Published in final edited form as:

Clin Cancer Res. 2008 December 15; 14(24): 8061–8069. doi:10.1158/1078-0432.CCR-08-1431.

DNA copy-number alterations underlie gene expression differences between microsatellite stable and unstable colorectal cancers

Robert N. Jorissen¹, Lara Lipton¹, Peter Gibbs¹, Matthew Chapman¹, Jayesh Desai¹, Ian T. Jones², Timothy J. Yeatman³, Philip East⁴, Ian P.M. Tomlinson^{5,6}, Hein W. Verspaget⁷, Lauri A. Aaltonen⁸, Mogens Kruhøffer⁹, Torben F. Ørntoft⁹, Claus Lindbjerg Andersen⁹, and Oliver M. Sieber¹

¹ Ludwig Colon Cancer Initiative Laboratory, Ludwig Institute for Cancer Research, Royal Melbourne Hospital, Parkville, Australia ² Department of Colorectal Surgery, Royal Melbourne Hospital, Parkville, Australia ³ Department of Surgery, H. Lee Moffitt Cancer Center and Research Institute, Tampa, Florida ⁴ Computational Genome Analysis Laboratory, London Research Institute, Cancer Research UK, London, UK ⁵ Molecular and Population Genetics Laboratory, London Research Institute, Cancer Research UK, London, UK ⁶ Institute of Cancer, Bart's and the London Medical School, Queen Mary College, University of London, London, UK ⁷ Department of Gastroenterology and Hepatology, Leiden University Medical Centre, Netherlands ⁸ Department of Medical Genetics, Biomedicum, University of Helsinki, Finland ⁹ Molecular Diagnostic Laboratory, Department of Clinical Biochemistry, Aarhus University Hospital, Skejby, Denmark

Abstract

Purpose—About 15% of colorectal cancers (CRCs) harbor microsatellite instability (MSI). MSI-associated gene expression changes have been identified in CRCs, but little overlap exists between signatures hindering an assessment of overall consistency. Little is known about the causes and downstream effects of differential gene expression.

Experimental Design—DNA microarray data on 89 MSI and 140 MSS CRCs from this study, and 58 MSI and 77 MSS cases from three published reports were randomly divided into test and training sets. MSI-associated gene expression changes were assessed for cross-study consistency using training samples, and validated as MSI classifier using test samples. Differences in biological pathways were identified by functional category analysis. Causation of differential gene expression was investigated by comparison to DNA copy-number data.

Results—MSI-associated gene expression changes in CRCs were found to be highly consistent across multiple studies of primary tumors and cancer cell lines from patients of different ethnicities

Request for reprints: Oliver Sieber, Ludwig Colon Cancer Initiative Laboratory, Ludwig Institute for Cancer Research, PO, Box 2008, Royal Melbourne Hospital, VIC, 3050, Australia, Ph.: +61 (0) 3 93413168, Fax: +61 (0) 3 93413104, E-mail: Oliver.Sieber@ludwig.edu.au.

Statement of Clinical Relevance

This study of MSI-associated gene expression changes in colorectal cancer (CRC) demonstrates the comparability and reproducibility of microarray data produced in different laboratories. This is an important issue, as cross-study consistency is key if gene expression-based classifiers are to be used in clinical practice. Consistent MSI-associated genes were successfully used for classification of additional CRC samples. Although gene expression-based MSI classification is unlikely to enter clinical use in isolation – given the ease of PCR-based MSI typing – it may find future application when combined with other signatures in a single assay. Our results improve our understanding of the MSI and MSS pathways of tumorigenesis, by demonstrating that DNA copy-number aberrations partly underlie MSI-associated gene expression changes. Genes associated with immune response were up-regulated in MSI cancers, whereas genes associated with cell-cell adhesion, ion-binding and regulation of metabolism were down-regulated. These candidate genes provide a good starting point for future study.

($P < 0.001$). Clustering based on consistent changes separated additional test cases by MSI status, and classification of individual samples predicted MSI status with a sensitivity of 96% and specificity of 85%. Genes associated with immune response were up-regulated in MSI cancers, whereas genes associated with cell-cell adhesion, ion-binding and regulation of metabolism were down-regulated. Differential gene expression was shown to reflect systematic differences in DNA copy-number aberrations between MSI and MSS tumors ($P < 0.001$).

Conclusions—Our results demonstrate cross-study consistency of MSI-associated gene expression changes in CRCs. DNA copy-number alterations partly cause the differences in gene expression between MSI and MSS cancers.

Keywords

colorectal cancer; microsatellite instability (MSI); gene expression; DNA copy-number alterations; cross-platform microarray analysis

Introduction

Gene expression profiling using DNA microarrays has been successfully applied in numerous studies of tumor classification, and is gradually being introduced into clinical practice (1). However, the comparability and reproducibility of microarray data produced in different laboratories continues to be debated (2). As more data become available, systematic comparisons of studies with similar research goals are therefore gaining high importance.

Colorectal cancer (CRC) is one of the most common malignancies and the second most common cause of cancer death in the western world (3). About 15% of sporadic CRCs exhibit microsatellite instability (MSI) caused by mutation or epigenetic silencing of DNA mismatch repair genes (4). MSI CRCs have characteristic clinical features including right-sided location in the colon, mucinous histology, poor differentiation and pronounced lymphocyte infiltration (5). In addition, MSI cancers tend to have a better prognosis than microsatellite stable (MSS) cancers (6). The majority of MSI tumors have a near-diploid karyotype and appear to follow a genetic pathway distinct from MSS tumors (7). For example, MSI cancers accumulate mutations at repeat sequences of genes including *TGFBR2* (8), *IGFR2* (9), *BAX* (10) and *E2F4* (11) which are rarely mutated in MSS cancers.

