Stony Brook University

Stochastic Modeling of Cell Dynamics,

Mutation Acquisition and Cancer Risk

A Dissertation Presented

by

Mu Tian

to

The Graduate School

in Partial Fulfillment of the

Requirements

for the Degree of

Doctor of Philosophy

in

Applied Mathematics and Statistics

Stony Brook University

August 2017

Stony Brook University

The Graduate School

Mu Tian

We, the dissertation committee for the above candidate for the

Doctor of Philosophy degree, hereby recommend

acceptance of this dissertation.

Song Wu – Dissertation Advisor
Associate Professor
Department of Applied Mathematics and Statistics

Wei Zhu – Co-advisor
Professor
Department of Applied Mathematics and Statistics

Pei Fen Kuan - Chairperson of Defense
Assistant Professor
Department of Applied Mathematics and Statistics

Yusuf Hannun – External Committee Member
Professor
Stony Brook University School of Medicine
Director, Stony Brook Cancer Center

This dissertation is accepted by the Graduate School

Charles Taber
Dean of the Graduate School

Abstract of the Dissertation

Stochastic Modeling of Cell Dynamics,

Mutation Acquisition and Cancer Risk

by

Mu Tian

Doctor of Philosophy

in

Applied Mathematics and Statistics

Stony Brook University

2017

It is well recognized that cancer results from multi-stage mutation acquisitions. To this end, both intrinsic and extrinsic factors contribute to mutagenesis in cancer and subsequently the risk of cancer. To better understand the process of cancer initiation and the contributions of various risk factors, we build stochastic models for carcinogenesis based on modern cancer stem cell theory with clonal expansion. In our extended risk model, we have incorporated all three types of cell lineages including stem cells, progenitor cells and terminal cells. We have also included major ingredients for cancer development, including general cell behaviors, tissue homeostasis, multi-stage mutation acquisition, as well as how driver mutations may alter cell behaviors through cell fitness or clonal expansion.

Our model provides a general framework for estimating cancer risk and cancer mutation distributions at any age in a lifetime. With these models, we can simulate and analyze the effect of different factors on the speed, magnitude and risk of cancer onset. In particular, for each cancer, based on observed cancer risk data, we can quantify (1) the amount of lifetime risk due to the intrinsic mutations alone, that is, the intrinsic risk, or as the media calls, the 'bad luck', and (2) the percent of mutations due to intrinsic risk alone. Applying our modeling in conjunction with the US and the World cancer registry data, we found that non-intrinsic risk accounts for not only the major percentage of lifetime cancer risk, but also the major proportion of lifetime cancer mutations.

Table of Contents

Chapter 1 Introduction 1

Chapter 2 Cell Dynamics and Mutation Acquisition 6

2.1 Cell and Cell Divisions 6
2.2 Mutation Acquisition 7
2.3 Cancer Risk 8
2.4 Cell Dynamics 8
2.5 Representative models 9
2.5.1 Original Stem Cell Model 10
2.5.2 Stem-Progenitor Cell Model 14
2.5.3 Discrete Time Stochastic Model 18
2.5.4 Continuous Time Model 19
2.5.5 Spatial Dynamics Model 22

Chapter 3 Intermediate Risk Model 24

3.1 Cell Dynamics and Homeostatic Condition 24
3.2 Mutation Acquisition Probability Propagation 29
3.3 Cancer Risk Computation 30

Chapter 4 Extended Risk Model 39

4.1 General Division Patterns 40
4.2 Cell Dynamics and Homeostatic Condition 43
4.3 Mutation Acquisition with different driver mutations 48
4.4 Extended theoretical lifetime intrinsic risk 51
4.4.1 Extended Risk Computation 56
4.4.2 Mutation Effects 62
4.4.3 Clonal Expansion and Regulation 63
4.4.4 Age Dependent Cancer Risk 65
4.5 General Mutation State Evolution 65
4.5.1 Cell Number Transition 66
4.5.2 Clonal Expansion 68
4.6 The Contributions of Intrinsic and Extrinsic Factors 70
4.6.1 Estimating Total Mutation Rate 71
4.6.2 Building Lower/Upper Bound Using the Original Stem Cell Model 74

Chapter 5. Results and Discussions 76

5.1 Preliminary Study with 31 Cancer Types 78
5.1.1 Age Dependent Cancer Risk 78
5.1.2 Intrinsic Risk and Intrinsic Contribution 85
5.2 Extensive Study with 18 Tissue Types 103
5.2.1 Lifetime Risk and Intrinsic Contribution 104
5.2.2 Age Dependent Risk and Intrinsic Contribution 112

5.2.3 Comparing Different Models *122*
5.3 Discussion and Future Work 129
Reference **133**

List of Figures

Figure 1 [23]: An illustration of stem cells, progenitor cells and terminal cells. The proposed locations of stem cells (purple), progenitor cells (green) and super-basal, terminally differentiating cells (pink) are shown in human interfollicular epidermis. Arrows represent the relationship between each cell compartment and the movement of cells to the surface of the skin as they undergo terminal differentiation. 3

Figure 2 [38]: Illustration of stem cells, progenitor cells and terminal cells with their general division patterns: 1 – symmetric stem cell division, 2 – asymmetric stem cell division, 3 – progenitor division, and 4 – terminal differentiation. 6

Figure 3 [34]: Stem cell division pattern and dynamics. Each original (Generation 0) stem cell first goes through nss symmetric divisions and then nas asymmetric divisions. 11

Figure 4 [34]: The propagation diagram of driver mutation states in one cell from one generation to the next. 12

Figure 5 [8]: The pattern of cell division giving rise to a total of N cells. The single, initial stem cell divides to produce a stem cell and a progenitor lineage. Each progenitor lineage divides np times, yielding 2np cells. The stem lineage divides nas times. The total number of cells produced are $N = nas2np$. 15

Figure 6 [1]: (A) Mutations: Schematic view of the multistep process of mutation acquisition. (B) Stem cell division modes: symmetric self-renewal division results in two daughter stem cells, asymmetric self-renewal division results in one stem cell and one progenitor cell, and symmetric differentiation division results in two progenitor cells..... 20

Figure 7: Cell division patterns and dynamics in our intermediate cancer risk model. Initial (Generation 0) stem cell will go through nss symmetric divisions and nas asymmetric divisions. At each stem cell symmetric division, two daughter stem cells are generated. At each stem cell asymmetric division, one stem cell and one progenitor cell are generated. Each progenitor cell generated from stem cell will go through np symmetric divisions. Each progenitor cell division will give birth to two daughter progenitor cells. Each npth generation progenitor cell will further divide into two progenitor cells that will eventually evolve to terminal cells and die. 25

Figure 8: Illustration of one stem cell asymmetric division lineage. This lineage includes one asymmetric division stem cell lineage along with all progenitor lineages that are originated from the asymmetric lineage stem cells. The stem cell asymmetric generation is the relative generation of a stem cell that has asymmetric divisions. The asymmetric generation 0 is total generation nss: the first generation of stem cells that start asymmetric divisions. 30

Figure 9: Illustration of the progenitor-terminal lineage originated from one single stem cell of some asymmetric generation $0 \leq g < nas$. 31

Figure 10: Mutation state notations in one progenitor-terminal lineage originated from one stem cell. Let $X = j$ be the mutation state of the stem cell. Let random variable Xnp be the mutation state of any final generation progenitor cell. Let random variable Xt be the mutation state of either one of daughter terminal cells generated from the progenitor cell. 33

Figure 11: A sensitivity analysis of different mutation rates on $ms = 3$. The figure provides computed cancer risk level using stem-cell-only model in [7] with mutation

rates: $1\times10\text{-}10$, $1\times10\text{-}9$, $1\times10\text{-}8$, and $2.5\times10\text{-}8$ and also the upper bound cancer risk level computed through our model for the setting: $\mathrm{us} = \mathrm{up} = 2.5\times10\text{-}8$ and $\mathrm{ms} = 3, \mathrm{mp} = 4, \mathrm{mt} = 5$. 38

Figure 12: General types of cell divisions. 40

Figure 13: General lifetime cell evolution structure. 42

Figure 14: Illustration of the multi-mutation state space, where the mutation state of each cell is represented by a vector of length equals to the total number of driver mutations considered. 50

Figure 15: Illustration of one general cell division, from Generation 0 to Generation 1. Parent cell (root) divides into two daughter cells, left child (lc) and right child (rc). The number of mutations (mutation state) on each cell is possibly from 0 to m. We use $\pi\mathrm{i}(\cdot)$ to represent the probability that the cell at the corresponding generation has i mutations. Given the mutation state in parent cell (root), each daughter cells will inherit all mutations in parent, and will independently acquire extra mutations according to the transition probability pij. 52

Figure 16: Illustration of binary tree structure in cell division 57

Figure 17: Probability recursion in lifetime general cell division structures. 62

Figure 18: Illustration of clonal expansion of factor $\alpha\mathrm{S} = 2$ for stem cells. For each division of normal cells, expansion cells will have two divisions. Note that after each division of normal cells, a proportion of daughter cells will become expansion cells, due to stochastic mutation acquisition. 63

Figure 19: This figure shows the log scale cancer risk from SEER and our model from age 1 to 80. At age 80, the observed SEER cancer risk is 0.001 while the computed

cancer risks from our mode are all below the observed risk, for mutation rates ranging from 1e-09 to 5e-08. In addition, we can see an overall trend of cancer risk is similar to SEER but the trend is different at both early age and middle age. The observation indicates significant contribution from age dependent potential extrinsic risk, such as alcohol, smoke, immune system, etc. 81

Figure 20: This figure shows the log scale cancer risk from SEER and our model from age 1 to 80 by varying the factor of mutation rate increase as an effect of certain mutations. The factor ranges from 1.0 (no effect) to 4.0. The baseline mutation rate was chosen to be 1e-08. 82

Figure 21: This figure shows the log10 scale cancer risk from SEER and our model from age 1 to 80 by varying the factor of clonal expansion. The factor ranges from 1.0 (no effect) to 3.0. The baseline mutation rate was chosen to be 1e-08. In here we restrict the clonal expansion to be effective only after 40 weeks, when homeostasis is approximately achieved. We see that a 2 or 3 fold clonal expansion would lead to increase of cancer risk but the overall risk is still well below the observed risk from SEER. This figure again indicates the potential contribution from extrinsic factors. 83

Figure 22: This figure shows the log10 scale cancer risk from SEER and our model from age 1 to 80 by varying the factor of both increasing mutation rate and clonal expansion. The mutation rate increase factor (emr) ranges from 1.0 (no effect) to 4.0 and the clonal expansion factor (fd) ranges from 1.0 to 3.0. The baseline mutation rate was chosen to be 1e-08. In here we restrict the clonal expansion to be effective only after 40 weeks, when homeostasis is approximately achieved. We see that the

increasing combined effects would lead to increase of cancer risk but the overall risk is still well below the observed risk from SEER. This figure again indicates the potential contribution from extrinsic factors..... 84

Figure 23: Lifetime intrinsic risk (log10 scale) computed from Extended Risk Model and statistics of NPCR observed risk in U.S. Tissue id/names are given below horizontal axis and the tissues are sorted in ascending order of "risk_observed", the average risk in U.S. Intrinsic mutation rate is selected to be {1e-08, 1.1e-08, 2.5e-08, 1e-07}; no mutation effects and clonal expansion were applied here ($\mathrm{emr} = 1.0$ and $\mathrm{fd} = 1.0$); also default required mutation hits were used ($\mathrm{mtS} = 00111, \mathrm{mtP} = 01111$ and $\mathrm{mtT} = 11111$). 104

Figure 24: Lifetime intrinsic risk (log10 scale) computed from Extended Risk Model and statistics of NPCR observed risk in U.S. Tissue id/names are given below horizontal axis and the tissues are sorted in ascending order of "risk_observed", the average risk in U.S. Intrinsic mutation rate is selected to be {1e-08, 1.1e-08, 2.5e-08, 1e-07}. Mutation effects and clonal expansion were applied here with factors $\mathrm{emr} = 2.0$ and $\mathrm{fd} = 2.0$; also default required mutation hits were used ($\mathrm{mtS} = 00111, \mathrm{mtP} = 01111$ and $\mathrm{mtT} = 11111$). 105

Figure 25: Lifetime intrinsic contribution (log10 scale) computed from Extended Risk Model and intrinsic contribution percentage, "Intrinsic_Contribution_TV", reported by Tomasetti, Li and Vogelstein [41] except Small Intestine and Head & Neck. Tissue id/names are given below horizontal axis and the tissues are sorted in ascending order of "risk_observed", the average risk in U.S. Intrinsic mutation rate is selected to be {1e-08, 1.1e-08, 2.5e-08, 1e-07}. No mutation effects and clonal

expansion were applied here ($\mathrm{emr} = 1.0$ and $\mathrm{fd} = 1.0$); also default required mutation hits were used ($\mathrm{mtS} = 00111, \mathrm{mtP} = 01111$ and $\mathrm{mtT} = 11111$). Note that in rare cases where computed intrinsic contribution is greater than 1.0, (there are more mutations acquired due to intrinsic rate than that due to estimated total rate), the intrinsic contribution was set to 1.0 (0.0 in log10 scale). 106

Figure 26: Lifetime intrinsic contribution (log10 scale) computed from Extended Risk Model and intrinsic contribution percentage, "Intrinsic_Contribution_TV", reported by Tomasetti, Li and Vogelstein [41] except Small Intestine and Head & Neck. Tissue id/names are given below horizontal axis and the tissues are sorted in ascending order of "risk_observed", the average risk in U.S. Intrinsic mutation rate is selected to be {1e-08, 1.1e-08, 2.5e-08, 1e-07}. Mutation effects and clonal expansion were applied here with factors $\mathrm{emr} = 2.0$ and $\mathrm{fd} = 2.0$; also default required mutation hits were used ($\mathrm{mtS} = 00111, \mathrm{mtP} = 01111$ and $\mathrm{mtT} = 11111$). Note that in rare cases where computed intrinsic contribution is greater than 1.0, (there are more mutations acquired due to intrinsic rate than that due to estimated total rate), the intrinsic contribution was set to 1.0 (0.0 in log10 scale). 107

Figure 27: Lifetime intrinsic risk (log10 scale) computed from Extended Risk Model and statistics of NPCR observed risk in U.S. Tissue id/names are given below horizontal axis and the tissues are sorted in ascending order of "risk_observed", the average risk in U.S. Intrinsic mutation rate is selected to be {1e-07}. Mutation effects and clonal expansion were applied here with factors ranging from $\mathrm{emr} = \{1.0,2.0\}$ and $\mathrm{fd} = \{1.0,2.0\}$; also we consider different required mutation hits: ($\mathrm{mtS} =$

$00111, \text{mtP} = 01111$ and $\text{mtT} = 11111$) vs. ($\text{mtS} = 01111, \text{mtP} = 11111$ and $\text{mtT} = 11111$). 108

Figure 28: Lifetime intrinsic contribution (log10 scale) computed from Extended Risk Model and intrinsic contribution percentage, "Intrinsic_Contribution_TV", reported by Tomasetti, Li and Vogelstein [41] except Small Intestine and Head & Neck. Tissue id/names are given below horizontal axis and the tissues are sorted in ascending order of "risk_observed", the average risk in U.S. Intrinsic mutation rate is selected to be {1e-07}. Mutation effects and clonal expansion were applied here with factors ranging from $\text{emr} = \{1.0,2.0\}$ and $\text{fd} = \{1.0,2.0\}$; also we consider different required mutation hits: ($\text{mtS} = 00111, \text{mtP} = 01111$ and $\text{mtT} = 11111$) vs. ($\text{mtS} = 01111, \text{mtP} = 11111$ and $\text{mtT} = 11111$). Note that in cases where computed intrinsic contribution is greater than 1.0, (there are more mutations acquired due to intrinsic rate than that due to estimated total rate), the intrinsic contribution was set to 1.0 (0.0 in log10 scale). 109

Figure 29: Lifetime intrinsic risk (log10 scale) computed from Extended Risk Model and statistics of NPCR observed risk in U.S. Tissue id/names are given below horizontal axis and the tissues are sorted in ascending order of "risk_observed", the average risk in U.S. Intrinsic mutation rate is selected to be {1e-08, 1.1e-08, 2.5e-08, 1e-07}; mutation effects and clonal expansion factors are selected from $\text{emr} = \{1.0, 2.0\}$ and $\text{fd} = \{1.0, 2.0\}$; the default required mutation hits are ($\text{mtS} = 00111, \text{mtP} = 01111$ and $\text{mtT} = 11111$); for small intestine, leukemia and colon, we add an additional set of mutation hits ($\text{mtS} = 01111, \text{mtP} = 11111$ and $\text{mtT} = 11111$). 110

Figure 30: Lifetime intrinsic contribution (log10 scale) computed from Extended Risk Model and intrinsic contribution percentage, "Intrinsic_Contribution_TV", reported by Tomasetti, Li and Vogelstein [41] except Small Intestine and Head & Neck. Tissue id/names are given below horizontal axis and the tissues are sorted in ascending order of "risk_observed", the average risk in U.S. Intrinsic mutation rate is selected to be {1e-08, 1.1e-08, 2.5e-08, 1e-07}; mutation effects and clonal expansion factors are selected from $\mathrm{emr} = \{1.0, 2.0\}$ and $\mathrm{fd} = \{1.0, 2.0\}$; the default required mutation hits are ($\mathrm{mtS} = 00111, \mathrm{mtP} = 01111$ and $\mathrm{mtT} = 11111$); for small intestine, leukemia and colon, we add an additional set of mutation hits ($\mathrm{mtS} = 01111, \mathrm{mtP} = 11111$ and $\mathrm{mtT} = 11111$). Note that in cases where computed intrinsic contribution is greater than 1.0, (there are more mutations acquired due to intrinsic rate than that due to estimated total rate), the intrinsic contribution was set to 1.0 (0.0 in log10 scale). 111

Figure 31: Comparison of Extended Risk Model, Original Stem Cell Model and Intermediate Model on computed lifetime intrinsic risk (log10 scale) computed from Extended Risk Model and statistics of NPCR observed risk in U.S. Tissue id/names are given below horizontal axis and the tissues are sorted in ascending order of "risk_observed", the average risk in U.S. Intrinsic mutation rate is selected to be {1e-08}; mutation effects and clonal expansion factors are selected from $\mathrm{emr} = \{1.0\}$ and $\mathrm{fd} = \{1.0\}$; the default required mutation hits are ($\mathrm{mtS} = 00111, \mathrm{mtP} = 01111$ and $\mathrm{mtT} = 11111$). 122

Figure 32: Comparison of Extended Risk Model, Original Stem Cell Model and Intermediate Model on computed lifetime intrinsic risk (log10 scale) computed from

Extended Risk Model and statistics of NPCR observed risk in U.S. Tissue id/names are given below horizontal axis and the tissues are sorted in ascending order of "risk_observed", the average risk in U.S. Intrinsic mutation rate is selected to be {1.1e-08}; mutation effects and clonal expansion factors are selected from $\mathrm{emr} = \{1.0\}$ and $\mathrm{fd} = \{1.0\}$; the default required mutation hits are ($\mathrm{mtS} = 00111, \mathrm{mtP} = 01111$ and $\mathrm{mtT} = 11111$). 123

Figure 33: Comparison of Extended Risk Model, Original Stem Cell Model and Intermediate Model on computed lifetime intrinsic risk (log10 scale) computed from Extended Risk Model and statistics of NPCR observed risk in U.S. Tissue id/names are given below horizontal axis and the tissues are sorted in ascending order of "risk_observed", the average risk in U.S. Intrinsic mutation rate is selected to be {2.5e-08}; mutation effects and clonal expansion factors are selected from $\mathrm{emr} = \{1.0\}$ and $\mathrm{fd} = \{1.0\}$; the default required mutation hits are ($\mathrm{mtS} = 00111, \mathrm{mtP} = 01111$ and $\mathrm{mtT} = 11111$). 124

Figure 34: Comparison of Extended Risk Model, Original Stem Cell Model and Intermediate Model on computed lifetime intrinsic risk (log10 scale) computed from Extended Risk Model and statistics of NPCR observed risk in U.S. Tissue id/names are given below horizontal axis and the tissues are sorted in ascending order of "risk_observed", the average risk in U.S. Intrinsic mutation rate is selected to be {1e-07}; mutation effects and clonal expansion factors are selected from $\mathrm{emr} = \{1.0\}$ and $\mathrm{fd} = \{1.0\}$; the default required mutation hits are ($\mathrm{mtS} = 00111, \mathrm{mtP} = 01111$ and $\mathrm{mtT} = 11111$). 125

Figure 35: Comparison of Extended Risk Model, Original Stem Cell Model and Intermediate Model on computed lifetime intrinsic risk (log10 scale) computed from Extended Risk Model and statistics of NPCR observed risk in U.S. Tissue id/names are given below horizontal axis and the tissues are sorted in ascending order of "risk_observed", the average risk in U.S. Intrinsic mutation rate is selected to be {1e-08}; mutation effects and clonal expansion factors are selected from $\mathrm{emr} = \{2.0\}$ and $\mathrm{fd} = \{2.0\}$; the default required mutation hits are ($\mathrm{mtS} = 00111, \mathrm{mtP} = 01111$ and $\mathrm{mtT} = 11111$). 126

Figure 36: Comparison of Extended Risk Model, Original Stem Cell Model and Intermediate Model on computed lifetime intrinsic risk (log10 scale) computed from Extended Risk Model and statistics of NPCR observed risk in U.S. Tissue id/names are given below horizontal axis and the tissues are sorted in ascending order of "risk_observed", the average risk in U.S. Intrinsic mutation rate is selected to be {1.1e-08}; mutation effects and clonal expansion factors are selected from $\mathrm{emr} = \{2.0\}$ and $\mathrm{fd} = \{2.0\}$; the default required mutation hits are ($\mathrm{mtS} = 00111, \mathrm{mtP} = 01111$ and $\mathrm{mtT} = 11111$). 127

Figure 37: Comparison of Extended Risk Model, Original Stem Cell Model and Intermediate Model on computed lifetime intrinsic risk (log10 scale) computed from Extended Risk Model and statistics of NPCR observed risk in U.S. Tissue id/names are given below horizontal axis and the tissues are sorted in ascending order of "risk_observed", the average risk in U.S. Intrinsic mutation rate is selected to be {2.5e-08}; mutation effects and clonal expansion factors are selected from $\mathrm{emr} =$

{2.0} and fd = {2.0}; the default required mutation hits are (mtS = 00111, mtP = 01111 and mtT = 11111). 128

Figure 38: Comparison of Extended Risk Model, Original Stem Cell Model and Intermediate Model on computed lifetime intrinsic risk (log10 scale) computed from Extended Risk Model and statistics of NPCR observed risk in U.S. Tissue id/names are given below horizontal axis and the tissues are sorted in ascending order of "risk_observed", the average risk in U.S. Intrinsic mutation rate is selected to be {1e-07}; mutation effects and clonal expansion factors are selected from emr = {2.0} and fd = {2.0}; the default required mutation hits are (mtS = 00111, mtP = 01111 and mtT = 11111). 129

List of Tables

Table 1 [31]. Thirty-one tissue/cancer types with their lifetime cancer risk and relevant parameters, from supplementary material for [31]. where id denotes the cancer type id number, following the same order as in Table S1 of the supplementary material for [31]. Here Ntotal is the total number of normal cells in the tissue of origin, Nstem is the number of stem cells in the tissue of origin, ntotal is the number of divisions of each stem cell per lifetime, lscd is the cumulative number of divisions of all stem cells per lifetime, and, cr is the observed cancer risk. 34

Table 2: Cancer risk observed in [31], computed using our intermediate model, and computed using our original stem-cell-only model [34]. Here crobserved is the observed cancer risk; cru is the upper bound cancer risk computed through our model and crstem is the cancer risk from the stem cell-only model in [34]. In addition, nupperp is the progenitor lineage length computed according to arguments above, Npl is the lower bound number of progenitor and terminal cells based on nupperp, nas and nss. $Ns = 2nss$. The total cell number from the model is $N = Npl + Ns$. Tissue types with significantly different cru and crstem are highlighted. Note that the id number here follows the same order with the tissue types in supplementary material of [31]. 36

Table 3.1: Basic configurations used in our analysis for 31 cancer types (from Table S1 in [31]). We list the cancer types in the same order as in [31], where Robs denotes observed lifetime risk, NH is the total homeostatic number of cells, NSH is the total homeostatic stem cell number, and nas represents number of asymmetric stem cell

divisions which equals d (Number of divisions of each stem cell per lifetime) in [31]. 76
Table 4: SEER (1973 - 2012) data for Hepatocellular carcinoma cancer risk for each age group. Rates are per 100,000 and age-adjusted to the 2000 US Standard Population (19 age groups - Census P25-1130) standard. 79
Table 5: Estimated excess mutation rate, total risk and intrinsic risk for selected cancer types with two settings of intrinsic mutation rate $\mathrm{uint}(1) = 10\text{-}8$ and $\mathrm{uint}(2) = 10\text{-}7$. Robs is the observed risk, $\mathrm{uexc}(1), \mathrm{uexc}(2)$ represent estimated excess rate with intrinsic rate settings $\mathrm{uint}(1)$ and $\mathrm{uint}(2)$ respectively. Similarly, Rtotal and Rint represent estimated total risk (from mutation rate $\mathrm{uint} + \mathrm{uexc}$), and intrinsic risk (from mutation rate uint). 86
Table 6: Estimated intrinsic contribution percentage for selected cancer types with two settings of intrinsic mutation rate $\mathrm{uint}(1) = 10\text{-}8$ and $\mathrm{uint}(2) = 10\text{-}7$. "Cint all cells" gives the estimated intrinsic contribution based on the ratio of number of acquired mutations among *all cells* due to uint vs. $(\mathrm{uint} + \mathrm{uext})$, while "Cint cancer cells" gives the same quantity considering only cancer cells, i.e. those cells acquired sufficient types of driver mutations for cancer onset. 87
Table 7: Estimated excess mutation rate, total risk and intrinsic risk for selected cancer types with two settings of mutation rate enlargement (as an effect of mutation M4) factor $\alpha(1) = 1$ and $\alpha(2) = 2$. We use intrinsic mutation rate 10-8. 89
Table 8: Estimated intrinsic contribution percentage for selected cancer types with two settings of mutation rate enlargement (as an effect of mutation M4) factor $\alpha(1) = 1$ and $\alpha(2) = 2$. We use intrinsic mutation rate 10-8. 90

Table 9: Estimated excess mutation rate, total risk and intrinsic risk for selected cancer types with two settings of mutation rate enlargement (as an effect of mutation M4) factor $\alpha(1) = 1$ and $\alpha(2) = 2$. We use intrinsic mutation rate 10-7..... 92

Table 10: Estimated intrinsic contribution percentage for selected cancer types with two settings of mutation rate enlargement (as an effect of mutation M4) factor $\alpha(1) = 1$ and $\alpha(2) = 2$. We use intrinsic mutation rate 10-7. 93

Table 11: Estimated excess mutation rate, total risk and intrinsic risk for selected cancer types with two settings of clonal expansion (as an effect of mutation M5) factor $\gamma(1) = 1$ and $\gamma(2) = 2$..... 95

Table 12: Estimated intrinsic contribution percentage for selected cancer types with two settings of mutation rate enlargement (as an effect of mutation M4) factor $\alpha(1) = 1$ and $\alpha(2) = 2$. 96

Table 13: Estimated excess mutation rate, total risk and intrinsic risk for selected cancer types with two settings of clonal expansion (as an effect of mutation M5) factor $\gamma(1) = 1$ and $\gamma(2) = 2$..... 97

Table 14: Estimated intrinsic contribution percentage for selected cancer types with two settings of mutation rate enlargement (as an effect of mutation M4) factor $\alpha(1) = 1$ and $\alpha(2) = 2$. 98

Table 15: Estimated excess mutation rate, total risk and intrinsic risk for selected cancer types with two settings of clonal expansion (as an effect of mutation M5) factor $\gamma(1) = 1$ and $\gamma(2) = 2$..... 99

Table 16: Estimated intrinsic contribution percentage for selected cancer types with two settings of mutation rate enlargement (as an effect of mutation M4) factor $\alpha(1) = 1$ and $\alpha(2) = 2$. 100

Table 17: Estimated excess mutation rate, total risk and intrinsic risk for selected cancer types with two settings of clonal expansion (as an effect of mutation M5) factor $\gamma(1) = 1$ and $\gamma(2) = 2$..... 101

Table 18: Estimated intrinsic contribution percentage for selected cancer types with two settings of mutation rate enlargement (as an effect of mutation M4) factor $\alpha(1) = 1$ and $\alpha(2) = 2$. 102

Table 19: Age dependent observed risk, intrinsic risk, and estimated intrinsic contribution at each of 5 years within an 80-year lifespan. Left side figures plot the average observed risk ("CR_MEAN_USA") and intrinsic risk under selected parameter settings; right side figures plot the estimated intrinsic contribution for the same tissue. The intrinsic contribution in above figures for age x, represents cumulative average intrinsic contribution percentage from age 0 to age x. All figures are in log10 scale. Intrinsic mutation rate is selected to be {1e-08, 1.1e-08, 2.5e-08, 1e-07}. No mutation effects and clonal expansion were applied here ($\mathrm{emr} = 1.0$ and $\mathrm{fd} = 1.0$); also default required mutation hits were used ($\mathrm{mtS} = 00111, \mathrm{mtP} = 01111$ and $\mathrm{mtT} = 11111$). Note that in rare cases where computed intrinsic contribution is greater than 1.0, (there are more mutations acquired due to intrinsic rate than that due to estimated total rate), the intrinsic contribution was set to 1.0 (0.0 in log10 scale)..... 112

Table 20: Age dependent observed risk, intrinsic risk, and estimated intrinsic contribution at each of 5 years within an 80-year lifespan. Left side figures plot the average observed risk ("CR_MEAN_USA") and intrinsic risk under selected parameter settings; right side figures plot the estimated intrinsic contribution for the same tissue. The intrinsic contribution in above figures for age x, represents cumulative average intrinsic contribution percentage from age 0 to age x. All figures are in log10 scale. Intrinsic mutation rate is selected to be {1e-08, 1.1e-08, 2.5e-08, 1e-07}. Mutation effects and clonal expansion were applied here with factors $\mathrm{emr} = 2.0$ and $\mathrm{fd} = 2.0$; also default required mutation hits were used ($\mathrm{mtS} = 00111, \mathrm{mtP} = 01111$ and $\mathrm{mtT} = 11111$). Note that in rare cases where computed intrinsic contribution is greater than 1.0, (there are more mutations acquired due to intrinsic rate than that due to estimated total rate), the intrinsic contribution was set to 1.0 (0.0 in log10 scale). 116

Acknowledgments

First and foremost, I would like to thank my advisor Professor Song Wu and my co-advisor Professor Wei Zhu; for your tremendous guidance during my Ph.D. study. I would like to thank you for encouraging my research. Your advice on both research as well as on my career development have been invaluable.

I would also like to thank my committee members, Professor Pei Fen Kuan and Professor Yusuf Hannun, for serving as my committee members and for providing your time and insightful comments on this work. I also want to thank you for letting my defense be an enjoyable moment, and for your sincere encouragement.

I also want to extend my thanks to all my friends for your suggestions and help along my doctoral study.

I want to express my special thanks to my parents for their constant support. I am most grateful for all of the sacrifices that you have made for me.

Chapter 1 Introduction

The risk of cancer is the probability that a tissue will develop cancer over a certain period in life, often the lifetime period. Much effort has been devoted to dissecting the factors behind cancer risk [31, 35]; however, conclusions regarding cancer risk factors may vary drastically from study to study [31, 34, 35]. Therefore, it is critical to model the inherent mechanics of cancer development for a more comprehensive and reliable analysis [34].

It is commonly accepted that cancer results from a series of somatic mutations [8] accumulated during cell divisions. Among different mutations, some can significantly influence the key features of a cell such as division pattern, fitness level, death rate and mutation rate [11, 14]. These mutations are known as driver mutations [5] and some of which can be the major causes for cancer. Once a tissue cell acquires sufficient number and types of driver mutations, it becomes a cancer cell.

Some mathematical models have been developed to describe the process of cancer initiation and progression. These models usually incorporate cell division dynamics among different types of cells along with mutation acquisition. Assumptions and types of models vary in literature. The complexity of each model depends mainly on the assumptions of the types of cells considered, the cell division structure and dynamics, and the effects of mutations.

Some assume that the cancer risk comes mainly from stem cells and therefore rule out other cells such as progenitor and terminal cells. Wu et al. [34] have developed the first discrete time probability propagation model to estimate the lifetime cancer risk due to the intrinsic mutations of cancer driver genes associated with cell divisions. Bozic et al. [5]

proposed a discrete time branching process model that considers the effect of mutations on cell fitness and provided the closed formula for the waiting time of mutation acquisition. They also assumed a single type of cells, that either divides symmetrically or dies. Frank et al. [8] incorporated progenitor cells and derived a simulation model for cancer risk. They also discussed the influence of different cell division structures on computed risk. However, they restricted each stem cell to have only asymmetric division that yields one daughter stem cell and one progenitor cell, which is unrealistic in real tissues.

As for continuous time models, Ashkenazi et al. [1] and, Gentry and Jackson [11] developed a dynamic system using differential equations that considered stem, progenitor and terminal cells. They allow the most general forms of cell division and impose the effect of different mutation pathways on system parameters. In addition, their model can simulate both the cell growth and the homeostasis state, and thus better capture the overall characteristics of tissue dynamics. However, their model basically focused on the quantities of cell numbers with each mutation acquisition state, instead of computing lifetime cancer risk.

Figure 1 below is an illustration of stem, progenitor and terminal cells and their relationships [23].

