Article

Cancer Etiology: Variation in Cancer Risk among Tissues is Poorly Explained by the Number of Gene Mutations

Miguel López-Lázaro

Department of Pharmacology, Faculty of Pharmacy, University of Seville, C/ Profesor Garcia Gonzalez 2, 41012 Sevilla, Spain; mlopezlazaro@us.es; Tel: +34-954-55-63-48. Fax: +34-954-23-7-65

Abstract: Recent evidence indicates that the risk of being diagnosed with cancer in a tissue is strongly correlated (0.80) with the number of stem cell divisions accumulated by the tissue. Since cell division can generate random mutations during DNA replication, this correlation has been used to propose that cancer is largely caused by the accumulation of unavoidable mutations in driver genes. However, no correlation between the number of gene mutations and cancer risk across tissues has been reported. Because many somatic mutations in cancers originate prior to tumor initiation and the number of cell divisions occurring during tumor growth is similar among tissues, here I use whole genome sequencing information from 22,086 cancer samples and incidence data from the largest cancer registry in each continent to study the relationship between the number of gene mutations and the risk of cancer across 33 tissue types. Results show a weak positive correlation (mean = 0.14) between these two parameters in each of the five cancer registries. The correlation became stronger (mean = 0.50) when gender-related cancers were excluded. Results also show that 1,003 samples from 29 cancer types have zero mutations in genes. These data suggest that cancer etiology can be better explained by the accumulation of stem cell divisions than by the accumulation of gene mutations. Possible mechanisms by which the accumulation of cell divisions in stem cells increases the risk of cancer are discussed.

Keywords: stem cells; somatic mutations; cancer prevention; carcinogenesis; whole genome sequencing; stem cell division theory of cancer; bad luck of cancer

1. Introduction

All cancer registries show large variations in cancer incidence by age and among tissues; cancer incidence increases dramatically with age, and some cancers occur much more frequently in some tissues than in others [1]. Understanding these differences in cancer incidence is crucial to understand the etiology of the disease. Analyses of cancer statistics by age carried out over sixty years ago revealed that cancer formation requires the multistep accumulation of DNA changes [2,3]. The striking increase in cancer incidence with age indicates that cancer is the result of the multistep accumulation of several cellular changes [3]. If cancer were caused by one cellular change or by several changes occurring in one step, these changes could occur at any moment in life, and cancer incidence would be similar at different ages. Because DNA is the only cellular component that can accumulate and transmit cellular changes throughout life, it is well known that carcinogenesis requires the gradual accumulation of DNA changes. According to the prevailing theory of carcinogenesis, these DNA changes are mutations in genes that control cell survival, cell fate and genome maintenance. Cancer would be the result of the stepwise acquisition of driver gene mutations that successively increase cell proliferation [4-7]. Evidence supporting and challenging this prevailing theory of cancer has been discussed elsewhere [4-14].

Recent analyses of the large variation in cancer incidence among tissues have provided another important clue on the etiology of cancer. Tomasetti and Vogelstein [15] studied the relationship between the lifetime number of stem cell divisions in a tissue and the risk of being diagnosed with cancer in the tissue. Using cancer incidence data from the US population, they found a strong correlation (Pearson's linear correlation = 0.804; $P < 5.15 \times 10^{-8}$) between these two parameters across 31 tissue types [15]. More recently, Tomasetti *et al.* [16] reported that this

1

correlation was similar across 17 tissue types in many countries of the world, even though their people are often exposed to different environmental carcinogens. These data suggest that the accumulation of cell divisions in stem cells is the main biological cause of cancer [14,17]. However, since cell division can generate unavoidable mutations during DNA replication, the authors considered that the number of stem cell divisions was equivalent to the number of random mutations arising during DNA replication, and proposed that the main biological cause of cancer is the accumulation of unavoidable mutations in driver genes [15,16]. This proposal has generated an intense and unresolved debate about the etiology of cancer and the preventability of the disease [18-21].

According to this controversial proposal, the number of stem cell divisions is equivalent to the number of mutations. Therefore, the variation in cancer risk across tissues could also be explained by the number of mutations. Unfortunately, whole-genome sequencing information for normal tissues is scarce due to the difficulty to obtain sufficient DNA for accurate analysis. To circumvent this limitation, Blokzijl et al. [22] recently used organoid cultures derived from 19 donors of different ages and observed similar mutation rates in three tissue types. However, the number of samples and tissues used in this investigation was relatively low and therefore insufficient to evaluate if the variation in cancer risk across tissues can be explained by the number of gene mutations. An alternative approach to test this proposal is to use sequencing information from cancer samples; large amounts of whole-genome sequencing data for many cancer types have become available in recent years. This approach has already been used to identify mutational rates and signatures in normal tissues [23]. It is accepted that many mutations found in cancer samples occur before malignant transformation [23,24]. It is also known that tumors from different tissues usually have similar sizes when they are diagnosed, removed and analyzed. Since tumor size depends on the number of cell divisions, the number of mutations occurring during tumor formation would be rather similar in different cancer types. This baseline number of mutations would not occult the number of mutations occurring before malignant transformation in tissues with different renewal capacities. Cancers arising from tissues with higher lifetime numbers of stem cell divisions would accumulate higher numbers of mutations. Most of these mutations would occur in non-protein-coding DNA (which may represent over 98% of the DNA), but other mutations would occur in genes (protein-coding DNA). Therefore, if the number of stem cell divisions is equivalent to the number of mutations arising during DNA replication, the variation in cancer risk across tissues would be explained not only by the number of stem cell divisions [15,16], but also by the number of gene mutations. Using whole genome sequencing information from 22,086 cancer samples and incidence data from five cancer registries, here I test this proposal.

2. Materials and Methods

2.1. Cancer registries

Cancer incidence data were obtained from 'Cancer Incidence in Five Continents Vol. X' [1]. This open access publication contains consistent cancer incidence data for numerous tissue types, including those with a high cancer risk (e.g., lung, breast and prostate) and those with a low cancer risk (e.g., small intestine, bone and eye). In this work, the largest cancer registry from each continent was selected: 1) USA, NPCR (National Program of Cancer Registries), 42 states (263 million people), 2) United Kingdom, England (50 million people), 3) Republic of Korea (48 million people), 4) Australia, New South Wales (6.9 million people) and 5) Egypt, Gharbiah (4 million people). Detailed information about these cancer registries, including geographical areas and age distribution of the population, can be found (on pages 218, 858, 582, 884 and 128, respectively) in reference [1].

2.2. Whole-genome sequenced samples

Whole genome sequencing information was obtained from the 'Catalogue of somatic mutations in cancer (COSMIC)' database [25] (v82). This database contains detailed gene mutation information for thousands of cancer samples across many tissue types. It also allows the selection of mutational data from whole-genome screens. The database uses the term 'Whole Genome Screen' to describe any study which has surveyed all genes in the genome (in contrast to 'Targeted Screen', which surveys a subset of genes). This term does not differentiate between whole genome sequencing and whole exome sequencing. This database is freely available at https://cancer.sanger.ac.uk/cosmic. Whole genome sequencing data were obtained from the 'Samples' section of the 'Cancer browser' tool.

