OPEN ACCESS Freely available online

PLoS one

Transitions at CpG Dinucleotides, Geographic Clustering of *TP53* Mutations and Food Availability Patterns in Colorectal Cancer

Fabio Verginelli[1], Faraz Bishehsari[1,2], Francesco Napolitano[3], Mahboobeh Mahdavinia[1,2], Alessandro Cama[1], Reza Malekzadeh[2], Gennaro Miele[4], Giancarlo Raiconi[3], Roberto Tagliaferri[3], Renato Mariani-Costantini[1]*

1 Department of Oncology and Neurosciences, ''G. d'Annunzio'' University, and Center of Excellence on Aging (CeSI), ''G. d'Annunzio'' University Foundation, Chieti, Italy, **2** Digestive Disease Research Center (DDRC), Shariati Hospital, University of Tehran, Tehran, Iran, **3** Department of Mathematics and Informatics, University of Salerno, Salerno, Italy, **4** Department of Physical Sciences, University of Naples, Naples, Italy

Abstract

Background: Colorectal cancer is mainly attributed to diet, but the role exerted by foods remains unclear because involved factors are extremely complex. Geography substantially impacts on foods. Correlations between international variation in colorectal cancer-associated mutation patterns and food availabilities could highlight the influence of foods on colorectal mutagenesis.

Methodology: To test such hypothesis, we applied techniques based on hierarchical clustering, feature extraction and selection, and statistical pattern recognition to the analysis of 2,572 colorectal cancer-associated *TP53* mutations from 12 countries/geographic areas. For food availabilities, we relied on data extracted from the Food Balance Sheets of the Food and Agriculture Organization of the United Nations. Dendrograms for mutation sites, mutation types and food patterns were constructed through Ward's hierarchical clustering algorithm and their stability was assessed evaluating silhouette values. Feature selection used entropy-based measures for similarity between clusterings, combined with principal component analysis by exhaustive and heuristic approaches.

Conclusion/Significance: Mutations clustered in two major geographic groups, one including only Western countries, the other Asia and parts of Europe. This was determined by variation in the frequency of transitions at CpGs, the most common mutation type. Higher frequencies of transitions at CpGs in the cluster that included only Western countries mainly reflected higher frequencies of mutations at CpG codons 175, 248 and 273, the three major *TP53* hotspots. Pearson's correlation scores, computed between the principal components of the datamatrices for mutation types, food availability and mutation sites, demonstrated statistically significant correlations between transitions at CpGs and both mutation sites and availabilities of meat, milk, sweeteners and animal fats, the energy-dense foods at the basis of ''Western'' diets. This is best explainable by differential exposure to nitrosative DNA damage due to foods that promote metabolic stress and chronic inflammation.

Citation: Verginelli F, Bishehsari F, Napolitano F, Mahdavinia M, Cama A, et al. (2009) Transitions at CpG Dinucleotides, Geographic Clustering of *TP53* Mutations and Food Availability Patterns in Colorectal Cancer. PLoS ONE 4(8): e6824. doi:10.1371/journal.pone.0006824

Editor: Irene Oi-Lin Ng, The University of Hong Kong, Hong Kong

Received May 20, 2009; **Accepted** July 14, 2009; **Published** August 31, 2009

Funding: The study was funded by the Italian Ministry of Education, University and Research (Ministero dell'Istruzione, dell'Università e della Ricerca, MIUR) and by the Faculties of Medicine and Pharmacy, G. d'Annunzio University, Chieti, Italy (years 2005-08). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: rmc@unich.it

Introduction

The *TP53* gene (OMIM no. 191117), which encodes a tumor-suppressor protein that drives multiple cellular responses to stress, including cell-cycle arrest, DNA repair, apoptosis, metabolism and autophagy, is frequently mutated in cancer [1,2,3,4,5,6]. *TP53* mutations are mostly missense and cluster in exons 5–8, the evolutionarily-conserved region of the DNA-binding domain that contains ≈90% of the known mutations and all mutation hotspots at CpG dinucleotides [7,8,9,10,11]. Laboratory models and data from tumors with established environmental risk factors show that *TP53* mutation patterns reflect primary mutagenic signatures of DNA damage by carcinogens, vulnerability of nucleotide positions in DNA secondary structure, efficiency of repair processing, and selection for loss of trans-activation properties [10,11,12,13,14,15,16].

Colorectal cancer (CRC), worldwide one of the most common malignancies, is mainly attributed to dietary risk factors [17,18,19,20,21,22,23,24]. *TP53* mutations are found in 50-60% of all CRCs and are thought to originate in precancerous lesions, where aberrantly proliferating colonocyte progenitors are directly exposed to dietary residue [25,26]. Nevertheless the *TP53* mutation pattern typical of CRC cannot be easily correlated to

diet, because it is characterized by a striking preponderance of G:C>A:T transitions [9,13,16]. These are the most frequent base substitutions induced by reactive oxygen species, byproducts of normal aerobic metabolism generated at high levels in all inflammatory processes and after exposure to a wide variety of carcinogens and toxicants [27,28,29,30,31,32,33,34]. Furthermore CRC development appears to depend on whole-life nutrition pattern [23], and *TP53* mutations may occur years before CRC diagnosis [25,35]. Thus the time-frame for the estimation of diet may not fully capture the period relevant for mutagenesis and carcinogenesis. This is complicated by the relatively limited variation in dietary habits within single populations, by biases in reporting and recording dietary intakes and by the problematic assessment of exposures to food-borne carcinogens and toxicants, natural and generated in foods production, processing, preservation, and preparation [17,23,36,37,38,39,40,41,42]. Adding to complexity, intestinal mutagenesis may be modified by nutrient/nutrient, nutrient/microflora, nutrient/cell metabolism, nutrient/gene and nutrient/DNA repair interactions, and affected by epigenetic modifications, transit time of dietary residue, inflammatory and endocrine responses, body mass and energy consumption through physical activity [23,40,43,44,45,46,47,48].

Geography strongly impacts on the ecological, cultural and economic factors that determine food systems and diets. CRCs from patients embedded in geographically diverse populations and cultures reflect substantially different dietary exposures, extended over the whole-life course and unbiased by estimation errors [17,21,23]. Thus food-related mutational signatures could be highlighted through the analysis of geographic variation in CRC-associated *TP53* mutations. To test such hypothesis, we analyzed 2,572 *TP53* mutations associated with primary CRCs from 12 countries or geographic areas. The mutations (Database S1) were extracted from the *TP53* database of the International Agency for Research on Cancer (IARC) (R10 update, July – 2005, http://www-p53.iarc.fr/Somatic.html), with the addition of an Iranian series [11,49]. To investigate correlations between geographic clustering of *TP53* mutations and foods, we relied on the food balance sheets (FBS) of the Food and Agriculture Organization of the United Nations (FAO, http://faostat.fao.org/site/368/DesktopDefault.aspx?PageID = 368), that provide unique comprehensive pictures of the patterns of national food supply, useful for international comparisons [50,51,52]. Food availability patterns (FPs) for the countries/geographic areas in the *TP53* database were derived from the mean *per caput* supplies, in percent of the total caloric value, of each major food group available for human consumption during the reference year 1990 [17] (Dataset S1). The datamatrices generated for mutation sites (MS), mutation types (MT) and FP (Datamatrices S1) were investigated for geographic variation by hierarchical clustering (HC). Factors underlying HC were defined by feature analysis (FA) through principal component analysis (PCA). Pearson's correlation scores were computed between the principal components of the mutation type, food availability pattern and mutation site datamatrices. These analyses demonstrated significant correlations between transitions at CpGs and both mutation sites as well as availabilities of meat, milk, sweeteners and animal fats. Our results could be best explainable with differential exposure to nitrosative DNA damage due to the consumption of energy-dense foods that promote metabolic stress and chronic low-grade inflammation.

Results

Geographic variation in mutation site and type

Panels A–B and C–D of Figure 1 respectively show hierarchical clustering (HC) by country/geographic area for *TP53* mutation sites (MS), based on 2,542 exonic mutations, and types (MT), based on 2,572 mutations in exons and intron-exon boundaries. The MS and MT trees showed similar structures, each with two major geographic clusters, one including only Western countries (I-MS, I-MT), the other Asia and parts of Europe (II-MS, II-MT). The main difference consisted in the position of West and Central Europe in II-MS and I-MT, respectively. Stability of clusters was assessed by silhouette values. Silhouette plots for different thresholds, applied to each dendrogram, were compared to assess the reliability of the clustering solutions. In both cases the tree structure showed two stable clusters. The low silhouette value of MT was related to the poor stability of the "Spain" branch, attributable to either I-MT or II-MT. The MS and MT tree structures were correlated by two-tailed Mantel test ($r = 0.581$, $P = 0.001$) (Figure 2).

By multivariate FA we next investigated the factors that determined clustering for MS (*i.e.*, codons) and MT (*i.e.*, mutation types), respectively using heuristic or exhaustive approaches. Feature selection aimed at identifying the minimum subset of features necessary to generate the clustering structure obtained using all the features. Sequential forward feature selection with two different rankings, respectively based on the number of mutations recorded for each codon (feature) and on the PC coefficients of each feature, was used to analyze the MS datamatrix by heuristic approach (Figure 3).

Stable MS clustering was obtained with 23 weight-ranked or 22 PCA-ranked codons, in both cases including the five *TP53* mutation hotspots (*i.e.*, CpG codons 175, 245, 248, 273, 282) [9,13,16], out of 173 mutated codons in the datamatrix. The variance contributed by the PCs of the MS datamatrix and their eigenvalues are shown in panels A and D of Figure 4, respectively. Total MS variance was explained by 11 components. Four components contributed 80% of the variance, and the first component, which accounted for 31%, had highly significant PC coefficients for the features corresponding to the five CpG hotspots, as detailed in File S1 and in Figure S1, panel A.

Exhaustive multivariate FA of the MT datamatrix is reported in Tables 1 and 2. In decreasing order, the most relevant features were G:C>A:T at CpGs, followed by A:T>C:G, G:C>A:T and G:C>C:G. The variance contributed by each PC of the MT datamatrix and their eigenvalues are shown in panels B and E of Figures 4 respectively. Total MT variance was explained by 4 components, the first of which accounted for 65%, and, as detailed in File S1 and in Figure S1, panel D, the highest PC feature loading among the 8 mutation types corresponded to transitions at CpGs. Other mutations, including transitions at non-CpGs, were associated to minor fractions of variance.

The frequency box-plots of the mutations at the 19 codons with highest weights and highest PCA variance coefficients in Figure 5, panel A, showed higher mutation frequencies at the three major hotspot codons 175, 248, and, particularly, 273, in I-MS versus II-MS. This reflected higher frequencies of transitions at CpGs in I-MT (range: 46.1–61.2%) versus II-MT (range: 41.2–43.3%) in the frequency box-plots of the 8 mutation types in Figure 5, panel B. Such most relevant features were used to geographically visualize MS and MT variation (Figure 6, panels A–B). Highlighted groupings of countries/geographic areas were similar to the MS and MT clusters in Figure 1, obtained by HC using all the features. Overall these results indicate that in CRC *TP53* transition mutagenesis at CpGs is modulated by geography-related factors. This might reflect differences in exposure(s) to specific food-associated mutagenic process(es) [53].

