Title: Mathematical modeling links pregnancy-associated changes and breast cancer risk

Authors: Daniel Temko[1-4+], Yu-Kang Cheng[1+], Kornelia Polyak[5*] and Franziska Michor[1*]

Authors' Affiliations: [1]Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute, and Department of Biostatistics, Harvard T. H. Chan of Public Health, Boston, MA, USA. [2]Centre for Mathematics and Physics in the Life Sciences and Experimental Biology (CoMPLEX), University College London, London, UK. [3]Department of Computer Science, University College London, London, UK. [4]Barts Cancer Institute, Queen Mary University of London, London, UK. [5]Department of Medical Oncology, Dana-Farber Cancer Institute, and Department of Medicine, Harvard Medical School, Boston, MA, USA. [+]Equal contribution.

Running title: Mammary progenitors and breast cancer risk

Keywords: breast cancer risk, pregnancy, progenitors

List of abbreviations: NHS- Nurses Health Study, ER+ - estrogen receptor positive, PR+ - progesterone receptor positive

Notes: *Financial support:* This work was supported by NCI U54CA193461 (to F. Michor and K. Polyak), and by EPSRC doctoral training grant EP/F500351/1 (to D. Temko)

Corresponding authors: Franziska Michor, Dana-Farber Cancer Institute, 450 Brookline Avenue, CLSB-11029, Boston, MA 02215, USA. Phone: 617-632-5045; Fax: 617-632-2444; E-mail: michor@jimmy.harvard.edu; Kornelia Polyak, Dana-Farber Cancer Institute, 450 Brookline Avenue, D740C, Boston, MA 02215, USA. Phone: 617-632-2106; Fax: 617-582-8490; E-mail: kornelia_polyak@dfci.harvard.edu;

ABSTRACT

Recent debate has concentrated on the contribution of bad luck to cancer development. The tight correlation between the number of tissue-specific stem cell divisions and cancer risk of the same tissue suggests that bad luck has an important role to play in tumor development, but the full extent of this contribution remains an open question. Improved understanding of the interplay between extrinsic and intrinsic factors at the molecular level is one promising route to identifying the limits on extrinsic control of tumor initiation, which is highly relevant to cancer prevention. Here we use a simple mathematical model to show that recent data on the variation in numbers of breast epithelial cells with progenitor features due to pregnancy are sufficient to explain the known protective effect of full-term pregnancy in early adulthood for estrogen receptor positive (ER+) breast cancer later in life. Our work provides a mechanism for this previously ill-understood effect and illuminates the complex influence of extrinsic factors at the molecular level in breast cancer. These findings represent an important contribution to the ongoing research into the role of bad luck in human tumorigenesis.

Major Findings

A mathematical model demonstrates that the pregnancy-associated reduction in $Ki67^+$ and $p27^+$ cells numbers in the human breast can explain the protective effect of pregnancy against ER^+ breast cancer.

Quick Guide to Equations and Assumptions

Cellular dynamics of the stem cell and proliferative progenitor cell populations

There are N stem cells per terminal end duct. The stem cells follow a stochastic process known as the Moran model. One cell division occurs during each time step of length t_{cycle} / N. In each time step a single stem cell is randomly chosen to divide proportional to the fitness of the cell, with the two daughter cells replacing the divided cell and another randomly chosen cell.

With probability *p*, stem cell divisions are asymmetric, giving rise to one stem cell that replaces the divided cell and one progenitor cell that forms the founder in a new cascade of progenitors. All cells in a progenitor cascade divide during every time step. In non-pregnant women, wild-type progenitor cells can divide a total of *z* times before becoming terminally differentiated (see below for the effects of mutations and effects of pregnancy). Cells that are terminally differentiated exit the simulation.

Cancer initiation

During each cell division, one of the two daughter cells in a division attains a new (epi)genetic mutation with probability μ. In stem cells, mutations increase the relative fitness of the cell by a factor of f_{mut}. In progenitor cells mutations increase the number of levels in the differentiation hierarchy by z_{mut} levels. Thus a stem cell with *n* mutations has relative fitness in the Moran model given by equation (1) and a progenitor cell with n mutations is able to divide a total number of times given by equation (2) before terminal differentiation:

(1) $(f_{mut})^n$

(2) $n * z_{mut}$

Additionally, progenitor cells must acquire the ability to self-renew. We assumed that the probability of a progenitor cell at differentiation level $0 \leq i \leq z + n * z_{mut}$ attaining self-renewal is given by equation (3):

(3) $\gamma = \gamma_{base} - (i * \gamma_{base}) / (2 * z)$

We assumed that cancer initiation occurs when a cell has accumulated a total of n_{mut} mutations and either retained (through being a stem cell) or attained (through a self-renewal event) the ability to self-renew.

Effect of pregnancy

Our model simulates an entire life-course over t_{total} years. The model takes into account possible changes to cellular dynamics during pregnancy, after pregnancy, and after menopause. During pregnancy we assumed that the stem cell cycle length decreases to $t_{cycle,preg}$, whereas the number of levels in the differentiation hierarchy of progenitor cells increases by z_{preg} levels. After menopause, the stem cell cyle length increases to $t_{cycle,menopause}$.

In parous scenarios, after the first birth the probability of asymmetric stem cell division changes by a multiplicative factor $p_{post,init}$ ($0 < p_{post,init} < 1$). After the second birth and subsequent births, the probability of asymmetric stem cell division changes by a factor of $p_{post,subs}$ ($p_{post,init} < p_{post,subs} < 1$).

INTRODUCTION

A recent study (1) by Tomasetti and Vogelstein analyzed the relationship between the number of stem cell divisions and cancer risk across tissues to investigate the role of “bad luck” in carcinogenesis. The authors demonstrated that the logarithm of lifetime cancer incidence in a tissue is closely correlated with the logarithm of the cumulative number of stem cell divisions in the same tissue ($R^2 = 0.64$). As a result, the authors claimed that the majority of the variance in cancer risk among tissues is due to bad luck (Fig. 1A).

In the reporting of the study and ensuing debate some commentators drew broader conclusions from the correlation found by Tomasetti and Vogelstein. While the initial study claimed that two thirds of the variation in cancer risk between tissues is due to bad luck, an accompanying commentary suggested that two thirds of all cancers, rather than two thirds of the variation, are due to random mutations in healthy cells (2). Subsequent analyses have shown that the initial correlation is not sufficient to imply a lower bound on the proportion of all cancers that are due to bad luck at 64%. To draw this conclusion from the study would require strong assumptions about the possible effects of controllable factors in the data set considered (3).

Importantly, the regression analysis used by Tomassetti and Vogelstein cannot quantify the possible effects of extrinsic factors that do not already vary within the data set used, which notably did not include breast cancer (4). Therefore, the regression cannot be used to draw conclusions about unavoidable bad luck, taking into account the variation of all possible extrinsic factors. To illustrate this point, consider the (perhaps unlikely) possibility that it is possible to safely alter the fitness advantage of mutations that can lead to cancer. The correlation analysis presented cannot tell us about the impact such variation could have on cancer risk.

The insufficiency of the current evidence to draw conclusions about the contribution of unavoidable bad luck to cancer demonstrates the important potential role of mechanistic

models in determining the contribution of controllable factors to different cancer types, and whether these factors can be harnessed for cancer prevention. The changes that lead to cancer are thought to develop in a complex molecular setting, which defies simple characterization. In this setting variation of any number of parameters may affect lifetime risk of cancer; these include but are not limited to the number of cells susceptible to transformation, the mutation rate of cells, and the fitness advantage conferred by those mutations when they occur (Fig. 1B).