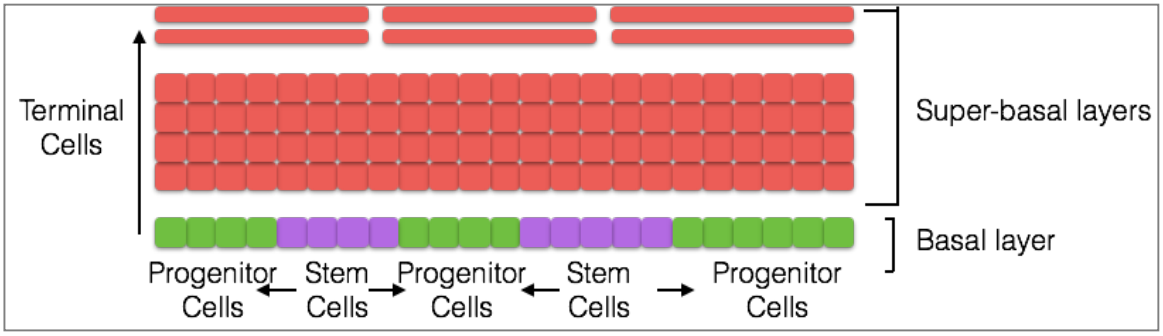


Figure 1 [23]: An illustration of stem cells, progenitor cells and terminal cells. The proposed locations of stem cells (purple), progenitor cells (green) and super-basal, terminally differentiating cells (pink) are shown in human interfollicular epidermis. Arrows represent the relationship between each cell compartment and the movement of cells to the surface of the skin as they undergo terminal differentiation.

There are other efforts that push the model complexity and flexibility into the next level by including spatial patterns and fluid dynamics into cell dynamics. For example, Hannezo et al. [12] considered the spatial variation of concentrations of each cell type in addition to non-spatial homeostasis regulation. Their focus is on pure cell dynamics without mutation acquisition. Waclaw et al. [33] built a 3D spatial tumor evolution model. Their model provided various aspects of tumor initiation and growth, as well as an analysis on the migratory activities. However, they focused more on the tumor progression than the initiation mechanics and cancer risk computation.

In our work, we built comprehensive discrete-time stochastic models for cell dynamics, mutation accumulation and cancer risk. We will start with the *original stem cell model*, which is identical to the Markov model by Wu et al. [34]; then we extend this model into our *intermediate risk model*, by considering not only stem cells, but also progenitor cell lineages and terminal cells. The intermediate risk model also contains algorithms for building homeostasis conditions with stem-progenitor-terminal cell division structures. Eventually, we developed a more realistic *extended risk model by incorporating mechanisms of tumour heterogeneity*. This extended model integrates conditional

dependency relations within cell division structures. It differentiates driver mutations in the mutation acquisition procedure, instead of simply using the number of acquired mutations as in previous models. This allows us to model different effects incurred by different types of mutations. In addition, we built algorithms for time dependent cell dynamics, homeostasis, clonal expansion as well as regulations. Our extended risk model allows us to obtain cell numbers of different cell types, mutation acquisition states, as well as estimated cancer risk at any time within a normal lifespan of 80 years.

During our work, we also developed a discrete simulation package that is able to model the activities of each type of cells. This simulation framework, based on the core algorithms in our theoretical models, features equal model complexity but superior computational efficiency in comparison to continuous time simulations [5]. Given that the mutation acquisition is an extremely rare event (mutation rate is normally around 10^{-8}), we usually need more than 10^{10} simulation runs to have a nonzero cancer risk estimate. Plus, the conjuncture epidemiological cancer population data usually come in the form of discrete age distribution in 1- or 5-year intervals. Thus it is both more efficient and more realistic to adopt the discrete-time models like our extended stochastic model.

In this work, we studied how mechanisms of tumor heterogeneity could impact the theoretical cancer risk. In addition, we compared for several applicable tissues the observed age dependent risk from SEER database with our theoretical risk for each age in years. Moreover, we use the ratio of the number of mutations acquired as a metric to estimate the percentage of contributions for cancer onset, from the intrinsic risk factors.

These analyses from different angles converge to a unanimous conclusion that non-intrinsic factors are the major causes for cancer onset.

The rest of this thesis is organized as follows. In Chapter 2 we introduce modern biological theories for cell dynamics, mutation acquisition and cancer onset. We review major mathematical carcinogenesis models in literature, including our *original stem cell model*. In Chapter 3 we describe our *intermediate cancer risk model* that could bridge the original model and the more complicated extended model. We devote Chapter 4 to the development of *the extended cancer risk model.* In Chapter 5, we provide our analysis results and demonstrate our conclusions regarding intrinsic risk and mutation contributions, age dependent risk, and the impact of mutation effects.

Chapter 2 Cell Dynamics and Mutation Acquisition

2.1 Cell and Cell Divisions

Human tissues and organs are composed of a heterogeneous mix of cells [11]. Stem cells refer to those cells that can self-renew and generate other types of cells of the organ. A stem cell division follows three basic division patterns: symmetrical self-renewal to form two daughter stem cells, asymmetrical self-renewal to generate one stem cell and one progenitor cell, or symmetrical differentiation to yield two progenitor cells [1,11,17,19,27].

A progenitor cell usually divides a limited number of times and eventually produces terminal cells which never divide and will eventually die [1, 8]. Figure 2 is an illustration of the most general division patterns of stem and progenitor cells in the literature.

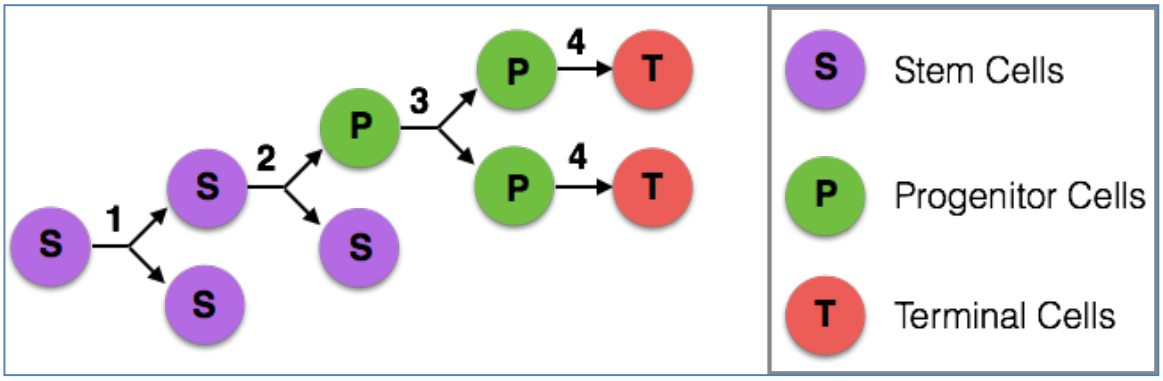


Figure 2 [38]: Illustration of stem cells, progenitor cells and terminal cells with their general division patterns: 1 – symmetric stem cell division, 2 – asymmetric stem cell division, 3 – progenitor division, and 4 – terminal differentiation.

Now, to model cell dynamics and its division behavior, the following parameters may be considered:

- $r_{die}^{<type>}$ is the death rate (average number of deaths per cell per day) of cell type <type>, which can be one of SC (Stem Cell), PC (Progenitor Cell) or TC (Terminal Cell).

- $r_{div}^{<type>}$ is the division rate (average number of divisions per cell per day). We must have $r_{die}^{<type>} + r_{div}^{<type>} \leq 1$ because at any moment each cell must has one of these three behaviors: death, division, or staying still.

Upon each division, we have the probabilities of each division pattern: $p_{sym}^{<type>}$, $p_{asym}^{<type>}$ and $p_{diff}^{<type>}$ representing probabilities of symmetric division, asymmetric division and symmetric differentiation, respectively, given that a division occurs. Here $p_{sym}^{<type>} + p_{asym}^{<type>} + p_{diff}^{<type>} = 1$.

2.2 Mutation Acquisition

The main stream theory is that cancer stems from a sequential accumulation of somatic mutations within tissue cells [1, 8]. A somatic mutation is genetic alteration acquired by a cell that can be passed to the progeny of the mutated cell in the course of cell division [37]. During each cell division, each daughter cell will inherit the mutation that was already obtained, if any, by its parent cell, and also has a probability to gain extra mutation(s). The probability that a daughter cell will obtain one extra mutation during a division is termed the mutation rate: $u^{<typeP><typeD>}$. Here $< typeP >$ and $< typeD >$ are parent cell type and daughter cell type respectively for the given division.

Although tumor cells often exhibit many mutations, only a relatively small subset is crucial for cancer development [1, 4, 15, 17, 19, 27]. Some literature classified mutations into "driver mutations", which have influences on a cell's fitness, and "passenger mutations", which do not affect a cell's growth behavior [5, 13, 20, 25]. Driver mutations dominate cancer initiation and progression. Quantitatively, acquired driver mutations can alter a cell's division rate, death rate, and the probability to acquire subsequent mutations.

For example, in [5] and [33], the authors assumed that each driver mutation reduces the death rate (or stagnate rate in [5] by their assumptions) of a cell at a rate of $\sim(1-s)$. A cell with k driver mutations will have a death/stagnation probability of $d = C(1-s)^k$ where s is the selective advantage provided by a driver mutation [5] and C is a constant depending on the assumption context. On the other hand, some literature [1, 11] listed the possible parameter changes due to mutations in advance and did not consider multiple mutations of the same type. Once the effect of mutations is quantitatively formularized, one can analyze more general mutation effects. For example, Ashkenazi et al. [1] and Gentry and Jackson [11] compared the effect of different orders of mutation acquisition sequences on the length of time until cancer initiation.

2.3 Cancer Risk

The lifetime risk of developing cancer refers to the chance a person has, over the course of his or her lifetime (from birth to death), of being diagnosed with cancer [36]. For simplicity, we can treat the cancer risk as the probability that at least one of the total cells will be a cancer cell at the end of a given time period. A cell becomes a cancer cell if it acquires a sufficient number and types of mutations [1, 8, 11, 34]. A progenitor cell usually needs more mutations to cause cancer than a stem cell [8] and has a much shorter lineage; hence some literature only considers stem cells in analyzing cancer risks [31, 34].

2.4 Cell Dynamics

To model cancer risk, a reasonable modeling of cell dynamics is required. Aside from different cell division patterns, cell dynamics modeling should also include the mechanism of how cells in a tissue get updated (turn over) and maintain homeostasis.

Specifically, the cells in a tissue maintain a dynamic balance such that the number of each type of cells remains approximately constant while old cells die and new cells being generated. There are basically two approaches in modeling homeostasis in literature. One is to specify a deterministic mechanism of cell division, growth and death. For example, Frank et al. [8] provides a stem-progenitor cell division pattern such that the total number of cells is fixed and their model has inspired our development of the homeostasis condition. However, since their focuses were on the analysis of cancer risk and the lengths of cell division lineages, they did not explicitly develop the cell number dynamics over time. Wu et al. [34] gave a very clear mechanism such that once the number of cells grows to the homeostasis stage, it will remain constant, and only stem cells were included in their model. The other approach to model homeostasis is to regulate parameters such as death rate and division rate in ODEs, so that the number of cells remains constant in time and the steady-state solution matches the homeostasis condition [1, 11].

2.5 Representative models

We now describe some representative models from the literature, which could provide valuable insight for our model development, current or future. First, we describe the stem cell mutation acquisition model proposed by Wu et al. [34], based on which we will develop and extend our own models. Second, the deterministic stem-progenitor cell model of Frank et al. [8] will be described in detail since it inspired us on the modeling of progenitor/progenitor cell dynamics. Then, we will give a brief introduction to the discrete time stochastic model by Bozic et al. [4] and the continuous time (ODE) model in [1, 11], which could serve as a reference on building more complicated simulation framework in the future. Finally, a more comprehensive model published recently [33],

which incorporated spatial dynamics, and focuses more on cancer progression and migration, will be briefly discussed also.

2.5.1 Original Stem Cell Model

The probability model proposed by Wu et al. [34] was built upon the assumption that stem cell dominates cancer initiation because of its self-renewal property. Cell dynamics in this model has two stages. On the first stage, symmetric division, the stem cell population, originated from a single stem cell, grows exponentially in that each stem cell will give birth to two daughter stem cells in each division. Once the cell number increases to approximately the homeostasis number, the dynamics switch to the second stage, asymmetric division, where each stem cell only generates one daughter stem cell in each division, thus the stem cell population remains constant. Figure 3 below illustrates the stem cell division and dynamics in [34] with n_s^s and n_a^s denoting the numbers of symmetric and asymmetric divisions of a stem cell, respectively.

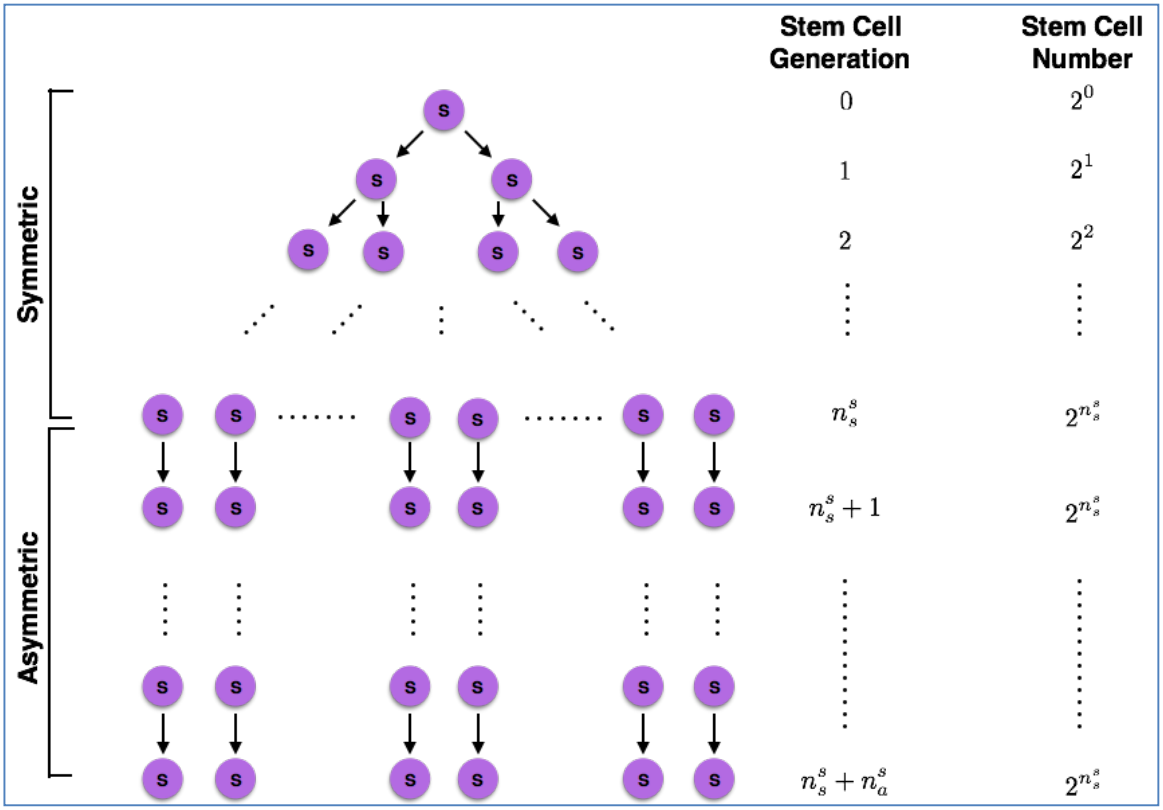


Figure 3 [34]: Stem cell division pattern and dynamics. Each original (Generation 0) stem cell first goes through n_s^s symmetric divisions and then n_a^s asymmetric divisions.

The stem cell number, $N(g) = 2^g$ if $g \leq \boldsymbol{n_s^s}$ and $N(g) = 2^{\boldsymbol{n_s^s}}$ if $g > \boldsymbol{n_s^s}$.

For mutation acquisition, this model assumes binomial distribution in additional driver mutations acquired by each daughter cell during each division. Specifically, suppose m_s driver mutations are required for cancer onset, and the current cell state (the number of mutations carried) is i, then the transition probability to its daughter cell state is $p_{ij} = \binom{m_s - i}{j - i} u^{j-i}(1-u)^{m_s - j}\mathbb{I}_{\{0 \leq i \leq j \leq m_s\}}$ where u is the mutation rate.

The cell state propagation along cell generations follows a Markov process with the above transition rules. Figure 4 below shows a diagram for the cell state propagation process in mutation acquisition.

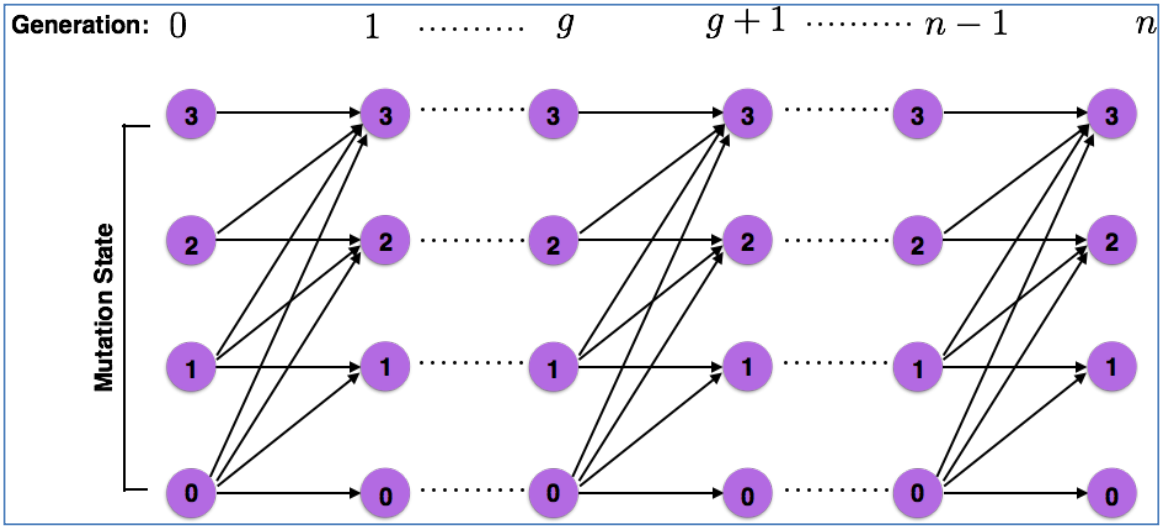


Figure 4 [34]: The propagation diagram of driver mutation states in one cell from one generation to the next.

Given the initial state, for example, $P[X_0 = 0] = 1$ and $P[X_i = 0] = 1$ for $i = 1 \ldots m_S$, where X_i represents the number of mutations acquired on a cell at Generation i; one can compute the cell state distribution at each generation. In addition, the mutation acquisition process is built upon individual cells, and the model assumes conditional independence according to the dependence diagram following Figure 3. Therefore, each cell at the same generation will have the same cell state distribution. Eventually, the cancer risk due to intrinsic driver gene mutations, referred to as the theoretical lifetime intrinsic risk (tLIR) [34] will be

$$tLIR = 1 - [1 - P[X_n = m_S]]^S$$

Here S is the total number of stem cells in the final generation. A key reason for this simple formula is that each cell state (number of driver mutations carried) is not lower than any of its ancestors, based on the probability transition rule. So, if none of the cells at the final generation is a cancer cell, then none of the cells throughout the entire

lifespan is cancerous. Cancer results if at least one cell acquired the sufficient number of driver gene mutations, in the last generation.

Here we can see that $P[X_n = m]$ plays a crucial rule in cancer risk computation. In fact, we have developed a closed form formula so that $P_n(m) \overset{\text{def}}{\equiv} P[X_n = m]$ can be directly computed from the initial state without going through transitions:

$$P_n(m) = \sum_{i=0}^{m} [1 - (1-u)^n]^{m-i} P_0(i)$$

Proof:

We prove a more general version $P_n(m) = \sum_{i=0}^{m} [1 - (1-u)^l]^{m-i} P_{n-l}(i)$ by induction. For $l = 1$, this holds because $P_n(m) = \sum_{i=0}^{m} u^{m-i} P_{n-1}(i) = \sum_{i=0}^{m} [1 - (1-u)^1]^{m-i} P_{n-1}(i)$. Now assume for some $l < n$, we have $P_n(m) = \sum_{i=0}^{m} [1 - (1-u)^l]^{m-i} P_{n-l}(i)$, then consider $(l+1)$:

$$P_n(m) = \sum_{i=0}^{m} [1 - (1-u)^l]^{m-i} P_{n-l}(i)$$

$$= \sum_{i=0}^{m} [1 - (1-u)^l]^{m-i} \sum_{j=0}^{i} \binom{m-j}{i-j} u^{i-j} (1-u)^{m-i} P_{n-(l+1)}(j)$$

$$= \sum_{j=0}^{m} \sum_{i=j}^{m} [1 - (1-u)^l]^{m-i} \binom{m-j}{i-j} u^{i-j} (1-u)^{m-i} P_{n-(l+1)}(j)$$

Note that

$$\sum_{i=j}^{m}[1-(1-u)^{l}]^{m-i}\binom{m-j}{i-j}u^{i-j}(1-u)^{m-i}$$

$$=\sum_{t=0}^{m-j}\binom{m-j}{t}[(1-(1-u)^{l})(1-u)]^{m-j-t}u^{t}$$

$$=[(1-(1-u)^{l})(1-u)+u]^{m-j}=[1-(1-u)^{l+1}]^{m-j}$$

So

$$P_n(m)=\sum_{i=0}^{m}[1-(1-u)^{l+1}]^{m-i}\,P_{n-(l+1)}(i)$$

Proof Done.

2.5.2 Stem-Progenitor Cell Model

Another representative model for cancer risk was proposed by Frank et al. [8]. They included both stem and progenitor cells in cell dynamics and developed a recursive formula for total cancer risk. Though they did not explicitly model the dynamic balance in homeostasis, their model provides insight for our model development. In this model, the single, initial stem cell divides to produce a stem cell lineage and a progenitor cell lineage. Each progenitor lineage divides symmetrically n^p times, yielding 2^{n^p} cells, while the stem lineage has n_a^s asymmetric divisions, producing a total of $N = n_a^s 2^{n^p}$ cells [8]. See Figure 5 below for an illustration of the stem-progenitor cell division pattern in [8].

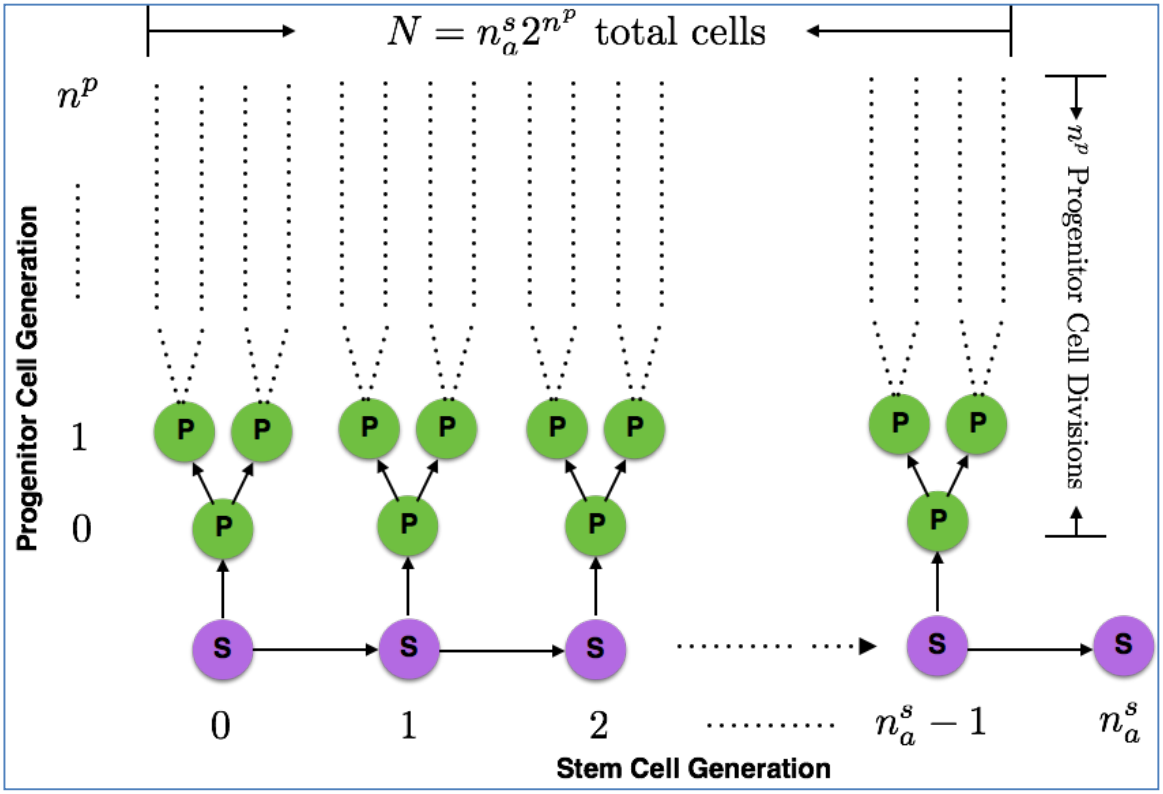


Figure 5 [8]: The pattern of cell division giving rise to a total of *N* cells. The single, initial stem cell divides to produce a stem cell and a progenitor lineage. Each progenitor lineage divides n^p times, yielding 2^{n^p} cells. The stem lineage divides n_a^s times. The total number of cells produced are $N = n_a^s 2^{n^p}$.

We now describe cancer risk computation of this model. In their model, m_s mutations to the stem lineage or m_p mutations to the progenitor lineage cause cancer. They provided cancer risk formulas for $m_s = m_p = 2$ and $m_s = m_p = 3$ without detailed mathematical build up.

Here we first give their formula for $m_s = m_p = 2$ with a brief derivation, and then generalize their model to any $m_s = m_p$.

The probability of $m_p = 2$ mutations to a cell in a progenitor lineage with n^p cell divisions, if the initial (stem) cell has no mutations, is given as in [8]:

$$T_2(n^p) \approx \sum_{i=1}^{n^p} 2^i (u_p^2 + 2u_p T_1(n^p - i))$$

Here u_p is the mutation rate of progenitor cells and $T_1(n^p - i)$ is the probability that at least one additional mutation will occur in the descendant progenitor lineage, given that the current cell at the ith generation suffers one mutation. We also have

$$T_1(n^p - i) = 1 - e^{-u_p d}$$

where $d = 2(2^{n^p - i} - 1)$ is the number of branches in the descendant cell lineage along which mutations can occur.

The total risk of two mutations accumulating along the entire stem-progenitor lineage is:

$$R_2(n_a^s, n^p) \approx \sum_{i=1}^{n_a^s} e^{-2u_s(i-1)} [2u_s R_1(n_a^s - i + 1, n^p) + (1 - 2u_s)T_2(n^p)]$$

where u_s is the progenitor cell mutation rate and

$$R_1(n_a^s - i + 1, n^p) = 1 - e^{-(n_a^s - i + 1)(u_s + u_p(2^{n^p + 1} - 1))}$$

is the risk that at least one additional mutation will occur in the descendant stem-progenitor branches including the current (stem) cell, given that the current cell suffers from one mutation already.

The formulas were basically built upon the underlying probability assumptions with Binomial and Poisson distributions. For a single stem or progenitor cell, given there are no mutations accumulated so far, the number of mutations acquired will follow a binomial distribution.

The risk of k hits, provided m hits will cause cancer, is $p_k = \binom{m}{k} u^k (1 - u)^{m-k}$ where u is the mutation rate. An approximate version is $p_k \approx \binom{m}{k} u^k$ for small u. Specifically, the risk of obtaining one mutation is $p_1 \approx mu$ and the probability that no mutation occurs is $p_0 = (1 - u)^m \approx 1 - mu$.

Also, for event A with small probability $p \ll 1$, the risk that A happens among N independent instances will be $1-(1-p)^N \approx Np$.

In addition, if a progenitor cell already suffers from $(m_p - 1)$ mutations, the number of extra mutations acquired by its descendant lineage will follow approximately $Poisson(u_p)$. If a stem cell already has $(m_s - 1)$ mutations, the number of extra mutations, in the descendant branches, including current cell and all descendant stem cells, will follow $Poisson(u_s)$ and all descendant progenitor branches will follow $Poisson(u_p)$ independently. Finally, the model also assumes Poisson distribution on mutation numbers of stem cells if the initial stem cell state is 0 with mean $p_1 \approx m_s u_s$.

Based on these assumptions, we now see that

$$T_1(n^p - i) = 1 - e^{-u_p d}$$

and

$$R_1(n_a^s - i + 1, n^p) = 1 - e^{-u_s(n_a^s - i + 1)} e^{-u_p(n_a^s - i + 1)(2^{n^p+1} - 1)}$$

where it is obvious that $d = \sum_{l=1}^{n^p - i} 2^l = 2\left(2^{n^p - i} - 1\right)$ is the corresponding descendant progenitor cell number and $(n_a^s - i + 1)\left(2^{n^p+1} - 1\right)$ is the corresponding descendant stem-progenitor cell number. Also, the probability that at stem cell division step i a first mutation in the stem lineage has not occurred is $e^{-2u_s(i-1)}$. So far we have replicated the derivation behind the formulas for $m_s = m_p = 2$.

Now for a general $m_s = m_p = m$, according to a reasonable extension based on the above assumptions, we have

$$R_m(n_a^s, n^p) \approx \sum_{i=1}^{n_a^s} e^{-mu_s(i-1)} \left[mu_s R_{m-1}(n_a^s - i + 1, n^p) + (1 - mu_s) T_m(n^p)\right]$$

and

$$T_m(n^p) \approx \sum_{i=1}^{n^p} 2^i [u_p^m + \sum_{k=1}^{m} \binom{m}{k} u_p^k T_{m-k}(n^p - i)]$$

2.5.3 Discrete Time Stochastic Model

In this section, we will briefly describe the discrete time stochastic model proposed by Bozic et al. [5]. They model tumor progression and mutation acquisition with a discrete time branching process [3, 5]. Basically, at each time step a cell with j mutations (j-cell) either divides into two cells with probability b_j or dies with probability d_j and $b_j + d_j = 1$. In addition, at every division, one of the daughter cells can acquire an additional mutation with probability u and they ignore the probability that both daughter cells will acquire extra mutations. Also, at most one extra mutation can be acquired upon each division. Therefore, the offspring of a j-cell could be (1) none, with probability d_j as the cell dies; (2) two j-cells with probability $b_j(1-u)$; or (3) one j-cell and one $(j+1)$-cell with probability $b_j u$.

On the population basis, let $N_j(t)$ be the number of j-cells at time step t, B_j be the number of j-cells that will give birth to two daughter j-cells, M_j be the number of j-cells that will give birth to one daughter j-cell and one daughter $(j+1)$-cell, and D_j be the number of j-cells that will die. Here B_j, M_j and D_j will follow a multinomial distribution:

$$P\big[(B_j, D_j, M_j) = (n_1, n_2, n_3)\big] = \frac{N_j(t)!}{n_1!\, n_2!\, n_3!} \big[b_j(1-u)\big]^{n_1} \big[d_j\big]^{n_2} \big[b_j u\big]^{n_3}$$

and it is easy to see that $N_j(t+1) = N_j(t) + B_j - D_j + M_{j-1}$

The model provides, explicitly, the effect of driver mutations on cell death rate: a cell with k driver mutations has a death rate $d_k = \frac{1}{2}(1-s)^k$ where s is the selective

advantage provided by a driver mutation [5]. In addition to this simulation framework, they also derived an approximate closed form formulas for the waiting times of 1 and k additional driver mutations, which we shall not include in this thesis.

2.5.4 Continuous Time Model

There are basically two classes of continuous time model. One is a continuous time branching process, proposed by Durrett and Moseley [7], which can be seen as an extension of the discrete branching process model in [5]. Simulations of this model can be very expensive as the time steps for updates can be smaller and smaller as cell population grows larger [5]. Here we will briefly describe another category of continuous time model based on ordinary differential equations [1, 11].

For cell dynamics, the model in [1] assumes all possible modes of stem cell divisions, symmetric and asymmetric [6]. Figure 6 provides an illustration of stem cell division modes and mutation acquisition process used in [1].

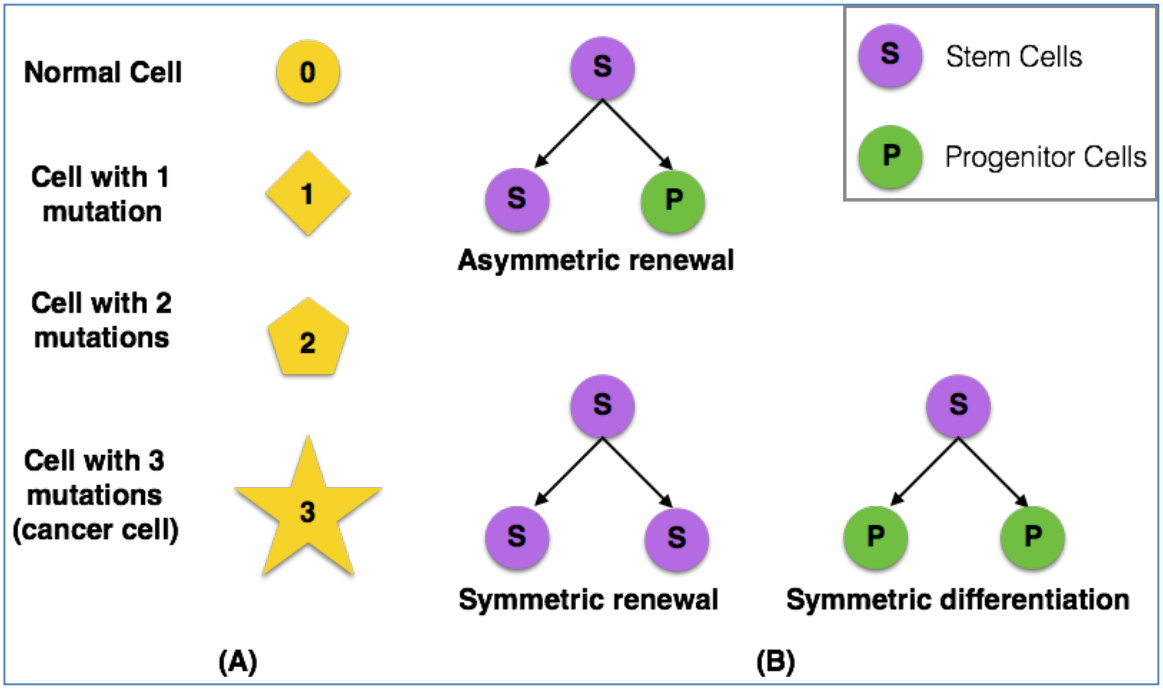


Figure 6 [1]: (A) Mutations: Schematic view of the multistep process of mutation acquisition. (B) Stem cell division modes: symmetric self-renewal division results in two daughter stem cells, asymmetric self-renewal division results in one stem cell and one progenitor cell, and symmetric differentiation division results in two progenitor cells.

