Genome Wide Expression Analysis of Colorectal Cancer and Cervical Cancer

Michal Sheffer

M.Sc Thesis submitted to the Feinberg Graduate School

Weizmann Institute of Science

Research conducted under the supervision of

Prof. Eytan Domany

March 2005

Acknowledgments

I would like to express my deep gratitude to my supervisor, Professor Eytan Domany, for his inspiring and supporting guidance, and for his credence to my work.

I also wish to thank Professor Moshe Oren for giving me the opportunity to work in his lab, providing full training and guidance throughout the project.

I would like to thank Prof. David Givol, a great teacher, for enriching me with his vast knowledge and ideas.

I am grateful to the members of 'Domany's group' and 'Oren's group', in particular Dafna Tsafrir and Dr. Yael Aylon.

My deep gratitude goes to my family and especially to my husband Uri, for their encouragement and support through out these fruitful years.

Abstract

In this study we aim to investigate a whole-genome view of gene expression patterns in colorectal cancer and cervical cancer, using clustering and sorting techniques developed in our lab. The purpose of the first section was to perform unsupervised analysis of colorectal cancer data in order to find the main clusters in this data. Two of the clusters that were found, the proliferation cluster and the ribosomal cluster, were especially interesting because of their possible connection to the initiation of chromosomal instability. Another cluster was related to metastases formation. One protein, B23, was of particular interest because of its possible connection to centrosome aberrations. B23, which was part of the ribosomal cluster, is known to have a central role in centrosome duplication. We designed a lab experiment in order to check whether over-expression of this protein in its phosphorylated form can cause amplification of centrosomes, which is assumed to cause chromosomal instability. The second section of this study examines a cervical cancer proliferation cluster (CCPC) which was previously found in the cervical cancer data. CCPC resembles the proliferation cluster found in the colorectal cancer section. We found that in tumors with an early relapse the average expression level of CCPC genes was higher than in tumors with a favourable course. Using a subset of 20 of these genes, we showed a positive correlation between high gene expression level (measured by RT-PCR) and unfavourable disease outcome in 70 tumor samples, suggesting that the CCPC may be indicative of disease outcome. Moreover, we found that the viral genes E6/E7 had mRNA expression levels with high variation among cases and was positively correlated with the level of the CCPC genes. These findings suggest that the expression level of gene controlling tumor cell proliferation is dependent on E6/E7 mRNA levels in invasive carcinoma of the uterine cervix and may account for the course of the disease.

Contents

1 Introduction

This work is divided into two sections; each section represents an analysis of a different gene expression data. The purpose of the first section was to perform unsupervised analysis of colorectal cancer data in order to find the main clusters. This work lead to the beginning of an experiment in a lab. The second section refers to cervical cancer data. The analysis was started by Tsafrir Dafna and Ilan, who found a very interesting cluster, on which my work was based.

2 Colorectal cancer

2.1 Materials and Methods

2.1.1 DNA microarray

Figure 1 describes the design of the Affymetrix HG-U133A GeneChip® microarray. (http://www.affymetrix.com/products/arrays/specific/hgu133.affx).

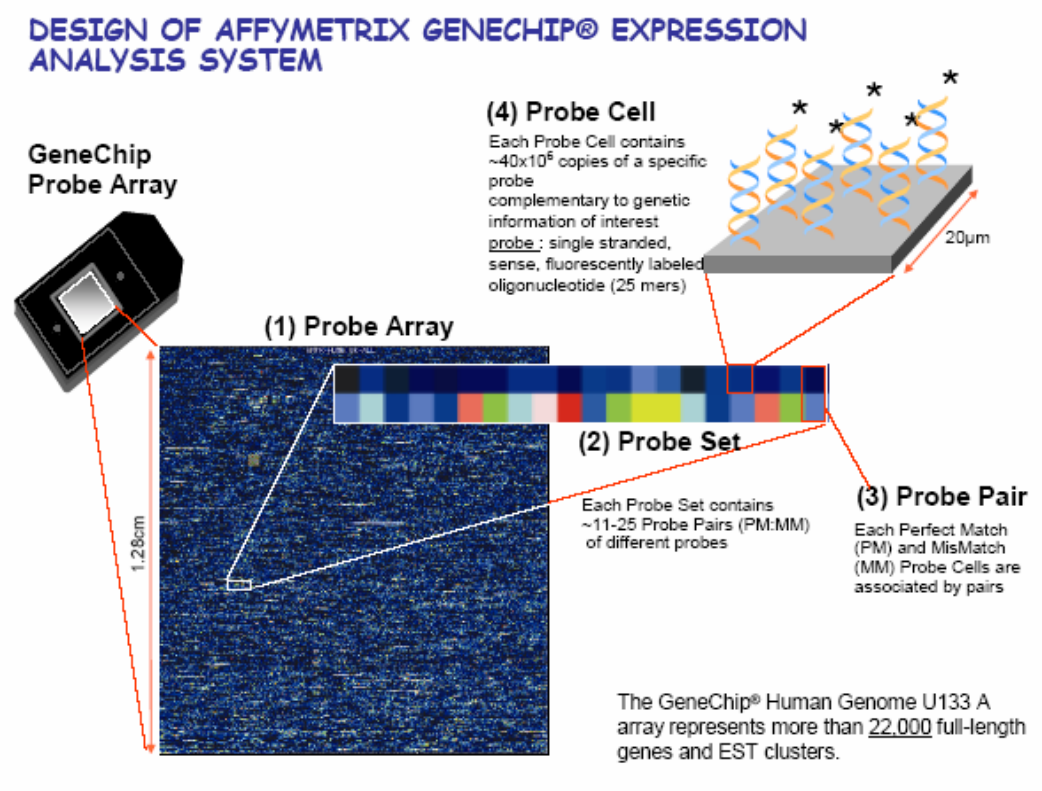

Figure 1: Design of Affymetrix GeneChip® Expression Analysis System**.**

(1) Probe array is the chip containing around 22,000 probe sets (genes or EST). (2) Probe set is a set of probes designed to detect one transcript. A probe set usually consists of 16-20 probe pairs. (3) Probe pair is two probe cells, a PM and its corresponding MM. (4) Probe cell is a single square.shaped feature on an array containing one type of probe. Each probe cell contains millions of probes molecules. Probe is a single 25 base long stranded DNA oligonucleotide complementary to a specific sequenece.

2.1.2 SPC

SPC belongs to the family of hierarchical clustering algorithms, based on the physical properties of inhomogeneous ferromagnets. Its main advantage is that it provides for each cluster of the hierarchy a stability index, whose value indicates the statistical significance or reliability of the cluster, which allows recognizing stable clusters. The clusters are drawn using the hierarchy view of a dendrogram (tree view). Another advantage of SPC is stability against noise. In addition, SPC does not need specification of the number of clusters in advance, a major advantage once working with large data sets, as microarray data (Blatt et al., 1996).

2.1.3 CTWC

Coupled Two-Way Clustering is a method for reducing noise by focusing on small subsets of genes and samples. This is achieved by using only genes (and samples) that were identified previously as a stable cluster for the clustering process. The procedure is iterative; each stable cluster of genes that was found is used, in the next iteration, for the clustering of each of the stable clusters of samples that were previously found and vice versa. This process is repeated until no more clusters are discovered. By focusing on correlated groups of genes and samples, CTWC is able to reduce the noise generated by the majority of "irrelevant" genes and identify specific biological processes involving specific genes or samples. CTWC can be used with a variety of clustering algorithms. I used CTWC with SPC because of its robustness against noise and because it is one of the few algorithms that provides a reliable stability index to each cluster (Getz et al., 2000).

2.1.4 SPIN

Another exploratory analysis method that uses groups of correlated genes for meaningful ordering of patients is SPIN (Sorting Points Into Neighborhoods) (Tsafrir et al., 2005) our recently proposed methodology for data organization and visualization. At the heart of this method is a presentation of the full pairwise distance matrix of the samples, viewed in pseudo-color. The samples are iteratively permuted in search of an optimal ordering, i.e. one that can be used to study embedded shapes. Hence, certain structures in the data (elongated, circular and compact) manifest themselves visually in a SPIN generated distance matrix.

2.1.5 Bonferroni

In order to address contamination with false positive genes associated with multiple comparisons I used the method of Bonferroni correction, which divides the test-wise significance level (α) by the number of tests (n): α / n to get the a new limit for the significant level (Bonferroni, 1936).

2.1.6 Two way ANOVA

ANOVA is the "Analysis of Variance", a statistical test for heterogeneity of means by analysis of group variances. ANOVA assumes random sampling of a random variable with equal variances, independent errors, and a normal distribution. Let n be the number of replicates (sets of identical observations) within each of K factor levels (treatment groups), and y_{ij} be the jth observation within factor level i. The hypothesis is that all population means are equal. These are the definitions of the sum of square (SS) terms:

SST Total SS = $\sum_{i=1}^{k}\sum_{j=1}^{n}\left(y_{ij}-\bar{\bar{y}}\right)^2$ $\bar{\bar{y}}$ - the "group" mean (i.e., mean of means)

SSA SS between groups = $\frac{1}{n}\sum_{i=1}^{k}\left(\sum_{j=1}^{n}y_{ij}\right)^2-\frac{1}{Kn}\left(\sum_{i=1}^{k}\sum_{j=1}^{n}y_{ij}\right)^2$

SSE SS within groups = $\sum_{i=1}^{k}\sum_{j=1}^{n}\left(y_{ij}-\bar{y}_i\right)^2$ $\bar{y}_i$ - the mean of observations within factor level i

= **SST - SSA**

The P-value is obtained in corresponding to the calculated F-ratio of the mean squared values, as described in the following table:

category	Degree of freedom	SS	mean squared	F-ratio
model	$K-1$	SSA	$MSA \equiv \frac{SSA}{K-1}$	$F=\frac{MSA}{MSE}$
error	$K(n-1)$	SSE	$MSE \equiv \frac{SSE}{K(n-1)}$	
total	$Kn-1$	SST	$MST \equiv \frac{SST}{Kn-1}$	

If the P-value is small, reject the null hypothesis that all means are the same for the different groups.

MANOVA ("multiple analysis of variance") is a procedure for testing the equality of mean vectors of more than two populations. The technique is analogous to ANOVA, except that, with two independent variables and the data arrayed in the form of a rows-by-columns matrix, **SSA** can be further divided, into three parts: rows, columns and interactions. Each of these three components is then converted into a corresponding value of **MS**, with the result that there are now three separate **F-ratios** to calculate and three separate tests of significance to perform.

2.1.7 Transfections

Transfections for immunostainings were carried out using a supplemented JetPEI (Polyplus Transfection, France) protocol. H1299 cells were seeded in 6-well plates, each well containing 120,000 cells and 4 cover slips, 24 hours prior to transfection. Before transfection, cells were washed and the medium was replaced with serum- free medium. JetPEI mixtures, at a ratio of 1µg:2µl were prepared in 0.15M NaCl with total volume of 75 µl per mixture, incubated for 15 minutes, and added to the cells. Four hours later, cells were again washed and the transfection medium was replaced with 10% culture medium.

2.1.8 Immunofluorescence

H1299 Cover slips were transferred into cold PBS, washed twice with PBS and fixed with 10% methanol for 20 min at 4°C. Fixed cover slips were washed 6 times in PBS, permeabilized in 0.5% Triton-x-100-PBS for 10 min at RT, washed twice with PBS and incubated for another 5 min at RT.

Cover slips were blocked with 3% BSA in PBS-T for 40 min in 4°C. Then, cover slips were incubated with the primary antibodies γ-tubulin (1:500; Sigma), α-tubulin (1:500; Sigma) for 1 hour at 37°C and washed (3x3 min) in PBS. Secondary antibodies Cy-2 (1:200), Cy-3 (1:500) and DAPI (1:3000) were used for incubating the cells for 30 min at 37°C in the dark. After 3x3min washes with PBS, cover slips were mounted on glass slides and were analyzed by fluorescence microscopy.

2.2 Biological background

Each year 130,000 new patients are found to have colorectal cancer in the United States, while 55,000 die with metastatic disease even after attempted surgical resection (Boring et al., 1994; Parker et al., 1996). Colorectal cancer is most common in people over the age of 50.

2.2.1 The colon

The large intestine is the long muscular tube that is part of the digestive system; its main role is to absorb water and electrolytes from ingesta and to change it from liquid to a solid form. The large intestine is 5 feet long and includes the appendix, cecum, colon and rectum (Figure 2).

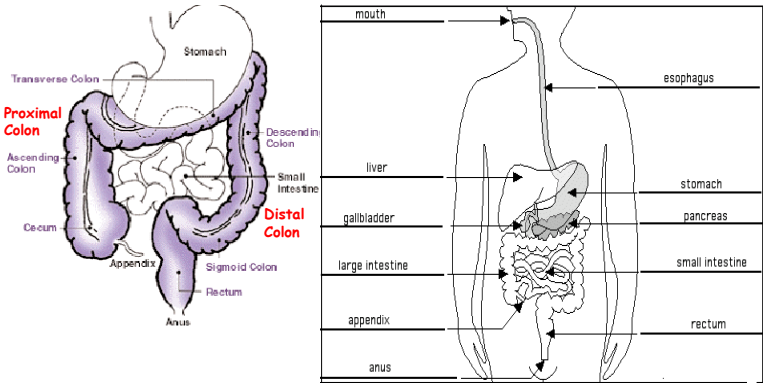

Figure 2: The large intestine

Histologically, the colon is composed of several layers (Figure 3). The inner surface of the colon is a mucosal coating, which is composed of an epithelial layer under which is a submucosa and muscles that function in forcing waste materials through the colon into the rectum.

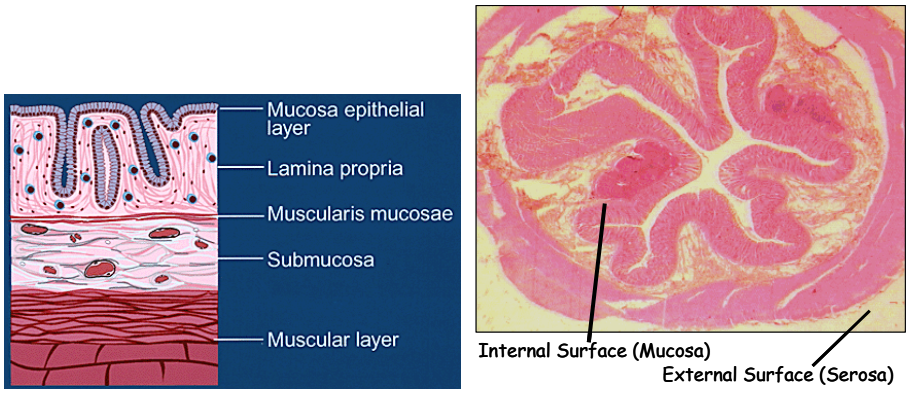

Figure 3: Cross-sectional Anatomy

2.2.2 Colorectal cancer

The development of colorectal cancer lasts for decades and requires multiple mutations throughout the process. As described in Figure 4, there are 5 main stages of the disease: starting from normal epithelium to pre-adenoma, adenoma, adeno-carcenoma and metastasis. Adenoma are neoplastic types of the colon polyps; these tumors have not undergone malignant transformation yet, but without proper surgical procedure they will become adeno-carcenoma, which later progress into metastasis that penetrate the blood system and initiate cancer in other organs.

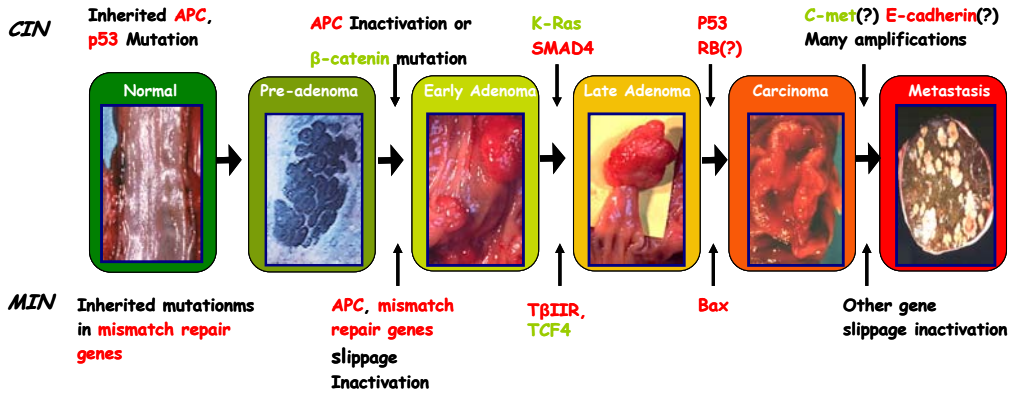

Figure 4: The clinical stages of the colorectal cancer

The progressive stages of the colorectal tumor: from normal epithelium to various stages of adenoma, to adeno-carcenoma and then the final stage of metastasis.

There are two molecular models for colorectal carcinogenesis: (i) Inherited familial adenomatous polyposis (FAP) and most sporadic forms of colon cancer (RER-), which account for 85% of the colorectal cancer cases, and (ii) inherited hereditary non-polyposis colorectal carcinoma (HNPCC) and a small portion of the sporadic colon cancers characterized by defects in the replication error repair pathway (RER+) that account for 15% the of colorectal cancer cases.

It is likely that RER- tumors begin with the inactivation of *APC* or the constitutive activation of *β-catenin*, followed by the activation of *K-ras*, and the mutational inactivation or loss of *P53*. Also, RER- tumors exhibit chromosomal instability (CIN) as evidenced by: Loss of heterozygosity (LOH), gene amplification, promoter hypermethylation (usually of *CDKN2A – ARF, p16*) (Lengauer et al., 1997; Breivik et al., 1999). LOH is the result of the deletion of a gene that already had one inactivated allele, according to Knudson's two hit hypothesis (Figure 5). Gene amplification results in over-expression of a gene. Promoter hypermethylation results in inactivation of a gene. These tumors occur commonly in the distal colon or rectum, progress

slowly and have a less favorable prognosis. In contrast, RER+ tumors have inactivated the replication repair genes which results in mononucleotide repeat slippage (for example *TBIIR*, *BAX*) (Perucho, 1998). RER+ tumors maintain chromosomal integrity and are diploid, but exhibit microsatellite instability (MIN) – the formation of new alleles in tumor DNA as compared with the two parental alleles in the normal DNA (Breivik et al., 1999). These tumors are found mostly in the proximal colon, progress rapidly but have favorable prognosis.

Cancer arises when essential cellular activities are perturbed by abnormal signaling. The accumulation of both genetic and epigenetic alterations (at least five or six) overrides normal mechanisms that regulate growth, differentiation and cell death. Genetic events include mutations, insertions, deletions, gene amplification, LOH. Epigenetic events include methylation and chromatin remodeling.

Loss of heterozygosity (LOH)

According to Knudson's 2-hit hypothesis (Figure 5), both alleles of a tumor suppressor gene must be inactivated in order to develop a tumor. Usually the first allele is mutationally inactivated. LOH is the loss of the second allele. For example in colorectal cancer, *APC*, *P53*, *SMAD4* are inactivated by LOH at 5q, 17p, 18q. 50% of the CIN tumors show LOH in regions of chromosomes 1, 4, 7, 8, 14, 22. LOH in chromosome 4 was associated with metastasis (Malkhosyan et al., 1998).

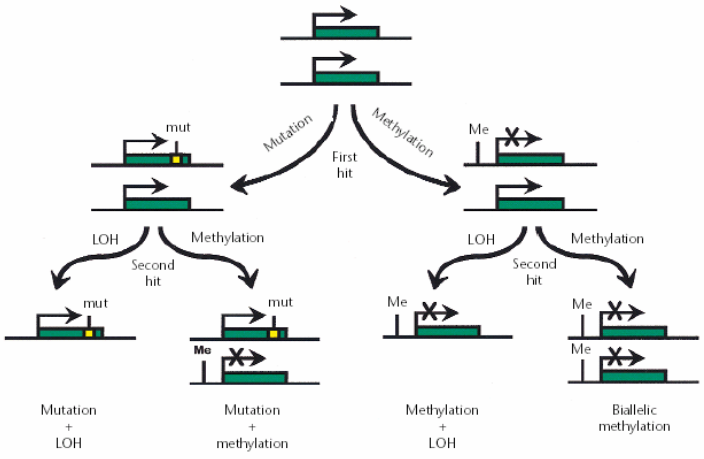

Figure 5: Knudson's 2-hit hypothesis (Jones & Laird, 1999)

Gene Amplification

The exact mechanism is not clear but probably involves defects in genes which function in homologous recombination or regulate cell cycle.

Analysis of colorectal cancer revealed that 50% of colorectal tumors show amplification in chromosomes 6, 20. Over 75% of colorectal tumors show amplification in chromosomes 8, 13. Metastasis is associated with amplification of chromosome 6 (Malkhosyan et al., 1998). The gene *c-Met*, related to cell motility, proliferation, angiogenesis and invasion, is amplified in more than 90% of metastasis tumors (10% in primary) (Takayama et al., 1997).

DNA hypermethylation

CpG islands - CG-rich areas of approximately 1kb are often found near the promotor. Methylation of cytosine residues 5' to Guanines (CpG) has been proposed to regulate both chromatin structure and transcription. Following DNA synthesis, *DNA methyltransferase* methylates the new strand according to hemi-methylated CpG on the old strand (Figure 6). Methylation inhibits transcription as described in Figure 7.

In tumor cells, hypermethylation of CpG islands has been proposed to result from aberrant *DNA methyltransferase* (*DNMT)* activity. Tumor cell lines and primary tumors often overexpress *DNMT1*, *DNMT3*. Tumor cells display higher levels of CpG islands hypermethylation when compared with normal cells (Jones & Laird, 1999).

In colorectal cancer the *CDKN2A* (*ARF, p16*, *p19*) promoter was found to be hypermethylated.

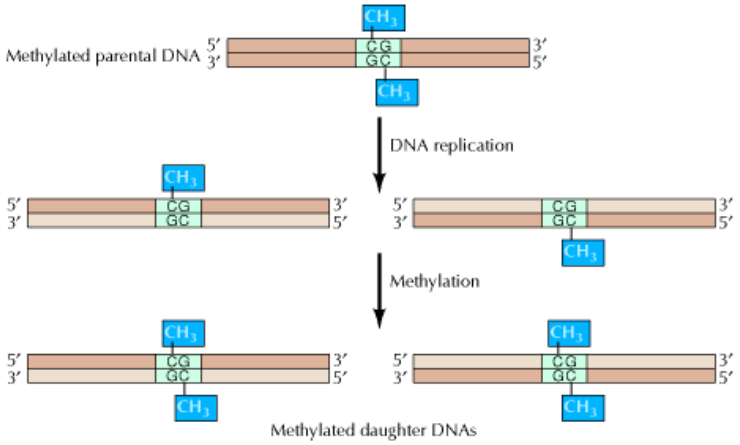

Figure 6: Inheritance of DNA methylation patterns (Alberts et al., 2000)

Once a pattern of DNA methylation is established, each site of methylation is inherited in the progeny DNA.