Both oligonucleotide and cDNA microarrays have been used to characterize gene expression profiles in MSI and MSS CRCs and in CRC cell lines. Mori *et al* (12) and Koinuma *et al* (13) found that MSI had a great impact on the global transcriptome, and Banerjee *et al* (14) identified a gene expression cluster in MSI tumors that correlated with an activated immune response. Kruhøffer *et al* (15) constructed a gene expression classifier that could identify sporadic MSI cancers as well as HNPCC cancers, and Watanabe *et al* (16) reported signatures that could predict MSI status and differentiate distal from proximal MSI cancers. Giacomini *et al* (17) constructed an MSI classifier in CRC cell lines that also predicted MSI status in primary CRCs and gastric tumors. Although these studies provide good evidence that MSI-associated gene expression changes do exist, there is little overlap between reported signatures which hinders an assessment of overall consistency.

In this study, we analyzed global gene expression in 89 MSI and 140 MSS primary CRCs. In addition, we retrieved microarray data on 58 MSI and 77 MSS cases from three published reports (13,16,17) for which complete results were deposited in the Gene Expression Omnibus database.¹⁰ The data under investigation were from primary CRCs and CRC cell lines derived

¹⁰<http://www.ncbi.nlm.nih.gov/geo/>

from different ethnic populations (European and Japanese) which had been using various microarray platforms. Our aims were to determine the extent of consistency of MSI-associated gene expression changes across independent studies, and to evaluate consistent changes as a MSI classifier in additional test samples. Furthermore, we wished to establish the underlying causes and downstream effects of differential gene expression to improve our understanding of the MSI and MSS pathways of tumorigenesis.

Materials and Methods

Colorectal cancer specimens

74 fresh-frozen CRCs were retrieved from the tissue bank of the Royal Melbourne Hospital in Melbourne, Australia. The study was approved by the hospital ethics committee, and all patients gave informed consent prior to surgery. The samples consisted of 6 Dukes stage A, 23 Dukes B, 30 Dukes C and 15 Dukes D cancers, 48 of which were localized to the colon, 3 to the colorectal junction and 23 to the rectum. Median patient age at surgery was 69 years (range 30 to 92 years). Tumor DNA was extracted from cancer tissue containing >75% tumor cells as judged by histological assessment. Control DNA was extracted from blood or from normal tissue derived from the resection margin. Microsatellite instability (MSI) status was determined using the Bethesda microsatellite panel (BAT25, BAT26, D2S123, D5S346 and D17S250) (18). MSI was scored as present, if instability was seen at two or more markers; 11 CRCs were MSI and 63 were MSS. Total RNA was extracted from cancer tissue using Trizol reagent (Invitrogen). The total RNA was labeled and hybridized to HG-U133Plus2.0 GeneChip arrays (Affymetrix) according to the manufacturer's instructions at the H. Lee Moffitt Cancer Center. The probe sets on this array represent over 47,000 transcripts.

Additional total RNA from 78 MSI and 77 MSS CRCs was collected as part of a retrospective international study involving 8 different centers in Denmark, the Netherlands and Finland. Informed consent was obtained from all patients according to local ethics regulations. The samples consisted of 9 Dukes stage A, 127 Dukes B, 11 Dukes C and 8 Dukes D cancers, 122 of which were localized to the colon, 3 to the colorectal junction and 30 to the rectum. Median patient age at surgery was 69.5 years (range 28 to 88 years). The total RNA was labeled and hybridized to HG-U133Plus2.0 GeneChip arrays (Affymetrix) at Aarhus University Hospital.

Further gene expression data were retrieved from two independent studies of sporadic CRCs with known MSI status. The first cohort consisted of 33 MSI and 51 MSS tumors which had been analyzed by Watanabe *et al* using HG-U133Plus2.0 GeneChip arrays (Affymetrix) (GSE4554) (16), and the second cohort of 10 MSI and 10 MSS tumors which had been analyzed by Koinuma *et al* using the HG-U133A and HG-U133B Gene Chip arrays (Affymetrix) (GSE2138) (13).

For each dataset, MAS5.0-calculated signal intensities were normalized using the quantile normalization procedure implemented in robust multiarray analysis (RMA) (19,20), and the normalized data were log transformed (base 2). Filtering was performed to exclude probe sets which were not expressed or probe sets which showed a low variability across samples. Expression values were required to be above the median of all expression measurements in at least 25% of samples, and the interquartile range (IQR) across the samples on the log base 2 scale was required to be at least 0.5. The statistical software package R was used for all subsequent statistical analyses.¹¹

¹¹<http://www.r-project.org/>

Colorectal cancer cell lines

Two-color cDNA microarray (Stanford Functional Genomics Facility) data on 15 MSI and 16 MSS CRC cell lines were retrieved from a study by Giacomini *et al* (GSE2591) (5). These arrays comprised 39,632 different human IMAGE clones, representing over 21,000 transcripts. Expression values were provided as log transformed (base 2) intensity ratios of test sample against a reference sample constituted from 11 human cell lines (17). To enable comparison across platforms, Affymetrix probe IDs corresponding to IMAGE clones were retrieved by matching Entrez identifiers, GenBank accession numbers and gene symbols. Further matching was performed using MatchMiner (6).

Array-based comparative genomic hybridization (CGH) data were available from a study by Douglas *et al* (21) for 10 MSI and 13 MSS CRC cell lines analyzed by Giancomini *et al*. CGH arrays consisted of 3,452 bacterial artificial chromosome (BAC) clones that covered the human genome at an average spacing of about 1 megabase (Mb). The threshold for scoring DNA copy number gain or loss had been defined as $\log_2$ tumor:normal ratio >0.2 or <-0.2 (21).

