Oncogene (2003) 22, 2226–2235

www.nature.com/onc

ORIGINAL PAPERS

Pathogenesis of DNA repair-deficient cancers: a statistical meta-analysis of putative Real Common Target genes

Stefan M Woerner[1], Axel Benner[2], Christian Sutter[1], Marian Schiller[1], Yan P Yuan[3], Gisela Keller[4], Peer Bork[3,5], Magnus von Knebel Doeberitz[1] and Johannes F Gebert*[,1]

[1]*Department of Molecular Pathology, Institute of Pathology, University of Heidelberg, Im Neuenheimer Feld 220/221, D-69120 Heidelberg, Germany;* [2]*Central Unit Biostatistics, German Cancer Research Center, Im Neuenheimer Feld 280, D-69120 Heidelberg, Germany;* [3]*European Molecular Biology Laboratory, Meyerhofstr. 1, D-69117 Heidelberg, Germany;* [4]*Institute of Pathology, Technical University München, Trogerstr. 18, D-81675 Munich, Germany;* [5]*Max-Delbrück-Centrum for Molecular Medicine, Robert-Rössle-Str. 10, D-13092 Berlin, Germany*

DNA mismatch repair deficiency is observed in about 15% of human colorectal, gastric, and endometrial tumors and in lower frequencies in a minority of other tumors thereby causing insertion/deletion mutations at short repetitive sequences, recognized as microsatellite instability (MSI). Evolution of tumors, including those with MSI, is a continuous process of mutation and selection favoring neoplastic growth. Mutations in microsatellite-bearing genes that promote tumor cell growth in general (Real Common Target genes) are assumed to be the driving force during MSI carcinogenesis. Thus, microsatellite mutations in these genes should occur more frequently than mutations in microsatellite genes without contribution to malignancy (ByStander genes). So far, only a few Real Common Target genes have been identified by functional studies. Thus, comprehensive analysis of microsatellite mutations will provide important clues to the understanding of MSI-driven carcinogenesis. Here, we evaluated published mutation frequencies on 194 repeat tracts in 137 genes in MSI-H colorectal, endometrial, and gastric carcinomas and propose a statistical model that aims to identify Real Common Target genes. According to our model nine genes including BAX and TGFβRII were identified as Real Common Targets in colorectal cancer, one gene in gastric cancer, and three genes in endometrial cancer. Microsatellite mutations in five additional genes seem to be counterselected in gastrointestinal tumors. Overall, the general applicability, the capacity to unlimited data analysis, the inclusion of mutation data generated by different groups on different sets of tumors make this model a useful tool for predicting Real Common Target genes with specificity for MSI-H tumors of different organs, guiding subsequent functional studies to the most likely targets among numerous microsatellite harboring genes.
Oncogene (2003) **22,** 2226–2235. doi:10.1038/sj.onc.1206421

*Correspondence: JF Gebert;
E-mail: Johannes_gebert@med.uni-heidelberg.de
Received 25 July 2002; revised 7 January 2003; accepted 22 January 2003

Keywords: mismatch repair; coding microsatellite; microsatellite instability; Real Common Target genes; ByStander genes

Introduction

Two major forms of genetic instability known as chromosomal instability (CIN) (Lengauer *et al*., 1997) and microsatellite instability (MSI) (Ionov *et al*., 1993; Thibodeau *et al*., 1993) have been observed in human malignancies. The type and spectrum of mutated genes markedly differ among CIN and MSI-H tumors (Ionov *et al*., 1993; Konishi *et al*., 1996; Lengauer *et al*., 1998), suggesting distinct but not mutually exclusive pathways of carcinogenesis (Perucho *et al*., 1994). MSI usually manifests as insertion or deletion mutations at short repetitive DNA sequences termed microsatellites and is caused by functional loss of cellular DNA mismatch repair (MMR) (Aaltonen *et al*., 1993; Ionov *et al*., 1993; Thibodeau *et al*., 1993). In the absence of any selection pressure these mutations should occur at a similar frequency in noncoding and coding microsatellites, primarily depending on repeat type and length (Sagher *et al*., 1999; Woerner *et al*., 2001).

It is generally accepted that MSI carcinogenesis like progression of other tumors is an evolutionary process driven by genetic instability with the generation of large numbers of random mutations and selection of clones that exhibit malignant properties (for a review see Loeb, 2001). Mutations in nonfunctional noncoding intronic or intergenic microsatellite sequences are unlikely to favor neoplastic growth of MMR-deficient cells. However, in a recent report substantial variation in the prevalence of mutations among noncoding mononucleotide tracts of different types but identical length was observed (Zhang *et al*., 2001). Although these repeats were located deep within intronic sequences of a number of different genes, evidence that these regions of the genome do not play any tumorigenic role has not been provided. In contrast, MSI-associated insertion/deletion mutations in coding microsatellites of expressed genes

inevitably confer a shift in the translational reading frame of encoded proteins ultimately abrogating or severely altering the normal protein function. At the cellular level, frameshift mutations in coding microsatellites might lead to different functional consequences: If the mutant proteins disrupt essential metabolic or signaling pathways, MMR-deficient cells will face growth arrest and eventually cell death. This in turn leads to counterselection and a bias toward decreased mutation frequencies in these particular cMS gene sequences in MSI-H tumors. Alternatively, mutations in coding microsatellites and their encoded proteins might not exert any tumorigenic effect and genetic alterations in these sequences should occur randomly and with similar frequency like mutations in nonfunctional microsatellites (ByStanders). Finally, some frameshift mutations in coding microsatellites will provide a growth advantage or an immune escape mechanism to affected cells. A positive selection for these mutations will lead to increased mutation frequencies in the corresponding genes in MMR-deficient tumors. It is generally believed that this latter subset of coding microsatellites defines critical targets of frameshift mutations specifically promoting MSI carcinogenesis in a large proportion of tumors (Real Common Targets).

Recent attempts therefore aimed to identify systematically cMS gene sequences in the human genome and to determine their frameshift mutation frequencies in MSI-H tumors (Duval *et al*., 2001). The proteins encoded by these genes participate in a variety of essential cellular processes like signal transduction (TGFβRII, IGFIIR, PTEN; Markowitz *et al*., 1995), apoptosis (BAX, caspase 5; Rampino *et al*., 1997; Schwartz *et al*., 1999), DNA repair (hMSH3, hMSH6, MBD4; Malkhosyan *et al*., 1996), transcriptional regulation (TCF-4; Duval *et al*., 1999), protein translocation and modification (SEC63, OGT; Woerner *et al*., 2001), or immune surveillance (β2M (Bicknell *et al*., 1996)). These studies revealed major differences in frameshift mutation frequencies among cMS sequences of identical type and length and different mutation frequencies for a given cMS sequence in MSI-H tumors of different organs. These findings strongly suggest that coding region MSI is a selective process and mutational inactivation of specific target genes and pathways should provide a growth advantage to such cells.