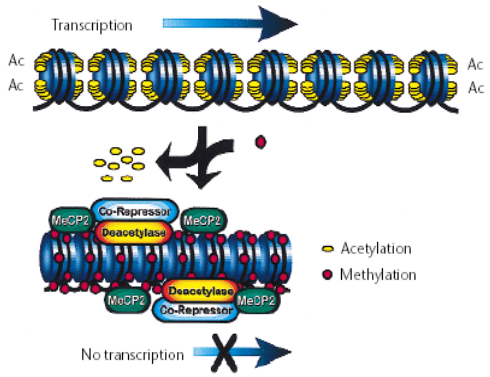

Figure 7: Methylation inhibits transcription (Jones & Laird, 1999)

Methylation of CpG dinucleotides recruits both methyl binding transcription repressors and histone deacetylases to the promotor, resulting in the condensation of DNA and repression of transcription

Microsatellite instability (MIN)

Microsatellites are short nucleotide sequences (2-5 base pairs) that are usually repeated 15-30 times and distributed throughout the genome (Figure 8). Although the length of these microsatellites is highly variable from person to person, each individual has microsatellites of a set length. These repeated sequences are common, and normal. In cells with mutations in DNA repair genes, however, some of these sequences accumulate errors and become longer or shorter. The appearance of abnormally long or short microsatellites in an individual's DNA is referred to as microsatellite instability. Microsatellite instabilities were found in patients with hereditary non-polyposis colorectal cancer (HNPCC) and in a subgroup of sporadic (non inherited) cancers. This phenotype was initially seen as the formation of new alleles in tumor DNA as compared with the two parental alleles in the normal DNA, and was shown to be caused by mutation in the highly conserved mismatch repair genes (Breivik et al., 1999; Lothe, 1997).

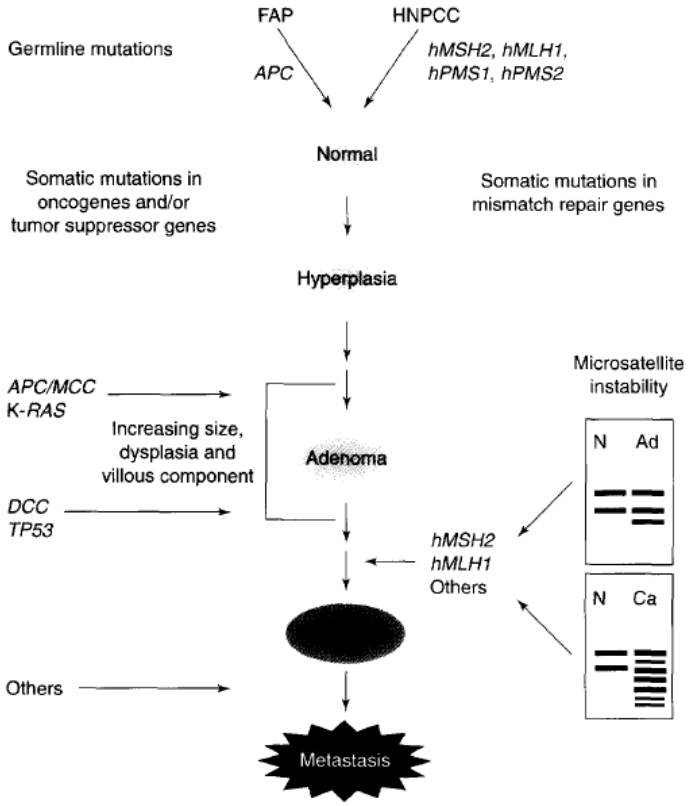

Figure 8: Microsatellite instabilities in Colorectal cancer (Lothe, 1997)

2.2.3 Chromosomal instability (CIN)

CIN is generated by: (i) the failure to repair DNA breakage, which is characterized by insertions, deletions, translocations and amplifications, (ii) the failure in separation of sister chromatids, which leads to **aneuploidy**: chromosomes (or part of them) are gained or lost (Pihan et al., 1999). Mis-regulation leads to loss of cell cycle checkpoint control which results in replication of damaged DNA.

Cell division overview

The M-phase of the cell cycle is separated into 5 stages as described in Figure 9.

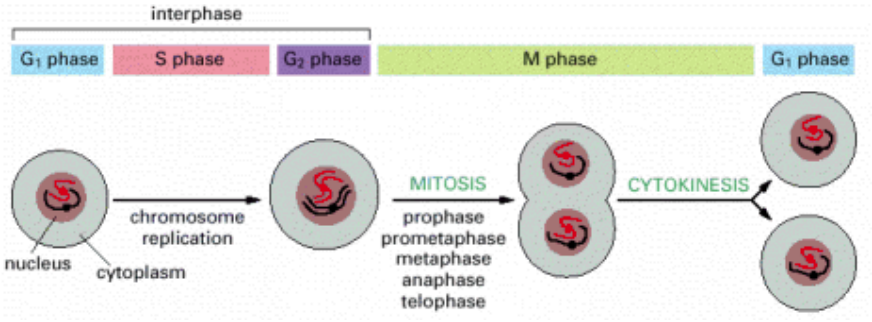

Figure 9: The cell cycle (Alberts et al., 2002)

Figure 10 (Scholey et al., 2003) describes the different stages of the M-phase. At prophase the duplicated centrosomes migrate around the nucleus (Figure 10a). The breaking down of the nuclear envelope occurs during prometaphase (Figure 10b), allowing congression (the process of chromosomes moving to the mitotic spindle). At metaphase (Figure 10c) the sister chromatids are aligned on the mitotic spindle, facing opposite poles. There are four types of microtubules that are involved at this stage, all have their plus ends away from the poles; **Astral** - link the spindle poles to the cell cortex, **Chromosomal** – link chromosome arms to poles, **Kinetochore** – link poles to the kinetochores (Structures at the centromere – the region of each chromatid, to which sister chromatids attach), **interpolar** – link the two poles. Segregation of the sister chromatids occur at anaphase (Figure 10d) and the poles move further apart from one another (Figure 10e) at late anaphase the contractile ring begins to assemble and contract. The separation of the daughter cell occurs at Telophase (Figure 10f), and the nuclear envelope reassembles around the pole of each daughter cell.

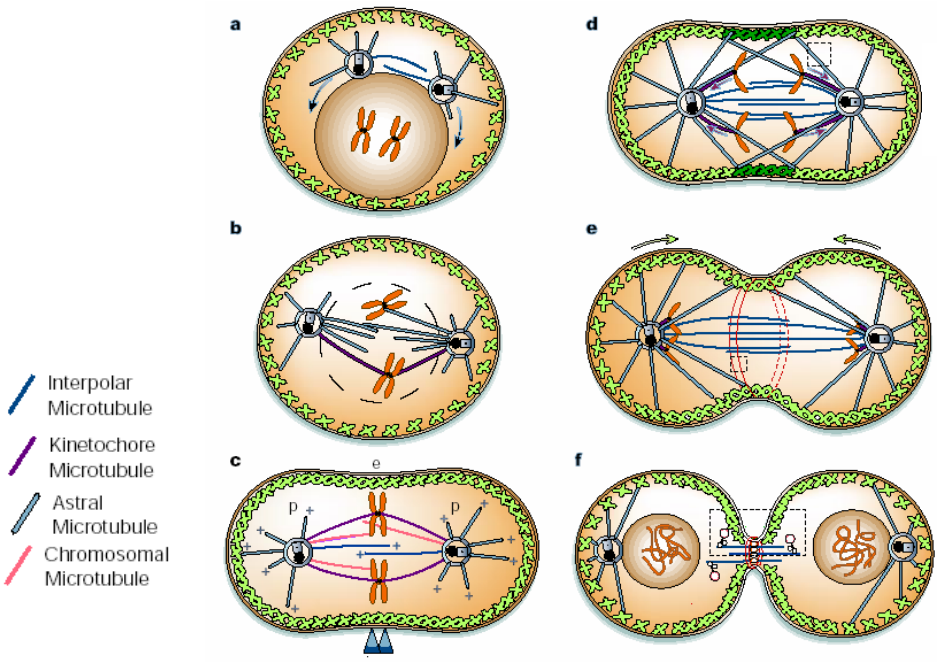

Figure 10: Mitosis and Cytokinesis (Scholey et al., 2003)

a. Prophase - the duplicated centrosomes move around around the nucleus. **b**. Prometaphase - nuclear envelope breaks down, chromosomes approach the mitotic spindle. **c**. Metaphase - sister chromatids are aligned on the mitotic spindle, facing opposite poles. **d**. Anaphase A – segregation of the sister chromatids. **e**. Anaphase B- poles move further apart from one another, contractile ring begins to assemble and contract. **f**. Telophase - separation of the daughter cell, nuclear envelope reassemble around the pole of each daughter cell.

BUB, MAD – The mitotic spindle checkpoints

The Mitotic spindle check point detects the precise spatial orientation of whole chromosomes, ensuring that all pairs of sister chromatids establish bilateral attachment to the mitotic spindle, to become aligned on the metaphase plate. Several of the spindle check point proteins attach to the kinetochore of the sister chromatids that haven't attached to the mitotic spindle. *MAD* and *BUB* were discovered as key component of this checkpoint. Anaphase promoting complex (*APC/C)* is a large ubiquitin protein ligase that associates with the mitotic cofactor *CDC20.* According to Figure 11, this complex is inhibited by two mechanisms: the first mechanism involves the binding of *MAD2*. The second mechanism involves BUB1. As long as the 'wait anaphase' signal (lack of tension) exists, *APC/C* is inhibited. When all kinetochores are attached, *APC/C* becomes active and degrades *Securin*, resulting in active *Separin* protease. *Separin* then cleaves the Cohesin complexes connecting the aligned sister chromatids. *PLK*, a polo like kinase, was found to enhance the ability of Cohesin to be cleaved by *Separin*, in a *Securin* independent mechanism (Jallepalli et al., 2001).

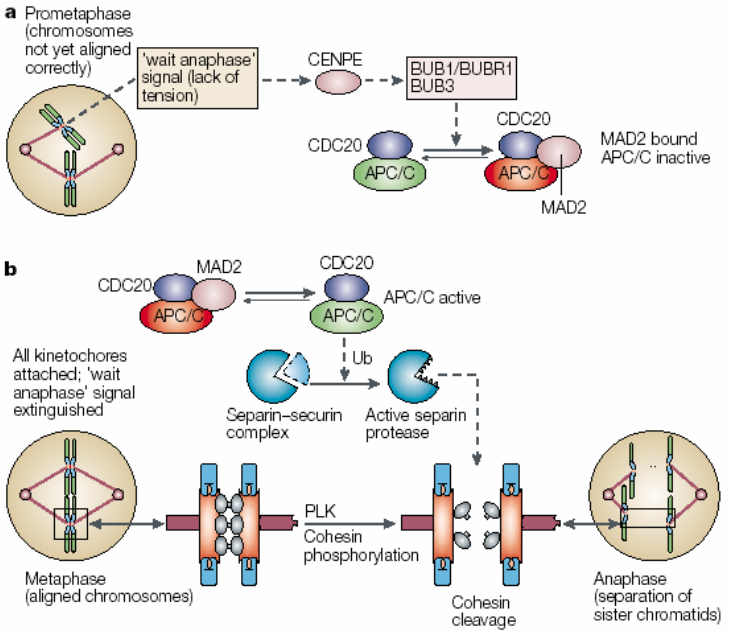

Figure 11**: Regulation of sister chromatid separation** (Jallepalli et al., 2001)

a. In prometaphase, Unattached chromosomes apparently generate a signal that delays progress to anaphase until all sister chromatids are attached to the spindle apparatus. This signal is transduced by a relay of spindle-checkpoint proteins that include *MAD/BUB* proteins. This results in inhibition of the anaphase-promoting complex (*APC/C*) that is associated with the mitotic cofactor *CDC20*. **b.** Following attachment of the last kinetochore to the mitotic spindle, the 'wait anaphase' signal is extinguished. This activates *APC/C–CDC20*, which results in the degradation of securin and liberation of active separin protease. This protease catalyses the cleavage of cohesin complexes that bridge the aligned sister chromatids. The newly separated sister chromatids can then migrate poleward along the spindle axis during anaphase.

Mutations in these check point proteins can cause CIN as described in Figure 12:

In cells with *MAD2* heterozygocity (i.e. with reduced amount of the active protein), the *APC/C* complex is not inhibited, resulting in premature destruction of securin and anaphase entry. This results in precocious sister-chromatids separation and chromosome loss. In cells with deficiency in *Securin*, the *separin* protease is not properly activated. This results in incomplete cleavage of cohesin bridges during anaphase, nondisjunction of sister chromatids and ultimately aneuploidy (Jallepalli et al., 2001).

HBub1 was found to be mutated in some colorectal cancers and over expression of the mutant induced aneuploidy (Cahill et al., 1998).

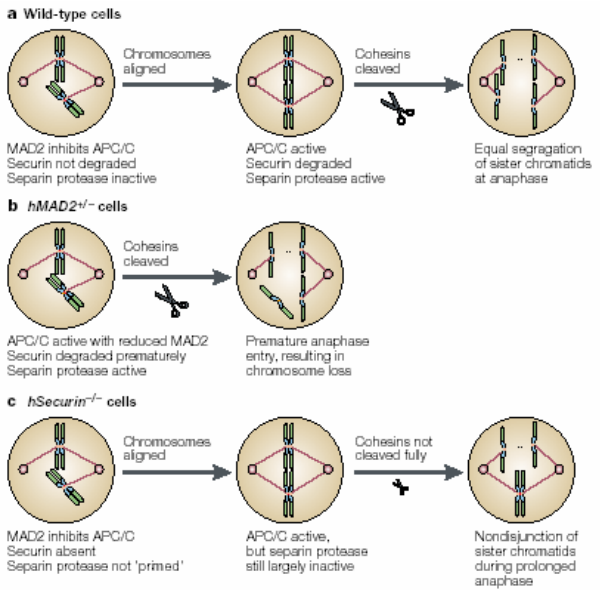

Figure 12: Defective spindle checkpoints (Jallepalli et al., 2001)

a. In wild-type human cells, *APC/C* is inhibited at prometaphase until all chromosomes have aligned on the metaphase plate, ensuring equal segregation of sister chromatids to the two daughter cells. **b.** *hMAD2*$^{+/-}$ cells, the *APC/C* is not inhibited, resulting in premature destruction of *securin* and anaphase entry. This results in precocious sister-chromatid separation and chromosome loss. **c.** In *Securin*$^{-/-}$ cells, the *separin* protease is not properly 'primed' for full biochemical activation. This results in incomplete cleavage of cohesin bridges during anaphase and nondisjunction of sister chromatids.

2.2.4 The centrosome

The centrosome is a small organelle with diameter of 1μm. It consist two centrioles that are surrounded by pericentriolar material (PCM). The centrioles are cylindrical structures made up of nine triplet microtubules (Figure 13b), displaying orthogonal orientation as described in Figure 13a.

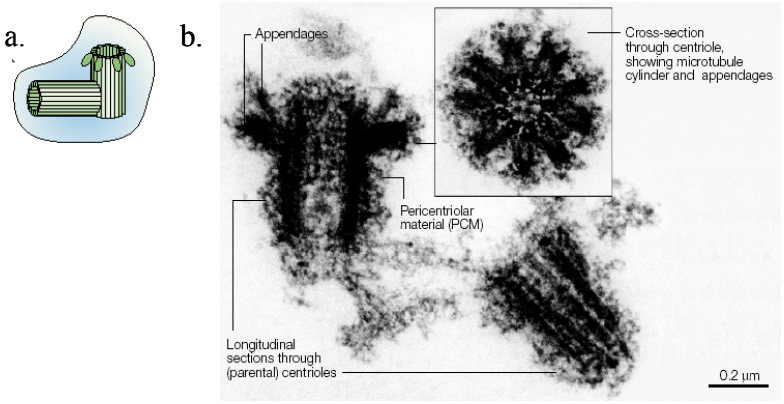

Figure 13: Centriole structure (Nigg, 2002)

a. Schematic drawing of the orthogonal orientation of the centrioles pair.
b. The centrioles pair surrounded with pericentriolar material.

The centrosome is a microtubules organizing center. During cell division the centrosomes form the poles of the bipolar mitotic spindle, to allow chromosome segregation. In addition the centrosomes are also needed for cytokinesis, by determining the position of the cleavage plane, which is essential for asymmetric divisions (Nigg, 2002). Centrosome duplication occurs only once and must be in coordination with DNA synthesis as shown in Figure 14. Extra copies of centrosomes can result in multipolar spindles, which cause aneuploidy.

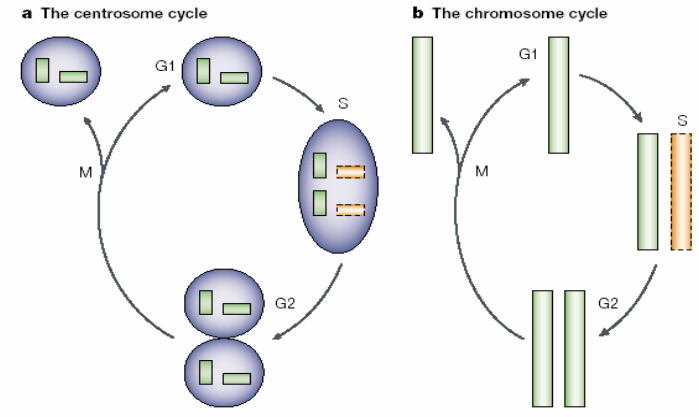

Figure 14: The cell cycle and Centrosome duplication cycle (Nigg, 2002)

Both the chromosomes and the centrosomes should be duplicated only once during one cell cycle, for this reason these cycles are co-regulated.

The key players in cell cycle regulation are the tumor suppressors *P53*, *P21*, *ARF* and *RB*. Over expression of *CDK2* and hyper phosphorylation of *RB* are required for S-phase entry and duplication of centrosomes as described in Figure 15.

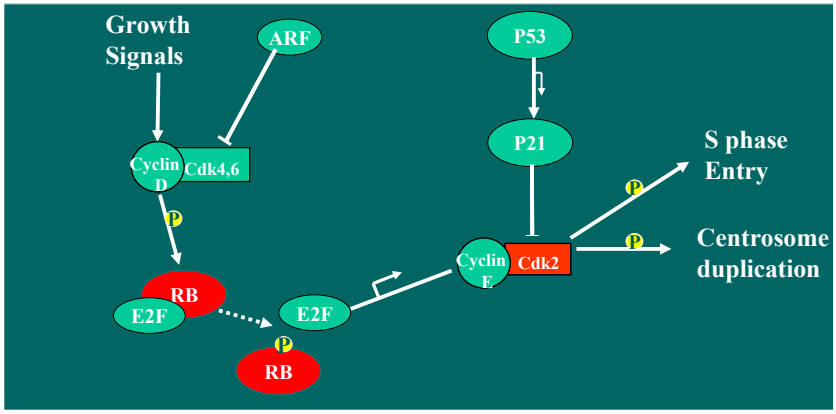

Figure 15: Regulation of the cell cycle check point and centrosome duplication

Both centrosome duplication and DNA replication require the hyper-phosphorylation of RB and the activation of CDK2.

Centrosome amplification

Centrosome amplification has been described for nearly all types of cancers, including: brain, breast, colon, head and neck, lung, pancreas, prostate, cervix. Centrosome amplification was also found in low-grade pre-invasive tumors, together with mitotic spindle defects, implying that centrosomal abnormalities are an important cause of chromosome instability (Pihan et al., 2002). There are three distinct mechanisms that may cause centrosome amplification, as described in Figure 16. The first mechanism is more then one duplication in the same cell cycle. This situation can occur in *RB*-deficient cells. This is consistent with the effect of the viral oncogene *E7*, which inhibits *RB*, giving rise to an excessive number of centrioles (Duensing & Munger, 2002). The second mechanism is aborted mitosis: without normal apoptosis pathway it can lead to cell with tetraploidization and two centrosomes. This phenomenon was seen in cells having the viral oncogene *E6*, which inactivates *P53*. The effect of *E6* produced centrosome amplificationin conjugated with multinucleation, indicating unsuccessful cell division (Duensing & Munger, 2002). The third mechanism, of cell fusion, is rather rare.

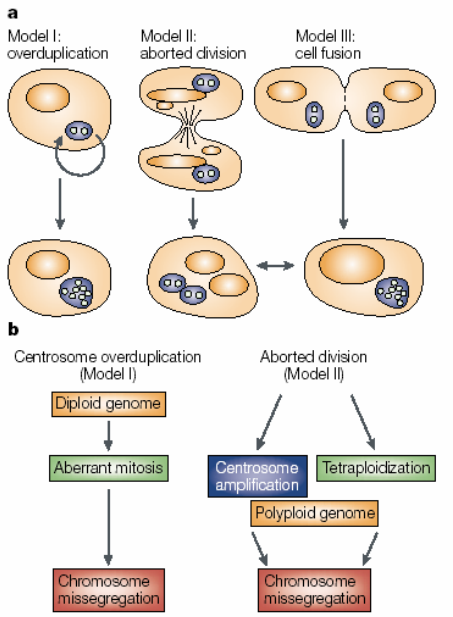

Figure 16: Mechanisms for centrosome amplification (Nigg, 2002)

a. Three models for centrosome amplification: overduplication within a single cell cycle, aborted division, cell fusion. **b**. centrosome amplification and ploidy: two different ways that form chromosome missegregation.

The effect of abnormal number of centrosome on cell is, unavoidably, the formation of multipolar spindles, which lead to aneuploidy. multipolar spindles are common in human cancers as shown in Figure 17.

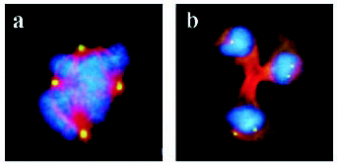

Figure 17: Multipolar spindles in P53$^{-/-}$ cells (Tarapore & Fukasawa, 2002)

B23 – a multifunctional protein

B23 (nucleophosmin/NPM) was originally identified as a nucleolar phosphoprotein. It is a multifunctional protein involved in ribosome biogenesis (Yung et al., 1985), able to shuttle between the nucleus and cytoplasm (Borer et al., 1989), and to bind to peptides containing nuclear and nucleolar localization signal (Szebeni et al., 1995). *B23* has been also found to serve as a chaperone in the very crowded environment of the nucleolus during ribosome assembly (Szebeni & Olson, 1999).

B23 was identified as a direct target for the transcriptional factor *c-Myc* (Zeller et al., 2001). There are several proteins shown to physically interact with *B23: RB* (Takemura et al., 1999), *P53* (Colombo et al., 2002), *ARF* (Bertwistle et al., 2004),

HDM2 (Kurki et al., 2004). It was shown that *B23* induces cell cycle arrest in cells with wild-type *P53*, whereas in cells with mutant *P53* it promotes S phase entry (Itahana et al., 2003). Also, *B23* has been found generally at higher levels in tumor cells than in normal cells (Chan et al., 1989).

B23 and Centrosome duplication

B23 was implicated in the centrosome duplication cycle, ensuring the coordination of centrosomes and DNA duplication (Okuda et al., 2000; Tokuyama et al., 2001). As described in Figure 18 (Okuda, 2002), *B23* dissociates from centrosomes upon phosphorylation by *cdk2/cyclinE* on T199, which in turn triggers initiation of centrosome duplication. During S and G2 phases, *B23* doesn't re-associate with centrosomes, potentially through the activity of *cdk2/cyclinA*, which also phosphorylates *B23* on T199. During mitosis, *B23* re-associates to the centrosomes and each daughter cell receives one centrosome bound by *B23* upon cytokinesis.

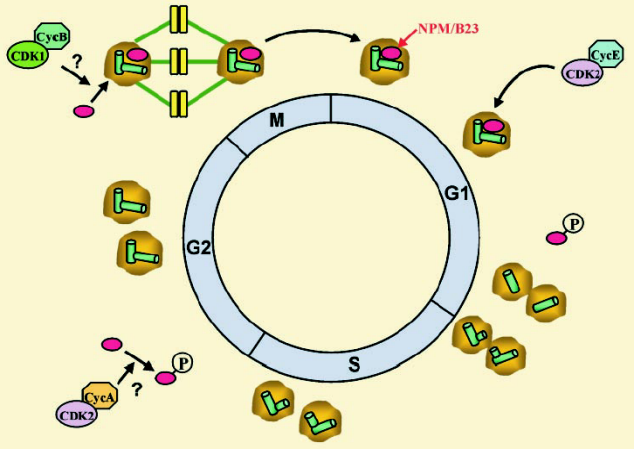

Figure 18: **NPM in the centrosome duplication cycle** (Okuda, 2002)

Centrosome-bound *B23* dissociates from centrosomes upon phosphorylation by *CDK2/cyclinE*, which in turn triggers initiation of centriole duplication. When the nuclear membrane breaks down during mitosis, *B23* re-localizes to the centrosomes. After mitosis, each daughter cell receives one centrosome bound by *B23*.

