Published in final edited form as:

Genes Chromosomes Cancer. 2011 April ; 50(4): 275–283. doi:10.1002/gcc.20853.

Sporadic Breast Cancer Patients' Germline DNA Exhibit an AT-Rich Microsatellite Signature

Cristi L. Galindo¹, Lauren J. McIver¹, Hongseok Tae¹, John F. McCormick¹, Michael A. Skinner^{3,8}, Ina Hoeschele², Cheryl M. Lewis³, John D. Minna^{4,5,6,7}, David A. Boothman⁷, and Harold R. Garner¹

¹Virginia Bioinformatics Institute, Blacksburg, VA

²Department of Statistics of Virginia Polytechnic Institute and State University, Blacksburg, VA

³Department of Surgery, Dallas, TX

⁴Department of Pharmacology, Dallas, TX

⁵Department of Internal Medicine, Dallas, TX

⁶Hamon Center for Therapeutic Oncology Research, Dallas, TX

⁷Simmons Comprehensive Cancer Center of the University of Texas Southwestern Medical Center, Dallas, TX

⁸Children's Medical Center, Dallas, TX

Abstract

Using a custom CGH-like oligonucleotide array to measure the global microsatellite content in the genomes of 72 cancer, cancer-free, and high risk patient and cell line samples (56 germline DNA and 16 in tumor or tumor cell line DNA) we found a unique, reproducible, and statistically significant pattern of 18 motif-specific microsatellite families (out of 962 possible 1-6 mer repeats) in breast cancer patient germline and tumor DNA, but not in germline DNA of cancer-free volunteer controls or in breast cancer patients with BRCA1/2 mutations. These high-similarity A/T rich repetitive motifs were also more pronounced in the germlines and tumors of colon cancer tumor patients (3/6 samples) and microsatellite unstable colon cancer cell lines; however, germline DNA of sporadic breast cancer patients exhibited the largest global content shift for those motifs with extreme AT/GC ratios. These results indicate that global microsatellite variability is complex, suggest the existence of a previously unknown genomic destabilization mechanism in breast cancer patients' germline DNA, and warrant further testing of such microsatellite variability as a predictor of future breast cancer development.

Introduction

Microsatellite length mutations are receiving increasing attention as both a marker and contributing factor to oncogenesis (Forgacs et al., 2001, Woerner et al., 2001). Microsatellite repeats (typically defined as tandemly repeated sequences (motifs) of one to six nucleotides) are ubiquitous and frequently polymorphic at rates that far exceed typical single-nucleotide mutation rates (Elegren, 2004) in mammalian genomes, and their polymorphism can generate significant phenotype variation (Rubinsztein et al., 1995, Fujisawa et al., 1999, Laidlaw et al., 2007). Somatic microsatellite length mutations are commonly observed in

colorectal, endometrial, breast, and gastric carcinomas, and are a common feature of some lung cancers (Girard et al., 2000, Wistuba et al., 2000, Forgacs et al., 2001). Microsatellite instability (MSI), defined as extreme hypervariability of microsatellites throughout the genome, has been shown to be a manifestation of defects in DNA mismatch repair genes (Jiricny, 2006). We hypothesize that variability in both somatic and germline microsatellites may play an important etiological role in the development and progression of some cancers. It is critical to have knowledge of their mutational frequency, complexity, and diversity among different types of epithelial-derived cancers, as well as an understanding of how they vary in different normal genetic backgrounds.

Microsatellite alterations occur in many tumors, but their frequency and spectra are variable, with certain types of tumors harboring significantly elevated rates of mutation at these loci (Imai et al., 2008). The occurrence of microsatellite mutations at several loci in multiple different cancers is strong evidence that these microsatellite mutations are indeed important events in the progression of these cancers. Even stronger evidence lies in the observation that there is likely some selection for these specific mutations, because microsatellite mutations in other loci with similar repeat sequences are not observed in these tumors (Riccio et al., 1999). Alterations in repeat unit number in and around coding sequences can have important quantitative and qualitative effects on gene expression (Tassone et al., 2000, Bontekoe et al., 2001, Di Marco et al., 2001, Fondon et al., 2004) and thus could potentially contribute directly to cancer progression. Elucidation of the nature and cause of somatic microsatellite mutations in cancer and how they are distinct from those operating in the germline can provide critical insights into the molecular underpinnings of the oncogenetic process. Furthermore, an investigation of global microsatellite differences in various cancers might provide cancer-specific signatures, as well as help identify individual cancer biomarkers.