Assessment of consistency of MSI-associated gene expression changes

Genes differentially expressed between MSI and MSS cases were identified using the Wilcoxon rank sum test and a P -value of <0.05 for the following training sample sets: 48 MSI and 47 MSS CRCs randomly selected from the samples analyzed at Aarhus University Hospital, 24 MSI and 36 MSS CRCs randomly selected from Watanabe *et al*, all 10 MSI and 10 MSS CRCs reported by Koinuma *et al*, and 10 MSI and 10 MSS CRC cell lines randomly selected from Giacomini *et al* (Supplementary Table 1). Separate lists were generated for genes significantly up- or down-regulated in MSI cancers for each comparison; genes mapping to sex chromosomes were excluded as cases were not matched by gender. For significant genes repeatedly identified between cohorts, consistency of up- or down-regulation was assessed using the chi-squared test.

Evaluation of consistent MSI-associated genes as a MSI classifier

Consistent MSI-associated genes were evaluated as classifiers of MSI status using independent test sample sets: Additional 30 MSI and 30 MSS primary CRCs analyzed at Aarhus University Hospital, 9 MSI and 15 MSS primary CRCs from Watanabe *et al*, and 5 MSI and 6 MSS CRC cell lines from Giacomini *et al*. Furthermore, all 11 MSI and 63 MSS CRCs from the Royal Melbourne Hospital were evaluated (Supplementary Table 1). Primary CRCs, which had been analyzed using oligonucleotide microarrays, and CRC cell lines, which had been analyzed using cDNA microarrays, were evaluated separately. For primary CRCs, quantile normalization was performed across studies. Expression values of MSI associated genes were mean-centered and scaled, followed by divisive hierarchical clustering using pair distances calculated as one minus the Spearman correlation coefficient as distance metric. The distribution of MSI and MSS cases within the two main branches of the resulting dendrogram was assessed for significance using the chi-squared test or Fisher's exact test.

Single-sample MSI classification against a common reference set

Single-sample MSI classification was performed by scoring individual test samples against a common reference set. Reference samples were selected using divisive hierarchical clustering from the MSI and MSS cases initially used for the identification of MSI-associated genes; CRC samples from Koinuma *et al* were omitted, as expression data were derived from two separate microarray platforms (HG-U133A and HG-U133B) thus not permitting reliable integration of results. Only cases which 'correctly' clustered into MSI or MSS branches were included in the reference set (Supplementary Table 1). Individual test samples were added to the reference set

and quantile normalization was performed for Affymetrix data followed by joint divisive hierarchical clustering as described above.

Functional category analysis of MSI-associated genes

Gene Ontology categories were analyzed using the Functional Annotation Clustering tool on the Database for Annotation, Visualization and Integrated Discovery (DAVID).¹² Genes were classified according to their annotated role in biological process, molecular function, and cellular components from Gene Ontology (The Gene Ontology Consortium). Category enrichment was tested against all human genes. *P*-values were adjusted using the Benjamini and Hochberg False Discovery Rate multiple testing correction.

Results

Consistency of MSI-associated gene expression changes across independent studies of CRCs

Consistency of MSI-associated gene expression changes in primary CRCs was assessed using oligonucleotide microarray data from three independent studies representing European and Japanese patients: 48 MSI and 47 MSS cases analyzed at Aarhus University Hospital, 24 MSI and 36 MSS cases from Watanabe *et al* (16), and 10 MSI and 10 MSS cases from Koinuma *et al* (13) (Supplementary Table 1). For each training cohort, genes (probe sets) differentially expressed between MSI and MSS cases were identified, and separate lists were generated for up- and down-regulated genes. Genes overlapping between studies were assessed for consistency of up- or down-regulation in MSI cancers (Table 1). All pair-wise comparisons between studies were found to be significant ($P < 0.001$, chi-squared test), with 98.0% (6600 of 6732, Aarhus University Hospital *versus* Watanabe), 93.9% (1081 of 1151, Aarhus University Hospital *versus* Koinuma) and 95.1% (1006 of 1058, Watanabe *versus* Koinuma) of genes showing consistent changes in expression. A total of 829 genes were consistently up- (424 genes) or down- (405 genes) regulated in MSI cancers when all three datasets were combined (Supplementary Table 2).

Consistency of MSI-associated gene expression changes between primary CRCs and CRC cell lines

Similar comparisons were performed to assess the consistency of MSI-associated gene expression changes between primary CRCs and CRC cell lines. Data for the latter were 10 MSI and 10 MSS cases randomly selected from a two-color cDNA microarray study performed by Giacomini *et al* (17) (Supplementary Table 1). Again, all pair-wise comparisons between studies were significant ($P < 0.001$, chi-squared test), with 69.3% (1641 of 2367, Giacomini *versus* Aarhus University Hospital), 69.2% (1660 of 2398, Giacomini *versus* Watanabe) and 78.1% (339 of 434, Giacomini *versus* Koinuma) of genes showing consistent changes in expression (Table 2). These proportions were significantly lower than those seen in the pair-wise comparisons between studies of primary tumors ($P < 0.03$ for all comparisons, chi-squared test). This finding may be partly due to differences in gene expression between primary and cultured tumor cells, increased noise levels for two-color cDNA microarrays as compared to oligonucleotide microarrays due to the use of a total RNA reference, or non-specific hybridization of probe sets and IMAGE clones. A total of 192 genes (229 IMAGE clones) were consistently up- (93 genes/117 IMAGE clones) or down- (99 genes/112 IMAGE clones) regulated when all four data sets were combined (Supplementary Table 2).