At each division, a stem cell can give birth to two daughter stem cells (symmetric renewal), one stem cell and one progenitor cell (asymmetric renewal), or two progenitor cells (symmetric differentiation). A progenitor cell will symmetrically divide into daughter progenitor cells until it becomes fully terminal (mature) cells and lose the ability to proliferate.

As for mutation acquisition, at each division, daughter cells have a probability to acquire one additional mutation. Any cell with 3 mutations will be a cancer cell. Their model also incorporates the mutation effect on cell behavior by varying model parameters. For example, they study three types of driver mutations, R-mutation, which could result in an increase of the cell replication rate or a shift in the balance of stem cell division modes; D-mutation, which could decrease the cell death rate; and G-mutation, which could

increase the cellular mutation rate [1, 14]. In addition, they also study impact of the order of the onset of three types of mutations on the speed of cancer initiation and progression. The development of ODEs follows from these very natural relations:

(1) *Rate of change in stem cell number of a certain mutation state = Increase due to symmetric renewal or mutation – decrease due to symmetric differentiation, death, or mutation.*

Since the division pattern of progenitor cells depends on generation, the progenitor cell state has both mutation state and generation number. For 0th generation progenitor cells, which are generated directly from stem cells:

(2) *Rate of change in* 0*th generation progenitor cell number of a certain mutation state = Increase due to stem cell asymmetric renewal or differentiation, or mutation – Decrease due to division, death or mutation*

Let n^p be the last generation of progenitor cells and $(n^p + 1)$ be the generation of mature cells, then for progenitor cells of generation $1 \leq n \leq n^p$:

(3) *Rate of change in nth generation progenitor cell number of a certain mutation state = Increase due to division or mutation of* $(n-1)$ *generation progenitor cells – Decrease due to divisions, death, or mutation*

Finally, we have:

(4) *Rate of change in mature cell number of a certain mutation state = Increase due to division or mutation of* n^p *generation progenitor cells – Decrease due to death.*

For appropriate rates of proliferation and death, the equations built upon (1) - (4) lead to homeostatic dynamics, that is, a steady state of a healthy tissue [1]. Due to space limit,

we will not copy detailed equations in this thesisas all ODEs were clearly presented in the supplemental material of [1].

The ODE model in [11] is very similar to the model in [1] except that the former considers regulation to stem cell division rate by chemical signaling and environmental (niche) constraints by a fixed functional term [11, 18, 21, 22, 26]. On the other hand, the model in [11] is simplified in that it did not include the intermediate progenitor cells; instead, fully terminal mature cells are directly generated from stem cells [11].

2.5.5 Spatial Dynamics Model

On a higher model complexity, spatial correlation and interaction among cells could be considered. Waclaw et al. [33] proposed a model for tumor evolution to address the concern on how genetic alterations expand within the spatially constrained three-dimensional architecture and come to dominate a large, pre-existing lesion [33].

Their model combines spatial growth and accumulation of multiple mutations. Note that their focus is not on cancer initiation, but cancer progression and migration.

The cell dynamics here follows a 3-D spatial model; in which each tumor cell occupy a site of a regular 3D square lattice while empty sites represent normal cells or extracellular matrix. Cell replication occurs stochastically, with rate proportional to the number of empty sites surrounding the replicating cell, and death occurs with constant rate [33]. Once a tumor cell successfully replicates, with some probability the cell will migrate and create a new micro-lesion. Detailed spatial dynamics can be found in Extended Data Figure 1 in [33].

As for mutation acquisition, when a cell replicates, each of the daughter cells receives i new mutations of each type, either driver or resistant, where i follows a Poisson distribution:

$$P(i) = \frac{e^{-\frac{\gamma_x}{2}}(\gamma_x/2)^i}{i!}$$

Here x denotes the type of mutation and γ_x is the average number of genetic alterations of type x in a single replication event. Like the model in [4], driver mutations can increase the net growth rate either by increasing the birth rate or decreasing the death rate by a constant factor of $(1 + s)$. They provided a simulation of 3D tumor progression and studied multiple indicators of tumor growth and migration with various groups of parameters.

Chapter 3 Intermediate Risk Model

We now extend our original stem cell model [34] to incorporate progenitor and terminal cells while keeping the overall stem cell division stages (symmetric division and asymmetric division). During asymmetric stem cell division, one stem cell divides into not only a daughter stem cell, but also an initial progenitor cell, which will grow a limited generations of progenitor lineage.

We provide a recursive/iterative computation for cancer risk. In addition, we also develop homeostatic conditions for the stem-progenitor lineage structures.

3.1 Cell Dynamics and Homeostatic Condition

We explicitly model cell dynamics that fits into homeostatic conditions and incorporates stem cells, progenitor cells, and terminal cells. We prove that in the long run, this model automatically guarantees constant number of normal cells of different types without further conditions on the model parameters.

The initial cell will be one single stem cell with no mutation. The cell dynamics has two stages. In stage one, each stem cell will go through symmetric divisions only and give birth to two daughter stem cells each time until the stem cell number reaches the homeostatic number. During the second stage, each stem cell will go through asymmetric divisions only, and generate one daughter stem cell and one progenitor cell through each division. Each progenitor cell, upon each division, will give birth to two daughter progenitor cells and finally evolve to terminal cells after a limited number of divisions. A terminal cell cannot divide and dies after several time steps.

Let n_s^s denote the number of stem cell symmetric divisions, n_a^s denote the number of stem cell asymmetric divisions, and n^p the number of progenitor cell divisions. The n^pth

generation progenitor cell will further divide into terminal cells. See Figure 7 below for an illustration of cell division patterns.

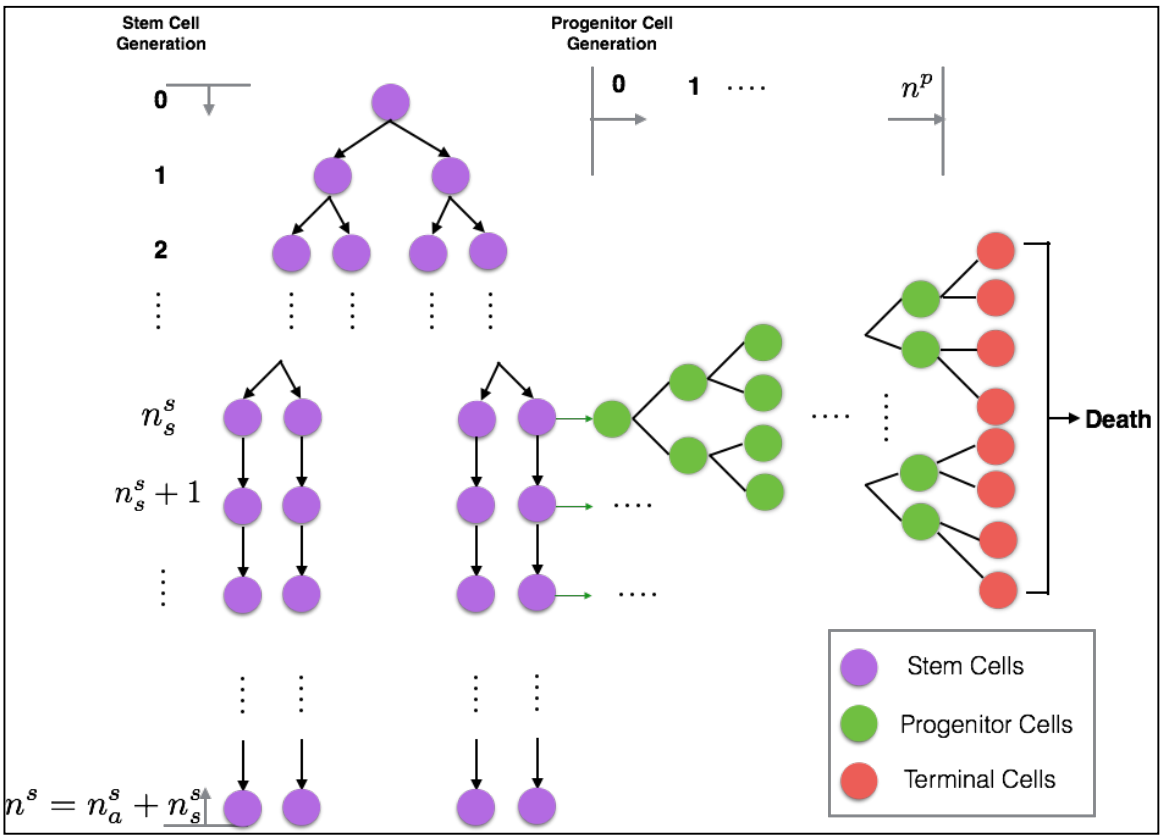


Figure 7: Cell division patterns and dynamics in our intermediate cancer risk model. Initial (Generation 0) stem cell will go through n_s^s symmetric divisions and n_a^s asymmetric divisions. At each stem cell symmetric division, two daughter stem cells are generated. At each stem cell asymmetric division, one stem cell and one progenitor cell are generated. Each progenitor cell generated from stem cell will go through n^p symmetric divisions. Each progenitor cell division will give birth to two daughter progenitor cells. Each n^p th generation progenitor cell will further divide into two progenitor cells that will eventually evolve to terminal cells and die.

Though this is a discrete time model, not all divisions happen simultaneously. Let r_s, r_p and r_t be the division rates of stem and progenitor cells, and the death rate of terminal cells respectively. Let $\Delta t_s = 1/r_s$, $\Delta t_p = 1/r_p$ be the time from the generation to the division completion of stem and progenitor cells, and $\Delta t_t = 1/r_t$ be the time from the birth to the death of a terminal cell. We now derive formulas for cell numbers.

Obviously, the stem cell number at time t, starting from the "birth" of the initial cell, is

$$N_s(t) = 2^{\lfloor t/\Delta t_s \rfloor}\mathbb{I}_{\{0\le t<n_s^s\Delta t_s\}} + 2^{n_s^s}\mathbb{I}_{\{n_s^s\Delta t_s\le t<(n_s^s+n_a^s)\Delta t_s\}}$$

For the progenitor cell, first consider just one lineage. Let τ be the time elapsed from the birth of the initial progenitor cell, the progenitor cell number in the lineage is

$$n_p(\tau) = 2^{\lfloor \tau/\Delta t_p \rfloor}\mathbb{I}_{\{0\le\tau<(n^p+1)\Delta t_p\}}$$

Since there are no progenitor cell when $t < n_s^s\Delta t_s$, Let $t_a = t - n_s^s\Delta t_s$ be the time elapsed from the birth of n_s^s generation stem cell. Now considering $k\Delta t_s \le t_a < (k+1)\Delta t_s$, for $k = 1 \dots n_a^s - 1$, we need to count the number of progenitor cells from all progenitor lineages initiated at time points $t_a = \Delta t_s, 2\Delta t_s \dots k\Delta t_s$. So as $k\Delta t_s \le t_a < (k+1)\Delta t_s$, we have $N_p(t_a) = \sum_{i=1}^{k} n_p(t_a - i\Delta t_s)$. Summing over all possible ranges of t_a, the progenitor cell number at t_a *for a single asymmetric stem cell lineage* is

$$N_p(t_a) = \sum_{k=1}^{n_a^s-1}\sum_{i=1}^{k} n_p(t_a - i\Delta t_s)\mathbb{I}_{\{k\Delta t_s\le t_a<(k+1)\Delta t_s\}} + \sum_{i=1}^{n_a^s} n_p(t_a - i\Delta t_s)\mathbb{I}_{\{t_a\ge n_a^s\Delta t_s\}}$$

For terminal cells, the argument is very similar. With the same definition of $t_a$and τ the number of terminal cells within one progenitor lineage is

$$n_t(\tau) = 2^{(n^p+1)}\mathbb{I}_{\{(n^p+1)\Delta t_p\le\tau<(n^p+1)\Delta t_p+\Delta t_t\}},$$

and the total terminal cell number is

$$N_t(t_a) = \sum_{k=1}^{n_a^s-1}\sum_{i=1}^{k} n_t(t_a - i\Delta t_s)\mathbb{I}_{\{k\Delta t_s\le t_a<(k+1)\Delta t_s\}} + \sum_{i=1}^{n_a^s} n_t(t_a - i\Delta t_s)\mathbb{I}_{\{t_a\ge n_a^s\Delta t_s\}}$$

We now show that for sufficiently large $t_a \ge k\Delta t_s$, N_p and N_t become constants at each discrete time points $k\Delta t_s$ and do not depend on k. Suppose that n_a^s is reasonably large such that $(n_a^s - 2)\Delta t_s \ge (n^p + 1)\Delta t_p$, and consider any particular k and time point $t_a = k\Delta t_s$. Then we have

$$N_p(k\Delta t_s) = \sum_{i=1}^{k} 2^{\lfloor (k-i)\Delta t_s/\Delta t_p \rfloor} \mathbb{I}_{\{0 \le (k-i)\Delta t_s < (n^p+1)\Delta t_p\}}$$

Case 1: for small k such that $(k-1)\Delta t_s < (n^p+1)\Delta t_p$, $N_p(k\Delta t_s)$ is strictly increasing with respect to k because

$$N_p(k\Delta t_s) = \sum_{i=1}^{k} 2^{\lfloor (k-i)\Delta t_s/\Delta t_p \rfloor} > \sum_{i=1}^{k-1} 2^{\lfloor (k-1-i)\Delta t_s/\Delta t_p \rfloor} = N_p((k-1)\Delta t_s)$$

This means that at the early stages of asymmetric stem cell division, progenitor cells will keep growing in size.

Case 2: for any k such that $(k-1)\Delta t_s \ge (n^p+1)\Delta t_p$, there must be some $1 < l \le k$ such that $0 \le (k-k)\Delta t_s < \cdots < (k-l)\Delta t_s < (n^p+1)\Delta t_p \le \big(k-(l-1)\big)\Delta t_s$. Now W.L.O.G, $k-l = \lfloor (n^p+1)\Delta t_p/\Delta t_s \rfloor$. In this case $N_p(k\Delta t_s) = \sum_{i=l}^{k} 2^{\lfloor (k-i)\Delta t_s/\Delta t_p \rfloor}$. But since

$$N_p((k+1)\Delta t_s) = \sum_{i=l+1}^{k+1} 2^{\lfloor (k+1-i)\Delta t_s/\Delta t_p \rfloor} = \sum_{i=l+1}^{k+1} 2^{\lfloor (k-(i-1))\Delta t_s/\Delta t_p \rfloor}$$

$$= \sum_{i=l+1}^{k+1} 2^{\lfloor (k-(i-1))\Delta t_s/\Delta t_p \rfloor} = \sum_{i=l}^{k} 2^{\lfloor (k-i)\Delta t_s/\Delta t_p \rfloor} = N_p(k\Delta t_s)$$

Now $N_p(k\Delta t_s)$ becomes a constant that is independent of k. This marks the homeostatic stage for progenitor cells where old cells become terminal cells and then die while new cells are generated from parent stem or progenitor cells as time goes by. Since $(k-i)\lfloor \Delta t_s/\Delta t_p \rfloor \le \lfloor (k-i)\Delta t_s/\Delta t_p \rfloor \le (k-i)\lceil \Delta t_s/\Delta t_p \rceil$, we can develop an upper bound and a lower bound of the constant by summing over the geometric series:

$$N_p^u(k\Delta t_s) = \frac{2^{\lceil \frac{\Delta t_s}{\Delta t_p} \rceil (\frac{\lfloor (n^p+1) \rfloor \Delta t_p}{\Delta t_s}+1)} - 1}{2^{\lceil \frac{\Delta t_s}{\Delta t_p} \rceil} - 1}$$

$$N_p^l(k\Delta t_s) = \frac{2^{\left\lfloor \frac{\Delta t_s}{\Delta t_p} \right\rfloor \left(\frac{\lfloor (n^p+1)\rfloor \Delta t_p}{\Delta t_s}+1\right)} - 1}{2^{\left\lfloor \frac{\Delta t_s}{\Delta t_p} \right\rfloor} - 1}$$

For terminal cells, following similar argument, considering a particular time point $k\Delta t_s$: $N_t(k\Delta t_s) = \sum_{i=1}^{k} 2^{(n^p+1)} \mathbb{I}_{\{(n^p+1)\Delta t_p \le (k-i)\Delta t_s < (n^p+1)\Delta t_p + \Delta t_t\}}$, we have

Case 1: if $(k-1)\Delta t_s < (n^p+1)\Delta t_p$, obviously $N_t(k\Delta t_s) \equiv 0$

Case 2: if $(n^p+1)\Delta t_p \le (k-1)\Delta t_s < (n^p+1)\Delta t_p + \Delta t_t$, then there is some l such that $(k-k)\Delta t_s < \cdots \big(k-(l+1)\big)\Delta t_s < (n^p+1)\Delta t_p \le (k-l)\Delta t_s < \cdots < (k-1)\Delta t_s < (n^p+1)\Delta t_p + \Delta t_t$. Then $N_t(k\Delta t_s) = \sum_{i=1}^{l} 2^{(n^p+1)} = l2^{(n^p+1)}$. As long as $(n^p+1)\Delta t_p \le (k-1)\Delta t_s < (n^p+1)\Delta t_p + \Delta t_t$ still holds, $N_t(k\Delta t_s)$ will be strictly increasing w.r.t. k. This marks the growth stage for terminal cells.

Case 3: if $(k-1)\Delta t_s \ge (n^p+1)\Delta t_p + \Delta t_t$, then there exist l and q such that $(k-k)\Delta t_s < \cdots \big(k-(l+1)\big)\Delta t_s < (n^p+1)\Delta t_p \le (k-l)\Delta t_s < \cdots < (k-q)\Delta t_s < (n^p+1)\Delta t_p + \Delta t_t \le \big(k-(q-1)\big)\Delta t_s \le \cdots \le (k-1)\Delta t_s$. In this case $N_t(k\Delta t_s) = \sum_{i=q}^{l} 2^{(n^p+1)} = (l-q+1)2^{(n^p+1)} = \sum_{i=q+1}^{l+1} 2^{(n^p+1)} = N_t((k+1)\Delta t_s)$

Therefore, terminal cell reaches homeostatic condition where old cells die and new cells keep being generated. We have also derived the upper bound and the lower bound for homeostatic terminal cell numbers in:

$$N_t^u(k\Delta t_s) = \left(\left\lceil \frac{\Delta t_t}{\Delta t_s} \right\rceil + 1\right) 2^{(n^p+1)}$$

$$N_t^l(k\Delta t_s) = \left(\left\lfloor \frac{\Delta t_t}{\Delta t_s} \right\rfloor - 1\right) 2^{(n^p+1)}$$

3.2 Mutation Acquisition Probability Propagation

We still assume that cell state propagation within a lineage of the same type of cells follows a Markov process with transition probability $p_{ij} = \binom{m-i}{j-i} u^{j-i}(1-u)^{m-j}\mathbb{I}_{\{0\le i\le j\le m\}}$, where m is the number of mutation hits required for cancer onset, and u is the mutation rate [34].

The transition matrix has the pattern (for $m = 2$):

$$M = \begin{bmatrix} p_{00} & p_{01} & p_{02} \\ 0 & p_{11} & p_{12} \\ 0 & 0 & p_{22} \end{bmatrix}$$

Thus cell state distribution for any generation can be computed given initial state.

Progenitor cells and terminal cells always require more mutation hits to become cancer cells, because they must first acquire enough mutations to enable sufficient self-renewal ability [8]. We can generalize the transition probability when a stem cell gives birth to a progenitor cell, or a progenitor gives birth to a terminal cell.

Suppose that the required number of mutations for cancer onset for stem, progenitor and terminal cell are $m_s \le m_p \le m_t$, then the transition probability from a stem cell to a progenitor cell is:

$$p_{ij}^{sp} = \binom{m_p - i}{j-i} u^{j-i}(1-u)^{m_p-j}\mathbb{I}_{\{0\le i\le j\le m_p\}}\mathbb{I}_{\{0\le i\le m_s\}}$$

For the case of $m_s = 2, m_p = 3$, the transition matrix looks like:

$$M^{sp} = \begin{bmatrix} p_{00} & p_{01} & p_{02} & p_{03} \\ 0 & p_{11} & p_{12} & p_{13} \\ 0 & 0 & p_{22} & p_{23} \end{bmatrix}$$

Likewise, $p_{ij}^{pt} = \binom{m_t - i}{j-i} u^{j-i}(1-u)^{m_t-j}\mathbb{I}_{\{0\le i\le j\le m_t\}}\mathbb{I}_{\{0\le i\le m_p\}}$ specifies the transition from a progenitor cell to a terminal cell.

3.3 Cancer Risk Computation

Let $P[nc]$ denote the probability that none of the cells in the entire cell lineage will become a cancer cell throughout their lifespan. Then the cancer risk is $tLIR = 1 - P[nc]$. To have no cancer, since the cell state can only go up from parent to child, it is sufficient to guarantee that each of the stem cell asymmetric division lineage, along with all its progenitor-terminal lineages, will not have any cancer cells. Denote this probability as p_{nc}. Then $P[nc] = p_{nc}^{N_s}$, where N_s is the number of such lineages, which is usually the same as the final stage stem cell number, or $2^{n_s^s}$, based on the structure. For an illustration of one such stem cell asymmetric division lineage, see Figure 8 below.

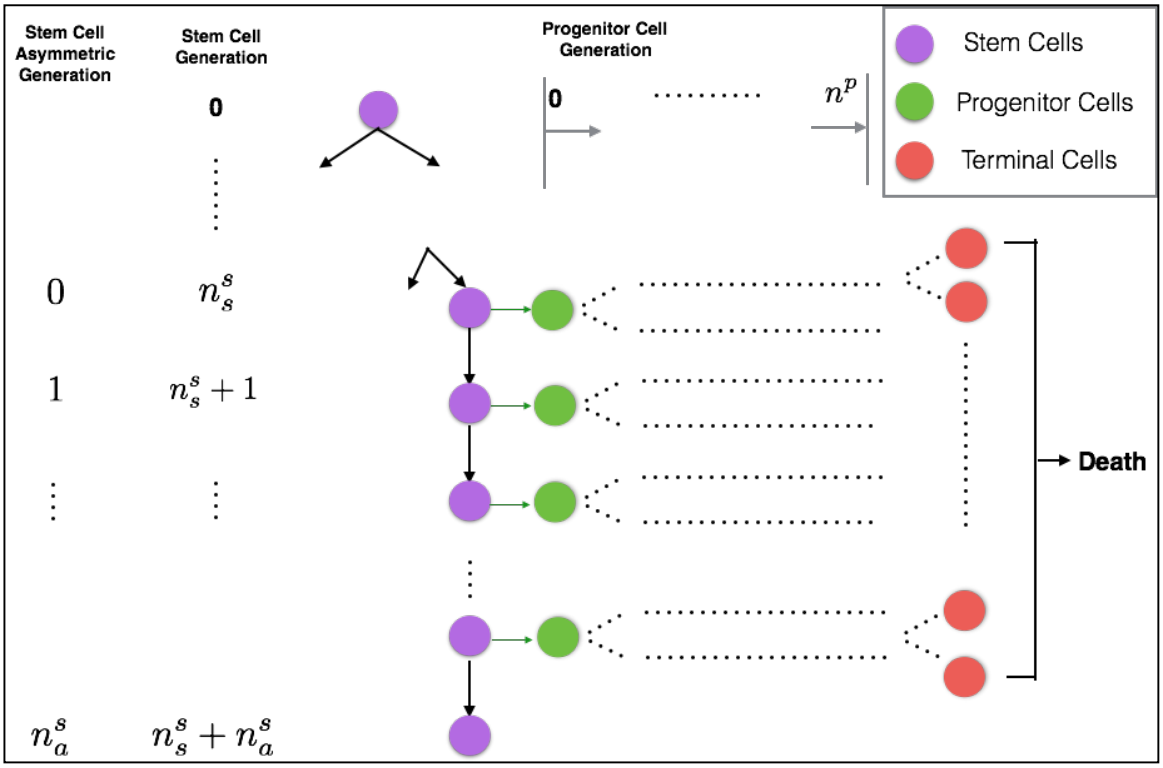


Figure 8: Illustration of one stem cell asymmetric division lineage. This lineage includes one asymmetric division stem cell lineage along with all progenitor lineages that are originated from the asymmetric lineage stem cells. The stem cell asymmetric generation is the relative generation of a stem cell that has asymmetric divisions. The asymmetric generation 0 is total generation n_s^s: the first generation of stem cells that start asymmetric divisions.

We now develop a recursive formula for p_{nc} according to the dependency structure indicated by the lineage. We let $nc(P|j) = \{nc(P)|X = j\}$ be the event that no cancer cell will occur in the progenitor-terminal lineage originated from a stem cell in state j. See illustration in Figure 9 below for one progenitor-terminal lineage originated from one stem cell:

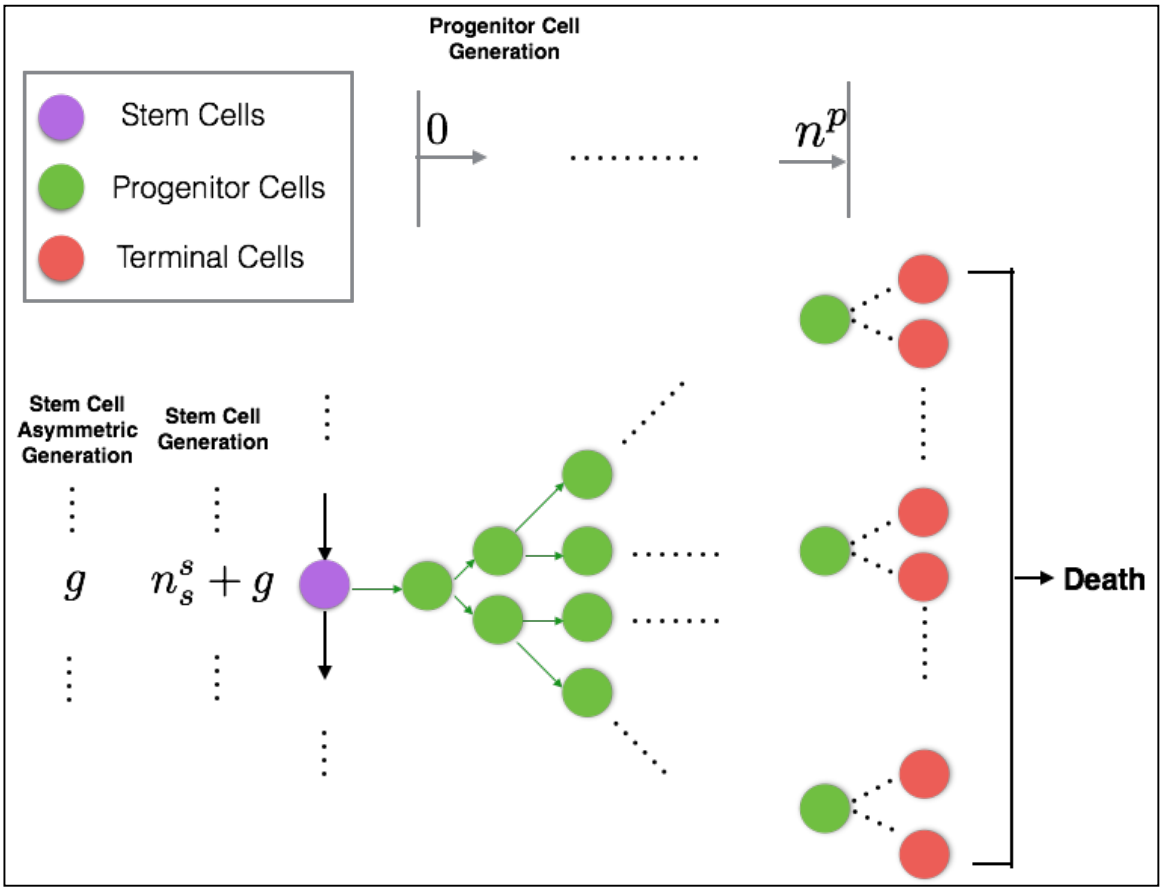


Figure 9: Illustration of the progenitor-terminal lineage originated from one single stem cell of some asymmetric generation $0 \leq g < n_a^s$.

Here we notice that $nc(P|j)$ does not depend on stem cell generations. Now define $A(n, j)$ to be the event that stem cell at generation n (we now define stem cell generation 0 to be the first stem cell that has asymmetric division, i.e. global generation n_s^s) will be

in state j, i.e. $X_n = j$, and no cancer cells occurred through progenitor-terminal lineages originated from stem cells in generation $0,1 \ldots n$. It is obvious that

$$p_{nc} = \sum_{j=0}^{m_s-1} P[A(n_a^s, j)]$$

For $n = 0$, we have

$$P[A(0,j)] = P[X_0 = j \cap nc(P|j)] = P[X_0 = j]P[nc(P|j)]$$

and for $n = 1$, by summing over all possible paths, we have

$$P[A(1,j)] = \sum_{l=0}^{j} P[X_1 = j \cap X_0 = l]\, P[nc(P|l)]P[nc(P|j)]$$

$$= P[nc(P|j)] \sum_{l=0}^{j} p_{lj}^{ss} P[X_0 = l]P[nc(P|l)]$$

$$= P[nc(P|j)] \sum_{l=0}^{j} p_{lj}^{ss} P[A(0,l)]$$

where p_{lj}^{ss} is the stem to stem state transition probability defined before.

For a general $n > 1$, we have

$$P[A(n,j)] = \sum_{i_{n-1}=0}^{j} \sum_{i_{n-2=0}}^{i_{n-1}} \ldots \sum_{i_0=0}^{i_1} P[X_n = j \cap X_{n-1} = i_{n-1} \cap \ldots \cap X_0$$

$$= i_0]P[nc(P|j)]P[nc(P|i_{n-1})] \ldots P[nc(P|i_0)]$$

$$= P[nc(P|j)] \sum_{i_{n-1}=0}^{j} p_{i_{n-1}j}^{ss} P[nc(P|i_{n-1})] \sum_{i_{n-2=0}}^{i_{n-1}} \ldots \sum_{i_0=0}^{i_1} P[X_{n-1}$$

$$= i_{n-1} \cap \ldots \cap X_0 = i_0]P[nc(P|i_{n-2})] \ldots P[nc(P|i_0)]$$

$$= P[nc(P|j)] \sum_{l=0}^{j} p_{lj}^{ss} P[A(n-1,l)]$$

We need to compute $P[nc(P|j)]$. Let X_{n^p} be the state of any n^p generation progenitor cell from the progenitor lineage originated from a j-state stem cell. Let X_t be the state of one of terminal cells generated by the corresponding progenitor cell. See Figure 10 below:

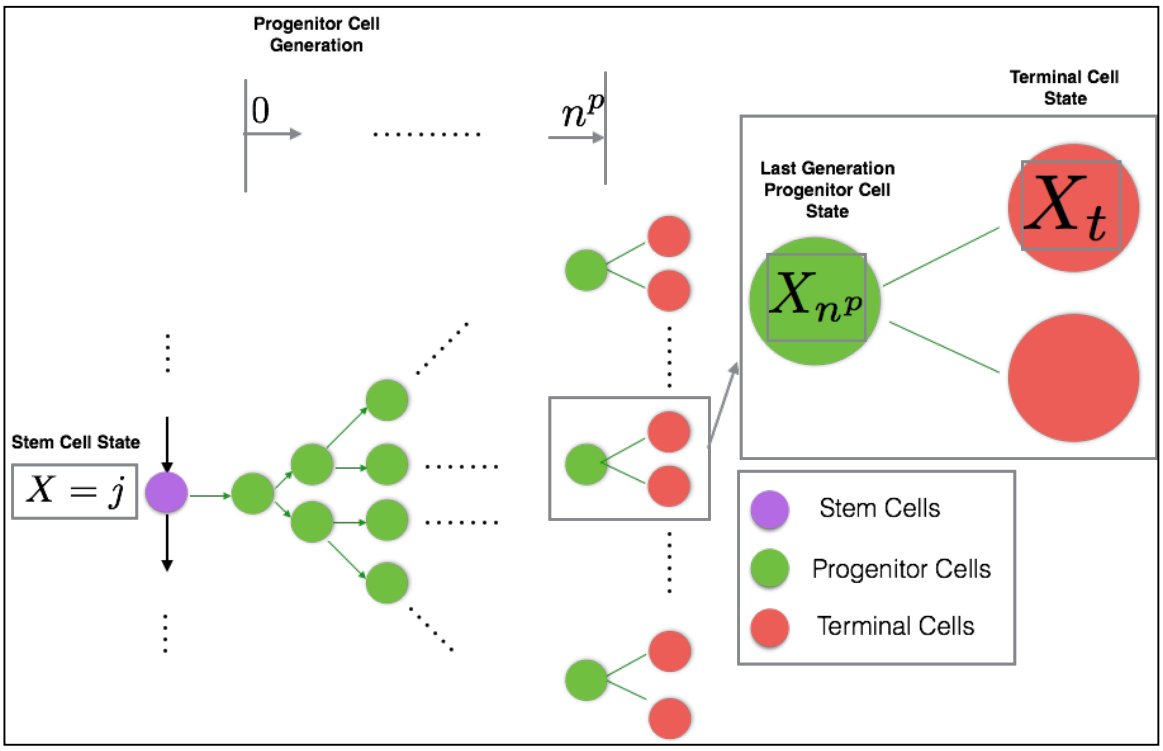


Figure 10: Mutation state notations in one progenitor-terminal lineage originated from one stem cell. Let $X = j$ be the mutation state of the stem cell. Let random variable X_{n^p} be the mutation state of any final generation progenitor cell. Let random variable X_t be the mutation state of either one of daughter terminal cells generated from the progenitor cell.

According to the propagation property of cell states, we have

$$P[nc(P|j)] = (\sum_{i=j}^{m_p-1} P[X_{n^p} = i|X = j](\sum_{l=i}^{m_t-1} P[X_t = l|X_{n^p} = i])^2)^{2^{n^p}}$$

Note that $P[X_{n^p} = i|X = j] = p_{ji}^{sp}$ and $P[X_t = l|X_{n^p} = i] = p_{il}^{pt}$.

With these arguments in place, we can finally compute the total cancer risk.