To investigate microsatellites on a global scale, we designed a custom hybridization array to measure genomic microsatellite content, which we previously used to phylogenetically classify a large number of diverse and closely related species (Galindo et al., 2009). The probe sequences were computationally-derived simple repeat DNA sequences (i.e. all possible 1- to 6-mer microsatellite motif combinations, including every cyclic permutation and corresponding complement sequence). This global microsatellite array directly compares hybridization intensity values that represent the summation across all individual microsatellite motif-containing loci irrespective of their locations in the genome; this distinguishes the process from arrays used to estimate copy variations at specific positions within the genome. For example, the intensity recorded on the probe for the AATT motif (and probes for its cyclic permutations, ATTA, TAAT, and TTAA) measures the contributions from the 886 AATT motif specific microsatellite loci spread throughout the human genome, as computationally identified from the published reference genomic sequence. The global microsatellite array can therefore be used to specifically and accurately measure significant motif-specific variations (polymorphisms), whether they are in the germline or arise as somatic mutations, in any DNA sample (Galindo et al., 2009).

Materials and Methods

Sample acquisition and preparation

A total of 72 genomic DNA samples (Supplementary Table 1) were acquired from 12 cancer-free volunteers (DNA obtained from peripheral blood leukocytes, PBL), 10 patients with sporadic breast cancer (breast tumor and normal tissue-derived DNA), 6 breast cancer patients with BRCA1 or BRCA2 mutation (PBL), and DNA derived from colon cancer patients, cell lines, and individuals with 'unknown' cancer status obtained from Coriell Cell Repositories (Camden, New Jersey). Volunteer blood samples were collected by the

McDermott Clinical Laboratory at the University of Texas Southwestern Medical Center (UTSW) in accordance with Institutional Review Board (UTSW IRB# 1287-355). Cell lines were provided by Drs. Girard, Minna, and Boothman and authenticated by fingerprinting through the McDermott Center Sequencing Core at UTSW. Patient samples were provided by Drs. Perou, Lewis, and the UTSW Tissue Repository, with each institution's review board approval. All other genomic DNA was purchased from Coriell Cell Repositories (Camden, New Jersey).

Microsatellite instability (MSI) assay

Genomic DNA samples were analyzed for microsatellite instability using the Promega MSI Analysis System, Version 1.2, in the McDermott Center Sequencing Core Facility, per manufacturer's instructions (Supplementary Table 2). MSI status was assigned according to the Bethesda Guidelines (Boland et al., 1998, Umar et al., 2004).

Array design, manufacture, and processing

Each array consisted of 53,735 unique probes, replicated 7X, including probes to measure repetitive DNA sequences for all possible 1-mers to 6-mers, a variety of non-repetitive DNA sequences, and a series of controls. Further details are available in reference 16. A database containing all raw array intensity data from these experiments are available for download at <http://discovery.vbi.vt.edu/gmc>. Roche NimbleGen (Madison, Wisconsin) manufactured each array and also performed DNA (~1 µg, 250 ng/µl) labeling, hybridization, and scanning following their standard protocols. All test samples (labeled with Cy3) were co-hybridized with Cy-5-labeled Promega (Madison, Wisconsin) human reference DNA. To measure array specificity, DNA from two volunteer samples before and after Epstein-Barr virus (EBV) transformation were hybridized to the array and microsatellite content examined for the presence of an EBV genomic microsatellite (GAGCAG). Additionally, a custom 60-mer (5' -

GCAAAGGGACCCACGGTGGAAACAGGAGCAGGAGCAGGAGCGGGAGGGGCAGGAGCAGGAG-3') and its complement were designed based on the GAGCAG repeat-containing EBV sequence (Integrated DNA Technologies, Coralville, Iowa), and 500 pmoles was spiked into a cancer-free volunteer DNA sample (Supplementary Figure 3).

Array data processing and statistical analysis

Background subtraction and quantile normalization was performed across all arrays using NimbleScan (Roche NimbleGen), followed by comparison of all reference sample signal intensity values ($R^2 = 0.93 \pm 0.06$ across the 72 arrays). Control probes were used to gauge background levels, reproducibility of reference samples, and final statistical output as previously described (Galindo et al., 2009). As expected, the intensity values decreased predictably among wild type (WT) and microsatellite-specific control (single mismatch (SM), double mismatch (DM) and deletion (DEL)) probes (Supplementary Fig. 1).