Full-term pregnancy in young adulthood is a well-documented natural protective factor for breast cancer (5,6). Estimates suggest that risk increases by 5% for every five-year increase in the age at first birth for women with one birth (6). The specific effects of parity vary by hormone-receptor status of the resulting tumors (7). Analysis of the Nurses Health Study (NHS) cohort showed that the risk for ER+ breast cancer decreases with the number of pre-menopausal years accumulated since first birth (7). Hence, early first birth confers the greatest protective effect; a woman with four births at age 20, 23, 26 and 29 years old has an estimated 29% reduced risk of ER+/PR+ breast cancer between the ages of 30 and 70, compared to a nulliparous woman during the same time period. The same study found that first birth causes a one-off increase in risk for PR- cancer compared to nulliparous women, with an effect size that increases with age at first birth. As a result, women with a first birth over the age of 35 can be at an increased risk of breast cancer.

In the absence of high-resolution single-cell data which are nearly impossible to obtain in large cohorts of humans, mathematical models have demonstrated the plausibility of general molecular explanations for the protective effects of pregnancy. An important study by Moolgavkar et al. explored a framework where breast cancer is caused by two cellular transitions occurring in normal cells (8). In this model, pregnancy increases the rate of differentiation of normal and partially transformed cells, decreasing the pool of cells susceptible to the cellular transitions leading to cancer. The study leads to a good fit

to the data of MacMahon et al. (5). The model of Pike et al. (9) uses a concept of breast tissue age: breast cancer incidence is modeled as a linear function of the logarithm of breast tissue age, and risk factors for breast cancer alter the rate of breast tissue aging. First full-term pregnancy causes a one-off increase in breast tissue age, but decreases its subsequent rate of increase. This study also demonstrated a good fit to the Moolgavkar et al. (8) data. Rosner and Colditz then adapted and extended the model developed by Pike et al. (9), including changes to further improve the fit and accommodate multiple births, and applied the adapted model to data from the NHS cohort (7,10,11). The fit of these models to epidemiological data provide support for the theory that pregnancy alters the number of cells that are at risk for accumulating changes leading to breast cancer. However, they do not identify the molecular mechanisms responsible, nor do they accommodate the effects of a cellular hierarchy of stem and progenitor cells.

Recently, single cell technology has made it possible to collect quantitative data on changes in individual mammary sub-populations, presenting the possibility to quantitatively assess the molecular-level changes, as well as the epidemiological incidence curves, associated with pregnancy. Studies in mice and humans provide evidence that p27+ mammary epithelial cells with progenitor features decrease in number with pregnancy, and are present in high numbers in *BRCA1* and *BRCA2* germline mutation carriers (12,13). Evidence was presented that a subset of p27+ cells with progenitor features are hormone-responsive quiescent luminal progenitors with proliferative potential, and that their variation could relate to breast cancer risk (12). Here, we use a simple mathematical model to test whether, given a role for p27+ progenitor cells as proliferative progenitors which can accumulate changes leading to breast cancer, the observed reduction in the populations of p27+ progenitor cells with pregnancy is sufficient to explain the protective effect of pregnancy.

MATERIALS AND METHODS

We aimed to test the hypothesis that a decreasing cell number and proliferative capacity of luminal progenitor cells after pregnancy can result in a protective effect against breast cancer and that the effect decreases with increasing age of pregnancy. To this end, we designed a mathematical model of the dynamics of proliferating cells in the breast tissue that can accumulate the changes leading to cancer initiation. We considered two types of cells: a self-renewing population of stem cells, and a population of proliferating luminal progenitor cells that result from differentiation of these stem cells and respond to hormonal stimuli. We first tested whether we could identify a biologically plausible parameter setting in our model under which the variation in progenitor cell numbers results in a risk decrease that fits the quantitative risk decreases observed with pregnancy. We then tested the robustness of the fit of our model in the surrounding parameter space.

We first studied the dynamics of stem cells in the breast ductal system. Given the population structure inherent to breast ducts, we considered the stem cells in each duct to act independently. As such, we investigated the dynamics of a single duct within the breast since the total probability of cancer initiation is given by the probability per niche times the number of niches; thus, the relative likelihood of cancer initiation is not altered by considering only one niche. The overall number of stem cells in the breast is estimated to be on the order of 5 to 10 cells per duct (14,15), and we denoted this number by N, although there is some uncertainty in these estimates. We defined a fundamental time unit of our system to be dictated by the division time of stem cells, t_{cycle}, which varies during pregnancy. In *in vivo* experiments, the mean cell cycle length of benign breast cancer cells was approximately 162 hours per cell (16). We assumed that even pre-cancerous cells divide faster than stem cells; thus, using t_{cycle} = 162 hours as the average pre-menopausal stem cell cycle length when not pregnant may be an overestimation of the number of stem cell divisions that occur in the normal breast, and we verified that our results were

unaffected at higher stem cell cell cycle lengths. Further, previous data by our lab (12) and several others (17-22) suggests that the percentage of cells in normal breast that stain positive for Ki67 are approximately 3% and 12% in the follicular and luteal phases of the menstrual cycle, respectively. Assuming that the duration of these two menstrual cycle phases is roughly the same, at two weeks per cycle, leads to an average Ki67 value of 7.5%. Considering that Ki67 is detectable for 24 hours during the active phases of the cell cycle (23,24), this translates to an estimate of 320 hours (24 / 0.075) for the average cell cycle length, which is also within the range tested (162 hours to 324 hours). Other studies have shown a broadly consistent range of Ki67 / KiS5 values (20) or lower values consistent with still longer cell cycle times (18,19).

Experimental data suggests that proliferation decreases 4-5 fold after menopause, irrespective of parity (12,25). To take this effect into account, we assumed that the cell cycle length increases by a factor of $\alpha_{menopause} = 4$ after menopause. In our model, a single stem cell in each duct is randomly chosen to divide during each time step, proportional to the fitness of the cell, following a stochastic process known as the Moran model (26). According to this model, the divided cell is replaced by one of the daughter cells of the division, while the other daughter replaces another stem cell that was randomly selected from the population to die. The use of this model ensures preservation of homeostasis in the normal breast epithelial cell population. Since the specific dynamics of stem cells in the breast are not known, we chose the Moran model as it has been used to model stem cell populations in other tissues (27-29). For each cell division, we allowed for a single mutation to arise in one of the two daughter cells of the division with a certain probability.

In the mature breast, stem cells divide primarily to maintain cellular integrity. However, differentiating events do occur, although rarely (30-32). In our model, with probability *p*, we allowed the cell division in the current time step to be asymmetric, producing one daughter stem cell to maintain the stem cell population and one progenitor

daughter to arise. Since the exact rate of differentiation is unknown, we tested $p = 10^{-1}$ to 10^{-3}. With the remaining $1 - p$ probability, the stem cell division is symmetric and follows the usual Moran division dynamics. In each time step thereafter, all cells resulting from the progenitor daughter divide and differentiate further until a total of z cell divisions are accumulated. The number of luminal epithelial progenitors in humans is unknown. As a result, we set $z = 10$ to fit data from mouse mammary fat pad transplantation experiments (33), and tested a wide range of alternate values for this parameter. After z_{pre} divisions, we considered the cells differentiated and at this point, they are no longer considered in our mathematical model. Thus, in the wild-type system, there are N stem cells per duct and $2^{z+1} - 1$ progenitor cells per differentiation cascade. Since the dynamics of progenitor cells in the human breast are not known, we have adopted the assumption that progenitor cells undergo a limited number of divisions, similar to what has been observed for transit-amplifying cells in the colon and other tissues. Fig. 2A describes the temporal dynamics of the system.

During each cell division, genetic alterations contributing to cancer initiation may arise with a small probability. We considered a number n_{mut} of mutations that, when combined, result in a single cell leading to cancer initiation. These mutations could each be any of the many mutations commonly found in breast cancer with initiation potential. As a simplifying assumption we considered a mutation rate on the order of 10^{-5} mutations per oncogenic mutation per cell division to limit the required number of simulations for detection to a reasonable number.