2.3. Selection of tissues

This study includes tissues for which both cancer incidence data and whole-genome sequencing information were available in the 'Cancer Incidence in Five Continents Vol. X' [1] and in the COSMIC database [25]. For most tissues, information was organized under the same headings in both data sources and tissue selection was straightforward. For other tissues, however, information was organized differently in each collection of data. For example, tissue types under the 'Head and neck' heading were different in the two datasets. Subheadings were therefore inspected to find tissue matches. This led, in some cases, to the selection of a part of the tissue instead of the whole tissue. For example, incidence data and whole-genome information were available for nasopharynx, but not for pharynx or hypoparynx; therefore, this study includes data for nasopharynx only. In other cases, minor assumptions were made to avoid the exclusion of important information. For example, data for corpus uteri were available in the cancer registries but not in the COSMIC database, and data for endometrial tissue were available in the COSMIC database but not in the cancer registries. Since the endometrium is one of the two main layers of the corpus of the uterus and virtually all cancers occurring in the corpus of this tissue arise from the endometrium, incidence data for corpus uteri were matched with whole-genome data for endometrium. Lung in the COSMIC database includes lung and bronchus, but not trachea (which in included in 'upper aerodigestive tract'). In the cancer registries, however, incidence data for trachea, bronchus and lung are grouped all together. Since tracheal cancer is extremely rare [26], lung data were not excluded from this study despite this not perfect match. Myeloid leukemia incidence data were matched with whole-genome information for acute myeloid leukemia and chronic myeloid leukemia. For lymphoid leukemia, sequencing data for acute lymphoblastic B cell leukemia, acute lymphoblastic T cell leukemia and chronic lymphocytic leukemia were used. Non-Hodgkin lymphoma incidence data were matched with whole-genome information for anaplastic large cell lymphoma, angioimmunoblastic T cell lymphoma, Burkitt lymphoma, diffuse large B cell lymphoma, follicular lymphoma, mantle cell lymphoma, marginal zone lymphoma and NK-T cell lymphoma.

2.4. Data acquisition and analysis

Cancer incidence data were obtained from section 'Age-standardized and cumulative incidence rates and standard errors' (pages 917- 1252) of reference [1]. Male and female cumulative (0-74) incidence (percent) rates were selected for analysis. Because the proportion of males and females is rather similar in each of the five populations included in this study [1], incidence data for males and females were averaged for all cancers occurring in both sexes, except for breast cancer (which is very rare in males).

Two types of data were obtained from the COSMIC database for each tissue type: number of samples with and without gene mutations, and number of mutations found in each mutant sample. Data were downloaded in CSV format and organized in excel datasheets. Then, the average number of gene mutations for each cancer type was calculated using two functions: mean and median. The mean was calculated by summing the number of mutations found in each sample of a tissue and then dividing by the total number of samples available for the tissue. The median is the middle number of a group of numbers. For example, if 5 samples are available for a tissue and the number of mutations is 1, 2, 3, 44 and 50, the mean is 20 and the median is 3. If the numbers are 1, 2, 3, 4 and 190, the mean is 40 and the median is 3. Using both functions is important to detect a distorted central tendency caused, for example, by tissue samples with unusually high numbers of mutations.

Pearson correlation coefficient was used to measure the strength and direction of the linear relationship between pairs 'cancer risk' and 'number of gene mutations' across the selected tissues. To determine whether the correlation was statistically significant, *P*-values (two sided) were used. One of the software used to calculate Pearson correlation and *P*-values can be freely accessed at https://www.wessa.net/rwasp_correlation.wasp/.

3. Results and Discussion

3.1. Whole-genome sequenced samples with and without gene mutations in 33 cancer types

This study includes 22,086 whole-genome sequenced samples from 33 cancer types. Of them, 21,083 samples had at least one gene mutation, and 1003 samples (4.5%) did not have any mutation in any gene. Data for each tissue are available in Table 1. This table shows a variable proportion of samples from 29 of the 33 cancer types with zero

gene mutations. These data do not support the widely accepted idea that cancer formation requires the accumulation of gene mutations. For example, it has been proposed that three driver gene mutations are required for the development of lung cancer [27]. However, Table 1 shows that 113 of the 1,499 lung cancer samples do not have any gene mutation. In addition, 36 lung cancer samples have 1 gene mutation and 13 samples have 2 mutations. A large variation in the number of gene mutations was found for the rest of lung cancer samples; 443 samples have 3-100 mutations, 840 samples have 101-1,000 mutations, and 64 samples have more than 1,000 mutations (data not shown). If a gene mutation is a change in the nucleotide sequence in protein-coding DNA, the information provided by the COSMIC database indicates that the accumulation of gene mutations in a cell is not mandatory for many cancers to occur. Although this idea is not new [13,28], this work provides the numbers for 33 cancer types (Table 1). In addition, the large variation in the number of gene mutations observed for most cancer types suggests that understanding cancer in terms of several mutations in specific genes may not be possible.

Table 1. Number of whole-genome sequenced samples with and without gene mutations in 33 cancer types.

Cancer type	Whole-genome sequenced samples			
	Total	with gene mutations	without gene mutations	% without gene mutations
Adrenal gland	358	356	2	0.6
Bladder	392	384	8	2.0
Bone	559	412	147	26.3
Brain, CNS	1,921	1,836	85	4.4
Breast	2,080	2,026	54	2.6
Cervix uteri	190	183	7	3.7
Corpus uteri	296	289	7	2.4
Esophagus	1,028	1,006	22	2.1
Eye	39	39	0	0.0
Gallbladder	71	67	4	5.6
Hodgkin lymphoma	6	0	6	100
Kidney	1,261	1,241	20	1.6
Large intestine	1,553	1,520	33	2.1
Larynx	26	26	0	0
Liver	2,136	2,097	39	1.8
Lung	1,499	1,386	113	7.5
Lymphoid leukemia	851	824	27	3.2
Melanoma	861	822	39	4.5
Mouth	367	358	9	2.5
Myeloid leukemia	814	789	25	3.1
Nasopharynx	167	167	0	0
Non-Hodgkin lymphoma	504	441	63	12.5
Ovary	698	671	27	3.9
Pancreas	1,441	1,321	120	8.3
Prostate	1,519	1,425	94	6.2
Salivary glands	79	76	3	3.8
Sinonasal and nasal cavity	3	2	1	33.3
Small intestine	52	46	6	11.5
Stomach	661	635	26	3.9
Testis	20	18	2	10
Thyroid	630	619	11	1.7
Tonsils	1	1	0	0
Vulva	3	0	3	100
TOTAL	22,086	21,083	1,003	4.5

3.2. Correlation between the number of gene mutations and cancer risk across 33 cancer types

Whole genome sequencing information from 22,086 cancer samples and incidence data from the largest cancer registry in each continent were used to study the relationship between the average number of gene mutations and the lifetime cancer risk across 33 cancer types. The average number of mutations for each cancer type was expressed as mean and median (see methods section). The mean and median numbers of mutations for the 22,086 cancer samples were, respectively, 177 and 61. Data for each cancer type are shown in Table 2.

Table 2. Average number of gene mutations and cumulative cancer risk in 33 cancer types. The cancer code used in reference [1] is shown in column 1. The number of samples for each cancer type can be found in Table 1.

