Genetic alterations in cervical cancer

Consequences for gene regulation and clinical outcome after chemoradiotherapy

by

Malin Lando

Department of Radiation Biology

Institute for Cancer Research – Norwegian Radium Hospital

Oslo University Hospital

Oslo, 2011

© Malin Lando, 2011

Series of dissertations submitted to the
Faculty of Medicine, University of Oslo
No. 1150

ISBN 978-82-8264-031-2

All rights reserved. No part of this publication may be
reproduced or transmitted, in any form or by any means, without permission.

Cover: Inger Sandved Anfinsen.
Printed in Norway: AIT Oslo AS.

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

Acknowledgements

The work presented in this thesis has been carried out at the Department of Radiation Biology, Institute for Cancer Research, Oslo University Hospital, from October 2007 to February 2011. The work was financed by Helse Sør-Øst, which is gratefully acknowledged.

I would like to express my greatest gratitude to my supervisor, Dr. Heidi Lyng, for the supervision and support, for always taking time to discuss my work and for always being enthusiastically involved in my project. I highly appreciate your extremely positive attitude.

I'm also thankful to the Department of Gynecologic Oncology, especially the PI of the clinical protocol Dr. Gunnar B. Kristensen, for providing clinical data, collecting clinical samples, and for useful discussions.

I would also like to thank Dr. Trond Stokke for sharing your broad knowledge, for giving valuable advice and suggestions, and last but not least for hosting annual boat-trips with the group in the Oslofjord.

Special thanks goes to the rest of my group, especially Kirsti, Debbie, Cathinka, Idun, and Kristin who I've been fortunate to share office with during the last years, and also to Sebastian for sharing your broad lab expertise. You have all participated in creating a fun and social working environment, making me look forward to go to work every day.

Last but not least I would like to thank my husband, Odd Martin Staal, for your support and for listening to all my presentations up to several times.

Oslo, February 2011

Malin Lando

Table of contents

ABBREVIATIONS ..... **VI**

AIMS OF THE STUDY ..... **VII**

LIST OF PAPERS ..... **VIII**

INTRODUCTION **1**

CERVICAL CANCER 1
Epidemiology *1*
Risk factors *1*
Histological subtypes *1*
Staging *2*
Treatment and prognosis *3*
Biomarkers *3*
GENERAL CHARACTERISTICS OF CANCER 3
Cancer development *4*
Genetic and epigenetic alterations *5*
Genetic instability *7*
Cancer genes *8*
CARCINOGENESIS IN CERVICAL CANCER 10
The multistep process *10*
HPV infection *11*
Genetic alterations in cervical cancer *13*
BIOLOGICAL CONSEQUENCES OF GENETIC ALTERATIONS 14
CLINICAL CONSEQUENCES OF GENETIC ALTERATIONS 17
MICROARRAY TECHNOLOGY 18
Principle *19*
Platforms *19*
Array comparative genomic hybridization *20*

SUMMARY OF PAPERS **22**

Paper I *22*
Paper II *22*
Paper III *23*

EXPERIMENTAL CONSIDERATIONS **24**

PATIENTS 24
MICROARRAY TECHNIQUES 25
aCGH *25*
Gene expression arrays *25*
Downstream analysis of microarray data *25*

DISCUSSION **27**

TOOL FOR DNA COPY NUMBER DETERMINATION IN CERVICAL CANCERS 27
DNA COPY NUMBER ALTERATIONS IN CARCINOGENESIS 28
DNA COPY NUMBER ALTERATIONS AND CHEMORADIORESISTANCE 29

FUTURE PERSPECTIVES **30**

IMPLEMENTATION IN THE CLINIC 30
Diagnosis *30*
Therapy *31*

Abbreviations

AC	Adenocarcinoma
aCGH	Array comparative genomic hybridization
ASC	Adenosquamous carcinoma
BAC	Bacterial artificial chromosome
bp	base pair
cDNA	Complementary DNA
CFS	Common fragile site
CGH	Comparative genomic hybridization
CIN	Cervical intraepithelial neoplasia
DNA	Deoxyribonucleic acid
eGOn	Explore gene ontology
FIGO	International Federation of Gynecology and Obstetrics
FISH	Fluorescence *in situ* hybridization
GO	Gene ontology
HPV	Human papilloma virus
HR-HPV	High-risk HPV
Kb	kilo bases
Mb	mega bases
miRNA	MicroRNA
PAC	P1-derived artificial chromosome
Pap smear	Papaniculaou smear
PCR	Polymerase chain reaction
RNA	Ribonucleic acid
SAM-GS	Significant Analysis of Microarrays for Gene Sets
SCC	Squamous cell carcinoma
SNP	Single nucleotide polymorphism

Aims of the study

The majority of patients with locally advanced cervical cancer are treated with chemoradiotherapy. Approximately 55% of these patients show progressive disease within 5 years after diagnosis, and side effects in organs within the pelvis after radiation treatment are frequent. It is therefore a lot to gain on finding biological characteristics of the disease that can be used to predict the outcome more accurately than the clinical information currently used. Chromosomal alterations, like gains and losses, are frequent in cervical cancer. Such DNA copy number alterations may deregulate the expression of affected genes and change cell behavior, however, their biological significance in cervical cancers have not been clarified.

The overall aim of this work was to explore the potential of DNA copy number alterations as biomarkers for identification of patients at risk for failure or for development of targets for therapeutic intervention. The preferred technique for detecting copy number alterations is array comparative genomic hybridization (aCGH). This technique provides relative copy numbers, which are influenced by tumor ploidy and the normal cell content of the clinical specimen, leading to unreliable results when data are compared across patients. An improvement of the current analysis methods was therefore required to handle these problems. The specific aims of the work were to:

- Develop a tool for genome-wide calculation of absolute DNA copy numbers from clinical aCGH data that enabled comparisons across patients despite differences in tumor ploidy and normal cell content of the clinical specimens.
- Explore the importance of DNA copy number alterations for the carcinogenesis and chemoradioresistance of cervical cancers by aCGH analysis.
- Investigate the functional role of selected DNA copy number alterations by integrating the data with gene expression and ontology information of the same patients.
- Evaluate the usefulness of DNA copy number alterations and their target genes as biomarkers for the clinical outcome.

List of papers

I Lyng H, **Lando M**, Brøvig RS, Svendsrud DH, Johansen M, Galteland E., Brustugun OT, Meza-Zepeda LA, Myklebost O, Kristensen GB, Hovig E, Stokke T (2008). GeneCount: genome-wide calculation of absolute tumor DNA copy numbers from array comparative genomic hybridization data. Genome Biol 9(5):R86.

II **Lando M**, Holden M, Bergersen LC, Svendsrud DH, Stokke T, Sundfør K, Glad IK, Kristensen GB, Lyng H (2009). Gene dosage, expression and ontology analysis identifies driver genes in the carcinogenesis and chemoradioresistance of cervical cancer. PLOS Genetics, 5(11):e1000719.

III **Lando M**, Snipstad K, Clancy T, Halle C, Holden M, Stokke T, Sundfør K, Holm R, Kristensen GB, Lyng H. Microarray based analysis of gene expression and pathways associated with loss on chromosome 3p in cervical cancer. *Manuscript*.

Introduction

Cervical cancer

Epidemiology

Cervical cancer is the second most common cause of cancer-related death among women globally, and the number-one cause of cancer-related death in women in developing countries (Lehoux, D'Abramo, and Archambault 2009). From 2004-2008, 284 new cases were diagnosed in Norway each year (Cancer Registry of Norway 2010). International Agency for Research on Cancer (IARC) estimated 493,000 new cases in 2002 worldwide, and the number of deaths was 274,000 (Parkin *et al.* 2005). The peak incidence of the disease occurs in women over 40 years of age. The introduction of the Pap smear has reduced the overall death rate by approximately 74% since the tests were first implemented (Cervical Cancer Statistics 2007). Therefore 80% of the incidences are occurring in developing countries, where cervical cancer prevention programs are not available.

Risk factors

The primary risk factor for cervical cancer is the Human Papilloma Virus (HPV). There are several types of HPV, which vary in their ability to transform the cervical epithelium (Cannistra and Niloff 1996). Low-risk varieties, such as HPV 6 and 11 are commonly associated with CIN I, which do not usually progress to invasive disease. The high-risk types, such as HPV 16, 18, 31, 33, and 35 are often associated with moderate dysplasia (CIN II) and severe dysplasia (CIN III) or carcinoma in situ and are observed in more than 99% of patients with invasive cervical cancer (Lehoux *et al.* 2009). The highest incidence of HPV infection is between the age of 20 and 30, suggesting a long latency period between infection and cancer appearance (Lazo 1999). Other risk factors have been identified, including poverty, high parity, sexual intercourse at an early age, multiple sexual partners, smoking, eating habits, and weakened immune system (Hulka 1982;Reeves, Rawls, and Brinton 1989).

Histological subtypes

The three major histological types of invasive cervical cancer are squamous cell carcinoma (SCC), adenocarcinoma (AC), and adenosquamous carcinoma (ASC), where SCC comprises 80-85% of the cases and AC and ASC together comprise approximately 15% of

the cases (Baalbergen *et al.* 2010). SCC arises from the multilayered squamous epithelium of the ectocervix and transformation zone, whereas AC arises from the glandular epithelium of the endocervix (Szalmas and Konya 2009).

Staging

Staging of invasive cervical cancer is usually described in terms of the FIGO (International Federation of Gynecology and Obstetrics) system based on clinical criteria like size, depth of penetration within the cervix, and spread within and beyond the cervix (Fig. 1) (Koyama, Tamai, and Togashi 2007). Stage I tumors are limited to the cervix, and are divided into stage 1A, $1B_1$, and $1B_2$ depending on the penetration depth. Stage II tumors extend beyond the cervix to the upper two thirds of the vagina (IIA) or the parametrial tissue (IIB), but not to the pelvic side wall. Stage III tumors have spread to the lower third of the vagina (IIIA) or to the pelvic side wall (IIIB). Stage IV tumors have invaded the mucosa of the bladder or rectum (IVA) or have spread to distant sites outside the pelvis (IVB).

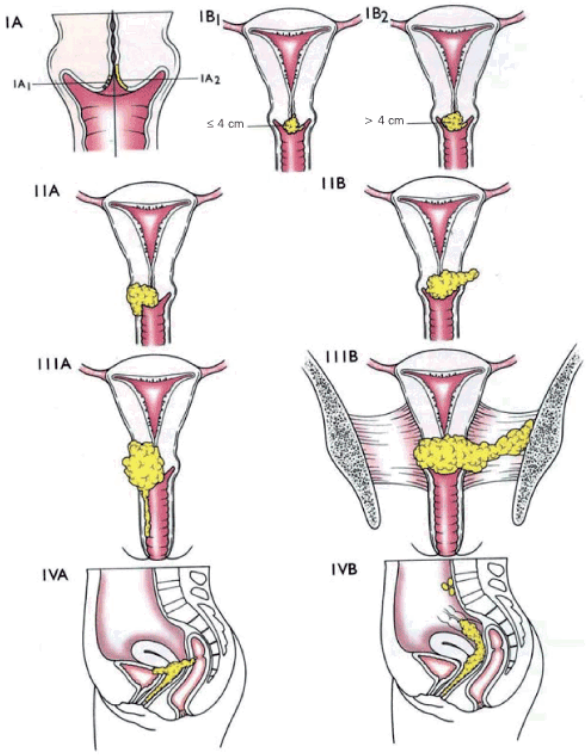


Fig. 1: Staging of cervical cancer according to FIGO. Adapted from (Camisπo *et al.* 2007).

Treatment and prognosis

The choice of treatment for invasive cervical cancer depends primarily on tumor stage, although factors like the location of the tumor within the cervix, histology, patient age, general health, childbearing plans, and pregnancy also affect the therapeutic options (Health MD: How is Cervical Cancer Treated? 2010). At stage IA, hysterectomy or conization is performed. The five-year survival rate for these patients exceeds 95%. Patients with stage IB or early stage IIA receive either surgery or chemoradiotherapy, which produces equivalent results with a five-year survival rate of 80%. At the locally advanced stages IIB, III, and IVA, chemoradiotherapy is the treatment of choice. Both external irradiation and brachytherapy are given, combined with adjuvant cisplatin-based chemotherapy, which has shown to increase the effect of the radiotherapy (Eifel 2006). The five-year survival rates are 65, 40, and less than 20% for stage IIB, III, and IVA. Patients with stage IVB may also benefit from local radiotherapy, which is combined with carboplatin and 5-FU-based chemotherapy. At this stage the disease is generally not considered to be curable.

Biomarkers

Standard diagnostic procedures for cervical tumors are currently based on clinical data, which, in most cases, provide precise information on tissue origin, tumor type, stage, grade, and the possibility for complete surgical tumor removal (Dietel and Sers 2006). Up to now, these data are the rational basis for therapy design. However, they do not fully reflect the aggressiveness of the disease and two patients harboring the apparently “same” type of tumor in a seemingly “identical” stage often show different therapeutic outcome. It is therefore a considerable need for biomarkers that can be used in addition to the standard clinical information, to predict the therapeutic outcome and thereby aid the clinical decision-making (Dietel and Sers 2006). Biomarkers for design of targeted therapeutics, directed towards the treatment resistant cells and/or interfering with key signaling pathways in the tumor response, are also required, in particular for the patients receiving chemoradiotherapy. Development of novel radiosensitizing agents can enable reduction in the radiation dose and thereby side effects while retaining the tumor control probability.

General characteristics of cancer

Cancer is characterized by a group of cells displaying uncontrolled growth, invasion and destruction of adjacent tissues, and sometimes spread to other locations in the body

(metastasis) (Alberts *et al.* 2002). These are malignant features and distinguish cancers from benign tumors, which are self-limited and do not invade or metastasize. Cancers are classified according to the tissue and cell type from which they arise. The main types are carcinoma (epithelial tissue), glioma (brain tissue), sarcoma (soft tissue and bone), and leukemia/lymphoma (haematopoietic tissue).

Cancer development

Cancer is, in essence, a genetic disease, and alterations in oncogenes, tumor suppressor genes, and stability genes are responsible for the initiation and progression of the disease in a process called carcinogenesis (Vogelstein and Kinzler 2004). Both genetic and epigenetic alterations are involved, leading to changes in the gene expression and thereby cell behavior.

Carcinogenesis is a multi-step process, typically a progression from benign lesions to malignant tumors (Fig. 2) (Stewart and Kleihues P. 2003). All living cells will probabilistically suffer from errors in the genome. Normally the body is protected against their potentially lethal effects. However, the cellular correction machinery may fail, especially in environments that make errors more likely to arise and propagate. The progressive errors then slowly accumulate until the cell begins to act contrary to its function in the organism and a malignant lesion is formed.

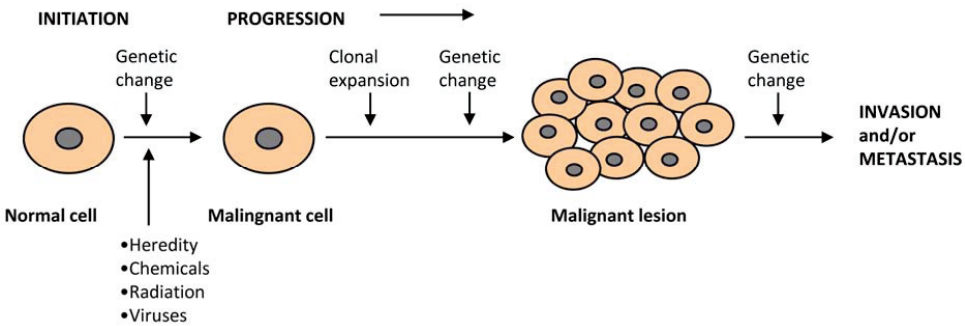


Fig. 2: Tumor initiation and progression

Malignant tumors are generally believed to be monoclonal, meaning that each tumor arises from a single cell (Nowell 1976). This cell will acquire sequential accumulation of genetic and epigenetic alterations in genes responsible for the control of cellular proliferation, cell

death and the maintenance of genetic integrity. Selective clonal expansion will then lead to progression towards cancerous growth.

Tumor progression is not only a process of alterations in cancer cells, but also the environment contributes substantially to the malignancy. This was hypothesized already in 1889 by Paget in his "seed and soil" hypothesis, suggesting that the cancer cells (seeds) need specific organ microenvironments (soils) in order to proliferate (Paget 1989). Tumor progression is the product of an evolving crosstalk between different cell types within the tumor and its surrounding supporting tissue, called the tumor stroma. Cancer cells can alter their stromal environment by producing growth factors and proteases to form a permissive and supportive environment for tumor progression (Mueller and Fusenig 2004). Such alterations involve induction of angiogenesis, degradation and remodeling of the extracellular matrix, inflammation and production of a range of factors that promote cell proliferation.

Genetic and epigenetic alterations

Genetic alterations include subtle sequence changes, amplifications, deletions, and translocations (Fig. 3). The genetic alterations may either be random, leading to no growth advantage, or important for the malignant progression, driving the tumor towards a more aggressive state (Ried *et al.* 1999). Subtle sequence changes involve base substitutions or deletions or insertions of nucleotides. The simplest alteration is a point mutation, where a single nucleotide is exchanged for another. If mutations in key regulatory genes are not repaired, the cells will alter their behavior and cancer may arise. Humans with defects in repair mechanisms, such as nucleotide-excision repair or mismatch repair, are shown to be more susceptible to certain types of cancers (Lengauer, Kinzler, and Vogelstein 1998).

Amplification is a copy number increase of a restricted region of a chromosome arm, or of the whole chromosome arm, and is often associated with overexpression of some of the affected genes. Amplifications can be divided into small chromosomal segments that may be present at high level, usually denoted high-level amplification or just amplification, and larger chromosomal regions generally present at lower levels, called gains. Deletion is a copy number loss of a chromosome arm or of a region of the chromosome arm. It may involve a loss on one of the chromosomes in a pair, called a heterozygous deletion, or loss of the same region on both chromosomes, called a homozygous deletion.

A consequence of severe changes in the copy number of chromosomes is aneuploidy, which is caused by mitotic abnormalities (Munger 2002). Aneuploidy occurs during cell division in cases where the chromosomes do not separate properly between the two cells. Cancer cells are frequently aneuploid and can contain either less or more than two copies of the chromosomes. The importance of gene copy number alterations and aneuploidy for the carcinogenic process is not clear and probably varies substantially between tumors. Due to differences in the mechanisms by which tumors are initiated, either in the individual genotype or in the particular cell type in which the tumor arises, some tumors accumulate a number of alterations, whereas others may evolve by mechanisms that result in little chromosomal changes (Albertson *et al.* 2003).

Translocations are caused by rearrangement of parts between nonhomologous chromosomes. They can involve two or more chromosomes and are either balanced, involving an even exchange of material, or unbalanced, where the exchange of chromosomal material is unequal, resulting in extra or missing chromosomal regions (Lengauer *et al.* 1998). Chromosomal translocations can result in the generation of new gene products, either by the fusion of genes or by bringing genes close to enhancer or promoter elements, hence leading to their altered expression (Nambiar, Kari, and Raghavan 2008). Translocations are considered the primary cause of many cancers, including lymphoma, leukemia, and some solid tumors.

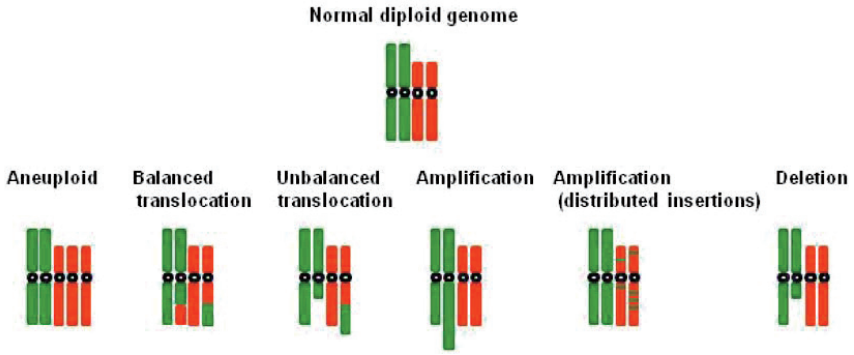


Fig. 3: Genetic alterations found in cancer cells

Epigenetic alterations are changes in gene expression caused by mechanisms other than changes in the DNA sequence itself. A key mechanism in epigenetic regulation of gene

expression is DNA methylation (Wajed, Laird, and DeMeester 2001). This is an enzyme-induced chemical modification of the DNA structure where a methyl group is added to the DNA, mostly at CpG sites, to convert cytosine to 5-methylcytosine. Epigenetic alteration also involves post translational modification of histone proteins (Herceg and Hainaut 2007). If the amino acids that make up the histones are changed, this can lead to a modified shape of the histone sphere and an incomplete unwinding of the DNA during replication.

Genetic instability

Genetic instability is the rate of alterations observed in a cell population at a certain time point, and is seen primarily at the chromosomal level (Lengauer *et al.* 1998). Due to genetic instability, subpopulations with different genetic alterations may emerge successively and coexist within tumors, reflecting clonal divergence (Fig. 4) (Georgiades *et al.* 1999). Assessment of this intratumor heterogeneity provides important clues of the genetic steps involved in carcinogenesis. Since the heterogeneous alterations have emerged after the homogeneous ones, the chronological order of the alterations indicates how the tumor has evolved (Lyng *et al.* 2004). Aneuploidy may be a consequence of high genomic instability, although it may also arise from a single genetic event.

The intratumor heterogeneity in genetic alterations is a major challenge in the treatment of cancer since different subpopulations may have different capacities for growth, differentiation and metastasis formation, as well as sensitivity to radiation and chemotherapeutic agents (Barranco *et al.* 1994;Vogelstein and Kinzler 1993). Treatment results in selective killing of sensitive populations, whereas the resistant ones survive (Barranco *et al.* 1988). Continued treatment may lead to total resistance, not only to the initial agent or combination, but to other agents as well (Rothenberg and Ling 1989). Knowledge of the genetic heterogeneity within tumors may therefore be crucial for the handling of the disease.

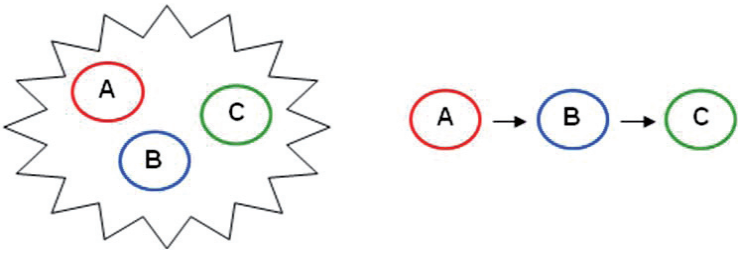


Fig. 4: Schematic diagram showing the evolution of the subpopulations A, B, and C with different genetic alterations. Subpopulation B has evolved from subpopulation A, and subpopulation C has evolved from subpopulation B.

Cancer genes

In the last decades, many genes involved in carcinogenesis of various cancers have been discovered. The regulation of these genes facilitates signal transduction, mediates cell division, differentiation or cell death, and maintains the integrity of genetic information by DNA repair and similar processes (Stewart and Kleihues P. 2003). The genes are believed to be targets for the genetic alterations and fall into three major categories; oncogenes, tumor suppressor genes, and stability genes. In addition, microRNA (miRNA) is a recently discovered class of small non-coding RNAs that regulate gene expression post-transcriptionally.

Oncogenes are genes that, when mutated or expressed at high levels, are capable of transforming a normal cell into a tumor cell and therefore contribute to the formation of cancer (Vogelstein and Kinzler 2004). The normal gene counterparts are referred to as proto-oncogenes. An activating somatic mutation in one allele of an oncogene is generally sufficient to confer a selective growth advantage on the cell, and the activation of an oncogene is therefore dominant. Oncogene activations can result from chromosomal translocations, gene amplifications, or from subtle intragenic mutations affecting crucial residues that regulate the activity of the gene product. The *RAS* family of oncogenes was among the first that was recognized as being mutated in a variety of human cancers, whereas the *MYC* oncogene is frequently overexpressed by gene amplification (Stewart and Kleihues P. 2003).

Tumor suppressor genes protect the cell from becoming a tumor cell. Their alteration during carcinogenesis results in the loss of functional property essential for the maintenance of normal cell proliferation (Vogelstein and Kinzler 2004). The loss of function of a tumor suppressor gene is a recessive mechanism, implying that both alleles must be affected to switch it off. Inactivation of tumor suppressor genes can arise from missense mutations, deletions, insertions, or from epigenetic silencing. *TP53* is the most commonly altered tumor suppressor gene in human cancers, being altered in over 50% of most tumor types (Stewart and Kleihues P. 2003).

Stability genes keep genetic alterations to a minimum, and when they are inactivated, mutations in other genes occur at a higher rate. Stability genes are involved in DNA repair and include mismatch repair genes (*MLH1*, *MSH2, PMS2*), nucleotide-excision repair genes (*XPA*, *DDB2*), and base-excision repair genes (*MUTYH*), which are responsible for repairing mistakes occurring during normal DNA replication or induced by mutagenic exposure (Vogelstein and Kinzler 2004). Stability genes are also involved in control processes, such as mitotic recombination and chromosomal segregation (*ATM*, *BRCA1*, *BLM, RECQL4*). As with tumor-suppressors, both alleles need to be inactivated to have full effect.

MicroRNAs are evolutionarily conserved, small, single-stranded molecules that do not encode proteins, but instead suppress the expression of protein-coding genes by blocking translation, degrading mRNA, or both (Anglicheau, Muthukumar, and Suthanthiran 2010). More than 700 miRNAs have so far been identified in the human genome. A single miRNA can regulate the expression of hundreds of mRNAs or proteins, and each mRNA can be targeted by multiple miRNAs that can interact with each other, either by synergism or competition. It is believed that most protein-coding genes are regulated by miRNAs, and it has been shown that miRNAs play an important role in diverse biological processes like development, differentiation, cell proliferation, apoptosis, and oncogenesis (Valencia-Sanchez *et al.* 2006). Up- or down-regulation can be due to amplification, deletion, regulation of a transcription factor, or methylation/demethylation. miRNA genes can function either as oncogenes or tumor suppressor genes depending on their targets in a specific tissue (Croce 2008), and many miRNAs have been shown to locate to chromosomal regions that undergo rearrangements, deletions, and amplifications in cancer cells (Calin *et al.* 2004). Dysregulated expression of miRNAs has already been associated with several

cancers. Of interest, *miR-127* has been shown to be overexpressed in cervical cancer patients with lymph node metastasis (Lee *et al.* 2008). Moreover, anti-*miR-199a* has been shown to inhibit cervical cancer cell growth, suggesting *miR-199a* as a potential therapeutic target for cervical cancer therapy (Lee *et al.* 2008).

Carcinogenesis in cervical cancer

The multistep process

Cervical cancers progress in a multistep process, from preinvasive cervical intraepithelial neoplasia (CIN) to invasive stages (Grubisic *et al.* 2009). The majority of cervical cancers originate from the squamous epithelium near the opening of the cervix (Alberts *et al.* 2002), where the proliferation normally occurs only in the basal layer (Fig. 5A). The newly generated cells move towards the surface, differentiate, and form flattened, keratin-rich, non-dividing cells that are sloughed off as they reach the surface. It is not unusual to find patches in which this organization is disturbed in a way that suggests the beginning of a cancerous transformation, called cervical intraepithelial neoplasia (CIN). These changes can be low-grade (CINI), moderate (CINII), or severe (CINIII) (Szalmas and Konya 2009).

In the low-grade lesions, dividing cells are no longer confined to the basal layer, but occupy the lower third of the epithelium (Fig. 5B). Most of these lesions will spontaneously regress; however, about 10% might progress to become high-grade lesions. Most of the epithelial layers will then be occupied by undifferentiated dividing cells, which are usually highly variable in cell size and shape (Fig. 5C). If remained untreated, the abnormal tissue may persist and stop progressing or it may regress spontaneously. However, in almost half of the cases, progression will occur, giving rise to an invasive carcinoma where cells cross or destroy the basal lamina, invade the underlying tissue, and metastasize (Fig. 5D).

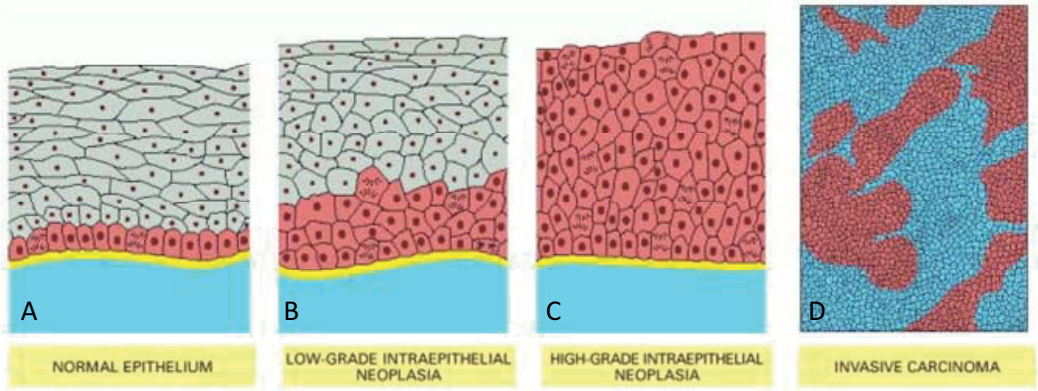


Fig. 5: The stages of progression in the development of squamous cell cervical carcinoma. Adapted from (Alberts *et al.* 2002).

HPV infection

Infection with high-risk HPV (HR-HPV) is a necessary, but insufficient, contribution to cervical cancer, with additional viral and host genetic events required to drive cells to the malignant phenotype (Pett and Coleman 2007). HR-HPV is detected in 99.7% of the cases and also in the vast majority of high-grade neoplasia (Lehoux *et al.* 2009). HPV 16 and 18 are the most prevalent types, accounting for about 70% of all cases. In the normal viral life cycle, HPV genomes exist in a circular or episomal state and are thought to be retained in basal cells of the squamous epithelium at approximately 50-100 copies per cell (Fig. 6) (Bedell *et al.* 1991;Stanley *et al.* 1989). In the episomal form, the expression of the viral oncogenes E6 and E7 is tightly regulated. However, in most cervical cancers (90%), truncated viral genomes are integrated into host DNA, resulting in deregulation of viral oncogene expression (Pett and Coleman 2007). Integration is not a part of the HR-HPV life cycle and represents a by-product of viral infection, leading to increased expression and stability of transcripts encoding the E6 and E7 proteins. This effect is restricted to HR-HPV. Integrated HPV may be viewed as selectable because it represents a form of the virus that is resistant to host mechanisms of viral clearance, enabling infected cells to maintain viral oncogene expression and avoid cell death.

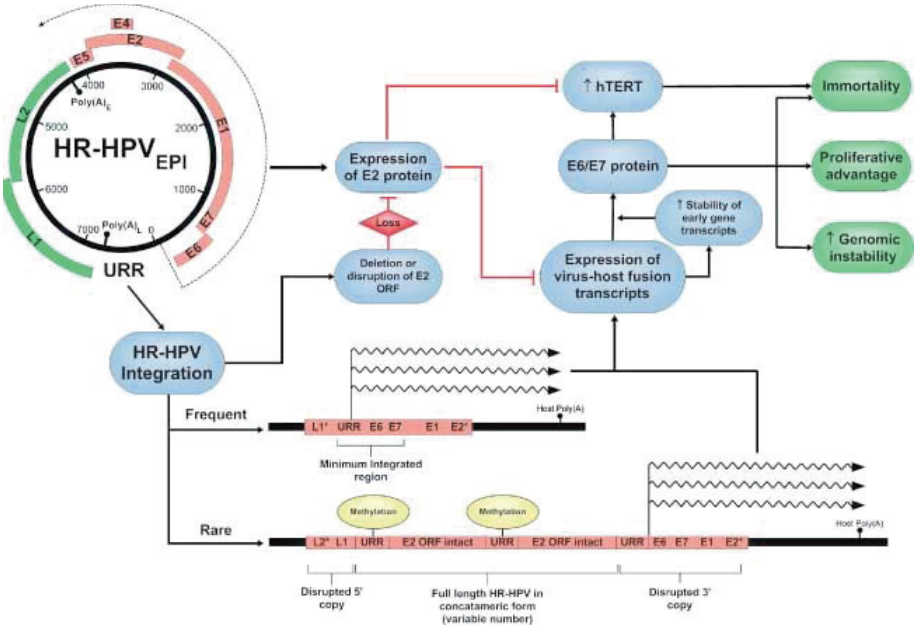


Fig. 6: Significance of HR-HPV integration events detected in cervical carcinomas. Adapted from (Pett and Coleman 2007).

The viral proteins E6 and E7 produced by HR-HPV types are critical for malignant transformation, due to their ability to bind and inactivate the tumor suppressor proteins TP53 and RB1 (Cannistra and Niloff 1996). E6 inactivates the function of wild-type TP53 by enhancing its degradation, resulting in prevention of apoptosis. It can also interfere with cellular apoptotic pathways by a TP53-independent mechanism by binding to the tumor necrosis factor receptor TNFRSF1A and FAS-associated death domain protein FADD, leading to the degradation of FADD and the caspase CASP8 (Garnett and Duerksen-Hughes 2006). E7 binds to RB1, relieves its repression of the E2F transcription factors, and causes the cell to reenter S-phase.

The mechanism by which HPV integrates is not fully understood, but there is a clear predilection for integration at chromosomal common fragile sites (CFSs) (Thorland *et al.* 2003;Thorland *et al.* 2000;Wentzensen, Vinokurova, and von Knebel 2004;Yu *et al.* 2005), which are DNA regions of profound genomic instability that span from several hundred kilobases to greater than 9 Mb (Richards 2001). The most frequently observed integration site, particularly in cervical cancers positive for HPV 18, is the region of the *MYC* oncogene at chromosomal band 8q24 (Durst *et al.* 1987;Peter *et al.* 2006;Wentzensen *et al.*

2002;Ferber *et al.* 2003b). Another active CFS is the *FRA3B* locus at 3p14.2, which is frequently altered in cervical cancers. Recurrent integration has also been observed at the *TERT* (Ferber *et al.* 2003a) and *FANCC* (Wentzensen *et al.* 2002;Ferber *et al.* 2003b) loci, at chromosomal bands 5p15 and 9q22, respectively.

HR-HPV infection also leads to genomic instability, which is reflected in the high proportion of aneuploid cervical tumors. The HPV E6 and E7 proteins disturb the centrosome replication during cell division, leading to increased centrosome numbers, aberrant mitotic spindle pole formation, and aneuploidy (Duensing *et al.* 2000). In fact, abnormal, tripolar mitoses have been described as diagnostic markers of HR-HPV positive lesions. E6 expressing cells show mostly gene amplifications, abnormalities consistent with TP53 loss, whereas E7 expressing cells often displays gains and losses of entire chromosomes and are aneuploid (White, Livanos, and Tlsty 1994). Hence, in addition to containing compromised surveillance functions, HR-HPV E6 and E7 expressing cells constantly produce mitotic abnormalities even when expressed from viral episomes, which greatly increases genomic instability and the probability of malignant progression (Duensing *et al.* 2001).

Genetic alterations in cervical cancer

Genome wide screening of chromosomal alterations in cervical cancer has revealed genetic changes on all chromosomes, and identified several chromosomal regions that are recurrently affected. The most frequent alterations are gain of 1q, 3q, 5p, and 20q and loss of 2q, 3p, 4p, 4q, 11q, and 13q (Heselmeyer *et al.* 1997;Lyng *et al.* 2004;Rao *et al.* 2004;Ried *et al.* 1999;Umayahara *et al.* 2002). The gain of 3q and loss of 3p are especially frequent, where the former has been found to define the transition from preinvasive to invasive cervical carcinoma (Heselmeyer *et al.* 1996;Umayahara *et al.* 2002). The gain of 20q has been described by a number of studies (Dellas *et al.* 1999;Kirchhoff *et al.* 1999;Narayan *et al.* 2003;Umayahara *et al.* 2002) and has in other solid tumors, such as breast, colon, and gastric carcinomas been associated with a selective growth advantage and a more aggressive metastatic phenotype (Hermsen *et al.* 2002;Hodgson *et al.* 2003;Kallioniemi *et al.* 1994;Muleris *et al.* 1994;Weiss *et al.* 2003).

Cervical cancers also show substantial intratumor genetic heterogeneity due to the high chromosomal instability (Ried *et al.* 1999). Loss of 3p and gain of 3q are seldom

heterogeneous and are therefore believed to play a role in early development of the disease (Guo *et al.* 2001;Lyng *et al.* 2004). On the other hand, loss of chromosome 4 and 6q, and gain of 2p have been shown to be both frequent and heterogeneous, indicating that they lead to a significant growth advantage and that they are important for tumor progression at a later stage (Lyng *et al.* 2004).

Although several genetic alterations have been suggested to play a role in the carcinogenesis of cervical cancer, most of the target genes are unknown and the biological meaning of the alterations is poorly understood. The low resolution of the CGH-technique used in many of the studies, have led to difficulties in limiting the gained and lost regions. This has made it difficult to distinguish between the genes playing a functional role during carcinogenesis and those that are passengers. Moreover, the number of samples included in the studies has often been limited, reducing the statistical reliability of the analyses and therefore the ability to discover genetic alterations important for the clinical outcome. Application of the aCGH technique on a large sample set of tumors may therefore lead to novel insight into the genetic alterations of cervical cancers and their potential as biomarkers.

Biological consequences of genetic alterations

Changes in the cancer genome are believed to manifest essential alterations, known as cancer hallmarks, associated with tumor cell phenotype and physiology that collectively promote malignant growth. These cancer hallmarks are self-sufficiency in growth signals, insensitivity to antigrowth signals, evasion of apoptosis, limitless replicative potential, sustained angiogenesis, and tissue invasion and metastasis (Hanahan and Weinberg 2000). Also metabolism and endocytosis are considered hallmarks of physiological changes in cancer cells. The genomic instability is a way of evolving populations of pre-malignant and malignant cells to reach these biological endpoints in the process of carcinogenesis.

Self-sufficiency in growth signals implies that the tumor cells generate many of their own growth signals by altering extracellular growth signals, transcellular transducers of those signals, or intracellular circuits that translate the signals into action (Mees, Nemunaitis, and Senzer 2009). This is in contrast to normal cells which require mitogenic growth signals to move from an inactive state into an active proliferative state (Hanahan and Weinberg 2000). Tumor cells are therefore less dependent on exogenous growth stimulation.

Insensitivity to antigrowth signals implies that cancer cells have the ability to evade antiproliferative signals, such as soluble growth inhibitors and immobilized inhibitors, in order to progress (Hanahan and Weinberg 2000). These signals block proliferation by either forcing the cells out of the active proliferative cycle into the G_0 phase of the cell cycle, or by inducing the cells to permanently give up their proliferative potential by making them enter into postmitotic states. At the molecular level, most antiproliferative signals are funneled through RB1 and its two relatives, RBL1 and RBL2. Disruption of the RB1 pathway allows proliferation, causing cells insensitive to antigrowth factors that normally operate along this pathway to block progression (Weinberg 1995).

Evasion of apoptosis is, in addition to proliferation, important for the tumor cell population to expand in number. Apoptosis, programmed cell death, is triggered by a variety of physiologic signals, and will unfold in a precisely choreographed series of steps, including disruption of cellular membranes, breakdown of cytoplasmic nuclear skeletons, degradation of chromosomes, and fragmentation of the nucleus (Fulda 2009). Apoptosis is essential to all types of cancer and the apoptotic program is present in a latent form in almost all cell types in the body.