After these pre-processing steps, a total of 962 different microsatellite motifs remained for statistical analysis, including 24 relevant comparisons (Supplementary Table 3). Differences among groups were initially assessed with the parametric Welch t-test and non-parametric Wilcoxon two-sample test. For our small sample sizes, these types of tests tend to have little power (e.g., Kooperberg et al.) (Kooperberg et al., 2005). We therefore also performed the moderated t-tests of Smyth (Smyth, 2004) implemented in the Bioconductor (<http://www.bioconductor.org>) package *Limma*. We obtained *P*-values from suitable reference distributions and applied several false discovery rate (FDR; Benjamini-Hochberg method (Hochberg et al., 1990)) controlling methods (at the 0.05 level) to the resulting *P*-values, including several versions of the q-value method (Storey et al., 2003), which produced very similar results. Only motifs with significant differences consistent across all

possible cyclic permutations were considered (25% difference between compared samples, q value = 0.05).

Computation of probe occurrences in the genome

The 5,356 wild type microsatellite probe sequences on the array were computationally aligned to the published human reference genome (NCBI Build Number 36, Version 3, Human Genome Sequencing Consortium release 4, March 24, 2008) using a Perl script to search for all 1-mer through 6-mer microsatellite motifs (minimum length of 18-20 bp). Genetic regions were constructed from the March 2006 assembly of the human reference genome from the UCSC Genome Table Browser (<http://genome.ucsc.edu>). Microsatellites were similarly associated with fragile sites, 'cancer' genes, defined as those 4,230 genes in NCBI Entrez associated with cancer, and previously identified cancer breakpoint genes (Burrow et al., 2009). All microsatellite occurrences were also aligned to the nearest SNP-associated comparative genomic hybridization value, as obtained from Illumina 109K SNP array (Illumina Inc., San Diego, California) data for 10 breast cancer patients. Global gain/loss in copy number, estimated as the average signal amplification ratio (tumor vs. normal, diploid DNA) for all SNPs associated with each individual microsatellite locus compared to the number present in the reference genome, was negligible (~2.6% variation on average).

Results

Global Microsatellite Content is a Sensitive Measure of Small Genomic Differences

Using the custom microarray, we assessed the global microsatellite content of a total of 72 genomic DNA samples. Our initial comparisons of the various sample groups resulted in the demonstration of the power of this technique to detect variations in global microsatellite content. The array specificity and sensitivity was confirmed by comparing the genomic content of blood leukocytes from two cancer-free volunteers, before and after Epstein-Barr virus (EBV) transformation. All motifs in primary and transformed leukocytes were virtually identical ($R^2 = 0.95$) with the exception of one motif, a GAGCAG repeat, the only microsatellite present in the EBV genome (Supplementary Fig. 2). Likewise, addition of a custom 70-mer oligonucleotide including the GAGCAG motif and flanking EBV genomic sequence into DNA samples recapitulated these results, demonstrating detection linearity ($R^2 = 0.99$, Supplementary Fig. 3).

There is an Increased Concentration of AT-Rich Microsatellites in the Genomes of Breast Cancer Patients

Based on analysis of germline DNA from 12 cancer-free volunteers versus 10 breast cancer patients, there were 18 statistically significant (using modified t-tests and FDR control by the q-value method with smoothing) microsatellite motifs (including cyclic permutations) that consistently differed between each cancer-free volunteer and all ten patient samples (Fig. 1). Comparisons of race (data not shown) and gender (Supplementary Fig. 4) indicated that within sample size limitations of this study, gender and ethnicity were not factors related to the higher incidence of these global microsatellite motifs in the germlines of breast cancer patients. Notably, there was no significant difference in the global microsatellite content between the cancer patient germlines and tumors (Fig. 1). More significant was the fact that BRCA1/2 positive breast cancer patients did not exhibit the microsatellite signature detected in women with sporadic breast cancer. Overall, these results suggest that these 18 microsatellite motifs might represent a biomarker for breast cancer predisposition that is orthogonal to known risk mutations.