Cancer type (code)	Average number of gene mutations		Cumulative risk of developing cancer (0-74 years)				
	Mean	Median	USA	England	Korea	Australia	Egypt
Adreanal gland (C74)	23.5	10	0.02	0.02	0.025	0.03	0.035
Bladder (C67)	216.2	140.5	1.575	1.44	0.655	0.625	1.52
Bone (C40-41)	48.6	16	0.08	0.07	0.07	0.075	0.08
Brain, CNS (C70-72)	44.3	28	0.535	0.545	0.28	0.575	0.365
Breast (C50)	95.0	54	9.87	9.66	3.2	9.41	4.95
Cervix uteri (C50)	261.2	127	0.63	0.62	1.34	0.52	0.18
Corpus uteri (C54)	869.9	121.5	2.07	1.57	0.42	1.3	0.48
Esophagus (C15)	152.9	135	0.435	0.75	0.52	0.38	0.15
Eye (C69)	16.3	13	0.07	0.055	0.02	0.095	0.02
Gallbladder (C23-24)	54.3	22	0.18	0.13	0.81	0.175	0.12
Hodgkin lymphoma (C81)	0	0	0.215	0.19	0.03	0.165	0.14
Kidney (C64)	74.0	65	1.215	0.725	0.5	0.885	0.21
Large intestine (C18-21)	463.5	138	3.665	3.5	3.93	4.82	0.65
Larynx (C32)	205.0	181.5	0.385	0.44	0.28	0.26	0.19
Liver (C22)	191.4	116	0.52	0.31	3.15	0.42	1.925
Lung (C33-34)	289.8	184	5.74	3.79	3.855	3.32	1
Lymphoid leukemia (C91)	23.3	18	0.455	0.375	0.1	0.455	0.235
Melanoma (C43)	645.0	207	1.415	1.115	0.07	3.755	0.035
Mouth (C03-06)	188.9	120	0.18	0.17	0.095	0.2	0.1
Myeloid leukemia (C92-94)	35.5	11	0.405	0.345	0.265	0.42	0.195
Nasopharynx (C11)	121.1	69	0.07	0.035	0.07	0.06	0.055
Non-Hodgkin lymphoma (C82-85,C96)	47.1	14.5	1.47	1.115	0.585	1.39	1.35
Ovary (C56)	77.0	64	1.05	1.46	0.55	0.88	0.59
Pancreas (C25)	71.1	54	0.855	0.72	0.76	0.725	0.4
Prostate (C61)	83.2	47	14.28	8.31	2.01	14.64	0.91
Salivary glands (C07-08)	35.9	16	0.09	0.06	0.06	0.11	0.045
Sinonasal and nasal cavity (C30-31)	20.0	24	0.055	0.045	0.06	0.04	0.03
Small intestine (C17)	221.3	5	0.165	0.09	0.1	0.115	0.055
Stomach (C16)	388.2	135	0.505	0.7	5.475	0.645	0.24
Testis (C62)	18.7	17	0.38	0.48	0.04	0.47	0.04
Thyroid (C73)	171.1	21	0.805	0.22	2.165	0.69	0.26
Tonsils (C09)	19.0	19	0.16	0.105	0.05	0.1	0.005
Vulva (C51)	0	0	0.16	0.17	0.03	0.15	0.11

Pearson linear correlation was used to study the relationship between the average number of gene mutations and cancer risk across 33 tissue types for each of the five cancer registries. Weak and non-statistically significant correlations were found for each cancer registry when the average number of mutations was expressed as mean (Table 3). Slightly higher correlations were found when the average number of mutations was expressed as median. In this case, the correlations remained weak and not statistically significant for 4 cancer registries. However, the correlation became moderate and statistically significant (P <0.05) for the "Republic of Korea" cancer registry (Table 3). The relatively low risk of gender-related cancers such as prostate cancer and breast cancer shown in this registry may explain this difference. Graphical representations are shown in Figure 1 for the largest cancer registry (USA, NPCR, 42 states) and in Figure 2 for the rest of cancer registries.

Table 3. Pearson correlation and statistical significance (*P*-value, 2 sided) between number of gene mutations and cancer risk across 33 tissue types. When the average number of mutations was expressed as mean, the mean correlation was 0.144 and the median correlation was 0.124. When the average number of mutations was expressed as median, the mean correlation was 0.202 and the median correlation was 0.146.

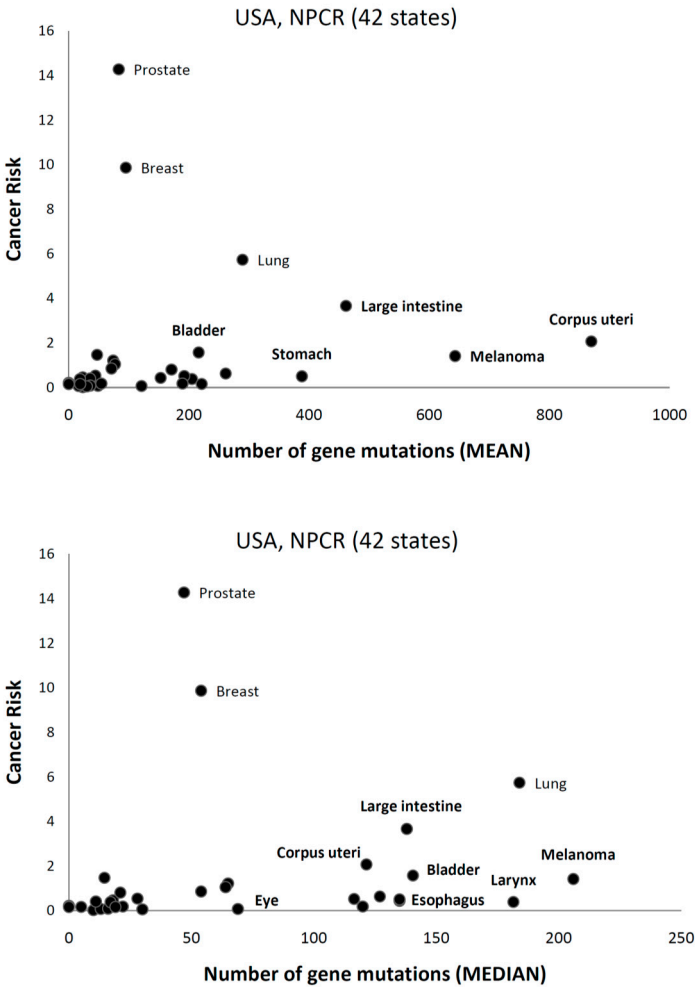| MEAN NUMBER OF GENE MUTATIONS *vs* CANCER RISK		
	Pearson correlation	*P*-value
USA, NPCR (42 states)	0.105	0.563
United Kingdom, England	0.124	0.492
Republic of Korea	0.329	0.062
Australia, New South Wales	0.138	0.443
Egypt, Gharbiah	0.025	0.892
MEDIAN NUMBER OF GENE MUTATIONS *vs* CANCER RISK		
	Pearson correlation	*P*-value
USA, NPCR (42 states)	0.145	0.419
United Kingdom, England	0.174	0.334
Republic of Korea	0.423	0.014
Australia, New South Wales	0.146	0.416
Egypt, Gharbiah	0.123	0.496

