BMC Medical Genomics

BioMed Central

Research article

Open Access

Classification of human cancers based on DNA copy number amplification modeling

Samuel Myllykangas*[1], Jarkko Tikka[2], Tom Böhling[1], Sakari Knuutila[1] and Jaakko Hollmén*[2]

Address: [1]Department of Pathology, Haartman Institute and HUSLAB, University of Helsinki and Helsinki University Central Hospital, P.O. Box 21, FI-00014, University of Helsinki, Helsinki, Finland and [2]Department of Information and Computer Science, Helsinki University of Technology, P.O. Box 5400, FI-02015 TKK, Espoo, Finland

Email: Samuel Myllykangas* - samuel.myllykangas@helsinki.fi; Jarkko Tikka - tikka@cis.hut.fi; Tom Böhling - tom.bohling@helsinki.fi; Sakari Knuutila - sakari.knuutila@helsinki.fi; Jaakko Hollmén* - jaakko.hollmen@hut.fi

* Corresponding authors

Published: 14 May 2008

BMC Medical Genomics 2008, **1**:15 doi:10.1186/1755-8794-1-15

Received: 12 December 2007
Accepted: 14 May 2008

This article is available from: http://www.biomedcentral.com/1755-8794/1/15

Abstract

Background: DNA amplifications alter gene dosage in cancer genomes by multiplying the gene copy number. Amplifications are quintessential in a considerable number of advanced cancers of various anatomical locations. The aims of this study were to classify human cancers based on their amplification patterns, explore the biological and clinical fundamentals behind their amplification-pattern based classification, and understand the characteristics in human genomic architecture that associate with amplification mechanisms.

Methods: We applied a machine learning approach to model DNA copy number amplifications using a data set of binary amplification records at chromosome sub-band resolution from 4400 cases that represent 82 cancer types. Amplification data was fused with background data: clinical, histological and biological classifications, and cytogenetic annotations. Statistical hypothesis testing was used to mine associations between the data sets.

Results: Probabilistic clustering of each chromosome identified 111 amplification models and divided the cancer cases into clusters. The distribution of classification terms in the amplification-model based clustering of cancer cases revealed cancer classes that were associated with specific DNA copy number amplification models. Amplification patterns – finite or bounded descriptions of the ranges of the amplifications in the chromosome – were extracted from the clustered data and expressed according to the original cytogenetic nomenclature. This was achieved by maximal frequent itemset mining using the cluster-specific data sets. The boundaries of amplification patterns were shown to be enriched with fragile sites, telomeres, centromeres, and light chromosome bands.

Conclusions: Our results demonstrate that amplifications are non-random chromosomal changes and specifically selected in tumor tissue microenvironment. Furthermore, statistical evidence showed that specific chromosomal features co-localize with amplification breakpoints and link them in the amplification process.

Background

Alterations that increase DNA copy number are frequently observed in a variety of human cancers [1,2]. An amplification is a mutation that increases the copy number of a specific DNA segment in a cancer cell [3,4]. A normal diploid genome contains two DNA copies, while amplification increases the DNA copy number [5]. High-level gene amplifications may significantly elevate the gene copy number, e.g., the amplifications of *MYC* and *EGFR* oncogenes have been shown to be more than hundred-fold in neuroblastoma [6] and gliomas [7]. Gene amplifications have clinical relevance as targets for therapy and in predictive diagnosis. For example, amplification of the *ERBB2* gene is an indicator for trastuzumab (Herceptin®) treatment of patients with metastatic breast cancer. In general, cancers with DNA copy number amplifications have worse prognosis and poorer survival than cancers that do not manifest amplifications. The amplifications have been shown to associate with adverse clinical outcomes, i.e., high grade and advanced stage, metastasis, and poor response to therapy [8].

The guidelines for classification of tumors have been established by the World Health Organization (WHO) [9]. The WHO classification is based on the evaluation of the primary organ site, morphology, cell type, histology, and malignancy state. In addition, the epidemiological, etiological, clinical, and genetic features of tumors have been evaluated. In a subset of hematologic malignancies, specific mutations and translocations have been used in classification but otherwise tumor classification is based on clinical, histological, and pathological parameters. Although a wide range of cancer subtype-specific genetic abnormalities are known, they are rarely used to classify cancers. Given that genomic changes underlie the cancer phenotype, DNA copy number amplification is a justified foundation for classification. Nonetheless, molecular properties underlie phenotypic changes in cancer cells and contribute to the clinical outcome. Thus, molecular classification of cancers is well-founded. DNA copy number amplifications are suitable classification targets, because they are relatively prevalent in a variety of cancers. Since 1992, it has been possible to screen DNA copy number amplifications in genome-wide coverage using comparative genomic hybridization (CGH) [10] and large amounts of DNA copy number data from different cancers have been published and collected [1,11].

In a previous amplification profiling study, we identified four separate clusters and showed that the clusters based on DNA copy number amplifications comprised anatomically similar neoplasms [1]. For example, gastrointestinal adenocarcinomas (gastric cancer, colorectal cancer, and Barrett's adenocarcinoma) clustered together. Similar clustering emerged when amplification-activated oncogenes were analyzed using hierarchical clustering [8]: cancers with similar embryonic background (hematopoietic, mesenchymal or epithelial) formed separate clusters. Even though cancer types inside the identified clusters showed similar biological backgrounds, using the profiling approach, specific amplifications could not be appointed for specific cancer classes and analytical assessment was not plausible. Here, we used probabilistic modeling and a collection of DNA copy number amplifications, and identified 111 specific amplification models. Modeling was performed based on mixtures of multivariate Bernoulli distributions (see sub-section of methods entitled "Probabilistic modeling of DNA copy number amplification" for details). Based on the amplification modeling, the cancer cases were divided into clusters. Specific cancer cases, either of the same type or etiology, were shown to associate with specific amplification model-based clusters. Compact and comprehensible presentations for probabilistic amplification models, amplification patterns, were extracted using maximal frequent itemset mining (see sub-section of methods entitled "Finite descriptions for continuous DNA copy number amplification models" for details). Amplification patterns represent the ranges and structures of the amplicons. We present statistical evidence showing that fragile sites, telomeres, centromeres, and light chromosome bands are enriched at the amplicon boundaries, linking them to the mechanisms of amplification.

Methods

DNA copy number amplification and cancer classification data

DNA copy number amplification data were retrieved from [12]. The compilation of DNA copy number amplification data contains curated data from more than 800 published CGH studies on 4590 cases [1]. The data set includes DNA copy number amplification data at chromosome sub-band resolution (393 bands). The original classification of 73 human neoplasms was redefined to contain 95 specific neoplasm types by sub-classifying B-cell neoplasms and neuroepithelial tumors. The studied neoplasms were grouped according to the guidelines provided in the WHO Classification of Tumors [9]. In addition to the WHO classification, the neoplasms were arranged according to cell-lineage, which included determination of system, organ, cell type, and embryonic lineages. Moreover, the neoplasm types were categorized using clinical and genetic attributes. Gender and age group specificity was determined based on the WHO classification. Various etiological factors were collected from the WHO classification: tobacco, alcohol, hormonal imbalance, ultraviolet radiation, obesity, diet, human immunodeficiency virus, AIDS, human papilloma virus, Epstein-Barr virus, polyoma virus, bacterial and parasite infections, radiation, and prosthetic implant, as well as asbestos exposure and toxin exposures. Inflammatory etiologies of neoplasms and underlying conditions were defined according to literature. Tumor behavior was defined according to the stage (cancer, benign, and border line). After the neoplasms were filtered to include only malignant cancers and to discard benign, borderline, and non-malignant tumors, the number of cancer types was 82.

