# Parallel routes of human carcinoma development: Implications of the age-specific incidence data

James P. Brody
*Chao Family Comprehensive Cancer Center*
*and*
*Department of Biomedical Engineering*
*University of California–Irvine*
*Irvine, CA 92697-2715*

## Abstract

The multi-stage hypothesis suggests that cancers develop through a single defined series of genetic alterations. This hypothesis was first suggested over 50 years ago based upon age-specific incidence data. However, recent molecular studies of tumors indicate that multiple routes exist to the formation of cancer, not a single route. This parallel route hypothesis has not been tested with age-specific incidence data. To test the parallel route hypothesis, I formulated it in terms of a mathematical equation, confirmed this equation with computer simulations, then tested whether this equation was consistent with age-specific incidence data compiled by the Surveillance Epidemiology and End Results (SEER) cancer registries since 1973. I used the chi-squared goodness of fit test to measure consistency. I found that the age-specific incidence data from most human carcinomas, including those of the colon, lung, prostate, and breast were consistent with the parallel route hypothesis. However, this hypothesis is only consistent if an immune sub-population exists, one that will never develop carcinoma. Furthermore, breast carcinoma has two distinct forms of the disease, and one of these occurs at significantly different rates in different racial groups. I conclude that the parallel route hypothesis is consistent with the age-specific incidence data only if carcinoma occurs in a distinct sub population.

## Introduction

The multi-stage hypothesis [1, 2] states that cancers develop through a series of genetic alterations. This hypothesis is schematically indicated by the following diagram,

$$\text{Gene1} \rightarrow \text{Gene2} \rightarrow \cdots \rightarrow \text{Gene}n \rightarrow \textbf{Cancer}. \qquad \text{(Scheme 1)}$$

Alterations occur successively in $n$ genes (Gene1, Gene2, . . . , Gene$n$) before a tissue specific cancer develops. The process is best studied in human colorectal cancers, where the first four genes in the sequence have been identified as *APC*, *K-ras*, *DCC*, and *p53* [2].

The multi-stage hypothesis was first suggested over 50 years ago based upon an analysis of the age-specific incidence data [3, 4]. This data consists of a histogram of the age at which a population develops cancer. It has long been interpreted to suggest that four to six rate-limiting events are required for the formation of cancer.

Some problems exist with the multi-stage hypothesis. The sequence *APC*, *K-ras*, *DCC*, and *p53* is not the only route to developing colon cancer. This particular route accounts for a subset: only about half of colorectal cancer patients have detectable mutations in the *APC* gene[5], and alternative routes have been identified that do not involve *APC* [6].

Anomalies also exist with the age-specific incidence data that cannot be explained by the multi-stage hypothesis. The incidence for several carcinomas drops at advanced ages[7, 8]. Breast carcinoma incidence data is very different than the other carcinomas[9], and it varies depending on the race and nationality of the population studied [10]. Prostate carcinoma incidence increases much more rapidly with age than other carcinomas, implying that 20 to 30 mutations are required for its development.

Based on several molecular studies[11, 12], it is now generally accepted that multiple parallel routes exist to the formation of tumors, as indicated by the following diagram:

$$\left.\begin{array}{c}
{}_1\text{Gene1} \rightarrow {}_1\text{Gene2} \rightarrow \cdots \rightarrow {}_1\text{Gene}n_1 \rightarrow \\
\text{or} \\
{}_2\text{Gene1} \rightarrow {}_2\text{Gene2} \rightarrow \cdots \rightarrow {}_2\text{Gene}n_2 \rightarrow \\
\text{or} \\
\vdots \\
\text{or} \\
{}_m\text{Gene1} \rightarrow {}_m\text{Gene2} \rightarrow \cdots \rightarrow {}_m\text{Gene}n_m \rightarrow
\end{array}\right\} \rightarrow \textbf{Cancer}. \qquad \text{(Scheme 2)}$$

This **parallel route hypothesis** is a generalization of the multi-stage hypothesis. The number of routes, $m$, and the number of genes involved in each route, $n_1, n_2, n_3, \ldots, n_m$ are not known for any cancer. Just as the multi-stage hypothesis was tested against the age-specific incidence data, so too can the parallel route hypothesis.

Mathematical models of the age-specific incidence of cancer have been an important tool to understand the tumorigenesis process [13, 14]. For instance, Knudson's two-hit hypothesis of retinoblastoma age-incidence data led to the identification of the first tumor suppressor gene, *Rb1* [15]. One of the few pieces of evidence about the human carcinogenesis process is epidemiological data on the incidence as a function of age [16].

I tested the multiple parallel routes hypothesis by comparing a mathematical representation of it to the age-specific incidence data for different forms of cancer. Following this, I examined the implications of the hypothesis and attempted to better understand some of the anomalies of the age-specific incidence data.

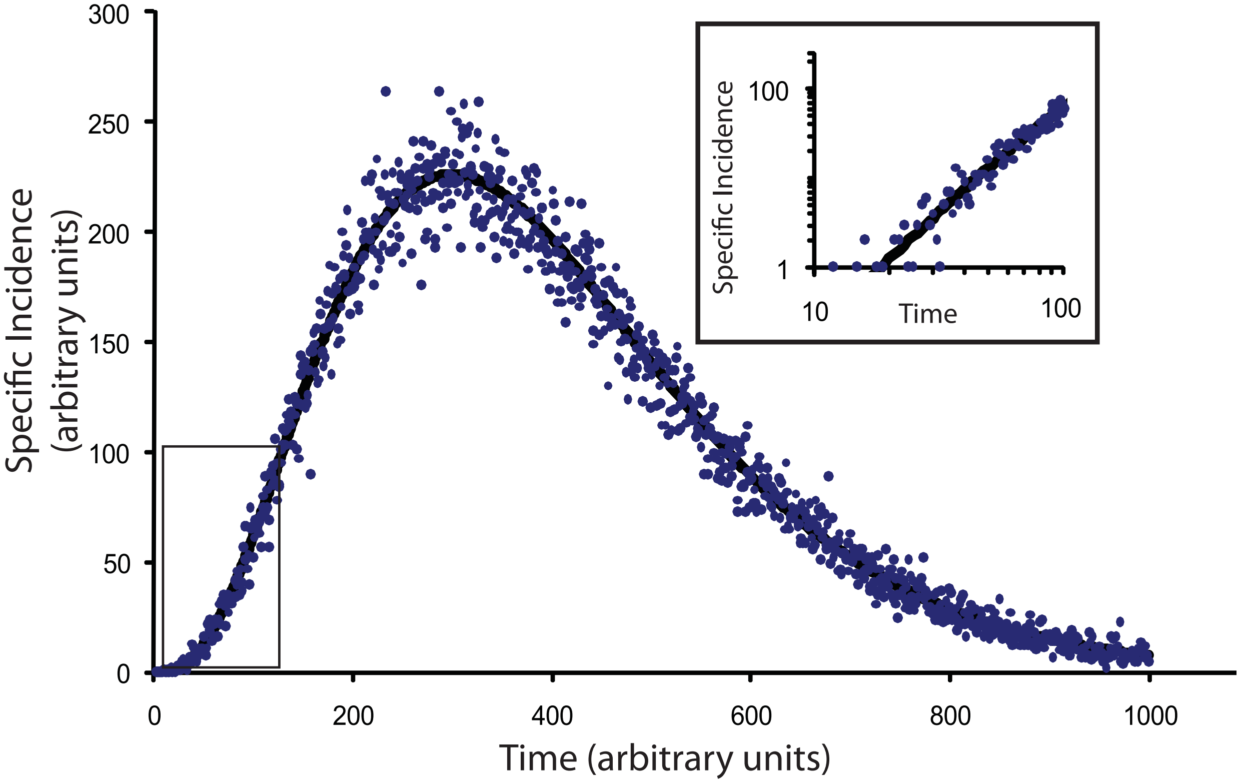

Figure 1: This graph presents the results of computer simulations based upon the assumption that carcinoma occurs through a single route, as depicted in Scheme 1. The points represent the simulation results, while the solid line represents Equation 1. The inset shows the log-log graph in the boxed region as the cancer age-incidence is usually shown

## Results

I first performed computer simulations of Scheme 1 and Scheme 2 to confirm that Equation 1 is the appropriate mathematical representation of the multi-hit hypothesis (Scheme 1) and that Equation 3 is the appropriate representation of the parallel route hypothesis (Scheme 2).