Limitless replicative potential is a characteristic of most cancer cells, which is due to their ability to maintain their telomeres, the ends of chromosomes, at a length above critical threshold. Telomeres are composed of several thousand repeats of a short 6 bp sequence element. In normal cells, DNA polymerases are unable to completely replicate the 3' ends of chromosomal DNA, resulting in progressive shortening of the telomeres. Telomeres will thereby lose their ability to protect the ends of chromosomal DNA, which eventually leads to death of the affected cell, a process called senescense (Counter *et al.* 1992). Most cancer cells maintain their telomeres by upregulating the expression of the telomerase enzyme, which adds hexanucleotide repeats onto the ends of telomeric DNA, while the remainder is activating a mechanism which maintains telomeres through recombination-based inter-chromosomal changes of sequence information (Bryan and Cech 1999).

Sustained angiogenesis is the process by which cancer cells induce and sustain the growth of new blood vessels. Oxygen and nutrients are crucial for cell function and survival, and are supplied by the vasculature. Cells within aberrant proliferative lesions initially lack the angiogenic ability, reducing their capability to expand. Cancer cells on the other hand

overcome this restriction by inducing and sustaining angiogenesis (Hanahan and Folkman 1996). This is done by activating the angiogenic switch by changing the balance of angiogenesis inducers and countervailing inhibitors, increasing the expression of pro-angiogenic factors like the vascular endothelial growth factor, VEGF, or reducing the expression of anti-angiogenic factors like thrombospondin-1 (Volpert, Dameron, and Bouck 1997;Singh *et al.* 1995).

Although cancer cells are able to regulate angiogenesis, the blood vessels they develop are aberrant and have poor blood flow. Tumors therefore often become hypoxic (Harris 2002), and cells that do not manage to adapt to oxygen and nutrient deprivation undergo cell death by apoptosis, necrosis, and/or autophagy (Zhou *et al.* 2006). Tumor cells mostly respond to reduced oxygen levels through activation of hypoxia inducible factors (HIFs), although the responses may also be HIF-independent (Harris 2002). Paradoxically, hypoxia can also affect tumor growth positively by inducing genetic and adaptive changes (Zhou *et al.* 2006). This will allow the cancer cells to survive and proliferate in a hypoxic environment, resulting in a more aggressive and treatment resistant tumor phenotype.

Tissue invasion and metastasis is a capability of cancer cells to move out from the primary tumor mass, invade adjacent tissues, and travel to distant sites where they settle down and found new colonies, called metastases. This is partly achieved through alteration of proteins involved in the tethering of cells to their surroundings in a tissue, including cell adhesion molecules (CAMs) and members of the immunoglobulin, cadherin, and integrin families.

Energy metabolism is required at a high rate in cancer cells due to increased proliferation rate and the motile nature of cancer cells, resulting in the lack of oxygen and nutrition (Furuta *et al.* 2010). Persistent glucose metabolism and generation of lactate is therefore thought to be an adaptation of tumor cells to hypoxia. Alteration in the expression of genes involved in metabolic pathways like glycolysis and lipogenesis, contribute to the progression of tumor cells to become more aggressive phenotypes. Transformed cells are distinguished from normal cells in the constitutive activation of growth factor signaling pathways, which directly control cellular metabolism (Yalcin *et al.* 2009).

Endocytosis is the process by which cells absorb molecules from outside the cell by engulfing them with their plasma membrane. A mechanism shared by tumor-initiating and

metastasizing cells is an aberrant tendency to disassemble signaling and adhesion complexes and sort them for either degradation or re-assembly, following vesicular trafficking (Mosesson, Mills, and Yarden 2008). Aberrant endocytosis thereby contributes to malignant transformation and is therefore looked upon as a crucial target of the processes driving cancer initiation and progression.

Clinical consequences of genetic alterations

The biological tumor characteristics caused by genetic alterations promote aggressive phenotypes and may, in turn, lead to chemoradioresistance and poor clinical outcome.
Radiotherapy and some types of chemotherapy cause damage of the DNA. The most serious DNA damage induced by radiation is DNA double-strand breaks, whereas chemotherapeutics like cisplatin, causes cross-linking of DNA (Borst, Rottenberg, and Jonkers 2008). In cervical cancer treatment, cisplatin is used in small doses in combination with radiotherapy to increase the cells sensitivity to radiation. However, the benefit of cisplatin is minor (Sundfor *et al.* 1996) and radiotherapy is therefore the most significant treatment for achieving local tumor control.

DNA damage caused by radiotherapy initiates several cellular responses, like inducing cell cycle arrest, promoting apoptosis or mitotic catastrophe, initiating DNA repair, and stopping transcription (Huerta, Gao, and Saha 2009;Sakata *et al.* 2007). Tumors can develop resistance to radiotherapy by deregulation of cell cycle, by evasion of apoptosis, or by adaptation to hypoxia.

Cell cycle delay or arrest occurs after exposure to radiation due to the activation of cell-cycle checkpoints (Kastan and Bartek 2004). When the check points are functional, they block further proliferation of the cells and suppress cancer development. In tumor cells, however, the checkpoints are often disabled because of genetic changes in cell cycle control genes like *TP53* or *BRCA1*, resulting in loss of apoptotic mechanisms during progression. Also tumors that retain a functional *TP53* pathway still frequently lose the basic apoptotic mechanisms and acquire resistance to apoptosis (Igney and Krammer 2002). This is probably due to enhanced expression of anti-apoptotic genes like *BCL2* or inactivation of pro-apoptotic genes like *BAX* and *APAF1*. Mitotic catastrophe is therefore the major cell death mechanism following radiation therapy (Eriksson and Stigbrand 2010). This is a

delayed type of cell death executed days after treatment initiation, and may be induced either as a consequence of DNA damage and deficient cell cycle checkpoints or through hyperamplification of centrosomes (Eriksson *et al.* 2007;Kawamura *et al.* 2004).

Hypoxia leads to resistance to both chemo- and radiotherapy. Cells in a hypoxic environment harbor a range of genetic changes that make them more invasive, more likely to metastasize, and eventually harder to treat. Moreover, hypoxia increases radioresistance due to the requirement of oxygen during irradiation to produce maximal DNA damage through the creation of free radicals (Koch, Kruuv, and Frey 1973). It has been shown that severely hypoxic cells require a 2-3 fold higher dose of radiation to achieve the same level of cell killing compared to well-oxygenated cells (Karar and Maity 2009). Hypoxic cells have also been found to be resistant to chemotherapy, possibly because drugs do not penetrate into regions with low vascularity, because the cells are slowed in their progression through cell cycle, or because of difficulties in the uptake and metabolism of drugs by hypoxic cells (Moulder and Rockwell 1987;Kennedy *et al.* 1980).

Microarray technology

For the past two decades, the development of genomic technology has revolutionized modern biological research (Chen, Jorgenson, and Cheung 2009). Genome wide analyses have made it possible for biologists to analyze genetic events on a global scale, and they have been used in gene discovery, biomarker determination, drug target identification and disease classification, as well as in understanding carcinogenesis. One of the genomic tools for high-throughput and large-scale genomic analysis is microarray technology. This technology allows the simultaneous analysis of thousands of variables in a single sample in a simple hybridization experiment (Jares 2006).

Microarrays were first described in 1995 and were originally designed for depositing DNA onto solid support for gene expression profiling (Schena *et al.* 1995). Since then, the technology has been extended to DNA copy number analysis (Pinkel *et al.* 1998;Snijders *et al.* 2001), protein measurements (Kingsmore 2006), microRNA analysis (Yin, Zhao, and Morris 2008), SNP genotyping (Gunderson *et al.* 2005), DNA methylation profiling (van and Henikoff 2003), and protein-DNA interactions and chromatin modifications using ChiIP-microarrays (Buck and Lieb 2004). In cancer research, the technology has resulted in

numerous of novel opportunities, for example detailed genomic comparison of different tumor types, discovery of novel aberrant chromosomal regions, and sub-classification of tumors according to differences in their gene expression (Rew 2001). Microarrays have also offered new approaches to drug development and therapeutic efficacy, for example by characterizing each cell, tissue, and tumor in terms of the mechanism of drug action, its predicted drug sensitivity, and side-effect profile to a number of reagents. These novel opportunities have aimed in understanding the initiation and development of the disease.

Principle

DNA microarrays consist of a solid matrix where thousands of known nucleic acid fragments (BACs, cDNAs, oligonucleotides) which constitute the microarray probes, are synthesized or spotted to a glass support (Jares 2006). The principle of DNA microarrays relies on the ability of single-strand nucleic acid fragments to hybridize with high specificity to a second complementary single strand to generate a double-stranded DNA molecule (Fig. 7). The sample or target is labeled by using either radioactivity or fluorescent dyes and hybridized to the array surface. After stringency washes have been performed, the complementary target-probe complexes remain tightly bound and the amount of the retained labeled target is quantified by fluorescent detector systems, followed by a data analysis process.

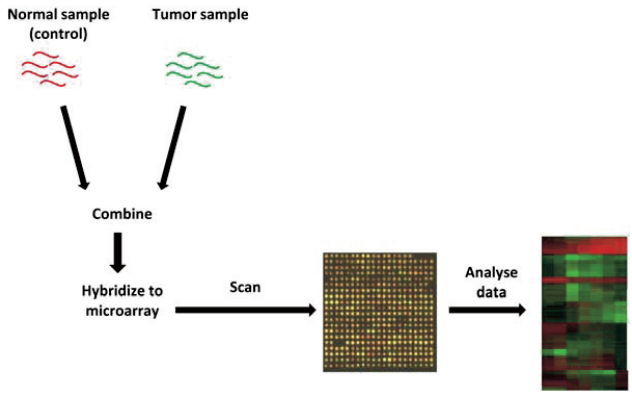


Fig. 7: The principle of two-channel microarray technology

Platforms

The microarray probes are large-insert clones like bacterial artificial chromosomes (BACs) (150-350 kb), P1-derived artificial chromosomes (PACs) (100-300 kb), formid clones (40

kb), cosmid clones (37-52 kb), smaller insert clones (1.5-4.5 kb), cDNA clones (0.5-2 kb), genomic PCR products (100 bp-1.5 kb), and oligonucleotides (25-80 bp). BACs are used for DNA copy number detection and provide the most comprehensive coverage of the genome as well as producing robust hybridizations with low noise. However, the resolution is generally only about 1 Mb and it is difficult to identify single copy number differences smaller than 50 kb (Carter 2007). Oligonucleotides provide the highest potential resolution for array CGH, with the possibility of achieving more than 2 million oligonucleotides per array (NimbleGen, WI). For gene expression analysis, the two most common spotted arrays are cDNA microarrays and oligo-based arrays.

Alternatives to the in-house produced arrays are commercial microarray platforms. The most common platform is the GeneChip system from Affymetrix (Santa Clara, CA). GeneChips are made by synthesizing matched sets of short oligonucleotide pairs, one that matches perfectly and one with a single mismatch, on a silicon-based substrate using a photolithographic process (Ness 2007;Pease *et al.* 1994). Other common platforms are Agilent (Santa Clara, CA) and Illumina (San Diego, CA). Agilent arrays are synthesized following the same principles as Affymetrix, whereas Illumina uses designed oligonucleotide probes attached to beads that are deposited randomly in a support. These are created by either impregnating beads with different concentrations of fluorescent dye, or by some type of barcoding technology. The beads are addressable and used to identify specific binding events that occur on their surface.

Array comparative genomic hybridization

CGH has been widely used as a genome-wide screening method to search for DNA copy numbers (Kallioniemi *et al.* 1992). The first approach was introduced in 1992, where total genomic DNA from a test and a reference cell population were co-hybridized to normal metaphase chromosomes. This limited the detection of events involving small regions of the genome, and the resolution was only of about 5-10 Mb (Carter 2007). The limitation has been resolved by the introduction of microarrays in the late 1990s (Pinkel *et al.* 1998;Solinas-Toldo *et al.* 1997). aCGH combines microarray technology with the CGH approach and the metaphase chromosomes have been replaced by defined nucleic acid fragments. The resolution is therefore limited only by the number, distribution and the length of the sequences spotted onto the array. A higher resolution allows a precise mapping of the boundaries of the detected genomic rearrangements.

Limitations in clinical applications of aCGH

Despite the improvements in the aCGH technique in the last decade, only relative copy numbers have been achieved. Absolute copy numbers have only been obtained on a single gene basis using fluorescence *in situ* hybridization (FISH). This may cause problems when using aCGH for analysis of clinical samples. The relative values are influenced by the ploidy, the normal cell fraction in the sample, the experimental bias, and the DNA copy numbers (Pinkel and Albertson 2005). The tumor ploidy and normal cell content often vary considerably among the samples, making comparison across the tumors unreliable. This is also the case for the intratumor heterogeneity in the DNA copy numbers of some tumors.

Different approaches to correct for the confounding effect caused by the normal cell content of the samples have been attempted, such as excluding these samples from the analysis or handling them separately (Weir *et al.* 2007), correcting the ratio levels based on histological examinations of the tumor sections (Chin *et al.* 2007), or reducing the normal cell fraction using laser capture microdissection (Hunt and Finkelstein 2004). Corrections based on histological examinations are not satisfactory since a different part of the biopsy is used, whereas microdissection is often expensive, difficult to perform, and time consuming. This has led to difficulties in interpreting the data and limited the clinical usefulness of the technique.

Summary of papers

Paper I: ***GeneCount: genome-wide calculation of absolute tumor DNA copy numbers from array comparative genomic hybridization data***

In this paper, GeneCount is presented, a method for genome-wide calculation of absolute DNA copy numbers from clinical aCGH data. GeneCount estimates and corrects for the proportion of normal cells in the samples, corrects for tumor ploidy, and considers possible intratumor heterogeneity in DNA copy numbers. Inputs to the model are tumor ploidy, experimental bias, and aCGH ratios. Predetermined measures of tumor ploidy and experimental bias are therefore needed. The absolute DNA copy numbers were compared with FISH data and showed 97% consistency. The data were used to generate gene dosage (copy number/ploidy) profiles that could be compared across patients regardless of ploidy and normal cell content. Significant improvements compared to existing methods for exploring gene dosages in cancers were demonstrated. In particular, we showed a higher sensitivity in the detection of cervical tumors with copy number changes. GeneCount was implemented in software packages to be used downstream of statistical methods for breakpoint detection, like GLAD and CGH-Explorer.

Paper II: ***Gene dosage, expression and ontology analysis identifies driver genes in the carcinogenesis and chemoradioresistance of cervical cancer***

Integrative analysis of gene dosage, expression, and ontology (GO) data was performed to discover driver genes in the carcinogenesis and chemoradioresistance of cervical cancers. Microarray techniques were used to generate gene dosage and expression profiles of 102 patients. Twenty-nine recurrent gains and losses and 3 losses (on 3p, 13q, 21q) associated with poor outcome after chemoradiotherapy were identified. The 3 predictive regions showed low intratumor heterogeneity, indicating that they had emerged prior to many other alterations and probably were early events in carcinogenesis. The genetic alterations were correlated with gene expression, which identified the genes regulated by the alterations. Correlation with GO data resulted in five biological processes; apoptosis, metabolism, macromolecule localization, translation, and transcription, that were overrepresented among the gene dosage regulated genes. Four genes on 3p (*RYBP*, *GBE1*) and 13q (*FAM48A*, *MED4*) correlated with survival at the gene dosage and expression level and were satisfactorily validated in the independent cohort. Fifty-seven candidate driver genes of 24

genetic events were identified by these integrated analyses, including novel loci responsible for chemoradioresistance. Further mapping of the connections among genetic events, drivers, and biological processes suggested that each individual event stimulates specific processes in carcinogenesis through the coordinated control of multiple genes. These results show that genetic alterations play a significant role in carcinogenesis, and may provide novel therapeutic opportunities of both early and advanced stage cervical cancers.

Paper III: ***Microarray based analyses of gene expression and pathways associated with loss on chromosome 3p in cervical cancer***

In this paper, microarray data was used for identification of genes and pathways affected by the frequent loss on chromosome 3p in cervical cancer. Pretreatment tumor samples from 160 cervical cancer patients who received chemoradiotherapy were included. Pairwise gene dosage (aCGH) and expression (Illumina beadarrays) analysis in 77 patients identified candidate target genes of the 3p loss. Gene expressions associated with the 3p loss were subjected to pathway, gene ontology, and gene set analyses. Protein expression was assessed by immunohistochemistry in 150 patients and cervical cancer cell lines. Loss of 3p11.2-p14.2 was the most frequent 3p event and was associated with poor clinical outcome. Fourteen candidate targets of the loss were identified within the region, and three of the genes, *RYBP*, *TMF1*, and *PSMD6*, generated large networks of interaction partners differentially expressed between patients with and without 3p loss. Apoptosis (*RYBP*) and cellular metabolism (*TMF1*, *PSMD6*) were the biological processes overrepresented in the networks. Gene set analysis showed that the pro-apoptotic gene set, involving several genes in the *RYBP* network, was downregulated in patients with 3p loss. Nuclear RYBP protein expression was found to be downregulated in tumors with 3p loss and was also associated with clinical outcome in stage 1 and 2 cancers ($p=0.034$). These results indicate that multiple target genes and their interaction partners are affected by the frequent 3p11.2-p14.2 loss in cervical cancer. Repression of one of these genes, *RYBP*, leads to evasion of apoptosis and disease progression.

Experimental considerations

Patients

The tumor material used in these studies was collected at the Norwegian Radium Hospital since 2001. The clinical protocol was approved by the regional ethical committee and informed consent was obtained from all patients. The samples were obtained from the primary tumors at the time of diagnosis and made anonymous by using a letter code followed by inclusion number. One to four biopsies were taken from different locations of the tumor, and for the microarray experiments, the biopsies were pooled. Squamous cell carcinoma, adenosquamous carcinoma, and adenocarcinoma were included in paper I and II, whereas paper III only included the squamous cell carcinomas. In paper II and III, only patients who received curative radiotherapy were included, involving tumors with stage 1b-4a, whereas paper I also included stage 4b.

There are several strengths of this tumor material. First, up to 156 tumor samples were included, making the statistical analysis more reliable in comparison to other studies with less samples. Second, all the patients received the same treatment, contributing to a more homogenous patient group, which reduces confounding effects caused by variation in efficiency of different treatments. Third, the long observation time ensured that most failures already had been detected, and fourth, up to four biopsies were pooled from each tumor, accounting for the possible intratumor heterogeneity of the samples.

A limitation of this tumor material is that in paper I and II, patients with different histology were included. This may have influenced the biological findings of paper II, since tumors with different histology might show distinct patterns of gains and losses. This was, however, accounted for in the supplementary material of the paper. In paper I, the difference in histology did not play a significant role, due to the methodological character of the study.

Microarray techniques

aCGH

BAC-arrays were used for the aCGH experiments (papers I-III). These arrays cover the whole genome and provide a high signal to noise ratio compared to oligoarrays, increasing the reliability of the data. The resolution of BAC-arrays is, however, limited. Moreover, only the average gene dosage of each DNA segment is obtained, and each segment may cover several genes. This may result in undetected gains and losses if the alterations involve short DNA segments. On the other hand, in cervical cancer, most gains and losses seem to include large chromosomal regions like the whole p- or q-arm (Wilting *et al.* 2008), resulting in the detection of most alterations although the start and stop positions may be inaccurate. Furthermore, in paper II and III we focused on chromosomal gains and losses with altered expression of the corresponding genes, increasing the reliability of the detected copy number changes.

Gene expression arrays

In paper I and II, cDNA expression arrays were used. These arrays do not cover the whole genome, incorrect annotation is often encountered, and for genes with different isoforms, an average expression value is given. Illumina gene expression bead arrays include all genes in the genome, provide separate data for most isoforms, and thereby give more accurate results. This technique was therefore applied in paper III and for validation in paper II.

Downstream analysis of microarray data

Instead of looking at individual genes, methods such as Network Analysis, Gene Ontology (GO) analysis, and Significant Analysis of Microarrays for Gene Sets (SAM-GS) take whole gene sets into account, incorporating biological knowledge regarding gene function and how genes work together. Analyses like these are useful when working with datasets containing a large number of genes.

Network Analysis allows genes to generate a network of their protein interactions from an integrated set of protein interaction databases (Stark *et al.* 2006;Keshava Prasad *et al.* 2009;Kerrien *et al.* 2007). In this work, we selected the genes that were differentially expressed between patients with and without 3p-loss, and generated second degree of interaction networks for the 3p genes that were significantly regulated by gene dosage. In

this way, we could provide evidences that several interaction partners of the candidate target genes *RYBP*, *TMF1*, and *PSMD6* were deregulated in tumors with 3p-loss.

GO is a collection of controlled vocabularies; molecular function, cellular component, and biological process, describing the biology of a gene product in any organism. The GO categories of two or more gene lists can be compared with statistical tests. eGOn (explore Gene Ontology) is a tool that can be used to annotate, display, and perform statistical hypothesis testing to assess the degree of similarity of GO categories between different gene lists (Beisvag *et al.* 2006). eGOn has advantages over other annotation databases in that it enables filtering of annotations by evidence code, it allows the entry of new annotations, and it provides a series of robust, statistical tests that are thoroughly validated and documented. In paper II, we used the eGOn to identify biological processes that were associated with the recurrent and predictive gene dosage alterations. These biological processes were closely related to known cancer hallmarks, and by using eGOn, we therefore provided evidences that genetic alterations lead to carcinogenesis. In paper III, eGOn was used to identify biological processes that were associated with the interaction networks of *RYBP*, *TMF1*, and *PSMD6*, pointing to the phenotypes resulting from the 3p-loss.

SAM-GS determines whether a defined set of genes shows statistically significant, concordant differences between two biological states (Liu *et al.* 2007;Dinu *et al.* 2007). Testing gene sets rather than genes moves the analysis towards biological themes, thus accelerating the discovery process. Further, the number of tests is reduced and thereby the multiple testing concerns. SAM-GS considers all of the genes in an experiment, not only the genes above an arbitrary cutoff in terms of fold-change or significance. In this work, SAM-GS showed that the pro-apoptotic signaling pathway was significantly downregulated in tumors with 3p-loss.

Discussion

Tool for DNA copy number determination in cervical cancers

GeneCount was developed to enable genome-wide calculation of absolute DNA copy numbers from clinical aCGH data. This method corrects for the normal cell fraction and tumor ploidy, and considers intratumor heterogeneity in DNA copy numbers, making it highly valuable when working with tumor samples. Especially cervical cancers show a high frequency of aneuploidy and by correcting for this and the normal cell content, GeneCount enabled reliable comparisons across experiments, as demonstrated in paper II and III.

The importance of taking tumor ploidy, normal cell content, and the genetic intratumor heterogeneity into account in analysis of aCGH data has also been noticed by others, and several statistical methods dealing with these aspects have later been developed. The ASCAT algorithm dissects allelic-specific copy numbers from SNP arrays (Van *et al.* 2010). A mathematical modeling approach is applied to estimate both the normal cell content and the tumor ploidy, which are utilized in the calculation of the copy numbers. A similar approach has also been proposed by Yau and colleagues (Yau *et al.* 2010).

To obtain absolute copy numbers by GeneCount, predetermination of tumor ploidy is required. This was performed by flow cytometric analysis, which is highly accurate in determining the ploidy. The mathematical models mentioned above estimate the tumor ploidy in addition to copy numbers and normal cell content. GeneCount may therefore result in more accurate copy numbers since fewer parameters have to be modeled. Allelic specific data achieved from SNP arrays are useful, and only minor modifications of GeneCount are required to use this method on the same SNP arrays.

Some of the new methods have also considered the intratumor heterogeneity in the copy number changes. Both Yau and colleagues (Yau *et al.* 2010) and Letouzé and colleagues (Letouze *et al.* 2010) have developed statistical methods that attempt to tackle intratumor heterogeneity. These methods may have advantages over GeneCount in that they are made for SNP arrays, which have high resolution. Even though GeneCount can be used to determine absolute DNA copy numbers from SNP arrays, tests need to be carried out to see

whether the signal-to-noise ratio of these arrays is satisfactory for determining heterogeneity in the copy numbers by GeneCount.

DNA copy number alterations in carcinogenesis

Until now, the overall importance of copy number alterations in cervical carcinogenesis and their relationship to known cancer hallmarks have remained unknown. In paper II we demonstrated how the recurrent gains and losses in cervical carcinoma were affecting the expression of certain genes that, in turn, were associated with five biological processes: apoptosis, metabolism, macromolecule localization, translation, and transcription. These processes were closely related to the known cancer hallmarks; evasion of apoptosis, metabolism and endocytosis. This was further supported by the findings in paper III, where candidate target genes of the loss on 3p were involved in apoptosis and cellular metabolism, indicating an association between the 3p-loss and the biological processes. This work therefore indicates that DNA copy number alterations play a significant role in reaching the hallmarks of cancer and thereby promote carcinogenesis in cervical cancer. The importance of copy number alterations for evasion of apoptosis in cervical cancer has also been demonstrated by Imoto and colleagues (Imoto *et al.* 2002). They showed that amplification of 11q21-q23 correlated with overexpression of the antiapoptotic gene *BIRC2*, resistance to apoptosis, and poor survival after radiotherapy.

To explore whether the recurrent genetic alterations were early or late events in carcinogenesis, the heterogeneity analysis in GeneCount was used. The recurrent alterations seemed to be early events since they showed minor intratumor heterogeneity compared to the other genetic alterations. This indicates that the hallmarks associated with these alterations arise early in the carcinogenesis and may be candidate biomarkers also in early stages of the disease. The more heterogeneous alterations, which arise later in the carcinogenesis, may either be associated with other hallmarks or they may be associated with the same hallmarks as the earlier events, reinforcing their effect and providing the tumor with alternative strategies to progress.

DNA copy number alterations and chemoradioresistance

In paper II it was shown that loss on 3p, 13q, and 21q were associated with chemoradioresistance. As mentioned earlier, tumors can develop resistance to radiotherapy by deregulation of cell cycle, evasion of apoptosis, or adaptation to hypoxia. In paper III it was suggested that loss of *RYBP* led to downregulation of other pro-apoptotic genes, and thereby evasion of apoptosis, which may have promoted chemoradioresistance and a more aggressive phenotype. Hence, a poorer clinical outcome was observed for patients with *RYBP* downregulation. Neither of the other 14 gene dosage regulated genes on 3p in paper III have been directly associated with any of the known biological processes leading to chemoradioresistance. A few of the genes, such as *FOXP1*, *GBE1*, *LRIG1*, *RYBP, TMF1*, and *SHQ1* have, however, shown reduced expression in different types of cancer, and have therefore been implicated as potential tumor suppressor genes.

The target genes on 13q and 21q are still unknown, and therefore also the mechanisms of their association to the clinical outcome. Two candidates on 13q, *MED4* and *FAM48A*, were suggested in paper II, but their potential role in chemoradioresistance is not clear. *FAM48A* is known to be required for activation of the MAPK p38 pathway, which represses cell proliferation, and may therefore be involved in deregulation of cell cycle, whereas loss of *MED4* may impair transcription of genes with anti-cancer effect. However, the use of cDNA expression arrays in paper II limited the number of genes included, and the targets of the alterations may not have been included in the data set. More studies are therefore needed to determine the target genes of 13q and 21q and their role in chemoradioresistance.

Future perspectives

Implementation in the clinic

Biomarkers can either be used in diagnosis to select patients for different treatment strategies, or in development of targeted therapeutics to improve the effect of the standard radiation treatment. This work proposes novel candidate biomarkers that may be used clinically, both in diagnosis and in therapy. To implement the candidate biomarkers in the cervical cancer clinic, stringent clinical and biological validation is needed (Chin and Gray 2008).

Diagnosis

Before the DNA copy numbers, gene expression or protein expression can be used for risk stratification and therapy decision, clinical validation in independent patient groups is needed to ensure reproducibility. This should preferably be carried out using multicenter trials. Multicenter trials have the benefits of recruiting a large number of patients at different geographic locations, enabling the inclusion of a wider range of population groups and the ability to compare results among different medical institutions. This will increase the generalizability, since significant variations between population groups with different genetic and environmental background may exist. Successful validation of the biomarkers proposed in this work, would be an important step towards implementing the copy number alterations on 3p, 13q, or 21q, *RYBP* gene expression, or RYBP protein expression as predictive, diagnostic biomarkers in cervical cancer. By adopting similar strategies on 13q and 21q as was used on 3p to find the target genes, even more biomarkers at the gene/protein expression level could be candidates for implementation. All three levels could potentially be used, but alterations at the DNA level might have the largest potential as biomarkers due to the stability of copy number alterations. Moreover, even though the three genomic alterations could be used separately, paper II suggests that it is more to gain if information about all three regions is combined. To date, no diagnostic biomarkers exist for cervical cancer, but other cancer types, like breast cancer, have the benefit of using for example the amplification of the *HER2* gene, to predict the response to trastuzumab and lapatinib (Esteva *et al.* 2010).

Therapy

To use biomarkers in therapy, pre-clinical and clinical trials are required to collect safety and efficacy data for health interventions. A better understanding of the gene or protein function and the effect of manipulation is also needed. Pre-clinical trials, starting with in vitro studies and continuing with in vivo studies on animal models, form a major part of the biological validation and are necessary before starting the clinical trials (Chin and Gray 2008;Yap *et al.* 2010). All phases of clinical trials will normally be carried out, starting with obtaining information regarding efficacy and toxicity, then assessing safety, tolerability, and how well the drug works. Finally, randomized, controlled multicenter trials on large patient groups have to be completed.

Targeting oncogenes can be accomplished by inhibition of oncogene transcription, reduction of mRNA translation, or by interference with oncoprotein interactions, transportation, and function (Hughes 2004). In this work, however, the candidate biomarkers have tumor suppressor function, and mechanisms for targeting suppressor genes are often more complicated. One strategy is to restore gene expression using adenoviral vectors. This has been tested in tumor cell lines for several tumor suppressor genes, including *RYBP* (Novak and Phillips 2008). Infecting cells with a generated adenovirus expressing *RYBP* resulted in induction of apoptosis. The effect was seen only in tumor cell lines, whereas normal cells were unaffected, suggesting that *RYBP* has a potential as a cancer gene therapy agent (Danen-van Oorschot *et al.* 2004).

References

1. Alberts B, Johnson A, Lewis J, Raff M, Roberts K, Walter P (2002) Cancer. In: *Molecular biology of the cell*. New York: Garland Science.

2. Albertson DG, Collins C, McCormick F, Gray JW (2003) Chromosome aberrations in solid tumors. *Nat. Genet.*, **34:** 369-376.

3. Anglicheau D, Muthukumar T, Suthanthiran M (2010) MicroRNAs: small RNAs with big effects. *Transplantation.*, **90:** 105-112.

4. Baalbergen A, Veenstra Y, Stalpers LL, Ansink AC (2010) Primary surgery versus primary radiation therapy with or without chemotherapy for early adenocarcinoma of the uterine cervix. *Cochrane. Database. Syst. Rev.*, **20;** CD006248.

5. Barranco SC, Perry RR, Durm ME, Werner AL, Gregorcyk SG, Bolton WE, Kolm P, Townsend CM, Jr. (1994) Intratumor variability in prognostic indicators may be the cause of conflicting estimates of patient survival and response to therapy. *Cancer Res.*, **54:** 5351-5356.

6. Barranco SC, Townsend CM, Jr., Jenkins VK, Koester SK, Ho BY, Reumont KJ (1988) Treatment-induced changes in sensitivity in a multiclonal human tumor mixture model in vitro. *Cancer Res.*, **48:** 2749-2755.

7. Bedell MA, Hudson JB, Golub TR, Turyk ME, Hosken M, Wilbanks GD, Laimins LA (1991) Amplification of human papillomavirus genomes in vitro is dependent on epithelial differentiation. *J. Virol.*, **65:** 2254-2260.

8. Beisvag V, Junge FK, Bergum H, Jolsum L, Lydersen S, Gunther CC, Ramampiaro H, Langaas M, Sandvik AK, Laegreid A (2006) GeneTools--application for functional annotation and statistical hypothesis testing. *BMC. Bioinformatics.*, **7:470.:** 470.

9. Borst P, Rottenberg S, Jonkers J (2008) How do real tumors become resistant to cisplatin? *Cell Cycle.*, **7:** 1353-1359.

10. Bryan TM, Cech TR (1999) Telomerase and the maintenance of chromosome ends. *Curr. Opin. Cell Biol.*, **11:** 318-324.

11. Buck MJ, Lieb JD (2004) ChIP-chip: considerations for the design, analysis, and application of genome-wide chromatin immunoprecipitation experiments. *Genomics.*, **83:** 349-360.

12. Calin GA, Sevignani C, Dumitru CD, Hyslop T, Noch E, Yendamuri S, Shimizu M, Rattan S, Bullrich F, Negrini M, Croce CM (2004) Human microRNA genes are frequently located at fragile sites and genomic regions involved in cancers. *Proc. Natl. Acad. Sci. U. S. A.*, **101:** 2999-3004.

13. Camisπo CC, Brenna SMF, Lombardelli KVP, Djahjah MC, Zeferino LC (2007) RessonΓncia magnθtica no estadiamento dos tumores de colo uterino. *Radiologia Brasileira*, **40:** 207-215.

14. Cancer Registry of Norway (2010) http://www.kreftregisteret.no/no/Registrene/Kreftstatistikk/

15. Cannistra SA, Niloff JM (1996) Cancer of the uterine cervix. *N. Engl. J. Med.*, **334:** 1030-1038.

16. Carter NP (2007) Methods and strategies for analyzing copy number variation using DNA microarrays. *Nat. Genet.*, **39:** S16-S21.

17. Cervical Cancer Statistics (2007) http://www.cervicalcancer.org/statistics.html

18. Chen X, Jorgenson E, Cheung ST (2009) New tools for functional genomic analysis. *Drug Discov. Today.*, **14:** 754-760.

19. Chin L, Gray JW (2008) Translating insights from the cancer genome into clinical practice. *Nature.*, **452:** 553-563.

20. Chin SF, Teschendorff AE, Marioni JC, Wang Y, Barbosa-Morais NL, Thorne NP, Costa JL, Pinder SE, van de Wiel MA, Green AR, Ellis IO, Porter PL, Tavare S, Brenton JD, Ylstra B, Caldas C (2007) High-resolution aCGH and expression profiling identifies a novel genomic subtype of ER negative breast cancer. *Genome Biol.*, **8:** R215.

21. Counter CM, Avilion AA, LeFeuvre CE, Stewart NG, Greider CW, Harley CB, Bacchetti S (1992) Telomere shortening associated with chromosome instability is arrested in immortal cells which express telomerase activity. *EMBO J.*, **11:** 1921-1929.

22. Croce CM (2008) Oncogenes and cancer. *N. Engl. J. Med.*, **358:** 502-511.

23. Danen-van Oorschot AA, Voskamp P, Seelen MC, van Miltenburg MH, Bolk MW, Tait SW, Boesen-de Cock JG, Rohn JL, Borst J, Noteborn MH (2004) Human death effector domain-associated factor interacts with the viral apoptosis agonist Apoptin and exerts tumor-preferential cell killing. *Cell Death. Differ.*, **11:** 564-573.

24. Dellas A, Torhorst J, Jiang F, Proffitt J, Schultheiss E, Holzgreve W, Sauter G, Mihatsch MJ, Moch H (1999) Prognostic value of genomic alterations in invasive cervical squamous cell carcinoma of clinical stage IB detected by comparative genomic hybridization. *Cancer Res.*, **59:** 3475-3479.

25. Dietel M, Sers C (2006) Personalized medicine and development of targeted therapies: The upcoming challenge for diagnostic molecular pathology. A review. *Virchows Arch.*, **448:** 744-755.

26. Dinu I, Potter JD, Mueller T, Liu Q, Adewale AJ, Jhangri GS, Einecke G, Famulski KS, Halloran P, Yasui Y (2007) Improving gene set analysis of microarray data by SAM-GS. *BMC. Bioinformatics.*, **8:242.:** 242.

27. Duensing S, Duensing A, Flores ER, Do A, Lambert PF, Munger K (2001) Centrosome abnormalities and genomic instability by episomal expression of human papillomavirus type 16 in raft cultures of human keratinocytes. *J. Virol.*, **75:** 7712-7716.

28. Duensing S, Lee LY, Duensing A, Basile J, Piboonniyom S, Gonzalez S, Crum CP, Munger K (2000) The human papillomavirus type 16 E6 and E7 oncoproteins cooperate to induce mitotic defects and genomic instability by uncoupling centrosome duplication from the cell division cycle. *Proc. Natl. Acad. Sci. U. S. A.*, **97:** 10002-10007.

29. Durst M, Croce CM, Gissmann L, Schwarz E, Huebner K (1987) Papillomavirus sequences integrate near cellular oncogenes in some cervical carcinomas. *Proc. Natl. Acad. Sci. U. S. A.*, **84:** 1070-1074.

30. Eifel PJ (2006) Concurrent chemotherapy and radiation therapy as the standard of care for cervical cancer. *Nat. Clin. Pract. Oncol.*, **3:** 248-255.

31. Eriksson D, Lofroth PO, Johansson L, Riklund KA, Stigbrand T (2007) Cell cycle disturbances and mitotic catastrophes in HeLa Hep2 cells following 2.5 to 10 Gy of ionizing radiation. *Clin. Cancer Res.*, **13:** 5501s-5508s.

32. Eriksson D, Stigbrand T (2010) Radiation-induced cell death mechanisms. *Tumour. Biol.*, **31:** 363-372.

33. Esteva FJ, Yu D, Hung MC, Hortobagyi GN (2010) Molecular predictors of response to trastuzumab and lapatinib in breast cancer. *Nat. Rev. Clin. Oncol.*, **7:** 98-107.

34. Ferber MJ, Montoya DP, Yu C, Aderca I, McGee A, Thorland EC, Nagorney DM, Gostout BS, Burgart LJ, Boix L, Bruix J, McMahon BJ, Cheung TH, Chung TK, Wong YF, Smith DI, Roberts LR (2003a) Integrations of the hepatitis B virus (HBV) and human papillomavirus (HPV) into the human telomerase reverse transcriptase (hTERT) gene in liver and cervical cancers. *Oncogene.*, **22:** 3813-3820.

35. Ferber MJ, Thorland EC, Brink AA, Rapp AK, Phillips LA, McGovern R, Gostout BS, Cheung TH, Chung TK, Fu WY, Smith DI (2003b) Preferential integration of human papillomavirus type 18 near the c-myc locus in cervical carcinoma. *Oncogene.*, **22:** 7233-7242.

36. Fulda S (2009) Tumor resistance to apoptosis. *Int. J. Cancer.*, **124:** 511-515.

37. Furuta E, Okuda H, Kobayashi A, Watabe K (2010) Metabolic genes in cancer: their roles in tumor progression and clinical implications. *Biochim. Biophys. Acta.*, **1805:** 141-152.

38. Garnett TO, Duerksen-Hughes PJ (2006) Modulation of apoptosis by human papillomavirus (HPV) oncoproteins. *Arch. Virol.*, **151:** 2321-2335.

39. Georgiades IB, Curtis LJ, Morris RM, Bird CC, Wyllie AH (1999) Heterogeneity studies identify a subset of sporadic colorectal cancers without evidence for chromosomal or microsatellite instability. *Oncogene.*, **18:** 7933-7940.

40. Grubisic G, Klaric P, Jokanovic L, Soljacic VH, Grbavac I, Bolanca I (2009) Diagnostic approach for precancerous and early invasive cancerous lesions of the uterine cervix. *Coll. Antropol.*, **33:** 1431-1436.