Now we provide the cancer risk computed through our model using the data in the supplementary material of [31], which provided a list of common types of tissues with their stem cell number, total cell number, cell division parameters and observed cancer risk. For example, see Table 1 blow for the first 3 rows of the data:

Table 1 [31]. Thirty-one tissue/cancer types with their lifetime cancer risk and relevant parameters, from supplementary material for [31]. where id denotes the cancer type id number, following the same order as in Table S1 of the supplementary material for [31]. Here N_{total} is the total number of normal cells in the tissue of origin, N_{stem} is the number of stem cells in the tissue of origin, n_{total} is the number of divisions of each stem cell per lifetime, $lscd$ is the cumulative number of divisions of all stem cells per lifetime, and, cr is the observed cancer risk.

id	*Tissue/cancer_ name*	*risk_observe d (cr)*	*num_total (N_{total})*	*num_stem (N_{stem})*	*num_stem_ generation (n_{total})*	*num_stem_div_ all_cells (lscd)*
1	Acute myeloid leukemia	0.0041	3.00E+12	1.35E+08	960	1.30E+11
2	Basal cell carcinoma	0.3	1.80E+11	5.82E+09	608	3.55E+12
3	Chronic lymphocytic leukemia	0.0052	3.00E+12	1.35E+08	960	1.30E+11
4	Colorectal adenocarcinoma	0.048	3.00E+10	2.00E+08	5840	1.17E+12
5	Colorectal adenocarcinoma with FAP	1	3.00E+10	2.00E+08	5840	1.17E+12
6	Colorectal adenocarcinoma with Lynch syndrome	0.5	3.00E+10	2.00E+08	5840	1.17E+12
7	Duodenum adenocarcinoma	0.0003	6.80E+08	4.00E+06	1947	7.80E+09
8	Duodenum adenocarcinoma with FAP	0.035	6.80E+08	4.00E+06	1947	7.80E+09
9	Esophageal squamous cell carcinoma	0.001938	3.24E+09	8.64E+05	1390	1.20E+09
10	Gallbladder non papillary adenocarcinoma	0.0028	1.60E+08	1.60E+06	47	7.84E+07
11	Glioblastoma	0.00219	8.46E+10	1.35E+08	0	2.70E+08
12	Head and neck squamous cell carcinoma	0.0138	1.67E+10	1.85E+07	1720	3.19E+10
13	Head and neck squamous cell carcinoma with HPV-16	0.07935	1.67E+10	1.85E+07	1720	3.19E+10
14	Hepatocellular carcinoma	0.0071	2.41E+11	3.01E+09	88	2.71E+11
15	Hepatocellular carcinoma with HCV	0.071	2.41E+11	3.01E+09	88	2.71E+11
16	Lung adenocarcinoma (nonsmokers)	0.0045	4.34E+11	1.22E+09	5.6	9.27E+09
17	Lung adenocarcinoma (smokers)	0.081	4.34E+11	1.22E+09	5.6	9.27E+09

18	Medulloblastoma	0.00011	8.50E+10	1.36E+08	0	2.72E+08
19	Melanoma	0.0203	3.80E+09	3.80E+09	199	7.64E+11
20	Osteosarcoma	0.00035	1.90E+09	4.18E+06	5	2.93E+07
21	Osteosarcoma of the arms	0.00004	3.00E+08	6.50E+05	5	4.55E+06
22	Osteosarcoma of the head	0.0000302	3.90E+08	8.60E+05	5	6.02E+06
23	Osteosarcoma of the legs	0.00022	7.20E+08	1.59E+06	5	1.11E+07
24	Osteosarcoma of the pelvis	0.00003	2.00E+08	4.50E+05	5	3.15E+06
25	Ovarian germ cell	0.000411	1.10E+07	1.10E+07	0	2.20E+07
26	Pancreatic ductal adenocarcinoma	0.013589	1.67E+11	4.18E+09	80	3.43E+11
27	Pancreatic endocrine (islet cell) carcinoma	0.000194	2.95E+09	7.40E+07	80	6.07E+09
28	Small intestine adenocarcinoma	0.0007	1.70E+10	1.00E+08	2920	2.92E+11
29	Testicular germ cell cancer	0.0037	2.16E+10	7.20E+06	463	3.35E+09
30	Thyroid papillary/follicular carcinoma	0.01026	1.00E+10	6.50E+07	7	5.85E+08
31	Thyroid medullary carcinoma	0.000324	1.00E+09	6.50E+06	7	5.85E+07

Here N_{total} is the total number of normal cells in the tissue of origin, N_{stem} is the number of stem cells in the tissue of origin, n_{yr} is the number of divisions of each stem cell per year, n_{total} is the number of divisions of each stem cell per lifetime, and, $lscd$ is the cumulative number of divisions of all stem cells per lifetime.

In this chapter we use the provided data to set up our model parameters and then compute an approximated cancer risk upper bound. In addition, we also present the cancer risk computed using our original stem-cell only model presented in [34] as a comparison.

The only parameters relevant to cancer risk computation are n_s^s, n_a^s and n^p. The division/death rates are not involved in our mutation acquisition model. The parameters n_s^s (upper bound) and n_a^s can be directly determined from the data provided with $n_s^s = \lceil \log_2(N_{stem}) \rceil$ and $n_a^s = N_{total}$.

According to the probability propagation and cancer risk computing formulas, the longer the cell lineage is, the higher the cancer risk will be. Therefore, to obtain the upper bound,

we can determine n^p, the length of progenitor lineage, so that the homeostatic progenitor and terminal cell number will not be lower than the given values even only one generation of stem cell are taken into account. Specifically, we can assume all cells except stem cells are progenitor cells, i.e. $N_p = N_{total} - N_{stem}$. Then we assume the contribution to N_p all come from one single generation of stem cells, such that we have $N_{stem} \times 2^{n^p} = N_p$ and $n^p = \lceil \log_2(N_p/N_{stem}) \rceil$. We chose the required mutation number for cancer onset to be: $m_s = 3$ for stem cells, $m_p = 4$ for progenitor cells and $m_t = 5$ for terminal cells. We chose the mutation rate for all cells to be $u = 2.5 \times 10^{-8}$.

Table 2 below contains the cancer risk computed using our model and the stem cell-only model in [34]. Here n^p_{upper} is the progenitor lineage length computed according to arguments above, N_p^l is the lower bound number of progenitor and terminal cells based on n^p_{upper}, n_a^s and, n_s^s. $N_s = 2^{n_s^s}$. The total cell number from the model is $N = N_p^l + N_s$. In addition, $cr_{observed}$ is the observed cancer risk, cr_u is the upper bound cancer risk, and cr_{stem} is the cancer risk from the stem cell-only model in [34]. The tissue types follow the same order as in the supplementary material of [31].

Table 2: Cancer risk observed in [31], computed using our intermediate model, and computed using our original stem-cell-only model [34]. Here $cr_{observed}$ is the observed cancer risk; cr_u is the upper bound cancer risk computed through our model and cr_{stem} is the cancer risk from the stem cell-only model in [34]. In addition, n^p_{upper} is the progenitor lineage length computed according to arguments above, N_p^l is the lower bound number of progenitor and terminal cells based on n^p_{upper}, n_a^s and n_s^s. $N_s = 2^{n_s^s}$. The total cell number from the model is $N = N_p^l + N_s$. Tissue types with significantly different cr_u and cr_{stem} are highlighted. Note that the id number here follows the same order with the tissue types in supplementary material of [31].

id	N_{total}	N_{stem}	n_a^s	$cr_{observed}$	n_s^s	n^p_{upper}	N_p^l	N_s	N	cr_u	cr_{stem}
1	3.00E+12	1.35E+08	960	4.100000E-03	28	14	8.80E+12	2.68E+08	8.80E+12	2.590000E-06	2.040000E-06
2	1.80E+11	5.82E+09	608	3.000000E-01	33	4	2.75E+11	8.59E+09	2.83E+11	6.300704E-03	2.390000E-05
3	3.00E+12	1.35E+08	960	5.200000E-03	28	14	8.80E+12	2.68E+08	8.80E+12	2.590000E-06	2.040000E-06
4	3.00E+10	2.00E+08	5840	4.800000E-02	28	7	6.87E+10	2.68E+08	6.90E+10	6.314950E-04	6.310960E-04

5	3.00E+10	2.00E+08	5840	1.000000E+00	28	7	6.87E+10	2.68E+08	6.90E+10	6.314950E-04	6.310960E-04
6	3.00E+10	2.00E+08	5840	5.000000E-01	28	7	6.87E+10	2.68E+08	6.90E+10	6.314950E-04	6.310960E-04
7	6.80E+08	4.00E+06	1947	3.000000E-04	22	7	1.07E+09	4.19E+06	1.08E+09	4.770000E-07	4.770000E-07
8	6.80E+08	4.00E+06	1947	3.500000E-02	22	7	1.07E+09	4.19E+06	1.08E+09	4.770000E-07	4.770000E-07
9	3.24E+09	8.64E+05	1390	1.938000E-03	20	11	4.29E+09	1.05E+06	4.30E+09	3.870000E-08	3.780000E-08
10	1.60E+08	1.60E+06	47	2.800000E-03	21	6	2.68E+08	2.10E+06	2.71E+08	7.860000E-12	7.860000E-12
11	8.46E+10	1.35E+08	0	2.190000E-03	28	9	2.75E+11	2.68E+08	2.75E+11	4.630000E-11	4.630000E-11
12	1.67E+10	1.85E+07	1720	1.380000E-02	25	9	3.44E+10	3.36E+07	3.44E+10	1.540000E-06	1.540000E-06
13	1.67E+10	1.85E+07	1720	7.935000E-02	25	9	3.44E+10	3.36E+07	3.44E+10	1.540000E-06	1.540000E-06
14	2.41E+11	3.01E+09	88	7.100000E-03	32	6	5.50E+11	4.29E+09	5.54E+11	3.340000E-07	8.130000E-08
15	2.41E+11	3.01E+09	88	7.100000E-02	32	6	5.50E+11	4.29E+09	5.54E+11	3.340000E-07	8.130000E-08
16	4.34E+11	1.22E+09	5.6	4.500000E-03	31	8	1.10E+12	2.15E+09	1.10E+12	8.890000E-10	8.890000E-10
17	4.34E+11	1.22E+09	5.6	8.100000E-02	31	8	1.10E+12	2.15E+09	1.10E+12	8.890000E-10	8.890000E-10
18	8.50E+10	1.36E+08	0	1.100000E-04	28	9	2.75E+11	2.68E+08	2.75E+11	4.660000E-11	4.660000E-11
19	3.80E+09	3.80E+09	199	2.030000E-02	32	0	8.59E+09	4.29E+09	1.29E+10	8.440000E-07	8.440000E-07
20	1.90E+09	4.18E+06	5	3.500000E-04	22	8	2.15E+09	4.19E+06	2.15E+09	1.290000E-12	1.290000E-12
21	3.00E+08	6.50E+05	5	4.000000E-05	20	8	5.37E+08	1.05E+06	5.38E+08	1.590000E-13	1.590000E-13
22	3.90E+08	8.60E+05	5	3.020000E-05	20	8	5.37E+08	1.05E+06	5.38E+08	2.100000E-13	2.100000E-13
23	7.20E+08	1.59E+06	5	2.200000E-04	21	8	1.07E+09	2.10E+06	1.08E+09	4.370000E-13	4.370000E-13
24	2.00E+08	4.50E+05	5	3.000000E-05	19	8	2.68E+08	5.24E+05	2.69E+08	9.720000E-14	9.720000E-14
25	1.10E+07	1.10E+07	0	4.110000E-04	24	0	3.36E+07	1.68E+07	5.03E+07	2.380000E-12	2.380000E-12
26	1.67E+11	4.18E+09	80	1.358900E-02	32	5	2.75E+11	4.29E+09	2.79E+11	9.180000E-08	9.180000E-08
27	2.95E+09	7.40E+07	80	1.940000E-04	27	5	8.59E+09	1.34E+08	8.72E+09	1.420000E-09	1.420000E-09
28	1.70E+10	1.00E+08	2920	7.000000E-04	27	7	3.44E+10	1.34E+08	3.45E+10	4.000000E-05	4.000000E-05
29	2.16E+10	7.20E+06	463	3.700000E-03	23	11	3.44E+10	8.39E+06	3.44E+10	1.280000E-08	1.280000E-08
30	1.00E+10	6.50E+07	7	1.026000E-02	26	7	1.72E+10	6.71E+07	1.72E+10	3.650000E-11	3.650000E-11
31	1.00E+09	6.50E+06	7	3.240000E-04	23	7	2.15E+09	8.39E+06	2.16E+09	7.220000E-10	2.740000E-12

For all tissue types, both cr_u and cr_{stem} are far below the observed risk $cr_{observed}$; cr_u and cr_{stem} are very close for most tissue types, except for tissues 2 (Basal cell carcinoma),14/15 (Hepatocellular carcinoma without/with HPC), and 31 (Thyroid medullary carcinoma).

Figure 11 below provides a sensitivity analysis on different mutation rates and two different scenarios: 1×10^{-10} (stem cell only with $m_s = 3$), 1×10^{-9} (stem cell only with

$m_s = 3$), 1×10^{-8} (stem cell only with $m_s = 3$), 2.5×10^{-8} (stem cell only with $m_s = 3$), and 2.5×10^{-8} (our model with $m_s = 3$, $m_p = 4$, and $m_t = 5$). We can see that the cancer risk level with progenitor and terminal cell involved is still below the observed level.

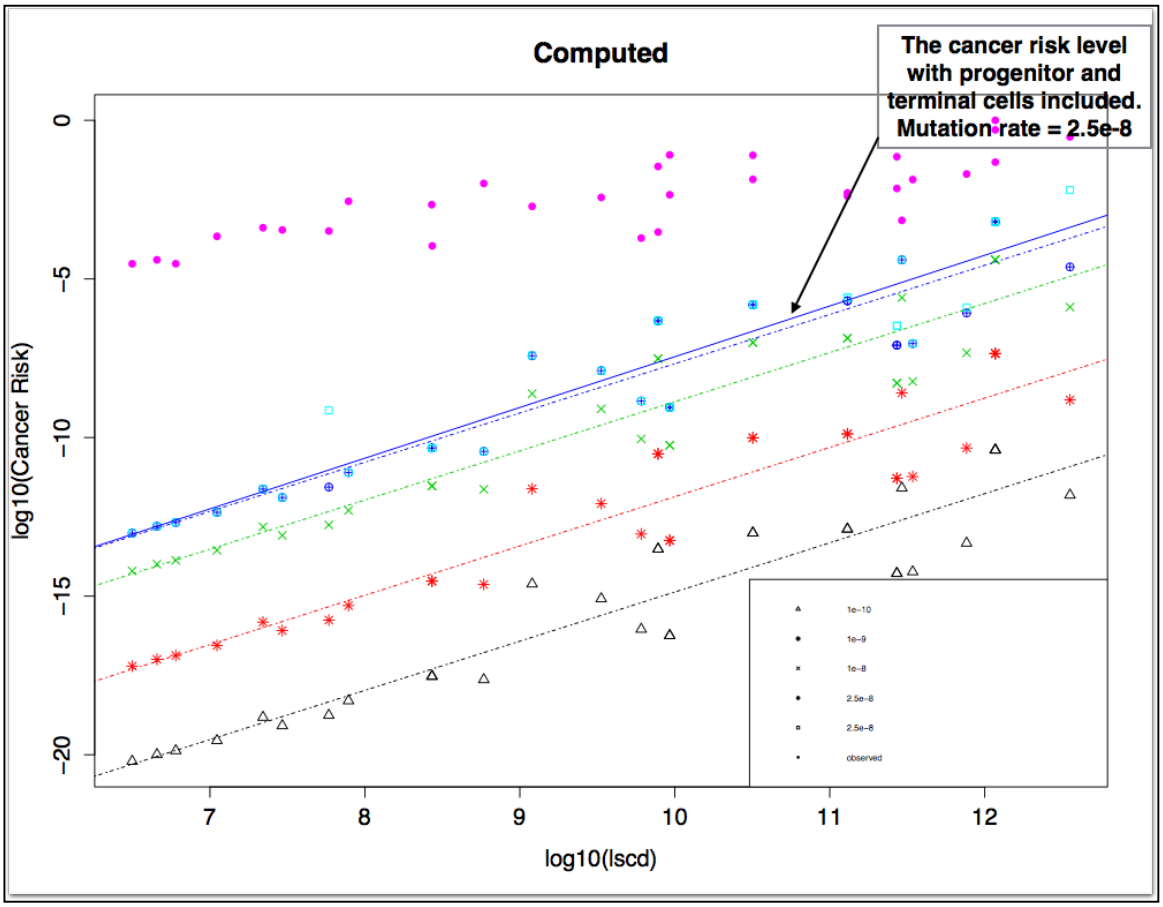


Figure 11: A sensitivity analysis of different mutation rates on $m_s = 3$. The figure provides computed cancer risk level using stem-cell-only model in [7] with mutation rates: 1×10^{-10}, 1×10^{-9}, 1×10^{-8}, and 2.5×10^{-8} and also the upper bound cancer risk level computed through our model for the setting: $u_s = u_p = 2.5\times10^{-8}$ and $m_s = 3, m_p = 4, m_t = 5$.

Chapter 4 Extended Risk Model

For the model in Section 2.5.1 and Chapter 3, the cell division structure is fixed, i.e. stem cells first go through symmetric self-renewals and then asymmetric divisions throughout lifetime, which is a strong assumption for cell dynamics. Hence, we need to extend the previous model to allow more general division patterns (e.g. self-renewal, differentiation, etc.) with a certain probability distribution, for stem and progenitor cells at each of their generations.

During mutation acquisitions, different driver mutations could have different effects on mutation rate or cell dynamics [1, 11, 39]. These effects might include increasing cell fitness, increasing mutation rate, reducing cell death probability, etc. [39]. Therefore, as opposed to using simply the number of mutations to label cell status in Chapter 3, we need to differentiate various driver mutations to sufficiently describe the multistage cancer development.

In addition, we will provide an extended theoretical lifetime intrinsic risk (eTLIR) computation by including all general division patterns, using a recursive framework. The eTLIR model will also be able to simulate different mutation effects, including clonal expansion.

This chapter will be structured as follows: First we will formulate cell dynamics model allowing generalized cell division patterns at each generation in Sections 4.1 and 4.2. Then we will describe how to model mutation acquisitions for different driver mutations and their effects on cell behaviors in Section 4.3. We will provide a detailed derivation of the extended theoretical lifetime risk computation model, including mutation effects and clonal expansion in Section 4.4. In Section 4.5, we discuss the algorithm for cell

dynamics under clonal expansion. Finally, we will describe how to quantify the contributions of intrinsic factors to the observed cancer risk, in Section 4.6.

4.1 General Division Patterns

Generally, a stem or progenitor cell could go through one of 6 behaviors, illustrated in Figure 12, at the beginning of a cell cycle at each generation.

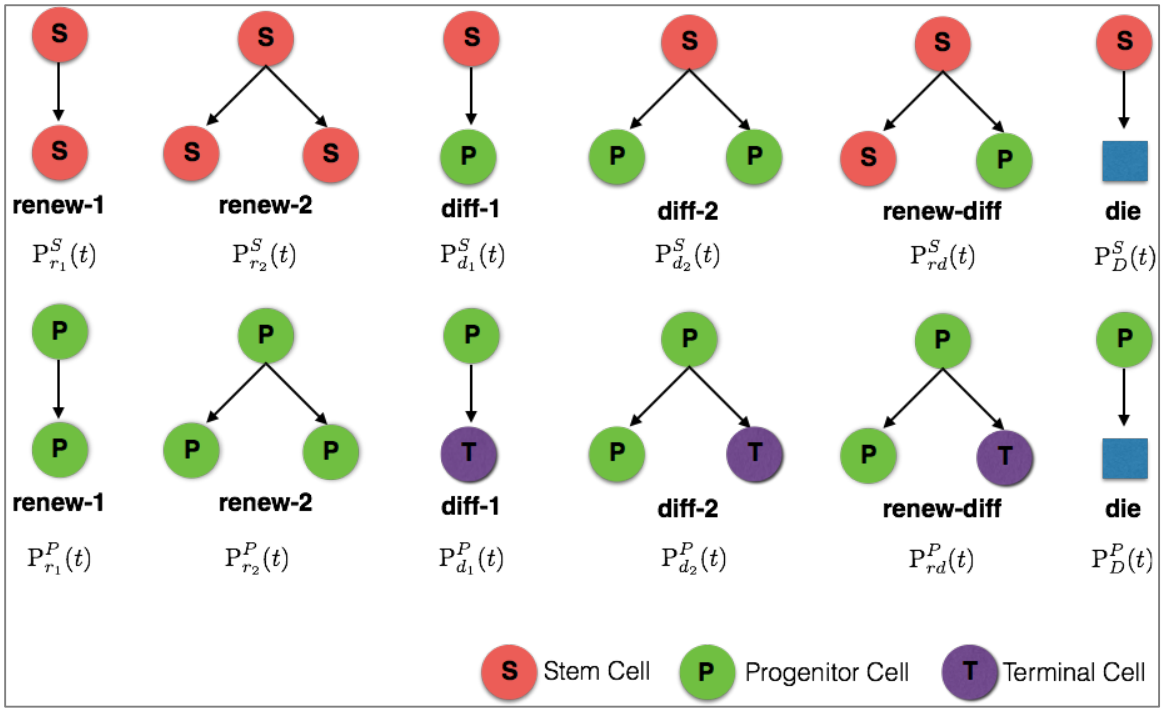


Figure 12: General types of cell divisions.

A stem/progenitor cell, at the moment of division, could self-renew to 1 or 2 daughter stem/progenitor cells, with probabilities $P^C_{r1}(t)$ and $P^C_{r2}(t)$, where C denotes the type of cells, i.e. $C = S$ for stem cells and $C = P$ for progenitor cells, where t is the time for the beginning of some cell cycle. In addition, a cell could differentiate to 1 or 2 differentiated cells (stem to progenitor, or progenitor to terminal) with probabilities $P^C_{d1}(t)$ and $P^C_{d2}(t)$. Also a cell could generate 1 renewed daughter cell and 1 terminal daughter cell, with probabilities $P^C_{rd}(t)$. Finally, a cell could die with probability $P^C_D(t)$. Later we will see

that these cell behavior probabilities, $P_b^C(t), b \in \{r1, r2, d1, d2, rd, D\}; C \in \{S, P\}$ could also vary depending on cell mutation acquisitions, by incorporating mutation effects on dynamics.

Let g_S and g_P be the maximum generation numbers (total number of divisions) of stem and progenitor cells, respectively, in their lineages. Also, let $\Delta t_S(t), \Delta t_P(t)$ and $\Delta t_T(t)$ be the cell cycle time for stem, progenitor, and terminal cells at some time t for the beginning of a cell cycle. We denote $t_S(h), h = 0,1 \dots g_S$ be the start time of stem cell cycles at generation h; $t_P(h_S, h_P), h_S = 0,1 \dots (g_S - 1); h_P = 0,1 \dots g_P$ be the start time of progenitor cell cycles, for generation h_P progenitor cells originated from h_S generation stem cells (progenitor lineage h_S); $t_T(h_S, h_P, 0)$ and $t_T(h_S, h_P, 1), h_S = 0,1 \dots (g_S - 1); h_P = 0,1 \dots g_P$ be the start and end (death) time points of terminal cell cycles, for terminal cells generated from h_P generation progenitor cells in lineage h_S. Figure 13 below provides an illustration for the general lifetime cell evolution.

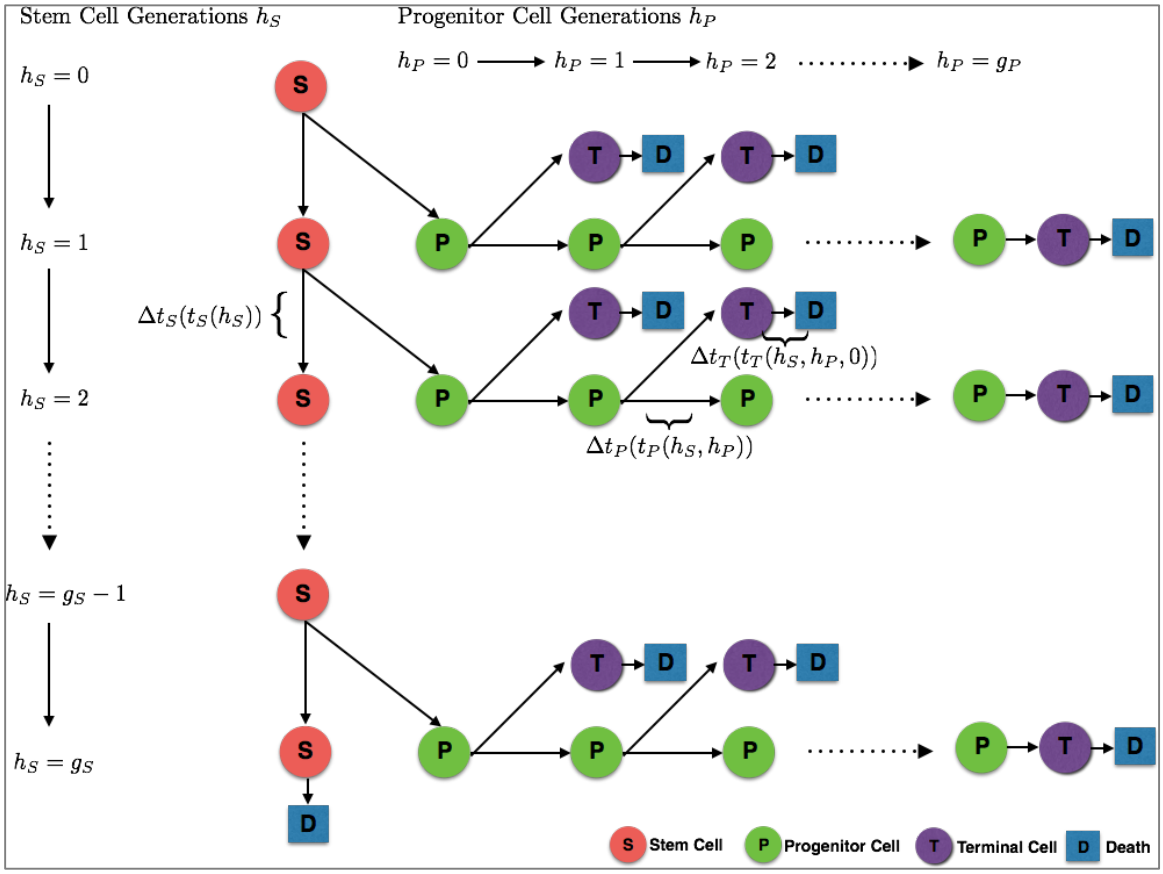


Figure 13: General lifetime cell evolution structure.

We can easily determine the start times for each cell cycles with $\Delta t_S(t), \Delta t_P(t)$ and $\Delta t_T(t)$. For stem cells, $t_S(0) = 0$ and $t_S(h) = t_S(h-1) + \Delta t_S\big(t_S(h-1)\big)$ for $h = 1 \dots g_S$; similarly for progenitor cells of an arbitrary branch $h_S = 0 \dots (g_S - 1)$, we have $t_P(h_S, 0) = t_S(h_S) + \Delta t_S\big(t_S(h_S)\big) = t_S(h_S + 1)$ and $t_P(h_S, h_P) = t_P(h_S, h_P - 1) + \Delta t_P(t_P(h_S, h_P - 1))$ for $h_P = 1 \dots g_P$; for terminal cells, we have its cell cycle start time $t_T(h_S, h_P, 0) = t_P(h_S, h_P) + \Delta t_P\big(t_P(h_S, h_P)\big)$ and cell death time $t_T(h_S, h_P, 1) = t_T(h_S, h_P, 0) + \Delta t_T\big(t_T(h_S, h_P, 0)\big)$, for terminal cells generated by h_P generation progenitor cells within branch h_S.

4.2 Cell Dynamics and Homeostatic Condition

We now describe how cell numbers changes stochastically with time based on the general division patterns in Section 4.1. Later we will derive homeostatic conditions on expected cell numbers.

Let $N_S(t), N_P(t)$ and $N_T(t)$ be the number of stem, progenitor and terminal cells at time t. To derive $N_P(t)$ and $N_T(t)$, we define auxiliary cell numbers for progenitor and terminal cells: $N_P(t, h_S)$ and $N_T(t, h_S)$ to be the progenitor/terminal cell number within branch h_S at time t; and $N_T(t, h_S, h_P)$ to be the number of terminal cells generated from h_P generation progenitor cells, among $N_T(t, h_S)$.

Since each cell will randomly go through one of these behaviors independently, the transition of cell numbers from one generation to the next should follow a multinomial distribution. Let $n_S(h_S, b), b \in \{r1, r2, d1, d2, rd, D\}$ be the number of stem cells, among all stem cells of Generation h_S, that will have behavior b during their cell cycles. Similarly, we can define $n_P(h_S, h_P, b)$ for progenitor cells.

We have

$$\{n_S(h_S, b)\}|_{b\in\{r1,r2,d1,d2,rd,D\}}$$

$$\sim \text{Multinomial}\left(N_S\left(t_S(h_S)\right), \{P_b^S\left(t_S(h_S)\right)\}|_{b\in\{r1,r2,d1,d2,rd,D\}}\right)$$

and

$$\{n_P(h_S, h_P, b)\}|_{b\in\{r1,r2,d1,d2,rd,D\}}$$

$$\sim \text{Multinomial}(N_P\left(t_P(h_S, h_P)\right), \{P_b^P\left(t_P(h_S, h_P)\right)\}|_{b\in\{r1,r2,d1,d2,rd,D\}})$$

Now the transitions are obvious; basically, part of the stem cells will renew to yield the next generation of stem cells:

$N_S\left(t_S(h_S + 1)\right) = 2n_S(h_S, r2) + n_S(h_S, r1) + n_S(h_S, rd)$ for $h_S = 0 \ldots g_S - 1$

and the rest of stem cells will differentiate into initial generation progenitor cells within the particular branch:

$N_P(t_P(h_S, 0), h_S) = 2n_S(h_S, d2) + n_S(h_S, d1) + n_S(h_S, rd)$

Similarly, some of the progenitor cells will renew to yield the next generation's progenitor cells, and the rest differentiate into terminal cells. The last generation progenitor cells will all become terminal cells. In general,

$N_P(t_P(h_S, h_P + 1), h_S) = 2n_P(h_S, h_P, r2) + n_P(h_S, h_P, r1) + n_P(h_S, h_P, rd)$ for h_P

$= 0 \dots g_P - 1$

and

$N_T(t_T(h_S, h_P, 0), h_S, h_P) = 2n_P(h_S, h_P, d2) + n_P(h_S, h_P, d1) + n_P(h_S, h_P, rd)$ for h_P

$= 0 \dots g_P$

With the above transition relations and the initial condition $N_S(0) = 1$, we can easily compute cell numbers at the start time of the corresponding cell cycles: $N_S\big(t_S(h_S)\big), N_P(t_P(h_S, h_P), h_S)$ and $N_T(t_T(h_S, h_P, 0), h_S, h_P)$.

Then for an arbitrary time t, we have

$$N_S(t) = \sum_{h_S=0}^{g_S-1} \mathbb{I}_{\{t_S(h_S) \le t < t_S(h_S+1)\}} N_S\big(t_S(h_S)\big) + \mathbb{I}_{\{t \ge t_S(g_S)\}} N_S\big(t_S(g_S)\big)$$

$$N_P(t, h_S) = \sum_{h_P=0}^{g_P-1} \mathbb{I}_{\{t_P(h_S,h_P) \le t < t_P(h_S,h_P+1)\}} N_P(t_P(h_S, h_P), h_S)$$

$$+ \mathbb{I}_{\{t_P(h_S,g_P) \le t < t_T(h_S,g_P,0)\}} N_P(t_P(h_S, g_P), h_S)$$

$$N_T(t, h_S, h_P) = \mathbb{I}_{\{t_T(h_S,h_P,0) \le t < t_T(h_S,h_P,1)\}} N_T(t_T(h_S, h_P, 0), h_S, h_P)$$

For progenitor and terminal cells, the total cell number can be obtained by summing over all sub-branches.

$$N_P(t) = \sum_{h_S=0}^{g_S-1} N_P(t, h_S)$$

$$N_T(t) = \sum_{h_S=0}^{g_S-1} \sum_{h_P=0}^{g_P} N_T(t, h_S, h_P)$$

Note that the above cell numbers are basically stochastic and can be obtained through simulation in practice. We now derive simplified expressions, in closed form, for expected values of $N_S(t), N_P(t)$ and $N_T(t)$ and then discuss the constraints on cell behavior probabilities, cell lineage structure and cell cycle time required for homeostatic states.

Let N^H be the homeostatic cell number, among which N_S^H, N_P^H and N_T^H are homeostatic stem, progenitor and terminal cell numbers. Note that it is possible that only some of these homeostatic numbers are available. For example, in [31] only N^H and N_S^H are provided, in which case we will have weaker constraints on progenitor and terminal cell parameters.

At the steady state, in general we need $E[N_S(t)] = N_S^H, E[N_P(t)] = N_P^H, E[N_T(t)] = N_T^H \; and \; E[N_S(t) + N_P(t) + N_T(t)] = N^H$. Also, it is reasonable to assume that in steady state the probability for different cell behaviors and cell cycle times remain unchanged with time, i.e. $P_b^S(t) \equiv P_b^S, P_b^P(t) \equiv P_b^P$ (except for the last progenitor generation where all progenitor cells evolve to terminal cells) and $\Delta t_S(t) \equiv \Delta t_S, \Delta t_P(t) \equiv \Delta t_P, \Delta t_T(t) \equiv \Delta t_T$.

We introduce notation $\lambda[A|B]$ to be the average number of A cells, generated from one B cell through renewal or differentiation in steady state. Then it is obvious that $\lambda[S|S] = 2P_{r2}^S + P_{r1}^S + P_{rd}^S, \lambda[P|S] = 2P_{d2}^S + P_{d1}^S + P_{rd}^S, \lambda[P|P] = 2P_{r2}^P + P_{r1}^P + P_{rd}^P$ and

$\lambda[T|P] = 2P_{d2}^{P} + P_{d1}^{P} + P_{rd}^{P}$. Note that for last generation progenitor cells, $\lambda[P|P] = 0$ and $\lambda[T|P] = 1$. Now we build the homeostatic conditions.

