UPTEC X 07 012 ISSN 1401-2138
JAN 2007

Johan Winquist

Integration of mRNA and gene copy number measurements for elucidation of drug resistance mechanisms

Master's degree project

UPPSALA
UNIVERSITET

Molecular Biotechnology Programme

Uppsala University School of Engineering

UPTEC X 07 012

Date of issue 2007-01

Author

Johan Winquist

Title (English)

Integration of mRNA and gene copy number measurements for elucidation of drug resistance mechanisms

Title (Swedish)

Abstract

Acquired cellular resistance to cytotoxic drugs is a major problem in modern cancer treatment. This project aimed to find a high throughput approach to identify candidate genes responsible for resistance, by the means of large-scale molecular analysis and data mining. These genes could be used as targets in future attempts to resensitize cancer cells to drugs or in diagnostics of resistance. Values of *in vitro* resistance for 39 drugs were collected with FMCA (fluorometric microculture cytotoxicity assay), gene copy number and gene expression for 11246 genes with microarrays.

Integration of the three characters by pairwise Pearson correlation studies (univariate and quasi-bivariate) identified new interesting gene transcripts as well as ones already known to be involved in resistance mechanisms. The approach proved to be robust and useful in this type of analysis.

Keywords

Cancer treatment, correlation, data mining, drug resistance, gene copy number and expression

Supervisors

Anders Isaksson
Department of Medical Sciences, Uppsala University

Scientific reviewer

Tomas Olofsson
Department of Engineering Sciences, Uppsala University

Project name	Sponsors
Language **English**	Security
ISSN 1401-2138	Classification
Supplementary bibliographical information	Pages **22**

Biology Education Centre Biomedical Center Husargatan 3 Uppsala
Box 592 S-75124 Uppsala Tel +46 (0)18 4710000 Fax +46 (0)18 555217

Integration of mRNA and gene copy number measurements for elucidation of drug resistance mechanisms

Johan Winquist

Populärvetenskaplig sammanfattning

Dagens sjukvård har stora problem med att cancerceller utvecklar resistens mot cytostatika; upp till 90 % av alla misslyckade behandlingar av cancerpatienter med metastaser tros bottna i läkemedelsresistens. I grund och botten kan resistensen bero på ett förändrat antal kopior av vissa gener i en cell; om gener vars produkter är ansvariga för transporten av giftiga ämnen ut ur cellen ökar i antal kan cellgifterna i en cancerbehandling aldrig nå verksamma intercellulära halter.

Om man kunde identifiera de resistensgivande generna och motverkade kopietalsförändringar av dessa skulle fler cancerpatienter kunna behandlas framgångsrikt. Det är till detta projektet syftar. Utgångsmaterialet var data för 9 cancercellinjer innehållande kopietal och aktivitet för drygt 12000 gener (bestämda med hjälp av microarrays) samt resistensgrad mot 39 läkemedel. Genom att studera hur kopietal/genuttryck/resistensgrad samvarierar (parvisa Pearson korrelationer) har gener som redan är kända att vara inblandade i resistensmekanismer hittats, men också nya intressanta gener. Värdet av dessa resultat skall nu undersökas vidare i laboratoriet.
Projektet har resulterat i en storskalig, hypotesgenererande metod som integrerar olika typer av information och möjliggör en simultan analys av människans samtliga gener.

Examensarbete 20p
Civilingenjörsprogrammet Molekylär Bioteknik

Uppsala Universitet januari 2007

Table of contents

1. Abbreviations

b	base(s)
CN	Copy number
CNAG	Copy Number Analyzer for GeneChip (Software tool)
CNAT	Chromosome Copy Number Analysis Tool (Software tool)
FMCA	Fluorometric microculture cytotoxicity assay
HMM	Hidden Markov model
IC	Inhibitory concentration
SNP	Single nucleotide polymorphism

2. Aim of the project

Drug resistance is a grave concern in modern cancer treatment. The cytostatics at hand often have a narrow therapeutic window and are limited in numbers. If more information on the mechanisms of resistance is learned, new pharmaceutical approaches to cancer treatment may be formulated. One goal is to identify genes responsible for resistance to be able to resensitize the cell to drugs.

DNA lesions are common in cancer cells and closely linked to the expression of genes. Thereby, the lesions can ultimately affect the cells' degree of resistance to various compounds; genes in affected regions often display highly altered expression levels and thus can be expected to contribute to drug resistance.

To study this we used four cancer cell lines previously made resistant to various drugs through continuously increased exposure levels. These pairs (parental cell lines and their sub-strains) form a so called cell panel. Theoretically, genomic variation within a pair is associated with the mechanism of resistance. Thus, amplified or deleted genes in altered regions may be directly linked to cancer drug resistance. We hypothesized that high pair-wise correlation between cell line profiles of drug resistance, expression levels and copy numbers for genes, could identify genes responsible for drug resistance in a high throughput fashion. It is possible to compare the correlations for all genes one by one, but that is rather time consuming.

Thus, the aim was to develop a high throughput technique for identification of genes responsible for resistance to cytostatics in cancer cells. This was undertaken by correlation studies of drug resistance, gene expression and copy number data already available in the lab of Dr Anders Isaksson (Department of Medical sciences, Uppsala University).

The project has aimed to be hypothesis generating rather than a thorough investigation of each single gene.

3. Introduction

We are all aware that drug resistance of bacteria is a challenging problem for the health-care. However, drug resistance is not an issue isolated to the realm of prokaryotes, but is also a demanding problem for modern cancer treatment. Resistance to anticancer drugs is a major dilemma in chemotherapy, although the mechanisms of resistance are not conferred as for prokaryotes where the sharing of plasmids is a common cause of resistance. Over 90 % of the failed treatments of patients with metastatic cancers are estimated to be caused by drug resistance [1].

Bacterial cells in selective growth conditions can sometimes develop resistance to the selective agent; an event called “adaptive mutation”. Often, a copy number (CN) amplification can be observed in these cells. New findings explain this as a way to “add mutational targets”, i.e., the chance to obtain an advantageous mutation increases with the number of gene copies. This is preferred instead of an increase of the mutational rate, which is more likely to damage the cell severely.[2] There is, of course, an initial expression elevation effect upon gene amplification. This initial effect has proven to be essential for resistance to the drug doxorubicin in the cancer cell line RPMI 8226/Dox40, as expounded in the next section.

3.1 Mechanisms of drug resistance

There are various mechanisms behind resistance to cancer drugs, e.g. an acquired insensitivity to apoptosis signals [3] and/or inhibition of topoisomerase II [4], and can be ordered in function-based groups: alterations in influx/efflux systems of cells, modification of drug target and cellular damage repair systems, etc [1]. Studies of human cell lines have revealed that an acquired resistance to one cytotoxic compound often implies multidrug resistance. Given the vast number of different cancers, and that the total number of cytotoxic drugs approved in Sweden by the year 2000 was only 44, this is gravely concerning [5]. The degree of resistance can be measured with the fluorometric microculture cytotoxicity assay (FMCA) method (see section 3.2) [6].

