til død af kræft i den industrialiserede del af verden. Forudsigelse af oprindelsesstedet muliggør mere målrettet terapi, som forhåbentlig kan forbedre behandlingseffekten. I det første projekt blev mRNA ekspressionsprofiler på mere end 2400 tumorprøver brugt til at træne klassificeringsmodellen. Klassificeringsmodellen er forholdsvis god til at forudsige oprindelsen af primære og metastatiske prøver. Imidlertid viste ekspressionsniveauerne fra profilerne af 60 UPT patienter sig at være forskellige fra den primære tumor og metastaser af kendt oprindelse. Resultaterne indikerer at en del af UPT tumorerne vil kræve en anden diagnostisk strategi og behandling. I det andet

UPT projekt anvendes DNA-copy number data til at opbygge en lignende klassificeringsmodel. Resultaterne fra primære tumorer og kræft-cellelinier er lovende og åbner op for udvikling af en ny klassificeringsmodel til at forudsige oprindelsesstedet for den primære tumor.

Research Papers

1. **Kaczkowski, B.*** , Rossing, M., Andersen, D. K., Dreher, A., Visser, M.V., Winther, O., Nielsen, F.C., and Norrild B.*
   Integrative analyses reveal novel strategies in HPV11,-16 and -45 early infection
   (*under review*)

2. Dreher, A., Rossing, M., **Kaczkowski, B.**, Andersen, D. K., Larsen, T. J., Christophersen, M. K., Nielsen, F. C., & Norrild B.*
   Differential expression of cellular microRNAs in HPV 11, -16, and -45 transfected cells.
   Biochemical and Biophysical Research Communications, 412(1), 2025. (2011).

3. Møller, A.K., **Kaczkowski, B.**, Borup, R., Henao, R., Vikeså, J., Krogh, A., Perell, K., Jensen, F., Winther, O., Nielsen, F.C., & Daugaard G.*
   Carcinomas of Unknown Primary Origin are Distinct from Metastasis of Known Origin.
   (*under review*)

4. **Kaczkowski, B.*** , Sinha, R., Nikolaus Schultz, N., Sander, C., Nielsen, F.C., and Winther, O.*
   Somatic copy-number alteration can help predict the tissue origin of cancers of unknown primary.
   (*draft*)

Corresponding author is marked with '*'. My name is listed in **bold** and the first authors' positions are underlined.

Research Papers not Part of the Thesis

1. Gaedcke, J., Grade, M., Camps, J., Søkilde, R., **Kaczkowski, B.**, Schetter, A. J., Difilippantonio, M.J., Harris, H.C., Ghadimi, B. M., Møller, S., Beissbarth, T. & Ried T.
   The rectal cancer microRNAome - microRNA expression in rectal cancer and matched normal mucosa
   (*under review*)

2. The Cancer Genome Atlas Research Network.
   Comprehensive Molecular Characterization of Human Colon and Rectal Cancer
   (*under review*)

3. Holst, L. M. B., **Kaczkowski, B.**, Glud, M., Futoma-Kazmierczak, E., Hansen, L. F., & Gniadecki, R.
   The microRNA molecular signature of atypic and common acquired melanocytic nevi: differential expression of miR-125b and let-7c.
   Experimental Dermatology, 20(3), 278280. (2011).

4. Podolska, A., **Kaczkowski, B.**, Kamp Busk, P., Søkilde, R., Litman, T., Fredholm, M., & Cirera, S.
   MicroRNA expression profiling of the porcine developing brain.
   PLoS ONE, 6(1), e14494. (2011)

5. Podolska, A., **Kaczkowski, B.**, Litman, T., Fredholm, M., & Cirera, S.
   How the RNA isolation method can affect microRNA microarray results.
   Acta biochimica Polonica, 58(4), 535540. (2011).

6. Søkilde, R., **Kaczkowski, B.**, Podolska, A., Cirera, S., Gorodkin, J., Møller, S., & Litman, T.
   Global microRNA analysis of the NCI-60 cancer cell panel.
   Molecular Cancer Therapeutics, 10(3), 375384. (2011).

7. Manfe, V., Holst, L. M., Rosbjerg, A., Kamstrup, M. R., **Kaczkowski, B.**, & Gniadecki, R.
   Changes in oncomiR expression in CTCL cell lines during apoptosis induced by Notch inhibition.
   Leukemia Research, 34(9), (2010).

8. Dreher, A., Rossing, M., **Kaczkowski, B.**, Nielsen, F. C., & Norrild, B.
   Differential expression of cellular microRNAs in HPV-11 transfected cells. An analysis by three different array platforms and qRT-PCR.
   Biochemical and Biophysical Research Communications, 403(3-4), 357362. (2010).

9. Rossing, M., **Kaczkowski, B.**, Futoma-Kazmierczak, E., Glud, M., Klausen, M., Faber, J., Nygaard, B., Kiss, K., Sørensen, C.H. , Nielsen, F.C., Bennedbæk, F.N., & Friis-Hansen L.
   A simple procedure for routine RNA extraction and miRNA array analyses from a single thyroid in vivofine needle aspirate.
   Scandinavian Journal of Clinical & Laboratory Investigation, 70(8), 529534. (2010).

10. Holst, L. M., **Kaczkowski, B.**, & Gniadecki, R.
    Reproducible pattern of microRNA in normal human skin.
    Experimental Dermatology, 19(8), e201e205. (2010).

11. Zibert, J. R., Løvendorf, M. B. L., Litman, T., Olsen, J. R., **Kaczkowski, B.**, & Skov, L.
    MicroRNAs and potential target interactions in psoriasis.
    Journal of Dermatological Science, 58(3), 177185. (2010).

12. **Kaczkowski, B.**, Torarinsson, E., Reiche, K., Havgaard, J. H., Stadler, P. F., & Gorodkin, J.
    Structural profiles of human miRNA families from pairwise clustering.
    Bioinformatics, 25(3), 291294. (2009).

Acknowledgements

I would like to thank all the people that have helped and supported me during the three-year PhD project; in particular I would like to express my gratitude to:

My supervisor Ole Winther, for day-to-day supervision, teaching me the machine learning and statictical methods, and always being optimistic about research.

Anders Krogh for being a great official supervisor, for the help with establishing contact and organizing external stay at cBio, MSKCC, NYC and keeping great atmosphere at BINF.

Jan Gorodkin and Thomas Litman, my previous supervisors, for supervising my master degree projects into first author publications in respectable journals. You paved my way to this PhD project.

Special thanks to my dear friend Angieszka Podolska for finding, and bringing my attention to the announcement of the PhD student position I finally occupied. For the research we did together, and for being a good friend.

I would like to thank the members of Computational Biology Center at MSKCC for welcoming me in NYC and six months of inspirational stay; in particular: Rileen Sinha, Nikolaus Schultz and Chris Sander for involving me in the TCGA project and for inspiriting me to use the TCGA data for CUP project. Anders Jacobsen for helping me with some practical aspects of my stay at MSKCC.

I would like to thank all my great collaborators, it has been a real pleasure to work and interact with you. Especially: Bodil Norrild and Maria Rossing for several successful projects and always-good atmosphere during our meetings. Anita Dreher for solid wet-lab work in HPV project. CUP project collaborators, especially Ricardo Henao, Rehannah Borup, Anne Møller, Finn Cilius Nielsen and Gedske Daugaard. Line Holst and Robert Gniadecki for collaborative research of miRNAs in melanoma. Lennart Friis Hansen and Kristina Døssing for involving me in carcinoid project.

I would like to thank Frederik Otzen Bagger for translacting the thesis summary to Danish and Nicolas Rapin for sharing some of his post doc experience.

Big thanks to Upper Binfers: Albin Sandelin, Frederik Bagger, Nicolas Rapin, Mette Boyd, Robin Andersson, Eivind Valen, Berit Lilje, Mette Jørgensen, Tomas Bertelsen, Yun Chen and Xiaobei Zhao for being social and awesome.

Tawny Abaniel, my friend from the finance world, who made a tremendous, non-profit work of copy editing the thesis introduction and some manuscripts.

Last but not least, I would like my girlfriend Karin, my family and my friends for company and support.

Contents

Chapter 1

Introduction

1.1 Cancer Biology

Cancer is a heterogeneous group of diseases characterized by uncontrolled growth of the cells. Cancers are generally classified by the type of cells or organ from which they originate. Since malignant growth can occur in virtually all locations of the body, there are over 100 different types of cancers. Cancer is an immensely complex and diverse disease; however, a set of characteristics are shared among almost all malignancies. Those characteristics, named hallmarks of cancer, are a unified set of capabilities that are acquired during tumorgenesis (Figure 1.1).

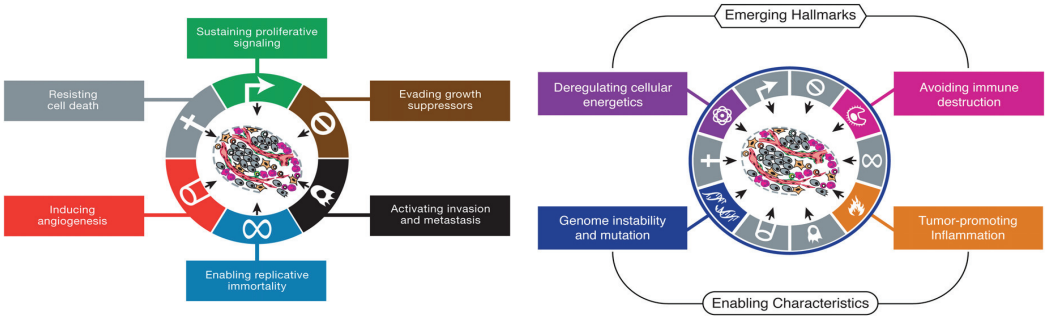


Figure 1.1: The Hallmarks of Cancer. Left: The original set of hallmarks of cancer (1). Right: Emerging Hallmarks and Enabling Characteristics. Reprinted by permission from Elsevier Ltd: Cell, 144(5), 646-674, copyright 2011 (2).

The originally proposed hallmarks of cancer are self-sufficiency in growth

signals, insensitivity to growth-inhibitory signals, evasion of programmed cell death, limitless replicative potential, sustained angiogenesis, and tissue invasion and metastasis (1). The list has been further extended with emerging hallmarks such as deregulating cellular energetics and avoiding immune response. Additionally, enabling characteristics were proposed, which are tumor promoting inflammation, and genome instability and mutation (2).

1.2 Causes of Cancer

Cancer is often described as the disease of the genome because it acquires the hallmarks of cancer through the accumulation of DNA mutations and genome instability (1). However, it is estimated that only 5-10% of cancer are caused by inherited traits and the remaining 90-95% are either caused or contributed to by environmental factors (Figure 1.2).

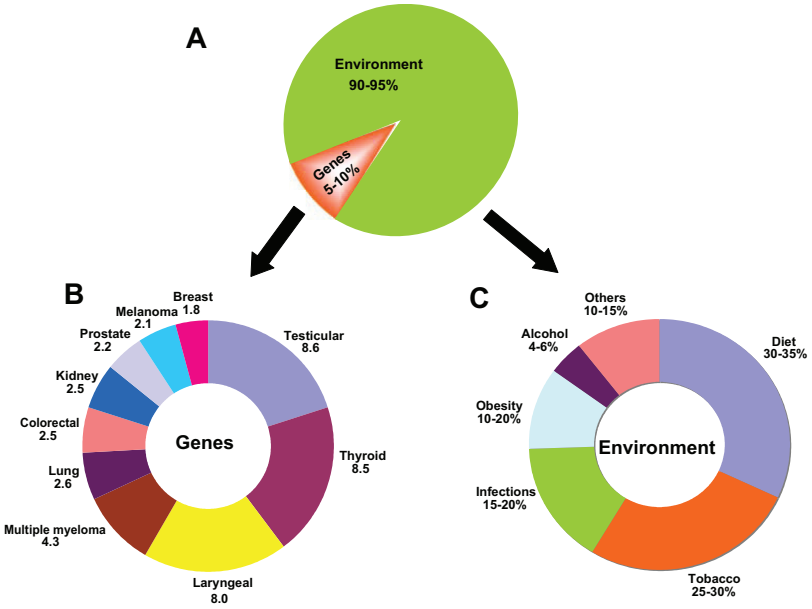


Figure 1.2: The impact of genes and environment on the development of cancer. A) The percentage contribution of genetic and environmental factors to cancer. B) Numbers represent familial risk ratios - an age-adjusted risk ratio to first-degree relatives of cases compared with the general population. C) Numbers represent the attributable-fraction of cancer deaths due to the specified environmental risk factor (3).

Worldwide, around 18% of cancers are caused by infections including viruses, bacteria, and parasites. Oncogenic viruses are responsible for most of those cancers. Human Papillomaviruses (HPV) cause 5.2% of all cancers, Hepatitis B and C viruses - 4.9%, Epstein-Barr virus - 1%, Human Immunodeficiency Virus (HIV) together with the human herpes virus 8 - 0.9% (4). The oncogenic viruses interfere with crucial cellular pathways of the host cells and affect cell growth, immortalization, genetic stability, cell cycle progression, and apoptosis (Figure 1.3). Thus, the infected cells acquire the hallmark of cancers, which lead to tumorgenesis. Chapter 2 of the thesis is focused on the effect of HPV genome on the gene expression of infected cells.

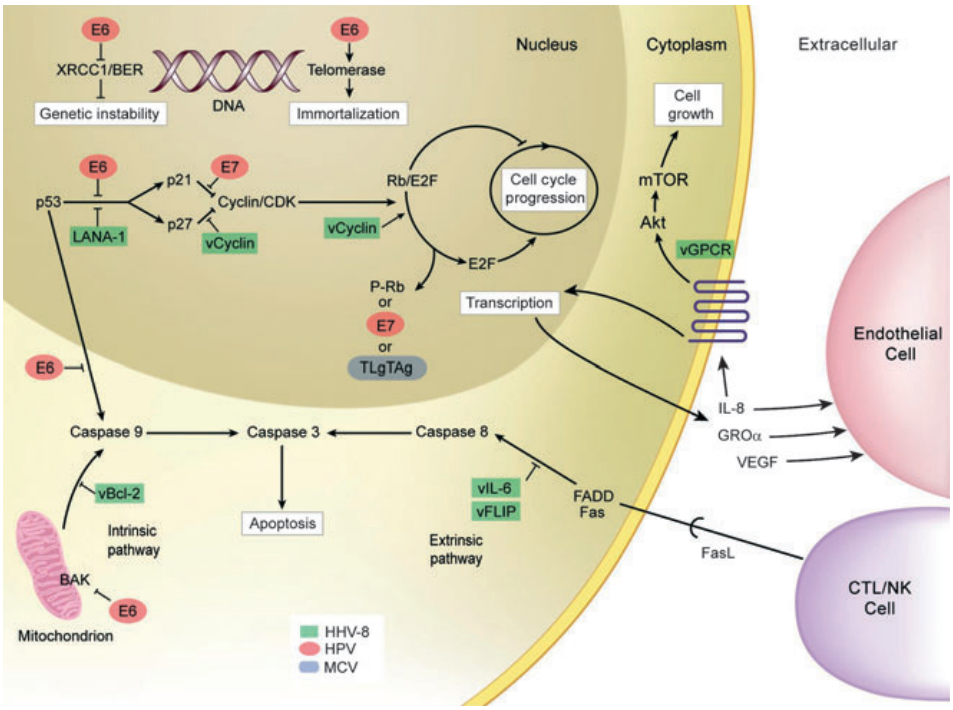


Figure 1.3: Example of how different viruses can impact cellular pathways and lead to carcinogenesis, here in nonmelanoma skin cancer. Reprinted by permission from John Wiley & Sons, Inc: Br. J. Dermatol. 164(6), 1201-1213, copyright 2011 (5).

1.3 Metastasis

Metastasis are the cause of more than 90% of cancer-related deaths (6). In order to create a metastatic lesion, the primary tumor cells invade the local environment and penetrate the walls of lymphatic and/or blood vessels, survive in the circulation, and invade and adapt to the environment of distant organ (Figure 1.4).

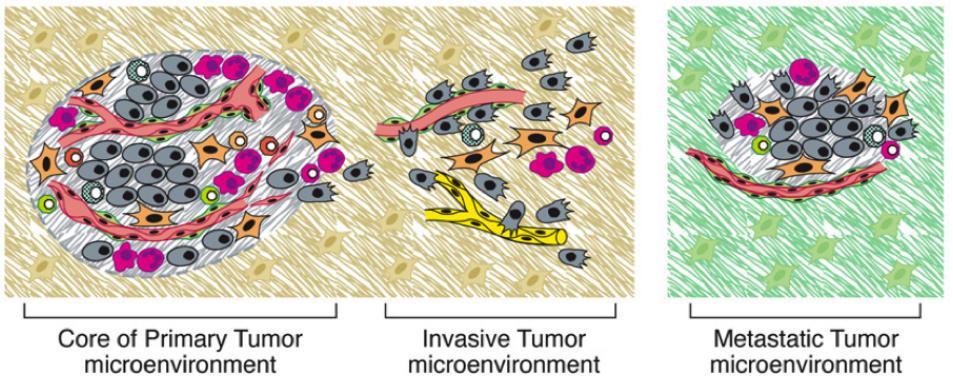


Figure 1.4: The cells of primary tumor and form metastasis in remote locations. Reprinted by permission from Elsevier Ltd: Cell, 144(5), 646-674, copyright 2011 (2).

Primary tumors differ on morphological and molecular level depending on the tissue origin and metastasis inherit the features of their primary sites. Different cancers respond differently to therapies, and the clinical management of cancers is chosen based on primary site. However, in 3-5% percent of all diagnosed cancers, no detectable primary site can be determined in the presence of metastatic lesions (7). The syndrome is called Cancer of Unknown Primary (CUP) and poses a significant challenge for the treatment (8). Chapter 3 of this thesis addresses the application of high-throughput genomic data to aid the diagnosis of CUP patients.

1.4 High-throughput Profiling in Cancers

An essential part of the results presented in this thesis comes from gene expression data. Microarray technology enables simultaneous measurement of thousands of messenger RNAs transcripts. Since all proteins in the cells

are produced by the translation of mRNA, the mRNA expression levels provide a good approximation of the abundance of proteins (Figure 1.5).

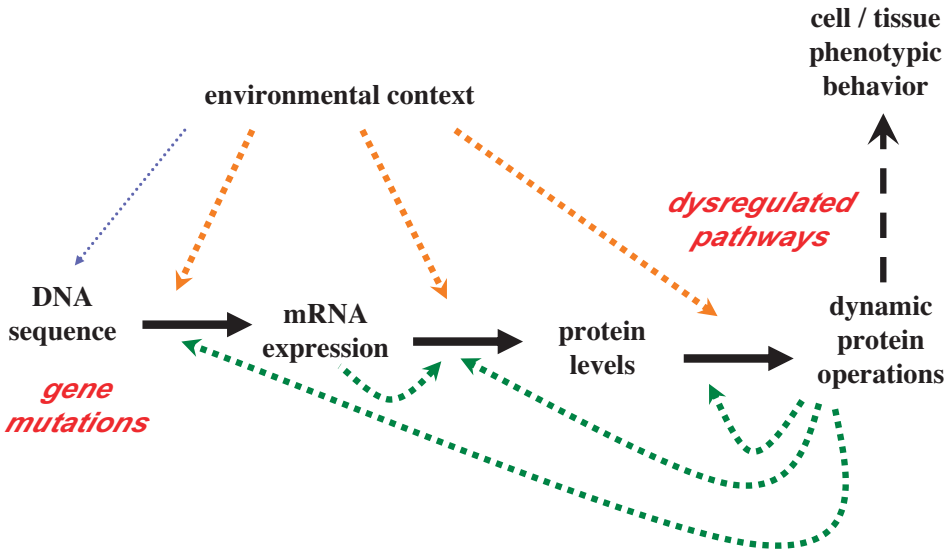


Figure 1.5: The transfer of genetic information from DNA, through mRNA to proteins. The genetic mutations are translated into dysregulation of cellular pathways, which in turn can have impact transcription and/or translation. The biological processes happen in and are influenced by environmental context. Reprinted by permission from Oxford University Press: Carcinogenesis, 2010, vol. 31(1), 2-8. (9)

Chapter 3 of the thesis shows how the microarray generated gene expression data set can be used to predict the primary site of Cancer of Unknown Primary. Additionally, the existing knowledge about protein-protein interaction and cellular pathways can be integrated in the analysis of gene expression data to obtain more biologically relevant and context-set results. This approach is presented in Chapter 2, where it is used to study the effect of Human Papilloma Virus genome of infected cells.

Chapter 2

The Cell Culture Model of Early HPV Infection

2.1 Introduction

Epidemiology of HPV

Cervical cancer is responsible for a quarter million deaths annually, which makes it the third largest cause of cancer deaths in women worldwide (http://globocan.iarc.fr/). Virtually all cervical cancers are caused by a persistent HPV infection (10). Human Papillomaviruses (HPVs) are a heterogeneous group of the papillomavirus family that consists of more than 120 species. They are small, DNA viruses that infect squamous epithelium of the skin or the mucous membranes and are best known for benign and malignant growth in the anogenital tract. However, HPVs can also cause less frequent neoplasms of airway mucosa i.e. Recurrent Respiratory Papillomas (RRP) and in the Head & Neck (H&N) region e.g. conjunctiva of the eyes, ear canal, nasal sinuses, and oral/pharyngeal cavity. Most HPV infections are cleared by the immune system within months and are often subclinical. The risk of a persistent and symptomatic HPV infection is significantly higher for individuals with a compromised immune system.

Life Cycle

The HPV life cycle starts with the virus particles infecting the exposed basal layer of epithelia. The virus is not able to penetrate the intact layers of epithelium; thus, a minor lesion is necessary for infection to occur. The reproductive cycle of HPV is tightly connected to differentiation stages of

the infected keratinocytes. HPV infects and persists in the basal layer where it also replicates its genome. The proteins are synthesized in the suprabasal and in the spinous layer. The viron assembly and release takes place in most differentiated layers of the epithelium i.e. granular and cornified layers (Figure 2.1) (11). The life cycle of the virus partially explains why it often takes months for the immune system to clear the infection. The HPV life cycle is restricted to keratinocytes (there is no viremia) and the lysis of the host cell and virion release happens in the upper cell layers prone for programmed cell death. This allows the infection to remain undetected for extended period of times leading to chronic and persistent infections (12).

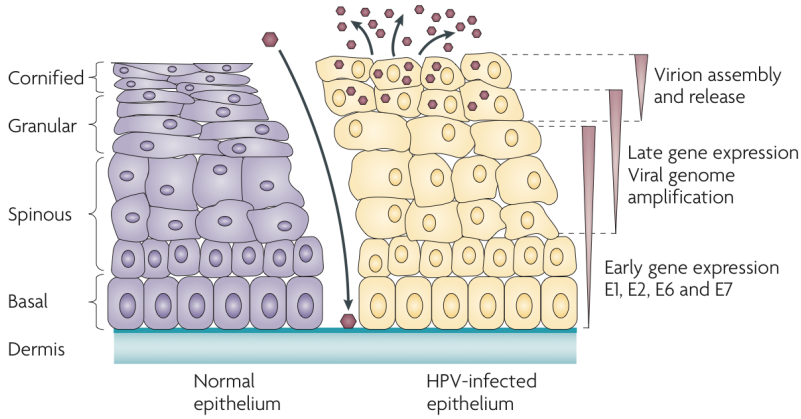


Figure 2.1: The cell cycle of HPV starts by infection of the basal layer of epithelium through microwound. The productive program of the virus is bound to the differentiation profile of keratinocytes. Early proteins E1, E2, E6 and E7 are expressed from early stages of infection. Reprinted by permission from Macmillan Publishers Ltd: Nat. Rev. Cancer, 10.1038/nrc2886, Copyright 2010 (11).

HPV-Driven Carcinogenesis of Cervical Cancer

While most of the HPV infections are spontaneously cleared by the immune system, a small percent of infection becomes persistent. The persistent HPV infection affects the host cells and induces the abnormal growth (dysplasia) and pre-malignant transformation (Figure 2.2). In cases of HPV infection in the cervix, this leads to low grade Cervical Intraepithelial Neoplasia (CIN-1). Most often, CIN-1 is cleared by the immune system and

regresses; however, in a fraction of cases it progresses to higher grades CIN-2 and CIN-3. The higher grade of CIN, the higher the risk the infection will persist. CIN-3 either regresses or it leads to chromosomal instability, accumulation of mutation, and integration of HPV DNA into host genome, which finally leads to the development of cancer.

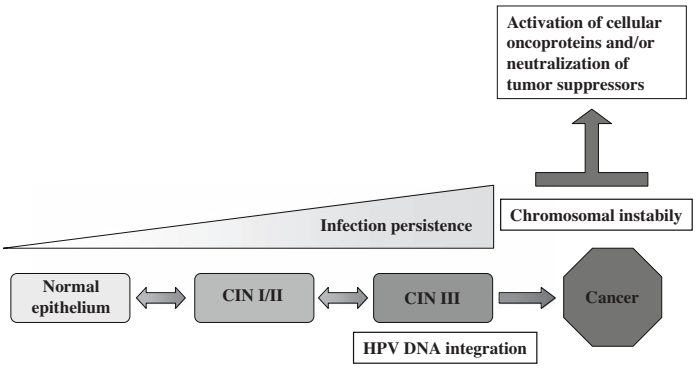


Figure 2.2: From Cervical Intraepithelial Neoplasia (CIN) to cancer. Most HPV infections are cleared by immune system within months. However in some cases, the infection persists leading to Cervical Intraepithelial Neoplasia (CIN) - premalignant transformation and abnormal growth (dysplasia) of epithelial cells. CIN can be eliminated by immune system or progress through more advanced stages and finally lead to cancer. Reprinted by permission from John Wiley & Sons, Inc: Virus Genes (2010) 40:1-13, (13)

Genomic Organization of HPV

The HPV genome consists of a double-stranded, circular DNA that is approximately 8,000 base pairs long. The genome of HPV has three main parts: Upstream Regulatory Region(URR), Early region (E) and Late region (L) (Figure 2.3). Early region encodes proteins expressed early during infection that are responsible for transformation of the host cell and replication of viral DNA genome (Figure 2.1). The function of two major oncogenes HPV E6 and HPV E7 are summarized in Figure 2.4 and Figure 2.5, respectively. The late region encodes the major (L1) and minor (L2) capsid proteins.

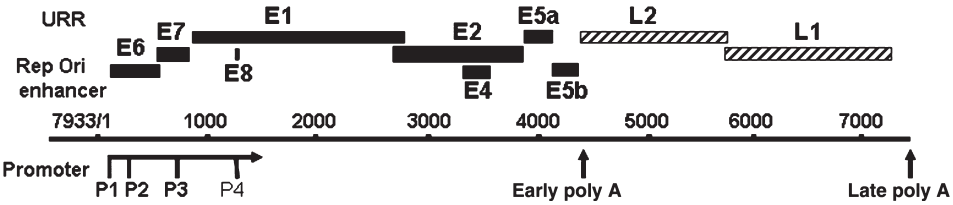


Figure 2.3: The genome of HPV11 consist of Upstream Regulatory Region (URR), Early region (E) and Late region (L). Reprinted by permission from John Wiley & Sons, Inc., APMIS, 118(6-7), 422:449 (14).

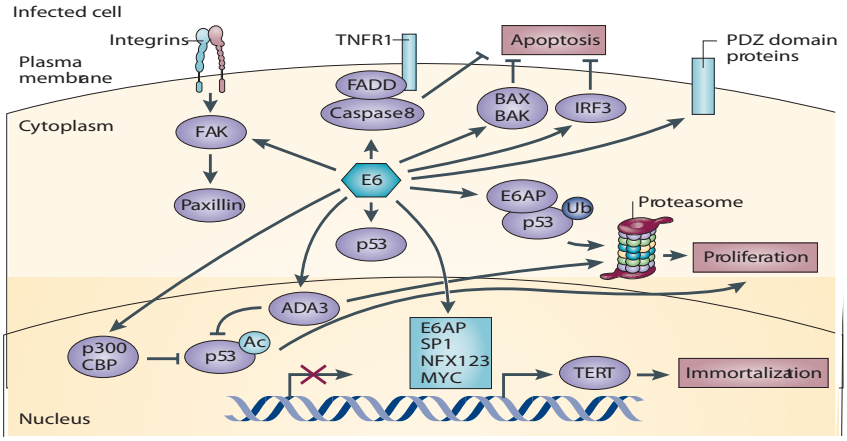


Figure 2.4: HPV E6 is a second major oncoprotein. It impacts several fundamental cellular activity such as apoptosis, proliferation and immortalization by targeting proteins like Caspase8, p53 and TERT. Reprinted by permission from Macmillan Publishers Ltd: Nat. Rev. Cancer, 10.1038/nrc2886, Copyright 2010 (11).

Prevention and Treatment

Despite the growing knowledge of HPV biology, there is still no treatment available to cure the infection. The current treatment is focused on managing the conditions caused by the virus such as cervical cell changes, genital warts, and papillomas. The infection of some HPV types can be prevented by vaccination. Cervarix and Gardasil protect against HPV16 and HPV18, which cause 70% of cervical cancers and anal, vaginal, and vulvar cancers (80%, 60%, and 40% of cases, respectively) (15). Additionally, Gardasil

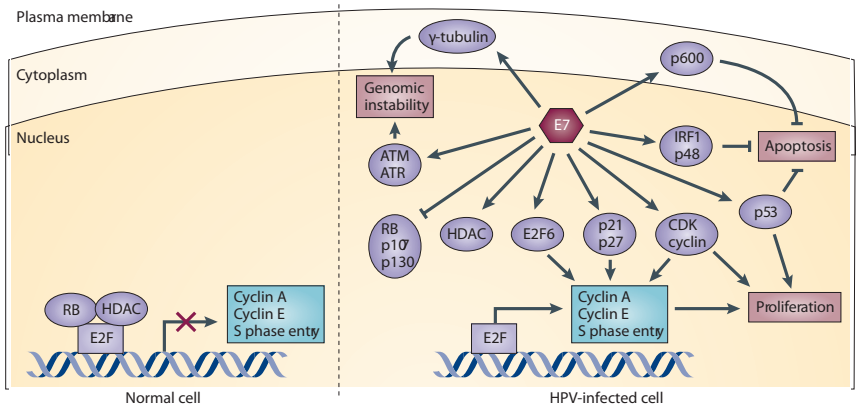


Figure 2.5: HPV E7 is a well-characterized oncoprotein. It interacts with several key cellular proteins e.g. HDAC, ATM, and p53 and influences crucial processes such as proliferation, apoptosis, and genomic stability. Reprinted by permission from Macmillan Publishers Ltd: Nat. Rev. Cancer , 10.1038/nrc2886 , Copyright 2010 (11).

protects against HPV6 and HPV11 that are responsible for 90% of genital warts. However, the vaccines do not protect from other high risk HPVs that cause the remaining 30% of cervical cancers.