Figure 1. Graphical representation of the relationship between number of gene mutations and cancer risk across 33 tissues for the largest cancer registry (USA, NPCR, 42 states). Non-labeled tissues can be identified from values in Table 2.
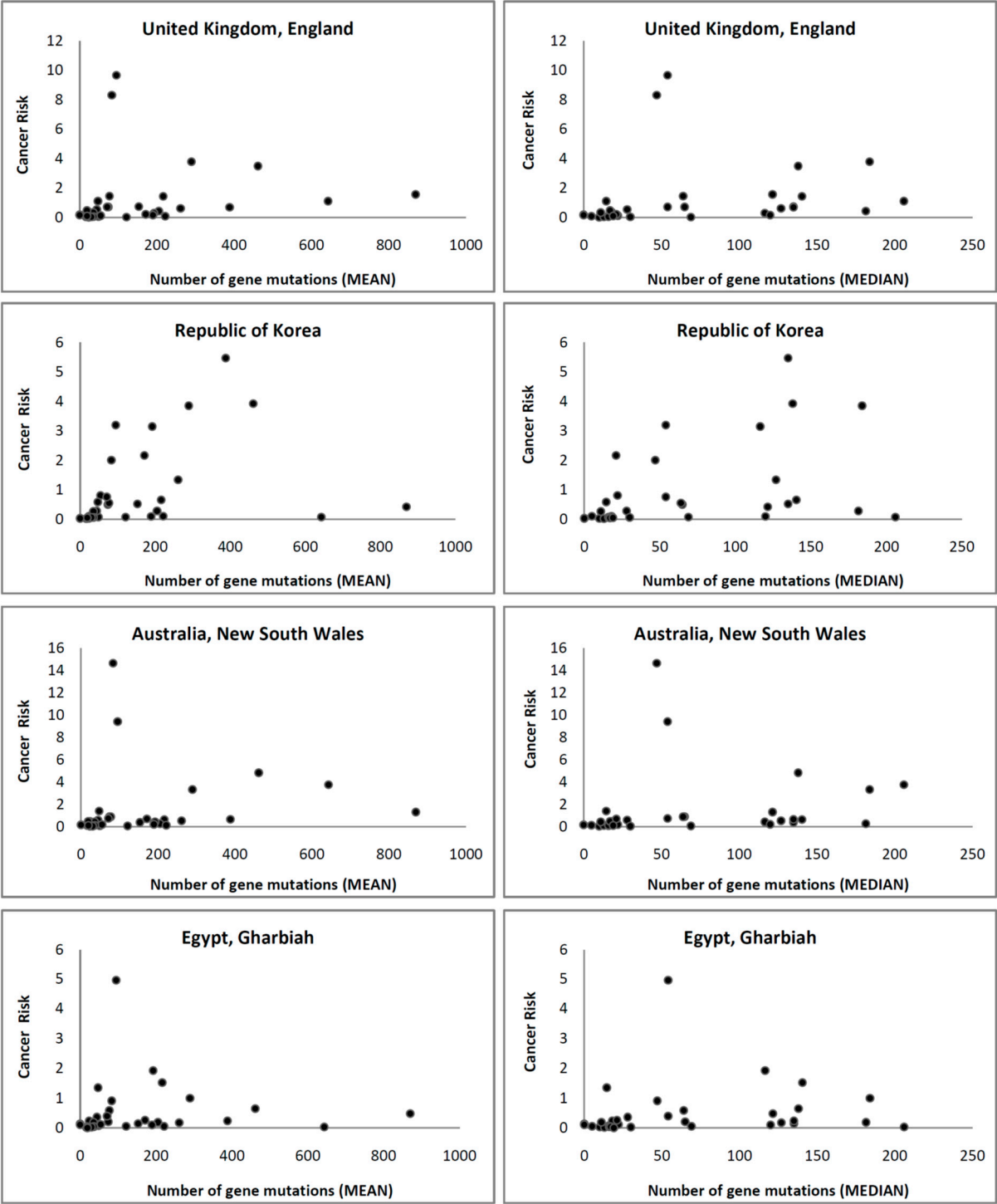
Figure 2. Graphical representation of the association between number of gene mutations and cancer risk across 33 tissues in four cancer registries. Values are from Table 2.

Because the number of samples was low for several cancer types, data were reanalyzed after excluding tissue types for which less than 10 samples were available (i.e., Hodgkin lymphoma, sinonasal and nasal cavity, tonsils, and vulva). All correlations became weaker after excluding these tissue types (data not shown). However, after excluding gender-related cancers (i.e., breast, cervix uteri, corpus uteri, ovary, prostate, testis and vulva), a positive association between the number of gene mutations and cancer risk across the rest of tissues was clearly observed (Figure 3); if the number of gene mutations increased, the risk of cancer also increased in many cases. All correlations became stronger and statistically significant for all the cancer registries but for the smallest one: Egypt, Gharbiah (Table 4).
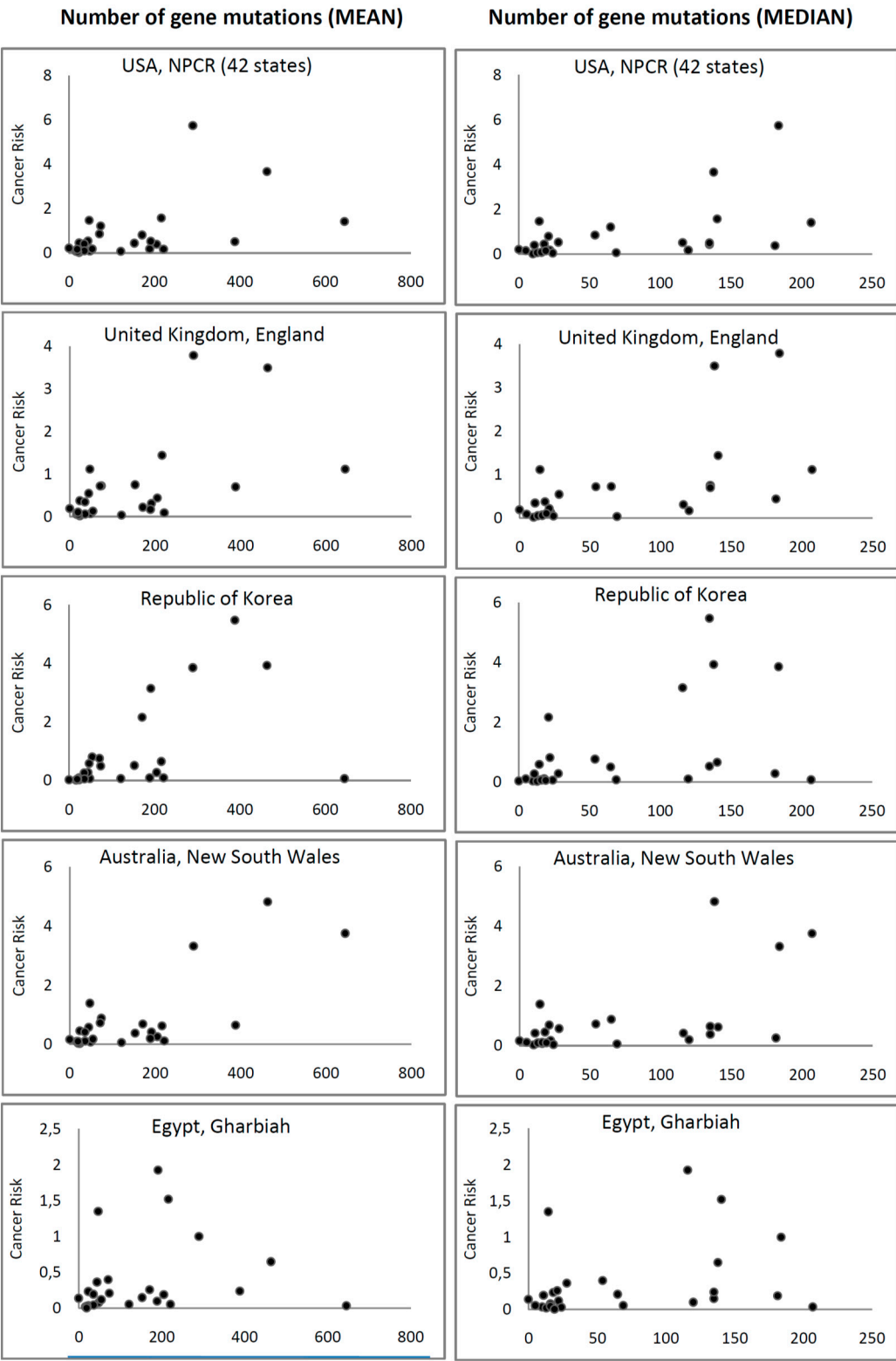
7