To exclude the possibility that the global microsatellite differences were related to segmental chromosomal duplications or insertions, we analyzed whole genomic SNP array

data on the ten breast cancer patient samples for amplifications and deletions of regions containing microsatellites. The copy number variations for each microsatellite at each locus were summed and compared. Based on this analysis, differences in variations in global microsatellite content as ascertained by the custom microsatellite array were not due to large gains or losses of chromosomal content. The contribution of segmental chromosomal duplications to the global microsatellite signature detected in breast cancer samples (compared to normal reference DNA) was negligible (2.6% for all differential microsatellite motifs).

Thirteen of the microsatellite motif families whose global content was more prevalent in sporadic breast cancer patients were notably AT-rich, and 5 families whose content was suppressed in these same patients were GC-rich (Fig. 1). Indeed, all low complexity repetitive DNA exhibited this striking AT/GC dependence (Figs. 2 and 3), as did high complexity sequences to a lesser extent (Supplementary Fig. 5). This pattern was not observed for breast cancer patients that tested positive for BRCA1/2 mutations (Supplementary Fig. 6) or any of the other sample groups examined in this study (Supplementary Table 3, typical example, colon cancer, Supplementary Fig. 7). Interestingly, the majority of microsatellites located in or near retrotransposons are AT-rich. Indeed, when we analyzed the human reference genome for correlations between AT-rich microsatellite regions and retrotransposons, we found a clear trend in the numbers of AT-rich repeats in relation to these mobile elements (Supplementary Fig. 8). Whereas 90.3% of all pure AT microsatellite repeats are located in or near retrotransposons, only 5.7% of microsatellites that are composed of pure GC motifs are within 50 bps of retrotransposons.

Global Microsatellite Content Distinguishes MSI-High and MSI-Low Colon Cancer

Examination of 3 primary colon cancer patient samples yielded only one motif (TGGGTC) that consistently distinguished the germlines and tumors of all three patients from cancer-free volunteers (Supplementary Fig. 9). Since this motif's higher global content was observed in patient germlines, it might possibly serve as a biomarker for colon cancer predisposition. The global content for this same motif was also much higher in all three colon cancer cell lines (RKO, HCT15, and HCT116), when compared to the 12 cancer-free volunteer samples. Interestingly, there were two additional motifs (poly A and AATAC) that were specific for the three colon cancer tumor lines (Supplementary Fig. 9). The global content of both of these motifs were much lower in all three of these tumor lines, which is consistent with their known high microsatellite instability (MSI) status (Ionov et al., 2004). We confirmed MSI status for our samples using the Promega MSI genotyping kit (Supplementary Table 2). This extensively used 'gold standard' for classification of MSI is based upon the analysis of only 5 intergenic poly-A repeats (Bacher et al., 2004), out of a total of 169,315 poly-A and poly-T repeats found within the genome sequence. Notably, primary colon cancer patient samples did not exhibit lower global poly A content (Supplementary Fig. 9) and were not identified as MSI-unstable using the kit (Supplementary Table 2).

Discussion

Microsatellites are understudied despite their known connection with cancer and other diseases, mainly because there has never previously been a method for assaying them *en masse*. In this study, we describe a global microsatellite cancer signature detected using a new array-based approach that we previously demonstrated to be both sensitive and specific for species differentiation (Galindo et al., 2009). We found a set of repetitive microsatellite motifs commonly differential in sporadic breast cancer patient germlines when compared to normal individuals. This pattern may represent a breast cancer risk biomarker with potential

to be translated into a new predisposition diagnostic or used to identify new therapeutic targets.

Our data indicate that microsatellites of the form A_nT_m were elevated in breast cancer patients. Further research will be needed to ascertain the underlying mechanism for these differences, but DNA slippage or crossing over cannot fully explain this data since we also observed enrichment for non-repetitive sequences. Other mechanisms, such as failures in the base excision repair pathway may account for some of these observations (Negrini et al., 2010). Consistent with this finding, the distribution of 13 statistically differential high AT-rich motifs were located almost exclusively outside gene coding regions, which are inherently GC-rich (Table 1).