Initial attempts to distinguish Real Common Target genes from randomly mutated ByStander genes led to the proposal of five criteria: (i) a high mutation frequency, (ii) biallelic inactivation, (iii) a role in a growth suppressor pathway, (iv) the occurrence of alterations within the same pathway in MSI-negative tumors, and (v) *in vitro* or *in vivo* functional suppressor studies (Boland *et al*., 1998). The third and fourth points are controversially discussed (Perucho, 1999), because not all important pathways can be assumed to be known yet, and the pathways and mutated genes involved in carcinogenesis appear to differ significantly between MSI-H and MSS tumors (Perucho *et al*., 1994). Only for a few of these genes like TGFβRII (Markowitz *et al*., 1995) and BAX (Rampino *et al*., 1997), functional studies have provided clear evidence for a central role in MSI carcinogenesis. Mutation frequency apart from functional studies thus remains the most simple parameter when comparing published data on known cMS gene sequences with mutation data gathered on novel cMS candidate sequences in MSI-H tumors.

In the present study, we performed a comprehensive literature search on published cMS and ncMS mutation frequencies in MSI-H colorectal, endometrial, and gastric cancer. Based on cumulative mutation data of 169 cMS and 25 ncMS, we propose a statistical model that might prove helpful in predicting Real Common Target genes. According to our model, 11 microsatellite sequences were predicted as putative Real Common Target genes in MSI-H colorectal, gastric, or endometrial tumors. Mutations in five other microsatellite bearing genes seem to be selected against in these MSI-H tumor entities. The results of the statistical model presented here are independent of functional considerations and irrespective of the sets of tumor samples analysed by different investigators.

Results

Evaluation of microsatellite mutation data

As a first step towards establishing a reliable statistical model for predicting Real Common Target and Bystander genes in MSI-H tumors we thoroughly reviewed all reported data on microsatellite mutations. Only mononucleotide repeats were considered since they represent the most simple type of microsatellites. This extensive survey of the literature (April 2002) revealed 110 publications referring to mutation analyses of 245 coding and noncoding microsatellites in 177 genes either in MSI-H colorectal, gastric, or endometrial cancer. For statistical analysis only those primary data were included which unambiguously assigned specific cMS mutations to individual MSI-H tumor samples. A minimum cumulative sample number ($n=10$) was defined as a study entry criterion for each mononucleotide repeat in order to reduce sampling errors. Accordingly, mutation data on 161 coding mononucleotide repeats in 108 genes originating from 101 publications met these criteria (Supplemental Tables 1–3 available at http://www.med.uni-heidelberg.de/patho/pathmol/woerner/model_real_targets). In addition, we included unpublished mutation data on eight novel cMS gene sequences not previously associated with MSI carcinogenesis (Supplemental Tables 1–3). These eight cMS containing genes originated from our previous systematic search for human cMS candidate sequences (Woerner *et al*., 2001) and were chosen for mutation analysis because they comprise particularly long coding mononucleotide repeats. We also calculated cumulative mutation frequencies in MSI-H tumors for 25 noncoding microsatellites (ncMS) in 22 different genes originating from 33 publications and own analyses (Supplemental Tables 1–3). The data set of cMS and ncMS sequences showed a disproportionate distribution

regarding repeat type and length. In particular, homopolymeric runs of A represented the most commonly investigated repeat in this data set [$n(A) = 131$, $n(C) = 20$, $n(G) = 17$, $n(T) = 25$] and thus excluded subsequent regression analysis by repeat type. We also noticed that especially short and long repeat tracts were under-represented [$n(N_{\leqslant 4}) = 9$, $n(N_5) = 21$, $n(N_6) = 24$, $n(N_7) = 10$, $n(N_8) = 59$, $n(N_9) = 38$, $n(N_{10}) = 21$, $n(N_{\geqslant 11}) = 12$]. Overall, 194 repeat tracts (169 cMS; 25 ncMS) in 137 genes (115 with cMS; 21 with ncMS, one with a cMS and an ncMS (Supplemental Tables 1–3*) entered our statistical analysis.

Cumulative mutation frequencies

Cumulative mutation frequencies were determined separately for each tumor entity. Compilation of published cMS mutation data qualifying for our statistical approach revealed the highest number of annotations for MSI-H colorectal tumors (CRC: 15.632) whereas recorded mutation analyses of cMS sequences in MSI-H gastric (GC: 4.232) or endometrial cancers (EC: 2.652) were strikingly lower. For a given repeat, both cumulative sample numbers (CRC: 10–1.236; GC: 10–399; EC: 12–275) and cumulative mutation frequencies (CRC: 0–100%; GC: 0–100%; EC: 0–92%) spanned over a wide range. The five most thoroughly investigated coding region microsatellites in all three tumor entities include the genes *TFGβIIR* (CRC: 1.236; GC: 399; EC: 275), *BAX* (CRC: 949; GC: 360; EC: 196), *MSH3* (CRC: 927; GC: 267; EC: 206), *MSH6* (CRC: 916; GC: 281; EC: 178), and *IGFIIR* (CRC: 637; GC: 325; EC: 191).

Regression analysis for MSI-H colorectal cancers

For MSI-H colorectal tumors, we investigated cMS ($n = 163$ within 114 genes) and ncMS ($n = 25$ within 22 genes, Supplemental Table 1 (available at http://www.med.uni-heidelberg.de/patho/pathomol/woerner/model_real_targets/)) sequences. Separate analysis of ncMS and cMS mutation data revealed different regression curves for both classes of repeats. (Figure 1). This difference was not unexpected because ncMS mutation frequencies most likely reflect the statistical mutation probability of ByStander genes, whereas the group of cMS sequences will include genes with impact on tumor progression. In order to exclude a biased statistical calculation combined analysis of ncMS and cMS mutation data was thus performed. Based on these data, we determined the 95% prediction interval (Figure 2). Using this approach, nine cMS harboring genes (PTHL3, HT001, TGFβIIR, AC1, ACVR2, SLC23A1, BAX, TCF-4, and MSH3) ranged above the upper prediction limit and thus were assumed to represent Real Common Targets in MSI-H colorectal tumors (Figure 2). In contrast, the cMS in CHD2 and RFC3 and the two ncMS in the BCL2 and WAF1 genes showed decreased mutation rates significantly below the lower prediction limit. Taken together, our model identified nine cMS gene sequences as Real Common Targets. Moreover, mutations in two cMS as well as in two ncMS sequences appeared to be counterselected in MSI-H colorectal tumors.