2.2 Paper I

Integrative analyses reveal novel strategies in HPV11,-16 and -45 early infection

The Objectives

The early stages of HPV infection are not well understood and it is not clear what factors decide whether the infection is cleared by immune system or becomes persistent and leads to disease. The carcinogenesis caused by HPV infection has been extensively studied in tissues affected by premalignant transformation and abnormal growth (dysplasia) ((Figure 2.2, CIN1-3) and in cancer. However, those results are biased towards cases or persistent HPV infection while the infections cleared by immune system are not represented.

The early HPV infection and its progress cannot be studied in humans as one cannot infect a healthy individual to study the progress of the disease

(12). Animal models have been used to study the progression of papillomavirus infection (16). However, papillomaviruses are strictly specific to their host and have co-evolved with their hosts. Therefore, it is debatable to what extent the knowledge gained from animal models is relevant to human papillomaviruses.

In our project, we use the HaCaT cells transfected with three mucosal HPV types: HPV11, HPV16, and HPV45 genomes which were grown for 3 weeks as a model for early HPV infection. We performed the global gene expression profiling of the transfected cells using microarray platforms. The model enables a direct comparison of the HPV types on the same cellular background. Additionally, it allows us to study the effect of the HPV genome on the host cells, which gives insight into how the HPV proteins work together to establish and maintain the viral infection. We focused on 3 types of HPV: HPV11, HPV16, and HPV45. HPV11 is a non-oncogenic, low-risk type that causes recurrent respiratory papillomatosis (RRP) in the respiratory tract. In anogenital region, HPV11 causes genital warts (condyloma acuminata) and in rare cases leads to Buschke-Lowenstein giant condylomas. Even though the growth is not malignant, RRP can be life threatening due to airway obstruction. The condylomas are disfiguring and also cause severe discomfort. Additionally, the treatment of HPV11 diseases, especially RRP, can be extremely expensive (17). The second type, HPV16, is an oncogenic, high-risk virus and is the most common cause of cervical cancer. Additionally, the HPV16 infection can lead to vulvar, anal, vaginal, penile and some Head & Neck cancers. HPV45, the third studied HPV, is a high-risk type that is the fifth most prevalent oncogenic virus type detected in cervical lesions. In contrast to HPV11 and HPV16, HPV45 is not covered by the available prophylactic vaccines with the exception of a minor cross-protection obtained by Cervarix.

The Results

The differential expression analysis reveals involvement of genes not previously implicated in HPV biology like pregnancy-specific glycoprotein family (PSG) and ANKRD1. Additionally, genes implicated in biology of other viruses e.g. IFI44 and DDX60 were differentially expressed. Carcinogenesis related genes e.g. ABL2, MGLL and CYR61 were found upregulated in high-risk HPV16 and HPV45. The integration of the protein-protein interaction data lead to the discovery of differentially expressed networks of genes. We observed the suppression of DNA damage repair by HPV11 and HPV16 as well as downregulation of various cytoskeleton genes in all HPV

types. The viral infection affected various signaling pathways: Interleukin-2 (IL-2) signaling in HPV11, JAK-STAT signaling in HPV16, and Transforming Growth Factor-beta (TGF-β), NOTCH, and tyrosine kinase signaling in HPV45. The increased activity of JUN transcription factor was observed in HPV16 and HPV45 infected cells.

2.3 Paper II
Differential expression of cellular microRNAs in HPV 11, -16, and -45 transfected cells

MicroRNAs a group of small, non-protein coding RNAs, that regulate gene expression at posttranscriptional level. The regulation is conducted through binding to 3′UTR of messengerRNA transcript. MicroRNA has been implicated in almost all cellular processes. Some microRNA regulate the expression of oncogenes and tumor-suppressor genes and have a impact on cell cycle, apoptosis, cell migration and angiogenesis (18). MicroRNA has also been shown to be involved in regulation of viral growth and cancer progression in cervical cancer (19).

In this project, we investigated the global miRNA expression profiles in our cell culture model of early HPV infection, as described before. We observed the differential expression of 50, 22 and 47 differentially expressed miRNA in HPV11, -16 and -45, respectively. Thirteen miRNAs were differentially expressed by all 3 HPV types. Notably, miR-886-3p was the most donwregulated miRNA in all types, and we decided to conduct more experimental experiments to elucidate its function in HPV infection in the future. . Remarkably, miR-886-3p was also shown to be repressed in cancer (20)

Interestingly, miR-886-3p may not be a true microRNA, but a fragment of vault RNA (21) and may therefore require experimental approach different from the one used for micorRNAs to validate its function in HPV.

The results add up to our understanding of miRNA involvement in HPV biology, however the exact mechanisms are not yet undersood. Single miRNA can target multiple protein coding transcripts, and single gene can be regulated by multiple miRNAs. Additionally, miRNAs are believed to fine-tune or modulate the expression of the gene rather than switching it on and off. Therefore the interpretation of de-regulation of miRNA remains much more challenging than messengerRNAs.

Chapter 3

Carcinoma of Unknown Origin

3.1 Introduction

The Definition and Epidemiology of CUP

Carcinoma of Unknown Primary (CUP) accounts for 3-5% of all new cancer cases and is among the 10 most common malignancies in developed societies (7). And due to the aggressiveness of the disease, it is the fourth most common cause of cancer deaths in both sexes (22). By definition, CUPs are metastatic cancers with no clinically detectable primary tumor site. Thus, the disease is a syndrome representing many types of cancer. Alternatively, CUP can be characterized by the limitation of current clinical procedures to find the primary tumor, since in 75% of CUP patients, the primary site can be found during an autopsy (23). However, CUPs do share a particular clinical behavior. In general, CUPs exhibit early dissemination despite that the primary tumor is too small to be detected, as well as multiple metastasis and an unpredictable metastatic pattern with high aggressiveness of the disease (7).

Clinical management of CUP

There is no unified consensus concerning the diagnosis and treatment of CUP patients and the management of the disease varies among hospitals. Figure 3.1 outlines the basics of diagnosis and management of CUPs. In general terms, the diagnosis of metastatic cancer begins with a initial evaluation that aims at determining the primary site of cancer. This includes

review of patient history and symptoms, thorough physical examination, radiology (x-ray, CT and PET scans) and the analysis of blood and urine samples(for detailed description see (24; 25) . If the initial procedure fails to determine the anatomical primary site, the patient is diagnosed with CUP. Further histological examination with basic immunohistochemistry and molecular profiling analysis is chosen for the CUP patients based on the history and results of performed diagnostic tests (24).

The identification of origin and primary site-specific treatment was reported to significantly improve the outcome of CUP patients and the current management of CUP disease is commonly focused on identifying the most probable primary site and tailoring the treatment accordingly (Figure 3.1 panel B). However, according to the European Society for Medical Oncology (ESMO) guidelines, it remains unproven that the administration of primary site specific based on molecular assays, improves the patients outcome. This is reflected in the clinical management of CUP as recommended by ESM0 (Figure 3.1 panel B), which focuses on excluding non-CUP neoplasms, recognizing a specific subset of CUP and defining prognosis for non-specific CUPs by measuring serum lactate dehydrogenase (LDH) levels. It should be noted, however, that ESMO encourages the participation in clinical trials for site-specific therapy in patients with primary site highly suspected from immunohistochemistry or microarray analysis (25).

Molecular Profiling Approaches to Predict the Tissue Origin

High-throughput expression data has been widely used both in classification of known cancer types as well as in discovery of novel types/subtypes (clustering). The machine learning methods has also been applied to gene expression data of primary and metastatic tumor and the trained classifiers were shown to be effective in predicting the origin of metastases and primary cancers (8). The examples of classifier that has already been implemented in clinics and commercialized include messenger RNA based Pathwork Tissue of origin test and CancerType ID test, and miRNA based miRview mets.

Pathwork Tissue of Origin Test

The test is based on a custom designed Affymetrix chip (Pathchip) and has been developed on fresh frozen specimens (27), which was then extended to formalin fixed paraffin embedded (FFPE)samples (28). Both are approved by the FDA and have similar accuracy of about 90%. However, only the

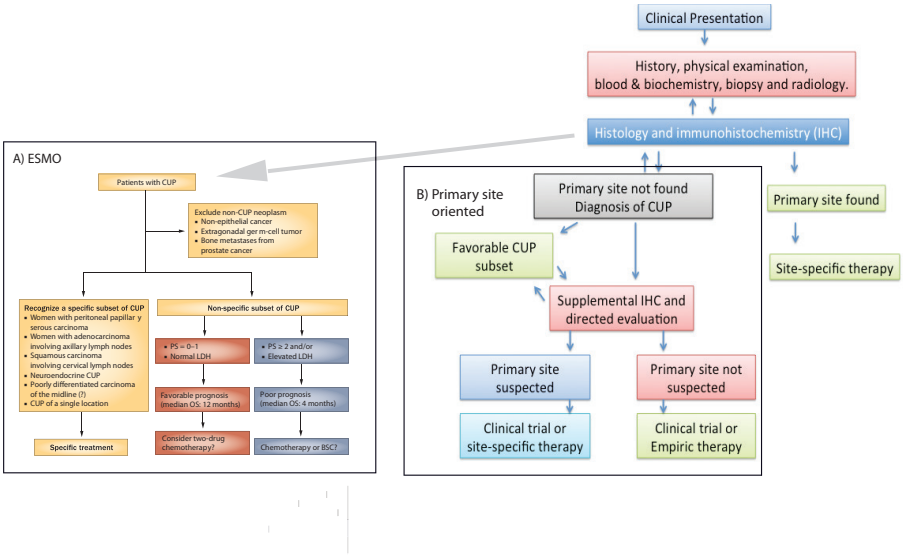


Figure 3.1: The outline of diagnosis and management of cancer of unknown origin patients. Panel A: reprinted by permission from Macmillan Publishers Ltd: Nat Rev Clin Oncol 2011 vol. 8 (12) pp. 701-710 (26). Panel B: based on figure from (24).

FFPE version has been commercialized, because paraffin embedding enables easier sample handling and shipping. The test covers 15 tissue types and uses 2000 mRNA transcripts out of 18,400 present on the chip. Recently, the Pathwork Tissue of origin test has also been used to predict the cancer origin from the cell blocks prepared from cytologic body fluid specimens (29).

CancerType ID Test

CancerType ID is another assay, which uses gene expression profiles to predict the origin of CUPs. Originally developed using fresh frozen samples on a micro-array platform, it was then adapted to be used with FFPE samples and qRT-PCR platform (30). The test uses 92 mRNA transcripts profiles and is claimed to distinguish 30 tumor type with overall accuracy of 82% (30). The test is offered as laboratory-developed tested (LTD).

miRview Mets

The miRview mets assay is based on the expression of 48 miRNAs measured by qRT-PCR and is offered as LTD. miRNAs are class of short (21nt) nonprotein-coding RNAs that regulate gene expression at post transcriptional level. The test predicts 17 different tissues. It uses two different classifiers (binary decision tree and k-nearest neighbors), and authors report accuracy of 90% in two-thirds of cases where the classifiers agree (31). In summary, the expression profiles of miRNA profiles yield lower performance of classification in comparison to messenger RNA. The authors highlight the 90% accuracy within the high confidence predictions (66% of cases), however the accuracy in remaining 33% of cases is 40% and the overall accuracy with single prediction is 59% (31; 8). It should be emphasized that the accuracies reported in for miRview mets and the mRNA based assays are calculated on cancers of known origins and are an (debatable) estimate of the expected accuracy in CUP patients.

Time

Time is an important factor in diagnosis and treatment of CUP patients due to very short life expectancy after the CUP diagnosis. The CUP patients have median overall survival of 4 or 12 months, depending on prognosis (favorable or poor respectively) (Figure 3.1). The current treatment of CUP patients in general and especially patients with poor prognosis is inefficient, and therefore best supportive care (BSC) is sometimes advised to avoid the devastating side effects of treatment. Due to complicated and time consuming diagnostic work-up, the patients may also not live long enough to receive the treatment. This leads to the situation where significant proportion of patients does not receive any treatment (Figure 3.2). Molecular assays, such as gene expression or miRNA profiling can offer an advantage, as those are usually high-throughput enough to test all the relevant genes/miRNAs at once, in contrast to laboratory medicine approaches which are performed sequentially. Additionally, standard CUP work-up is both labor intensive and requires highly qualified personnel. The molecular assays offer the possibility of greater automatization and can be performed by trained technician. In summary, if the molecular assays can match accuracy of standard work-up, they can be a faster alternative to traditional diagnostic work-up, and can lead to timely and more cost effective diagnosis and treatment for the patient.

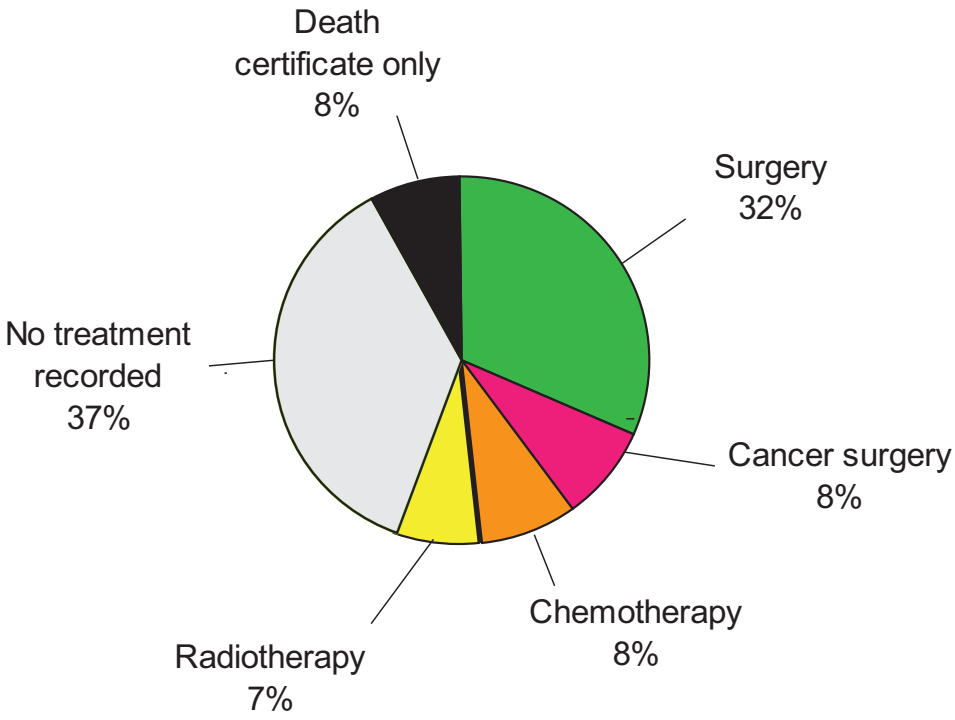


Figure 3.2: Treatment modalities of CUP patients received within 6 months of diagnosis (patients registered 2002-2006, collected by UK cancer registries). Remarkably, 37% of patients received no treatment. Evidently curative treatments i.e. chemotherapy, radiotherapy or cancer surgery constituted only 23% of recorded treatments. Eight percent were registered through their death certificate. Surgery represented 32% of treatments were surgical, however it is unknown if they were for diagnostic or therapeutic purposes. Modified from (32).

3.2 Paper III

Carcinomas of Unknown Primary Site are Distinct from Metastases of Known Origin

This publication is an outcome of an ongoing collaboration effort between the Bioinformatics Centre, the Department of Oncology and Center for Genomic Medicine at Rigshospitalet, Copenhagen University. The project aims at improving the diagnosis and treatment of CUP patients by devel-

oping statistical model based on high-throughput genomic data.

The aim of the project was to build a classifier that can predict the origin of Carcinoma of Unknown Primary (CUP) based on the gene expression data from a Affymetrix microarray platform and provide information about the biology of CUP. For this purpose, a data set was assembled which consisted of gene expression profiles with 1690 primary, 250 metastatic cancer, and 268 normal tissue samples. The majority of the data set was obtained from online depositories like GEO and ArrayExpress. The remaining profiles were generated at the University Hospital. Additionally, we performed the expression profiling of 60 samples from CUP patients from Danish hospitals.

During building and testing the classifier, we noticed that the expression profiles from CUP patients differ from the cancers of known origin. We reasoned that "forcing" a prediction of origin on some CUP expression profiles, which did not resemble any class present in our classifier, might be ill posed. The primary-site specific treatment, following such a prediction, may not be optimal for some of the CUP patients. To address the problem methodologically, we calculated the outlier score as a similarity measure of the sample in question to the closest cancer origin class. The rationale behind this is that the new sample deemed for prediction by the classifier, are compared to the pre-defined classes (cancer origins) of the classifier. If the expression profile of a new sample resemble one of the classes from the classifier, the outlier score should be low, if the new sample does not resemble any class from the classifier the outlier score should be high. We noticed the relationship between the outlier scores and the accuracy of the prediction. The training samples that had high outlier scores tended to be predicted with lower accuracy (more errors) than the samples with low outlier scores. Furthermore, the CUP patient samples had generally higher outlier scores than the metastases of known origins. We therefore concluded that CUPs are distinct from cancers of known primary and that the expected accuracy of the prediction of the origin is lower that the one estimated on the training set. These results suggest that the commonly used accuracy in primary and metastases of known origin, as an estimate for expected accuracy in CUP may not be true. We speculate that CUP with low outlier scores (similar to primary or metastasis) may be similar to cancer of known origins, where the standard work-up fails to determine the cancer origin. In those cases, the patient may benefit from the primary site specific therapy targeted against the predicted origin. On the other hand, CUPs with high outlier scores may be different biological entities

that require treatment specifically developed for CUP patients. However, more research is needed to support this claim. Please see the manuscript for more details on he methods, results and outcome of the study.

3.3 Paper IV

Somatic copy-number alteration can help predict the tissue origin of cancers of unknown primary

In this project, we propose the application of DNA copy number profiles to predict the origin of CUP. In normal cells the DNA sequence is present in two copies. Some loci show copy number polymorphisms, which means that the copy number of some DNA segments differs between germlines of different individuals (33). Copy number polymorphism affects around 12% of human DNA sequence (34). Copy number can also be acquired de Novo through mutations, and are referred to as Somatic Copy Number Alterations (SCNA). This happens rarely in normal cells, however the rate of mutation can increase drastically during cancer developments (Figure 3.3).

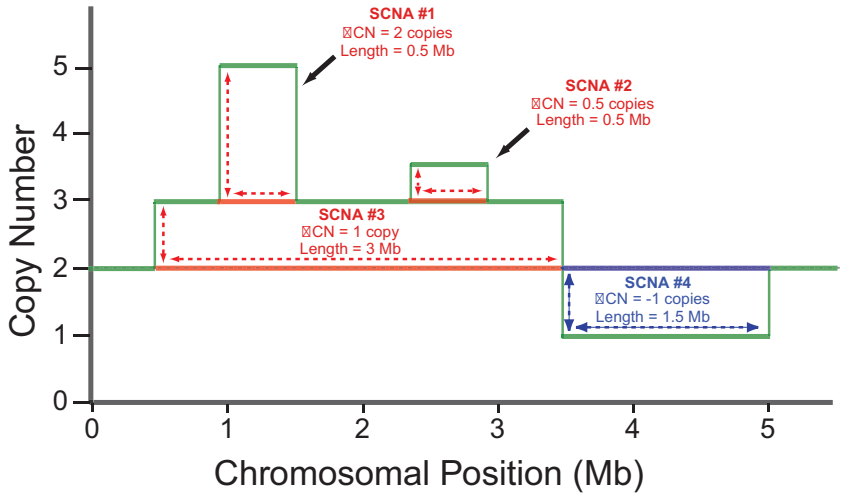


Figure 3.3: Somatic Copy Number Alterations (SCNA) are accumulated during development of cancer. Normally, two copies of DNA sequence are present in each cell. Reprinted by permission from Macmillan Publishers Ltd: Nature, 463(7283), 899:905. (33).

In this project, we used the DNA copy number data from The Cancer Genome Atlas project (TCGA). The data set consists of around 3500 cancer patient samples, representing 19 cancer origins. Similar number of normal tissue samples was obtained; those samples represent healthy tissue copy number profiles from the cancer patients. Additionally, we use the data from 639 samples from The Cancer Cell Line Encyclopedia. The aim of the project is to train classifier for CUP patients. Since the primary site of CUP patients is by definition unknown, there is no way of knowing how will the classifier, that is trained on the primary data samples, perform on CUP samples. The question is how universal/robust the classifier is. Or maybe is just suited, and will perform well on primary tumors. Cancer cell lines represent a biological entity that is quite different from primary tumors. First, cell lines are grown in culture, and they may also potentially represent cancer subtypes growable on petri dish. Second, most cancer cell lines are grown and passaged for extended period of time. Since cancer cells are often unstable and mutation-prone, cancer cell lines tend to accumulate a substantial amount of mutation on top of the mutations that were already present in the cancer, the cell line originated from. Indeed, our results show much higher rate of copy number changes in cancer compared to primary tumors. However, the classifier, which was trained on primary tumors, could still make high-confidence prediction for one third of the cell lines and was very accurate for several origins, such as glioma, kidney, head & neck, breast and colon. This result indicates that the primary site specific of acquired DNA copy number pattern is retained, at least in some origins. Therefore, the results suggest that copy number profile has a great potential to classify the primary origin of CUP cancers.

Chapter 4

Conclusion

In this thesis, I make use of the gene profiles to uncover biological relevant mechanisms that are disrupted in the case of viral infection or cancer. In the first part about HPV infection, the studies reveals valuable insights in the early infectious stages. There is a substantial amount of original results that broadens our understanding of HPV infection and parts of the results recapitulate what has been previously published for HPV. Other findings have been previously implicated in biology of other viruses. I believe that it is important to understand the mechanisms of early stages of HPV infection in order to develop treatments that can eradicate the viral infection before it leads to premalignant transformation and cancer. The current vaccination and treatment of HPV-caused conditions can be extremely costly, which limits its accessibility to most of the population worldwide. The results of vaccination will not be known until 20-30 years from now and treatment modalities are important for currently infected people. The development of small molecule drugs could substantially reduce the costs and help individuals already infected with the virus. This project leads to a substantial amount of results in the form of lists of differentially expressed microRNA, mRNA transcripts and affected protein-protein interaction network, which can be used as an inspiration for a more targeted, experimental research.
In the second part of the thesis, I developed a model that can predict the origin of cancer of unknown primary. The model uses a so called outlier score, and I believe that it would be interesting to examine if there is a relationship between the outlier scores from mRNA analysis and the overall survival of the CUP patients. Additionally, I would like to see if the CUPs are distinct from cancers of known origins in other aspects of biology, for example by using other high throughput data such as miRNA expression,

DNA copy number or mutations. There are no DNA copy number data available for CUP samples yet, and I look forward to get them someday. Secondly, finding the primary site of CUP patients and administering site-specific treatment is an indirect link of diagnosis to treatment. We are in need for more directed and personalized approaches i.e. prediction of optimal drug for a patient rather than the origin of the tumor. Based on our results, CUP patients with high outlier scores may benefit greatly from such an approach.

The microarray technology has now reached a mature state in terms of deployment, ease of use, costs and speed of analysis that makes it suitable for routine clinical use. Together with next generation sequencing, and other advanced data generating techniques, I believe that personalized medicine, and computer aided diagnostics is a reality that is close by.

Bibliography

[1] Hanahan D, Weinberg RA (2000) The hallmarks of cancer. Cell 100: 57–70.

[2] Hanahan D, Weinberg RA (2011) Hallmarks of cancer: the next generation. Cell 144: 646–674.

[3] Anand P, Kunnumakkara AB, Kunnumakara AB, Sundaram C, Harikumar KB, et al. (2008) Cancer is a preventable disease that requires major lifestyle changes. Pharmaceutical research 25: 2097–2116.

[4] Parkin DM (2006) The global health burden of infection-associated cancers in the year 2002. International journal of cancer Journal international du cancer 118: 3030–3044.

[5] Arron ST, Jennings L, Nindl I, Rosl F, Bouwes Bavinck JN, et al. (2011) Viral oncogenesis and its role in nonmelanoma skin cancer. The British journal of dermatology 164: 1201–1213.

[6] Sethi N, Kang Y (2011) Unravelling the complexity of metastasis — molecular understanding and targeted therapies. Nature reviews Cancer 11: 735–748.

[7] Pavlidis N, Pentheroudakis G (2010) Cancer of unknown primary site: 20 questions to be answered. Annals of oncology : official journal of the European Society for Medical Oncology / ESMO 21 Suppl 7: vii303–7.

[8] Takei H, Monzon FA (2011) Gene-expression assays and personalized cancer care: tissue-of-origin test for cancer of unknown primary origin. http://dxdoiorg/102217/pme1137 8: 429–436.

[9] Kreeger PK, Lauffenburger DA (2010) Cancer systems biology: a network modeling perspective. Carcinogenesis 31: 2–8.

[10] Walboomers JM, Jacobs MV, Manos MM, Bosch FX, Kummer JA, et al. (1999) Human papillomavirus is a necessary cause of invasive cervical cancer worldwide. The Journal of pathology 189: 12–19.

[11] Moody CA, Laimins LA (2010) Human papillomavirus oncoproteins: pathways to transformation. Nature reviews Cancer 10: 550–560.

[12] Stanley MA (2009) Immune responses to human papilloma viruses. The Indian journal of medical research 130: 266–276.

[13] Ghittoni R, Accardi R, Hasan U, Gheit T, Sylla B, et al. (2010) The biological properties of E6 and E7 oncoproteins from human papillomaviruses. Virus genes 40: 1–13.

[14] CHOW LT, BROKER TR, STEINBERG BM (2010) The natural history of human papillomavirus infections of the mucosal epithelia. APMIS 118: 422–449.

[15] De Vuyst H, Clifford GM, Nascimento MC, Madeleine MM, Franceschi S (2009) Prevalence and type distribution of human papillomavirus in carcinoma and intraepithelial neoplasia of the vulva, vagina and anus: a meta-analysis. International journal of cancer Journal international du cancer 124: 1626–1636.

[16] Nicholls PK, Moore PF, Anderson DM, Moore RA, Parry NR, et al. (2001) Regression of canine oral papillomas is associated with infiltration of CD4+ and CD8+ lymphocytes. Virology 283: 31–39.

[17] Lacey CJN, Lowndes CM, Shah KV (2006) Chapter 4: Burden and management of non-cancerous HPV-related conditions: HPV-6/11 disease. Vaccine 24 Suppl 3: S3/35–41.

[18] Blenkiron C, Miska EA (2007) miRNAs in cancer: approaches, aetiology, diagnostics and therapy. Human molecular genetics 16 Spec No 1: R106–13.

[19] Wang X, Tang S, Le SY, Lu R, Rader JS, et al. (2008) Aberrant expression of oncogenic and tumor-suppressive microRNAs in cervical cancer is required for cancer cell growth. PLoS ONE 3: e2557.

[20] Lee K, Kunkeaw N, Jeon SH, Lee I, Johnson BH, et al. (2011) Precursor miR-886, a novel noncoding RNA repressed in cancer, associates with PKR and modulates its activity. RNA (New York, NY) 17: 1076–1089.

[21] Stadler PF, Chen JJL, Hackermüller J, Hoffmann S, Horn F, et al. (2009) Evolution of vault RNAs. Molecular biology and evolution 26: 1975–1991.

[22] Pavlidis N, Pentheroudakis G (2012) Cancer of unknown primary site. Lancet .

[23] Pentheroudakis G, Greco FA, Pavlidis N (2009) Molecular assignment of tissue of origin in cancer of unknown primary may not predict response to therapy or outcome: a systematic literature review. Cancer treatment reviews 35: 221–227.

[24] Greco F (2010) Evolving understanding and current management of patients with cancer of unknown primary site. Community Oncology .

[25] Fizazi K, Greco FA, Pavlidis N, Pentheroudakis G, ESMO Guidelines Working Group (2011). Cancers of unknown primary site: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up.

[26] Massard C, Loriot Y, Fizazi K (2011) Carcinomas of an unknown primary origin–diagnosis and treatment. Nature reviews Clinical oncology 8: 701–710.

[27] Monzon FA, Lyons-Weiler M, Buturovic LJ, Rigl CT, Henner WD, et al. (2009) Multicenter validation of a 1,550-gene expression profile for identification of tumor tissue of origin. Journal of Clinical Oncology 27: 2503–2508.

[28] Pillai R, Deeter R, Rigl CT, Nystrom JS, Miller MH, et al. (2011) Validation and reproducibility of a microarray-based gene expression test for tumor identification in formalin-fixed, paraffin-embedded specimens. The Journal of molecular diagnostics : JMD 13: 48–56.

[29] Stancel GA, Coffey D, Alvarez K, Halks-Miller M, Lal A, et al. (2012) Identification of tissue of origin in body fluid specimens using a gene expression microarray assay. Cancer cytopathology 120: 62–70.

[30] Ma XJ, Patel R, Wang X, Salunga R, Murage J, et al. (2006) Molecular classification of human cancers using a 92-gene real-time quantitative polymerase chain reaction assay. Archives of pathology & laboratory medicine 130: 465–473.

[31] Rosenfeld N, Aharonov R, Meiri E, Rosenwald S, Spector Y, et al. (2008) MicroRNAs accurately identify cancer tissue origin. Nature biotechnology 26: 462–469.