The results of these computer simulations are presented in Figure 1 for the multi-hit hypothesis, and in Figure 2 for the parallel routes hypothesis. These results confirm that the appropriate mathematical representation was used. In the case of Scheme 2, the results also show the relative effect of the number of routes on the width of the age-specific incidence curve. Thus, Equation 3 is an appropriate representation of the parallel routes hypothesis.

To test the validity of the parallel routes hypothesis, I compared Equation 3 with the Surveillance Epidemiology and End Results 17 registries (SEER-17) data collected in the year 2000 for the age-specific incidence of colon carcinoma. The SEER-17 data is the most statistically powerful set of age-specific incidence data available. The year 2000 data monitored 73 million people in the United States and recorded over 22,000 cases of colon carcinoma. The population under surveillance, an important factor, was directly measured by the 2000 US Census data. Thus, this is the best data to use to test this hypothesis.

I found that the parallel route hypothesis is consistent with the colon carcinoma age-specific incidence data, while the multi-stage hypothesis is not. To determine this, I compared the mathematical representation of the parallel route hypothesis, Equation 3, with the age specific incidence

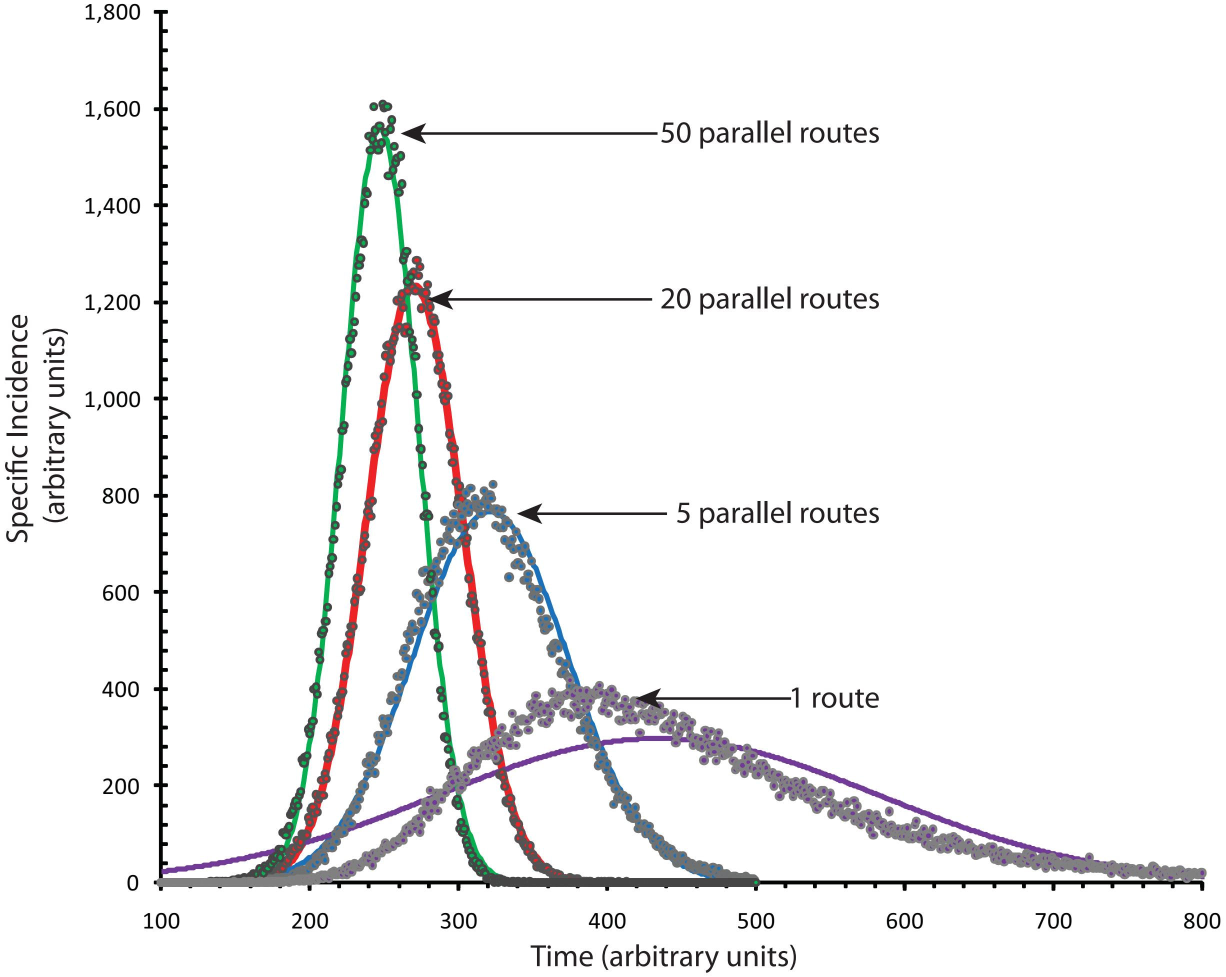

Figure 2: This presents the results of computer simulations based upon the assumption that carcinoma may occur through multiple parallel routes. The four situations represent different assumptions in the computer simulation of 1, 5, 20, or 50 different routes along which carcinoma may occur. The solid lines represent the best fit to each, from Equation 3. The single route simulation, similar to that shown in Figure 1, is clearly not fit by the equation. The 5 parallel routes exhibit slight systematic deviations from the equation, but the 20 and 50 route assumptions are well described by the equation.

data. In the same manner, I also compared the mathematical representation of the multi-stage hypothesis (the Armitage-Doll model Equation 2 [4]) with the age-specific incidence data. See Figure 3. The specific results are that the probability one should accept the parallel route hypothesis is $10^{-3}$ ($\chi^2 = 88$, with 52 degrees of freedom), while the probability that one should accept the multi-stage hypothesis (Equation 2 is $10^{-48}$ ($\chi^2 = 316$, with 53 degrees of freedom), see Figure 3. I also compared the colon carcinoma data to the exact representation of the multi-stage hypothesis, Equation 1, and found that one should not accept this either, ($\chi^2 = 172$, with 52 degrees of freedom, $p = 10^{-17}$).