B23 and Centrosome amplification

Only few articles linked B23 to centrosome amplification (Saavedra et al., 2003; Ussar & Voss, 2004; Zhang et al, 2004). The most common explanation for this link, given by Saavedra et al., was through early phosphorylation of B23 at G1; when the kinase CDK2/cyclinE was over-expressed, this early phosphorylation could lead to more than one duplication of centrosome in the same cell cycle, which results in centrosome amplification. In my work I tried to answer the question whether B23 mis-regulation is related to centrosome amplification.

2.2.5 Major key players in colorectal cancer

APC

The *APC* gene is a tumor suppressor found mutated in 85% of the colorectal tumors evidenced by CIN (Powell et al., 1992). In normal cells, *APC* is known to prevent the accumulation of *β-catenin*, in order to avoid the interaction of *β-catenin* with the transcription factor *TCF-4*, as described in Figure 19. In the absence of *APC*, *β-catenin* is free to activate transcripionally number of oncogenes, including *c-Myc, cyclinD1* (Goss & Groden, 2000). Tumors that lack *APC* mutations often have mutations in *β-catenin* (Morin et al., 1997). 40% of MIN tumors show mutant *TCF-4* (Duval et al., 1999). Another role for *β-catenin* is to bind *Cadherin*, a tumor suppressor gene for invasive carcinoma, required for cell-cell adhesion (Figure 19). Some oncogenes are known to phosphorylate *β-catenin*, such that it is unabled to bind Cadherin (Hiscox & Jiang, 1999).

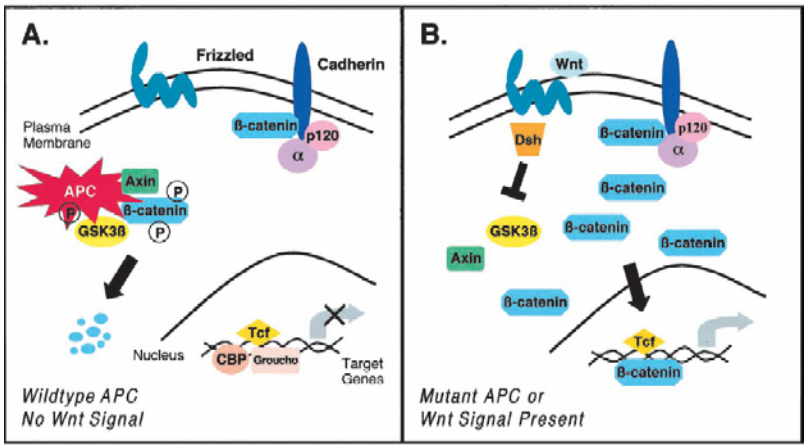

Figure 19: Wnt signaling pathway (Goss & Groden, 2000)

(**a**) In the presence of *APC* or in the absence of Wnt ligand, *β-catenin* is localized to the adherens junction where it is associated with *E-cadherin*. When phosphorylated, *β-catenin* is rapidly degraded by ubiquination of the *APC* complex. (**b**) When *APC* is mutated, or in the presence of the Wnt ligand, *β-catenin* accumulates in the cytoplasm and the nucleus and associates with members of the Tcf family of transcription factors to increase transcription of oncogenes.

Another role associated with *APC* is in regulating chromosomal segregation during cell division: APC localizes to the kinetochores at the end of spindle microtubules through interaction with microtubule associated protein *EB-1* on their c-terminus (Fodde et al., 2001). Thus APC functions to maintain chromosomal integrity. It was also found that APC interacts with the mitotic check point proteins hBUB1, hBUB3, both of them phosphorylated APC in vitro (Kaplan et al., 2001).

P53

The *P53* tumor suppressor is a transcription factor that functions at the center of the cell cycle and apoptosis signaling networks. It is activated in response to DNA damage, different types of stress, and unregulated growth signals. *P53* inhibits cell cycle progression in G1, G2 checkpoints through activation of *P21* and *Gadd45a*. *P53* induces apoptosis through the activation of pro-apoptotic genes such as *BAX*, *IGFBP-3*. *MDM2* and *ARF* tightly regulate *P53*: over-expression of MDM2 or inactivation of *P19* cause degradation of *P53* (Figure 20). Alterations of the *P53* gene occur relatively late in colorectal cancer development, about 50% of the adeno-carcinomas exhibit mutations in *P53* (Nigro et al., 1989). Complete inactivation of *P53* has been associated with a more aggressive disease and poor prognosis (Heide et al., 1997). Usually, CIN is observed prior to *p53* inactivation (Potter, 1999).

P19 is hyper-methylated in more then 60% of the tumors. *Mdm2* is over-expressed in 5-10% of the tumors. 50% of the MIN tumors have mutation in *Bax* (Rampino et al., 1997).

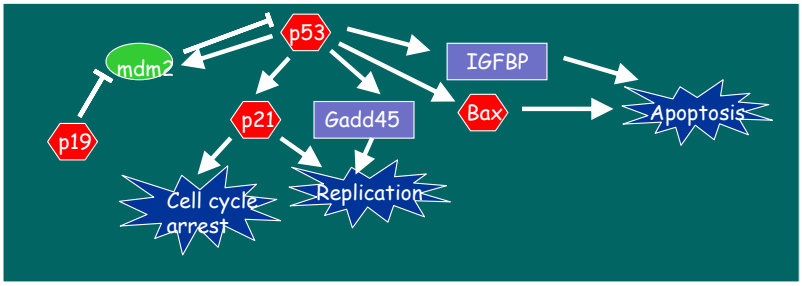

Figure 20: P53 – the gate keeper

P16-cycD-Cdk-RB pathway

The pathway involoving *P16*, *cyclinD*, *RB* regulate progression into G1 checkpoint. *P16* functions as a negative regulator of the G1-S transition by binding to the *cyclinD/CDK4* complex, and inhibiting its kinase activity. In the absence of P16, *cyclinD/CDK4* phsphorylates *RB*, thus, allowing progression into cell cycle (Figure 21). *P16* is hyper-methylated in half of the colorectal tumors

More than 90% of CIN tumors have inactivated *APC* or mutational inactivated *β-catenin* (Powell et al., 1992; Morin et al., 1997), which is associated with upregulation of *cyclinD1*, *c-MYC* expression. 80% of colorectal tumors upregulate the expression of *cyclinD1*. *c-MYC* has been proposed to positively regulate progression through G1 and to directly induce the expression of *CDK4*, which is known to lead to unregulated proliferation.

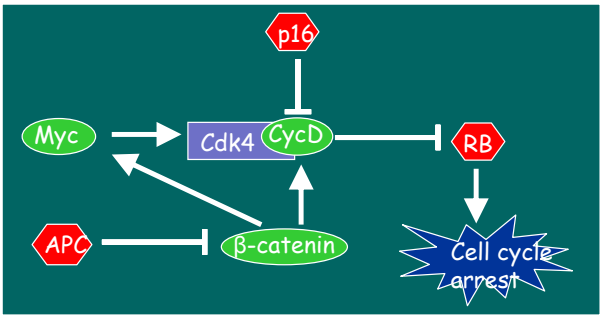

Figure 21: P16-cycD-Cdk-RB pathway

RAS

Hyper activation of *RAS* leads to shortening of the G1 phase. In colorectal cancers, *RAS* mutations were observed in 20-40% of small adenoma, 35-60% of adeno-carcinomas and 50-70% of metastases. *RAS* mutations are associated with hypermethylation of the *p16* promotor (Guan et al., 1999). *RAS* activation can also perturb the TGFβ anti-proliferative signaling pathway.

TGFβ signaling pathway

TGFβ signaling inhibits cellular proliferation through the up regulation of the cell cycle inhibitors *P15* and *P21* and the repression of the positive regulators of the cell cycle, *c-MYC* and *Cdc25a*. TGFβ signaling contributes to apoptotic signaling by inducing *PTEN* expression. Loss of TGFβ often occurs in colorectal cancer: *TβRII* is mutated in most of MIN tumors and in some of CIN (Grady et al., 1999), *SMAD* is often mutated in CIN tumors, higher percent (35%) is found in metastases (Miyaki et al., 1999).

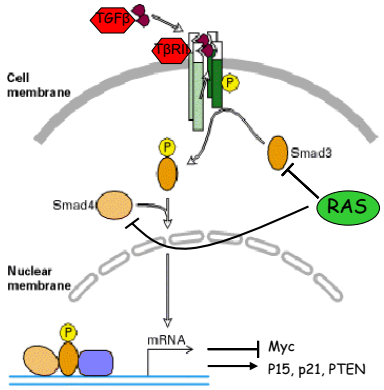

Figure 22: TGFβ signaling pathway (Lodish et al., 2000)
The ligand complexes with the receptor, leading to phosphorylation of *SMAD3* or *SMAD2* and together with *SMAD4* they enter the nucleos and activate transcription.

2.2.6 Summary

	CIN – Chromosomal instability RER-	**MIN – microsatellite instability RER+**
% of colorectal tumor	85	15
Evidenced with	LOH, gene amplification, hypermethylation	Slippage mutations
Progress rate	slowly	Rapidly
prognosis	bad	good
Normal Epithelium mutations	Inherited mutations in: ***APC*** – inhibits ***β-catenin*** from transcriptionally activating oncogenes, Bind *BUB*, *EB-1* ***P53*** – transcription factor activated in response to cellular stress and DNA damage.	Inherited mutations in **mismatch repair genes**
Early-adenoma Mutations	Inactivation of ***APC*** or ***β-catenin*** mutation	***APC*** slippage inactivation **mismatch repair genes** slippage inactivation and methylation silencing
Late-adenoma mutations	***RAS*** mutations - shortening of the G1 phase ***SMAD4*** mutations – transcription factor activated by the *TGFβ* signaling	***TBIIR*** slippage inactivation - *TGFβ* signaling receptor ***TCF4*** slippage gain of function – transcription factor that binds ***β-catenin***
Carcinoma mutations	***P53*** inactivation ***RB(?)*** deletion – binds to and represses *E2F* transcription factor to inhibit G1 progression.	***Bax*** slippage inactivation - pro-apoptotic protein
Metastasis mutations	***C-met(?)*** amplification - related to cell motility, proliferation, angiogenesis and invasion ***E-cadherin(?)*** deletion – cell adhesion molecule, interacts with ***β-catenin*** Other amplifications and deletions	Other genes slippage inactivation

Table 1: Summary of MIN, CIN colorectal tumors

Oncogenes are marked in green, tumor suppressors are marked in red.

2.3 Results

2.3.1 Gene expression analysis

In this work I performed several analyses to find clusters of genes with biological interest on the colorectal cancer expression data. To identify these clusters, I used the first step of the CTWC method, which simply executes the SPC algorithm. The clusters found are detailed in the following chapters and in Appendix A.

The data

The data of this analysis includes Affymetrix U133A GeneChips that were used on 144 samples taken from colon cancer patients. The samples comprise: 22 normal colon epithelium, 24 polyps (before transformation), 47 primary tumors, 16 liver metastases, 19 lung metastases, 11 normal liver and 5 normal lung.

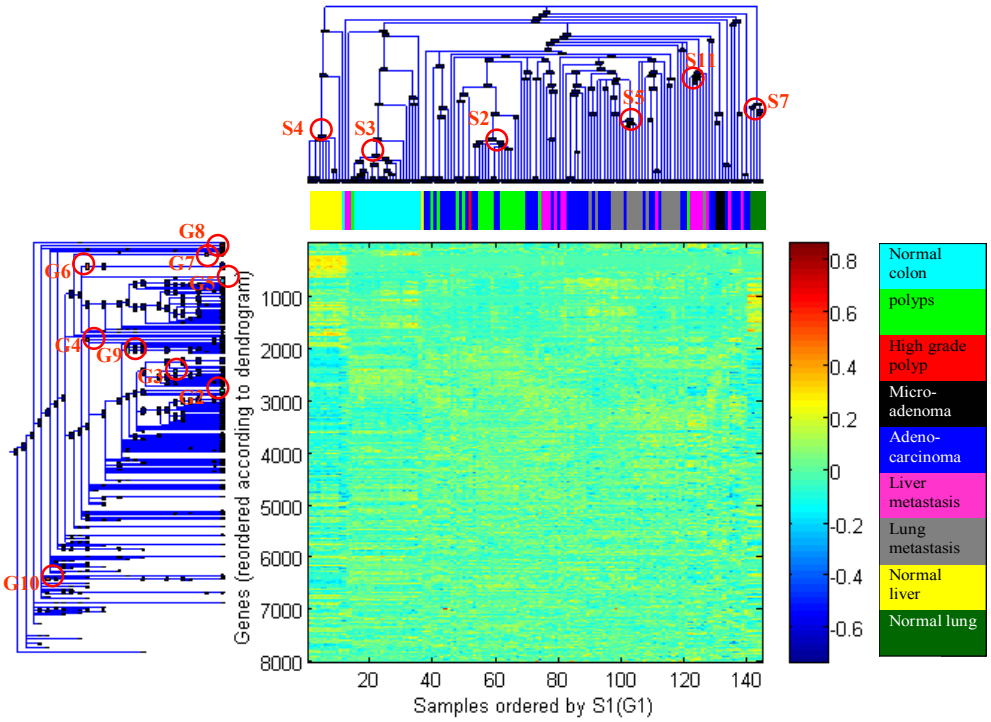

Figure 23: Super paramagnetic clustering (SPC) of the 8000 probe sets with highest variance and all samples

Description of the 10 clusters of probe sets found by SPC. The ordering of the samples is shown in the color bar above the expression matrix, the index of the colors is shown on the right.

CTWC analysis – 8000 probe sets

The analysis was performed on the 8000 probe-sets with the highest variance in the space of all 144 samples. The preprocessing procedures were:

- Microarray Suite 5.0 software of Affymetrix was used to scale the raw data
- Probe sets marked as "Absent" in all samples were removed (4959)
- Affymetrix markers were removed (47)
- Threshold data to T=70 and log2 transform
- Probe sets that had values that equals log2(threshold) in more than 90% of their samples were removed
- 8000 probe sets with the highest standard deviation were selected.
- Data was centered and normalized

The results are described in Figure 23 and Table 2.

Cluster name	Title	Size	Description	GO Analysis
Genes				
G2	colon related	60	High in adeno-carcinoma, medium in normal colon samples and low in normal liver and lung	Cell cycle, RNA splicing, protein folding, nucleotide binding
G3	Excretion	70	High in normal colon, medium in polyps	excretion, anion transport, antimicrobial humoral response
G4	Individual characteristic	88	High in 7 samples, 3 of them belong to the same person, mixed types of samples	G-protein coupled receptor activity
G5	Skeletal muscle	84	elongated cluster, high in metastasis and adeno-carcinoma, low in polyps and normal samples	skeletal development
G6	Liver related	389	High in liver	amino acid metabolism, amine catabolism, blood coagulation
G7	Muscle contamination, not related to disease	65	elongated for normal colon and for some of the adeno-carcinoma	cell motility, muscle contraction
G8	heparin binding	82	high in many of the clean metastasis, medium in polyps	heparin binding, positive regulation of cell cycle
G9	ribosomes	148	low in normal, high in polyps, metastasis and some of the adeno-carcinoma	protein biosynthesis, ribosome
G10	immunoglobulin	65	high in normal colon, low in adeno-carcinoma	antigen binding, defense response
Samples				
S2	adeno-carcinoma, polyps	13		
S3	Normal colon	17		
S4	Normal liver	10		
S5	metastasis	5		
S11	metastasis	6		
S7	Normal lung	5		

Table 2: Clusters found by the first step of CTWC analysis on the 8000 probe sets with highest variance

Cluster G3 – excretion

The expression profile of this cluster is high in the normal colon samples, medium in the polyps and low in the adeno-carcinoma tumors, metastasis, normal liver and normal lung (Figure 24). This cluster may contain tumor suppressor genes that are inactivated in cancer. One candidate gene is *Solute carrier family 26, member 3* (*SLC26A3*), which was found to be down regulated in colon adenoma (Schweinfest et al., 1993). Another explanation for this cluster could be loss of colon functions in tumor cells. The biological annotations that were most statistical significant: anion transport and excretion – the elimination of waste products that arise as a result of a metabolic activity, anion transport, antimicrobial humoral response, chemotaxis. These annotations are both related to the main roles of the colon.

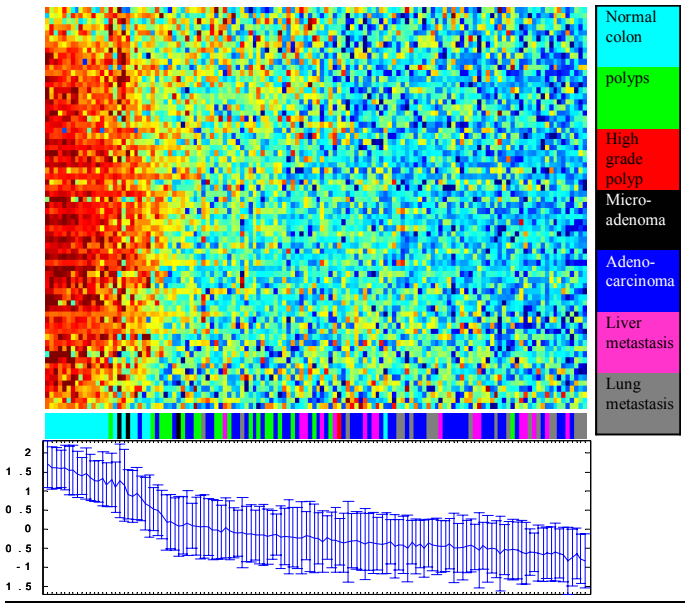

Figure 24: SPIN presentation of Cluster G3 – Excretion

This cluster is highest in normal colon, with gradual decrease in expression according to the transformation stages from polyps to metastasis. The expression matrix on top is ordered by SPIN in the space of all colon related samples (128), below is the color bar representing the ordering of the samples, using the index of colors on the right. At the bottom is the plot of the average expression levels of the cluster for patients ordered above; The error bars indicate one standard deviation.

Cluster G5 – skeletal musle

Figure 25 shows the expression levels of this cluster as well as the PCA image of the samples. The figure shows that this cluster is elongated, with lowest expression in the polyps, increasing in the normal tissues (including liver, lung) and highest in the tumors. The GO annotation of this cluster is mainly skeletal development. A similar cluster was also found by Tsafrir Dafna. According to her analysis, one explanation for this cluster is that it represents some sort of muscle contamination (Dissection of the relevant tissue without its surrounding) that doesn't appear in polyps (since the polyps are easily dissected). The normal samples seem to have this contamination as well, according to the first PCA, but in a slightly different manner according to the second PCA which is relevant only to the normal samples. Another possibility is that this skeletal development is part of the transformation process that is essential for the tumor and therefore it behaves slightly differently for the normal samples.

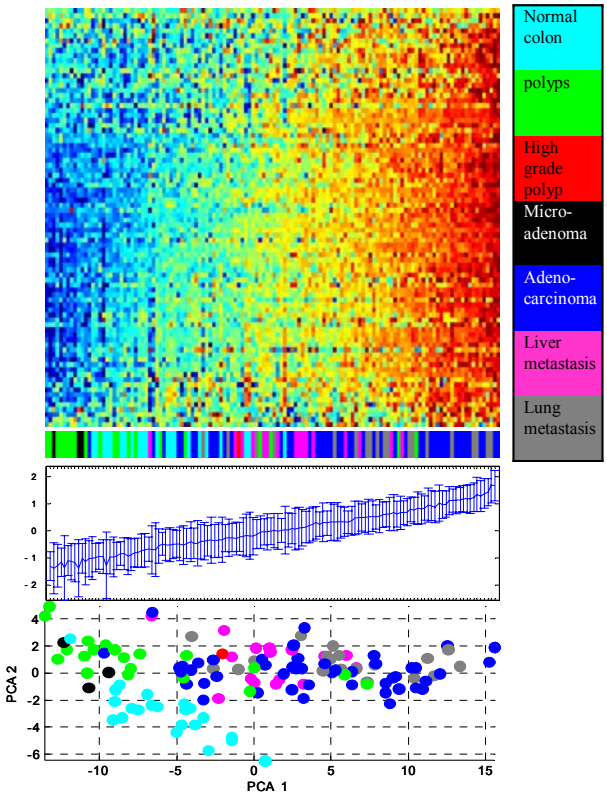

Figure 25: SPIN presentation of Cluster G5 - skeletal musle

This cluster is highest in tumors and lowest in polyps. The expression matrix on top is ordered by SPIN in the space of all colon related samples (128), below is the color bar representing the ordering of the samples, using the index of colors on the right. Underneath the color bar is the plot of the average expression levels of the cluster for patients ordered above; The error bars indicate one standard deviation. At the bottom is the PCA image of the samples. The first PCA is the degree of skeletal muscles in the tissue; the second PCA is related to normal samples only.

Cluster G8 – heparin binding

This cluster might be related to metastasis since the 15 samples with highest expression are all metastasis except for one adeno-carcinoma (Figure 26). This group of 15 samples contains most of the "clean metastasis" samples previously found by Tsafrir Dafna in her analysis. The phrase "clean" relates to metastasis samples that had the lowest contamination of liver/lung related genes. This means that the dissection process was more accurate for these samples. The GO annotation group that had the highest statistical significant was heparin binding. Heparin is a polysaccharide acid found in certain body tissues and organs used in medicine as an anti-coagulant. Exploring the genes that belong to this annotation group reveals several interesting genes known to be related to cancer:

- *VEGFB* – Vascular endothelial growth factor B, was shown to be related to angiogenesis (Silvestre et al., 2003)
- *PGF* – Placental growth factor, was shown to stimulate angiogenesis (Luttun et al., 2002)
- *FGFR* – Fibroblast growth factor receptor, this family is receptor protein tyrosin kinases that bind FGF and transduce a mitogen signal and angiogenic signal (Chen et al., 1999).

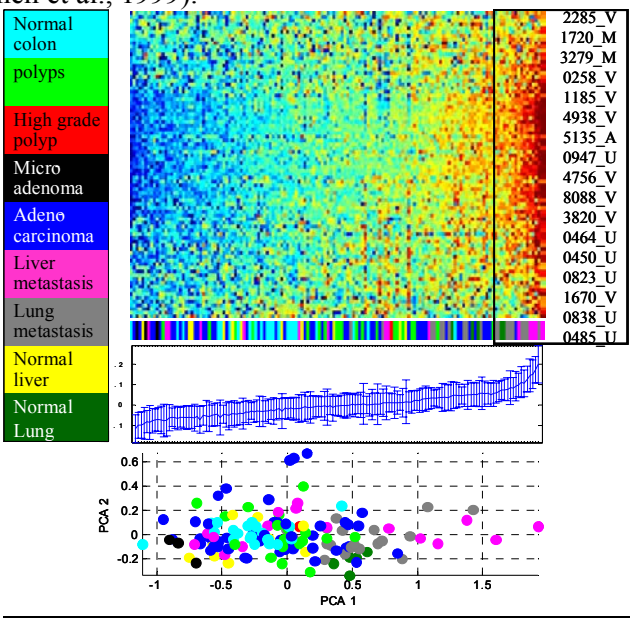

Figure 26: SPIN presentation of Cluster G8 – Heparin binding

This cluster is highest in metastasis. The expression matrix on top is ordered by SPIN in the space of all 144 samples, below is the color bar representing the ordering of the samples, using the index of colors on the right. Underneath the color bar is the plot of the average expression levels of the cluster for patients ordered above; The error bars indicate one standard deviation. At the bottom is the PCA image of the samples.