The most common reason for multidrug resistance is overexpression of certain transporter proteins. These are usually energy-dependent channel proteins that detect and eject cytostatic drugs from the cell (i.e. influx/efflux system modification). One example is the ATP-binding cassette protein family, abbreviated ABC proteins [3]. The myeloma cell line RPMI 8226/Dox40, resistant to doxorubicin and used in this project, has previously been proven to overexpress the integral membrane P-glycoprotein due to gene amplification [7]. All cell lines used, along with their respective selective agents, can be found in table 1, section 4.1.

With a better understanding of the underlying mechanisms of resistance, new and more potent cancer drugs can hopefully be found. One approach is to induce resensitization of already existing treatments by affecting the mechanisms of resistance. One of the cell lines used in this project is, as previously mentioned, the RPMI 8226/Dox40 strain. The cell line has its origin in human myeloma cells normally associated with an incurable disease due to frequent development of drug resistance [7]. An example of an attempt to overcome the resistance, in this case to radiation and the drug doxorubicin, is the treatment with fludarabine. It has proven to resensitize cells by affecting the STAT1 signalling; a lowered STAT1 expression is linked to an increased sensitivity to radiation [8].

3.2 Fluorometric microculture cytotoxicity assay – in vitro drug resistance evaluation

The fluorometric microculture cytotoxicity assay (FMCA) was developed at the University Hospital of Uppsala in the early 1990s by Larsson *et al.* (described in detail in ref [9]). This method is designed to evaluate the cytotoxicity of compounds through IC50 measurements under certain conditions, i.e. the concentration at which half of the cells die. The method is a 72 h assay based on fluorescein diacetate (FDA), which is hydrolyzed to the strongly fluorescent fluorescein. The hydrolysis is only performed by living cells and in the article of Larsson *et al.*, it is stated that “the FDA fluorescence was linearly related to viable cell number within a wide range of cell densities (3-4 logs) as well as in the presence of different added proportions of dead cells”. This makes the method robust and it is widely used.

3.3 Single Nucleotide Polymorphisms and copy numbers

A single nucleotide polymorphism (SNP) is defined as a DNA sequence variation of one nucleotide found in at least 1% of the allele pool. The variation can be found in any type of sequence: exons, introns, enhancers etc. Although fairly recently discovered, SNP is the most common type of genetic sequence variation, estimated to be present, on average, every 0.3-1 kb within the genome. Today, SNPs are, among other fields of application, used with microarrays for genotyping purposes, as genetic markers for diseases, and for linkage disequilibrium studies. [10] In addition, SNPs can also be used to determine the number of copies of a certain DNA fragment by interpretation of the signal intensities from different loci. Since the most common copy number is 2 (i.e. one loci at each of the two chromosomes in a pair) a normalization and subsequent comparisons between different loci are possible. This way copy numbers can be determined.

3.4 The microarray technique

In 1995, modern microarray technique saw daylight in a laboratory at the Stanford University [11]. The new approach of gene expression quantification made high throughput analysis possible and was thereby an important methodological breakthrough. The microarray itself usually consists of a matrix of spots (so called features) of single stranded RNA or DNA on, for instance, a slide of glass. Examples of other types are protein- and immuno-arrays.

A microarray slide of today for genotyping purposes assesses up to 250.000 genomic positions[1]. To interpret data, hidden Markov model (HMM) based computer software can be used. A gene expression array assesses 47.000 transcripts through 1.3 million features [12]. This large number of independent loci prevents the use of standard t-tests for

[1] For reasons explained in section 3.4.2, the number of features on the array is greater than 250.000.

comparison of signal intensities during clustering due to the ‘multiple testing’ problem[2]. The number of replicates is usually too few to tackle this issue. In 2002, Lönnstedt and Speed (ref [13]) addressed this by the introduction of the so called b-test, which uses the empirical Bayes approach. It is, however, possible to compare the signal intensities of two separate genes from a single microarray. Algorithms for genotype calling and similar methods for interpretation of microarray data are constantly refined (see e.g. Rabbee *et al.*, ref [14]).

3.4.1 The manufacturing of microarrays

One of the leading companies of microarray manufacturing is Affymetrix. This company uses photolithography to produce their oligonucleotide arrays (so called GeneChips): First, a photo-labile blocking compound is linked to a slide of quartz. Chosen features are then activated by an exposure to radiation through a mask, which directs the light where to shine. Nucleotides with photo-labile protection groups, which prevent polymerization, are added and let to bind covalently to the growing oligonucleotide of the activated features. The array is then ready for a new activation procedure. Notably, only one type of nucleotide can be attached in each round, which results in the need of many masks to produce 25-mers (schematic picture is found in figure 1).

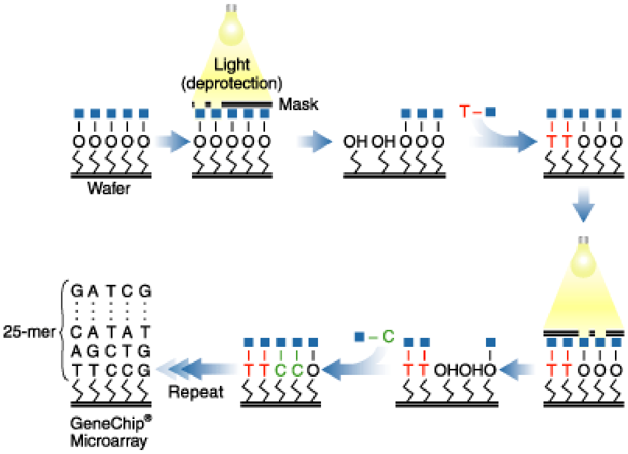


Figure 1. *The manufacturing of an oligonucleotide microarray through photolithography.* Nucleotides are immobilized to an array-slide (called a wafer). Photo-labile groups preventing polymerization are removed through light exposure at certain locations of the array. A mask directs the light where to shine. Nucleotides with new protection groups are let to bind, elongating the oligonucleotides. The procedure is repeated until single stranded 25-mers have been created. (Courtesy of Affymetrix, www.Affymetrix.com)

However, the photolithography technique has an important limitation due to the use of masks; besides the high manufacturing cost, light diffraction phenomena limit the lower boundary of miniaturization. Recently, Affymetrix introduced a microarray chip with 250.000 spots. With the use of two different restriction enzymes, which are used to fragment the genome before analysis, 500.000 SNPs can be genotyped by the same array. The next generation of GeneChips has been announced to contain 500.000 features on a

[2] The total significance, α, increases with n as $\alpha = 1 - (1 - \alpha_{comp})^{n}$ where α_{comp} is the significance level for each comparison and n the number of comparisons. Thus, the risk of false positives is higher with a large number of comparisons.

single slide. However, this goal has already proven hard to achieve due to the problem mentioned [15].

3.4.2 Genotyping and gene expression analysis with microarrays

Classically, sample and reference mRNA were labelled and let to competitively hybridize on expression arrays. Nowadays, commercial microarrays usually analyse samples and reference in separate runs. To compare arrays, normalization and reference features are necessary.