[32] National Collaborating Centre for Cancer (UK) (2010) Diagnosis and Management of Metastatic Malignant Disease of Unknown Primary Origin. National Institute for Health and Clinical Excellence: Guidance. Cardiff (UK): National Collaborating Centre for Cancer (UK).

[33] Beroukhim R, Mermel CH, Porter D, Wei G, Raychaudhuri S, et al. (2010) The landscape of somatic copy-number alteration across human cancers. Nature 463: 899–905.

[34] Stankiewicz P (2010) Structural variation in the human genome and its role in disease. Annual review of medicine .

Appendix A

Paper I
Integrative analyses reveal novel strategies in HPV11,-16 and -45early infection

Integrative analyses reveal novel strategies in HPV11,-16 and -45 early infection

Bogumil Kaczkowski[a*], Maria Rossing[b], Ditte Andersen[c], Anita Dreher[c], Melissa A. Visser[c], Ole Winther[a,d], Finn Cilius Nielsen[b], Bodil Norrild[c*]

a) The Bioinformatics Centre, Department of Biology and Biomedical Research and Innovation Centre, Copenhagen University, Ole Maaloes Vej 5, 2200 Copenhagen, Denmark

b) Department of Clinical Biochemistry, Copenhagen University Hospital, Blegdamsvej 5, 2100 Copenhagen, Denmark

c) Institute of Cellular and Molecular Medicine, DNA Tumor Virus Laboratory, University of Copenhagen, Panum Institute, Blegdamsvej 3, 2200 Copenhagen, Denmark

d) DTU Informatics, Technical University of Denmark, 2800 Lyngby, Denmark

* Corresponding authors. Address:

a) Bogumil Kaczkowski. The Bioinformatics Centre, Department of Biology and Biomedical Research and Innovation Centre, Copenhagen University, Ole Maaloes Vej 5, 2200 Copenhagen, Denmark. E-mail address: bok@binf.ku.dk

b) Bodil Norrild. Panum Institute, Blegdamsvej 3C, Building 22.3., 2200 Copenhagen N, Denmark. E-mail address: bnorrild@sund.ku.dk

Running title: Early infection of HPV11, -16 and -45.
Keywords: HPV11 / HPV16 / HPV45 / cell model / gene expression
The final character count: 39,846

Abstract

The interaction between human papillomavirus (HPV) and host cells is not well understood. We investigate the early stage of HPV infections by global expression profiling in a cell model, in which HaCaT cells were transfected with low-risk HPV11 or high-risk HPV16 or HPV45 genomes. We report the differential expression of genes not previously implicated in HPV biology, such as the PSG family and ANKRD1, and of genes implicated in the biology of other viruses, e.g. MX1, IFI44 and DDX60. Carcinogenesis-related genes, e.g. ABL2, MGLL and CYR61, were upregulated by high-risk HPV16 and -45. The integrative analysis revealed the suppression of DNA repair by HPV11 and -16, and downregulation of cytoskeleton genes by all HPV types. Various signalling pathways were affected by the HPVs: IL-2 by HPV11; JAK-STAT by HPV16; and TGF-β, NOTCH and tyrosine kinase signalling by HPV45. Additionally, we observed increased activity of the transcription factor JUN in HPV16 and -45 infected cells. This study has uncovered novel strategies employed by low- and high-risk viruses to establish persistent infection and promote uncontrolled growth.

Introduction

The early stages of human papillomavirus (HPV) infection and virus-host cell interaction are not well understood. The DNA genomes of HPVs encode six or seven early proteins E7, E6, E1, E2, E4 and E5 (HPV11 encodes two E5 genes) and two late structural genes L2 and L1 which assemble into the viral capsid (Hausen 1999; CHOW et al. 2010; Doorbar et al. 1990). The two major oncoproteins E6 and E7 have been extensively studied. They are multifunctional and are mainly involved in the deregulation of cell cycle control (Antinore et al. 1996). The E7 of malignant viruses inactivates the tumour-suppressor protein pRB and E6 degrades the tumour-suppressor protein p53 via interaction with E6AP. This leads to cell cycle progression from the G1 into the S-phase (Ghittoni et al. 2010). However, other aspects of HPV biology remain uncharacterised. For example, there is only fragmented knowledge concerning the function of the remaining early proteins. Additionally, since most research focuses on one protein at a time, little is known about how the viral proteins work together to establish and maintain viral infection.
Several studies have used global gene expression profiling to study the impact of HPV on the cellular transcriptome. Some of these studies used biopsy specimens infected with low- or high-risk HPVs (Santegoets et al. 2011; Santin et al. 2005; DeVoti et al. 2008), while others used cell culture models bearing HPV genomes, e.g. W12 cell line with the HPV16 genome from natural infection (Alazawi et al. 2002), keratinocytes transfected with HPV31 (Chang & Laimins 2001) or HPV11 episomes (Thomas et al. 2001) and keratinocytes with integrated or episomal HPV18 genomes (Karstensen et al. 2006). In yet another approach, cell cultures have been transfected with HPV DNA encoding viral oncogenes. Gene expression analysis has been reported in cervical keratinocytes infected with retroviruses encoding HPV16 E6 and E7 (Nees et al. 2001), HPV16 E7-expressing keratinocytes (Boccardo et al. 2010) and keratinocytes infected with retroviruses encoding HPV18 E6 and E7 genes (Garner-Hamrick et al. 2004).
In the present study, we transfected a human keratinocyte cell line, HaCaT, with the genomes of three HPV types: HPV11, HPV16 and HPV45. HPV11, which causes benign papillomas, is a prototype for non-malignant, low-risk virus. HPV16 is the most common high-risk virus, which is responsible for the majority of cervical and some head and neck cancers. HPV45 is the fifth most prevalent malignant virus type found in cervical lesions (Guan et al. 2012) and is absent from the prophylactic vaccine implemented in Western countries. We conducted global expression profiling of the transfected cells and the control, followed by bioinformatics analysis integrating available knowledge of cellular pathways, protein-protein interactions and transcription factor binding sites. It is, to the best of our knowledge, the first direct comparison of global gene expression profiles of HPV11, -16 and -45 infected cells. Our study addresses the unexplored interactions between HPV and the host cell at early stages of infection. Secondly, we explore the contribution of the virus to uncontrolled growth and carcinogenesis. The expression profiles of 20,000 genes within our model are freely available online and constitute a valuable resourse for the HPV research community (data are being submitted to GEO).

Results

Cell culture growth is slowed by viral infection

The cell cultures were transfected with circular DNA genomes of HPV11, -16 or -45 and the control cultures were transfected with the plasmid encoding only the neomycin resistance. All three cell cultures transfected with HPV genomes grew more slowly than control-transfected cells (Figure 1). The reduction in absorbance within the first 24 hours in HPV positive cells indicates death of part of the cell population. This can be a result of virus killing the cells or cells switching on the apoptosis programme in response to viral infection. HPV11 slowed the growth of the cell culture to the greatest extent, while the HPV16 and -45 infected cultures grew at a similar rate. The presence of viral mRNA transcription was verified by qRT-PCR.

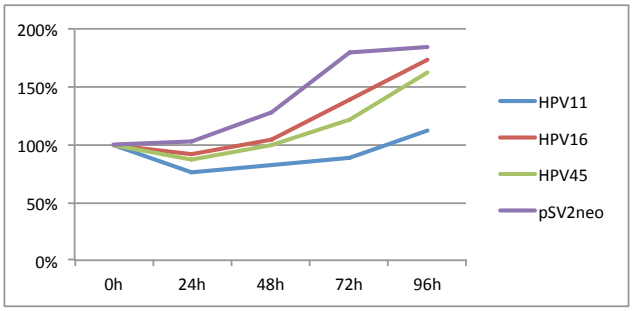


Figure 1 HaCaT cells were transfected with full genomic DNA and grown under G-418 selection for 2 weeks. Cells were seeded for measurement of proliferation using the MTT assay. Cultures were harvested for proliferation measurement every 24 hours for a total of 96 hours. The cell-cultures transfected with HPV genomes grew slower than control, and some cells died following the transfection.

Differential expression analysis

We profiled the global gene expression in the transfected cells and performed the analysis of differential expression between the cells transfected with each virus versus control. Throughout the article, we refer to the mRNA transcripts by their HUGO gene symbols. HPV11, -16 and -45 differentially expressed 391, 338 and 75 transcripts, respectively. The lower number of differentially expressed transcripts in HPV45 infected cells was due to failure of one array sample, which led to some loss of statistical power. Due to generally modest fold changes observed in differential expression (ranging from −4.1 to 4.5), we used a relatively stringent significance threshold (adjusted p-value <0.01). Tables of all the differentially expressed genes in cells infected with HPV11, -16 or -45 are available in Supplement 2. In order to improve the robustness of our results, we first focused on the genes that were differentially expressed by at least two HPV types and, in a further analysis, we integrated the prior knowledge of protein-protein interactions. This approach reduces the chances of false positives because the probability of observing a set of differentially expressed genes that interact at the protein level by chance only is considerably smaller.

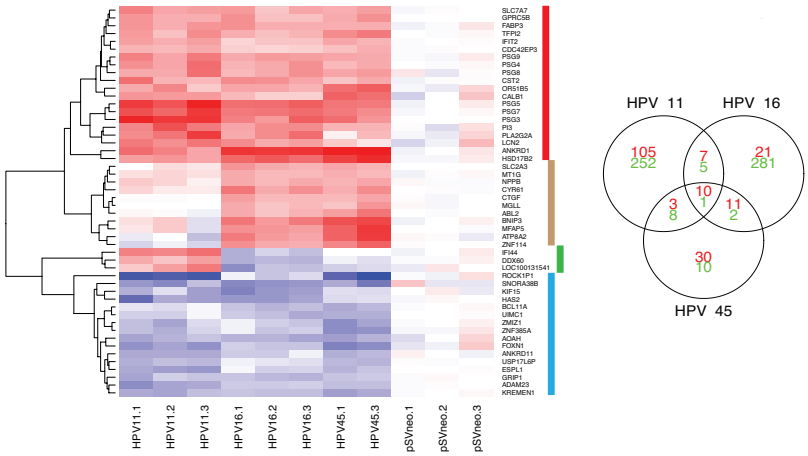


Figure 2 Venn diagram of differentially expressed genes (right panel) and the heatmap of the expression of genes (left panel) that are differentially expressed in at least two virus types. PSGs family, ANKRD1 and IFIT2 are upregulated by all three HPV types. ABL2, MGLL and CYR61, which have angiogenic and oncogenic potential, are upregulated in HPV16 and -45. IFI44 and DDX60 are upregulated in HPV11 and downregulated in HPV16 and -45. Genes downregulated in all HPVs include ANKRD11, AOAH, FOXN1, and oncogene BCL11A. The lists of all differentially expressed are available in Supplement 2.

We analysed the genes differentially expressed by more than one virus (Figure 2). Six members of the human pregnancy-specific glycoprotein (PSG) family (PSG3, 4, 5, 7, 8 and 9) were upregulated by all three HPV types. PSGs are known to be produced by placental syncytiotrophoblasts during pregnancy and have been shown to have an immune modulatory function and possibly proangiogenic effect (Lisboa et al. 2011). ANKRD1 and IFIT2 were also found to be upregulated by all the studied HPVs. ANKRD1 is induced by IL-1 and TNF-α (Chu et al. 1995), while IFIT2 is interferon-induced (Wathelet et al. 1988). This suggests that these genes were upregulated by the host cell in response to the infection.

The group of genes upregulated upon HPV16 and -45 infection included ABL2 and MGLL, which can promote cancer cell migration, invasion and tumour growth by regulating the levels of fatty acids that serve as signalling molecules (Nomura et al. 2011). Other upregulated genes were: BNIP3, involved in protecting cells from virally-induced cell death; CTGF, a pro-inflammatory cytokine also upregulated by hepatitis C virus E2 protein (Ming-Ju et al. 2011); and CYR61 (cysteine-rich, angiogenic inducer 61), a promoter of cell proliferation, chemotaxis and angiogenesis. This upregulation of proliferation, angiogenic and oncogenic genes by the high-risk HPV types suggests that these viruses may already exhibit their oncogenic potential during the early stages of infection.

Notably, IFI44 (interferon-induced protein 44) and DDX60 were upregulated by HPV11 and downregulated by HPV16 and -45 infections. IFI44 has previously been associated with hepatitis C virus infection (Kitamura et al. 1994) and exhibits anti-proliferative activity (Hallen et al. 2007). DDX60 is a helicase which has recently been shown to have anti-viral function (Schoggins et al. 2011; Miyashita et al. 2011). The group of genes that was downregulated by all the HPV types included: ANKRD11, which interacts with and enhances the activity of p53; AOAH, an immune response gene upregulated in response to swine fever virus (Gladue et al. 2010); and FOXN1, whose mutation in mice and rats causes hairlessness and a severely compromised immune system, and regulates keratin gene expression, which is consistent with the downregulation of a group of keratins in the HPV45 network (Figure 4C). Interestingly, the oncogene BCL11A was also downregulated by all the HPV types.

Integrative functional analysis

In order to ascertain the biological relevance of the differentially expressed mRNA transcripts, we integrated protein-protein interaction data to identify networks of differentially expressed genes that interact at the protein level. Throughout the article, we use the term 'gene', without referring specifically to mRNA or protein unless the distinction is necessary for clarification.

Assuming that functionally related genes will share an expression pattern within our model, we clustered all the genes that were differentially expressed by at least one virus type into six clusters. We then performed an enrichment analysis to ascertain the biological function of the genes within the clusters. We visualised the expression of differentially expressed genes across all three HPV types (Figure 3). Lists of genes for each of the six clusters are available in Supplement 3.

HPV11

The protein-protein interaction (PPI) network of genes differentially expressed upon HPV11 infection (Figure 4A; later referred to as the HPV11 network) was dominated by downregulated genes involved in the (mitotic) cell cycle (26 genes), mostly centred around CENPA, PIN1, PLK1 and MCM2, -5, -6, -7 proteins. We also observed that the genes of clusters 2 and 5 (Figure 3), which were downregulated by HPV11, were enriched in cell cycle genes. This result is consistent with the observed growth curves demonstrating the slowest growth of HPV11 infected cells.

The group of downregulated genes surrounding TP53 is involved in both the cell cycle and DNA damage repair. This group also connected the cell cycle group of genes to a group of genes involved in double-strand DNA repair, centred around BRCA1, RAD51 and FA family genes, FANCA, FANCE and FANCG, all of which were downregulated. In concordance with this result, clusters 2 and 5 (downregulated by HPV11) were enriched with 21 and 52 genes, respectively, from the BRCA1 signature. The signature of genes correlated with BRCA1 has been reported by others (Pujana et al. 2007).

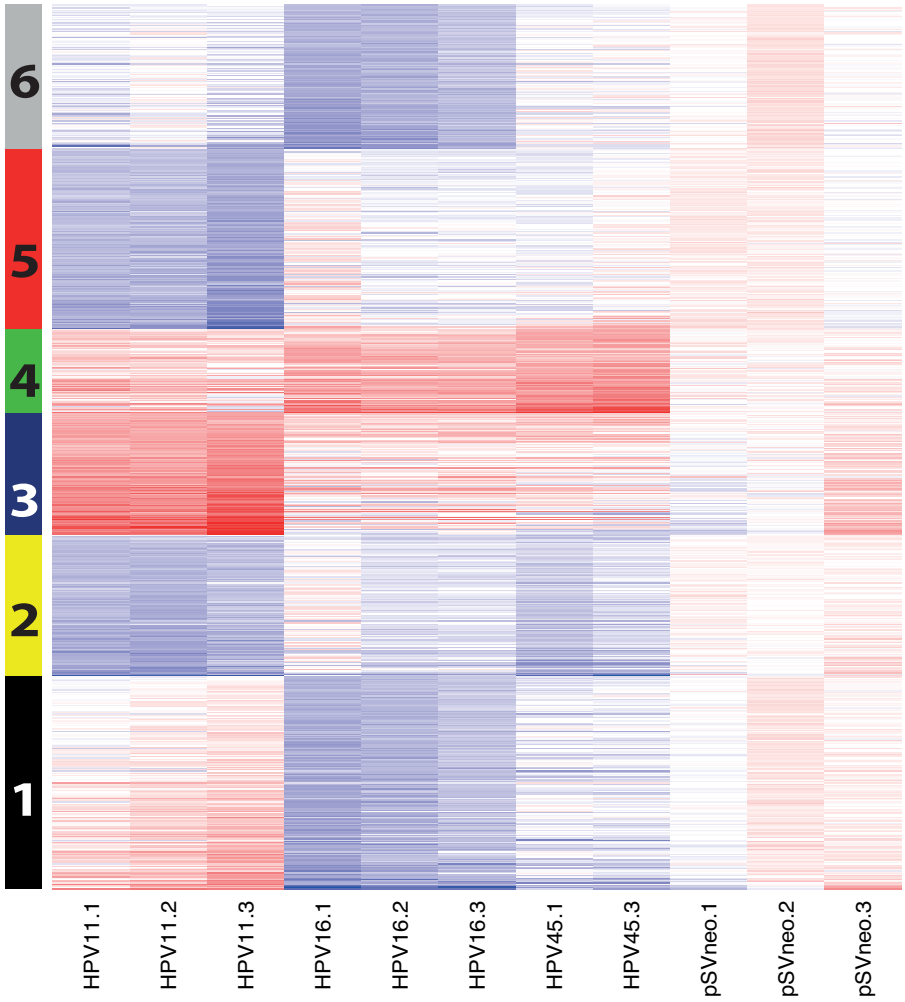


Figure 3 Heatmap showing the expression of differentially expressed genes upon transfection with HPV-11, -16 and -45 genomes. The genes were grouped into 6 clusters. Clusters 2, 5 and 6 are enriched in cell cycle and DNA repair genes from BRCA1 (cluster 2 and 5) and BRCA2 (cluster6) networks {Pujana:2007gl}. Cluster 4 is enriched in JUN transcription factor target genes and cluster 3 is enriched in interferon response genes. The genes of each of the six clusters are listed in supplement 3.

Additionally, cluster 2 contained five genes induced by BRCA1 (WELCSH_BRCA1_TARGETS_1_UP).
Interestingly, PARP1, which plays a role in the repair of single-stranded DNA, was downregulated and was connected to CENPA in a cell cycle sub-network. The lowered expression of BRCA1 and genes interacting with it or correlated with its expression suggests a reduction of the activity of DNA damage mediated by BRCA1 and related proteins. BRCA1 regulates transcription of POLD1 and CHAF1A, present in the network. POLD1, CHAF1A and other genes surrounding PCNA were all downregulated and their biological functions include DNA replication (several DNA polymerases), but also DNA repair.
Some of the upregulated genes present in the network encode proteins involved in the cellular response to HPV infection, including MX1 (an interferon-induced anti-viral protein), TLR4 (an activator of innate immunity) and pro-inflammatory PTGS2. Interestingly, MX1 interacts with FA family proteins and IFI44 (STRING 9.0). Other upregulated groups included extracellular proteins IGFBP3, CP and TFPI2, which interacted with plasminogen (PLG). Their function in HPV infection is elusive.

HPV16

The PPI network of genes affected by HPV16 infection was enriched with metabolic genes, most of them being downregulated (Figure 4B). Forty-six genes of the network are involved in biopolymer metabolic processes, 25 of which are involved in transcription/RNA metabolism. Within the network, the RNA metabolism genes were placed around RB1, SP1 and NCOR1, but also around JUN and SMAD2. SMAD2 was also connected to cell cycle genes STAG1 and STAG2. There was also a group of downregulated cell cycle genes surrounding YWHAG. Supporting results come from the cluster analysis, where cluster 6, downregulated by HPV16 (Figure 3), was enriched with genes involved in the mitotic cell cycle. This downregulation of metabolic and cell cycle genes suggests the slowing down of growth and proliferation processes. Thus, this result is consistent with the growth curves (Figure 1). The role of JUN in the network is unclear, being surrounded by nine genes involved in RNA metabolism, three of which were upregulated. JUN itself was upregulated, and we observed the upregulation of its targets in cluster 4.
Secondly, we observed a group of 11 genes associated with signalling pathways for cytokines and growth factors. The genes were centred around PIK3R1, RASA1 and JAK2. Seven of these genes are involved in JAK-STAT signalling. The downregulation of these genes may represent the reduction in growth signalling, which is consistent with the observed downregulation of metabolic and cell cycle genes. However, the presence of IL7 and IL7R, which also stimulate the B- and T-cell response, may also have implications for the immune response. Only one gene in the sub-network, GRAP, was upregulated.
Genes grouped around APC, NIN, VIM, KIAA1377 and ATRX are related to the cytoskeleton and vimentin (VIM) was connected to the actin regulators ROCK1 and ROCK2, which are involved in the formation of stress fibres (Katoh et al. 2011). Cluster 6 was enriched with genes involved in cytoskeleton structures.
Additionally, we observed a group of downregulated genes responsible for DNA repair. The genes were centred around NBN, ATR, BRCA1, BRCA2 and RAD50.

Further, cluster 6 (downregulated by HPV16) included 21 genes correlated with CHEK2 and 14 genes correlated with BRCA2.
Notably, several members of the PSG family, PSG3, 9 and 5, interacted with genes throughout the network.

HPV45

In the PPI network of genes differentially expressed upon HPV45 infection (Figure 4C), GRB2, which plays a role in signalling, was linked to two groups of surface proteins and to downstream signalling pathways. The first group consisted of membrane proteins EPHB2, EFNB1, SDCBP and MAGI3, and extracellular TGFA. EPHB2 was also connected to the upregulated oncogene ABL2 (tyrosine kinase). The second group included genes interacting with PLG, i.e. matrix remodelling proteins TFPI2 and MMP3 and cell surface glycoprotein F3, and genes interacting with PRNP, i.e. ADAM23, PVRL1 and PVRL4. Of note, the same interaction NEU3-GRB2-PRNP-PLG-TFPI2 was also observed in the HPV11 network. The genes interacting with PLG were upregulated, while the genes interacting with PRNP were downregulated by both viruses.
GRB2 was linked to a group of downregulated genes centred around members of the keratin family (KRT5, KRT6A and KRT14). Seven out of the nine proteins in this group are part of the cytoskeleton and four of these cytoskeletal genes are specifically involved in epidermis development. Additionally, four out of the nine genes (TRADD, TNFRSF1A (TNF receptor), PKP1 and DSP) are involved in apoptosis. GRB2, SHC3, AKT3, ADCY7 and PRKCD are part of the TRKA signalling from the plasma membrane.
The second part of the network was enriched with genes from the TGF-β pathway, which were placed around the INHBA-FST-SMAD9-SMAD3-CREBBP signalling chain. SMAD9 was linked to a group of eight extracellular/secreted proteins, seven of which were upregulated. SMAD9 interacted with SMAD3 which in turn was connected to a group of downregulated genes centred around CREBBP. Additionally, SMAD3 was linked to four downregulated members of the NOTCH pathway and three upregulated genes: JUN, SQSTM1 and IL1F7. JUN has binding sites on the promoters of four upregulated genes in the network: DDIT3 and HDAC9 in the SMAD3/CREBBP sub-network and F3 and TFPI2 in the PLG sub-network. Additionally, cluster 4, which was upregulated by HPV45, was enriched with genes having JUN binding sites in their promoter regions. The interaction chain SQSTM1-SMAD2/3-JUN-DDIT3 was shared between the high-risk types HPV16 and HPV45.
PRKCD, linked between CREBBP and PTK2B-GRB2, was also directly linked to plasma membrane proteins and the upregulated oncogene AKT3.

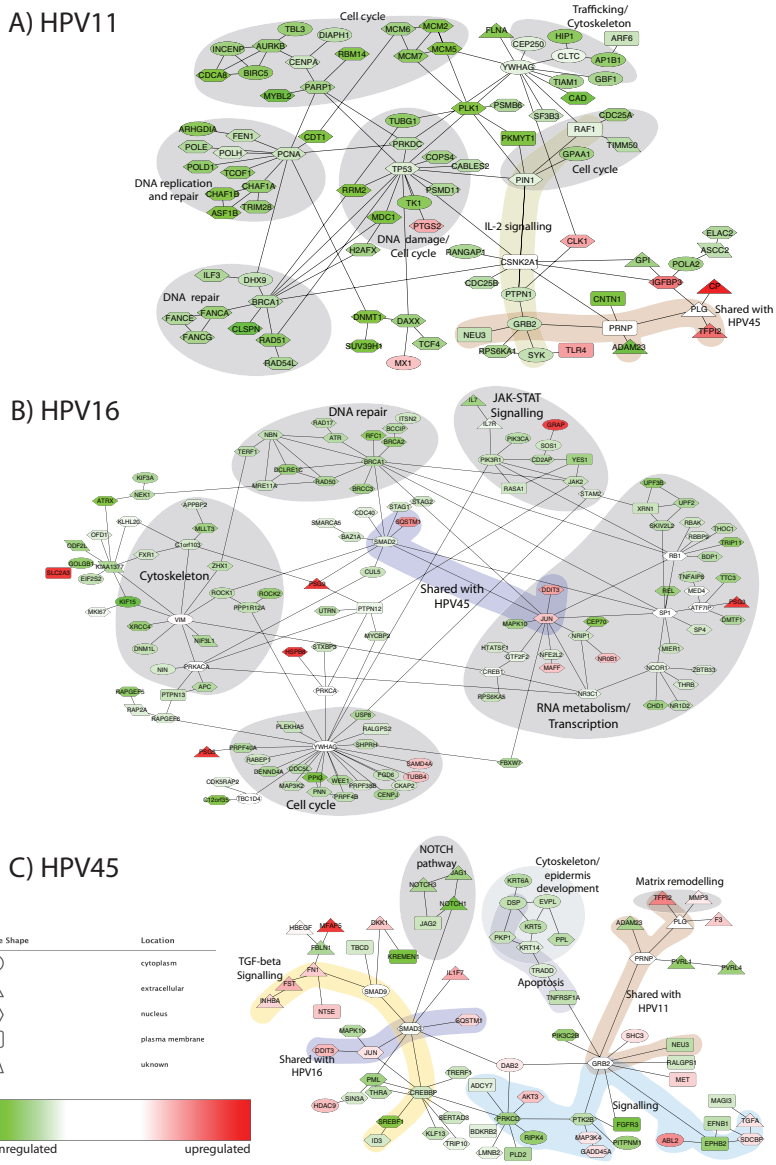


Figure 4 The networks represent the differentially expressed regions in human protein-protein interaction (PPI) network upon infection of A) HPV11, B) HPV16 and C) HPV45. The fold changes of differential expression are represented by color of the nodes: upregulated - red, downregulated - green and not changed – white. The subcellular localization is illustrated by shape of the nodes: ellipse – cytoplasm, triangle – extracellular, hexagon – nucleus, rectangular – plasma membrane, and unknown – parallelogram.

Discussion

Experimental model

The experimental model used here allows us to study and directly compare the differential expression of mRNAs mediated by different HPV types on the same cellular background. The model represents the early stages of viral infection and fills the gap between studies based on virus-induced cancers in human specimens and studies focused on single viral oncogenes like E6 and E7. Since only less than one percent of high-risk HPV infections leads to cancer, studies based on persistently infected tissues and infected cancers are biased towards the situations and mechanisms when the virus is not cleared by the immune system. We believe that a better understanding of the early stages of HPV infection can aid the development of treatment to clear out the infection in cases where it is not cleared spontaneously before it develops into malignancy. In contrast to studies which focus on single viral proteins like E6 or E7, our model represents the combined effect of all the early proteins working together in the infected cell, thus providing the possibility to observe the full effect of HPV genomes on host cells. It is especially useful for viral activities that rely on more than one viral protein. We have previously used the experimental model to study the impact of HPV11, -16 and -45 on cellular microRNA expression (Dreher et al. 2011; Dreher et al. 2010).

Early stage HPV infection slows the growth of host cells

Contrary to our expectations, transfection of host cells with the three studied HPV genomes has a negative impact on their growth. In concordance with this finding, the cluster analysis and PPI network show the downregulation of cell cycle and metabolic genes by HPV11 and -16, achieved by different sets of genes, e.g. genes of RNA metabolism are downregulated by HPV16 and genes of DNA metabolism are downregulated by HPV11. The downregulation of cell cycle genes is most profound in response to HPV11, which also induces the slowest cell culture growth. Notably, direct downregulation of cell cycle genes by HPV45 is not observed. We conclude that early HPV infection disrupts normal cellular processes, which hinders growth.

HPV11 and -16 suppress DNA repair

BRCA1, BRCA2 and CHEK2 are responsible for DNA repair, cell cycle arrest and apoptosis. The downregulation of BRCA1, genes correlated to BRCA1 expression as well as genes interacting with BRCA1 at the protein level, such as FANCA and RAD51, is mediated by HPV11. Transfection with HPV16 results in the downregulation of genes correlated to the expression of CHEK2 and BRCA2. Additionally, BRCA2, BRCA1, RAD50 and surrounding genes are also downregulated in the HPV16 PPI network. Different DNA repair genes are affected by HPV11 and -16. This suggests that HPVs can target different genes to achieve the same goal of lowering the capability of the host cell to repair its DNA.

The lowered expression of genes involved in DNA repair suggests a reduction of the activity of DNA damage repair and consequently a susceptibility to accumulation of mutations during repeated mitosis/DNA replication cycles. We propose that the

increased activity of the DNA repair system in late stages of HPV infection and during HPV-driven carcinogenesis reported in previous studies (Santegoets et al. 2011) is a secondary effect of substantial DNA damage present in the host cells and not induced by HPV. Contrarily, we suggest that HPV impairs the DNA damage detection and repair system early in infection, which allows the accumulation of mutations that can be beneficial for the initiation of a persistent viral infection and carcinogenesis process. The HPV16 E6 protein has also been found to interfere with single-strand repair by binding to XRCC1, leading the authors to conclude that the virus contributes to genomic instability (Iftner et al. 2002)

HPV infection and host cell immune response

It takes several months for the immune system to clear an HPV infection. Apart from avoiding immune surveillance by infecting only the basal layer of the epithelium and low expression of viral proteins, HPVs actively suppress the immune response (reviewed by (Stanley 2009)). It is not clear which factors lead to persistent infection that is considered a prerequisite for progression into cancer.