Geographic variation in food supply patterns

To address this issue, we analyzed the FP datamatrix by HC and FA through PCA. HC for FP was based on the mean *per caput*

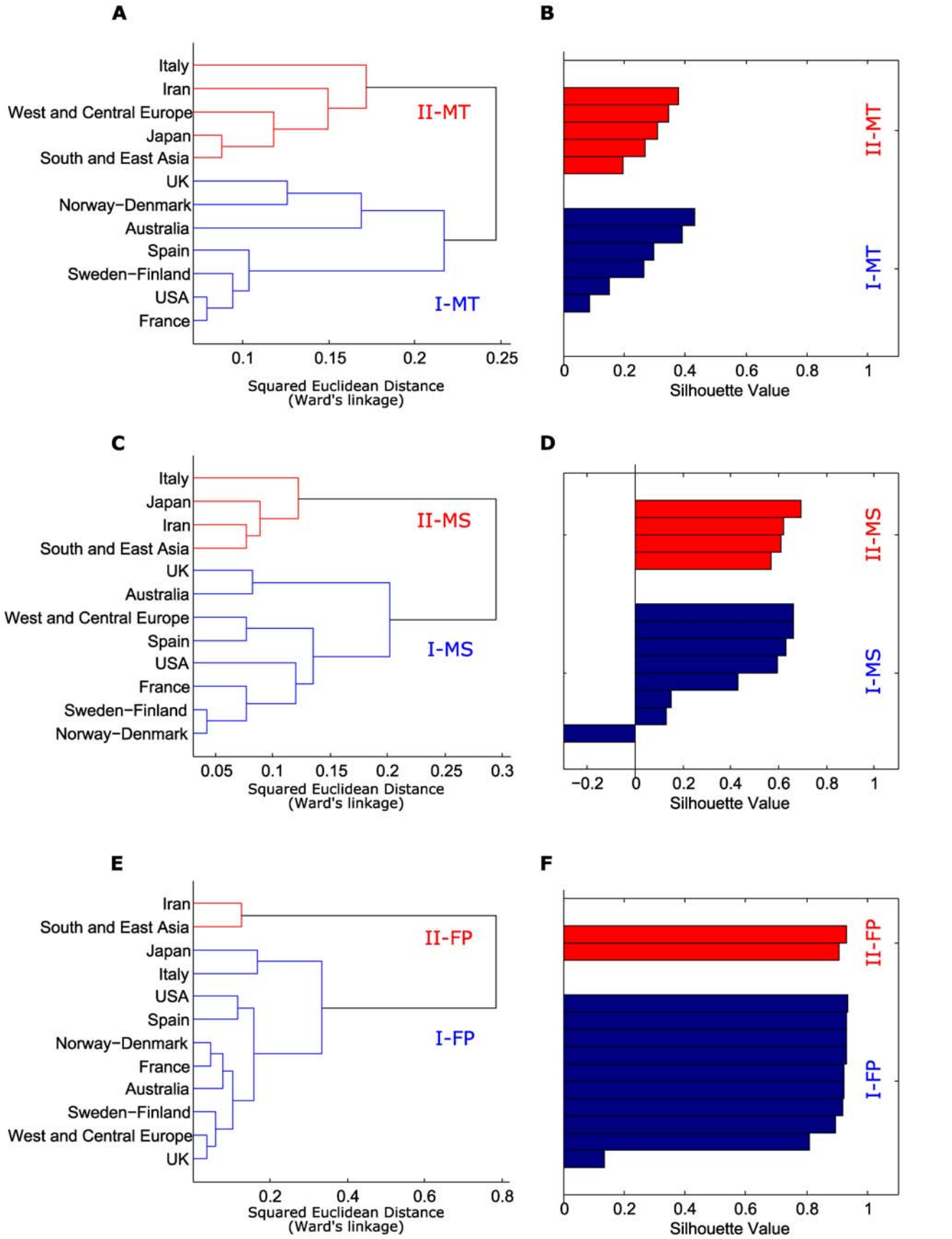


Figure 1. Hierarchical clusterings for *TP53* mutation sites, *TP53* mutation types and food patterns. Hierarchical clusterings (HC) by country/geographic area and silhouette plots for: A–B, *TP53* mutation sites (MS); C–D, *TP53* mutation types (MT); E–F, food patterns (FP).
doi:10.1371/journal.pone.0006824.g001

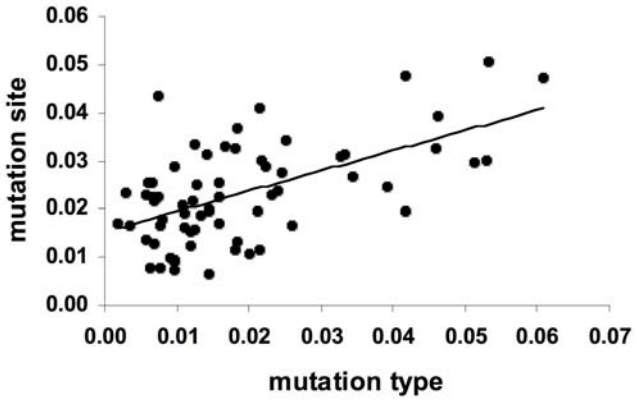


Figure 2. Mantel test correlation between *TP53* mutation sites and *TP53* mutation types. Mantel test shows correlation between the distance matrices of mutation sites (MS) and mutation types (MT), with regression line parameter: $r = 0.581$, $c = 0.015$; $R^2 = 0.338$, $P = 0.001$ after 10,000 permutations.
doi:10.1371/journal.pone.0006824.g002

supply values, in percent of total available calories, of each major food group in the relevant countries/geographic areas during the reference year 1990 [17]. HC yielded two major clusters, I-FP, with Western countries and Japan, and II-FP, with South and East Asia plus Iran (Figure 1, panels E–F). The clusterization of Japan in I-FP had a low silhouette value and contrasted with the previous assignments of Japan to clusters II-MS and II-MT. To verify Japan's assignment, we generated all the possible subsets of the 13 FP features (food groups), *i.e.*, 8,192 subsets. HC trees, cut to obtain two clusters, were then generated based on each of these subsets. Dendrograms were classified as A or B when Japan clusterized respectively in II-FP or I-FP, and as C, when different from A and B. Overall 2,405 clusterings, classified as A, assigned Japan to cluster II-FP with Iran and South and East Asia; 4,178, classified as B, assigned Japan to cluster I-FP, with Western countries; and 1,609 were classified as C, being different from A and B. The histograms in Figure S2, panels A–B, that visualize the number of times that each of the 13 features was present when type A or B clusterings respectively were obtained, readily show that feature cereals was almost always absent in type A clusterings and almost always present in type B clusterings. Thus Japan joined I-FP only because of the low availability of cereals.

Tables 3 and 4 show the results of exhaustive FA of the FP datamatrix. In decreasing order, the most relevant features were cereals, milk, and meat. PCA showed that total FP variance was explained by 3 components, the first of which accounted for a major fraction of 87.3% (Figure 4, panels C and F). The variance of this component, which, in loading order, included the features cereals, meat, milk, sweeteners, animal fats (File S1 and Figure S1, panel G), explained the tree structure, determined by lower cereals and higher meat, milk, sweeteners and animal fats in I-FP relative to II-FP, as shown in panels C and F of Figure 4.

Correlations between mutation pattern and food supply pattern

The data from the MS, MT and FP datamatrices were projected on the 1-dimensional space spanned by their respective PCs. Pairwise Pearson correlations were then computed for the three datamatrices in all the projected spaces. Tables 5 to 7 show the correlation scores, and the corresponding *P*-values, obtained for the first 3 PCs of each datamatrix, that, except for MS, accounted for most of the variance. Pearson's correlation between the PCs for MT and for FP (Table 5) showed that the first PC for MT was correlated with the first PC for FP, with $r = -0.60$ ($P = 0.039$). Availabilities of meat, milk, sweeteners and animal fats were directly correlated to transitions at CpGs, availability of cereals to transitions at non-CpGs (File S1 and Figure S1, panels D and G). As detailed in File S1 and in Figure S1, other less important correlations involved second and third PCs that accounted for minor fractions of variance. With the same analysis, the first PCs for MS and for MT resulted again strongly correlated, with $r = -0.87$ ($P = 0.0002$, Table 6), which supported Mantel's test results (Figure 2). However, in spite of the correlation between MT and FP, there were no significant correlations between the PCs of MS and FP (Table 7).

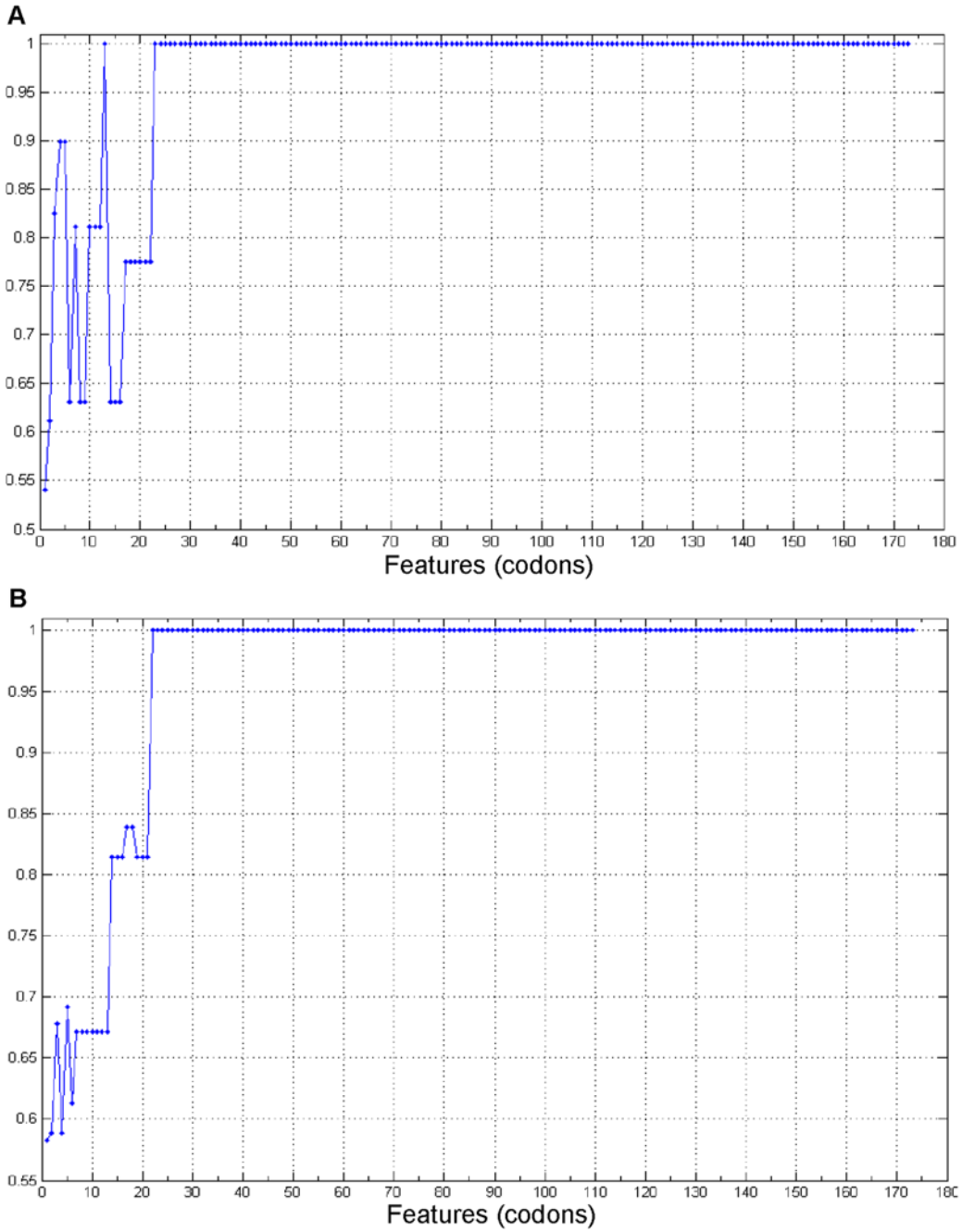
Scatter plots with superimposed linear regression showing the global trend of correlations were built for the countries/geographic areas as projected on the 2-dimensional spaces spanned by the first PCs of MS and MT (Figure 7) and of MT and FP (Figure 8). As shown in Figure 7, Italy, Iran, South and East Asia and West and Central Europe had relatively lower frequencies of mutations at CpG hotspot codons, compensated by higher frequencies of mutations at all other sites (see also box-plots in Figure 2). Mutation frequencies at CpG hotspots increased in other countries, with highest frequencies in Australia and UK. As shown in Figure 8, transitions at CpGs correlated with countries/geographic areas characterized by higher availabilities of energy-dense, Western-style foods, while South and East Asia, Iran, Japan and, to a lesser extent, Italy, where cereals were higher and meat, milk, sweeteners and animal fats lower, had lower frequencies of such mutations.