41. Gunderson KL, Steemers FJ, Lee G, Mendoza LG, Chee MS (2005) A genome-wide scalable SNP genotyping assay using microarray technology. *Nat. Genet.*, **37:** 549-554.

42. Guo Z, Wu F, Asplund A, Hu X, Mazurenko N, Kisseljov F, Ponten J, Wilander E (2001) Analysis of intratumoral heterogeneity of chromosome 3p deletions and genetic evidence of polyclonal origin of cervical squamous carcinoma. *Mod. Pathol.*, **14:** 54-61.

43. Hanahan D, Folkman J (1996) Patterns and emerging mechanisms of the angiogenic switch during tumorigenesis. *Cell.*, **86:** 353-364.

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

45. Harris AL (2002) Hypoxia--a key regulatory factor in tumour growth. *Nat. Rev. Cancer.*, **2:** 38-47.

46. Health MD: How is Cervical Cancer Treated? (2010) http://www.ehealthmd.com/library/cervicalcancer/CC_treatment.html#chemo

47. Herceg Z, Hainaut P (2007) Genetic and epigenetic alterations as biomarkers for cancer detection, diagnosis and prognosis. *Mol. Oncol.*, **1:** 26-41.

48. Hermsen M, Postma C, Baak J, Weiss M, Rapallo A, Sciutto A, Roemen G, Arends JW, Williams R, Giaretti W, De GA, Meijer G (2002) Colorectal adenoma to carcinoma progression follows multiple pathways of chromosomal instability. *Gastroenterology.*, **123:** 1109-1119.

49. Heselmeyer K, Macville M, Schrock E, Blegen H, Hellstrom AC, Shah K, Auer G, Ried T (1997) Advanced-stage cervical carcinomas are defined by a recurrent pattern of chromosomal aberrations revealing high genetic instability and a consistent gain of chromosome arm 3q. *Genes Chromosomes. Cancer.*, **19:** 233-240.

50. Heselmeyer K, Schrock E, du MS, Blegen H, Shah K, Steinbeck R, Auer G, Ried T (1996) Gain of chromosome 3q defines the transition from severe dysplasia to invasive carcinoma of the uterine cervix. *Proc. Natl. Acad. Sci. U. S. A.*, **93:** 479-484.

51. Hodgson JG, Chin K, Collins C, Gray JW (2003) Genome amplification of chromosome 20 in breast cancer. *Breast Cancer Res. Treat.*, **78:** 337-345.

52. Huerta S, Gao X, Saha D (2009) Mechanisms of resistance to ionizing radiation in rectal cancer. *Expert. Rev. Mol. Diagn.*, **9:** 469-480.

53. Hughes RM (2004) Strategies for cancer gene therapy. *J. Surg. Oncol.*, **85:** 28-35.

54. Hulka BS (1982) Risk factors for cervical cancer. *J. Chronic. Dis.*, **35:** 3-11.

55. Hunt JL, Finkelstein SD (2004) Microdissection techniques for molecular testing in surgical pathology. *Arch. Pathol. Lab Med.*, **128:** 1372-1378.

56. Igney FH, Krammer PH (2002) Death and anti-death: tumour resistance to apoptosis. *Nat. Rev. Cancer.*, **2:** 277-288.

57. Imoto I, Tsuda H, Hirasawa A, Miura M, Sakamoto M, Hirohashi S, Inazawa J (2002) Expression of cIAP1, a target for 11q22 amplification, correlates with resistance of cervical cancers to radiotherapy. *Cancer Res.*, **62:** 4860-4866.

58. Jares P (2006) DNA microarray applications in functional genomics. *Ultrastruct. Pathol.*, **30:** 209-219.

59. Kallioniemi A, Kallioniemi OP, Piper J, Tanner M, Stokke T, Chen L, Smith HS, Pinkel D, Gray JW, Waldman FM (1994) Detection and mapping of amplified DNA sequences in breast cancer by comparative genomic hybridization. *Proc. Natl. Acad. Sci. U. S. A.*, **91:** 2156-2160.

60. Kallioniemi A, Kallioniemi OP, Sudar D, Rutovitz D, Gray JW, Waldman F, Pinkel D (1992) Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors. *Science.*, **258:** 818-821.

61. Karar J, Maity A (2009) Modulating the tumor microenvironment to increase radiation responsiveness. *Cancer Biol. Ther.*, **8:** 1994-2001.

62. Kastan MB, Bartek J (2004) Cell-cycle checkpoints and cancer. *Nature.*, **432:** 316-323.

63. Kawamura K, Fujikawa-Yamamoto K, Ozaki M, Iwabuchi K, Nakashima H, Domiki C, Morita N, Inoue M, Tokunaga K, Shiba N, Ikeda R, Suzuki K (2004) Centrosome hyperamplification and chromosomal damage after exposure to radiation. *Oncology.*, **67:** 460-470.

64. Kennedy KA, Teicher BA, Rockwell S, Sartorelli AC (1980) The hypoxic tumor cell: a target for selective cancer chemotherapy. *Biochem. Pharmacol.*, **29:** 1-8.

65. Kerrien S, am-Faruque Y, Aranda B, Bancarz I, Bridge A, Derow C, Dimmer E, Feuermann M, Friedrichsen A, Huntley R, Kohler C, Khadake J, Leroy C, Liban A, Lieftink C, Montecchi-Palazzi L, Orchard S, Risse J, Robbe K, Roechert B, Thorneycroft D, Zhang Y, Apweiler R, Hermjakob H (2007) IntAct--open source resource for molecular interaction data. *Nucleic Acids Res.*, **35:** D561-D565.

66. Keshava Prasad TS, Goel R, Kandasamy K, Keerthikumar S, Kumar S, Mathivanan S, Telikicherla D, Raju R, Shafreen B, Venugopal A, Balakrishnan L, Marimuthu A, Banerjee S, Somanathan DS, Sebastian A, Rani S, Ray S, Harrys Kishore CJ, Kanth S, Ahmed M, Kashyap MK, Mohmood R, Ramachandra YL, Krishna V, Rahiman BA, Mohan S, Ranganathan P, Ramabadran S, Chaerkady R, Pandey A (2009) Human Protein Reference Database--2009 update. *Nucleic Acids Res.*, **37:** D767-D772.

67. Kingsmore SF (2006) Multiplexed protein measurement: technologies and applications of protein and antibody arrays. *Nat. Rev. Drug Discov.*, **5:** 310-320.

68. Kirchhoff M, Rose H, Petersen BL, Maahr J, Gerdes T, Lundsteen C, Bryndorf T, Kryger-Baggesen N, Christensen L, Engelholm SA, Philip J (1999) Comparative

genomic hybridization reveals a recurrent pattern of chromosomal aberrations in severe dysplasia/carcinoma in situ of the cervix and in advanced-stage cervical carcinoma. *Genes Chromosomes. Cancer.*, **24:** 144-150.

69. Koch CJ, Kruuv J, Frey HE (1973) Variation in radiation response of mammalian cells as a function of oxygen tension. *Radiat. Res.*, **53:** 33-42.

70. Koyama T, Tamai K, Togashi K (2007) Staging of carcinoma of the uterine cervix and endometrium. *Eur. Radiol.*, **17:** 2009-2019.

71. Lazo PA (1999) The molecular genetics of cervical carcinoma. *Br. J. Cancer.*, **80:** 2008-2018.

72. Lee JW, Choi CH, Choi JJ, Park YA, Kim SJ, Hwang SY, Kim WY, Kim TJ, Lee JH, Kim BG, Bae DS (2008) Altered MicroRNA expression in cervical carcinomas. *Clin. Cancer Res.*, **14:** 2535-2542.

73. Lehoux M, D'Abramo CM, Archambault J (2009) Molecular mechanisms of human papillomavirus-induced carcinogenesis. *Public Health Genomics.*, **12:** 268-280.

74. Lengauer C, Kinzler KW, Vogelstein B (1998) Genetic instabilities in human cancers. *Nature.*, **396:** 643-649.

75. Letouze E, Allory Y, Bollet MA, Radvanyi F, Guyon F (2010) Analysis of the copy number profiles of several tumor samples from the same patient reveals the successive steps in tumorigenesis. *Genome Biol.*, **11:** R76.

76. Liu Q, Dinu I, Adewale AJ, Potter JD, Yasui Y (2007) Comparative evaluation of gene-set analysis methods. *BMC. Bioinformatics.*, **8:431.:** 431.

77. Lyng H, Beigi M, Svendsrud DH, Brustugun OT, Stokke T, Kristensen GB, Sundfor K, Skjonsberg A, De Angelis PM (2004) Intratumor chromosomal heterogeneity in advanced carcinomas of the uterine cervix. *Int. J. Cancer.*, **111:** 358-366.

78. Mees C, Nemunaitis J, Senzer N (2009) Transcription factors: their potential as targets for an individualized therapeutic approach to cancer. *Cancer Gene Ther.*, **16:** 103-112.

79. Mosesson Y, Mills GB, Yarden Y (2008) Derailed endocytosis: an emerging feature of cancer. *Nat. Rev. Cancer.*, **8:** 835-850.

80. Moulder JE, Rockwell S (1987) Tumor hypoxia: its impact on cancer therapy. *Cancer Metastasis Rev.*, **5:** 313-341.

81. Mueller MM, Fusenig NE (2004) Friends or foes - bipolar effects of the tumour stroma in cancer. *Nat. Rev. Cancer.*, **4:** 839-849.

82. Muleris M, Almeida A, Gerbault-Seureau M, Malfoy B, Dutrillaux B (1994) Detection of DNA amplification in 17 primary breast carcinomas with homogeneously staining regions by a modified comparative genomic hybridization technique. *Genes Chromosomes. Cancer.*, **10:** 160-170.

83. Munger K (2002) The role of human papillomaviruses in human cancers. *Front Biosci.*, **7:d641-9.:** d641-d649.

84. Nambiar M, Kari V, Raghavan SC (2008) Chromosomal translocations in cancer. *Biochim. Biophys. Acta.*, **1786:** 139-152.

85. Narayan G, Pulido HA, Koul S, Lu XY, Harris CP, Yeh YA, Vargas H, Posso H, Terry MB, Gissmann L, Schneider A, Mansukhani M, Rao PH, Murty VV (2003) Genetic analysis identifies putative tumor suppressor sites at 2q35-q36.1 and 2q36.3-q37.1 involved in cervical cancer progression. *Oncogene.*, **22:** 3489-3499.

86. Ness SA (2007) Microarray analysis: basic strategies for successful experiments. *Mol. Biotechnol.*, **36:** 205-219.

87. Novak RL, Phillips AC (2008) Adenoviral-mediated Rybp expression promotes tumor cell-specific apoptosis. *Cancer Gene Ther.*, **15:** 713-722.

88. Nowell PC (1976) The clonal evolution of tumor cell populations. *Science.*, **194:** 23-28.

89. Paget S (1989) The distribution of secondary growths in cancer of the breast. 1889. *Cancer Metastasis Rev.*, **8:** 98-101.

90. Parkin DM, Bray F, Ferlay J, Pisani P (2005) Global cancer statistics, 2002. *CA Cancer J. Clin.*, **55:** 74-108.

91. Pease AC, Solas D, Sullivan EJ, Cronin MT, Holmes CP, Fodor SP (1994) Light-generated oligonucleotide arrays for rapid DNA sequence analysis. *Proc. Natl. Acad. Sci. U. S. A.*, **91:** 5022-5026.

92. Peter M, Rosty C, Couturier J, Radvanyi F, Teshima H, Sastre-Garau X (2006) MYC activation associated with the integration of HPV DNA at the MYC locus in genital tumors. *Oncogene.*, **25:** 5985-5993.

93. Pett M, Coleman N (2007) Integration of high-risk human papillomavirus: a key event in cervical carcinogenesis? *J. Pathol.*, **212:** 356-367.

94. Pinkel D, Albertson DG (2005) Array comparative genomic hybridization and its applications in cancer. *Nat. Genet.*, **37 Suppl:S11-7.:** S11-S17.

95. Pinkel D, Segraves R, Sudar D, Clark S, Poole I, Kowbel D, Collins C, Kuo WL, Chen C, Zhai Y, Dairkee SH, Ljung BM, Gray JW, Albertson DG (1998) High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. *Nat. Genet.*, **20:** 207-211.

96. Rao PH, rias-Pulido H, Lu XY, Harris CP, Vargas H, Zhang FF, Narayan G, Schneider A, Terry MB, Murty VV (2004) Chromosomal amplifications, 3q gain and deletions of 2q33-q37 are the frequent genetic changes in cervical carcinoma. *BMC. Cancer.*, **4:5.:** 5.

97. Reeves WC, Rawls WE, Brinton LA (1989) Epidemiology of genital papillomaviruses and cervical cancer. *Rev. Infect. Dis.*, **11:** 426-439.

98. Rew DA (2001) DNA microarray technology in cancer research. *Eur. J. Surg. Oncol.*, **27:** 504-508.

99. Richards RI (2001) Fragile and unstable chromosomes in cancer: causes and consequences. *Trends Genet.*, **17:** 339-345.

100. Ried T, Heselmeyer-Haddad K, Blegen H, Schrock E, Auer G (1999) Genomic changes defining the genesis, progression, and malignancy potential in solid human tumors: a phenotype/genotype correlation. *Genes Chromosomes. Cancer.*, **25:** 195-204.

101. Rothenberg M, Ling V (1989) Multidrug resistance: molecular biology and clinical relevance. *J. Natl. Cancer Inst.*, **81:** 907-910.

102. Sakata K, Someya M, Matsumoto Y, Hareyama M (2007) Ability to repair DNA double-strand breaks related to cancer susceptibility and radiosensitivity. *Radiat. Med.*, **25:** 433-438.

103. Schena M, Shalon D, Davis RW, Brown PO (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science.*, **20;270:** 467-470.

104. Singh RK, Gutman M, Bucana CD, Sanchez R, Llansa N, Fidler IJ (1995) Interferons alpha and beta down-regulate the expression of basic fibroblast growth factor in human carcinomas. *Proc. Natl. Acad. Sci. U. S. A.*, **92:** 4562-4566.

105. Snijders AM, Nowak N, Segraves R, Blackwood S, Brown N, Conroy J, Hamilton G, Hindle AK, Huey B, Kimura K, Law S, Myambo K, Palmer J, Ylstra B, Yue JP, Gray JW, Jain AN, Pinkel D, Albertson DG (2001) Assembly of microarrays for genome-wide measurement of DNA copy number. *Nat. Genet.*, **29:** 263-264.

106. Solinas-Toldo S, Lampel S, Stilgenbauer S, Nickolenko J, Benner A, Dohner H, Cremer T, Lichter P (1997) Matrix-based comparative genomic hybridization: biochips to screen for genomic imbalances. *Genes Chromosomes. Cancer.*, **20:** 399-407.

107. Stanley MA, Browne HM, Appleby M, Minson AC (1989) Properties of a non-tumorigenic human cervical keratinocyte cell line. *Int. J. Cancer.*, **43:** 672-676.

108. Stark C, Breitkreutz BJ, Reguly T, Boucher L, Breitkreutz A, Tyers M (2006) BioGRID: a general repository for interaction datasets. *Nucleic Acids Res.*, **34:** D535-D539.

109. Stewart BW, Kleihues P. World Cancer Report. 2003. Lyon, IARC Press.
Ref Type: Report

110. Sundfor K, Trope CG, Hogberg T, Onsrud M, Koern J, Simonsen E, Bertelsen K, Westberg R (1996) Radiotherapy and neoadjuvant chemotherapy for cervical carcinoma. A randomized multicenter study of sequential cisplatin and 5-fluorouracil and radiotherapy in advanced cervical carcinoma stage 3B and 4A. *Cancer.*, **77:** 2371-2378.

111. Szalmas A, Konya J (2009) Epigenetic alterations in cervical carcinogenesis. *Semin. Cancer Biol.*, **19:** 144-152.

112. Thorland EC, Myers SL, Gostout BS, Smith DI (2003) Common fragile sites are preferential targets for HPV16 integrations in cervical tumors. *Oncogene.*, **22:** 1225-1237.

113. Thorland EC, Myers SL, Persing DH, Sarkar G, McGovern RM, Gostout BS, Smith DI (2000) Human papillomavirus type 16 integrations in cervical tumors frequently occur in common fragile sites. *Cancer Res.*, **60:** 5916-5921.

114. Umayahara K, Numa F, Suehiro Y, Sakata A, Nawata S, Ogata H, Suminami Y, Sakamoto M, Sasaki K, Kato H (2002) Comparative genomic hybridization detects genetic alterations during early stages of cervical cancer progression. *Genes Chromosomes. Cancer.*, **33:** 98-102.

115. Valencia-Sanchez MA, Liu J, Hannon GJ, Parker R (2006) Control of translation and mRNA degradation by miRNAs and siRNAs. *Genes Dev.*, **20:** 515-524.

116. Van LP, Nordgard SH, Lingjaerde OC, Russnes HG, Rye IH, Sun W, Weigman VJ, Marynen P, Zetterberg A, Naume B, Perou CM, Borresen-Dale AL, Kristensen VN (2010) Allele-specific copy number analysis of tumors. *Proc. Natl. Acad. Sci. U. S. A.*, **107:** 16910-16915.

117. van SB, Henikoff S (2003) Epigenomic profiling using microarrays. *Biotechniques.*, **35:** 346-4, 356.

118. Vogelstein B, Kinzler KW (1993) The multistep nature of cancer. *Trends Genet.*, **9:** 138-141.

119. Vogelstein B, Kinzler KW (2004) Cancer genes and the pathways they control. *Nat. Med.*, **10:** 789-799.

120. Volpert OV, Dameron KM, Bouck N (1997) Sequential development of an angiogenic phenotype by human fibroblasts progressing to tumorigenicity. *Oncogene.*, **14:** 1495-1502.

121. Wajed SA, Laird PW, DeMeester TR (2001) DNA methylation: an alternative pathway to cancer. *Ann. Surg.*, **234:** 10-20.

122. Weinberg RA (1995) The retinoblastoma protein and cell cycle control. *Cell.*, **81:** 323-330.

123. Weir BA, Woo MS, Getz G, Perner S, Ding L, Beroukhim R, Lin WM, Province MA, Kraja A, Johnson LA, Shah K, Sato M, Thomas RK, Barletta JA, Borecki IB, Broderick S, Chang AC, Chiang DY, Chirieac LR, Cho J, Fujii Y, Gazdar AF, Giordano T, Greulich H, Hanna M, Johnson BE, Kris MG, Lash A, Lin L, Lindeman N, Mardis ER, McPherson JD, Minna JD, Morgan MB, Nadel M, Orringer MB, Osborne JR, Ozenberger B, Ramos AH, Robinson J, Roth JA, Rusch V, Sasaki H, Shepherd F, Sougnez C, Spitz MR, Tsao MS, Twomey D, Verhaak RG, Weinstock GM, Wheeler DA, Winckler W, Yoshizawa A, Yu S, Zakowski MF, Zhang Q, Beer DG, Wistuba II, Watson MA, Garraway LA, Ladanyi M, Travis WD, Pao W, Rubin

MA, Gabriel SB, Gibbs RA, Varmus HE, Wilson RK, Lander ES, Meyerson M (2007) Characterizing the cancer genome in lung adenocarcinoma. *Nature.*, **450:** 893-898.

124. Weiss MM, Snijders AM, Kuipers EJ, Ylstra B, Pinkel D, Meuwissen SG, van Diest PJ, Albertson DG, Meijer GA (2003) Determination of amplicon boundaries at 20q13.2 in tissue samples of human gastric adenocarcinomas by high-resolution microarray comparative genomic hybridization. *J. Pathol.*, **200:** 320-326.

125. Wentzensen N, Ridder R, Klaes R, Vinokurova S, Schaefer U, Doeberitz MK (2002) Characterization of viral-cellular fusion transcripts in a large series of HPV16 and 18 positive anogenital lesions. *Oncogene.*, **21:** 419-426.

126. Wentzensen N, Vinokurova S, von Knebel DM (2004) Systematic review of genomic integration sites of human papillomavirus genomes in epithelial dysplasia and invasive cancer of the female lower genital tract. *Cancer Res.*, **64:** 3878-3884.

127. White AE, Livanos EM, Tlsty TD (1994) Differential disruption of genomic integrity and cell cycle regulation in normal human fibroblasts by the HPV oncoproteins. *Genes Dev.*, **8:** 666-677.

128. Wilting SM, de WJ, Meijer CJ, Berkhof J, Yi Y, van Wieringen WN, Braakhuis BJ, Meijer GA, Ylstra B, Snijders PJ, Steenbergen RD (2008) Integrated genomic and transcriptional profiling identifies chromosomal loci with altered gene expression in cervical cancer. *Genes Chromosomes. Cancer.*, **..**

129. Yalcin A, Telang S, Clem B, Chesney J (2009) Regulation of glucose metabolism by 6-phosphofructo-2-kinase/fructose-2,6-bisphosphatases in cancer. *Exp. Mol. Pathol.*, **86:** 174-179.

130. Yap TA, Sandhu SK, Workman P, de Bono JS (2010) Envisioning the future of early anticancer drug development. *Nat. Rev. Cancer.*, **10:** 514-523.

131. Yau C, Mouradov D, Jorissen RN, Colella S, Mirza G, Steers G, Harris A, Ragoussis J, Sieber O, Holmes CC (2010) A statistical approach for detecting genomic aberrations in heterogeneous tumor samples from single nucleotide polymorphism genotyping data. *Genome Biol.*, **11:** R92.

132. Yin JQ, Zhao RC, Morris KV (2008) Profiling microRNA expression with microarrays. *Trends Biotechnol.*, **26:** 70-76.

133. Yu T, Ferber MJ, Cheung TH, Chung TK, Wong YF, Smith DI (2005) The role of viral integration in the development of cervical cancer. *Cancer Genet. Cytogenet.*, **158:** 27-34.

134. Zhou J, Schmid T, Schnitzer S, Brune B (2006) Tumor hypoxia and cancer progression. *Cancer Lett.*, **237:** 10-21.

I

Method

Open Access

GeneCount: genome-wide calculation of absolute tumor DNA copy numbers from array comparative genomic hybridization data

Heidi Lyng[*], Malin Lando[*], Runar S Brøvig[*], Debbie H Svendsrud[*], Morten Johansen[†], Eivind Galteland[*], Odd T Brustugun[*‡], Leonardo A Meza-Zepeda[†§], Ola Myklebost[†§], Gunnar B Kristensen[¶¥], Eivind Hovig[†¥#] and Trond Stokke[*]

Addresses: [*]Department of Radiation Biology, Institute for Cancer Research, Norwegian Radium Hospital, Montebello, NO-0310 Oslo, Norway. [†]Department of Tumor Biology, Institute for Cancer Research, Norwegian Radium Hospital, Montebello, NO-0310 Oslo, Norway. [‡]Department of Oncology, Norwegian Radium Hospital, Montebello, NO-0310 Oslo, Norway. [§]Norwegian Microarray Consortium, Department of Molecular Bioscience, University of Oslo, NO-0316 Oslo, Norway. [¶]Department of Gynecologic Oncology, Norwegian Radium Hospital, Montebello, NO-0310 Oslo, Norway. [¥]Department of Medical Informatics, University of Oslo, NP-0316 Oslo, Norway. [#]Institute of Informatics, University of Oslo, NO-0316 Oslo, Norway.

Correspondence: Heidi Lyng. Email: heidi.lyng@rr-research.no

Published: 23 May 2008

Genome ***Biology*** 2008, **9:**R86 (doi:10.1186/gb-2008-9-5-r86)

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2008/9/5/R86

Received: 18 January 2008
Revised: 23 April 2008
Accepted: 23 May 2008

© 2008 Lyng *et al.*; licensee BioMed Central Ltd.
This is an open access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/2.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Abstract

Absolute tumor DNA copy numbers can currently be achieved only on a single gene basis by using fluorescence *in situ* hybridization (FISH). We present GeneCount, a method for genome-wide calculation of absolute copy numbers from clinical array comparative genomic hybridization data. The tumor cell fraction is reliably estimated in the model. Data consistent with FISH results are achieved. We demonstrate significant improvements over existing methods for exploring gene dosages and intratumor copy number heterogeneity in cancers.

Background

Array comparative genomic hybridization (aCGH) is widely used for genome-wide mapping of DNA copy number changes in malignant cells [1,2]. Genetic gains and losses impact gene expression levels, and thereby promote tumor growth and progression [3-5]. Numerous clinical studies have been performed to find tumor characteristics and to classify patients with respect to their prognosis based on the copy number changes [6,7]. The usefulness of the aCGH data is limited, however, because only relative and not absolute copy numbers are achieved, making the interpretation of the data and comparisons across experiments difficult. Absolute DNA copy numbers can be obtained only on a single gene basis by the use of fluorescence *in situ* hybridization (FISH). Development of genome-wide methods for this purpose would enable generation of universal gene copy number databases of individual diseases that could be utilized more widely, as is the goal of several public repositories like the Mitelman Database of Chromosome Aberrations in Cancer [8].

The relative values achieved in aCGH experiments are influenced by the total DNA content (ploidy) of the tumor cells, the proportion of normal cells in the sample, and the experimental bias, in addition to the DNA copy numbers. The values are presented as intensity ratios between tumor and normal DNA [2]. The data are normalized so that the ratio of 1.0 is the baseline for the analysis, and corresponds to two DNA copies in near diploid (2*n*) tumors. The copy number changes are

identified from the ratios deviating from the baseline, using statistical methods for ratio smoothing and breakpoint detection [9-12]. To assign an absolute copy number to each ratio level identified by the statistical analysis and thereby score genetic aberrations are, however, challenging. In aneuploid tumors with gross alterations in the DNA content, the baseline represents a copy number other than 2, like 3 or 4 in tri- or tetraploid tumors, or a non-integer value when the DNA content differs from n, $2n$, $3n$, ... mn [13]. The presence of normal cells within the sample and experimental bias reduce the ratio dynamics. Moreover, in many tumors, several subpopulations of malignant cells with different genetic characteristics exist, leading to intratumor heterogeneity in the DNA copy numbers [14-16] and increased complexity in the data. Unreliable results occur, therefore, when common ratio levels are used to score gains and losses in tumors with different ploidy and normal cell content.

The confounding effect caused by normal cells within tumor samples is recognized as a problem in aCGH analyses and has been handled by excluding low purity samples [17,18] or correcting the ratio levels based on histological examination of tumor sections [6]. The latter approach is not satisfactory because only the proportion of connective tissue surrounding the tumor parenchyma, and not the infiltrating immune cells, is precisely quantified. Moreover, the measurements cannot be performed on exactly the same tissue as used in the aCGH experiment and may, therefore, not be representative. A model including the CGH ratios, ploidy, and experimental bias has been proposed for estimation of absolute DNA copy numbers in tumor cell lines [19]. To our knowledge, no method exists that also considers the normal cell content and, thus, is suited for analyses of clinical tumor samples.

We here present a new model, GeneCount, where the proportion of normal cells is estimated and corrected for and possible intratumor heterogeneity in DNA copy numbers is considered. Inputs to our model are the DNA index (DI, where DI = 1/2·tumor ploidy), tumor cell fraction, experimental bias, and aCGH ratios. Predetermined measures of tumor ploidy, determined either by flow or image based cytometry, are needed. The tumor cell fraction can be determined by, for example, flow cytometry on the same part of the sample as used in the aCGH experiment. In cases of unknown normal cell content, the tumor cell fraction is estimated in the model. The experimental bias is determined from the X-chromosome ratio in aCGH experiments where male and female DNA is compared. Smoothed ratio levels from any existing statistical analysis tools for breakpoint detection can be used.

We show that the model enabled automatic and genome-wide calculation of DNA copy numbers from aCGH data of both hematopoietic and solid tumors. The feasibility of GeneCount was demonstrated by analysis of 94 lymphomas, for which the DNA index and tumor cell fraction had been determined by use of flow cytometry and an extensive exploration of DNA copy numbers had been performed by the use of FISH in previous studies [20-25]. The GeneCount results, both based on the pre-determined tumor cell fraction and that determined by the model, were compared with the FISH data of 362 genes with and without gains and losses, showing 97% consistency in both cases. In particular, we explored the copy numbers achieved in the t(14;18) translocated chromosomal region involving *BCL2*. We further demonstrated the potential of GeneCount in analysis of solid tumors without pre-determined tumor cell fractions by relating the copy number of selected genes in 93 cervical cancers to gene expression and treatment outcome. By use of GeneCount we obtained a higher sensitivity in detecting cervix tumors with copy number changes than was obtained in analysis based directly on the ratio levels. Finally, we identified intratumor heterogeneity of DNA copy numbers in the lymphomas and cervical cancers, and showed how this information could be used to draw conclusions about the evolution of the genetic aberrations in the tumors. GeneCount was implemented in a software package to be used downstream of statistical methods for breakpoint detection, and results based on both the GLAD and CGH-Explorer packages are presented [9,11]. We supply our method through the open-source and free web-based database BioArray Software Environment (BASE) [26].

Results

Basis of GeneCount

Our model utilizes the fact that the normalized aCGH ratio increases with increasing DNA copy number in a stepwise manner, where the step size is dependent on the DI, the tumor cell fraction, and the experimental bias (Figure 1). In near diploid tumors (DI = 1) without a contribution from normal cells or affected by experimental bias, an increment of 1 in the copy number increases the ratio by a value of 0.5, leading to a normalized ratio of 0.5, 1, 1.5, 2, and so on (-1, 0, 0.69, 1 on a $\log_2$ scale) for a copy number of 1, 2, 3, and 4, respectively (see Equation 2 in Materials and methods). The corresponding increase in tetraploid tumors (DI = 2) is 0.25, whereas an increase between 0.25 and 0.5 occurs in tumors with a DI between 1 and 2. Baseline, at a $\log_2$ ratio of 0, corresponds to 2, 3, and 4 DNA copies in near diploid (Figure 1a), triploid, and tetraploid (Figure 1b) tumors, respectively. For DIs between 1 and 1.5 or between 1.5 and 2, baseline represents a copy number between 2 and 3 (Figure 1c) or between 3 and 4. The presence of normal cells within the tumor sample reduces the increase in aCGH ratio with incremental copy number (Equation 3), as can be seen when comparing the ratios of two near diploid lymphomas with different tumor cell fractions (Figures 1a,d). Using common ratio levels for scoring gains and losses in tumors like those presented in Figures 1a-d leads, therefore, to different results with respect to copy number changes.

A further reduction in the ratio dynamics occurs due to experimental bias (Equation 4). The bias, as represented by the

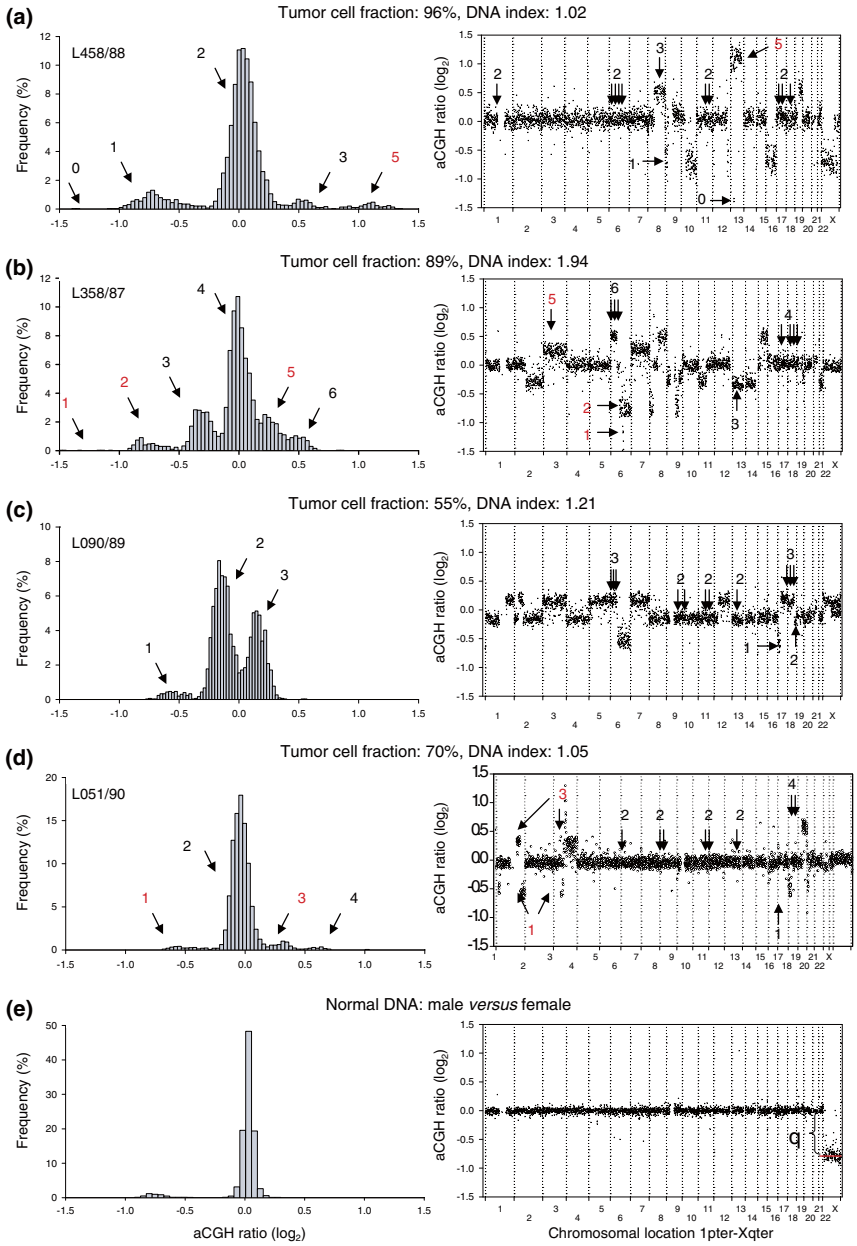


Figure 1 *(see legend on following page)*

Figure 1 *(see previous page)*
Illustration of the stepwise increase in aCGH ratios with increasing DNA copy number. Frequency histograms (% array probes) of aCGH ratios (left panels) and plot of aCGH ratio versus chromosomal location (right panels) are shown for a lymphoma with a DNA index (*DI*) of **(a)** 1.02, **(b)** 1.94, **(c)** 1.21, and **(d)** 1.05, and **(e)** for normal DNA comparing male and female. The tumor cell fraction, measured by flow cytometry, is indicated for each tumor. DNA copy numbers estimated by GeneCount are marked; those in black were consistent with FISH data, whereas those in red have not been subjected to FISH measurements in the specific tumors shown. The arrows in the right panels point to the locations of the FISH probes. At a *DI* close to 1 and 2 (a,b,d,e) the ratio distribution shows a major peak at a median $\log_2$ value of approximately zero, representing the most frequent DNA copy numbers of 2 and 4, respectively. At a *DI* of 1.21 (c) the baseline at a $\log_2$ ratio of 0 represents a number between 2 and 3 DNA copies. Note the smaller increase in the ratios with increasing DNA copy number at a tumor cell fraction of 70% (d) than of 96% (a). In (e), determination of the dynamic factor, *q*, as the absolute value of the X-chromosome $\log_2$ ratio level is indicated.

dynamic factor, *q*, can be determined from control experiments, where normal DNA from males and females is cohybridized (Figure 1e). Theoretically, the X-chromosome ratio is 0.5 (-1 on a $\log_2$ scale), but the experimental bias reduces the ratio dynamics, leading to a ratio level closer to zero. The absolute value of the $\log_2$-transformed ratio level was used as a measure of *q* (Figure 1e). This value differed little among the slide series used here, ranging from 0.75-0.85 with a mean ± standard deviation of 0.80 ± 0.04 based on 8 control experiments. A *q*-value of 0.8 and range of 0.7-0.9 was used in the GeneCount calculations in the cases of known and unknown tumor cell fraction, respectively.

To enable automatic calculation of the copy number associated with each array probe, we implemented GeneCount in a program to be run on top of statistical analysis packages for aCGH ratio smoothing and breakpoint detection (Additional data file 1). A separate algorithm was developed for samples with unknown tumor cell fraction, where the fraction was estimated based on two ratio levels and *DI* (panel B in Additional data file 1), as described in Materials and methods. One decimal was included in the calculated DNA copy numbers when evaluating the results in comparison with FISH data. Otherwise, the numbers were rounded off to the nearest integer values.

GeneCount copy numbers in comparison with FISH data

We compared the GeneCount results of 94 lymphomas with previously published FISH data from the same tumors [20-25]. The FISH probes were located at chromosomal regions with frequent copy number changes (Figure 1 and Additional data file 2), and copy numbers within the range of 0-8 had been measured. The *DI*s, ranging from 0.95-2.23, and the tumor cell fractions, ranging from 27% to 98%, were used as inputs to GeneCount, together with the smoothed aCGH ratios from the GLAD and CGH-Explorer packages. CGH-Explorer applied a more extensive ratio smoothing than GLAD, and this led occasionally to differences in the ratio levels and breakpoint detection between the two programs.

GeneCount with known tumor cell fraction

In most cases, we found an excellent agreement between the DNA copy number determined by GeneCount and FISH,

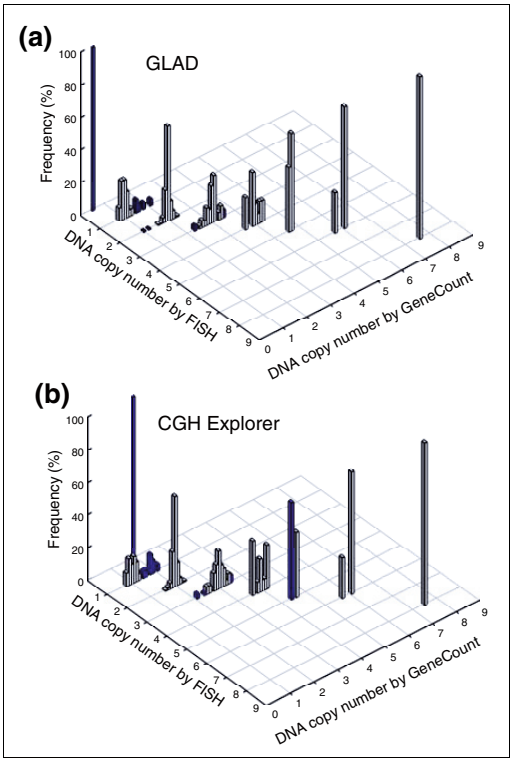


Figure 2
GeneCount calculations with known tumor cell fraction. DNA copy number calculated by GeneCount is plotted against the corresponding FISH result for 9 genes in 94 lymphomas. The smoothed aCGH ratios from **(a)** GLAD and **(b)** CGH-explorer, a *q*-value of 0.8, and a *DI* and tumor cell fraction determined by flow cytometry were inputs to GeneCount. Grey and blue columns represent GeneCount results that were consistent and inconsistent with the FISH data, respectively, after rounding off the GeneCount number to the nearest integer value. Frequency distributions are shown for each copy number, containing 1, 25, 246, 66, 15, 5, 4, and 1 value at a FISH copy number of 0, 1, 2, 3, 4, 5, 6, and 8, respectively.

regardless of whether GLAD or CGH-Explorer was used for breakpoint detection (Figure 2). The correlation between the data sets was considerably better than when the ratio levels were used in the comparison (Additional data file 3). Based on GLAD, 350 out of 362 GeneCount values were consistent with the FISH data (97%), whereas the corresponding number based on CGH-Explorer was 340 out of 362 (94%) (Figure 2). The few discrepancies between the GeneCount and FISH results occurred mainly for two reasons. First, GLAD and/or CGH-Explorer failed to detect the ratio change of some of the genes that had a copy number change by FISH (panel A in Additional data file 4). Second, the ratio level, and therefore the copy number, was inaccurately determined for some aberrations involving only a few array probes (panel B in Additional data file 4). This was primarily the case for aberrations with less than three probes, like the homozygote deletion involving two probes that covered *RB1* in one of the tumors (FISH copy number of 0 in Figure 2 and panel B in Additional data file 4). The discrepancies between the GeneCount and FISH data were related, therefore, to the software used for breakpoint detection and not due to errors in the GeneCount algorithm.