For stem cells, since

$$E[N_S(t)] = \sum_{h_S=0}^{g_S-1} \mathbb{I}_{\{t_S(h_S)\le t<t_S(h_S+1)\}} E[N_S(t_S(h_S))] + \mathbb{I}_{\{t\ge t_S(g_S)\}} E[N_S(t_S(g_S))]$$

we only need $E[N_S(t_S(h_S))]$ to stay unchanged w.r.t. h_S. Since it is not hard to see that $E[N_S(t_S(h_S))] = \lambda[S|S]E[N_S(t_S(h_S-1))]$, we just require $\lambda[S|S] = 1$ and $E[N_S(t_S(h_S))] \equiv N_S^H$, which is the condition for stem cell homeostasis.

For progenitor cells, similarly we can express the transition in steady state as:

$$E[N_P(t_P(h_S,0),h_S)] = \lambda[P|S]E[N_S(t_S(h_S))] \equiv \lambda[P|S]N_S^H$$

and

$$E[N_P(t_P(h_S,h_P),h_S)] = \lambda[P|P]E[N_P(t_P(h_S,h_P-1),h_S)] \text{ for } h_P = 1 \dots g_P$$

Therefore, we have

$$E[N_P(t_P(h_S,h_P),h_S)] = \lambda[P|P]^{h_P}\lambda[P|S]N_S^H \text{ for } h_P = 0 \dots g_P$$

As a fundamental difference to stem cell lineage, for progenitor cells we could include multiple branches when computing total cell number. At an arbitrary time t, we can show that the approximate average number of h_P generation progenitor cell cycles, across all branches, that include time t, equal to the ratio of progenitor cell cycle time and stem cell cycle time, i.e.

$$\sum_{h_S=0}^{g_S-1} \mathbb{I}_{\{t_P(h_S,h_P)\le t<t_P(h_S,h_P+1)\}} \approx \frac{\Delta t_P}{\Delta t_S} \text{ for } h_P = 0 \dots g_P - 1$$

and

$$\sum_{h_S=0}^{g_S-1} \mathbb{I}_{\{t_P(h_S,g_P)\le t<t_T(h_S,g_P,0)\}} \approx \frac{\Delta t_P}{\Delta t_S}$$

Therefore,

$$E[N_P(t)] = \sum_{h_S=0}^{g_S-1} E[N_P(t,h_S)]$$

$$= \sum_{h_S=0}^{g_S-1}\sum_{h_P=0}^{g_P-1} \mathbb{I}_{\{t_P(h_S,h_P)\le t<t_P(h_S,h_P+1)\}} E[N_P(t_P(h_S,h_P),h_S)]$$

$$+ \sum_{h_S=0}^{g_S-1} \mathbb{I}_{\{t_P(h_S,g_P)\le t<t_T(h_S,g_P,0)\}} E[N_P(t_P(h_S,g_P),h_S)]$$

$$= \sum_{h_P=0}^{g_P} \frac{\Delta t_P}{\Delta t_S}\lambda[P|P]^{h_P}\lambda[P|S]N_S^H$$

$$= \begin{cases} \dfrac{\Delta t_P}{\Delta t_S}\lambda[P|S]N_S^H & \text{if } \lambda[P|P]=0 \\ (g_P+1)\dfrac{\Delta t_P}{\Delta t_S}\lambda[P|S]N_S^H & \text{if } \lambda[P|P]=1 \\ \dfrac{\Delta t_P}{\Delta t_S}\lambda[P|S]N_S^H\dfrac{1-\lambda[P|P]^{g_P+1}}{1-\lambda[P|P]} & \text{otherwise} \end{cases}$$

For terminal cells, it is easy to show that

$$N_T(t_T(h_S,h_P,0),h_S,h_P) = \lambda[T|P]E[N_P(t_P(h_S,h_P),h_S)]$$

$$= \lambda[T|P]\lambda[P|P]^{h_P}\lambda[P|S]N_S^H \text{ for } h_P = 0 \dots g_P - 1$$

and

$$N_T(t_T(h_S,h_P,0),h_S,h_P) = E[N_P(t_P(h_S,h_P),h_S)] = \lambda[P|P]^{h_P}\lambda[P|S]N_S^H \text{ for } h_P = g_P$$

Similar to the arguments for progenitor cells, we have

$$\sum_{h_S=0}^{g_S-1} \mathbb{I}_{\{t_T(h_S,h_P,0)\le t<t_T(h_S,h_P,1)\}} \approx \frac{\Delta t_T}{\Delta t_S} \text{ for } h_P = 0 \dots g_P$$

Therefore,

$$E[N_T(t)] = \sum_{h_S=0}^{g_S-1} \sum_{h_P=0}^{g_P} E[N_T(t, h_S, h_P)]$$

$$= \sum_{h_S=0}^{g_S-1} \sum_{h_P=0}^{g_P} \mathbb{I}_{\{t_T(h_S,h_P,0) \le t < t_T(h_S,h_P,1)\}} E[N_T(t_T(h_S, h_P, 0), h_S, h_P)]$$

$$= \sum_{h_P=0}^{g_P-1} \frac{\Delta t_T}{\Delta t_S} \lambda[T|P]\lambda[P|P]^{h_P}\lambda[P|S]N_S^H + \frac{\Delta t_T}{\Delta t_S}\lambda[P|P]^{g_P}\lambda[P|S]N_S^H$$

$$= \begin{cases} \frac{\Delta t_T}{\Delta t_S}\lambda[P|S]N_S^H\lambda[T|P] & \text{if } \lambda[P|P] = 0 \\ \frac{\Delta t_T}{\Delta t_S}\lambda[P|S]N_S^H(g_P\lambda[T|P] + 1) & \text{if } \lambda[P|P] = 1 \\ \frac{\Delta t_T}{\Delta t_S}\lambda[P|S]N_S^H(\lambda[T|P]\frac{1-\lambda[P|P]^{g_P}}{1-\lambda[P|P]} + \lambda[P|P]^{g_P}) & \text{otherwise} \end{cases}$$

It is straightforward to derive conditions for progenitor and terminal cell parameters that satisfies homeostatic conditions $E[N_P(t)] = N_P^H$ and $E[N_T(t)] = N_T^H$. In cases where we do not know specifically the number of progenitor/terminal cells each in non-stem cells, we can use $E[N_P(t)] + E[N_T(t)] = N - N_S^H$ which will result in weaker constraints on parameters. Normally we assume some fixed parameters and tune others; for example, we can fix progenitor cell lineage length g_P and cell division probabilities and compute Δt_P and Δt_T that satisfies homeostasis.

4.3 Mutation Acquisition with different driver mutations

In our previous models, the state space was characterized by the number of mutations acquired $\{0,1,2 \dots m\}$. This setting is simple and efficient but unable to model the ordering of mutation acquisition and incorporate mutation effects. Now we expand the state space to consider the acquisition of each driver mutation from a pool.

In our model we assume 5 different driver mutations $(M_1, M_2, M_3, M_4, M_5)$ that are crucial for cancer onset (in fact an arbitrary number of mutations can be added to this framework), then the state space (the status of mutation acquisition in a cell) can be represented as an array $\theta = \{\mathbb{I}_{M_1}, \mathbb{I}_{M_2}, \mathbb{I}_{M_3}, \mathbb{I}_{M_4}, \mathbb{I}_{M_5}\}$ where $\mathbb{I}_{M_i} = 1$ if M_i has been acquired by the current cell and $\mathbb{I}_{M_i} = 0$ otherwise. There are totally 32 different states in the state space: $\theta_0 = \{0,0,0,0,0\}, \theta_1 = \{0,0,0,0,1\} \dots \theta_{31} = \{1,1,1,1,1\}$. Note that this state space does not consider re-hits of the same mutation.

Now we describe how to compute the transition probabilities,

$$U_{xy} = P\left[\text{child cell have state } \theta_y \middle| \text{parent cell have state } \theta_x\right] x, y = 0 \dots 31$$

There are two steps necessary to compute all the transition probabilities from $\theta_x = \left\{\mathbb{I}_{M_1}^{(x)} \dots \mathbb{I}_{M_5}^{(x)}\right\}$ to $\theta_y = \left\{\mathbb{I}_{M_1}^{(y)} \dots \mathbb{I}_{M_5}^{(y)}\right\}$, $x, y = 0 \dots 31$. First, we define a helper "transition" probability, V_{xy} to be the conditional probability that a cell of state θ_x will ***acquire extra mutations*** *specified in* θ_y in its daughter cell during the division. Specifically,

$$V_{xy} = P\left[\text{child cell has state } \theta_z = \theta_x + \theta_y \middle| \text{parent cell have state } \theta_x\right] x, y = 0 \dots 31$$

where we define $\theta_x + \theta_y \stackrel{\text{def}}{=} \theta_x | \theta_y = \{\mathbb{I}_{M_1}^{(x)} \text{ or } \mathbb{I}_{M_1}^{(y)}, \dots \mathbb{I}_{M_5}^{(x)} \text{ or } \mathbb{I}_{M_5}^{(y)}\}$.

Let $r[M_i|\theta_x] = P[\text{child cell will acquire mutation } M_i | \text{parent cell has state } \theta_x]$, for $i = 1 \dots 5$ and $x = 0 \dots 31$. Normally $r[M_i|\theta_x]$ is called the mutation rate, and in our context, it can depend on the current cell state θ_x so that we can incorporate mutation effects on mutation rates later on.

Since we assume a cell acquires each individual mutation independently during a division, we have $V_{xy} = \prod_{i=1}^{5} [\mathbb{I}_{M_i}^{(y)} r[M_i|\theta_x] + \left(1 - \mathbb{I}_{M_i}^{(y)}\right)(1 - r[M_i|\theta_x])]$. For example, if $\theta_y = \{0, 0, 1, 0, 1\}$ then we have

$$V_{xy} = (1 - r[M_1|\theta_x])(1 - r[M_2|\theta_x])r[M_3|\theta_x](1 - r[M_4|\theta_x])r[M_5|\theta_x]$$

Figure 14 below provides an illustration.

f

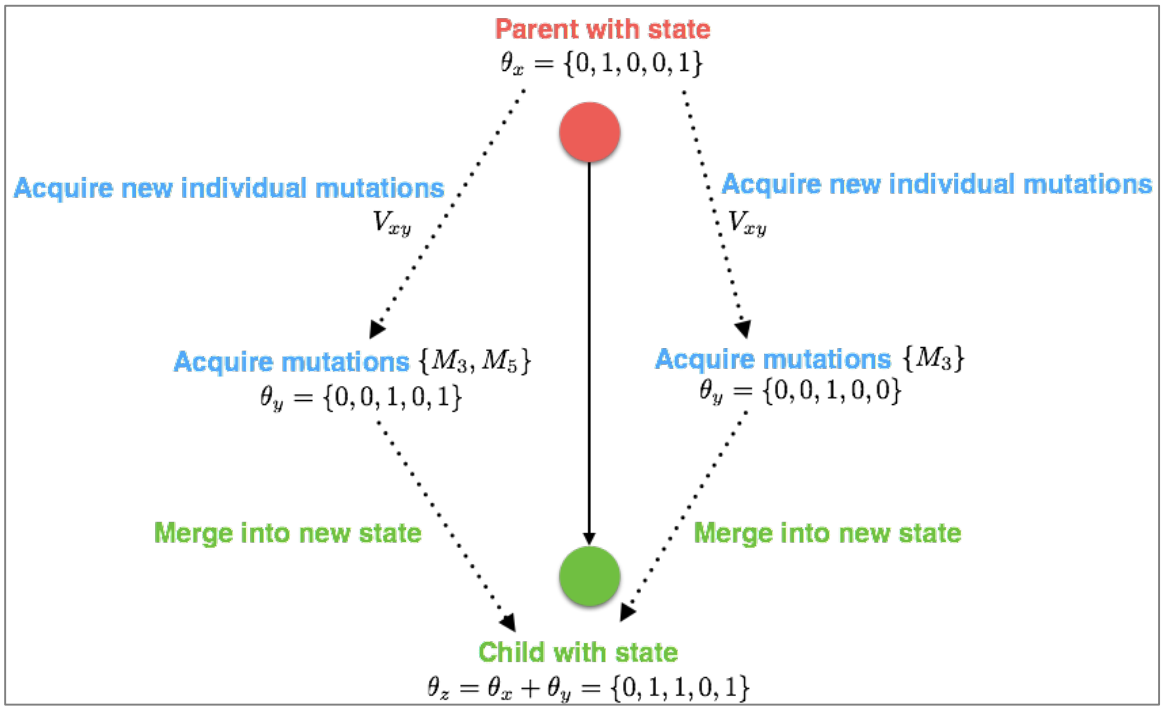


Figure 14: Illustration of the multi-mutation state space, where the mutation state of each cell is represented by a vector of length equals to the total number of driver mutations considered.

We now compute the transition probabilities U_{xz} with the help of V_{xy}. Since we do not consider re-hits, the already acquired mutations and new acquired mutations can be merged, by the bitwise OR operation as shown above, into a new state. As in Figure 14, there could be multiple "paths" from state θ_x to θ_z and the transition probability should sum over all the cases for the intermediate state θ_y.

Therefore we have:

$$U_{xz} = \mathbb{I}_{\{\theta_x \in \theta_z\}} \sum_{y:\theta_x + \theta_y = \theta_z} V_{xy}$$

We say $\theta_x \in \theta_z$ if $\theta_x + \theta_z = \theta_z$. This model allows us to consider mutation effects on mutation rate [1, 5, 11]. For example, if we let any cell that acquired M_4 to have an enlarged mutation rate, then we can specify $r[M_i|\theta_x] = \alpha r[M_i|\theta_0]$ if $\mathbb{I}_{M_4}^{(x)} = 1$, where $\alpha > 1$ is the factor of enlargement. In addition, we can specify the set of mutations required for cancer onset for stem, progenitor and terminal cells, for example, a stem cell could require (M_3, M_4, M_5) for cancer onset, while a progenitor cell may require extra mutation M_2, etc.

4.4 Extended theoretical lifetime intrinsic risk

In this section, we will develop a new theoretical lifetime cancer risk computation model, the extended TLIR. Comparing to our models in Chapters 2 and 3, this extended model will provide the most realistic computation for a general cell division structure based on current knowledge of carcinogenesis.

In fact, aside from the simulation models, very few discrete stochastic models in literature considered the general cell division structures for cancer risk computation, though in some cases they could provide close approximates. The model by Frank et al. [8] assumes a stem-progenitor division structure that is similar to our case, yet they did not derive precise solutions of their model; instead, they derived an approximate formula for cancer risk without featuring probability propagations. Also, their model is hard to generalize into the situation of dynamic cell evolution as stem and progenitor division may have different cycle length.

Little and Hendry [40] generalized both models from Wu et al. [34] and Frank et al. [8] by adding an additive extrinsic risk. They assumed that during each stem cell division, the mutation acquired in the daughter cell is either from extrinsic risk (mutagen-induced)

or intrinsic risk (spontaneous) with certain extrinsic/intrinsic mutation rates. However, this is not realistic as the intrinsic and extrinsic risk factors may interact, and thus, their approach ignores such interactions. In contrast, in our model we first compute the intrinsic risk caused by the intrinsic mutations alone and then the non-intrinsic risk (including extrinsic risk and risk due to interactions between intrinsic and extrinsic factors) as the remainder of the observed risk subtract the intrinsic risk. We also able to quantify the contributions of intrinsic factors to total cancer mutations based on the expected accumulated number of mutations within cancer cells, which will be discussed in detail in Section 4.6.

Though the extended model follows the same probability propagation mechanism inherently as the previous models, it provides a different way to compute the theoretical risk (extended risk). The extended risk is calculated based on more realistic dependency structures within cell divisions. We now demonstrate the major difference between the extended model and the original stem cell model, in the risk computation algorithms. First, we derive a more exact final step to risk computation for our original stochastic cancer stem cell model.

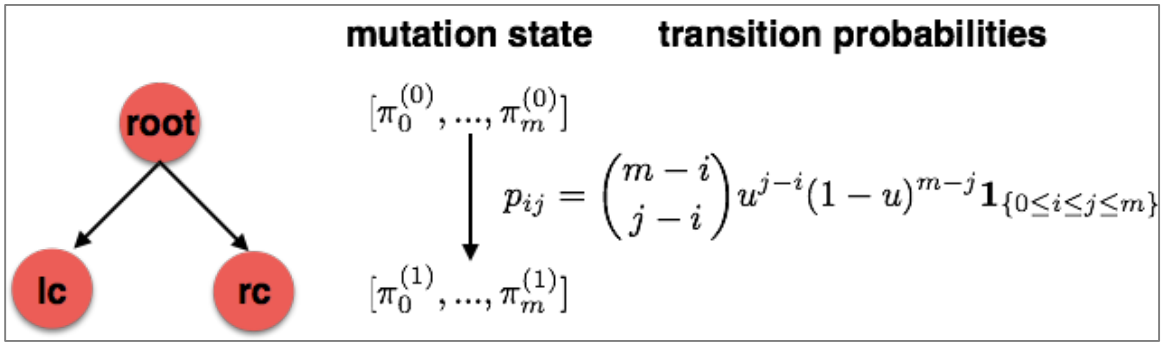


Figure 15: Illustration of one general cell division, from Generation 0 to Generation 1. Parent cell ($root$) divides into two daughter cells, left child (lc) and right child (rc). The number of mutations (mutation state) on each cell is possibly from 0 to m. We use $\pi_i^{(\cdot)}$ to represent the probability that the cell at the corresponding generation has i mutations. Given the mutation state in parent cell ($root$),

each daughter cells will inherit all mutations in parent, and will independently acquire extra mutations according to the transition probability p_{ij}.

We present a simple diagram in Figure 15 to illustrate a simple cell division, a parent cell ($root$) divides into two daughter cells, left child (lc) and right child (rc). The transition rule in Figure 15 is the fundamental assumption for probability propagation in our original stem cell model, intermediate model as well as exact risk model inherently. According to our original stem cell model in Section 2.5.1 the theoretical risk is

$$tLIR = 1 - \left[1 - P[X_1 = m]\right]^2 = 1 - \left[1 - \pi_m^{(1)}\right]^2$$

We denote the probability that no cancer cell occurs in the structure in Figure 15, according to original stem cell model is:

$$\text{Prob}_{\text{original}}(\text{no cancer}) = \left[1 - \pi_m^{(1)}\right]^2 = \left[\sum_{i=0}^{m-1} \pi_i^{(1)}\right]^2 = P[X_{lc} < m]P[X_{rc} < m]$$

according to our probability propagation rules, where $X.$ denotes the mutation state $(0,1,\cdots,m)$ in the corresponding cell. Note that since $\pi_i^{(1)} = \sum_{k=0}^{i} \pi_k^{(0)} p_{ki}$, therefore

$$\text{Prob}_{\text{original}}(\text{no cancer}) = \left[\sum_{i=0}^{m-1}\sum_{k=0}^{i} \pi_k^{(0)} p_{ki}\right]^2 = \left[\sum_{k=0}^{m-1}\sum_{i=k}^{m-1} \pi_k^{(0)} p_{ki}\right]^2$$

$$= \left[\sum_{k=0}^{m-1} \pi_k^{(0)} \sum_{i=k}^{m-1} p_{ki}\right]^2 = \left[\sum_{k=0}^{m-1} \pi_k^{(0)}(1 - p_{km})\right]^2$$

Note that $\sum_{i=k}^{m-1} p_{ki} = 1 - p_{km}$ and $1 - p_{km} = 0$ for $k = m$; so, we can add the term $\pi_m^{(0)}(1 - p_{mm})$ above:

$$\text{Prob}_{\text{original}}(\text{no cancer}) = \left[\sum_{k=0}^{m} \pi_k^{(0)}(1 - p_{km})\right]^2$$

On the other hand, we compute the above probability from scratch and obtain $\text{Prob}_{\text{exact}}(\text{no cancer})$. In fact, due to the inheritance property of mutation acquisition:

$$\begin{aligned}\text{Prob}_{\text{exact}}(\text{no cancer}) &= P[X_{lc} < m \cap X_{rc} < m \cap X_{root} < m] \\ &= P[X_{root} < m | X_{lc} < m \cap X_{rc} < m] P[X_{lc} < m \cap X_{rc} < m] \\ &= P[X_{lc} < m \cap X_{rc} < m]\end{aligned}$$

Then

$$\text{Prob}_{\text{exact}}(\text{no cancer}) = P[X_{lc} < m \cap X_{rc} < m] = \sum_{i=0}^{m-1} \sum_{j=0}^{m-1} P[X_{lc} = i \cap X_{rc} = j]$$

$$= \sum_{i=0}^{m-1} \sum_{j=0}^{m-1} \sum_{k=0}^{\min\{i,j\}} P[X_{lc} = i \cap X_{rc} = j | X_{root} = k] P[X_{root} = k]$$

$$= \sum_{i=0}^{m-1} \sum_{j=0}^{m-1} \sum_{k=0}^{\min\{i,j\}} P[X_{lc} = i | X_{root} = k] P[X_{lc} = j | X_{root} = k] P[X_{root}$$

$$= k] = \sum_{i=0}^{m-1} \sum_{j=0}^{m-1} \sum_{k=0}^{\min\{i,j\}} p_{ki} p_{kj} \pi_k^{(0)} = \sum_{k=0}^{m-1} \sum_{i=k}^{m-1} \sum_{j=k}^{m-1} \pi_k^{(0)} p_{ki} p_{kj}$$

$$= \sum_{k=0}^{m-1} \pi_k^{(0)} \sum_{i=k}^{m-1} p_{ki} \sum_{j=k}^{m-1} p_{kj} = \sum_{k=0}^{m-1} \pi_k^{(0)} (1 - p_{km})^2$$

$$= \sum_{k=0}^{m-1} \pi_k^{(0)} (1 - p_{km})^2 + 0 = \sum_{k=0}^{m-1} \pi_k^{(0)} (1 - p_{km})^2 + \pi_m^{(0)} (1 - p_{mm})^2$$

$$= \sum_{k=0}^{m} \pi_k^{(0)} (1 - p_{km})^2$$

We now use Jensen's inequality to prove that

$$\text{Prob}_{\text{original}}(\text{no cancer}) \leq \text{Prob}_{\text{exact}}(\text{no cancer})$$

Proof:

Jensen's inequality states that for a real convex function ϕ, numbers $x_1, x_2, \cdots, x_n$ in its domain, and positive weights a_i such that $\sum a_i = 1$, then we have

$$\phi\left(\sum a_i x_i\right) \leq \sum a_i \phi(x_i)$$

The equality holds if $a_i = 1_{\{i=k\}}$ for some k (special case).

Let $\phi(x) = x^2, x \in [0,1]$ be the quadratic function and it is obviously convex; $a_i = \pi_i^{(0)}$ and $x_i = (1 - p_{im})$ for $i = 0, \cdots, m$; since $\sum a_i = \sum_{i=0}^{m} \pi_i^{(0)} = 1$, we must have

$$\left[\sum_{k=0}^{m} \pi_k^{(0)}(1 - p_{km})\right]^2 \leq \sum_{k=0}^{m} \pi_k^{(0)} (1 - p_{km})^2$$

Therefore

$$\text{Prob}_{\text{original}}(\text{no cancer}) \leq \text{Prob}_{\text{exact}}(\text{no cancer})$$

Proof Done.

A natural conclusion from this inequality is that theoretical risk estimated from original stem cell model, and thus with intermediate risk model, will be larger than the exact theoretical risk,

$$tLIR_{\text{original}} = 1 - \text{Prob}_{\text{original}}(\text{no cancer}) \geq 1 - \text{Prob}_{\text{exact}}(\text{no cancer}) = tLIR_{\text{exact}}$$

If the initial stem cell has no mutation, i.e. $\pi_k^{(0)} = 1_{\{k=0\}}$ then we have $tLIR_{\text{original}} = tLIR_{\text{exact}}$ up to the second generation according to the special case of Jensen's inequality. However, beyond the second generation, $\pi_k^{(0)} > 0$ for each $k = 0, \cdots, m$ according to probability transitions; therefore, we will always have $tLIR_{\text{original}} > tLIR_{\text{exact}}$.

In other words, the algorithm to compute cancer risk in our original models, though simple, would tend to over-estimate the exact theoretical risk. Alternatively, if we

compute the risk exactly in original models, we would have to use mutation state probabilities in each stem cell generation from the beginning, due to the parent-children conditional dependency, during symmetric divisions; which would be much more complicated and computational expensive. In the following, we provide a succinct and efficient algorithm to compute the extended theoretical risk, based on recursion structure in binary tree and dynamic programming.

4.4.1 Extended Risk Computation

We first define a generic cell division structure and then develop a recursive formula that will be the fundamental part for theoretical cancer risk computation. Assume a parent cell prt give birth to two daughter cells: left child lc and right child rc. Note that prt could be a stem or progenitor cell and lc, rc could each be stem, progenitor or terminal cells according to general division patterns. In cases where there is only one daughter cell, we set lc to be the only child and mark $rc = Nil$.

Now we define a conditional probability

$$p_{root}^{nc}(\theta) \stackrel{\text{def}}{=} P[\text{no cancer in the subtree rooted at root, } Tr_{root} | \text{root state is } \theta]$$

See Figure 16 below for an explanation.

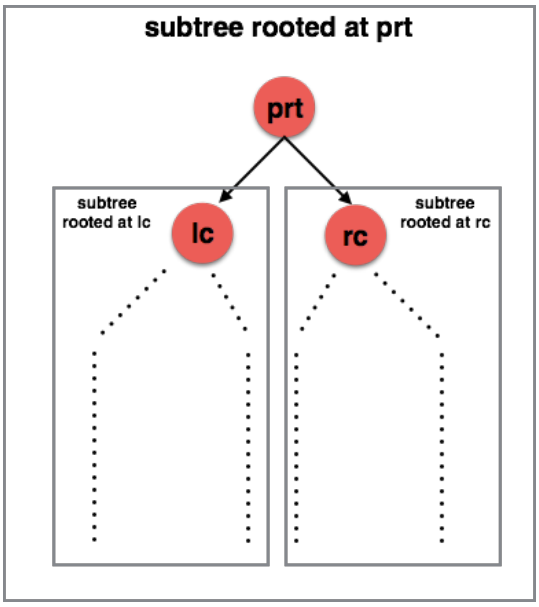


Figure 16: Illustration of binary tree structure in cell division

Note that the parent cell can be any cell in the entire cell lineage and unless lc/rc are leaves, they both are again roots of some sub-trees. Then $p_{prt}^{nc}(\theta)$ is the probability that there will be no cancer cells in prt with state θ along with all descendants of prt and likewise for $p_{lc}^{nc}(\theta)$ and $p_{rc}^{nc}(\theta)$. Note that $p_{Nil}^{nc}(\theta) \equiv 1$ for any θ.

We now develop a recursion that computes $p_{prt}^{nc}(\theta_x)$ from $p_{lc}^{nc}(\theta_x)$ and $p_{rc}^{nc}(\theta_x)$ for $x = 0 \dots 31$. Let $\theta_{\text{cancer}}^{prt}, \theta_{\text{cancer}}^{lc}, \theta_{\text{cancer}}^{rc}$ be the combinations of mutation hits required for initial cancer onset for cells prt, lc and rc.

Note that

$p_{prt}^{nc}(\theta_x) = P\left[\text{no cancer in } Tr_{prt} \middle| prt \text{ state} = \theta_x\right]$

$= P[\{prt \text{ is not a cancer cell}\} \textbf{ and } \{\text{no cancer in } Tr_{lc} \text{ and } Tr_{rc}\} | prt \text{ state} = \theta_x]$

$= P[prt \text{ is not a cancer cell} | prt \text{ state} = \theta_x]$

$\times P[\text{no cancer in } Tr_{lc} | \{prt \text{ is not a cancer cell}\} \text{ and } \{prt \text{ state} = \theta_x\}]$

$\times P[\text{no cancer in } Tr_{rc} | \{prt \text{ is not a cancer cell}\} \text{ and } \{prt \text{ state} = \theta_x\}]$

$= \mathbb{I}_{\left\{\theta_{\text{cancer}}^{prt} \notin \theta_x\right\}} \times P[\text{no cancer in } Tr_{lc} | prt \text{ state} = \theta_x] \times P[\text{no cancer in } Tr_{rc} | prt \text{ state} = \theta_x]$

Now

$$P[\text{no cancer in } Tr_{lc} | prt \text{ state} = \theta_x]$$

$$= \sum_{y=0}^{31} P\left[\{\text{no cancer in } Tr_{lc}\} \text{ and } \{lc \text{ state} = \theta_y\} \middle| prt \text{ state} = \theta_x\right]$$

$$= \sum_{y=0}^{31} P\left[\text{no cancer in } Tr_{lc} \middle| \{lc \text{ state} = \theta_y\} \text{ and } \{prt \text{ state} = \theta_x\}\right]$$

$$\times P\left[lc \text{ state} = \theta_y \middle| prt \text{ state} = \theta_x\right]$$

$$= \sum_{y=0}^{31} P\left[\text{no cancer in } Tr_{lc} \middle| lc \text{ state} = \theta_y\right] U_{xy}^{prt \to lc}$$

With similar arguments for the right child, we now have the recursion:

$$P\left[\text{no cancer in } Tr_{prt} \middle| prt \text{ state} = \theta_x\right]$$

$$= \mathbb{I}_{\left\{\theta_{\text{cancer}}^{prt} \notin \theta_x\right\}} \times \left(\sum_{y=0}^{31} P\left[\text{no cancer in } Tr_{lc} \middle| lc \text{ state} = \theta_y\right] U_{xy}^{prt \to lc} \right)$$

$$\times \left(\sum_{y=0}^{31} P\left[\text{no cancer in } Tr_{rc} \middle| rc \text{ state} = \theta_y\right] U_{xy}^{prt \to rc} \right)$$

i.e.

$$p_{prt}^{nc}(\theta_x) = \mathbb{I}_{\left\{\theta_{\text{cancer}}^{prt} \notin \theta_x\right\}} \times \left(\sum_{y=0}^{31} p_{lc}^{nc}(\theta_y) U_{xy}^{prt \to lc}\right) \times \left(\sum_{y=0}^{31} p_{rc}^{nc}(\theta_y) U_{xy}^{prt \to rc}\right)$$

Next, we will extend this generic recursion to include all cell division patterns.

The above formula is only useful when we know exactly the types of lc and rc, yet due to the stochastic nature of cell divisions, we have to consider all possible division patterns. We consider an extended conditional probability that no cancer cell onset will take place in subtree Tr_{root} rooted at cell $root$ given the cell state θ and cell behavior $b \in \{r1, r2, d1, d2, rd, D\}$:

$p_{root}^{nc}(\theta, b)$

$\stackrel{\text{def}}{=} P[\text{no cancer in the subtree rooted at root, } Tr_{root} | \text{root state is } \theta, root \text{ cell behavior is } b]$

We can compute this probability using the arguments above since a fixed cell behavior b determines the daughter cells, $lc(root, b)$ and $rc(root, b)$. For example, $lc(S, r2) = rc(S, r2) = S$, $lc(S, rd) = S$ and $rc(S, rd) = P$, etc. If $b = D$(death), then both lc and rc will be Nil. Therefore, we have

$$p_{prt}^{nc}(\theta_x, b) = \mathbb{I}_{\left\{\theta_{\text{cancer}}^{prt} \notin \theta_x\right\}} \times \left(\sum_{y=0}^{31} p_{lc(prt,b)}^{nc}(\theta_y) U_{xy}^{prt \to lc(prt,b)}\right)$$

$$\times \left(\sum_{y=0}^{31} p_{rc(prt,b)}^{nc}(\theta_y) U_{xy}^{prt \to rc(prt,b)}\right)$$

Then we can compute the general conditional probability considering all possible divisions:

$p_{root}^{nc}(\theta) \overset{\text{def}}{=} P[\text{no cancer in } Tr_{root} | \text{root state is } \theta]$

$$= \sum_{b\in\{r1,r2,d1,d2,rd,D\}} P[\{\text{no cancer in } Tr_{root}\} \; and \; \{root \text{ cell behavior is } b\} | \text{root state is } \theta]$$

$$= \sum_{b\in\{r1,r2,d1,d2,rd,D\}} P[\text{no cancer in } Tr_{root} | \{root \text{ cell behavior is } b\} \text{ and } \{\text{root state is } \theta\}]$$

$$\times P[root \text{ cell behavior is } b | \text{root state is } \theta] = \sum_{b\in\{r1,r2,d1,d2,rd,D\}} p_{root}^{nc}(\theta, b)\, P_b^{root}(\theta)$$

Where $P_b^{root}(\theta)$ is the probability that cell $root$ has behavior b given mutation state θ. This dependency on θ is useful, as we will see later, to model the mutation effects on cell dynamics. Combine the above arguments, we have

$$p_{root}^{nc}(\theta_x) = \sum_{b\in\{r1,r2,d1,d2,rd,D\}} P_b^{root}(\theta_x) \mathbb{I}_{\left\{\theta_{\text{cancer}}^{prt} \notin \theta_x\right\}} \times \left(\sum_{y=0}^{31} p_{lc(prt,b)}^{nc}(\theta_y) U_{xy}^{prt\to lc(prt,b)} \right)$$

$$\times \left(\sum_{y=0}^{31} p_{rc(prt,b)}^{nc}(\theta_y) U_{xy}^{prt\to rc(prt,b)} \right)$$

Now we show that the formula is a fundamental part to compute exact lifetime cancer risk.

Suppose a tissue starts from a single stem cell with no mutations, then the lifetime cancer risk $R = 1 - p_S^{nc}(\theta_0, h_s = 0)$ where $\theta_0 = \{0,0,0,0,0\}$ and h_s denotes the stem cell generation. Upon first division, with cell division probabilities $\{P_b^S(t_S(1), \theta_0)\}$ (note cell division probabilities can depend both on time and mutations state), we have

$$p_S^{nc}(\theta_0, h_s = 0)$$

$$= \sum_{b\in\{r1,r2,d1,d2,rd,D\}} \{P_b^S(h_s = 1, \theta_0)\}\mathbb{I}_{\{\theta^S_{\text{cancer}}\notin\theta_0\}}$$

$$\times\left(\sum_{y=0}^{31} p_{lc(S,b)}^{nc}(\theta_y)U_{0y}^{S\to lc(S,b)}\right)\times\left(\sum_{y=0}^{31} p_{rc(S,b)}^{nc}(\theta_y)U_{0y}^{S\to rc(S,b)}\right)$$

within the left and right child lc and rc throughout all division patterns there are only two types of cells, stem cell (S) and progenitor cell (P) as a result of renewal and differentiation respectively. Therefore we only need to compute $p_S^{nc}(\theta_0, h_s = 1)$ and $p_P^{nc}(\theta_0, h_P = 0)$ where h_P is the progenitor generation along its lineage.