Overall, variation in the frequency of transitions at CpGs reflected variation in the availabilities of the energy-dense foods that form the basis of "Western-style" diets and that are linked to overweight and obesity [18,20,21,22,23]. Transitions at non-CpGs balanced decreases in transitions at CpGs in the countries/geographic areas where cereals compensated for lower availabilities of such foods.

Discussion

Several studies addressed the issue of CpG transition mutagenesis in cancer, with particular regard to *TP53* mutations in CRC. Being exonic CpGs constitutively hyper-methylated, C to T mutations at coding CpGs in *TP53* should be scored as direct transitions from hypermutable 5-methylcytosine to thymine [54,55,56,57,58,59]. Dietary folate is a defined environmental determinant of genomic methylation [23,60,61]. Laboratory models and data on CRCs in patients carrying a germline methylenetetrahydrofolate reductase (*MTHFR*) gene variant that results in reduced plasma and serum folate suggest that low folate, by inducing global hypomethylation, may decrease *TP53* transition mutagenesis at CpGs [62,63,64]. Folate-rich foods include fresh vegetables, pulses (legumes) and relatively unprocessed cereals [65,66]. Little is known about DNA methylation variation among individuals and populations [67], [68]. We did not find any correlation between availability of vegetables or pulses and *TP53* mutation pattern, while cereals, relatively unprocessed in most Asian countries [69,70], inversely correlated with transitions at CpGs. Thus folate availability may not account for our results. This conclusion agrees with studies showing that, in absence of interacting genetic effects, folate alone does not influence *TP53* mutation patterns in CRC (although it may affect TP53 protein expression) [44,71,72].

The hypermutability of endogenous 5-methylcytosine does not *per se* explain the unique role of transitions at CpGs in geographic



C

Weight ranking	PCA ranking
135	135
151	151
152	152
157	157
158	
173	173
175	175
176	
179	179
	195
196	196
	205
213	213
220	220
238	238
	241
244	244
245	245
248	248
249	249
266	266
272	
273	273
278	
282	282
306	306

Figure 3. Feature selection by heuristic approach for the mutation sites data. Similarity values are on the y axis; number of features (*i.e.*, codons) on the x axis, respectively ranked in decreasing order of weight (number of mutations, panel A), and decreasing order of the first 11 PC coefficients of each feature (where 11 was the number of PCs contributing 100% of the data variance, panel B). Panel C lists (in order of number) the codons (features) with highest variance in the mutation sites (MS) data, selected by weight ranking (23/173 codons) and by PC coefficients ranking (22/173 codons). Overall 19 codons, including the five *TP53* mutation hotspots (highlighted in yellow), were selected by both methods.
doi:10.1371/journal.pone.0006824.g003

clustering of *TP53* mutations [57,58,59]. However transitions at CpGs in *TP53* are efficiently induced by nitrosative DNA damage [31,58,59], [73,74,75]. Nitric oxide (NO), a critical signalling molecule implicated in the regulation of peristalsis, gut vasomotor functions and mucosal inflammation, may contribute to transition mutagenesis at CpGs acting directly at 5-methylcytosines, by nitrosative deamination in oxidizing environments, and, indirectly, at guanines, by base alkylation after conversion to nitrate, bacterial reduction to nitrite and endogenous formation of N-nitroso compounds [73,74,75,76,77,78,79,80,81,82]. Mutagenesis at CpGs may be facilitated by NO-induced inhibition of DNA repair [75,80]. Furthermore, NO promotes apoptosis via *TP53* and therefore exerts a critical selective pressure for *TP53* mutation [83,84,85,86].

NO is produced at mutagenic concentrations by inducible NO synthase (iNOS), the widespread enzyme isoform upregulated by inflammatory cytokines [76,82,87]. It has already been suggested that the excess of *TP53* transitions at CpGs found in cancers

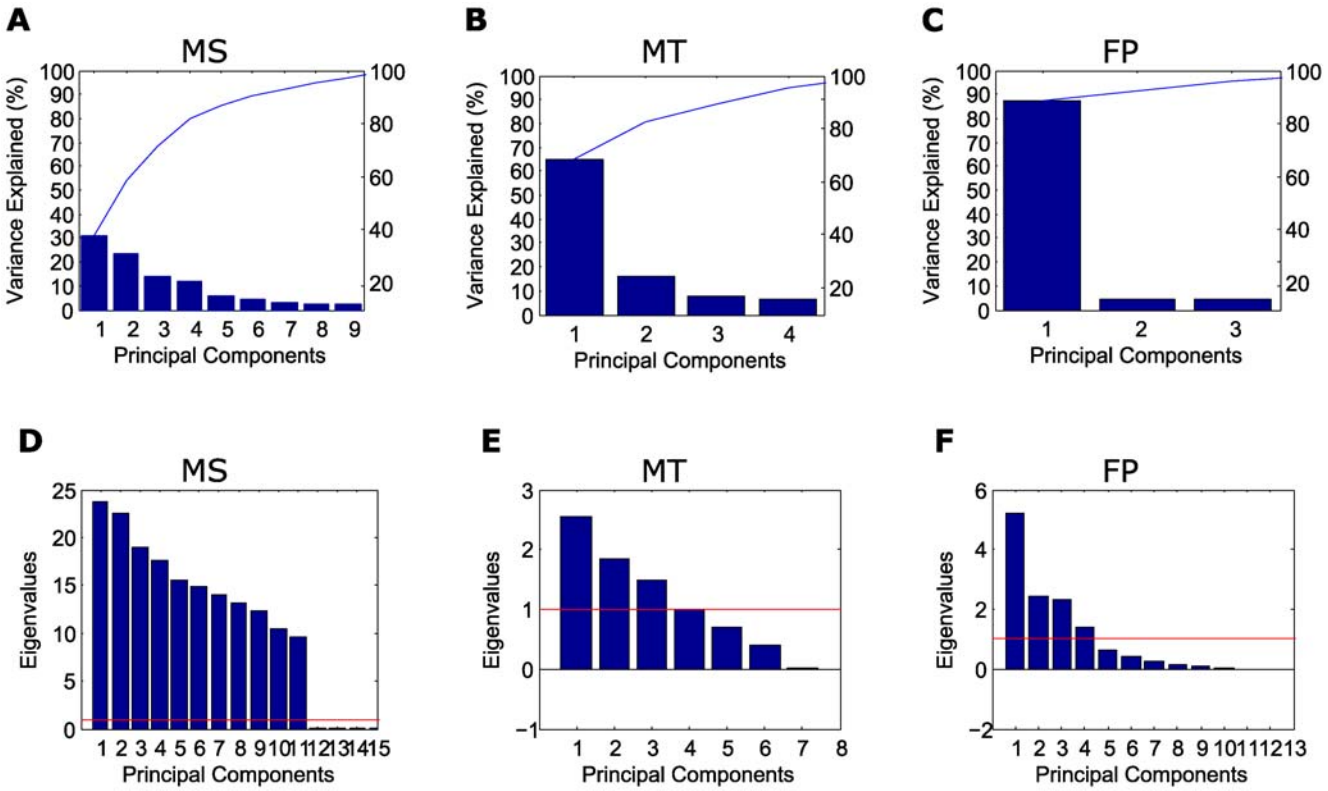


Figure 4. Scree and Kaiser's tests applied to the analysis of variance by PCA of the datamatrices for mutation sites, mutation types and food patterns. Results obtained using Scree test are shown in panels A–C. Total mutation sites (MS) variance (panel A) is explained by 11 components, of which only 9 are visualized, being the Scree test cut at the 98% level. Four components contribute 80% of the MS variance, the first accounting for 31%. Total mutation types (MT) variance (B) is explained by 4 components, the first of which, by far the most relevant, contributes 65% of variance. Total food patterns (FP) variance (C) is explained by 3 components, the first of which contributes 87.3% of variance. Results obtained using Kaiser's test are shown in panels D–F. The first 11 PCs for mutation sites (MS) (A), the first 3 PCs for mutation types (MT) (B) and the first 4 PCs for food patterns (FP) (C) have eigenvalues above 1 (red line).
doi:10.1371/journal.pone.0006824.g004

arising on a chronic inflammatory background, such as CRC in ulcerative colitis and bladder cancer associated with *Schistosomiasis*, results from nitrosative stress [74,88]. Moreover transitions at CpGs are strongly related to iNOS expression in both CRC and adenocarcinoma of Barrett's esophagus [89,90]. Arginine, the substrate for NO synthesis and a potential CRC-related dietary factor [87,91,92,93], is contained in a variety of protein-rich foods of animal and vegetable origin [65,66] and may not *per se* explain why variation in the frequency of transitions at CpGs correlated with variation in the availabilities of meat, milk, sweeteners and

Table 1. Worst-case exhaustive feature analysis of the mutation types datamatrix.

|S| = number of selected features (j)	1	2	3	4	5	6	7	8
Feature (i)								
1. A:T>C:G	0.42	0.39	0.32	0.38	0.31	0.38	0.38	1.00
2. A:T>G:C	0.49	0.44	0.34	0.39	0.31	0.39	0.38	1.00
3. A:T>T:A	0.44	0.40	0.32	0.40	0.38	0.38	0.38	1.00
4. FS	0.50	0.40	0.34	0.33	0.31	0.38	0.38	1.00
5. G:C>A:T	0.42	0.40	0.37	0.33	0.38	0.38	0.38	1.00
6. G:C>A:T at CpG	**0.92**	**0.80**	**0.80**	**0.79**	**0.77**	**0.80**	**0.81**	**1.00**
7. G:C>C:G	0.36	0.37	0.32	0.33	0.31	0.38	0.38	1.00
8. G:C>T:A	0.51	0.37	0.37	0.33	0.31	0.38	0.38	1.00

Worst similarity values for each subset S of the MT features, including the selected feature in the left column, are computed using the exhaustive multivariate feature analysis. Bold characters highlight the worst similarity values of the features that most influence cluster structure. Entry (i,j) reports the minimum similarity value obtained using the i-th feature together with any other j-1 features. For example, feature A:T>C:G gives a similarity value that is at least 0.38 when coupled with any 6 of the features in the left column.
doi:10.1371/journal.pone.0006824.t001

Table 2. Best-case exhaustive feature analysis of the mutation types datamatrix.