A microarray for genotyping, a so called mapping array, is an oligonucleotide array with 25-base probes. They have probes for perfect match (PM) and mismatch (MM) for all alleles that are screened for. A PM probe is 100% complementary to its target, whereas a MM probe has an intentional mismatch in the central, 13th, position. The latter probe helps to estimate the degree of cross-hybridization. To further improve the trustworthiness of the results, six offset positions are analysed (at 1, 2 and 4 bases from the SNP in both directions), resulting in 14 probe quartets tested for each SNP (see figure 2). This causes a demand for at least 14 times more features than the number of SNPs analysed. Likelihood ratios are calculated for all combinations of model and probe quartet, to finally let Wilcoxon's signed-rank test make the gene call. [16]

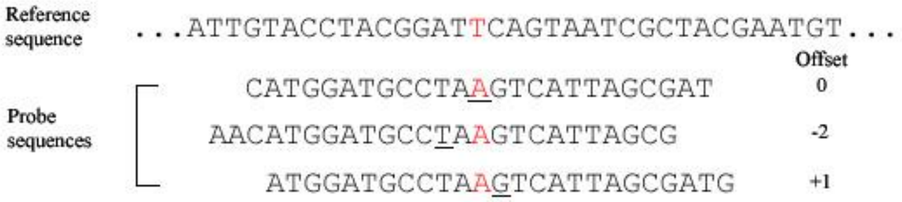

Reference sequence
...ATTGTACCTACGGATTCAGTAATCGCTACGAATGT...
Offset
Probe sequences
CATGGATGCCTAAGTCATTAGCGAT
0
AACATGGATGCCTAAGTCATTAGCG
-2
ATGGATGCCTAAGTCATTAGCGATG
+1

Figure 2. *Offset interrogation positions of PM probes.* The interrogation position is underlined and the SNP site is marked red. In a mismatch probe, the red base would in this case be a 'T'. (Modified from Matsuzaki *et al*., ref [16])

As an alternative, bacterial artificial chromosome (BAC) can be used for comparative gene hybridization in a microarray format to assess DNA copy number [17]. However, the use of SNP mapping microarrays might be a less cumbersome and more time efficient approach.

Microarrays for gene expression measurements are technically identical to mapping arrays. The only difference is the hybridization targets of the oligonucleotides; sequences on mapping arrays are complementary to DNA segments in the vicinity of SNPs whereas the on expression arrays are to mRNAs.

4. Materials and Methods

To identify genes putatively involved in cellular resistance to cancer drugs, a number of cancer cell lines were assayed with viability tests (FMCA) and microarrays to determine degrees of drug resistance, gene expression levels and CN; the latter ascribed through an in-house developed procedure based on SNP CN values. The identification was performed through correlation studies, univariate as well as quasi-bivariate. The results were finally evaluated visually by comparisons of resistant and their parental cell lines. This validated any genomic alterations. The whole process is summarized in figure 3.

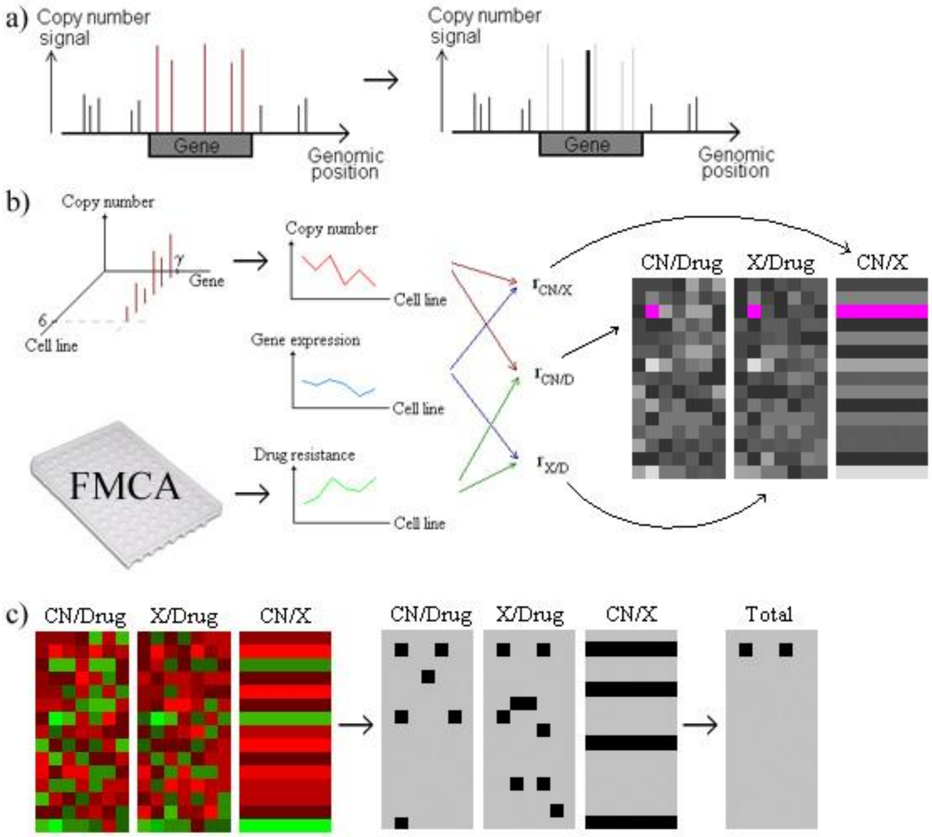


Figure 3. *Overview of project (univariate approach).* (**a**) A CN was ascribed each gene based on the SNP data from microarrays and knowledge about the transcriptional frames of genes. (**b**) Profiles for each gene were put together from CN values for all strains in the cell line panel (the figure shows the CN-profile for a gene γ from a cell panel of 6 strains). Gene expression profiles were created analogously and profiles of drug resistance were created from FMCA data. The pairwise correlations between the three characters were calculated and stored in matrices where each cell represents a correlation coefficient and each row a gene ('X' for gene expression). The correlation of CN and gene expression between different genes is not interesting why the coefficients could be summarized in a vector instead. In (**c**), the matrices are visualized as heatmaps where bright colours signifies strong positive (red) or negative (green) correlations. These matrices were transformed into indicator matrices through the use of cut-off values. A logic AND-operator applied on these matrices, cell by cell, resulted in a final indicator matrix where each row corresponded to a gene, and each column to a drug. Thus, genes putatively involved in cellular resistance of cytotoxic drugs were identified as rows with positive (black) signals, along with the compounds resistant to (in figure: one gene's involvement in resistance to two drugs is considered interesting). More extensive explanations can be found in following sections.

4.1 Strains

CN and gene expression data used in this project were collected from a cell panel summarized in table 1 below. All resistant strains were developed from each respective parental by continuously increased concentration of selective agent [7, 18].

Table 1. *Parental and resistant strains in cell panel used in analysis.*

Parental	Resistant	Origin	Selection agent
RPMI 8226/S	8226/Dox40	Myeloma	Doxorubicin
RPMI 8226/S	8226/LR5	Myeloma	Melphalan
CCRF-CEM	CEM/VM1	T-cell leukemia	Tenposide
NCI-H69	H69/AR	Small cell lung cancer	Doxorubicin
U-937-GTB	GTB/VCR10	Histolytic lymphoma	Vincristine

4.2 Software

Analysis of microarray data was performed in CNAT viewer (a module of the GTYPE software package from Affymetrix. For version see table 2 below) with its Integrated Genome Browser [19], and with CNAG [20, 21]. The latter uses HMM to identify putative CN alterations.