Our results from the PPI networks show the host cell response to infection both as an upregulation of anti-viral genes and a downregulation of genes involved in the immune response, which we deem to be part of the HPV strategy to avoid destruction by the immune system. In HPV11 positive cells, the pro-inflammatory PTGS2 and MX1 (which shows activity against influenza virus and VSV rhabdovirus and Hepatitis B virus (Haller et al. 2007)) are upregulated. However, DAXX, a protein which interacts with MX1, and TP53 which interacts with PTGS2 are downregulated. A similar pattern is seen for TLR4 (toll-like receptor 4). TLR4 is upregulated, yet SYK, GRB2 and other signalling genes functioning downstream in the signalling cascade are downregulated. We suggest that HPV may counteract the cellular response by downregulation of genes interacting with activated anti-viral proteins. Both MX1 and TLR4 have recently been shown to be involved in HPV infection (Reiser et al. 2011; Daud et al. 2011). However, MX1 was reported to be downregulated by high-risk HPVs (Reiser et al. 2011); MX1 expression is not changed by HPV16 and -45 in our experimental model. Notably, in HPV16 positive cells, a group of genes from the JAK-STAT signalling pathway is downregulated, including IL7, IL7R and JAK2. This finding is supported by a previous report showing that JAK2 is impaired by the E6 oncoprotein of HPV18 (Li et al. 1999). Our results point to downregulation of JAK-STAT signalling pathway genes as the viral action to hinder the interferon-driven immune response. Another upregulated gene likely to be involved in the anti-viral response is SQSTM1 (EBI3-associated protein of 60 kDa where EBI3 represents Epstein-Barr virus induced 3) (Devergne et al. 1996). In the HPV45 network, IL1F7, which suppresses the immune response, is upregulated. This is likely caused by the virus. IL1F7 requires SMAD3 for its function, to which it is connected in the network and SMAD3 is further connected to SQSTM1 and JUN.

Interestingly, the PSG family (PSG3, 4, 5, 7, 8 and 9) is upregulated by all three HPV types. We suggest that HPV uses the immune modulatory function of the PSG family to suppress the activity of the immune system. However, PSGs have also been suggested to be possible receptors for mouse hepatitis virus (Scanga et al. 1996;

Chen et al. 1995) and HIV (Blinov et al. 1994). The exact function of the PSGs in HPV infection remains unknown. Our study identified two further genes upregulated by all three HPV types, namely ANKRD1 and IFIT2. These are likely to be activated by the host cell in response to infection. In high-risk HPV16 and -45, BNIP3 and CTGF, which are known for their anti-viral activity (Ming-Ju et al. 2011), are also upregulated.
IFI44 and DDX60 are very interesting because they are upregulated by low-risk HPV11 and downregulated by high-risk HPV16 and -45 infections. Both genes have anti-viral activity reported in other viruses (Kitamura et al. 1994; Schoggins et al. 2011; Miyashita et al. 2011). IFI44 is induced by IFN-α and leads to microtubule aggregates in hepatitis C virus infected cells and overexpression inhibits cell proliferation (Kitamura et al. 1994; Hallen et al. 2007). DDX60 is a helicase and an integral part of the exosome. It is important for RNA stability (Miyashita et al. 2011). The mechanism of action of these genes in HPV infection is currently unknown, but if elucidated, could lead to a better understanding of the differences between low- and high-risk viruses.

The impact of HPV on the cytoskeleton

In epithelial cells, the cytoskeleton fibres maintain the cell structure. During viral infections, the actin fibres and microtubules are involved in both the uptake and release of virus particles (Taylor et al. 2011), whereas the function of the cytokeratins is less well understood. However, cytokeratins are important for the differentiation of epithelial cells and the fibres are composed of pairs of keratins defining the grade of differentiation. Keratins 5 and 14 are characteristic of the fibres in the stratified epithelial layer. Interestingly, these keratins together with keratin 6A are downregulated in HPV45 positive cells, indicating a transition of the HPV positive cells into a more simple epithelial cell (Cooper & Schermer 1985). In HPV16 positive cells, there is no differential regulation of cytokeratins; however, a cluster of genes linking to vimentin (VIM) includes the actin regulators ROCK1 and ROCK2 involved in the formation of stress fibres (Katoh et al. 2011). There are no differentially regulated cytoskeleton genes in the HPV11 PPI network. However, the pseudogene ROCK1P1 is the most downregulated gene in HPV11 positive cells. The downregulation of cytokeratins 6A, 10 and 13 (all from the stratified epithelium) has also been described in high-grade cervical lesions (HSIL) (Arnouk et al. 2009), thus validating the cell model system. Interestingly, FOXN1 induces keratin gene expression (Janes et al. 2004) and is downregulated by all three HPV types.

The impact of HPV on cellular signalling pathways

Our results imply that HPVs affect cellular signalling pathways by changing the expression of genes involved in signalling. Several signalling pathways are represented in the HPV45 PPI network. First, NOTCH pathway genes are downregulated. The downregulation of NOTCH1 expression has been reported in cervical carcinoma cells and is thought to be important in the late stages of HPV-induced carcinogenesis (Talora et al. 2002). We propose that the NOTCH pathway may already be modulated in the early stages of HPV infection.

Secondly, in the HPV45 network, TGF-α is upregulated and several genes of the TGF-β/SMAD signalling pathway are present. Interestingly, the genes of the extracellular part of TGF-β are upregulated, whereas the nuclear genes are downregulated. The TGF-β/SMAD signalling pathway is involved in many cellular processes; however, the molecular changes in the TGF-β/SMAD pathway in HPV infection are unclear. TGF-α is involved in mitogenesis and angiogenesis and is upregulated in several cancers (Hose et al. 2009).
An interesting chain of interaction proteins, TFPI2-PLG-PRNP-GRB2-NEU3, is present in the PPI networks of both low-risk HPV11 and high-risk HPV45. The limited information on these genes and their involvement in HPV biology makes it difficult to deduce their function. The presence of upregulated extracellular PLG, TFPI2 and MMP3, which are involved in matrix remodelling, may lead to changes on the cell surface and of the extracellular matrix, and possibly more invasive growth. The presence of a GRB2-interacting sub-network, which is involved in TLR4 and tyrosine kinase receptor signalling, points to the possible implication of all five proteins in the immune response signalling. Interestingly, GRAP, a GRB2-related adaptor protein, is upregulated in the HPV16 PPI network and interacts with downregulated genes of the JAK-STAT and PIK3R1 signalling pathway.

The impact of HPV on carcinogenesis

The aim of the study was to look for the early changes in HPV genome-bearing cells and not at HPV-driven carcinogenesis. However, we have identified some expression changes which we believe imprint the cells for future progression to malignant growth. For example, in response to high-risk HPV16 and -45, we observe increased activity of known oncogenes: JUN, upregulation of oncogene ABL2, cancer-promoting MGLL, angiogenic CYR61 and histone deacetylase HDAC9 (HPV45 only). Additionally, the impact of HPV45 infection on signalling pathways such as NOTCH, TGF-α and TGF-β could be implicated in cancer development. Finally, the observed downregulation of DNA repair genes could lead to the accumulation of mutations and contribute to the acquisition of other hallmarks of cancer.

Conclusion

The data presenting the genes differentially expressed in HPV11, -16 and -45 positive HaCaT cells reveal that the different viruses regulate essential pathways using different strategies. We believe that our experimental model is an interesting and useful link between profiling studies based on the oncogenes E7 and/or E6 and tissue studies. Furthermore, the data obtained in the present study open up new avenues of research where genes not previously identified as involved in HPV infections, e.g. IFI44 and DDX60 can be studied in depth. We believe that an understanding of the early stages of HPV infection could lead to the development of treatment strategies that would aid the clearing of persistent HPV infections.

Material and Methods

Transfection of cells and RNA extraction

HaCaT cells were cultured and transfected as previously described by Dreher et al. (Dreher et al. 2011). We conducted the transfection in three replicates for each virus type and control; however, one sample for HPV45 failed during microarray profiling. The transfected cells were grown under G418 sulphate (Invitrogen) selection for three weeks. Total RNA was extracted in TriZOL reagent according to the protocol of the manufacturer (Invitrogen).

MTT growth assay

We performed the measurement of cell proliferation with the Cell Proliferation Kit 1 (MTT) (Roche A/S, Hvidovre, Denmark) according to the manufacturer's protocol. HaCaT cells were transfected with HPV11, -16 or -45 full genomes and after a two-week selection cells were plated in five 96-well plates (10,000 cells per well). At time points 0h, 24h, 48h, 72h and 96h, 10µL MTT was added to each well of one plate and incubated at 37^0C for four hours. Thereafter, 100µL solubilising buffer was added to each well and the plate incubated overnight. Absorbance was measured at 570nm and 690nm for reference using a Synergy HT Multi-Mode Microplate Reader (Bio-Tek, Winooski, USA).

qRT-PCR validation

We used qRT-PCR to validate E7 transcription in RNA batches prepared from HPV11 transfected cells and E7 and E6 transcripts in RNA batches prepared from HPV16 transfected cells. The primers used were: **11-E7fwd**: 5'-gctggaagacttgttaccc-3' and **11-E7rev**: 5'-tcggacgttgctgtcacatcc-3', **16-E7fwd**: 5'-ttcggttgtgcgtacaaagc-3' and **16-E7rev**: 5'-agtgtgcccattaacaggtcttc-3', **16-E6fwd**: 5'-ctgcgacgtgaggtatatgacttt-3' and **16-E6rev**: 5'-acatacagcatatggattcccatct-3'. The qRT-PCR reaction was performed according to standard procedures with the following hybridisation temperatures: for 11-E7 55^0C, for 16-E6 and 16-E7 56^0C. The PCR reaction was continued for 40 cycles.

RNA labelling and hybridisation to microarray

Total RNA (100ng) was amplified and labelled using the Ambion WT Expression Kit (Applied Biosystems**)** according to the manufacturer's instructions. The labelled samples were hybridised to the Human Gene 1.0 ST GeneChip Array (Affymetrix, Santa Clara, CA, USA). Arrays were washed and stained with phycoerythrin-conjugated streptavidin (SAPE) using an Affymetrix Fluidics Station® 450 and subsequently scanned in an Affymetrix GeneArray® 2500 Scanner to generate fluorescent images according to the Affymetrix GeneChip® protocol. Cell intensity files (CEL files) were generated using Affymetrix GeneChip® Command Console® (AGCC) software.

Pre-processing of CEL files

The CEL files containing the array raw data were imported to the R environment and pre-processed using the *OLIGO* package from Bioconductor (Carvalho & Irizarry

2010). The expression data were normalised using the RMA (robust multi-array averaging) method and their probes were summarised by 'core' genes, resulting in log2 expression values for ~20,000 genes. A detailed description of all the analyses performed in R, as well as the source code to reproduce the results can be found in Supplement 1.

Differential expression

The analysis of differential expression was performed by means of the *Limma* package (Smyth 2004), which uses moderated t-statistics. Samples transfected by each virus type were compared to controls transfected with empty vectors. The empirical Bayes approach employed in *Limma* 'borrows' information about variance across samples and results in stable inference when the number of arrays is small (Smyth 2004). Therefore, we were able to perform the analysis of differential expression for HPV45 based on two samples. We did, however, observe some loss of statistical power, which resulted in a smaller number of differentially expressed genes for HPV45 compared to HPV11 and -16 (see Supplement 1 for the exact procedure and R code). The p-values were adjusted for multiple testing by the Benjamini-Hochberg correction method. The significance thresholds for differential expression were set to a) absolute log2 fold change above 0.6 (fold change >~1.5) and b) adjusted p-value <0.01.

Cluster analysis

All the genes that were differentially expressed in at least one virus type were clustered based on their expression profiles. $\mathbf{d = 1 - r}$ was used as the distance measure, where $\mathbf{r}$ is the Pearson correlation coefficient. The genes were clustered into six groups by PAM, a robust k-means-like clustering method. Each cluster was investigated for overlap with signatures from the Molecular Signatures Database v3.0, which includes the KEGG, GO, BioCarta and Reactome terms, as well as curated signatures of chemical and genetic perturbations. The significance of overlap/enrichment was calculated using the hypergeometric test implemented on the website (http://www.broadinstitute.org/gsea/msigdb/).

Integrated network analysis

The networks of differentially expressed networks of genes interacting at the protein-protein level were computed by means of *BioNet* (Beisser et al. 2010). This method identifies the differentially expressed functional module by integrating the p-values derived from the differential expression analysis and the human PPI network, as described in detail by (Dittrich et al. 2008). The heuristic approach was used to calculate an approximation to the optimal scoring sub-network. The significance thresholds were HPV11, FDR = 0.001; HPV16, FDR = 0.002; HPV45, FDR = 0.02. The networks were exported to and visualised by Cytoscape (Cline et al. 2007). The fold changes of differential expression were used to colour the nodes of the network; red was used for upregulated genes and green for downregulated genes. The function of the parts of the network was inferred using the enrichment/overlap analysis (http://www.broadinstitute.org/gsea/msigdb/) and literature search.

Acknowledgements

We wish to thank Susanne Smed for technical assistance, Tawny Abaniel for help with proofreading the manuscript and Nicolas Rapin for help with the preparation of the figures. Bogumil Kaczkowski was supported by the Novo Nordisk Foundation. Technician Melissa A. Visser was supported by a grant from the Læge Sophus Carl Emil Friis and hustru Olga Friis' Foundation. Ditte Andersen and Anita Dreher were supported by a student grant from the Danish Cancer Society. The work was funded by the following foundations: Beckett-Fonden, Fabrikant Einar Willumsens, Family Hede Nielsens, ML Jørgensen and Gunnar Hansen, and Lykfeldts.

Author contribution

BK, MR and BN designed the project. MR, DA, AD and MV performed the wet-lab experiment. BK performed the statistical/bioinformatics analysis and has been instrumental in the biological interpretation of the data. BK prepared figures 2-4 and supplementary materials. BK, MR and BN wrote the manuscript with ideas and contributions from OW and FN. OW and FN critically revised the manuscript. BN supervised the project. All authors approved the final version of the manuscript.

Competing interests

The authors declare no competing interests.

References

Alazawi, W., Pett, M., Arch, B., Scott, L., Freeman, T., Stanley, M. A., & Coleman, N. (2002). Changes in cervical keratinocyte gene expression associated with integration of human papillomavirus 16. *Cancer research*, *62*(23), 6959–6965.

Antinore, M. J., Birrer, M. J., Patel, D., Nader, L., & McCance, D. J. (1996). The human papillomavirus type 16 E7 gene product interacts with and trans-activates the AP1 family of transcription factors. *The EMBO journal*, *15*(8), 1950–1960.

Arnouk, H., Merkley, M. A., Podolsky, R. H., Stöppler, H., Santos, C., Alvarez, M., Mariategui, J., et al. (2009). Characterization of Molecular Markers Indicative of Cervical Cancer Progression. *Proteomics. Clinical applications*, *3*(5), 516–527. doi:10.1002/prca.200800068

Beisser, D., Klau, G. W., Dandekar, T., Muller, T., & Dittrich, M. T. (2010). BioNet: an R-Package for the functional analysis of biological networks. *Bioinformatics*, *26*(8), 1129–1130. doi:10.1093/bioinformatics/btq089

Blinov, V., Resenchuk, S., & Chirikova, G. (1994). *Possible role of pregnancy-specific glycoprotein (PSG) in mother-child HIV infection transfer.* … Conference On Aids.

Boccardo, E., Manzini Baldi, C. V., Carvalho, A. F., Rabachini, T., Torres, C., Barreta, L. A., Brentani, H., et al. (2010). Expression of human papillomavirus type 16 E7

oncoprotein alters keratinocytes expression profile in response to tumor necrosis factor. *Carcinogenesis*, *31*(3), 521–531. doi:10.1093/carcin/bgp333

Carvalho, B. S., & Irizarry, R. A. (2010). A framework for oligonucleotide microarray preprocessing. *Bioinformatics*, *26*(19), 2363–2367. doi:10.1093/bioinformatics/btq431

Chang, Y. E., & Laimins, L. A. (2001). Interferon-inducible genes are major targets of human papillomavirus type 31: insights from microarray analysis. *Disease markers*, *17*(3), 139–142.

Chen, D. S., Asanaka, M., Yokomori, K., Wang, F., Hwang, S. B., Li, H. P., & Lai, M. M. (1995). A pregnancy-specific glycoprotein is expressed in the brain and serves as a receptor for mouse hepatitis virus. *Proceedings of the National Academy of Sciences of the United States of America*, *92*(26), 12095–12099.

CHOW, L. T., BROKER, T. R., & STEINBERG, B. M. (2010). The natural history of human papillomavirus infections of the mucosal epithelia. *APMIS*, NATURAL HISTORY OF HPV INFECTIONS AND PATHOLOGY, *118*(6-7), 422–449. doi:10.1111/j.1600-0463.2010.02625.x

Chu, W., Burns, D. K., Swerlick, R. A., & Presky, D. H. (1995). Identification and characterization of a novel cytokine-inducible nuclear protein from human endothelial cells. *The Journal of biological chemistry*, *270*(17), 10236–10245.

Cline, M. S., Smoot, M., Cerami, E., Kuchinsky, A., Landys, N., Workman, C., Christmas, R., et al. (2007). Integration of biological networks and gene expression data using Cytoscape. *Nature protocols*, *2*(10), 2366–2382. doi:10.1038/nprot.2007.324

Cooper, D., & Schermer, A. (1985). Biology of disease. *Laboratory investigation*.

Daud, I. I., Scott, M. E., Ma, Y., Shiboski, S., Farhat, S., & Moscicki, A.-B. (2011). Association between toll-like receptor expression and human papillomavirus type 16 persistence. *International journal of cancer. Journal international du cancer*, *128*(4), 879–886. doi:10.1002/ijc.25400

Devergne, O., Hummel, M., Koeppen, H., Le Beau, M. M., Nathanson, E. C., Kieff, E., & Birkenbach, M. (1996). A novel interleukin-12 p40-related protein induced by latent Epstein-Barr virus infection in B lymphocytes. *Journal of Virology*, *70*(2), 1143–1153.

DeVoti, J. A., Rosenthal, D. W., Wu, R., Abramson, A. L., STEINBERG, B. M., & Bonagura, V. R. (2008). Immune dysregulation and tumor-associated gene changes in recurrent respiratory papillomatosis: a paired microarray analysis. *Molecular medicine (Cambridge, Mass.)*, *14*(9-10), 608–617. doi:10.2119/2008-00060.DeVoti

Dittrich, M. T., Klau, G. W., Rosenwald, A., Dandekar, T., & Muller, T. (2008). Identifying functional modules in protein-protein interaction networks: an integrated exact approach. *Bioinformatics*, *24*(13), i223–i231. doi:10.1093/bioinformatics/btn161

Doorbar, J., Parton, A., Hartley, K., Banks, L., Crook, T., Stanley, M., & Crawford, L. (1990). Detection of novel splicing patterns in a HPV16-containing keratinocyte cell line. *Virology*, *178*(1), 254–262.

Dreher, A., Rossing, M., Kaczkowski, B., Andersen, D. K., Larsen, T. J., Christophersen, M. K., Nielsen, F. C., et al. (2011). Differential expression of cellular microRNAs in

HPV 11, -16, and -45 transfected cells. *Biochemical and Biophysical Research Communications*, *412*(1), 20–25. Elsevier Inc. doi:10.1016/j.bbrc.2011.07.011

Dreher, A., Rossing, M., Kaczkowski, B., Nielsen, F. C., & Norrild, B. (2010). Differential expression of cellular microRNAs in HPV-11 transfected cells. An analysis by three different array platforms and qRT-PCR. *Biochemical and Biophysical Research Communications*, *403*(3-4), 357–362. Elsevier Inc. doi:10.1016/j.bbrc.2010.11.035

Garner-Hamrick, P. A., Fostel, J. M., Chien, W.-M., Banerjee, N. S., CHOW, L. T., BROKER, T. R., & Fisher, C. (2004). Global effects of human papillomavirus type 18 E6/E7 in an organotypic keratinocyte culture system. *Journal of Virology*, *78*(17), 9041–9050. doi:10.1128/JVI.78.17.9041-9050.2004

Ghittoni, R., Accardi, R., Hasan, U., Gheit, T., Sylla, B., & Tommasino, M. (2010). The biological properties of E6 and E7 oncoproteins from human papillomaviruses. *Virus genes*, *40*(1), 1–13. doi:10.1007/s11262-009-0412-8

Gladue, D. P., Zhu, J., Holinka, L. G., Fernandez-Sainz, I., Carrillo, C., Prarat, M. V., O'Donnell, V., et al. (2010). Patterns of gene expression in swine macrophages infected with classical swine fever virus detected by microarray. *Virus research*, *151*(1), 10–18. doi:10.1016/j.virusres.2010.03.007

Guan, P., Jones, R. H., & Li, N. (2012). Human papillomavirus (HPV) types in 115,789 HPV-positive women: A meta-analysis from cervical infection to cancer - Guan - International Journal of Cancer - Wiley Online Library. *... Journal of Cancer*.

Hallen, L. C., Burki, Y., Ebeling, M., Broger, C., Siegrist, F., Oroszlan-Szovik, K., Bohrmann, B., et al. (2007). Antiproliferative activity of the human IFN-alpha-inducible protein IFI44. *Journal of interferon & cytokine research : the official journal of the International Society for Interferon and Cytokine Research*, *27*(8), 675–680. doi:10.1089/jir.2007.0021

Haller, O., Staeheli, P., & Kochs, G. (2007). Interferon-induced Mx proteins in antiviral host defense. *Biochimie*, *89*(6-7), 812–818. doi:10.1016/j.biochi.2007.04.015

Hausen, zur, H. (1999). Immortalization of human cells and their malignant conversion by high risk human papillomavirus genotypes. *Seminars in cancer biology*, *9*(6), 405–411. doi:10.1006/scbi.1999.0144

Hose, D., Moreaux, J., Meissner, T., Seckinger, A., Goldschmidt, H., Benner, A., Mahtouk, K., et al. (2009). Induction of angiogenesis by normal and malignant plasma cells. *Blood*, *114*(1), 128–143. doi:10.1182/blood-2008-10-184226

Iftner, T., Elbel, M., Schopp, B., Hiller, T., Loizou, J. I., Caldecott, K. W., & Stubenrauch, F. (2002). Interference of papillomavirus E6 protein with single-strand break repair by interaction with XRCC1. *The EMBO journal*, *21*(17), 4741–4748.

Janes, S. M., Ofstad, T. A., Campbell, D. H., Watt, F. M., & Prowse, D. M. (2004). Transient activation of FOXN1 in keratinocytes induces a transcriptional programme that promotes terminal differentiation: contrasting roles of FOXN1 and Akt. *Journal of cell science*, *117*(Pt 18), 4157–4168. doi:10.1242/jcs.01302

Karstensen, B., Poppelreuther, S., Bonin, M., & Walter, M. (2006). Gene expression profiles reveal an upregulation of E2F and downregulation of interferon targets by HPV18 but no changes between keratinocytes with integrated or …. *Virology*.

Katoh, K., Kano, Y., & Noda, Y. (2011). Rho-associated kinase-dependent contraction

of stress fibres and the organization of focal adhesions. *Journal of the Royal Society, Interface / the Royal Society*, *8*(56), 305–311. doi:10.1098/rsif.2010.0419

Kitamura, A., Takahashi, K., Okajima, A., & Kitamura, N. (1994). Induction of the human gene for p44, a hepatitis-C-associated microtubular aggregate protein, by interferon-alpha/beta. *European journal of biochemistry / FEBS*, *224*(3), 877–883.

Li, S., Labrecque, S., Gauzzi, M. C., Cuddihy, A. R., Wong, A. H., Pellegrini, S., Matlashewski, G. J., et al. (1999). The human papilloma virus (HPV)-18 E6 oncoprotein physically associates with Tyk2 and impairs Jak-STAT activation by interferon-alpha. *Oncogene*, *18*(42), 5727–5737. doi:10.1038/sj.onc.1202960

Lisboa, F. A., Warren, J., Sulkowski, G., Aparicio, M., David, G., Zudaire, E., & Dveksler, G. S. (2011). Pregnancy-specific glycoprotein 1 induces endothelial tubulogenesis through interaction with cell surface proteoglycans. *The Journal of biological chemistry*, *286*(9), 7577–7586. doi:10.1074/jbc.M110.161810

Ming-Ju, H., Yih-Shou, H., Tzy-Yen, C., & Hui-Ling, C. (2011). Hepatitis C virus E2 protein induce reactive oxygen species (ROS)-related fibrogenesis in the HSC-T6 hepatic stellate cell line. *Journal of cellular biochemistry*, *112*(1), 233–243. doi:10.1002/jcb.22926

Miyashita, M., Oshiumi, H., Matsumoto, M., & Seya, T. (2011). DDX60, a DEXD/H box helicase, is a novel antiviral factor promoting RIG-I-like receptor-mediated signaling. *Molecular and cellular biology*, *31*(18), 3802–3819. doi:10.1128/MCB.01368-10

Nees, M., Geoghegan, J. M., Hyman, T., Frank, S., Miller, L., & Woodworth, C. D. (2001). Papillomavirus type 16 oncogenes downregulate expression of interferon-responsive genes and upregulate proliferation-associated and NF-kappaB-responsive genes in cervical keratinocytes. *Journal of Virology*, *75*(9), 4283–4296. doi:10.1128/JVI.75.9.4283-4296.2001

Nomura, D. K., Lombardi, D. P., Chang, J. W., Niessen, S., Ward, A. M., Long, J. Z., Hoover, H. H., et al. (2011). Monoacylglycerol lipase exerts dual control over endocannabinoid and fatty acid pathways to support prostate cancer. *Chemistry & biology*, *18*(7), 846–856. doi:10.1016/j.chembiol.2011.05.009

Pujana, M. A., Han, J.-D. J., Starita, L. M., Stevens, K. N., Tewari, M., Ahn, J. S., Rennert, G., et al. (2007). Network modeling links breast cancer susceptibility and centrosome dysfunction. *Nature genetics*, *39*(11), 1338–1349. doi:10.1038/ng.2007.2

Reiser, J., Hurst, J., Voges, M., Krauss, P., Münch, P., Iftner, T., & Stubenrauch, F. (2011). High-risk human papillomaviruses repress constitutive kappa interferon transcription via E6 to prevent pathogen recognition receptor and antiviral-gene expression. *Journal of Virology*, *85*(21), 11372–11380. doi:10.1128/JVI.05279-11

Santegoets, L. A. M., van Baars, R., Terlou, A., Heijmans-Antonissen, C., Swagemakers, S. M. A., van der Spek, P. J., Ewing, P. C., et al. (2011). Different DNA damage and cell cycle checkpoint control in low- and high-risk human papillomavirus infections of the vulva. *International journal of cancer. Journal international du cancer*. doi:10.1002/ijc.26345

Santin, A., Zhan, F., Bignotti, E., Siegel, E., & Cané, S. (2005). Gene expression profiles of primary HPV16- and HPV18-infected early stage cervical cancers and normal cervical epithelium: identification of novel candidate molecular markers for cervical cancer diagnosis and therapy. *Virology*.

Scanga, C., Cardellichio, C., & Holmes, K. (1996). Expression of the recombinant anchorless N-terminal domain of mouse hepatitis virus (MHV) receptor makes hamster of human cells susceptible to MHV infection. *Journal of ….*

Schoggins, J. W., Wilson, S. J., Panis, M., Murphy, M. Y., Jones, C. T., Bieniasz, P., & Rice, C. M. (2011). A diverse range of gene products are effectors of the type I interferon antiviral response. *Nature*, *472*(7344), 481–485. doi:10.1038/nature09907

Smyth, G. K. (2004). Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Statistical applications in genetics and molecular biology*, *3*, Article3. doi:10.2202/1544-6115.1027

Stanley, M. A. (2009). Immune responses to human papilloma viruses. *The Indian journal of medical research*, *130*(3), 266–276.

Talora, C., Sgroi, D., & Crum, C. (2002). Specific down-modulation of Notch1 signaling in cervical cancer cells is required for sustained HPV-E6/E7 expression and late steps of malignant transformation. *Genes & development*.

Taylor, M. P., Koyuncu, O. O., & Enquist, L. W. (2011). Subversion of the actin cytoskeleton during viral infection. *Nature reviews. Microbiology*, *9*(6), 427–439. doi:10.1038/nrmicro2574

Thomas, J. T., Oh, S. T., Terhune, S. S., & Laimins, L. A. (2001). Cellular changes induced by low-risk human papillomavirus type 11 in keratinocytes that stably maintain viral episomes. *Journal of Virology*, *75*(16), 7564–7571. doi:10.1128/JVI.75.16.7564-7571.2001

Wathelet, M. G., Clauss, I. M., Content, J., & Huez, G. A. (1988). Regulation of two interferon-inducible human genes by interferon, poly(rI).poly(rC) and viruses. *European journal of biochemistry / FEBS*, *174*(2), 323–329.