The dramatic enrichment of A:T (or depletion of G:C) may manifest itself through chromosomal fragile sites, which contain AT-rich flexibility islands that are especially prone to DNA breakage, potentially via the formation of secondary structures (Zlotorynski et al., 2003). Microsatellites commonly form such structures (Wells et al., 2005), and more than half of the breakpoints in the 444 genes involved in cancer-specific chromosomal translocations are associated with fragile sites (Burrow et al., 2009), supporting a mechanism for heritable breast cancers that are not caused by known mutations (e.g., BRCA1/2). Another possible consequence for the observed pattern is increased activity of LINE-1 endonuclease, which cleaves AT-rich sequences and generates double-strand breaks (Hedges et al., 2007). Although the effects of LINE-1 endonuclease on microsatellites has not been directly tested, decreased levels of LINE-1 methylation are correlated with the development and pathogenesis of several cancer types (Yu et al., 2001, Pattamadilok et al., 2008), and genomic loss of methylation results in microsatellite instability and cancer in mice (Dion et al., 2008, Howard et al., 2008). Further, our data indicate that there are a disproportionate number of AT-rich microsatellites in/near retrotransposons, compared to GC-rich repeat sequences (Supplementary Fig. 8); we speculate that an increase in AT-rich repeats near retrotransposons might increase the probability of transposable element activity and thus the chance of an insertion event that may lead to cancer.

A larger scale study may be merited to determine if global microsatellite content signatures can also be used as a reliable biomarker for tumor sub-type classification and prediction of prognosis or response to therapy. Only the analysis of a statistically significant number of other heritable cancer types will reliably determine the true prevalence of this signature, so future studies will be aimed at evaluating other cancer types, such as lung cancers arising in non-smokers, which presumably have a larger genetic predisposition component relative to environmentally induced lung cancers. Further, to obtain a complete picture of where and how these AT-rich microsatellites vary, it will be necessary to employ deep sequencing of microsatellite target-enriched regions. This experiment will have the added advantage of simultaneously generating hundreds of thousands of genotypes at individual loci that could yield informative biomarkers and causative lesions in cancer. The abnormal microsatellite signatures potentially implicate thousands of genetic loci, suggesting that there may be many more important repeat-containing loci affecting cancer development or progression that are yet to be identified.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was funded by Virginia Bioinformatics Institute director's funds, the P.O'B. Montgomery Distinguished Chair and the Hudson Foundation. Cristi L. Galindo received support from an NIH cardiology fellowship (5-T32-

HL07360-28), Cardiology Department, University of Texas Southwestern Medical Center (UTSW). This work was also partially supported by the University of Texas NIH/NCI SPORE in Lung Cancer (P50CA70907). We would like to extend a special thanks to Linda Gunn, Zhaohui Sun, Jennifer Sayne and Tanishia Choice, M.D. for their valuable technical assistance and Charles M. Perou of the Lineberger Comprehensive Cancer Center of the University of North Carolina who provided luminal sub-type and basal sub-type breast cancer samples.

Supported by: NIH 5-T32-HL07360-28, NIH/NCI SPORE in Lung Cancer P50CA70907, P. O'B. Montgomery, M.D. Distinguished Chair, direct funds from Virginia Tech