Probabilistic modeling of DNA copy number amplification

We applied machine learning techniques to model DNA copy number amplifications. The goal of probabilistic modeling is to estimate an unknown probability distribution based on observations to describe the inherent structure in the data. Finite mixture models are powerful and widely used in the estimation of complex probability distributions [13,14]. The advantage of a mixture model is that its components can represent different parts of the true distribution, which would be impossible to estimate by a single parametric distribution. In this work, we concentrated on the mixtures of multivariate Bernoulli distributions, since the representation of our DNA copy number amplification data was binary.

Analysis of DNA copy number amplification data was carried out separately for each human chromosome (except chromosome Y). Due to the low number of observations and insufficient resolution, the Y chromosome was excluded from the analysis. The mixture models were constructed separately for each chromosome, since only 10 percent of the observations showed amplification in more than one chromosome. The observations with amplification in several chromosomes were included in the analysis of corresponding chromosomes. In brief, DNA copy number amplification data of cancers including $N = 4402$ cases were used in the modeling. DNA copy number amplification data can be presented as binary vectors $x \in \{0,1\}^d$ in which $x_i = 1$ denotes an amplified chromosome band and $x_i = 0$ stands for a non-amplified band and d is the number of bands. The probabilities of the outcomes of observation $\boldsymbol{x} = (x_1,..., x_d)$ were modelled as $\theta_i = P(x_i = 1), i = 1,..., d$. Probability of the observed vector $\boldsymbol{x}$ was estimated using the finite mixture of multivariate Bernoulli distributions

$$p(\mathbf{x} \mid \boldsymbol{\Theta}) = \sum_{j=1}^{J} \pi_j p(\mathbf{x} \mid \boldsymbol{\theta}_j) = \sum_{j=1}^{J} \pi_j \prod_{i=1}^{d} \theta_{ji}^{x_i} (1 - \theta_{ji})^{1-x_i},$$

where $\boldsymbol{\Theta} = \left\{ J, \left\{ \pi_j, \boldsymbol{\theta}_j \right\}_{j=1}^{J} \right\}$ denotes the parameters of the model. The multivariate Bernoulli distributions $p(\boldsymbol{x}|\theta_j), j = 1,..., J$, also called the component distributions, are parameterized by $\theta_j = (\theta_{j1},..., \theta_{jd})$ and π_j are mixture proportions with the properties $\pi_j \geq 0$ and $\sum_{j=1}^{J} \pi_j = 1$. Although the finite mixture of multivariate Bernoulli distributions have been shown to be non-identifiable [15], they are useful in practical estimation problems [16].

In the case of N observations $\boldsymbol{x}^n$, $n= 1,..., N$ and J mixture components, the maximum likelihood estimates of the parameters $\left\{ \hat{\pi}_j, \hat{\boldsymbol{\theta}}_j \right\}_{j=1}^{J}$ are obtained by maximizing the log-likelihood of the observations $l = \sum_{n=1}^{N} \log \left[\sum_{j=1}^{J} \pi_j \prod_{i=1}^{d} \theta_{ji}^{x_{ni}} (1 - \theta_{ji})^{1-x_{ni}} \right]$. The optimization was carried out using the Expectation-Maximization (EM) algorithm [17-19]. The derivation of the EM algorithm for the finite mixture of multivariate Bernoulli distributions has been explained in detail by Everitt and Hand [14]. In the E-step, the posterior probabilities that the jth component distribution has generated the data point $\boldsymbol{x}_n$ are evaluated. In the M-step, the values of parameters $\left\{ \hat{\pi}_j, \hat{\boldsymbol{\theta}}_j \right\}_{j=1}^{J}$ are updated using the evaluated posterior probabilities. The iteration between E- and M-steps gives monotonically increasing series of the values for the log-likelihood. The EM algorithm was terminated when the relative change in log-likelihood was smaller than10^{-4}. According to Carreira-Perpiran and Renals, and Tikka et al. [16,20], the initial values of parameters θ_{ji}, $j = 1,..., J$, $i = 1,..., d$ were selected randomly from range 0.25–0.75 and the initial values of mixture proportions were $\pi_j = 1/J$.

In order to select a model with an appropriate complexity, the number of component distributions J was selected using 5-fold cross validation [21] that was repeated 10 times varying J from 2 to 30. The selected number of components maximized the validation log-likelihood, except in five cases, when less complex models were chosen to achieve validation log-likelihood that was in practice as good as the estimated optimum. Less complex models were chosen for chromosomes 3, 7, 8, 17, and 21. The final mixture model, with the number of component distributions based on cross validation, was trained 5 times and the model maximizing the log-likelihood was selected. The repetitions were done to avoid the local maxima in the log-likelihood. The obtained component distributions were regarded as DNA copy number amplification models.

Data mining from DNA copy number amplification model-based clusters

We applied data mining techniques and WHO derived cancer classifications as background data to explore the amplification model-based clustering of cancer cases. Cancer cases were divided into separate clusters for each chromosome using the inherent structure of DNA copy number amplification models. The component distributions of mixture model define clusters. Following the probabilistic approach, each observation was allocated to cluster $j*$, which maximizes the posterior probability according to Bayes's theorem

$$j^{*} = \arg\max_{j} \frac{p(j)p(\mathbf{x}|j)}{p(\mathbf{x})} = \arg\max_{j} \pi_{j} \prod_{i=1}^{d} \theta_{ji}^{x_i} \left(1-\theta_{ji}\right)^{1-x_i} .$$

Then, the observations belonging to cluster j were characterized by the corresponding component distribution θ_j. After clustering, the data in each cluster were divided into a test group and a reference group according to the collected classification terms. The test group contained those cases that were associated with a specific classification term (e.g., tobacco-related) and the reference group contained all other cases (not related to tobacco). Proportions of observations in each cluster f_{j1} and f_{j0} were calculated for the test and the reference group, respectively. The differences in cluster-specific observation proportions were compared by performing a pooled proportions statistical test [22]. The null hypothesis H_0 stated that the proportions are equal $f_{j1} = f_{j0}$ and alternative hypothesis H_1 was that the proportion in the test group is larger than that in the reference group $f_{j1} > f_{j0}$ or vice versa. The test statistic for the pooled proportions test is defined as

$$z = \frac{f_{j1} - f_{j0}}{\sqrt{\hat{f}\left(1-\hat{f}\right)\left(1/n_1 + 1/n_0\right)}}, \text{ where } \hat{f} = \frac{f_{j1}n_1 + f_{j0}n_0}{n_1 + n_0} \text{ and } n_1$$

and n_0 are the number of observations in the test group and in the reference group, respectively. The test statistic z is approximately normally distributed with zero mean and unit variance. The described test was carried out for each amplification pattern and classification term.

Due to the large number of hypotheses, we used the following procedure for control of the false discovery rate [23]. Let $p_{r_1} \le p_{r_2} \le \ldots \le p_{r_m}$ denote the observed ordered unadjusted p-values, where m is the number of hypotheses ($m = 13659$). For control of the false discovery rate at level α search $i^{*} = \max\left\{ i : p_{r_i} \le \frac{i}{m}\alpha \right\}$. The null hypotheses are rejected in the case of $i \le i*$. In the experiments, we used the significance level $\alpha = 0.001$, which corresponded to the unadjusted p-value 0.00005. Thus, the alternative hypothesis was accepted, i.e., the difference was regarded statistically significant, when the p-value of the test was lower than 0.00005.