Table 2. *Software versions.*

Software	Version
GTYPE	4.0.0.22
CNAG	2.0

Sorting-procedures and linkage of different data types were performed with help of MatLab, whereas Microsoft Office Excel was used for analysis and visualization of CN and gene expression data.

4.3 Microarray data

CodeLink arrays from GE Healthcare were used for the analysis of gene expression. To assess the CN of the strains, oligonucleotide mapping arrays from Affymetrix were used and analysed with the CNAT Viewer software (also from Affymetrix). The actual SNP CN values were obtained through a comparison of data from the cell lines to the 50K mapping-reference file supplied by Affymetrix. The reference represents a “normal” human population, which in this case is made up by 100 healthy persons from diverse ethnic groups [22]. This approach is used to be able to compare the cell lines. In contrast, the use of a drug sensitive parental cell line as reference, where the genomic cause of resistance theoretically easily could be identified, hinders this comparison. Thus, the parental cell lines were used as a reference in validation of single amplifications in their sub-strains. Since all cell lines have cancer origin, vast differences in genomic structure could be expected when compared to the normal population reference.

A gaussian smoothing factor of 0.5 Mb was used for the calculation of SNP CN values to reduce noise from single markers [22].

4.4 Preparation of input data – Ascribing copy numbers to genes

A list of genes with GenBank Accession Numbers[3] was extracted from previous gene expression analysis of the cell panel. With help of these tags, the web based service Match Miner [23] could provide the location of transcription start/stop and the chromosome number for each gene. Through the combination of the genomic positions of the SNPs, their CN signal and the transcriptional interval of the genes, each gene could be ascribed an average CN. For genes lacking SNPs in their interval of transcription, the value of the nearest SNP was used (see figure 4). This is considered appropriate since 92% of the genome is within 0.1 Mb of an SNP marker with an average intermarker distance of 23.6 kb. [16]

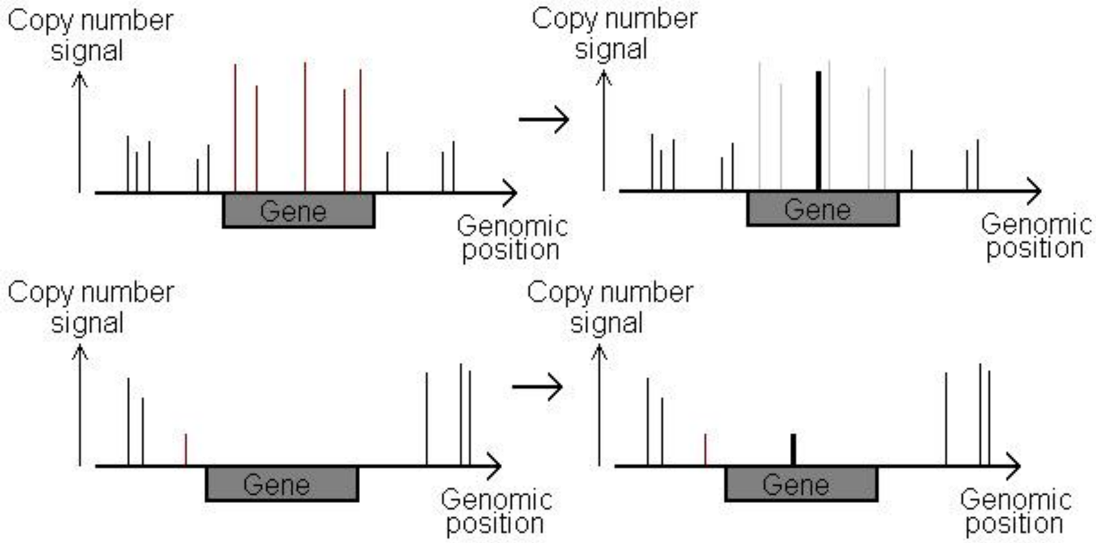


Figure 4. *A method to ascribe copy numbers to genes.* The figure exemplifies an amplified gene. CNs are in arbitrary signal intensity units. With knowledge of the transcriptional frames of genes, SNP CN data and their genomic position, a pseudo-CN can be ascribed to a gene (fat bar in figure). This CN is the arithmetic mean of SNP values (top) and if none is situated within the frame of the gene, the CN value of the nearest SNP is used (bottom). Length of vertical lines represent SNP CN signal. However, the signal intensity/CN ratio is not linear but rather a natural logarithmic function with an offset due to noise and a declining tendency for high intensities due to signal saturation [22].

4.5 Correlation analysis

To investigate potential correlation between gene expression levels, gene CN and drug effects, two measurements were used: First, the Pearson's coefficient [24], r:

(Eqn. 1)
$$r = \frac{\Sigma XY - \frac{\Sigma X \Sigma Y}{n}}{\sqrt{\left(\Sigma X^2 - \frac{(\Sigma X)^2}{n}\right)\left(\Sigma Y^2 - \frac{(\Sigma Y)^2}{n}\right)}} = \frac{Cov(X,Y)}{\sqrt{Var(X)\cdot Var(Y)}}, r \in [-1,1],$$

where X and Y are stochastical variables and n the number of data points.

[3] a unique identity tag for genes

As stated in (eqn. 1), r may take any value between -1 and +1 (see note in section 4.7) and assesses the linear nature of the correlation between the two variables; a value near 1 implies a very strong positive linear relationship. The use of Pearson's correlation measurement assumes continuous variables with a linear relationship. By a zero-mean statistical normalization of the vector X, the data obtains a unit variance and an average of zero. The normalization can be thought of as a vertical shift of a graph so that the function average is zero and can be performed by subtracting the mean (μ) from a vector (X), divided by the standard deviation (σ):

(Eqn. 2) $$\tilde{X} = \frac{(X - \mu)}{\sigma}$$

This simplifies the expression of the correlation coefficient to:

(Eqn. 3) $$r = Cov(\tilde{X}, \tilde{Y})$$

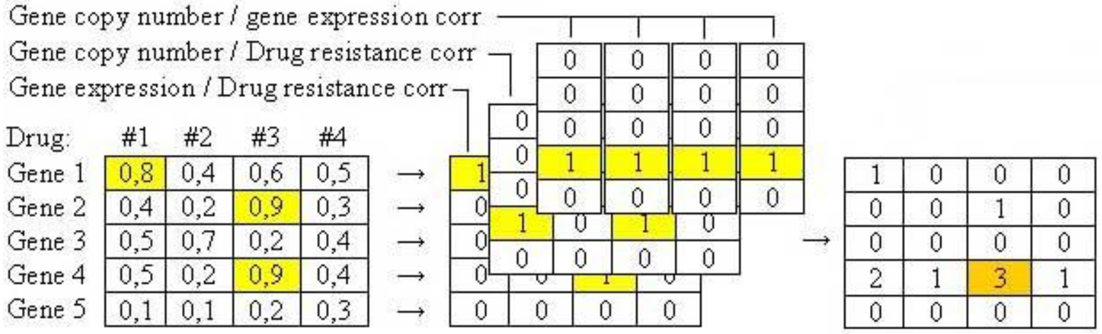
The assumption of linear relationships in Pearson's correlation measurement contributes to a risk of missing non-linear relations. Therefore, secondly, a Spearman's rank correlation test was performed, since it better can detect non-linear relationships. [25] The two tests are executed in a very similar way. The Spearman technique has an extra step which ranks both data sets from the highest to the lowest values before the calculation; the lowest value is given a rank of 1, the second lowest the rank 2 and so on. These rank-values are then used in the calculation of Spearman correlation as if measured values. Thus, this method makes it possible to use ordinal data. Often, the Spearman coefficient of correlation is denoted with the Greek letter ρ (rho) and can be calculated:

(Eqn. 4) $$\rho = 1 - 6 \cdot \sum_{n} \frac{d^2}{n(n^2 - 1)}, \rho \in [-1,1],$$

where d is the arithmetic difference in rank of corresponding variables and n is the number of data points.

Correlation coefficients were pair-wise calculated between gene expression levels, gene CN and drug effects. To identify the interesting genes a binary matrix was created for each correlation matrix through the use of a cut-off value (see figure 5). These indicator matrices were added and subsequently screened for genes with all three correlation coefficient values above the threshold, i.e. for rows with the value '3'. Since the correlation values between gene expression and CN are summarized in a vector[4], the vector was added to each column of the final indicator matrix.

[4] Secondary downstream effects of gene regulation are far too complex and poorly understood to be considered here, why only the correlation between the expression and the CN of each individual gene is interesting.


Gene copy number / gene expression corr
Gene copy number / Drug resistance corr
Gene expression / Drug resistance corr

Drug:	#1	#2	#3	#4
Gene 1	0,8	0,4	0,6	0,5
Gene 2	0,4	0,2	0,9	0,3
Gene 3	0,5	0,7	0,2	0,4
Gene 4	0,5	0,2	0,9	0,4
Gene 5	0,1	0,1	0,2	0,3

1	0	0	0
0	0	1	0
0	0	0	0
2	1	3	1
0	0	0	0

Figure 5. *The creation of an indicator matrix.* In this example, the cut-off for correlation was 0,75, resulting in one interesting gene to be studied further (Gene 4, resistant to drug #3).

Finally, the identified genes could be linked to one or several cell lines with help of the original gene expression and CN data, resulting in a list with the gene names and ID numbers, which cell lines they affect resistance in and to which drugs (partially displayed in Appendix II).

4.6 Quasi-bivariate correlation analysis

Scatter plots of genes expected to have higher correlation scores than observed in the univariate analysis, presented a, to some degree, expected problem; if resistance to one specific drug in two different cell lines depends on two different genes, there is a risk of observing low correlation scores (see figure 6, section 5.2). Ultimately, this would result in failure to detect either one of these interesting genes. To avoid this, a bivariate correlation study was preformed in the same manner as the previously described univariate. This time genes were paired and their expression and CN data summed before the analysis. However, all permutations were not analysed due to the vast number of pairs (more than $4.9 \cdot 10^9$ combinations). Instead, an initial filtering based on CN / gene expression correlation was performed (as described in section 4.5), with a subsequent bivariate analysis of all significant pairs.

Although two genes are analysed at the time, this is really a quasi-bivariate method. In a real bivariate analysis, a correlation-plane would be determined, not as in our case a 2D curve to a pseudogene. We deemed the description of a 3D plane with 9 values too inexact since the density of data-points decrease quickly when the number of dimensions is increased. Our method, i.e. the description of a linear slope with 9 coordinates in two dimensions, appears, however, to be robust.

4.7 A mathematical note to covariance matrices

In this project the correlation of three characters were analyzed: drug resistance, gene expression and CN. If these characters are denoted by the variables X, Y, Z, the correlation coefficients can be expressed through a correlation matrix:

(Eqn. 5) $$\overset{=}{R} = \begin{bmatrix} 1 & r_{XY} & r_{XZ} \\ r_{YX} & 1 & r_{YZ} \\ r_{ZX} & r_{ZY} & 1 \end{bmatrix}$$

A covariance matrix is always positively definite and r_{XY} is equal to r_{YX}. From this follows:

(Eqn. 6) $$\left|\overset{=}{R}\right| \geq 0 \Leftrightarrow \left\| \begin{bmatrix} 1 & r_{XY} & r_{XZ} \\ r_{YX} & 1 & r_{YZ} \\ r_{ZX} & r_{ZY} & 1 \end{bmatrix} \right\| = 1 + 2 \cdot r_{XY} r_{YZ} r_{ZX} - r_{XY}^2 - r_{XZ}^2 - r_{YZ}^2 \geq 0$$

It is easily realized that if X and Y correlate perfectly[5] when Y and Z do, all pairwise correlations are perfect. However, if, for instance, r_{XY} and r_{YZ} has a correlation of 0.9, r_{XZ} can vary between 0.62 and 1. Thus, two stringent thresholds guarantee the final correlation value to be above a certain level. To enable r_{XZ} to vary in its full range, i.e. from 0 to 1, the other pairwise correlations must not be higher than $2^{-1/2}$.

4.8 Visualization of chromosomes to verify results

To evaluate our results visually, graphs with gene CN signals as a function of genomic position were created for all chromosomes of all cell lines. Each resistant cell line was this time analyzed with its respective parental cell line as a reference since the genomic differences, in theory, are responsible for the drug resistance. As seen in figure 7, section 5.3, the cell line pairs (parental and resistant sub-strain) were evaluated in non-random groups. The motive was to support the results through a comparison of cell line pairs with the same parental strain (to minimize the risk of misinterpretations due to strain variations) and pairs exposed to the same selective agent (to identify any common characters of genomic alterations).

With the bare eye, putatively altered regions were noted and compared to previous results. As a complement, the software CNAG [21] was used to identify abnormal genomic sections by a hidden Markov model (HMM)-approach.

[5] has a correlation coefficient of 1 or -1

5. Results

5.1 Univariate analysis – correlation between copy number values, gene expression and drug effects

Pearson’s and Spearman’s correlation coefficients were pair-wise calculated for gene expression, CN and drug resistance data. Coefficients with values above certain thresholds (0.7; 0.8; 0.9 and 0.95) were noted, along with the number of corresponding unique genes (see table 3 and 4 for Pearson and Spearman analysis, respectively).