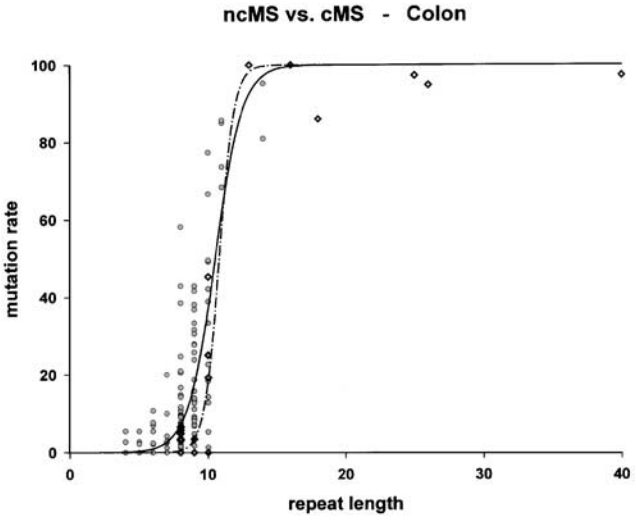

Figure 1 Comparison of noncoding (diamonds) and coding (circles) microsatellite harboring genes based on their mutation rate in MSI-H colorectal cancer

Regression analysis for MSI-H gastric cancers

Similar regression analyses were performed for repeat mutation data of MSI-H gastric ($n = 73$ cMS in 62 genes and $n = 7$ ncMS in seven genes, Supplemental Table 2 (available at http://www.med.uni-heidelberg.de/patho/pathomol/woerner/model_real_targets/)) and endometrial carcinoma ($n = 60$ cMS in 52 genes and $n = 2$ ncMS in two genes, Supplemental Table 3 (available at http://www.med.uni-heidelberg.de/patho/pathomol/woerner/model_real_targets/)). In MSI-H gastric cancer, mutation frequencies for the A_{11} repeat in the gene for the TATA box binding associated factor B of RNA polymerase 1 (TAF1B) reached 87%. The cumulative mutation frequency of this repeat tract ranged above the upper prediction threshold thus claiming it to represent a Real Common Target gene (Figure 3). In addition, the T_{14} repeat tract within a yet uncharacterized cDNA sequence (U79260) showed a significantly decreased mutation frequency of less than 10% (1/15) in MSI-H gastric cancers compared to a mutation frequency of about 80% (17/21) in colorectal tumors ($P = 0.005$, Fisher's exact test, P-value adjusted according to Bonferroni–Holm). Thus, its mutation rate below the lower prediction threshold strongly suggests a negative selection pressure on this repeat tract in MSI-H gastric cancer.

Regression analysis for MSI-H endometrial cancers

In MSI-H endometrial tumors, three cMS located within the genes TAF1B, AIM2, and SLC23A1 appeared above the upper prediction interval reasoning them as Real Common Targets (Figure 4). Data points for the four well-known genes BAX, MSH3, MSH6, and PTEN clearly mapped above the average mutation frequency of