Finite descriptions for continuous DNA copy number amplification models

Fusing of amplification models with relevant genomic mapping data required that continuous models were transformed into compact representations in the original nomenclature of the chromosomal bands. Finite descriptions, namely amplification patterns, were formed using maximal frequent itemset mining as presented earlier [24]. For the definitions and notations used in the following brief technical description of maximal frequent itemsets, we refer to Burdick et al. [25]. In a binary database with $\boldsymbol{x}_i = \{0,1\}^d$ and index set $I = \{1,...d\}$, an itemset is a subset of the index set I and a k-itemset is an itemset with cardinality k [25]. Support for an itemset X is defined as the frequency of the rows in the database including the items in X, i.e., how often X occurs in the database. σ is a predefined parameter that sets a threshold for selecting frequent itemsets. The mining task is to find all itemsets that have a frequency higher than σ. These are regarded as frequent itemsets. If an itemset X is frequent and no superset of X is frequent, we say that X is a maximal frequent itemset. The implementation by Burdick et al. [25] integrates a depth-first traversal of the itemset lattice with effective pruning mechanisms that significantly improve mining performance. We use maximal frequent itemsets in summarizing the marginal distribution of the clusters in a compact and understandable manner. In mining for the amplification patterns in the clustered data sets, we used a frequency threshold of $\sigma = 0.5$, so that the amplification patterns would be representative of the clusters in question. In simple terms, amplification patterns portray amplification models using finite ranges that capture the chromosomal structure of the amplified DNA element (also referred to as amplicon). The resulting amplification patterns are collections of the largest sets of chromosomal bands that occur jointly (together) in more than half of the data cases. The amplification patterns are thought to be representative of the whole cluster. As such, they represent the most probable amplicon structures and were used to map amplicon boundaries and putative DNA double-strand breakpoints.

Data mining from amplification patterns

A way to investigate the nature of DNA copy number amplification patterns is to compare them with relevant cytogenetic background data. For data mining, amplification patterns were used to depict the amplicon structures and identify putative ends of the amplified regions.

Cytogenetic features on amplicon boundaries were characterized to elucidate the genomic features that predispose to DNA double-strand breaks and enable amplification. Mechanistic models of amplification predict that DNA double-strand breaks must occur at the ends of the amplified chromosomal element. Enrichment of specific chromosomal regions in the ends of the amplification patterns (putative amplicons) was tested. The tested chromosomal regions included fragile sites, telomeres, centromeres, light and dark G-bands, and variable regions. Fragile sites were mapped according to the National Center for Biotechnology Information database annotations. Chromosome bands mapping to telomeric, centromeric and variable regions, as well as dark and light bands were extracted from the International System for Human Cytogenetic Nomenclature (2005) annotations [26]. We were interested in knowing whether the ends of amplification patterns were involved in an unexpectedly large number of specific chromosomal features, which would provide a possible explanation for DNA breakage associated with amplification. The test statistics were the differences between frequencies of chromosomal features in different sites of a given amplification pattern (borders, inside, and in general). If many patterns are based on one component distribution of the Bernoulli mixture model, they are very likely to overlap. In cases like this, the hypothesis testing includes the band in both sets with equal importance. We executed a permutation test of 10000 iterations to the hypothesis of comparing the proportions of chromosomal features. When executing the permutation test, an equal number of random bands to the number of bands of the chromosomal feature that was tested were distributed along the amplification patterns. Then, the corresponding proportions of the randomized sites occurring in the border and inside bands were calculated. This can be seen as a means to obtain empirical samples of the test statistic under the null hypothesis. The difference in proportions of the true test statistic and the randomized reference could then be calculated. The *p*-value for the one-tailed test was calculated using the difference in the proportions of the test statistic and randomized reference values. The threshold for significant findings was $p < 0.05$.

Results

Probabilistic models of DNA copy number amplification

We identified 111 amplification models (Figure 1). The number of clusters in chromosomes 1–22 and X varied between 2 and 7: chr1 (6), chr2 (4), chr3 (7), chr4 (2), chr5 (5), chr6 (6), chr7 (6), chr8 (7), chr9 (4), chr10 (3), chr11 (7), chr12 (6), chr13 (6), chr14 (3), chr15 (2), chr16 (4), chr17 (7), chr18 (4), chr19 (4), chr20 (6), chr21 (4), chr22 (3), and chrX (5). The Y chromosome was omitted from the analysis due to the low number of cases. Figure 1 shows the number of cancer cases (N_c) included in each amplification model. Vectors of probability parameters θ_j represent the resulting amplification models, where each mixture component assigns a continuous probability value θ_{ji} to each chromosome band. One chromosome band may have non-zero probabilities in different components, since components may correspond to different amplicon structures. Figure 2 shows that the clustering based on the probabilistic model of DNA copy number amplification manages to predetermine the structure of amplifications. It represents a typical amplification pattern and the general properties observed in an amplification, which may encompass additional bystander areas around the target gene locus.

DNA copy number amplification model-based clustering of cancer cases

The identified DNA copy number amplification models can be used to divide cases into clusters with similar molecular aberration. Figure 2 shows the data from chromosome 1 for reference. The clusters based on DNA copy number amplification models can be fused with known background data of the cancer types to study the underlying specificity of the amplifications.

Classification data of 95 human neoplasms were collected from the literature [see Additional file 1]. The analysis was restricted to malignant cancers of 82 different cancer types. Figure 3 presents the classification distribution based on cell lineages, age, and gender. Classification based on cell lineage contained anatomical system, organ, tissue, differentiation, and embryonic lineages. These attributes were divided into classification terms, e.g., nervous system (anatomical system), brain (organ), and glioma (cell). The classification terms can partially overlap with different attributes. Differentiation lineage (e.g., adenocarcinoma) refers to the histological type of the malignancy. Embryonic lineage divides cases into four main developmental compartments: epithelial, mesenchymal, hematopoietic, and neuroepithelial. The clinical attributes were age (pediatric, young adults, and adults) and gender specifications. In addition, 19 different etiological factors were collected (Figure 4). In all, 29 attributes and 100 classification terms were accumulated. Classification terms were appointed for cancers as primary data of individual cases was not available in the amplification data compilation. The compilation of DNA copy number amplification data was revised regarding the new annotations [12].