Column 1 (Feature)	Column 2 (Features that paired with feature in column 1 give the same clustering obtained with all features)
A:T>C:G	G:C>A:T at CpG
A:T>G:C	none
A:T>T:A	none
FS	none
G:C>A:T	G:C>A:T at CpG
G:C>A:T at CpG	A:T>C:G and/or G:C>A:T and/or G:C>C:G
G:C>C:G	G:C>A:T at CpG
G:C>T:A	none

Exhaustive multivariate feature analysis is used to highlight pairs of MT features giving similarity value equal to 1. Features in column 1 give similarity 1 if and only if coupled with one of the features in column 2. For example, feature A:T>C:G gives a similarity value 1 when coupled with G:C>A:T at CpGs. Relevance of individual features in column 1 is based on the number of other features in column 2 with which it can yield unitary value after pairing. The most relevant feature is G:C>A:T at CpGs, followed by A:T>C:G, G:C>A:T and G:C>C:G.

doi:10.1371/journal.pone.0006824.t002

animal fats. However it is known that these energy-dense foods promote a pro-inflammatory milieu that increases iNOS expression and NO production [23,78,94,95,96,97,98,99,100], [101]. In addition red meat is a major exogenous source of nitrogen compounds and haem, which contribute to N-nitrosation in the intestinal environment [23,102,103,104,105,106,107,108]. Such considerations are supported by the fact that our data point to a key role of the ubiquitously methylated major *TP53* hotspot codons 175, 248 and 273 in geographic clustering. In fact, the vast majority of the mutations at these 3 codons reported in human cancer are compatible with nitrosative deamination [9,11,32,54,74,109]. Moreover, transitions at codon 248 were experimentally induced with an NO-releasing compound [110] while mutations at codon 273 were found to be strongly associated with diets high in red meat and fat [44].

In conclusion, we recognize the difficulties inherent in interpreting causes and mechanisms responsible for CRC-associated *TP53* mutations, which are the end result of complex cascades of events. It is important to keep in mind the limitations of our analyses, based on a single, albeit large, database of mutations. Furthermore FAO FBS, the only standardized comprehensive food data available for international comparisons, approximate food supply patterns. Nevertheless, geographic variation in CRC-associated *TP53* mutation patterns appears to be due to transitions at CpGs and mainly related to differential mutation frequencies at the major *TP53* hotspots. This could be explainable by differential exposure to nitrosative DNA damage, linked to the consumption of foods promoting metabolic stress and chronic low-grade inflammation.

Materials and Methods

Databases, Datasets and Datamatrices

We analyzed 2,572 mutations in *TP53* exons 5–8 retrieved from primary CRCs, including 2,475 from 12 countries or geographic areas, extracted from the *TP53* database of the International Agency for Research on Cancer (IARC) (R10 update, July 2005, http://www-p53.iarc.fr/Somatic.html), and 97 from Iran [11,49]. Mutations in adenomas, metastatic CRC and cell lines were excluded, as their spectrum could differ from that of primary CRC [111]. Analyses were based on 2,542 mutations in coding regions for MS, and on 2,572 (*i.e.*, all) mutations for MT (Database S1). Mutations were grouped according to country or geographic area, the latter including geographically and ethnically related countries with low mutation numbers. Countries and number of mutations for MS and MT, were: Australia (including 6 mutations from New Zealand), MS:302, MT:302; USA, MS:233, MT:237; France, MS:215, MT:221; Italy, MS:181, MT:182; Spain, MS:181, MT:182; UK (including 3 mutations from Ireland), MS:131, MT:134; Iran, MS:94, MT:97; Japan, MS:323, MT:326. Geographic areas were: West and Central Europe (Germany, Austria, Switzerland, The Netherlands, Luxembourg), MS:174, MT:178; South and East Asia (China, Hong Kong, Taiwan, Singapore), MS:315, MT:318; Norway-Denmark, MS:162, MT:162; Sweden-Finland, MS:231, MT:233. Mutations from Brazil, Chile, Israel, Turkey, Korea and Eastern European countries listed in the R10 update of the IARC *TP53* database were excluded because of low numbers.

The FP dataset (Dataset S1) was extracted from the FAO FBS [50,51] compiled for the reference year 1990 (http://faostat.fao.org/site/368/DesktopDefault.aspx?PageID = 368), as used in reference [17]. Year selection tended to exclude the most recent and current international variations in food availabilities and nutrition, as CRC develops over several years and is mostly diagnosed in patients aged 65 years or older [112], while the IARC *TP53* database compiles mutations since 1989 [11]. The FP dataset included the following major food groups: animal fats, animal products, cereals, fish/seafood, fruit, meat, milk, oilcrops, pulses (legumes), starchy roots, sweeteners, vegetable oils and vegetables. For the purpose of this study alcohol was excluded, being much of the data on average availability of alcoholic drinks not informative and potentially confounding, due to large inter-individual variability [23]. Spices and stimulants, which account for low percentages of the total available daily energy supply, were also excluded. Statistical analyses were therefore conducted using the estimated percent (%) contribution of each considered food group to mean *per caput* daily energy availability [17]. Weighted average availabilities were calculated for geographic areas by adjusting for the 1990 population size of each included country. The MS, MT and FP datamatrices were normalized converting absolute numbers into frequencies (Datamatrices S1).

All standard techniques, including hierarchical clustering (HC), principal components analysis (PCA), Pearson correlation and linear regression, were used in their implementations from Matlab (2007b, The Mathworks and Matlab Statistics Toolbox).

Statistical Pattern Recognition

Statistical pattern recognition allowed the integrated analysis of the MS, MT, and FP datamatrices to investigate relations between *TP53* mutation sites, *TP53* mutation types, and food supply patterns. The first analytical step consisted in clustering the 12 analyzed coutries/geographic areas by HC with respect to the data contained in the MS, MT and FP datamatrices. The stability of the obtained clusterings was assessed using the silhouette values. The similarities between the obtained clustering solutions, represented by dendrograms, were assessed using an entropy-based similarity measure. Feature analysis and selection, which is the process of studying the contribution of single features, or subsets of features, to dataset properties, was the next relevant processing step. Exact feature analysis can be performed testing the ability of each single subset of features to maintain a chosen property. In practice, this is feasible only when

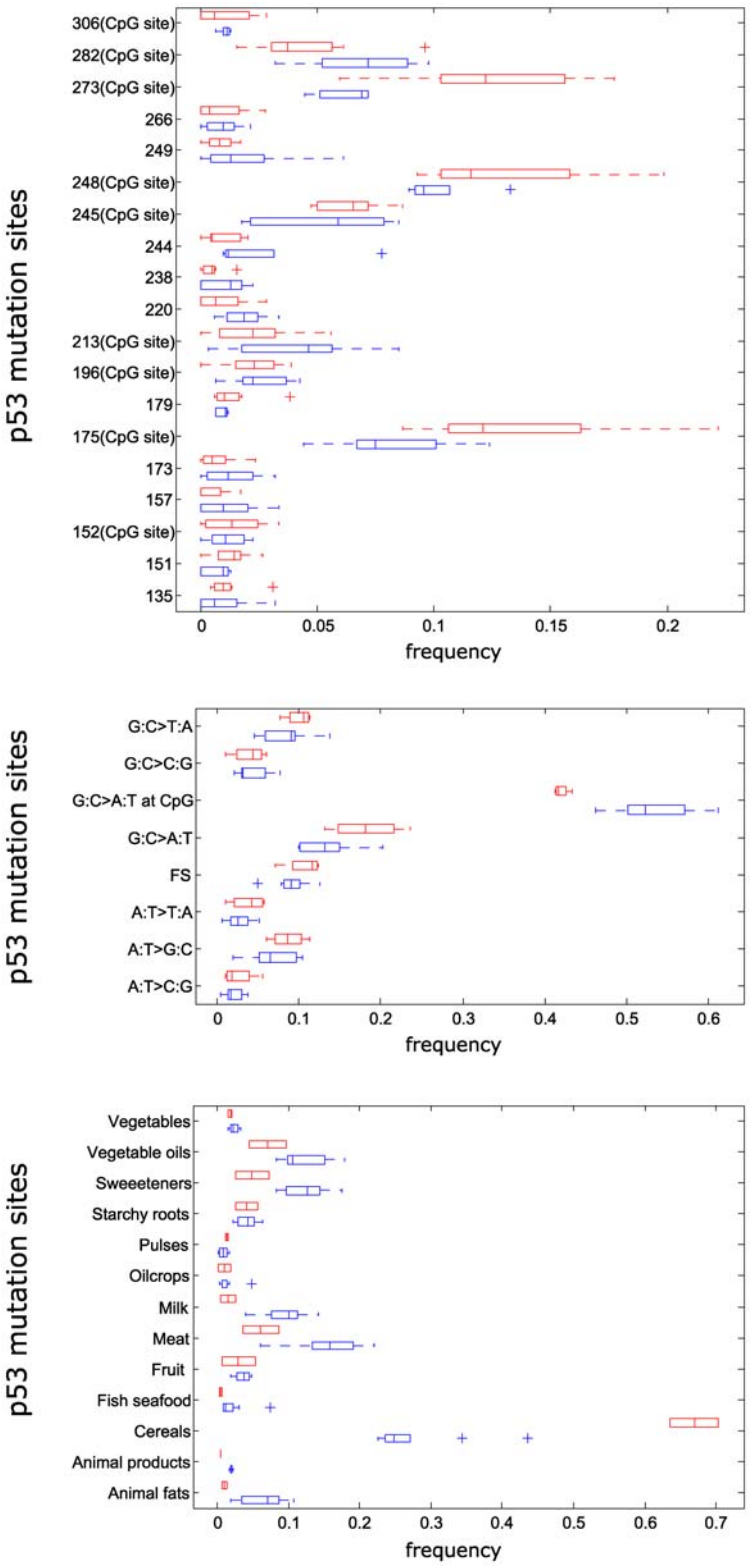


Figure 5. Box-plots of feature relevance for mutation sites, mutation types and food patterns. In each panel, box plots pertaining to clusters I versus II of mutation sites (MS), mutation types (MT) and food patterns (FP) obtained by hierarchical clustering are color-coded in red (cluster I) and blue (cluster II), respectively. Panel A: frequency box-plots of mutations at the 19 codons with highest weights and highest PCA variance coefficients (identified in Figure 4). Higher mutation frequencies at the three major CpG hotspot codons 175, 248, and 273 in I-MS versus II-MS are evident. Panel B: frequency box-plots of the 8 mutation types, showing higher frequencies of transitions at CpGs in I-MT (range: 46.1–61.2%) versus II-MT (range: 41.2–43.3%). Panel C: box-plots of the mean percent of the total available caloric value from each major food group in the relevant countries/geographic areas, showing lower cereals for the countries/geographic areas in I-FP versus those in II-FP, balanced by higher meat, milk, sweeteners and animal fats.
doi:10.1371/journal.pone.0006824.g005