For progenitor cell, we have

$$p_P^{nc}(\theta_x, h_P = 0)$$

$$= \sum_{b\in\{r1,r2,d1,d2,rd,D\}} P_b^P(h_P = 0, \theta_x)\mathbb{I}_{\{\theta^P_{\text{cancer}}\notin\theta_x\}}$$

$$\times\left(\sum_{y=0}^{31} p_{lc(P,b)}^{nc}(\theta_y)U_{xy}^{P\to lc(P,b)}\right)\times\left(\sum_{y=0}^{31} p_{rc(P,b)}^{nc}(\theta_y)U_{xy}^{P\to rc(P,b)}\right)$$

for $x = 0 \ldots 31$

Similarly, we only need $p_P^{nc}(\theta_x, h_P = 1)$ and $p_T^{nc}(\theta_x)$ for $x = 0 \ldots 31$. Note that for terminal cells $p_T^{nc}(\theta) = \mathbb{I}_{\{\theta^T_{\text{cancer}}\notin\theta\}}$.

See Figure 17 below, since we know the "leaf probabilities" such as $p_T^{nc}(\theta)$ and final generation stem cell $p_S^{nc}(\theta_x, h_s = g_S) = \mathbb{I}_{\{\theta^S_{\text{cancer}}\notin\theta_x\}}$, we can propagate the probabilities bottom up, from leaf to root, with a dynamic programming approach.

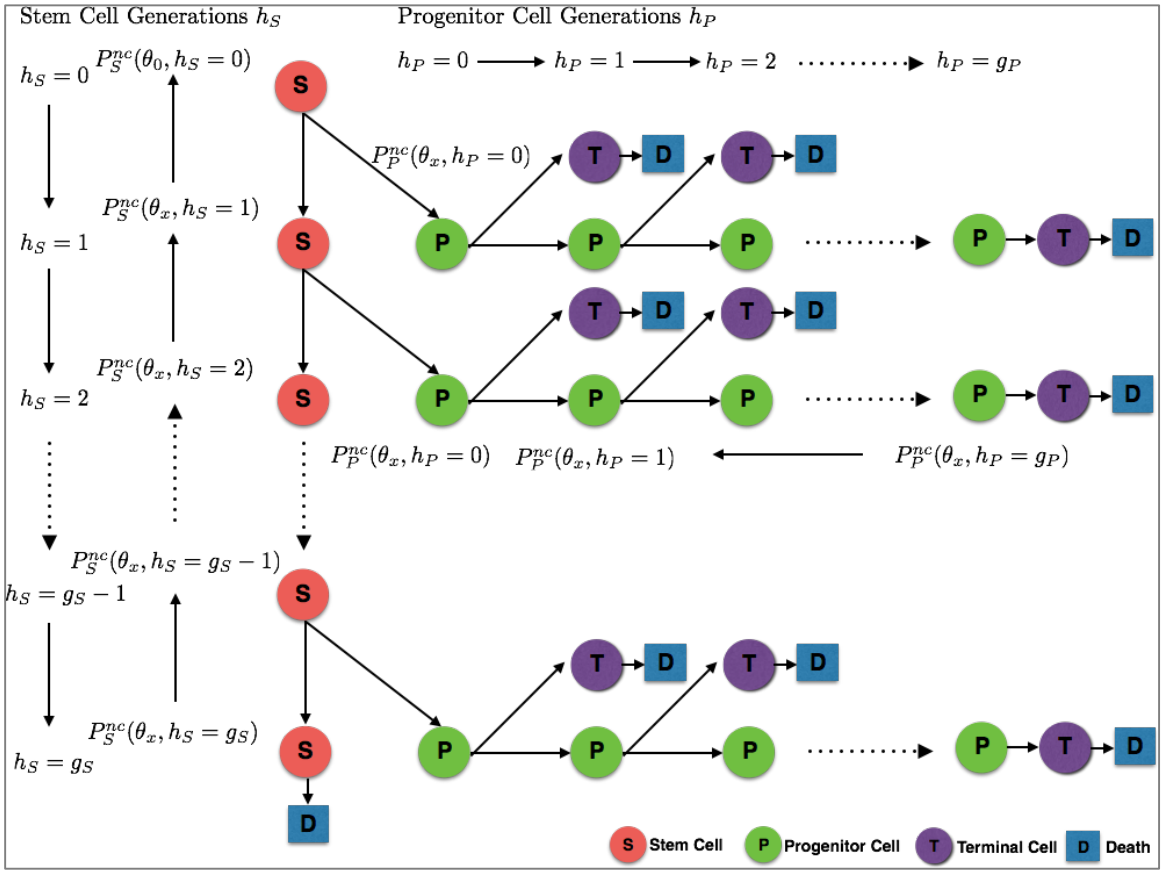


Figure 17: Probability recursion in lifetime general cell division structures.

4.4.2 Mutation Effects

Certain driver mutations can encourage the cell growth. In our model, $P_b^{cell}(t, \theta)$ specifies how cell division probabilities depend on time and mutation state. Suppose M_3 can promote cell growth through reducing death rate by a constant factor, then for any cell with mutation M_3, $P_{b=D}^{cell}(t, \theta | (0,0,1,0,0) \in \theta) = \alpha P_{b=D}^{cell}(t), 0 < \alpha < 1$ and all other probabilities should be renormalized, where $P_{b=D}^{cell}(t)$ is the base division probability independent of θ.

In addition, certain mutations can further increase the chance of acquiring additional mutations. For example, if M_4 increases mutation rates by a constant factor, then in our

model we set $r[M_i|\theta_x] = \alpha r[M_i],\ \alpha > 1$ and re-compute the transition matrix U_{xy} based on Section 4.3.

Other driver mutations could incur clonal expansion in which cells divide faster with a reduced cycle time. As this will require a fundamental extension in our model, we will dedicate the next session to clonal expansion.

4.4.3 Clonal Expansion and Regulation

With clonal expansion, some cell divides faster and therefore cell divisions will not stay on the same pace. In this section, we extend the recursive cancer risk computation in Section 4.4.1 to accommodate different cell cycles. Let M_5 be the mutation that causes cells divides faster than normal by a factor of $\alpha^{cell} > 1$, then on average during each division of normal cells (cells with no M_5 mutations), there are α^{cell} divisions for expansion cells (cells that acquired M_5). Figure 18 below provides an illustration for $\alpha^S = 2$.

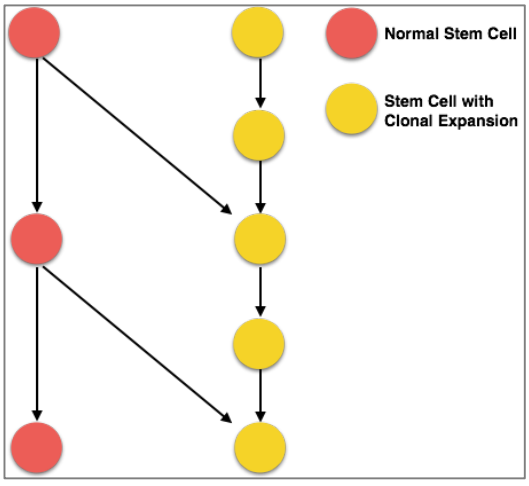


Figure 18: Illustration of clonal expansion of factor $\alpha^S = 2$ for stem cells. For each division of normal cells, expansion cells will have two divisions. Note that after each division of normal cells, a proportion of daughter cells will become expansion cells, due to stochastic mutation acquisition.

Now we develop cancer risk computation for integer expansion factor $\alpha > 1$. We define a function $f(\cdot)\colon [0,1]^{n_1} \to [0,1]^{n_2}$ that maps in general the $p^{nc}(\theta)$ in child generation to parent generation, with n_1 and n_2 to be the total number of p^{nc} considering all possible cell type and mutation combinations. For a stem parent cell, we have $n_1 = 32\times 2$ and $n_2 = 32$ considering stem and progenitor daughter cells and all 32 mutation combinations. In general based on our model in Section 4.4.1, $p^{nc}_{parent}(\cdot) = f(p^{nc}_{children}(\cdot))$ in that

$$p^{nc}_{prt}(\theta_x) = \sum_{b\in\{r1,r2,d1,d2,rd,D\}} P^{prt}_b(\theta_x)\mathbb{I}_{\left\{\theta^{prt}_{\text{cancer}}\notin\theta_x\right\}}\times\left(\sum_{y=0}^{31} p^{nc}_{lc(prt,b)}(\theta_y)U^{prt\to lc(prt,b)}_{xy}\right)$$
$$\times\left(\sum_{y=0}^{31} p^{nc}_{rc(prt,b)}(\theta_y)U^{prt\to rc(prt,b)}_{xy}\right)$$

for each θ_x.

For clonal expansion, we separate parent cells into two classes, one with clonal expansion inducing mutation M_5 and the other one not. Let α be the integer expansion factor and $p^{nc}_{children}(\cdot)$ be the probabilities of children cells in any normal generation. Note that the cells in $children$ generation includes both normal cells and expansion cells as normal cells will give birth to expansion cells with mutation acquisition upon the completion of the cell cycle, corresponding to α divisions of expansion cells. We update each class of cells separately and then merge them together. For expansion cells, $p^{nc}_{parent}\left(\{\theta\}|_{\{M_5\in\theta\}}\right) = f^{(\alpha)}(p^{nc}_{children}(\cdot))$ that applies $f(\cdot)$ α times on $p^{nc}_{children}(\cdot)$. Note that only the expansion part of $children$ cells actively participate in the computation as

$U_{xy} \equiv 0$ for any $\{\theta_x\}|_{\{M_5 \in \theta_x\}}$ and $\{\theta_y\}|_{\{M_5 \notin \theta_y\}}$. For normal cells, we simply have $p^{nc}_{parent}\left(\{\theta\}|_{\{M_5 \notin \theta\}}\right) = f(p^{nc}_{children}(\cdot))$.

For arbitrary expansion factor $\alpha > 1$, we estimate the process by taking the lower and upper closest integers of α with probability p and $(1-p)$ respectively, where $p = \lceil \alpha \rceil - \alpha$.

4.4.4 Age Dependent Cancer Risk

Our model can estimate cancer risk at any age (0 - 80 years) in a lifetime. Let t be any time (in days) from 0 to the end of 80 years ($80 \times 365 = 29200$ days), we can determine the maximum number of stem cell divisions along with progenitor cell divisions for each branch, up to time t.

In Section 4.1, we have determined the time point for each stem/progenitor generation, $t_S(h_S)$, $t_P(h_S, h_P)$ and the birth/death time for terminal cells, $t_T(h_S, h_P, 0)$ and $t_T(h_S, h_P, 1)$. Given time t, we can first determine the stem cell generation $h_S^\star$ such that $t_S(h_S^\star) \leq t < t_S(h_S^\star + 1)$, then determine the progenitor cell generations $h_P^\star(h_S)$ on branches $h_S = h_S^\star$, $h_S^\star - 1$, $\cdots$ such that $t_P(h_S, h_P^\star(h_S)) \leq t < t_P(h_S, h_P^\star(h_S) + 1)$. With these approaches, we could determine the entire cell division structure up to any given time t and then compute the cancer risk with our recursive formula in Section 4.4.1, with the last generation stem, progenitor and terminal cells treated as leaf cells.

4.5 General Mutation State Evolution

In Section 4.2 we described cell number evolution and homeostatic conditions without differentiating cells by their mutation state. In this section we build algorithms to

compute cell numbers of each mutation state at an arbitrary time t. In fact, similar to the arguments in Section 4.2, we basically need to model the state transition from one cell generation to the next; then cell numbers at any time t can be computed by aggregating multiple piecewise constant quantities. We first build state transition in Section 4.5.1 and later develop algorithms to include clonal expansion.

4.5.1 Cell Number Transition

We categorize cell division patterns into two classes: "renewal", where the daughter cell has the same type (stem or progenitor) with its parent cell, and "differentiation", where the daughter cell is the differentiated version of its parent. For example, stem cell can differentiate to progenitor cells, which can further differentiate into terminal cells.

Let $N_S(t(h_S), \theta_x)$ be the number of stem cells at generation h_S with mutation state θ_x, then their children at the next generation $h_S + 1$ could be stem or progenitor cells with several different mutation states. We model this process by two consecutive multinomial distribution samplings.

First

$$\{n_S(t_S(h_S), \theta_x, b)\}|_{b\in\{r1,r2,d1,d2,rd,D\}}$$

$$\sim \text{Multinomial}\big(N_S(t_S(h_S), \theta_x), \{P_b^S\big(t_S(h_S)\big)\}|_{b\in\{r1,r2,d1,d2,rd,D\}}\big)$$

where $n_S(t_S(h_S), \theta_x, b)$ is the cells among $N_S(t_S(h_S), \theta_x)$ that have action b. Then for the next generation $h_S + 1$, the number of daughter stem cells, including all possible mutation states, as a result of self-renewal from $N_S(t_S(h_S), \theta_x)$ is:

$$N_S(t_S(h_S + 1), \theta.|\theta_x) = 2n_S(t_S(h_S), \theta_x, r2) + n_S(t_S(h_S), \theta_x, r1) + n_S(t_S(h_S), \theta_x, rd)$$

and similarly, the number of daughter progenitor cells, as a result of differentiation is

$$N_P(t_P(h_S, 0), \theta.|\theta_x) = 2n_S(t_S(h_S), \theta_x, d2) + n_S(t_S(h_S), \theta_x, d1) + n_S(t_S(h_S), \theta_x, rd)$$

Secondly,

$$\{N_S\big(t_S(h_S+1),\theta_y|\theta_x\big)\}|_{y=0,\cdots,31} \sim \text{Multinomial}\big(N_S(t_S(h_S+1),\theta_\cdot|\theta_x),U_{xy}|_{y=0,\cdots,31}\big)$$

where $N_S\big(t_S(h_S+1),\theta_y|\theta_x\big)$ is the number of stem cells at generation h_S+1 with mutation state θ_y as a result of self-renewal from $N_S(t_S(h_S),\theta_x)$, and U_{xy} is the transition probability from θ_x to θ_y. Similarly,

$$\{N_P\big(t_P(h_S,0),\theta_y|\theta_x\big)\}|_{y=0,\cdots,31} \sim \text{Multinomial}\big(N_P(t_P(h_S,0),\theta_\cdot|\theta_x),U_{xy}|_{y=0,\cdots,31}\big)$$

Finally, we have

$$N_S(t_S(h_S+1),\theta_x)=\sum_{z=0}^{31}N_S(t_S(h_S+1),\theta_x|\theta_z)$$

and

$$N_P(t_P(h_S,0),\theta_x)=\sum_{z=0}^{31}N_P(t_P(h_S,0),\theta_x|\theta_z)$$

We can use the same arguments for progenitor cells at each branch and therefore we omit its derivation here.

It is not hard to derive the transition in terms of expected cell numbers,

$$E[N_S(t_S(h_S+1),\theta_x)]=\sum_{z=0}^{31}E[N_S(t_S(h_S+1),\theta_x|\theta_z)=\sum_{z=0}^{31}E[N_S(t_S(h_S+1),\theta_\cdot|\theta_z)]U_{zx}$$

$$=\sum_{z=0}^{31}E[2n_S(t_S(h_S),\theta_z,r2)+n_S(t_S(h_S),\theta_z,r1)+n_S(t_S(h_S),\theta_z,rd)]U_{zx}$$

$$=\sum_{z=0}^{31}E[N_S(t_S(h_S),\theta_z)](2P_{r2}^S\big(t_S(h_S)\big)+P_{r1}^S\big(t_S(h_S)\big)+P_{rd}^S\big(t_S(h_S)\big))U_{zx}$$

Similarly,

$E[N_P(t_P(h_S,0),\theta_x)]$

$$= \sum_{z=0}^{31} E[N_S(t_S(h_S),\theta_z)](2P_{d2}^S(t_S(h_S)) + P_{d1}^S(t_S(h_S)) + P_{rd}^S(t_S(h_S)))U_{zx}$$

And for progenitor cell parents, we have for an arbitrary progenitor lineage branch h_S:

$$E[N_P(t_P(h_S,h_p+1),\theta_x)]$$

$$= \sum_{z=0}^{31} E[N_P(t_P(h_S,h_p),\theta_z)](2P_{r2}^P(t_P(h_S,h_P)) + P_{r1}^P(t_P(h_S,h_P))$$

$$+ P_{rd}^P(t_P(h_S,h_P)))U_{zx}$$

for terminal cell children:

$$E[N_T(t_T(h_S,h_p,0),\theta_x)]$$

$$= \sum_{z=0}^{31} E[N_P(t_P(h_S,h_p),\theta_z)](2P_{d2}^P(t_P(h_S,h_P)) + P_{d1}^P(t_P(h_S,h_P))$$

$$+ P_{rd}^P(t_P(h_S,h_P)))U_{zx}$$

Using the same arguments in Section 4.2, we can obtain cell numbers at an arbitrary time in a lifespan.

4.5.2 Clonal Expansion

Clonal expansion could break homeostasis with cell numbers continue to increase until regulations take place. Cells with certain mutations that cause clonal expansion would divide in a faster pace. Similar to the discussion in Section 4.4.3, assuming M_5 is the mutation that causes cells to divide faster than normal by a factor of $\alpha^{cell} > 1$, we can model the cell number transition separately for cells with or without M_5, according to the diagram presented in Figure 19.

Again, we define a mapping $f_{A\to B}(\cdot)\colon [0,\infty]^{32} \mapsto [0,\infty]^{32}$ for the cell number transition in 4.5.1 from parent cell A to children cell B with all 32 mutation states. For example, we have $f_{S\to S}(E[N_S(t_S(h_S),\theta_\cdot)]) = E[N_S(t_S(h_S+1),\theta_\cdot)]$ following the equations in Section 4.5.1. We discuss stem cell transitions first with progenitor lineages following similar arguments.

From generation h_S to h_S+1, given an integer clonal expansion factor α, cells with M_5 divide α times on average, at time points $t_S(h_S(i)) \triangleq t_S(h_S) + i\frac{t_S(h_S+1)-t_S(h_S)}{\alpha}, i = 0,\cdots,\alpha-1$ assuming equal cell cycle time. At h_S+1, cells with M_5: $N_S(t_S(h_S+1),\theta|M_5\in\theta)$ are the combinations of self-renewal from both normal cells at $t(h_S)$: $N_S(t_S(h_S),\theta|M_5\notin\theta)$ and clonal expansion cells at $t_S(h_S(\alpha-1))$: $N_S(t_S(h_S(\alpha-1)),\theta|M_5\in\theta)$. Therefore, $E[N_S(t_S(h_S+1),\theta|M_5\in\theta)] = f_{S\to S}(E[N_S(t_S(h_S),\theta|M_5\notin\theta)]) + f_{S\to S}^{\alpha}(E[N_S(t_S(h_S),\theta|M_5\in\theta)])$ where $f_{S\to S}^{\alpha}(\cdot)$ indicates applying the mapping $f_{S\to S}(\cdot)$ α times.

It is not hard to include progenitor cell branches independently from each stem cell (sub) generations. For example, we can obtain the first-generation progenitor cells:

$E[N_P(t_P(h_S(i),0),\theta_\cdot)] = f_{S\to P}(E[N_S(t_S(h_S(i)),\theta|M_5\in\theta)])$ for $i = 0,\cdots,\alpha-2$.

Note that progenitor cells originated from stem sub-generation $h_S(\alpha-1)$ need to be merged with those originated from normal stem cells at h_S, so

$E[N_P(t_P(h_S,0),\theta_\cdot)] = f_{S\to P}(E[N_S(t_S(h_S),\theta|M_5\notin\theta)]) + f_{S\to P}(E[N_S(t_S(h_S(\alpha-1)),\theta|M_5\in\theta)])$

Then we can use the same arguments for the evolution within each progenitor branch and aggregate the cell numbers at each (sub) generation to an arbitrary time t.

4.6 The Contributions of Intrinsic and Extrinsic Factors

According to Tomasetti and Vogelstein [31] and their more recent work in 2017 [41], significant cancer risk was due to intrinsic mutation acquisition while extrinsic factors only have limited contributions to the cancer onset. However, according to our original model, random mutation acquisition only accounts for a small portion of total cancer risk, while extrinsic factors (such as environmental changes, smoking, radiation exposure, etc.) are the major causes. Therefore, it is important to quantify the contribution of intrinsic factors to cancer risk.

We use the number of accumulated driver mutations as a metric to evaluate contributions to cancer risk since multi-stage mutation acquisition is the cause of cancer onset.

We assume a general lifespan of 80 years. At the end of age 80, we obtain the *expected* number of cells of each type and mutation state, $N_S(t,\theta_x), N_P(t,\theta_x), N_T(t,\theta_x)$ where $t = 80\times365$ and $x = 0\cdots31$ assuming 5 different driver mutations. For each mutation state θ_x, we can easily obtain the number of mutations acquired as $|\theta_x| \triangleq \sum_{i=0}^{4}\theta_x[i]$. For example, mutation state "01011" has 3 mutations acquired and "00011" has 2 mutations, etc. Therefore, the *expected* total number of mutations acquired is just

$M_{\text{total}} = \sum_{i=0}^{31}|\theta_x|(N_S(t,\theta_x) + N_P(t,\theta_x) + N_T(t,\theta_x))$.

If we only consider cancer cells, then

$M_{\text{cancer}} = \sum_{i=0}^{31}|\theta_x|(N_S(t,\theta_x)1_{\{\theta^S_{\text{cancer}}\in\theta_x\}} + N_P(t,\theta_x)1_{\{\theta^P_{\text{cancer}}\in\theta_x\}} + N_T(t,\theta_x)1_{\{\theta^T_{\text{cancer}}\in\theta_x\}})$.

Suppose we have an intrinsic mutation rate is u_{int} (normally between 10^{-9} to10^{-6}), then let $M(u_{\text{int}})$ be the expected number of acquired mutations due to intrinsic risk. If we can estimate a total mutation rate u_{total} based on observed risk, then we can get $M(u_{\text{total}})$, the expected number of mutations due to all factors. Then the ratio $C_{\text{int}} = \frac{M(u_{\text{int}})}{M(u_{\text{total}})}$ indicates

the fraction of cancer risk due to intrinsic factors. In our study, we will use both M_{total} and M_{cancer} to compute C_{int}. Now we need to estimate total mutation rate u_{total} given observed risk R_{obs}.

We also define the estimated excess mutation rate $u_{\text{exc}} = u_{\text{total}} - u_{\text{int}}$ to be the rate due to non-intrinsic factors.

4.6.1 Estimating Total Mutation Rate

We need to find mutation rate u_{total} such that the computed risk matches observed risk. It is not feasible to directly estimate u_{total} in our extended risk model due to non-linearity in transition matrix, mutation effects and clonal expansion. Therefore, we use our original stem cell model in Section 2.5.1 to provide an initial estimate.

In Section 2.5.1, we have a formula

$$P_n(m) = \sum_{i=0}^{m} [1 - (1-u)^n]^{m-i} P_0(i)$$

the probability that a stem cell at generation n has m mutations, where $P_n(m) = P[X_n = m]$. Note that here m mutations are required for a stem cell to become a cancer cell and m is the absorbing state in the Markov evolution for $P_n(\cdot)$. Let n_1 be the number of symmetric stem divisions and $n_2 = n - n_1$ is the number of asymmetric divisions. The theoretical lifetime cancer risk due to mutation rate u is $R(u) = 1 - (1 - P_n(m))^{2^{n_1}}$. Assuming we start with one stem cell with no mutations so $P_0(i) = 1_{\{i=0\}}$, and set $R(u) = R_{\text{obs}}$, then we have $P_n(m) = 1 - \left(1 - R(u)\right)^{2^{-n_1}} = [1 - (1-u)^n]^m$, therefore the estimated total mutation rate is

$$\hat{u}^0_{\text{total}} = 1 - [1 - (1 - (1 - R_{\text{obs}})^{2^{-n_1}})^{-m}]^{-n}$$

As mentioned in Section 4.4, our original model tends to over-estimate the theoretical risk comparing to our extended risk model. In this case, the $\hat{u}_{\text{total}}^{0}$ computed from original stem cell model would be underestimated. For cancer types with relatively long stem cell lineages, such as Hepatocellular carcinoma, Acute myeloid leukemia and Colorectal adenocarcinoma, etc, $\hat{u}_{\text{total}}^{0}$ usually provides a close estimate.

In addition, the extended risk model also includes progenitor cell lineages which accounts for the majority of the cell population, though they contribute little to cancer risk as they usually require more mutations for cancer onset. Therefore, with the above $\hat{u}_{\text{total}}^{0}$ the extended risk model will give a total risk close to or larger than the observed risk.

As we will see in the next chapter, clonal expansion has the most significant impact on cancer risk. With clonal expansion, using the $\hat{u}_{\text{total}}^{0}$ above will yield a higher total risk than the observed. Therefore, we can iteratively reduce $\hat{u}_{\text{total}}$ in this case to reach a lower bound or a close estimate for the total rate.

To estimate $\hat{u}_{\text{total}}$ for our Extended Risk Model, we can either take an iterative approach to provide a rough interval estimated with an upper/lower bound, or use binary search to give an accurate estimate.

The way we tune $\hat{u}_{\text{total}}$ for the first approach is by adjusting $\hat{u}_{\text{exc}} = \hat{u}_{\text{total}} - u_{\text{int}}$ as the intrinsic rate should be fixed. In particular, we use $\hat{C}_{\text{int}} = \frac{M(u_{\text{int}})}{M(u_{\text{int}}+\hat{u}_{\text{exc}})}$ as an estimated intrinsic contribution when $R(u_{\text{int}} + \hat{u}_{\text{exc}}) \approx R_{\text{obs}}$.

In cases where $R(u_{\text{int}} + \hat{u}_{\text{exc}})$ is much larger than R_{obs}, we use the interval estimate $[\ \hat{C}_{\text{int}}^{L} = \frac{M(u_{\text{int}})}{M(u_{\text{int}}+\hat{u}_{\text{exc}}/2)}, \hat{C}_{\text{int}}^{U} = \frac{M(u_{\text{int}})}{M(u_{\text{int}}+\hat{u}_{\text{exc}})}]$ for $\hat{C}_{\text{int}}$ providing $R(u_{\text{int}} + \hat{u}_{\text{exc}}/2) < R_{\text{obs}} < R(u_{\text{int}} + \hat{u}_{\text{exc}})$. Note that for some tissues with large number of stem divisions, we cannot

guarantee that the interval $\hat{C}_{\text{int}}^{L} \leq \hat{C}_{\text{int}}$; however, if $R(u_{\text{int}} + \hat{u}_{\text{exc}}/2)$ is close to R_{obs}, we can still use $\hat{C}_{\text{int}}^{L}$ as a close estimate as $\hat{C}_{\text{int}}$. Alternatively, we could further use $\hat{u}_{\text{ext}}/4$ as a smaller external risk to find a lower bound.

On the other hand, in rare cases where $R(u_{\text{int}} + \hat{u}_{\text{exc}})$ is much smaller than R_{obs} and $R(u_{\text{int}} + \hat{u}_{\text{exc}}) < R_{\text{obs}} < R(u_{\text{int}} + 2\hat{u}_{\text{exc}})$, we use $[\hat{C}_{\text{int}}^{L} = \frac{M(u_{\text{int}})}{M(u_{\text{int}}+\hat{u}_{\text{exc}})}, \hat{C}_{\text{int}}^{U} = \frac{M(u_{\text{int}})}{M(u_{\text{int}}+2\hat{u}_{\text{exc}})}]$ as an interval estimate.

In Section 4.6.2 we will also provide a heuristic lower/upper bound for cancer risk and intrinsic fraction of contribution, using our original stem cell model.

As we will see in first part of Chapter 5, the above approach to estimate total/excess rate based on observed risk cannot always guarantee accuracy. Therefore, we also use binary search to find $\hat{u}_{\text{total}}$ and provide a much more accurate estimate of intrinsic fraction of contribution.

Specifically, we can use $\hat{u}_{\text{total}}^{0}$ estimated from original stem cell model as an initial value, and then choose $[\hat{u}_{\text{low}}, \hat{u}_{\text{high}}]$ based on $\hat{u}_{\text{total}}^{0}$ as an initial lower/upper bound for $\hat{u}_{\text{total}}$. During binary search, we iteratively take the mean value $\hat{u}_{\text{mean}}$ of $\hat{u}_{\text{low}}, \hat{u}_{\text{high}}$; if the total risk based on $\hat{u}_{\text{mean}}$, $R(\hat{u}_{\text{mean}}) \approx R_{\text{obs}}$, then we take $\hat{u}_{\text{total}} = \hat{u}_{\text{mean}}$. Otherwise, if $R(\hat{u}_{\text{mean}}) < R_{\text{obs}}$, then we update $\hat{u}_{\text{low}} = \hat{u}_{\text{mean}}$; on the other hand, if $R(\hat{u}_{\text{mean}}) > R_{\text{obs}}$, we take $\hat{u}_{\text{upper}} = \hat{u}_{\text{mean}}$.

The first approach is computational efficient but not as accurate as the binary search algorithm. In Section 5.1, we will provide results for the 31 cancer types listed in Tomasetti and Vogelstein [31] and provide estimated intrinsic contribution based on the first tuning approach. In Section 5.2, we provide results for selected 18 organs similar to Tomasetti et al.'s work in 2017 [41], with intrinsic contribution estimated based on

binary search algorithm. We will see that binary search provides a more accurate estimate and is very effective in longitudinal study, i.e. the cancer risk and intrinsic contribution at each age.

4.6.2 Building Lower/Upper Bound Using the Original Stem Cell Model

We can use our original stem cell model to build lower and upper bounds for theoretical risk and mutation contributions from intrinsic factors. The bounds can be used as a validation of our extended risk model, as well as a heuristic interval estimated for the intrinsic contribution $\hat{C}_{\text{int}}$.

Since the original model only considered stem cells, while ignoring all other cell populations, it should already be a lower bound model that always underestimates the cancer risk. But due to the assumption of independence, inherently we are assuming more stem cell divisions during the symmetric division stage (see Section 4.4). Therefore, to build the absolute lower model we take $n_1 = n_s^s - 1 = \lfloor \log_2 N_S^H \rfloor$, as the number of symmetric stem divisions and keep n_1 the same as n_a^s, where n_s^s and n_a^s are the numbers of symmetric/asymmetric divisions for the extended risk model. Under this setting the original stem cell model will always has less inherent total stem cell divisions than extended risk model, therefore it will provide a lower bound for cancer risk.

We use the similar argument to build the upper bound model. We simply set $n_1 = \lceil \log_2 N^H \rceil$, meaning that we assume all cells are stem cells (which is an extreme case). This will provide us an absolute upper bound since stem cells requires less mutations for cancer onset for all other cells, and this model gives larger inherent cell divisions than extended risk model at any generation.

In addition, the lower/upper bound models based on the original stem cell model can provide approximate lower/upper bound for intrinsic contributions. Intuitively, since we force the total risk from both LB and UB models to reach the same R_{obs}, the lower bound model indicates minimum intrinsic contribution, while the upper bound model gives the maximum possible intrinsic contribution. Therefore, approximately the lower/upper bound models will also provide lower/upper bound for intrinsic contribution from our extended risk model, in cases where $R(u_{\text{int}} + \hat{u}_{\text{exc}})$ is close to R_{obs}.

However, if $R(u_{\text{int}} + \hat{u}_{\text{exc}}) \gg R_{\text{obs}}$, i.e. we have a significantly over-estimated $\hat{u}_{\text{exc}}$, $\hat{C}_{\text{int}}$ might be slightly smaller than the value provided by the lower bound model. In this case we can tune $\hat{u}_{\text{ext}}$ as described in Section 4.6.1.