Frequencies of classification annotation terms were compared between cases in the studied cluster and a reference group that contained all other cases. The statistical significance of difference in the frequencies observed in each amplification model based cluster was determined using a hypothesis test. The significance threshold was set to

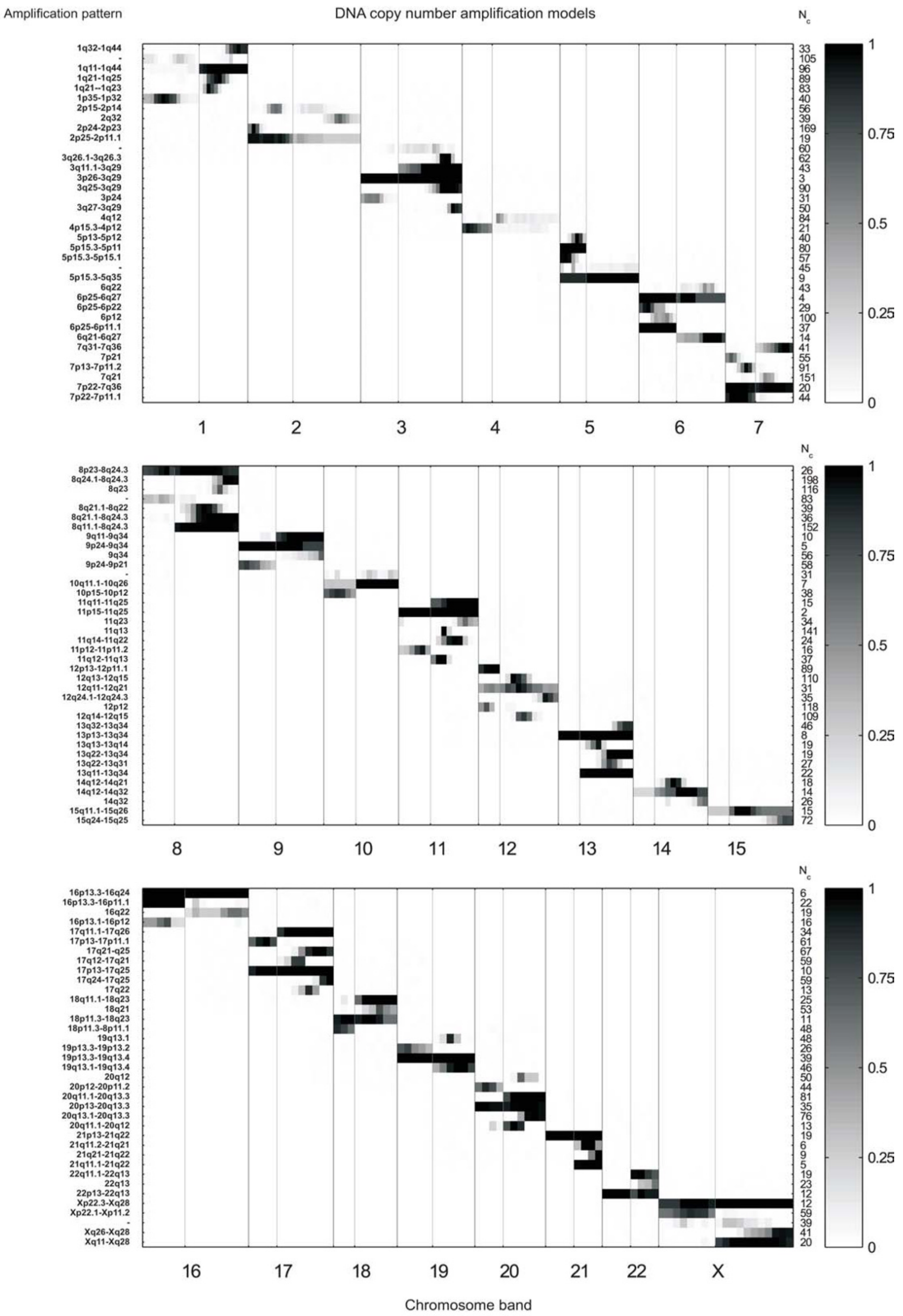

Figure 1

Probabilistic models of DNA copy number amplification. The models are component distributions of chromosome specific mixture models. Models are marked in the figure on separate lines. The probability of an amplification in each chromosome band is denoted using white to black scaling, where black indicates a chromosome band with a high probability of an amplification (p = 1) and white indicates a chromosome band with a low probability of an amplification (p = 0). Probabilities between 0 and 1 have been linearly scaled as shades of gray. Amplification patterns (based on the maximal frequent itemsets) are reported on the left side of the amplification models. Prevalence of the amplification patterns in terms of the number of cancer cases (N_c) is shown on the right side of the models.

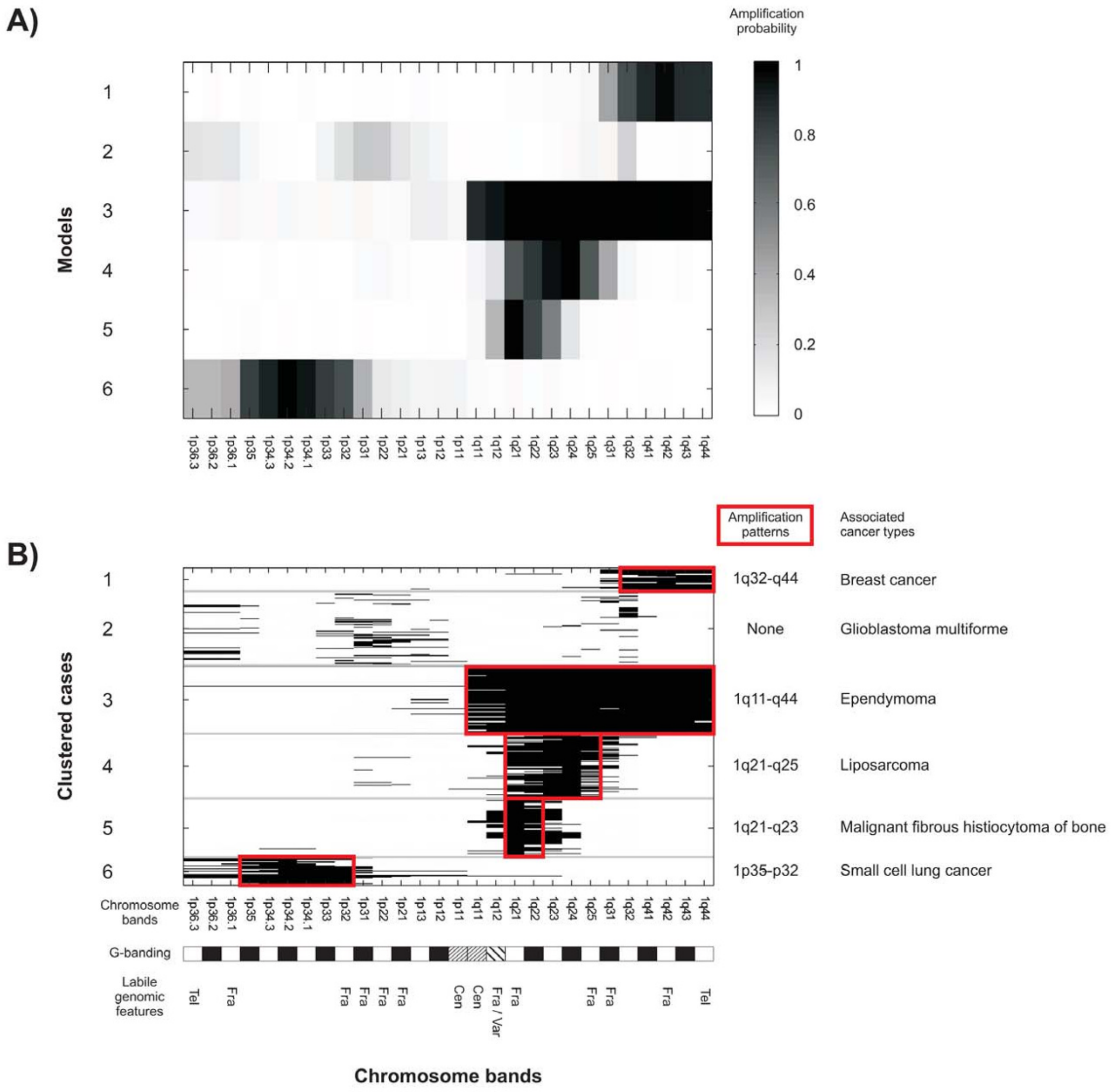

Figure 2

Amplification models and clustered data from chromosome 1. A) Component distributions of the amplification model and B) Clustered DNA copy number amplification data are shown for chromosome 1. Amplification models are presented as amplification probabilities in each chromosome band. White to black scaling represents amplification probabilities from zero to one. Six component distributions of the amplification model (panel A) have been used to cluster data into six clusters (panel B). In panel B, each line represents a single cancer case in the data set and amplified chromosome bands are marked black. Models and clusters are separated using gray horizontal lines. Amplification patterns are marked in the figure by red boxes and cancer types that are associated with the amplification model-based cancer clusters are denoted in the figure. Chromosome band annotations were collected from the International System for Human Cytogenetic Nomenclature (2005) [26] and are marked below corresponding chromosome bands. Black and white chromosome bands according to G-banding are marked. In addition, chromosome bands with specific chromosomal properties, namely, centromeres (Cen), telomeres (Tel), fragile sites (Fra), and variable regions (Var), are marked.