Figure 3. Graphical representation of the association between number of gene mutations and cancer risk across non-gender-related tissues in the five cancer registries. Values are from Table 2, after excluding data for breast, cervix uteri, corpus uteri, ovary, prostate, testis and vulva.

Table 4. Pearson correlation and statistical significance (*P*-value, 2 sided) between number of gene mutations and cancer risk across 26 non-gender related cancer types. Data for breast, cervix uteri, corpus uteri, ovary, prostate, testis and vulva were excluded. When the average number of mutations was expressed as mean, the mean correlation was 0.502 and the median correlation was 0.533. When the average number of mutations was expressed as median, the mean correlation was 0.496 and the median correlation was 0.535.

MEAN NUMBER OF GENE MUTATIONS *vs* CANCER RISK		
	Pearson correlation	*P*-value
USA, NPCR (42 states)	0.497	0.010
United Kingdom, England	0.562	0.003
Republic of Korea	0.533	0.005
Australia, New South Wales	0.755	8.39×10^{-6}
Egypt, Gharbiah	0.166	0.825
MEDIAN NUMBER OF GENE MUTATIONS *vs* CANCER RISK		
	Pearson correlation	*P*-value
USA, NPCR (42 states)	0.535	0.005
United Kingdom, England	0.588	0.002
Republic of Korea	0.462	0.018
Australia, New South Wales	0.582	0.002
Egypt, Gharbiah	0.315	0.117

Data shown in Figure 3 and Table 4 suggest that the number of gene mutations occurring during tumor growth may be rather similar among tissues. This baseline number of mutations does not seem to occult the number of mutations occurring before malignant transformation in tissues with different renewal capacities and cancer risks. These data also support the strategy of using cancer genomes to gain mutational information about normal tissues [23].

3.3. Proportion of the variation in cancer risk among tissues explained by the number of gene mutations

Tomasetti and Vogelstein found that the Pearson correlation coefficient between the number of stem cell divisions and cancer risk across tissues was 0.804 [15]. Afterwards, they used the square correlation coefficient to determine the proportion of cancer risk that could be explained by the number of stem cell divisions. Since 0.804 x 0.804 = 0.646, they proposed that 65% (two-thirds) of the variation in cancer risk among tissues could be explained by the number of stem cell divisions. The authors then assumed that the number of stem cell divisions was equivalent to the number of random mutations arising during DNA replication, and proposed that cancer is largely caused (two-thirds) by unavoidable mutations in driver genes; environmental factors and heredity would only contribute one-third to cancer development [15,16]. This is, in essence, the scientific basis of the 'bad luck' hypothesis of cancer [15,16,18].

Using whole genome sequencing information from 22,086 cancer samples and incidence data from five cancer registries, this work does not support the assumption made by Tomasetti and Vogelstein (i.e., the number of stem cell divisions is equivalent to the number of gene mutations arising during DNA replication) [15,16]. The strong correlation (0.80) between cancer risk and number of stem cell divisions found by Tomasetti and Vogelstein is not observed between cancer risk and number of gene mutations (mean = 0.14, Table 3). Following their approach, only 2% (0.14 x 0.14 ~ 2%) of the variation in cancer risk among tissues would be explained by the number of gene mutations. Even after excluding all gender-related cancers, the correlation between cancer risk and number of gene mutations only becomes moderate (mean = 0.50, Table 4). Even in this case, only 25% of the variation in cancer risk among tissues (0.50 x 0.50 = 0.25) would be explained by the number of gene mutations; the rest (75%) would not be explained by unavoidable mutations in driver genes. These data indicate that the variation in cancer risk among tissues is poorly explained by the number of gene mutations.

However, it is important to note that correlation does not necessarily mean causation. In addition, using the square correlation coefficient to determine the proportion of explained variance may be misleading, because the inclusion or exclusion of selected data can notably change the value of the correlation coefficient and, therefore, induce bigger changes in the correlation square. For example, if we excluded breast and prostate cancers from the USA data, the correlation between median number of gene mutations and cancer risk would increase from 0.14 to

0.54. Without excluding these two cancers, only 2% (0.14 x 0.14) of the variation in cancer risk among tissues would be explained by the number of gene mutations. After excluding breast and prostate data, however, the proportion of the variation in cancer risk among tissues explained by the number of gene mutations would be 14.5 times higher (29%; 0.54 x 0.54). These data show that the approach used by Tomassetti and Vogelstein to calculate the proportion of unavoidable cancer risk may be misleading, because results can be very different depending on which tissues are selected for analysis. Therefore, the proportion of cancer risk that can be explained by the number of stem cell divisions or by the number of gene mutations need to be interpreted cautiously.

3.4. Possible mechanisms by which the accumulation of cell divisions in stem cells increases the risk of cancer

The raw data provided by Tomasetti and Vogelstein show a strong positive association between the number of stem cell divisions accumulated by a tissue and the risk of cancer in the tissue [15,16]. The data presented in this article indicate that the association between the number of gene mutations and cancer risk is also positive, but much weaker. Data also show that 1,003 of the 22,806 whole-genome sequenced samples have zero gene mutations. These data indicate that the carcinogenic potential of cell division is not limited to the generation of gene mutations, which leads to the following question: What other mechanisms can explain the highly positive association between the number of stem cell divisions and cancer risk observed by Tomasetti and Vogelstein? Below I discuss several general mechanisms by cell division increases the risk of cancer. Then, I briefly give my opinion on the etiology of the disease.