The baseline mutation rate is roughly 5×10^{-9} per base pair per cell division (34,35). It is estimated that there are roughly 34,000 possible driver base pairs in the genome (36), thus it may be reasonable to assume that there are on the order of 10,000 possible ways to achieve each oncogenic mutation, which would lead to the above rates on the order of 10^{-5} mutations per oncogenic mutation. However, it is important to note that not all driver

loci are relevant in breast cancer, and in particular the exact combinations of driver loci that could cause breast cancer are unknown, thus the 10^{-5} figure can only be a broad approximation. For this reason, we also tested our model at other mutation rates, and found that our main conclusions were also consistent at lower mutation rates.

We studied the following mutational effects for each mutation: under the default assumptions in stem cells, mutant cells had a relative fitness of $f_{mut} = 1.1$, i.e. a fitness increase of 10%, resulting in an increased probability of dividing, while mutant progenitor cells divided an additional $z_{mut} = 1$ times (Fig. 2B). Since the number of stem cells per duct is small, the fitness of mutant alleles has little effect on cancer initiation probabilities, as the fixation time of mutations is much smaller than the mutation accumulation time (27); we also tested our results at other values of f_{mut} and z_{mut}. Additionally, progenitor cells must accumulate some propensity towards self-renewal: we defined a parameter $\gamma = \gamma_{base} - (i * \gamma_{base}) / (2 * z)$ as the probability of a progenitor cell at differentiation level $0 \leq i \leq z + n * z_{mut}$ to acquire self-renewal. We chose this functional form to capture a decrease in the probability of attaining self-renewal as progenitor cells differentiate, and explored different values of γ_{base} within this framework. We defined cancer initiation as a single cell that accumulated all required mutations and either retained or acquired the ability to self-renew, either through being a stem cell or through acquiring a genetic or epigenetic self-renewal event.

As we were interested in the effects of the timing of pregnancy, we considered the phenotypic alterations that occur in the breast during pregnancy and as a result of pregnancy. For the purposes of this simulation, we considered the 280 day period of time for the pregnancy itself as the time period during which parameters are altered by pregnancy. Evidence suggests that pregnancy results in the differentiation of mammary epithelial cells (37,38) as well as their increased proliferation (19,39). To model these effects, we allowed further differentiation of progenitor cells during pregnancy by an

additional z_{preg} differentiation levels, and a decrease in the cell cycle length of stem cells (Figure 2C). There is a 4.5 to 8.5-fold increase in the number of Ki67+ cells during pregnancy (19,39). Thus, we allowed a 4-fold to 8-fold increase in progenitor cells during pregnancy, corresponding to z_{preg} = 2 to 3. The remaining ~1.1 fold increase in proliferation was modeled as a decrease in stem cell cycle length, specifically a change by a factor of α_{preg} = (1/1.1). Importantly, we considered that pregnancy reduces the progenitor population in our model. We simulated this change in population structure by decreasing the rate of asymmetric division of stem cells giving rise to progenitor cells by a factor of $p_{post,init}$ after an initial pregnancy. Our experiments suggested a 2-3 fold drop in $p27^+$ expressing progenitor cells, which suggests a value of $p_{post,init}$ = 0.5 (12).

We also modeled the effects of later pregnancies. In runs of the model with more than one birth, we considered the effect of the period of subsequent pregnancies to be the same as for the first birth. That is, the number of levels in the differentiation hierarchy of progenitor cells increases by z_{preg} levels, and the cell cycle length of stem cells decreases to $t_{cycle,preg}$ = 147 hours. Regarding the lasting effects of pregnancy on the structure of the breast epithelium, we allowed for the possibility of a smaller decrease in the probability of asymmetric stem cell division after later births compared to the decrease after the first birth, and defined a separate parameter, $p_{post,subs}$, for the decrease in asymmetric divisions after subsequent births.

Our simulation spanned from menarche to death or initiation of cancer within the duct. Our total simulation time was calculated from the average woman's life expectancy in the US, which was 81.2 years in 2014 (40), and the average age of menarche, which ranged between 12.2 – 12.8 years of age for different ethnic groups in 2007 (41). We used the mean age of menarche between the groups, which was 12.5 years and thus resulted in a total of 68.7 years of simulation time.

The parameters in Table S1 were set at fixed values from the literature. The parameters in Table S2 were set at values that fit to epidemiological data, as described below. We tested the robustness of the fit by varying each of these parameters individually.

RESULTS

We first investigated whether our model could quantitatively match the epidemiological data available on the protective effect of early pregnancy on breast cancer risk, within the space of biologically plausible parameters. From the literature, a woman with one birth at age 20 has a cumulative relative risk of ER+/PR+ breast cancer of 0.88 (C.I = 0.81 to 0.96) between the ages of 30 and 70, compared to a nulliparous woman, while a woman with four births at ages 20, 23, 26 and 29 has a cumulative relative risk of 0.71 (C.I. = 0.60 to 0.84) over the same age range (7). To match these rates, we varied the probability that a progenitor cell acquires the ability to self-renew, γ_{base}, and the reduction of the size of the p27+ progenitor cell population after the second pregnancy and later pregnancies, $p_{post,subs}$. We found that with $\gamma = 3.2 \times 10^{-3}$ and $p_{post,subs} = p_{post,init} = 0.5$, the modeled relative risk were within the confidence intervals reported in the literature for these two data points, at 0.86 and 0.73, respectively (Fig. 3A). Due to binning of modeled incidence into annual groups we considered risk during the 40-year period from age 30.5 to 70.5. Note that there are likely other parameter settings that could fit the data, in addition to those that we used; the ones presented here serve as an example of how our model can explain the data, rather than as an exact parameter estimation approach.

Using the fitted model, we first tested the effects of varying model parameters in the nulliparous simulations to test the behavior of the model. As expected, we found that the rate of cancer initiation per duct was increased by increasing the number of stem and progenitor cells per duct, the rate of asymmetric stem cell division, the mutation rate, the

probability of progenitor cells attaining self-renewal capacity, and the fitness advantage of mutated progenitor cells compared to wild type cells. By contrast, the rate of cancer initiation per duct was increased by decreasing the number of mutations required for cancer initiation. Also, as expected, changes in the proliferative capacity of progenitor cells during pregnancy, and the effects of subsequent pregnancies, have no effect in the nulliparous state (Fig. 3B, C).

We then tested the robustness of the fit of our model to the result that early pregnancy protects against breast cancer in the surrounding parameter space. We compared the relative likelihood of cancer initiation with pregnancy occurring at five year intervals during a woman's childbearing years as compared to the nulliparous simulations. We tested for the effects of pregnancy occurring from the age of menarche until immediately before menopause at the average age of 51.3 in 1998 (42). We tested the effects of varying the simulation parameters independently for each pregnancy age t_{preg}. All fixed value parameters are listed in Table S1, while Table S2 lists the values of all other parameters. We found that the probability of cancer initiation in a duct increases as the age of first pregnancy increases within the range of all simulated parameters (Fig. 4A). Additionally, the average probability of cancer initiation across birth ages was lower than the nulliparous risk for all parameter settings. Both of these effects were less marked under parameter settings in which most of the cancers resulted from the stem cells under the nulliparous scenario ($p = < 4x10^{-4}$ Spearman's rank correlation coefficient, in both cases). Indeed, the $z = 6$ setting had the highest proportion of stem cell cancers under the nulliparous setting.

We also investigated the effects of multiple births on cancer risk. We tested model runs with one, two, three, and four total births. For each of these cases, we investigated varying the age at first birth in five year intervals as above from the age of menarche to the age of menopause, assuming that all subsequent births were distributed evenly across the

intervening years between the first birth and the age of menopause. For all numbers of total births, risk increased with increasing age at first birth. Additionally, as expected, scenarios with a larger total number of births were at a lower risk compared to scenarios with fewer births (Fig. 4B).