Cluster G9 – ribosomes

The cluster revealed by CTWC analysis contains 148 probe sets. SPIN was used to sharpen the expression pattern leaving a group of 91 highly correlated probe sets; almost all of them are ribosomal proteins. The expression profile of this cluster is lowest for all normal samples (including liver and lung) and higher for polyps and tumors, as shown in Figure 27. Several proto-oncogenes and tumor suppressors have recently been shown to directly regulate ribosomal activity. One of these oncogenes is *B23* (NPM1/ nucleophosmin 1), a nucleolar phosphoprotein which is over-expressed in a wide range of tumors, and functions in ribosome biogenesis (Ruggero & Pandolfi, 2003). This gene is part of this cluster. I chose to focus on this protein to investigate its connection to centrosome amplification, as described in 2.2.4. *C-myc*, another oncogene, is found to enhance the expression of a large set of genes that function in ribosome biogenesis and protein synthesis (Boon,et al,. 2001); a large group of these genes are part of this cluster. Many studies have reported correlated deregulation of protein biosythesis with cancer, but whether this is a by-product of highly proliferating cells, or whether this high expression of cluster of genes can really cause cancer, still remains to be established.

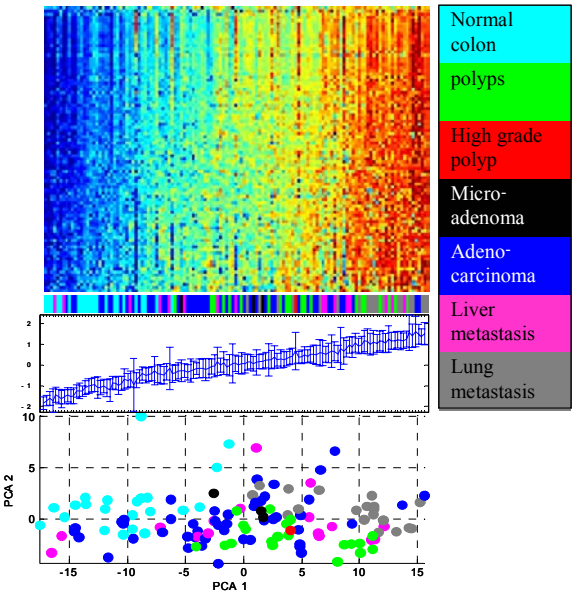

Figure 27: SPIN presentation of Cluster G9 – Ribosomes

This cluster is lowest for normal colon, and gradually increases for polyps and tumor samples, showing highest expression for lung metastasis. The expression matrix on top is ordered by SPIN in the space of all colon related samples (128), below is the color bar representing the ordering of the samples, using the index of colors on the right. Underneath the color bar is the plot of the average expression levels of the cluster for patients ordered above; The error bars indicate one standard deviation. At the bottom is the PCA image of the samples.

Cluster G10 – immunoglobulins

This cluster is enriched in Immunoglobulin proteins (antibodies). Antibodies are proteins produced by a type of white blood cell (B-cells or B-lymphocytes). These proteins bind to foreign materials (antigens) in our body and aid in their elimination. Cancer cells may sometimes express proteins on their surface that may be targets for antibodies. There are few other interesting proteins that appeared in this cluster:

- *TNFRSF7* – an antigen that belong to the Tumor necrosis factor superfamily, which play an important role in cell growth and differentiation, as well as in apoptosis or programmed cell death. This antigen was shown to induce apoptosis (Prasad et al., 1997).
- *PIM2* – a member of a family of serine/threonine protein kinases, known as an anti-apoptotic protein (White, 2003)
- *POU2AF1* - POU domain, class 2, associating factor 1, shown to directly enhance immunoglobulins transcription (Casellas et al., 2002).
- *PACAP* - proapoptotic caspase adaptor protein.

The expression profile of this cluster (Figure 28) is somewhat confusing since the highest expression belongs to the normal colon and lung metastasis, and lowest expression belong to adeno-carcinoma and liver metastasis. In between there are the normal lung, normal liver and polyps. It seems that this cluster represent some kind of defense mechanism that might be related to apoptosis, and indeed it is higher for normal colon, but then, it doesn't make sense to have the PIM2 protein in this cluster. It seems also that the tissue source of the sample plays a role in this cluster since the liver metastasis behaves very differently from the lung metastasis.

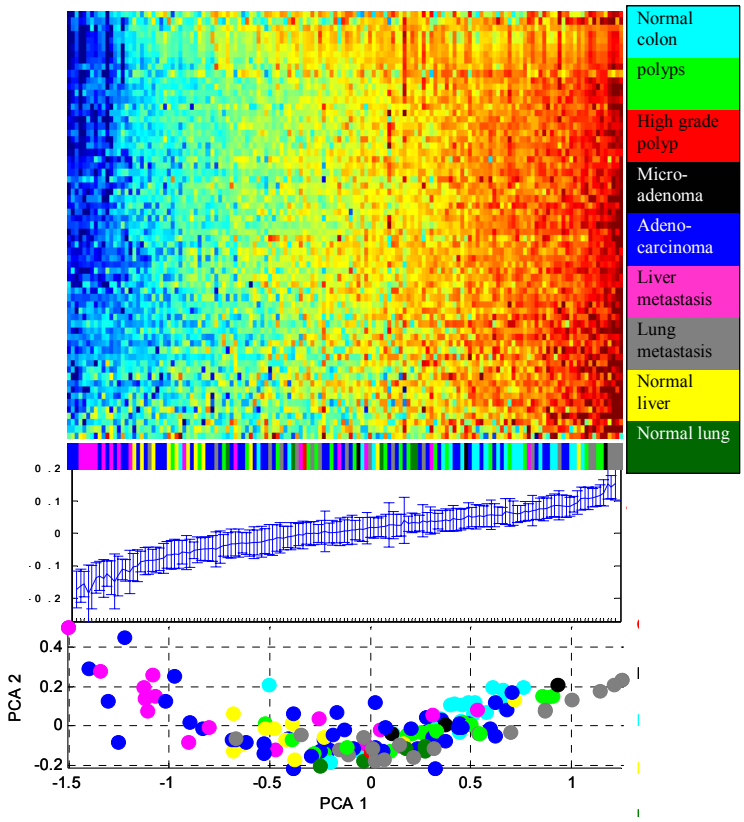

Figure 28: SPIN presentation of Cluster G10 – Immunoglobulins

The expression matrix on top is ordered by SPIN in the space of all 144 samples, below is the color bar representing the ordering of the samples, using the index of colors on the right. Underneath the color bar is the plot of the average expression levels of the cluster for patients ordered above; The error bars indicate one standard deviation. At the bottom is the PCA image of the samples.

CTWC analysis of 'clean data'– 8000 probe sets

Electronic micro-dissection is a new method, developed by Dafna Tsafrir, which identifies the purest metastasis samples in terms of contamination by surrounding normal tissue, using the SPIN algorithm (Tsafrir et al., 2005). The expression profile of normal liver tissue was used to recognize the 6 best liver metastases out of 16 available samples, while normal lung tissue provided the information needed to identify the 9 out of 19 best-dissected lung metastasis samples. The analysis was performed on the 8000 probe-sets with the highest variance in the space of colon related samples, excluding the 15 contaminated metastases and 1 normal colon sample that had characteristics of a normal liver, leaving all together 107 samples. The preprocess procedures were:

- Microarray Suite 5.0 software of Affymetrix was used to scale the raw data
- Probe sets marked as "Absent" in all samples were removed (4959)
- Affymetrix markers were removed (47)
- Threshold data to T=70 and log2 transform
- Exclude samples.
- Probe sets that have values that equals log2(threshold) in more than 90% of their samples were removed
- Based on Standard deviation the highest 8000 probe sets were chosen.
- Data was Centered and normalized

The results are shown in Figure 29. Some of the clusters were already presented in the previous section; therefore I elaborated only on the new clusters (Table 3).

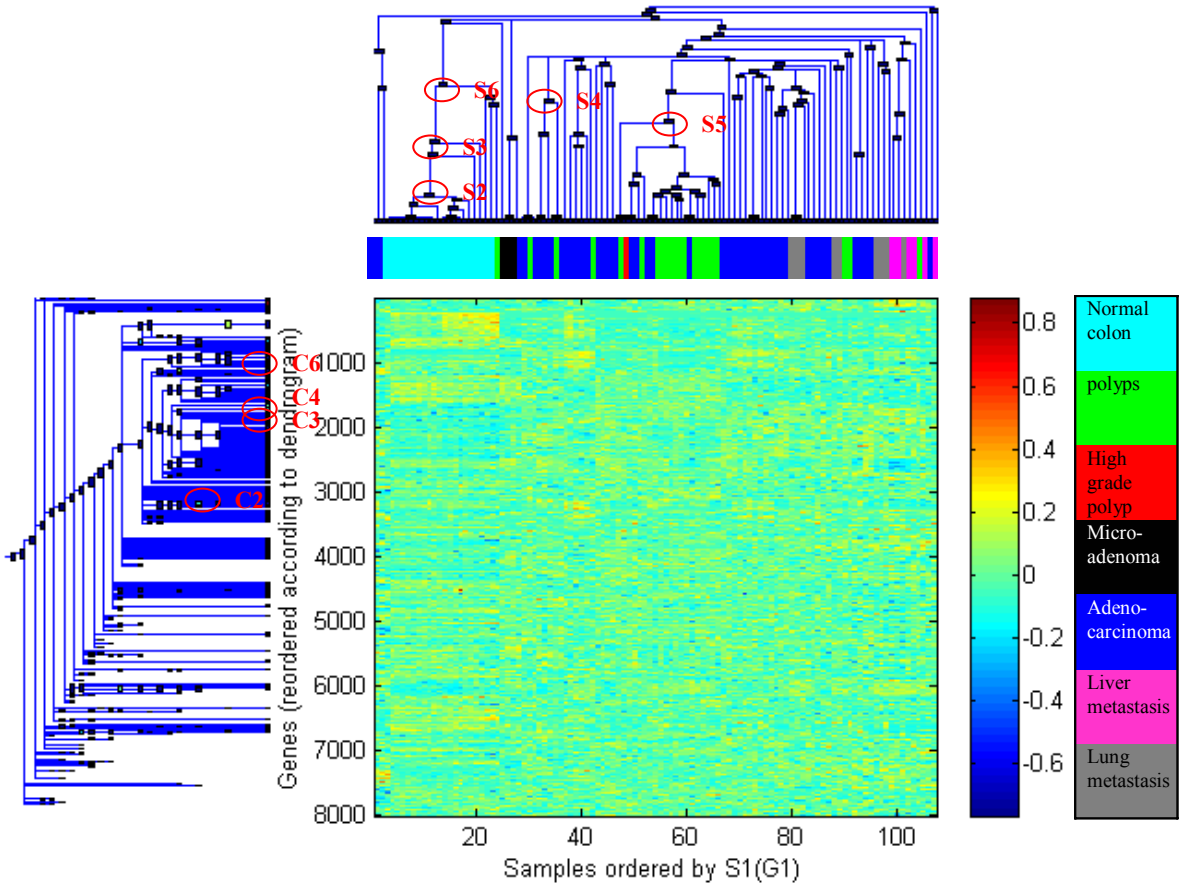

Figure 29: Super paramagnetic clustering on the 8000 probe sets with highest variance and 'clean samples'

Description of the clusters of probe sets found by SPC in the space of the selected 107 samples. The ordering of the samples is shown in the color bar above the expression matrix using the index of colors on the right.

Cluster name	Title	Size	Description	GO Analysis
Genes				
C2	immune response to pathogen/parasite	54	High in some adeno-carcinoma, normal lung and 'contaminated lung metastasis', medium in normal colon ,normal liver and 'contaminated liver metastasis'. Lowest in 'clean metastasis'.	Immune response to pest/pathogen/parasite, inflammatory response, response to wounding
C3	transcription	64	low in some adeno-carcinome and in normal colon, normal liver. Medium for normal lung, high in polyps and metastases	regulation of transcription, transcription, metabolism
C4	RNA processing	71	low in normal colon, liver and lung	RNA processing, ribosome biogenesis, RNA splicing, small nucleolar ribonucleoprotein complex
C6	energy	53	high in normal colon, polyps and some of the adeno-carcinome. Low in all the rest	Mitochondrion, oxidative phosphorylation
Samples				
S2	Normal colon	16		
S3	Normal colon	18		
S4	Adeno-carcinoma, 1 polyp	5		
S5	Adeno-carcinoma, polyps	21		
S6	Normal colon	21		

Table 3: clusters found by the first step of CTWC analysis of the 'clean data'

The analysis was performed on the 8000 probe sets with highest variance in the space of the selected 107 samples.

Cluster C4 – RNA processing

This cluster is probably linked to the ribosomal cluster (cluster G9) since its annotation is mainly related to RNA processing and ribosome biogenesis, however, it behaves a little different then the ribosomal cluster in that the metastases have lower expression levels (Figure 30).

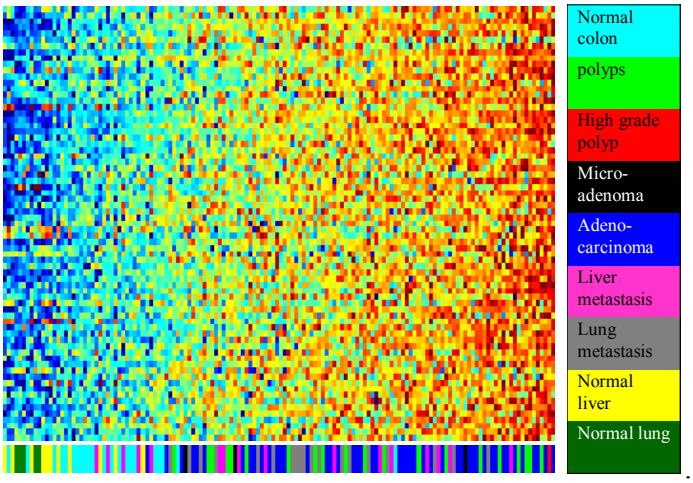

Figure 30: SPIN presentation of Cluster C4 – RNA processing

The expression matrix is ordered by SPIN in the space of all 144 samples, below is the color bar representing the ordering of the samples, using the index of colors on the right.

Cluster C6 – Energy

The annotation of this cluster was related to mitochondrion processes, including oxidative phosphorylation and synthesis of ATP. The expression level of this cluster is higher for normal colon, polyps and some adeno-carcinoma, and lower for all the rest of the samples (Figure 31). It could be a result of loss of colon functions for the transformed samples.

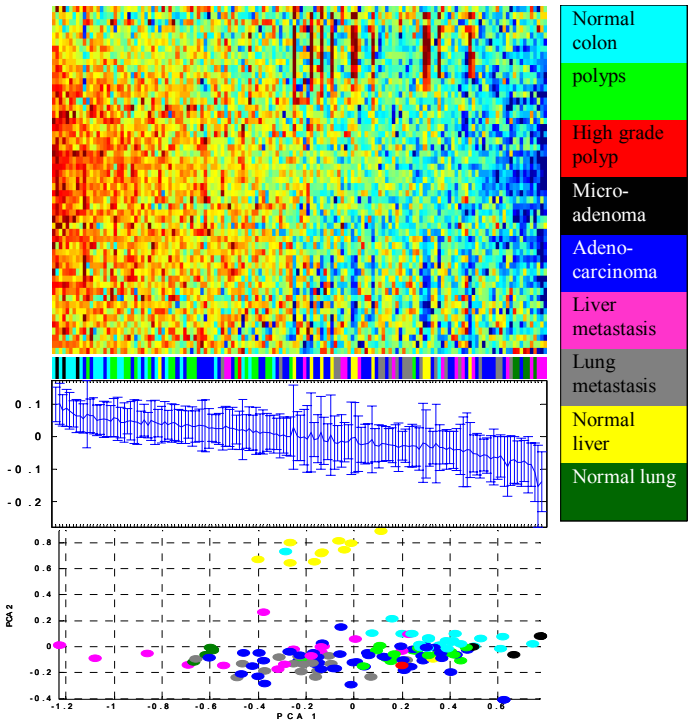

Figure 31: presentation of Cluster C6 – Energy

This cluster is highest in normal colon and polyps. The expression matrix on top is ordered by SPIN in the space of all 144 samples, below is the color bar representing the ordering of the samples, using the index of colors on the right. Underneath the color bar is the plot of the average expression levels of the cluster for patients ordered above; The error bars indicate one standard deviation. At the bottom is the PCA image of the samples.

Proliferation cluster

During my work in I came across more clusters that appeared in other CTWC runs. Among them I found the proliferation cluster, which I will discuss in detail in the section on cervical cancer. Figure 32A shows the original cluster (69 probe sets) revealed by CTWC in the space of all samples. Since the metastases seemed to be too scattered, I expanded the cluster only for normal colon, polyps and adeno-carcinoma. The expression profile of the expansion (207 probe sets) is shown in Figure 32B. It seems that the expression of this cluster reflects the stages of transformation, from normal to polyps and then to primary tumor. The intersection an analogous cluster found in the cervical cancer data (163 probe sets) is 97 probe sets. The most significant annotations for this cluster are: mitotic cell cycle, M phase, DNA metabolism, regulation of cell cycle and S-phase.

This cluster contains many probe sets related to cell cycle regulation; some of them are described in the biological background section (2.1):

Bub1, *mad2*, *PTTG1* (securing), *CDC20* are mitotic spindle proteins active at prometaphase in mitosis (Jallepalli and Lengauer, 2001). *CSE1L* is another protein responsible for sister chromatids separation. *DNMT1* is responsible for DNA methylation. *TTK* (*Mps1p*) has a role in centrosome duplication together with *B23 (NPM1)* (Hinchcliffe and Sluder, 2001), *PLK* (Zhang et al., 2004), while *RAN*, *Nek2* are centrosome localized proteins (Saavedra et al., 2003). *TOP2A*, *CKS2*, *CDK4* and many other proteins in this cluster are related to cancer.

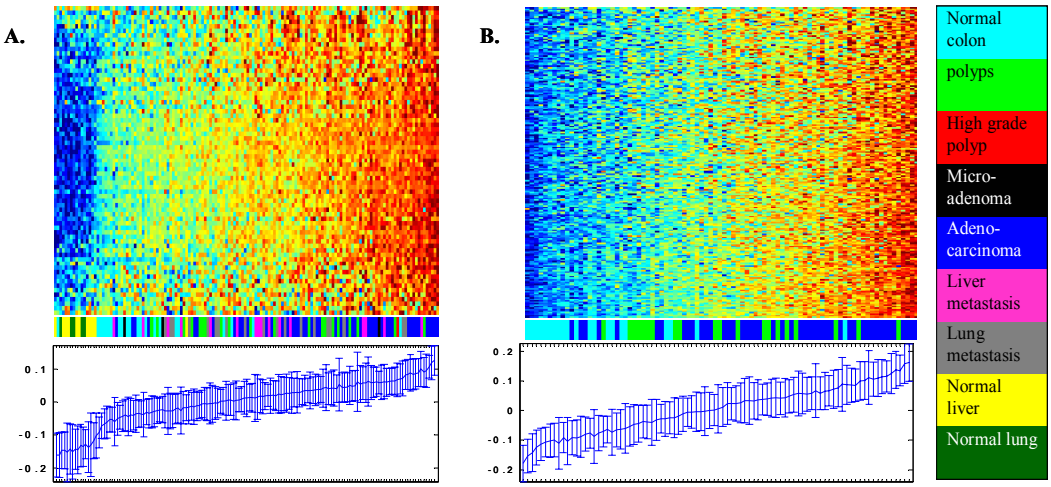

Figure 32: SPIN presentation of the proliferation cluster

A. The cluster found by CTWC. The expression matrix on top is ordered by SPIN in the space of all 144 samples. Below the matrix is the color bar representing the ordering of the samples, using the index of colors on the right. Underneath the color bar is the plot of the average expression levels of the cluster for patients ordered above; The error bars indicate one standard deviation. B. This is the expansion of the cluster in A, in the space of normal colon, polyps and adeno-carcinoma samples. Lowest expression level is observed in normal samples, medium in polyps and highest in adeno-carcinomas.

Two-way Anova

The two-way Anova test was designed as described in Table 4: Normal colon, Normal lung and Normal liver versus colon adeno-carcinoma, liver and lung metastasis.

	Normal	cancer
colon	22	47
liver	11	16
lung	5	19

Table 4: Two-way Anova

The Anova test produces three groups of probe sets: Pathology related, Tissue type related and interaction, as described in Figure 33. To avoid false positive genes I used 1% Bonferroni statistical confidence. The main purpose of the Anova test was to identify genes that separate the samples on their pathology basis, without being influenced by tissue type separation.

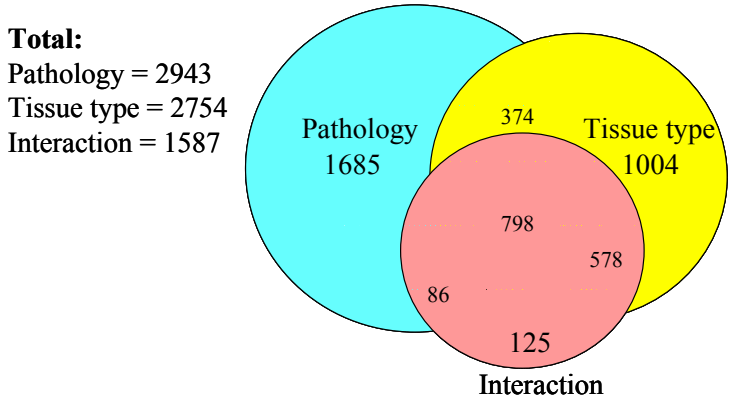

Figure 33: Two-way Anova results

Pathology

SPC was used to analyze the group of probe sets that separated according to pathology (the blue part in Figure 33), revealing clusters that were already discussed. Figure 34 shows the results of SPC.

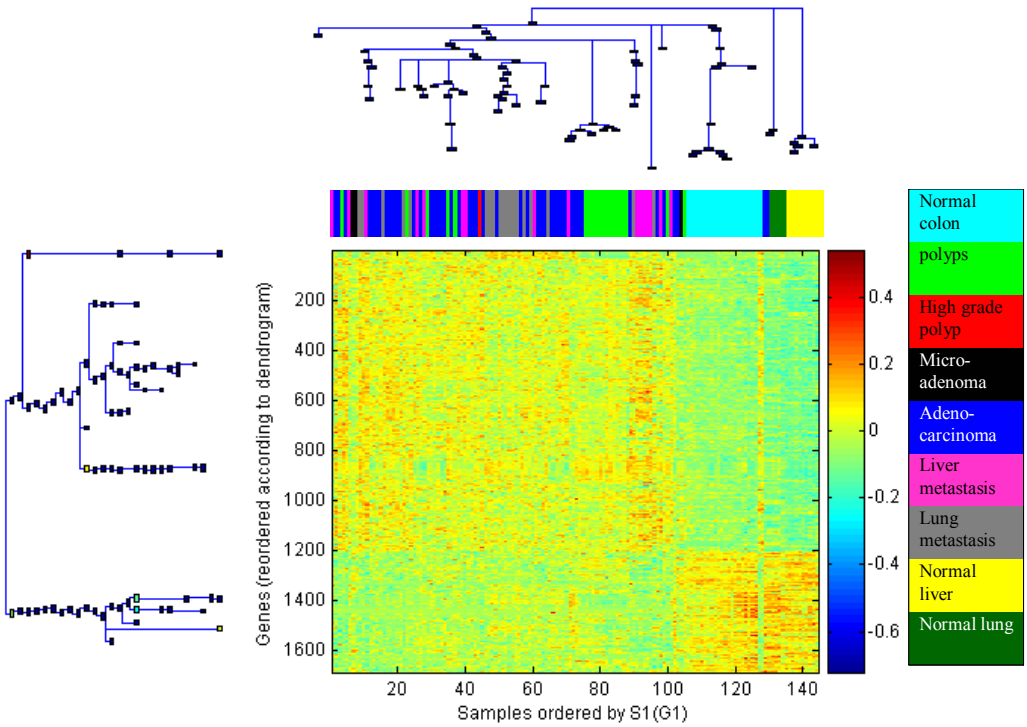

Figure 34: Super paramagnetic clustering (SPC) of the pathology probe sets and all samples

The dendrogram drawn by SPC in the space of all 144 samples. The ordering of the samples is shown in the color bar above the expression matrix, using the index of colors on the right.