¹²<http://david.abcc.ncifcrf.gov/home.jsp>

Consistent MSI-associated gene expression changes as a MSI classifier

The 829 and 192 gene sets found to be consistently up- or down-regulated in MSI cases across multiple studies of primary CRCs and across primary CRCs and CRC cell lines were assessed as MSI classifiers in independent test samples. These consisted of a separate set of additional primary cancers analyzed at Aarhus University Hospital (30 MSI and 30 MSS) and Watanabe *et al* (9 MSI and 15 MSS), as well as the 74 primary cancers from the Royal Melbourne Hospital (11 MSI and 63 MSS). Additional CRC cell lines were derived from Giacomini *et al* (5 MSI and 6 MSS) (Supplementary Table 1). Given that classification of a binary outcome, MSI or MSS, was desired, samples were clustered using divisive hierarchical clustering. Clustering was performed separately for primary CRC samples, which had been analyzed using oligonucleotide microarrays, and CRC cell lines, which had been analyzed using cDNA microarrays.

For primary CRCs, clustering using either the 829 or 192 gene set produced two main branches comprising predominantly MSI or MSS cases, respectively (Figure 1). For the 829 gene set, one branch contained 48 MSI and 14 MSS cases, the other 2 MSI and 94 MSS cases ($P < 0.001$, chi-squared test); for the 192 gene set, one branch contained 49 MSI and 11 MSS cases, the other 1 MSI and 97 MSS cases ($P < 0.001$, chi-squared test). Clustering by study was also evident in both dendrograms, but this was secondary to the MSI and MSS branches (Figure 1). For CRC cell lines, clustering using the 192 (229 IMAGE clone) gene set separated all 5 MSI and all 6 MSS cases into two main clusters ($P < 0.002$, Fisher's exact test) (Figure 1).

Single-sample MSI classification against a common reference set

In clinical practice, classification of individual patient samples is required. Given that divisive hierarchical clustering successfully clustered test samples from independent studies and from patients of different ethnicities by MSI status, we modified this approach to permit scoring of individual CRC cases against a common reference set.

Reference samples were selected from the primary CRCs of Aarhus University Hospital and Watanabe *et al*, and the CRC cell lines of Giacomini *et al* which had initially been used to identify MSI-associated gene expression changes (samples from Koinuma *et al* were excluded). Divisive hierarchical clustering was performed using the 829 or 192 gene set, and only samples 'correctly' segregating into MSI and MSS branches were chosen, resulting in reference sets of 67 MSI and 77 MSS primary CRCs for the 829 gene set, and 68 MSI and 77 MSS primary CRCs and 10 MSI and 10 MSS CRC cell lines for the 192 gene set (Supplementary Table 1).

For single-sample MSI classification, the test samples were added one at a time to the above reference set. Divisive hierarchical clustering was performed, and a score was given as to whether the test sample clustered within the MSI or the MSS branch of the resulting dendrogram (Table 3). In all cases, the reference samples in the smallest branch containing the test and at least five other reference samples were either all MSI or all MSS cases. Taking PCR-based MSI typing as the gold standard, classification of primary CRCs using the 829 gene set had an overall sensitivity of 96.0% (48 of 50) and specificity of 83.3% (90 of 108), with a positive predictive value 72.7% (48 of 66) and negative predictive value of 97.8% (90 of 92). Classification using the 192 gene set appeared to slightly increase performance, with an overall sensitivity of 96.0% (48 of 50), specificity of 88.9% (96 of 108), positive predictive value of 80.0% (48 of 60) and negative predictive value of 98.0% (98 of 100). Importantly, classification using either the 829 or 192 gene set showed similar sensitivity and specificity for test samples from Aarhus University Hospital/Watanabe *et al* (sensitivity 94.9% (37 of 39) and 94.9% (37 of 39), specificity 86.7% (39 of 45) and 88.9% (40 of 45), respectively) and the Royal Melbourne Hospital (sensitivity 100.0% (11 of 11) and 100.0% (11 of 11), specificity 81.0%

(51 of 63) and 88.9% (56 of 63), respectively), despite samples from the former studies, but not from the latter, contributing to the reference set. These results demonstrate the potential utility of this approach for single-sample MSI classification irrespective of study origin, and further show the reproducibility and comparability of microarray data produced in different laboratories.

Similar results were obtained for CRC cell lines using the 192 gene (229 IMAGE clone) set, suggesting that this classification approach can also be applied to two-color cDNA microarray data provided that the same total RNA reference is being used. The sensitivity of MSI prediction was 100.0% (5 of 5) and specificity 83.3% (5 of 6).

Functional category analysis of discriminating genes between MSI and MSS cancers

For the 829 gene set, functional category analysis identified five significant annotation clusters, correlating with cell-cell adhesion, ion-binding, positive and negative regulation of metabolism and immune response (Supplementary Table 3). When the 829 gene set was separated into genes showing up- and down-regulation in MSI cancers, the immune response cluster was specifically up-regulated in the MSI cancer group. In contrast, the cell-cell adhesion, ion-binding, positive and negative regulation of metabolism clusters were specifically associated with the down-regulated genes. Functional category analysis of the smaller 192 gene set did not reveal any significant associations, but down-regulated genes from the cell-cell adhesion, ion-binding, positive and negative regulation of metabolism clusters were represented at the expected ratios (observed 33/116; expected 99/405; $P=0.58$, chi-squared test). In contrast, there were significantly fewer up-regulated genes from the immune response cluster than expected (observed 10/105, expected 93/424; $P<0.001$, chi-squared test), consistent with an absence of such a response in cell culture. A total of 20 MSI-associated genes were identified by at least two Affymetrix probe sets and one IMAGE clone, and may therefore be regarded as good candidates for further study (Table 4).