We also tested for robustness of the quantitative fit to the two data points considered. As expected, we found that for some parameters the decrease in risk in the two modeled scenarios remained within the bounds of the confidence intervals for all settings tested, whereas for other parameters, there were some settings where the risk decrease did not match the literature values (Supplementary Fig. S1). In particular the quantitative fit to both data points was robust to changes in the cell cycle time of stem cells, the number of stem cells per duct, the fitness effects of mutations in stem cells, the number of additional progenitor cell divisions during pregnancy and the reduction in numbers of progenitors with subsequent births, within the range of values tested. The quantitative fit was also robust to decrease in the mutation rate in the range of values tested. Thus, our analysis demonstrates that the hypothesis can explain these two observed quantitative decreases in breast cancer risk, under some, but not all, plausible biological settings. Our hypothesis is thus one possible explanation for the observed protective effect of parity. However, we cannot rule out other possible explanations for the relatively limited amount of available data on the quantitative risk reduction.

Another interesting result is the specificity of the effect of the decrease in the progenitor pool with pregnancy to decrease the risk of cancers initiating from the progenitor compartment. We noted that the risk of cancers initiating from the progenitor cell compartment increased with age at first birth, while the risk of cancers where the final mutation occurs in the stem cell compartment showed a (smaller) decrease ($p < 4x10^{-5}$ in both cases under linear regression). Similarly, under the default parameter settings, whereas the risk of cancers initiated from the progenitor compartment was lower under all

parous scenarios compared to the nulliparous scenarios, the risk of cancers initiated from the stem compartment was slightly higher under all parous scenarios.

This result raises one possible explanation for the specificity of the protective effect of early pregnancy to ER+/PR+ cancer (7). Mounting experimental evidence suggests that the typical cell of origin of breast carcinomas is a stem or progenitor cell (43). The specificity of the protective effects in our model to a single cellular compartment poses the question of whether other breast cancer molecular subtypes may have a different cell of origin as a possible explanation for the observed specificity of protective effects. Relatedly it is also possible that changes during carcinogenesis render other breast cancer subtypes insensitive to hormone-driven growth or that some of the molecular parameters considered differ between breast cancer subtypes. By the same token, our model is agnostic on whether the pregnancy should protect against other histological breast cancer types, such as lobular cancers. Whether or not protective effects would be expected for these subtypes depend on the extent to which the etiology of these cancer types, in terms of cell of origin and other molecular parameters, corresponds to ER+ cancers.

As a further test of our framework, we investigated whether our model reproduced the known effect that breast cancer risk is increased for a short period immediately following pregnancy (6). For these purposes we investigated an extended model including a variable delay between initiation of cancer within the duct and clinical presentation. We investigated two scenarios, first birth at age 20, and first birth at age 40, and calculated the relative risk compared to nulliparous women of matched age in the years following pregnancy for varying average waiting times to clinical presentation between 0 and 5 years. We found that with an average waiting time of one year, relative risk in both parous scenarios was greater than one during the two years following the pregnancy (Fig. S2)

DISCUSSION

Here we investigated whether variation in the size of the progenitor cell population is sufficient to explain the protective effects of pregnancy. We used a simple mathematical model of the steps leading to cancer initiation, which included both stem cells and progenitor cells. We found that within the range of biologically plausible parameters, our model matches the observed decrease in ER+/PR+ cancer risk for a woman with a birth at age 20 and a woman with four births in her 20's compared to a nulliparous woman. Using these parameter settings, we found that the risk of cancer in our model decreased with increasing age of first birth in scenarios with one birth. Moreover, the risk of cancer was lower in all scenarios with one birth compared to the nulliparous case. This behavior was robust to variation in key model parameters. The ability of our model to robustly recreate the effect on cancer risk when varying the progenitor population size with pregnancy is striking given the modeled assumption that progenitor cells terminally differentiate after a finite number of divisions, so that mutations arising in progenitor cells are liable to leave the population without any functional impact. Taken together, these results support the hypothesis that a subset of p27+ cells represents quiescent hormone-responsive luminal progenitor cells with proliferative potential.

Our mathematical modeling approach for breast cancer can be useful in understanding the contribution of unavoidable bad luck to cancer risk. We have presented evidence that, in the setting of breast cancer, the size of a sub-population of progenitor cells may vary safely over the course of a life to alter breast cancer risk, independent of the probability of mutations. While it is possible that the mechanisms explored here are specific to the breast cancer setting, our results highlight the possibility that extrinsic factors can interact with molecular parameters to affect cancer risk in ways that are not yet fully mapped out. These results therefore further motivate the use of complementary approaches to assess the contribution of bad luck to cancer risk that do not rely on strong assumptions about the effects of extrinsic factors, which may still be subject to revision.

The modeling approach developed here is one such possible complementary approach. Therefore, the main implications of our study are support for a mechanism in the breast cancer setting, with potential implications for other cancers with an important role for hormone-driven growth, including endometrial and ovarian cancers. And, in addition, the current approach may be usefully applied in a range of cancer types.

In conclusion, our results demonstrate that variation in the size of the pool of progenitor cells with proliferative potential is capable of explaining the protective effect of early pregnancy against breast cancer. We obtained good agreement between our simple model's predictions and specific epidemiological data points within the range of plausible parameters. Intense recent debate, prompted by the work of Tomasetti and Vogelstein (1), has indicated the limits of regression techniques for determining the ultimate contribution of bad luck to cancer incidence. Continuing improvements in our mechanistic understanding of the etiology of different cancers can help elucidate the contribution of bad luck to cancer risk and the limits of cancer prevention strategies. Given the complexity of the molecular setting in which cancer develops, mathematical models can be a useful tool in developing such a mechanistic understanding. Our work has developed this approach for the case of breast cancer to provide evidence for a possible mechanism for the protective effect of early pregnancy against the disease.

ACKNOWLEDGMENTS

We thank members of the Michor and Polyak lab for their critical reading of our manuscript.

REFERENCES

1. Tomassetti C, Vogelstein B. Variation in cancer risk among tissues can be explained by the number of stem cell divisions. Science 2015;347:78-81
2. Couzin-Frankel J. Biomedicine. The bad luck of cancer. Science 2015;347:12
3. Weinberg CR, Zaykin D. Is bad luck the main cause of cancer? J Natl Cancer Inst 2015;107