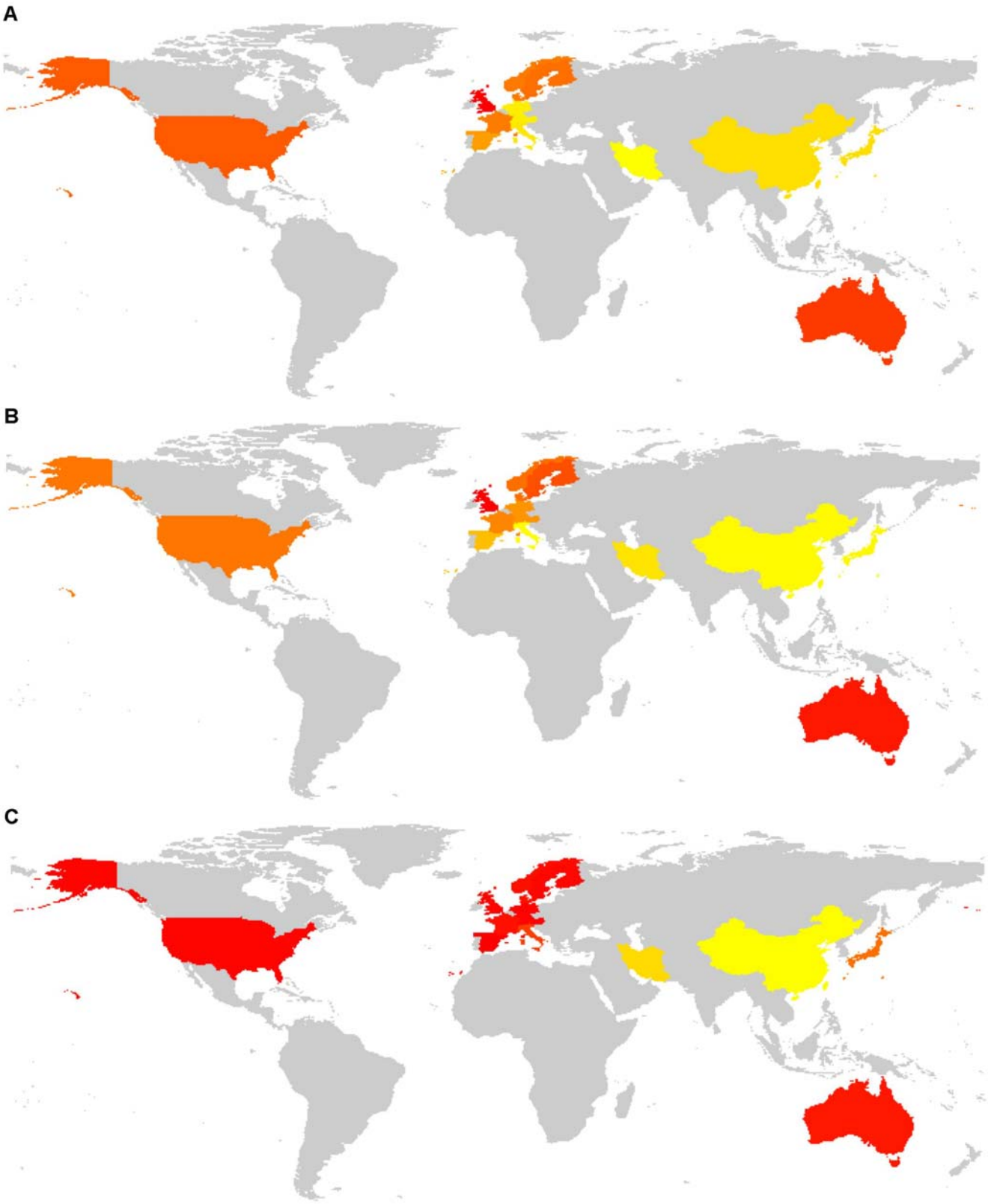


Figure 6. Geographic visualization of the most relevant features evidenced in the box-plots. Selected features of the mutation sites (MS, A), mutation types (MT, B) and food availability patterns (FP, C) datamatrices were mutations at the three major *TP53* hotspot codons 175, 248, and 273 for MS; G:C>A:T mutations at CpGs for MT; meat/milk/sweeteners/animal fats (added), cereals (subtracted) for FP. Feature frequencies were summed and projected in yellow to red color range onto the geographic profiles of the relevant countries/geographic areas.
doi:10.1371/journal.pone.0006824.g006

Table 3. Worst-case exhaustive feature analysis of the food patterns datamatrix.

|S| = number of selected features (j)	1	2	3	4	5	6	7	8	9	10	11	12	13
Feature (i)													
1. Animal fats	0.53	0.43	0.43	0.45	0.54	0.54	0.54	0.54	0.54	0.71	1.00	1.00	1.00
2. Animal products	1.00	0.51	0.51	0.51	0.51	0.53	0.53	0.54	0.54	0.71	1.00	1.00	1.00
3. Cereals	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
4. Fish/seafood	0.47	0.34	0.32	0.36	0.40	0.53	0.53	0.54	0.54	0.71	1.00	1.00	1.00
5. Fruit	0.65	0.52	0.52	0.51	0.49	0.53	0.53	0.54	0.54	0.71	1.00	1.00	1.00
6. Meat	0.71	0.71	0.71	0.71	0.71	0.71	0.72	0.72	0.72	0.80	1.00	1.00	1.00
7. Milk	1.00	0.80	0.71	0.71	0.71	0.71	0.71	0.71	0.71	0.71	1.00	1.00	1.00
8. Oilcrops	0.55	0.37	0.39	0.36	0.40	0.53	0.53	0.54	0.54	0.71	1.00	1.00	1.00
9. Pulses	0.40	0.40	0.35	0.36	0.40	0.53	0.53	0.54	0.54	0.71	1.00	1.00	1.00
10. Starchy roots	0.40	0.31	0.32	0.36	0.40	0.53	0.53	0.54	0.54	0.71	1.00	1.00	1.00
11. Sweeteners	0.58	0.56	0.54	0.54	0.54	0.54	0.54	0.54	0.54	0.71	1.00	1.00	1.00
12. Vegetable oils	0.51	0.51	0.51	0.51	0.51	0.54	0.54	0.70	0.70	1.00	1.00	1.00	1.00
13. Vegetables	0.35	0.31	0.32	0.49	0.40	0.53	0.53	0.54	0.54	0.71	1.00	1.00	1.00

Worst similarity values for each subset S of the FP features, including the selected feature in the left column, are computed using the exhaustive multivariate feature analysis. Bold characters highlight the worst similarity values of the features that most influence cluster structure (cereals). Entry (i,j) reports the minimum similarity value obtained using the i-th feature together with any other j-1 features. For example, feature "animal products" (i = 2) gives a similarity at least equal to 0.71 combined with any other remaining 9 features (j = 10).

doi:10.1371/journal.pone.0006824.t003

the number of features is low. In comparing the MT and FP datasets, because of the relatively low number of features, such exhaustive analysis could be carried out. With regard to the MS dataset, the number of possible subsets of features was too high, and therefore a heuristic approach, *i.e.*, sequential forward selection, was used to select feature subsets. The principal components and the relative weights of the features were used as ranking criteria. Results were visualized on geographic maps with the relevant areas colored according to the most relevant features. Finally, multivariate correlations between the datasets were computed exploiting their PC projections. All these analytical steps are detailed below.

Table 4. Best-case exhaustive feature analysis of the food patterns datamatrix.

Column 1 (Feature)	Column 2 (Features that paired with feature in column 1 give the same clustering obtained with all features)
Animal fats	Cereals, Meat, Milk
Animal products	Cereals, Meat, Oilcrops, Vegetable Oils
Cereals	All
Fish/Seafood	Cereals, Meat,Milk
Fruit	Cereals, Milk
Meat	Animal fats, Cereals, Fish/Seafood, Milk, Oilcrops, Starchy roots, Vegetable oils
Milk	All but Sweeteners
Oilcrops	Animal Products, Cereals, Meat, Milk
Pulses	Cereals, Milk
Starchy roots	Cereals, Meat, Milk
Sweeteners	Cereals
Vegetable oils	Cereals, Meat, Milk
Vegetables	Animal products, Cereals, Milk

Exhaustive multivariate feature analysis is used to highlight pairs of FP features giving similarity value equal to 1. Features in column 1 give similarity equal to 1 if and only if coupled with one of the features in column 2. For example, feature cereals (i = 3) gives a similarity that is always equal to 1 alone or together with any set of other features. Relevance of individual features in column 1 is based on the number of other features in column 2 with which it can yield unitary value after pairing. The most relevant features are cereals, milk, and meat.

doi:10.1371/journal.pone.0006824.t004

Table 5. Pearson's correlation scores between the PCs of mutation types and food patterns.

PCs	1th PC (MT)	2nd PC (MT)	3rd PC (MT)
1th PC (FP)	**−0.6003*** [*0.0391*]	0.2534 [*0.4268*]	**0.6194*** [*0.0317*]
2nd PC (FP)	**−0.6021*** [*0.0383*]	−0.1735 [*0.5897*]	−0.4654 [*0.1273*]
3rd PC (FP)	0.1724 [*0.5921*]	**0.6050*** [*0.0371*]	−0.1058 [*0.7434*]

Pearson's correlation scores computed between the principal components (PC) of mutation types (MT) and food availability patterns (FP), with the corresponding *P*-values (square brackets). Significant correlations are highlighted in bold and the corresponding r-coefficient values are marked with an asterisk (*).

doi:10.1371/journal.pone.0006824.t005

Table 6. Pearson's correlation scores between the PCs of mutation types and mutation sites.

PCs	1th PC (MT)	2nd PC (MT)	3rd PC (MT)
1th PC (MS)	**−0.8742*** [*0.0002*]	−0.0908 [*0.7790*]	−0.0118 [*0.9709*]
2nd PC (MS)	0.0975 [*0.7630*]	−0.1414 [*0.6612*]	−0.0568 [*0.8609*]
3rd PC (MS)	−0.0069 [*0.9830*]	0.2340 [*0.4642*]	0.3828 [*0.2194*]

Pearson's correlation scores computed between the principal components (PC) of mutation types (MT) and mutation sites (MS), with the corresponding *P*-values (square brackets). Significant correlations are highlighted in bold and the corresponding r-coefficient values are marked with an asterisk (*).

doi:10.1371/journal.pone.0006824.t006

Table 7. Pearson's correlation scores between the PCs of mutation sites and food patterns.

PCs	1th PC (MS)	2nd PC (MS)	3rd PC (MS)
1th PC (FP)	0.4176 [*0.1767*]	−0.4246 [*0.1688*]	0.4168 [*0.1777*]
2nd PC (FP)	0.3858 [*0.2155*]	0.1226 [*0.7042*]	−0.2474 [*0.4382*]
3rd PC (FP)	−0.2826 [*0.3735*]	−0.4014 [*0.1960*]	−0.1824 [*0.5705*]

Pearson's correlation scores computed between the principal components (PC) of mutation sites (MS) and food availability patterns (FP), with the corresponding *P*-values (square brackets). Significant correlations are highlighted in bold and the corresponding r-coefficient values are marked with an asterisk (*).

doi:10.1371/journal.pone.0006824.t007

Hierarchical clustering

Distance matrices for MS, MT and FP were computed by pairwise comparison between *TP53* countries/geographic areas using the squared Euclidean distance. Dendrograms were constructed through Ward's hierarchical clustering algorithm [113]. Stability of clusters was assessed evaluating the silhouette values [114] that measure how close each point in one cluster is to the points in the neighboring clusters. This measure ranges from +1, indicating points very distant from neighboring clusters, through 0, indicating points not distinctly in one cluster or another, to −1, indicating points probably assigned to the wrong cluster.

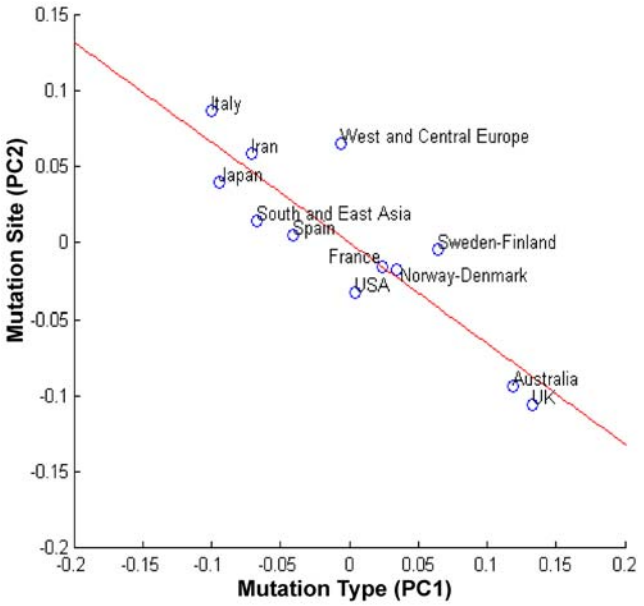


Figure 7. Scatter plot of the correlation between mutation sites and types according to countries/geographic areas. Scatter plot of the countries/geographic areas in the *TP53* database projected on the 2-dimensional space spanned by the first principal components of mutation sites (Mutation Site PC2) and mutation types (Mutation Type PC1). Italy, Iran, South and East Asia and West and Central Europe have relatively lower frequencies of mutations at CpG hotspot codons, compensated by higher frequencies of mutations at all other sites (see also box-plots in Figure 5). Mutation frequencies at CpG hotspots increase in other countries, with highest frequencies in Australia and UK. A linear regression shows the global trend of the correlation ($r = -0.8742$).

doi:10.1371/journal.pone.0006824.g007

Matrices for MS, MT and FP were tested for correlation by Mantel's test [115]. The program Mantel version 3.1 was used to estimate Pearson correlation coefficients. Significance was assessed by 10,000 random permutations.