Molecular basis of MSI-associated gene expression changes

MSI CRCs generally have near-diploid karyotypes, whereas MSS tumors tend to be aneuploid (22). In addition, MSS CRCs and CRC cell lines show particularly high frequencies of DNA copy-number changes at certain chromosomal regions, such as loss of chromosome 17p and 18q or gain of chromosome 13 and 20q (23). Previous gene expression studies on primary cancers have shown that chromosomal losses and gains are associated with corresponding changes in gene expression (24,25). We therefore hypothesized that systematic differences in the frequencies of DNA copy-number changes between MSI and MSS tumors might underlie the MSI-associated gene expression changes in CRCs.

Array-based CGH data were available from a previous study by Douglas *et al* (21) for 10 MSI and 13 MSS CRC cell lines analyzed for gene expression by Giancomini *et al*. The CGH data were used to determine frequencies of DNA copy number changes for MSI and MSS cases across the genome, measured as the fraction of cases gained or lost at each BAC clone represented on the array. Differential frequencies between MSS and MSI cases were determined by subtracting frequencies in MSI from those in MSS cancers, and plotted against BAC chromosome position (Figure 2 A–B). Similarly for gene expression data, frequencies of genes significantly up- or down-regulated in MSS cases were determined for 5Mb windows spaced at 1Mb intervals across the genome. Differential frequencies between MSS and MSI cases were obtained by subtracting frequencies of down-regulated genes from those of up-regulated genes, and plotted against chromosome position (Figure 2 A–B).

There was good evidence that gene expression changes at least partly reflected differences in frequencies of DNA-copy number alterations between MSI and MSS CRC cell lines. At

chromosomal regions for which MSS cases showed high frequencies of loss as compared to MSI cases, genes tended to show reduced levels of expression, whereas at chromosomal regions for which MSS cases showed high frequencies of gain as compared to MSI cases, genes tended to show increased levels of expression ($r = 0.66$; $P < 0.001$, Pearson's product-moment correlation test). These data suggest that DNA copy number changes have profound effects on gene expression *in vivo*. Overall, 44.4% (1345 of 3028) of up-regulated and 72.1% (1110 of 1539) of down-regulated genes showed association with corresponding DNA copy-number changes ($P < 0.001$, chi-squared test).

We then analyzed MSI-associated gene expression changes in primary CRCs analyzed at Aarhus University Hospital, the Royal Melbourne Hospital, from Watanabe *et al* and Koinuma *et al* for evidence of causation by underlying DNA copy-number changes (Figure 2 C–G). As DNA copy-number data were not available for these tumors, alternative array-CGH data were retrieved for 7 MSI and 102 MSS tumors published by Nakao *et al* (26). Again, there was good evidence that differential frequencies in gene expression between MSS and MSI cases strongly resemble systematic frequencies in DNA copy-number changes. The Pearson's product-moment correlation coefficients for DNA copy-number frequencies against gene expression frequencies were 0.72 for Aarhus University Hospital, 0.69 for Royal Melbourne Hospital, 0.71 for Watanabe *et al* and 0.41 for Koinuma *et al* ($P < 0.001$ for all comparisons).

Discussion

We have found a high level of consistency of MSI-associated gene expression changes across independent studies of CRCs from different ethnic populations. This finding demonstrates that despite differences in genetic background, limited study sizes and the use of various analysis protocols, consistent changes in gene expression can be readily identified. A high level of consistency of MSI-associated gene expression changes was also found when comparing primary CRCs and CRC cell lines, despite the former having been run on oligonucleotide microarrays and the latter on two-color cDNA microarrays. This concordance across platforms suggests that gene expression patterns in CRC cell lines broadly reflect those in primary tumors. Furthermore, divisive hierarchical clustering based on consistent gene sets successfully separated additional primary CRCs from multiple studies into MSI and MSS cases, further demonstrating the cross-laboratory reproducibility of this gene expression signature.

Single-sample classification of additional CRC samples from multiple studies was successfully achieved by divisive analysis clustering against a common MSI/MSS reference set. Compared to PCR-based MSI typing and irrespective of study origin of the test sample, the sensitivity and specificity of this approach were high, being approximately 96% and 85%, respectively. Although unlikely to be applied for classification of MSI status in clinical practice unless combined with other signatures – given that current PCR-based MSI typing is technically less demanding and more cost-effective than DNA microarray analysis – this approach may provide a more general avenue for single-sample classification using gene expression signatures.

About 4% of MSI and 12% of MSS cases were not correctly classified using consistent sets of MSI-associated genes. This probably partly reflected experimental noise inherent to gene expression data from primary tissues (for example, due to the presence of contaminating normal cells) and partly underlying genetic heterogeneity between cancers. For example, it has been reported that a small proportion of MSI cancers show aneuploid rather than near-diploid karyotypes, similar to MSS cancers (23). Conversely, a subset of MSS tumors appears to harbor only a few chromosomal changes. Given the observed association between MSI-associated gene expression changes and DNA copy-number, this variation may account for some of misclassifications. Furthermore, there is the possibility that some misclassified MSI samples were not from sporadic cases, but instead derived from patients with hereditary non-polyposis

colorectal cancer (HNPCC). The latter tumors may follow a pathway of tumorigenesis distinct from sporadic MSI tumors (27). However, there was no evidence for this from our two misclassified MSI patients from Aarhus University Hospital, both of which had presented late in life (66 and 55 years) and neither of whom had a family history of HNPCC-associated cancers.