References

- Bacher JW, Flanagan LA, Smalley RL, Nassif NA, Burgart LJ, Halberg RB, Megid WM, Thibodeau SN. Development of a fluorescent multiplex assay for detection of MSI-High tumors. *Dis Markers*. 2004; 20:237–50. [PubMed: 15528789]
- Boland CR, Thibodeau SN, Hamilton SR, Sidransky D, Eshleman JR, Burt RW, Meltzer SJ, Rodriguez-Bigas MA, Fodde R, Ranzani GN, Srivastava S. A National Cancer Institute Workshop on Microsatellite Instability for cancer detection and familial predisposition: development of international criteria for the determination of microsatellite instability in colorectal cancer. *Cancer Res*. 1998; 58:5248–57. [PubMed: 9823339]
- Bontekoe CJ, Bakker CE, Nieuwenhuizen IM, van der Linde H, Lans H, de Lange D, Hirst MC, Oostra BA. Instability of a (CGG)₉₈ repeat in the Fmr1 promoter. *Hum Mol Genet*. 2001; 10:1693–9. [PubMed: 11487573]
- Burrow AA, Williams LE, Pierce LC, Wang YH. Over half of breakpoints in gene pairs involved in cancer-specific recurrent translocations are mapped to human chromosomal fragile sites. *BMC Genomics*. 2009; 10:59. [PubMed: 19183484]
- Di Marco S, Hel Z, Lachance C, Furneaux H, Radzioch D. Polymorphism in the 3'-untranslated region of TNF α mRNA impairs binding of the post-transcriptional regulatory protein HuR to TNF α mRNA. *Nucleic Acids Res*. 2001; 29:863–71. [PubMed: 11160917]
- Dion V, Lin Y, Hubert L Jr, Waterland RA, Wilson JH. Dnmt1 deficiency promotes CAG repeat expansion in the mouse germline. *Hum Mol Genet*. 2008; 17:1306–17. [PubMed: 18252747]
- Ellegren H. Microsatellites: simple sequences with complex evolution. *Nat Rev Genet*. 2004; 5:435–45. [PubMed: 15153996]
- Fondon JW 3rd, Garner HR. Molecular origins of rapid and continuous morphological evolution. *Proc Natl Acad Sci U S A*. 2004; 101:18058–63. [PubMed: 15596718]
- Forgacs E, Wren JD, Kamibayashi C, Kondo M, Xu XL, Markowitz S, Tomlinson GE, Muller CY, Gazdar AF, Garner HR, Minna JD. Searching for microsatellite mutations in coding regions in lung, breast, ovarian and colorectal cancers. *Oncogene*. 2001; 20:1005–9. [PubMed: 11314036]
- Fujisawa T, Ikegami H, Kawaguchi Y, Yamato E, Nakagawa Y, Shen GQ, Fukuda M, Ogihara T. Length rather than a specific allele of dinucleotide repeat in the 5' upstream region of the aldose reductase gene is associated with diabetic retinopathy. *Diabet Med*. 1999; 16:1044–7. [PubMed: 10656235]
- Galindo CL, McIver LJ, McCormick JF, Skinner MA, Xie Y, Gelhausen RA, Ng K, Kumar NM, Garner HR. Global microsatellite content distinguishes humans, primates, animals, and plants. *Mol Biol Evol*. 2009:msp192.
- Girard L, Zochbauer-Muller S, Virmani AK, Gazdar AF, Minna JD. Genome-wide allelotyping of lung cancer identifies new regions of allelic loss, differences between small cell lung cancer and non-small cell lung cancer, and loci clustering. *Cancer Res*. 2000; 60:4894–906. [PubMed: 10987304]
- Hedges DJ, Deininger PL. Inviting instability: Transposable elements, double-strand breaks, and the maintenance of genome integrity. *Mutat Res*. 2007; 616:46–59. [PubMed: 17157332]
- Hochberg Y, Benjamini Y. More powerful procedures for multiple significance testing. *Stat Med*. 1990; 9:811–8. [PubMed: 2218183]
- Howard G, Eiges R, Gaudet F, Jaenisch R, Eden A. Activation and transposition of endogenous retroviral elements in hypomethylation induced tumors in mice. *Oncogene*. 2008; 27:404–8. [PubMed: 17621273]