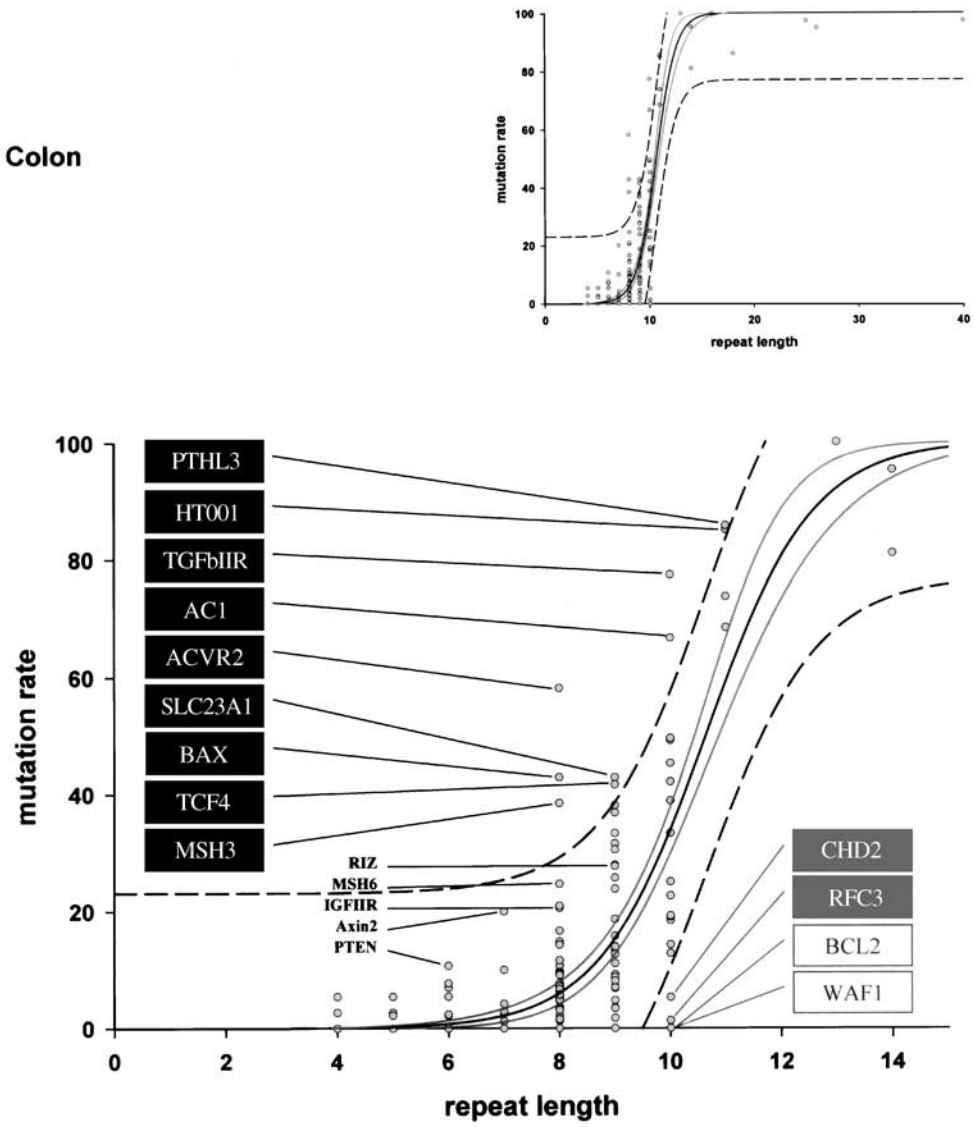

Figure 2 Regression analysis of ncMS and cMS (circles) for MSI-H colorectal cancer. The lower part is an enlarged clipping of the small inset above. The fitted regression line (solid black), the corresponding 95% confidence limits (gray lines), and the 95% prediction intervals (dashed black lines) are shown. Real Common Target genes characterized by high mutation frequency in cMS (black filled boxes) are identified above the upper prediction curve whereas genes displaying particularly low mutation frequencies in cMS (gray filled boxes) or ncMS (open gray boxes) due to counterselection reside below the lower prediction curve

repeats of the same length and type, but still below the upper prediction curve. Finally, the cMS mutation frequencies for IGFIIR, TCF-4, and TGFβIIR placed these genes quite close to or even below the fitted regression line. In conclusion, our statistical analysis revealed three cMS harboring genes as Real Common Target genes for MSI-H endometrial cancers.

Discussion

A new model for predicting genes involved in MSI tumorigenesis

Recently, several computer-based approaches have been initiated to identify and analyse mutations systematically in coding microsatellites and to identify new common target genes contributing to MSI carcinogenesis (Duval *et al*., 2001; Mori *et al*., 2001; Woerner *et al*., 2001). These studies have provided a large number of data for a new statistical model that allows prediction about genes whose mutations are either selected for or against during MSI tumorigenesis. This model is generally applicable to all experimentally observed mutation data generated by different groups on different sets of MSI-H tissues and samples, accounting for a variety of different parameters.

Our model accounts for the known correlation between repeat length and mutation frequency (Sia *et al*., 1997) and relies on three assumptions: First, the mutation rate of a repeat tract with $n=1$ repeat unit is identical to the somatic mutation rate in eucaryotic cells

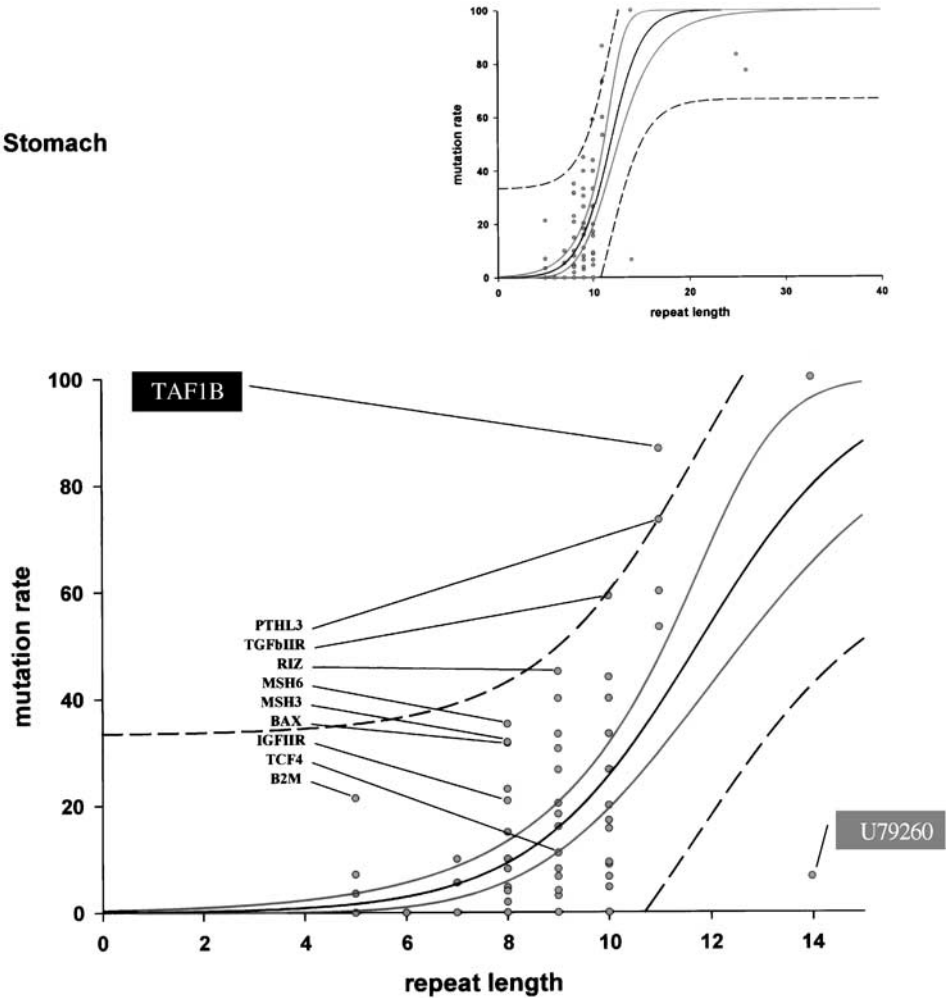

Figure 3 Regression analysis of ncMS and cMS for MSI-H gastric cancer. For details, see legend of Figure 2

(10^{-10} mutations per base pair replicated per generation (Kunkel and Bebenek, 2000)); accordingly, an approximate value of zero was used in our model. Second, for a repeat tract with a length of $n \rightarrow \infty$ the mutation rate in mismatch repair-deficient cells would reach 100%. Interestingly, cMS sequences exceeding 14 repetitive units have not been detected in the human genome (Woerner *et al*., 2001), presumably because of a limited MMR capacity in normal cells (Sia *et al*., 1997) and a high susceptibility of long repeats to somatic mutations. Third, the number of Real Common Target genes promoting MSI tumorigenesis is expected to be low, implicating that the majority of cMS and ncMS mutations would not provide a growth advantage. In such presumably neutral microsatellites mutations should arise at a mutation rate only depending on repeat type and length (Zhang *et al*., 2001). The mutation frequencies of an expected small number of Real Common Target genes are unlikely to have a significant influence on the fit of the regression curve. Therefore, we intentionally did not weight statistically the single data on the individual repeat tracts proportionate to the total number of analysed samples, since some known Real Common Target genes like TGFβIIR and BAX have been investigated more thoroughly yielding higher numbers of samples analysed. Weighted data points would increase the influence of genes of interest and of the few Real Common Target genes on the course of the regression curve. In contrast, genes of less scientific interest or more recently published microsatellite harboring genes would have been represented inadequately. This would have altered regression plots and biased statistical analysis.