The mathematical equation (Equation 3) of the parallel routes hypothesis has three parameters, and each has a well defined meaning. The first parameter, $\alpha$ the susceptible population, describes the fraction of the total population susceptible to this carcinoma. This parameter is always greater than the actual fraction who develop the disease. The parameter $\alpha$ can vary from $0.0$ to $1.0$. The second parameter, the mean time, $\tau$, indicates the average time, measured from birth, for the for-

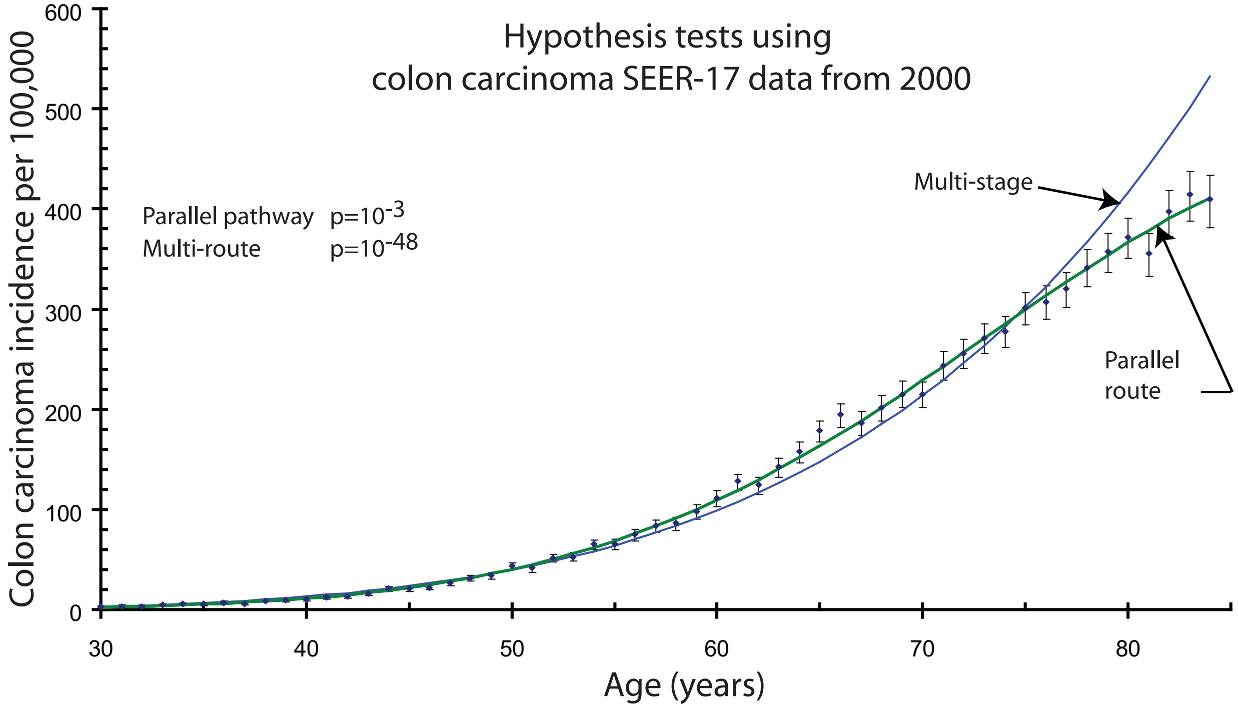

Figure 3: This graph compares two different hypotheses for the development of carcinoma with the age-specific incidence data. The multi-stage hypothesis, as represented by the Armitage-Doll model, Equation 2 [17], and parallel route hypothesis, represented by Equation 3. The Armitage-Doll model is an approximation of the differential form of the Poisson process, Equation 1. Using this equation gives $p = 10^{-17}$, a substantially better, but still clearly unacceptable fit. Thus, the parallel route hypothesis is clearly acceptable, while the multi-stage hypothesis is not. The error bars represent 95% confidence intervals in the measured values.

mation and detection of the cancer. This is a theoretical average, the actual average will always be less, since many die from other causes before getting cancer. One would expect that environmental influences can affect this parameter, and that may be one explanation for the variation of times observed in the population. The third parameter, the standard deviation of the time, $\sigma$, quantifies this variation in the population. This variation can be attributed to either intrinsically random processes, genetic variation, or environmental differences within the population. It is also a function of the number of parallel routes, $m$, as shown in Scheme 2 and Figure 2.

Surprisingly, the age-specific incidence data implies that only about one of every five could ever develop colon carcinoma through this process. This suggests the existence of two distinct sub-populations that are determined either before birth or at a very young age, one is destined to develop colon carcinoma, while the other will never develop it. Competing risk, or death due to other diseases, does not confuse this interpretation, since this age-specific data corrects for reductions in the population.

For the next test of the hypothesis, I used the SEER-9 data [18] to perform 31 similar tests (one for each year from 1973 to 2003) on the four most common types of carcinoma: lung, colon, breast, and prostate carcinoma. The SEER-9 data monitors fewer people than the SEER-17 data, but has been collected since 1973. Using this data allows more independent tests of the hypothesis.

I found that the parallel routes hypothesis was consistent with the age-specific incidence date for lung and colon carcinomas for all 32 years. However, the results for prostate and breast carcinoma were more complicated. See Figure 4. Detailed graphs of the data, along with the parallel routes hypothesis, are shown for 2003 in Figure 5. A table presenting the fraction of the population that is susceptible to the disease is shown in Table 1.

Table 1: This table shows an estimate of the fraction of the population susceptible to each form of carcinoma, as measured by the parameter $\alpha$. The estimate is given as the mean value measured in five recent years, with the standard deviation given in parenthesis.

| Tissue | Susceptible population ($\alpha$) |
|---|---:|
| Colon | $20.0(1.3)\%$ |
| Lung (female) | $9.4(0.4)\%$ |
| Lung (male) | $16.6(0.6)\%$ |
| Breast A | $5.1(0.7)\%$ |
| Breast B | $17.2(1.4)\%$ |
| Prostate | $25(1)\%$ |

Prostate carcinoma age-specific incidence data was consistent with the parallel route hypothesis until 1991, see Figure 4. The 1991 time period corresponds to the widespread adoption of the PSA serum test for prostate cancer screening. The PSA test was first approved in 1986, but it was initially used only to monitor the progress of prostate cancer patients. In 1991 a study [19] showed that the best method of screening for prostate cancers is measurement of serum PSA levels combined with digital rectal exams. This radically changed the diagnosis procedure for prostate carcinoma.

Breast carcinoma has never (1973–2003) been consistent with Equation 3. Instead, I found that breast carcinoma could only be consistent with the parallel routes hypothesis, if one assumes that **three** different sub-populations exist, as expressed in Equation 4. One of these sub-populations is not susceptible to the disease, while the other two sub-populations can develop breast carcinoma. These distinct sub-populations probably develop distinct forms of the disease.

Other studies also suggest that two distinct forms of breast carcinomas exist. Early-onset breast carcinoma is already a recognized subclass of the disease, typically being described as occurring in women before the age of 35 [20–22]. My results indicate significant overlap between the early and late onset forms, and age itself is insufficient to determine whether a woman has one form of the disease or another. (Hence, I refer to these as breast carcinoma **A** and breast carcinoma **B**). However, most cases (about 80%) diagnosed before the age of 35 will be breast carcinoma **A**. Early-onset disease is biologically distinct [23, 24] from the late onset version and also a more potent form of the disease [25, 26]. Furthermore, recent work comparing the age distribution patterns for different histopathologic types of breast cancer using smoothed density plots[27, 28] found results similar to ours [29], including a bimodal distribution of age at diagnosis, with one mode centered about 50 years, and the second about 70 years. Hence, my conclusion is consistent with an emerging view of breast cancer [22].

The origin of these two distinct sub-populations that develop breast carcinoma is unclear. One possibility that these correspond to inherited and sporadic cancers seems unlikely. Three of 211 (95% confidence intervals, 0%–7.2%) breast cancer patients in a population based study were found to have inherited mutations in BRCA1 [30]. These small numbers of BRCA1 mutation breast cancers are unlikely to be apparent in this data.