Our finding that immune response genes are up-regulated genes in MSI cancers is consistent with a previous report by Banerjee *et al* (14) and with histopathological data showing that MSI cancers tend to have more pronounced lymphocyte infiltration than MSS cancers (28). Furthermore, immune response associated changes were not seen in the CRC cell lines. Although results from functional category analysis must be interpreted with caution, the novel observed down-regulation of genes involved in cell-cell adhesion, ion-binding and regulation of cellular processes in MSI cancers is intriguing, and perhaps accounts for some differences in tumor behavior between MSI and MSS cases. Notably, 20 MSI-associated genes were identified by at least two Affymetrix probe sets and one IMAGE clone, and may therefore be regarded as good candidates for further study. These genes include *ZNF148*, which, when overexpressed, has been shown to suppress adenoma growth in multiple intestinal neoplasia (*Apc^{Min}*) mice, a widely used model of intestinal tumorigenesis (29). Loss of another candidate gene, *TFAP2A*, has been shown to deregulate E-cadherin and MMP-9 and to increase tumorigenicity of colon cancer cells *in vivo* (30).

The comparison of array-CGH and gene expression microarray data for primary CRCs and CRC cell lines showed that MSI-associated gene expression changes broadly reflect systematic DNA copy-number differences between MSI tumors, which tend to be near-diploid, and MSS tumors, which tend to be aneuploid. These data demonstrate that DNA copy number changes in cancer cells have profound effects on gene expression and therefore the potential to impact on tumor cell behavior and phenotype. Taken together, our results suggest that this mechanism contributes to the clinical differences between MSI and MSS tumors.

In conclusion, we found that MSI-associated gene expression changes were highly consistent across multiple independent studies of CRCs. Consistency was observed across different ethnic populations and between primary CRCs and cancer cell lines. Consistent MSI-associated genes were successfully used to predict MSI status of additional individual CRC samples with high sensitivity and specificity. Together, these results suggest that microarray data are broadly comparable and reproducible across different laboratories. Our study provides novel insights into the MSI and MSS pathways of tumorigenesis, by demonstrating that DNA copy-number aberrations at least partly underlie MSI-associated gene expression changes. Genes associated with immune response were found to be up-regulated in MSI cancers, whereas genes associated with cell-cell adhesion, ion-binding and regulation of metabolism were found to be down-regulated. The candidate genes identified provide a starting point for further study of MSI and MSS cancers, and may ultimately elucidate the clinical differences between these two types of CRCs.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

We are grateful to the Australian Genome Research Facility (AGRF) for their excellent technical support and to the patients for participating in this study.

Financial Support: Lara Lipton is supported by a fellowship from the Victorian Cancer Agency.

References

1. Buyse M, Loi S, van't Veer L, et al. Validation and clinical utility of a 70-gene prognostic signature for women with node-negative breast cancer. *Journal of the National Cancer Institute* 2006;98:1183–92. [PubMed: 16954471]
2. Shi L, Perkins RG, Fang H, Tong W. Reproducible and reliable microarray results through quality control: good laboratory proficiency and appropriate data analysis practices are essential. *Current opinion in biotechnology* 2008;19:10–8. [PubMed: 18155896]
3. Jemal A, Siegel R, Ward E, Murray T, Xu J, Thun MJ. Cancer statistics, 2007. *CA: a cancer journal for clinicians* 2007;57:43–66. [PubMed: 17237035]
4. Peltomaki P. Role of DNA mismatch repair defects in the pathogenesis of human cancer. *J Clin Oncol* 2003;21:1174–9. [PubMed: 12637487]
5. Lynch HT, Smyrk TC, Watson P, et al. Genetics, natural history, tumor spectrum, and pathology of hereditary nonpolyposis colorectal cancer: an updated review. *Gastroenterology* 1993;104:1535–49. [PubMed: 8482467]
6. Popat S, Hubner R, Houlston RS. Systematic review of microsatellite instability and colorectal cancer prognosis. *J Clin Oncol* 2005;23:609–18. [PubMed: 15659508]
7. Rowan A, Halford S, Gaasenbeek M, et al. Refining molecular analysis in the pathways of colorectal carcinogenesis. *Clin Gastroenterol Hepatol* 2005;3:1115–23. [PubMed: 16271343]
8. Markowitz S, Wang J, Myeroff L, et al. Inactivation of the type II TGF-beta receptor in colon cancer cells with microsatellite instability. *Science (New York, NY)* 1995;268:1336–8.
9. Souza RF, Appel R, Yin J, et al. Microsatellite instability in the insulin-like growth factor II receptor gene in gastrointestinal tumours. *Nature genetics* 1996;14:255–7. [PubMed: 8896552]
10. Rampino N, Yamamoto H, Ionov Y, et al. Somatic frameshift mutations in the BAX gene in colon cancers of the microsatellite mutator phenotype. *Science (New York, NY)* 1997;275:967–9.
11. Souza RF, Yin J, Smolinski KN, et al. Frequent mutation of the E2F-4 cell cycle gene in primary human gastrointestinal tumors. *Cancer research* 1997;57:2350–3. [PubMed: 9192806]
12. Mori Y, Selaru FM, Sato F, et al. The impact of microsatellite instability on the molecular phenotype of colorectal tumors. *Cancer research* 2003;63:4577–82. [PubMed: 12907634]
13. Koinuma K, Yamashita Y, Liu W, et al. Epigenetic silencing of AXIN2 in colorectal carcinoma with microsatellite instability. *Oncogene* 2006;25:139–46. [PubMed: 16247484]
14. Banerjee A, Ahmed S, Hands RE, et al. Colorectal cancers with microsatellite instability display mRNA expression signatures characteristic of increased immunogenicity. *Molecular cancer* 2004;3:21. [PubMed: 15298707]
15. Kruhoffer M, Jensen JL, Laiho P, et al. Gene expression signatures for colorectal cancer microsatellite status and HNPCC. *Br J Cancer* 2005;92:2240–8. [PubMed: 15956967]
16. Watanabe T, Kobunai T, Toda E, et al. Distal colorectal cancers with microsatellite instability (MSI) display distinct gene expression profiles that are different from proximal MSI cancers. *Cancer research* 2006;66:9804–8. [PubMed: 17047040]
17. Giacomini CP, Leung SY, Chen X, et al. A gene expression signature of genetic instability in colon cancer. *Cancer research* 2005;65:9200–5. [PubMed: 16230380]
18. Boland CR, Thibodeau SN, Hamilton SR, et al. A National Cancer Institute Workshop on Microsatellite Instability for cancer detection and familial predisposition: development of international criteria for the determination of microsatellite instability in colorectal cancer. *Cancer research* 1998;58:5248–57. [PubMed: 9823339]
19. Bolstad BM, Irizarry RA, Astrand M, Speed TP. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics (Oxford, England)* 2003;19:185–93.
20. Irizarry RA, Bolstad BM, Collin F, Cope LM, Hobbs B, Speed TP. Summaries of Affymetrix GeneChip probe level data. *Nucleic acids research* 2003;31:e15. [PubMed: 12582260]
21. Douglas EJ, Fiegler H, Rowan A, et al. Array comparative genomic hybridization analysis of colorectal cancer cell lines and primary carcinomas. *Cancer research* 2004;64:4817–25. [PubMed: 15256451]