- Ionov Y, Matsui S, Cowell JK. A role for p300/CREB binding protein genes in promoting cancer progression in colon cancer cell lines with microsatellite instability. *Proc Natl Acad Sci U S A*. 2004; 101:1273–8. [PubMed: 14732695]
- Imai K, Yamamoto H. Carcinogenesis and microsatellite instability: the interrelationship between genetics and epigenetics. *Carcinogenesis*. 2008; 29:673–80. [PubMed: 17942460]
- Jiricny J. The multifaceted mismatch-repair system. *Nat Rev Mol Cell Biol*. 2006; 7:335–46. [PubMed: 16612326]
- Kooperberg C, Aragaki A, Strand AD, Olson JM. Significance testing for small microarray experiments. *Stat Med*. 2005; 24:2281–98. [PubMed: 15889452]
- Laidlaw J, Gelfand Y, Ng KW, Garner HR, Ranganathan R, Benson G, Fondon JW 3rd. Elevated basal slippage mutation rates among the Canidae. *J Hered*. 2007; 98:452–60. [PubMed: 17437958]
- Negrini S, Gorgoulis VG, Halazonetis TD. Genomic instability--an evolving hallmark of cancer. *Nat Rev Mol Cell Biol*. 2010; 11:220–8. [PubMed: 20177397]
- Pattamadilok J, Huapai N, Rattanatanyong P, Vasurattana A, Triratanachai S, Tresukosol D, Mutirangura A. LINE-1 hypomethylation level as a potential prognostic factor for epithelial ovarian cancer. *Int J Gynecol Cancer*. 2008; 18:711–7. [PubMed: 17944913]
- Riccio A, Aaltonen LA, Godwin AK, Loukola A, Percepe A, Salovaara R, Masciullo V, Genuardi M, Paravatou-Petsotas M, Bassi DE, Ruggeri BA, Klein-Szanto AJ, Testa JR, Neri G, Bellacosa A. The DNA repair gene MBD4 (MED1) is mutated in human carcinomas with microsatellite instability. *Nat Genet*. 1999; 23:266–8. [PubMed: 10545939]
- Rubinsztein DC, Leggo J, Coetzee GA, Irvine RA, Buckley M, Ferguson-Smith MA. Sequence variation and size ranges of CAG repeats in the Machado-Joseph disease, spinocerebellar ataxia type 1 and androgen receptor genes. *Hum Mol Genet*. 1995; 4:1585–90. [PubMed: 8541843]
- Smyth GK. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol*. 2004; 3 Article3.
- Storey JD, Tibshirani R. Statistical significance for genomewide studies. *Proc Natl Acad Sci U S A*. 2003; 100:9440–5. [PubMed: 12883005]
- Tassone F, Hagerman RJ, Chamberlain WD, Hagerman PJ. Transcription of the FMR1 gene in individuals with fragile X syndrome. *Am J Med Genet*. 2000; 97:195–203. [PubMed: 11449488]
- Umar A, Boland CR, Terdiman JP, Syngal S, de la Chapelle A, Rüschoff J, Fishel R, Lindor NM, Burgart LJ, Hamelin R, Hamilton SR, Hiatt RA, Jass J, Lindblom A, Lynch HT, Peltomäki P, Ramsey SD, Rodriguez-Bigas MA, Vasen HF, Hawk ET, Barrett JC, Freedman AN, Srivastava S. Revised Bethesda Guidelines for hereditary nonpolyposis colorectal cancer (Lynch syndrome) and microsatellite instability. *J Natl Cancer Inst*. 2004; 96:261–8. [PubMed: 14970275]
- Wells RD, Dere R, Hebert ML, Napierala M, Son LS. Advances in mechanisms of genetic instability related to hereditary neurological diseases. *Nucleic Acids Res*. 2005; 33:3785–98. [PubMed: 16006624]
- Wistuba II, Behrens C, Virmani AK, Mele G, Milchgrub S, Girard L, Fondon JW 3rd, Garner HR, McKay B, Latif F, Lerman MI, Lam S, Gazdar AF, Minna JD. High resolution chromosome 3p allelotyping of human lung cancer and preneoplastic/preinvasive bronchial epithelium reveals multiple, discontinuous sites of 3p allele loss and three regions of frequent breakpoints. *Cancer Res*. 2000; 60:1949–60. [PubMed: 10766185]
- Woerner SM, Gebert J, Yuan YP, Sutter C, Ridder R, Bork P, von Knebel Doeberitz M. Systematic identification of genes with coding microsatellites mutated in DNA mismatch repair-deficient cancer cells. *Int J Cancer*. 2001; 93:12–9. [PubMed: 11391615]
- Yu F, Zingler N, Schumann G, Strätling WH. Methyl-CpG-binding protein 2 represses LINE-1 expression and retrotransposition but not Alu transcription. *Nucleic Acids Res*. 2001; 29:4493–501. [PubMed: 11691937]
- Zlotorynski E, Rahat A, Skaug J, Ben-Porat N, Ozeri E, Hershberg R, Levi A, Scherer SW, Margalit H, Kerem B. Molecular basis for expression of common and rare fragile sites. *Mol Cell Biol*. 2003; 23:7143–51. [PubMed: 14517285]

Fig. 1.

Comparison of normalized signal values for primary breast cancer tumors and blood lymphocytes and cancer-free volunteers reveals a consistent pattern of microsatellite motif changes. Each colored line represents the average normalized signal intensity value (ratio of sample signal to reference DNA signal) for all possible permutations of the listed motif (e.g., TTA, ATT, TAT), each of which was identified as significantly differential between breast cancer patients and cancer-free volunteers (25%, p value = 0.05, calculated using modified t-tests and FDR control by the q-value method with bootstrap). Each signal value for a probe representing a particular motif results from the additive hybridization of all of fluorescently labeled DNA sequences. For example, there are 14,460 genomic loci containing the TTA microsatellite motif, and the value obtained is a 'collective' or 'summed' measurement of all TTA-containing loci. Normalized hybridization signal intensity and individual samples are displayed on the ordinate and abscissa, respectively. As shown, there were 18 microsatellite motifs that clearly distinguished the primary tumors and germlines of the 10 sporadic breast cancer patients from the 12 cancer-free volunteers and 6 BRCA1/2+ patients. There was a pattern to these motifs, 13 of which were AT-rich and more prevalent in sporadic breast cancer patients and 5 of which were suppressed and GC-rich.