4. Potter JD, Prentice RL. Cancer risk: tumors excluded. Science 2015;347:727
5. MacMahon B, Cole P, Lim T, Lowe A, Mirra B, Ravnihar B, et al. Age at First Birth and Breast Cancer Risk. Bull Wld Hlth Org 1970;43:209-21
6. Albrektsen G, Heuch I, Hansen S, Kvale G. Breast cancer risk by age at birth, time since birth and time intervals between births: exploring interaction effects. Br J Cancer 2005;92:167-75
7. Colditz GA, Rosner BA, Chen WY, Holmes MD, Hankinson SE. Risk Factors for Breast Cancer According to Estrogen and Progesterone Receptor Status. JNCI Journal of the National Cancer Institute 2004;96:218-28
8. Moolgavkar S, Day N, Stevens R. Two-Stage Model for Carcinogenesis- Epidemiology of Breast Cancer In Females. JNCI Journal of the National Cancer Institute 1980;65:559-69
9. Pike M, Krailo M, Henderson B, Casagrande J, Hoel D. 'Hormonal' risk factors, 'breast tissue age' and the age-incidence of breast cancer. Nature 1983;303:769-70
10. Rosner BA, Colditz GA, Willett C. Reproductive risk factors in a prospective study of breast cancer - the nurses' health study. American Jounral of Epidemiology 1994;139:819-35
11. Rosner B, Colditz GA. Nurses' health study: log-incidence mathematical model of breast cancer incidence. J Natl Cancer Inst 1996;88:359-64
12. Choudhury S, Almendro V, Merino VF, Wu Z, Maruyama R, Su Y, et al. Molecular profiling of human mammary gland links breast cancer risk to a p27(+) cell population with progenitor characteristics. Cell Stem Cell 2013;13:117-30
13. Huh SJ, Clement K, Jee D, Merlini A, Choudhury S, Maruyama R, et al. Age- and pregnancy-associated DNA methylation changes in mammary epithelial cells. Stem Cell Reports 2015;4:297-311
14. Eirew P, Stingl J, Raouf A, Turashvili G, Aparicio S, Emerman JT, et al. A method for quantifying normal human mammary epithelial stem cells with in vivo regenerative ability. Nat Med 2008;14:1384-9
15. Villadsen R, Fridriksdottir AJ, Ronnov-Jessen L, Gudjonsson T, Rank F, LaBarge MA, et al. Evidence for a stem cell hierarchy in the adult human breast. The Journal of cell biology 2007;177:87-101
16. Schiffer LM, Braunschweiger PG, Stragand JJ, Poulakos L. The cell kinetics of human mammary cancers. Cancer 1979;43:1707-19
17. Popnikolov N, Yang J, Liu A, Guzman R, Nandi S. Reconstituted normal human breast in nude mice: effect of host pregnancy environment and human chorionic gonadotropin on proliferation. The Journal of endocrinology 2001;168:487-96
18. Taylor D, Pearce CL, Hovanessian-Larsen L, Downey S, Spicer DV, Bartow S, et al. Progesterone and estrogen receptors in pregnant and premenopausal non-pregnant normal human breast. Breast cancer research and treatment 2009;118:161-8
19. Chung K, Hovanessian-Larsen LJ, Hawes D, Taylor D, Downey S, Spicer DV, et al. Breast epithelial cell proliferation is markedly increased with short-term high levels of endogenous estrogen secondary to controlled ovarian hyperstimulation. Breast cancer research and treatment 2012;132:653-60
20. Olsson H, Jernstrom H, Alm P, Kreipe H, Ingvar C, Jonsson PE, et al. Proliferation of the breast epithelium in relation to menstrual cycle phase, hormonal use, and reproductive factors. Breast cancer research and treatment 1996;40:187-96
21. Going JJ, Anderson TJ, Battersby S, MacIntyre CC. Proliferative and secretory activity in human breast during natural and artificial menstrual cycles. The American journal of pathology 1988;130:193-204
22. Anderson TJ, Battersby S, King RJ, McPherson K, Going JJ. Oral contraceptive use influences resting breast proliferation. Human pathology 1989;20:1139-44
23. Scholzen T, Gerdes J. The Ki-67 protein: from the known and the unknown. Journal of cellular physiology 2000;182:311-22

24. Cooper GM. The cell a molecular approach. NCBI bookshelf. 2nd ed. Sunderland, Mass.: Sinauer Associates; 2000.
25. Huh SJ, Oh H, Peterson MA, Almendro V, Hu R, Bowden M, et al. The Proliferative Activity of Mammary Epithelial Cells in Normal Tissue Predicts Breast Cancer Risk in Premenopausal Women. Cancer Res 2016;76:1926-34
26. Moran PAP. The statistical processes of evolutionary theory. Clarendon Press; 1962.
27. Hambardzumyan D, Cheng YK, Haeno H, Holland EC, Michor F. The probable cell of origin of NF1- and PDGF-driven glioblastomas. PloS one 2011;6:e24454
28. Traulsen A, Lenaerts T, Pacheco JM, Dingli D. On the dynamics of neutral mutations in a mathematical model for a homogeneous stem cell population. Journal of the Royal Society, Interface 2013;10:20120810
29. Foo J, Liu LL, Leder K, Riester M, Iwasa Y, Lengauer C, et al. An Evolutionary Approach for Identifying Driver Mutations in Colorectal Cancer. PLoS computational biology 2015;11:e1004350
30. Bresciani F. Cell proliferation in cancer. European journal of cancer 1968;4:343-66
31. Daniel CW, Young LJ. Influence of cell division on an aging process. Life span of mouse mammary epithelium during serial propagation in vivo. Experimental cell research 1971;65:27-32
32. Faulkin LJ, Jr., Deome KB. Regulation of growth and spacing of gland elements in the mammary fat pad of the C3H mouse. J Natl Cancer Inst 1960;24:953-69
33. Kordon EC, Smith GH. An entire functional mammary gland may comprise the progeny from a single cell. Development (Cambridge, England) 1998;125:1921-30
34. Jones S, Chen WD, Parmigiani G, Diehl F, Beerenwinkel N, Antal T, et al. Comparative lesion sequencing provides insights into tumor evolution. Proc Natl Acad Sci U S A 2008;105:4283-8
35. Salk JJ, Fox EJ, Loeb LA. Mutational heterogeneity in human cancers: origin and consequences. Annual review of pathology 2010;5:51-75
36. Bozic I, Antal T, Ohtsuki H, Carter H, Kim D, Chen S, et al. Accumulation of driver and passenger mutations during tumor progression. Proc Natl Acad Sci U S A 2010;107:18545-50
37. Russo J, Moral R, Balogh GA, Mailo D, Russo IH. The protective role of pregnancy in breast cancer. Breast Cancer Res 2005;7:131-42
38. Russo J, Rivera R, Russo IH. Influence of age and parity on the development of the human breast. Breast cancer research and treatment 1992;23:211-8
39. Suzuki R, Atherton AJ, O'Hare MJ, Entwistle A, Lakhani SR, Clarke C. Proliferation and differentiation in the human breast during pregnancy. Differentiation; research in biological diversity 2000;66:106-15
40. Statistics NCFH. Health, United States, 2015: With Special Features on Racial and Ethnic Health Disparities. 2016
41. Cabrera SM, Bright GM, Frane JW, Blethen SL, Lee PA. Age of thelarche and menarche in contemporary US females: a cross-sectional analysis. Journal of pediatric endocrinology & metabolism : JPEM 2014;27:47-51
42. Kato I, Toniolo P, Akhmedkhanov A, Koenig KL, Shore R, Zeleniuch-Jacquotte A. Prospective study of factors influencing the onset of natural menopause. Journal of clinical epidemiology 1998;51:1271-6
43. Visvader JE. Cells of origin in cancer. Nature 2011;469:314-22

FIGURE LEGENDS

Figure 1. Multiple factors can affect cancer risk in a complex setting. A, An analysis by Tomassetti and Vogelstein demonstrated a close correlation between the log of lifetime cancer incidence in a tissue and the cumulative number of stem cell divisions in the same tissue. Plot shown is a schematic using simulated data. **B,** Variation in multiple molecular factors may affect cancer risk when they change from the homeostatic state (**top left**), including the number of progenitor cells (**top right**), the mutation rate (**bottom left**), and the fitness effect conferred by mutations (**bottom right)**.

Figure 2. Schematic representation of the mathematical model. A, Initially, there are N wild-type stem cells (blue), which give rise to a differentiation cascade of $2^{z+1} - 1$ wild-type luminal progenitor cells (purple). At each time step, all progenitor cells as well as one randomly selected stem cell divide. With probability 1 - p, the stem cell divides symmetrically and one daughter cell replaces another randomly chosen stem cell. With probability p, the stem cell divides asymmetrically and one daughter cell remains a stem cell while the other daughter cell becomes committed to the progenitor population (light pink). Regardless of the dividing stem cell's fate, all existing progenitor cells divide symmetrically for a total of z times to give rise to successively more differentiated cells (progressively darker shades of purple) before becoming terminally differentiated. In the figure, the darkening purple gradations refer to successively more differentiated cells and serve to clarify a single time step of the stochastic process. **B,** The acquisition of mutations leading to breast cancer initiation all result in an increased relative fitness (i.e. growth rate) f_{mut} in stem cells (red) as compared to wild-type cells (blue) and an additional number of divisions z_{mut} progenitor cells can undergo before terminally differentiating. **C,** During pregnancy, progenitor cells experience an expansion in proliferative capacity through an

additional number of divisions z_{preg} in order to form terminally differentiated milk-producing cells (dotted triangle) and a decrease in cell cycle length.