0.00005 and p-values were corrected for multiple testing using the Benjamini-Hochberg false discovery rate method [23]. Statistically significant observations are presented in Figure 5. Individual cancers (Figure 5A) and sample groups with specific classification terms (Figure 5B) were tested against all other samples. Our results show that a subset of amplifications is associated with specific cancer type, whereas some amplifications are

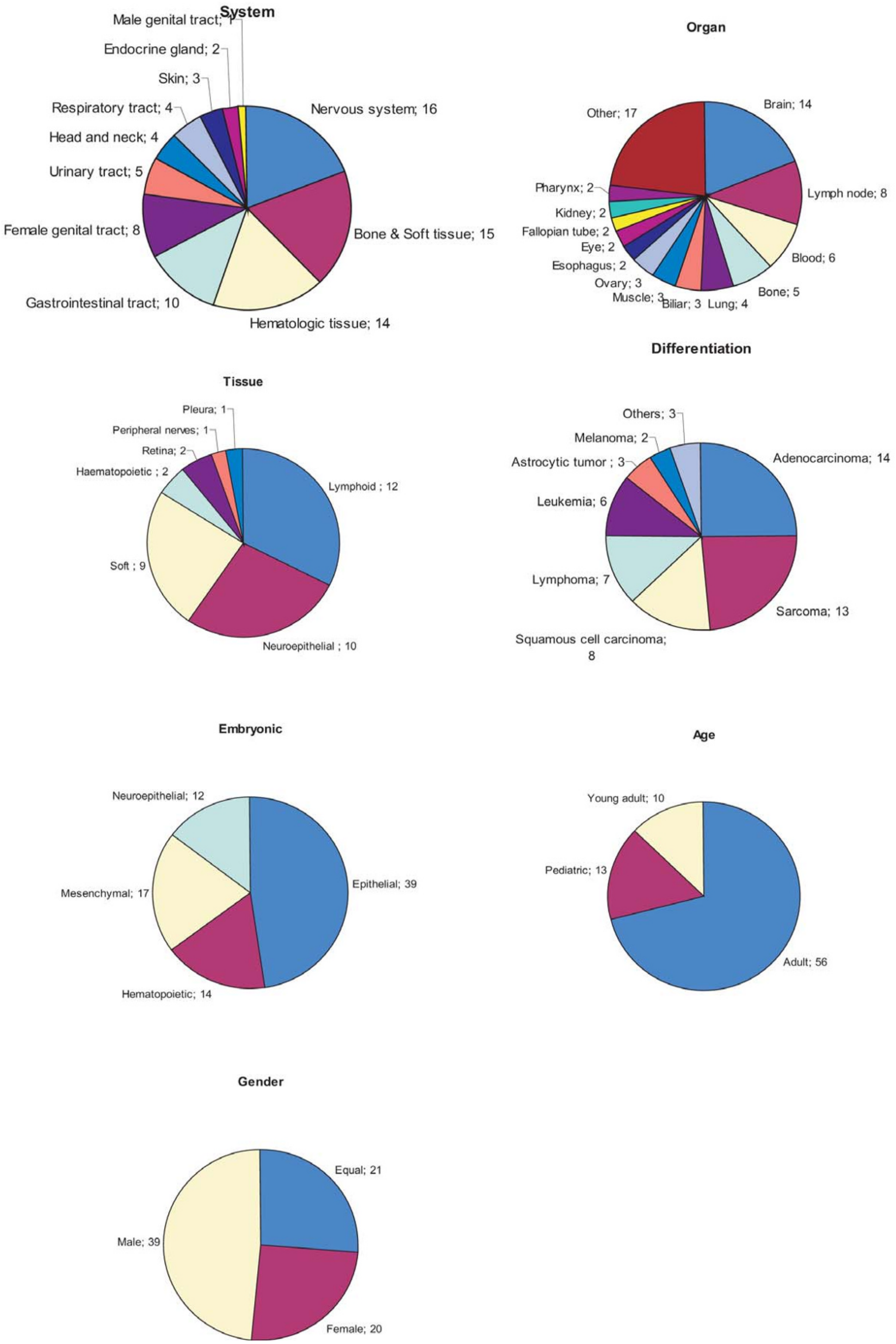

Figure 3

Distribution of cancer classification attributes and classification terms. Classification data was compiled from the WHO sources [9]. Figure is divided to individual pie charts according to different classification attributes. Each pie chart describes the numbers of cancer types in specific classification terms.