DNA alterations that involve changes in the nucleotide sequence: DNA mutations

Cell division increases the risk of cancer because it can lead to DNA mutations, i.e., permanent changes in the nucleotide sequence of the DNA. A mutation can affect from one nucleotide to a large DNA segment; one gene can have several mutations and one mutation can affect several genes. Mutations can occur in genes (protein-coding DNA), but they can also affect non-protein-coding DNA sequences. Mutations are more frequent in non-protein-coding DNA than in genes, because genes may represent less than 2% of the total DNA. Despite insufficient recognition, mutations in non-protein-coding DNA can have a major impact on cancer development [29]. Cell division can generate mutations during DNA replication, e.g., as a result of the limited fidelity of DNA polymerases [30]. Cell division can also generate DNA mutations during mitosis; chromosome segregation errors can lead to chromosome rearrangements and mutations that integrate into the DNA [31-33]. In addition, cell division can transform unrepaired DNA damages occurring during quiescence (e.g., oxidative lesion such as 8-hydroxyguanine, spontaneous deamination of 5-methylcytosine, and UV-induced pyrimidine dimers) into mutations. The accumulation of DNA mutations promotes carcinogenesis by altering cellular mechanisms that keep cell division under control.

DNA gains and losses

Cell division can lead to gains and losses of DNA in the daughter cells. For example, the daughter cells can gain or lose large segments of DNA following chromosome segregation mistakes occurring during mitosis [34,35]. Chromosome aberrations can result, e.g., in losses of tumor suppressor genes. Mutations are not required for these carcinogenic events to occur. Like mutations, DNA gains and losses can promote carcinogenesis by disrupting cellular mechanisms that maintain cell division under control.

DNA alterations that do not involve changes in the nucleotide sequence: DNA disorganization

Cell division can disorder epigenetic mechanisms that organize the DNA and control which DNA sequences are used. Structural and functional DNA organization is crucial to restrain the intrinsic tumorigenic potential of our cells. It is important to realize that all the information that a cell needs to give rise to a tumor is in our DNA; no change in the nucleotide sequence in any part of the genome is mandatory for cancer to occur. The accumulation of mutations and other DNA aberrations simply activates and amplifys the natural tumorigenic potential of our cells. We cannot forget that our life starts when one normal cell produces several Kg of cells in nine months; few tumor cells proliferate so fast. Normal embryonic stem cells actually share cellular and molecular phenotypes with tumor cells, including rapid proliferation rates, lack of contact inhibition, high telomerase activity and high expression of

10

oncogenes such as *MYC* [36]. Embryonic stem cells also behave like metastatic cancer cells. During embryogenesis, stem cells invade tissues, migrate through the interior of the embryo, travel long distances, and establish in new places to participate in the formation of organs and tissues. During metastasis, cells invade tissues, migrate through the lymphatic and circulatory systems, travel long distances, and establish in new tissues to form tumors [37,38]. No mutations are needed to activate the migration activity of stem cells and, despite intense efforts, no mutation has been shown to be required for metastasis [7]. When injected into immunodeficient mice, normal human embryonic stem cells form tumors [36]. After embryonic development, the tumorigenic potential of our cells is repressed by epigenetic modifications that organize the DNA into chromatin structures and repress specific DNA sequences in a tissue-specific manner. Histone proteins play a crucial role in DNA compaction and organization. Covalent DNA methylation (usually at the 5-carbon position of the DNA base cytosine in CpG dinucleotides) is an essential mechanism to stably control the gene expression of a cell; about 70-80% of CpGs are known to be methylated in human adult cells [39,40]. Every time a cell divides, however, chromatin must be disassembled and reassembled to allow DNA replication. In addition, the two new DNA strands formed during DNA replication are born naked and must be precisely methylated. In each cell division, the passage of the replication fork during DNA replication not only challenges the integrity of the genetic information, but also the integrity of the epigenetic information [39]. Since cellular states are governed by epigenetic mechanisms, the temporal loss of epigenetic memory occurring during cell division can also lead to a dedifferentiation process caused by a pathological reprogramming of the epigenome. Cell division increases the risk of cancer because it can lead to epigenetic errors that disorganize the DNA and bring back the innate tumorigenic potential of our cells.

As an aside, it is currently unknown whether quiescence or proliferation is the natural state of human cells. Knowing what the natural state of the cell is has important implications for understanding cancer. If quiescence is the natural state of the cell, carcinogenesis would be the process by which the multistep accumulation of DNA alterations in our cells leads to the acquisition of new capabilities, generally known as the hallmarks of cancer [41,42]. If proliferation is the natural state of the cell [9,43,44], carcinogenesis would be the process by which the multistep accumulation of DNA alterations in our cells releases the mechanisms that restrain cell proliferation. Since nearly all our cells have the same DNA and some cells can proliferate and others cannot, it is not easy to uncover the natural state of the cell. In my opinion, the natural state of the cell is the state in which the DNA of the cell is naked, that is, the state in which the DNA of the cell is not altered by covalent modifications such as DNA methylations. Since DNA methylation (e.g., at CpG sites) can alter the expression of genes involved in proliferation and quiescence, these DNA changes can camouflage the native state of the cell. Importantly, a global methylation loss is known to occur in the zygote after fertilization; this DNA methylation removal is maintained during the initial states of embryonic development [40,45]. As discussed above, embryonic stem cells have a natural proliferation activity and tumorigenic potential. This analysis indicates that proliferation is the natural state of human cells, and suggests that some widely-accepted concepts in cancer need reconsideration.

Amplification of DNA alterations

Cell division can increase the risk of cancer because some DNA alterations that are relatively benign during quiescence exert an amplifying carcinogenic effect during cell division. For example, since DNA methylation is a stable DNA modification, a mutation affecting a DNA methylating enzyme (e.g. DNMT1) may have a limited effect on the DNA during quiescence. However, the new strands formed during DNA replication require DNMT1 to copy the methylation pattern of the parental DNA strands. If a cell with defective DNMT1 activity enters cell division, the new DNA strands may not be properly methylated, and major changes in DNA organization and gene expression may occur. Likewise, the mistaken epigenetic silencing of a mismatch repair gene (e.g., *MLH1*) may have a limited effect during quiescence. During cell division, however, deficiencies in the DNA mismatch repair machinery may lead to numerous point mutations, because the cell would have a limited ability to repair the mis-incorporation of nucleotides occurring when the whole genome of the cell is replicated. These two examples also seek to show that DNA mutations and epigenetic alterations are not independent carcinogenic mechanisms; they can promote each other [46].

Loss of physical interactions with other tissue components: Gain of independence

It has been known for many years that the physical interactions of a cell with other cells and non-cellular tissue components constitute a key tumor-suppressor mechanism that keeps cell division and migration under control [47-50]. Cell division leads to a temporal loss of these physical interactions. This temporal freedom can lead to a permanent gain of independence and self-control if the cells involved in the recovery of these physical interactions have accumulated sufficient changes in the mechanisms responsible for recovering these interactions. Note that these interactions not only depend on the cell that gives rise to cancer, but also on other tissue cells. Cell division increases the risk of cancer because the release of contact inhibition occurring during cell division gives the cell an opportunity to break relationships with the tissue components that limit its proliferation and migration potential.