Feature selection

Feature selection involved the use of a similarity measure between hierarchical clusterings, visualized as dendrograms, respectively built on the entire feature set and on the feature subset(s) to be tested. The higher the similarity, the higher the rank of the chosen feature subset. The entropy-based similarity measure used is defined below.

Two clusterings are identical if there is one-to-one correspondence between their clusters. The more a cluster of one clustering is filled with objects from different clusters of the other clustering (disorder), the less is the concordance between clusterings. All the information needed to summarize this phenomenon is the corresponding confusion matrix. Given two clusterings, A and B, where A is made of n clusters and B of m clusters, the confusion matrix M between A and B is an $n \times m$ matrix, in which the entry (i,j) reports the number of objects in the cluster i of A falling into the cluster j of B. Entropy is the obvious tool to measure such disorder. If R_i is the i-th row of M and C_j is the j-th column of M, then $H(R_i)$ measures the disorder of the i-th cluster of A with respect to B, and $H(C_j)$ measures the disorder of the j-th cluster of B with respect to A.

A way to compute the similarity between B and A is the mean entropy of the clusters of B versus A, where the *a priori* probability of a cluster X, $p(X)$, can be approximated as $number-of-objects-in-X/total-number-of-objects$, giving the formula:

$$S(M) = \sum_i p(X_i) \cdot H(R_i)$$

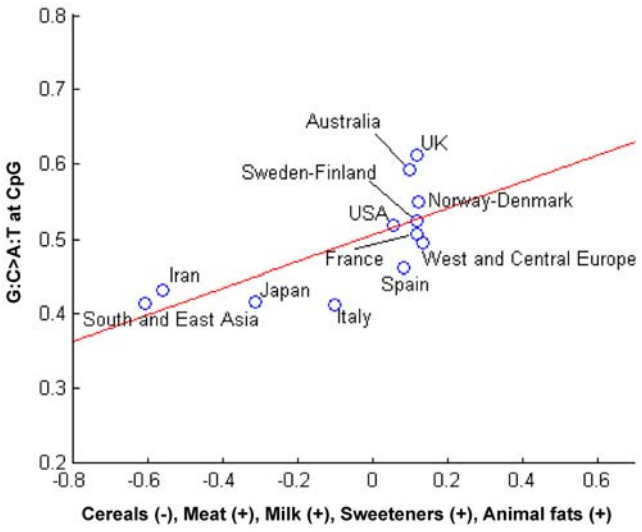


Figure 8. Scatter plot of the correlation between mutation types and food patterns according to countries/geographic areas. Scatter plot of the countries/geographic areas in the *TP53* database projected on the 2-dimensional space spanned by the highest coefficient features of the first principal component (PC) of food availability pattern (FP), *i.e.*, cereals, meat, milk, sweeteners, animal fats, and of the first PC of mutation type (MT), *i.e.*, G:C>A:T at CpGs. A linear regression shows the global correlation trend.

doi:10.1371/journal.pone.0006824.g008

expressing the similarity of B versus A, while the similarity of A versus B can be obtained with the analogue formula on C_j, which turns to be $S(M^T)$. The measure of similarity between clusterings is in the trade-off between $S(M)$ and $S(M^T)$. We define the final similarity measure:

$$S_a(M) = S(M) + a \cdot S(M^T)$$

where a, $0 \leq a \leq 1$, can be used to set the acceptable level of 'sub-clusteringness' of B with respect to A. When $a=0$, no importance is given to the fragmentation level of the clusters in B. When $a=1$ only exact matching between A and B will produce a maximum for S_a.

Basing on such similarity measure between clusterings, useful comparisons between dendrograms can be easily performed. Given a solution obtained from a dendrogram (the target solution), it is possible to assess how much such solution can be approximated by another dendrogram.

Given a dendrogram D, let $\delta(D)$ be the clustering solution obtained applying a cutting threshold δ to D. We define *complete threshold set* for a dendrogram any minimal set of threshold values, applying which all the possible clusterings for the dendrogram can be obtained. We indicate any such set for a dendrogram D by $\Delta(D)$. It can be easily shown that $|\Delta(D)| = N(D)+1$, where $N(D)$ is the number of nodes in D.

Given a dendrogram D', a target solution T can be derived applying a cutting threshold. The similarity between D' and another dendrogram, D, can be approximated using the dendrogram similarity procedure.

Dendrogram similarity procedure (T, D, a)

$i \leftarrow 1$

for each δ ***in*** $\Delta(D)$

Build M***, the confusion matrix between*** T ***and*** $\delta(D)$

$S(i) \leftarrow S_a(M)$

$i \leftarrow i+1$

return $\min(S)$

Exhaustive approach to feature selection for the MT and FP datamatrices

Feature analysis studies the properties of single features or subsets of features of the analyzed data. Exact feature analysis can be performed testing the properties of each possible subset of features. In this study, the property of interest was the ability to maintain the groups obtained in the clustering analysis phase. Such exhaustive approach was successfully performed on the MT and FP datamatrices.

Given a set of features F and a scoring function $f : \wp(F) \rightarrow \Re$, the exhaustive feature analysis approach consists in computing $f(A)$, $\forall A \in \wp(F)$. We performed this analysis using the features of the MT and FP datamatrices in turn for F and *dendrogramSimilarity*$(T, D(A), 1)$ for $f(A)$, where T is the solution obtained in the clustering analysis phase of the data and $D(A)$ is the dendrogram built using the features subset $A \subseteq F$.

The results of exhaustive feature analysis are reported in Tables 1 and 2 for the MT datamatrix and in Tables 3 and 4 for the FP datamatrix. In Tables 1 and 3 the entry i, j reports the worst score obtained using $A = \{x_i \cup B\}$, where $x_i \in F$ and $|B| = i$. In Tables 2 and 4 the *i-th* entry reports the set of features C such that $f(\{i,j\}) = 1$, $\forall j \in C$.

Heuristic approach to feature selection for the MS datamatrix

Being the number of MS feature subsets equal to 2^{173}, we used the sequential forward selection approach for MS feature selection. A filter method was used.

Given a feature set F and a scoring function $f : x \rightarrow \Re$, $x \in F$, a ranking of features can be obtained computing and sorting $\{f(x), \forall x \in F\}$. Let such ordered set be: $\{x_{i_1}, x_{i_2}, \ldots, x_{i_n}\}$, $f(x_{i_1}) \geq f(x_{i_2}) \geq \ldots \geq f(x_{i_n}), x_{i_j} \in F$.

Instead of producing all possible subsets of F, we produce the sets $S_1, \ldots S_n$ such that:

$$S_1 = \{x_{i_1}\}$$

$$S_k = S_{k-1} \cup \{x_{i_k}\}, k = 2, \ldots, n.$$

Substituting $\wp(F)$ with $\{S_1, \ldots, S_n\}$ in the exhaustive approach completes the definition of the heuristic approach.

We used this method with two different rankings, respectively based on the number of mutations recorded for each codon (feature); and on the sum of the first 11 Principal Components (PCs) coefficients of each feature (where 11 was the number of PCs contributing 100% of the data variance). Panels A and B of Figure 3 report the f values obtained for the two different ranking functions. Panel C of the same Figure compares the best feature sets (minimal stable subsets giving $f=1$).

Geographic visualization of feature relevance

Geographic visualizations of the most relevant features of the MS, MT and FP datamatrices were obtained by respectively summing feature frequencies (for MS and MT) and *per caput* supply of each food group expressed as % of the total available calories, as detailed above [50,17] (for FP). Resulting values were projected into yellow to red color range onto the geographic profiles of the countries and geographic areas contributing to the *TP53* mutation database.

Correlation analyses

To perform a multivariate correlation analysis between the PCs of MS, MT and FP, we exploited their projections on the respective PCs. Both the Scree and the Kaiser [116] tests provided clear support for extracting the first 11 components for MS. When applied to MT, these tests supported the extraction of 4 and 3 PCs respectively, being the eigenvalue of the fourth PC near the lower limit value (*i.e.*, 0.9). For FP the Scree and Kaiser tests indicated 3 and 4 PCs, respectively. Pairwise Pearson correlations were then computed between the PCs in all the projected spaces.

Supporting Information

File S1 Coefficient loadings of the three most relevant PCs of mutation sites (MS), mutation types (MT) and food patterns (FP) and Pearson's correlation scores computed between the PCs of MS, MT and FP.

Found at: doi:10.1371/journal.pone.0006824.s001 (0.04 MB DOC)

Figure S1 Coefficient loadings of the first three PCs of the mutation sites, mutation types and food patterns datamatrices. Coefficient loadings of the three most relevant principal components (PCs) of the mutation sites (MS, A–C), mutation types (MT, D–F) and food availability patterns (FP, G–I) datamatrices are projected on their 1-dimensional space (see File S1 for discussion).
Found at: doi:10.1371/journal.pone.0006824.s002 (0.71 MB TIF)

Figure S2 Assignment of Japan to clusters I or II in cluster analysis for food availability patterns. The food category "cereals" determined clusterization of Japan with Western countries for food availability patterns. Histograms visualize the number of times that each of the 13 features was present in the 2,405 clusterings classified as type A, i.e, where Japan joined Iran and South and East Asia in cluster II-FP (A), or in the 4,178 clusterings classified as type B, i.e, where Japan joined Western countries in cluster I-FP (B). It is readily evident that feature 3 (cereals) was almost always absent in type A clusterings and almost always present in type B clusterings. This reflects the estimated low mean per caput supply of cereals available for human consumption in Japan, compared to the countries/geographic areas in the II-FP cluster (i.e., Iran and South and East Asia). Features 1 to 13 represent the following food categories: 1, animal fats; 2, animal products; 3, cereals; 4, fish/seafood; 5, fruit; 6, meat; 7, milk; 8, oilcrops; 9, pulses (legumes); 10, starchy roots; 11, sweeteners; 12, vegetable oils; 13, vegetables.
Found at: doi:10.1371/journal.pone.0006824.s003 (0.15 MB TIF)

Dataset S1 1990 Food Balance Sheet data - Estimated mean per caput supply of each major food group available for human consumption in the *TP53* database countries, as extracted from the Food Balance Sheets (FBS) of the Food and Agriculture Organization of the United Nations (FAO), year 1990.
Found at: doi:10.1371/journal.pone.0006824.s004 (0.02 MB XLS)

Datamatrices S1 Normalized datamatrix for *TP53* mutation sites (MS) and mutation types (MT) assigned to 12 countries/geographic areas and normalized datamatrix for the estimated food availability patterns (FP) of the 12 countries/geographic areas in the *TP53* database. Data from 2,572 *TP53* exons 5–8 mutations associated with primary CRCs were retrieved from the *TP53* database of the International Agency for Research on Cancer (IARC), R10 update, July - 2005, and from Mahdavinia et al., J Cell Physiol. 2008; 216(2):543–550.
Found at: doi:10.1371/journal.pone.0006824.s005 (0.07 MB XLS)

Database S1 List of 2,572 *TP53* exons 5–8 mutations reported in association with primary colorectal cancers, including 2,475 from 11 countries/geographic areas, extracted from the *TP53* database of the International Agency for Research on Cancer (IARC), R10 update, July 2005, and 97 from Iran (Mahdavinia et al., J Cell Physiol. 2008;216:543–550).
Found at: doi:10.1371/journal.pone.0006824.s006 (0.61 MB XLS)

Acknowledgments

We acknowledge the International Agency for Research on Cancer (IARC) and the Food and Agriculture Organization of the United Nations (FAO), for their investment in the collection and free distribution of the data that made the present study possible. We also thank Dr. Angela Polito, Human Nutrition Unit, National Research Institute for Food and Nutrition (*INRAN*), Rome, Italy, for assisting with FAO data.