Tissue type analysis

The results of the tissue type analysis were as expected: the Anova test discovered genes that separated the three types of tissues, as shown in Figure 35.

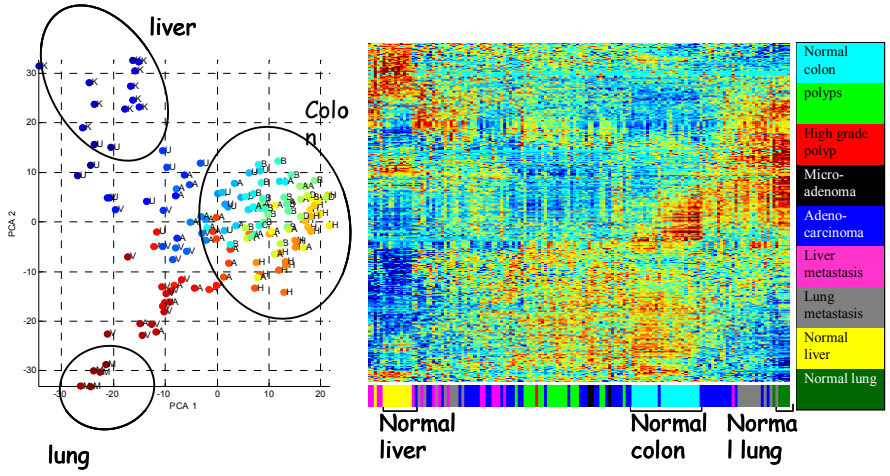

Figure 35: SPIN presentation of the Tissue type separating probe sets

Left: First and second PCA, showing the separation of the samples into three major groups. Right: The sorted expression matrix of the probe sets, the index of colors on the right.

2.3.2 B23 and Centrosome amplification

The protein B23 caught my attention when I first saw it as part of the ribosomal cluster; it was among a few other proteins that were not ribosomal. Reading more on the ribosomal cluster revealed that this protein is considered to be an oncogene (Ruggero & Pandolfi, 2003).

Another article that caught my attention was about the *Myc* oncogene, a transcription factor, which was found to cause centrosomes amplification when it was over-expresseded together with DNA damage (Sugihara et al, 2004). *Myc* was found to enhance the expression of genes that function in ribosome biogenesis and protein synthesis, including *B23* (Boon et al., 2001). Also, *B23* was identified as a direct target for *c-Myc* (Zeller et al., 2001).

Since *B23* has a role also in centrosomes duplication (for more details check section 2.2.4), I started to think that maybe *B23* causes centrosomes amplification when it is over-expressed.

Only few articles linked *B23* to centrosome amplification (Saavedra et al., 2003; Ussar & Voss, 2004; Zhang et al, 2004). The most common explanation for this link, given by Saavedra et al., was through early phosphorylation of *B23* at G1, when the kinase *CDK2*/*cyclinE* was over-expressed. This early phosphorylation could lead to more than one duplication of centrosome in the same cell cycle, which results in centrosome amplification.

Another possible option, briefly discussed by Ussar & Voss is that the over-expression of *CDK4/cyclinD* could cause early phosphorylation of *B23*. They showed that the activation of *CDK4*, together with over-expression of *B23* lead to centrosome amplification (Ussar & Voss, 2004).

In the expression data one of the probe sets of *B23* was part of the ribosomal cluster and another probe sets was part of the proliferation cluster (These two cluster are also correlated with one another since both show similar profiles: low in normal samples and high in the rest). *CDK4* is highly correlated with *B23* (mean correlation with all 3 probe sets of *B23* is 0.62), perhaps because they are both transcriptionaly activated by *Myc* (Zeller et al., 2001; Menssen and Hermeking, 2002).

When analyzing the colon cancer expression data it is obvious that *B23* is over-expressed in colon cancer (Figure 36) and it was shown that B23 is over expressed in other types of cancer (Chan et al., 1989; Kondo et al., 1997; Ruggero et al., 2003). The question is why is it over-expressed? Does the cancer cell gain any advantage

from this over-expression, or is it just a by-product of over-expression of its transcriptional activator (*c-MYC*)? I wanted to answer this question by checking whether over-expression of *B23* can cause centrosome amplification.

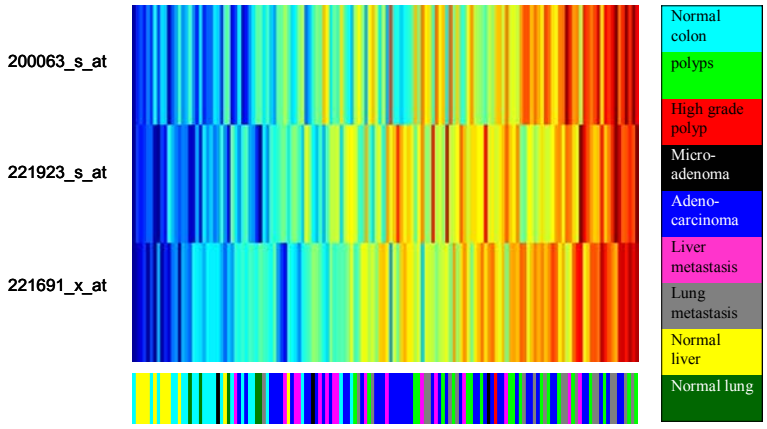

Figure 36: SPIN presentation of the B23 probe sets

The expression matrix is ordered by SPIN in the space of all 144 samples. Below the matrix is the color bar representing the ordering of the samples, the index of colors on the right.

To summarize the statements made above, Figure 37 displays the "suggested hypothesis", showing the relationships between some of the genes of interest in colorectal cancer. APC is the first mutation that starts the transformation process and might be involved in causing the chromosomal instability (CIN) of the cancer. I show here two possible ways to get to CIN: mutant *APC* can cause amplification of *Myc* by elevating *β-catenin* levels. *Myc*, as a transcription factor, causes the over-expression of *cycE*, *B23*, *CDK4* that might be involved in the amplification of centrosomes. *APC* can also cause miss-segregation of sister chromatids by the interaction with *Bub1*. At last, both centrosomes amplification and mis-segregation can cause CIN.

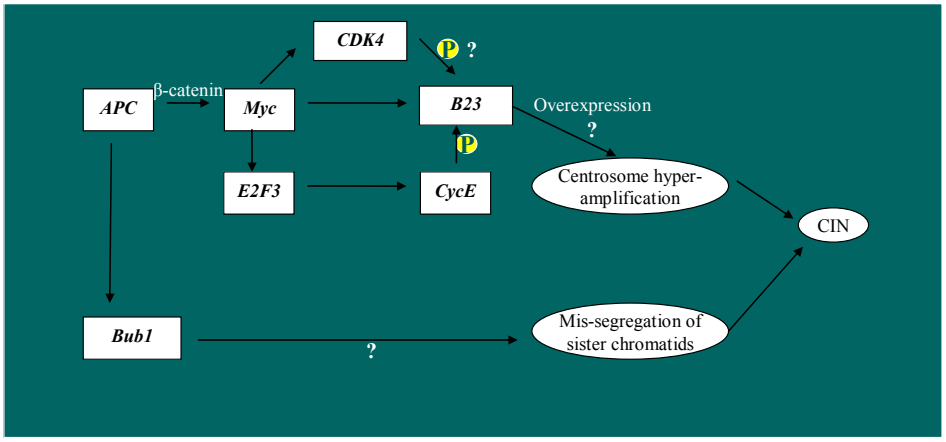

Figure 37: Suggested hypothesis

The experimental question I wanted to check was: does hyper-phosphorylation of *B23* contribute to centrosome amplification? Or is it just the timing of the phosphorylation that causes the amplification? In order to answer this I started to study lab techniques in Prof. Moshe Oren's lab. One of the methods is transfection of DNA plasmids that are injected into the cells. The transfected DNA is then translated in the cell causing over-expression of the plasmid's genes. After the transfection (~48 hours) the cells are fixated and stained for gamma-tubulin in order to see if there are centrosomes amplifications. The goal was to perform a transfection of two types of plasmid into cancer cell-line that lack *P53*, *RB*: one plasmid containing *B23* and *CDK2/cyclinE* DNA, and the second plasmid containing only *CDK2*/*cyclinE* DNA. If *B23* hyper-phosphorylation is actually important for the amplification of centrosomes, I would expect to see the more amplification in the cells containing the transfected *B23*.

Due to technical problems and lack of time, I was able to get only to the stage of being able to see good staining of centrosomes (Figure 38). I hope to be able to complete this experiment in the near future.

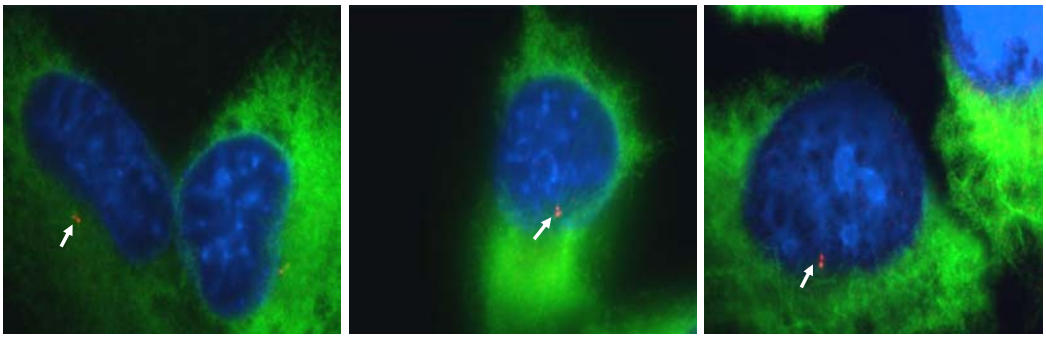

Figure 38: Centrosome staining in H1299 cells

H1299 cells were fixed and immunostained with anti-γ-tubulin antibodies to view centrosomes (red), as indicated by arrows. The nuclei were visualized by DAPI staining (blue), α-tubulin was stained using anti-α-tubulin antibodies (green).

3 Cervical cancer

3.1 Introduction

This work was done in collaboration with the Curie Institute, Paris, France and was submitted as an article to Oncogene magazine on March 2005. Our collaborators from the Curie Institute supplied the gene expression and qRT-PCR data, my part in this work was to analyze the data. The first part of this work (Global data overview) was done by Tsafrir Dafna and Ilan, I continued my work based on their findings. This chapter consists, to a large extent, of an early version of the submitted manuscript.

DNA sequences of specific HPV types are detected in the vast majority of invasive cervical carcinoma (Bosch et al., 1995), a worldwide and frequent disease (Ferlay et al., 2001). HPV 16 and 18, corresponding to highly oncogenic genotypes, are detected in 59% and 15% of the cases, respectively (Munoz et al., 2003). E6 and E7 viral oncoproteins are major contributors to neoplastic progression by interfering with cell cycle G1-S checkpoint (for review, (zur Hausen et al., 2002)). Among a variety of cellular targets, E6 binds and degrades TP53 protein by forming a complex with the ligase E6AP. E7 abrogates pRB protein function through its ubiquitination-mediated degradation, which leads to activation of E2F regulated genes and genetic instability. Integration of viral sequences into the host genome interrupts E2 open reading frame, leading to the constitutive expression of E6/E7 in the transformed cells (Romanczuk et al., 1992).

Most cases of early stage cervical carcinoma can be cured by a combination of surgery, radiotherapy and chemotherapy (Gerbaulet et al., 1992; Morris et al., 1999). However, some tumors relapse at short term and are lethal in most of the cases, despite chemotherapy (Omura et al., 1997). Little is known about the biological mechanisms which could account for these differences in clinical behavior. Viro-clinical studies have reported that the outcome of cervical cancer was related to the type of HPV associated to the tumor. A favourable course was observed for tumors associated with HPV58 and related types (Lai et al., 1999) whereas association with HPV18 was found to be indicative of poor outcome (Lombard et al., 1998; Burger et al., 1996).

To get insight into the molecular mechanisms controlling the progression of invasive cervical carcinoma, we have designed a gene expression study on cases selected according to viral and clinical parameters. HPV16- and HPV18-associated tumors

were included in order to determine whether specific gene expression profile could characterize these HPV types. To determine whether a characteristic pattern of gene expression could be linked to the disease course, we also analyzed cases with favourable outcome and tumors which presented an early relapse uncontrolled by the treatment. A combination of unsupervised Coupled Two-Way Clustering (CTWC) (Getz et al., 2000) and Sorting Points Into Neighborhoods (SPIN) (Tsafrir et al., 2005) methods was employed to mine the expression data, together with the use of rigorous statistical tests, thus combining the benefits of both knowledge and data driven approaches. One major finding of our analysis was the identification of a 'Cervical Cancer Proliferation Cluster' (CCPC) composed of 163 highly correlated transcripts, many of which corresponded to genes controlling cell proliferation. We found that tumors with an early relapse had an average expression level of CCPC genes higher than that of tumors with a favourable course, suggesting that the CCPC may be indicative of disease outcome. Moreover, we showed that E6/E7 mRNA expression was positively correlated with the expression level of the CCPC genes and to viral DNA load. Altogether, these findings suggest that tumor cell proliferation is dependent on E6/E7 mRNA levels and that HPV DNA load, positively correlated to E6/E7 mRNA level, may be associated with the outcome of invasive carcinoma of the uterine cervix.

3.2 Materials and Methods

3.2.1 Cervical Tissue Samples and Cell Lines

Tumor samples (n=60) from invasive cervical carcinoma were selected from the Institut Curie tumor bank. Patient's age ranged from 23 to 78 year-old (median, 46). Clinical stage (International Federation of Gynecology Obstetrics), available for 56 cases, was IB for 32 cases, IIA for 3 cases, IIB for 14 cases, IIIA for 2 cases, and IIIB for 5 cases. Histological analysis showed that 46 tumors were squamous cell carcinomas (SSC) and 14 adenocarcinomas (AC). HPV status characterization, determined by PCR (Rosty et al., 2004), showed that 35 tumors were associated with HPV16 (29 SSCs and 6 ACs), 13 tumors with HPV18 (6 SSCs and 7 ACs), and 1 tumor with both HPV16 and HPV18. The remaining cases corresponded to HPV33 (n=1), HPV58 (n=1), HPVx (undetermined type, other than 6, 11, 16, 18, 31, 33, 35, 39, 42, 45, 52 and 58) (n=5), and HPV-negative tumors (n=5). Relapse-free survival (RFS), known for 58 patients, was defined as the interval elapsed between the date of

the first symptoms and that of local recurrence and/or distant metastasis. Cases with RFS >5 years (n=30) were classified as diseases with favourable outcome and those with RFS <3 years (n=27) as diseases with unfavourable outcome.

All tumor samples had been flash-frozen and stored at -80°C. Histological analysis of tumor tissues adjacent to the selected samples showed that tumor samples contained >50% of invasive carcinoma cells.

Normal exocervical mucosa has been sampled and flash-frozen from 5 hysterectomy specimens, removed for non-cervical diseases. Cell lines (IC1, IC3-8) were established from human primary invasive cervical carcinoma (Couturier et al., 1991; Sastre-Garau et al., 2000). IC1 (HPV18), IC3 (HPV18), IC6 (HPVx) and IC8 (HPVx) were derived from AC (3 HPV18- and 1 HPV16-associated), whereas IC4, IC5 and IC7, derived from SSC, were found associated with HPV45, 1 HPV18 and HPV16, respectively. The primary tumors corresponding to IC5, IC6, and IC8 were included among the 60 tumor samples.

3.2.2 Labeling and Microarray Hybridization

A total of 45 cervical samples were analyzed: 30 taken from invasive carcinoma with 5 duplicates, 5 from normal mucosa, and 5 corresponding to carcinoma-derived cell lines (IC1, IC3, IC5-7). Invasive carcinoma samples were composed of 20 SSCs and 10 ACs; HPV type was 16 for 16 tumors, 18 for 12 tumors, 33 for 1 tumor and undetermined for 1 tumor; disease outcome was favourable for 15 cases, unfavourable for 13 cases, and unknown for 2 cases. Total RNAs were extracted from each sample by caesium chloride ultracentrifugation. RNA quality was assessed by visualization of the 28S/18S ribosomal RNA ratio on electrophoresis gel. Complementary RNA target was prepared and labelled as described in the Affymetrix GeneChip Expression Analysis Technical Manual. The labelled target was hybridized to Affymetrix HG-U133A oligonucleotide microarray, representing 22 215 probe sets. To control the reproducibility of the results, hybridization was performed in duplicate for 5 tumor samples.

3.2.3 Quantitative Real-Time PCR

Reverse transcription was performed using 1 μg of total RNA, random hexamer primer, and the SuperScript II reverse transcription kit (Invitrogen). Real-time PCR was performed in the SYBR green format for STK6, H2AFZ, KPNA2, CDC20 and to amplify E6 and E7 HPV transcripts for HPV16 and HPV18 tumors. Primers can be

obtained from the corresponding author upon request. For normalization, TBP expression was used. The Applied Biosystems Assays-on-Demand™ Gene Expression system was used to analyze gene expression of 16 human genes: ANKT, GGH, CCNB2, BUB1B, FEN1, CCNB1, OIP5, MELK, MCM4, UBE2C, PLK, CDC2, ZWINT, CCNA2, TOPK, RRM2. All samples were tested in duplicate. Analysis was performed using SDS v2.1 software (Applied Biosystems) according to the manufacturer's instructions. For each mRNA sample, a difference in CT values (ΔCT) was calculated by taking the mean CT of duplicate reaction and subtracting the mean CT of the duplicate reaction of the reference (TBP) RNA. A normal cervical sample was used as the calibrator. The $2^{-\Delta\Delta CT}$ method was used for quantification of gene expression.

HPV viral load was quantified using E7 primers specific for HPV16 and HPV18, in 35 HPV16 tumors and in 18 HPV18 tumors. We chose PSA as the reference gene. Comparative genomic hybridization analysis of the same tumor samples (manuscript in preparation) showed that the chomosomal location where PSA maps (19q13) has little if any variation in DNA copy number.

3.2.4 Data Analysis

Data Preprocessing. The Microarray Suite 5.0 software (MAS v5.0, Affymetrix) was used to scale the raw data and produce an expression matrix, where each value is the expression level of one transcript measured in one sample. In order to avoid working with unreliably small numbers, the remaining genes were threshold to 10 (i.e. all signal values below 10 were set to 10) and a log2 transformation was applied (Tsafrir et al., 2005). When the Affymetrix data was used to measure correlation with E7 PCR measurments, log2 and thresholding were not applied.

Genes were chosen for unsupervised analysis on the basis of their standard deviations. For duplicated tumor samples, the assigned value was the average of the two duplicates (except for the unsupervised analysis in the global overview, where both duplicates were represented).

Unsupervised analysis: Since hypothesis testing can not reveal unexpected partitions, unsupervised techniques, such as clustering, are more suited for such a task. The CTWC (Coupled Two-Way Clustering) method (Getz et al., 2000) focuses on correlated subsets of genes and samples, such that when one is used to cluster the other, stable and significant partitions emerge. The underlying algorithm is based on

iterative clustering, enabling identification of biologically relevant subsets of the data. (see also 2.1.3) This reveals partitions and correlations that are masked when the full dataset is used in the analysis. For example, when a particular set of genes is used to cluster the samples we find that they divide into 2 groups: a relatively tight cluster of predominantly favourable outcome tumors, and a larger cluster containing both favourable and unfavourable outcome tumors. The statistical significance of this 'favourable outcome group' was measured with Fisher exact test (Fisher et al., 1935).
Another exploratory analysis method that uses groups of correlated genes for meaningful ordering of patients is SPIN (Sorting Points Into Neighbourhoods) (Tsafrir et al., 2005), check section 2.1.4.

Supervised analysis. Supervised methods were employed in order to expand and refine the list of genes that was obtained by the unsupervised step. In order to control contamination with false positive genes associated with multiple comparisons, we used the method of Benjamini and Hochberg (Benjamini et al., 1995) that defines the average false discovery rate (FDR); namely, the fraction of false positives among the list of differentiating genes.

qRT-PCR analysis

The data analysis for the qRT-PCR for the selected genes was based on samples previously analyzed with Affymetrix array with 30 additional primary tumors and 2 additional cell lines. Missing values were completed using a K-nearest neighbours algorithm (Troyanskaya et al., 2001).

E7 analysis

The samples used for this qRT-PCR analysis were 34 HPV16 tumors (including 2 cell-lines), 16 of which were used previously for the Affymetrix arrays, and 17 HPV18 tumors (including 4 cell-lines), of which 14 were used for the Affymetrix arrays. Normal samples were added with assigned values of 0 for E7 mRNA and DNA expression. Correlations of E7 expression with other genes were calculated using Spearman's Rho correlation.

Gene Ontology Annotation

For Gene Ontology (GO) annotation we used the web site http://apps1.niaid.nih.gov/David/upload.asp (Dennis et al., 2003) that produces p-values according to Fisher exact test for the statistical significance of the measured over (or under) representation of particular annotation among members of a particular group of probe sets. Another web site used for GO annotation is the Affymetrix Analysis Center http://www.affymetrix.com/analysis/netaffx/index.affx (Liu et al., 2003).

3.3 Results

3.3.1 Global data overview

Unsupervised analysis separated tumor samples according to their histological type

Gene expression profiling was performed on 45 samples (5 normal mucosa, 5 cell lines, and 35 primary tumors including 5 duplicates) with Affymetrix oligonucleotide microarray (HG-U133A). Samples in duplicates exhibited high similarity in expression profiles (average correlation of 0.95). In order to generate an overview of the data and to identify major partitions and relationships, we filtered the genes with highest variance and ordered the resulting expression matrix in SPIN (Figure 39). Two separate ordering operations were performed: one on the genes (rows; Figure 39c) and another on the samples (columns; Figure 39b). The two-way organized expression matrix permitted thus to study concurrently the structure of both samples and genes. Unsupervised ordering in the context of the most varying transcripts separated the samples in complete agreement with the nature of the 3 types of samples: normal mucosa, primary tumors and cell lines. Furthermore, a clear distinction was seen within the tumor samples according to their histological type (SSC versus AC). All this information is visually displayed in the SPIN permutated distance matrix for the samples (Figure 39b). While the PCA image (Figure 39a) provides only the top principal directions (here – 3), the distance matrices and the reordered expression matrix contain the full high-dimensional relationships (Tsafrir et al., 2005).

At this stage we did not detect an expression signal associated with differences in viral type or disease outcome.