Appendix B

Paper II
Differential expression of cellular microRNAs in HPV 11, -16, and -45 transfected cells

Biochemical and Biophysical Research Communications 412 (2011) 20–25

Contents lists available at ScienceDirect

Biochemical and Biophysical Research Communications

journal homepage: www.elsevier.com/locate/ybbrc

Differential expression of cellular microRNAs in HPV 11, -16, and -45 transfected cells

Anita Dreher [a,1], Maria Rossing [b,1], Bogumil Kaczkowski [c,1], Ditte K. Andersen [a], Therese Juhlin Larsen [a], Mikael Kronborg Christophersen [a], Finn Cilius Nielsen [b], Bodil Norrild [a,*]

[a] Institute of Cellular and Molecular Medicine, DNA Tumor Virus Laboratory, University of Copenhagen, Panum Institute, Blegdamsvej 3, 2200 Copenhagen, Denmark
[b] Department of Clinical Biochemistry, Copenhagen University Hospital, Blegdamsvej 5, 2100 Copenhagen, Denmark
[c] The Bioinformatics Centre, Department of Biology and Biomedical Research and Innovation Centre, Copenhagen University, Ole Maaloes Vej 5, 2200 Copenhagen, Denmark

ARTICLE INFO

Article history:
Received 11 June 2011
Available online 18 July 2011

Keywords:
miRNA
HPV 11
HPV 16
HPV 45
Microarray

ABSTRACT

Human papillomaviruses (HPVs) are highly prevalent giving rise to both benign and malignant lesions why they are classified as high- and low-risk viruses. In this study we selected one low-risk (HPV 11) and two high-risk (HPV 16 and -45) types for genomewide miRNA analysis to investigate possible common and distinct features in the expression profiles. For this purpose we developed a cell culture model system in HaCaT cells for expression of the viral genomes under standardized conditions. We identified 25 miRNAs which were differentially regulated in two or three HPV types where 13 miRNAs were in common for all three types. Among the miRNAs identified, miR-125a-5p, miR-129-3p, miR-363, and miR-145 are related to human cancers. Noteworthy, miR-145 is found upregulated in the miRNA profiles of both high-risk HPV types. For selected differentially expressed miRNAs in HPV 16 predicted miRNA target transcript involved in signal transduction, RNA splicing and tumor invasive growth were validated by qRT-PCR. In addition, our results imply that the early 3′ untranslated region (3′UTR) of the three HPV genomes were not a target for miRNA regulation.

1. Introduction

The last decades have shown that human papillomaviruses (HPVs) are involved in several malignant and benign diseases. They cause highly proliferative infections and are dependent on the differentiation of their host cells where the virus infects the basal cell layer in cutaneous and mucosal epithelium. Dependent on their degree of oncogenecity they are divided into high-risk and low-risk types [6,16]. The high-risk types, such as HPV 16, -18, -45, and -33 belong to the genus α-papillomavirus species 7 and 9 and infections with these viruses can cause malignant transformation of the host cell. HPV 16 and -18 are the most frequently detected HPV types in anogenital cancers, especially in the cervix [20] and in head and neck cancers [8,10]. The low-risk types, such as HPV 6 and -11 cause benign condylomas and papillomas which can be highly distressing for the infected individuals [4]. The oncogenecity of the high-risk types is dependent on the constitutive expression of the oncogenes E6 and E7 and their following inactivation of p53 and pRB respectively [15,32]. Due to low binding affinity, E6 and E7 do not induce degradation of p53 and pRB, in low-risk HPV types [22]. HPV dependent malignancy is believed to depend on integration of the viral genome into the host cell chromosomes. This leads to disruption of the viral E2 gene, encoding a transcription factor able to regulate the expression of the viral oncoproteins E6 and E7. The upstream regulatory region (URR) has four binding sites for E2 and the transcription factor can up-regulate or inhibit transcription of the early HPV mRNAs [21,29,30].

MicroRNAs (miRNAs) interact with important cellular processes such as signal transduction, apoptosis and cell cycle progression. MiRNAs influence target mRNAs by binding to the 3′UTR and hereby regulating gene expression [3]. The regulatory effects of miRNAs are likely to be involved in both regulation of viral growth and cancer progression [33]. Several human cancers, e.g. breast cancer (for review see [12]), colorectal cancer [1], and head and neck cancer [9,23], show changes in the expression profiles of miRNAs. *In vitro* studies have shown certain miRNAs to be differentially expressed in HPV infected cells and cell lines [33]. Interestingly, several studies have demonstrated differential expression of miRNAs in invasive HPV positive squamous cell carcinomas and HPV positive cell lines compared to control tissue [13,14].

We have recently developed a model system in HaCaT cells in order to analyze the HPV mediated changes in cellular pathways and miRNA expression profiles in optimized growth conditions [7]. In the present study we examined the differentially expressed miRNAs in three selected HPV types; the low-risk HPV 11 and the

* Corresponding author. Address: Panum Institute, Blegdamsvej 3C, Building 22.3, 2200 Copenhagen, Denmark.
E-mail address: bnorrild@sund.ku.dk (B. Norrild).
[1] These authors contributed equally to this publication.

0006-291X/$ - see front matter © 2011 Elsevier Inc. All rights reserved.
doi:10.1016/j.bbrc.2011.07.011

high-risk HPV 16 and -45. We found common characteristics in the miRNA expression profiles as well as distinctly expressed miRNAs among the three examined HPV types. Moreover, specific HPV regulated miRNAs were analyzed for regulatory function on target transcripts. Finally, a functional luciferase experiment was carried out to explore possible effects of selected miRNAs on the 3′UTR of the viruses.

2. Materials and methods

2.1. Cell culture and transfections

The human keratinocyte cell line HaCaT was chosen for the present study. The cells were grown at 37 °C and 5% CO_2. For cell cultivation Dulbecco's Modified Eagle Medium (DMEM 1965, Invitrogen) supplemented with 10% Fetal Calf Serum, 1% Penicillin/Streptomycin, 1% Glutamate and 1% 0.1 M sodium pyruvate was used. Transfections were done using jetPEI™ transfection reagent (Polyplus transfections). For stable transfection the circularized genome (6 μg) was transfected together with the pSV2-neo selection vector (2 μg) into HaCaT cells seeded 2×10^6 cell per 10 cm plate. As control, cells were transfected with the selection vector alone. The cells were grown under standard conditions using the same passage number and maintained under selection with G418 sulfate (Invitrogen) in a concentration of 500 μg per ml.

2.2. RNA and microarray analysis

The transfected cells grown under selection for 3 weeks were harvested for total RNA using the TriZOL reagent (Invitrogen) according to the manufacturer's protocol. RNA was analyzed by the Affymetrix and Exiqon_V.11.0 miRNA platforms. Details regarding labeling, hybridization, scanning procedures, and data pre-processing were carried out as described previously [7]. A heatmap was generated based on analysis of variance (ANOVA), for variance filtering ($q = 0.2$).

2.3. Statistical analysis

Differentially expressed miRNAs between each HPV type versus control were identified by means of empirical Bayes moderated *t*-statistics as implemented in LIMMA [26]. *p*-Values were adjusted for multiple testing by Benjamini and Hochberg False Discovery Rate method and *p*-values <0.05 were considered significant.

2.4. Cloning HPV and control 3′UTRs into the luciferase vector

The pCDNA3.1puro-dsLuc2cp *Firefly* and corresponding pCDNA3.1puro-hRluc-cp *Renilla* vector were kindly supplied by Christopher S. Sullivan, University of Texas at Austin. The 3′UTRs of HPV 11 (nt. 4370–4640), HPV 16 (nt. 4100–4310) and HPV 45 (nt. 3875–4323) were PCR-amplified from plasmids given as a kind gift from Prof. Harald zur Hausen, DKFZ, Heidelberg. The primer sequences used for PCR amplification were as follows (restrictions sites for *XhoI* and *XbaI* are underlined).

HPV 11 FW, 5′-GCATGGACCTCGAGGAGTAAACCTTTTTTATACAG-3′ and RV, 5′-GCATGGACTCTAGACTTCCCAAGGGTATATACC-3′;

HPV 16 FW, 5′-GCATGGACCTCGAGTGTATATGTACATAATG-3′ and RV, 5′-GCATGGACTCTAGACCTGCCTGTTTGCATGTTTTAT-3′;

HPV 45 FW, 5′-GCATGGACCTCGAGAATCTGTATATTGTATAC-3′ and RV, 5′-GCATGGACTCTAGACAGGGGGGCACGTACC-3′ (Eurofins MWG Operon).

After cloning all constructs were sequenced (Eurofins MWG Operon).

2.5. Target gene analysis using qRT-PCR

Primers for the genes IKKα, IKKβ, ASF/SF2 and MTSS1 were designed and validated. Primers used were as follows.

IKKα FW, 5′-CGAAAGCTGCTCAACAAAC-3′ and RV, 5′-CCACCAACATCCTGAAGAAC-3′;

IKKβ FW, 5′-AAACCAGCATCCAGATTGAC-3′ and RV, 5′-AGCCATCATCCGTTCTACC-3′;

ASF/SF2 FW, 5′-TCCAGACATGCGAACCAAGGAC-3′ and RV, 5′-TCGAACTCAACGAAGGCGAAG-3′;

MTSS1 FW, 5′-GAAATAACCCACCTTCAGACC-3′ and RV, 5′-CCTTTCAAGTCCAGAATCACC-3′.

qRT-PCR was done with 90 ng of RNA extracted from HPV 16 transfected HaCaT cells selected for three weeks. qRT-PCR was done using the QuantiTect SYBR Green RT-PCR Kit (Qiagen). Relative quantification of the miRNA expression was calculated with the $2^{-\Delta\Delta C(T)}$ method [25].

2.6. Target analysis using Luciferase assay

HaCaT cells were seeded in 24-well plates at 1×10^5 cells per well 24 h prior transfection. For transfection 1 ng of pCDNA3.1-puro-dsLuc2cp with or without HPV 3′UTR insert, 1 ng pCDNA3.1-puro-hRluc-cp vector and 1 μg pBLUESCRIPT II SK (+) vector were mixed and co-transfected with 30 nM specific pre-miRNA (Ambion) or negative control#1 (scrambled miR) (Ambion) using jetPEI™ transfection reagent following the manufacturer's protocol (Polyplus Transfection). Luciferase activity was measured 24 h post transfection using the Dual-Luciferase® Reporter Assay System (Promega).

3. Results

3.1. Comparison of differentially expressed miRNAs

HaCaT cells were transfected with circularized HPV 11,-16, or -45 genomes as previously described [7]. Same batch and passage of transfected HaCaT cells were used for all studies. RNA obtained from each HPV type was analyzed on Affymetrix and Exiqon microarray platforms. Only miRNAs expressed with a fold change (FC) >1.5 and $p < 0.05$ were considered significant. For comparative studies we mainly focused on results based on the Affymetrix platform as this provided the most consistent expression values. The recently published HPV 11 miRNA expression profile (Affymetrix platform) is used for comparison of profiles between HPV 16 and -45 [7]. Our microarray experiment resulted in 50, 22, and 47 differentially expressed miRNAs in HPV 11, -16, and -45, respectively. Among these, only 13 miRNAs are shared between all three HPV types. Of the shared miRNAs, miR-181a, -125a-5p, -502-3p, -923, -92a-1*, and -500* are upregulated and miR-558, -576-3p, -606, -886-3p, -888, -1255a, and 1274b are downregulated. HPV 16 and -45 have three upregulated miRNAs in common; miR-145, -29b-1*, and -1246. HPV 16 and -11 share three downregulated miRNAs; miR-454*, -363, and -129*. Lastly, HPV 11 and -45 share five upregulated miRNAs; miR-455-3p, -331-3p, 15b, -1231, and -1180 and one; miR-129-3p is downregulated. Results of shared differentially expressed miRNAs are depicted in Fig. 1 and Table 1. The complete lists of differentially expressed miRNAs, including results from the Exiqon microarray platform are listed in Supplementary Tables S1A and S1B. The numbers of miRNAs identified by both platforms show some similarities. However, different array systems are based on different probe design and intensity of features will vary between platforms. The ANOVA analysis ($q = 0.2$) show 65 miRNAs representing the largest variance among the three HPV types and the HPV negative controls, as illustrated in

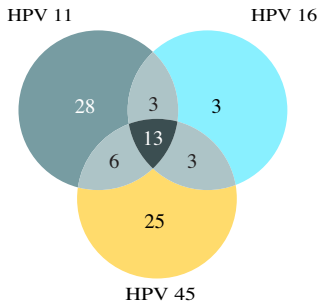


Fig. 1. Venn diagram showing the number of significantly changed miRNAs ($p < 0.05$) in HPV 11, -16 and -45. For the complete list of differentially expressed miRNAs see Supplementary Table S1A. The overlap between the individual types is seen and annotated miRNAs are listed in Table 1.

Fig. 2. It clearly shows several miRNA clusters unique for the various examined HPV-types. In particular it is noticeable that there are more clusters in common between HPV 16 and -45, reflecting the oncogenic potential of these two high-risk viruses opposed to the benign HPV 11.

3.2. Biological function of differentially expressed miRNA

As HPV 16 is the most prevalent type in premalignant and malignant diseases we searched for biological targets among differentially expressed miRNAs. By using computationally predicted mRNA targets in the TargetScan database (www.targetscan.org). Genes involved mainly in signal transduction, RNA splicing, post-transcriptional modifications and tumor metastasis. The chosen miRNA target genes, the miRNAs and their respective Fold Change are listed in Table 2. The expected inverse expression of the miRNA target genes was confirmed by qRT-PCR for all transcripts (IKKα, IKKβ, ASF/SF2 and MTSS1).

To study the possible effect of differentially expressed miRNAs on the 3′UTR of the selected HPV genomes, bioinformatic analysis of potential miRNA seed sequences (6-mer target sites) in HPV 11, -16 and -45 was done by Prof. Anders Lund's Research group. This analysis resulted in identification of a high number of potential binding sites for miRNAs. Only binding sites matching the differentially expressed miRNAs obtained by the array analysis were considered for further studies. After insertion of the 3′UTR sequences from each HPV type downstream of the luciferase reporter gene (U-LUC) we analyzed the putative miRNA regulation on HPV gene expression. Interestingly, none of the examined miRNAs showed any regulatory function. The results obtained for HPV 11 which was potentially targeted by miR-331-3p, -637, -874, -1274b, and -1275 are shown in Supplementary Fig. S1 (similar experiments were done for HPV 16 and -45, data not shown).

4. Discussion

Several studies have focused at the analysis of miRNA profiles in cultured HPV positive and negative cells as well as in clinical specimens from cervical cancer [24,33,34], nasopharyngeal cancer and oral cancer [2,9]. One study has reported that HPV modulates miRNA in a differentiation dependent process in order to maintain the viral replication capacity [14]. The mechanisms behind HPV mediated deregulation of the host cells are not completely understood and we therefore developed a reference cell culture model system.

The miRNA profiling of HPV 11, -16 and -45 transfected cells identified differentially expressed miRNAs. Interestingly, there were just 13 miRNAs regulated in common by all three types. However, among the most upregulated miRNAs, miR-125a-5p is known to be regulated in several cancers such as head and neck and gastric cancer. In gastric cancer a high level of the miRNA is related to a better prognosis for survival and can therefore be used as a prognostic factor [18].

Among the commonly downregulated miRNAs miR-886-3p is intriguing as this non-coding RNA molecule has recently been reclassified into a vault RNA even though it is still debated whether

Table 1
Significantly changed miRNAs shared between two or three HPV-types. Listed fold changes and expression levels derive from the Affymetrix platform.

	miRNA ID	Fold change			Average expression
		HPV11	HPV16	HPV45	
HPV16/45/11	hsa-miR-181a	1.8	1.7	2.1	8.5
	hsa-miR-125a-5p	3.3	2.6	3.8	8.4
	hsa-miR-502-3p	1.7	1.8	2	6.6
	hsa-miR-923	2.2	2.3	2.7	9.2
	hsa-miR-92a-1*	1.7	2.1	2.8	5.4
	hsa-miR-500*	2	2	2.2	6.5
	hsa-miR-558	−1.7	−1.9	−1.6	3.9
	hsa-miR-576-3p	−3	−3.4	−3.2	5.4
	hsa-miR-606	−1.9	−2	−1.7	4.9
	hsa-miR-886-3p	−10.3	−3.4	−4.4	7.2
	hsa-miR-888	−1.9	−2.3	−1.8	4
	hsa-miR-1255a	−1.9	−2.1	−2.2	4.3
	hsa-miR-1274b	−2.3	−2	−2.1	5.7
HPV16/45	hsa-miR-145		2.2	2.3	7.1
	hsa-miR-29b-1*		1.8	1.8	6
	hsa-miR-1246		2	2.7	8
HPV16/11	hsa-miR-454*	−2.9	−2.5		
	hsa-miR-363	−1.7	−1.6		
	hsa-miR-129*	−1.8	−1.9		
HPV11/45	hsa-miR-455-3p	2.1		1.7	7.5
	hsa-miR-331-3p	1.6		1.6	5.3
	hsa-miR-15b	2.1		2.6	9.8
	hsa-miR-1231	3.8		1.6	5.8
	hsa-miR-1180	1.8		2.1	5.7
	hsa-miR-129-3p	−1.7		−2	4.9

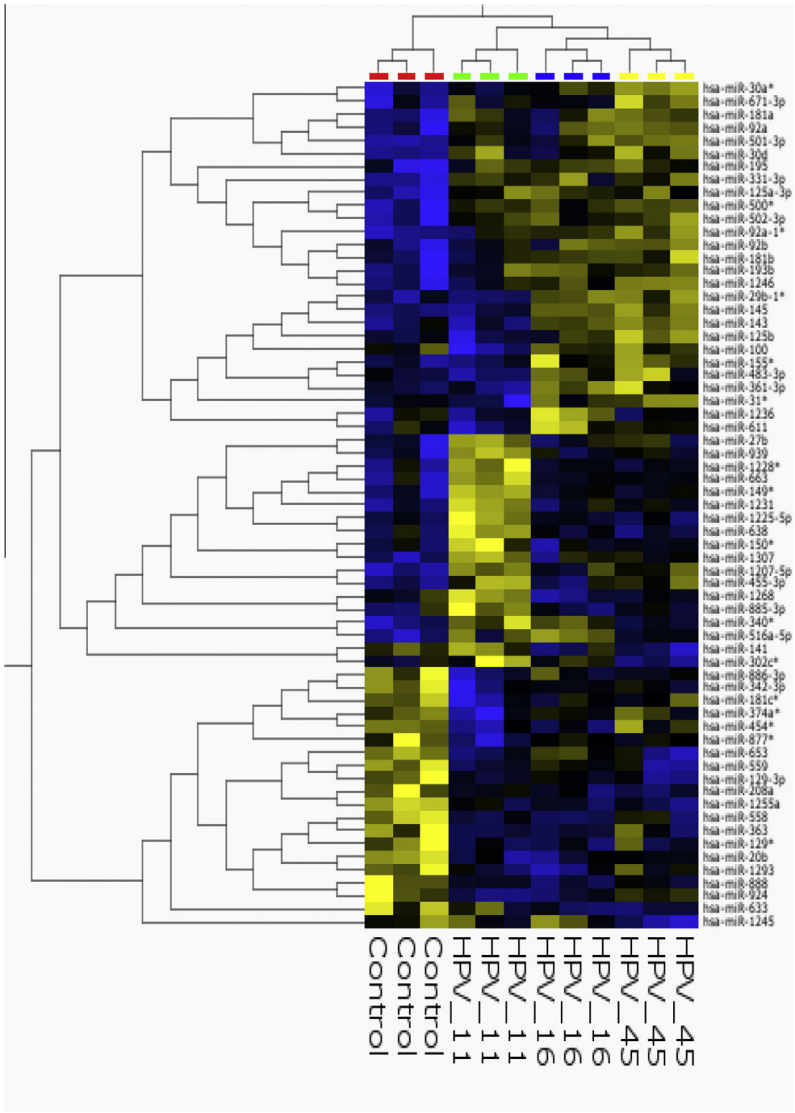


Fig. 2. Heatmap showing the result of the ANOVA analysis ($q = 0.2$). The 65 miRNAs correspond to the miRNAs with the largest variance between HPV 11, -16, -45, and control. Relatively low expression levels are represented as blue and relatively high expression levels are represented in yellow.

it is a true vault RNA as it behaves differently from other vault RNAs [11]. This non-coding RNA molecule is also regulated by other viruses such as Herpes viruses. Epstein-Barr-Virus infected cells increases, whereas Cytomegalovirus decreases the expression level of the molecule ("miR-886-3p") [17]. Moreover, suppressed expression of "miR-886-3p" is often seen in human cancer [11]. It will be interesting to learn more about the biological function of vault RNAs which could unveil why the three HPV types down-regulate the "miR-886-3p" up to 10-fold.

In the high-risk HPV 16 and -45 a twofold upregulation was shown for miR-145. Previous studies on HPV 16 positive cell lines showed no significant expression of miR-145 but when examined in raft cultures this miRNA was upregulated and therefore miR-145 may well correlate to cell differentiation [33]. In a recent study

Table 2
qRT-PCR validation of predicted miRNA target transcripts for four selected genes. Fold changes (FC) for miRNAs originate from the Exiqon microarray platform results.

Target mRNA	miRNA ID	FC miRNA	FC target mRNA
IKKα	miR-23a	1.5	−1.7
IKKβ	miR-16	1.4	−1.3
	miR-198	1.7	
ASF/SF2	miR-27b	−1.6	1.4
	miR-542-3p	−1.7	
MTSS1	miR-23a	1.5	−1.5

on head and neck cancers miR-145 was downregulated which might reflect dedifferentiation of tissue [10]. In line with these observations, previous studies revealed miR-145 as a common oncomiR in human cancers [1].

Among the downregulated miRNAs shared between HPV 11 and -16, miR-363 is downregulated 1.7-fold whereas studies in HPV positive pharyngeal squamous cancers show minor upregulation of miR-363 [10]. These opposing results could reflect the differences between cell culture models and *ex-vivo* cancer specimens. As a final point, miR-129-3p is differentially expressed in both HPV 11 and -45 and a recent study show miR-129–3p to be silenced by methylation as the DNA sequences contains CpG-islands upstream of the miR-locus in gastric cancer tissue [31].

To understand the biological function of changes in miRNA profiles, it is important to identify relevant target genes involved in regulation of either the virus life cycle or malignant progression of the host cell. In this study we selected genes involved in signal transduction, RNA splicing and tumor invasive growth. The gene ASF/SF2 has previously been shown to be of importance for expression of the late HPV 16 mRNAs [27] which is in line with the observed inverse correlation to the levels of miRs-27b and -542-3p in our model. The MTSS1 gene encodes a metastasis suppressor, and it is interesting that this gene is downregulated by miR-23a in the present study. Although it has not been discussed in HPV carcinogenesis it correlates to previous observations in bladder cancer [19]. Target analysis of miR-23a, -16 and -198 predicted the two IKK subunits to be target genes and our analysis confirmed the inverse mRNA expression of IKKα and IKKβ. The IKK complex is known to be modified by protein binding of HPV16 E7 oncoprotein [17,28]. The E7 protein binding to the IKK subunits reduces the NF-κB activity which is involved in cell proliferation [28]. It will be of interest to study the interplay of specific miRNAs and E7 for the regulation of the IKK subunits. The miRNA might confer a default regulatory pathway for fine tuning of the IKK level. Moreover, decreased NF-κB activity leads to tumor progression in skin cells, shown in a mouse model [5] and this finding could relate to our present HaCaT cell model as these cells originate from human skin. In summary, the present miRNA profiling provides a basis for further analysis of unknown biological pathways regulated by HPV. In this way new knowledge on common and individual features of the high- and low-risk HPV types may be discovered by analyzing the biological function of the distinct miRNAs found in this study. Finally, we found no indications that the 3'UTR of examined HPVs are targeted by miRNAs.

Acknowledgments

The technician Melissa Visser was supported by a grant from Laege Sophus Carl Emil Friis og hustru Olga Friis' Legat. Her technical assistance and that of Lucia Gavnholdt is acknowledged. The work was funded by, Lykfeldts Legat, Brødrene Hartmanns Fond, Snedkermester Sophus Jacobsen og hustru Astrid Jacobsens Fond, and Gangstedfonden, Dagmar Marshalls Fond og Fabrikant Einar Willumsens Fond. The Novo Nordisk foundation (Bogumil Kaczkowski) and student grants from the Danish Cancer Society supported Anita Dreher, Ditte K. Andersen and Mikael Kronborg Christophersen. Professor Anders Lund, Ulf A. Ørum, Lea Gregersen, BRIC, University of Copenhagen, Dr. Christopher S. Sullivan, University of Texas at Austin, and Professor Harald zur Hausen, DKFZ, Heidelberg are acknowledged for their help, useful discussions as well as for reporter constructs.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.bbrc.2011.07.011.

References

[1] Y. Akao, Y. Nakagawa, T. Naoe, MicroRNAs 143 and 145 are possible common onco-microRNAs in human cancers, Oncol. Rep. 16 (2006) 845–850.
[2] M. Avissar, B.C. Christensen, K.T. Kelsey, C.J. Marsit, MicroRNA expression ratio is predictive of head and neck squamous cell carcinoma, Clin. Cancer Res. 15 (2009) 2850–2855.
[3] R.W. Carthew, E.J. Sontheimer, Origins and Mechanisms of miRNAs and siRNAs, Cell 136 (2009) 642–655.
[4] L.T. Chow, T.R. Broker, B.M. Steinberg, The natural history of human papillomavirus infections of the mucosal epithelia, APMIS 118 (2010) 422–449.
[5] M. Dajee, M. Lazarov, P. Khavari, NF-kappa B blockade and oncogenic Ras induce invasive human epidermal neoplasia by circumventing cell cycle arrest, J. Invest. Dermatol. 119 (2002) 220.
[6] E.M. de Villiers, C. Fauquet, T.R. Broker, H.U. Bernard, H. zur Hausen, Classification of papillomaviruses, Virology 324 (2004) 17–27.
[7] A. Dreher, M. Rossing, B. Kaczkowski, F.C. Nielsen, B. Norrild, Differential expression of cellular microRNAs in HPV-11 transfected cells. An analysis by three different array platforms and qRT-PCR, Biochem. Biophys. Res. Commun. 403 (2010) 357–362.
[8] M.J. Lace, J.R. Anson, A.J. Klingelhutz, J.H. Lee, A.D. Bossler, T.H. Haugen, L.P. Turek, Human papillomavirus (HPV) type 18 induces extended growth in primary human cervical tonsillar or foreskin keratinocytes more effectively than other high-risk mucosal HPVs, J. Virol. 83 (2009) 11784–11794.
[9] C.B. Lajer, F.C. Nielsen, L. Friis-Hansen, B. Norrild, R. Borup, E. Garnaes, M. Rossing, L. Specht, M.H. Therkildsen, B. Nauntofte, S. Dabelsteen, B.C. von, Different miRNA signatures of oral and pharyngeal squamous cell carcinomas: a prospective translational study, Br. J. Cancer 104 (2011) 830–840.
[10] C.B. Lajer, B.C. von, The role of human papillomavirus in head and neck cancer, APMIS 118 (2010) 510–519.
[11] K. Lee, N. Kunkeaw, S.H. Jeon, I. Lee, B.H. Johnson, G.Y. Kang, J.Y. Bang, H.S. Park, C. Leelayuwat, Y.S. Lee, Precursor miR-886, a novel noncoding RNA repressed in cancer associates with PKR and modulates its activity, RNA 17 (2011) 1076–1089.
[12] N. Lynam-Lennon, S.G. Maher, J.V. Reynolds, The roles of microRNA in cancer and apoptosis, Biol. Rev. Camb. Philos. Soc. 84 (2009) 55–71.
[13] I. Martinez, A.S. Gardiner, K.F. Board, F.A. Monzon, R.P. Edwards, S.A. Khan, Human papillomavirus type 16 reduces the expression of microRNA-218 in cervical carcinoma cells, Oncogene 27 (2008) 2575–2582.
[14] M. Melar-New, L.A. Laimins, Human papillomaviruses modulate expression of microRNA 203 upon epithelial differentiation to control levels of p63 proteins, J. Virol. 84 (2010) 5212–5221.
[15] K. Munger, W.C. Phelps, V. Bubb, P.M. Howley, R. Schlegel, The E6 and E7 genes of the human papillomavirus type 16 together are necessary and sufficient for transformation of primary human keratinocytes, J. Virol. 63 (1989) 4417–4421.
[16] N. Munoz, f.x. Bosch, S.S. de, R. Herrero, X. Castellsague, K.V. Shah, P.J. Snijders, C.J. Meijer, Epidemiologic classification of human papillomavirus types associated with cervical cancer, N. Engl. J. Med. 348 (2003) 518–527.
[17] C. Nandy, J. Mrazek, H. Stoiber, F.A. Grasser, A. Huttenhofer, N. Polacek, Epstein-barr virus-induced expression of a novel human vault RNA, J. Mol. Biol. 388 (2009) 776–784.
[18] N. Nishida, K. Mimori, M. Fabbri, T. Yokobori, T. Sudo, F. Tanaka, K. Shibata, H. Ishii, Y. Doki, M. Mori, MicroRNA-125a-5p is an independent prognostic factor in gastric cancer and inhibits the proliferation of human gastric cancer cells in combination with Trastuzumab, Clin. Cancer Res. 17 (2011) 2725–2733.
[19] S. Nixdorf, M.O. Grimm, R. Loberg, A. Marreiros, P.J. Russell, K.J. Pienta, P. Jackson, Expression and regulation of MIM (missing in metastasis), a novel putative metastasis suppressor gene, and MIM-B, in bladder cancer cell lines, Cancer Lett. 215 (2004) 209–220.
[20] W.C. Phelps, J.A. Barnes, D.C. Lobe, Molecular targets for human papillomaviruses: prospects for antiviral therapy, Antivir. Chem. Chemother. 9 (1998) 359–377.
[21] W.C. Phelps, P.M. Howley, Transcriptional transactivation by the human papillomavirus type-16 E2 gene-product, J. Virol. 61 (1987) 1630–1638.

[22] D. Pim, L. Banks, Interaction of viral oncoproteins with cellular target molecules: infection with high-risk vs low-risk human papillomaviruses, APMIS 118 (2010) 471–493.

[23] L. Ramdas, U. Giri, C.L. Ashorn, K.R. Coombes, A. El-Naggar, K.K. Ang, M.D. Story, miRNA expression profiles in head and neck squamous cell carcinoma and adjacent normal tissue, Head Neck 31 (2009) 642–654.

[24] Q. Rao, H. Zhou, Y. Peng, J. Li, Z. Lin, Aberrant microRNA expression in human cervical carcinomas, Med. Oncol. (2011) [Epub ahead of print].

[25] T.D. Schmittgen, K.J. Livak, Analyzing real-time PCR data by the comparative C(T) method, Nat. Protoc. 3 (2008) 1101–1108.

[26] G.K. Smyth, Linear models and empirical bayes methods for assessing differential expression in microarray experiments, Stat. Appl. Genet. Mol. Biol. 3 (2004) Article3.