Chapter 5. Results and Discussions

In this chapter, we present analysis results on cancer risk, intrinsic contribution factors, as well as the impact from mutation effects and clonal expansion for both 31 cancer types listed in the supplementary material of Tomasetti and Vogelstein [31], and 18 selected tissues similar to those in Tomasetti et al. [41], within an 80-year lifespan. The basic configurations used in our study for the 31 cancer types and 18 tissues are listed in Tables 3.1 and 3.2 below:

Table 3.1: Basic configurations used in our analysis for 31 cancer types (from Table S1 in [31]). We list the cancer types in the same order as in [31], where R_{obs} denotes observed lifetime risk, N^H is the total homeostatic number of cells, N_S^H is the total homeostatic stem cell number, and n_a^s represents number of asymmetric stem cell divisions which equals d (Number of divisions of each stem cell per lifetime) in [31].

id	cancer type	R_{obs}	N^H	N_S^H	n_a^s
1	Acute myeloid leukemia	0.0041	3.00E+12	1.35E+08	960
2	Basal cell carcinoma	0.3	1.80E+11	5.82E+09	608
3	Chronic lymphocytic leukemia	0.0052	3.00E+12	1.35E+08	960
4	Colorectal adenocarcinoma	0.048	3.00E+10	2.00E+08	5840
5	Colorectal adenocarcinoma with FAP	1	3.00E+10	2.00E+08	5840
6	Colorectal adenocarcinoma with Lynch syndrome	0.5	3.00E+10	2.00E+08	5840
7	Duodenum adenocarcinoma	0.0003	6.80E+08	4.00E+06	1947
8	Duodenum adenocarcinoma with FAP	0.035	6.80E+08	4.00E+06	1947
9	Esophageal squamous cell carcinoma	0.001938	3.24E+09	8.64E+05	1390
10	Gallbladder non papillary adenocarcinoma	0.0028	1.60E+08	1.60E+06	47
11	Glioblastoma	0.00219	8.46E+10	1.35E+08	0
12	Head and neck squamous cell carcinoma	0.0138	1.67E+10	1.85E+07	1720
13	Head and neck squamous cell carcinoma with HPV-16	0.07935	1.67E+10	1.85E+07	1720
14	Hepatocellular carcinoma	0.0071	2.41E+11	3.01E+09	88
15	Hepatocellular carcinoma with HCV	0.071	2.41E+11	3.01E+09	88
16	Lung adenocarcinoma (nonsmokers)	0.0045	4.34E+11	1.22E+09	5.6
17	Lung adenocarcinoma (smokers)	0.081	4.34E+11	1.22E+09	5.6
18	Medulloblastoma	0.00011	8.50E+10	1.36E+08	0
19	Melanoma	0.0203	3.80E+09	3.80E+09	199
20	Osteosarcoma	0.00035	1.90E+09	4.18E+06	5
21	Osteosarcoma of the arms	0.00004	3.00E+08	6.50E+05	5

22	Osteosarcoma of the head	0.0000302	3.90E+08	8.60E+05	5
23	Osteosarcoma of the legs	0.00022	7.20E+08	1.59E+06	5
24	Osteosarcoma of the pelvis	0.00003	2.00E+08	4.50E+05	5
25	Ovarian germ cell	0.000411	1.10E+07	1.10E+07	0
26	Pancreatic ductal adenocarcinoma	0.013589	1.67E+11	4.18E+09	80
27	Pancreatic endocrine (islet cell) carcinoma	0.000194	2.95E+09	7.40E+07	80
28	Small intestine adenocarcinoma	0.0007	1.70E+10	1.00E+08	2920
29	Testicular germ cell cancer	0.0037	2.16E+10	7.20E+06	463
30	Thyroid papillary/follicular carcinoma	0.01026	1.00E+10	6.50E+07	7
31	Thyroid medullary carcinoma	0.000324	1.00E+09	6.50E+06	7

Table 3.2: Basic configurations used in our analysis for selected 18 tissue types where National Program of Cancer Registries (NPCR) age dependent risk data is available. We list the tissue types in tissue id from 1 to 20 except 9 and 18 (to keep consistency), where R_{obs} denotes observed lifetime risk which equals the average value of observed risk within U.S. according to NPCR, N^H is the total homeostatic number of cells, N_S^H is the total homeostatic stem cell number, and n_a^s represents number of asymmetric stem cell divisions (Number of divisions of each stem cell per lifetime).

id	tissue type	R_{obs}	N^H	N_S^H	n_a^s
1	Brain	6.11E-03	8.50E+10	1.36E+08	1.4
2	Thyroid medullary	1.98E-04	1.00E+09	6500000	7
3	Bone	9.02E-04	1.90E+09	4180000	5
4	Ovarian germ cell	2.82E-04	1.10E+07	1.10E+07	0
5	Esophageal	5.85E-03	3.24E+09	864000	1390
6	leukemia	1.21E-02	3.00E+12	1.35E+08	960
7	liver	6.71E-03	2.41E+11	3.01E+09	88
8	Testicular	3.84E-03	2.16E+10	7200000	463
10	Thyroid follicular	8.31E-03	1.00E+10	6.50E+07	7
11	Pancreatic	1.21E-02	1.70E+11	4.25E+09	80
12	Head & neck	1.62E-02	1.67E+10	18500000	1720
13	Melanoma	1.78E-02	3.80E+09	3.80E+09	199
14	Colon	4.55E-02	3.00E+10	2.00E+08	5840
15	Lung	8.09E-02	4.34E+11	1.22E+09	5.6
16	Breast	1.19E-01	6.80E+11	8.70E+09	345.5621302
17	Prostate	1.86E-01	3.00E+10	2.10E+08	240
19	Gallbladder	1.11E-03	1.60E+08	1600000	47
20	Small intestine	2.10E-03	1.70E+10	1.00E+08	2920

In Section 5.1, we focus on the 31 cancer types in Table 3.1 and give their intrinsic risk, intrinsic contribution to observed risk, computed from extended risk model under different settings of parameters. In addition, we visualize the effect of certain mutation effects such as mutation rate enlargement and clonal expansion, by displaying the computed intrinsic risk for each age for hepatocellular carcinoma, whose age dependent observed risk is readily available through SEER 1973-2012 database. The purpose of Section 5.1 is to provide a preliminary and quick demonstration on how much intrinsic/non-intrinsic factors contributes to overall cancer risk, and the impact of different mutation effects on intrinsic risk.

In Section 5.2, we will use the selected 18 tissues in Table 3.2 for a more detailed study. In particular, we will use binary search algorithm to accurately estimate a total mutation rate corresponding to observed risk given by NPCR database, and then give an accurate estimation of intrinsic contribution factor. In addition, we perform a longitudinal study on the age-dependent risk and intrinsic contribution within an 80-year lifetime. Also, for some tissues we also study the sensitivity of intrinsic risk with respect to the mutation hits required for cancer onset.

5.1 Preliminary Study with 31 Cancer Types

5.1.1 Age Dependent Cancer Risk

It is commonly accepted that cancer risk could depend on age, for example, the incidences of cancer increases faster after 40~50 years in a lifetime. The SEER 1973-2012 database provided cancer incidences and cancer risk for 19 age groups (spanning from 0 ~ 85) as well as each year from 0 to 85 of the US population. In this experiment, we compare the cancer risk computed from our extended risk model and the SEER data.

Table 4 lists the SEER cancer incidence data for 19 age groups for Hepatocellular carcinoma (Liver), and we use the ratio of cancer incidence count vs. the population size as an average cancer risk for each age within the age group. Note that though the table displays average incidence counts for each 5-year age group, we use the incidence counts for each of 80 years, which is also available in SEER database.

Table 4: SEER (1973 - 2012) data for Hepatocellular carcinoma cancer risk for each age group. Rates are per 100,000 and age-adjusted to the 2000 US Standard Population (19 age groups - Census P25-1130) standard.

Age Group	Rate	Count	Population
00 years	0	6	14,353,623
01-04 years	0	9	56,138,688
05-09 years	0	21	70,244,197
10-14 years	0.1	37	72,307,118
15-19 years	0.1	42	73,435,145
20-24 years	0.1	73	74,598,991
25-29 years	0.1	116	77,895,377
30-34 years	0.3	204	77,308,099
35-39 years	0.5	395	72,951,615
40-44 years	1.2	813	68,321,306
45-49 years	3.2	1,991	62,601,425
50-54 years	6.6	3,764	56,797,613
55-59 years	10.1	4,995	49,315,629
60-64 years	11.6	4,816	41,464,247
65-69 years	12.8	4,372	34,194,373
70-74 years	15.4	4,245	27,589,401
75-79 years	17	3,671	21,652,916
80-84 years	15.3	2,315	15,092,287
85+ years	11.2	1,450	12,976,530
Unknown	~	0	0

We now compute the cancer risk for each year (1 - 80) computed by our extended cancer risk model and plot against the SEER data. Based on the data in supplementary material of [31], Hepatocellular has totally 120 number of stem cell divisions, among which the first 32 were considered symmetric divisions. The homeostatic total cell number is 2.41×10^{11} and the stem cell number is 3.01×10^{9}. Assuming the symmetric division stage (tissue growth before maturing) takes ~40 weeks, then we could decide the normal stem cell cycle time at each year. In later simulation on clonal expansion, cells with certain driver mutation will have shorter cell cycle time. Note that the homeostatic cell number is also used to determine the progenitor lineage length, which is 5 in our case assuming equal cell cycle time. As for mutation acquisition, we consider five driver mutations $(M_1, M_2, M_3, M_4, M_5)$; a stem cell needs (M_3, M_4, M_5) to become a cancer cell, while a progenitor cell needs (M_2, M_3, M_4, M_5), and a terminal cell needs $(M_1, M_2, M_3, M_4, M_5)$ to form a cancer onset. We define mutation probability for 4 general types of transitions: stem - stem, stem - progenitor, progenitor - progenitor and progenitor - terminal. In this our base intrinsic mutation rate were set to be 1×10^{-8} unless otherwise noted.

Figure 19 illustrates the log10 scale cancer risk from SEER and our extended risk model from age 1 to 80, with a set of different intrinsic mutation rates $1\times10^{-9}, 5\times10^{-9}, 1\times10^{-8}, 5\times10^{-8}$.

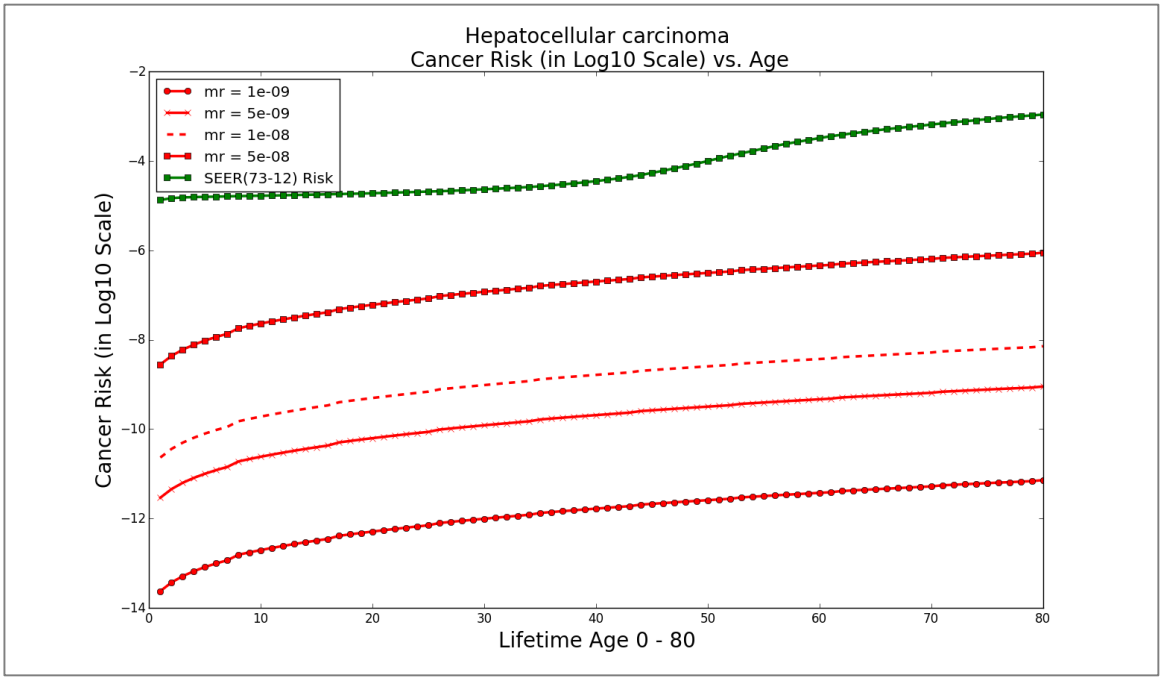


Figure 19: This figure shows the log scale cancer risk from SEER and our model from age 1 to 80. At age 80, the observed SEER cancer risk is 0.001 while the computed cancer risks from our mode are all below the observed risk, for mutation rates ranging from 1e-09 to 5e-08. In addition, we can see an overall trend of cancer risk is similar to SEER but the trend is different at both early age and middle age. The observation indicates significant contribution from age dependent potential extrinsic risk, such as alcohol, smoke, immune system, etc.

From Figure 19 we can see similar overall increasing trend with age for both theoretical risk and observed risk, except that SEER has a higher rate of increasing after age 40 and a higher relative risk at very early ages. This is a clear indication that age-varying extrinsic risk, such as alcohol, smoke, immune system, would be a major factor for the age-dependent patterns in cancer risk.

However, to quantify how much the contribution of intrinsic and extrinsic risks, we need to first add enough flexibility into our experiment, such as clonal expansion and other mutation effects.

As mentioned previously, our model integrates the impact of driver mutations on the process of cell evolution as well as mutation acquisition. In the following we present

results that incorporate the effect of enlarged mutation rate as well as clonal expansion. This makes the model closer to the real dynamics of a tissue.

We assume driver mutation M_4 to be the one that increases mutation rate of descendant cells. Varying the enlarging factor from 1 to 4, we observe that the cancer risk trend curve simply shift upwards in log10 scale, as shown in Figure 20 below:

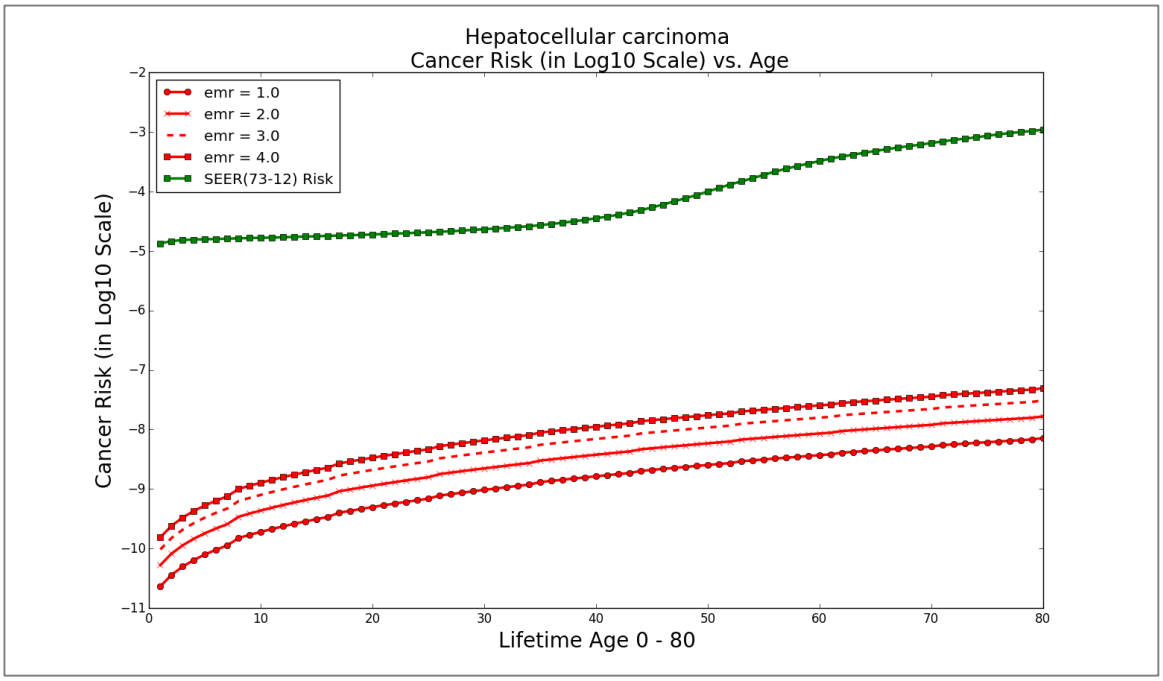


Figure 20: This figure shows the log scale cancer risk from SEER and our model from age 1 to 80 by varying the factor of mutation rate increase as an effect of certain mutations. The factor ranges from 1.0 (no effect) to 4.0. The baseline mutation rate was chosen to be 1e-08.

We see that this mutation effect increases the chance of cancer onset at each age at an equal scale. However it does not significantly match closer to the shape of SEER curve, therefore there must be additional factors that may contribute to the shift of cancer risk increase at the middle age. Next, we will include the clonal expansion effect.

The factor of clonal expansion here is defined as how many more times a cell with certain driver mutations will divide faster than the rest group of cells. Note that this effect can break the tissue homeostasis, i.e. make the tissue size grow infinite large if without

restriction. Therefore in reality there is a regulation mechanism. In here, once the cell size grows to its homeostatic number, the clonal expansion will be regulated and cells return to its normal division speed. Also it is reasonable to restrict that clonal expansion can only occur after the symmetric stem division stage.

Figure 21 displays the cancer risk vs. age considering clonal expansion resulting from the acquisition of mutation M_5. We see that a 2 or 3 fold clonal expansion would lead to the overall increase of cancer risk.

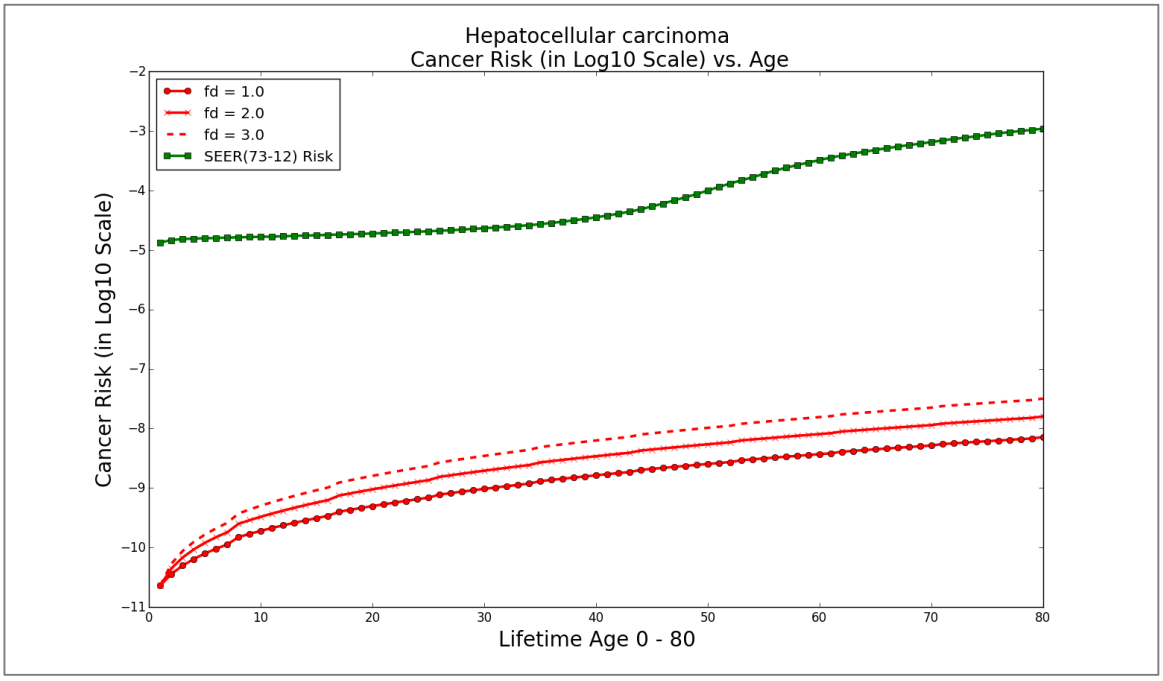


Figure 21: This figure shows the log10 scale cancer risk from SEER and our model from age 1 to 80 by varying the factor of clonal expansion. The factor ranges from 1.0 (no effect) to 3.0. The baseline mutation rate was chosen to be 1e-08. In here we restrict the clonal expansion to be effective only after 40 weeks, when homeostasis is approximately achieved. We see that a 2 or 3 fold clonal expansion would lead to increase of cancer risk but the overall risk is still well below the observed risk from SEER. This figure again indicates the potential contribution from extrinsic factors.

We can also see that the cancer risk grows slightly faster with clonal expansion and the overall risk is still far below the level in SEER data.

In Figure 22 we can visualize the combined effect of mutation rate enlargement and clonal expansion. The increasing combined effects would lead to increase of cancer risk but the overall risk is still well below the observed risk from SEER. In addition, the discrepancy in the rate of cancer risk increase after age 40 also indicates fundamental extrinsic influence.

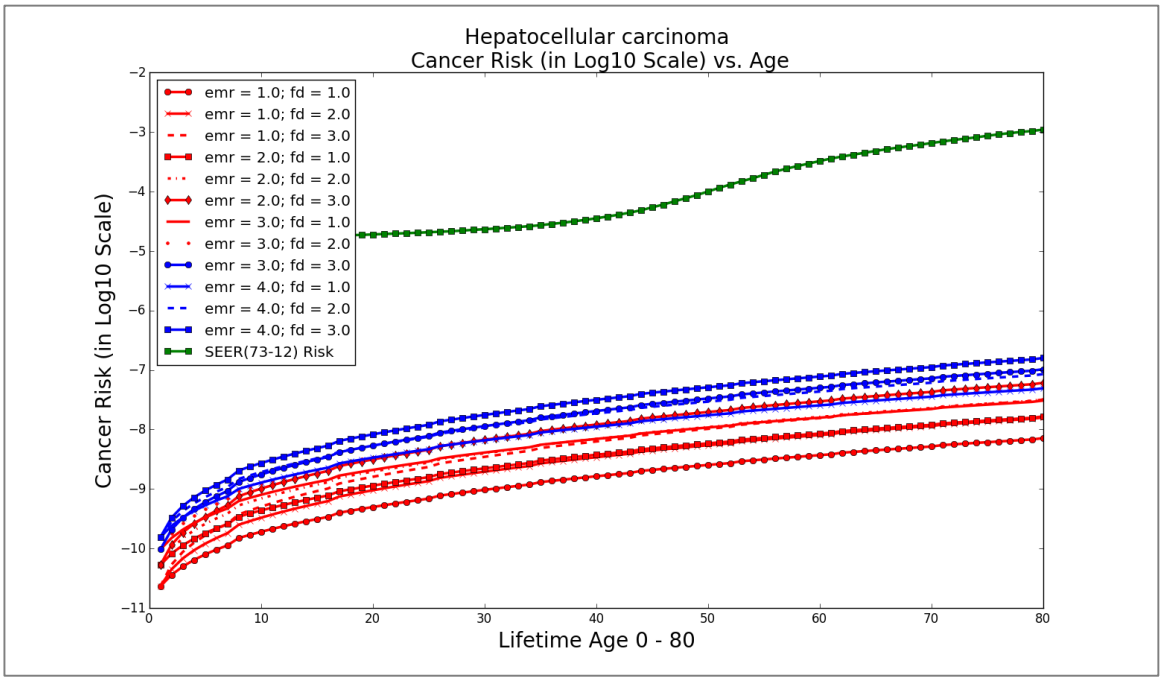


Figure 22: This figure shows the log10 scale cancer risk from SEER and our model from age 1 to 80 by varying the factor of both increasing mutation rate and clonal expansion. The mutation rate increase factor (emr) ranges from 1.0 (no effect) to 4.0 and the clonal expansion factor (fd) ranges from 1.0 to 3.0. The baseline mutation rate was chosen to be 1e-08. In here we restrict the clonal expansion to be effective only after 40 weeks, when homeostasis is approximately achieved. We see that the increasing combined effects would lead to increase of cancer risk but the overall risk is still well below the observed risk from SEER. This figure again indicates the potential contribution from extrinsic factors.

5.1.2 Intrinsic Risk and Intrinsic Contribution

In this section, we will provide results about intrinsic contribution as discussed in Section 4.6, for 31 cancer types by Tomasetti and Vogelstein [31]. We will give the lifetime contribution percentage based on our extended model with estimated excess mutation rate, along with the lower/upper bound from original stem cell model (Section 4.6.2). Many parameters for our model, such as intrinsic mutation rate and mutation effects, were not precisely predetermined; therefore, we need to provide different parameter settings in following sections to obtain a reasonable range/sensitivity for the intrinsic contributions.

In each subsection below, we will provide two tables for each set of parameter configurations. The first table gives the estimated excess mutation rate $\hat{u}_{\text{exc}}$ along with intrinsic and total estimated cancer risk from our extended risk model, as a reference to see how whether $\hat{u}_{\text{ext}}$ could be a close estimate or serve as a lower bound. In cases where $\hat{u}_{\text{ext}}$ is highly overestimated, we provide the results with $\hat{u}_{\text{ext}}/2$ because we need to ensure the intrinsic contribution is not underestimated to obtain a reliable conclusion. The second table gives the corresponding intrinsic contribution percentage values. Note that it is important to interpret the results *using both tables* as a pair as the first table usually marks some special occasions that might lead to a misleading conclusion if not considered.

5.1.2.1 Intrinsic Mutation Rate

According to most literatures the intrinsic mutation rate should be fall between $10^{-9}\sim10^{-6}$. In our setting, we use two different mutation rates, $u_{\text{int}}^{(1)} = 10^{-8}$ and $u_{\text{int}}^{(2)} = 10^{-7}$as an intrinsic rate greater than 10^{-6} will give unreasonably large intrinsic

risk (sometimes close to 1.0). We currently do not consider any mutation effects and clonal expansion, which will be discussed later.

Table 5: Estimated excess mutation rate, total risk and intrinsic risk for selected cancer types with two settings of intrinsic mutation rate $u_{int}^{(1)} = 10^{-8}$ and $u_{int}^{(2)} = 10^{-7}$. R_{obs} is the observed risk, $\hat{u}_{exc}^{(1)}, \hat{u}_{exc}^{(2)}$ represent estimated excess rate with intrinsic rate settings $u_{int}^{(1)}$ and $u_{int}^{(2)}$ respectively. Similarly, $\hat{R}_{total}$ and $\hat{R}_{int}$ represent estimated total risk (from mutation rate $u_{int} + \hat{u}_{exc}$), and intrinsic risk (from mutation rate u_{int}).

Note:

***1. In cases where $\hat{R}_{total}$ is much higher than R_{obs} with originally estimated $\hat{u}_{exc}$, we take $\hat{u}_{exc}/2$ instead.**

***2. In cases where $\hat{R}_{int}$ is already bigger than R_{obs}, which usually occurs for tissues with very long stem lineages, we could have a negative $\hat{u}_{exc}$. This indicates only a smaller intrinsic mutation rate could be reasonable and therefore we ignore the interpretations for this situation.**

***3. This is a special case where observed risk is 1.0, in which our algorithm for estimation $\hat{u}_{exc}$ does not apply.**

id	cancer type	R_{obs}	$\hat{R}_{total}^{(1)}$	$\hat{R}_{total}^{(2)}$	$\hat{u}_{exc}^{(1)}$	$\hat{u}_{exc}^{(2)}$	$\hat{R}_{int}^{(1)}$	$\hat{R}_{int}^{(2)}$
1	Acute myeloid leukemia	4.10E-03	1.26E-03	3.42E-03	1.21E-07*1	7.57E-08*1	2.91E-07	5.12E-04
2	Basal cell carcinoma	3.00E-01	2.92E-01	2.92E-01	5.30E-07	4.40E-07	2.26E-06	2.22E-03
3	Chronic lymphocytic leukemia	5.20E-03	1.63E-03	4.15E-03	1.31E-07*1	8.61E-08*1	2.91E-07	5.12E-04
4	Colorectal adenocarcinoma	4.80E-02	5.04E-02	5.04E-02	8.68E-08	-3.18E-09*2	5.45E-05	5.54E-02
5	Colorectal adenocarcinoma with FAP	1.00E+00*3	1.00E+00	1.00E+00	1.00E+00	1.00E+00	5.45E-05	5.54E-02
7	Duodenum adenocarcinoma	3.00E-04	3.12E-04	3.12E-04	2.01E-07	1.11E-07	3.21E-08	3.26E-05
10	Gallbladder non papillary adenocarcinoma	2.80E-03	2.91E-03	2.91E-03	1.62E-05	1.61E-05	6.44E-13	6.42E-10
14	Hepatocellular carcinoma	7.10E-03	6.13E-03	6.13E-03	9.77E-07	8.87E-07	7.16E-09	7.04E-06
15	Hepatocellular carcinoma with HCV	7.10E-02	5.90E-02	5.90E-02	2.14E-06	2.05E-06	7.16E-09	7.04E-06
16	Lung adenocarcinoma (nonsmokers)	4.50E-03	1.61E-03	1.61E-03	3.55E-06	3.46E-06	4.55E-11	4.32E-08
17	Lung adenocarcinoma (smokers)	8.10E-02	2.67E-02	2.67E-02	9.44E-06	9.35E-06	4.55E-11	4.32E-08
19	Melanoma	2.03E-02	1.87E-02	1.87E-02	7.19E-07	6.29E-07	5.24E-08	5.17E-05
20	Osteosarcoma	3.50E-04	2.45E-04	2.45E-04	1.62E-05	1.61E-05	4.85E-14	4.83E-11
21	Osteosarcoma of the arms	4.00E-05	2.98E-05	2.98E-05	1.35E-05	1.34E-05	1.02E-14	1.02E-11
22	Osteosarcoma of the head	3.02E-05	2.21E-05	2.21E-05	1.23E-05	1.22E-05	1.02E-14	1.02E-11
23	Osteosarcoma of the legs	2.20E-04	1.66E-04	1.66E-04	1.81E-05	1.80E-05	2.22E-14	2.22E-11
24	Osteosarcoma of the pelvis	3.00E-05	2.43E-05	2.43E-05	1.60E-05	1.60E-05	4.62E-15	4.63E-12
26	Pancreatic ductal adenocarcinoma	1.36E-02	1.14E-02	1.14E-02	1.30E-06	1.21E-06	5.79E-09	5.68E-06
27	Pancreatic endocrine (islet cell) carcinoma	1.94E-04	1.83E-04	1.83E-04	1.05E-06	9.57E-07	1.61E-10	1.59E-07
30	Thyroid papillary/follicular carcinoma	1.03E-02	5.08E-03	5.08E-03	1.62E-05	1.61E-05	1.46E-12	1.42E-09
31	Thyroid medullary	3.24E-04	1.94E-04	1.94E-04	1.13E-05	1.12E-05	1.48E-13	1.46E-10

	carcinoma							

Table 6: Estimated intrinsic contribution percentage for selected cancer types with two settings of intrinsic mutation rate $u_{\text{int}}^{(1)} = 10^{-8}$ and $u_{\text{int}}^{(2)} = 10^{-7}$. "$\widehat{C}_{\text{int}}$ all cells" gives the estimated intrinsic contribution based on the ratio of number of acquired mutations among *all cells* due to u_{int} vs. $(u_{\text{int}} + u_{\text{ext}})$, while "$\widehat{C}_{\text{int}}$ cancer cells" gives the same quantity considering only cancer cells, i.e. those cells acquired sufficient types of driver mutations for cancer onset.
Note:
***1. In cases where $\widehat{R}_{\text{int}}$ is already bigger than R_{obs}, which usually occurs for tissues with very long stem lineages, we could have an intrinsic contribution larger than 100%. This indicates only a smaller intrinsic mutation rate could be reasonable and therefore we ignore the interpretations for this situation.**

id	cancer type	$\widehat{C}_{\text{int}}^{(1)}$ all cells	$\widehat{C}_{\text{int}}^{(2)}$ all cells	$\widehat{C}_{\text{int}}^{(1)}$ cancer cells	$\widehat{C}_{\text{int}}^{(2)}$ cancer cells
1	Acute myeloid leukemia	7.65%	56.92%	0.01%	12.17%
2	Basal cell carcinoma	1.85%	18.50%	0.00%	0.63%
3	Chronic lymphocytic leukemia	7.09%	53.74%	0.01%	9.79%
4	Colorectal adenocarcinoma	10.32%	103.28%[*1]	0.10%	110.48%[*1]
5	Colorectal adenocarcinoma with FAP	0.00%	0.02%	0.00%	0.00%
7	Duodenum adenocarcinoma	4.74%	47.41%	0.01%	10.30%
10	Gallbladder non papillary adenocarcinoma	0.06%	0.62%	0.00%	0.00%
14	Hepatocellular carcinoma	1.01%	10.13%	0.00%	0.10%
15	Hepatocellular carcinoma with HCV	0.47%	4.65%	0.00%	0.01%
16	Lung adenocarcinoma (nonsmokers)	0.28%	2.81%	0.00%	0.00%
17	Lung adenocarcinoma (smokers)	0.11%	1.06%	0.00%	0.00%
19	Melanoma	1.37%	13.72%	0.00%	0.26%
20	Osteosarcoma	0.06%	0.62%	0.00%	0.00%
21	Osteosarcoma of the arms	0.07%	0.74%	0.00%	0.00%
22	Osteosarcoma of the head	0.08%	0.82%	0.00%	0.00%
23	Osteosarcoma of the legs	0.06%	0.55%	0.00%	0.00%
24	Osteosarcoma of the pelvis	0.06%	0.62%	0.00%	0.00%
26	Pancreatic ductal adenocarcinoma	0.76%	7.61%	0.00%	0.04%
27	Pancreatic endocrine (islet cell) carcinoma	0.95%	9.46%	0.00%	0.08%
30	Thyroid papillary/follicular carcinoma	0.06%	0.62%	0.00%	0.00%
31	Thyroid medullary carcinoma	0.09%	0.89%	0.00%	0.00%

From Tables 5 and 6, we see for the majority of cancer types under both intrinsic mutation rates, the percentage of intrinsic contributions are very small. When we have a larger mutation rate 10^{-7}, the intrinsic contribution from all cells appears significant in tissues with long stem cell lineages such as bone marrow, colon and duodenum, etc.

However, the real intrinsic contributions in this case are very likely to be overestimated for three reasons. First, the predetermined intrinsic rate could be unreasonably higher than reality, for example, for Colorectal adenocarcinoma, the theoretical intrinsic risk is greater than observed risk under $u_{\text{int}} = 10^{-7}$ which means a smaller u_{int} is more valid. Second, one tissue have very different observed risks for different cancer types or different environmental conditions, such as colon and liver, which indicates significant non-intrinsic contribution [31]. Finally, the intrinsic contribution percentage with only cancer cells is a far more valid indicator comparing to that including all cells, because the former one is directly correlated to cancer onset. For example, a tissue could have many stem cells with 1 or 2 driver mutations under intrinsic and total rate both, but these cells does not contribute to the risk of cancer; also it is much easier for a stem cell to acquire just 1 mutations than 3 under small mutation rate, therefore the computed intrinsic contribution from all cells has an "offset effect" and thus tends to be larger. In this and all following experiments, we use $\hat{C}_{\text{int}}$ from only cancer cells as a primary indicator.

5.1.2.2 Mutation Effects Factors

We compare the results with and without the mutation effect that causes increased mutation rate. We assume any cell with mutation M_4 will have higher rate, $\alpha \times u$ of acquiring additional mutations, where u is the initial mutation rate and α is the factor of enlargement. We compare the results with $\alpha = 1$ (no effect) and $\alpha = 2$.

In this experiment, we will not include clonal expansion effects, but will provide results under both mutation rates of 10^{-8} and 10^{-7}.

Tables 7/8 below provide the results under intrinsic mutation rate 10^{-8}, and Tables 9/10 provide the results under intrinsic mutation rate 10^{-7}.

Table 7: Estimated excess mutation rate, total risk and intrinsic risk for selected cancer types with two settings of mutation rate enlargement (as an effect of mutation M_4) factor $\alpha^{(1)} = 1$ and $\alpha^{(2)} = 2$. We use intrinsic mutation rate 10^{-8}.

R_{obs} is the observed risk, $\hat{u}_{ext}^{(1)}, \hat{u}_{ext}^{(2)}$ represent estimated excess rate with two settings respectively. Similarly, $\hat{R}_{total}$ and $\hat{R}_{int}$ represent estimated total risk (from mutation rate $u_{int} + \hat{u}_{exc}$), and intrinsic risk (from mutation rate u_{int}).