GeneCount with unknown tumor cell fraction

The tumor cell fraction could be estimated for 55 and 43 out of 94 lymphomas based on GLAD and CGH-Explorer, respectively. The remaining tumors lacked aberrations or two different ratio levels that could be used for the estimation (Materials and methods). The estimated tumor cell fractions correlated significantly with those measured by flow cytometry (Figure 3). Moreover, the estimates had a coefficient of variance (CV) of less than 11% (Figure 3), and were therefore fairly stable. The mean *q*-value determined in the calculation differed little across the tumors, ranging from 0.73-0.84 (GLAD) and 0.74-0.82 (CGH-Explorer) (data not shown).

The consistency between the GeneCount and FISH data (Figure 4) was comparable to when the known tumor cell fraction was used (Figure 2) and much better than when the ratio levels and FISH data were compared (Additional data file 3). Based on GLAD, 218 out of 231 DNA copy numbers were in agreement with the FISH data (94%), whereas the corresponding numbers based on CGH-Explorer were 173 out of 179 (97%) (Figure 4). Most differences between the GeneCount and FISH results occurred for the same reasons as when the known tumor cell fraction was used (Additional data file 4). Additionally, a discrepancy was seen for some of the highest copy numbers based on GLAD (Figure 4), due to a large discrepancy between the estimated and measured tumor cell fraction in one of the cases (Figure 3a).

DNA copy numbers in translocated chromosomal regions

The relationship between the GeneCount estimates and FISH data in translocated chromosomal regions was explored by using *BCL2*, which is involved in the translocation t(14;18) in lymphomas, as an example. The aCGH probe covering *BCL2* is located telomeric of the breakpoint. The aCGH data and GeneCount results of *BCL2* were therefore not affected by the translocation. For FISH analysis, we selected a *BCL2* probe covering the breakpoint. The probe signal was split in tumors with translocation, leading to a signal from both der(14)t(14;18) and der(18)t(14;18), although *BCL2* is located on the former chromosome. The FISH signal was therefore higher than the actual *BCL2* copy number, and differed from the GeneCount result in all 38 tumors with translocation (Figure 5a). After recalculating the FISH copy numbers as described [22], the consistency in the data was excellent, except in one case at a corrected FISH value of five copies (Figure 5b). This discrepancy was due to failure of GLAD and

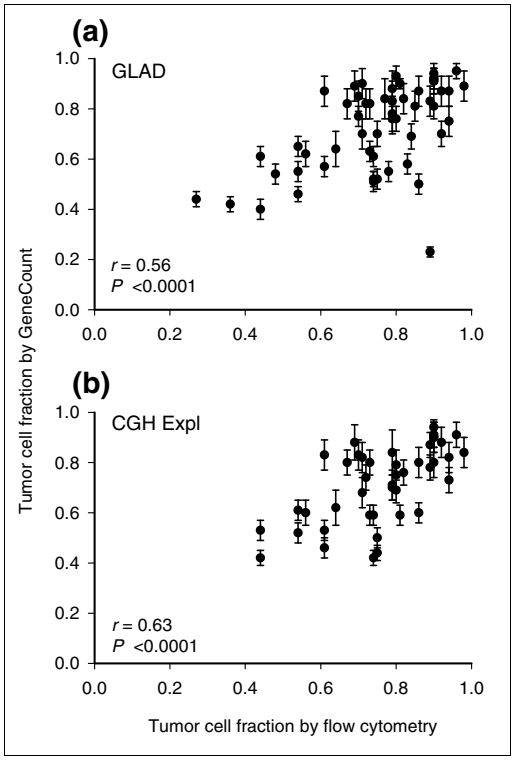


Figure 3
GeneCount estimations of tumor cell fraction. Tumor cell fraction of lymphomas estimated by GeneCount is plotted against tumor cell fraction measured by flow cytometry. Each point represents mean ± standard deviation based on the values achieved for *q* within the range 0.7-0.9. The smoothed aCGH ratios from **(a)** GLAD and **(b)** CGH-explorer, the *q* range 0.7-0.8, and a *DI* determined by flow cytometry were inputs to GeneCount. The calculations were based on 55 (a) and 43 (b) tumors for which suitable ratio levels for the calculations existed. Correlation coefficients and *P*-values from Pearson product moment correlation analyses are indicated.

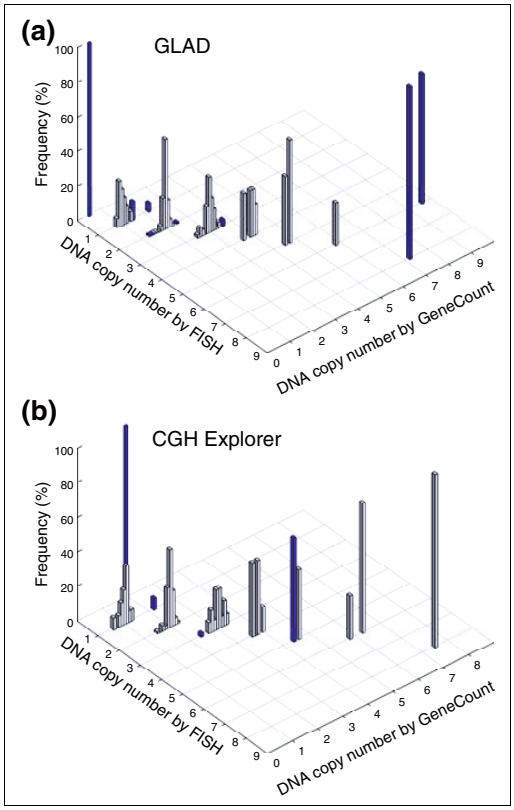


Figure 4
GeneCount estimations with unknown tumor cell fraction. DNA copy number calculated by GeneCount, using a *q*-value within the range 0.7-0.9, a *DI* determined by flow cytomery, and the tumor cell fraction estimated by GeneCount in Figure 3, is plotted against the corresponding FISH result for 9 genes in **(a)** 55 and **(b)** 43 lymphomas. The smoothed array CGH ratio derived from GLAD and CGH-explorer was used in (a) and (b), respectively. Grey and blue columns represent GeneCount results that were consistent and inconsistent with the FISH data, respectively, after rounding off the GeneCount value. Frequency distributions are shown for each copy number, containing 1, 19, 134, 56, 11, 5, 4, and 1 value at a FISH copy number of 0, 1, 2, 3, 4, 5, 6, and 8, respectively, based on GLAD. The corresponding numbers based on CGH Explorer were 1, 15, 98, 48, 7, 5, 4, and 1.

CGH-Explorer in detecting a narrow amplicon involving *BCL2* (panel C in Additional data file 4).

GeneCount analysis of solid tumors

The feasibility of our method for analysis of solid tumors without information of tumor cell fraction was explored in 99 cervical cancers, for which the *DI* ranged from 1.00-3.16. The tumor cell fraction could be estimated for 93 and 89 tumors based on GLAD and CGH-Explorer, respectively, fulfilling the requirements for this estimation (Materials and methods). The tumor cell fractions were poorly correlated with the values determined by analysis of histological sections (Additional data file 5). In most cases, the histology result was higher than the GeneCount estimate, probably because immune cells infiltrating the tumor parenchyma were not properly quantified by the histological examination. In a few cases, however, the histology result was higher, probably reflecting that different parts of the sample were used in the aCGH and histology analyses. The tumors for which the tumor cell fraction could be estimated by GeneCount were included in the further analyses.

A higher number of genetic aberrations were generally found in the cervical cancers than in the lymphomas. High level amplifications with more than 2.5-fold increases in gene dosage (that is, copy number, N, relative to total DNA content given by two times the DNA index ($N/(2.DI)$), were found in about half of the tumors and most frequently on chromosomes 5p and 11q. GeneCount analysis showed copy numbers within the range of 5-80 in these regions, which were often surrounded by gains at lower levels.

The GeneCount results were compared with the outcomes of existing analysis methods, where gains and losses were scored from the smoothed ratio levels and breakpoints obtained by GLAD and CGH-Explorer. The $\log_2$ transformed ratio levels of ± 0.2 (that is, approximately two times the ratio standard deviation (Additional data file 6)) were applied as cut-off levels for scoring aberrations. We selected genes that were shown to be affected by gains and losses in previous studies on a subgroup of the patients [27]. Some of the genes showed only a small variation in the aCGH ratios, often within the level of ± 0.2, and only a few tumors with aberrations were identified (Figure 6a and panel A in Additional data file 7). A higher number of patients with changes in gene copy numbers and in the corresponding gene dosages were identified with GeneCount, using the cut-off levels of ± 0.2 for scoring gene dosage changes (Figure 6b,c and panels B and C in Additional data file 7). The gene dosage correlated significantly with gene expression (Figure 6c and panel C in Additional data file 7), making the copy number changes determined by GeneCount plausible.

The copy number changes of *MRPS23* have previously been shown to correlate with survival probability [27]. Survival analysis based on the GeneCount data of *MRPS23* identified more patients with poor outcome than the corresponding analysis based on ratio levels (Figure 6d,e). Hence, 15 high risk patients were identified based on the GeneCount results, whereas only 5 patients were classified with high risk based on the ratio levels. Nine of the ten patients that were not identified based on ratio levels (blue curve in Figure 6d) had aneuploid tumors with a DNA index ranging from 1.10-1.92. The remaining diploid tumor had a relatively low tumor cell fraction of 23%.

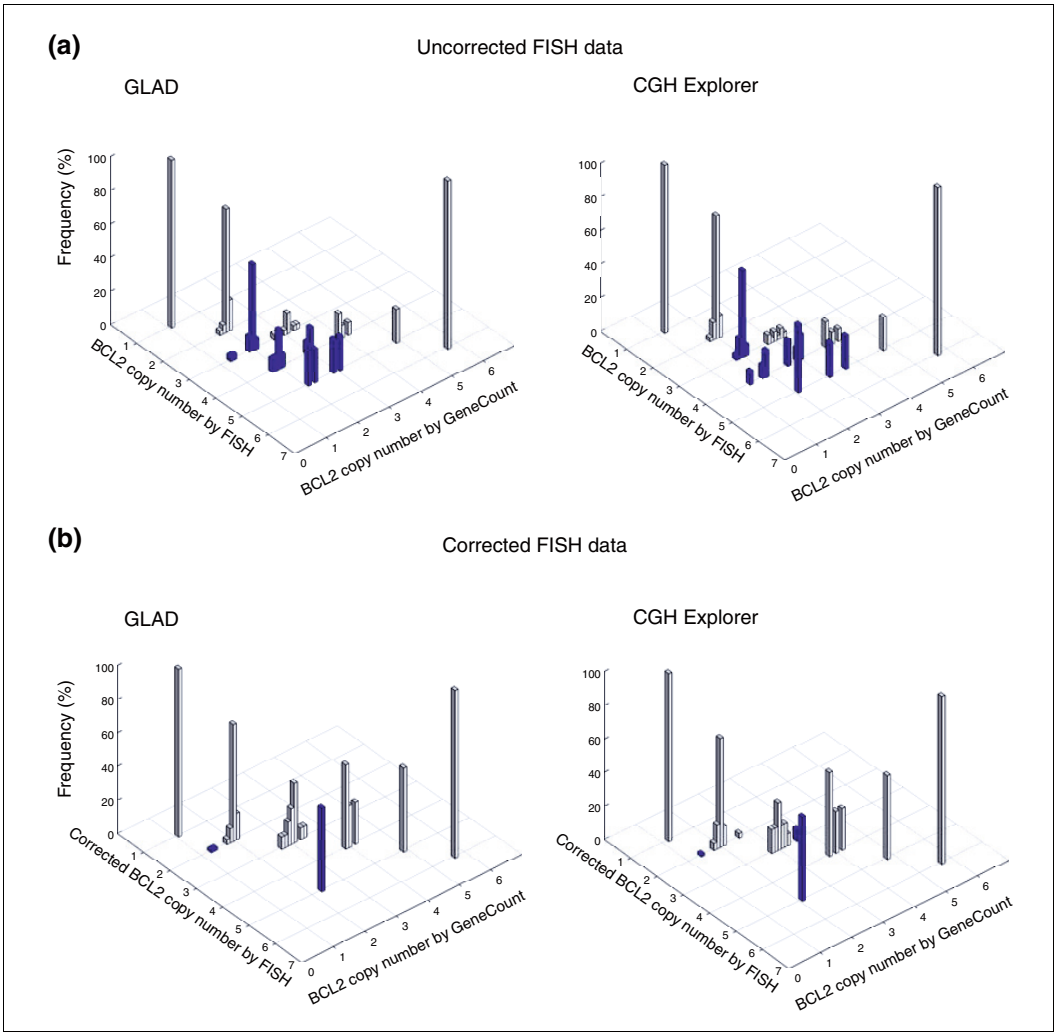


Figure 5
GeneCount estimations in the t(14;18) translocated region involving *BCL2*. *BCL2* copy number estimated by GeneCount, using a *q*-value of 0.8 and a *DI* and tumor cell fraction determined by flow cytometry, is plotted against the corresponding FISH result in 94 lymphomas. The smoothed array CGH ratios derived from GLAD and CGH-explorer were used in the left and right panels, respectively. Grey and blue columns represent GeneCount calculations that were consistent and inconsistent with the FISH measurements, respectively, after rounding off the GeneCount value. **(a)** Uncorrected FISH data are plotted; **(b)** these data were corrected as described in [22]. Frequency distributions are shown for each copy number, containing 1, 38, 33, 13, 5, and 1 value for a red spot FISH copy number of 1, 2, 3, 4, 5, and 6. The corresponding number of measurements for the corrected FISH data of 1, 2, 3, 4, 5, and 6 were 1, 69, 14, 4, 2 and 1.

Intratumor heterogeneity in DNA copy numbers

Some tumors had genome regions for which the aCGH ratio was clearly different from that corresponding to an integer copy number. This probably reflected intratumor heterogeneity in the DNA copy numbers, that is, the existence of subpopulations with copy number changes that are not common for all tumor cells in the sample. The common aberrations can thus be considered homogeneous. Lymphomas and cervical cancers with heterogeneous DNA regions had ratio levels that fell in between, and were significantly different from, those

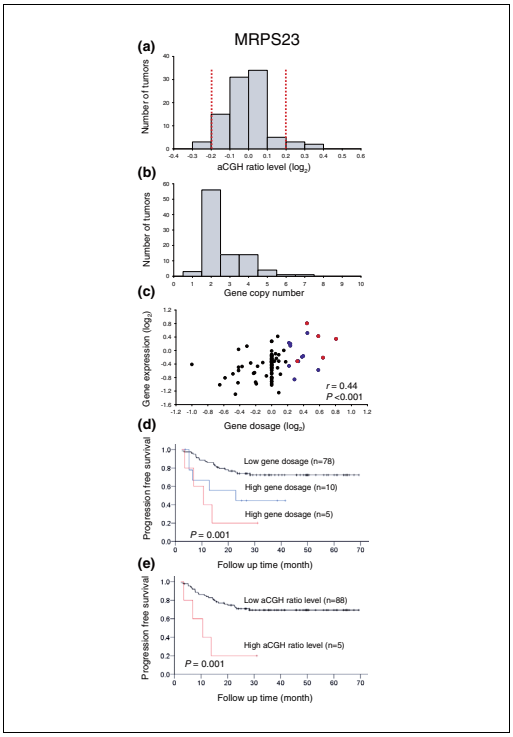


Figure 6

GeneCount analyses in cervical cancers. **(a)** Frequency histogram (number of tumors) of smoothed aCGH ratios (GLAD) for *MRPS23* (BAC clone ID RP11-19F16). Dotted lines indicate the cut off ratio levels of ± 0.2, identifying 5 tumors with genetic gain and 3 tumors with loss. **(b)** Frequency histogram (number of tumors) of *MRPS23* copy number calculated by GeneCount. The GLAD ratio levels, the *DI* measured by flow cytometry, and the tumor cell fraction estimated by GeneCount were used in the calculation. Similar results were achieved based on the CGH-Explorer ratio levels. **(c)** Plot of gene expressions against gene dosage; that is, the *MRPS23* copy number divided by the total DNA content ($N/(2 \cdot DI)$). Increased gene dosage with more than 15% of the total DNA content ($\log_2$ transformed gene dosage of at least 0.2) were seen in 15 tumors (red and blue symbols). Red symbols represent the five tumors with gain in (a), whereas blue symbols represent the remaining ten tumors with increased gene dosage that were not identified in (a). The correlation coefficient and *P*-value from Pearson product moment correlation analysis are indicated. **(d)** Kaplan Meier analysis based on GeneCount results for *MRPS23*. Plots of the survival probability are shown for 5 patients with high gene dosage in (c), who also had gain in (a) (red line), 10 patients with high gene dosage in (c) and without gain in (a) (blue line), and 78 patients with low gene dosage in (c). **(e)** Kaplan Meier analysis based on the *MRPS23* ratio levels. The survival probability of 5 patients with gain in (a) (red line) and 88 patients without gain in (a) (black line) is plotted. Only five high risk patients were identified in (e), whereas ten more patients were identified by GeneCount in (d). *P*-value in log-rank test is indicated in (d,e). Panels (a,b,d) are based on 93 tumors, for which the tumor cell fraction could be estimated by GeneCount. Panel (c) is based on 89 of these tumors, for which both DNA copy number and gene expression were available.

corresponding to integer values (Figure 7). The actual ratio level reflected the proportion of cells with that aberration (Equation 1).

Nineteen (20%) lymphomas and 44 (50%) cervical cancers had one or more heterogeneous DNA regions with copy numbers 1&2, 2&3, or 3&4 (Additional data files 8 and 9). Reliable detection of heterogeneity required tumor cell fractions above 24% (Additional data file 10) and 5 out of 93 cervical cancers were therefore excluded from this analysis. Lymphoma L309/89 (Figure 7) had previously been identified as heterogeneous by FISH, showing one population with three and another with four copies of *MYC* and centromeres 8 and 17 [20]. Moreover, several of the heterogeneous aberrations in the cervical cancers, such as loss on chromosome 4 and X and gain on 11q and 17 in C005/01, loss on 6q and gain of 11q in C006/01, and loss on 4 in C023/01, were similar to those detected earlier by conventional CGH [14]. The previous study was, however, based on a different set of biopsies, which probably explains the lack of consistency for some of the tumors.

In a few of the heterogeneous tumors, two different ratio levels were identified between one and two copies (Figure 8 and Additional data file 11). Thus, it appeared that the corresponding aberrations were present in different fractions of the tumor cell population. Lymphoma L008/92 had two intermediate ratio levels between one and two copies, corresponding to 70% and 30% of the tumor cells (Figure 8b, blue and red ratios, respectively), leading to the possible tumor evolutionary schemes depicted in Figure 8c. As the sum of the two fractions did not exceed 100%, the heterogeneous aberrations may be found in non-overlapping subpopulations of the tumor, where the subpopulations have evolved differently from a predicted common population containing the homogenous aberrations (parallel sequence). A serial sequence, where the populations have evolved in a linear manner from a common population, was also possible. In C024/01, however, the heterogeneous ratio levels corresponded to 78% and 44% of the tumor cells, and a serial sequence was the only one suggested (panel C in Additional data file 11).

Discussion

We have shown that GeneCount is a reliable method for genome-wide calculation of DNA copy numbers in clinical tumor samples. Such data are biologically interesting in themselves but may also lead to improved prediction of treatment outcome and aid in the identification of novel tumor suppressors and oncogenes. We applied the method to lymphomas, for which accurate measures of tumor cell fraction and DNA copy numbers have been obtained by other techniques that could be compared with the GeneCount results. We further used the method on cervical cancers, for which tumor cell fractions representative of the aCGH data are more difficult to achieve by a separate technique. The

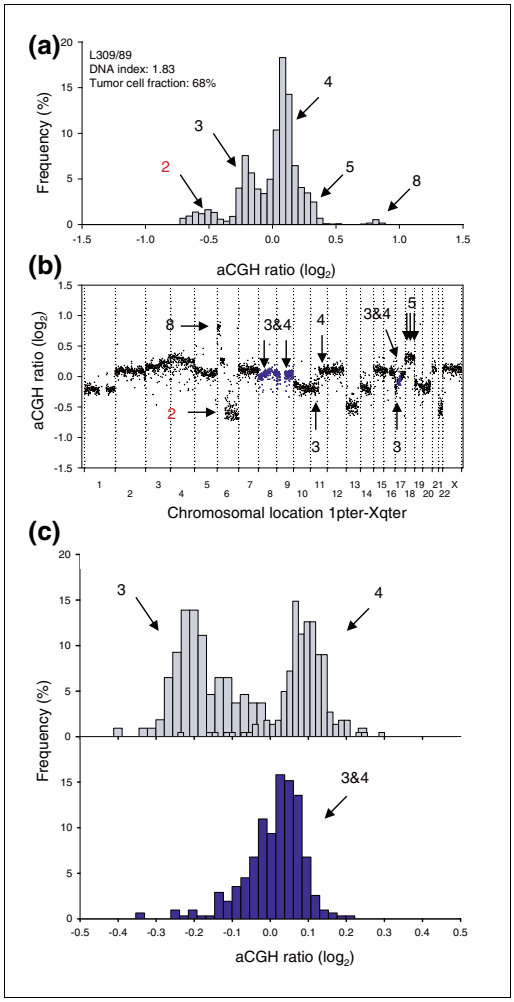


Figure 7
GeneCount identification of DNA copy number heterogeneity within tumors. **(a)** Frequency histogram (% array probes) of aCGH ratios in a heterogeneous lymphoma, including data for the entire genome. **(b)** aCGH ratios are plotted against chromosomal location, showing the heterogeneous regions on chromosomes 8, 9, and 17 with a DNA copy number of 3&4 in blue. **(c)** Frequency histogram (% array probes) of aCGH ratios for two homogeneous DNA regions with a copy number of 3&4 (upper panel) and the heterogeneous region depicted in (b) with a copy number of 3&4 (lower panel). The ratio distributions of copy number 3, 4, and 3&4 were significantly different ($p < 0.001$, ANOVA). DNA copy numbers estimated by GeneCount from the *DI* and tumor cell fractions measured by flow cytometry are marked; those in black were consistent with FISH experiments, whereas those in red have not been subjected to FISH measurements in the specific tumors shown. The arrows in (b) point to the locations of the FISH probes. Note that the 3&4 copy number of the heterogeneous region has been confirmed with FISH.

GeneCount model is simple, due to the use of normal cells with two DNA copies throughout the genome as a reference sample. Moreover, the estimated copy numbers are restricted to positive integers, increasing the robustness of the method. A requirement for achieving the absolute quantification format is the use of pre-determined tumor ploidies, whereas the tumor cell fractions, if not known, and experimental bias can be estimated from the aCGH data.

The experimental bias is mainly caused by signals from unsuppressed repetitive sequences and nonspecific hybridization [2]. The bias influences the test and reference sample equally and independently of the DNA copy number, since it is generated by sequences distributed throughout the genome. The bias could, therefore, be summed up in an array specific factor, q, representing the dynamics of the log-transformed ratios. Mohapatra *et al.* [19] included the bias as a constant factor affecting the absolute, rather than the log-transformed, ratios in their model for pure tumor cells. Our approach seems justified because the noise (width) of the log-transformed ratios was independent of the ratios and, therefore, of the DNA copy numbers (Additional data file 6). We allowed for a small variation in q when calculating the tumor cell fraction to account for minor differences in the bias across the tumors. This q-value, optimized for each tumor, was highly similar to the mean q determined from control experiments, indicating that the bias was stable across experiments. Moreover, the discrepancies between the GeneCount and FISH results were related to the specific genetic aberration involved and, therefore, to the breakpoint detection algorithm, rather than to possible uncertainties in q. Recent developments in array CGH technology, utilizing oligonucleotides rather than bacterial artificial chromosome (BAC) clones, led to improved ratio dynamics and reduction in the experimental bias due to less repetitive sequences [28]. Ongoing work in our laboratory shows that by using oligoarrays, GeneCount can be applied with a q value close to 1.

Inclusion of the tumor cell fraction is a prerequisite for the calculation of absolute DNA copy numbers in clinical tumor samples. The lymphoma data were based on single cell suspensions made from the entire lymph nodes. A tumor cell fraction representative of the lymph node could, therefore, be determined with high accuracy by a separate technique like flow cytometry. In solid tumors such as cervical cancers, the normal cells consist of stroma, which is highly heterogeneously distributed within the tissue, and immune cells, which infiltrate the tumor parenchyma. A measure of the tumor cell fraction achieved by, for example, histological examination, which is based on a part of the sample different from that used for the aCGH experiment and/or fails to quantify the proportion of immune cells accurately, is, therefore, not precise enough for the calculation of DNA copy numbers. Histology data may, however, be useful for preselecting tumor enriched samples for the aCGH analysis. Fairly stable estimates of the tumor cell fraction, consistent with the values

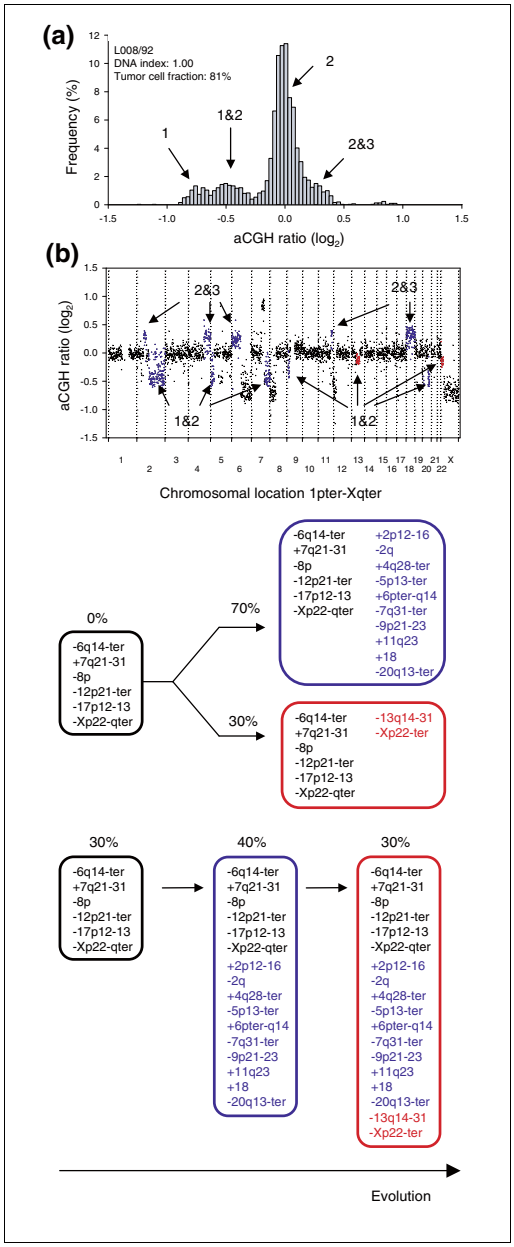


Figure 8

Figure 8
Evolutionary sequences of subpopulations in heterogeneous tumors. **(a)** Frequency histogram (% array probes) of aCGH ratios in a heterogeneous lymphoma is shown, including data for the entire genome. **(b)** The aCGH ratios are plotted against chromosomal location. The heterogeneous regions on chromosomes 2q, 5p, 7q, 9p, 13q, 20q, and Xp with a DNA copy number of 1&2 and on chromosomes 2p, 4q, 6p, 11q, and 18 with a DNA copy number of 2&3 are shown in blue and red. The blue and red colors represent aberrations that are present in different fractions of the tumor cells; 70% and 30%, respectively. The heterogeneous aberrations are listed in Additional data file 8 except those with a copy number of 2&3, since the lack of 3 DNA copies in this tumor prevented statistical analysis to identify 2&3 heterogeneity. **(c)** Schematic diagram of two possible evolutionary sequences for the aberrations, one parallel and one serial sequence, are shown. The blue and red circles represent the blue and red aberrations in (b). The percentages indicate the fractions of tumor cells with the listed aberrations, as calculated by GeneCount, showing that the aberrations in blue and red are present in 70% and 30% of the tumor cells, respectively.

measured by flow cytometry, were achieved by the use of GeneCount. The estimates led to DNA copy numbers in agreement with the FISH data, suggesting that the accuracy of the tumor cell fractions was sufficient for reliable data analysis. Selection of appropriate ratio levels for the estimation was crucial for achieving this accuracy. We required that the tumors had at least two aberrations with different copy numbers and with more than ten array probes each to reduce errors caused by poorly defined ratio levels and breakpoints. Moreover, only ratio levels deviating more than 0.15 ($\log_2$ scale) from the baseline were selected, implying that tumor fractions higher than 24% (diploid) and 36% (tetraploid) were needed when copy numbers were changed to 3 or 5 copies, respectively (Additional data file 12).

The few discrepancies between the GeneCount and FISH data were not related to our model, but rather to the ability of the statistical methods to detect some of the aberrations. Hence, the consistency between the GeneCount and FISH results was similar to the reliability of GLAD in detecting breakpoints in simulated data [9]. The highest accuracy of the GeneCount results was obtained for well defined aberrant regions containing at least three array probes. In these cases a ratio level representative of the corresponding copy number was achieved and the probability of detecting the aberration was high. The increased uncertainty in the results of narrow aberrations implies that they should be confirmed by a separate technique like FISH. Moreover, to ensure sufficient ratio dynamics and, therefore, a high probability of breakpoint detection, a tumor cell fraction higher than a certain value, which depends on the experimental noise and tumor ploidy, is needed. With the noise of our experiments (Additional data file 6), a tumor cell fraction above 23% in diploid, and somewhat higher in hyperdiploid cases, enabled separation of an aberration with more than three array probes (Additional data file 13). This fraction also enabled detection of heterogeneous DNA copy numbers involving more than ten array probes (Additional data file 10). In experiments with more

noise, caused by, for example, poor DNA quality, higher tumor cell fractions are required. In comparison, at least 50% tumor cells is suggested for optimal detection of gains and losses by conventional CGH [29].

The DNA copy number of genes involved in translocations cannot be directly assessed by FISH when a probe covering the breakage region is used, because signals from both the original chromosomes are detected in the translocated derivatives. Correction of the probe signal to achieve the true copy number requires knowledge of the breakpoint and genes involved in the translocation. Reliable FISH analysis in solid tumors, where the translocations are not well identified and may occur throughout the genome [30] is, therefore, particularly challenging. By aCGH, the probe signal is measured independently of the actual genome organization of the DNA covered by the probe. Hence, in the case of balanced translocations, a correct result will be obtained even if the probe covers the breakpoint. If the probe is located at the start or end of an amplified or deleted region (unbalanced translocation), the aCGH ratios of the adjacent probes ensure that the correct copy number is calculated. Our model therefore provides a novel method for assessment of copy numbers both in balanced and unbalanced translocated regions and without knowing that the translocation exists.

Current methods for analysis of aCGH data generally score genetic gains and losses based on ratio levels [31-36]. The breakpoints in individual tumors can be detected with high accuracy by use of statistical algorithms like GLAD and CGH-Explorer. However, the existing downstream analyses, using common ratio levels for scoring aberrations across tumors, fail to identify gains and losses in cases of high ploidy and normal cell content. By the use of GeneCount, the ratio levels are replaced with the absolute copy numbers relative to the total DNA content as measures of gene dosage, which can be compared across tumors regardless of ploidy and normal cell content. Hence, copy number changes that were not detected by analysis based on ratio levels, but showed significant correlation with gene expression, were found in cervical cancers, suggesting that improved results were achieved. Moreover, many patients with poor outcome that had *MRPS23* gain by GeneCount had no gain based on ratio levels. In the latter case, the gain was masked by high content of normal cells or high ploidy, showing that GeneCount is more sensitive in detecting patients with genetic aberrations. The finding further demonstrates that GeneCount applies well to solid tumors for which the tumor cell fraction is generally unknown and must be estimated by the method. Advances in current statistical analysis methods may utilize adjustable ratio levels for scoring gains and losses, optimizing the cut-off ratios for each tumor based on a mathematical evaluation of the ratio dynamics. Such methods may account for varying ploidy and normal cell content across diploid, triploid, and tetraploid tumors. However, the strategy is not useful for tumors with an intermediate ploidy like 1.25 (Figure 1c). In contrast, the absolute DNA copy number relative to the total DNA content, or gene dosage, is comparable also across such tumors.

We also showed that GeneCount can provide genome-wide and high resolution information of intratumor heterogeneity in the DNA copy numbers. Such heterogeneity has previously been detected only on a single gene basis by FISH or at low resolution by conventional CGH analyses [14,15,20,37,38], probably reflecting a high genomic instability [39]. Detection of heterogeneity involving two DNA copies by the use of FISH is challenging, since the heterogeneous tumor population is difficult to distinguish from normal cells. The probability to detect heterogeneity with GeneCount depends on the fraction of tumor cells with the heterogeneous aberration. Obviously, the probability is largest at a fraction of 50%, but fractions higher than 70% and lower than 30% were also identified. Heterogeneity in low copy numbers, like 1&2 and 2&3, are more easily detected, since the separation between the log-transformed ratio levels are larger. At higher copy numbers, the possibility to detect heterogeneity decreases, depending on the ploidy and normal tissue content. However, we also identified heterogeneous regions with copy number 3&4 in several tumors and 4&5 in one tumor. Finally, the probability to detect heterogeneity also depends on the proportion of the genome that is affected. In our data severe heterogeneity affecting up to 40% of the genome could be analyzed with GeneCount (C002/01; Additional data file 9). With an increasingly larger part of the genome affected, difficulties in finding breakpoints and even homogeneous aberrations eventually occur, leading to unreliable results regardless of analysis method.

The heterogeneity data led to insight into the evolutionary sequence of the copy number changes. The homogeneous aberrations had probably occurred prior to the heterogeneous ones [14]. Moreover, in cases where the heterogeneous aberrations appeared to be present in different fractions of the tumor cell population, these aberrations could be ordered chronologically in a serial and/or parallel sequence. It was not always possible to identify the correct sequence among the proposed ones, as could be done by comparing data for several biopsies from the same tumor [14]. However, identification of the heterogeneous as well as the homogeneous aberrations suggests a further possible investigation of the exact combination of aberrations in each subpopulation, employing, for example, triple-color FISH with one probe for a homogeneous aberrant region and two for the heterogeneous ones.

In the heterogeneity analysis we assumed that the ploidy was the same for all subpopulations of malignant cells. This assumption was justified because no cases were observed with two aneuploid populations by flow cytometry. A possible difference in the ploidy of two aneuploid populations within a tumor was therefore probably smaller than 10%, leading to less than 10% uncertainty in the copy numbers calculated by

GeneCount (Equation 4; data not shown). The same uncertainty also applied to near diploid and heterogeneous cervical cancers. These tumors often showed a broad G_1 peak with a CV up to 10% by flow cytometry, probably reflecting the existence of several subpopulations with ploidy within the range of 1.0-1.1. Moreover, few or no light chain positive cells were observed in the diploid population of the aneuploid lymphomas, suggesting that the diploid population contained primarily normal cells. It is possible, however, that the diploid population of the aneuploid cervical cancers contained malignant cells, as we have previously shown for aneuploid colorectal cancers [40]. This might have led to larger uncertainties in the heterogeneous copy numbers due to the use of an erroneous DNA index of the diploid population. The data of such tumors can be improved by sorting the diploid and aneuploid fractions by flow cytometry [40] for separate aCGH and GeneCount analyses.

Conclusion

GeneCount provides reliable DNA copy numbers, both when based on the tumor cell fractions determined by flow cytometry and those estimated by the method. Accurate data are also achieved in translocated chromosomal regions, as demonstrated for the t(14;18) translocation involving *BCL2*. Our method is the only one to provide genome-wide information of absolute DNA copy numbers. Moreover, the method represents a significant improvement compared to existing methods in the study of gene dosages and intratumor copy number heterogeneities. The robustness of GeneCount implies that the method can be utilized widely in the genomic exploration of both hematopoietic and solid tumors, addressing DNA copy number aspects in a reliable manner, regardless of possible translocations. This may lead to improved assays for disease classification and outcome prediction and aid the identification of efficient targets for new cancer therapies.

Materials and methods

Tumor samples, DNA index, and tumor cell fraction

Samples from 94 patients with B-cell non-Hodgkin's lymphoma and 99 patients with squamous cell carcinoma of the uterine cervix were analyzed. We used fresh frozen lymphoma cell suspensions for which the tumor subtype, stage, patient treatment, and follow-up have been presented previously [25]. The cervical cancers were of FIGO (Fédération Internationale des Gynaecologistes et Obstetristes) stage 1b-4b, treated with radiotherapy. Tumor biopsies taken before the start of treatment were used.

The *DI* of the lymphomas and cervical cancers and the tumor cell fraction of the lymphomas were determined by use of flow cytometry, and most of these data have been published earlier [14,23,25]. The lymphoma cells were labeled with phycoerythrin-labeled antibodies to the tumor characteristic light chains for identifying the tumor cells and Hoechst 33258 for assessment of DNA content. The *DI* was determined from the G_1 peak position of the light chain positive cells relative to the light chain negative cells. Tumor cell fraction was determined as the fraction of light chain positive cells. The *DI* of the cervical cancers was assessed by preparing clean nuclei, stained with propidium iodide, using the detergent-trypsin method [41]. Cells from a diploid cell line were used as an internal reference. Samples showing two distinct G_1 peaks in the DNA histogram were classified as aneuploid, and the *DI* was determined from the position of the G_1 peak of the aneuploid cells relative to the corresponding peak of the diploid cells. Samples with a single G_1 peak were classified as near diploid. An estimate of the tumor cell fraction was achieved for each cervical cancer sample by histological examination of hematoxylin and eosin stained sections derived from the middle part of the biopsies. These values were used to compare with the tumor cell fractions estimated by GeneCount.

Array CGH

Genomic array slides produced by the Microarray Facility at the Norwegian Radium Hospital were used [42]. The arrays contained 4,549 unique genomic clones of BACs and P1 artificial chromosomes (PACs) (Wellcome Trust Sanger Institute, Cambridge, UK) that covered the whole genome with a resolution of approximately 1 Mb. The 1 Mb clone collection was supplemented with tiling path probes between 1q12 and 1q25, using overlapping BACs and PACs. The clones were from the RPCI-11 (BAC) and the RPC1-1, -3, -4, and -5 (PAC) libraries. Each clone was printed in 4-8 array spots. The genes covered by the clones were found from Ensembl [43].