Figure 3. Model fitting and effect of parameter variation on cancer initiation in nulliparous simulations. A, Evolution of initiation-free ducts with age under the default parameter settings for three birth scenarios (nulliparous, a single birth at age 20, and four births at ages 20, 23, 26 and 29). **B,** Effects of varying individual parameters of the model on nulliparous cancer initiation. **C,** Evolution of intiation-free ducts with age under the nulliparous scenario for different settings of the probability of asymmetric division **(top)**, the mutation rate **(middle)**, and the number of mutations required for cancer **(bottom)**. Default values were $N = 8$, $z = 10$, $p = 10^{-2}$, $\mu = 2 \times 10^{-6}$, $f_{mut} = 1.1$, $z_{mut} = 1$, $n_{mut} = 2$, $z_{preg} = 2$,

Figure 4. Relative probability of cancer initiation per duct as compared to nulliparous simulations. A), Variation in cancer initiation relative to nulliparous for different ages at first birth under default parameter settings (green lines), and when varying individual model parameters upwards (red lines) or downwards (blue lines). Left to right from top left, effects of varying stem cell cell cyle time, number of stem cells, number of progenitors**,** probabilities of stem cell differentiation*,* mutation rate, probability of progenitor cells attaining ability to self-renew, fitness effects of mutations, number of mutations required for cancer initiation, and additional pregnancy divisions, are shown. **B)**, Variation in cancer initiation relative to nulliparous for different ages at first birth and different numbers of total births

Figure 1

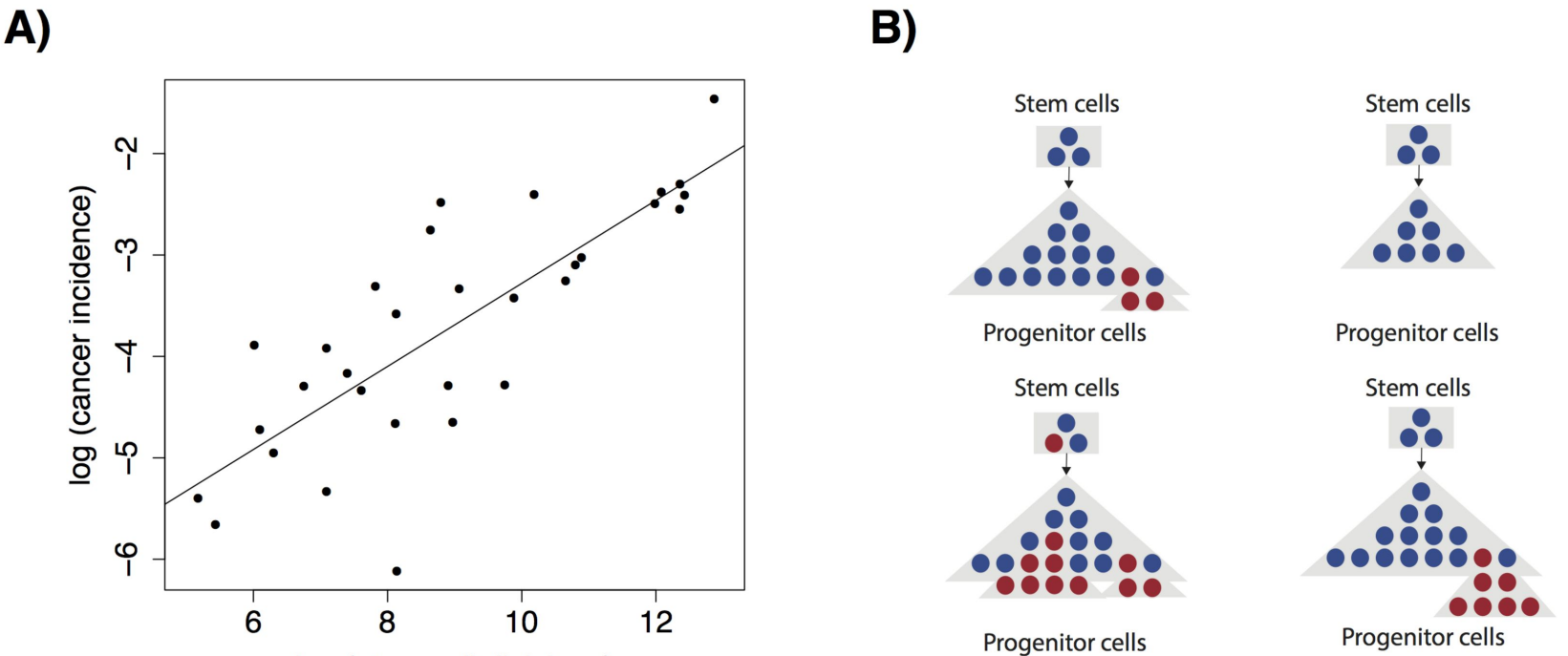
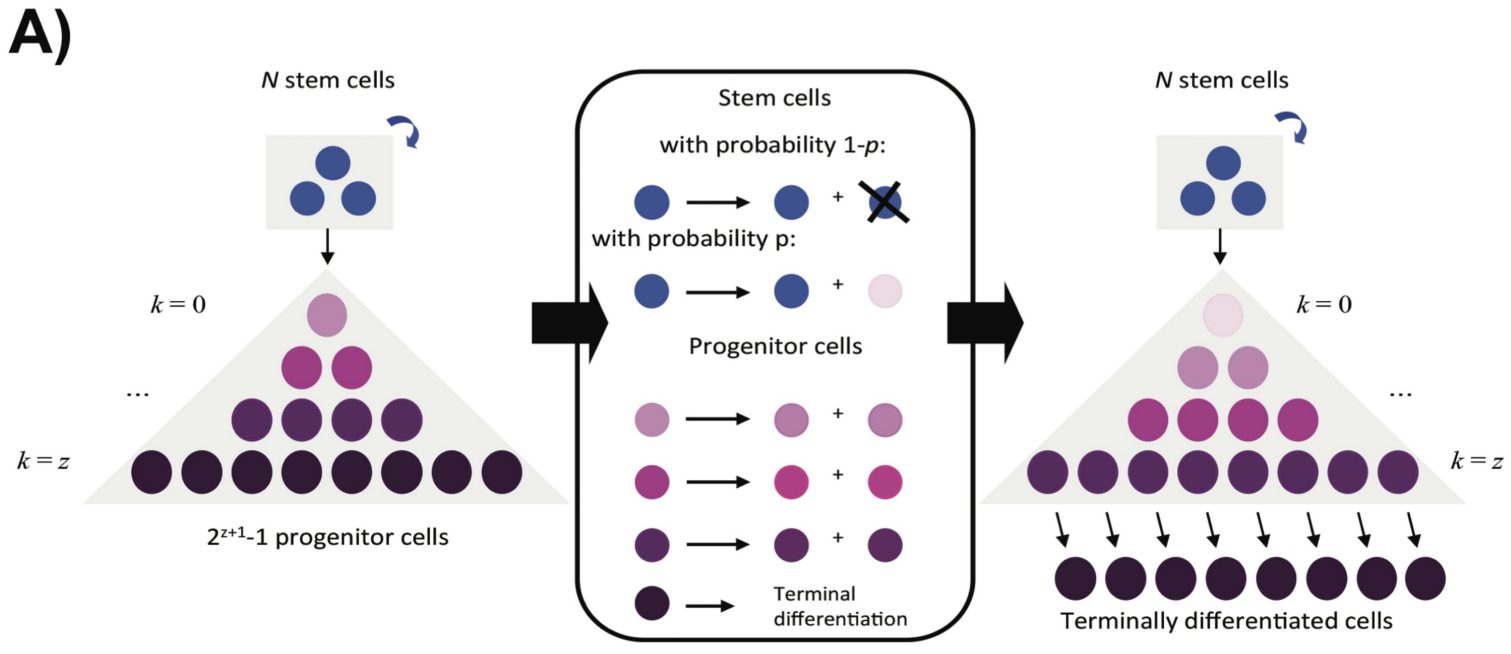
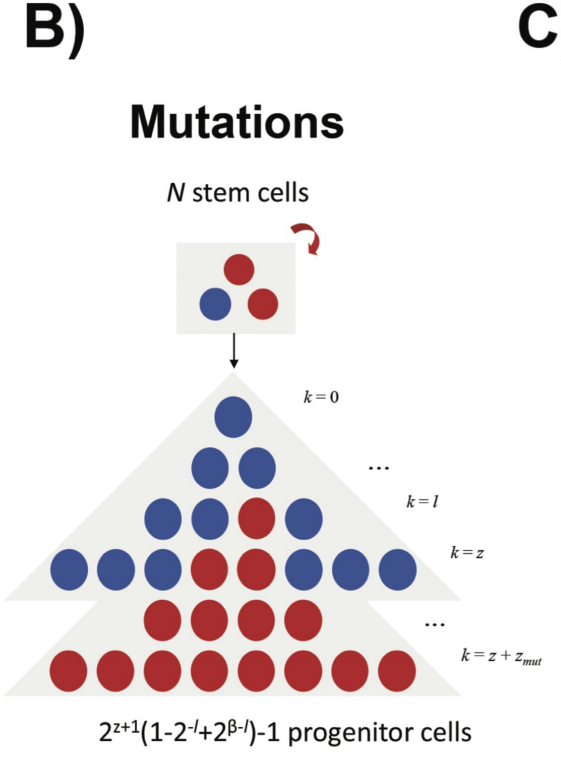