Recently, a different statistical approach to discriminate Real Common Target genes from randomly mutated ByStander genes was proposed by Duval *et al*. (2001). However, the microsatellite harboring genes analysed in this study have not been assorted to different groups of genes by clustering analyses, but by functional considerations on well-characterized genes. Subsequently, predictive values for the relevance of a gene in MSI-associated carcinogenesis have been calculated based on mutational analyses of these cMS-harboring genes using the same set of tumors. Assuming that the true number of Real Common Target genes is very small, a mathematical problem occurs: the larger the

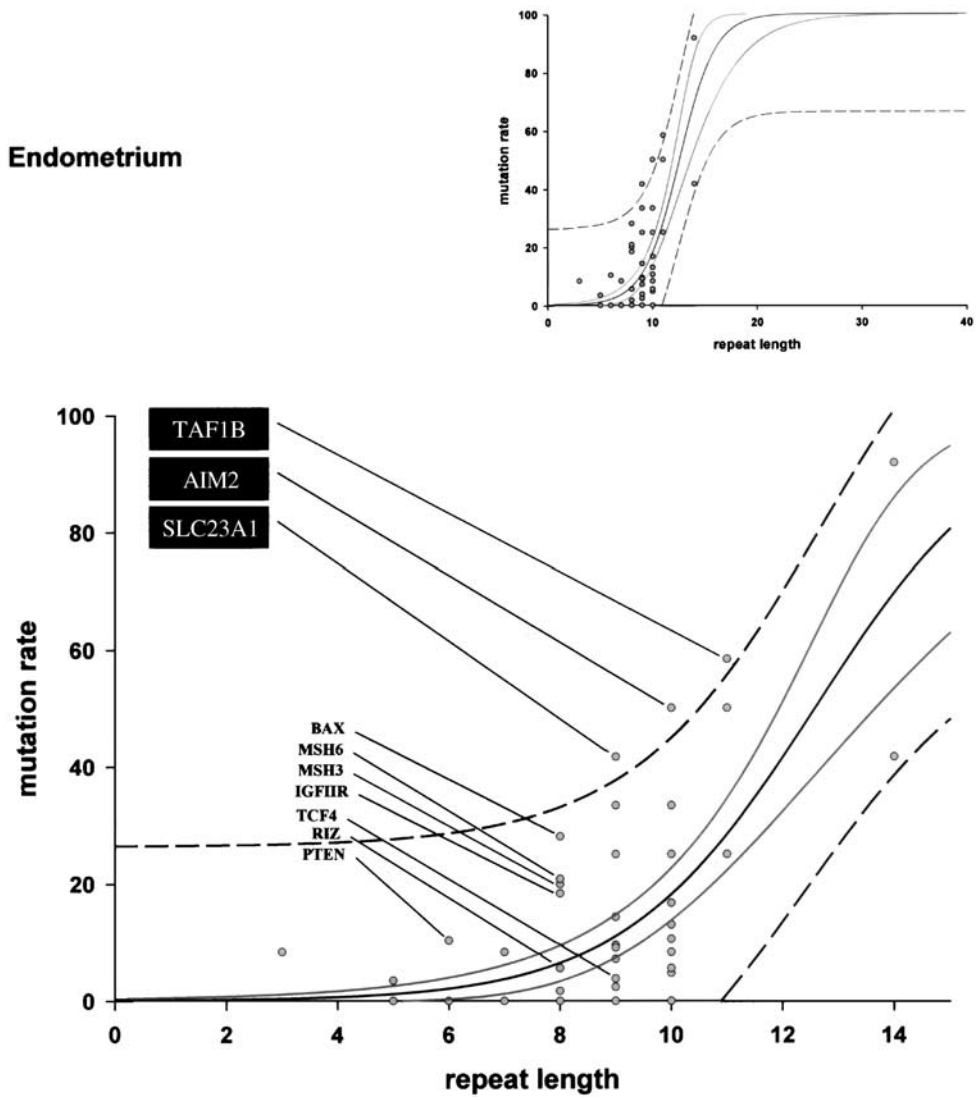

Figure 4 Regression analysis of ncMS and cMS for MSI-H endometrial cancer. For details, see legend of Figure 2

number of candidate genes analysed, the lower is the percentage of mutations in each tumor. Consequently, the discriminatory power of these predictive values decreases. Moreover, mutation data for each tumor for a given panel of candidate genes is essential for the calculation of the aforementioned predictive values, but the number of putative targets that can be analysed is limited by the amount of available tumor tissue. Hence, in contrast to our model, this approach seems rather inappropriate for the systematic identification of Real Common Target genes in MSI-H tumors if analysis of all microsatellite harboring genes is envisioned. Recently, a comparison based on the same statistical approach between MSI-H colorectal, gastric, and endometrial cancer has been published (Duval *et al*., 2002). The authors observed qualitative and quantitative differences in target gene mutation profiles among gastrointestinal and endometrial MSI-H cancers, similar to the tissue-specific predictions made by our approach.

Predictions made by our model

According to our model, nine genes with coding region MSI (TGFβRII, BAX, TCF-4, MSH3, ACVR2, PTHL3, HT001, AC1, and SLC23A1) were predicted to represent Real Common Targets in mismatch repair-deficient colon cancers. At least for the TGFβRII and BAX genes cMS frameshift mutations have been reported to confer a growth advantage to MSI-H colon tumor cells (Markowitz *et al*., 1995; Ionov *et al*., 2000), confirming the predictive power of our model. Such functional evidence is not shown at least for cMS mutations in ACVR2, PTHL3, HT001, AC1, and SLC23A1. TCF-4, involved in the wnt signal transduction pathway, is discussed to be a real target gene in MSI-H colorectal cancer (Duval *et al*., 1999). The MMR gene MSH3 was reported to play an important role by increasing the instability phenomenon characterizing these cancers as a result of the statistical approach done

by Duval *et al*. (2001). Published observations suggest that inactivation of the activin receptors (ACVR2 and others) is associated with tumorigenesis in the gastrointestinal tract (Liu *et al*., 2000). PTHL3, a secreted hormone involved in lactation and Ca-turnover, responsible for most cases of humoral hypercalcemia of malignancy, has been proposed as an antiproliferative factor (OMIM*168470).