Table 3. *The number of Pearson’s correlation coefficients (r) whose absolute value was above given thresholds and the number of corresponding genes.* ‘X’ for gene expression, ‘CN’ for copy number and ‘Drug’ for drug resistance. The number of sets with all three correlation coefficients above given threshold can be found in the ‘Triple’ column, and corresponding number of genes are found in the rightmost column. 11246 genes and 39 drugs were analyzed.

abs(r)	X vs. Drug	CN vs. Drug	CN vs. X	Triple	Genes
> 0,7	28074	33742	1328	2017	608
> 0,8	12931	13085	555	498	165
> 0,9	4768	2544	122	100	24
> 0,95	2032	505	25	46	11
Out of:	438594	438594	11246	438594	11246

Table 4. *The number of Spearman’s correlation coefficients (ρ) whose absolute value was above given thresholds and the number of corresponding genes.* ‘X’ for gene expression, ‘CN’ for copy number and ‘Drug’ for drug resistance. The number of sets with all three correlation coefficients above given threshold can be found in the ‘Triple’ column, and corresponding genes are found in the rightmost column. 11246 genes and 39 drugs were analyzed.

abs(ρ)	X vs. Drug	CN vs. Drug	CN vs. X	Triple	Genes
> 0,7	31200	32394	1194	1661	774
> 0,8	10584	10698	471	197	129
> 0,9	1618	1497	83	6	7
> 0,95	400	350	18	1	1
Out of:	438594	438594	11246	438594	11246

Table 5. *The number of genes found with both Pearson and Spearman-correlation studies.*

Cut-Off	Pearson	Spearman	Shared
> 0,7	608	774	275
> 0,8	165	129	26
> 0,9	24	7	0
> 0,95	11	1	0

5.2 Bivariate analysis – correlation between copy number values, gene expression and drug effects

To evaluate and enhance our method, a bivariate version of the analysis was performed for genes with a high CN / gene expression correlation. Only the Pearson’s coefficient of correlation was used due to the poor outcome of the Spearman approach in the univariate analysis. Figure 6 below exemplifies how two genes that would be overlooked with stringent cut-off values when studied individually can be detected by our method when looked upon together.

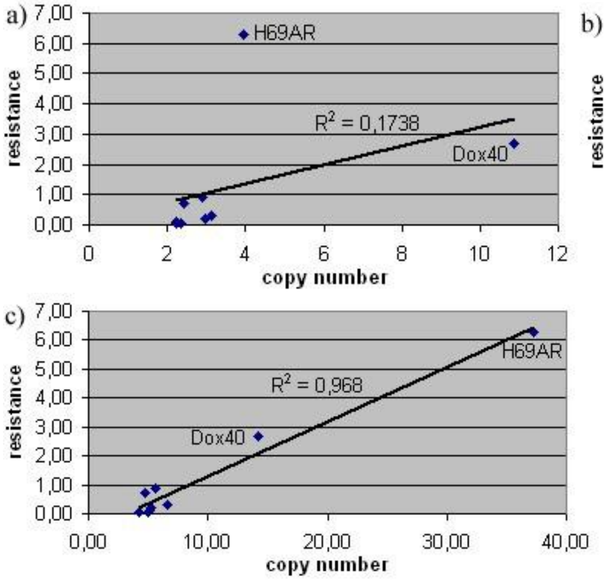


Figure 6. *Scatter plots of correlation coefficients (R^2) between resistance to the drug doxorubicin and gene copy number for the MDR (figure a) and MRP genes (figure b), and their sum (figure c).* From **a** and **b**, one can conclude that MDR (ATP-binding cassette, sub-family B member 4) is amplified in the 8226/Dox40 cell line as MRP (ATP-binding cassette, sub-family C, member 1) is in H69AR. This suggests them to be involved in drug resistance mechanisms. Note the significant increase in correlation when pairwise comparisons are made.

Results from the different correlation measurements used in this project are summarized in table 6 for the most stringent thresholds.

Table 6. *The number of genes whose correlation coefficients were above given thresholds.* 'Pearson' for standard and 'Pairwise' for bivariate analysis. 11246 genes and 39 drugs were analyzed.

abs(r)	**Pearson**	**Pairwise**	**Shared**
> 0,9	24	121	22
> 0,95	11	24	11

5.3 Multiple copy number variations – multiple mechanisms?

As can be expected from microarray data, the CN data was noisy and frequently drifting from the presupposed baseline of 2. However, some distinct regions with genomic variations could be observed when SNP signals from comparisons of resistant and parental cell lines were visualized (see figure 7).

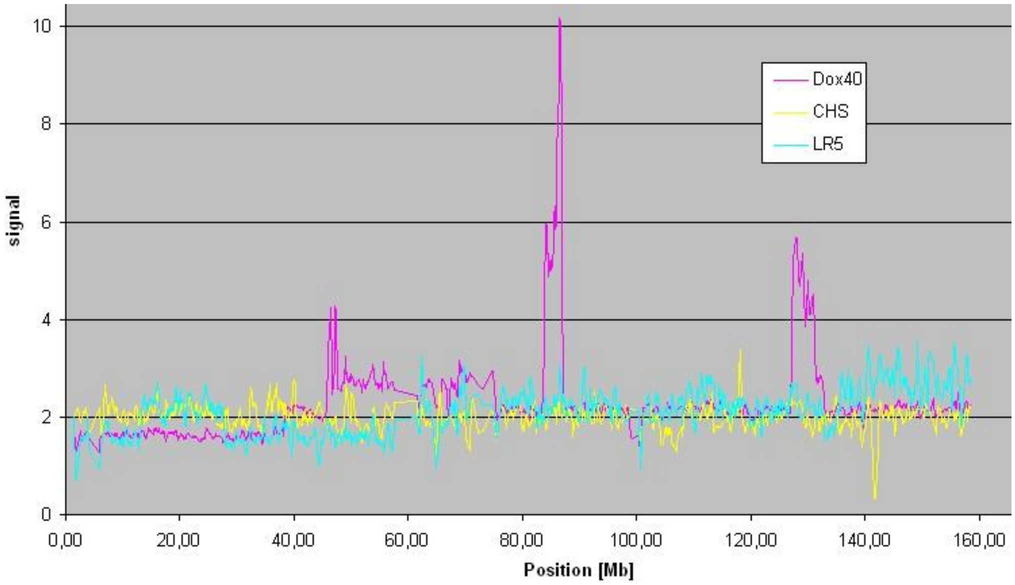


Figure 7. *Example of copy number analysis in Microsoft Excel.* Copy numbers for chromosome 7 in three cell lines resistant to doxorubicin (Dox40), CHS and melphalan (LR5), respectively, are plotted against genomic position. The cell lines are compared to their parental cell line (i.e. RPMI-8226), from which they are originally derived. The theoretical CN signal for an unaltered region is 2. Three interesting regions for the 8226/Dox40 strain can be discerned: at 47, 85 and 130 Mb. If these regions indeed are the cause of resistance, the other cell lines shown in the diagram must have other mechanisms of resistance.

Clustering of putatively interesting genes to amplified genomic regions were observed. This was expected since the approach used demanded a variation in CN to correlate well with, for instance, a for the cell panel altered drug resistance; gene expressions can vary despite constant CN. Interestingly, more than one altered region often contained genes detected by the correlation studies in the same cell line pair. This could be a sign of multiple mechanisms of resistance.

The approach to visualize the comparison of pairs of strains could be performed on all genes, one by one. This would however be very cumbersome. As a visualization tool to evaluate the high throughput method presented in this article, it is, however, valuable.

6. Discussion

Data mining has become somewhat of a prerequisite for many fields of 21st century science. New techniques constantly demand increasingly efficient methods of data analysis. This project has handled vast amounts of data in a fairly straightforward and basic fashion, but has still been able to deliver interesting results. This exemplifies that interesting results can be obtained by fairly basic data mining techniques.