Genomic DNA was isolated from the lymphoma cell suspensions and cervical cancer biopsies according to a standard protocol, including proteinase K, phenol, chloroform, and isoamylalcohol [44]. DNA (1 μg) was digested overnight, using *Dpn*II endonuclease (New England Biolabs, Beverly, MA, USA), and purified using the QIAquick PCR Purification Kit (Qiagen, Valencia, CA, USA). Digested and purified DNA and normal reference DNA (0.5 μg each) were labeled by a random primer reaction (BioPrime DNA Labeling System, Invitrogen, Carlsbad, CA, USA) with Cy3-dCTP and Cy5-dCTP (Perkin-Elmer Life Sciences, Foster City, CA, USA), respectively, and co-hybridized to the array slides [42]. Scanning and image analysis were performed by use of an Agilent scanner (Agilent Technologies Inc., Palo Alto, CA, USA) and the GenePix 6.0 image analysis software (Axon Instruments Inc., Union City, CA, USA). The microarray management and preprocessing software BASE [26] was used for spot filtering and ratio normalization. The mean value of the 4-8 spots of each genomic clone was used, provided that the standard deviation was less than 0.2. Lowess normalization was performed so that the mean log-transformed ratio of all clones was equal to 0. The GLAD and CGH-Explorer algorithms were used for ratio smoothing and breakpoint detection [9,11]. Default values of 8 (GLAD) and 1.5 (CGH-Explorer) for

the statistical penalty, λ, were used. The smoothed ratios were inputs to GeneCount.

Principle of GeneCount

For a heterogeneous test sample consisting of several cell populations, like normal cells and distinct populations of malignant cells, the DNA of each cell population contributes to the aCGH ratio. Ideally (that is, in cases of no experimental bias), the normalized ratio of each array probe is given by:

$$R_{ideal} = \frac{\sum_{i=1}^{n} \frac{N_i}{2 \cdot DI_i} \cdot F_i \cdot DI_i}{\sum_{i=1}^{n} F_i \cdot DI_i} \quad (1)$$

where R_{ideal} is the aCGH ratio of a sample with n cell populations, and N_i, DI_i, and F_i are the DNA copy number, DNA index, and tissue fraction of cell population i, respectively. We assume that: the reference sample is normal DNA with a copy number of 2 throughout the genome, except for the X and Y chromosomes in males; sex-matched hybridizations are performed; and DI is given relative to the DNA content of normal cells.

In cases of a homogeneous sample with a single cell population, for example, a cancer cell line, Equation 1 is reduced to:

$$R_{ideal} = \frac{N}{2 \cdot DI} \quad (2)$$

In clinical samples with two cell populations, that is, malignant and normal cells, the ratio is given by:

$$R_{ideal} = \frac{N_T}{2 \cdot DI_T} \cdot \frac{F_T \cdot DI_T}{F_T \cdot DI_T + (1 - F_T)} + \frac{1 - F_T}{F_T \cdot DI_T + (1 - F_T)} \quad (3)$$

where N_T, DI_T, and F_T are the DNA copy number, DNA index, and fraction of malignant cells in the sample, respectively. 1 - F_T represents the fraction of normal cells, which have a DI of 1 and DNA copy number (N) of 2.

It was clear from experiments where normal male DNA was hybridized against female DNA that the ratio dynamics were somewhat reduced (Figure 1e). A dynamic factor, q, was included in Equation 3 to compensate for this effect. Since the experimental noise was independent of the logarithm of the ratio (Additional data file 6), Equation 3 was rewritten to account for the reduced dynamic in the following way:

$$Log_2(R) = q \cdot Log_2\left(\frac{N_T}{2 \cdot DI_T} \cdot \frac{F_T \cdot DI_T}{F_T \cdot DI_T + (1 - F_T)} + \frac{1 - F_T}{F_T \cdot DI_T + (1 - F_T)} \right) \quad (4)$$

The dynamic factor represents the systematic, non-random reduction in the log-transformed ratios caused by the experimental bias and has a value between 0 and 1, where the latter value occurs in the ideal situation without any reduction in the ratio dynamics. The factor is a characteristic of the array slide series and the laboratory protocol and was determined from the ratio of the X chromosome in a control experiment hybridizing male versus female normal DNA (Figure 1e). Equation 4 was used in GeneCount to calculate F_T and N_T from the ratio profile of the sample.

Intratumor heterogeneity in the DNA copy numbers, that is, the cases of several populations of malignant cells in addition to the normal cells, was identified by selecting the tumors for which one or more of the aCGH ratio levels were different from that corresponding to an integer value by visual inspection. The ratio distributions of the potential heterogeneous regions were compared to the distributions of the adjacent homogeneous aberrations by ANOVA analysis, and a P-value of 0.05 was required to classify the aberration as heterogeneous. The fraction of tumor cells with a heterogeneous aberration was calculated, employing the more general Equation 1. The DI was assumed to be the same for all subpopulations of malignant cells.

Implementation of GeneCount in BASE

We used BASE as a platform for GeneCount and linked the algorithm to the output of the GLAD and CGH-Explorer packages, which were implemented in our BASE version. The method can also be developed as a separate program or integrated in other aCGH analysis packages. The algorithm consists of three major steps: data input for all samples; estimation of tumor cell fraction in the cases when this parameter is unknown; and estimation of DNA copy number for each array probe (panel A in Additional data file 1). The smoothed aCGH ratios served as input, together with the DI, the q-value from control experiments with its lower and upper limits (q_{min}, q_{max}) and, if available, the tumor cell fraction.

In cases of unknown tumor cell fraction, this value was estimated in a simulation procedure based on two selected ratio levels, using the tumor cell fraction and DNA copy numbers as independent and q as dependent variables. The copy numbers and tumor cell fraction were increased in steps of 1 and 0.01, respectively, and the corresponding q-value was calculated (panel B in Additional data file 1). To ensure high accuracy in the estimated fractions, it was required that the absolute value of the selected ratio levels was larger than 0.15. This implied that samples with a tumor cell fraction lower than 24% in diploid and 36% in tetraploid tumors could not be analyzed when only aberrations involving one copy number change existed (Additional data file 12). Moreover, a minimum absolute difference of 0.2 - that is, approximately two times the standard deviation of the log-transformed ratio levels (Additional data file 6) - between the two selected ratio

levels was needed. To further increase the reliability of the estimation, only ratio levels with more than ten probes were selected. We optimized q for each tumor by allowing the value to vary within the limited range of q_{min} to q_{max}, typically $q \pm 10\%$, leading to fairly stable estimates of the tumor cell fraction. The mean tumor cell fraction based on these estimates and the corresponding mean q-value was used in Equation 4 to estimate the DNA copy numbers of the tumor. In cases of known tumor cell fraction, this fraction and q from control experiments were used in Equation 4. The source code of the module is provided by communication to the authors. A demo version of GeneCount in BASE is also available [45].

Fluorescence *in situ* hybridization

GeneCount estimates for the lymphomas were compared with direct assessments of gene copy numbers by use of FISH. All FISH analyses have been published previously [20-25]. Dual-color FISH was applied to all 94 tumors. We used spectrum orange labeled locus-specific propidium iodide DNA probes for genes commonly aberrant in lymphomas (*CCND3*, *BMP6*, *PIM1*, *MYC*, *CDKN2A*, *RB1*, *TP53*, *PMAIP1*, and *MALT1*) and spectrum green labeled centromer probes (centromere 1, 6, 8, 17, and 18) (Vysis Inc., Downers Grove, IL, USA) for assessing the quality of the experiment. For exploring DNA copy number calculations in translocated chromosomal regions, *BCL2*, which is frequently involved in the translocation t(14;18)(q32;q21) in lymphomas, was considered. A dual-color translocation probe involving *BCL2* and covering the breakpoint region was used (LSI *IGH* Spectrum Green/LSI *BCL2* Spectrum Orange, Vysis Inc.). Due to splitting of the probe signal in cases of translocation, erroneous high *BCL2* copy numbers were derived directly with this probe. The *BCL2* copy number was therefore corrected based on the signals from the *IGH* and centromere 18 probes, as described [22].

Gene expression microarrays

Gene expressions were determined by microarray analysis of 89 of the cervical cancers and related to the GeneCount estimates. We used array slides produced at the Microarray Facility at the Norwegian Radium Hospital, containing 15,000 cDNA clones. The data from 48 of the patients, with a detailed description of the experimental procedures, have been presented [27]. Cy3- and Cy5-labeled cDNA was synthesized from total RNA by anchored oligo(dT)-primed reverse transcription and co-hybridized with a reference sample (Universal Human Reference RNA, Stratagene, La Jolla, CA, USA) to the array slides overnight at 65°C. Scanning and image analysis were performed with an Agilent scanner and the GenePix 4.1 image analysis software, respectively. Data preprocessing, including correction of saturated intensities, filtering of weak and bad spots, and lowess normalization, was performed in BASE. All hybridizations were performed twice in a dye-swap design, and the average expression ratio based on the two experiments was used in the further analyses.

ArrayExpress accession

The array CGH raw data have been deposited to the ArrayExpress repository (E-TABM-398, E-TABM-399).

Abbreviations

aCGH, array comparative genomic hybridization; BAC, bacterial artificial chromosome; BASE, Bioarray Software Environment; DI, DNA index; FISH, fluorescence *in situ* hybridization; PAC, P1 artificial chromosome.

Authors' contributions

HL and TS conceived and designed the study and analyzed data. HL wrote the article, ML, RSB, DHS, EG, and OTB carried out the aCGH, FISH, and flow cytometry experiments and participated in data analysis, LAMZ and OM contributed to the aCGH experiments, MJ and EH contributed to the implementation of GeneCount in BASE, GBK provided clinical samples and data, and TS helped to draft the manuscript. All authors read and approved the final manuscript.

Additional data files

The following additional data are available with the online version of this paper. Additional data file 1 is a figure showing the calculation steps in GeneCount. Additional data file 2 is a figure showing an example of FISH probe locations. Additional data file 3 is a figure comparing FISH DNA copy numbers and smoothed aCGH ratio levels in non-Hodgkin's lymphomas. Additional data file 4 is a figure illustrating discrepancies between GeneCount and FISH DNA copy numbers. Additional data file 5 is a figure comparing tumor cell fractions derived by histological examination and by GeneCount estimation in cervical cancers. Additional data file 6 is a figure showing the standard deviation (noise) of the log-transformed aCGH ratios. Additional data file 7 is a figure comparing results from ratio level and GeneCount analyses in cervical cancers. Additional data file 8 is a table listing regions with DNA copy number heterogeneity in non-Hodgkin's lymphomas. Additional data file 9 is a table listing regions with DNA copy number heterogeneity in cervical cancers. Additional data file 10 is a figure showing tumor cell fraction required for detection of heterogeneous copy number changes. Additional data file 11 is a figure illustrating analysis of the evolutionary sequence of subpopulations in heterogeneous tumors. Additional data file 12 is a figure showing the minimum tumor cell fraction that can be calculated in GeneCount. Additional data file 13 is a figure showing the tumor cell fraction required for detection of homogeneous copy number changes.

Acknowledgements

We would like to acknowledge Vegard Nygaard from The Norwegian Microarray Consortium for help with implementing the BASE plug-in module. The study was supported by The Norwegian Cancer Society and the

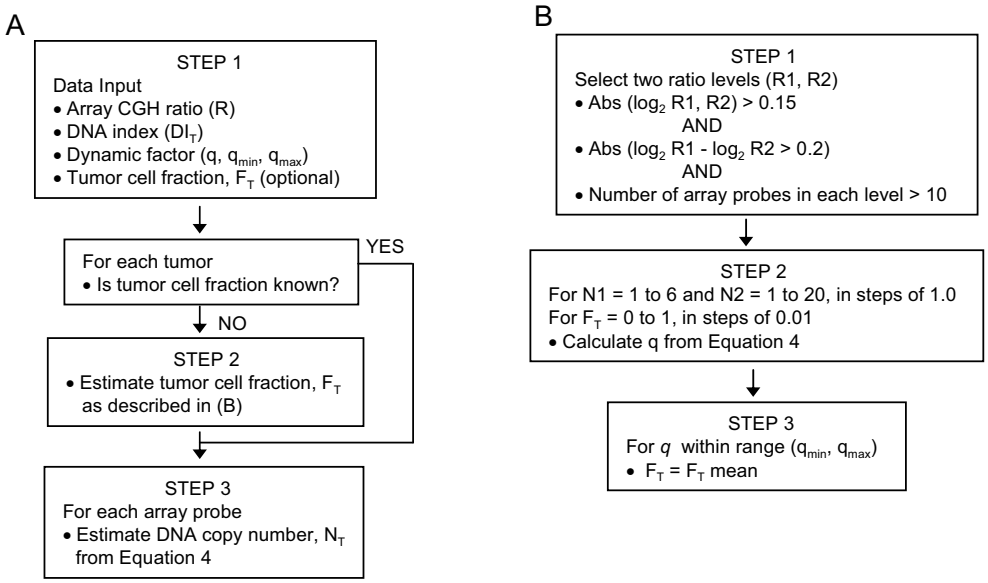
Microarray Platform of The National Programme for Research in Functional Genomics (FUGE) in the Research Council of Norway.

References

1. Mantripragada KK, Buckley PG, de Stahl TD, Dumanski JP: **Genomic microarrays in the spotlight.** *Trends Genet* 2004, **20:**87-94.
2. Pinkel D, Albertson DG: **Array comparative genomic hybridization and its applications in cancer.** *Nat Genet* 2005, **37(Suppl):**S11-S17.
3. Albertson DG, Collins C, McCormick F, Gray JW: **Chromosome aberrations in solid tumors.** *Nat Genet* 2003, **34:**369-376.
4. Albertson DG: **Gene amplification in cancer.** *Trends Genet* 2006, **22:**447-455.
5. Knuutila S, Aalto Y, Autio K, Bjorkqvist AM, El-Rifai W, Hemmer S, Huhta T, Kettunen E, Kiuru-Kuhlefelt S, Larramendy ML, Lushnikova T, Monni O, Pere H, Tapper J, Tarkkanen M, Varis A, Wasenius VM, Wolf M, Zhu Y: **DNA copy number losses in human neoplasms.** *Am J Pathol* 1999, **155:**683-694.
6. Chin SF, Teschendorff AE, Marioni JC, Wang Y, Barbosa-Morais NL, Thorne NP, Costa JL, Pinder SE, Wiel MA van de, Green AR, Ellis IO, Porter PL, Tavare S, Brenton JD, Ylstra B, Caldas C: **High-resolution array-CGH and expression profiling identifies a novel genomic subtype of ER negative breast cancer.** *Genome Biol* 2007, **8:**R215.
7. Chen W, Houldsworth J, Olshen AB, Nanjangud G, Chaganti S, Venkatraman ES, Halaas J, Teruya-Feldstein J, Zelenetz AD, Chaganti RS: **Array comparative genomic hybridization reveals genomic copy number changes associated with outcome in diffuse large B-cell lymphomas.** *Blood* 2006, **107:**2477-2485.
8. **The Mitelman Database of Chromosome Aberrations in Cancer** [http://cgap.nci.nih.gov/Chromosomes/Mitelman]
9. Hupe P, Stransky N, Thiery JP, Radvanyi F, Barillot E: **Analysis of array CGH data: from signal ratio to gain and loss of DNA regions.** *Bioinformatics* 2004, **20:**3413-3422.
10. Jong K, Marchiori E, Meijer G, Vaart AV, Ylstra B: **Breakpoint identification and smoothing of array comparative genomic hybridization data.** *Bioinformatics* 2004, **20:**3636-3637.
11. Lingjaerde OC, Baumbusch LO, Liestol K, Glad IK, Borresen-Dale AL: **CGH-Explorer: a program for analysis of array-CGH data.** *Bioinformatics* 2005, **21:**821-822.
12. Olshen AB, Venkatraman ES, Lucito R, Wigler M: **Circular binary segmentation for the analysis of array-based DNA copy number data.** *Biostatistics* 2004, **5:**557-572.
13. Rosenberg C, Schut TB, Mostert MC, Tanke HJ, Raap AK, Oosterhuis JW, Looijenga LH: **Comparative genomic hybridization in hypotriploid/hyperdiploid tumors.** *Cytometry* 1997, **29:**113-121.
14. Lyng H, Beigi M, Svendsrud DH, Brustugun OT, Stokke T, Kristensen GB, Sundfor K, Skjonsberg A, De Angelis PM: **Intratumor chromosomal heterogeneity in advanced carcinomas of the uterine cervix.** *Int J Cancer* 2004, **111:**358-366.
15. Harada K, Nishizaki T, Ozaki S, Kubota H, Ito H, Sasaki K: **Intratumoral cytogenetic heterogeneity detected by comparative genomic hybridization and laser scanning cytometry in human gliomas.** *Cancer Res* 1998, **58:**4694-4700.
16. Harada T, Okita K, Shiraishi K, Kusano N, Kondoh S, Sasaki K: **Interglandular cytogenetic heterogeneity detected by comparative genomic hybridization in pancreatic cancer.** *Cancer Res* 2002, **62:**835-839.
17. Beroukhim R, Getz G, Nghiemphu L, Barretina J, Hsueh T, Linhart D, Vivanco I, Lee JC, Huang JH, Alexander S, Du J, Kau T, Thomas RK, Shah K, Soto H, Perner S, Prensner J, Debiasi RM, Demichelis F, Hatton C, Rubin MA, Garraway LA, Nelson SF, Liau L, Mischel PS, Cloughesy TF, Meyerson M, Golub TA, Lander ES, Mellinghoff IK, *et al.*: **Assessing the significance of chromosomal aberrations in cancer: methodology and application to glioma.** *Proc Natl Acad Sci USA* 2007, **104:**20007-20012.
18. Weir BA, Woo MS, Getz G, Perner S, Ding L, Beroukhim R, Lin WM, Province MA, Kraja A, Johnson LA, Shah K, Sato M, Thomas RK, Barletta JA, Borecki IB, Broderick S, Chang AC, Chiang DY, Chirieac LR, Cho J, Fujii Y, Gazdar AF, Giordano T, Greulich H, Hanna M, Johnson BE, Kris MG, Lash A, Lin L, Lindeman N, *et al.*: **Characterizing the cancer genome in lung adenocarcinoma.** *Nature* 2007, **450:**893-898.
19. Mohapatra G, Moore DH, Kim DH, Grewal L, Hyun WC, Waldman FM, Pinkel D, Feuerstein BG: **Analyses of brain tumor cell lines confirm a simple model of relationships among fluorescence *in situ* hybridization, DNA index, and comparative genomic hybridization.** *Genes Chromosomes Cancer* 1997, **20:**311-319.
20. Galteland E, Holte H, Stokke T: **c-MYC, RB-1, TP53, and centromere 8 and 17 copy number in B-cell non-Hodgkin's lymphomas assessed by dual-color fluorescence *in situ* hybridization.** *Cytometry* 1999, **38:**53-60.
21. Galteland E, Smedshammer L, Suo Z, DeAngelis P, Stokke T: **Proliferation-dependent expression and phosphorylation of pRB in B cell non-Hodgkin's lymphomas: dependence on RB1 copy number.** *Leukemia* 2002, **16:**1549-1555.
22. Galteland E, Sivertsen EA, Svendsrud DH, Smedshammer L, Kresse SH, Meza-Zepeda LA, Myklebost O, Suo Z, Mu D, DeAngelis PM, Stokke T: **Translocation t(14;18) and gain of chromosome 18/BCL2: effects on BCL2 expression and apoptosis in B-cell non-Hodgkin's lymphomas.** *Leukemia* 2005, **19:**2313-2323.
23. Sivertsen EA, Galteland E, Mu D, Holte H, Meza-Zepeda L, Myklebost O, Patzke S, Smeland EB, Stokke T: **Gain of chromosome 6p is an infrequent cause of increased PIM1 expression in B-cell non-Hodgkin's lymphomas.** *Leukemia* 2006, **20:**539-542.
24. Stokke T, Galteland E, Holte H, Smedshammer L, Suo Z, Smeland EB, Borresen-Dale AL, DeAngelis P, Steen HB: **Oncogenic aberrations in the p53 pathway are associated with a high S phase fraction and poor patient survival in B-cell Non-Hodgkin's lymphoma.** *Int J Cancer* 2000, **89:**313-324.
25. Stokke T, DeAngelis P, Smedshammer L, Galteland E, Steen HB, Smeland EB, Delabie J, Holte H: **Loss of chromosome 11q21-23.1 and 17p and gain of chromosome 6p are independent prognostic indicators in B-cell non-Hodgkin's lymphoma.** *Br J Cancer* 2001, **85:**1900-1913.
26. Saal LH, Troein C, Vallon-Christersson J, Gruvberger S, Borg A, Peterson C: **BioArray Software Environment (BASE): a platform for comprehensive management and analysis of microarray data.** *Genome Biol* 2002, **3:**SOFTWARE0003.
27. Lyng H, Brovig RS, Svendsrud DH, Holm R, Kaalhus O, Knutstad K, Oksefjell H, Sundfor K, Kristensen GB, Stokke T: **Gene expressions and copy numbers associated with metastatic phenotypes of uterine cervical cancer.** *BMC Genomics* 2006, **7:**268.
28. Davies JJ, Wilson IM, Lam WL: **Array CGH technologies and their applications to cancer genomes.** *Chromosome Res* 2005, **13:**237-248.
29. Kallioniemi OP, Kallioniemi A, Piper J, Isola J, Waldman FM, Gray JW, Pinkel D: **Optimizing comparative genomic hybridization for analysis of DNA sequence copy number changes in solid tumors.** *Genes Chromosomes Cancer* 1994, **10:**231-243.
30. Mitelman F, Johansson B, Mertens F: **The impact of translocations and gene fusions on cancer causation.** *Nat Rev Cancer* 2007, **7:**233-245.
31. Kim SW, Kim JW, Kim YT, Kim JH, Kim S, Yoon BS, Nam EJ, Kim HY: **Analysis of chromosomal changes in serous ovarian carcinoma using high-resolution array comparative genomic hybridization: Potential predictive markers of chemoresistant disease.** *Genes Chromosomes Cancer* 2007, **46:**1-9.
32. Nigro JM, Misra A, Zhang L, Smirnov I, Colman H, Griffin C, Ozburn N, Chen M, Pan E, Koul D, Yung WK, Feuerstein BG, Aldape KD: **Integrated array-comparative genomic hybridization and expression array profiles identify clinically relevant molecular subtypes of glioblastoma.** *Cancer Res* 2005, **65:**1678-1686.
33. Pierga JY, Reis-Filho JS, Cleator SJ, Dexter T, Mackay A, Simpson P, Fenwick K, Iravani M, Salter J, Hills M, Jones C, Ashworth A, Smith IE, Powles T, Dowsett M: **Microarray-based comparative genomic hybridisation of breast cancer patients receiving neoadjuvant chemotherapy.** *Br J Cancer* 2007, **96:**341-351.
34. Ruano Y, Mollejo M, Ribalta T, Fiano C, Camacho FI, Gomez E, de Lope AR, Hernandez-Moneo JL, Martinez P, Melendez B: **Identification of novel candidate target genes in amplicons of Glioblastoma multiforme tumors detected by expression and CGH microarray profiling.** *Mol Cancer* 2006, **5:**39.
35. van Wieringen WN, Belien JA, Vosse SJ, Achame EM, Ylstra B: **ACE-it: a tool for genome-wide integration of gene dosage and RNA expression data.** *Bioinformatics* 2006, **22:**1919-1920.
36. van Duin M, van Marion R, Watson JE, Paris PL, Lapuk A, Brown N, Oseroff VV, Albertson DG, Pinkel D, de Jong P, Nacheva EP, Dinjens W, van Dekken H, Collins C: **Construction and application of a full-coverage, high-resolution, human chromosome 8q genomic microarray for comparative genomic hybridization.** *Cytometry A* 2005, **63:**10-19.
37. Di VA, Infusini E, Peveri C, Sciutto A, Orecchia R, Geido E, Monaco

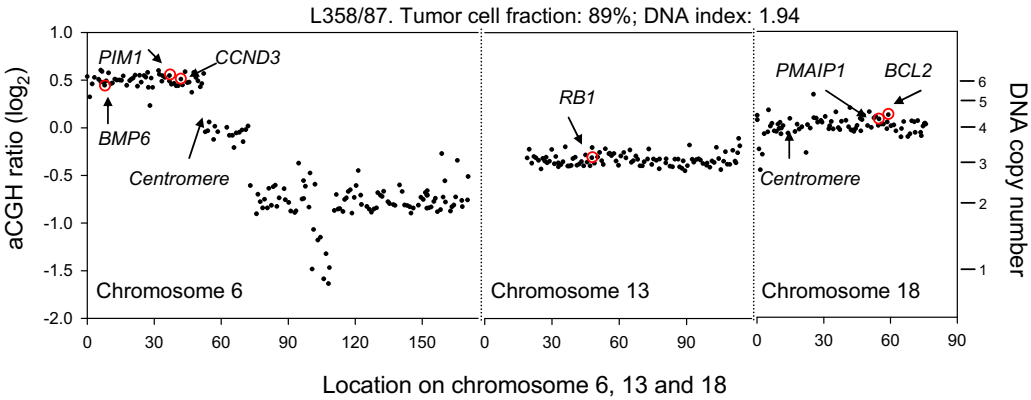
R, Giaretti W: **Intratumor heterogeneity of chromosome 1, 7, 17, and 18 aneusomies obtained by FISH and association with flow cytometric DNA index in human colorectal adenocarcinomas.** *Cytometry* 1999, **35:**369-375.
38. Jenkins RB, Qian J, Lieber MM, Bostwick DG: **Detection of c-myc oncogene amplification and chromosomal anomalies in metastatic prostatic carcinoma by fluorescence *in situ* hybridization.** *Cancer Res* 1997, **57:**524-531.
39. Lengauer C, Kinzler KW, Vogelstein B: **Genetic instabilities in human cancers.** *Nature* 1998, **396:**643-649.
40. De Angelis PM, Stokke T, Beigi M, Flatberg G, Enger M, Haug K, Aass HC, Schjolberg A, Andresen PA, Ariansen S, Bo AS, Mjaland O, Clausen OP: **Chromosomal 20q gain in the DNA diploid component of aneuploid colorectal carcinomas.** *Int J Cancer* 2007, **120:**2734-2738.
41. Vindelov LL, Christensen IJ, Nissen NI: **A detergent-trypsin method for the preparation of nuclei for flow cytometric DNA analysis.** *Cytometry* 1983, **3:**323-327.
42. Meza-Zepeda LA, Kresse SH, Barragan-Polania AH, Bjerkehagen B, Ohnstad HO, Namlos HM, Wang J, Kristiansen BE, Myklebost O: **Array comparative genomic hybridization reveals distinct DNA copy number differences between gastrointestinal stromal tumors and leiomyosarcomas.** *Cancer Res* 2006, **66:**8984-8993.
43. **Ensembl** [http://www.ensembl.org/Homo_sapiens/searchview]
44. De Angelis PM, Clausen OP, Schjolberg A, Stokke T: **Chromosomal gains and losses in primary colorectal carcinomas detected by CGH and their associations with tumour DNA ploidy, genotypes and phenotypes.** *Br J Cancer* 1999, **80:**526-535.
45. **Demo version of GeneCount in BASE** [http://alba.uio.no/base/local/genecountdemo/]

Additional data file 1



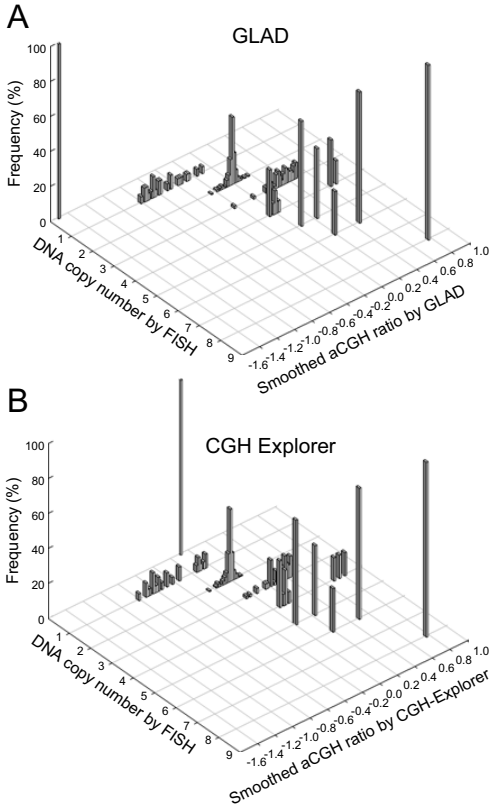
Calculation steps in GeneCount.
Float diagrams of the entire GeneCount procedure (A) and the estimation of tumor cell fraction (B) are shown. Data input in (A) includes the aCGH ratio levels, *R*, derived from statistical analysis tools, the DNA index (DI_T) of the tumor cells, the dynamic factor, *q*, of the experiment, and the tumor cell fraction, F_T. F_T can either be measured by a separate technique like flow cytometry, or estimated by the procedure in (B). In the former case, a fixed *q*-value, as determined from control experiments, is used. Otherwise, *q* is estimated for each tumor as indicated in (B), allowing for a deviation of typically 10% (range *q*min - *q*max) from the value determined in control experiments. To calculate tumor cell fraction, two ratio levels, *R*1 and *R*2, are selected, and the DNA copy numbers *N*1 and *N*2, corresponding to *R*1 and *R*2, respectively, are predicted in a stepwise manner by increasing *N*1 from 1 to 6 and *N*2 from 1 to 20 in steps of 1. *R*1, *R*2, and the incremental values of *N*1 and *N*2 are used in a simulation procedure based on Equation 4, allowing F_T to vary from 0 to 1 in steps of 0.01, and calculating the corresponding *q*. The F_T values obtained for *q* within the range *q*min – *q*max are used to calculate a mean F_T representative of the tumor. The mean F_T is further used as input to the algorithm in (A) to calculate the DNA copy numbers of all array probes. The criteria used to select *R*1 and *R*2 for estimation of tumor cell fraction are described in the Materials and methods section of the paper. The source code of the GeneCount module is available by communication to the authors.

Additional data file 2



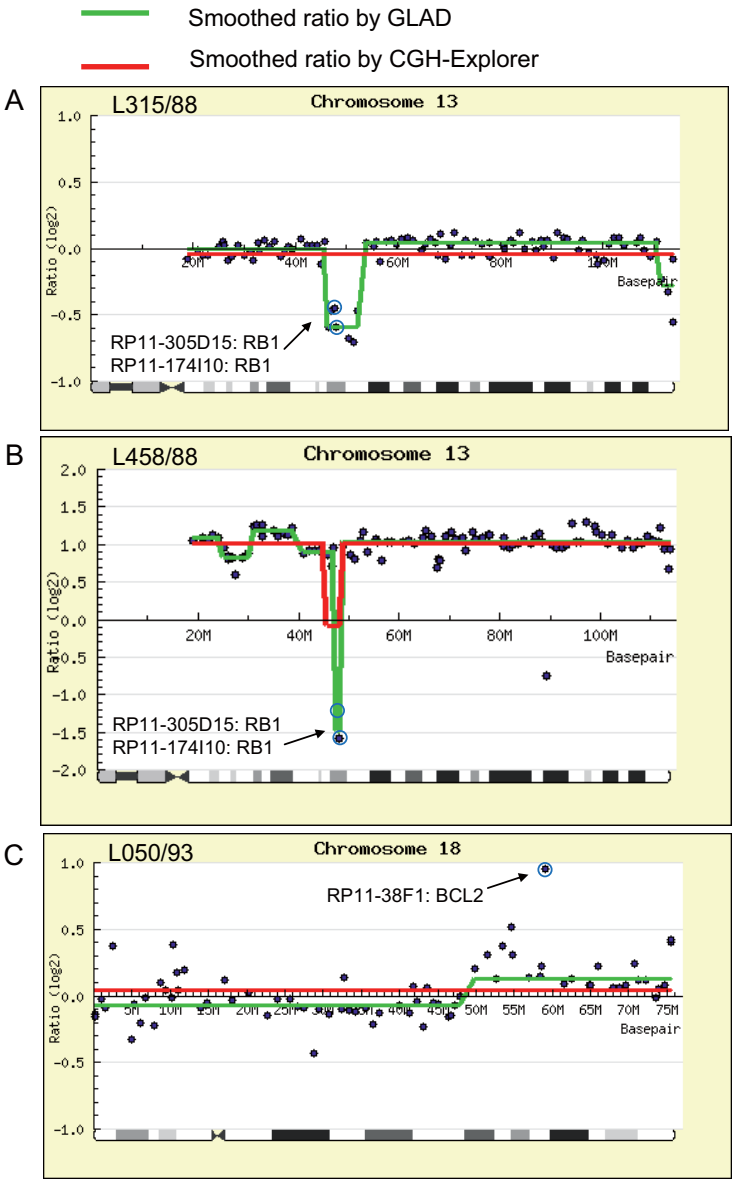
Example of FISH probe locations.
ACGH ratios are plotted against chromosomal location on chromosomes 6, 13, and 18 in a tetraploid lymphoma with a tumor cell fraction of 89%. FISH probes for *BMP6, PIM1, CCND3*, centromere 6, *RB1*, *PMAIP1*, *BCL2,* and centromere 18 were used in this particular tumor, and their location is marked. The copy numbers determined with FISH were 3 (*RB1*), 4 (*PMAIP1*, *BCL2*), and 6 (*BMP6*, *PIM1*, *CCND3*).

Additional data file 3



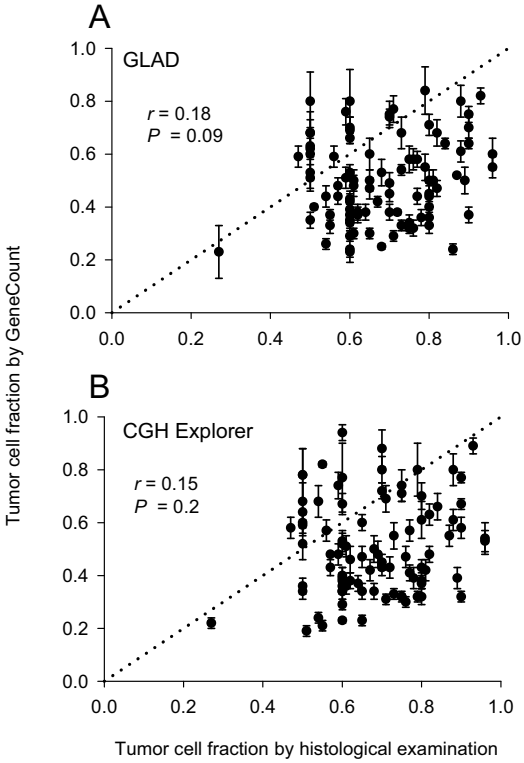
FISH DNA copy numbers in relation to smoothed aCGH ratio levels.
Smoothed aCGH ratio levels derived from GLAD (A) and CGH-Explorer (B) is plotted against the corresponding FISH results for 9 genes in 94 lymphomas. Frequency distributions are shown for each copy number, containing 1, 25, 246, 66, 15, 5, 4, and 1 values at a FISH copy number of 0, 1, 2, 3, 4, 5, 6, and 8, respectively.

Additional data file 4



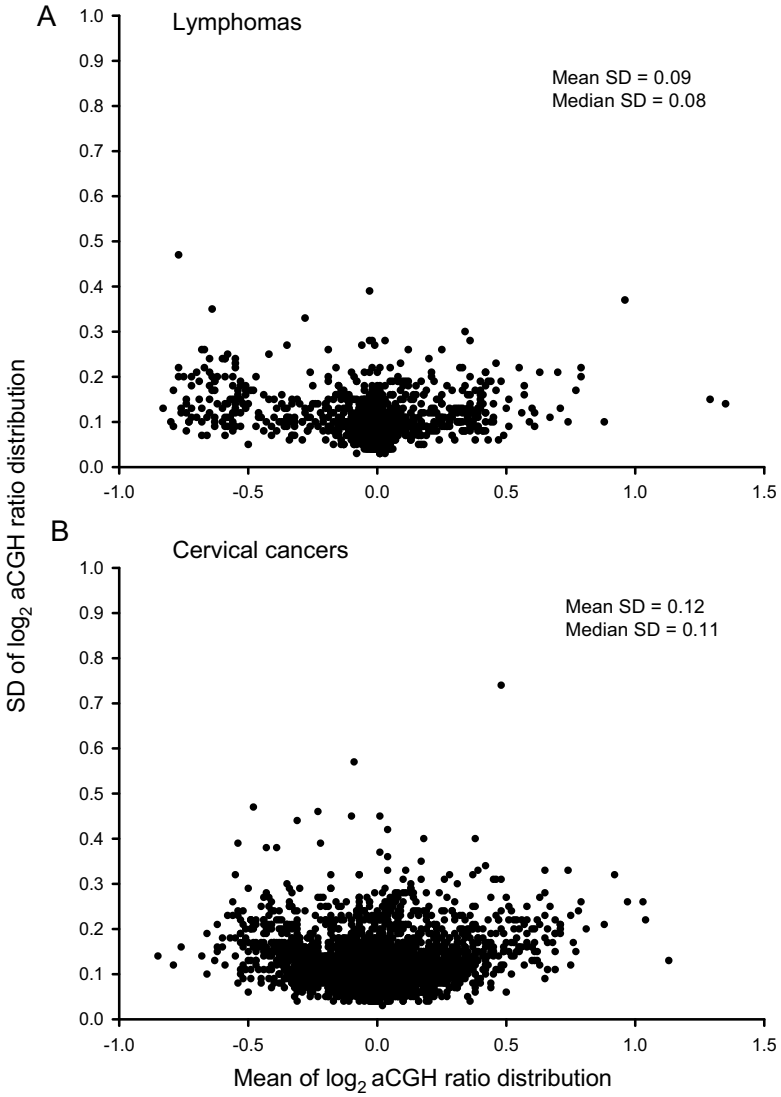
Discrepancies between GeneCount and FISH DNA copy numbers.
(A) shows a deletion involving *RB1* that was detected by GLAD but not by CGH-Explorer. The GeneCount estimate was 1.1 and 1.9 based on GLAD and CGH-Explorer, respectively. The *RB1* copy number determined by FISH was 1. (B) shows a homozygote deletion involving *RB1*, where the ratio level was inappropriately determined since the deleted region contained only two array probes. The GeneCount estimate was 0.5 and 1.9 based on GLAD and CGH-Explorer, respectively. The *RB1* copy number determined by FISH was zero. (C) shows a gain involving *BCL2*, which was not detected by GLAD or CGH-Explorer. The GeneCount estimate was 2.3 and 2.1 based on GLAD and CGH-Explorer, respectively. The *BCL2* copy number determined by FISH was 5. There was no t(14;18) translocation in this tumor.