However, for HT001, an unknown cDNA, expressed in hypothalamus and other tissues, as well as for AC1, an unknown cDNA differentially expressed in neuroblastoma, evidence for their association with tumorigenesis is lacking. Mutations in AC1 like those in TGFβRII (Myeroff *et al*., 1995) are more common in colorectal carcinomas, but rare in endometrial cancers with MSI indicating a positive selection pressure on mutated AC1 only in MSI-H colorectal cancer.

SLC23A1 (SVCT2) encodes a solute carrier protein associated with tissue-specific uptake of vitamin C thereby preventing cells from free radical damage. Knockout mice for the SVCT2 gene died shortly after birth as a result of respiratory failure and extensive bleeding in the brain. However, the role of SVCT2 in adulthood and scurvy remains unclear and requires additional investigation (Hediger, 2002; Sotiriou *et al*., 2002). Interestingly, SLC23A1 is mutated at a comparable mutation rate of about 40% in all three tumor entities presumably indicating a Real Common Target gene for all three organs.

In contrast to the high mutation frequencies of the predicted Real Common Target genes, the mutation rate of the A_{10} repeat tract in the RFC3 gene is nearly 0%, suggesting a strong negative selection pressure on cMS mutations in this gene. RFC3 is an important cofactor of DNA polymerases delta and epsilon in DNA replication and repair. Together with four other proteins, it forms a clamp loading complex proposed to organize the higher-order architecture of the replication machinery. Frameshift mutations in the A_{10} repeat of RFC3 would lead to a truncated protein unable to form stable RFC complexes (O'Donnell *et al*., 2001). These functional considerations might provide an explanation why RFC3 remains protected from cMS mutations during MSI carcinogenesis.

CHD2 (chromodomain helicase DNA binding protein 2) is a DNA helicase. Genes of this family are suggested to be involved in the regulation of gene expression and modification of chromatin structure. One might hypothesize that global changes in chromatin structure as a consequence of CHD2 mutations might exert pleiotropic effects and thus could be envisioned to be counterselected during MSI tumorigenesis.

Unexpectedly, two noncoding but transcribed microsatellites (a T_{10} in the 5′ UTR of WAF1 and an A_{10} in the 3′ UTR of BCL2) displayed mutation frequencies below the lower prediction limit arguing for counterselection. Sequence elements in 5′ or 3′ UTR regions are known to control the rate of synthesis, stability, and translational efficiency of mRNA (Mignone *et al*., 2002). Such control mechanisms could be responsible for the decreased mutation rate of these two UTR repeat tracts. A strong negative selection pressure on these two genes seems to be very plausible.

There are some genes whose mutation frequencies are different from the average mutation frequency, but still range inside the upper and lower prediction limits. Examples of this group include the Rb-interacting zinc-finger gene RIZ1, the gene for the insulin-like growth factor II receptor gene IGFIIR and Axin2 genes as members of the wnt signal transduction pathway, the PTEN gene encoding a dual specificity tyrosine phosphatase, and the mismatch repair gene MSH6. For some of these mutated gene products, functional support for promoting tumor cell growth has been provided (Souza *et al*., 1999; Chadwick *et al*., 2000; Liu *et al*., 2000; Piao *et al*., 2000; Sakurada *et al*., 2001).

The predictions made by our model are unbiased and do not rely on any functional assumptions. However, they can be nicely combined with the recently proposed classification of target genes into four different categories (Duval and Hamelin, 2002). Accordingly, the nine cMS-harboring genes TGFβRII, BAX, TCF-4, MSH3, ACVR2, PTHL3, HT001, AC1, and SLC23A1 can be assigned to the category of Transformators. In contrast, the exceedingly rare mutations in the cMS sequences of CHD2, RFC3 (MSI-H CRC), and U79260 (MSI-H GC), and in the ncMS sequences of WAF1 and BCL2 would mark these genes as examples of the Survivor category. At this point one should remind that genes localized outside the prediction lines are suspicious and therefore can be declared as Real Common Target genes or survivors, but none of the genes localized within the prediction interval are allowed to be assigned to any of these groups. These genes only show a statistically non significant mutation rate at the time of investigation. Thus, genes localized between upper and lower prediction lines cannot be assigned to either of the Cooperator or Hibernator categories proposed by the Duval model.

Limitations of the model and future work

Although our model is able to make predictions about Real Common Target and Survivor genes based upon positively and negatively selected mutations in microsatellite sequences, these predictions are specific for a single repeat and do not account for mutations in multiple repeats within the same gene. Reported mutation data on multiple repeats within genes like RIZ, AXIN2, or PRKDC, however, suggest that one repeat is preferentially mutated (Supplemental Tables 1–3). A further shortcoming of the model is the restriction to microsatellite mutation frequencies in single genes rather than cumulative mutation frequencies in different genes involved in the same pathway. In the case of a Real Target pathway, the mutation frequency of independent genes within such a pathway might be too low to be recognized as Real Common Target genes. A mutational analysis of entire pathways, that is, of all genes constituting one pathway is expected to result in high cumulative mutation frequencies for Real Target pathways and significantly lower mutation

frequencies for pathways not involved in MSI-associated carcinogenesis. The anticorrespondence of IGFIIR and TGFβIIR mutations (Souza *et al*., 1996) might reflect such a situation. Cluster analysis of all gene-specific mutation data of a given pathway would overcome the above-mentioned restriction, but requires mutation analysis of large numbers of cMS sequences in the same set of tumor samples often available only at limited amounts. Most likely, both theories – Real Common Target genes as well as Real Common Target pathways – even can coexist and overlap. Finally, our model is based on the analysis of mutation rates in samples, where multiple selection processes during the course of carcinogenesis have already manifested in MSI-associated carcinoma and thus, this analysis represents an end point view. Consequently, the Real Common Target genes predicted by our model, might rather reflect the selection pressure under which a cancer phenotype is maintained than the situation in premalignant cell clones during MSI-driven carcinogenesis.

Overall, our model is capable of detecting a large number of but probably not all, Real Common Target genes or Survivor genes. Future studies should try to determine the exact repeat mutation frequencies for a given gene in large numbers of tumor samples. Currently, such comprehensive cumulative information is only available for some genes (e.g. TGFβIIR), thereby increasing the reliability of single data points in the regression analysis. Additional studies on larger numbers of repeat harboring genes, as well as examination of multiple repeat tracts within a particular gene will reveal a true global view of repeat mutation frequencies. Especially the investigation of a large series of noncoding microsatellites mainly representing ByStander genes, accounting for the background mutation rate in MSI-H tumors, will further specify the model's predictions and, thus, will help to focus on genes of very high interest.