Finally, I applied this hypothesis to better understand the racial disparity in breast cancer. This striking disparity exists in the prognosis for breast cancer patients in the United States. Although African-Americans are less likely to contract the disease, a significantly larger percentage of these patients die from it, compared to white patients. Furthermore this gap has been increasing over the past few decades [31]. While the obvious cause, unequal treatment, may be responsible for some of this disparity, it is not responsible for all. A detailed study of over 20,000 breast cancer

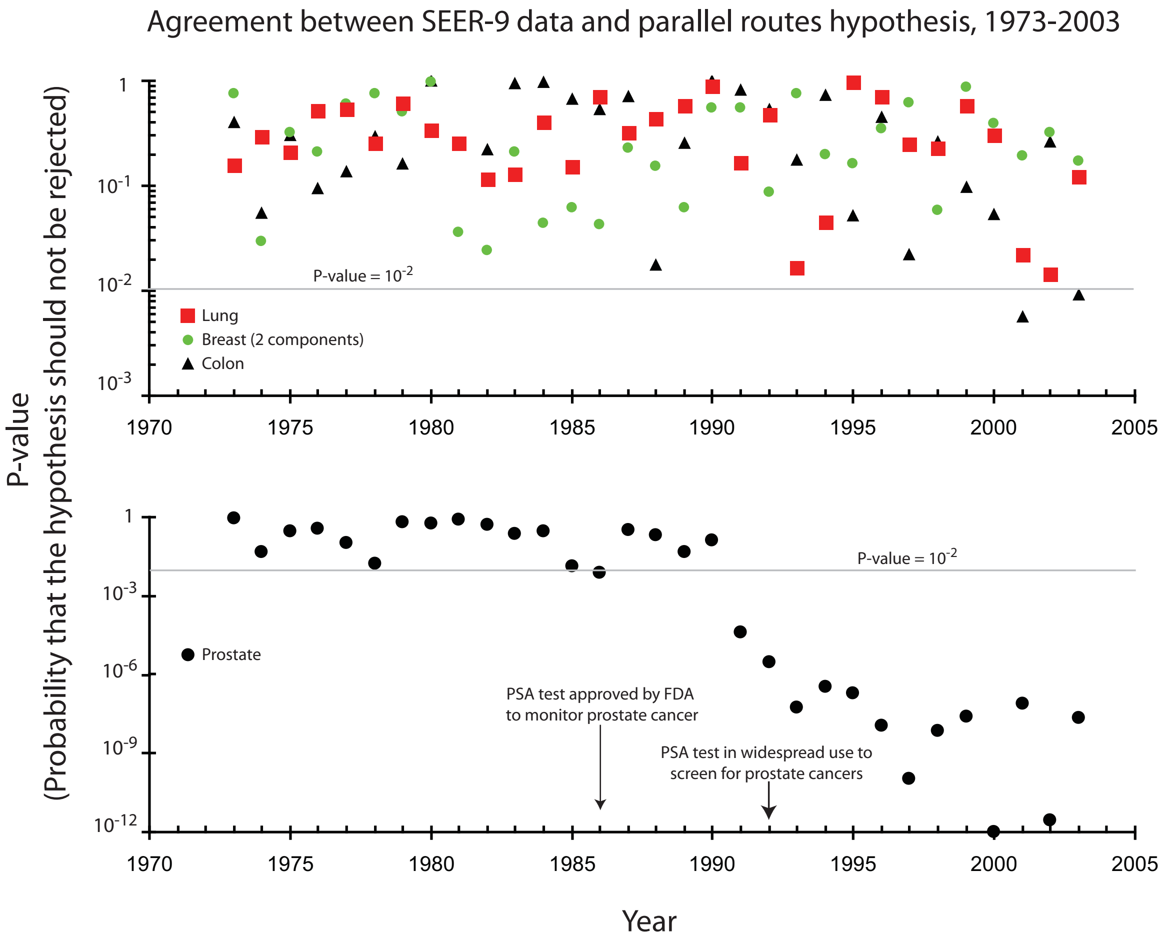

Figure 4: This graph quantitatively shows the agreement between the SEER-9 data and the hypotheses (Equation 3 and Equation 4). The top panel displays the agreement, as measured by a p-value, between lung and colon carcinoma and Equation 3, and between breast carcinoma and Equation 4. In all cases the p-value, representing the probability that one should accept the hypothesis, is greater than 0.001, and in most cases it exceeds 0.01. In contrast, the corresponding graph for prostate carcinoma and Equation 3 shows that the p-value always exceeded 0.01, until 1991 when it plunged below that level. Prostate carcinoma, post 1991, clearly cannot be explained by Equation 3, but it is in agreement with Equation 4. The 1991 change corresponds to the widespread implementation of screening for prostate carcinoma using the PSA test.

patients treated in the equal access Department of Defense Health care system between 1980 and 1999 also revealed a consistent and growing disparity [32].

I reasoned that since early-onset breast cancer is known to be more deadly [25, 26] and that 80% of early-onset cases are breast carcinoma **A** then perhaps the racial disparity exists because African-American women may be less likely to develop the breast carcinoma **B**. To test this, I extracted age-specific incidence data for the year 2000 for both African-American and white women from the SEER-17 dataset. I simultaneously fit the mathematical equation of breast carcinoma age-specific incidence (Equation 4) to each data set. I used the constraint that the breast carcinoma **A** parameters were identical for both African American and white women, while the breast carcinoma **B** parameters could vary.

I found that the age-specific incidence is consistent with the hypothesis that breast carcinoma **A**