Additional data file 5



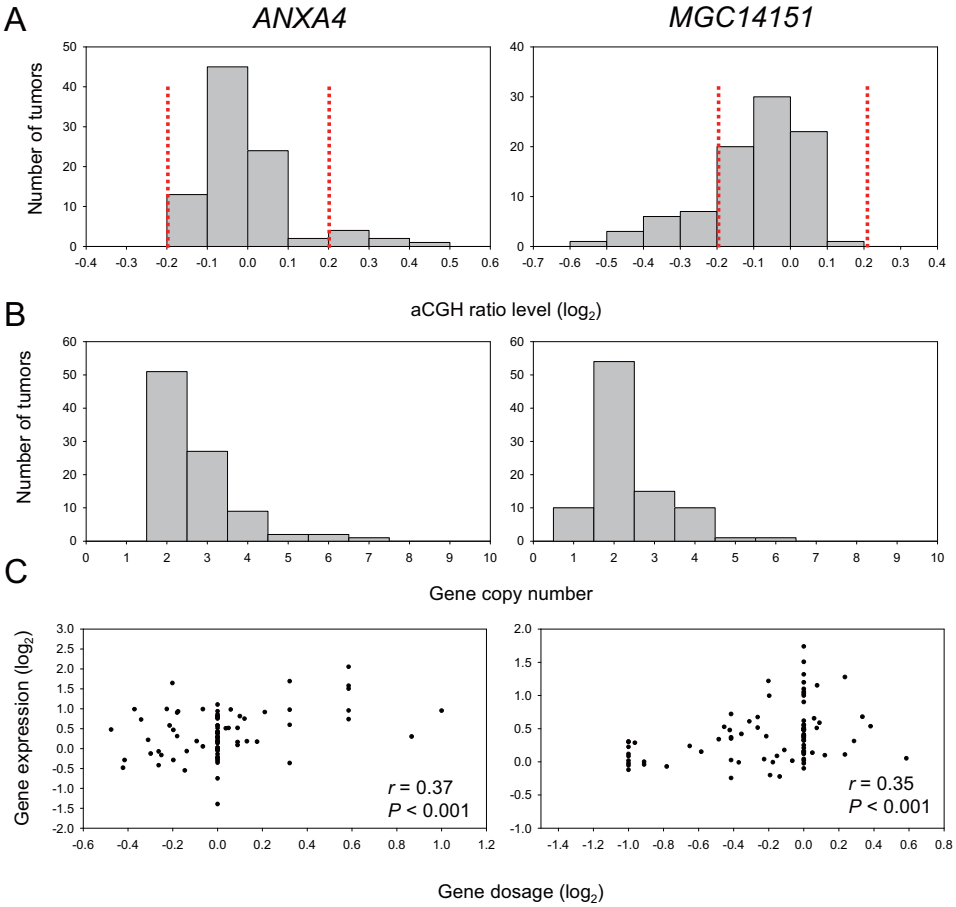
Histological examination and GeneCount estimation of tumor cell fraction.
Plots of tumor cell fraction of cervical cancer samples estimated by GeneCount versus tumor cell fraction measured by histological examination of tumor sections. Each point represents mean ± SD based on the values achieved for *q* within the range 0.7 – 0.9. The smoothed aCGH ratios from GLAD (A) and CGH-explorer (B), the *q* range 0.7 – 0.8, and a *DI* determined by flow cytometry were inputs to GeneCount. The calculations were based on 93 (A) and 89 (B) tumors, for which suitable ratio levels for the calculations existed. Correlation coefficients and *P*-values from Pearson product moment correlation analyses are indicated.

Additional data file 6



Standard deviation (noise) of log transformed aCGH ratios.
Standard deviation (SD) *versus* mean of the log-transformed ratio distributions achieved for 94 lymphomas (A) and 99 cervical cancers (B) is shown. The individual ratio levels determined by the breakpoint detection algorithm in GLAD were considered. The median and mean SD is indicated. Note that the SD was independent of the mean ratio and therefore of the DNA copy number.

Additional data file 7



Ratio level and GeneCount analyses in cervical cancers.
In (A) frequency histogram (number of tumors) of smoothed aCGH ratios (GLAD) is shown for *ANXA4* (BAC clone ID RP11-304A15) (left panel) and *MGC14151* (RP11-144K9) (right panel). Stippled lines indicate the cut off ratios of ±0.2, identifying 7 (*ANXA4*) and 17 (*MGC14151*) tumors with genetic gain or loss. In (B) frequency histogram (number of tumors) of DNA copy number calculated by GeneCount is shown for *ANXA4* (left panel) and *MGC14151* (right panel). The GLAD ratio levels, the *DI* measured by flow cytometry, and the tumor cell fraction estimated by GeneCount were used in the calculation. Similar results were achieved based on the CGH-Explorer ratio levels. In (C) plot of gene expressions against gene dosage; *i.e.*, the *ANXA4* (left panel) and *MGC14151* (right panel) copy number divided by the total DNA content ($N/(2 \cdot DI)$), is shown. The gene dosage was changed; *i.e.,* higher than 0.2 or lower than -0.2, in 26 (*ANXA4*) and 40 (*MGC14151*) tumors; *i.e.,* in considerable more tumors than when only the ratio levels were used. Correlation coefficient and *P*-value from Pearson product moment correlation analysis are indicated. Panels A and B are based on 92 (ANXA4) and 91 (MGC14151) tumors, for which GeneCount data existed. Panel C is based on 85 of these tumors, for which both DNA copy numbers and expression were available.

Additional data file 8

Table 1. DNA copy number heterogeneity in non-Hodgkin's lymphomas

Patient	DNA index	DNA region[a]	DNA copy number
L122/84	1.23	17q24-ter	1&2
L255/85	1.04	12p	1&2
L358/87	1.94	9pter-q21, 9q31-ter	3&4
L47/88	1.00	11q14-22	1&2
L154/88	0.95	4	1&2
L399/88	1.00	11q13-14, X (male)	1&2
		22	2&3
L064/89	1.26	17	2&3
L309/89	1.83	8, 9, 17q21-ter	3&4
L339/89	1.00	17q21-25	2&3
L034/90	1.16	X (male)	1&2
L472/90	1.16	X	2&3
L577/90	1.97	4	2&3
		11	3&4
		3	4&5
L382/91	1.02	13q32-ter	1&2
L383/91	0.97	11q22-23, 15q25-ter	1&2
L436/91	1.04	Xq26-ter (male)	1&2
L462/91	1.00	3pcen-14, 3p21-22, 3p24-ter, 15qcen-15, 17p11-ter	1&2
L008/92	1.00	2q, 5p13-ter, 7q31-ter, 9p21-23, 13q14-31, 20q13-ter, Xp22-ter	1&2
L037/92	1.00	19	1&2
L117/92	1.00	4	1&2

[a]The aCGH ratios of the heterogeneous DNA region were significantly different from those of the homogeneous regions in all cases, as verified from ANOVA analysis.

Additional data file 9

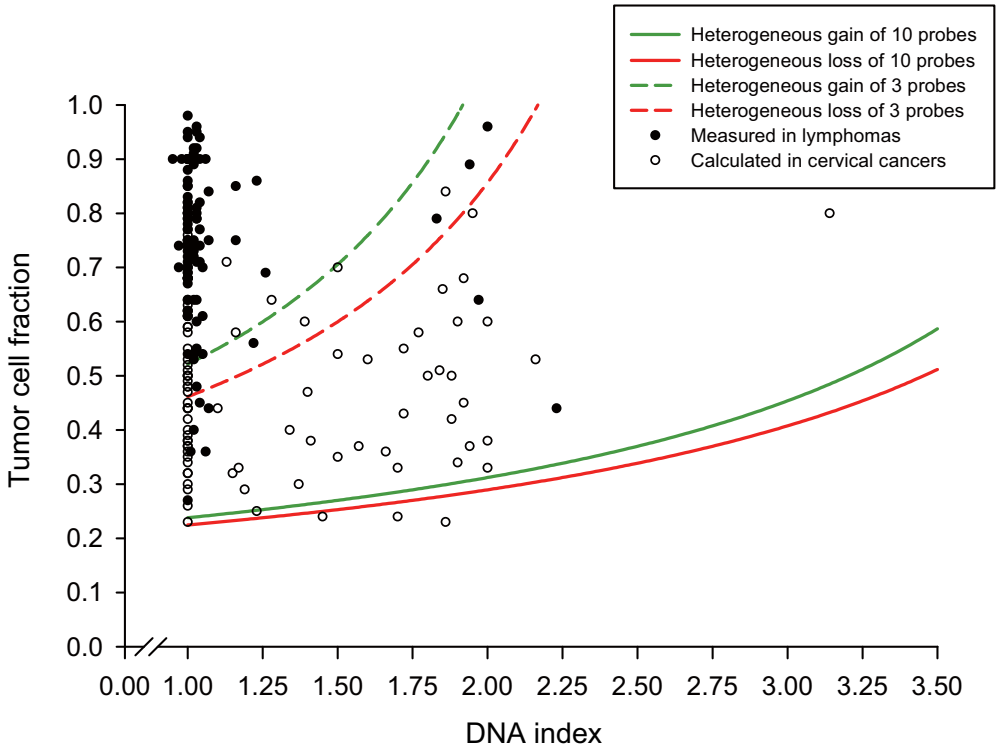
Table 2. DNA copy number heterogeneity in cervical cancers

Patient	DNA index	DNA region[a]	DNA copy number
C002/01	1.00	1p34-ter, 10, 16, 17, 19, 20	1&2
		2pter-q33, 5p14-q14, 6, 7p11-ter, 8, 12pter-q23, X	2&3
C003/01	1.90	5pter-q12, 6, 11pter-q12, 12, 13q21-31, Xp	3&4
C004/01	1.00	17q23-ter, X	2&3
C005/01	1.70	1p33-ter, 4, 7q, Xpter-q27	2&3
		3q, 6q13-ter, 7p, 10pter-q24, 11pter-q22, 17, 18, 20	3&4
C006/01	1.30	6q24-ter	1&2
		1p32-ter, 11q12-13, 17q, 18, 19, 22q	2&3
C008/01	1.00	9q	2&3
C014/01	1.94	3p	3&4
C023/01	1.00	2q33-ter, 4, 8p	1&2
		1p32-qter, 3, 7q21-ter	2&3
C024/01	1.00	1p31-ter, 2q21-ter, 4p, 8p, 13qcen-34	1&2
		6p	2&3
C025/01	1.00	21qcen-21, X	1&2
		9	2&3
C028/01	1.00	X	1&2
		9	2&3
C041/01	1.34	8	3&4
C045/01	1.00	3p	1&2
C053/01	1.00	1p34-ter	1&2
C054/01	1.00	2q33-ter, 9qcen-22, Xq22-ter	1&2
		15	2&3
C078/02	1.70	5q, Xpter-q28	3&4
C092/02	1.00	3p, 8p12-ter, 17, 19, 20, 22	1&2
		1p34-ter, 2p13-ter, 8p12-q11, 8q13-ter, 13q, 18	2&3
C093/02	2.00	17pter-q25, 18	3&4
C101/02	1.50	2, 4, 20q11-ter	2&3
		1, 3p	3&4
C115/03	1.00	2p21-11, 17q21-ter, 18q21-ter	2&3
C120/03	1.00	22	1&2
		8, 18, X	2&3

C130/03	1.60	4, 5qcen-14, 7, 17pter-q21	2&3
		12q, Xq	3&4
C136/03	1.40	4p15-ter, 19pter-q13	1&2
		6p12-ter, 8q22-ter, X	2&3
C138/03	1.90	7p11-22, 12, 15	3&4
C141/03	1.10	5q23-ter, 20q12-ter	2&3
C142/03	1.00	3q24-ter, X	2&3
C144/03	1.50	3p, 19	2&3
		3q, 4q13-ter, 7, Xp22-q21	3&4
C152/03	1.00	3q	2&3
C162/04	2.16	10q, 12q, 18q, 21	3&4
C169/04	1.00	18q12-ter, 21qcen-21	1&2
C174/04	1.95	1p32-ter, 16, 19, 20, 22	3&4
C178/04	1.00	5q, 17p, 21	1&2
		1q, 13q22-ter, 16, 20	2&3
C180/04	1.92	11p, 18q11-21,	2&3
		3p, 7, 13q32-ter, 18q21-ter	3&4
C184/04	1.39	3p, 14, 17p, Xpter-q25	2&3
		1pter-q32, 8q21-ter	3&4
C188/04	1.00	2, 5q, 6pter-q25, 10, 11, 15, 18	1&2
		9q22-ter, 12p11-ter, 12q13-ter, 16	2&3
C189/04	1.50	4q, Xp11-ter	1&2
		17pter-q24	2&3
C190/04	1.00	3p, 4, 8, 13	1&2
C193/04	1.00	1pcen-31, 3p, 4p15-qter, 5, 6p21-q24, 8, 14qcen -21, 18	1&2
C194/04	1.00	5q, 9p, 11p15-q13, 18p12-ter, 22q13-ter	1&2
		16q, 19p, 20	2&3
C195/04	1.90	1p32-q12, 2p11-ter, 12q, 14, 15	3&4
C196/04	1.00	2q12-q33, 4p13-qter, 17p, X	1&2
C202/04	1.00	13q21-q32	1&2
C206/04	1.00	4, 13q21-q31	1&2
		1p31-ter, 19, 22	2&3
C215/04	1.00	11p, 13q21-ter	1&2
		1p31-ter	2&3

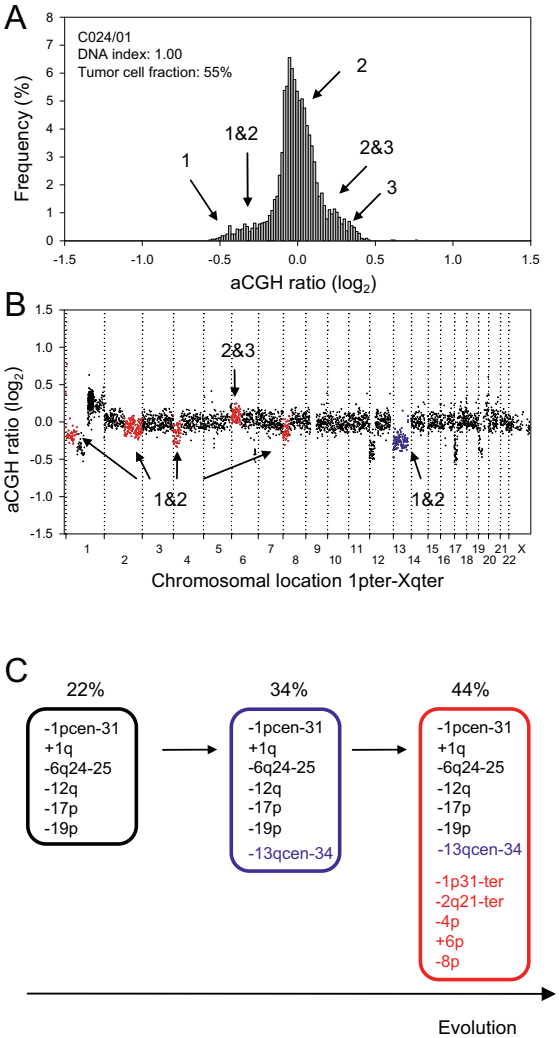
[a]The aCGH ratios of the heterogeneous DNA region were significantly different from those of the homogeneous regions in all cases, as verified from ANOVA analysis.

Additional data file 10



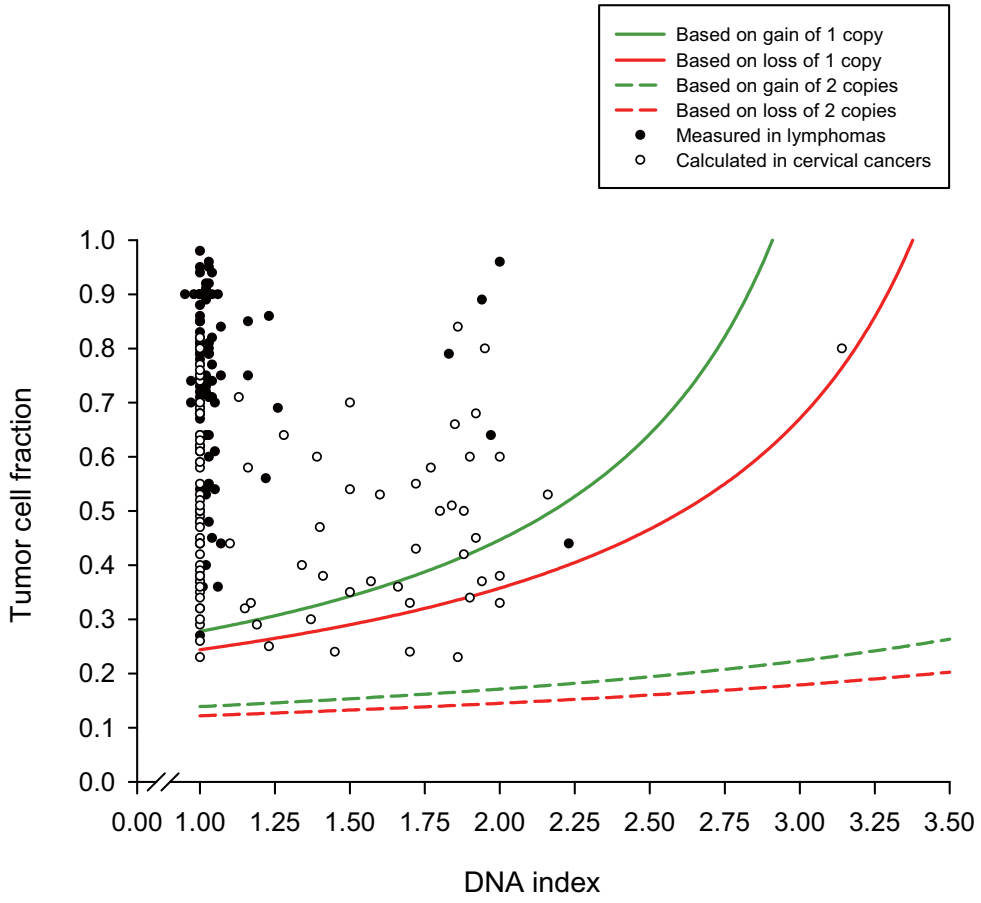
Tumor cell fraction required for detection of heterogeneous copy number changes.
The tumor fraction needed for statistically significant separation of a heterogeneous chromosomal region involving more than ten (solid lines) or three (stippled lines) array probes from a homogeneous region without aberration is shown as a function of the DNA index (*DI*). The curve for heterogeneous gain (copy number $2 \cdot DI + 0.5$) and loss (copy number $2 \cdot DI - 0.5$) is shown in green and red color, respectively. A Student t-test and a standard deviation of 0.1 for the log-transformed ratio distributions were used to estimate the curves. It was assumed that the heterogeneous gain and loss affected 50% of the tumor cells in the sample. Data showing the tumor cell fraction of 94 lymphomas (closed symbols) and 93 cervical cancers (open symbols), as determined by flow cytometry (lymphomas) and estimated by GeneCount from the GLAD ratio levels (cervical cancers) are included.

Additional data file 11



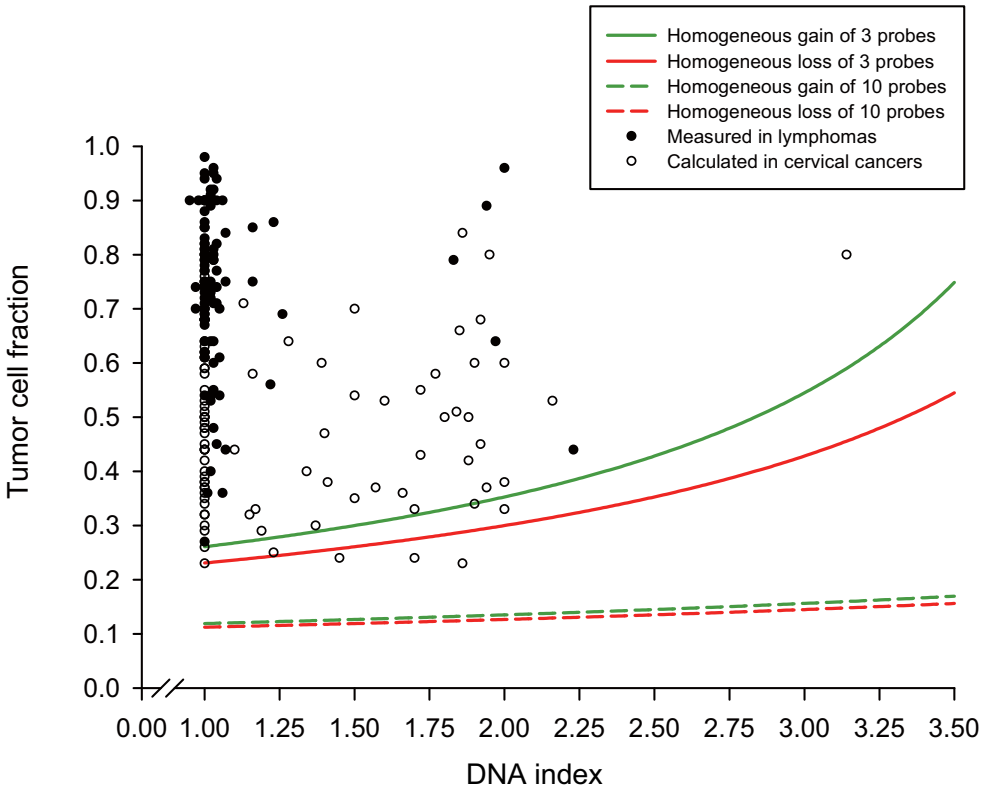
Evolutionary sequence of subpopulations in heterogeneous tumors.
In (A) frequency histogram (% array probes) of aCGH ratios in a heterogeneous cervical cancer is shown. Data of the entire genome are included. In (B) the aCGH ratios are plotted against chromosomal location. The heterogeneous regions on chromosomes 1p, 2q, 4p, 8p, and 13q with a DNA copy number of 1&2 and on 6p with a DNA copy number 2&3 are shown in blue and red. The blue and red colors represent aberrations of two different fractions of the tumor cells, 78% and 44%, respectively. The heterogeneous aberrations are listed in Additional Data File 9. In (C) schematic diagram of the possible evolutionary sequence for the aberrations are shown. The blue and red aberrations refer to the blue and red aberrations in (B). The percentages indicate the fractions of tumor cells with the listed aberrations, as calculated by GeneCount, showing that the aberrations in blue and red are present in 78% and 44% of the tumor cells, respectively.

Additional data file 12



Minimum tumor cell fraction that can be calculated in GeneCount.
The tumor fraction needed for achieving ratio levels of 0.15 above the baseline, assuming a copy number gain (green) or loss (red) of 1 (stippled lines) and 2 (dotted lines) copies, is shown as a function of the DNA index (*DI*). The curves were calculated from Equation 4. Data showing the tumor cell fractions of 94 lymphomas (closed symbols) and 93 cervical cancers (open symbols), as determined by flow cytometry (lymphomas) and estimated by GeneCount from the GLAD ratio levels (cervical cancers) are included.

Additional data file 13



Tumor cell fraction required for detection of homogeneous copy number changes.
The tumor fraction needed for statistically significant separation of an aberrant chromosomal region involving more than three (solid lines) or ten (stippled lines) array probes from a region without aberration is shown as a function of the DNA index (*DI*). The curve for gain and loss of one DNA copy (copy number $2 \cdot DI \pm 1$) is shown in green and red color, respectively. A Student t-test and a standard deviation of 0.1 for the log-transformed ratio distributions were used to estimate the curves. Data showing the tumor cell fractions of 94 lymphomas (closed symbols) and 93 cervical cancers (open symbols), as determined by flow cytometry (lymphomas) and estimated by GeneCount from the GLAD ratio levels (cervical cancers) are included.

II

Gene Dosage, Expression, and Ontology Analysis Identifies Driver Genes in the Carcinogenesis and Chemoradioresistance of Cervical Cancer

Malin Lando[1], Marit Holden[2], Linn C. Bergersen[3], Debbie H. Svendsrud[1], Trond Stokke[1], Kolbein Sundfør[4], Ingrid K. Glad[3], Gunnar B. Kristensen[4,5], Heidi Lyng[1]*

1 Department of Radiation Biology, Norwegian Radium Hospital, Oslo, Norway, **2** Norwegian Computing Center, Oslo, Norway, **3** Department of Mathematics, University of Oslo, Oslo, Norway, **4** Department of Gynecologic Oncology, Norwegian Radium Hospital, Oslo, Norway, **5** Department of Medical Informatics, University of Oslo, Oslo, Norway

Abstract

Integrative analysis of gene dosage, expression, and ontology (GO) data was performed to discover driver genes in the carcinogenesis and chemoradioresistance of cervical cancers. Gene dosage and expression profiles of 102 locally advanced cervical cancers were generated by microarray techniques. Fifty-two of these patients were also analyzed with the Illumina expression method to confirm the gene expression results. An independent cohort of 41 patients was used for validation of gene expressions associated with clinical outcome. Statistical analysis identified 29 recurrent gains and losses and 3 losses (on 3p, 13q, 21q) associated with poor outcome after chemoradiotherapy. The intratumor heterogeneity, assessed from the gene dosage profiles, was low for these alterations, showing that they had emerged prior to many other alterations and probably were early events in carcinogenesis. Integration of the alterations with gene expression and GO data identified genes that were regulated by the alterations and revealed five biological processes that were significantly overrepresented among the affected genes: apoptosis, metabolism, macromolecule localization, translation, and transcription. Four genes on 3p (*RYBP*, *GBE1*) and 13q (*FAM48A*, *MED4*) correlated with outcome at both the gene dosage and expression level and were satisfactorily validated in the independent cohort. These integrated analyses yielded 57 candidate drivers of 24 genetic events, including novel loci responsible for chemoradioresistance. Further mapping of the connections among genetic events, drivers, and biological processes suggested that each individual event stimulates specific processes in carcinogenesis through the coordinated control of multiple genes. The present results may provide novel therapeutic opportunities of both early and advanced stage cervical cancers.

Citation: Lando M, Holden M, Bergersen LC, Svendsrud DH, Stokke T, et al. (2009) Gene Dosage, Expression, and Ontology Analysis Identifies Driver Genes in the Carcinogenesis and Chemoradioresistance of Cervical Cancer. PLoS Genet 5(11): e1000719. doi:10.1371/journal.pgen.1000719

Editor: Yoshihide Hayashizaki, RIKEN Genomic Sciences Center, Japan

Received May 26, 2009; **Accepted** October 14, 2009; **Published** November 13, 2009

Copyright: © 2009 Lando et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Funding: The study was supported by The National Programme for Research in Functional Genomics (FUGE) in the Research Council of Norway, The Norwegian Cancer Society, and the South-Eastern Norway Regional Health Authority. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

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

* E-mail: heidi.lyng@rr-research.no

Introduction

Cervical cancer is one of the most common malignancies affecting women worldwide and a major cause of cancer death for women globally [1]. Radiotherapy combined with cisplatin is the treatment of choice at the locally advanced stages [2]. Improved therapy is needed, since more than 30% of the patients show progressive disease within 5 years after diagnosis and treatment related side effects to organs within the pelvis are frequent. Tumor stage, size, and lymph node involvement are the most powerful markers of aggressive disease, but do not fully account for the observed variability in outcome and are not biologically founded. A better handling of the disease may be provided by the discovery of efficient biomarkers for therapeutic planning and intervention, but requires more insight into the mechanisms underlying cervical carcinogenesis and treatment relapse.

During carcinogenesis, genetic and epigenetic alterations drive the evolution of tumor towards increased malignancy and treatment resistance. The changes enable tumor cells to overcome microenvironmental constraints, sustain proliferation, and invade adjacent tissues and distinct organs [3–5]. Gene dosage alterations like gains and losses regulate the expression of genes and are motive forces for this evolution [6,7]. Tumor cells bearing an increasing number of gains and losses successively emerge and are selected for based on the growth advantage caused by the genetic changes. Discovery and functional assessment of gene dosage alterations involved in carcinogenesis are therefore essential for understanding the biology of the disease.

At the locally advanced stages of cervical cancer, numerous gene dosage alterations and severe aneuploidy are frequently seen [8–10]. Moreover, pronounced intratumor heterogeneity in the gains and losses exists within the tumors, reflecting a high genetic instability [9]. The consequences of these alterations for the tumor phenotype are difficult to predict, since large chromosomal regions involving multiple genes are generally affected and some aberrations may be random events without biological significance [11]. Genome wide screening of DNA copy numbers in a decent number of patients enables identification of recurrent gene dosage

Author Summary

Genetic gains and losses, i.e. changes in gene dosages, are common abnormalities of human cancers. Discovering these defects and understanding the biological meaning can lead to improved therapeutic opportunities. This paper reports a large scale screening of gene dosage alterations in cervical cancer and gives a broader exploration of the expression and function of genes with gains or losses. We have focused on the most frequent gene dosage alterations and the alterations associated with survival after chemoradiotherapy, since these defects are likely to be of major importance for developing disease. The most notable finding was the discovery of a set of biological processes that are known hallmarks of cancer and were associated with gains and losses of specific genes. Moreover, novel loci associated with chemoradioresistance independent of existing clinical markers were found, and the genes involved were depicted. Our results indicated that gene dosage alterations play a causative role in the carcinogenesis and chemoradioresistance of cervical cancer and pinpointed candidate biomarkers of the disease.

alterations; *i.e.*, alterations characteristic of the disease, and alterations associated with the clinical outcome [12], which are likely to be important in carcinogenesis and treatment resistance. Combining the data with expression profiles of the same tumors reveals the genes that are regulated primarily by the genetic events. The potential of this integrative strategy was recently demonstrated in a study on 15 early stage cervical cancers, where genes affected by aberrations on 1q, 3q, 11q, and 20q were reported [13]. Genetic events promoting tumor evolution and treatment resistance have, however, not been explored on a genome wide scale, and their biological meaning has not been addressed.

The present work was conducted to discover candidate driver genes and assess their function in the carcinogenesis and chemoradioresistance of cervical cancers. Genome wide screening of DNA copy numbers and expressions was performed in 102 patients with locally advanced disease. Of these, pairwise data were available for 95 patients. Reliable comparison of gains and losses across the patients was ensured by using the tumor ploidy, as determined by flow cytometry, and the GeneCount method to correct for the normal cell content of the samples and extract the absolute copy numbers and thereby the gene dosages [14]. The use of GeneCount also enabled mapping of the intratumor heterogeneity in the gene dosage alterations, providing information of the chronological order in which they had occurred during tumor evolution [14]. The recurrent gene dosage alterations, the alterations associated with outcome after chemoradiotherapy, and the genes that were regulated by these alterations were identified. Further analysis of gene ontology (GO) categories [15] was performed to identify biological processes that were overrepresented among the affected genes and therefore probably regulated by the gene dosage alterations. Such large scale and combined genomic, transcriptional, and functional analysis is powerful in detection of driver genes and their biological meaning, but has not been presented before. We demonstrate the potential of this approach by the identification of five biological processes in carcinogenesis that were associated with recurrent and predictive gains and losses of a set of genes. The set included four genes within the predictive losses for which repressed expression was related to poor outcome in the patient group and in an independent cohort of 41 patients. The genes are candidate drivers of the genetic events and novel biomarkers of cervical cancers.

Results

Recurrent Gene Dosage Alterations

Cervical cancer patients subjected to curative chemoradiotherapy were included in the study (Table 1). Most cases were squamous cell carcinoma and human papillomavirus (HPV) positive. Aneuploidy was seen in about half of the tumors, including some of the adenosquamous carcinomas and HPV negative cases (Figure S1A, S1B). Based on 97 patients, we generated an absolute gene dosage profile of the cancer genome by the use of array comparative genomic hybridization (aCGH) and

Table 1. Patient and tumor characteristics.

Characteristic	Basic cohort (n = 102)	Validation cohort (n = 41)
Histology (n)		
Squamous	96	40
Adenocarcinoma	1	0
Adenosquamous carcinoma	5	1
HPV status (n)[a,b]		
HPV16	65	35
HPV18	7	0
HPV16+18	11	1
HPV other	10	4
HPV negative	8	1
FIGO stage (n)		
1B	6	2
2	57	27
3	35	9
4A	4	3
Tumor size[c]: vol (cm^3)[d], diameter (cm)[e]		
Median	45.1, 4.4	36.6, 4.1
Range	2.8–321, 1.8–8.5	8.7–192, 2.5–7.2
Pelvic lymph node status[c] (n)		
Positive	37	12
Negative	65	29
Age (years)		
Median	56	55
Range	28–85	25–81
Observation time (months)		
Median	42	31
Range	21–71	24–46
Relapse	32	12

[a]PCR on DNA was performed, using the primers listed in [9]. The products were detected by polyacrylamide gel electrophoresis or the Agilent DNA 1000 kit (Agilent Technologies Inc., Germany).
[b]HPV status was not determined for one patient in the basic cohort due to lack of DNA for analysis.
[c]Tumor size and lymph node status were determined from pretreatment magnetic resonance (MR) images.
[d]Volume was calculated based on 3 orthogonal diameters (a,b,c) as $(\pi/6)*abc$.
[e]Diameter was calculated from tumor volume $(4\pi/3)*r^3$.
doi:10.1371/journal.pgen.1000719.t001

the GeneCount analysis tool (Figure 1A). All chromosomes were affected with gains and losses, however, some regions were more frequently found to be aberrant than others (Figure 1B). Clustering of the patients based on gene dosages revealed no clear groups with characteristic aberrations.

The recurrent gains and losses were identified by considering both the amplitude and frequency of each alteration in Figure 1B [16]. Hence, a larger weight was given to high-amplitude events that are less likely to be random aberrations without biological significance. The recurrent alterations comprised more than 42% of the genome, and consisted of 14 regions (528 Mb) with gain and 15 (734 Mb) with loss (Figure 1C). Most of these alterations were also seen in the adenosquamous carcinomas and the HPV negative tumors (Figure S1C, S1D). The most common alterations were gain on 1q, 3q, 5p, 20q, and Xq and loss on 2q, 3p, 4p, 11q, and 13q, each involving 44–76% of the patients (Figure 1C, Table 2). High level amplification (seven regions) and homozygote deletion (six regions) helped to depict the peak of five recurrent gains and two recurrent losses (Table 2, Table S1). The frequency of the homozygote deletions was low (1–3%, Table S1), and none

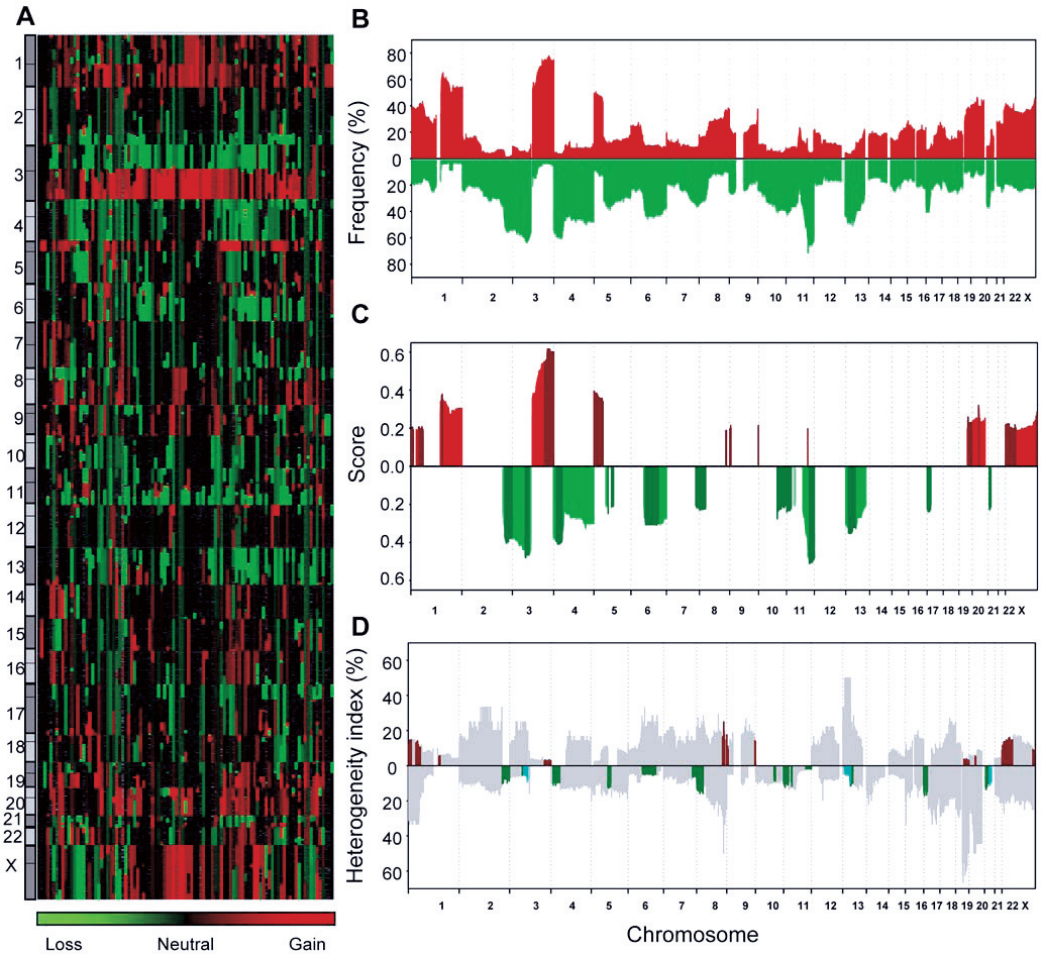


Figure 1. Gene dosage alterations of locally advanced cervical cancers. (A) Absolute gene dosage profile of 97 patients. Patients are shown in columns and gene dosages are ordered by DNA location in rows. The color scale ranges from green (loss) through black (neutral) to red (gain). Grey indicates missing values. (B) Frequency of patients with gains (red) and losses (green) along chromosome 1 to X for the patients in (A). Gene dosage alterations above 1.1 and below 0.9 were classified as gains and losses, respectively. (C) Score of recurrent gains (red) and losses (green) along chromosome 1 to X for the patients in (A). Peak regions, defined in Table 2, are shown in darker colors. (D) Intratumor heterogeneity in gene dosage alterations along chromosome 1 to X for the patients in (A). The heterogeneity index is shown for gains (above the zero line) and losses (below the zero line) separately, and was calculated as the number of heterogeneous cases relative to the total number of cases with alteration at each DNA location. The peak regions shown in (C) are marked in red (recurrent gains) and green (recurrent losses). The predictive losses are indicated in light blue.

doi:10.1371/journal.pgen.1000719.g001

Table 2. Gene dosage alterations and correlating genes in locally advanced cervical cancer.