Materials and methods

Statistical Methods

Building data tables Publications reporting mutation frequencies of microsatellite genes were examined and data on investigated tumor samples as well as mutated microsatellites were sorted by gene and repeat type in a spreadsheet. Individual sample numbers and numbers of mutated samples from each single publication for each repeat tract within a gene were summarized. Subsequently, tissue-specific cumulative mutation rates were calculated. If no single data for individual repeats were available – a correlation of mutation rate and repeat length is not possible if data from at least two repeats were combined – or if cell lines and tumor samples were shown as one item, these data were omitted. In order to minimize sampling errors only repeat mutation data obtained from at least 10 tissue samples were considered for statistical analysis. A detailed listing of all mutation data together with corresponding references is available at http://www.med.uni-heidelberg.de/patho/pathomol/woerner/model_real_targets/as supplemental information.

Statistical analysis To model the dependency of the mutation rate y_i on the repeat length x_i in publication i a nonlinear regression model was chosen: $y_{ij} = f(x_i, \theta) + \varepsilon_{ij}$, where $f(x_i, \theta)$ describes the nonlinear relation between repeat length and mutation rate. The errors are assumed to be centered random variables, $E(\varepsilon_{ij}) = 0$, having homogeneous variance. We chose the logistic regression model since it can be seen as a direct probability model if the regression function is the logistic function $[1 + \exp(-x)]^{-1}$, which is restricted to range from 0 to1.

A general four-parameter logistic regression model that includes arbitrary lower and upper asymptotes is given by

$$f(x, \theta) = y_0 + \frac{\alpha}{1 + e^{-((x-\delta)/\beta)}}$$

with nonzero lower asymptote y_0. The upper asymptote $y_0 + \alpha$ represents the maximum mutation rate possible. Both parameters, y_0 and α, can be assumed to be 0 and 1, respectively, since the mutation rate of repeat tracts depends on repeat unit size increasing with repeat unit size (Sia *et al*., 1997) and can be assumed to be asymptotically equal to 1. The model now simplifies to

$$f(x, \theta) = [1 + \exp(-(x - \delta)/\beta)]^{-1}$$

with parameter vector $\theta = \binom{\beta}{\delta}$, which is estimated by the maximum likelihood method (ML). The fitted curve is skew-symmetric with an inflection point at $x = \delta$ and $f(x, \theta) = 1/2$. To determine the accuracy of the parameter estimates, $\hat{\theta}$, and estimates for functions of parameters, $\hat{\lambda} = \lambda(\hat{\theta})$, we calculated the Wald test and the likelihood ratio test (Huet *et al*., 1996). In addition, asymptotic 95% confidence and prediction intervals were computed. Owing to particularly small numbers of some single nucleotide groups for ncMS (n(A) = 17, n(C) = 0, n(G) = 5, and n(T) = 3) and cMS (n(A) = 115, n(C) = 20, n(G) = 12, and n(T) = 22) leading to very unequal groups, all four nucleotides were combined for ncMS and cMS analysis each. Based on the assumption that the majority of mutations affecting noncoding as well as coding microsatellites are unlikely to play any tumorigenic role and thus are expected to show medial mutation frequencies, we also combined ncMS and cMS for logistic regression analysis.

Additional comparisons of mutation rates of microsatellite gene sequences between MSI-H colorectal, endometrial, and gastric carcinomas have been performed using Fisher's exact test. To account for multiple testing the *P*-values were adjusted according to the method of Bonferroni–Holm.

Data analysis and visualization was done using SigmaPlot 2001 for Windows Version 7.0 (SPSS Inc., Chicago, IL, USA) and S-Plus, Version 3.4 for Unix (Insightful Corporation, Cambridge, MA, USA) using the software library nls2 for nonlinear regression (Huet *et al*., 1996). For all analyses two-sided tests were used and the significance level α was set to 5%.

Mutation analysis A group of 15 genes containing cMS had been investigated previously for mutations in MSI-H colorectal tumors and cell lines (Woerner *et al*., 2001). These repeat tracts were analysed in the present study for mutations in MSI-H gastric and endometrial cancers. Nine additional genes harboring A_{11} repeats (TAF1B, MACS, HT001), A_{10} repeats (CHD2, UVRAG, TCF6L1, ABCF1, AIM2), and one G_9 repeat (ELAVL3) that had not yet been analysed for frameshift mutations in MSI-H colorectal, gastric, and endometrial tumors were examined in the present study. During preparation of this manuscript mutations in the A_{10} repeat of one of these candidate genes (AIM2) have been reported (Mori *et al*., 2001). These cMS genes were selected from our list of human coding microsatellites (Woerner *et al*.,

2001) because they represent the longest coding repeat tracts. Primer sequences are available from the authors upon request. The mutation status of the ncMS APdelta3 was assessed in a set of 32 MSI-H colorectal tumor samples using published primer sequences (Ionov *et al*., 1993).

Tumor samples and microsatellite analysis MSI-H primary tumors from the colon ($n=21$, Department of Surgery, University of Heidelberg, Germany), stomach ($n=15$, Department of Surgery, TU Munich, Germany), and endometrium ($n=12$; Institute of Pathology, Mannheim, Germany) were analysed. MSI classification was performed as previously described (Woerner *et al*., 2001) using the ICG-HNPCC microsatellite reference marker panel (Boland *et al*., 1998). Fragment analysis was carried out on ABI 310/3100 genetic analyzers (Applied Biosystems, Darmstadt, Germany) using the Genescan Analysis Software (Applied Biosystems, Darmstadt, Germany).

Acknowledgements
We thank J Lacroix for helpful discussion and critical reading of the manuscript, and G Dallenbach-Hellweg for kindly providing endometrium carcinoma tissues. Technical assistance is acknowledged to G Russel, S Bielau, and B Kuchenbuch. Financial support was obtained from the Deutsche Krebshilfe and from the Verein zur Foerderung der Krebsforschung in Deutschland e.V. This article is dedicated to Harald zur Hausen on the occasion of his retirement as head of the German Cancer Research Center.

References

Aaltonen LA, Peltomaki P, Leach FS, Sistonen P, Pylkkanen L, Mecklin JP, Jarvinen H, Powell SM, Jen J and Hamilton SR. (1993). *Science,* **260,** 812–816.

Bicknell DC, Kaklamanis L, Hampson R, Bodmer WF and Karran P. (1996). *Curr. Biol.,* **6,** 1695–1697.