[27] M. Somberg, S. Schwartz, Multiple ASF/SF2 sites in the human papillomavirus type 16 (HPV-16) E4-coding region promote splicing to the most commonly used 3′-splice site on the HPV-16 genome, J. Virol. 84 (2010) 8219–8230.

[28] D. Spitkovsky, S.P. Hehner, T.G. Hofmann, A. Moller, M.L. Schmitz, The human papillomavirus oncoprotein E7 attenuates NF-kappa B activation by targeting the I kappa B kinase complex, J. Biol. Chem. 277 (2002) 25576–25582.

[29] S.H. Tan, B. Gloss, H.U. Bernard, During negative regulation of the human papillomavirus-16 E6 promoter, the viral E2 protein can displace Sp1 from a proximal promoter element, Nucleic Acids Res. 20 (1992) 251–256.

[30] F. Thierry, P.M. Howley, Functional-analysis of E2-mediated repression of the Hpv18 P105 promoter, New Biologist 3 (1991) 90–100.

[31] K.W. Tsai, C.W. Wu, L.Y. Hu, S.C. Li, Y.L. Liao, C.H. Lai, H.W. Kao, W.L. Fang, K.H. Huang, W.C. Chan, W.C. Lin, Epigenetic regulation of miR-34b and miR-129 expression in gastric cancer, Int. J. Cancer (2011) [Epub ahead of print].

[32] K. Vousden, Interaction of human papillomavirus transforming proteins with the products of tumor suppressor genes, FASEB J. 7 (1993) 872–879.

[33] X. Wang, S. Tang, S.Y. Le, R. Lu, J.S. Rader, C. Meyers, Z.M. Zheng, Aberrant expression of oncogenic and tumor-suppressive microRNAs in cervical cancer is required for cancer cell growth, PLoS ONE 3 (2008) e2557.

[34] Z.M. Zheng, X. Wang, Regulation of cellular miRNA expression by human papillomaviruses, Biochim. Biophys. Acta (2011) [Epub ahead of print].

Appendix C

Paper III
Carcinomas of Unknown Primary Site are Distinct from Metastases of Known Origin

Carcinomas of Unknown Primary Site are Distinct from Metastases of Known Origin

[1,2*]Anne Kirstine Hundahl Møller, [4*]Bogumil Kaczkowski, [1*]Rehannah Borup, [5,6]Ricardo Henao, [1]Jonas Vikeså, [4]Anders Krogh, [2]Katharina Perell, [3]Flemming Jensen, [4,5]Ole Winther, [1#]Finn Cilius Nielsen and [2#]Gedske Daugaard

[1]*Center for Genomic Medicine,* [2]*Department of Oncology and* [3]*Department of Radiology Copenhagen University Hospital ,Rigshospitalet, Blegdamsvej 9, DK-2100 Copenhagen Ø, Denmark,* [4]*Bioinformatics Centre*, Department of Biology *and* Biotech Research and Innovation Centre *, University of Copenhagen, DK-2200 Copenhagen Denmark.* [6]*Technical University of Denmark (DTU), DK-2800 Lyngby, Denmark*

*These first authors contributed equally
*The two last authors contributed equally

Running title: CUP are distinct from metastases of known origin

Key words: Cancer of unknown origin, biology, array analysis

Financial grant: This study was supported by grants from the Danish National Advanced Technology Foundation and the Danish Cancer Society, the Svend Andersen Foundation, the Toyota Foundation and the NOVO-Nordisk Foundation.

Words 4013, Tables 2, Figures 4, Supplementary Tables 3, Supplementary Figures 4

Correspondence
Gedske Daugaard
Department of Oncology 5073
Copenhagen University Hospital
University of Copenhagen
Blegdamsvej 9
DK-2100 Copenhagen Ø
Denmark
Telephone: +45 3545 4677
Email: gedske.daugaard@rh.regionh.dk

TRANSLATIONAL RELEVANCE

Cancer of unknown primary site (CUP) have an aggressive biological and clinical behavior, but it is currently unknown if they exhibit distinct molecular features. Employing transcriptome based Linear Discriminant Analysis (LDA) and Quadratic Discriminant Analysis (QDA) on a set of primary cancers, metastases and CUP; we demonstrate that CUP are distinct from metastases of known origin. CUP exhibit an inconsistent expression of conventional cancer biomarkers and from the QDA derived outlier score, we show they are more distantly related to 16 predefined tumor classes than corresponding metastases of known origin. We conclude that CUP exhibit biological features that distinguish them from metastases of known origin, which may warrant selective treatment and diagnostic strategies for these aggressive tumors.

ABSTRACT

Cancer of unknown primary site (CUP) constitute ~5% of all cancers. The tumors have an aggressive biological and clinical behavior, but it is currently unknown if they exhibit distinct molecular features. Employing transcriptome based Linear Discriminant Analysis (LDA) and Quadratic Discriminant Analysis (QDA) on a set of primary cancers, metastases and CUP; we demonstrate that CUP are distinct from metastases of known origin. CUP exhibit an inconsistent expression of conventional cancer biomarkers and from the QDA derived outlier score, we show they are more distantly related to 16 predefined tumor classes than corresponding metastases of known origin. Based on the LDA prediction, we performed a paired comparison of CUP and metastases and defined a CUP core set of 1117 up- and 934 down-regulated genes. CUP were enriched in networks promoting DNA damage response, DNA double strand break repair and chromatin remodeling that may translate into early dissemination and poor outcome. We conclude that CUP exhibit biological features that distinguish them from metastases of known origin, which may warrant selective treatment and diagnostic strategies for these aggressive tumors.

INTRODUCTION

Carcinomas of unknown primary site (CUP) are a heterogeneous group of cancers with variable clinical and histological features for which no primary site of the tumor can be identified despite an extensive diagnostic work-up (1, 2). CUP accounts for 3-5 % of all cancer diagnoses and about 85% of the patients have a very poor prognosis (3). Although a primary tumor cannot be identified in about two-thirds of the cases, CUP are generally considered to represent metastases. The elusive origin may partly be related to limitations in our diagnostic procedures, but it may also indicate that CUP follows a distinct molecular mechanism.

The prevalent model of metastasis is that cells from a primary tumor invade the local environment and spread to distant locations. Metastases may derive from more or less differentiated cancer cells at different stages of tumor growth and this may provide a substantial heterogeneity in the clinical presentation and nature of metastases. Although micrometastases are enriched in cells expressing stem cell markers, macrometastases share many similarities to the primary tumor, so newly settled cancer stem cells not only self-renew, but also foster differentiated colonies of cancer cells (4). Because metastases retain some of the characteristics of the primary cancer, transcriptome signatures have been employed to depict the origin of CUP.

It is currently unknown if CUP exhibit particular genetic and phenotypic characteristics compared to metastases of known origin. The challenge in addressing this problem is obviously that CUP per definition are of unknown origin. To circumvent this problem, we generated a genetic signature that could classify a wide number of known primary tumor classes and their metastases with high accuracy and used Quadratic Discriminant Analysis (QDA) to estimate an outlier score, which measures how dissimilar a particular sample is from the group of primary tumors.

The outlier score does not require knowledge about the origin of the individual tumors because the outlier score expresses the distance of a sample to the nearest known tumor class. Subsequently, we used the LDA predictions to define a CUP core set of differentially expressed genes that could provide leads to the molecular pathology of CUP.

MATERIALS AND METHODS

Gene expression profiles for tumor classification

Expression profiles of more than 2400 tumor samples were downloaded from the Gene Expression Omnibus (GEO) (http://www.ncbi.nlm.nih.gov/geo/) or generated from samples collected and processed at our own facility at Rigshospitalet. The material comprised 15 classes of carcinomas from thyroid, lung, stomach, colon/rectum, pancreas, bile duct/gallbladder, liver, kidney, urinary tract, prostate, breast, ovary, endometrium, cervix uteri, testis cancer, a group of malignant melanomas and a group with pooled normal tissue samples from various organs that was included in order to allow detection of samples without sufficient tumor tissue. Sample IDs are indicated in the enclosed Supplemental Table 1. The pathology descriptions were reviewed in order to group the samples into tumor classes and this ultimately resulted in a set of 1466 expression profiles from well-defined primary tumors (1299) and normal tissue (167) (Table S1). The classifier was tested on an independent validation set including 641 tumor samples (391 primary tumors and 250 metastases) from all 16 tumor classes (Table S1).

CUP Samples

CUP patients were consecutively enrolled between November 2004 and September 2010 for diagnostic work-up and treatment. A schematic representation of the CUP patients and the inclusion of samples are shown in Supplemental Figure S1. Patients were included when the diagnostic work-up, as recommended by the European Society of Medical Oncology (ESMO) (5), failed to identify a primary site of origin. At least two ultra-sonography-guided biopsies – one for histopathological work-up and one for gene expression

profiling – were obtained from all patients. The study was approved by the local ethical committee and patients had given their written informed consent.

Microarray analysis and expression values

Total RNA was isolated, labeled and hybridized as described (6). Cell files were pre-processed using the *Robust multi-chip average* (RMA) method (7) and evaluated for quality parameters with the Simpleaffy functionality of the *R /Bioconductor* packages. The data sets were filtered so probe sets with Interquartile Range (IQR) below 0.8 were omitted.

Tumor classification and outlier analysis

Linear discriminant analysis (LDA) was used for classification as implemented in the R language. Briefly, in LDA the predictive probability of class c given input x is computed using Bayes' theorem $p(c|x) = p(x|c)\, p(c) / p(x)$, where $p(x|c)$ is a normal density specific for the class, $p(c)$ the a priori probability of class c and $p(x)=\text{sum\_c}\, p(x|c)\, p(c)$ the density of the input according to the model. Maximum likelihood is used to fit $p(x|c)$ and $p(c)$, $c=1,\ldots,17$ on the training data. In order to construct a gene signature for our classifier we used leave‐one‐out cross validation (LOOCV), where for each split, feature selection by F-test were applied prior to LDA. A grid search over p-value cut-offs yielded the cut-off with the optimal LOOCV accuracy. The signature was eventually selected by an F-test using the optimal p‐value cut‐off on the full set of 1466 training samples, resulting in 428 probes (311 unique genes). The performance of this first (428 probe) classifier was then assessed using the independent 641 sample validation set. We merged the original training and validation set and used the found p-value cut-off (giving 641 probes) to generate a second classifier optimized for for CUP prediction. The performance of this classifier was assessed using LOOCV. Finally, the LDA classifier was made sex-specific by setting the prior probabilities to zero for sex specific cancers (ovary, cervical and prostate) not occurring and in the sex in question renormalizing the remaining prior probabilities accordingly.

A low model density p(x) implied that the input x was not similar to those in the training data. We therefore defined an outlier score OS = –log p(x) and calculated the OS for each sample in the LOOCV loop. We used QDA (individual covariance of normals) rather than LDA (shared covariance of normals) in this step.

RESULTS

Tumor Classification

To define the probable origins of CUP, we generated a transcriptome-based signature that could distinguish 16 common tumor classes and metastases of known origin (Detailed in Table S1 and S2). A group of normal tissues was moreover included, to allow detection of samples without sufficient tumor tissue and because all CUP data were generated at our facility, we also examined a series of primary cancers and metastases from Rigshospitalet to exclude possible site effects. The first 428 probe (F-test p-value cut-off 10^-180) had a validation set accuracy of 90 % and 83% for primary tumors and known metastases, respectively (Table S2). The second (641 probe) classifier, used in the analysis of CUP, training on the merged training and validation had a LOOCV accuracy in primary tumors, known metastases and normal samples of 92 %, 87 % and 89 %, respectively (Table S2) The principal component analysis is shown in Figure 1 and the ten most selective transcripts and their gene ontology for each tumor class are listed in Figure S2.

CUP Patients and Samples

Sixty eight consecutive CUP patients were enrolled in the study, but since eleven samples did not meet the quality criteria the number of CUP samples ended at 57. During the diagnostic work-up a primary tumor site was identified in 23 patients and a consensus diagnosis was obtained in 5 patients as described (Figure S1). In 29 patients the primary tumor site remained unknown. The histological features of the 57 CUP samples that underwent expression profiling are summarized in Table 1. To provide a systematic overview of the expression of conventional biomarkers in the CUP samples, we compiled 45 common histopatological markers and depicted their expression in a two-way hierarchal cluster (Figure 2). Whereas, the histomarkers exhibited a characteristic expression pattern in about 85% of the primary cancers, only 10 of the 28 (35%) CUP - where a putative primary site was identified and 3 of the 29 (10%) CUP - where the primary site remained unknown expressed one or more characteristic markers at significant levels. The

strongest overlap with the 641 LDA based CUP predictions described below was observed for WT1 and CEA that were positive in 4 and 3 samples predicted as ovary (6%) and colorectal (4%) cancers, respectively. Moreover, 6 samples were positive for TP63 and 2 samples were positive for surfactant proteins. Finally, one sample was positive for PAX2 in agreement with the LDA prediction as renal carcinoma. Compared to the primary cancers there was moreover a limited concordance between markers characteristic for the same tumor category. Only two of the WT1 positive cancers were positive for CA125, and only 3 of the TP63 positive samples expressed CK17 and CK5, characteristic of squamous carcinoma. If the histological markers were combined and used in an LDA based fashion, the concordance with the 641 signature LDA predictions or Standard of Reference was about 66% indicating that systematical application of the patomarkers may compensate for the modest predictive power of individual markers. Finally, we collected a stem cell marker panel and repeated the two way clustering (data not shown). Compared with metastases of known origin, we did not detect any markers that were selective to CUP. The only enriched (1.6 fold) marker was HNF4A and this was related to contamination of normal liver tissue in some of the samples (see below).

LDA based CUP Classification

When we applied the 641 classifier on the CUP samples, a tumor class was confined to 48 of the samples. Tumor calls were in general robust and relevant diagnoses were obtained in the presence of up to 75% normal tissue. We ascribe this to the fact that all data were logarithmic and that many of the tumor markers are expressed at high levels in the neoplastic cells. Nine samples were classified as normal tissue (Table 1 and Figure S3) and to determine if the latter represented mistaken biopsies, with little or no tumor cells, we examined the level of liver (*APOA2, ALB*), muscle (*ACTA1*), lymph node (*IGJ, IGHA1, IGKV3-20*) and skin (*KRT2, TYRP1*) specific transcripts in the samples (8). Six of the nine samples consisted almost entirely of normal cells from the site of biopsy, whereas the last three contained about 10, 15 and 40% normal cells.

These samples were considered to either exhibit low tumor marker expression or originate from a tumors that were not represented in our training set (see below).

To estimate the fidelity of the tumor calls, we considered the 28 CUP where the diagnostic work-up or *Standard of Reference* (see legend to Table 1) had provided a possible primary tumor site (Table 1). Eighteen (64 %) samples were in accordance with *the Standard of Reference.* In the cases where a histological diagnosis was provided there was complete agreement with the LDA prediction except in one case. Three of the samples misclassified as normal or breast or were verified as angiosarcoma, adnex carcinoma and desmoplastic small round cell tumor that were not included in the classifier. Moreover, two cholangiocarcinoma were predicted as HCC and cervical cancer, respectively and two lung cancers were predicted as stomach- and breast- cancers, respectively. Among the CUP where no primary tumors had been proposed, six samples were called as normal tissue and in two cases there was no agreement between the LDA and the histological diagnosis. Finally, we noted that 16% of the CUP was classified as breast cancer.

QDA based outlier analysis

To determine the similarity between primary cancers, metastases of known origin and CUP, we employed a QDA to determine the likelihood that a particular sample belonged to one of the predefined tumor classes. Outlier scores were calculated in LOOCV fashion that is for one sample at a time using all remaining samples i.e. primary tumors and metastases to represent the classes. The outlier scores of the samples from normal tissues are not comparable to the primary tumors and metastases because of the heterogeneity among the many different tissues in the class.

Based on the results from primary tumors and metastases, we plotted the error rates versus the outlier scores and demonstrated a clear relation between errors and outlier scores (Figure 3). Samples with outlier scores below 800 exhibited less than 10% risk of being erroneous, whereas, outlier scores above 1000 had

more than 25% risk of being incorrect. However, even in the high end of outlier scores with only 75% accuracy, prediction is far from random, since we are working with 16 different classes. As shown in the box plot (Figure 3), CUP samples had significantly higher outlier values than primary tumors and metastases. To ensure that the difference was not related to our platform, we compared our own samples of known metastases and primary tumors and observed the same difference. CUP moreover, consisted of biopsies that may contain more normal tissue than samples obtained during surgery. We therefore plotted the percentage of normal tissue as estimated from the relative expression of markers of lymphoid, liver, and muscle tissue versus the outlier scores, but observed no correlation between the amount of normal tissue in the biopsies and the outlier scores (Figure S3). A number of the samples that expressed conventional histopathological biomarkers exhibited low scores, but if we compared CUP where a primary cancer were identified during the clinical processing with CUP where no primary site could be identified, there was no difference between the outlier scores (mean 991 vs mean 1031, P=0.24). Taken together, the results demonstrate that CUP are more distantly related to the predefined tumor classes, than known metastases.

mRNA Expression and Gene Set Enrichment in CUP

To identify differentially expressed transcripts, we performed a class comparison between CUP and metastases of known origin. To eliminate differences between tumor classes, the analysis was performed as a paired analysis with respect to the LDA predictions. Metastases from uterine, testis, prostate, melanoma and thyroid cancers were excluded from the analysis because no CUP had been allocated to these groups by the LDA. CUP predicted as normal tissue were also excluded. Moreover, cholangiocarcinoma were omitted from the calculations because they were not represented in the LDA predicted metastases group. In total 41 CUP and 186 metastases comprising 10 different cancer groups were included in the analysis. To define the most up- and down-regulated CUP transcripts, a cut-off of $p<10^{-8}$ corresponding to a false discovery rate of $q< 1.9*10^{-7}$ was used. This resulted in 1550 down- and 1390 up-regulated probe sets corresponding to 1117 and 934 unique annotated genes, respectively. We

defined these two lists as our CUP core set of differentially expressed transcripts. The 40 most down- or up-regulated mRNAs are shown in Table S4. We subsequently performed a Gene Set Enrichment Analysis (GSEA) on our CUP core set using the Broad Institute's GSEA database (http://www.broadinstitute.org/gsea/msigdb). Initially, we searched for enriched gene ontology terms, and this revealed that up-regulated transcripts were associated with GO-terms (P<0.01): DNA_INTEGRITY_CHECKPOINT,DNA_DAMAGE_CHECKPOINT,DNA_REPLICATION_INITIATION,DNA_PACKAGING,NEGATIVE_REGULATION_OF_DNA_METABOLIC_PROCESS,CELL_CYCLE_CHECKPOINT;NEGATIVE_REGULATION_OF_DNA_REPLICATION,CHROMATIN_REMODELING,DNA_DAMAGE_RESPONSESIGNAL_TRANSDUCTION. There were no particular enrichments among the down-regulated mRNAs.

To depict CUP enriched molecular pathways, we examined if the CUP core set exhibited overlaps with the Molecular Signature Database (MSigDB) curated gene sets. Overlaps between the CUP core set ($P<10^{-8}$) were computed by submission of up- and down-regulated probe sets separately (Table 2). Gene sets consisting of transcripts that were positively correlated to BRCA1, ATM and CHECK2 expression were highly enriched in the up-regulated CUP core set. The down-regulated CUP mRNAs showed fewer significant overlaps but SHEN_SMARCA2_TARGETS_DN gene set, which depict transcripts that are negatively correlated with SMARCA2 expression in prostate cancer was clearly overlapping with the CUP set.

To examine the BRCA1 and SMARCA2 pathway networks defined by the SHEN_SMARCA2_TARGET_DN, SHEN_SMARCA2_TARGET_UP and PUJANA_BRCA1_PCC_NETWORK in greater detail, we generated two way clusters using the complete gene sets on our CUP core set (Figure 4). The clusters were based on a paired analysis with respect to their LDA predictions and with the same inclusion criteria, as described above. The SHEN_SMARCA2_TARGET_DN; SHEN_SMARCA2_TARGET_UP and PUJANA_BRCA1_PCC_NETWORK gene symbols were translated into probe sets and to exclude non-functional redundant probe sets, only the probe sets with the 50% highest variance were included. We moreover applied a p-value of 0.001 to select for probe sets that differed among the two groups (Figure 4).

The PUJANA_BRCA1_PCC_NETWORK set of genes consists of 1671 gene symbols that translated into 3897 probe sets. Following filtering 705 probe sets corresponding to 519 up-regulated and 66 down-regulated genes were clustered (Figure 4). From the cluster it is apparent that the BRCA1 profile is strongly enriched in CUP compared to the corresponding metastases. A schematic representation of the BRCA1 and non-homologous repair networks showing the enriched factors is depicted in Figure S4. Following the same procedure, we subsequently looked at the SMARCA2 networking (Figure 4). The SHEN sets consists of 360 SMARCA2 negatively- and 430 SMARCA2 positively- correlated genes that translated into 772 and 1211 probe sets respectively. In the SMARC2A negatively correlated group, we observed 20 genes that were up-regulated and 95 that were down-regulated in CUP compared to metastases, and amongst the SMARCA2 positive correlated genes we saw 161 up-regulated genes and 19 down regulated after filtering (top 50% variance probes and $p>0.001$). Taken together, the GSEA shows that CUP are characterized by enrichment of the double strand break DNA repair system and the SMARCA2/BRM chromatin dependent remodeling system.

DISCUSSION

We developed a robust LDA classifier to compare metastasis of known origin with CUP. The LDA agreed with the clinical consensus in almost two-thirds of the patients, underscoring the clinical relevance of molecular CUP classification and in line with a number of previous molecular prediction studies (9-23) CUP mainly originated from bile duct/cholangiocarcinoma, breast, lung and colorectal cancers. Breast cancers are almost never encountered during autopsy, but CUP predicted as breast cancer have also been observed in previous array studies (reviewed in (24)), where ~15% of the CUP cases were supposed to originate from the breast. From a principal component analysis (data not shown) it is evident that CUP predicted as breast cancers have similarities to both ductal and basal like breast cancers and it is possible that the invasive nature and rapid dissemination from breast - or other sites – of the latter cancer may lead to a clinical presentation as CUP.

We employed quadratic discrimant analysis (QDA) to calculate the distance of primary tumors, metastases and our CUP samples to the nearest tumor class. In agreement with the acquisition or loss of phenotypic traits compared to their origin, CUP were more distantly related to the predefined tumor classes, than known metastases. A nearby explanation for the disparity between CUP and known metastases could be that CUP were derived from types or subclasses of cancers not represented among our 16 classes, but a number of arguments speak against this. Firstly, autopsy and previous molecular classification studies support that the vast majority of CUP are likely to originate from the included tumor classes (24). Secondly, the genetic signature was selected by means of an F-test considering the entire class, so class specific transcripts are supposed to be present even in putative subclasses. Thirdly, high outlier scores were also observed among classes such as colorectal cancers that are not known to contain subclasses. Finally, if a number of CUP represented rare cancers, the majority of the CUP scores should have overlapped with

metastases of known origin. So taken together, we infer that the observed difference in outlier scores is likely to reflect that CUP have a distinct molecular features.

Attempts to elucidate the molecular biology of CUP have been hampered by the heterogeneity of the cancers and so far it has not been possible to identify common genetic aberrations. Based on the LDA predictions it became however possible to define CUP enriched transcripts and molecular pathways. One of the most consistently down-regulated factors was the transcription factor early growth response 1 (*EGR1*). *EGR1* is involved in cell growth and differentiation and suppressed *EGR1* levels have previously been reported in breast carcinoma (25), glioblastoma (26) and lung (27) cancer, where it was predictive of poor outcome. Although *EGR1* promotes growth of prostate cancer, it is generally considered to function as a tumor suppressor (reviewed in (28)). Among the up-regulated transcripts *eIF5B* was noted because it is an important regulator of protein synthesis, which is pivotal for rapidly growing cancer cells. The associated GTPase activating eIF5A has previously been shown to be highly expressed in various other cancers and together with eIF5b it is considered an attractive target for cancer treatment (29).

To obtain further leads to the pathogenesis of CUP, we searched for gene set enrichments in the molecular signatures database. The gene ontology collection showed that factors involved in DNA -damage and –integrity were enriched in CUP and this was reinforced by the observed enrichment of BRCA1, ATM and CHEK2 and SWI/SNF networks controlling DNA damage response and DNA double strand break repair and chromatin remodeling. DNA double strand break repair and chromatin remodeling are functionally coupled and intimately associated with carcinogenesis (30, 31). BRCA1, ATM and CHEK2 are well established tumor suppressor genes in e.g. hereditary breast-ovarian cancer and we were surprised to discover that CUP were enriched in DNA double strand repair networks. Recent observations from cancers such as melanoma (32, 33) and breast cancer (34) have shown that DNA repair systems are frequently up-regulated in metastases. It has been hypothesized that over-expression of DNA repair genes promotes metastases because efficient

repair increase the overall fitness and viability of the malignant cells (34). In pancreatic cancer increased expression of BRCA1 was confined to the most invasive cellst hat thrived because of genomic stability(35). Recent findings moreover suggest that oncogenes stimulate stalling and collapse of DNA replication forks that in turn leads to formation of DNA double strand break (36). Finally, induction of the double strand break repair may increase the number of mutations (37) and thereby increase the acquisition of independent phenotypic traits of the metastases that may translate into atypical presentation and poor outcome.

Systemic cancer progression has been proposed to occur via two models. The prevailing model states that cancer progression occurs within the primary tumor before metastatic dissemination of fully malignant cells, whereas the second put forward that cells disseminate from the primary tumor at an early stage and pursues a parallel and independent progression of metastases (reviewed in (38)). The two models provides an appealing rationale for the observed difference between metastases of known origin and CUP, because the parallel progression predicts greater disparity between metastatic founders and primary tumor cells than does linear progression. By inference, parallel progression may also be characterized by the accumulation of distinct genetic and epigenetic alterations in the primary tumor compared to the metastases. Moreover, tumor cells are predicted to settle at unconventional sites due to their independent selection and spread before the primary cancer causes clinical symptoms.

In conclusion, we report that CUP exhibit distinct gene expression patterns that distinguish them from metastases of known origin. We propose that enrichment of DNA damage repair pathways and parallel metastatic behavior may be implicated in early dissemination and poor outcome of CUP.

ACKNOWLEDGMENTS

Elisabeth Schiefloe and Susanne Smed are thanked for their technical assistance and Leila Majdanac for secretarial assistance and proof-reading. This study was supported by grants from the Danish National Advanced Technology Foundation and the Danish Cancer Society, the Svend Andersen Foundation, the Toyota Foundation and the NOVO-Nordisk Foundation.

REFERENCES

1. Pavlidis N, Briasoulis E, Hainsworth J, and Greco FA, *Diagnostic and therapeutic management of cancer of an unknown primary.* Eur J Cancer, 2003. **39**(14): 1990-2005.
2. Pavlidis N and Fizazi K, *Carcinoma of unknown primary (CUP).* Crit Rev Oncol Hematol, 2009; **69**(3): 271-8.
3. Daugaard G, Møller A, and Petersen B, *Carcinoma of Unknown Primary*, in *Textbook of Medical Oncology*, F. Cavalli HH, S. Kaye, J. Amritage and M. Pickard, Editor 2009. p. 313-322.
4. Wicha MS and Hayes DF, *Circulating tumor cells: not all detected cells are bad and not all bad cells are detected.* J Clin Oncol. 2011;**29**(12): 1508-11.
5. Briasoulis E, Tolis C, Bergh J, and Pavlidis N, *ESMO Minimum Clinical Recommendations for diagnosis, treatment and follow-up of cancers of unknown primary site (CUP).* Ann Oncol, 2005. **16**:75-76.
6. Borup R, Rossing M, Henao R, Yamamoto Y, Krogdahl A, Godballe C, et al., *Molecular signatures of thyroid follicular neoplasia.* Endocr Relat Cancer.2010; **17**(3): p. 691-708.
7. Bolstad BM, Irizarry RA, Astrand M, and Speed TP, *A comparison of normalization methods for high density oligonucleotide array data based on variance and bias.* Bioinformatics, 2003. **19**(2): p. 185-93.
8. Greco D, Somervuo P, Di Lieto A, Raitila T, Nitsch L, Castren E, et al., *Physiology, pathology and relatedness of human tissues from gene expression meta-analysis.* PLoS One, 2008. **3**(4): p. e1880.
9. Talantov D, Baden J, Jatkoe T, Hahn K, Yu J, Rajpurohit Y, et al., *A quantitative reverse transcriptase-polymerase chain reaction assay to identify metastatic carcinoma tissue of origin.* J Mol Diagn, 2006. **8**(3): p. 320-9.
10. Tothill RW, Kowalczyk A, Rischin D, Bousioutas A, Haviv I, van Laar RK, et al., *An expression-based site of origin diagnostic method designed for clinical application to cancer of unknown origin.* Cancer Res, 2005. **65**(10): p. 4031-40.

11. Ma XJ, Patel R, Wang X, Salunga R, Murage J, Desai R, et al., *Molecular classification of human cancers using a 92-gene real-time quantitative polymerase chain reaction assay.* Arch Pathol Lab Med, 2006. **130**(4): p. 465-73.

12. Su AI, Welsh JB, Sapinoso LM, Kern SG, Dimitrov P, Lapp H, et al., *Molecular classification of human carcinomas by use of gene expression signatures.* Cancer Res, 2001. **61**(20): p. 7388-93.

13. Bloom G, Yang IV, Boulware D, Kwong KY, Coppola D, Eschrich S, et al., *Multi-platform, multi-site, microarray-based human tumor classification.* Am J Pathol, 2004. **164**(1): p. 9-16.

14. Al-Brahim N, Ross C, Carter B, and Chorneyko K, *The value of postmortem examination in cases of metastasis of unknown origin-20-year retrospective data from a tertiary care center.* Ann Diagn Pathol, 2005. **9**(2): p. 77-80.