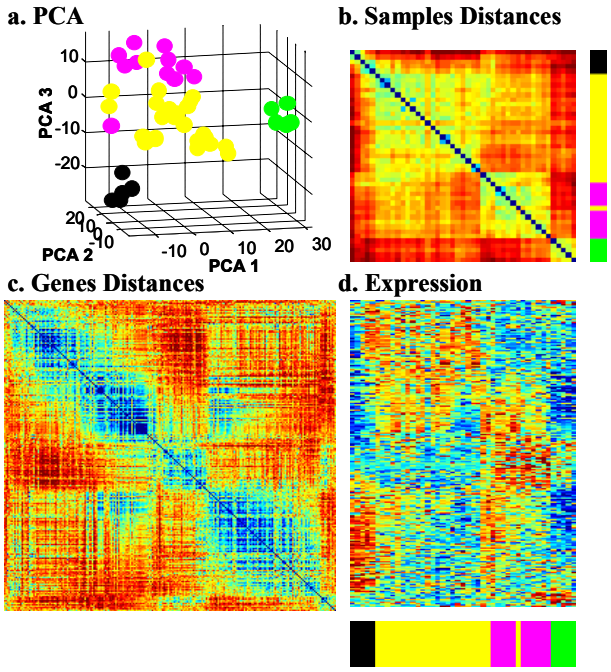

Figure 39: Global data overview

The data presented in this figure includes the 1,000 genes with highest variance over the 45 samples. Focusing on the most relevant genes by means of variance filtration facilited the computations and still gave a good overall picture of the layout of the data. (a) Projection of the samples onto the first (x-axis), second (y-axis) and third (z-axis) principal components (PC), calculated in gene-space. The nature of samples is indicated by color: normal cervical mucosa (black); carcinoma cell lines (green); primary squamous cell carcinoma (SCC, yellow) and adenocarcinoma (AC, magenta). The first PC is dominated by the differences between the cell lines and all other samples. The second PC is dominated by the differences between SSC and AC tumors. The third PC is dominated by the differences between normal samples and tumors. (b) SPIN-ordered distance matrix for the samples. Colors in the distance matrix depict dissimilarity levels between points, with red (blue) indicating large (small) distances. Hence, clusters of highly similar samples are manifested as bluish squares around the main diagonal. Note that the cell lines are a distinct, homogeneous group (marked by green in the colored bar on the right), while the normal cervical samples are clearly separated but are rather heterogeneous. (c) SPIN-ordered distance matrix for the genes. Note the grouping into several distinctive profiles. (d) Two-way SPIN-ordered expression matrix. Here colors depict relative expression intensities after centering and normalization of genes (rows), where red (blue) denotes relatively high (low) expression. Rows represent genes and columns represent samples. The colored bar below the matrix provides the tissues clinical identity.

Supervised analysis

Supervised hypothesis testing corroborated our observations regarding the grouping of samples. Using t-test with 5% FDR statistical confidence, we found that 2507 of 22,215 probe sets (11.3%) were differentially expressed in the tumor samples as compared with the normal samples, with 1206 probe sets (which include 849 unique annotated genes and 178 ESTs) showing overexpression in tumor samples. Among these, the major and most significant functional groups were: mitotic cell cycle (93),

DNA metabolism (96), DNA replication and chromosome cycle (56), regulation of cell cycle (59), and DNA repair (35). 94 of the known genes, overexpressed in tumors as compared to normal cervix, were also found to be overexpressed when compared to the cell lines. The majority of these genes were involved in immune response and were related to stroma cells. Among the tumor samples, 6.86% of the probe sets were differentially expressed in SSC as compared to AC.

As in unsupervised analysis, no single gene separated the tumor samples according to either viral type or disease outcome (using the constraint of 5% FDR significance level).

3.3.2 A gene cluster associated with disease outcome includes mostly proliferation genes

In a second step, we focused on one gene cluster, including 163 probe sets, identified by using CTWC on the 5000 probe sets with the highest variance (full list is given in appendix B). The expression profile of the genes of this cluster separated the samples into four groups: a group composed of all normal samples, a second group including 7 primary tumors among which 6 presented a favourable outcome ('favourable outcome group'), a third group containing the remaining primary tumors, and a fourth group composed of all cell lines (Figure 2). The p-value for having only 6 favourable outcome tumors and 1 unfavourable outcome tumor in one group is 0.06, according to one-tail Fisher exact test. It should be stressed that this not a separation according to tumor outcome, since only a subset of tumors with favourable outcome belongs to the 'favourable outcome group'. This makes it impossible to identify these groups of genes by using a supervised test designed according to tumor outcome. Figure 2 shows the expression matrix of the corresponding dendrogram: the normal samples with the lowest expression levels, the 'favourable outcome group' closest to the normal samples and the cell lines with the highest expression levels.

The 163 probe sets included in this gene cluster correspond to 123 unique genes and 16 ESTs (appendix B). Looking at Gene Ontology biological process annotations, we found that 55 of these genes were related to cell cycle, 30 to nuclear division, 28 to regulation of cell cycle, 29 to M-phase of mitotic cell cycle, and 22 to DNA replication and chromosome cycle. All these annotations have p-value $< 10^{-10}$ according to Fisher Exact test. Consequently, we refer to this gene cluster as the 'Cervical Cancer Proliferation Cluster' (CCPC).

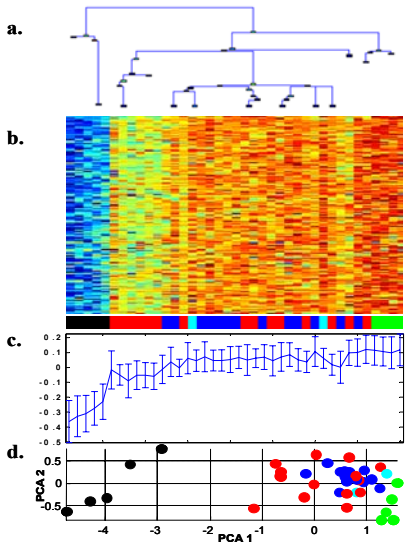

Figure 40: Pattern of expression of the 'Cervical Cancer Proliferation Cluster'(CCPC)

(a) The dendrogram above the matrix represents the clusters of the samples identified by CTWC. (b) In the expression matrix the samples were ordered according to the dendrogram (generated by clustering), whereas the genes were ordered by SPIN. The samples in clued: 5 normals, 5 cell lines, 15 favourable outcome tumors, 13 unfavourable outcome tumors, 2 unknown outcome tumors. The color bar below the matrix displays the different origin of samples: primary tumors with unfavorable outcome (blue), favorable outcome (red), unknown outcome (cyan); cell-line (green), normal samples (black). (c) Plot of the average expression levels of the CCPC genes; the error bars indicate one standard deviation. (d) PCA diagram. Data is after log, centering and normalization. The six tumors closests to the normal samples are all with favorable outcome (marked in red).

3.3.3 The 'Cervical Cancer Proliferation Cluster' in other datasets

To further support the correlation with disease outcome, we checked our CCPC on breast cancer gene expression data reported by Van't Veer et al (van't Veer et al., 2002). This study was based on a different DNA microarray, in which only 55 genes corresponding to the 163 probe sets of the CCPC were represented. Among those, we searched for genes that could separate unfavourable outcome tumors from favourable outcome tumors, using t-test.

Figure 41 displays the sorted expression data of 31 of the 49 genes that passed the 5% FDR threshold. The samples are presented in the order obtained by sorting them, using SPIN (Tsafrir et al., 2005).

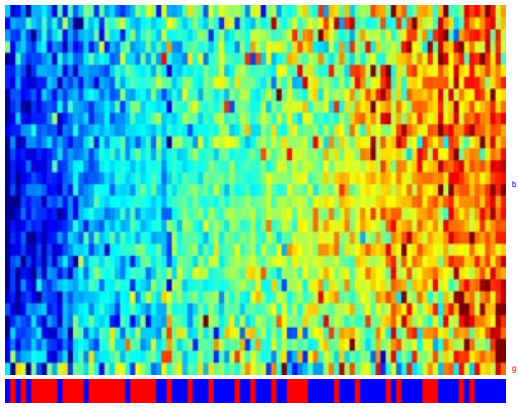

Figure 41: Correlation between the CCPC and outcome in breast cancer.

The breast data is taken from Van't veer et al. Expression levels of the 34 genes, that passed 5% FDR for t-test of favourable outcome vs. unfavourable outcome, were used to sort the samples using SPIN. The color bar below the matrix displays the different labels for outcome: unfavorable (blue), favorable (red). Low expression levels correspond predominantly to favourable outcome, and high levels to unfavourable outcome.

3.3.4 Validation of the 'Cervical Cancer Proliferation Cluster' by qRT-PCR

The identification of a proliferation gene cluster that potentially correlated with clinical outcome led us to hypothesize that a representative set of genes from this cluster could be used as a molecular signature of the course of the disease. To validate this hypothesis, we selected 20 genes from this cluster and, using qRT-PCR, we analyzed their expression level in 70 samples: the 5 cells lines and 28 of 30 invasive carcinomas with known disease outcome previously analyzed by Affymetrix array, 30 additional invasive carcinomas and 2 additional cell lines (IC4 and IC8). To select these genes, probe sets from the CCPC were sorted according to their ability to separate the 6 tumors with favourable outcome from the 13 tumors with unfavourable outcome, using a T-test and fold-change ratio between the average expressions of the two groups. The 20 genes that showed the best combination of low p-value and high fold-change were selected (Table 5). A high Pearson Correlation between qRT-PCR gene expression level and Affymetrix signal for the corresponding probe sets was observed (Figure 42).

The qRT-PCR data were analyzed in order to validate the separation based on the 'favourable outcome group'. Cluster analysis of the samples, using the qRT-PCR results revealed a group that contained the same 6 favourable outcome tumors

previously identified ('favourable outcome group'), as well as 3 additional favourable tumors with favourable outcome, 1 normal sample, and 3 unfavourable outcome tumors. The p-value for having 9 favourable outcome tumors and 3 unfavourable outcome tumors in one group is 0.076 according to one-tail Fisher exact test. These results indicated that some of the genes chosen for the qRT-PCR could be potential markers for cervical cancer outcome.

Probe Set ID	P-value	Fold-change	Title	Gene Symbol
218039_at	6.35E-06	2.06	nucleolar protein ANKT	ANKT
219978_s_at	5.36E-05	2.88	nucleolar protein ANKT	ANKT
208079_s_at	4.99E-07	2.83	serine/threonine kinase 6	STK6
203560_at	5.18E-06	2.27	gamma-glutamyl hydrolase (conjugase, folylpolygammaglutamyl hydrolase)	GGH
212141_at	7.01E-06	2.51	MCM4 minichromosome maintenance deficient 4 (S. cerevisiae)	MCM4
202705_at	7.03E-06	2.54	cyclin B2	CCNB2
204767_s_at	8.81E-06	1.64	flap structure-specific endonuclease 1	FEN1
200853_at	1.66E-05	2.11	H2A histone family, member Z	H2AFZ
204825_at	2.50E-05	2.40	maternal embryonic leucine zipper kinase	MELK
214710_s_at	2.88E-05	2.91	cyclin B1	CCNB1
209773_s_at	2.95E-05	2.63	ribonucleotide reductase M2 polypeptide	RRM2
202954_at	3.39E-05	2.07	ubiquitin-conjugating enzyme E2C	UBE2C
202240_at	6.61E-05	2.14	polo-like kinase (Drosophila)	PLK
202870_s_at	0.000161	2.69	CDC20 cell division cycle 20 homolog (S. cerevisiae)	CDC20
211762_s_at	0.000179	1.96	karyopherin alpha 2 (RAG cohort 1, importin alpha 1)	KPNA2
222036_s_at	0.000205	2.06	MCM4 minichromosome maintenance deficient 4 (S. cerevisiae)	MCM4
203418_at	0.000219	2.40	cyclin A2	CCNA2
222037_at	0.000311	2.87	MCM4 minichromosome maintenance deficient 4 (S. cerevisiae)	MCM4
219148_at	0.000325	2.55	T-LAK cell-originated protein kinase	TOPK
203213_at	0.000342	2.14	cell division cycle 2, G1 to S and G2 to M	CDC2
204026_s_at	0.000436	2.03	ZW10 interactor	ZWINT
201890_at	0.000545	2.13	ribonucleotide reductase M2 polypeptide	RRM2
213599_at	0.001371	2.93	Opa-interacting protein 5	OIP5
203755_at	0.001608	1.56	BUB1 budding uninhibited by benzimidazoles 1 homolog beta (yeast)	BUB1B
210559_s_at	0.002303	1.83	cell division cycle 2, G1 to S and G2 to M	CDC2

Table 5: The 20 genes showing separation between the 'favourable outcome tumor group' and the remaining tumors.

The second column describes the score of the T-test between the two groups; the third column describes the fold change ratio between the average expressions of the two groups. These probe sets were selected out of 84 probe sets that passed the t-test at 5% FDR.

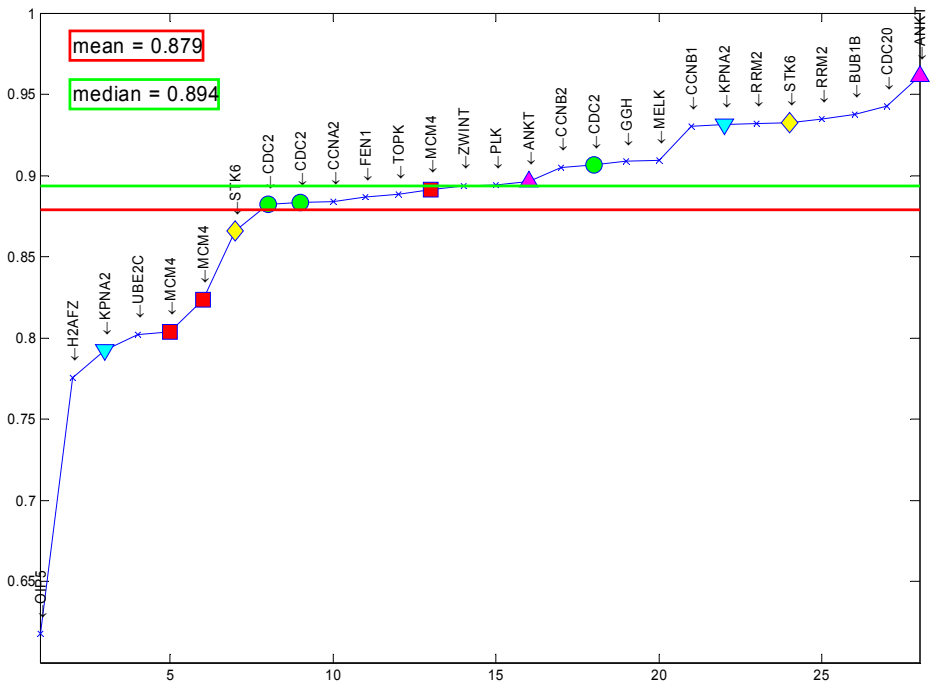

Figure 42: Correlation between gene expression measured by Affymetrix probe sets and by qRT-PCR genes.

The Pearson correlation between Affymetrix probe sets and their corresponding PCR genes ordered according to the value of the correlation. The selected 28 probe sets were those that had gene symbols identical to the 20 qRT-PCR genes. Some gene symbols had more then one probe-set. Log2 was taken on both Affymetrix and qRT-PCR data. The green line depicts the median of the correlations and the red line depicts their mean. The genes that have more than one probe set are marked in colors.

3.3.5 E6 and E7 expression correlates with the 'Cervical Cancer Proliferation Cluster' expression level and with viral load

The viral proteins E6 and E7 bind to and inhibit TP53 and pRB, respectively, driving the cell into proliferation (see Figure 43a). We therefore hypothesized that variation in expression level of the CCPC genes might correlate with E6/E7 mRNA levels. E6 and E7 mRNA expression was measured by qRT-PCR for HPV16 and HPV18 separately. Quantitative RT-PCR showed great variations in E7 expression levels among tumor samples. $2^{-\Delta\Delta CT}$ expression values ranged from 0.002 to 12.46 (mean 2.75 ±2.58) in HPV16 tumors and from 0.22 to 9.77 (mean 1.54 ±2.14) in HPV18 tumors. E6 expression was highly correlated with E7 expression in HPV16 and HPV18 tumors ($R = 0.792$, $p<0.0001$, linear regression). We thus used only E7 expression for further correlation analysis. Considering the large variations in E7 expression among the different samples, we hypothesized that the mRNA levels of E7 depended on the number of HPV genomes per neoplastic cells. To test this hypothesis, E7 DNA load

was also measured by qRT-PCR, for HPV16 and HPV18. We found that E7 mRNA expression level was correlated with E7 DNA load (Spearman correlation of 0.47 for HPV16, 0.66 for HPV18). Furthermore, E7 mRNA and DNA levels were highly correlated with the expression of the CCPC, both for the Affymetrix and the qRT-PCR measurements (Table 6 and Figure 43). We haven't found any correlation between disease outcome and E6/E7 RNA expression levels.

To evaluate the extent to which high correlation with E7 is characteristic of genes that belong to the CCPC, Spearman's Rho correlation was measured between E7 expression levels and the levels of all probe sets of the Affymetrix microarray. This was done separately for HPV-16 and HPV-18 tumors. Both highly correlated (>0.7) and anti-correlated (<-0.7) genes were considered. Table 7 summarizes the GO annotation of these lists of probe sets. The analysis for both HPV16 and HPV18 tumors showed (see Table 7) that the CCPC genes were over-represented among the probe sets whose correlation with E7 exceeded 0.7. 33 probe sets had correlation >0.7 for both HPV16 and HPV18 (p-value < 10^{-29}, hypergeometric test); 230 had correlation>0.6 for both types (out of 637 for HPV16 and 700 for HPV18, p-value < 10^{-100}, hypergeometric test).

		Mean correlation with E7 mRNA	**Mean correlation with E7 DNA**
PCR	**HPV-16**	0.553	0.335
	HPV-18	0.670	0.552
Affymetrix	**HPV-16**	0.629	0.537
	HPV-18	0.683	0.562

Table 6: Summary of the Spearman's Rho average correlation between E7 mRNA, E7 DNA and the 'Cervical Cancer Proliferation Cluster', for both the Affymetrix data and qRT-PCR data. To perform the correlation calculation, normal samples were included (E7 expression was set to 0).

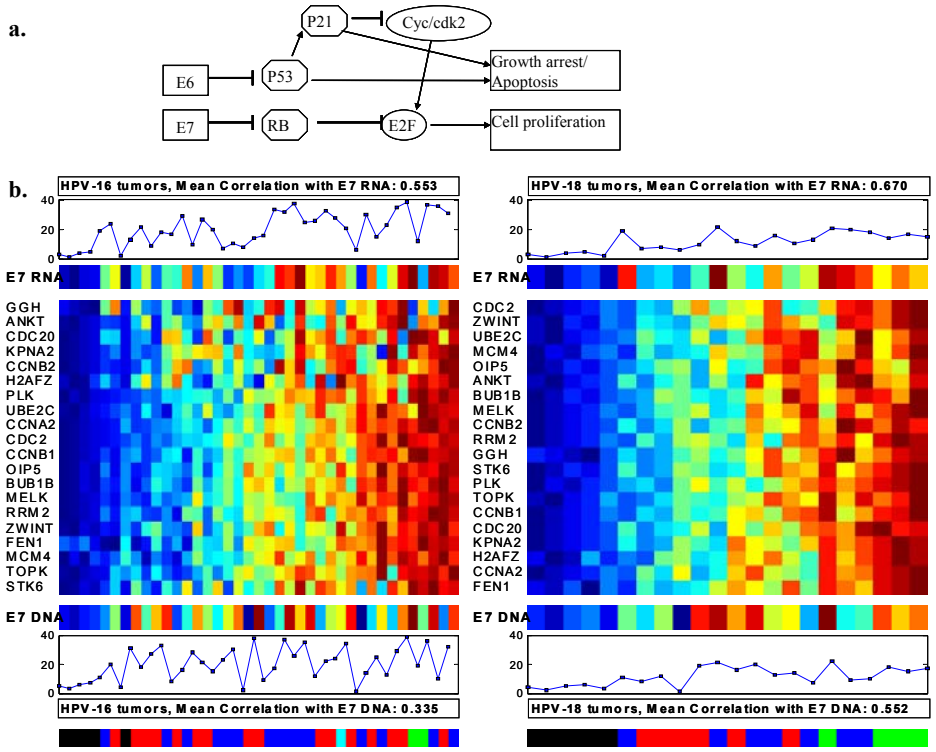

Figure 43: Correlations of qRT-PCR genes with E7 mRNA and E7 DNA. (a) Schematic drawing of the network that controls expression of the CCPC genes, indicating the manner in which the viral proteins E6 and E7 affect the network. (b) For each gene (including E7 mRNA and DNA) the measured qRT-PCR values were ranked, separately for HPV16 and HPV18 tumors. The resulting "rank matrix" was used to sort the samples; it is presented on the left for HPV16 tumors and on the right for HPV18 tumors. Rows represent genes and columns represent samples; the entry in row g and column s represents the color code for the rank of the expression level of gene g in sample s; blue entries denote low rank and red high rank. For the sake of clarity, the ranks of the samples according to E7 mRNA and DNA measurements are also represented in the two graphs at the top and bottom. The color bar at the very bottom displays the different labels for samples, using the same color scheme as in figure 2. Spearman's Rho mean correlation with E7 mRNA and DNA levels, as described in **Table 6** is presented separately for HPV-16 and HPV-18.

	Correlated	**Anti-correlated**
Hpv 16	**195** probe sets had correlation above **0.7,** 55 belong to the proliferation cluster (33.74%, p-value < 10-^{30}, hypergeometric test). **Go Annotation includes:** Cell cycle, M-phase, S-phase, regulation of cell cycle, DNA replication, DNA metabolism	**190** probe sets had correlation below **-0.7** **Go Annotation includes:** cell matrix adhesion, cell growth**,** negative regulation of cell proliferation, cell communication, transcription.
Hpv 18	**230** probe sets had correlation above **0.7,** 37 belong to the proliferation cluster (22.7%, p-value < 10-^{30}, hypergeometric test). **Go Annotation includes:** Cell cycle, M-phase, S-phase, regulation of cell cycle, DNA replication, DNA metabolism	**219** probe sets had correlation below **-0.7** **Go Annotation includes:** cell growth**,** transcription, regulation of transcription, cell differentiation, muscle development

Table 7: The GO annotations of all the probe-sets whose Spearman's correlation (or anti-correlation) with E7 is higher (or lower) then the specified threshold. Total number of probe sets on the chip is 22,215. p-value was calculated using hypergeometric test. Log2 was not taken.

3.4 Discussion

Gene expression profiling in cervical carcinoma specimens using CTWC analysis identified a cluster of 163 transcripts most of which were related to cell proliferation (CCPC genes) and were found differently expressed according to disease outcome. The expression level assessment of 20 of these genes in 70 tumor samples showed a correlation at the limit of statistical significance between high gene expression level and unfavorable disease outcome, indicating that some of the CCPC genes may be molecular markers of the clinical course of cervical cancer. Importantly, expression level of the CCPC genes were found positively correlated to E6/E7 expression level and viral load. These results suggest that high HPV DNA copy number may be related to unfavourable disease outcome by regulating the expression level of proliferation genes.

Interestingly, most of the CCPC genes are known E2F targets. Thierry et al. studied gene expression before and after infection of E2 expressing adenovirus in HeLa cervical cancer cell lines. They showed that many genes repressed by E2 are E2F-regulated mitotic genes (Thierry et al., 2004). Among the 28 mitotic genes reported to be repressed by E2, 19 (68%) are common with the CCPC, 12 of which are known E2F targets (AURKB, CDC20, CCNA2, CCNB2, MAD2L1, MKI67, MYBL2, NEK2, PTTG1, RRM2, TOP2A, UBE2C). Wells et al reported similar results (Wells et al., 2003). Large scale gene expression analysis of cervical carcinoma have been previously reported (Wong et al., 2003; Chen et al., 2003; Sopov et al., 2004}. Chen at al. aimed to identify molecular markers of high grade squamous intra-epithelial lesions and invasive carcinoma from normal cervix and low grade lesions. Four of their 62 discrimating genes were common to the CCPC genes (TK1, MYBL2, MCMC4, TOP2A). However, none of these studies focused on correlation between gene expression and clinico-pathological characteristics.