The stem cell division theory of cancer

According to this theory [14,17], the main cause of cancer is the accumulation of cell divisions in stem cells. Cell division, process by which a cell copies and separates its cellular components to finally split into two cells, is the key event in carcinogenesis. Cell division is necessary not only for cancer cell proliferation and tumor growth, but also for the transformation of a normal cell into a cancer cell. Since cell division is a major source of cancer-promoting errors (discussed in the previous paragraphs), the number of cell divisions accumulated by a cell largely determines its risk of malignant transformation. The accumulation of cell divisions drives the accumulation of the DNA alterations required for carcinogenesis; these DNA alterations are necessary to break the cellular mechanisms controlling the natural tumorigenic potential of the cell. Cell division also breaks the physical interactions between the cell and other tissue components that restrain its proliferation and migration potential; in the absence of cell division, a cell with a heavy load of mutations and epigenetic alterations may still be controlled by these physical restrictions. According to the stem cell division theory of cancer, the DNA alterations required for carcinogenesis must develop in stem cells, or end up in a stem cell after a dedifferentiation process. Cancer will not occur if these DNA changes end up in a terminally differentiated cell or in a cell programmed for terminal differentiation (progenitor cell). Cancer formation requires cells with the ability to transmit their DNA alterations indefinitely. Cancer formation also requires cells with the ability to produce an unlimited number of cells. Only stem cells, normal and malignant (cancer stem cells), meet these two linked requirements. Stem cells have the potential to produce an unlimited number of cells because they retain the natural proliferative state of the cell. Under normal conditions, stem cells use this potential to produce the huge number of cells required to form, maintain and repair our body tissues. After malignant transformation, stem cells use this potential to produce the cancer cells required for tumor formation and growth [14,17].

The stem cell division theory proposes that the main determinant of carcinogenesis is the number of cell divisions that the DNA of a stem cell has accumulated in any type of cell from the zygote. This number is not the total number of stem cell divisions accumulated by a tissue after birth divided by the number of stem cells of the tissue [15]; stem cells start accumulating cell divisions from the zygote, and stem cell hierarchies are complex and probably different among tissues [14,51]. In addition, the number of cell divisions required for carcinogenesis is not the number of cell divisions accumulated by a stem cell under normal physiological conditions. Cancer probably arises when the physiological accumulation of cell divisions in a cell leads to the accumulation of DNA alterations that change its behavior and fate; this leads to a pathological increase in the number of cell divisions accumulated by the cell. This pathological accumulation of cell divisions generates additional DNA alterations, which in turn leads to additional cell divisions. This vicious circle is the driving force of carcinogenesis. The number of cell divisions required for carcinogenesis is influenced by numerous factors that are not identical for all tissues; there is not an 'evil number of cell divisions' above which a normal cell gives rise to cancer. However, analyses of the striking differences in cancer incidence by age and among tissues [14] indicate that we may never find a better parameter to integrate all the factors involved in the formation of all cancer types into one principle. The stem cell division theory of cancer provides a new framework for understanding the disease and has important implications for cancer prevention and therapy [14,17].

4. Conclusion

Recent findings indicate that the risk of being diagnosed with cancer in a tissue is strongly determined by the number of stem cell divisions accumulated by the tissue. Because cell division can generate random mutations

during DNA replication, these findings have been used to propose that cancer is largely caused by the accumulation of unavoidable mutations in driver genes. This proposal is based on the assumption that the number of cell divisions is equivalent to the number of gene mutations. To test this proposal, this study has evaluated if the risk of being diagnosed with cancer in a tissue is also strongly determined by the number of gene mutations. Because whole-genome sequencing information for normal tissues is scarce, cancer samples have been used. Cancer samples provide information about the number of mutations occurring in normal tissues because many somatic mutations found in cancer samples occur before tumor initiation [23,24]. In addition, since different tumor types reach similar sizes before surgical resection and since tumor size depends on the number of cell divisions, the number of mutations arising during tumor growth is probably similar among different cancer types. The baseline number of mutations occurring during tumor growth would not occult the number of mutations occurring before tumor initiation in tissues with different stem cell division rates. Using whole genome sequencing information from 22,086 cancer samples and incidence data from several cancer registries, this study shows that the correlation between the number of gene mutations and cancer risk is weaker than the correlation between the number of stem cell divisions and cancer risk. Results also show that 1,003 cancer samples do not have any gene mutation. These data suggest that cancer etiology can be better explained by the accumulation of stem cell divisions than by the accumulation of gene mutations. Possible mechanisms by which the accumulation of cell divisions increases the risk of cancer have been discussed. Cell division increases the risk of cancer because it can lead to changes in the nucleotide sequence of the DNA; these DNA mutations can occur in genes, but also in non-protein-coding DNA sequences. Mutations can arise during DNA replication and during mitosis, and they can be spontaneous or promoted by endogenous and exogenous DNA damage occurring during quiescence. Cell division can also lead to other DNA alterations that do not involve changes in the nucleotide sequence, for example, DNA gains and losses occurring during mitosis and epigenetic aberrations occurring during DNA replication. Cell division also leads to a temporal loss of physical interactions with other tissue components that limit the proliferation and migration potential of the cell. Finally, this article has provided a brief overview of the stem cell division theory of cancer, which proposes that the driving force of carcinogenesis is the accumulation of cell divisions in stem cells.

Conflicts of Interest

The author declares no conflict of interest.

References

1. Forman D, Bray F, Brewster DH, Gombe Mbalawa C, Kohler B, Piñeros M, Steliarova-Foucher E, Swaminathan R, Ferlay J, editors (2014). Cancer Incidence in Five Continents, Vol. X. IARC Scientific Publication No.164. Lyon: International Agency for Research on Cancer. Available online: https://www.iarc.fr/en/publications/pdfs-online/epi/sp164/CI5volX_Full.pdf

2. Nordling CO. A new theory on cancer-inducing mechanism. Br. J Cancer. 1953; 7(1): 68-72.

3. Armitage P, Doll R. The age distribution of cancer and a multi-stage theory of carcinogenesis. Br. J Cancer. 1954; 8(1): 1-12.

4. Vogelstein B, Kinzler KW. Cancer genes and the pathways they control. Nat. Med. 2004; 10(8): 789-799.

5. Vogelstein B, Papadopoulos N, Velculescu VE, Zhou S, Diaz LA, Jr., Kinzler KW. Cancer genome landscapes. Science. 2013; 339(6127): 1546-1558.

6. Stratton MR, Campbell PJ, Futreal PA. The cancer genome. Nature. 2009; 458(7239): 719-724.

7. Vogelstein B, Kinzler KW. The Path to Cancer --Three Strikes and You're Out. N. Engl. J. Med. 2015; 373(20): 1895-1898.

8. Vaux DL. In defense of the somatic mutation theory of cancer. Bioessays. 2011; 33(5): 341-343.

9. Soto AM, Sonnenschein C. The somatic mutation theory of cancer: growing problems with the paradigm? Bioessays 2004; 26(10): 1097-1107.

10. Soto AM, Sonnenschein C. One hundred years of somatic mutation theory of carcinogenesis: is it time to switch? Bioessays. 2014; 36(1): 118-120.