15. Varadhachary GR, Edmonston, T. B., Karanth, S., Carlson, H. R., Lebanony, D. , Rosenwald, S., Lenzi, R., Spector, Y., Cohen, D., and Raber, M. N. , *Prospective gene signature study using microRNA to predict the tissue of origin (ToO) in pts with cancer of unknown primary site (CUP).* J. Clini. Oncol., 2011;17(12):4063-70 .

16. van Laar RK, Ma XJ, de Jong D, Wehkamp D, Floore AN, Warmoes MO, et al., *Implementation of a novel microarray-based diagnostic test for cancer of unknown primary.* Int J Cancer, 2009. **125** : p. 1390-7.

17. Varadhachary GR, Talantov D, Raber MN, Meng C, Hess KR, Jatkoe T, et al., *Molecular profiling of carcinoma of unknown primary and correlation with clinical evaluation.* J Clin Oncol, 2008.;**26**(27): p. 4442-8.

18. Morawietz L, Floore A, Stork-Sloots L, Folprecht G, Buettner R, Rieger A, et al., *Comparison of histopathological and gene expression-based typing of cancer of unknown primary.* Virchows Arch. 2009;**456**(1): p. 23-9.

19. Monzon FA, Medeiros F, Lyons-Weiler M, and Henner WD, *Identification of tissue of origin in carcinoma of unknown primary with a microarray-based gene expression test.* Diagn Pathol. 2010;**5**: p. 3.

20. Horlings HM, van Laar RK, Kerst JM, Helgason HH, Wesseling J, van der Hoeven JJ, et al., *Gene expression profiling to identify the histogenetic origin of metastatic adenocarcinomas of unknown primary.* J Clin Oncol, 2008. **26**(27): p. 4435-41.

21. Greco FA, Spigel DR, Yardley DA, Erlander MG, Ma XJ, and Hainsworth JD, *Molecular profiling in unknown primary cancer: accuracy of tissue of origin prediction.* Oncologist. 2010;**15**(5): p. 500-6.

22. Hainsworth JD, Henner, W. D. , Pillai, R., and Greco, F. A., *Molecular tumor profiling in the diagnosis of patients with carcinoma of unknown primary (CUP): Retrospective evaluation of the Tissue of Origin Test (Pathwork Diagnostics).* J. Clini. Oncol., 2010; 15(5):500-6

23. Bridgewater J, van Laar R, Floore A, and Van TVL, *Gene expression profiling may improve diagnosis in patients with carcinoma of unknown primary.* Br J Cancer, 2008. **98**(8): p. 1425-30.

24. Pentheroudakis G, Golfinopoulos V, and Pavlidis N, *Switching benchmarks in cancer of unknown primary: from autopsy to microarray.* Eur J Cancer, 2007. **43**(14): p. 2026-36.

25. Huang RP, Fan Y, de Belle I, Niemeyer C, Gottardis MM, Mercola D, et al., *Decreased Egr-1 expression in human, mouse and rat mammary cells and tissues correlates with tumor formation.* Int J Cancer., 1997. **72**(1): p. 102-9.

26. Calogero A, Arcella A, De Gregorio G, Porcellini A, Mercola D, Liu C, et al., *The early growth response gene EGR-1 behaves as a suppressor gene that is down-regulated independent of ARF/Mdm2 but not p53 alterations in fresh human gliomas.* Clin Cancer Res., 2001. **7**(9): p. 2788-96.

27. Levin WJ, Press MF, Gaynor RB, Sukhatme VP, Boone TC, Reissmann PT, et al., *Expression patterns of immediate early transcription factors in human non-small cell lung cancer. The Lung Cancer Study Group.* Oncogene., 1995. **11**(7): p. 1261-9.

28. Baron V, Adamson ED, Calogero A, Ragona G, and Mercola D, *The transcription factor Egr1 is a direct regulator of multiple tumor suppressors including TGFbeta1, PTEN, p53, and fibronectin.* Cancer Gene Ther, 2006. **13**(2): p. 115-24.

29. Silvera D, Formenti SC, and Schneider RJ, *Translational control in cancer.* Nature.2010; **10**(4): p. 254-66.

30. Bochar DA, Wang L, Beniya H, Kinev A, Xue Y, Lane WS, et al., *BRCA1 is associated with a human SWI/SNF-related complex: linking chromatin remodeling to breast cancer.* Cell., 2000. **102**(2): p. 257-65.

31. Osley MA, Tsukuda T, and Nickoloff JA, *ATP-dependent chromatin remodeling factors and DNA damage repair.* Mutat Res., 2007. **618**(1-2): p. 65-80. .

32. Kauffmann A, Rosselli F, Lazar V, Winnepenninckx V, Mansuet-Lupo A, Dessen P, et al., *High expression of DNA repair pathways is associated with metastasis in melanoma patients.* Oncogene, 2008. **27**(5): p. 565-73.

33. Winnepenninckx V, Lazar V, Michiels S, Dessen P, Stas M, Alonso SR, et al., *Gene expression profiling of primary cutaneous melanoma and clinical outcome.* J Natl Cancer Inst, 2006. **98**(7): p. 472-82.

34. Sarasin A and Kauffmann A, *Overexpression of DNA repair genes is associated with metastasis: a new hypothesis.* Mutat Res., 2008. **659**(1-2): p. 49-55..

35. Mathews LA, Cabarcas SM, Hurt EM, Zhang X, Jaffee EM, and Farrar WL, *Increased expression of DNA repair genes in invasive human pancreatic cancer cells.* Pancreas. 2011;**40**(5): p. 730-9.

36. Halazonetis TD, Gorgoulis VG, and Bartek J, *An oncogene-induced DNA damage model for cancer development.* Science., 2008. **319**(5868): p. 1352-5.

37. Hicks WM, Kim M, and Haber JE, *Increased mutagenesis and unique mutation signature associated with mitotic gene conversion.* Science. 2010;**329**(5987): p. 82-5.

38. Klein CA, *Parallel progression of primary tumours and metastases.* Nat Rev Cancer., 2009. **9**(4): p. 302-12.

LEGENDS TO FIGURES

Figure 1. **A**. One-way hierachial cluster of 16 tumor classes by the 428 transcript signature. The tumor classes are shown at the top of the cluster and the transcripts are clustered at the left side. The scale bar shows the color coding of the relative expression values. **B**. Principal component analysis (PCA) of primary tumors and known metastases based on the 641 transcript signature. The tumor classes are colored and indicated in association with the corresponding tumor samples.

Figure 2. Patomarkers in primary tumors and CUP. Probeset Ids for 45 common histopatological markers were collected and used to generate a two-way hierarchal cluster with a selection of primary tumors **(Panel A)** or CUP **(Panel B)**. The variance of the individual markers is show to the left and the scale is indicated at the top of the clusters. Gene symbols are shown to the right and the different tumor classes are shown below (**(Panel A)**, primary tumors). For the CUP samples **(Panel B),** groups of markers corresponding to different tumor classes are indicated by the boxes around the gene symbols at the right side of the cluster. The number below the cluster indicated the number of the CUP sample corresponding to the annotation in Table 1.

Figure 3. QDA derived outlier scores in CUP. A) To determine the relationship between prediction error and outlier scores, the primary cancers and metastases were divided into ten bins according to the outlier scores and the error rate was calculated for each bin. Each point represents the error rate plotted versus the median outlier score of the bin. The vertical lines show the span of outlier scores within the bins. The plot shows that higher outlier score translate into higher error rate. We modeled the relationship between outlier scores and prediction error by fitting polynomial function to the data points (the orange line), the function allows us to estimate the expected error rate for new samples of unknown origin, once their outlier scores have been determined. **B)** Samples from CUP patients tend to have higher outlier scores than other cancer patients. The box plot summarizes the distributions of outlier scores within metastases (MET), primary (PRIM) and CUP tumors. There is a clear tendency for CUP samples to have higher outlier score than metastases and primary cancers. The median outlier score of CUP samples of >1000 suggest the origin prediction error above 30 %. On the other hand, most primary cancers and metastases have outlier scores below 800, hence the estimated prediction error from 2-10 % (see panel A). Since data for CUP and some primary tumors and metastases were generated at Rigshospitalet, the non-CUP samples from Rigshospitalet are presented as separate group (RH_MET and RH-PRIM), this is to show that the shift in outlier scores was not caused by technical bias. Additionally, the normal, non-cancerous tissue group (NORMAL) is included, and shows the whole range of outlier scores.

Figure 4. Two way hierachial clusters of BRCA1 and SMARCA2 networks in metastases and CUP. A. The PUJANA_BRCA1_PCC_NETWORK was downloaded from the MSig database (http://www.broadinstitute.org/gsea/msigdb) and used to generate a paired two way hierachial cluster with known metastases and CUP. Gene symbols were translated into probe sets and because of the probe set redundancy the data were filtered by a $p<0.001$ before clustering. Following filtering 1297 probe sets were included in the clustering. Known metastases are indicated in green and CUP samples are labeled with pink above the

cluster. The scale is shown at the right side of the cluster. (B) Two-way cluster of the SHEN_SMARCA2_ TARGETS including both up- and down-regulated transcripts. The set consists of 360 down- and 430 up-regulated genes that translated into 772 and 1211 probe sets, respectively. As above known metastases are indicated in pink and CUP samples are labeled with green below the cluster. The scale is shown at the right side of the cluster.

ID	Sex/age	Biopsy site	Histology	Path Diag.	Stand of Ref	LDA Pred	Outlier score
14.	F/56	LN neck	PDC	Lung	Lung (CD)	Lung	975
17.	F/57	LN neck	Adenoc.	Lower GI	Colon (RD)	Colon	746
22.	M/55	LN neck	Adenoc.	CUP	Stomach (RD)	Normal	934
23.	M/39	LN retro	PDC	CUP	Kidney (RD)	Kidney	1085
28.	F/58	Peritoneum	PDA	Ovary	Ovary (RD)	Ovary	810
31.	F/40	LN neck	PDA	CUP	Lung (CD)	Stomach	985
34	M/74	Skin	PDA	Lung	Lung (RD)	Lung	898
39.	M/71	Liver	Adenoc.	CUP	Pancreas (CD)	Pancreas	1097
40.	F/44	Liver	Adenoc.	Colon	Colon (RD)	Colon	729
44.	F/43	Kidney	Carc.	CUP	Bladder (RD)	Bladder	1286
49.	M/60	LN neck	PDA	Kidney	Kidney (RD)	Kidney	1223
51.	F/42	LN pelvis	SCC	CUP-SCC	Cervical (RD)	Cervix	828
52.	M/53	Liver	PDA	CUP	CCC (RD)	CCC	923
53.	M/70	Liver	Adenoc.	Lung	Lung (RD)	Lung	1047
57.	M/67	Liver	Adenoc.	CCC	CCC (RD)	HCC	965
66.	F/68	Liver	PDA	CUP	CCC (RD)	Cervix	1100
70.	M/38	Peritoneum	Adenoc.	Stomach	Stomach (CD)	Colon	842
74.	M/62	Leg	Carc.	Adnex tumor	Adnex tumor (RD)	Normal	1010
76.	M/64	Liver	Adenoc.	Lower GI	Small intestine (RD)	Colon	912
77.	M/59	LN axilla	PDC	CUP	Lung (CD)	Breast	978
86.	F/61	LN axilla	Adenoc.	CUP	Lung (RD)	Stomach	1108
88.	F/36	Peritoneum	Adenoc.	Ovary	Ovary (RD)	Cervix	1033
89.	F/57	Liver	PDA	CCC	CCC (RD)	CCC	916
90.	F/71	Peritoneum	Adenoc	Ovary	Ovary (RD)	Ovary	781
92.	M/62	Liver	Malignant tumor	Angiosarcoma	Angiosarcoma (RD)	Normal	1097
95.	M/45	Peritoneum	PDC	DSRCT	DSRCT (RD)	Breast	1098
71+72	M/61	Bone + Kidney	PDC	Kidney	Kidney (RD)	Kidney	1096 1277
75+87	F/43	Liver	PDA	CCC	CCC (RD)	CCC	925 1030

ID	Sex/age	Biopsy site	Histology	Path Diag.	Stand of Ref	LDA Pred	Outlier score
11.	F/58	LN neck	PDA	CUP	CUP (SD)	Ovary	756
13.	F/72	Peritoneum	PDA	CUP	CUP (NSD)	Pancreas	1193
21.	M/63	LN neck	PDC	CUP	CUP (NSD)	Breast	1108
26.	F/67	Skin	PDA	CUP	CUP (NSD)	Breast	971
32.	M/53	LN neck	PDSCC	CUP-SCC	CUP (NSD)	Normal	926
33.	M/58	Skin	PDA	CUP	CUP (NSD)	Colon	1098
41.	M/74	Liver	PDA	Pancreas	CUP (NSD)	Stomach	1040
42.	M/56	Liver	Adenoc.	CUP	CUP (NSD)	Pancreas	994
43.	F/50	LN retro	PDA	CUP	CUP (NSD)	Stomach	797
45.	M/44	Liver	PDC	CUP	CUP (NSD)	Colon	1245
46.	F/76	Liver	Adenoc.	CUP	CUP (NSD)	Normal	1027
47.	F/59	Liver	Adenoc.	CUP	CUP (SD)	CCC	932
48.	F/59	LN neck	PDC	CUP	CUP (NSD)	Ovary	1032
54.	F/67	Liver	Adenoc.	CUP	CUP (NSD)	Normal	1068
55	F/55	Liver	Adenoc.	CUP	CUP (NSD)	Normal	962
58.	F/67	Liver	PDC	CUP	CUP (SD)	CCC	995
61.	M/72	Liver	Carc.	HCC	CUP (NSD)	CCC	1102
64.	F/65	LN inguien	PDA	CUP	CUP (SD)	Lung	1168
65.	M/62	LN neck	PDSCC	CUP-SCC	CUP (NSD)	Breast	929
73.	M/43	LN retro	PDC	CUP	CUP (NSD)	Normal	1020
78.	F/59	Lung	Adenoc.	Lower GI	CUP (NSD)	Lung	1062
80.	F/58	Liver	Adenoc.	CUP	CUP (SD)	CCC	1111
81.	F/71	Liver	PDA	CUP	CUP (NSD)	Breast	1212
82.	F/56	Bone	Adenoc.	CUP	CUP (NSD)	CCC	1209
83.	F/59	Liver	PDA	CUP	CUP (SD)	CCC	1061
91.	F/65	LN axilla	Adenoc.	CUP	CUP (SD)	Lung	939
93.	M/58	Bone	PDSCC	CUP-SCC	CUP (NSD)	Breast	940
94.	F/55	Liver	PDA	CUP	CUP (NSD)	Normal	984
50. + 68	M/41	Adr gl	PDC	CUP	CUP (NSD)	Stomach Pancreas	978 1079

Table 1. Prediction results in CUP patients. A validation of the LDA predicted diagnoses was performed by comparing with a Standard of Reference (SR). SR was established by an experienced pathologist and two experienced oncologists. In addition to the 23 patients where a primary tumor site was identified (Reference Diagnosis (RD)) within the study period, the Standard of Reference reached a Consensus Diagnosis (CD) in 5 patients based on patient demographics, metastatic pattern, results of clinical and laboratory tests, imaging data and pathologic evaluations (Samples labeled in red). In the 29 remaining CUP patients (labeled in blue), the results from gene expression profiling were compared with clinicopathological features and the predictions were categorized as Supportive (SD) or Non-Supportive (NSD). LN: lymph node; n: neck LN; m: mediastinal LN; a: axilla LN; r: retroperitoneal LN; p: pelvis LN; adr gl: adrenal gland; Adenoc: adenocarcinoma, PDA: poorly differentiated adenocarcinoma; Carc: carcinoma; PDC: poorly differentiated carcinoma; SCC: squamous cell carcinoma; PDSCC: poorly differentiated SCC; CCC: cholangiocarcinoma; HCC: hepatocelluar carcinoma; DSRCT: desmoplastic small round cell tumor. Path Diag: pathological diagnosis; Stand of ref: Standard of reference; LDA pred: Linear discriminant analysis prediction; RD: Reference Diagnosis; CD: Consensus Diagnosis, SD: Supportive Diagnosis; NSD: Non-Supportive Diagnosis

Up-regulated in CUP

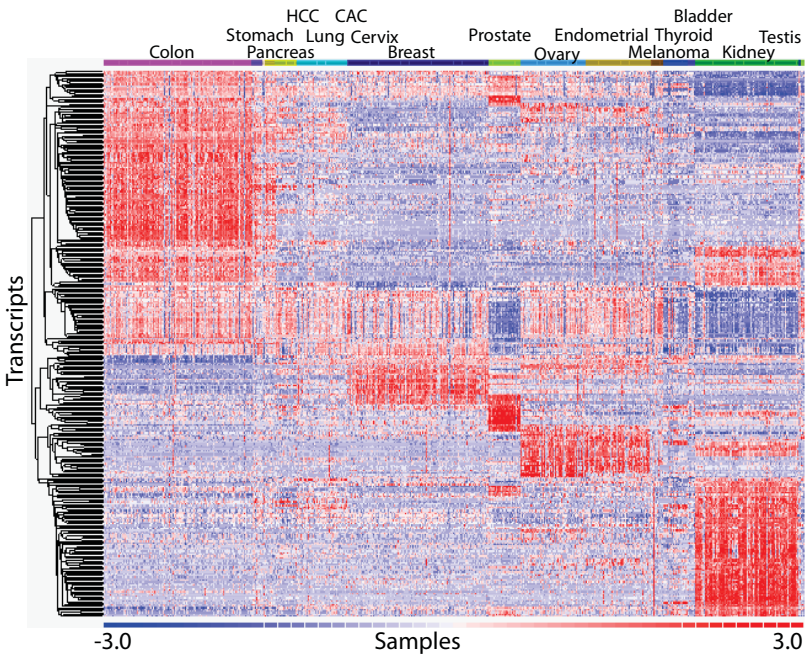
Gene Set Name	Transcripts	Overlap	k/K	p value
PUJANA_BRCA1_PCC_NETWORK Genes constituting the BRCA1-PCC network of transcripts whose expression positively correlated (Pearson correlation coefficient, PCC >= 0.4) with that of BRCA1	1671	159	0.0952	0.00E+00
KINSEY_TARGETS_OF_EWSR1_FLII_FUSION_UP Genes up-regulated in TC71 and EWS502 cells (Ewing's sarcoma) upon knockdown of the EWSR1-FLII fusion	1281	133	0.1038	0.00E+00
PUJANA_ATM_PCC_NETWORK Genes constituting the ATM-PCC network of transcripts whose expression positively correlated (Pearson correlation coefficient, PCC >= 0.4) with that of ATM	1461	152	0.104	0.00E+00
PUJANA_CHEK2_PCC_NETWORK Genes constituting the CHEK2-PCC network of transcripts whose expression positively correlates (Pearson correlation coefficient, PCC >= 0.4) with that of CHEK2	782	89	0.1138	0.00E+00
DODD_NASOPHARYNGEAL_CARCINOMA_DN Genes down-regulated in nasopharyngeal carcinoma (NPC) compared to the normal tissue.	1375	157	0.1142	0.00E+00
RODRIGUES_THYROID_CARCINOMA_ANAPLASTIC_UP Genes up-regulated in anaplastic thyroid carcinoma (ATC) compared to normal thyroid tissue.	721	93	0.129	0.00E+00
MILI_PSEUDOPODIA_HAPTOTAXIS_UP Transcripts enriched in pseudopodia of NIH/3T3 cells (fibroblast) in response to haptotactic migratory stimulus by fibronectin, FN1	552	74	0.1341	0.00E+00
RODRIGUES_THYROID_CARCINOMA_POORLY_DIFFERENTIATED_UP Genes up-regulated in poorly differentiated thyroid carcinoma (PDTC) compared to normal thyroid tissue.	640	94	0.1469	0.00E+00
DACOSTA_UV_RESPONSE_VIA_ERCC3_DN Genes down-regulated in fibroblasts expressing mutant forms of ERCC3 [Gene ID=2071] after UV irradiation.	855	126	0.1474	0.00E+00
DACOSTA_UV_RESPONSE_VIA_ERCC3_COMMON_DN Common down-regulated transcripts in fibroblasts expressing either XP/CS or TDD mutant forms of ERCC3 [Gene ID=2071], after UVC irradiation.	420	64	0.1524	0.00E+00
OSMAN_BLADDER_CANCER_UP Genes up-regulated in blood samples from bladder cancer patients.	402	57	0.1418	5.55E-16
SENGUPTA_NASOPHARYNGEAL_CARCINOMA_WITH_LMP1_UP Genes up-regulated in nasopharyngeal carcinoma (NPC) positive for LMP1 [Gene ID=9260], a latent gene of Epstein-Barr virus (EBV).	399	56	0.1404	1.55E-15
SENGUPTA_NASOPHARYNGEAL_CARCINOMA_UP Genes up-regulated in nsopharyngeal carcinoma relative to the normal tissue.	286	46	0.1608	3.33E-15
PUJANA_XPRSS_INT_NETWORK Genes constituting the XPRSS-Int network: intersection of genes whose expression correlates with BRCA1, BRCA2, ATM, and CHEK2 [Gene ID=672, 675, 472, 11200] in a compendium of normal tissues.	167	34	0.2036	1.21E-14

Down-regulated in CUP

Gene Set Name	Transcripts	Overlap	k/K	p value
SHEN_SMARCA2_TARGETS_DN Genes whose expression negatively correlated with that of SMARCA2 [Gene ID=6595] in prostate cancer samples.	360	73	0.2028	0.00E+00
GINESTIER_BREAST_CANCER_ZNF217_AMPLIFIED_DN Genes down-regulated in non-metastatic breast cancer tumors having type 1 amplification in the 20q13 region; involves ZNF217 [Gene ID=7764] locus only.	336	49	0.1458	7.71E-11

Table 2. Enriched or depleted gene sets in CUP compared to metastases of known origin. Gene set enrichments among up or down regulated mRNAs in the CUP core set were examined in the molecular signatures database (MSig) among the C2 curated gene sets comprising profiles from chemical and genetic perturbations, canonical pathways, BIOCARTA, KEGG and the reactome collections.

A.


One-way hierachial cluster of primary tumors - 428 signature
Colon
Stomach
Pancreas
HCC
Lung
CAC
Cervix
Breast
Prostate
Ovary
Endometrial
Thyroid
Melanoma
Bladder
Kidney
Testis
Transcripts
-3.0
Samples
3.0

B.

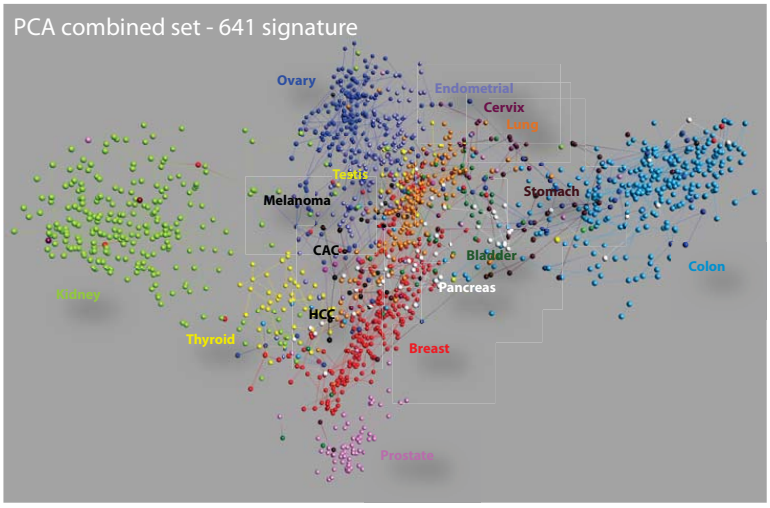

PCA combined set - 641 signature
Ovary
Endometrial
Cervix
Lung
Testis
Melanoma
Stomach
CAC
Bladder
Colon
Pancreas
Kidney
HCC
Thyroid
Breast
Prostate

Figure 1

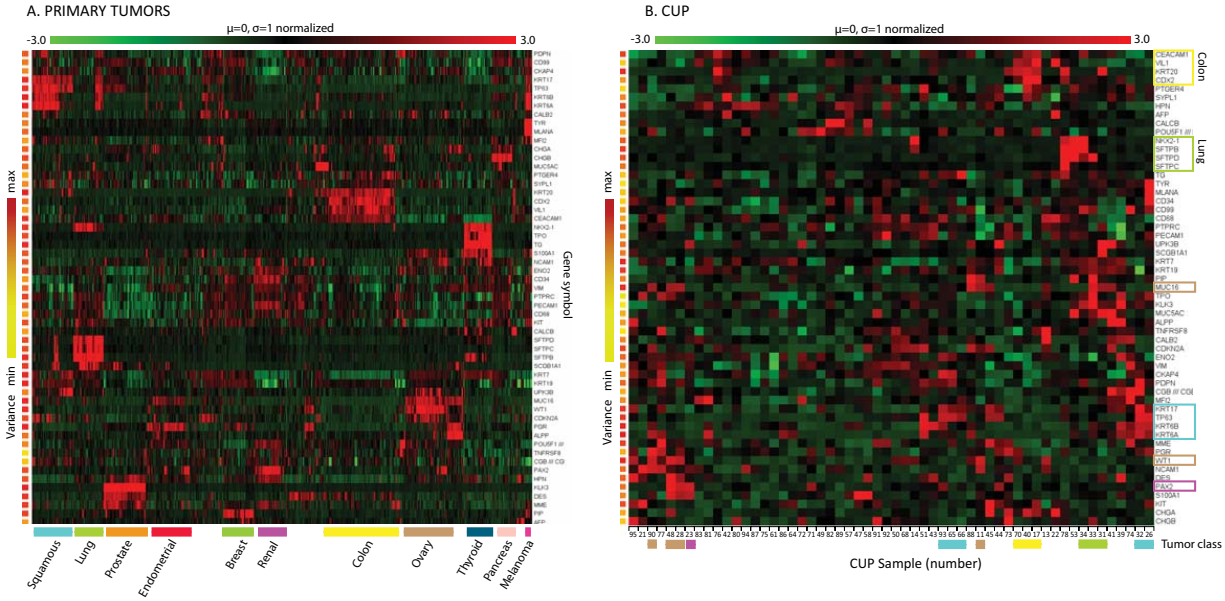

A. PRIMARY TUMORS
μ=0, σ=1 normalized
-3.0
3.0
max
min
Variance
Gene symbol
Squamous
Lung
Prostate
Endometrial
Breast
Renal
Colon
Ovary
Thyroid
Pancreas
Melanoma
B. CUP
μ=0, σ=1 normalized
-3.0
3.0
max
min
Variance
Colon
Lung
Tumor class
CUP Sample (number)

Figure 2

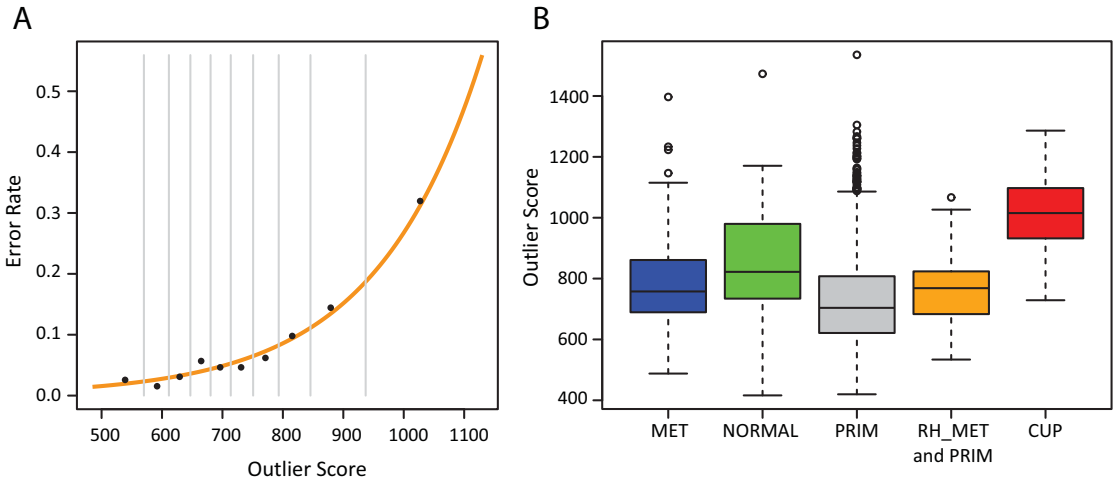

A
Error Rate
0.0
0.1
0.2
0.3
0.4
0.5
500
600
700
800
900
1000
1100
Outlier Score
B
Outlier Score
400
600
800
1000
1200
1400
MET
NORMAL
PRIM
RH_MET and PRIM
CUP

Figure 3

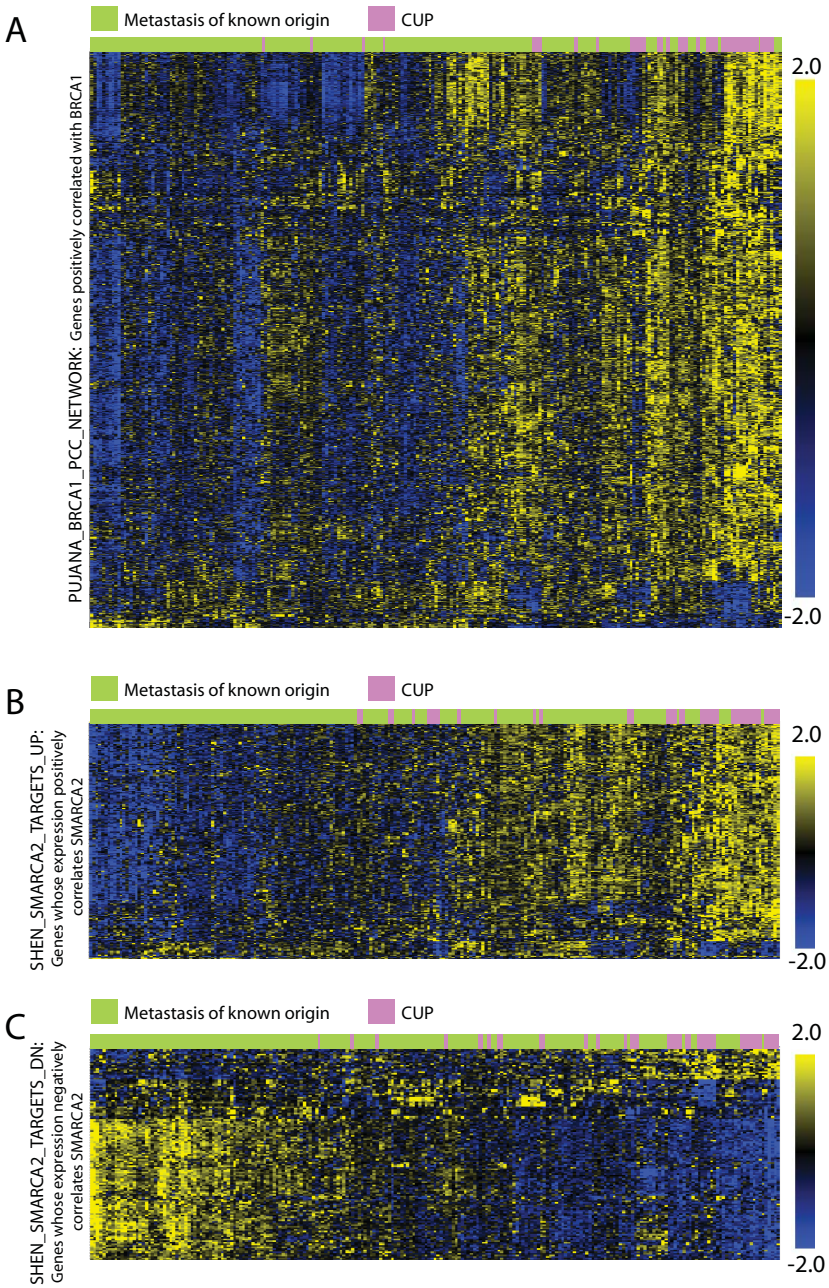

A
Metastasis of known origin
CUP
PUJANA_BRCA1_PCC_NETWORK: Genes positively correlated with BRCA1
2.0
-2.0
B
Metastasis of known origin
CUP
SHEN_SMARCA2_TARGETS_UP: Genes whose expression positively correlates SMARCA2
2.0
-2.0
C
Metastasis of known origin
CUP
SHEN_SMARCA2_TARGETS_DN: Genes whose expression negatively correlates SMARCA2
2.0
-2.0

Figure 4

Appendix D

Paper IV
Somatic copy-number alteration can help predict the tissue origin of cancers of unknown primary

Somatic copy-number alteration can help predict the tissue origin of cancers of unknown primary

Bogumil Kaczkowski[a*], Rileen Sinha[b] , Nikolaus Schultz[b], Chris Sander[b], Finn Cilius Nielsen[c] and Ole Winther[d*]

a) The Bioinformatics Centre, Department of Biology and Biomedical Research and Innovation Centre, Copenhagen University, Ole Maaloes Vej 5, 2200 Copenhagen, Denmark
b) Computational Biology Center, Memorial Sloan-Kettering Cancer Center, New York, New York 10065, USA
c) Department of Clinical Biochemistry, Copenhagen University Hospital, Blegdamsvej 5, 2100 Copenhagen, Denmark
d) DTU Informatics, Technical University of Denmark, 2800 Lyngby, Denmark

* Corresponding authors

DRAFT

Abstract

Cancer of Unknown Primary (CUP) is a heterogeneous disease that represents 3-5% of all new cancer cases. The unknown origin poses a challenge for treatment of CUP patients. Here, we investigated the DNA copy number profiles from 3573 tumors of 19 origin, 3796 normal tissue samples and 639 cancer cell lines. Subsequently, we built the classifier that could predict the origins of 72% primary tumors with 95% accuracy. The tumors with low number of somatic copy-number alterations were predicted as normal tissue. We applied the classifier to predict the origin of the cell lines and 33% of cancer cell lines were predicted with high confidence. Glioma, kidney, head & neck, breast and colon cancer cell lines were predicted with very high accuracy, whereas lung and pancreas cancer cell lines were mostly misclassified. We propose, that DNA copy number profiles can be used to predict the primary site of CUP and can complement available messengerRNA and miRNA based classifiers.