In an in vitro experiment, Tabach et al identified a cluster of genes overexpressed as a consequence of RB downregulation in cancer cell lines (Tabach et al.}. In these cell lines, p53 was regulated directly by exposing the cells to GSE56 (Milyavsky et al., 2003), while RB activity was tuned in an indirect way, by spontaneous inactivation of the INK4A gene. Cluster analysis of gene expression profiles at several time points, both before and after RB downregulation, identified a cluster of proliferation genes, analogous to the CCPC (51 out of the 163 probe sets appeared in both clusters). Our

work nicely complements this in vitro study showing that gene expression variations in cervical tumors with various E6/E7 expression levels mirror gene expression variations in cancer cell lines with varying p53 and RB expression. Furthermore, Tabach et al have also found that the expression levels of their proliferation cluster are correlated with unfavourable outcome in the van t'Veer et al (van 't Veer et al., 2002) breast cancer data.

We found a strong positive correlation between the CCPC genes expression level and E7 expression levels in invasive cervical carcinoma. Large differences in expression levels were observed between cases. This correlation may be linked to E6 and/or E7 since those genes are co-expressed in tumor cells at the same level. We found a direct correlation between E6/E7 mRNA expression levels and viral load in tissue specimens. It is unlikely that this correlation was related to differences in tumor cell density or tumor differentiation. For all cases, histological sections of tumor specimens have been analyzed showing more than 50% of invasive carcinoma cells in all samples.

Little data are available about the impact of viral DNA load and clinical course of invasive cervical carcinoma. It has been reported that a high viral load was related to a higher risk of progression from low grade intraepithelial neoplasia to high grade lesions (references) and to persistence of high grade lesions. In invasive carcinomas, a high viral load has been found positively correlated to the degree of tumor differentiation (Ikenberg et al., 1994) and negatively correlated to clinical stages (Berumen et al., 1994). However, no link has been reported between viral load and proliferation index or disease outcome. In our study, an association between disease outcome and viral load is suggested but the number of cases is too low to demonstrate such a correlation. Further studies with multivariate analysis on cases with different clinical stages is necessary to document whether viral load is a biological marker of disease outcome.

In all extensive studies reporting HPV types in invasive cervical carcinoma, 5-10% of cases were not found to be associated with HPV, using blot hybridization or PCR. This suggests that a small subset of cervical carcinoma may be HPV-negative and may have a different biology. Five HPV-negative tumors were included in our study. We have not found any difference in gene expression profiling of these tumors compared to that of HPV-associated tumors. These results suggest that these HPV-

negative tumors can not be distinguished from HPV-positive tumors and that HPV may be present in these tumors but was not detected (false negative cases).

In summary, CTWC analysis identified a proliferation gene cluster which may be correlated to disease outcome in invasive cervical carcinoma. We demonstrated a positive correlation between level of expression of the CCPC genes and E6/E7 mRNA expression levels and HPV DNA copy number. Viral load may thus correspond to a biological parameter which has to be considered in the perspective of immunotherapy or specific targeted therapy using gene silencing.

4 References

1. Alberts Molecular Biology of the Cell, 4th edition. New York: Garland Publishing, 2002.
2. Benjamini Y, Hochberg Y: Controlling the false discovery rate: A practical and powerful approach to multiple testing. J. R. Statist. Soc. 1995, 57:289-300.
3. Bertwistle, D., Sugimoto, M., and Sherr, C. J. Physical and functional interactions of the Arf tumor suppressor protein with nucleophosmin/B23. Mol Cell Biol, 24: 985-996, 2004.
4. Berumen J, Casas L, Segura E, Amezcua JL, Garcia-Carranca A: Genome amplification of human papillomavirus types 16 and 18 in cervical carcinomas is related to the retention of e1/e2 genes. Int J Cancer 1994, 56:640-645.
5. Blatt, M., Wiseman, S., and Domany, E. Superparamagnetic clustering of data. Physical Review Letters, 76: 3251-3254, 1996.
6. Bonferroni Teoria statistica delle classi e calcolo delle probabilita. Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze, 8: 3-62, 1936.
7. Boon, K., Caron, H. N., van Asperen, R., Valentijn, L., Hermus, M. C., van Sluis, P., Roobeek, I., Weis, I., Voute, P. A., Schwab, M., and Versteeg, R. N-myc enhances the expression of a large set of genes functioning in ribosome biogenesis and protein synthesis. Embo J, 20: 1383-1393, 2001.
8. Borer, R. A., Lehner, C. F., Eppenberger, H. M., and Nigg, E. A. Major nucleolar proteins shuttle between nucleus and cytoplasm. Cell, 56: 379-390, 1989.
9. Boring, C. C., Squires, T. S., Tong, T., and Montgomery, S. Cancer statistics, 1994. CA Cancer J Clin, 44: 7-26, 1994.
10. Bosch FX, Manos MM, Munoz N, Sherman M, Jansen AM, Peto J, Schiffman MH, Moreno V, Kurman R, Shah KV: Prevalence of human papillomavirus in cervical cancer: A worldwide perspective. International biological study on cervical cancer (ibscc) study group. J Natl Cancer Inst 1995, 87:796-802.
11. Breivik, J. and Gaudernack, G. Genomic instability, DNA methylation, and natural selection in colorectal carcinogenesis. Semin Cancer Biol, 9: 245-254, 1999.
12. Burger RA, Monk BJ, Kurosaki T, Anton-Culver H, Vasilev SA, Berman ML, Wilczynski SP: Human papillomavirus type 18: Association with poor prognosis in early stage cervical cancer. J Natl Cancer Inst 1996, 88:1361-1368.
13. Cahill, D. P., Lengauer, C., Yu, J., Riggins, G. J., Willson, J. K., Markowitz, S. D., Kinzler, K. W., and Vogelstein, B. Mutations of mitotic checkpoint genes in human cancers. Nature, 392: 300-303, 1998.
14. Casellas, R., Jankovic, M., Meyer, G., Gazumyan, A., Luo, Y., Roeder, R., and Nussenzweig, M. OcaB is required for normal transcription and V(D)J recombination of a subset of immunoglobulin kappa genes. Cell, 110: 575-585, 2002.
15. Chan, W. Y., Liu, Q. R., Borjigin, J., Busch, H., Rennert, O. M., Tease, L. A., and Chan, P. K. Characterization of the cDNA encoding human nucleophosmin and studies of its role in normal and abnormal growth. Biochemistry, 28: 1033-1039, 1989.
16. Chen Y, Miller C, Mosher R, Zhao X, Deeds J, Morrissey M, Bryant B, Yang D, Meyer R, Cronin F, et al: Identification of cervical cancer markers by cdna and tissue microarrays. Cancer Res 2003, 63:1927-1935.
17. Chen, W. S., Kung, H. J., Yang, W. K., and Lin, W. Comparative tyrosine-kinase profiles in colorectal cancers: enhanced arg expression in carcinoma as compared with adenoma and normal mucosa. Int J Cancer, 83: 579-584, 1999.
18. Colombo, E., Marine, J. C., Danovi, D., Falini, B., and Pelicci, P. G. Nucleophosmin regulates the stability and transcriptional activity of p53. Nat Cell Biol, 4: 529-533, 2002.

19. Couturier J, Sastre-Garau X, Schneider-Maunoury S, Labib A, Orth G: Integration of papillomavirus DNA near myc genes in genital carcinomas and its consequences for proto-oncogene expression. J Virol 1991, 65:4534-4538.
20. Dalstein V, Riethmuller D, Pretet JL, Le Bail Carval K, Sautiere JL, Carbillet JP, Kantelip B, Schaal JP, Mougin C: Persistence and load of high-risk hpv are predictors for development of high-grade cervical lesions: A longitudinal french cohort study. Int J Cancer 2003, 106:396-403.
21. Dennis G, Jr., Sherman BT, Hosack DA, Yang J, Gao W, Lane HC, Lempicki RA: David: Database for annotation, visualization, and integrated discovery. Genome Biol 2003, 4:P3.
22. Duensing, S. and Munger, K. Human papillomaviruses and centrosome duplication errors: modeling the origins of genomic instability. Oncogene, 21: 6241-6248, 2002.
23. Duval, A., Gayet, J., Zhou, X. P., Iacopetta, B., Thomas, G., and Hamelin, R. Frequent frameshift mutations of the TCF-4 gene in colorectal cancers with microsatellite instability. Cancer Res, 59: 4213-4215, 1999.
24. Ferlay F, Bray F, Pisani P, Parkin DM: Globocan 2000: Cancer incidence, mortality and prevalence worldwide, version 1.0. Iarc cancerbase no.5. Lyon: IARCPress; 2001.
25. Fisher RA: Journal of the Royal Statistical Society 1935, 98:39-82.
26. Fodde, R., Kuipers, J., Rosenberg, C., Smits, R., Kielman, M., Gaspar, C., van Es, J. H., Breukel, C., Wiegant, J., Giles, R. H., and Clevers, H. Mutations in the APC tumour suppressor gene cause chromosomal instability. Nat Cell Biol, 3: 433-438, 2001.
27. Gerbaulet AL, Kunkler IH, Kerr GR, Haie C, Michel G, Prade M, Lhomme C, Masselot M, Albano M, Dutreix A, et al.: Combined radiotherapy and surgery: Local control and complications in early carcinoma of the uterine cervix--the villejuif experience, 1975-1984. Radiother Oncol 1992, 23:66-73.
28. Getz, G., Levine, E., and Domany, E. Coupled two-way clustering analysis of gene microarray data. Proc Natl Acad Sci U S A, 97: 12079-12084, 2000.
29. Goss, K. H. and Groden, J. Biology of the adenomatous polyposis coli tumor suppressor. J Clin Oncol, 18: 1967-1979, 2000.
30. Grady, W. M., Myeroff, L. L., Swinler, S. E., Rajput, A., Thiagalingam, S., Lutterbaugh, J. D., Neumann, A., Brattain, M. G., Chang, J., Kim, S. J., Kinzler, K. W., Vogelstein, B., Willson, J. K., and Markowitz, S. Mutational inactivation of transforming growth factor beta receptor type II in microsatellite stable colon cancers. Cancer Res, 59: 320-324, 1999.
31. Guan, R. J., Fu, Y., Holt, P. R., and Pardee, A. B. Association of K-ras mutations with p16 methylation in human colon cancer. Gastroenterology, 116: 1063-1071, 1999.
32. Heide, I., Thiede, C., Sonntag, T., de Kant, E., Neubauer, A., Jonas, S., Peter, F. J., Neuhaus, P., Herrmann, R., Huhn, D., and Rochlitz, C. F. The status of p53 in the metastatic progression of colorectal cancer. Eur J Cancer, 33: 1314-1322, 1997.
33. Hinchcliffe, E. H. and Sluder, G. Centrosome duplication: three kinases come up a winner! Curr Biol, 11: R698-701, 2001.
34. Hiscox, S. and Jiang, W. G. Association of the HGF/SF receptor, c-met, with the cell-surface adhesion molecule, E-cadherin, and catenins in human tumor cells. Biochem Biophys Res Commun, 261: 406-411, 1999.
35. Ho GY, Burk RD, Klein S, Kadish AS, Chang CJ, Palan P, Basu J, Tachezy R, Lewis R, Romney S: Persistent genital human papillomavirus infection as a risk factor for persistent cervical dysplasia. J Natl Cancer Inst 1995, 87:1365-1371.
36. Ikenberg H, Sauerbrei W, Schottmuller U, Spitz C, Pfleiderer A: Human papillomavirus DNA in cervical carcinoma--correlation with clinical data and influence on prognosis. Int J Cancer 1994, 59:322-326.

37. Itahana, K., Bhat, K. P., Jin, A., Itahana, Y., Hawke, D., Kobayashi, R., and Zhang, Y. Tumor suppressor ARF degrades B23, a nucleolar protein involved in ribosome biogenesis and cell proliferation. Mol Cell, 12: 1151-1164, 2003.
38. Jallepalli, P. V. and Lengauer, C. Chromosome segregation and cancer: cutting through the mystery. Nat Rev Cancer, 1: 109-117, 2001.
39. Jones, P. A. and Laird, P. W. Cancer epigenetics comes of age. Nat Genet, 21: 163-167, 1999.
40. Kaplan, K. B., Burds, A. A., Swedlow, J. R., Bekir, S. S., Sorger, P. K., and Nathke, I. S. A role for the Adenomatous Polyposis Coli protein in chromosome segregation. Nat Cell Biol, 3: 429-432, 2001.
41. Kondo, T., Minamino, N., Nagamura-Inoue, T., Matsumoto, M., Taniguchi, T., and Tanaka, N. Identification and characterization of nucleophosmin/B23/numatrin which binds the anti-oncogenic transcription factor IRF-1 and manifests oncogenic activity. Oncogene, 15: 1275-1281, 1997.
42. Kurki, S., Peltonen, K., Latonen, L., Kiviharju, T. M., Ojala, P. M., Meek, D., and Laiho, M. Nucleolar protein NPM interacts with HDM2 and protects tumor suppressor protein p53 from HDM2-mediated degradation. Cancer Cell, 5: 465-475, 2004.
43. Lai HC, Sun CA, Yu MH, Chen HJ, Liu HS, Chu TY: Favorable clinical outcome of cervical cancers infected with human papilloma virus type 58 and related types. Int J Cancer 1999, 84:553-557.
44. Lengauer, C., Kinzler, K. W., and Vogelstein, B. Genetic instability in colorectal cancers. Nature, 386: 623-627, 1997.
45. Liu G, Loraine AE, Shigeta R, Cline M, Cheng J, Valmeekam V, Sun S, Kulp D, Siani-Rose MA: Netaffx: Affymetrix probesets and annotations. Nucleic Acids Res 2003, 31:82-86.
46. Lodish Molecular Cell Biology, 4th edition. New York: W. H. Freeman & Co, 1999.
47. Lombard I, Vincent-Salomon A, Validire P, Zafrani B, de la Rochefordiere A, Clough K, Favre M, Pouillart P, Sastre-Garau X: Human papillomavirus genotype as a major determinant of the course of cervical cancer. J Clin Oncol 1998, 16:2613-2619.
48. Lothe, R. A. Microsatellite instability in human solid tumors. Mol Med Today, 3: 61-68, 1997.
49. Luttun, A., Tjwa, M., Moons, L., Wu, Y., Angelillo-Scherrer, A., Liao, F., Nagy, J. A., Hooper, A., Priller, J., De Klerck, B., Compernolle, V., Daci, E., Bohlen, P., Dewerchin, M., Herbert, J. M., Fava, R., Matthys, P., Carmeliet, G., Collen, D., Dvorak, H. F., Hicklin, D. J., and Carmeliet, P. Revascularization of ischemic tissues by PlGF treatment, and inhibition of tumor angiogenesis, arthritis and atherosclerosis by anti-Flt1. Nat Med, 8: 831-840, 2002.
50. Malkhosyan, S., Yasuda, J., Soto, J. L., Sekiya, T., Yokota, J., and Perucho, M. Molecular karyotype (amplotype) of metastatic colorectal cancer by unbiased arbitrarily primed PCR DNA fingerprinting. Proc Natl Acad Sci U S A, 95: 10170-10175, 1998.
51. Markey MP, Angus SP, Strobeck MW, Williams SL, Gunawardena RW, Aronow BJ, Knudsen ES: Unbiased analysis of rb-mediated transcriptional repression identifies novel targets and distinctions from e2f action. Cancer Res 2002, 62:6587-6597.
52. Menssen, A. and Hermeking, H. Characterization of the c-MYC-regulated transcriptome by SAGE: identification and analysis of c-MYC target genes. Proc Natl Acad Sci U S A, 99: 6274-6279, 2002.
53. Miyaki, M., Iijima, T., Konishi, M., Sakai, K., Ishii, A., Yasuno, M., Hishima, T., Koike, M., Shitara, N., Iwama, T., Utsunomiya, J., Kuroki, T., and Mori, T. Higher frequency of Smad4 gene mutation in human colorectal cancer with distant metastasis. Oncogene, 18: 3098-3103, 1999.

54. Morin, P. J., Sparks, A. B., Korinek, V., Barker, N., Clevers, H., Vogelstein, B., and Kinzler, K. W. Activation of beta-catenin-Tcf signaling in colon cancer by mutations in beta-catenin or APC. Science, 275: 1787-1790, 1997.
55. Morris M, Eifel PJ, Lu J, Grigsby PW, Levenback C, Stevens RE, Rotman M, Gershenson DM, Mutch DG: Pelvic radiation with concurrent chemotherapy compared with pelvic and para-aortic radiation for high-risk cervical cancer. N Engl J Med 1999, 340:1137-1143.

56. Muller H, Bracken AP, Vernell R, Moroni MC, Christians F, Grassilli E, Prosperini E, Vigo E, Oliner JD, Helin K: E2fs regulate the expression of genes involved in differentiation, development, proliferation, and apoptosis. Genes Dev 2001, 15:267-285.
57. Munoz N, Bosch FX, de Sanjose S, Herrero R, Castellsague X, Shah KV, Snijders PJ, Meijer CJ: Epidemiologic classification of human papillomavirus types associated with cervical cancer. N Engl J Med 2003, 348:518-527.
58. Nigg, E. A. Centrosome aberrations: cause or consequence of cancer progression? Nat Rev Cancer, 2: 815-825, 2002.
59. Nigro, J. M., Baker, S. J., Preisinger, A. C., Jessup, J. M., Hostetter, R., Cleary, K., Bigner, S. H., Davidson, N., Baylin, S., Devilee, P., and et al. Mutations in the p53 gene occur in diverse human tumour types. Nature, 342: 705-708, 1989.
60. Okuda, M. The role of nucleophosmin in centrosome duplication. Oncogene, 21: 6170-6174, 2002.
61. Okuda, M., Horn, H. F., Tarapore, P., Tokuyama, Y., Smulian, A. G., Chan, P. K., Knudsen, E. S., Hofmann, I. A., Snyder, J. D., Bove, K. E., and Fukasawa, K. Nucleophosmin/B23 is a target of CDK2/cyclin E in centrosome duplication. Cell, 103: 127-140, 2000.
62. Omura GA, Blessing JA, Vaccarello L, Berman ML, Clarke-Pearson DL, Mutch DG, Anderson B: Randomized trial of cisplatin versus cisplatin plus mitolactol versus cisplatin plus ifosfamide in advanced squamous carcinoma of the cervix: A gynecologic oncology group study. J Clin Oncol 1997, 15:165-171.
63. Parker, S. L., Tong, T., Bolden, S., and Wingo, P. A. Cancer statistics, 1996. CA Cancer J Clin, 46: 5-27, 1996.
64. Perucho, M. Correspondence re: C.R. Boland et al., A National Cancer Institute workshop on microsatellite instability for cancer detection and familial predisposition: development of international criteria for the determination of microsatellite instability in colorectal cancer. Cancer Res., 58: 5248-5257, 1998. Cancer Res, 59: 249-256, 1999.
65. Pihan, G. A. and Doxsey, S. J. The mitotic machinery as a source of genetic instability in cancer. Semin Cancer Biol, 9: 289-302, 1999.
66. Potter, J. D. Colorectal cancer: molecules and populations. J Natl Cancer Inst, 91: 916-932, 1999.
67. Powell, S. M., Zilz, N., Beazer-Barclay, Y., Bryan, T. M., Hamilton, S. R., Thibodeau, S. N., Vogelstein, B., and Kinzler, K. W. APC mutations occur early during colorectal tumorigenesis. Nature, 359: 235-237, 1992.
68. Prasad, K. V., Ao, Z., Yoon, Y., Wu, M. X., Rizk, M., Jacquot, S., and Schlossman, S. F. CD27, a member of the tumor necrosis factor receptor family, induces apoptosis and binds to Siva, a proapoptotic protein. Proc Natl Acad Sci U S A, 94: 6346-6351, 1997.
69. Rampino, N., Yamamoto, H., Ionov, Y., Li, Y., Sawai, H., Reed, J. C., and Perucho, M. Somatic frameshift mutations in the BAX gene in colon cancers of the microsatellite mutator phenotype. Science, 275: 967-969, 1997.

70. Ren B, Cam H, Takahashi Y, Volkert T, Terragni J, Young RA, Dynlacht BD: E2f integrates cell cycle progression with DNA repair, replication, and g(2)/m checkpoints. Genes Dev 2002, 16:245-256.
71. Romanczuk H, Howley PM: Disruption of either the e1 or the e2 regulatory gene of human papillomavirus type 16 increases viral immortalization capacity. Proc Natl Acad Sci U S A 1992, 89:3159-3163.
72. Rosty C, Couturier J, Vincent-Salomon A, Genin P, Freneaux P, Sigal-Zafrani B, Sastre-Garau X: Overexpression/amplification of her-2/neu is uncommon in invasive carcinoma of the uterine cervix. Int J Gynecol Pathol 2004, 23:13-17.
73. Ruggero, D. and Pandolfi, P. P. Does the ribosome translate cancer? Nat Rev Cancer, 3: 179-192, 2003.
74. Saavedra, H. I., Maiti, B., Timmers, C., Altura, R., Tokuyama, Y., Fukasawa, K., and Leone, G. Inactivation of E2F3 results in centrosome amplification. Cancer Cell, 3: 333-346, 2003.
75. Scholey, J. M., Brust-Mascher, I., and Mogilner, A. Cell division. Nature, 422: 746-752, 2003.
76. Schweinfest, C. W., Henderson, K. W., Suster, S., Kondoh, N., and Papas, T. S. Identification of a colon mucosa gene that is down-regulated in colon adenomas and adenocarcinomas. Proc Natl Acad Sci U S A, 90: 4166-4170, 1993.
77. Silvestre, J. S., Tamarat, R., Ebrahimian, T. G., Le-Roux, A., Clergue, M., Emmanuel, F., Duriez, M., Schwartz, B., Branellec, D., and Levy, B. I. Vascular endothelial growth factor-B promotes in vivo angiogenesis. Circ Res, 93: 114-123, 2003.
78. Sopov I, Sorensen T, Magbagbeolu M, Jansen L, Beer K, Kuhne-Heid R, Kirchmayr R, Schneider A, Durst M: Detection of cancer-related gene expression profiles in severe cervical neoplasia. Int J Cancer 2004, 112:33-43.
79. Szebeni, A. and Olson, M. O. Nucleolar protein B23 has molecular chaperone activities. Protein Sci, 8: 905-912, 1999.
80. Szebeni, A., Herrera, J. E., and Olson, M. O. Interaction of nucleolar protein B23 with peptides related to nuclear localization signals. Biochemistry, 34: 8037-8042, 1995.
81. Tabach Y, Milyavsky M, Zuk O, Yitzhaki A, Shats I, Domany E, Rotter V, Pilpel Y: Genome-wide transcription regulatory circuits controlling cellular malignant transformation. submitted.
82. Takayama, H., LaRochelle, W. J., Sharp, R., Otsuka, T., Kriebel, P., Anver, M., Aaronson, S. A., and Merlino, G. Diverse tumorigenesis associated with aberrant development in mice overexpressing hepatocyte growth factor/scatter factor. Proc Natl Acad Sci U S A, 94: 701-706, 1997.
83. Takemura, M., Sato, K., Nishio, M., Akiyama, T., Umekawa, H., and Yoshida, S. Nucleolar protein B23.1 binds to retinoblastoma protein and synergistically stimulates DNA polymerase alpha activity. J Biochem (Tokyo), 125: 904-909, 1999.
84. Tarapore, P. and Fukasawa, K. Loss of p53 and centrosome hyperamplification. Oncogene, 21: 6234-6240, 2002.
85. Thierry F, Benotmane MA, Demeret C, Mori M, Teissier S, Desaintes C: A genomic approach reveals a novel mitotic pathway in papillomavirus carcinogenesis. Cancer Res 2004, 64:895-903.
86. Tokuyama, Y., Horn, H. F., Kawamura, K., Tarapore, P., and Fukasawa, K. Specific phosphorylation of nucleophosmin on Thr(199) by cyclin-dependent kinase 2-cyclin E and its role in centrosome duplication. J Biol Chem, 276: 21529-21537, 2001.
87. Troyanskaya O, Cantor M, Sherlock G, Brown P, Hastie T, Tibshirani R, Botstein D, Altman RB: Missing value estimation methods for DNA microarrays. Bioinformatics 2001, 17:520-525.