11.   Gatenby RA. Is the Genetic Paradigm of Cancer Complete? Radiology. 2017; 284(1): 1-3.

12.   Prehn RT. Cancers beget mutations versus mutations beget cancers. Cancer Res. 1994; 54(20): 5296-5300.

13.   Baker SG. A cancer theory kerfuffle can lead to new lines of research. J. Natl. Cancer Inst. 2014; 107(2): dju405.

14.   Lopez-Lazaro M. The stem cell division theory of cancer. Preprints, 2017, doi:10.20944/preprints201707.0074.v1; https://www.preprints.org/manuscript/201707.0074/v1

15.   Tomasetti C, Vogelstein B. Cancer etiology. Variation in cancer risk among tissues can be explained by the number of stem cell divisions. Science. 2015; 347(6217): 78-81.

16.   Tomasetti C, Li L, Vogelstein B. Stem cell divisions, somatic mutations, cancer etiology, and cancer prevention. Science. 2017; 355(6331): 1330-1334.

17.   Lopez-Lazaro M. Stem cell division theory of cancer. Cell Cycle 2015; 14(16): 2547-2548.

18.   Couzin-Frankel J. Biomedicine. The bad luck of cancer. Science. 2015; 347(6217): 12.

19.   Tomasetti C, Vogelstein B. Cancer risk: role of environment-response. Science. 2015; 347(6223): 729-731.

20.   Wu S, Powers S, Zhu W, Hannun YA. Substantial contribution of extrinsic risk factors to cancer development. Nature. 2016; 529(7584): 43-47.

21.   Kelly-Irving M, Delpierre C, Vineis P. Beyond bad luck: induced mutations and hallmarks of cancer. Lancet Oncol. 2017; 18(8): 999-1000.

22.   Blokzijl F, de Ligt J, Jager M, Sasselli V, Roerink S, Sasaki N, Huch M, Boymans S, Kuijk E, Prins P, Nijman IJ, Martincorena I, Mokry M et al. Tissue-specific mutation accumulation in human adult stem cells during life. Nature. 2016; 538(7624): 260-264.

23.   Alexandrov LB, Jones PH, Wedge DC, Sale JE, Campbell PJ, Nik-Zainal S, Stratton MR. Clock-like mutational processes in human somatic cells. Nat Genet. 2015; 47(12): 1402-1407.

24.   Tomasetti C, Vogelstein B, Parmigiani G. Half or more of the somatic mutations in cancers of self-renewing tissues originate prior to tumor initiation. Proc. Natl. Acad. Sci. U. S. A. 2013; 110(6): 1999-2004.

25.   Forbes SA, Beare D, Gunasekaran P, Leung K, Bindal N, Boutselakis H, Ding M, Bamford S, Cole C, Ward S, Kok CY, Jia M, De T et al. COSMIC: exploring the world's knowledge of somatic mutations in human cancer. Nucleic Acids Res. 2015; 43(Database issue): D805-D811.

26.   Nouraei SM, Middleton SE, Nouraei SA, Virk JS, George PJ, Hayward M, Sandhu GS. Management and prognosis of primary tracheal cancer: a national analysis. Laryngoscope. 2014; 124(1): 145-150.

27.   Tomasetti C, Marchionni L, Nowak MA, Parmigiani G, Vogelstein B. Only three driver gene mutations are required for the development of lung and colorectal cancers. Proc. Natl. Acad. Sci. U. S. A. 2015; 112(1): 118-123.

28.   Versteeg R. Cancer: Tumours outside the mutation box. Nature. 2014; 506(7489): 438-439.

29.   Diederichs S, Bartsch L, Berkmann JC, Frose K, Heitmann J, Hoppe C, Iggena D, Jazmati D, Karschnia P, Linsenmeier M, Maulhardt T, Mohrmann L, Morstein J et al. The dark matter of the cancer genome: aberrations in regulatory elements, untranslated regions, splice sites, non-coding RNA and synonymous mutations. EMBO Mol. Med. 2016; e201506055.

30.   Arana ME, Kunkel TA. Mutator phenotypes due to DNA replication infidelity. Semin. Cancer Biol. 2010; 20(5): 304-311.

31.   Natarajan AT. Chromosome aberrations: past, present and future. Mutat. Res. 2002; 504(1-2): 3-16.

32.   Crasta K, Ganem NJ, Dagher R, Lantermann AB, Ivanova EV, Pan Y, Nezi L, Protopopov A, Chowdhury D, Pellman D. DNA breaks and chromosome pulverization from errors in mitosis. Nature. 2012; 482(7383): 53-58.

33.   Stephens PJ, Greenman CD, Fu B, Yang F, Bignell GR, Mudie LJ, Pleasance ED, Lau KW, Beare D, Stebbings LA, McLaren S, Lin ML, McBride DJ et al. Massive genomic rearrangement acquired in a single catastrophic event during cancer development. Cell. 2011; 144(1): 27-40.

34. Rajagopalan H, Lengauer C. Aneuploidy and cancer. Nature. 2004; 432(7015): 338-341.

35. Gordon DJ, Resio B, Pellman D. Causes and consequences of aneuploidy in cancer. Nat. Rev. Genet. 2012; 13(3): 189-203.

36. Ben David U, Benvenisty N. The tumorigenicity of human embryonic and induced pluripotent stem cells. Nat Rev. Cancer. 2011; 11(4): 268-277.

37. Nieto MA. Epithelial plasticity: a common theme in embryonic and cancer cells. Science. 2013; 342(6159): 1234850.

38. Lopez-Lazaro M. The migration ability of stem cells can explain the existence of cancer of unknown primary site. Rethinking metastasis. Oncoscience. 2015; 2(5): 467-475.

39. Probst AV, Dunleavy E, Almouzni G. Epigenetic inheritance during the cell cycle. Nat Rev. Mol. Cell Biol. 2009; 10(3): 192-206.

40. Kohli RM, Zhang Y. TET enzymes, TDG and the dynamics of DNA demethylation. Nature. 2013; 502(7472): 472-479.

41. Hanahan D, Weinberg RA. The hallmarks of cancer. Cell 2000; 100(1): 57-70.

42. Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. Cell. 2011; 144(5): 646-674.

43. Soto AM, Sonnenschein C. The tissue organization field theory of cancer: a testable replacement for the somatic mutation theory. Bioessays. 2011; 33(5): 332-340.

44. Sonnenschein C, Soto AM. Carcinogenesis explained within the context of a theory of organisms. Prog. Biophys. Mol. Biol. 2016; 122(1): 70-76.

45. Lee HJ, Hore TA, Reik W. Reprogramming the methylome: erasing memory and creating diversity. Cell Stem Cell. 2014; 14(6): 710-719.

46. You JS, Jones PA. Cancer genetics and epigenetics: two sides of the same coin? Cancer Cell. 2012; 22(1): 9-20.

47. Abercrombie M. Contact inhibition and malignancy. Nature. 1979; 281(5729): 259-262.

48. Ribatti D. A revisited concept: Contact inhibition of growth. From cell biology to malignancy. Exp. Cell Res. 2017; S0014-4827(17): 10.

49. Levin M. Molecular bioelectricity: how endogenous voltage potentials control cell behavior and instruct pattern regulation in vivo. Mol. Biol. Cell. 2014; 25(24): 3835-3850.

50. Sun Y, Chen CS, Fu J. Forcing stem cells to behave: a biophysical perspective of the cellular microenvironment. Annu. Rev. Biophys. 2012; 41:519-42.

51. Visvader JE, Clevers H. Tissue-specific designs of stem cell hierarchies. Nat. Cell Biol. 2016; 18(4): 349-355.