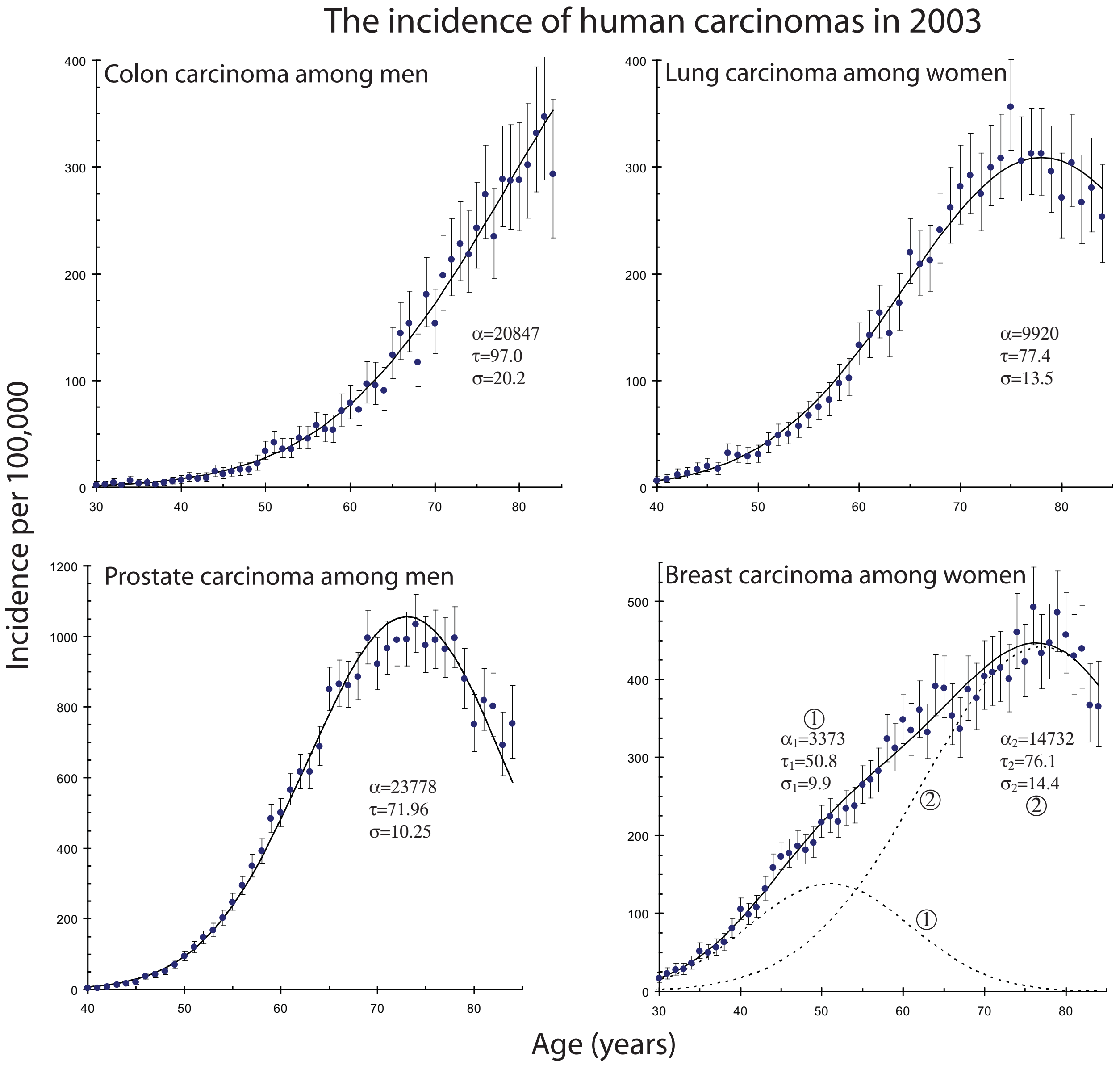

Figure 5: This is a comparison between the hypotheses and the observed age-specific incidence of colon, lung, prostate, and breast carcinomas in 2003 as recorded by the SEER-9 registries. In each case, the measured incidence is represented by a point and the 95% confidence intervals by error bars. A solid line represents the hypothesis, (Equation 3 for colon, lung, and prostate carcinomas and Equation 4 breast carcinomas) and the parameters for the model are indicated on the graph.

has no racial disparity, while breast carcinoma **B** incidence has substantial racial disparity. See Figure 6. Breast carcinoma **B** accounts for three quarters of the cases of breast cancer in the United States. A test that could distinguish breast carcinoma **B** from breast carcinoma **A** might give promising prognosis information to many patients.

The results presented here indicate that the age-specific incidence data for most human carcinomas is consistent with the parallel route hypothesis as expressed in Equation 3 and Equation 4. The parallel routes hypothesis, then, implies (through the age specific-incidence data) that only a subset of the population is susceptible to developing each carcinoma. Other hypotheses may be consistent with this equation, but the implications are not dependent upon the hypothesis.

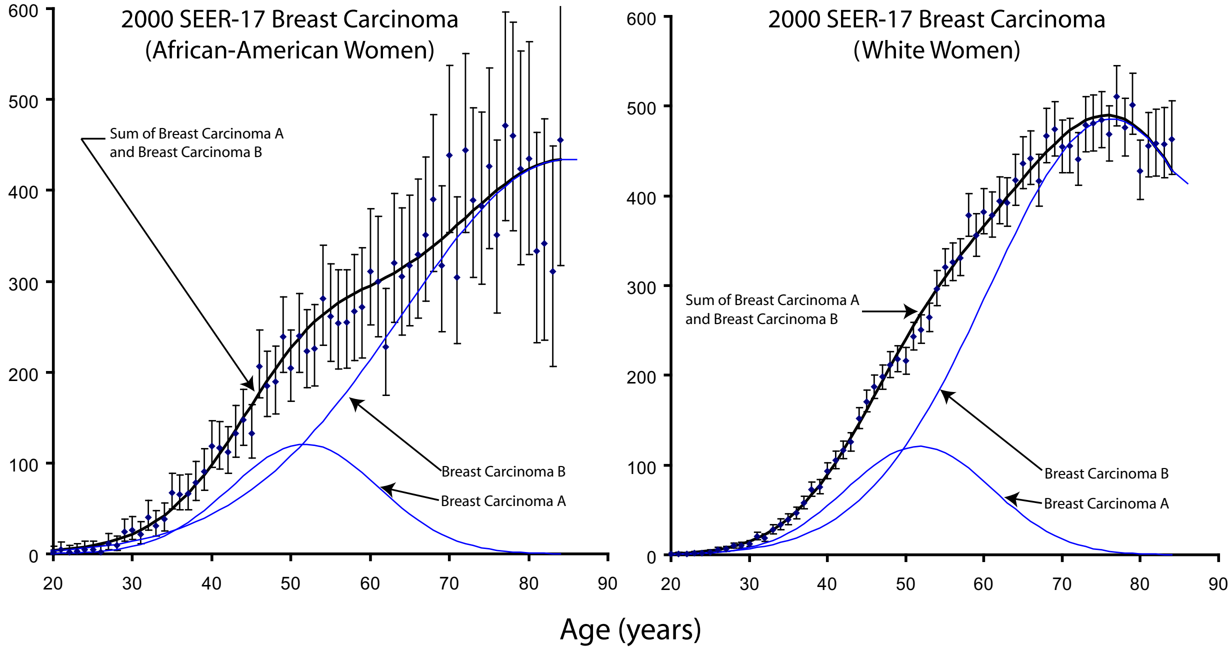

Figure 6: The breast carcinoma disparity between African-American and white women is not due to breast carcinoma **A**, which occurs at exactly the same rate, but solely due to a difference in breast carcinoma **B**. The data are from the SEER-17 database during the year 2000 [18]. In each case, the measured incidence is represented by a point and the 95% confidence intervals by error bars. The dark solid lines represents predicted incidence levels based upon the hypothesis, Equation 4, which is the sum of breast carcinoma **A** and breast carcinoma **B** indicated by the other curves.

This may provide useful guidance for whole genome analysis of carcinomas. Although whole genome analysis has identified genetic causes of some diseases [33], it has not shown similar success when applied to carcinomas [34, 35]. One issue with whole genome analysis, which compares genomes of those who have cancer with those who do not, is the proper identification of samples without cancer. These samples may be from people who simply did not yet develop cancer. A second issue is that samples are usually drawn from peripheral blood and these would only be sensitive to germ-line mutations. About 40% of children who develop retinoblastoma have a germline mutation[36]. If these statistics are true for carcinomas, the current approach to whole genome analysis will not be successful. Thus, this analysis may be useful in the design of whole genome analysis of carcinomas by quantifying the population that will never develop carcinomas.

Finally, similar conclusions have been drawn by others [37–40] using different methods. Other work on age-specific incidence has focused on the two stage with clonal expansion model [41–45] and on developing novel techniques for the analysis of this data [8, 46, 47]. Different mechanisms could explain the existence of a sub-populations susceptible to carcinoma. For instance, members of this sub-population may have inherited susceptibility conferred through low-penetrance alleles [48] or may have acquired a somatic mutation early in life [49].

In conclusion, the parallel routes hypothesis is consistent with the age-specific incidence data for most common forms of human carcinoma. Furthermore, the age-specific incidence data suggest that a only a measurable sub-population is likely to contract carcinoma.

## Methods

**Computer Simulations** I wrote a computer program to simulate Scheme 1 and Scheme 2. The program tracked 100,000 subjects (people). For each subject, it kept a list of genes and routes. Each route contained 50 genes, for Scheme 2. For each time step, a random integer was chosen from 1 to 100 and a gene was said to be mutated if this number was equal to 1. If 50 genes were found to be mutated in sequence, the subject was recorded as having developed cancer at that time step. I chose 50 since recent sequencing studies have shown that 50-100 genes are mutated in some forms of cancer [11, 12]. Similar results can be obtained using different numbers of genes per route, mutation rates, and mutations required, although the time scales will be altered. I also investigated cases in which different routes are not equally probable and where mutation rates vary for different genes. In both cases, I obtained results similar to those in Figure 2. One complication that needs further study is that some genes are almost certainly present in multiple routes.