6.1 Reliability of data and results

Optimally, technical as well as biological replicates would be run for all experiments to strengthen the reliability of the data, and thereby the results. However, the analysis is based on trends of profiles and should not be greatly disturbed by cell lines with abnormal values. Furthermore, the data has been collected under controlled conditions and has shown small variation when replicated. Nonetheless, since a handful of different mechanisms of drug resistance are represented in the data used, interesting genes risk to be masked by signals from others in the univariate analysis.

The gene expression data suffered from some missing entries. Instead of excluding these genes from the analysis, missing values could be approximated with a Nearest Neighbour (NN) approach for each gene profile: The most similar expression profiles could be found by least square measurements of zero-mean normalized data. Each missing entry is then substituted by the conditional mean of these profiles. [26] There is a small, but obvious, risk to leave out possibly interesting genes from the analysis. The small fraction of genes excluded (about 1.8 %) strengthens our belief that nothing has been overlooked.
In this project, all genes with multiple genomic locations have also been excluded.

6.2 The different approaches and their results

Although a fairly basic approach, the use of Microsoft Excel spreadsheets to visualize CN data and manually identify interesting genes was successful. With support from gene expression data, lists with interesting genes for a number of chromosomes for each cell line were matched with the results from the correlation studies. For instance, in an obvious CN alteration at chromosome 7 of RPMI-8226/Dox40 (peak at 85 Mb in figure 7, section 5.3), 19 genes were observed. With the most stringent cut-off level (0.95) of the Pearson approach, 4 of these were picked out again. 2 of these are ATP-binding cassettes of the B-subfamily (abbreviated ABCB), known to confer resistance when overexpressed [27].However, the manual/visual approach is cumbersome and subjective, and thereby highlights the needs of a high-throughput method like ours. The manual approach is best used as a complement to the correlation studies.

A model of perfect linear correlation between the expression and the CN of a gene assumes a far too simplistic view of gene regulation. Downstream effects such as those of genes, whose transcription factors are overexpressed, are not at all considered. Nevertheless, the primary link between drug resistance and gene expression can be expected to be strong, since a mere alteration of CN cannot be expected to convey resistance in any higher degree. This speaks for the use of different cut-off thresholds in the correlation studies. Another issue is whether all three correlation pairs should be ascribed equal importance or not. Weight-coefficients could balance the importance of the characters with respect to one another, but their values are hard to choose. Therefore weight coefficients have not been used in this study.

The considerable difference in the results between the Pearson and Spearman correlation measurements is somewhat surprising. One explanation is the power of large values in linear relations: In the search of linear correlations with help of the Pearson approach, single extreme values in one or a couple of cell lines breaks through to the results. In the Spearman analysis, the numerical value is replaced with a rank, thereby removing the dimension of size. The ranking enables the technique to detect non-linear correlations. Thus, the approach as a whole used in this project is somewhat based on its innate inability to detect small variations in noisy data, i.e. it uses outliers for analytical purposes.

The pseudo-bivariate analysis can be seen as a simplified unit-weight regression analysis, i.e. with equal weight for both genes. A more advanced analysis is in that sense somewhat uncalled for, since we hypothesize e.g. found protein pumps to, to some degree, be interchangeable.

Univariate trials where strains were systematically left out of the correlation analysis (data not shown) had similar results as the quasi-bivariate approach. This suggests that outliers are overrepresented in mainly two of the cell lines of our panel. The positive effect of quasi-multivariate analysis seen in this project raises the question of how many variables that are fruitful, and feasible, to study simultaneously. This seems to depend on the samples at hand.

Pearson correlation studies of gene expression, copy number and drug resistance has been performed in the past (e.g. ref [17]), but have never been integrated in a high throughput manner. Our analysis is, as explained, based on the behaviour of linear correlations in the presence of data outliers. This approach is shared by several other research groups (i.e. the COPA -Cancer Outlier Profile Analysis tool introduced in 2005, ref [28]), indicating the usefulness of outlier analysis. Analysis with profiles of, e.g., gene expression and drug resistance, is popular and has proven fruitful during the last couple of years [29].

6.3 Future prospects

In this project, many interesting genes possibly drowned in the noise of others since a couple of different resistance mechanisms are analysed at the same time. This might to some extent have been avoided through the bivariate analysis. A higher form of multivariate analysis or the use of a more uniform cell line panel, i.e. with a single type of resistance, might improve the discovery rate further. However, a natural first step of further investigations would be to study the possible use of different cut-off threshold values for the correlation coefficients and the weight between themselves.

The multivariate approach might benefit from the analysis of non-linear relations rather than linear ones since this is, most likely, closer to the biological truth. The importance of the data outliers will be compensated by the support of the increased number of dimensions.

In a longer perspective, verification of the results in the lab, by the means of e.g. siRNA experiments, is needed. Despite all positive sides of data mining, the final step of the process (i.e. verification of results) must still take place in the laboratory.

7. Acknowledgements

First of all, I would like to acknowledge my supervisor Anders Isaksson for giving me the opportunity to do my Master Thesis in the Computational Medicine group (department of Medical Sciences, Uppsala University), but also for valuable discussions and help.
I wish to convey special thanks to Hanna Göransson and Maria Rydåker (Uppsala Array Platform) for their continuous help, and would also like to thank Claes Andersson, Mats Gustafsson, Mårten Fryknäs and Linda Rickarsson for valuable contributions to the project.

Finally, I would like to acknowledge Tomas Olofsson (department of Engineering Sciences) for important remarks and helpful ideas.