22. Eshleman JR, Casey G, Kochera ME, et al. Chromosome number and structure both are markedly stable in RER colorectal cancers and are not destabilized by mutation of p53. *Oncogene* 1998;17:719–25. [PubMed: 9715273]
23. Jones AM, Douglas EJ, Halford SE, et al. Array-CGH analysis of microsatellite-stable, near-diploid bowel cancers and comparison with other types of colorectal carcinoma. *Oncogene* 2005;24:118–29. [PubMed: 15531920]
24. Cifola I, Spinelli R, Beltrame L, et al. Genome-wide screening of copy number alterations and LOH events in renal cell carcinomas and integration with gene expression profile. *Molecular cancer* 2008;7:6. [PubMed: 18194544]
25. Andersen CL, Wiuf C, Kruhoffer M, Korsgaard M, Laurberg S, Orntoft TF. Frequent occurrence of uniparental disomy in colorectal cancer. *Carcinogenesis* 2007;28:38–48. [PubMed: 16774939]
26. Nakao K, Mehta KR, Fridlyand J, et al. High-resolution analysis of DNA copy number alterations in colorectal cancer by array-based comparative genomic hybridization. *Carcinogenesis* 2004;25:1345–57. [PubMed: 15001537]
27. Johnson V, Volikos E, Halford SE, et al. Exon 3 beta-catenin mutations are specifically associated with colorectal carcinomas in hereditary non-polyposis colorectal cancer syndrome. *Gut* 2005;54:264–7. [PubMed: 15647192]
28. Young J, Simms LA, Biden KG, et al. Features of colorectal cancers with high-level microsatellite instability occurring in familial and sporadic settings: parallel pathways of tumorigenesis. *The American journal of pathology* 2001;159:2107–16. [PubMed: 11733361]
29. Law DJ, Labut EM, Merchant JL. Intestinal overexpression of ZNF148 suppresses ApcMin/+ neoplasia. *Mamm Genome* 2006;17:999–1004. [PubMed: 17019648]
30. Schwartz B, Melnikova VO, Tellez C, et al. Loss of AP-2alpha results in deregulation of E-cadherin and MMP-9 and an increase in tumorigenicity of colon cancer cells in vivo. *Oncogene* 2007;26:4049–58. [PubMed: 17224907]

Fig. 1. Divisive hierarchical clustering of test colorectal cancers and cancer cell lines using the 829 and 192 consistent MSI-associated genes

A and B, primary CRCs clustered using the 829 and 192 gene sets, respectively. C, CRC cell lines clustered using the 192 gene (229 IMAGE clone) set. Samples are arranged along the x-axis and genes along the y-axis. Each square represents the expression level of a given gene in an individual sample. Red represents increased expression and green represents decreased expression relative to the mean-centered and scaled expression of the gene across the samples following quantile normalization across studies. Genes are grouped into those down-regulated (top) and those up-regulated (bottom) in MSI cases. For the dendrogram, orange lines represent MSI cases, blue lines MSS cases. Test samples included additional 30 MSI and 30 MSS cancers analyzed at the Aarhus University Hospital, 9 MSI and 15 MSS cancers from Watanabe *et al* (16), 11 MSI and 63 MSS cancers from the Royal Melbourne Hospital, and 5 MSI and 6 MSS CRC cell lines from Giacomini *et al* (17).

Fig. 2. Comparison of DNA copy number and gene expression differences between matched MSI and MSS colorectal cancer cell lines and unmatched primary tumors

A and B, differential DNA copy-number and gene expression frequencies for 10 MSI and 13 MSS CRC cell lines; data from Douglas *et al.* (21) and Giacomini *et al.* (17). C, differential DNA copy-number frequencies for 7 MSI and 102 MSS primary CRCs from Nakao *et al.* (26). D, E, F and G, differential gene expression frequencies for 78 MSI and 77 MSS primary CRCs analyzed at Aarhus University Hospital, 11 MSI and 63 MSS primary CRCs analyzed at the Royal Melbourne Hospital, 33 MSI and 51 MSS primary CRCs from Watanabe *et al.* (16), and 10 MSI and 10 MSS primary CRCs from Koinuma *et al.* (13). For differential DNA copy number frequencies, lower bars represent losses or deletions, and the upper bars represent gains or amplifications. For differential gene expression frequencies, lower bars represent regions for which genes in MSS cases show predominant down-regulation, and the upper bars represent regions for which genes in MSS cases show predominant up-regulation. The dashed lines represent the location of the centromeres.