**Mathematical Models** The multi-step hypothesis implies that the cancer age-specific incidence should follow a Poisson process. The differential form of the Poisson process is,

$$I(t) = \frac{\lambda \mathrm{e}^{-\lambda t}}{k!}\left(k(\lambda t)^{k-1} - (\lambda t)^{k}\right) \tag{1}$$

where $\lambda$ is the expected number of events that occur per unit time and $k$ is the number of rate limiting events required before cancer occurs. However, the age-specific incidence is typically modeled by the Armitage Doll equation[4],

$$I(t) \approx a t^{k-1}, \tag{2}$$

where $a$ is an arbitrary constant and $k$ is again the number of rate limiting events required before cancer occurs. This is an approximation, valid for $t \ll 1/\lambda$, of Equation 1. The condition $t \ll 1/\lambda$ is equivalent to saying that only a small percentage of the population develops cancer.

The parallel route hypothesis is a combination of an unknown number of Poisson processes. Thus, the Central Limit Theorem suggests that the age-specific incidence, $I(t)$, of carcinoma for the parallel route hypothesis is given by

$$I(t) = \alpha \frac{1}{\sqrt{2\pi\sigma^2}} \mathrm{e}^{-\left(\frac{(t-\tau)^2}{2\sigma^2}\right)}, \tag{3}$$

where $\alpha$ represents the fraction of people susceptible to the cancer. This is the well-known normal distribution, also called the Gaussian distribution, or bell-shaped curve.

If two distinct sub-populations can develop carcinoma through a different set of processes, the observed age-specific incidence will be a linear combination of two independent functions,

$$I(t) = \alpha_1 \left(\frac{1}{\sqrt{2\pi\sigma_1^2}} \mathrm{e}^{-\left(\frac{(t-\tau_1)^2}{2\sigma_1^2}\right)}\right) + \alpha_2 \left(\frac{1}{\sqrt{2\pi\sigma_2^2}} \mathrm{e}^{-\left(\frac{(t-\tau_2)^2}{2\sigma_2^2}\right)}\right). \tag{4}$$

**Hypothesis Testing** To test the hypothesis that carcinoma age-specific incidence data follow Equation 2, or Equation 3, or Equation 4, two steps are required. First, the unknown parameters, ($\alpha, \tau$, and $\sigma$, in the case of Equation 3) must be determined; I used maximum likelihood estimation to do so. Second, the probability that the observed data was generated by the postulated equation is determined. I used the chi-squared test for goodness-of-fit to determine this probability.

The maximum likelihood estimator was

$$\chi^2 = \sum_{k=i}^{84} \frac{(O_k - E_k)^2}{E_k}, \tag{5}$$

where $O_k$ is the observed number of carcinoma cases and $E_k$ is the expected number of carcinoma cases in each of the $k = i$ to $84$ age ranges. The SEER dataset provides counts of cancer cases (and population data) in one year intervals from 0 to 84 years. It also includes data on all cases for those greater than 85 years old, this was excluded. The minimum age used, $i$ years, was chosen so that at least 10 cases were present in that year. It was typically in the 20s or 30s, depending on the cancer. The expected number of cases, $E_k$, was obtained by multiplying the age-corrected incidence function $I(t)$, by the population under surveillance in that age range.

Once the parameters were estimated, the chi-squared value was determined, from Equation 5. A p-value, or probability that the fit should be accepted, was calculated based as the one-tailed probability of the chi-squared distribution.

I took several measures to ensure over fitting was not a problem. Over fitting can be caused by fitting an arbitrary mathematical function with multiple free parameters to a dataset. The defining characteristic of over fitting is the inability of the model to fit multiple independent datasets. The first step I took is to fit independent data sets from successive years. This gives me confidence that the model is representative of the underlying data.

**SEER Data** I tested the models, Equation 3 and Equation 4, by fitting them to age-specific incidence data from different carcinomas. The age-specific incidence data were recorded by the SEER 9 registries [18]. The SEER registries have compiled cancer incidence information on a large representative sub-population of US residents since 1973. From this database, I selected patients diagnosed in a particular year with carcinoma in the indicated tissue. This excludes the small number with other types of cancers, sarcomas for instance, which probably arise through a different process. The calculation of confidence intervals are based upon the method of Fay and Feuer [50].

I chose to analyze data for patients who were diagnosed in the same year (2000, for instance), rather than those born in the same year (a birth-cohort). Different factors could distort either data set. Birth-cohort analysis is significantly distorted by changes in medical practice and diagnostic technology. On the other hand, changes in environmental carcinogens may distort period analysis [43–45], like the data presented here. Detection technology for the carcinomas presented here have dramatically improved over the past 50 years. Hence, I focus on patients diagnosed in a single year, while recognizing that their environmental exposure may be different.

**Model Comparison** I used SEER-17 colon carcinoma data from 2000 for both men and women. This gave cancer cases and population under surveillance by individual years. I excluded those ages where fewer than ten cases were observed, ages less than 30. The data then included 22,344 cases in patients from 30 to 84 years of age, 55 independent data points. I tested the hypothesis that each was consistent with the data by first determining the parameters using the maximum likelihood method, then determining the goodness-of-fit by minimizing the $\chi^2$ value.

## Acknowledgments

This work was supported by grant HG-00047 from the National Institutes of Health.

[1] E. Farber. The multistep nature of cancer development. *Cancer Res*, 44(10):4217–4223, Oct 1984.

[2] B. Vogelstein and K. W. Kinzler. The multistep nature of cancer. *Trends Genet*, 9(4):138–141, Apr 1993.

[3] C O Nordling. A new theory on cancer-inducing mechanism. *Br J Cancer*, 7(1):68–72, Mar 1953.

[4] P. Armitage and R. Doll. The age distribution of cancer and a multi-stage theory of carcinogenesis. *Br J Cancer*, 8(1):1–12, Mar 1954.

[5] Giovanni Traverso, Anthony Shuber, Bernard Levin, Constance Johnson, Louise Olsson, David J Schoetz, Stanley R Hamilton, Kevin Boynton, Kenneth W Kinzler, and Bert Vogelstein. Detection of APC mutations in fecal DNA from patients with colorectal tumors. *N Engl J Med*, 346(5):311–320, Jan 2002. doi: 10.1056/NEJMoa012294. URL `http://dx.doi.org/10.1056/NEJMoa012294`.

[6] S. Olschwang, R. Hamelin, P. Laurent-Puig, B. Thuille, Y. De Rycke, Y. J. Li, F. Muzeau, J. Girodet, R. J. Salmon, and G. Thomas. Alternative genetic pathways in colorectal carcinogenesis. *Proc Natl Acad Sci U S A*, 94(22):12122–12127, Oct 1997.

[7] Francesco Pompei and Richard Wilson. A quantitative model of cellular senescence influence on cancer and longevity. *Toxicol Ind Health*, 18(8):365–376, Sep 2002.

[8] Steven A Frank. Age-specific acceleration of cancer. *Curr Biol*, 14(3):242–246, Feb 2004. doi: 10.1016/j.cub.2003.12.026. URL `http://dx.doi.org/10.1016/j.cub.2003.12.026`.

[9] Theodore R Holford, Kathleen A Cronin, Angela B Mariotto, and Eric J Feuer. Changing patterns in breast cancer incidence trends. *J Natl Cancer Inst Monogr*, (36):19–25, 2006. doi: 10.1093/jncimonographs/lgj016. URL `http://dx.doi.org/10.1093/jncimonographs/lgj016`.

[10] Ruth M Pfeiffer, Aya Mitani, Rayna K Matsuno, and William F Anderson. Racial differences in breast cancer trends in the united states (2000-2004). *J Natl Cancer Inst*, 100(10):751–752, May 2008. doi: 10.1093/jnci/djn112. URL `http://dx.doi.org/10.1093/jnci/djn112`.