Fig. 2.

The ratio of global microsatellite content (gmc) in breast cancer patients compared to that of cancer-free volunteers is a function of AT percentage. Array hybridization intensity ratios (test sample / reference sample) representing global microsatellite content for 10 breast cancer patient and 12 cancer-free volunteer blood leukocyte DNA samples were used to compute a log fold difference. The resulting 962 values, which represented all possible 1-mer to 6-mer microsatellite motif families (excepting the spike-in control GAGCAG motif present in the EBV genome) were binned based on percentage AT content and graphed accordingly. As shown, there was a clear increase in the global microsatellite signal intensities of breast cancer patient germlines relative to the genomes of cancer-free volunteers with increasing AT content. The same trend was observed in the tumors of breast cancer patients. The 13 AT-rich and 5 AT-poor/GC-rich motif families determined to be significantly different between breast cancer patients and cancer-free volunteers are denoted in red.

Fig. 3.

Hierarchical clustering of all normalized signal ratios representing microsatellites composed of 100% AT content and 0% AT content (212 motif families) differentiates sporadic breast cancer patients (tumors and germlines) from the majority of healthy volunteers (9 out of 12), BRCA1+ and BRCA2+ breast cancer patients, and familial breast cancer patients that tested negative for BRCA1/2 mutations. Note that the 106 microsatellites composed purely of AT nucleotides clustered apart from the 106 0% AT microsatellite families (vertical nodes). The tumor and germline genomes of sporadic breast cancer patients were clearly defined by an overall higher content of 100% AT microsatellite motifs, as compared to cancer-free volunteers and BRCA1/2+ breast cancer patients (horizontal nodes). Bright red indicates the

highest normalized intensity value, bright blue indicates the lowest, and yellow represents median values.

Table 1

Motif	Up stream	Down stream	5' UTR	3' UTR	Intron	Exon	Intergenic	Total Loci	Total RefSeq Genes	Total Cancer Genes
Sporadic Breast Cancer Patient Motifs										
TTA	74	67	732	134	4,588	0	8,865	14,460	3,508	774
TATAT	3	1	11	2	99	0	363	479	98	22
TTTAGT	0	1	3	0	22	0	34	60	26	5
TATT	139	170	1,609	233	9,411	0	18,063	29,625	5,919	1,362
TTTTCA	0	0	8	1	23	0	49	81	32	7
TATTCT	1	0	1	0	18	0	39	59	18	4
TATTTC	0	0	2	0	18	0	50	70	18	3
TATATT	1	1	17	6	154	0	383	562	147	31
TATCTT	0	0	1	0	7	0	10	18	6	0
ACTTTT	0	0	2	0	8	0	17	27	9	2
AATTT	2	2	35	6	193	0	452	690	227	57
AATTTT	3	2	38	8	246	0	462	759	277	64
TATTTT	63	79	496	85	3,173	0	5,639	9,535	2,832	662
ACGGGC	1	0	1	0	1	0	0	3	4	1
TGGCGA	0	0	0	0	1	1	0	2	2	1
GCGGT	0	0	1	1	2	0	1	5	4	0
CGGCCA	1	1	2	1	2	1	1	9	9	2
GAGCGG	6	1	8	0	1	7	12	35	23	7
Motifs that Differentiate Colon Cancer Only										
TGGGTC	1	0	1	0	5	1	13	21	9	2
poly A	1,080	1,660	10,215	1,721	63,241	0	91,398	169,315	13,375	2,864
AATAC	0	2	9	0	42	0	94	147	57	12

Only genes in the RefSeq database were included. A “count” is defined as a complete tandem repeat at least 18 bp (for 3-mers and 6-mers) or 20 bp (for 1-, 2-, 4-, and 5-mers), in length. Upstream and downstream were defined as 1,000 bp distal from the transcribed gene. There are a total of 4,230 ‘cancer’ genes and 31,118 RefSeq genes.