88. Tsafrir I, Tsafrir D, Ein-Dor L, Zuk O, Notterman DA, Domany E: Sorting points into neighborhoods (spin): Data analysis and visualization by ordering distance matrices. Bioinformatics In press.
89. Ussar, S. and Voss, T. MEK1 and MEK2, different regulators of the G1/S transition. J Biol Chem, 279: 43861-43869, 2004.
90. van 't Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AA, Mao M, Peterse HL, van der Kooy K, Marton MJ, Witteveen AT, et al: Gene expression profiling predicts clinical outcome of breast cancer. Nature 2002, 415:530-536.
91. Wells SI, Aronow BJ, Wise TM, Williams SS, Couget JA, Howley PM: Transcriptome signature of irreversible senescence in human papillomavirus-positive cervical cancer cells. Proc Natl Acad Sci U S A 2003, 100:7093-7098.
92. White, E. The pims and outs of survival signaling: role for the Pim-2 protein kinase in the suppression of apoptosis by cytokines. Genes Dev, 17: 1813-1816, 2003.
93. Wong YF, Selvanayagam ZE, Wei N, Porter J, Vittal R, Hu R, Lin Y, Liao J, Shih JW, Cheung TH, et al: Expression genomics of cervical cancer: Molecular classification and prediction of radiotherapy response by DNA microarray. Clin Cancer Res 2003, 9:5486-5492.
94. Yung, B. Y., Busch, H., and Chan, P. K. Translocation of nucleolar phosphoprotein B23 (37 kDa/pI 5.1) induced by selective inhibitors of ribosome synthesis. Biochim Biophys Acta, 826: 167-173, 1985.
95. Zeller, K. I., Haggerty, T. J., Barrett, J. F., Guo, Q., Wonsey, D. R., and Dang, C. V. Characterization of nucleophosmin (B23) as a Myc target by scanning chromatin immunoprecipitation. J Biol Chem, 276: 48285-48291, 2001.
96. Zhang, H., Shi, X., Paddon, H., Hampong, M., Dai, W., and Pelech, S. B23/nucleophosmin serine 4 phosphorylation mediates mitotic functions of polo-like kinase 1. J Biol Chem, 279: 35726-35734, 2004.
97. zur Hausen H: Papillomaviruses and cancer: From basic studies to clinical application. Nat Rev Cancer 2002, 2:342-350.

5 Appendix

5.1 Appendix A

List of the remaining clusters found in the analysis of the colorectal cancer

Cluster G2 – colon related

The expression pattern of this cluster is highest for adeno-carcinoma, medium for normal colon and lowest for normal liver and lung. Metastasis and polyp samples are scattered over all expression levels, therefore I have shows the expression pattern of this cluster in the space of normal samples and adeno-carcinoma. The groups of GO annotation that were statistically significant for this cluster are cell cycle, RNA splicing, protein folding, nucleotide binding.

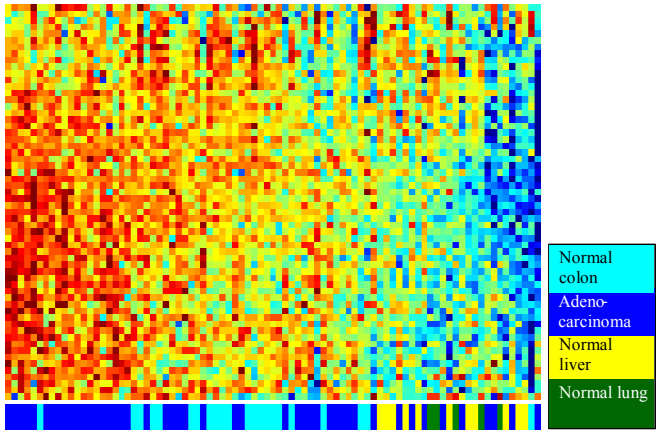

Figure 44: SPIN presentation of Cluster G2 – normal colon

This cluster is highest in adeno-carcinoma samples, medium for normal colon, and lowest in normal liver and lung. The expression matrix is ordered by SPIN in the space of adeno-carcinoma and normal samples (85). Below the matrix is the color bar representing the ordering of the samples, using the index of colors on the right.

Cluster G4 – Individual characteristic

This cluster is relevant for 7 specific samples (Figure 45), 3 of them belong to the same person, and therefore, this cluster is not related to the disease.

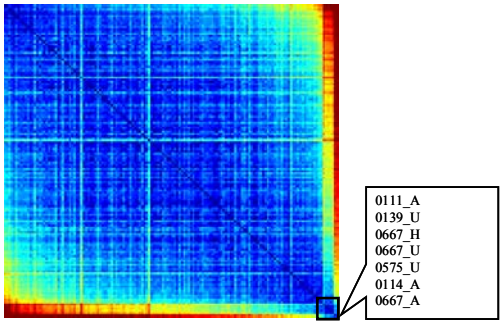

Figure 45: Cluster G4 – Individual characteristic

The distance matrix of the samples, only 7 samples show unique characteristics, 3 of them belong to the same person.

Cluster G6 – Liver related

This cluster corresponds to liver related genes as shown in Figure 46. This cluster is not related to disease.

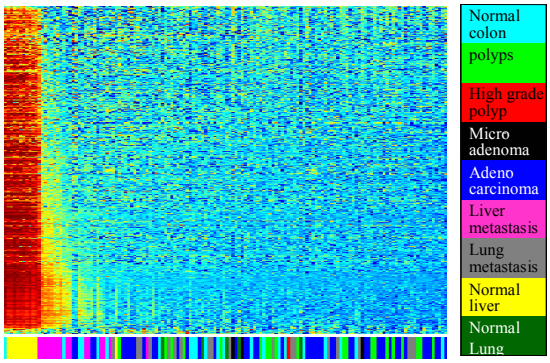

Figure 46: SPIN presentation of Cluster G6 – Liver related

The expression matrix is ordered by SPIN in the space of all 144 samples, below is the color bar representing the ordering of the samples, using the index of colors on the right.

Cluster G7 – muscle contamination

This cluster has highest expression levels for the normal samples and some of the adeno-carcinoma tumors, as described in Figure 47. The PCA image shows elongation for these samples. A similar cluster was found by Tsafrir Dafna as well and was reported as the muscle contamination cluster, again, because of dissection problems. Therefore it seems that this cluster is not related to the disease.

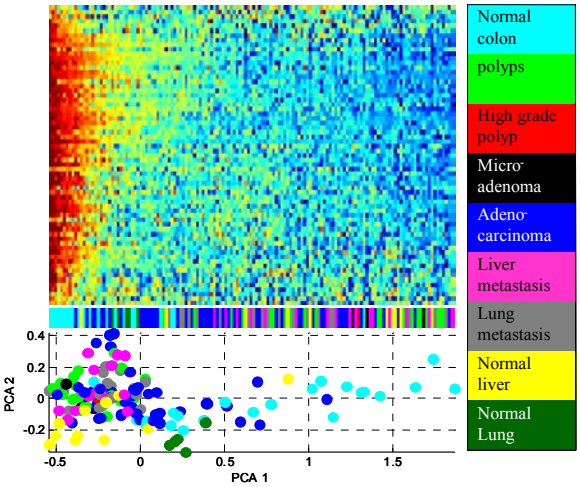

Figure 47: SPIN presentation of Cluster G7 – Muscle contamination

This cluster is highest in normal samples and some of the adeno-carcinoma tumors. The expression matrix on top is ordered by SPIN in the space of all 144 samples, below is the color bar representing the ordering of the samples, using the index of colors on the right. Underneath the color bar is the PCA image of the samples. The first PCA shows the degree of contamination in the normal colon samples.

Cluster C2 – immune response to pathogen/parasite

This cluster represents an immune response that seems to be linked to some adeno-carcinoma and normal samples. Then main annotation of this cluster is the response to pathogen/parasite. The expression pattern of this cluster, as shown in Figure 48, is high in some adeno-carcinoma, medium in normal colon and low in 'clean' metastases. Looking at the expression pattern of all samples reveals that normal lung and the 'contaminated' lung metastasis have high expression levels while normal liver and the 'contaminated' liver metastases have medium expression levels. It seems that this cluster is related to colon tissues, and it might also be related to the disease.

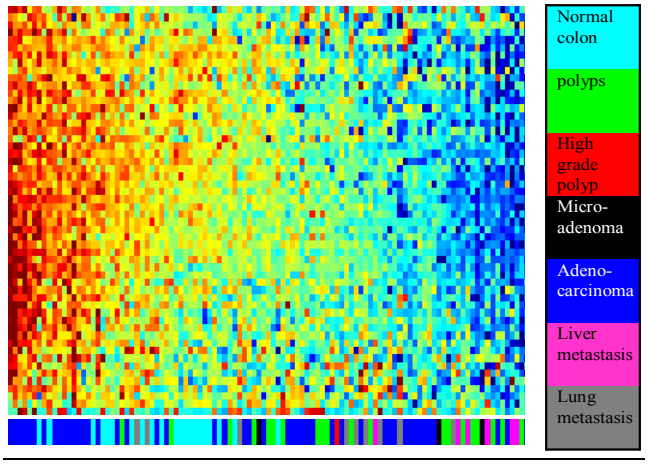

Figure 48: SPIN presentation of Cluster C2 – Immune response

The expression matrix is ordered by SPIN in the space of the 107 samples, chosen for this analysis. Below is the color bar representing the ordering of the samples, using the index of colors on the right.

Cluster C3 – transcription

This cluster's annotation is mainly related to transcription and metabolism. As shown in Figure 49 it is higher for metastases, polyps and some adeno-carcinoma. This could be related to the higher rate of protein production in tumor cells, which proliferate more often then normal cells.

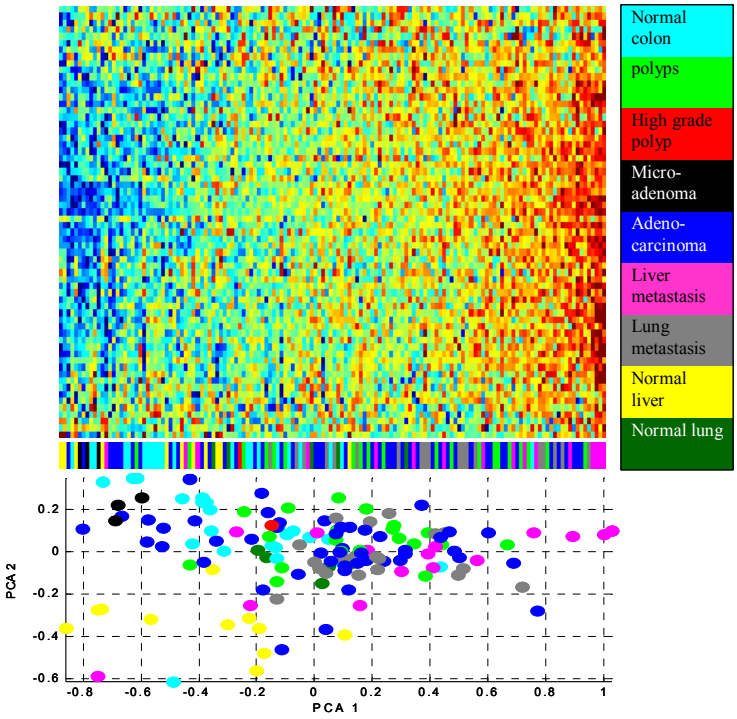

Figure 49: SPIN presentation of Cluster C3 – Transcription

This cluster is highest in metastases and polyps. The expression matrix on top is ordered by SPIN in the space of all 144 samples, below is the color bar representing the ordering of the samples, using the index of colors on the right. At the bottom is the PCA image of the samples.

5.2 Appendix B

List of the 163 probe sets of the cervical cancer proliferation cluster (CCPC)

Probe Set ID	Gene Symbol	Title
212186_at	ACACA	acetyl-Coenzyme A carboxylase alpha
218039_at	ANKT	nucleolar protein ANKT
208103_s_at	ANP32E	acidic (leucine-rich) nuclear phosphoprotein 32 family, member E
206632_s_at	APOBEC3B	apolipoprotein B mRNA editing enzyme, catalytic polypeptide-like 3B
218115_at	ASF1B	ASF1 anti-silencing function 1 homolog B (S. cerevisiae)
204244_s_at	ASK	activator of S phase kinase
219918_s_at	ASPM	asp (abnormal spindle)-like, microcephaly associated (Drosophila)
209464_at	AURKB	aurora kinase B
202094_at	BIRC5	baculoviral IAP repeat-containing 5 (survivin)
204531_s_at	BRCA1	breast cancer 1, early onset
212949_at	BRRN1	barren homolog (Drosophila)
209642_at	BUB1	BUB1 budding uninhibited by benzimidazoles 1 homolog (yeast)
203755_at	BUB1B	BUB1 budding uninhibited by benzimidazoles 1 homolog beta (yeast)
209301_at	CA2	carbonic anhydrase II
203418_at	CCNA2	cyclin A2
214710_s_at	CCNB1	cyclin B1
202705_at	CCNB2	cyclin B2
205034_at	CCNE2	cyclin E2
204826_at	CCNF	cyclin F
203213_at	CDC2	cell division cycle 2, G1 to S and G2 to M
202870_s_at	CDC20	CDC20 cell division cycle 20 homolog (S. cerevisiae)
203967_at	CDC6	CDC6 cell division cycle 6 homolog (S. cerevisiae)
221436_s_at	CDCA3	cell division cycle associated 3
221520_s_at	CDCA8	cell division cycle associated 8
207039_at	CDKN2A	cyclin-dependent kinase inhibitor 2A (melanoma, p16, inhibits CDK4)
205165_at	CELSR3	cadherin, EGF LAG seven-pass G-type receptor 3 (flamingo homolog, Drosophila)
204962_s_at	CENPA	centromere protein A, 17kDa
205046_at	CENPE	centromere protein E, 312kDa
207828_s_at	CENPF	centromere protein F, 350/400ka (mitosin)
204775_at	CHAF1B	chromatin assembly factor 1, subunit B (p60)
205394_at	CHEK1	CHK1 checkpoint homolog (S. pombe)
204170_s_at	CKS2	CDC28 protein kinase regulatory subunit 2
202532_s_at	DHFR	dihydrofolate reductase
203764_at	DLG7	discs, large homolog 7 (Drosophila)
213647_at	DNA2L	DNA2 DNA replication helicase 2-like (yeast)
220668_s_at	DNMT3B	DNA (cytosine-5-)-methyltransferase 3 beta
218567_x_at	DPP3	dipeptidylpeptidase 3
217901_at	DSG2	desmoglein 2
203270_at	DTYMK	deoxythymidylate kinase (thymidylate kinase)
202779_s_at	E2-EPF	ubiquitin carrier protein
204947_at	E2F1	E2F transcription factor 1
202735_at	EBP	emopamil binding protein (sterol isomerase)
219787_s_at	ECT2	epithelial cell transforming sequence 2 oncogene
221539_at	EIF4EBP1	eukaryotic translation initiation factor 4E binding protein 1
204817_at	ESPL1	extra spindle poles like 1 (S. cerevisiae)
203358_s_at	EZH2	enhancer of zeste homolog 2 (Drosophila)
218875_s_at	FBXO5	F-box only protein 5
204767_s_at	FEN1	flap structure-specific endonuclease 1
202580_x_at	FOXM1	forkhead box M1

203560_at	GGH	gamma-glutamyl hydrolase (conjugase, folylpolygammaglutamyl hydrolase)
218350_s_at	GMNN	geminin, DNA replication inhibitor
204318_s_at	GTSE1	G-2 and S-phase expressed 1
205436_s_at	H2AFX	H2A histone family, member X
200853_at	H2AFZ	H2A histone family, member Z
218663_at	HCAP-G	chromosome condensation protein G
204162_at	HEC	highly expressed in cancer, rich in leucine heptad repeats
220085_at	HELLS	helicase, lymphoid-specific
206074_s_at	HMGA1	high mobility group AT-hook 1
208808_s_at	HMGB2	high-mobility group box 2
207165_at	HMMR	hyaluronan-mediated motility receptor (RHAMM)
217755_at	HN1	hematological and neurological expressed 1
204444_at	KIF11	kinesin family member 11
206364_at	KIF14	kinesin family member 14
218755_at	KIF20A	kinesin family member 20A
204709_s_at	KIF23	kinesin family member 23
209408_at	KIF2C	kinesin family member 2C
218355_at	KIF4A	kinesin family member 4A
209680_s_at	KIFC1	kinesin family member C1
219306_at	KNSL7	kinesin-like 7
201088_at	KPNA2	karyopherin alpha 2 (RAG cohort 1, importin alpha 1)
203276_at	LMNB1	lamin B1
208433_s_at	LRP8	low density lipoprotein receptor-related protein 8, apolipoprotein e receptor
202736_s_at	LSM4	LSM4 homolog, U6 small nuclear RNA associated (S. cerevisiae)
203362_s_at	MAD2L1	MAD2 mitotic arrest deficient-like 1 (yeast)
210059_s_at	MAPK13	mitogen-activated protein kinase 13
220651_s_at	MCM10	MCM10 minichromosome maintenance deficient 10 (S. cerevisiae)
202107_s_at	MCM2	MCM2 minichromosome maintenance deficient 2, mitotin (S. cerevisiae)
212141_at	MCM4	MCM4 minichromosome maintenance deficient 4 (S. cerevisiae)
204825_at	MELK	maternal embryonic leucine zipper kinase
212020_s_at	MKI67	antigen identified by monoclonal antibody Ki-67
205235_s_at	MPHOSPH1	M-phase phosphoprotein 1
221437_s_at	MRPS15	mitochondrial ribosomal protein S15
201710_at	MYBL2	v-myb myeloblastosis viral oncogene homolog (avian)-like 2
204641_at	NEK2	NIMA (never in mitosis gene a)-related kinase 2
218888_s_at	NETO2	neuropilin (NRP) and tolloid (TLL)-like 2
213599_at	OIP5	Opa-interacting protein 5
203228_at	PAFAH1B3	platelet-activating factor acetylhydrolase, isoform Ib, gamma subunit 29kDa
201202_at	PCNA	proliferating cell nuclear antigen
204146_at	PIR51	RAD51-interacting protein
212858_at	PKMYT1	membrane-associated tyrosine- and threonine-specific cdc2-inhibitory kinase
218644_at	PLEK2	pleckstrin 2
202240_at	PLK	polo-like kinase (Drosophila)
213226_at	PMSCL1	polymyositis/scleroderma autoantigen 1, 75kDa
204441_s_at	POLA2	polymerase (DNA-directed), alpha (70kD)
213007_at	POLG	polymerase (DNA directed), gamma
207746_at	POLQ	polymerase (DNA directed), theta
218009_s_at	PRC1	protein regulator of cytokinesis 1
218782_s_at	PRO2000	PRO2000 protein
203554_x_at	PTTG1	pituitary tumor-transforming 1
222077_s_at	RACGAP1	Rac GTPase activating protein 1
218585_s_at	RAMP	RA-regulated nuclear matrix-associated protein
209507_at	RPA3	replication protein A3, 14kDa
201890_at	RRM2	ribonucleotide reductase M2 polypeptide

219493_at	SHCBP1	likely ortholog of mouse Shc SH2-domain binding protein 1
205339_at	SIL	TAL1 (SCL) interrupting locus
218653_at	SLC25A15	solute carrier family 25 (mitochondrial carrier; ornithine transporter) member 15
218237_s_at	SLC38A1	solute carrier family 38, member 1
213253_at	SMC2L1	SMC2 structural maintenance of chromosomes 2-like 1 (yeast)
201663_s_at	SMC4L1	SMC4 structural maintenance of chromosomes 4-like 1 (yeast)
203145_at	SPAG5	sperm associated antigen 5
204092_s_at	STK6	serine/threonine kinase 6
218308_at	TACC3	transforming, acidic coiled-coil containing protein 3
202338_at	TK1	thymidine kinase 1, soluble
203432_at	TMPO	thymopoietin
217733_s_at	TMSB10	thymosin, beta 10
201291_s_at	TOP2A	topoisomerase (DNA) II alpha 170kDa
219148_at	TOPK	T-LAK cell-originated protein kinase
210052_s_at	TPX2	TPX2, microtubule-associated protein homolog (Xenopus laevis)
204033_at	TRIP13	thyroid hormone receptor interactor 13
204822_at	TTK	TTK protein kinase
202589_at	TYMS	thymidylate synthetase
202954_at	UBE2C	ubiquitin-conjugating enzyme E2C
204026_s_at	ZWINT	ZW10 interactor
219978_s_at	ANKT	nucleolar protein ANKT
202095_s_at	BIRC5	baculoviral IAP repeat-containing 5 (survivin)
218542_at	C10orf3	chromosome 10 open reading frame 3
217851_s_at	C20orf45	chromosome 20 open reading frame 45
218741_at	C22orf18	chromosome 22 open reading frame 18
211814_s_at	CCNE2	cyclin E2
203214_x_at	CDC2	cell division cycle 2, G1 to S and G2 to M
210559_s_at	CDC2	cell division cycle 2, G1 to S and G2 to M
203968_s_at	CDC6	CDC6 cell division cycle 6 homolog (S. cerevisiae)
209644_x_at	CDKN2A	cyclin-dependent kinase inhibitor 2A (melanoma, p16, inhibits CDK4)
40020_at	CELSR3	cadherin, EGF LAG seven-pass G-type receptor 3 (flamingo homolog, Drosophila)
209172_s_at	CENPF	centromere protein F, 350/400ka (mitosin)
48808_at	DHFR	dihydrofolate reductase
213616_at	DKFZP586M1523	DKFZP586M1523 protein
218726_at	DKFZp762E1312	hypothetical protein DKFZp762E1312
38158_at	ESPL1	extra spindle poles like 1 (S. cerevisiae)
219650_at	FLJ20105	hypothetical protein FLJ20105
221685_s_at	FLJ20364	hypothetical protein FLJ20364
218351_at	FLJ20502	hypothetical protein FLJ20502
218802_at	FLJ20647	hypothetical protein FLJ20647
219990_at	FLJ23311	hypothetical protein FLJ23311
218883_s_at	FLJ23468	hypothetical protein FLJ23468
209709_s_at	HMMR	hyaluronan-mediated motility receptor (RHAMM)
205449_at	HSU79266	protein predicted by clone 23627
202503_s_at	KIAA0101	KIAA0101 gene product
206102_at	KIAA0186	KIAA0186 gene product
211762_s_at	KPNA2	karyopherin alpha 2 (RAG cohort 1, importin alpha 1)
222039_at	LOC146909	hypothetical protein LOC146909
203960_s_at	LOC51668	HSPCO34 protein
222036_s_at	MCM4	MCM4 minichromosome maintenance deficient 4 (S. cerevisiae)
222037_at	MCM4	MCM4 minichromosome maintenance deficient 4 (S. cerevisiae)
211767_at	MGC14799	hypothetical protein MGC14799
212022_s_at	MKI67	antigen identified by monoclonal antibody Ki-67
212023_s_at	MKI67	antigen identified by monoclonal antibody Ki-67

219510_at	POLQ	polymerase (DNA directed), theta
209773_s_at	RRM2	ribonucleotide reductase M2 polypeptide
201664_at	SMC4L1	SMC4 structural maintenance of chromosomes 4-like 1 (yeast)
208079_s_at	STK6	serine/threonine kinase 6
201292_at	TOP2A	topoisomerase (DNA) II alpha 170kDa
211725_s_at	BID	BH3 interacting domain death agonist

חקר ביטוי גנטי בסרטן המעי הגס ובסרטן צוואר הרחם

מיכל שפר

תזה להדרכת מוסמך מוגש למועצה המדעית של

מכון ויצמן למדע

בהדרכת

פרופסור איתן דומני

מרץ 2005