Introduction

Cancer of Unknown Primary (CUP) represents 3-5% of all new cancer cases, which places it among the 10 most common malignancies in developed societies (Pavlidis & Pentheroudakis, 2010). CUP is a highly aggressive disease and it is the fourth most common cause of cancer deaths in both sexes (Pavlidis & Pentheroudakis, 2012). The disease is a syndrome representing many types of cancer, which exhibit early dissemination even though the primary tumor is too small to be detected, have multiple metastasis, and have an unpredictable metastatic pattern (Pavlidis & Pentheroudakis, 2010).The management of CUP is often focused on identifying the most probable primary site and adapting the treatment accordingly (Greco, 2010)

There exist several molecular assays designed to predict the primary site of CUPs. The Pathwork Tissue of origin test (Monzon et al., 2009) and CancerType ID test(Ma et al., 2006) predict the primary site based on the messengerRNA expression level, using microarray and qRT-PCR platform, respectively. Another assay, miRview mets, uses the expression levels of 48 miRNAs measured by qRT-PCR (Rosenfeld et al., 2008). Those methods were reported to have overall accuracies of 82-90%, which is based on cancers of know origin. The performance on CUP patients is unknown.

Cancer samples are often contaminated with normal tissue. This has as impact on miRNA and mRNA expression profiles. If the classifier is trained on primary cancer samples, components of normal tissue specific expression are used to predict the origin. In case of metastases, contamination with normal tissue biases the expression. The contamination with normal tissue also affects the prediction. This can be especially problematic when working with a biopsy of small metastatic cancers, which is a common scenario for CUP.

Copy Number Variations (CNV) are a form of structural variation of genomic DNA, exhibited by a changing number of copies of a DNA segments that range from kilobase to several megabases of nucleotides (Stankiewicz, 2010). Germline CNVs are observed when a DNA sequence is found at different copy numbers in different individuals, which is due to Copy Number Polymorphism (CNP). De novo copy number changes, for example acquired during cancer developments are referred to as Somatic Copy Number Alteration (SCNA) (Beroukhim et al., 2010).

Here, we use the DNA copy-number profiles of 3573 tumors from The Cancer Genome Atlas (TCGA) project to build a classifier that could predict the primary site based on the CNV pattern. Additionally, we tested the classier on the cancer cell line data from the Cancer Cell Line Encyclopedia (CCLE) project.

Material and Methods

Data Sets

The pre-processed, segmented DNA copy number data (.seg) were downloaded from The Cancer Genome Atlas (TCGA) project and The Cancer Cell Line Encyclopedia (CCLE) project The seg files were generated by the Circular Binary Segmentation algorithm (Olshen, Venkatraman, & Lucito, 2004) using the hybridization data from The Genome-Wide Human SNP Array 6.0 microarray platform. For each segment of DNA in every patient, the seg files contains information about chromosome, start location, end location of the segment as well as log2 ratio, which represents the changes in DNA copy number compared to control. The positive log2 ratios correspond to amplification and the negative values correspond to deletion of DNA fragments. The segmented copy number data of 3573 primary cancers, representing 19 cancer types, were downloaded from the data portal of the TCGA project (https://tcga-data.nci.nih.gov/tcga/). The list of cancer types and the number of samples is presented in Table 1. Additionally, the 3796 normal tissue profiles of TCGA project were downloaded.

The segmented copy number data of 883 cancer cell lines were downloaded from the CCLE project website (http://www.broadinstitute.org/ccle/home). Sixteen cancer types of the CCLE data set, that were closest to the types present in TCGA data, were used for the analysis, leaving 639 samples. The types and number of cases are shown in Table 2.

Copy-Number Values per Gene

Samples differ in the number and length of altered regions and the altered segments rarely overlap between the samples. Therefore, the data in seg format cannot be used directly with common statistical and machine learning methods. In order to transform the seg file into the matrix format, the genome was divided by all annotated sequence taken from the Reference Sequence (RefSeq) collection (hg18), which was a downloaded from the UCSC Genome Browser (http://genome.ucsc.edu/). Each annotated sequence, referred further to as a "gene", was assigned the log2 ratio of overlapping segment, which resulted in 20647 gene copy-number values for each patient sample. This and all following analysis were performed in R.

Clustering and Cluster Averaging

In order to reduce the dimensionality and redundancy of the dataset of copy number per gene values, the genes were clustered into groups of gene of similar copy number profile across 3573 primary cancers and the values were averaged to represent one log2 ratio per cluster. 1-correlation of the copy number profiles was used as a distance measure between the genes. Clustering was performed using bottom-up, hierarchical clustering and "ward" agglomeration method. Each chromosome was clustered into the number of chromosomal regions (CR) that corresponded to a number of features on the chromosome divided by 20. The gene values of each CR were summarized using mean of genes within it. This procedure yielded 1,030 CR copy number ratios for each patient.

The clusters of genes from TCGA tumor data were used to summarize the TCGA normal and CCLE cell line data.

Building a Classifier

The CR log2 values were used as features in the classification. The CR values, that had less than 5 genes or had a standard deviation of values below .20, were excluded, which resulted in 790 features (Figure 1).

The classifier was trained on the TCGA tumor data set, which represents 19 primary tumor types. Colon and rectal tumor classes were merged into colorectal class due to high similarity and shared biology. Additionally, a "normal" class was created by adding 500 samples from the TCGA normal tumor data set.

Two different methods were used for the classification: K-Nearest Neighbor (KNN, k= 3) and Linear Discriminant Analysis (LDA). In cases where the two methods agreed, the prediction is referred to as a high confidence prediction. In cases of disagreement, the prediction is regarded as having low confidence.

The F-test was applied as a feature selection method, and a threshold of p-value < 10^-70, which led to about 563 features used to predict the tumor type of the CCLE cancer cell line panel. The accuracy of the prediction on primary cancers was estimated through 10-fold cross validation. The feature selection was included in the cross validation loop to avoid over fitting.

Results

The copy number data of 3573 primary tumors, 3796 normal tissues and 639 cancer cell lines were downloaded in seg file format. The cancer origins and the number of samples per origin are listed in Table 1 and Table 2, for primary tumors and cancer cell lines, respectively.

Table 1 List of cancer types in the training set. The DNA copy number variation profiles for 3573 primary tumor samples were downloaded from The Cancer Genome Atlas (TCGA) project data portal.

Acronym	Cancer type	Origin	Number of samples	Color
BLCA	Bladder Urothelial Carcinoma	bladder	32	yellow
BRCA	Breast invasive carcinomaæ	breast	506	red
CESC	Cervical squamous cell carcinoma and endocervical adenocarcinoma	cervix	23	orange
COREAD	Colon and rectum adenocarcinoma	colorectal	472	black
GBM	Glioblastoma multiforme	glioblastoma	516	blue
HNSC	Head and Neck squamous cell carcinoma	head_and_neck	50	deeppink
KIRC	Kidney renal clear cell carcinoma	kidney_RC	91	brown
KIRP	Kidney renal papillary cell carcinoma	kidney_RP	494	beige
LGG	Brain Lower Grade Glioma	glioma	52	cyan
LIHC	Liver hepatocellular carcinoma	liver	44	slateblue
LUAD	Lung adenocarcinoma	lung_AD	98	greenyellow
LUSC	Lung squamous cell carcinoma	lung_SC	167	green
OV	Ovarian serous cystadenocarcinoma	ovary	513	pink
PAAD	Pancreatic adenocarcinoma	pancreas	26	grey
PRAD	Prostate adenocarcinoma	prostate	82	lightblue
STAD	Stomach adenocarcinoma	stomach	107	magenta
THCA	Thyroid carcinoma	thyroid	35	beige
UCEC	Uterine Corpus Endometrioid Carcinoma	uterus	265	purple

Table 2 List of cancer cell line origins from The Cancer Cell Line Encyclopedia (CCLE) project.

Origin	No of samples	Color on heatmap
breast	52	red
glioma	44	cyan
head & neck	26	deeppink
kidney	20	brown
colon	51	black
liver	26	slateblue
lung	158	green
melanoma	56	blue
oesophagus	25	violet
ovary	43	pink
pancreas	40	grey
prostate	7	lightblue
stomach	37	magenta
thyroid	8	beige
urinary tract	23	yellow
uterus	23	purple

The seg files were converted to a matrix format with chromosomal regions (CR) as features. The copy number value per CR (CR value), was calculated as a mean of log2 ratios of genes with the region. CRs covered from 1 to more that 100 genes, the majority covered between 5 and 35 genes. The standard deviation of CR values within tumor samples ranged form 0.1 to 1.5 genes (Figure 1).

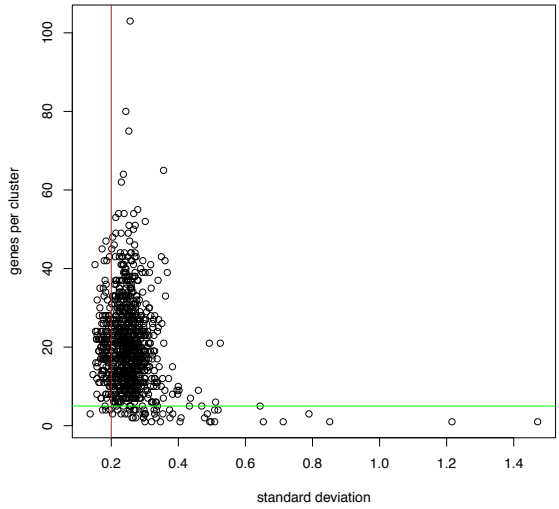


Figure 1 Chromosomal Regions (CR) as defined by the clustering of genes. Each chromosomal region (cluster) is represented as a circle. The y-axis shows the number of genes per region. Only regions of more than 5 genes were used for classification and the green line represent that threshold. The y-axis indicates the standard deviation of the CR values across the primary tumor samples. The samples with SD>0.2 were used for classification, as indicated by the red line.

Copy number changes show different patterns in the primary tumor, normal samples and cell lines (Figure 2). In normal samples, the observed variation of CR values is due to Copy Number Polymorphism (CNP), a structural variation that affects about 12% of human genomic DNA. CNVs can cover from a thousand to several million nucleotide bases (Stankiewicz, 2010). CNPs are a form of normal variation within the population and they are inherited, i.e. found in germline DNA. In this analysis, the variation of CR values within the normal tissue samples can be regarded as a background for acquired (somatic) copy number changes observed in tumor and cancer cell lines.

The variation in CR values observed in tumor samples comes from germline CNVs (visible in the same genomic location as CNV in normal samples) and Somatic Copy Number Alterations (SCNA) that were acquired during the cancer development (Figure 2A).

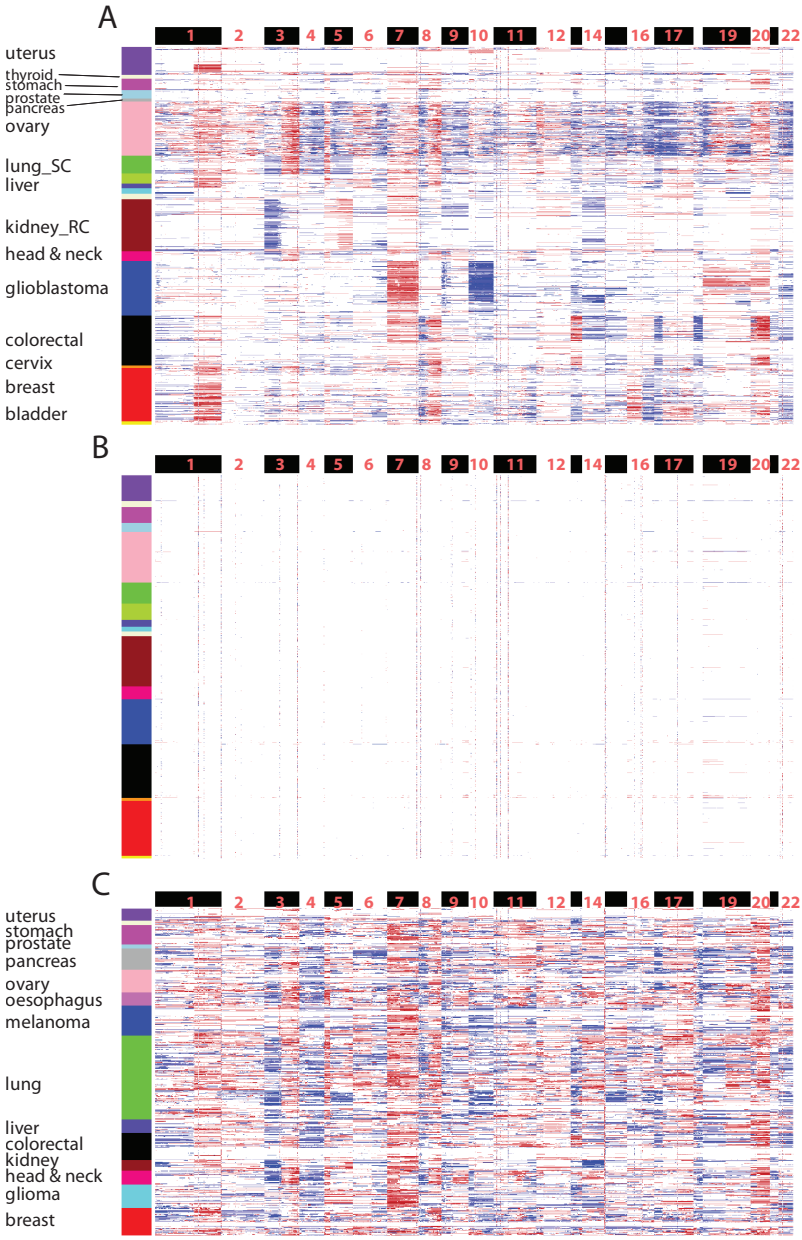


Figure 2 Heatmap representing the Copy Number Variations (CNV) within 3573 primay tumors of 18 origins (panel A), 3796 normal profiles from TCGA cancer patients (panel B) and 639 cell line profiles of Cancer Cell Line Encyclopedia (panel C). The presented cell lines are from 16 cancer types. Red and blue represent amplification and deletion, respectively.

Different patterns of SCNAs are observed in different tumor origins. For example, thyroid and prostate samples show minimal number of SCNAs, whereas ovarian cancers have vast numbers of SCNAs.

Clear patterns are visible in some origins, such as deletion of 3p chromosome arm in kidney cancer, or amplification of chromosome 7 and deletion of chromosome 10 in glioblastomas. However, in most cases, the amplifications and deletions pattern observed within some cancer should be considered enriched or more probable rather than a definite rule. Therefore a probabilistic classifier is needed in order to classify the cancer origin based on CNV data. It should be noted that in all origins, there are cases with minimal number of SCNA as well as cases with most of the genome altered. In some origin, most clearly in colorectal and uterine cancer, there seem to be two group, one resembling normal tissues i.e. just CNP present and the other group being heavily affected by SCNAs.

Copy number changes detected in cancer cell line (Figure 2C) originate from a) copy number polymorphism, b) SCNA acquired during carcinogenesis and c) SCNA acquired during growth and multiple passaging of the cells in the cell culture. As expected, cell lines tend to have much more CNVs than primary tumors. The deletion of 3p chromosome arm, amplification of chromosome 7 and deletion of chromosome 10 is now observed across almost all cancer origins. There also seem to be fewer differences between cell lines of different origin. Noticeably, there are subgroups within cancer cell lines of uterine, colorectal and melanoma origin that show "flat", normal tissue like, CNV profile.

In the next step, we investigate if CNV data can be successfully used to predict the primary site of cancer. We build classifier using the CR values as a feature set. We use two classifier methods: LDA and KNN. High confidence prediction is made when both methods agree and low confidence in case of disagreement. During 10-fold cross validation, 72% of primary tumors are predicted with high confidence with accuracy of 85%. A number of tumor samples from different origins are predicted as "normal". If cancers predicted as normal are not considered a prediction error, the accuracy is 95% (Table 3). Most origins were predicted with high accuracy. A number of cancers from each class were predicted as normal, most notably, majority of thyroid, pancreas, prostate and stomach cancers. This is a result of low number of SCNA in those cancers. Interestingly, the classifier is able to distinguish subtypes of the same cancer origin, such as renal papillary cell carcinoma from clear cell carcinoma and lung adenocarcinoma from squamous cell carcinoma.

The classification accuracy of low confidence prediction was 27% and 58% when cancers predicted as normal were excluded.

Table 3 The confusion matrix of high confidence predictions within the TCGA tumor data set. Columns represent the true origin and rows the predicted origin.

	bladder	breast	cervix	colorectal	glioblastoma	glioma	head & neck	kidney_RC	kidney_RP	liver	lung_AD	lung_SC	normal	ovary	pancreas	prostate	stomach	thyroid	uterus
bladder	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
breast	0	278	0	1	1	0	0	0	1	4	1	1	0	3	0	1	0	0	7
cervix	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
colorectal	0	3	0	319	0	0	0	0	0	0	1	0	0	1	0	1	7	0	0
glioblastoma	1	1	0	1	414	7	0	2	0	0	1	0	0	1	0	0	0	0	0
glioma	0	0	0	0	4	18	0	0	0	0	0	0	0	1	0	0	0	0	0
head & neck	1	1	1	0	0	0	26	0	0	0	0	5	0	0	0	0	0	0	0
kidney_RC	1	3	2	2	6	2	4	428	5	1	6	0	0	1	0	1	0	1	0
kidney_RP	0	0	0	0	0	0	0	4	25	0	0	0	0	0	0	0	0	0	0
liver	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0
lung_AD	0	1	0	0	0	0	0	0	0	0	12	2	0	1	0	0	0	0	0
lung_SC	1	3	3	0	0	0	8	0	0	0	5	90	0	1	0	0	1	0	0
normal	4	16	0	42	9	5	5	12	4	3	4	1	487	3	17	27	19	23	74
ovary	0	0	0	0	0	0	0	0	0	0	0	0	0	322	0	0	0	0	7
pancreas	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
prostate	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	19	0	0	0
stomach	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5	0	0
thyroid	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
uterus	0	5	1	3	0	0	0	0	0	1	4	1	0	6	0	0	4	1	75

The classifier was also used to predict the origin of cancer cell lines. Only, 33% of cancer cell lines were predicted with high confidence (Table 4). Remarkably, glioma, kidney, head & neck, breast and colon cancer cell lines were predicted with very high accuracy. Melanoma cell lines were predicted to glioma, which is the closest biological class within the classes represented in classifier. Evan though melanoma is a skin cancer, it origins from melanocytes, that are derived from neural crest cells which also give rise to Glial cells. The overall low accuracy of 43% is mostly caused by very poor predictions of lung and pancreas cancer cell lines (Table 4).

Table 4 The confusion matrix of high confidence predictions of CCLE cancer cell line panel. The classifier was trained on primary tumors. Columns represent the true origin and rows the predicted origin.

	urinary tract	breast	colon	glioma	melanoma	head & neck	oesophagus	kidney	liver	lung	ovary	pancreas	prostate	stomach	thyroid	uterus
bladder	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
breast	0	10	0	0	1	0	0	0	0	2	0	0	0	0	0	0
cervix	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
colorectal	0	0	13	0	0	2	0	0	2	1	0	3	0	0	1	0
glioblastoma	1	0	0	15	14	0	0	0	0	7	0	1	0	1	0	0
glioma	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
head & neck	2	0	0	0	1	9	4	0	0	1	0	2	0	1	0	0
kidney_RC	1	1	3	3	3	0	0	15	1	14	2	1	0	0	1	2
kidney_RP	0	0	0	0	0	0	0	2	0	1	0	0	0	0	0	0
liver	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
lung_AD	0	1	0	0	0	0	0	0	0	9	1	1	0	0	0	0
lung_SC	0	1	0	1	0	0	1	0	0	8	0	2	0	0	0	0
normal	1	5	9	0	3	0	0	0	1	2	1	0	0	0	0	4
ovary	0	0	0	0	0	0	0	0	0	1	9	0	0	0	0	0
pancreas	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
prostate	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
stomach	0	0	0	0	0	0	1	0	0	1	0	0	0	2	0	0
thyroid	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
uterus	0	0	1	0	1	0	0	0	0	2	3	0	0	1	0	1

Discussion

The difference in CNV pattern between different cancer origins has been previously reported (Baudis, 2007) and there are reports of building classifiers on CNV data (Liu, Ranka, & Kahveci, 2008). However, the CNV data has not been previously proposed to build a classifier to predict the primary cite of CUP patients. Here, we report a classifier that has been trained on the CNV data profile from TCGA project (~4000 tumor samples). The classifier is able to make high confidence prediction with 95% accuracy for 72% of primary tumor samples. The robustness of the classifier is shown by its ability to predict the origin of cell line cancers, which represent distinct biological entity and higher rate of somatic CNVs, which are acquired during extended growth in the cell culture.

The DNA copy number variation (CNV) data offers several advantages that can aid the prediction of cancer primary site. First, DNA is more stable than RNA, which makes the predictor less vulnerable to improper handling of cancer samples. Secondly, cancer samples obtained by surgery or biopsy are often contaminated with one or a mixture of surrounding tissues, which can skew the RNA based methods. In contrast, virtually all cells within a body share the same DNA. Cancers on the other hand, display somatic copy number alteration (SCNA) that have been accumulated during carcinogenesis. Therefore, DNA of surrounding tissue is a "blank" background. The signal from the cancer cells' DNA can be diluted but not confounded by DNA from normal tissues. If the biopsy was "missed" and no cancer tissue was present in the tissue, the sample will be recognized as "normal" tissue, due to lack of accumulated somatic CNVs.
Secondly, no current method is able to correctly classify all CUP samples. The classifiers based on mRNA and microRNA can only predict the origins covers in their training set and the accuracy differs significantly for different origins. The higher number of cases in a cancer class, the easier the training of the classifier and the higher chances of good accuracy of specific class. The overall accuracy is commonly reported and the accurately predicted cancer classes that are represented by hundreds of samples can overshadow the poor accuracy of smaller classes, where not enough data were available for successful training. It is unknown if CUPs originate mostly from the most common cancers, or if they originate from more rare and obscure cancers. Therefore the accuracy of prediction for CUP patients can be much lower than the reported overall accuracy. Therefore, the composition of the training and test set should be closely considered rather than to rely solemnly on the overall performance.

Additionally, the prediction of the primary site of CUP, should be considered in the bigger picture. A single method may not be optimal for all CUPs. As the results here show, CNVs can be successfully used to predict the cancer that does show significant changes in CNV. However, some cancers are not affected and are classified as normal. Those cases with "flat" CNV profile are likely driven by other factors and therefore other kind of data, which represent different layer of biology, are needed to predict the origin. Similar effect can be possible for messengerRNA or microRNAs. Therefore, the combination of various molecular assays may offer the performance that cannot be matched by any single method. The RNA and DNA can be easily and simultaneously extracted from the same cancer samples. Therefore, the CNV based classifier provides

the opportunity to aid the classification based on RNA expression profiles, if applied together. The abundance of already available CNV profiles data enables timely and cost effective development of clinically relevant classifier.

Perspectives

The project opens the way to a development of more comprehensive classifier that will be able to predict the primary site of CUPs. To achieve this goal, a bigger training set is required, which will cover more histological types of cancer. Additionally, a validation set of metastatic cancers is necessary to prove that the classier can make clinically valid predictions for metastatic samples. A set of samples from CUP patients is also desirable. This will allow investigating the behavior of the classifier when presented with true CUPs' CNV profiles. Finally, the CNA based classifier can be combined with messengerRNA or miRNA based classifier to obtain optimal predictions.

References

Baudis, M. (2007). Genomic imbalances in 5918 malignant epithelial tumors: an explorative meta-analysis of chromosomal CGH data. *BMC cancer*, *7*, 226. doi:10.1186/1471-2407-7-226

Beroukhim, R., Mermel, C. H., Porter, D., Wei, G., Raychaudhuri, S., Donovan, J., Barretina, J., et al. (2010). The landscape of somatic copy-number alteration across human cancers. *Nature*, *463*(7283), 899–905. Nature Publishing Group. doi:10.1038/nature08822

Greco, F. (2010). Evolving understanding and current management of patients with cancer of unknown primary site. *Community Oncology*.

Liu, J., Ranka, S., & Kahveci, T. (2008). Classification and feature selection algorithms for multi-class CGH data. *Bioinformatics*, *24*(13), i86–95. doi:10.1093/bioinformatics/btn145

Ma, X.-J., Patel, R., Wang, X., Salunga, R., Murage, J., Desai, R., Tuggle, J. T., et al. (2006). Molecular classification of human cancers using a 92-gene real-time quantitative polymerase chain reaction assay. *Archives of pathology & laboratory medicine*, *130*(4), 465–473. doi:10.1043/1543-2165(2006)130[465:MCOHCU]2.0.CO;2

Monzon, F. A., Lyons-Weiler, M., Buturovic, L. J., Rigl, C. T., Henner, W. D., Sciulli, C., Dumur, C. I., et al. (2009). Multicenter validation of a 1,550-gene expression profile for identification of tumor tissue of origin. *Journal of Clinical Oncology*, *27*(15), 2503–2508. doi:10.1200/JCO.2008.17.9762

Olshen, A., Venkatraman, E., & Lucito, R. (2004). Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics*.

Pavlidis, N., & Pentheroudakis, G. (2010). Cancer of unknown primary site: 20 questions to be answered. *Annals of oncology : official journal of the European Society for Medical Oncology / ESMO*, *21 Suppl 7*, vii303–7. doi:10.1093/annonc/mdq278

Pavlidis, N., & Pentheroudakis, G. (2012). Cancer of unknown primary site. *Lancet*. doi:10.1016/S0140-6736(11)61178-1

Rosenfeld, N., Aharonov, R., Meiri, E., Rosenwald, S., Spector, Y., Zepeniuk, M., Benjamin, H., et al. (2008). MicroRNAs accurately identify cancer tissue origin. *Nature biotechnology*, *26*(4), 462–469. doi:10.1038/nbt1392

Stankiewicz, P. (2010). Structural variation in the human genome and its role in disease. *Annual review of medicine*.