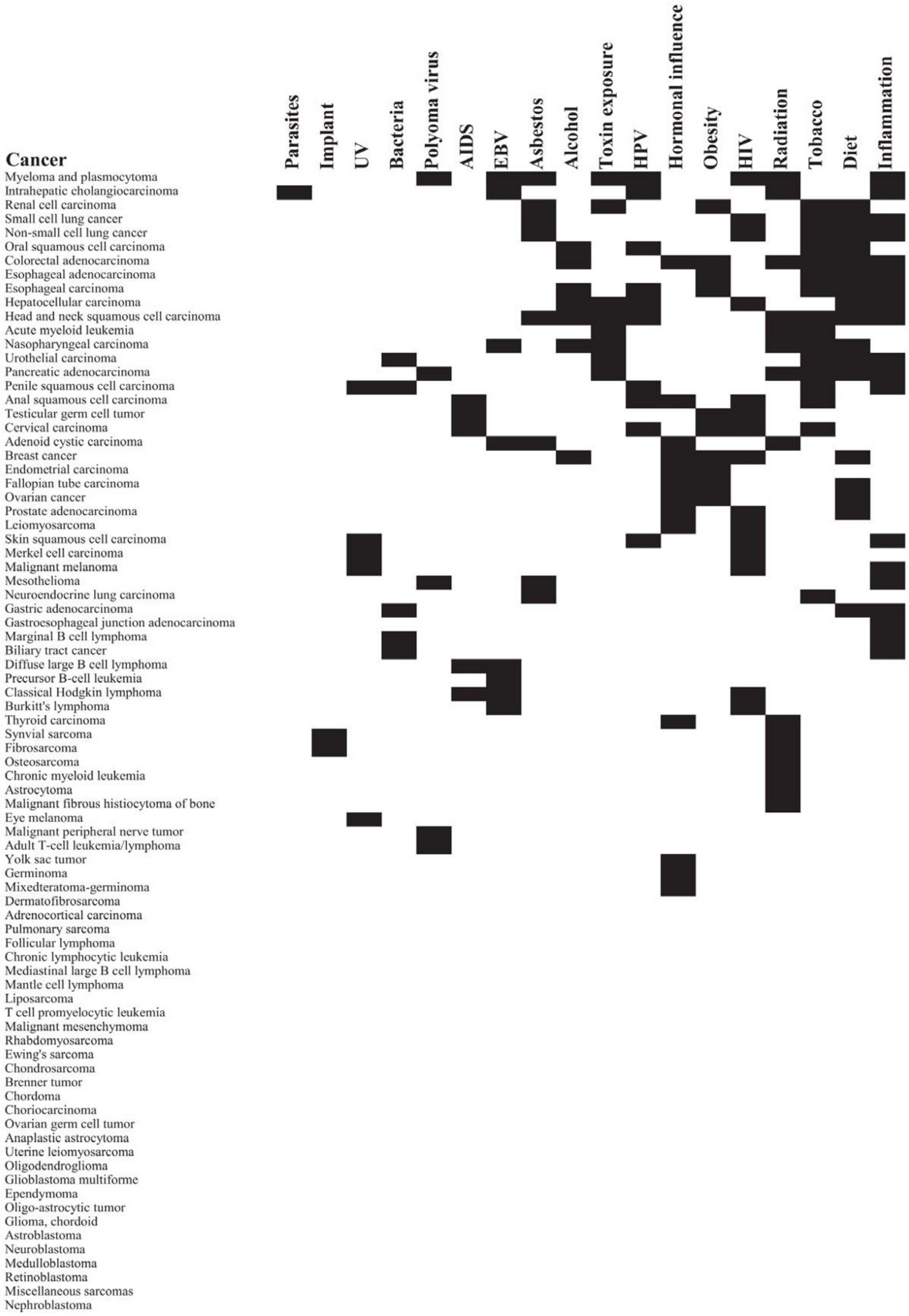

Figure 4

Etiological factors of cancers. Etiological data was compiled from the WHO sources [9]. Each row describes a cancer type and the etiological factors that have been associated with it (indicated by black boxes). Cancer type rows and etiological factor columns are sorted according to hierarchical clustering. Between groups-linkage method and Squared Euclidean distance measure for binary classification terms were used in clustering.

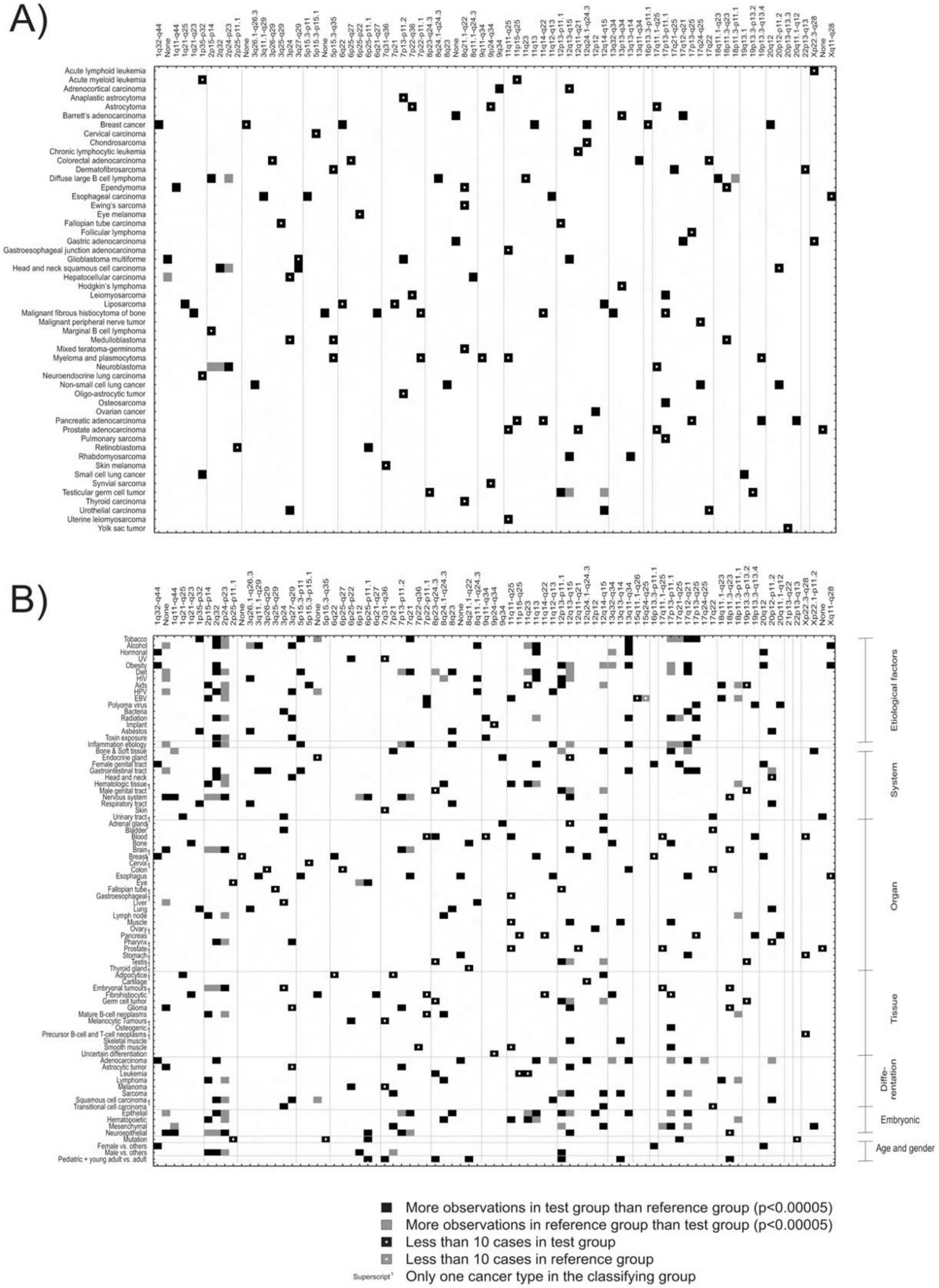

Figure 5

Associations between DNA copy number amplification models and classification terms. Data are presented for A) individual cancer types and B) classification terms. Data was collected from the WHO sources [9]. Rows show A) cancer types or B) specific classification terms. Each column represents one amplification pattern. Statistically significant differences in the prevalence of the test group (cancer cases or classification terms) and reference group (other cases) are marked with black and gray boxes. Findings that are based on a small sample (<10) are marked with a white circle inside the black box. Only amplification patterns, cancer types and classification terms with significant findings are shown.

more commonly shared by cancers of similar etiology or cell-lineage. For example, 1q33-q44 amplification is specific to breast cancers and 17q12-q21 region is specifically amplified in gastric cancer and Barrett's adenocarcinoma (Figure 5A). Similarly, 1q32-q44 amplification is specifically present in cancers associated with hormonal imbalance, obesity, female genital tract, breast tissue, and adenocarcinoma as well as cancers with female overrepresentation (Figure 5B). On the other hand, 17q12-q21 amplicon is enriched in cancers that associate with tobacco, obesity, diet, bacterial infections, inflammation, gastrointestinal tract, esophagus, stomach, adenocarcinoma, epithelial origin, and ethnic prevalence (Figure 5B).