A)
log (cancer incidence)
−2
−3
−4
−5
−6
6
8
10
12
log (stem cell divisions)
B)
Stem cells
Progenitor cells
Stem cells
Progenitor cells
Stem cells
Progenitor cells
Stem cells
Progenitor cells

Figure 2


A)
N stem cells
k = 0
...
k = z
2z+1-1 progenitor cells
Stem cells
with probability 1-p:
with probability p:
Progenitor cells
Terminal differentiation
N stem cells
k = 0
...
k = z
Terminally differentiated cells


B)
Mutations
N stem cells
k = 0
...
k = l
k = z
...
k = z + zmut
2z+1(1-2-l+2β-l)-1 progenitor cells

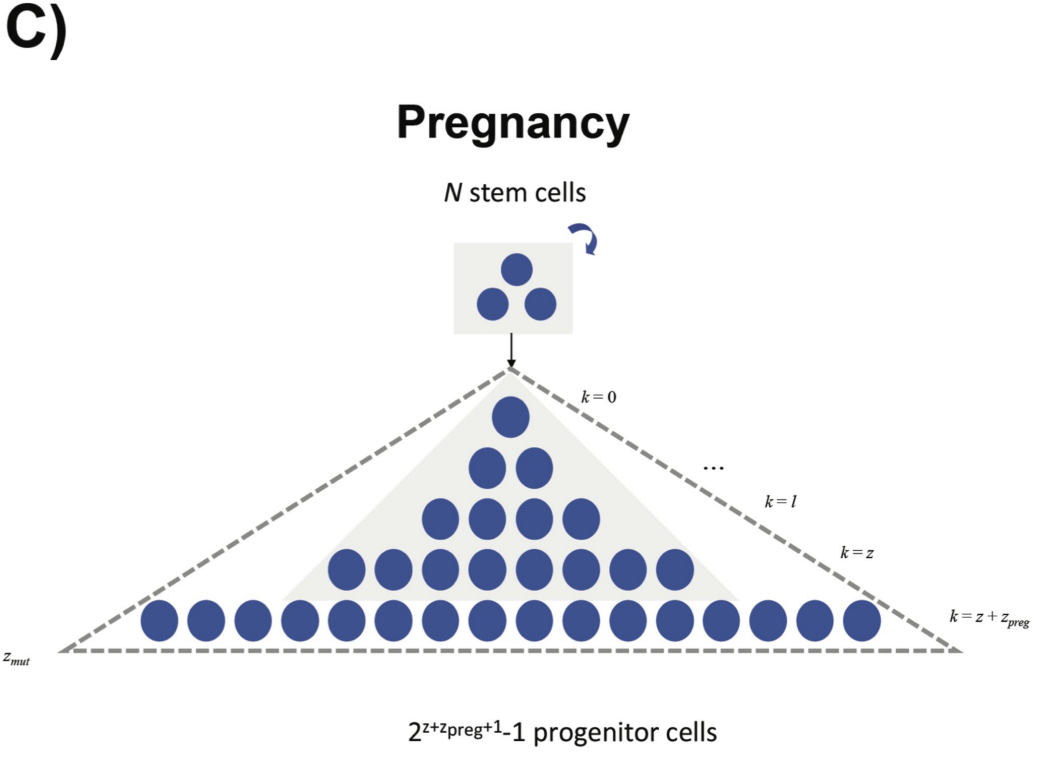

C)
Pregnancy
N stem cells
k = 0
...
k = l
k = z
k = z + zpreg
2z+zpreg+1-1 progenitor cells