Boland CR, Thibodeau SN, Hamilton SR, Sidransky D, Eshleman JR, Burt RW, Meltzer SJ, Rodriguez-Bigas MA, Fodde R, Ranzani GN and Srivastava S. (1998). *Cancer Res.,* **58,** 5248–5257.

Chadwick RB, Jiang GL, Bennington GA, Yuan B, Johnson CK, Stevens MW, Niemann TH, Peltomaki P, Huang S and de la Chapelle A. (2000). *Proc. Natl. Acad. Sci. USA,* **97,** 2662–2667.

Duval A, Gayet J, Zhou XP, Iacopetta B, Thomas G and Hamelin R. (1999). *Cancer Res.,* **59,** 4213–4215.

Duval A, Rolland S, Compoint A, Tubacher E, Iacopetta B, Thomas G and Hamelin R. (2001). *Hum. Mol. Genet.,* **10,** 513–518.

Duval A and Hamelin R. (2002). *Cancer Res.,* **62,** 2447–2454.

Duval A, Reperant M, Compoint A, Seruca R, Ranzani GN, Iacopetta B and Hamelin R. (2002). *Cancer Res.,* **62,** 1609–1612.

Hediger MA. (2002). *Nat. Med.,* **8,** 445–446.

Huet S, Bouvier A, Gruet M and Jolivet E. (1996). Statistical tools for nonlinear regression. A Practical Guide with δ-Plus Examples. Springer, New York.

Ionov Y, Peinado MA, Malkhosyan S, Shibata D and Perucho M. (1993). *Nature,* **363,** 558–561.

Ionov Y, Yamamoto H, Krajewski S, Reed JC and Perucho M. (2000). *Proc. Natl. Acad. Sci. USA,* **97,** 10872–10877.

Konishi M, Kikuchi-Yanoshita R, Tanaka K, Muraoka M, Onda A, Okumura Y, Kishi N, Iwama T, Mori T, Koike M, Ushio K, Chiba M, Nomizu S, Konishi F, Utsunomiya J and Miyaki M. (1996). *Gastroenterology,* **111,** 307–317.

Kunkel TA and Bebenek K. (2000). *Annu. Rev. Biochem.,* **69,** 497–529.

Lengauer C, Kinzler KW and Vogelstein B. (1997). *Nature*, **386,** 623–627.

Lengauer C, Kinzler KW and Vogelstein B. (1998). *Nature,* **396,** 643–649.

Liu W, Dong X, Mai M, Seelan RS, Taniguchi K, Krishnadath KK, Halling KC, Cunningham JM, Boardman LA, Qian C, Christensen E, Schmidt SS, Roche PC, Smith DI and Thibodeau SN. (2000). *Nat. Genet.,* **26,** 146–147.

Loeb LA. (2001). *Cancer Res.,* **61,** 3230–3239.

Malkhosyan S, Rampino N, Yamamoto H and Perucho M. (1996). *Nature,* **382,** 499–500.

Markowitz S, Wang J, Myeroff L, Parsons R, Sun L, Lutterbaugh J, Fan RS, Zborowska E, Kinzler KW and Vogelstein B. (1995). *Science,* **268,** 1336–1338.

Mignone F, Gissi C, Liuni, S and Pesole G. (2002). *Genome Biol.,* **3,** REVIEWS0004, 1–10.

Mori Y, Yin J, Rashid A, Leggett BA, Young J, Simms L, Kuehl PM, Langenberg P, Meltzer SJ and Stine OC. (2001). *Cancer Res.,* **61,** 6046–6049.

Myeroff LL, Parsons R, Kim SJ, Hedrick L, Cho KR, Orth K, Mathis M, Kinzler KW, Lutterbaugh J and Park K. (1995). *Cancer Res.,* **55,** 5545–5547.

O'Donnell M, Jeruzalmi D and Kuriyan J. (2001). *Curr. Biol.,* **11,** R935–R946.

Perucho M, Peinado MA, Ionov Y, Casares S, Malkhosyan S and Stanbridge E. (1994). *Cold Spring Harb. Symp. Quant. Biol.,* **59,** 339–348.

Perucho M. (1999). *Cancer Res.,* **59,** 249–256.

Piao Z, Fang W, Malkhosyan S, Kim H, Horii A, Perucho M and Huang S. (2000). *Cancer Res.,* **60,** 4701–4704.

Rampino N, Yamamoto H, Ionov Y, Li Y, Sawai H, Reed JC and Perucho M. (1997). *Science,* **275,** 967–969.

Sagher D, Hsu A and Strauss B. (1999). *Mutat. Res.,* **423,** 73–77.

Sakurada K, Furukawa T, Kato Y, Kayama T, Huang S and Horii A. (2001). *Genes Chromosomes Cancer,* **30,** 207–211.

Schwartz Jr S, Yamamoto H, Navarro M, Maestro M, Reventos J and Perucho M. (1999). *Cancer Res.,* **59,** 2995–3002.

Sia EA, Kokoska RJ, Dominska M, Greenwell P and Petes TD. (1997). *Mol. Cell. Biol.,* **17,** 2851–2858.

Sotiriou S, Gispert S, Cheng J, Wang Y, Chen A, Hoogstraten-Miller S, Miller GF, Kwon O, Levine M, Guttentag SH and Nussbaum RL. (2002). *Nat. Med.,* **8,** 514–517.

Souza RF, Appel R, Yin J, Wang S, Smolinski KN, Abraham JM, Zou TT, Shi YQ, Lei J, Cottrell J, Cymes K, Biden K, Simms L, Leggett B, Lynch PM, Frazier M, Powell SM, Harpaz N, Sugimura H, Young J and Meltzer SJ. (1996). *Nat. Genet.,* **14,** 255–257.

Souza RF, Wang S, Thakar M, Smolinski KN, Yin J, Zou TT, Kong D, Abraham JM, Toretsky JA and Meltzer SJ. (1999). *Oncogene*, **18,** 4063–4068.

Thibodeau SN, Bren G and Schaid D. (1993). *Science,* **260,** 816–819.

Woerner SM, Gebert J, Yuan YP, Sutter C, Ridder R, Bork P and von Knebel Doeberitz M. (2001). *Int. J. Cancer,* **93,** 12–19.

Zhang L, Yu J, Willson JK, Markowitz SD, Kinzler KW and Vogelstein B. (2001). *Cancer Res.,* **61,** 3801–3805.

References of Supplemental Tables

(available at http://www.med.uni-heidelberg.de/patho/pathomol/woerner/model_real_targets/)