[11] Si?n Jones, Xiaosong Zhang, D. Williams Parsons, Jimmy Cheng-Ho Lin, Rebecca J Leary, Philipp Angenendt, Parminder Mankoo, Hannah Carter, Hirohiko Kamiyama, Antonio Jimeno, Seung-Mo Hong, Baojin Fu, Ming-Tseh Lin, Eric S Calhoun, Mihoko Kamiyama, Kimberly Walter, Tatiana Nikolskaya, Yuri Nikolsky, James Hartigan, Douglas R Smith, Manuel Hidalgo, Steven D Leach, Alison P Klein, Elizabeth M Jaffee, Michael Goggins, Anirban Maitra, Christine Iacobuzio-Donahue, James R Eshleman, Scott E Kern, Ralph H Hruban, Rachel Karchin, Nickolas Papadopoulos, Giovanni Parmigiani, Bert Vogelstein, Victor E Velculescu, and Kenneth W Kinzler. Core signaling pathways in human pancreatic cancers revealed by global genomic analyses. *Science*, Sep 2008. doi: 10.1126/science.1164368. URL `http://dx.doi.org/10.1126/science.1164368`.

[12] Laura D Wood, D. Williams Parsons, Si?n Jones, Jimmy Lin, Tobias Sj?blom, Rebecca J Leary, Dong Shen, Simina M Boca, Thomas Barber, Janine Ptak, Natalie Silliman, Steve Szabo, Zoltan Dezso, Vadim Ustyanksky, Tatiana Nikolskaya, Yuri Nikolsky, Rachel Karchin, Paul A Wilson, Joshua S Kaminker, Zemin Zhang, Randal Croshaw, Joseph Willis, Dawn Dawson,

Michail Shipitsin, James K V Willson, Saraswati Sukumar, Kornelia Polyak, Ben Ho Park, Charit L Pethiyagoda, P. V Krishna Pant, Dennis G Ballinger, Andrew B Sparks, James Hartigan, Douglas R Smith, Erick Suh, Nickolas Papadopoulos, Phillip Buckhaults, Sanford D Markowitz, Giovanni Parmigiani, Kenneth W Kinzler, Victor E Velculescu, and Bert Vogelstein. The genomic landscapes of human breast and colorectal cancers. *Science*, 318(5853): 1108–1113, Nov 2007. doi: 10.1126/science.1145720. URL http://dx.doi.org/10.1126/science.1145720.

[13] A. G. Knudson. Two genetic hits (more or less) to cancer. *Nat Rev Cancer*, 1(2):157–162, Nov 2001. doi: 10.1038/35101031. URL http://dx.doi.org/10.1038/35101031.

[14] Graham A Colditz and Bernard A Rosner. What can be learnt from models of incidence rates? *Breast Cancer Res*, 8(3):208, 2006. doi: 10.1186/bcr1414. URL http://dx.doi.org/10.1186/bcr1414.

[15] A. G. Knudson. Mutation and cancer: statistical study of retinoblastoma. *Proc Natl Acad Sci U S A*, 68(4):820–823, Apr 1971.

[16] William C Hahn and Robert A Weinberg. Rules for making human tumor cells. *N Engl J Med*, 347(20):1593–1603, Nov 2002. doi: 10.1056/NEJMra021902. URL http://dx.doi.org/10.1056/NEJMra021902.

[17] P. Armitage and R. Doll. A two-stage theory of carcinogenesis in relation to the age distribution of human cancer. *Br J Cancer*, 11(2):161–169, Jun 1957.

[18] Surveillance, Epidemiology, and End Results (SEER) Program. Seerstat database: Incidence–SEER 9 regs public-use (www.seer.cancer.gov). Nov 2002 Sub (1973-2000), National Cancer Institute, DCCPS, Surveillance Research Program, Cancer Statistics Branch, released April 2003, based on the November 2002 submission, 2003.

[19] W. J. Catalona, D. S. Smith, T. L. Ratliff, K. M. Dodds, D. E. Coplen, J. J. Yuan, J. A. Petros, and G. L. Andriole. Measurement of prostate-specific antigen in serum as a screening test for prostate cancer. *N Engl J Med*, 324(17):1156–1161, Apr 1991.

[20] F. de Waard. Premenopausal and postmenopausal breast cancer: one disease or two? *J Natl Cancer Inst*, 63(3):549–552, Sep 1979.

[21] R. A. Walker, E. Lees, M. B. Webb, and S. J. Dearing. Breast carcinomas occurring in young women ($<$ 35 years) are different. *Br J Cancer*, 74(11):1796–1800, Dec 1996.

[22] William F Anderson and Rayna Matsuno. Breast cancer heterogeneity: a mixture of at least two main types? *J Natl Cancer Inst*, 98(14):948–951, Jul 2006. doi: 10.1093/jnci/djj295. URL http://dx.doi.org/10.1093/jnci/djj295.

[23] Suzanne M Johnson, Jacqueline A Shaw, and Rosemary A Walker. Sporadic breast cancer in young women: prevalence of loss of heterozygosity at p53, BRCA1 and BRCA2. *Int J Cancer*, 98(2):205–209, Mar 2002.

[24] S. Weber-Mangal, H.P. Sinn, S. Popp, R. Klaes, R. Emig, M. Bentz, U. Mansmann, G. Bastert, C.R. Bartram, and A. Jauch. Breast cancer in young women ($<$ or $=$ 35 years): Genomic aberrations detected by comparative genomic hybridization. *Int J Cancer*, 107(4):583–592, Nov 2003.

[25] M. Chung, H. R. Chang, K. I. Bland, and H. J. Wanebo. Younger women with breast carcinoma have a poorer prognosis than older women. *Cancer*, 77(1):97–103, Jan 1996.

[26] Q Xiong, V Valero, V Kau, S W Kau, S Taylor, T L Smith, A U Buzdar, G N Hortobagyi, and R L Theriault. Female patients with breast carcinoma age 30 years and younger have a poor prognosis: the M.D. Anderson Cancer Center experience. *Cancer*, 92(10):2523–2528, Nov 2001.

[27] William F Anderson, Nilanjan Chatterjee, William B Ershler, and Otis W Brawley. Estrogen receptor breast cancer phenotypes in the surveillance, epidemiology, and end results database. *Breast Cancer Res Treat*, 76(1):27–36, Nov 2002.

[28] W. F. Anderson, K. C. Chu, N. Chatterjee, O. Brawley, and L. A. Brinton. Tumor variants by hormone receptor expression in white patients with node-negative breast cancer from the surveillance, epidemiology, and end results database. *J Clin Oncol*, 19(1):18–27, Jan 2001.

[29] William F Anderson, Ruth M Pfeiffer, Grata M Dores, and Mark E Sherman. Comparison of age distribution patterns for different histopathologic types of breast carcinoma. *Cancer Epidemiol Biomarkers Prev*, 15(10):1899–1905, Oct 2006. doi: 10.1158/1055-9965.EPI-06-0191. URL `http://dx.doi.org/10.1158/1055-9965.EPI-06-0191`.

[30] B. Newman, H. Mu, L. M. Butler, R. C. Millikan, P. G. Moorman, and M. C. King. Frequency of breast cancer attributable to BRCA1 in a population-based series of American women. *JAMA*, 279(12):915–921, Mar 1998.

[31] J W Eley, H A Hill, V W Chen, D F Austin, M N Wesley, H B Muss, R S Greenberg, R J Coates, P Correa, and C K Redmond. Racial differences in survival from breast cancer. Results of the National Cancer Institute Black/White Cancer Survival Study. *JAMA*, 272(12):947–954, Sep 1994.