Figure 3

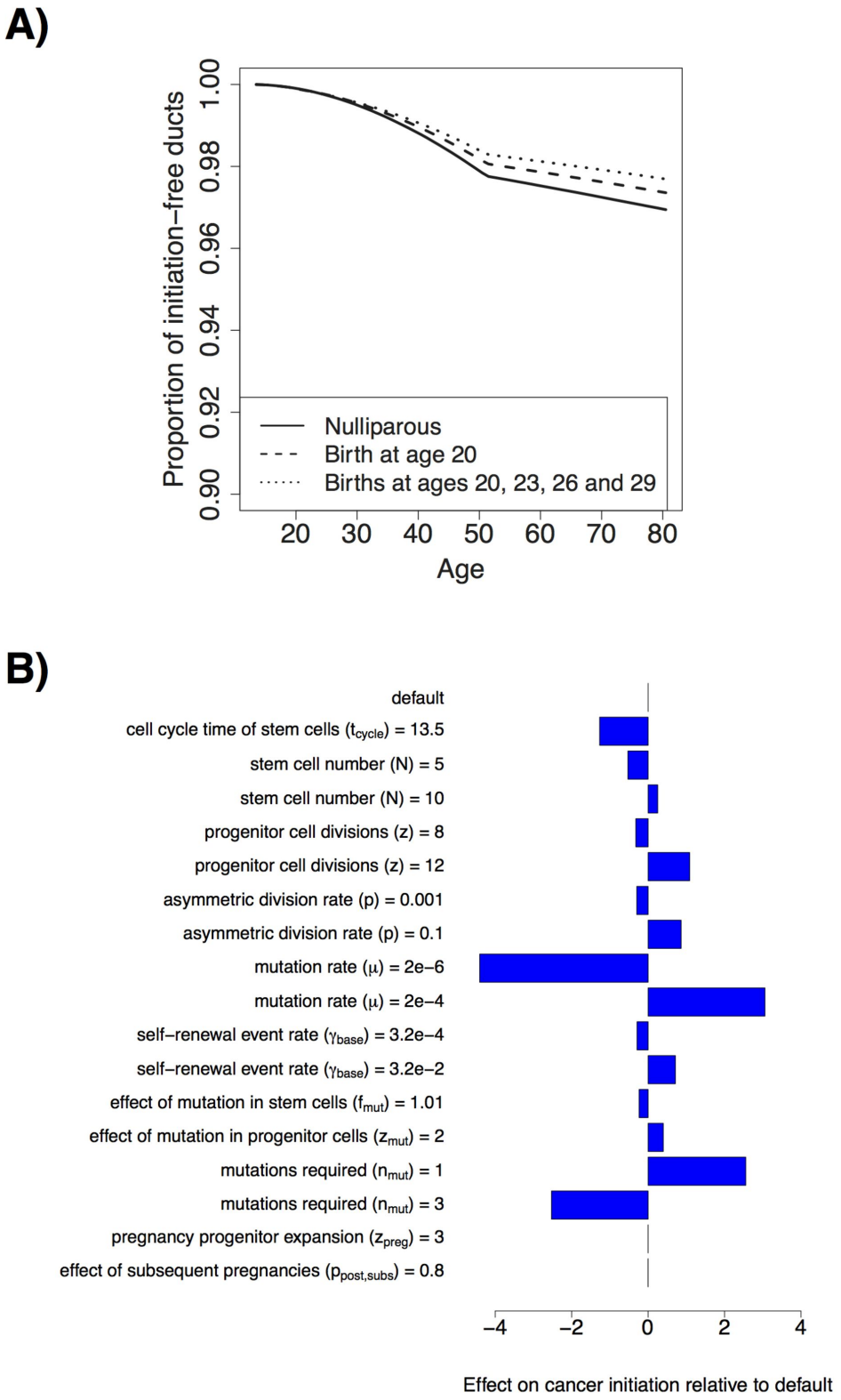
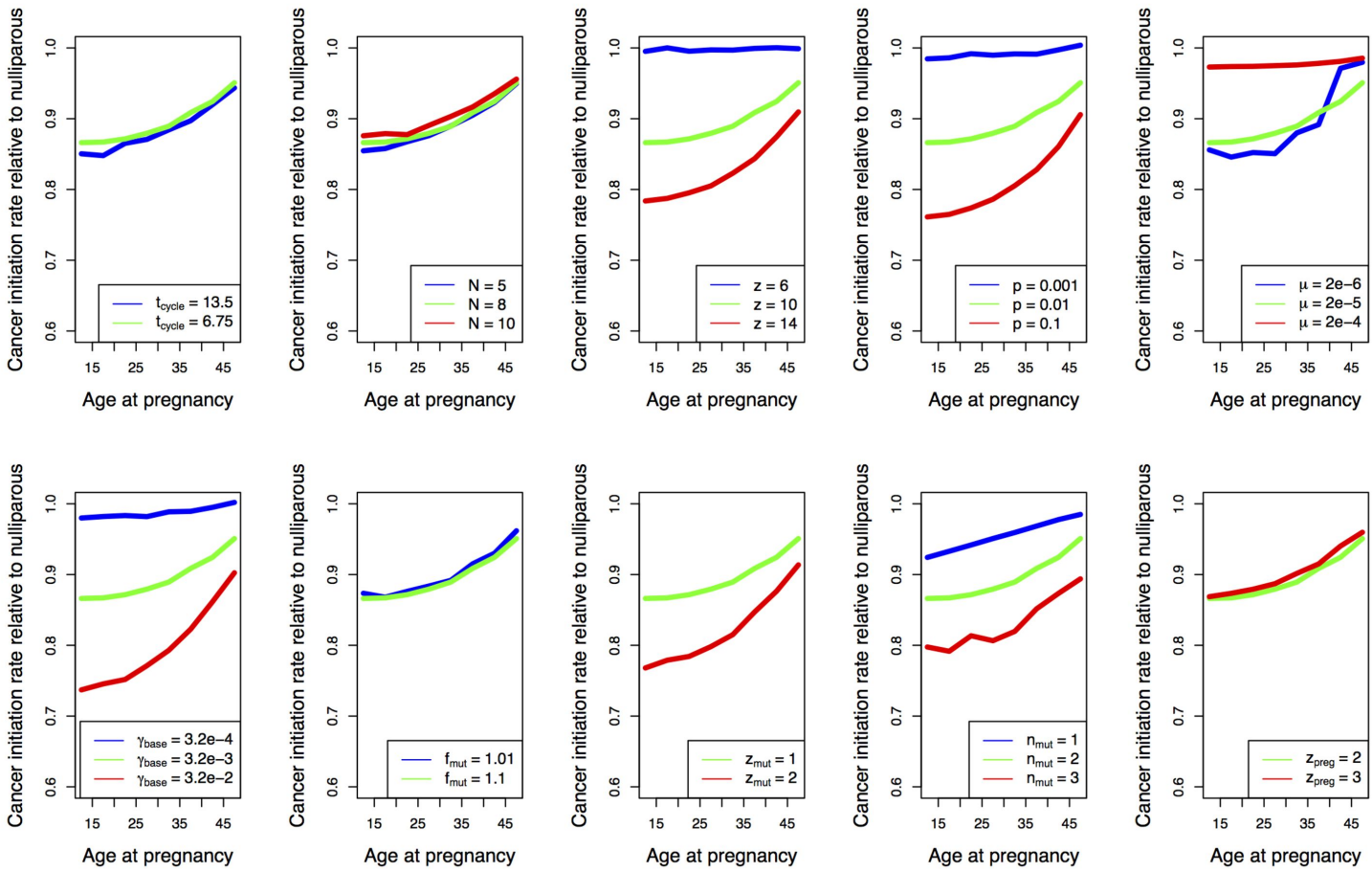
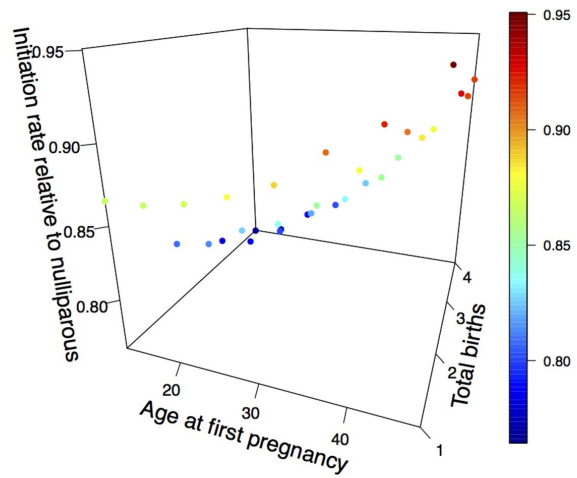

A)
Proportion of initiation−free ducts
0.90
0.92
0.94
0.96
0.98
1.00
20
30
40
50
60
70
80
Age
Nulliparous
Birth at age 20
Births at ages 20, 23, 26 and 29
B)
default
cell cycle time of stem cells (t_{cycle}) = 13.5
stem cell number (N) = 5
stem cell number (N) = 10
progenitor cell divisions (z) = 8
progenitor cell divisions (z) = 12
asymmetric division rate (p) = 0.001
asymmetric division rate (p) = 0.1
mutation rate (μ) = 2e−6
mutation rate (μ) = 2e−4
self−renewal event rate (γ_{base}) = 3.2e−4
self−renewal event rate (γ_{base}) = 3.2e−2
effect of mutation in stem cells (f_{mut}) = 1.01
effect of mutation in progenitor cells (z_{mut}) = 2
mutations required (n_{mut}) = 1
mutations required (n_{mut}) = 3
pregnancy progenitor expansion (z_{preg}) = 3
effect of subsequent pregnancies ($p_{post,subs}$) = 0.8
−4
−2
0
2
4
Effect on cancer initiation relative to default
C)
Proportion of initiation−free ducts
0.90
0.92
0.94
0.96
0.98
1.00
20
30
40
50
60
70
80
Age
asymmetric division rate (p) = 0.001
asymmetric division rate (p) = 0.01
asymmetric division rate (p) = 0.1
Proportion of initiation−free ducts
Age
mutation rate (μ) = 2e−6
mutation rate (μ) = 2e−5
mutation rate (μ) = 2e−4
Proportion of initiation−free ducts
Age
mutant progenitor divisions (z_{mut}) = 1
mutant progenitor divisions (z_{mut}) = 2

Figure 4

A)


Cancer initiation rate relative to nulliparous
Age at pregnancy
t_{cycle} = 13.5
t_{cycle} = 6.75
N = 5
N = 8
N = 10
z = 6
z = 10
z = 14
p = 0.001
p = 0.01
p = 0.1
μ = 2e−6
μ = 2e−5
μ = 2e−4
γ_{base} = 3.2e−4
γ_{base} = 3.2e−3
γ_{base} = 3.2e−2
f_{mut} = 1.01
f_{mut} = 1.1
z_{mut} = 1
z_{mut} = 2
n_{mut} = 1
n_{mut} = 2
n_{mut} = 3
z_{preg} = 2
z_{preg} = 3

B)


Initiation rate relative to nulliparous
Age at first pregnancy
Total births