Pattern description for probabilistic models of DNA copy number amplification

In order to facilitate the interpretation, we generated compact and understandable descriptions of the amplification models. The descriptions, i.e., amplification patterns, are local, finite, and plausible representations of the chromosomal areas of amplification. The amplification patterns were identified as maximal frequent itemsets and the chromosome bands are expressed following the original cytogenetic nomenclature. Amplification patterns depict the structures of the amplicons. One amplification model may result in many patterns. In fact, we extracted 140 maximal frequent itemsets from the 111 amplification models. When multiple patterns were identified for a specific amplification model, the most frequent pattern was chosen. Amplification patters are reported alongside their representative models in Figure 1. Amplification patterns in chromosome 1 are marked in Figure 2.

Mechanisms of DNA copy number amplification

Amplification patterns were used to investigate the mechanisms of DNA copy number amplification. Our hypothesis was that patterns represent the general structures of amplicons and can thus be applied to map amplicon boundaries and genomic loci that are susceptible to DNA double-strand damage. Differences in the proportions of labile chromosomal features in amplification patterns and within border bands or inside amplification patterns were determined using a hypothesis test (Table 1). Noteworthy is that the same band can occur both inside and on the border of the patterns, which can overlap. This is exemplified in Figure 2, where the ends of one pattern are inside another pattern. To be precise, the telomeric ends of the amplification patterns for models four and five (1q23 and 1q25, respectively) are inside the amplification pattern of model three (1q11-1q44). Statistically significant differences in proportions (*p*-value < 0.05) were identified when the proportions of fragile sites within the borders of amplification patterns were compared with the proportions of fragile sites in the amplification patterns. Similarly, light chromosome bands, telomeres, and centromeres were more frequent in the border bands than in the patterns in general. Dark chromosome bands inside the amplification patterns were more frequent than those in patterns as a whole.

Discussion

A machine learning approach was utilized to model DNA copy number amplifications in a landscape of cancers. The current modeling approach disregarded the cancer type information and modeled amplifications based on case-specific data vectors. This resulted in identification of 111 amplification models (Figure 1). The identified mod-

Table 1: Hypothesis testing of proportions of labile chromosomal sites within the amplification patterns.

Labile chromosomal site	Proportion in patterns	Proportion in amplification pattern borders	Proportion inside amplification patterns	*p*-value
Fragile sites	0.3086	0.3693	-	0.0069*
	0.3086	-	0.2975	0.8448
Dark chromosome bands	0.3739	0.3182	-	0.9896
	0.3739	-	0.4301	0.0000*
Light chromosome bands	0.4629	0.5114	-	0.0405*
	0.4629	-	0.4122	1.0000
Telomere bands	0.1217	0.2330	-	0.0000
	0.1217	-	0.0000	1.0000**
Centromere bands	0.1187	0.1477	-	0.0378*
	0.1187	-	0.1147	0.7481
Variable bands	0.0445	0.0227	-	0.9388
	0.0445	-	0.0430	0.7154

The proportion of chromosomal attributes in patterns in general is compared to the proportion of chromosomal attributes on the borders of the amplification patterns as well as inside amplification patterns (excluding the borders). Statistical significance of the difference in proportions was determined using permutation tests and *p*-values are marked in the table. Significant findings are marked with an asterisk.
*Statistically significant findings ($p < 0.05$).
**By definition, telomeres can not be located inside the amplification patterns and therefore this test could not yield any significant findings.

els could be viewed as specific cancer classes that allow more refined dissection of amplification processes. Our hypothesis was that cancers with a common amplification model might exhibit dependency on a specific oncogene amplification and therefore share common biological background. We tested this hypothesis by analyzing the distribution of WHO classification terms in the amplification modeling-based clustering of cancer cases. Specific amplification models were shown to associate with specific cancer types and classification terms (Figure 5). The non-random structure in the spectrum of DNA copy number amplifications in cancer suggests that cancer etiology and tumor microenvironment could manifest as specific amplification signatures. According to our results, amplifications are selected according to the anatomical locations and biological background of the cancers. Theoretically, carcinogenesis could be viewed as an evolutionary process that involves the selection of cancer cells in the somatic tissue by specific mutations. In the Darwinian perspective, the classification based on DNA copy number amplifications reflects the differences in the selective properties in different anatomical locations and in specific adaptation of cancers with similar biological backgrounds.

DNA copies generated in amplification manifest as concatenated homogenously staining regions and extra-chromosomal acentric DNA fragments, double minutes and episomes [3,27]. Models of DNA amplification mechanisms, the breakage-fusion-bridge cycle and excision of extrachromosomal DNA segments, state that two independent DNA double-strand breaks that flank the amplified region are required to initiate the amplification pathway [28]. Using the amplification modeling and pattern discovery approach, presented in the current study, fixing of finite amplicon structures became feasible and amplification patterns, representations of amplicon structures, were determined from the DNA copy number amplification models. The amplification patterns could then be used to identify specific chromosomal sites that associate with amplicon boundaries. By fusing the boundaries of amplification patterns with cytogenetic annotations of the genome it was possible to elucidate the features in the chromosomal structure and genomic architecture that predispose the genome to amplifications. The hypothesis was that specific genomic regions may be damage-prone and susceptible to DNA double-strand breaks. The statistical hypothesis testing demonstrated that fragile sites, light chromosome bands, telomeres, and centromeres were enriched in the ends of the amplicons (Table 1). This suggests that these sites might be associated with the amplification mechanisms and DNA double-strand breakage at amplicon boundaries. Fragile sites are damage-prone genomic regions when cells are treated with chemicals that interfere with replication [29], which makes it likely that they are often found at amplification breakpoints. In addition to fragile sites, chromosome ends are unstable and may produce double-strand DNA breaks due to telomere shortening during replication and cell division [30]. Similarly, centromere regions have been shown to be unstable and damage-prone upon replication stress [31], which might explain the accumulation of amplicon boundaries on them. Light chromosome bands were also enriched at amplicon boundaries. The light bands contain euchromatin and are gene-rich, G/C-rich and late-replicating, whereas dark bands correspond to gene-poor, A/T-rich and early replicating heterochromatin. Due to its high gene content, the structure of euchromatin is more open than that of heterochromatin [32], which may affect its physical protection and render euchromatin more susceptible to DNA damage than the gene-poor heterochromatin. We hypothesize that open chromatin reduces the protection of chromosomal DNA and serves as preferential target for DNA damage. Open chromatin might therefore expose light chromosome bands to DNA double-strand breaks that initiate the amplification pathways.

Conclusion

We classified human cancers based on DNA copy number amplification models. Cancer cases were fused with the WHO classification annotations. The inherent structure in the probabilistic clustering suggests that amplifications are non-randomly selected according to biological backgrounds of cancers. Amplification patterns were extracted and probed using cytogenetic annotations. We show statistical evidence that connects fragile sites, telomeres, centromeres, and light chromosome bands to the amplification mechanism. These results suggest that labile chromosomal features are involved in the amplification process by promoting the formation of DNA double-strand breaks at amplicon margins.

Abbreviations

CGH: comparative genomic hybridization; WHO: World Health Organization; EM: Expectation-Maximization.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

SM, SK and JH conceived and designed the research and set goals. SM and TB collected clinical and biological background data. SM, JT and JH decided on the methodologies and planned data analysis. JT and JH performed computational work. SM, TB and SK interpreted the results. All authors participated in drafting of the manuscript and read and approved the final version. SM coordinated the research.

Additional material

Additional file 1

Classification annotations for cancer types. Supplement 1_Classification data.txt is a tab-delimited text-file containing all collected cancer classification data.