8. References

1. Longley, D.B., and Johnston, P. G., *Molecular mechanisms of drug resistance.* J Path, 2005. **205**: p. 275-292.
2. Kugelberg, E., Kofoid, E., Reams, A. B., Anderson, D. I., and Roth, J. R., *Multiple pathways of selected gene amplification during adaptive mutation.* PNAS, 2006. **103**: p. 17319-17324.
3. Gottesman, M.M., *Mechanisms of cancer drug resistance.* Annu Rev Med, 2002. **53**: p. 615-627.
4. Hazlehurst, L., Foley, N., Gleason-Guzman, M., Hacker, M., Cress, A., Greengerger, L., De Jong, M., and Dalton, W., *Multiple mechanisms confer drug resistance to mitoxantrone in the human 8226 myeloma cell line.* Cancer research, 1999. **59**: p. 1021-1028.
5. Nygren, P., *What is cancer chemotherapy?* Acta Oncologica, 2001. **40**: p. 166-174.
6. Larsson, R., Fridborg H., Kristensen J., Sundström C., and Nygren P., *In vitro testing of chemotherapeutic drug combinations in acute myelocytic leukemia using the fluorometric microculture cytotoxicity assay (FMCA).* Br. J. cancer, 1993. **67**: p. 969-974.
7. Dalton, W., Durie, B., Alberts, D., Gerlach, J., and Cress, A., *Characterization of a new drug-resistant human myeloma cell line that expresses P-glycoprotein.* Cancer research, 1986. **46**: p. 5125-5130.
8. Fryknas, M., Dhar, S., Oberg, F., Rickardson, L., Rydaker, M., Goransson, H., Gustafsson, M., Pettersson, U., Nygren, P., Larsson, R., and Isaksson, A., *STAT1 signaling is associated with acquired crossresistance to doxorubicin and radiation in myeloma cell lines.* Int J Cancer, 2007. **120**: p. 189-195.
9. Larsson, R., Nygren, P., Ekberg, M., and Slater, L., *Chemotherapeutic drug sensitivity testing of human leukemia cells in vitro using a semiautomated fluorometric assay.* Leukemia, 1990. **4**: p. 567-571.
10. Schork, N.J., Fallin, D., and Lanchbury, J. S., *Single nucleotide polymorphisms and the future of genetic epidemiology.* Clin Genet, 2000. **58**: p. 250-264.
11. Schena, M., Shalon, D., Davies, R. W., and Brown, P. O., *Quantitative monitoring of gene expression patterns with a complementary DNA microarray.* Science, 1995. **270**: p. 467-470.
12. Affymetrix, *Data sheet GeneChip Human Genome Arrays.* 2004.
13. Lönnstedt, I., and Speed, T., *Replicated microarray data.* Statistica Sinica, 2002. **12**: p. 31-46.
14. Rabbee, N., and Speed, T., *A genotype calling algorithm for Affymetrix SNP arrays.* Bioinformatics, 2006. **22**: p. 7-12.
15. Gräns, H., *Field application specialist, Affymetrix Scandinavia.* 2006.
16. Matsuzaki, H., Dong, S, Loi, H., Di, X., Liu, G., Hubbell, E., Law, J., Berntsen, T., Chadha, M., Hui, H., Yang, G., Kennedy, G. C., Webster, T. A., Cawley, S., Walsh, P. S., Jones, K. W., Fodor, S. P. A., and Mei, R., *Genotyping over 100.000 SNPs on a pair of oligonucleotide arrays.* Nature methods, 2004. **1**: p. 109-111.
17. Bussey, K.J., Chin, K., Lababidi, S., Reimers, M., Reinhold, W. C., Kuo, W-L., Gwadry, F., Ajay, Kouros-Mehr, H., Fridlyand, J., Jain, A., Collins, C.,

Nishizuka, S., Tonon, G., Roschke, A., Gehlhaus, K., Kirsch, I., Scudiero, D. A., Gray, J. W., and Weinstein, J. N., *Integrating data on DNA copy number with gene expression levels and drug sensitivities in the NCI-60 cell line panel.* Mol Cancer Ther, 2006. **5**: p. 853-867.

18. Dhar, S., Nygren, P., Csoka, K., Botling, J., Nilsson, K., and Larsson, R., *Anti-cancer drug characterisation using a human cell line panel representing defined types of drug resistance.* Br J Cancer, 1996. **74**: p. 888-896.
19. Affymetrix, *Integrated Genome Browser*. 2006.
20. Nannya, Y., Sanada, M., Nakazaki, K., Hosoya, N., Wand, L., Hangaishi, A., Kurokawa, M., Chiba, S., Bailey, D. K., Kennedy, G. C., and Ogawa, S., *A robust algorithm for copy number detection using high-density oligonucleotide single nuvleotide polymorphism genotyping arrays.* Cancer research, 2005. **65**: p. 6071-6079.
21. Genome Laboratory., D.o.R.M.f.H., Tokyo University, *CNAG*. 2005.
22. Affymetrix, *Affymetrix GeneChip® Chromosome Copy Number Analysis Tool User Guide, Version 3.0.* 2006.
23. Bussey, K.J., Kane, D., Sunshine, M., Narasimhan, S., Nishizuka, S., Reinhold, W. C., Zeeberg, B., Ajay and Weinstein, J. N., *MatchMiner: a tool for batch navigation among gene and gene product identifiers.* Genome biology, 2003. **4**: p. R27.
24. Weisstein, E.W., *Correlation Coefficient.*, Wolfram MathWorld.
25. Weisstein, E.W., *Spearman Rank Correlation Coefficient.*, Wolfram MathWorld.
26. Wasito, I., and Mirkin, B., *Nearest neighbours in least-squares data imputation algorithms with different missing patterns.* Comp stat data anal, 2004. **50**: p. 926-949.
27. Michieli, M., Damiani, D., Geromin, A., Michelutti, A., Fanin, R., Raspadori, D., Russo, D., Visani, G., Dinota, A., Pileri, S., *et al.*, *Overexpression of multidrug resistance-associated p170-glycoprotein in acute non-lymphocytic leukemia.* Eur J Haematol, 1992. **48**: p. 87-92.
28. MacDonald, J.W., and Ghosh, D., *COPA-cancer outlier profile analysis.* Bioinformatics, 2005. **22**: p. 2950-2951.
29. Lamb, J., Crawford, E. D., Peck, D., Modell, J. W., Blat, I. C., Wrobel, M. J., Lerner, J., Brunet, J. P., Subramanian, A., Ross, K. N., Reich, M., Hieronymus, H., Wei, G., Armstrong, S. A., Haggarty, S. J., Clemons, P. A., Wei, R., Carr, S. A., Lander, E. S., and Golub, T. R., *The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease.* Science, 2006. **313**: p. 1929-1935.

Appendix

Appendix I – Cancer drugs used in analysis

4-HC
5-Aza-2-cytidine
5-Azacytidine
6-Mercaptopurine
6-Thioguanine
Acivicin
Aclarubicin
Amsacrine
Anguidine
Bisantrene
Bortezomib
Camptothecin
Chlorambucil
Cisplatin
Cyclohexamide
Daunorubicin
Doxorubicin
Epirubicin
Etoposide
Hoechst 33342
Idarubicin
J1
Lactacystin
Mechlorethamine
Melphalan
MG132
MG262
MIBG
Mitomycin C
Mitoxantrone
Paclitaxel
P2
Sarcolysin
SN-38
Spirogermanium
Teniposide
Topotecan
Vinblastine
Vinorelbine

Appendix II – Genes from univariate Pearson analysis (correlation threshold 0,95)

Full gene name	Chrom	Approx pos. [Mb]	Pair of strains
ATP-binding cassette, sub-family B (MDR/TAP), member 4	7	86	8226/Dox
ATP-binding cassette, sub-family B (MDR/TAP), member 1	7	86	8226/Dox
Cyclin D binding myb-like transcription factor 1	7	86	8226/Dox
Chromosome 7 open reading frame 23	7	86	8226/Dox
Calumenin	7	128	8226/Dox
ATP-binding cassette, sub-family C (CFTR/MRP), member 1	16	16	NCI-H69/AR
Dystrobrevin, alpha	18	30	NCI-H69/AR
Solute carrier family 39 (zinc transporter), member 6	18	31	NCI-H69/AR
Microtubule-associated protein, RP/EB family, member 2	18	31	NCI-H69/AR
Hypothetical protein FLJ10656	18	31	NCI-H69/AR
Signal transducer and activator of transcription 3 interacting protein 1	18	31	NCI-H69/AR
Chromosome 18 open reading frame 21	18	31	NCI-H69/AR