[32] Ismail Jatoi, Heiko Becher, and Charles R Leake. Widening disparity in survival between white and African-American patients with breast carcinoma treated in the U. S. Department of Defense Healthcare system. *Cancer*, 98(5):894–899, Sep 2003. doi: 10.1002/cncr.11604. URL `http://dx.doi.org/10.1002/cncr.11604`.

[33] Travis Dunckley, Matthew J Huentelman, David W Craig, John V Pearson, Szabolcs Szelinger, Keta Joshipura, Rebecca F Halperin, Chelsea Stamper, Kendall R Jensen, David Letizia, Sharon E Hesterlee, Alan Pestronk, Todd Levine, Tulio Bertorini, Michael C Graves, Tahseen Mozaffar, Carlayne E Jackson, Peter Bosch, April McVey, Arthur Dick, Richard Barohn, Catherine Lomen-Hoerth, Jeffrey Rosenfeld, Daniel T O'connor, Kuixing Zhang, Richard Crook, Henrik Ryberg, Michael Hutton, Jonathan Katz, Ericka P Simpson, Hiroshi Mitsumoto, Robert Bowser, Robert G Miller, Stanley H Appel, and Dietrich A Stephan. Whole-genome analysis of sporadic amyotrophic lateral sclerosis. *N Engl J Med*, 357(8):775–788, Aug 2007. doi: 10.1056/NEJMoa070174. URL `http://dx.doi.org/10.1056/NEJMoa070174`.

[34] David J Hunter, Peter Kraft, Kevin B Jacobs, David G Cox, Meredith Yeager, Susan E Hankinson, Sholom Wacholder, Zhaoming Wang, Robert Welch, Amy Hutchinson, Junwen Wang, Kai Yu, Nilanjan Chatterjee, Nick Orr, Walter C Willett, Graham A Colditz, Regina G Ziegler, Christine D Berg, Saundra S Buys, Catherine A McCarty, Heather Spencer Feigelson, Eugenia E Calle, Michael J Thun, Richard B Hayes, Margaret Tucker, Daniela S Gerhard, Joseph F Fraumeni, Robert N Hoover, Gilles Thomas, and Stephen J Chanock. A genome-wide association study identifies alleles in fgfr2 associated with risk of sporadic postmenopausal breast cancer. *Nat Genet*, 39(7):870–874, Jul 2007. doi: 10.1038/ng2075. URL `http://dx.doi.org/10.1038/ng2075`.

[35] Meredith Yeager, Nick Orr, Richard B Hayes, Kevin B Jacobs, Peter Kraft, Sholom Wacholder, Mark J Minichiello, Paul Fearnhead, Kai Yu, Nilanjan Chatterjee, Zhaoming Wang, Robert Welch, Brian J Staats, Eugenia E Calle, Heather Spencer Feigelson, Michael J Thun, Carmen Rodriguez, Demetrius Albanes, Jarmo Virtamo, Stephanie Weinstein, Fredrick R Schumacher, Edward Giovannucci, Walter C Willett, Geraldine Cancel-Tassin, Olivier Cussenot, Antoine Valeri, Gerald L Andriole, Edward P Gelmann, Margaret Tucker, Daniela S Gerhard, Joseph F Fraumeni, Robert Hoover, David J Hunter, Stephen J Chanock, and Gilles Thomas. Genome-wide association study of prostate cancer identifies a second risk locus at 8q24. *Nat Genet*, 39(5):645–649, May 2007. doi: 10.1038/ng2022. URL `http://dx.doi.org/10.1038/ng2022`.

[36] Tamara Marees, Annette C Moll, Saskia M Imhof, Michiel R de Boer, Peter J Ringens, and Flora E van Leeuwen. Risk of second malignancies in survivors of retinoblastoma: more than 40 years of follow-up. *J Natl Cancer Inst*, 100(24):1771–1779, Dec 2008. doi: 10.1093/jnci/djn394. URL `http://dx.doi.org/10.1093/jnci/djn394`.

[37] J. Peto and T. M. Mack. High constant incidence in twins and other relatives of women with breast cancer. *Nat Genet*, 26(4):411–414, Dec 2000. doi: 10.1038/82533. URL `http://dx.doi.org/10.1038/82533`.

[38] Paul D P Pharoah, Antonis Antoniou, Martin Bobrow, Ron L Zimmern, Douglas F Easton, and Bruce A J Ponder. Polygenic susceptibility to breast cancer and implications for prevention. *Nat Genet*, 31(1):33–36, May 2002. doi: 10.1038/ng853. URL `http://dx.doi.org/10.1038/ng853`.

[39] P. Herrero-Jimenez, G. Thilly, P. J. Southam, A. Tomita-Mitchell, S. Morgenthaler, E. E. Furth, and W. G. Thilly. Mutation, cell kinetics, and subpopulations at risk for colon cancer in the United States. *Mutat Res*, 400(1-2):553–578, May 1998.

[40] P Herrero-Jimenez, A Tomita-Mitchell, E E Furth, S Morgenthaler, and W G Thilly. Population risk and physiological rate parameters for colon cancer. The union of an explicit model for carcinogenesis with the public health records of the United States. *Mutat Res*, 447(1):73–116, Jan 2000.

[41] S. H. Moolgavkar and G. Luebeck. Two-event model for carcinogenesis: biological, mathematical, and statistical considerations. *Risk Anal*, 10(2):323–341, Jun 1990.

[42] S. H. Moolgavkar and E. G. Luebeck. Multistage carcinogenesis: population-based model for colon cancer. *J Natl Cancer Inst*, 84(8):610–618, Apr 1992.

[43] Suresh H Moolgavkar and E. Georg Luebeck. Multistage carcinogenesis and the incidence of human cancer. *Genes Chromosomes Cancer*, 38(4):302–306, Dec 2003. doi: 10.1002/gcc.10264. URL `http://dx.doi.org/10.1002/gcc.10264`.

[44] E. Georg Luebeck and Suresh H Moolgavkar. Multistage carcinogenesis and the incidence of colorectal cancer. *Proc Natl Acad Sci U S A*, 99(23):15095–15100, Nov 2002. doi: 10.1073/pnas.222118199. URL `http://dx.doi.org/10.1073/pnas.222118199`.

[45] Rafael Meza, Jihyoun Jeon, Suresh H Moolgavkar, and E. Georg Luebeck. Age-specific incidence of cancer: Phases, transitions, and biological implications. *Proc Natl Acad Sci U S A*, 105(42):16284–16289, Oct 2008. doi: 10.1073/pnas.0801151105. URL `http://dx.doi.org/10.1073/pnas.0801151105`.

[46] Steven A Frank, Peng-Chieh Chen, and Steven M Lipkin. Kinetics of cancer: a method to test hypotheses of genetic causation. *BMC Cancer*, 5:163, 2005. doi: 10.1186/1471-2407-5-163. URL `http://dx.doi.org/10.1186/1471-2407-5-163`.

[47] Steven A. Frank. *Dynamics of Cancer: Incidence, Inheritance, and Evolution*. Princeton University Press, 2007.

[48] Richard S Houlston and Julian Peto. The search for low-penetrance cancer susceptibility alleles. *Oncogene*, 23(38):6471–6476, Aug 2004. doi: 10.1038/sj.onc.1207951. URL `http://dx.doi.org/10.1038/sj.onc.1207951`.

[49] Steven A Frank and Martin A Nowak. Cell biology: Developmental predisposition to cancer. *Nature*, 422(6931):494, Apr 2003. doi: 10.1038/422494a. URL `http://dx.doi.org/10.1038/422494a`.

[50] M. P. Fay and E. J. Feuer. Confidence intervals for directly standardized rates: a method based on the gamma distribution. *Stat Med*, 16(7):791–801, Apr 1997.