Links: TP53 database of the International Agency for Research on Cancer (IARC), http://www-p53.iarc.fr/Somatic.html; Food Balance Sheets (FBS) of the Food and Agriculture Organization of the United Nations (FAO), http://faostat.fao.org/site/368/DesktopDefault.aspx?PageID=368.

Author Contributions

Conceived and designed the experiments: FV AC RT RMC. Performed the experiments: FV FB FN MM. Analyzed the data: FV FB FN MM AC GM GR RT RMC. Contributed reagents/materials/analysis tools: FV AC RM GM GR RT RMC. Wrote the paper: FV FB FN MM RM GR RT RMC.

References

1. Vousden KH, Lu X (2002) Live or let die: the cell's response to p53. Nat Rev Cancer 2(8): 594–604.
2. Sengupta S, Harris CC (2005) P53: traffic cop at the crossroads of DNA repair and recombination. Nat Rev Mol Cell Biol 6: 44–55.
3. Crighton D, Wilkinson S, O'Prey J, Syed N, Smith P, et al. (2006) DRAM, a p53-induced modulator of autophagy, is critical for apoptosis. Cell 126: 121–134.
4. Royds JA, Iacopetta B (2006) P53 and disease: when the guardian angel fails. Cell Death Differ 13: 1017–1026.
5. Matoba S, Kang JG, Patino WD, Wragg A, Boehm M, et al. (2006) P53 regulates mitochondrial respiration. Science 312: 1650–1653.
6. Bensaad K, Vousden KH (2007) P53: new roles in metabolism. Trends Cell Biol 17: 286–291.
7. Hainaut P, Hollstein M (2000) P53 and human cancer: the first ten thousand mutations. Adv Cancer Res 77: 81–137.
8. Olivier M, Eeles R, Hollstein M, Khan MA, Harris CC, et al. (2002) The IARC TP53 database: new online mutation analysis and recommendations to users. Hum Mutat 19: 607–14.
9. Soussi T, Kato S, Levy PP, Ishioka C (2005) Reassessment of the TP53 mutation database in human disease by data mining with a library of TP53 missense mutations. Hum Mutat 25: 6–17.
10. Petitjean A, Achatz MI, Borresen-Dale AL, Hainaut P, Olivier M (2007) TP53 mutations in human cancers: functional selection and impact on cancer prognosis and outcomes. Oncogene 26: 2157–65.
11. Petitjean A, Mathe E, Kato S, Ishioka C, Tavtigian SV, et al. (2007) Impact of mutant p53 functional properties on TP53 mutation patterns and tumor phenotype: lessons from recent developments in the IARC TP53 database. Hum Mutat 28: 622–9.
12. Pfeifer GP, Denissenko MF (1998) Formation and repair of DNA lesions in the p53 gene: relation to cancer mutations? Environ Mol Mutagen 31: 197–205.
13. Hussain SP, Harris CC (2000) Molecular epidemiology and carcinogenesis: endogenous and exogenous carcinogens. Mutat Res 462: 311–22.
14. Hainaut P (2002) Tumor-specific mutations in p53: the acid test. Nat Med 8: 21–3.
15. Wright BE, Reimers JM, Schmidt KH, Reschke DK (2002) Hypermutable bases in the p53 cancer gene are at vulnerable positions in DNA secondary structures. Cancer Res 62: 5641–4.
16. Olivier M, Hussain SP, Caron de Fromentel C, Hainaut P, Harris CC (2004) TP53 mutation spectra and load: a tool for generating hypotheses on the etiology of cancer. IARC Sci Publ 157: 247–70.
17. World Cancer Research Fund/American Institute for Cancer Research (1997) Food, nutrition and the prevention of cancer: a global perspective. Washington DC: AICR press.
18. Potter JD (1999) Colorectal cancer: molecules and populations. J Natl Cancer Inst 9: 916–32.
19. Lichtenstein P, Holm NV, Verkasalo PK, Iliadou A, Kaprio J, et al. (2000) Environmental and heritable factors in the causation of cancer–analyses of cohorts of twins from Sweden, Denmark, and Finland. N Engl J Med 343: 78–85.
20. Bingham S, Riboli E (2004) Diet and cancer–the European Prospective Investigation into Cancer and Nutrition. Nat Rev Cancer 4: 206–15.
21. Kolonel LN, Altshuler D, Henderson BE (2004) The multiethnic cohort study: exploring genes, lifestyle and cancer risk. Nat Rev Cancer 4: 519–27.
22. Johnson IT, Lund EK (2007) Review article: nutrition, obesity and colorectal cancer. Aliment Pharmacol Ther 26: 161–81.
23. World Cancer Research Fund/American Institute for Cancer Research (2007) Food, nutrition, physical activity, and the prevention of cancer: a global perspective. Washington DC: AICR press.
24. Parkin DM, Bray F, Ferlay J, Pisani P (2005) Global cancer statistics, 2002. CA Cancer J Clin 55: 74–108.
25. Fearon ER, Vogelstein B (1990) A genetic model for colorectal tumorigenesis. Cell 61: 759–67.
26. Pinto D, Clevers H (2005) Wnt control of stem cells and differentiation in the intestinal epithelium. Exp Cell Res 306: 357–63.