Peak region[a] (Cytoband)	Peak region[a] (MB)	Freq.[b] (%)	Max./min. gene dosage[c] (copy no.)	Correlating genes[d]
Recurrent gain				
1p36.21-pter	0–14.6	38	2 (4)	*SLC35E2, UBE4B, AGTRAP*
1p32.1-p34.3	37.3–59.9	40	2 (4)	*C1orf149, YRDC, RLF, EBNA1BP2, TACSTD2*
1q21.1-q22	148.0–153.7	61	2.5 (6)	*SF3B4, ENSA, GOLPH3L, ARNT, LASS2, ANXA9, POGZ, CGN, SNX27, C1orf77, ILF2, DENN4B, SLC39A1, UBE2Q1, EFNA1, KRTCAP2, MUC1, FDPS*
3q26.1-qter[e]	166.2–199.5	75	4.5 (9)	*PDCD10, PHC3, ZNF639, FXR1, PARL, DVL3, ABCF3, ALG3, EIF4G1, SFRS10, DGKG, EIF4A2, RFC4, CCDC50, PPP1R2, PAK2, NCBP2, DLG1, BDH1, FLYTTD1*
5p15.2-pter[e]	1.0–12.1	47	4 (15)	*CLPTM1L , MED10, FASTKD3, CCT5, DAP*
8q24.13-22	125.7–134.1	37	2 (4)	None
8q24.3-qter	144.5–146.3	38	2 (4)	*TSTA3, FAM83H, CYC1*
9p24.1-2[e]	2.7–6.0	22	13.5 (27)	*KIAA0020, RCL1*
9q34.2-qter	135.6–138.2	35	3.5 (7)	*MRPS2*
11q22.1-2[e]	100.2–102.0	14	36 (72)	*YAP1, BIRC3, BIRC2*
19q13.11-qter	40.3–63.8	36	10 (29)	*SPINT2, PSMD8, CAPN12, MRPS12, RPS16, AP2S1, KDELR1, NUP62, ATF5, NKG7, ZNF787*
20q11.21-22[e]	30.0–33.0	45	3.4 (9)	*POFUT1, KIF3B, MAPRE1, SNTA1, EIF2S2, AHCY*
Xp11.22-pter[f]	0–54.1	38	2.5 (5)	*SLC25A6, CD99, ARSD, PNPLA4, PRPS2, PIR, CXorf15, PHKA2, PDHA1, RPS6KA3, PRDX4, EIF2S3, USP9X, DDX3X, NDUFB11, UBA1, EBP, PLP2, JARID1C, SMC1A, HUWE1*
Xq28-qter	148.5–154.9	47	4 (8)	*NSDHL, BCAP31, IDH3G, IRAK1, TAZ, LAGE3, UBL4A, FAM34, MTCP1*
Recurrent loss				
2q33.3-qter	206.2–243.0	54	0.26 (1)	*NDUFS1, SPAG16, MREG, SMARCAL1, AAMP, WNT10A, ZFAND2B, ANKZF1, STK11IP, FARSB, ACSL3, HRB, SP100, EIF4E2, COPS8, HDAC4, MTERFD2, PPP1R7*
3p12.3-p14.2	60.9–81.6	61	0.26 (1)	*RYBP, GBE1*
4p13-p16.1	8.3–42.3	58	0.42 (1)	*WDR1, UBE2K, PDS5A*
5q13.2[g]	67.4–71.7	38	0 (0)	*SMN2*
5q14.2-q15	82.5–96.9	35	0.5 (1)	*COX7C, TTC37, GLRX*
6q12-q23.2	67.0–132.9	42	0.43 (1)	*LMBRD1, MYO6, HMGN3, SYNCRIP, MAP3K7, CCNC, C6orf203, FOXO3, AMD1, HDAC2, NT5DC1, DSE, NUS1, ECHDC1*
7q34-qter	139.3–158.8	35	0.43 (1)	*PDIA4*
8p12-pter	0–31.9	32	0.34 (1)	*XPO7, BIN3, BNIP3L, EPHX2, CCDC25, DCTN6, PPP2CB*
10q23.31[g]	88.2–92.1	38	0 (0)	None
11p14.3-pter	0–24.4	40	0.5 (1)	*COPB1, PSMA1, GTF2H1, TSG101*
11p12	37.8–40.2	37	0.5 (1)	None
11q22.3-qter	105.1–134.5	63	0.35 (1)	*PPP2R1B, C11orf57, TIMM8B, REXO2, C11orf60, TRAPPC4, H2AFX, POU2F3, ARHGEF12, SC5DL, ZNF202, CHEK1, APLP2, ZBTB44, SNX19*
13q12.2-q21.32	27.5–67.4	46	0.33 (1)	*ALG5, FAM48A, COG6, KIAA1704, GTF2F2, MED4, RNASEH2B*
17p11.2-pter	0–19.1	38	0.27 (1)	*SPAG7, MPDU1, LSMD1, CYB5D1, COPS23*
21q21.1-3	18.3–28.6	35	0.28 (1)	*ATP5J*
Predictive loss				
3p11.2-p14.1	67.0–87.6	58	0.26 (1)	*RYBP, GBE1*
13q13.1-q21.1	30.0–56.5	46	0.41 (1)	*ALG5, FAM48A, COG6, KIAA1704, GTF2F2, MED4, RNASEH2B*
21q22.2-3	38.0–46.4	23	0.28 (1)	*PCP4, RIPK4, PDXK*

[a]Peak region of the recurrent gains and losses is the minimum shared region surrounded by at least three patients. In cases of recurrent high level amplification or homozygote deletion, this event determines the peak region. Peak region of the predictive losses is the region selected by LASSO.
[b]Frequency is the median percentage of tumors with the alteration.
[c]Gene dosage is absolute DNA copy number divided by ploidy. Maximum (gain) or minimum (loss) gene dosage and corresponding copy number are listed.
[d]Genes within the peak region showing a correlation between gene dosage and expression are ordered by DNA location.
[e]Recurrent high level amplification detected within recurrent gain. Peak region is the region with more than 25% higher amplitude than surrounding region.
[f]Probably two different peak regions.
[g]Homozygote deletion within recurrent loss. Peak region is the region with a gene dosage of zero.
doi:10.1371/journal.pgen.1000719.t002

of the tumors had more than one of them. Homozygote alteration is therefore probably not a common mechanism of gene regulation in cervical cancers, in contrast to the highly frequent heterozygote deletion. The highest gene dosage of 36 was found in a diploid tumor with a copy number of 72 on 11q22.1-2 (Table 2).

Intratumor heterogeneity of the recurrent alterations. Intratumor heterogeneity in one or more of the gene dosage alterations was seen in about half of the patients [14]. The ploidy and genetic alterations of the heterogeneous tumors were similar to that of the homogeneous ones (Figure S2). It is reasonable to assume that homogeneous alterations have emerged before the heterogeneous ones during tumor evolution [9]. To order the recurrent alterations chronologically in relation to the less common alterations, we therefore mapped the intratumor heterogeneity along the chromosomes based on the absolute data achieved with GeneCount [14]. The heterogeneity was low for the recurrent alterations compared to others, like gain on 2q and 13q and loss on 1q, 19q, and 20q (Figure 1D). The recurrent aberrations had therefore probably occurred prior to many of these less common events.

Gene Dosage Alterations in Relation to Outcome after Chemoradiotherapy

Gene dosage alterations responsible for poor clinical outcome may not be as common as the recurrent ones. All alterations in Figure 1B were therefore included in the survival analysis. The LASSO method identified three regions with loss, 3p11.2-p14.1, 13q13.1-q21.1, and 21q22.2-3, which jointly showed the strongest association to progression free survival (Table 2). The 3p11.2-p14.1 and 13q13.1-q21.1 regions overlapped with the recurrent 3p12.3-p14.2 and 13q12.2-q21.32 losses, whereas the predictive loss of 21q22.2-3 was distal of the recurrent loss of 21q21.1-3. The predictive losses were not correlated and were related to poor outcome also when analyzed separately (Figure 2A–2C). The intratumor heterogeneity of the losses was low and similar to that of the recurrent losses (Figure 1D).

Most patients had more than one of the predictive 3p, 13q, and 21q losses. We therefore investigated whether there was an increased risk of relapse in cases of two or three losses. Kaplan-Meier plots for patients with different combinations of the predictive losses revealed three major groups with different outcome (Figure S3). Patients without any of the losses had a low risk of relapse and a survival probability of 91% (Figure 2D). Patients with 3p and/or 13q loss, without 21q loss, had an intermediate survival probability of 68%, whereas those with 21q loss had the lowest survival probability of 44%. The risk of relapse therefore seemed to be particularly high when loss of 21q22.2-3 was involved.

The predictive impact of the 3p, 13q, and 21q losses were assessed by multivariate analysis together with tumor size, stage, and lymph node status. Histological type, HPV status, and heterogeneity status showed no correlation to outcome in univariate analysis and were therefore not included. The losses and tumor size had independent predictive value (Table 3), showing that the gene data contained information of the progression free survival that was not covered by tumor size. Since tumor size is a strong predictor (Figure 3A), we also investigated the predictive impact of the three losses for small and large tumors separately. About 20% of the patients with tumor size less than the median had relapse and all of them had one or more of the losses (Figure 3B). In the cases of tumors larger than the median, about 47% of the patients progressed and all except two of them had one or more of the losses (Figure 3C). None of the patients with loss involving 21q were disease free after 28 months, suggesting a particularly high risk of relapse in cases of a large

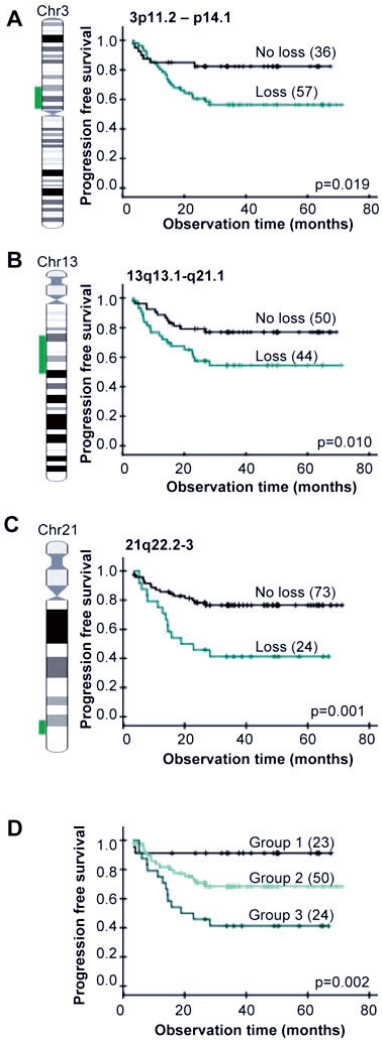


Figure 2. Gene dosage alterations and outcome after chemoradiotherapy. Kaplan-Meier curves of progression free survival for cervical cancer patients with (green) and without (black) loss of 3p11.2-p14.1 (A), 13q13.1-q21.1 (B), 21q22.2-3 (C), and for patients with different combinations of the three losses (D). P-values in log-rank test and number of patients are indicated. Data of the most significant genomic clone within each region were used; *i.e*, BAC clone ID RP11-118O11 (3p), RP11-408L13 (13q), and RP1-128M19 (21q). Total number of patients in (A, B) is less than 97 due to missing gene dosage data. (A–C) The lost DNA region is indicated on the chromosome (*left*). (D) Group 1: patients without loss of 3p11.2-p14.1, 13q13.1-q21.1, or 21q22.2-3, group 2: patients with loss of 3p11.2-p14.1 and/or 13q13.1-q21.1, but not 21q22.2-3, group 3: patients with loss of 21q22.2-3 only or loss of 21q22.2-3 combined with loss of 3p11.2-p14.1 and/or 13q13.1-q21.1. The groups were determined from data of each possible combination of the losses (Figure S3).
doi:10.1371/journal.pgen.1000719.g002

Table 3. Cox regression analysis of genetic losses and clinical variables.

	Univariate analysis[a]			Multivariate analysis[a]		
Covariate	**P**	**HR**	**95% CI**	**P**	**HR**	**95% CI**
Loss of 3p11.2-p14.1[b]	0.003	0.27	0.11–0.66	0.018	0.33	0.13–0.83
Loss of 13q13.1-q21.1[b]	0.006	0.32	0.14–0.72	0.015	0.35	0.14–0.82
Loss of 21q22.2-3[b]	0.004	0.34	0.16–0.71	0.019	0.32	0.12–0.84
Tumor size[c]	0.001	4.5	1.9–10.5	0.001	5.5	1.9–15.5
FIGO stage[d]	0.004	2.9	1.4–5.9	0.072	-	-
Total lymph node status[e]	0.030	0.46	0.22–0.93	0.285	-	-

[a]P-value (P), hazard ratio (HR), and 95% confidence interval (CI) are listed.
[b]Semi-discrete gene dosage data of the most significant genomic clone within each region were used.
[c]Tumor size was divided in two groups based on the median size of 45.1 cm^3, corresponding to a median diameter of about 4.4 cm.
[d]FIGO stage was divided in two groups; 1b–2b and 3a–4a.
[e]Total includes pelvic and para aortal lymph nodes.
doi:10.1371/journal.pgen.1000719.t003

tumor bearing loss of 21q22.2-3. There was no difference in tumor size for patients with and without loss in Figure 3B or in Figure 3C (data not shown). The gene data therefore enabled identification of high and low risk patients both in cases of a small and a large tumor.

Integration of Gene Expression

To find genes regulated by the recurrent and predictive gene dosage alterations, we used cDNA microarrays and generated a cancer gene expression profile. The profile was based on 100 patients, including 95 of those analyzed with aCGH. Expression data were available for 1357 of the about 4000 known genes within the altered regions, and a significant correlation to gene dosage was found for 191 of them (Table 2). Several correlating genes were identified for each region, except for 8q24.13-22, 10q23.31, and 11p12, where no genes were found. Typical examples of correlation plots are shown in Figure S4. The results were confirmed with the Illumina gene expression assay on 52 patients. Although the Illumina analysis was based on a lower number of patients, an excellent correlation between the Illumina and cDNA data and between the Illumina and gene dosage data was found for almost all of the genes, as demonstrated in Table S2. We also performed a second cDNA analysis, including only tumors with more than 70% tumor cells in hematoxylin and eosin (HE) stained sections. Totally 179 of the genes (94%) were identified, suggesting few false positive results due to normal cells in the samples. The observations supported our conclusion that the genes in Table 2 were gene dosage regulated. The latter analysis identified 26 genes that were not depicted when all patients were considered. These genes were not considered further, since the results were based on only half of the data set.

Expression of known oncogenes and tumor suppressor genes within the depicted regions, like *MYC* (8q24.21), *BRCA2* (13q13.1), *RB1* (13q14.2), and *TP53* (17p13.1), was not significantly correlated to gene dosage. These genes are therefore probably not regulated primarily by gains and losses. The *TP53* and *RB1* results were consistent with the high frequency of HPV positive tumors (Table 1).

The predictive losses on 3p and 13q involved the same correlating genes as the corresponding recurrent ones, whereas *PCP4*, *RIPK4*, and *PDXK* were correlating genes within the

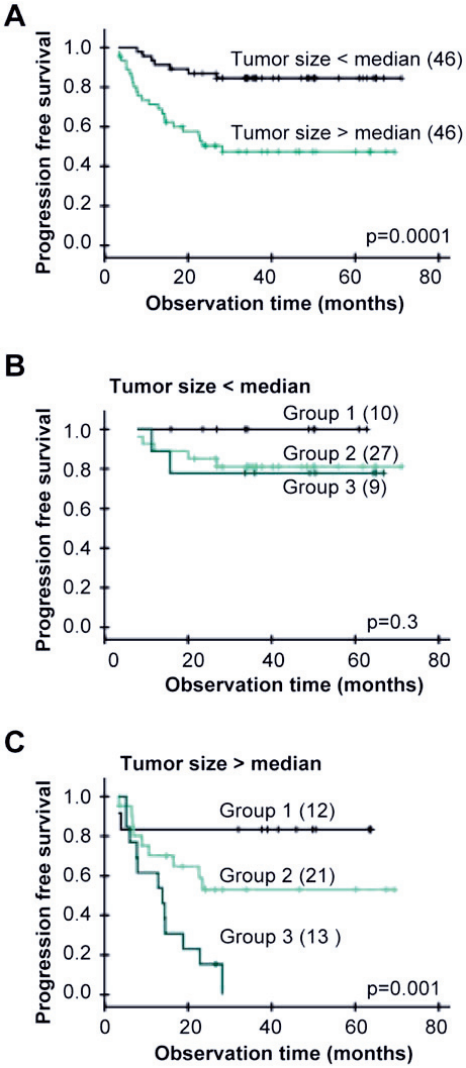


Figure 3. Gene dosage alterations and outcome after chemoradiotherapy for patients with different tumor size. (A) Kaplan-Meier curves of progression free survival for cervical cancer patients with tumor size above (green) and below (black) median. Ninety-two patients with tumor size determined from diagnostic MR images were included. Median size was 45.1 cm^3, corresponding to a diameter of 4.4 cm. (B,C) Kaplan-Meier curves for patients in (A) with tumor size below median (B) and above median (C). Group 1: patients without loss of 3p11.2-p14.1, 13q13.1-q21.1, or 21q22.2-3, group 2: patients with loss of 3p11.2-p14.1 and/or 13q13.1-q21.1, but not 21q22.2-3, group 3: patients with loss of 21q22.2-3 only or loss of 21q22.2-3 combined with loss of 3p11.2-p14.1 and/or 13q13.1-q21.1. The groups were determined from data of each possible combination of the losses (Figure S3). P-values in log-rank test and number of patients are indicated.
doi:10.1371/journal.pgen.1000719.g003

predictive 21q region (Table 2). To depict the correlating genes that most probably were involved in development of chemoradioresistance, we required that the gene was significantly associated with clinical outcome both at the gene dosage and expression level. Moreover, a clear difference in the survival curves should also be seen in an independent cohort of 41 patients when based on the Illumina gene expression data. The criteria were fulfilled for four genes; *RYBP* and *GBE1* on 3p and *MED4* and *FAM48A* on 13q, which were termed predictive genes (Figure 4). Two more genes, *GTF2F2* and *RNASEH2B* on 13q, were correlated to outcome based on the cDNA data, but were not considered further since the tendency based on the Illumina data was weak ($p>0.15$). The relationship to outcome was not strong enough for *PCP4*, *RIPK4*, and *PDXK* on 21q to be included among the predictive genes either.

Gene Ontology Analysis

Biological processes associated with the recurrent and predictive gene dosage alterations were found by comparing the GO categories of the affected genes with those of all genes in the data set [15]. One or more biological processes were annotated to 155 of the correlating and predictive genes and to 5824 of all genes. The categories apoptosis, carbohydrate metabolism, translation, and RNA-protein complex biogenesis and assembly were significantly overrepresented among the correlating genes within the recurrent gains, whereas macromolecule localization, generation of precursor metabolites and energy, transcription from RNA polymerase II promoter, and establishment or maintenance of chromatin architecture were overrepresented among those within the recurrent and predictive losses (Table 4). Fifty-six genes were included in the significant categories and were candidate drivers of the biological processes. In addition, we included the predictive gene *FAM48A*, which was not associated to any process in the GO database, as a potential driver of chemoradioresistance together with *RYBP* and *MED4* (transcription) and *GBE1* (generation of precursor metabolites and energy).

We generated a map to visualize the connections between genetic events, affected genes, and biological processes (Figure 5). The processes carbohydrate metabolism and generation of precursor metabolites and energy were combined in metabolism, translation and RNA-protein complex biogenesis and assembly were combined in translation, and transcription from RNA polymerase II promoter was combined with establishment or maintenance of chromatin architecture in transcription. The combined categories were closely related, justifying this strategy. All but six of the recurrent alterations were associated with a process and represented in the map. The predictive 3p and 13q losses were merged with the corresponding recurrent losses, since the regions overlapped, and linked to metabolism (*GBE1*) and transcription (*RYBP*, *MED4*) in addition to chemoradioresistance. The predictive 21q loss was not connected to any known gene, but associated with chemoradioresistance. The map revealed features that seemed to be characteristic of recurrent and predictive alterations in cervical cancer. First, many of the genetic events were associated with clusters of genes in the same biological

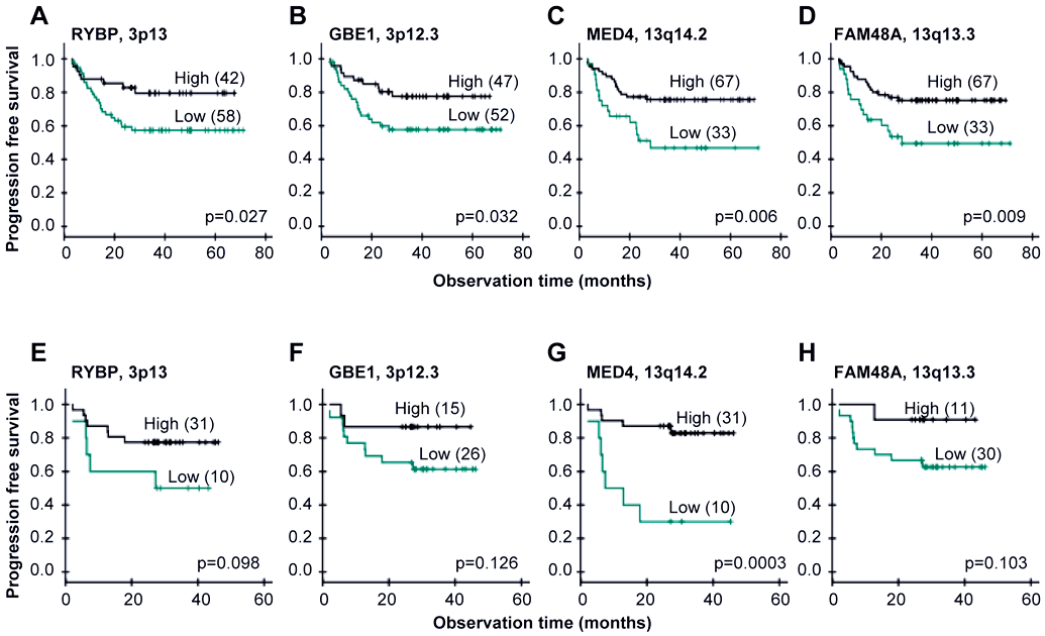


Figure 4. Gene expressions and outcome after chemoradiotherapy. Kaplan-Meier curves of progression free survival for cervical cancer patients with low (green) and high (black) expression of *RYBP* (A,E), *GBE1* (B,F), *MED4* (C,G), and *FAM48A* (D,H). cDNA data of 100 patients is used in (A–D), and Illumina data of an independent cohort of 41 patients is used in (E–H) for validation. P-value in log-rank test and number of patients are indicated. The number of patients in each group was chosen to achieve the largest difference in survival between the groups, approximately reflecting the number of patients with and without loss in (A–D). Total number of patients is less than 100 in (B) due to missing gene expression data.
doi:10.1371/journal.pgen.1000719.g004

Table 4. Biological processes overrepresented among the correlating genes within recurrent and predictive regions.

GO number	GO category	No. correlating genes	No. genes on the array	p-value	Correlating genes
Gains					
GO: 000815	Biological process	93[a]	5824[a]		
GO: 0006915	Apoptosis	13 (14.0%)	434 (7.5%)	0.026	*UBE4B, BIRC2, BIRC3 , ATF5, BCAP31, CLPTM1L, DAP, FASTKD3, FXR1, NUP62, PAK2, PDCD10, SLC25A6*
GO: 0005975	Carbohydrate metabolism	7 (7.5%)	198 (3.4%)	0.038	*PPP1R2, ARNT, PHKA2, POFUT1, PDHA1, TSTA3, IDH3G*
GO: 0006412	Translation	7 (7.5%)	163 (2.8%)	0.015	*EIF4G1, EIF4A2, EIF2S2, MRPS12, RPS16, EIF2S3, MRPS2*
GO: 0022613	RNA-protein complex biogenesis and assembly	7 (7.5%)	89 (1.5%)	0.001	*EIF4G1, EIF4A2, EIF2S2, EIF2S3, EBNA1BP2, NCBP2, RCL1*
Losses					
GO: 000815	Biological process	62[a]	5824[a]		
GO: 0033036	Macromolecule localization	10 (16.1%)	427 (7.3%)	0.022	*BIN3, COPB1, COG6, XPO7, HRB, MYO6, PDIA4, SNX19, TIMM8B, TSG101*
GO: 0006091	Generation of precursor metabolites and energy	4 (6.5%)	117 (2.0%)	0.035	*ATP5J, COX7C, GBE1, NDUFS1*
GO: 0006366	Transcription from RNA polymerase II promoter	10 (16.1%)	357 (6.1)	0.004	*RYBP, FOXO3, GTF2F2, GTF2H1, MED4, MYO6, POU2F3, SMARCAL1, ZNF202, HDAC4*
GO: 0006325	Establishment or maintenance of chromatin architecture	5 (8.1%)	140 (2.4%)	0.016	*DSE, H2AFX, HDAC2, SMARCAL1, HDAC4*

[a]Genes with GO annotation (biological process).
doi:10.1371/journal.pgen.1000719.t004

process. For example, gain on 3q affected three genes in apoptosis and three in translation, gain on 5p was linked to tree apoptosis genes, and loss on 6q was associated with four genes in transcription. Second, several events, like gain on 3q, 19q, 20q and loss on 2q, 6, and 11q, were connected to more than one biological process, either through the regulation of several genes or because some genes had multiple functions.

Discussion

This work presents the first coupling of gene dosage and expression profiles in a large sample set of cervical cancers. We based our study on absolute gene dosages, which are more sensitive than the commonly used aCGH ratios in detecting gains and losses and enable comparisons across tumors with differences in ploidy and normal cell content [14]. This strategy and the large number of patients ensured reliable identification of recurrent gene dosage alterations, events associated with clinical outcome, and their intratumor heterogeneity. Further analysis based on GO categories provided an objective way of organizing the numerous correlating genes into biological meaningful information. We demonstrate a large potential of the integrative approach by the discovery and functional assessment of candidate driver genes that represent novel biomarkers of the disease. In particular, novel loci associated with clinical outcome were identified, providing the first evidence that gene dosage can be responsible for developing chemoradioresistance in cervical cancers.

The recurrent gene dosage alterations were consistent with earlier reports on advanced stage cervical cancer based on conventional CGH [8,9,17]. However, a more precise definition of the altered regions was achieved here due to the improved resolution of the array technique. The high frequency of the alterations suggests that they play a causative role in carcinogenesis. Hence, many of the alterations are common also in other squamous cell carcinomas, like head and neck cancers [18,19]. Moreover, the recurrent loss on 3p and 13q overlapped with the losses associated with poor clinical outcome, strengthening the hypothesis of a central role in tumor evolution. Less frequent alterations can, however, also be crucial for tumor evolution, as was demonstrated by the recurrent gain on 11q22 in 14 patients and predictive loss on 21q in 23 patients.

The low intratumor heterogeneity of the recurrent and predictive gene dosage alterations indicated that they had occurred prior to many of the other alterations. The result was consistent with our previous cervical cancer study based on conventional CGH [9], showing a homogeneous intratumor distribution of the frequent gains on 3q, 5p, and 20q and losses on 3p and 11q14-qter. Moreover, regions overlapping with the 1p, 1q, 3q, 8q, 9q, and 20q recurrent gains and 2q, 3p, 4p, 11q, and 17p losses have been found to be altered in precancerous cervical intraepithelial lesions [17,20–23], suggesting that the events had occurred at an early stage. It is therefore likely that the alterations identified here, and the consequently control of biological processes and development of chemoradioresistance, emerge early during carcinogenesis. It should be noted that a low heterogeneity was seen for some of the less common alterations as well, implying that they had occurred early. The affected genes in these regions may also be crucial for tumor evolution, however, other mechanisms than gene dosage alterations, such as epigenetic events or mutations, probably play the major role in their regulation. Moreover, some of the highly heterogeneous alterations may be important for disease progression a later stage, being a result of the continuing tumor evolution towards increased aggressiveness.

The gene dosage alterations were associated with specific biological processes that are closely related to known cancer

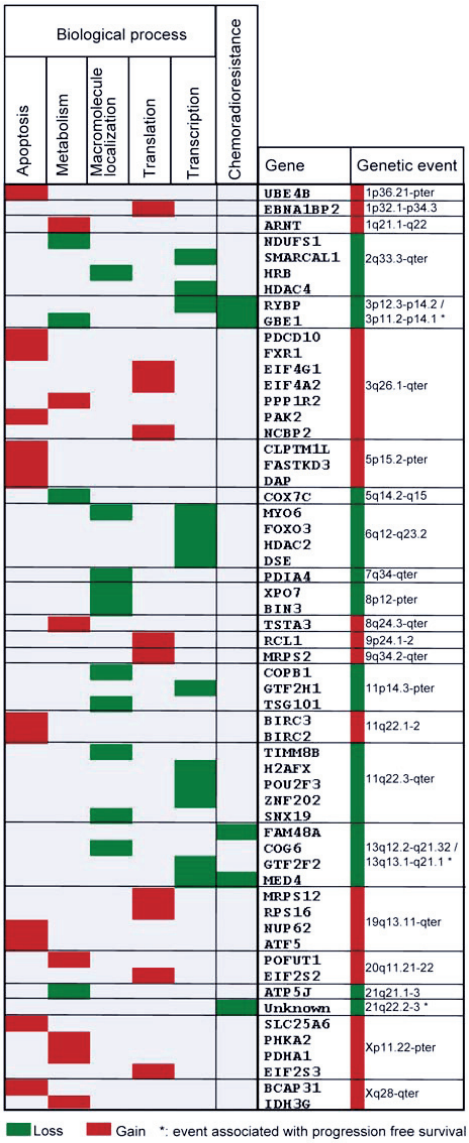


Figure 5. Genetic events, correlating genes, and biological processes in carcinogenesis and chemoradioresistance of cervical cancers. Recurrent and predictive gene dosage alterations, correlating genes, and biological processes overrepresented among the genes are listed. Only the genetic events associated with a process or chemoradioresistance (*) are included; six of the recurrent alterations are therefore not shown. The genes are ordered by DNA location. Correlating genes connected to chemoradioresistance were associated with clinical outcome both at the gene dosage and expression level and validated in an independent patient cohort. Gains and losses are indicated with red and green color, respectively.
doi:10.1371/journal.pgen.1000719.g005

hallmarks [3–5], indicating that the genes involved are drivers of carcinogenesis. Hence, gain of the genes in apoptosis, including the anti-apoptosis genes *BIRC2*, *BIRC3*, and *ATF5*, can help carcinoma cells to evade apoptosis [3]. Aberrations of the genes in metabolism, like gain of *ARNT* and *IDH3G* in carbohydrate metabolism, and loss of *COX7C* and *ATP5J* in oxidative phosphorylation, can be part of a metabolic reprogramming towards increased glycolysis and decreased mitochondrial function to meet the high energy demand linked to tumor growth [4]. In particular, gain of *ARNT* may increase hypoxia and hypoglycemia tolerance by signaling through the *HIF1A* pathway [24]. Loss of the genes in molecular localization, including *HRB* and *TSG101*, can lead to abnormal protein internalization and recycling and thereby abrogated degradation of proteins like growth factor receptors [25,26]. Finally, aberrations of the genes in translation and transcription, such as gain of the translation initiation factors *EIF4A2*, *EIF4G1*, *EIF2S2*, and *EIF2S3* and loss of the transcriptional repressors *HDAC2* and *HDAC4*, can be a way to control the formation and activity of essential proteins. The EIF-proteins are central in adaptation to hypoxia and can stimulate *MYC* translation and thereby oncogenic processes like cell proliferation [27,28]. Improper function of *HDAC2* and *HDAC4* may also increase proliferation [29]. Many of the genes, including *BIRC2*, *BIRC3*, *ATF5*, *NUP62*, *FASTKD3*, *IDH3G*, and *POFUT1*, have been found to be regulated by gains or losses in previous cervical cancer studies [30–33]. Our findings link each gene to one or more specific biological processes, and thereby indicate the functional meaning of the genetic events in carcinogenesis.

Loss and down regulation of *GBE1* and *RYBP* on 3p and *MED4* and *FAM48A* on 13q were associated with poor clinical outcome, suggesting that the genes are drivers of chemoradioresistance. The mechanisms underlying these findings and possible associations to known aggressive phenotypes like hypoxia and rapid proliferation [34–36] are not clear, but a tumor suppressor function of the genes has been indicated. *GBE1*, which plays a role in carbohydrate metabolism, has been found to be down regulated in ovarian cancers [37]. Loss of the transcriptional repressor *RYBP* may impair death receptor-mediated apoptosis [38,39], and the encoded protein has been shown to be down regulated in many tumor types, including cervical cancer [40]. Loss of the transcriptional activators *MED4* may impair transcription of genes with anti-cancer effect, like the vitamin D receptor [41,42]. The function of *FAM48A* is less clear, but some studies indicate that loss of this gene can promote aggressiveness. Hence, *FAM48A* is required for activation of the MAPK p38 pathway [43], which represses cell proliferation [44]. We found no candidate driver gene of chemoradioresistance within the predictive loss on 21q. Only a few tumor suppressor genes have been identified in this region. One candidate is the transcriptional regulator *PRDM15*, which was not included in our cDNA data set [45]. Our data showed, however, no correlation between *PRDM15* expression, assessed with the Illumina method in 52 patients, and gene dosage (data not shown), suggesting that the gene is not regulated by genetic loss. Further investigation with denser microarrays or possibly microRNA screening would be needed to find the drivers in this region.

The connection between genetic events, genes, and biological processes may provide insight into more general aspects of cervical carcinogenesis. Several genes were often associated with a single genetic event, supporting the hypothesis that there can be multiple drivers of an event that coordinately promote tumor evolution [11]. In cases of genes in the same biological process, like the anti-apoptosis genes *BIRC2* and *BIRC3* on 11q22, a broad and therefore efficient control of the process may be obtained. Hence, *BIRC2* and *BIRC3* may play complementary roles in apoptosis

evasion, since upregulation of *BIRC3*, but probably not *BIRC2*, may impair hypoxia induced apoptosis [46,47]. In cases of genes in different biological processes, such as metabolism (*NDUFS1*), macromolecule transport (*HRB*), and transcription (*SMARCAL1*, *HDAC4*) on 2q, the collective control of these processes through a single event is likely to give a growth advantage that is selected for in carcinogenesis. One or more genes in all biological processes were affected in most tumors due to the high frequency of the recurrent gene dosage alterations. All processes were therefore probably important, and the control of them through gains and losses seems to be a common feature of the disease.

The candidate driver genes represent novel biomarkers that may be utilized in the handling of cervical cancers. Diagnostic assessment of the biomarkers may help to understand the evolutionary status and therefore the biology of the cancer in individual patients. In particular, the predictive biomarkers may be used in addition to tumor size for classification of patients into risk groups in a personalized treatment regime. The biomarkers also open for the possibility to specifically repress biological processes in carcinogenesis by molecular targeting, and thereby interfere with tumor evolution. The use of drugs to inhibit translation by interaction with EIF-proteins has shown promising results [48] and been suggested as a tool to target tumor hypoxia [49]. The approach may be applied at all stages of the disease, since the genetic events probably emerge early. Moreover, improved outcome after chemoradiotherapy might be achieved by targeting the predictive biomarkers. Hence, viral-mediated delivery of *RYBP* has been shown to induce apoptosis in a number of cancer cell lines [38], and could be a useful strategy for the patients with loss of this gene.

Materials and Methods

Patients

A cohort of 102 patients was included for basic analyses to identify gene dosage alterations with aCGH (97 patients), affected transcripts with cDNA microarrays (100 patients), and to confirm the affected transcripts with the Illumina method (52 patients) (Table 1). An independent cohort of 41 patients was used to validate relationships between gene expression and outcome with the Illumina method (Table 1). All patients received external irradiation and brachytherapy combined with adjuvant cisplatin and were followed up as described previously [50]. Eighteen patients received extended radiation field due to enlarged common iliac and para-aortal lymph nodes. Progression free survival, defined as the time between diagnosis and the first event of locoregional and/or distant relapse, was used as end point. Six patients died of causes not related to cancer and were therefore censored. Tumor samples were collected at the time of diagnosis. One – four biopsies, approximately 5×5×5 mm in size, were taken at different locations of the tumor, immediately snap-frozen in liquid nitrogen and stored at −80°C until used for analyses. The study was approved by the regional committee of medical research ethics in southern Norway, and written informed-consent was achieved from all patients.

Array Comparative Genomic Hybridization

The aCGH experiments and generation of absolute gene dosage profiles have been described previously for all 97 patients (ArrayExpress accession no. E-TABM-398) [14]. The array slides were produced at the Microarray Facility at the Norwegian Radium Hospital and contained 4549 unique genomic BAC and PAC clones that covered the whole genome with a resolution of approximately 1 Mb. Genomic DNA was isolated from the biopsies, labeled, and co-hybridized with normal female DNA to the array slides. DNA from different biopsies of the same tumor was pooled. The biopsies of all except two patients had more than 50% tumor cells in HE stained sections from the middle part of the sample. Median tumor cell fraction was 70% (range 30–90%). After array scanning, image analysis, spot filtering, and ratio normalization, the GLAD algorithm was applied for ratio smoothing and breakpoint detection [51].

Absolute gene dosages. The smoothed ratios were transferred to absolute DNA copy numbers in GeneCount by utilizing tumor ploidy data and correcting for the normal cell content of the samples [14]. The tumor ploidy was determined from a separate piece of the biopsy by flow cytometry, and tumor cell fraction was estimated by the program prior to the copy number calculations. The ploidy data and tumor cell fractions have been presented previously [14]. The tumor cell fractions, ranging from 27% to 84%, were in general lower than the results based on HE stained sections, probably because the amount of immune cells infiltrating the tumor parenchyma are difficult to quantify by histological examination [14]. The copy numbers were rounded off to the nearest integer values.

The absolute gene dosage profile of each tumor was generated by dividing each copy number by the ploidy. A gene dosage of 1 therefore implied no change in the copy number. The gene dosage thresholds for scoring gains and losses were 1.1 and 0.9, respectively, taking into account an uncertainty in the ploidy measurement of approximately 10%. For scoring high level amplification, a gene dosage of 2.5 or higher; *i.e.* 5 DNA copies in diploid tumors, was required. Homozygote deletions had a gene dosage of 0.

Intratumor heterogeneity. The intratumor heterogeneity in the copy numbers was assessed by comparing the aCGH ratio distributions of the possible heterogeneous regions with the distributions of the adjacent homogeneous regions by ANOVA analysis [14]. Totally 86 patients had a tumor cell fraction sufficiently high for reliable detection of heterogeneity, and the remaining eleven patients were excluded from this analysis. The heterogeneous regions have been listed previously [14]. A heterogeneity index was calculated for gains and losses separately, as the number of heterogeneous cases relative to the total number of cases with alteration at each DNA location. The copy number of the heterogeneous region was 0.5 above (gain) or below (loss) the nearest integer value.

The GeneCount method has been extensively validated based on the cervical cancer samples included in this study and a cohort of 94 lymphoma samples [14]. In particular, we used lymphoma samples to show that the estimated tumor cell fractions correlate significantly with the highly accurate values determined by flow cytometry [14].

cDNA Microarrays

The cDNA microarray experiments have been presented previously for 48 of the 100 patients [50]. The array slides were produced at the Microarray Facility at the Norwegian Radium Hospital and contained more than 12000 unique cDNA clones, including most known oncogenes and tumor suppressor genes. Total RNA was isolated from the biopsies, labeled, and co-hybridized with reference RNA (Universal Human Reference RNA, Stratagene, La Jolla, CA) to the array slides. RNA from different biopsies of the same tumor was pooled. Only biopsies with more than 50% tumor cells in HE stained sections were utilized. Median tumor cell fraction was 70% (range 50–90%). All hybridizations were performed twice in a dye-swap design (ArrayExpress accession no. E-TABM-817). After array scanning, image analysis, spot filtering, and ratio normalization, the average expression ratios were calculated from the two data sets and used in the further analyses. The gene expressions were mapped to the

gene dosages based on the exact chromosomal position of the cDNA and genomic clones, as derived from Ensembl (http://www.ensembl.org/Homo_sapiens/searchview).

Illumina Gene Expression Beadarrays

Results based on cDNA data were validated with Illumina gene expression beadarrays in 52 of the patients subjected to aCGH and in the independent cohort of 41 patients. HumanWG-6 v3 beadchips (Illumina Inc., San Diego, CA) with 48000 transcripts were used. RNA was isolated from the biopsies as described above and amplified using the Illumina TotalPrep RNA amplification kit (Ambion Inc., Austin, TX) with 500 ng of total RNA as input material. cRNA was synthesized overnight (14 hr), labelled, and hybridized to the chips at 58°C overnight, according to the standard protocol. The hybridized chip was stained with streptavidin-Cy3 (AmershamTM, PA43001, Buckinghampshire, UK) and scanned with an Illumina beadarray reader. The scanned images were imported into BeadStudio 3.1.3.0 (Illumina Inc.) for extraction, quality control, and quintile normalization. The annotation file HumanWG-6_V3_0_R0_11282955_A was used.

Statistics

The recurrent gene dosage alterations were identified based on a score that was calculated for each genomic clone by multiplying the average gene dosage amplitude with its frequency [16]. Gains and losses were considered in two separate procedures. Semi-discrete data were used, for which amplitudes lower than 1.1 were set to 1 when searching for gains and amplitudes higher than 0.9 were set to 1 when searching for losses. The score significance was assessed by comparison to similar scores obtained after data permutation [16], adjusting the p-value by a multiple testing procedure to control the false discovery rate (FDR) [52]. Recurrent alterations with an FDR q-value $<5\%$ were reported.