Click here for file

[http://www.biomedcentral.com/content/supplementary/1755-8794-1-15-S1.txt]

Acknowledgements

The authors of this article would like to thank Dr. Outi Monni (Institute of Biomedicine/Biochemistry, Genome-Scale Biology Research Program and Biomedicum Biochip Center, University of Helsinki) and Dr. Kai Puolamäki (Department of Information and Computer Science, Helsinki University of Technology) for critical comments on the manuscript. This work was supported by the Academy of Finland in the SYSBIO (Systems Biology and Bioinformatics) Research Program, the State Appropriations of Helsinki University Central Hospital, the Sigrid Jusélius Foundation, and the Finnish Cancer Organizations.

References

1. Myllykangas S, Himberg J, Böhling T, Nagy B, Hollmén J, Knuutila S: **DNA copy number amplification profiling of human neoplasms.** *Oncogene* 2006, **25(55)**:7324-7332.
2. Mitelman F, Johansson B, Mertens F: **Catalog of Chromosome Aberrations in Cancer.** *Volume 2*. New York , Wiley-Liss; 1994.
3. Albertson DG, Collins C, McCormick F, Gray JW: **Chromosome aberrations in solid tumors.** *Nat Genet* 2003, **34(4)**:369-376.
4. Lengauer C, Kinzler KW, Vogelstein B: **Genetic instabilities in human cancers.** *Nature* 1998, **396(6712)**:643-649.
5. Brodeur GM, Hogarty MD: **Gene amplification in human cancers: biological and clinical significance.** In *The genetic basis of human cancer* Edited by: Vogelstein B, Kinzler KW. New York , McGraw-Hill; 1998:161-172.
6. Schwab M, Westermann F, Hero B, Berthold F: **Neuroblastoma: biology and molecular and chromosomal pathology.** *Lancet Oncol* 2003, **4(8)**:472-480.
7. Vogt N, Lefevre SH, Apiou F, Dutrillaux AM, Cor A, Leuraud P, Poupon MF, Dutrillaux B, Debatisse M, Malfoy B: **Molecular structure of double-minute chromosomes bearing amplified copies of the epidermal growth factor receptor gene in gliomas.** *Proc Natl Acad Sci U S A* 2004, **101(31)**:11368-11373.
8. Myllykangas S, Böhling T, Knuutila S: **Specificity, selection and significance of gene amplifications in cancer.** *Semin Cancer Biol* 2007, **17(1)**:42-55.
9. Kleihues P, Sobin LH: **World Health Organization Classification of Tumours.** In *World Health Organization Classification of Tumours* Lyon , IARCPress. 2000 - 2006
10. Kallioniemi A, Kallioniemi OP, Sudar D, Rutovitz D, Gray JW, Waldman F, Pinkel D: **Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors.** *Science* 1992, **258(5083)**:818-821.
11. Baudis M, Cleary ML: **Progenetix.net: an online repository for molecular cytogenetic aberration data.** *Bioinformatics* 2001, **17(12)**:1228-1229.
12. **Laboratory of Cytomolecular Genetics (CMG)** [http://www.helsinki.fi/cmg/cgh_data.html]
13. MacLachlan GJ, Peel D: **Finite Mixture Models (Wiley Series in Probability and Statistics).** New York , John Wiley & Sons; 2000.
14. Everitt BS, Hand DJ: **Finite Mixture Distributions (Monographs on Applied Probability and Statistics).** Boca Raton , Chapman & Hall; 1981.
15. Gyllenberg M, Koski T, Reilink E, Verlaan M: **Non-uniqueness in probabilistic numerical identification of bacteria.** *Journal of Applied Probability* 1994, **31**:542-548.
16. Carreira-Perpinan MA, Renals S: **Practical identifiability of finite mixtures of multivariate Bernoulli distributions.** *Neural Computation* 2000, **12**:141-152.
17. Dempster AP, Laird NM, Rubin DB: **Maximum likelihood from incomplete data via the EM algorithm.** *Journal of the Royal Statistical Society, Series B* 1977, **39**:1-39.
18. Redner RA, Walker HF: **Mixture densities, maximum likelihood and the EM algorith.** *SIAM Review* 1984, **26(2)**:195-234.
19. MacLachlan GJ, Thiriyambakam K: **The EM Algorithm and Extensions (Wiley Series in Probability and Statistics).** New York , John Wiley & Sons; 1996.
20. Tikka J, Hollmén J, Myllykangas S: **Mixture modeling of DNA copy number amplification patterns in cancer.** In *Proceedings of the 9th International Work-Conference on Artificial Neural Networks (IWANN'2007): June 2007; San Sebastián* Edited by: Sandoval F, Prieto A, Cabestany J, Graña M. Heidelberg , Springer-Verlag; 2007:972-979.
21. Efron B, Tibshirani RJ: **An Introduction to the Bootstrap (Monographs on Statistics and Applied Probability).** Boca Raton , Chapman & Hall; 1993.
22. Milton JS, Arnold JC: **Introduction to Probability and Statistics: Principles and Applications for Engineering and the Computing Sciences.** 2nd edition. New York , McGraw-Hill; 1990.
23. Benjamini Y, Hochberg Y: **Controlling the False Discovery Rate: a Practical and Powerful Approach to Multiple Testing.** *Journal of Royal Statistical Society* 1995, **57(1)**:289-300.
24. Hollmén J, Tikka J: **Compact and Understandable Descriptions of Mixtures of Bernoulli Distributions.** In *Proceedings of the 7th International Symposium on Intelligent Data Analysis (IDA 2007): September 2007; Ljubljana* Edited by: Berthold MR, Shawe-Taylor J, Lavrac N. Heidelberg , Springer-Verlag; 2007:1-12.
25. Burdick D, Calimlim M, Flannick J, Gehrke J, Yiu T: **MAFIA: A Maximal Frequent Itemset Algorithm.** *IEEE Transactions on Knowledge and Data Engineering* 2005, **17(11)**:1490-1504.
26. Shaffer LG, Tommerup N: **ISCN (2005): An International System for Human Cytogenetic Nomenclature.** Basel , S. Karger; 2005.
27. Schwab M: **Amplification of oncogenes in human cancer cells.** *Bioessays* 1998, **20(6)**:473-479.
28. Myllykangas S, Knuutila S: **Manifestation, mechanisms and mysteries of gene amplifications.** *Cancer Lett* 2006, **232(1)**:79-89.
29. Schwartz M, Zlotorynski E, Kerem B: **The molecular basis of common and rare fragile sites.** *Cancer Lett* 2006, **232(1)**:13-26.
30. Murnane JP, Sabatier L: **Chromosome rearrangements resulting from telomere dysfunction and their role in cancer.** *Bioessays* 2004, **26(11)**:1164-1174.
31. Hellman A, Zlotorynski E, Scherer SW, Cheung J, Vincent JB, Smith DI, Trakhtenbrot L, Kerem B: **A role for common fragile site induction in amplification of human oncogenes.** *Cancer Cell* 2002, **1(1)**:89-97.
32. Gilbert N, Boyle S, Fiegler H, Woodfine K, Carter NP, Bickmore WA: **Chromatin architecture of the human genome: gene-rich domains are enriched in open chromatin fibers.** *Cell* 2004, **118(5)**:555-566.

Pre-publication history

The pre-publication history for this paper can be accessed here:

http://www.biomedcentral.com/1755-8794/1/15/prepub