27. Wagner JR, Hu CC, Ames BN (1992) Endogenous oxidative damage of deoxycytidine in DNA. Proc Natl Acad Sci USA 89: 3380–4.
28. Feig DI, Sowers LC, Loeb LA (1994) Reverse chemical mutagenesis: identification of the mutagenic lesions resulting from reactive oxygen species-mediated damage to DNA. Proc Natl Acad Sci USA 91: 6609–13.
29. Kreutzer DA, Essigmann JM (1998) Oxidized, deaminated cytosines are a source of C → T transitions *in vivo*. Proc Natl Acad Sci USA 95: 3578–82.
30. Marnett LJ (2000) Oxyradicals and DNA damage. Carcinogenesis 21: 361–70.
31. Ohshima H, Tatemichi M, Sawa T (2003) Chemical basis of inflammation-induced carcinogenesis. Arch Biochem Biophys 417: 3–11.
32. Iacopetta B (2003) TP53 mutation in colorectal cancer. Hum Mutat 21: 271–6.
33. Russo A, Bazan V, Iacopetta B, Kerr D, Soussi T, et al. (2005) TP53-CRC Collaborative Study Group. The TP53 colorectal cancer international collaborative study on the prognostic and predictive significance of p53 mutation: influence of tumor site, type of mutation, and adjuvant treatment. J Clin Oncol 23: 7518–28.
34. Iacopetta B, Russo A, Bazan V, Kerr D, Soussi T, et al. (2006) Functional categories of TP53 mutation in colorectal cancer: results of an international collaborative study. Ann Oncol 17: 842–7.
35. Baker SJ, Preisinger AC, Jessup JM, Paraskeva C, Markowitz S, et al. (1990) P53 gene mutations occur in combination with 17p allelic deletions as late events in colorectal tumorigenesis. Cancer Res 50: 7717–22.
36. Day N, McKeown N, Wong M, Welch A, Bingham S (2001) Epidemiological assessment of diet: a comparison of a 7-day diary with a food frequency questionnaire using urinary markers of nitrogen, potassium and sodium. Int J Epidemiol 30: 309–17.
37. Johansson G, Wikman A, Ahrén AM, Hallmans G, Johansson I (2001) Underreporting of energy intake in repeated 24-hour recalls related to gender, age, weight status, day of interview, educational level, reported food intake, smoking habits and area of living. Public Health Nutr 4: 919–27.
38. Asbeck I, Mast M, Bierwag A, Westenhöfer J, Acheson KJ, et al. (2002) Severe underreporting of energy intake in normal weight subjects: use of an appropriate standard and relation to restrained eating. Public Health Nutr 5: 683–90.
39. Ferrari P, Slimani N, Ciampi A, Trichopoulou A, Naska A, et al. (2002) Evaluation of under- and overreporting of energy intake in the 24-hour diet recalls in the European Prospective Investigation into Cancer and Nutrition (EPIC). Public Health Nutr 5: 1329–45.
40. Huycke MM, Gaskins HR (2004) Commensal bacteria, redox stress, and colorectal cancer: mechanisms and models. Exp Biol Med (Maywood) 229: 586–97.
41. Goldman R, Shields PG (2003) Food mutagens. J Nutr 133: 965S–973S.
42. Jägerstad M, Skog K (2005) Genotoxicity of heat-processed foods. Mutat Res 574: 156–72.
43. Ishibe N, Freedman AN (2001) Understanding the interaction between environmental exposures and molecular events in colorectal carcinogenesis. Cancer Invest 19: 524–39.
44. Slattery ML, Curtin K, Ma K, Edwards S, Schaffer D, et al. (2002) Diet activity, and lifestyle associations with p53 mutations in colon tumors. Cancer Epidemiol Biomarkers Prev 11: 541–8.
45. Collins AR, Harrington V, Drew J, Melvin R (2003) Nutritional modulation of DNA repair in a human intervention study. Carcinogenesis 24: 511–5.
46. Gunter MJ, Leitzmann MF (2006) Obesity and colorectal cancer: epidemiology, mechanisms and candidate genes. J Nutr Biochem 17: 145–56.
47. Slattery ML (2008) Defining dietary consumption: is the sum greater than its parts? Am J Clin Nutr 88: 14–5.
48. Arasaradnam RP, Commane DM, Bradburn D, Mathers JC (2008) A review of dietary factors and its influence on DNA methylation in colorectal carcinogenesis. Epigenetics 3: 193–8.
49. Mahdavinia M, Bishehsari F, Verginelli F, Cumashi A, Lattanzio R, et al. (2008) P53 mutations in colorectal cancer from northern Iran: Relationships with site of tumor origin, microsatellite instability and K-ras mutations. J Cell Physiol 216: 543–50.
50. Food and Agriculture Organization of the United Nations (2001) Food Balance Sheets: a handbook. Rome, FAO press.
51. Serra-Majem L, MacLean D, Ribas L, Brulé D, Sekula W, et al. (2003) Comparative analysis of nutrition data from national, household, and individual levels: results from a WHO-CINDI collaborative project in Canada, Finland, Poland, and Spain. J Epidemiol Community Health 57: 74–80.
52. Tornaletti S, Pfeifer GP (1995) Complete and tissue-independent methylation of CpG sites in the p53 gene: implications for mutations in human cancers. Oncogene 10: 1493–9.
53. Hu FB. Dietary pattern analysis: a new direction in nutritional epidemiology (2002) Curr Opin Lipidol 13: 3–9.
54. Tornaletti S, Pfeifer GP (1995) Complete and tissue-independent methylation of CpG sites in the p53 gene: implications for mutations in human cancers. Oncogene 10: 1493–9.
55. You YH, Halangoda A, Buettner V, Hill K, Sommer S, et al. (1998) Methylation of CpG dinucleotides in the lacI gene of the Big Blue transgenic mouse. Mutat Res 420: 55–65.
56. Esteller M (2008) Epigenetics in cancer. N Engl J Med 358: 1148–59.
57. Rideout WM 3rd, Coetzee GA, Olumi AF, Jones PA (1990) 5-Methylcytosine as an endogenous mutagen in the human LDL receptor and p53 genes. Science 249: 1288–90.
58. Pfeifer GP (2000) P53 mutational spectra and the role of methylated CpG sequences. Mutat Res 450: 155–66.
59. Pfeifer GP. Mutagenesis at methylated CpG sequences. Curr Top Microbiol Immunol 2006; 301: 259–81.
60. Friso S, Choi SW (2005) Gene-nutrient interactions in one-carbon metabolism. Curr Drug Metab 6: 37–46.
61. Kim YI (2005) Nutritional epigenetics: impact of folate deficiency on DNA methylation and colon cancer susceptibility. J Nutr 135: 2703–9.
62. Ulrich CM, Curtin K, Samowitz W, Bigler J, Potter JD, et al. (2005) MTHFR variants reduce the risk of G:C->A:T transition mutations within the p53 tumor suppressor gene in colon tumors. J Nutr 135: 2462–7.
63. Quinlivan EP, Davis SR, Shelnutt KP, Maneval DR, Ghandour H, et al. (2005) Methylenetetrahydrofolate reductase 677C->T polymorphism and folate status affect one-carbon incorporation into human DNA deoxynucleosides. J Nutr 135: 389–96.
64. Arasaradnam RP, Commane DM, Bradburn D, Mathers JC (2008) A review of dietary factors and its influence on DNA methylation in colorectal carcinogenesis. Epigenetics 3: 193–8.
65. Paul AA, Southgate DAT (1978) McCance and Widdowson's The Composition of Foods. Amsterdam/New York/London: HMSO and Elsevier/North-Holland Biomedical Press.
66. USDA nutrient database for standard reference, release 20 (accessed 2007) Available at: http://www.nal.usda.gov/fnic/foodcomp.
67. Bernstein BE, Meissner A, Lander ES (2007) The mammalian epigenome. Cell 128: 669–81.
68. Bock C, Lengauer T (2008) Computational epigenetics. Bioinformatics 24: 1–10.
69. Drewnowski A, Popkin BM (1997) The nutrition transition: new trends in the global diet. Nutr Rev 55: 31–43.
70. Ghassemi H, Harrison G, Mohammad K (2002) An accelerated nutrition transition in Iran. Public Health Nutr 5: 149–55.
71. Curtin K, Slattery ML, Ulrich CM, Bigler J, Levin TR, et al. (2007) Genetic polymorphisms in one-carbon metabolism: associations with CpG island methylator phenotype (CIMP) in colon cancer and the modifying effects of diet. Carcinogenesis 28: 1672–9.
72. Schernhammer ES, Ogino S, Fuchs CS (2008) Folate and vitamin B(6) intake and risk of colon cancer in relation to p53 expression. Gastroenterology 135: 770–80.
73. Wink DA, Kasprzak KS, Maragos CM, Elespuru RK, Misra M, et al. (1991) DNA deaminating ability and genotoxicity of nitric oxide and its progenitors. Science 254: 1001–3.
74. Goodman JE, Hofseth LJ, Hussain SP, Harris CC (2004) Nitric oxide and p53 in cancer-prone chronic inflammation and oxyradical overload disease. Environ Mol Mutagen 44: 3–9.
75. Sawa T, Ohshima H (2006) Nitrative DNA damage in inflammation and its possible role in carcinogenesis. Nitric Oxide 14: 91–100.
76. Wink DA, Hanbauer I, Grisham MB, Laval F, Nims RW, et al. (1996) Chemical biology of nitric oxide: regulation and protective and toxic mechanisms. Curr Top Cell Regul 34: 159–87.
77. Grisham MB, Pavlick KP, Laroux FS, Hoffman J, Bharwani S, et al. (2002) Nitric oxide and chronic gut inflammation: controversies in inflammatory bowel disease. J Investig Med 50: 272–83.
78. Cross RK, Wilson KT (2003) Nitric oxide in inflammatory bowel disease. Inflamm Bowel Dis 9: 179–89.
79. Gal A, Wogan GN (1996) Mutagenesis associated with nitric oxide production in transgenic SJL mice. Proc Natl Acad Sci USA 93: 15102–7.
80. Lala PK, Chakraborty C (2001) Role of nitric oxide in carcinogenesis and tumour progression. Lancet Oncol 2: 149–56.
81. Lewin MH, Bailey N, Bandaletova T, Bowman R, Cross AJ, et al. (2006) Red meat enhances the colonic formation of the DNA adduct O6-carboxymethyl guanine: implications for colorectal cancer risk. Cancer Res 66: 1859–65.
82. Hughes MN (2008) Chemistry of nitric oxide and related species. Methods Enzymol 436: 3–19.
83. Forrester K, Ambs S, Lupold SE, Kapust RB, Spillare EA, et al. (1996) Nitric oxide-induced p53 accumulation and regulation of inducible nitric oxide synthase expression by wild-type p53. Proc Natl Acad Sci USA 93: 2442–7.
84. Ambs S, Hussain SP, Harris CC (1997) Interactive effects of nitric oxide and the p53 tumor suppressor gene in carcinogenesis and tumor progression. FASEB J 11: 443–8.
85. Ambs S, Merriam WG, Bennett WP, Felley-Bosco E, Ogunfusika MO, et al. (1998) Frequent nitric oxide synthase-2 expression in human colon adenomas: implication for tumor angiogenesis and colon cancer progression. Cancer Res 58: 334–41.
86. Mihara M, Erster S, Zaika A, Petrenko O, Chittenden T, et al. (2003) P53 has a direct apoptogenic role at the mitochondria. Mol Cell 11: 577–90.
87. Yerushalmi HF, Besselsen DG, Ignatenko NA, Blohm-Mangone KA, Padilla-Torres JL, et al. (2006) The role of NO synthases in arginine-dependent small intestinal and colonic carcinogenesis. Mol Carcinog 45: 93–105.
88. Warren W, Biggs PJ, el-Baz M, Ghoneim MA, Stratton MR, et al. (1995) Mutations in the p53 gene in schistosomal bladder cancer: a study of 92 tumours from Egyptian patients and a comparison between mutational spectra from schistosomal and non-schistosomal urothelial tumours. Carcinogenesis 16: 1181–9.

89. Ambs S, Bennett WP, Merriam WG, Ogunfusika MO, Oser SM, et al. (1999) Relationship between p53 mutations and inducible nitric oxide synthase expression in human colorectal cancer. J Natl Cancer Inst 91: 86–8.
90. Vaninetti NM, Geldenhuys L, Porter GA, Risch H, Hainaut P, et al. (2008) Inducible nitric oxide synthase, nitrotyrosine and p53 mutations in the molecular pathogenesis of Barrett's esophagus and esophageal adenocarcinoma. Mol Carcinog 47: 275–85.
91. Castillo L, DeRojas TC, Chapman TE, Vogt J, Burke JF, et al. (1993) Splanchnic metabolism of dietary arginine in relation to nitric oxide synthesis in normal adult man. Proc Natl Acad Sci USA 90: 193–197.
92. Zell JA, Ignatenko NA, Yerushalmi HF, Ziogas A, Besselsen DG, et al. (2007) Risk and risk reduction involving arginine intake and meat consumption in colorectal tumorigenesis and survival. Int J Cancer 120: 459–68.
93. Gerner EW (2007) Impact of dietary amino acids and polyamines on intestinal carcinogenesis and chemoprevention in mouse models. Biochem Soc Trans 35: 322–5.
94. Broughton KS, Wade JW (2002) Total fat and (n-3):(n-6) fat ratios influence eicosanoid production in mice. J Nutr 132: 88–94.
95. Joseph SB, Castrillo A, Laffitte BA, Mangelsdorf DJ, Tontonoz P (2003) Reciprocal regulation of inflammation and lipid metabolism by liver X receptors. Nat Med 9: 213–9.
96. Lopez-Garcia E, Schulze MB, Fung TT, Meigs JB, Rifai N (2004) Major dietary patterns are related to plasma concentrations of markers of inflammation and endothelial dysfunction. Am J Clin Nutr 80: 1029–35.
97. Giugliano D, Ceriello A, Esposito K (2006) The effects of diet on inflammation: emphasis on the metabolic syndrome. J Am Coll Cardiol 48: 677–85.
98. Innis SM, Jacobson K (2007) Dietary lipids in early development and intestinal inflammatory disease. Nutr Rev 65: S188–93.
99. Kallio P, Kolehmainen M, Laaksonen DE, Pulkkinen L, Atalay M, et al. (2008) Inflammation markers are modulated by responses to diets differing in postprandial insulin responses in individuals with the metabolic syndrome. Am J Clin Nutr 87: 1497–503.
100. Devaraj S, Wang-Polagruto J, Polagruto J, Keen CL, Jialal I (2008) High-fat, energy-dense, fast-food-style breakfast results in an increase in oxidative stress in metabolic syndrome. Metabolism 57: 867–70.
101. Arulampalam V (2008) Gastrointestinal inflammation: lessons from metabolic modulators. J Intern Med 263: 607–12.
102. Bingham SA, Pignatelli B, Pollock JR, Ellul A, Malaveille C, et al. (1996) Does increased endogenous formation of N-nitroso compounds in the human colon explain the association between red meat and colon cancer? Carcinogenesis 17: 515–23.
103. Tricker AR (1997) N-nitroso compounds and man: sources of exposure, endogenous formation and occurrence in body fluids. Eur J Cancer Prev 6: 226–68.
104. Rhodes JM, Campbell BJ (2002) Inflammation and colorectal cancer: IBD-associated and sporadic cancer compared. Trends Mol Med 8: 10–16.
105. Cross AJ, Pollock JR, Bingham SA (2003) Haem, not protein or inorganic iron, is responsible for endogenous intestinal N-nitrosation arising from red meat. Cancer Res 63: 2358–60.
106. Cross AJ, Sinha R (2004) Meat-related mutagens/carcinogens in the etiology of colorectal cancer. Environ Mol Mutagen 44: 44–55.
107. Oates PS, West AR (2006) Heme in intestinal epithelial cell turnover, differentiation, detoxification, inflammation, carcinogenesis, absorption and motility. World J Gastroenterol 12: 4281–95.
108. MacFarlane AJ, Stover PJ (2007) Convergence of genetic, nutritional and inflammatory factors in gastrointestinal cancers. Nutr Rev 65: S157–66.
109. Magewu AN, Jones PA (1994) Ubiquitous and tenacious methylation of the CpG site in codon 248 of the p53 gene may explain its frequent appearance as a mutational hot spot in human cancer. Mol Cell Biol 14: 4225–32.
110. Souici AC, Mirkovitch J, Hausel P, Keefer LK, Felley-Bosco E (2000) Transition mutation in codon 248 of the p53 tumor suppressor gene induced by reactive oxygen species and a nitric oxide-releasing compound. Carcinogenesis 21: 281–7.
111. Hao XP, Frayling IM, Sgouros JG, Du MQ, Willcocks TC, et al. (2002) The spectrum of p53 mutations in colorectal adenomas differs from that in colorectal carcinomas. Gut 50: 834–839.
112. Parkin DM, Whelan SL, Ferlay J, Raymond L, Young J, eds. (1997) Cancer incidence in five continents, vol. VII. IARC Scientific Publications No. 143, Lyon: IARC Press.
113. Ward JH (1963) Hierarchical Grouping to optimize an objective function. J Am Stat Assoc 58: 236–244.
114. Kaufman L, Rousseeuw PJ (1990) Finding groups in data: an introduction to cluster analysis. Wiley Press.
115. Mantel N, Valand R (1970) A technique of non parametric multivariate analysis. Biometrics 26: 547–558.
116. Kaiser HF (1960) The application of electronic computers to factor analysis. Educational and Psychological Measurement 20: 141–151.