Gene dosage alterations associated with clinical outcome were identified with the LASSO method in the Cox proportional hazards model [53], as implemented in [54]. The LASSO is a method for variable selection and shrinkage in regression models when the number of covariates is larger than the number of individuals. By performing shrinkage in addition to selection, the LASSO is more stable than stepwise procedures where variables are either retained or discarded from the model sequentially, one at a time. In groups of highly correlated variables the LASSO tends to select only one variable in the group [55], and reported therefore one representative of each DNA region that jointly explained the outcome. Each region was thereafter found by selecting neighbouring genomic clones with strong correlation ($r>0.9$) to the one reported. Survival curves were generated by Kaplan-Meier analysis and compared by using log-rank test.

Spearman's rank correlation analysis with an FDR q-value $<5\%$ was used to search for significant correlations between gene dosage and expression. The analysis was based on semi-discrete data, retrieved as described above. To identify biological processes that were overrepresented among the correlating genes, the GO categories of the genes were compared with those of all genes on the array by using the master-target procedure with the Fisher's exact test in the eGOn software [15]. The GO categories were found in eGOn from public databases, based on the gene reporter EntrezGeneID.

Supporting Information

Figure S1 Tumor ploidy and gene dosage alterations in relation to histological type and HPV status. (A) Ploidy distribution of 97 patients. Tumors with a ploidy within the range of 1.8–2.2 were considered as near diploid. (B) Ploidy of patients with adenosquamous carcinoma or HPV negative tumor. (C, D) Frequency of patients with gains (red) and losses (green) along chromosome 1 to X for patients with adenosquamous carcinoma (C) and HPV negative tumor (D). Gene dosage alterations above 1.1 and below 0.9 were classified as gains and losses, respectively. (A–D) Tumors in the basic cohort subjected to aCGH analysis were included.
Found at: doi:10.1371/journal.pgen.1000719.s001 (0.30 MB TIF)

Figure S2 Tumor ploidy and gene dosage alterations in homogeneous and heterogeneous tumors. (A) Ploidy distribution of patients with homogeneous (left) and heterogeneous (right) tumors. (B,C) Frequency of patients with gains (red) and losses (green) along chromosome 1 to X for patients with homogeneous (B) and heterogeneous (C) tumor. Gene dosage alterations above 1.1 and below 0.9 were classified as gains and losses, respectively. Totally 86 patients with a tumor cell fraction sufficiently high for reliable detection of heterogeneity were included in the analysis.
Found at: doi:10.1371/journal.pgen.1000719.s002 (0.29 MB TIF)

Figure S3 Clinical outcome for patients with different combinations of predictive losses. Kaplan-Meier curves showing progression free survival after chemoradiotherapy of 97 cervical cancer patients with different combinations of 3p11.2-p14.1, 13q13.1-q21.1, and 21q22.2-3 loss. The different combinations and number of patients in each group are listed (right). P-value in log-rank test is indicated.
Found at: doi:10.1371/journal.pgen.1000719.s003 (0.24 MB TIF)

Figure S4 Correlations between gene dosage and expression. Typical correlation plots of gene dosage and expression for 9 correlating genes within the recurrent and predictive regions; 6 with gain and 3 with loss. Spearman's rank correlation analysis on semi-discrete data was performed, for which amplitudes lower than 1.1 were set to 1 for gains and amplitudes higher than 0.9 were set to 1 for losses. Correlation coefficient (R) and p-value are indicated.
Found at: doi:10.1371/journal.pgen.1000719.s004 (0.27 MB TIF)

Table S1 Recurrent high-level amplifications and homozygous deletions in locally advanced cervical cancer.
Found at: doi:10.1371/journal.pgen.1000719.s005 (0.03 MB PDF)

Table S2 Relationships among Illumina, cDNA, and gene dosage data for correlating genes.
Found at: doi:10.1371/journal.pgen.1000719.s006 (0.07 MB PDF)

Acknowledgments

We wish to thank Thea Smedsrud at the Microarray Facility, the Norwegian Radium Hospital, for help with the Illumina experiments.

Author Contributions

Conceived and designed the experiments: ML HL. Performed the experiments: ML DHS. Analyzed the data: ML MH LCB TS KS IKG GBK HL. Contributed reagents/materials/analysis tools: MH LCB KS IKG GBK. Wrote the paper: ML MH LCB DHS TS KS IKG GBK HL.

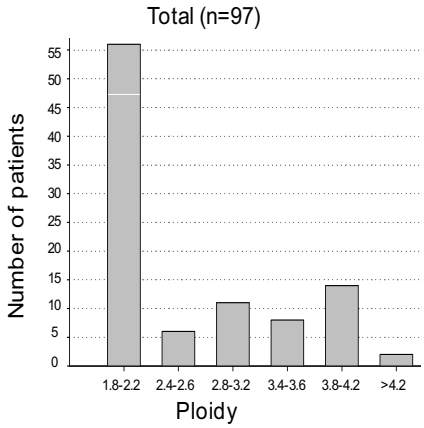
References

1. Cannistra SA, Niloff JM (1996) Cancer of the uterine cervix. N Engl J Med 334: 1030–1038.
2. Eifel PJ (2006) Concurrent chemotherapy and radiation therapy as the standard of care for cervical cancer. Nat Clin Pract Oncol 3: 248–255.
3. Hanahan D, Weinberg RA (2000) The hallmarks of cancer. Cell 100: 57–70.
4. Kroemer G, Pouyssegur J (2008) Tumor cell metabolism: cancer's Achilles' heel. Cancer Cell 13: 472–482.
5. Mosesson Y, Mills GB, Yarden Y (2008) Derailed endocytosis: an emerging feature of cancer. Nat Rev Cancer 8: 835–850.
6. Albertson DG (2006) Gene amplification in cancer. Trends Genet 22: 447–455.

7. Knuutila S, Aalto Y, Autio K, Bjorkqvist AM, El-Rifai W, et al. (1999) DNA copy number losses in human neoplasms. Am J Pathol 155: 683–694.
8. Heselmeyer K, Macville M, Schrock E, Blegen H, Hellstrom AC, et al. (1997) Advanced-stage cervical carcinomas are defined by a recurrent pattern of chromosomal aberrations revealing high genetic instability and a consistent gain of chromosome arm 3q. Genes Chromosomes Cancer 19: 233–240.
9. Lyng H, Beigi M, Svendsrud DH, Brustugun OT, Stokke T, et al. (2004) Intratumor chromosomal heterogeneity in advanced carcinomas of the uterine cervix. Int J Cancer 111: 358–366.
10. Rao PH, Arias-Pulido H, Lu XY, Harris CP, Vargas H, et al. (2004) Chromosomal amplifications, 3q gain and deletions of 2q33-q37 are the frequent genetic changes in cervical carcinoma. BMC Cancer 4: 5.
11. Albertson DG, Collins C, McCormick F, Gray JW (2003) Chromosome aberrations in solid tumors. Nat Genet 34: 369–376.
12. Davies JJ, Wilson IM, Lam WL (2005) Array CGH technologies and their applications to cancer genomes. Chromosome Res 13: 237–248.
13. Wilting SM, de WJ, Meijer CJ, Berkhof J, Yi Y, et al. (2008) Integrated genomic and transcriptional profiling identifies chromosomal loci with altered gene expression in cervical cancer. Genes Chromosomes Cancer 47: 890–905.
14. Lyng H, Lando M, Brovig RS, Svendsrud DH, Johansen M, et al. (2008) GeneCount: genome-wide calculation of absolute tumor DNA copy numbers from array comparative genomic hybridization data. Genome Biol 9: R86.
15. Beisvag V, Junge FK, Bergum H, Jolsum L, Lydersen S, et al. (2006) GeneTools - application for functional annotation and statistical hypothesis testing. BMC Bioinformatics 7: 470.
16. Beroukhim R, Getz G, Nghiemphu L, Barretina J, Hsueh T, et al. (2007) Assessing the significance of chromosomal aberrations in cancer: methodology and application to glioma. Proc Natl Acad Sci U S A 104: 20007–20012.
17. Kirchhoff M, Rose H, Petersen BL, Maahr J, Gerdes T, et al. (1999) Comparative genomic hybridization reveals a recurrent pattern of chromosomal aberrations in severe dysplasia/carcinoma in situ of the cervix and in advanced-stage cervical carcinoma. Genes Chromosomes Cancer 24: 144–150.
18. Bauer VL, Braselmann H, Henke M, Mattern D, Walch A, et al. (2008) Chromosomal changes characterize head and neck cancer with poor prognosis. J Mol Med 86: 1353–1365.
19. van den Broek GB, Wreesmann VB, van den Brekel MW, Rasch CR, Balm AJ, et al. (2007) Genetic abnormalities associated with chemoradiation resistance of head and neck squamous cell carcinoma. Clin Cancer Res 13: 4386–4391.
20. Heselmeyer K, Schrock E, du MS, Blegen H, Shah K, et al. (1996) Gain of chromosome 3q defines the transition from severe dysplasia to invasive carcinoma of the uterine cervix. Proc Natl Acad Sci U S A 93: 479–484.
21. Kersemaekers AM, van de Vijver MJ, Kenter GG, Fleuren GJ (1999) Genetic alterations during the progression of squamous cell carcinomas of the uterine cervix. Genes Chromosomes Cancer 26: 346–354.
22. Umayahara K, Numa F, Suehiro Y, Sakata A, Nawata S, et al. (2002) Comparative genomic hybridization detects genetic alterations during early stages of cervical cancer progression. Genes Chromosomes Cancer 33: 98–102.
23. Wilting SM, Steenbergen RD, Tijssen M, van Wieringen WN, Helmerhorst TJ, et al. (2009) Chromosomal signatures of a subset of high-grade premalignant cervical lesions closely resemble invasive carcinomas. Cancer Res 69: 647–655.
24. Maltepe E, Schmidt JV, Baunoch D, Bradfield CA, Simon MC (1997) Abnormal angiogenesis and responses to glucose and oxygen deprivation in mice lacking the protein ARNT. Nature 386: 403–407.
25. Chaineau M, Danglot L, Proux-Gillardeaux V, Galli T (2008) Role of HRB in Clathrin-dependent Endocytosis. J Biol Chem 283: 34365–34373.
26. Tanaka N, Kyuuma M, Sugamura K (2008) Endosomal sorting complex required for transport proteins in cancer pathogenesis, vesicular transport, and non-endosomal functions. Cancer Sci 99: 1293–1303.
27. Lin CJ, Cencic R, Mills JR, Robert F, Pelletier J (2008) c-Myc and eIF4F are components of a feedforward loop that links transcription and translation. Cancer Res 68: 5326–5334.
28. Wouters BG, Koritzinsky M (2008) Hypoxia signalling through mTOR and the unfolded protein response in cancer. Nat Rev Cancer 8: 851–864.
29. Wade PA (2001) Transcriptional control at regulatory checkpoints by histone deacetylases: molecular connections between cancer and chromatin. Hum Mol Genet 10: 693–698.
30. Imoto I, Tsuda H, Hirasawa A, Miura M, Sakamoto M, et al. (2002) Expression of cIAP1, a target for 11q22 amplification, correlates with resistance of cervical cancers to radiotherapy. Cancer Res 62: 4860–4866.
31. Narayan G, Bourdon V, Chaganti S, rias-Pulido H, Nandula SV, et al. (2007) Gene dosage alterations revealed by cDNA microarray analysis in cervical cancer: identification of candidate amplified and overexpressed genes. Genes Chromosomes Cancer 46: 373–384.
32. Scotto L, Narayan G, Nandula SV, rias-Pulido H, Subramaniyam S, et al. (2008) Identification of copy number gain and overexpressed genes on chromosome arm 20q by an integrative genomic approach in cervical cancer: potential role in progression. Genes Chromosomes Cancer 47: 755–765.
33. Scotto L, Narayan G, Nandula SV, Subramaniyam S, Kaufmann AM, et al. (2008) Integrative genomics analysis of chromosome 5p gain in cervical cancer reveals target over-expressed genes, including Drosha. Mol Cancer 7: 58.: 58.
34. Fyles A, Milosevic M, Hedley D, Pintilie M, Levin W, et al. (2002) Tumor hypoxia has independent predictor impact only in patients with node-negative cervix cancer. J Clin Oncol 20: 680–687.
35. Lyng H, Sundfor K, Trope C, Rofstad EK (2000) Disease control of uterine cervical cancer: relationships to tumor oxygen tension, vascular density, cell density, and frequency of mitosis and apoptosis measured before treatment and during radiotherapy. Clin Cancer Res 6: 1104–1112.
36. Tsang RW, Wong CS, Fyles AW, Levin W, Manchul LA, et al. (1999) Tumour proliferation and apoptosis in human uterine cervix carcinoma II: correlations with clinical outcome. Radiother Oncol 50: 93–101.
37. Birch AH, Quinn MC, Filali-Mouhim A, Provencher DM, Mes-Masson AM, et al. (2008) Transcriptome analysis of serous ovarian cancers identifies differentially expressed chromosome 3 genes. Mol Carcinog 47: 56–65.
38. Novak RL, Phillips AC (2008) Adenoviral-mediated Rybp expression promotes tumor cell-specific apoptosis. Cancer Gene Ther 15: 713–722.
39. Zheng L, Schickling O, Peter ME, Lenardo MJ (2001) The death effector domain-associated factor plays distinct regulatory roles in the nucleus and cytoplasm. J Biol Chem 276: 31945–31952.
40. Sanchez-Beato M, Sanchez E, Gonzalez-Carrero J, Morente M, Diez A, et al. (2006) Variability in the expression of polycomb proteins in different normal and tumoral tissues. A pilot study using tissue microarrays. Mod Pathol 19: 684–694.
41. Haussler MR, Haussler CA, Bartik L, Whitfield GK, Hsieh JC, et al. (2008) Vitamin D receptor: molecular signaling and actions of nutritional ligands in disease prevention. Nutr Rev 66: S98–112.
42. Rachez C, Lemon BD, Suldan Z, Bromleigh V, Gamble M, et al. (1999) Ligand-dependent transcription activation by nuclear receptors requires the DRIP complex. Nature 398: 824–828.
43. Zohn IE, Li Y, Skolnik EY, Anderson KV, Han J, et al. (2006) p38 and a p38-interacting protein are critical for downregulation of E-cadherin during mouse gastrulation. Cell 125: 957–969.
44. Han J, Sun P (2007) The pathways to tumor suppression via route p38. Trends Biochem Sci 32: 364–371.
45. Bashyam MD, Bair R, Kim YH, Wang P, Hernandez-Boussard T, et al. (2005) Array-based comparative genomic hybridization identifies localized DNA amplifications and homozygous deletions in pancreatic cancer. Neoplasia 7: 556–562.
46. Kilic M, Kasperczyk H, Fulda S, Debatin KM (2007) Role of hypoxia inducible factor-1 alpha in modulation of apoptosis resistance. Oncogene 26: 2027–2038.
47. Dong Z, Venkatachalam MA, Wang J, Patel Y, Saikumar P, et al. (2001) Up-regulation of apoptosis inhibitory protein IAP-2 by hypoxia. Hif-1-independent mechanisms. J Biol Chem 276: 18702–18709.
48. Low WK, Dang Y, Schneider-Poetsch T, Shi Z, Choi NS, et al. (2005) Inhibition of eukaryotic translation initiation by the marine natural product pateamine A. Mol Cell 20: 709–722.
49. Ron D, Hinnebusch AG (2006) Targeting translation in hypoxic tumors. ACS Chem Biol 1: 145–148.
50. Lyng H, Brovig RS, Svendsrud DH, Holm R, Kaalhus O, et al. (2006) Gene expressions and copy numbers associated with metastatic phenotypes of uterine cervical cancer. BMC Genomics 7: 268.
51. Hupe P, Stransky N, Thiery JP, Radvanyi F, Barillot E (2004) Analysis of array CGH data: from signal ratio to gain and loss of DNA regions. Bioinformatics 20: 3413–3422.
52. Benjamini Y, Hochberg Y (2008) Controlling the false discoveryrate: a practical and powerful approach to multiple testing. Journal of R Stat Soc B 57: 289–300.
53. Tibshirani R (1997) The lasso method for variable selection in the Cox model. Stat Med 16: 385–395.
54. Bovelstad HM, Nygard S, Storvold HL, Aldrin M, Borgan O, et al. (2007) Predicting survival from microarray data - a comparative study. Bioinformatics 23: 2080–2087.
55. Zou H, Hastie T (2005) Regularization and variable selection via the elastic net. J R Statist Soc B 67: 301–320.

Supplementary Figure 1

A

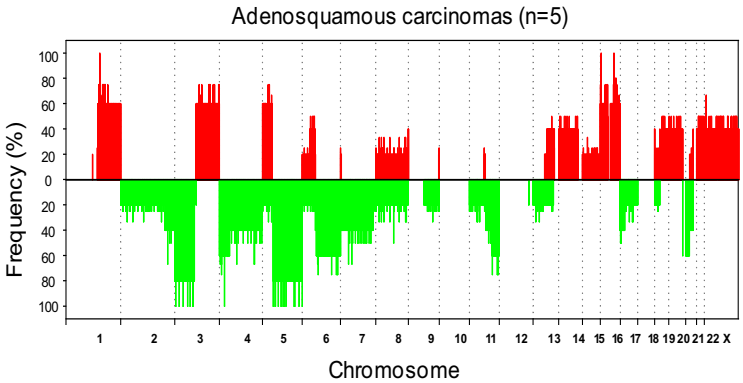


B

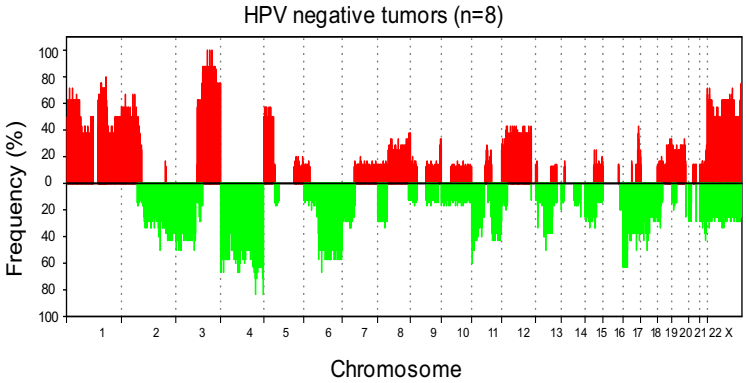
Ploidy	Adenosquamous carcinomas (n=5)	HPV negative tumors (n=8)
1.8-2.2	3[a]	5
2.4-2.6	0	0
2.8-3.2	1	1
3.4-3.6	0	0
3.8-4.2	1	2
>4.2	0	0

[a]Number of patients is indicated

C

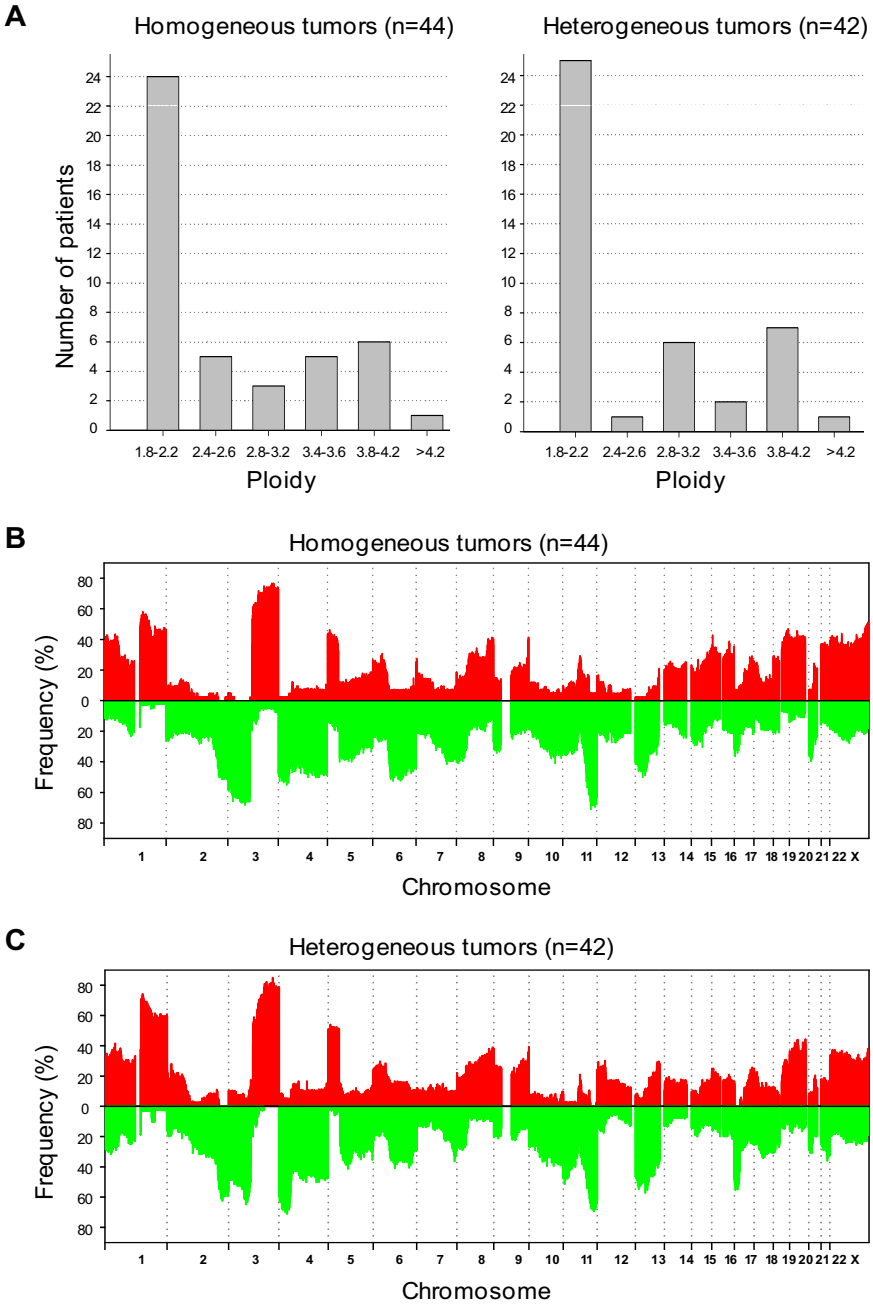


D



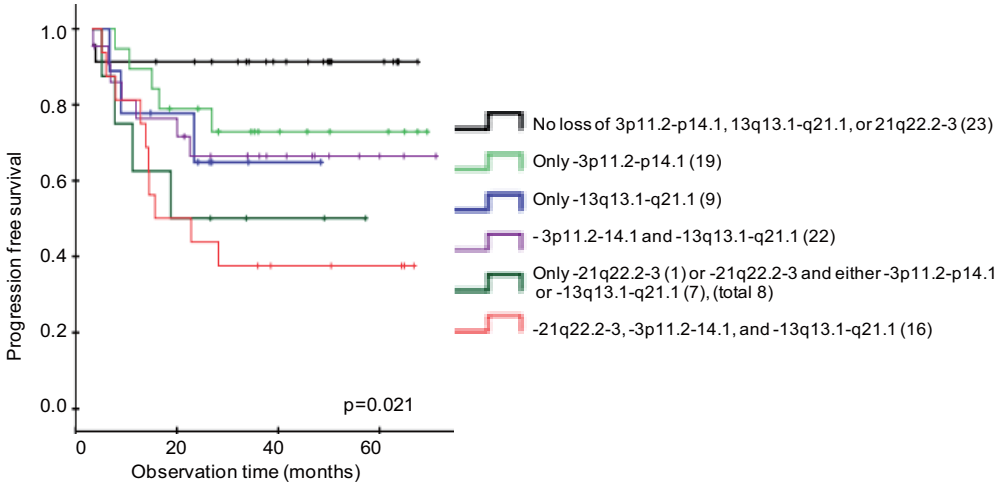
Supplementary Figure 1
Tumor ploidy and gene dosage alterations in relation to histological type and HPV status. (A) Ploidy distribution of 97 patients. Tumors with a ploidy within the range of 1.8–2.2 were considered as near diploid. (B) Ploidy of patients with adenosquamous carcinoma or HPV negative tumor. (C, D) Frequency of patients with gains (red) and losses (green) along chromosome 1 to X for patients with adenosquamous carcinoma (C) and HPV negative tumor (D). Gene dosage alterations above 1.1 and below 0.9 were classified as gains and losses, respectively. (A–D) Tumors in the basic cohort subjected to aCGH analysis were included.

Supplementary Figure 2



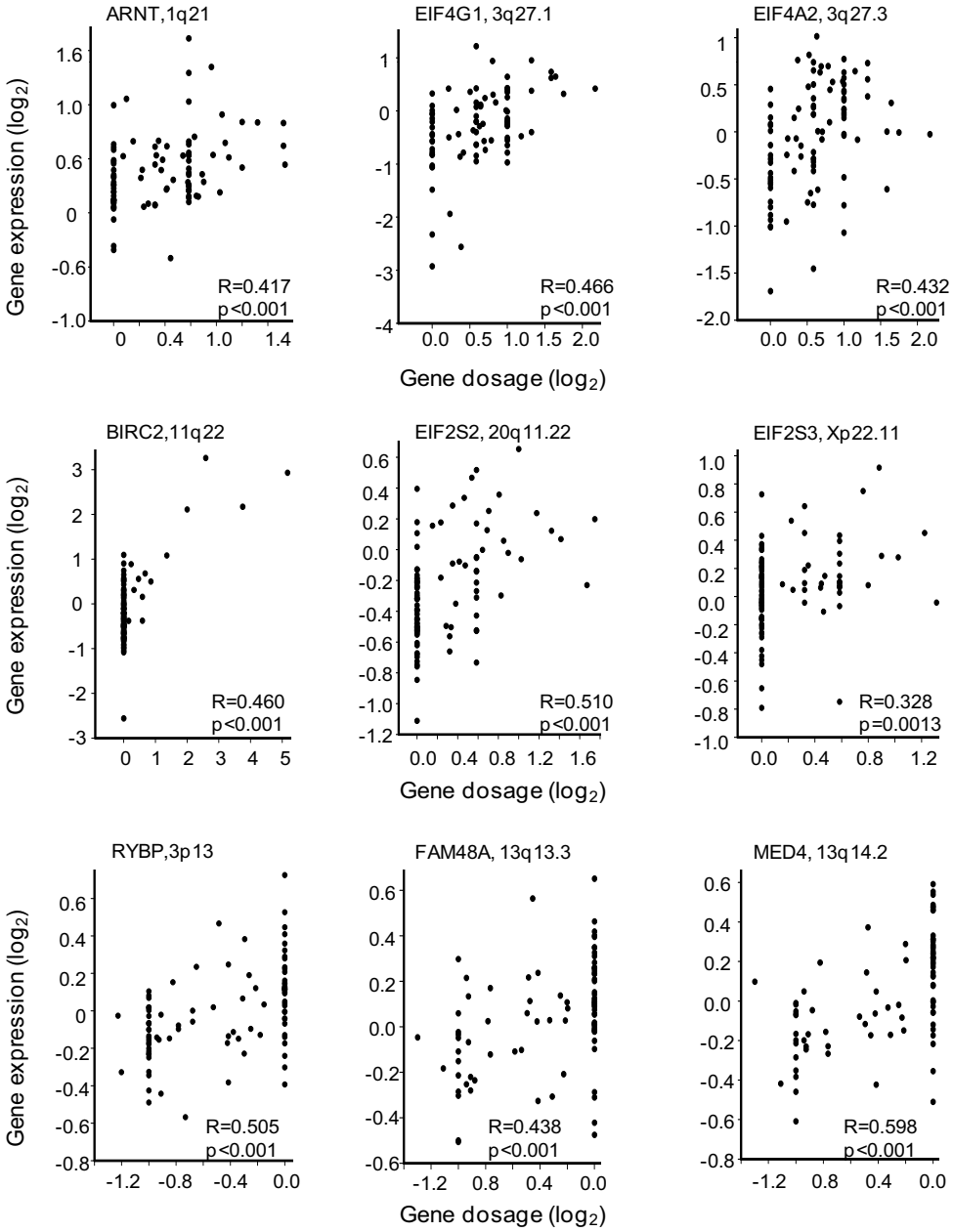
Supplementary Figure 2
Tumor ploidy and gene dosage alterations in homogeneous and heterogeneous tumors. (A) Ploidy distribution of patients with homogeneous (left) and heterogeneous (right) tumors. (B,C) Frequency of patients with gains (red) and losses (green) along chromosome 1 to X for patients with homogeneous (B) and heterogeneous (C) tumor. Gene dosage alterations above 1.1 and below 0.9 were classified as gains and losses, respectively. Totally 86 patients with a tumor cell fraction sufficiently high for reliable detection of heterogeneity were included in the analysis.

Supplementary Figure 3



Supplementary Figure 3
Clinical outcome for patients with different combinations of predictive losses. Kaplan-Meier curves showing progression free survival after chemoradiotherapy of 97 cervical cancer patients with different combinations of 3p11.2-p14.1, 13q13.1-q21.1, and 21q22.2-3 loss. The different combinations and number of patients in each group are listed (right). P-value in log-rank test is indicated.

Supplementary Figure 4



Supplementary Figure 4
Correlations between gene dosage and expression. Typical correlation plots of gene dosage and expression for 9 correlating genes within the recurrent and predictive regions; 6 with gain and 3 with loss. Spearman's rank correlation analysis on semi-discrete data was performed, for which amplitudes lower than 1.1 were set to 1 for gains and amplitudes higher than 0.9 were set to 1 for losses. Correlation coefficient (R) and p-value are indicated.

Table S1. Recurrent high level amplifications and homozygous deletions in locally advanced cervical cancer.

Peak region[a] (Cytoband)	Peak region[a] (MB)	Freq.[b] (%)	Max./min. gene dosage[c] (copy no.)	Correlating genes[d]
Recurrent high level amplification				
3q26.1-qter	166.2-199.5	8	4.5 (9)	*PDCD10, PHC3, ZNF639, FXR1, PARL, DVL3, ABCF3, ALG3, EIF4G1, SFRS10, DGKG, EIF4A2, RFC4, CCDC50, PPP1R2, PAK2, NCBP2, DLG1, BDH1, FLYTTD1*
5p15.2-pter	1.0-12.1	8	4 (15)	*CLPTM1L , MED10, FASTKD3, CCT5, DAP*
9p24.1-2	2.7-6.0	4	13.5 (27)	*KIAA0020, RCL1*
11q13.2-3	68.6-70.6	4	10 (20)	*FADD*
11q22.1-2	100.2-102.0	5	36 (72)	*YAP1, BIRC3, BIRC2*
20q11.21-22	30.0-33.0	5	3.4 (9)	*POFUT1, KIF3B, MAPRE1, SNTA1, EIF2S2, AHCY*
21q22.11-2	32.9-39.6	4	7.5 (15)	*TTC3, BRWD1*
Homozygous deletion[e]				
5q13.2	67.4-71.7	1	0 (0)	*SMN2*
6p21.1-p12.1	44.1-54.1	1	0 (0)	-
8q24.23	136.6-139.3	1	0 (0)	-
9p21.1-3	22.6-29.6	1	0 (0)	*MOBKL2B*
10q23.31	88.2-92.1	3	0 (0)	-
13q34	111.7-114.1	1	0 (0)	-

[a]Peak region of high level amplifications is the region with more than 25% higher amplitude than surrounding region. Peak region of homozygote deletions is the region with a gene dosage of zero.
[b]Frequency is the median percentage of tumors with the alteration.
[c]Gene dosage is absolute DNA copy number divided by ploidy. Maximum (gain) or minimum (loss) gene dosage and the corresponding copy number are listed.
[d]Genes within the peak region showing a correlation between gene dosage and expression are ordered by DNA location.
[e]Homozygote deletions were seen in only few tumors and were not detected as recurrent in statistical analysis.

Table S2. Relationships between Illumina, cDNA, and gene dosage data for correlating genes[a].

Reporter ID	IlluminaID	Gene	cDNA vs gene dosage		cDNA vs gene dosage		cDNA vs Illumina		Illumina vs gene dosage	
			95 patients		**52 patients**					
			R	**p**	**R**	**p**	**R**	**p**	**R**	**p**
129563	ILMN_1762582	*ARNT*	0.407	0.000	0.403	0.004	0.412	0.003	0.411	0.003
814158	ILMN_1669113	*ATF5*	0.347	0.001	0.520	0.000	0.812	0.000	0.488	0.000
825312	ILMN_1772929	*ATP5J*	0.369	0.002	0.421	0.007	0.611	0.000	0.368	0.010
877832	ILMN_1853837	*BCAP31*	0.327	0.001	0.225	0.111	0.866	0.000	0.280	0.047
782748	ILMN_1708485	*BIN3*	0.502	0.000	0.577	0.000	0.345	0.013	0.415	0.003
34852	ILMN_1768194	*BIRC2*	0.475	0.000	0.467	0.000	0.697	0.000	0.573	0.000
201890	ILMN_2405684	*BIRC3*	0.476	0.000	0.445	0.001	0.636	0.000	0.237	0.091
249618	ILMN_1752802	*CLPTM1L*	0.367	0.001	0.484	0.002	0.517	0.000	0.701	0.000
343352	ILMN_1779530	*COG6*	0.403	0.000	0.510	0.000	0.622	0.000	0.451	0.002
897971	ILMN_1699112	*COPB1*	0.345	0.001	0.579	0.000	0.468	0.000	0.556	0.000
884480	ILMN_1798189	*COX7C*	0.500	0.000	0.341	0.017	0.609	0.000	0.407	0.004
814381	ILMN_2112493	*DAP*	0.404	0.000	0.565	0.000	0.790	0.000	0.621	0.000
487082	ILMN_1706498	*DSE*	0.378	0.000	0.365	0.010	0.586	0.000	0.444	0.002
22918	ILMN_1768127	*EBNA1BP2*	0.293	0.004	0.242	0.086	0.696	0.000	0.352	0.012
469151	ILMN_1798014	*EIF2S2*	0.508	0.000	0.571	0.000	0.444	0.001	0.439	0.001
810237	ILMN_1665717	*EIF2S3*	0.427	0.000	0.248	0.076	0.492	0.000	0.320	0.021
307532	ILMN_1685722	*EIF4A2*	0.434	0.000	0.459	0.000	0.633	0.000	0.552	0.000
25988	ILMN_2370772	*EIF4G1*	0.468	0.000	0.314	0.038	0.636	0.000	0.446	0.002
809453	ILMN_1802376	*FAM48A*	0.415	0.000	0.349	0.016	0.492	0.000	0.344	0.018
133158	ILMN_1750160	*FASTKD3*	0.619	0.000	0.782	0.000	0.645	0.000	0.608	0.000
82171	ILMN_1687940	*FOXO3*	0.335	0.002	0.315	0.033	0.618	0.000	0.359	0.015
289551	ILMN_2389273	*FXR1*	0.435	0.000	0.530	0.000	0.752	0.000	0.554	0.000
127509	ILMN_1789702	*GBE1*	0.347	0.001	0.271	0.057	0.465	0.000	0.365	0.009
754085	ILMN_1745798	*GTF2F2*	0.344	0.001	0.473	0.000	0.700	0.000	0.480	0.000
811942	ILMN_2157957	*GTF2H1*	0.341	0.001	0.323	0.021	0.449	0.000	0.215	0.130
256664	ILMN_2200331	*H2AFX*	0.388	0.000	0.305	0.036	0.495	0.000	0.332	0.021
502669	ILMN_1767747	*HDAC2*	0.439	0.000	0.471	0.000	0.521	0.000	0.652	0.000
1606829	ILMN_1764396	*HDAC4*	0.357	0.001	0.397	0.005	0.395	0.004	0.301	0.036
843319	ILMN_1792497	*HRB*	0.425	0.000	0.466	0.000	0.292	0.036	0.415	0.002
810942	ILMN_1802706	*IDH3G*	0.428	0.000	0.486	0.000	0.789	0.000	0.275	0.051
795282	ILMN_1664641	*MED4*	0.568	0.000	0.502	0.000	0.445	0.001	0.446	0.003
131653	ILMN_2371964	*MRPS12*	0.455	0.000	0.521	0.000	0.785	0.000	0.548	0.000
810979	ILMN_1815043	*MRPS2*	0.365	0.001	0.333	0.018	0.495	0.000	0.507	0.000
470216	ILMN_1727080	*MYO6*	0.373	0.000	0.219	0.122	0.298	0.032	0.373	0.007
26711	ILMN_1720442	*NCBP2*	0.494	0.000	0.442	0.002	0.590	0.000	0.682	0.000
753457	ILMN_1728810	*NDUFS1*	0.431	0.000	0.405	0.004	0.383	0.005	0.476	0.000
795439	ILMN_2323491	*NUP62*	0.318	0.002	0.438	0.002	0.281	0.044	0.451	0.001
134439	ILMN_1712687	*PAK2*	0.304	0.003	0.245	0.093	0.414	0.002	0.568	0.000
137836	ILMN_2269002	*PDCD10*	0.588	0.000	0.646	0.000	0.775	0.000	0.705	0.000
80374	ILMN_1772369	*PDHA1*	0.327	0.001	0.177	0.214	0.743	0.000	0.207	0.144
248454	ILMN_1815261	*PDIA4*	0.357	0.001	0.289	0.049	0.621	0.000	0.445	0.002
454475	ILMN_1814074	*PHKA2*	0.470	0.000	0.298	0.036	0.656	0.000	0.418	0.003
112131	ILMN_1776076	*POFUT1*	0.303	0.003	0.476	0.000	0.662	0.000	0.457	0.000
2191807	ILMN_1773613	*POU2F3*	0.399	0.000	0.257	0.113	0.295	0.052	0.120	0.431
769657	ILMN_1683044	*PPP1R2*	0.524	0.000	0.551	0.000	0.558	0.000	0.637	0.000
125148	ILMN_1813766	*RCL1*	0.603	0.000	0.492	0.000	0.719	0.000	0.515	0.000
853151	ILMN_1651850	*RPS16*	0.440	0.000	0.418	0.002	0.042	0.766	0.169	0.236
795453	ILMN_1721842	*RYBP*	0.494	0.000	0.606	0.000	0.522	0.000	0.480	0.000
450131	ILMN_1752111	*SMARCAL1*	0.437	0.000	0.614	0.000	0.521	0.000	0.625	0.000
295255	ILMN_1788211	*SNX19*	0.422	0.000	0.417	0.006	0.671	0.000	0.440	0.003
884657	ILMN_1738938	*TIMM8B*	0.432	0.000	0.649	0.000	0.717	0.000	0.587	0.000
878744	ILMN_1747146	*TSG101*	0.538	0.000	0.554	0.000	0.632	0.000	0.549	0.000
739126	ILMN_1697777	*TSTA3*	0.387	0.000	0.336	0.020	0.336	0.000	0.222	0.128
346292	ILMN_1675674	*UBE4B*	0.304	0.003	0.460	0.001	0.661	0.000	0.367	0.011
206545	ILMN_2174884	*XPO7*	0.400	0.000	0.434	0.001	0.742	0.000	0.614	0.000
126702	ILMN_1792990	*ZNF202*	0.333	0.001	0.355	0.013	0.289	0.038	0.387	0.006

[a]Data for the genes in Figure 5 are shown, except for SLC25A6 which was not included in the Illumina data set.

III