Table 1**Comparison of gene expression differences between MSI and MSS colorectal cancers across multiple studies**

Analysis was performed on 48 MSI and 47 MSS cases analyzed at Aarhus University Hospital, 24 MSI and 36 MSS cases from Watanabe *et al* (16), and 10 MSI and 10 MSS cases from Koinuma *et al* (13). For each cohort, genes (probe sets) differentially expressed between MSI and MSS cases were identified using the Wilcoxon rank sum test and a P-value of <0.05. For genes overlapping between cohorts, consistency of up- or down-regulation in MSI cancers was assessed using the chi-squared test.

Primary CRCs - MSI versus MSS			
Aarhus University Hospital			
Watanabe <i>et al</i>	Up	Down	P-value
Up	3448 (51.2%)	49 (0.7%)	<0.001
Down	83 (1.2%)	3152 (46.8%)	
Aarhus University Hospital			
Koinuma <i>et al</i>	Up	Down	
Up	553 (48.0%)	24 (2.1%)	<0.001
Down	46 (4.0%)	528 (45.9%)	
Watanabe <i>et al</i>			
Koinuma <i>et al</i>	Up	Down	
Up	528 (49.9%)	25 (2.4%)	<0.001
Down	27 (2.6%)	478 (45.2%)	

Table 2**Comparison of MSI-associated gene expression changes between primary colorectal cancers and cancer cell lines**

Analysis was performed on 48 MSI and 47 MSS CRCs analyzed at Aarhus University Hospital, 24 MSI and 36 MSS CRCs from Watanabe *et al* (16), 10 MSI and 10 MSS CRCs from Koinuma *et al* (13), and 10 MSI and 10 MSS CRC cell lines from Giacomini *et al* (17). For each cohort, genes (probe sets/IMAGE clones) differentially expressed between MSI and MSS cases were identified using the Wilcoxon rank sum test and a P-value of <0.05. For genes overlapping between cohorts of primary CRCs and CRC cell lines, consistency of up- or down-regulation in MSI cases was assessed using the chi-squared test.

MSI versus MSS				
Primary CRCs		CRC cell lines		
		Giacomini <i>et al</i>		P-value
		Up	Down	
Aarhus University Hospital	Up	859 (36.3%)	305 (12.9%)	<0.001
	Down	421 (17.8%)	782 (33.0%)	
Watanabe <i>et al</i>	Up	913 (38.1%)	267 (11.1%)	<0.001
	Down	471 (19.6%)	747 (31.2%)	
Koinuma <i>et al</i>	Up	192 (44.2%)	44 (10.1%)	<0.001
	Down	51 (11.8%)	147 (33.9%)	

Single-sample MSI classification of test colorectal cancers and cancer cell lines using the 829 and 192 consistent MSI-associated genes

Individual test samples were clustered against a common MSI/MSS reference set of primary CRCs or CRC cell lines (Supplementary Table 1). Test samples included additional 30 MSI and 30 MSS cancers analyzed at the Aarhus University Hospital, 9 MSI and 15 MSS cancers from Watanabe *et al* (16), 11 MSI and 63 MSS cancers from the Royal Melbourne Hospital, and 5 MSI and 6 MSS CRC cell lines from Giacomini *et al* (17).

	MSI-typing (reference)	829-gene classifier		192-gene (229 IMAGE clone) classifier	
		MSS	MSI	MSS	MSI
Primary CRCs					
Aarhus University Hospital	MSS	25	5	25	5
	MSI	2	28	2	28
Watanabe <i>et al</i>	MSS	14	1	15	0
	MSI	0	9	0	9
Royal Melbourne Hospital	MSS	51	12	56	7
	MSI	0	11	0	11
CRC cell lines					
Giacomini <i>et al</i>	MSS	NA	NA	5	1
	MSI	NA	NA	0	5

Table 4

Consistent MSI-associated candidate genes identified by at least two Affymetrix probe sets and one IMAGE clone.

Gene ID	Gene symbol	Gene name	MSI vs. MSS
NM_025113	C13orf18	chromosome 13 open reading frame 18	Down
NM_003671	CDC14B	CDC14 cell division cycle 14 homolog B (S. cerevisiae)	Down
BC003064	DAB2	disabled homolog 2, mitogen-responsive phosphoprotein (Drosophila)	Down
NM_013974	DDAH2	dimethylarginine dimethylaminohydrolase 2	Down
NM_019114	EPB41L4B	erythrocyte membrane protein band 4.1 like 4B	Down
NM_018267	H2AFJ	H2A histone family, member J	Down
BE566023	KIAA0372	KIAA0372	Down
NM_016436	PHF20	PHD finger protein 20	Down
AF131790	SHANK2	SH3 and multiple ankyrin repeat domains 2	Down
AB018322	TMCC1	transmembrane and coiled-coil domain family 1	Down
NM_020182	TMEPAI	transmembrane, prostate androgen induced RNA	Down
NM_005783	TXNDC9	thioredoxin domain containing 9	Down
NM_021964	ZNF148	zinc finger protein 148	Down
AA551142	PHACTR2	phosphatase and actin regulator 2	Down
NM_030920	ANP32E	acidic (leucine-rich) nuclear phosphoprotein 32 family, member E	Up
BC000751	EIF5A	eukaryotic translation initiation factor 5A	Up
R60018	RABEP1	rabaptin, RAB GTPase binding effector protein 1	Up
BF343007	TFAP2A	transcription factor AP-2 alpha (activating enhancer binding protein 2 alpha)	Up
AK021741	TMF1	TATA element modulatory factor 1	Up
M61715	WARS	tryptophanyl-tRNA synthetase	Up