Note:

***1. In cases where $\hat{R}_{total}$ is much higher than R_{obs} with originally estimated $\hat{u}_{exc}$, we take $\hat{u}_{exc}/2$ instead.**

***2. This is a special case where observed risk is 1.0, in which our algorithm for estimation $\hat{u}_{exc}$ does not apply.**

id	cancer type	R_{obs}	$\hat{R}_{total}^{(1)}$	$\hat{R}_{total}^{(2)}$	$\hat{u}_{exc}^{(1)}$	$\hat{u}_{exc}^{(2)}$	$\hat{R}_{int}^{(1)}$	$\hat{R}_{int}^{(2)}$
1	Acute myeloid leukemia	4.10E-03	1.26E-03	1.37E-03	1.21E-07*1	1.21E-07*1	2.91E-07	7.13E-07
2	Basal cell carcinoma	3.00E-01	2.92E-01	2.96E-01	5.30E-07	5.30E-07	2.26E-06	5.26E-06
3	Chronic lymphocytic leukemia	5.20E-03	1.63E-03	1.76E-03	1.31E-07*1	1.31E-07*1	2.91E-07	7.13E-07
4	Colorectal adenocarcinoma	4.80E-02	5.04E-02	5.57E-02	8.68E-08	8.68E-08	5.45E-05	1.27E-04
5	Colorectal adenocarcinoma with FAP	1.00E+00*2	1.00E+00	1.00E+00	1.00E+00	1.00E+00	5.45E-05	1.27E-04
7	Duodenum adenocarcinoma	3.00E-04	3.12E-04	3.27E-04	2.01E-07	2.01E-07	3.21E-08	7.49E-08
10	Gallbladder non papillary adenocarcinoma	2.80E-03	2.91E-03	2.91E-03	1.62E-05	1.62E-05	6.44E-13	1.49E-12
14	Hepatocellular carcinoma	7.10E-03	6.13E-03	6.19E-03	9.77E-07	9.77E-07	7.16E-09	1.66E-08
15	Hepatocellular carcinoma with HCV	7.10E-02	5.90E-02	5.92E-02	2.14E-06	2.14E-06	7.16E-09	1.66E-08
16	Lung adenocarcinoma (nonsmokers)	4.50E-03	1.61E-03	1.62E-03	3.55E-06	3.55E-06	4.55E-11	1.04E-10
17	Lung adenocarcinoma (smokers)	8.10E-02	2.67E-02	2.67E-02	9.44E-06	9.44E-06	4.55E-11	1.04E-10
19	Melanoma	2.03E-02	1.87E-02	1.89E-02	7.19E-07	7.19E-07	5.24E-08	1.22E-07
20	Osteosarcoma	3.50E-04	2.45E-04	2.45E-04	1.62E-05	1.62E-05	4.85E-14	1.11E-13
21	Osteosarcoma of the arms	4.00E-05	2.98E-05	2.99E-05	1.35E-05	1.35E-05	1.02E-14	2.32E-14
22	Osteosarcoma of the head	3.02E-05	2.21E-05	2.22E-05	1.23E-05	1.23E-05	1.02E-14	2.32E-14
23	Osteosarcoma of the legs	2.20E-04	1.66E-04	1.66E-04	1.81E-05	1.81E-05	2.22E-14	5.09E-14
24	Osteosarcoma of the pelvis	3.00E-05	2.43E-05	2.43E-05	1.60E-05	1.60E-05	4.62E-15	1.06E-14
26	Pancreatic ductal adenocarcinoma	1.36E-02	1.14E-02	1.15E-02	1.30E-06	1.30E-06	5.79E-09	1.34E-08
27	Pancreatic endocrine (islet cell) carcinoma	1.94E-04	1.83E-04	1.85E-04	1.05E-06	1.05E-06	1.61E-10	3.73E-10
30	Thyroid papillary/follicular carcinoma	1.03E-02	5.08E-03	5.08E-03	1.62E-05	1.62E-05	1.46E-12	3.35E-12
31	Thyroid medullary carcinoma	3.24E-04	1.94E-04	1.95E-04	1.13E-05	1.13E-05	1.48E-13	3.38E-13

Table 8: Estimated intrinsic contribution percentage for selected cancer types with two settings of mutation rate enlargement (as an effect of mutation M_4) factor $\alpha^{(1)} = 1$ and $\alpha^{(2)} = 2$. We use intrinsic mutation rate 10^{-8}.
"$\hat{C}_{int}$ all cells" gives the estimated intrinsic contribution based on the ratio of number of acquired mutations among all cells due to u_{int} vs. $(u_{int} + u_{exc})$, while "$\hat{C}_{int}$ cancer cells" gives the same quantity considering only cancer cells, i.e. those cells acquired sufficient types of driver mutations for cancer onset.
Note:
***1. We list the values in the format of scientific numbers as all of them are below 0.01%**

id	cancer type	$\hat{C}_{int}^{(1)}$ all cells	$\hat{C}_{int}^{(2)}$ all cells	$\hat{C}_{int}^{(1)}$ cancer cells*1	$\hat{C}_{int}^{(2)}$ cancer cells*1
1	Acute myeloid leukemia	7.65%	7.65%	1.43E-04	3.33E-04
2	Basal cell carcinoma	1.85%	1.85%	6.26E-06	1.43E-05
3	Chronic lymphocytic leukemia	7.09%	7.09%	1.08E-04	2.52E-04
4	Colorectal adenocarcinoma	10.32%	10.32%	1.02E-03	2.14E-03
5	Colorectal adenocarcinoma with FAP	0.00%	0.00%	9.99E-16	2.22E-15
7	Duodenum adenocarcinoma	4.74%	4.74%	1.00E-04	2.23E-04
10	Gallbladder non papillary adenocarcinoma	0.06%	0.06%	1.98E-10	4.58E-10
14	Hepatocellular carcinoma	1.01%	1.01%	1.03E-06	2.37E-06
15	Hepatocellular carcinoma with HCV	0.47%	0.47%	9.86E-08	2.28E-07
16	Lung adenocarcinoma (nonsmokers)	0.28%	0.28%	2.20E-08	5.02E-08
17	Lung adenocarcinoma (smokers)	0.11%	0.11%	1.15E-09	2.64E-09
19	Melanoma	1.37%	1.37%	2.58E-06	5.92E-06
20	Osteosarcoma	0.06%	0.06%	2.32E-10	5.28E-10
21	Osteosarcoma of the arms	0.07%	0.07%	4.05E-10	9.18E-10
22	Osteosarcoma of the head	0.08%	0.08%	5.37E-10	1.22E-09
23	Osteosarcoma of the legs	0.06%	0.06%	1.61E-10	3.66E-10
24	Osteosarcoma of the pelvis	0.06%	0.06%	2.38E-10	5.40E-10
26	Pancreatic ductal adenocarcinoma	0.76%	0.76%	4.38E-07	1.01E-06
27	Pancreatic endocrine (islet cell) carcinoma	0.95%	0.95%	8.44E-07	1.94E-06
30	Thyroid papillary/follicular carcinoma	0.06%	0.06%	2.16E-10	4.95E-10
31	Thyroid medullary carcinoma	0.09%	0.09%	6.65E-10	1.52E-09

In our implementation of mutation rate enlargement effect, we restrict this effect to only influence intrinsic mutation rate. For example, cells with M_4 could have an intrinsic mutation rate $2u_{\text{int}}$ and the total rate $2u_{\text{int}} + u_{\text{exc}}$, as excess rate can only be altered due to possible changes in non-intrinsic factors. We can see an obvious demonstration in Table 7, where $\hat{R}_{\text{int}}^{(2)}$ for each cancer type is nearly twice as $\hat{R}_{\text{int}}^{(1)}$, while $\hat{R}_{\text{total}}^{(2)}$ is only slightly bigger than $\hat{R}_{\text{total}}^{(1)}$ as $\hat{u}_{\text{exc}}$ dominates the rate.

In Table 8, we noticed that $\hat{C}_{\text{int}}$ from total cells appears no difference under two conditions, $\alpha = 1$ and $\alpha = 2$. In fact the total numbers of mutations under two conditions are very close and non-differentiable under ordinary precision. Since intrinsic rate are very small, it is very unlikely that a cell will acquire ≥ 2 extra mutations even with mutation rate doubled. In addition, cells without M_4 should be on average $1/u_{\text{int}}$ times of cells with M_4. Also due to the fact that we do not count duplicated mutations, i.e. if a cell acquires the same mutation multiple times, we only count as one. Therefore on average, the total number of mutations will only have a slight difference.

On the other hand, $\hat{C}_{\text{int}}$ from cancer cells under $\alpha = 2$ is significantly larger (almost two fold) than the quantity at $\alpha = 1$, though most of them both are very small. This is because we are enforcing a condition that the cells have already acquired requested set of mutations. In our case M_4 is among the set of required driver mutations for cancer onset for all cell types: stem, progenitor and terminal cells. Therefore for the small population of cancer cells, the change of acquiring new mutations is nearly doubled, which explains the significant change in $\hat{C}_{\text{int}}$. This means that the mechanism of mutation effect actually increases intrinsic factor contributions, assuming constant non-intrinsic influence.

We list the same results for intrinsic rate 10^{-7} in Tables 9 and 10 below:

Table 9: Estimated excess mutation rate, total risk and intrinsic risk for selected cancer types with two settings of mutation rate enlargement (as an effect of mutation M_4) factor $\alpha^{(1)} = 1$ and $\alpha^{(2)} = 2$. We use intrinsic mutation rate 10^{-7}.

R_{obs} is the observed risk, $\hat{u}_{exc}^{(1)}, \hat{u}_{exc}^{(2)}$ represent estimated excess rate with two settings respectively. Similarly, $\hat{R}_{total}$ and $\hat{R}_{int}$ represent estimated total risk (from mutation rate $u_{int} + \hat{u}_{exc}$), and intrinsic risk (from mutation rate u_{int}).

Note:

***1. In cases where $\hat{R}_{total}$ is much higher than R_{obs} with originally estimated $\hat{u}_{exc}$, we take $\hat{u}_{exc}/2$ instead.**

***2. This is a special case where observed risk is 1.0, in which our algorithm for estimation $\hat{u}_{exc}$ does not apply.**

***3. In cases where $\hat{R}_{int}$ is already bigger than R_{obs}, which usually occurs for tissues with very long stem lineages, we could have a negative $\hat{u}_{exc}$. This indicates only a smaller intrinsic mutation rate could be reasonable and therefore we ignore the interpretations for this situation.**

id	cancer type	R_{obs}	$\hat{R}_{total}^{(1)}$	$\hat{R}_{total}^{(2)}$	$\hat{u}_{exc}^{(1)}$	$\hat{u}_{exc}^{(2)}$	$\hat{R}_{int}^{(1)}$	$\hat{R}_{int}^{(2)}$
1	Acute myeloid leukemia	4.10E-03	3.42E-03	5.97E-03	7.57E-08*1	7.57E-08*1	5.12E-04	1.32E-03
2	Basal cell carcinoma	3.00E-01	2.92E-01	3.38E-01	4.40E-07	4.40E-07	2.22E-03	5.13E-03
3	Chronic lymphocytic leukemia	5.20E-03	4.15E-03	7.02E-03	8.61E-08*1	8.61E-08*1	5.12E-04	1.32E-03
4	Colorectal adenocarcinoma	4.80E-02	5.04E-02	1.19E-01	-3.18E-09*3	-3.18E-09*3	5.54E-02	1.27E-01
5	Colorectal adenocarcinoma with FAP	1.00E+00*2	1.00E+00	1.00E+00	1.00E+00	1.00E+00	5.54E-02	1.27E-01
7	Duodenum adenocarcinoma	3.00E-04	3.12E-04	2.22E-04	1.11E-07	5.54E-08*1	3.26E-05	7.67E-05
10	Gallbladder non papillary adenocarcinoma	2.80E-03	2.91E-03	2.93E-03	1.61E-05	1.61E-05	6.42E-10	1.48E-09
14	Hepatocellular carcinoma	7.10E-03	6.13E-03	6.74E-03	8.87E-07	8.87E-07	7.04E-06	1.62E-05
15	Hepatocellular carcinoma with HCV	7.10E-02	5.90E-02	6.16E-02	2.05E-06	2.05E-06	7.04E-06	1.62E-05
16	Lung adenocarcinoma (nonsmokers)	4.50E-03	1.61E-03	1.65E-03	3.46E-06	3.46E-06	4.32E-08	9.80E-08
17	Lung adenocarcinoma (smokers)	8.10E-02	2.67E-02	2.69E-02	9.35E-06	9.35E-06	4.32E-08	9.80E-08
19	Melanoma	2.03E-02	1.87E-02	2.13E-02	6.29E-07	6.29E-07	5.17E-05	1.19E-04
20	Osteosarcoma	3.50E-04	2.45E-04	2.46E-04	1.61E-05	1.61E-05	4.83E-11	1.10E-10
21	Osteosarcoma of the arms	4.00E-05	2.98E-05	3.01E-05	1.34E-05	1.34E-05	1.02E-11	2.32E-11
22	Osteosarcoma of the head	3.02E-05	2.21E-05	2.23E-05	1.22E-05	1.22E-05	1.02E-11	2.32E-11
23	Osteosarcoma of the legs	2.20E-04	1.66E-04	1.67E-04	1.80E-05	1.80E-05	2.22E-11	5.08E-11
24	Osteosarcoma of the pelvis	3.00E-05	2.43E-05	2.45E-05	1.60E-05	1.60E-05	4.63E-12	1.06E-11
26	Pancreatic ductal adenocarcinoma	1.36E-02	1.14E-02	1.22E-02	1.21E-06	1.21E-06	5.68E-06	1.31E-05
27	Pancreatic endocrine (islet cell) carcinoma	1.94E-04	1.83E-04	2.01E-04	9.57E-07	9.57E-07	1.59E-07	3.69E-07
30	Thyroid papillary/follicular carcinoma	1.03E-02	5.08E-03	5.11E-03	1.61E-05	1.61E-05	1.42E-09	3.24E-09
31	Thyroid medullary carcinoma	3.24E-04	1.94E-04	1.96E-04	1.12E-05	1.12E-05	1.46E-10	3.33E-10

Table 10: Estimated intrinsic contribution percentage for selected cancer types with two settings of mutation rate enlargement (as an effect of mutation M_4) factor $\alpha^{(1)} = 1$ and $\alpha^{(2)} = 2$. We use intrinsic mutation rate 10^{-7}.
"$\hat{C}_{int}$ all cells" gives the estimated intrinsic contribution based on the ratio of number of acquired mutations among all cells due to u_{int} vs. $(u_{int} + u_{exc})$, while "$\hat{C}_{int}$ cancer cells" gives the same quantity considering only cancer cells, i.e. those cells acquired sufficient types of driver mutations for cancer onset.
Note:
***1. In cases where $\hat{R}_{int}$ is already bigger than R_{obs}, which usually occurs for tissues with very long stem lineages, we could have an intrinsic contribution larger than 100%. This indicates only a smaller intrinsic mutation rate could be reasonable and therefore we ignore the interpretations for this situation.**

id	cancer type	$\hat{C}_{int}^{(1)}$ all cells	$\hat{C}_{int}^{(2)}$ all cells	$\hat{C}_{int}^{(1)}$ cancer cells	$\hat{C}_{int}^{(2)}$ cancer cells
1	Acute myeloid leukemia	56.92%	56.92%	12.17%	19.03%
2	Basal cell carcinoma	18.50%	18.50%	0.63%	1.22%
3	Chronic lymphocytic leukemia	53.74%	53.74%	9.79%	15.87%
4	Colorectal adenocarcinoma	103.28%*1	103.28%*1	110.48%*1	107.90%*1
5	Colorectal adenocarcinoma with FAP	0.02%	0.02%	0.00%	0.00%
7	Duodenum adenocarcinoma	47.41%	64.32%	10.30%	34.31%
10	Gallbladder non papillary adenocarcinoma	0.62%	0.62%	0.00%	0.00%
14	Hepatocellular carcinoma	10.13%	10.13%	0.10%	0.22%
15	Hepatocellular carcinoma with HCV	4.65%	4.65%	0.01%	0.02%
16	Lung adenocarcinoma (nonsmokers)	2.81%	2.81%	0.00%	0.00%
17	Lung adenocarcinoma (smokers)	1.06%	1.06%	0.00%	0.00%
19	Melanoma	13.72%	13.72%	0.26%	0.53%
20	Osteosarcoma	0.62%	0.62%	0.00%	0.00%
21	Osteosarcoma of the arms	0.74%	0.74%	0.00%	0.00%
22	Osteosarcoma of the head	0.82%	0.82%	0.00%	0.00%
23	Osteosarcoma of the legs	0.55%	0.55%	0.00%	0.00%
24	Osteosarcoma of the pelvis	0.62%	0.62%	0.00%	0.00%
26	Pancreatic ductal adenocarcinoma	7.61%	7.61%	0.04%	0.09%
27	Pancreatic endocrine (islet cell) carcinoma	9.46%	9.46%	0.08%	0.18%
30	Thyroid papillary/follicular carcinoma	0.62%	0.62%	0.00%	0.00%
31	Thyroid medullary carcinoma	0.89%	0.89%	0.00%	0.00%

Again, we see from Table 10 that $\hat{C}_{int}$ from all cells displays no difference between two conditions while $\hat{C}_{int}$ from only cancer cells nearly doubled when applying mutation rate enlargement effect. We see that $\hat{C}_{int}$ from only cancer cells is below 10% for most cancer types; the largest intrinsic contribution here is 34.31% for Duodenum adenocarcinoma

with mutation enlargement effect. Again, under this parameter setting most of the cancer risk should be due to non-intrinsic factors.

5.1.2.3 Clonal Expansion Factors

A very typical characteristic in the process of cancer development is clonal expansion, in which certain mutations could give the cell the ability to divide faster and expand in population. Clonal expansion expedites mutation acquisition and therefore increases the chance of cancer onset.

We assume any cell with M_5 divide faster by a factor of α than otherwise. In our experiment, we restrict the clonal expansion effect to take place only after stem cell symmetric division stage (the first 40 weeks on average). To incorporate the regulation procedure, once the total cell number grows larger than the homeostatic cell number, we stop the clonal expansion. Since the regulation mechanism is hardly well defined, we will remove the regulation effect based on total cell population in Section 5.2, to see the maximum possible intrinsic risk and contribution under certain clonal expansion factors.

We compare the results with γ (no clonal expansion) and $\gamma = 2$ with all other parameter configurations considering different intrinsic mutation rates 10^{-8} and 10^{-7}, mutation rate enlargement factor $\alpha = 1$ and $\alpha = 2$. We summarize the results in Tables 11 - 18 below.

Tables 11 and 12 provide the results with intrinsic mutation rate $\mathbf{10^{-8}}$ and mutation enlargement factor 1.

Table 11: Estimated excess mutation rate, total risk and intrinsic risk for selected cancer types with two settings of clonal expansion (as an effect of mutation M_5) factor $\gamma^{(1)} = 1$ and $\gamma^{(2)} = 2$.
We use intrinsic mutation rate 10^{-8} and mutation enlargement factor 1.
R_{obs} is the observed risk, $\hat{u}_{exc}^{(1)}, \hat{u}_{exc}^{(2)}$ represent estimated excess rate with two settings respectively. Similarly, $\hat{R}_{total}$ and $\hat{R}_{int}$ represent estimated total risk (from mutation rate $u_{int} + \hat{u}_{exc}$), and intrinsic risk (from mutation rate u_{int}).
Note:
***1. In cases where $\hat{R}_{total}$ is much higher than R_{obs} with originally estimated $\hat{u}_{exc}$, we take $\hat{u}_{exc}/2$ instead.**
***2. This is a special case where observed risk is 1.0, in which our algorithm for estimation $\hat{u}_{exc}$ does not apply.**
***3. In cases where $\hat{R}_{total}$ is much higher than R_{obs} with originally estimated $\hat{u}_{ext}$; and even with $\hat{u}_{exc}/2$ taken as excess rate, $\hat{R}_{total}$ is still ~ 2 times larger than R_{obs}, we mark these particular cancer types/settings and use the intrinsic contribution values just as a reference but not solid evidence for making conclusions.**

id	cancer type	R_{obs}	$\hat{R}_{total}^{(1)}$	$\hat{R}_{total}^{(2)}$	$\hat{u}_{ext}^{(1)}$	$\hat{u}_{ext}^{(2)}$	$\hat{R}_{int}^{(1)}$	$\hat{R}_{int}^{(2)}$
1	Acute myeloid leukemia	4.10E-03	1.26E-03*1	1.78E-03*1	1.21E-07	1.21E-07	2.91E-07	3.38E-05
2	Basal cell carcinoma	3.00E-01	2.92E-01	1.08E-01*1	5.30E-07	2.65E-07	2.26E-06	5.26E-06
3	Chronic lymphocytic leukemia	5.20E-03	1.63E-03*1	2.19E-03*1	1.31E-07	1.31E-07	2.91E-07	3.38E-05
4	Colorectal adenocarcinoma	4.80E-02	5.04E-02	4.17E-02*1	8.68E-08	4.34E-08	5.45E-05	1.78E-04
5	Colorectal adenocarcinoma with FAP	1.00E+00	1.00E+00*2	1.00E+00*2	1.00E+00	1.00E+00	5.45E-05	1.78E-04
7	Duodenum adenocarcinoma	3.00E-04	3.12E-04	2.04E-04*1	2.01E-07	1.00E-07	3.21E-08	8.62E-08
10	Gallbladder non papillary adenocarcinoma	2.80E-03	2.91E-03	3.11E-03*1	1.62E-05	8.10E-06	6.44E-13	1.41E-12
14	Hepatocellular carcinoma	7.10E-03	6.13E-03	2.03E-03*1	9.77E-07	4.88E-07	7.16E-09	1.60E-08
15	Hepatocellular carcinoma with HCV	7.10E-02	5.90E-02	2.19E-02*1	2.14E-06	1.07E-06	7.16E-09	1.60E-08
16	Lung adenocarcinoma (nonsmokers)	4.50E-03	1.61E-03	1.10E-03*1	3.55E-06	1.77E-06	4.55E-11	7.25E-11
17	Lung adenocarcinoma (smokers)	8.10E-02	2.67E-02	3.35E-02*1	9.44E-06	4.72E-06	4.55E-11	7.25E-11
19	Melanoma	2.03E-02	1.87E-02	1.87E-02	7.19E-07	7.19E-07	5.24E-08	5.24E-08
20	Osteosarcoma	3.50E-04	2.45E-04	1.47E-03*3	1.62E-05	8.09E-06	4.85E-14	8.40E-14
21	Osteosarcoma of the arms	4.00E-05	2.98E-05	1.73E-04*3	1.35E-05	6.73E-06	1.02E-14	1.77E-14
22	Osteosarcoma of the head	3.02E-05	2.21E-05	1.21E-04*3	1.23E-05	6.13E-06	1.02E-14	1.77E-14
23	Osteosarcoma of the legs	2.20E-04	1.66E-04	1.09E-03*3	1.81E-05	9.07E-06	2.22E-14	3.87E-14
24	Osteosarcoma of the pelvis	3.00E-05	2.43E-05	1.60E-04*3	1.60E-05	8.02E-06	4.62E-15	8.11E-15
26	Pancreatic ductal adenocarcinoma	1.36E-02	1.14E-02	3.45E-03*1	1.30E-06	6.52E-07	5.79E-09	1.28E-08
27	Pancreatic endocrine (islet cell) carcinoma	1.94E-04	1.83E-04	5.47E-05*1	1.05E-06	5.23E-07	1.61E-10	3.58E-10
30	Thyroid papillary/follicular carcinoma	1.03E-02	5.08E-03	3.68E-03*1	1.62E-05	8.11E-06	1.46E-12	2.45E-12
31	Thyroid medullary carcinoma	3.24E-04	1.94E-04	1.17E-04*1	1.13E-05	5.63E-06	1.48E-13	2.50E-13

Table 12: Estimated intrinsic contribution percentage for selected cancer types with two settings of mutation rate enlargement (as an effect of mutation M_4) factor $\alpha^{(1)} = 1$ and $\alpha^{(2)} = 2$.
We use intrinsic mutation rate 10^{-8} and mutation enlargement factor 1.
"$\hat{C}_{int}$ all cells" gives the estimated intrinsic contribution based on the ratio of number of acquired mutations among all cells due to u_{int} vs. $(u_{int} + u_{ext})$, while "$\hat{C}_{int}$ cancer cells" gives the same quantity considering only cancer cells, i.e. those cells acquired sufficient types of driver mutations for cancer onset.

id	cancer type	$\hat{C}_{int}^{(1)}$ all cells	$\hat{C}_{int}^{(2)}$ all cells	$\hat{C}_{int}^{(1)}$ cancer cells	$\hat{C}_{int}^{(2)}$ cancer cells
1	Acute myeloid leukemia	7.65%	7.65%	1.43E-04	3.00E-04
2	Basal cell carcinoma	1.85%	3.63%	6.26E-06	4.13E-05
3	Chronic lymphocytic leukemia	7.09%	7.09%	1.08E-04	2.25E-04
4	Colorectal adenocarcinoma	10.32%	18.70%	1.02E-03	2.14E-03
5	Colorectal adenocarcinoma with FAP	0.00%	0.07%	9.99E-16	4.33E-15
7	Duodenum adenocarcinoma	4.74%	9.05%	1.00E-04	2.21E-04
10	Gallbladder non papillary adenocarcinoma	0.06%	0.12%	1.98E-10	2.07E-10
14	Hepatocellular carcinoma	1.01%	2.01%	1.03E-06	6.63E-06
15	Hepatocellular carcinoma with HCV	0.47%	0.93%	9.86E-08	5.37E-07
16	Lung adenocarcinoma (nonsmokers)	0.28%	0.56%	2.20E-08	1.27E-07
17	Lung adenocarcinoma (smokers)	0.11%	0.21%	1.15E-09	4.64E-09
19	Melanoma	1.37%	1.37%	2.58E-06	2.58E-06
20	Osteosarcoma	0.06%	0.12%	2.32E-10	1.46E-09
21	Osteosarcoma of the arms	0.07%	0.15%	4.05E-10	2.67E-09
22	Osteosarcoma of the head	0.08%	0.16%	5.37E-10	3.59E-09
23	Osteosarcoma of the legs	0.06%	0.11%	1.61E-10	8.09E-10
24	Osteosarcoma of the pelvis	0.06%	0.12%	2.38E-10	1.53E-09
26	Pancreatic ductal adenocarcinoma	0.76%	1.51%	4.38E-07	3.23E-06
27	Pancreatic endocrine (islet cell) carcinoma	0.95%	1.87%	8.44E-07	6.27E-06
30	Thyroid papillary/follicular carcinoma	0.06%	0.12%	2.16E-10	5.04E-10
31	Thyroid medullary carcinoma	0.09%	0.18%	6.65E-10	2.06E-09

We could see that none of intrinsic contribution values from only cancer cells exceed 1% and the highest intrinsic contribution from all cells is 18.70%. In addition, $\hat{C}_{int}$ increased for all cancer types when clonal expansion takes effects, except Melanoma. With the constraints considering only cancer cells, the shift on $\hat{C}_{int}$ became more significant as cancer cells are guaranteed to already have mutation M_5 so all of them divide faster with the clonal expansion factor.

According to Table 3, all cells in Melanoma are stem cells; therefore our model assumes no progenitor lineages in the entire lifespan. Therefore clonal expansion for Melanoma only increases the number of divisions for a very small portion of stem cells and does not expand the cell population. Since we take the upper bound $\lceil \log_2 N_S^H \rceil$ as stem cell symmetric division generations, the computed cell number from our model 4.29E+09 is slightly larger than the given number 3.80E+09. Therefore according to the regulation mechanism the clonal expansion did not take place.

Tables 13 and 14 below provide results with intrinsic mutation rate $\mathbf{10^{-8}}$ and mutation enlargement factor 2.

Table 13: Estimated excess mutation rate, total risk and intrinsic risk for selected cancer types with two settings of clonal expansion (as an effect of mutation M_5) factor $\gamma^{(1)} = 1$ and $\gamma^{(2)} = 2$. We use intrinsic mutation rate 10^{-8} and mutation enlargement factor 2. R_{obs} is the observed risk, $\hat{u}_{\text{exc}}^{(1)}, \hat{u}_{\text{exc}}^{(2)}$ represent estimated excess rate with two settings respectively. Similarly, $\hat{R}_{\text{total}}$ and $\hat{R}_{\text{int}}$ represent estimated total risk (from mutation rate $u_{\text{int}} + \hat{u}_{\text{exc}}$), and intrinsic risk (from mutation rate u_{int}).
Note:
***1. In cases where $\hat{R}_{\text{total}}$ is much higher than R_{obs} with originally estimated $\hat{u}_{\text{exc}}$, we take $\hat{u}_{\text{exc}}/2$ instead.**
***2. This is a special case where observed risk is 1.0, in which our algorithm for estimation $\hat{u}_{\text{exc}}$ does not apply.**
***3. In cases where $\hat{R}_{\text{total}}$ is much higher than R_{obs} with originally estimated $\hat{u}_{\text{exc}}$; and even with $\hat{u}_{\text{ext}}/2$ taken as excess rate, $\hat{R}_{\text{total}}$ is still ~ 2 times larger than R_{obs}, we mark these particular cancer types/settings and use the intrinsic contribution values just as a reference but not solid evidence for making conclusions.**

id	cancer type	R_{obs}	$\hat{R}_{\text{total}}^{(1)}$	$\hat{R}_{\text{total}}^{(2)}$	$\hat{u}_{\text{exc}}^{(1)}$	$\hat{u}_{\text{exc}}^{(2)}$	$\hat{R}_{\text{int}}^{(1)}$	$\hat{R}_{\text{int}}^{(2)}$
1	Acute myeloid leukemia	4.10E-03	1.37E-03*1	1.93E-03*1	1.21E-07	1.21E-07	7.13E-07	9.01E-05
2	Basal cell carcinoma	3.00E-01	2.96E-01	1.11E-01*1	5.30E-07	2.65E-07	5.26E-06	1.05E-05
3	Chronic lymphocytic leukemia	5.20E-03	1.76E-03*1	2.35E-03*1	1.31E-07	1.31E-07	7.13E-07	9.01E-05
4	Colorectal adenocarcinoma	4.80E-02	5.57E-02	4.72E-02*1	8.68E-08	4.34E-08	1.27E-04	3.66E-04
5	Colorectal adenocarcinoma with FAP	1.00E+00	1.00E+00*2	1.00E+00*2	1.00E+00	1.00E+00	1.27E-04	3.66E-04
7	Duodenum adenocarcinoma	3.00E-04	3.27E-04	2.18E-04*1	2.01E-07	1.00E-07	7.49E-08	1.77E-07
10	Gallbladder non papillary adenocarcinoma	2.80E-03	2.91E-03	3.11E-03*1	1.62E-05	8.10E-06	1.49E-12	2.85E-12
14	Hepatocellular carcinoma	7.10E-03	6.19E-03	2.06E-03*1	9.77E-07	4.88E-07	1.66E-08	3.21E-08
15	Hepatocellular carcinoma with HCV	7.10E-02	5.92E-02	2.20E-02*1	2.14E-06	1.07E-06	1.66E-08	3.21E-08

16	Lung adenocarcinoma (nonsmokers)	4.50E-03	1.62E-03	1.11E-03[*1]	3.55E-06	1.77E-06	1.04E-10	1.56E-10
17	Lung adenocarcinoma (smokers)	8.10E-02	2.67E-02	3.36E-02[*1]	9.44E-06	4.72E-06	1.04E-10	1.56E-10
19	Melanoma	2.03E-02	1.89E-02	1.89E-02	7.19E-07	7.19E-07	1.22E-07	1.22E-07
20	Osteosarcoma	3.50E-04	2.45E-04	1.47E-03[*3]	1.62E-05	8.09E-06	1.11E-13	1.83E-13
21	Osteosarcoma of the arms	4.00E-05	2.99E-05	1.73E-04[*3]	1.35E-05	6.73E-06	2.32E-14	3.86E-14
22	Osteosarcoma of the head	3.02E-05	2.22E-05	1.21E-04[*3]	1.23E-05	6.13E-06	2.32E-14	3.86E-14
23	Osteosarcoma of the legs	2.20E-04	1.66E-04	1.10E-03[*3]	1.81E-05	9.07E-06	5.09E-14	8.42E-14
24	Osteosarcoma of the pelvis	3.00E-05	2.43E-05	1.60E-04[*3]	1.60E-05	8.02E-06	1.06E-14	1.76E-14
26	Pancreatic ductal adenocarcinoma	1.36E-02	1.15E-02	3.49E-03[*1]	1.30E-06	6.52E-07	1.34E-08	2.57E-08
27	Pancreatic endocrine (islet cell) carcinoma	1.94E-04	1.85E-04	5.56E-05[*1]	1.05E-06	5.23E-07	3.73E-10	7.20E-10
30	Thyroid papillary/follicular carcinoma	1.03E-02	5.08E-03	3.69E-03[*1]	1.62E-05	8.11E-06	3.35E-12	5.20E-12
31	Thyroid medullary carcinoma	3.24E-04	1.95E-04	1.17E-04[*1]	1.13E-05	5.63E-06	3.38E-13	5.30E-13

Table 14: Estimated intrinsic contribution percentage for selected cancer types with two settings of mutation rate enlargement (as an effect of mutation M_4) factor $\alpha^{(1)} = 1$ and $\alpha^{(2)} = 2$.
We use intrinsic mutation rate 10^{-8} and mutation enlargement factor 2.
"$\hat{C}_{int}$ all cells" gives the estimated intrinsic contribution based on the ratio of number of acquired mutations among all cells due to u_{int} vs. $(u_{int} + u_{ext})$, while "$\hat{C}_{int}$ cancer cells" gives the same quantity considering only cancer cells, i.e. those cells acquired sufficient types of driver mutations for cancer onset.

id	cancer type	$\hat{C}_{int}^{(1)}$ all cells	$\hat{C}_{int}^{(2)}$ all cells	$\hat{C}_{int}^{(1)}$ cancer cells	$\hat{C}_{int}^{(2)}$ cancer cells
1	Acute myeloid leukemia	7.65%	7.65%	3.33E-04	6.23E-04
2	Basal cell carcinoma	1.85%	3.63%	1.43E-05	8.03E-05
3	Chronic lymphocytic leukemia	7.09%	7.09%	2.52E-04	4.71E-04
4	Colorectal adenocarcinoma	10.32%	18.70%	2.14E-03	4.35E-03
5	Colorectal adenocarcinoma with FAP	0.00%	0.07%	2.22E-15	1.08E-14
7	Duodenum adenocarcinoma	4.74%	9.05%	2.23E-04	4.51E-04
10	Gallbladder non papillary adenocarcinoma	0.06%	0.12%	4.58E-10	4.20E-10
14	Hepatocellular carcinoma	1.01%	2.01%	2.37E-06	1.32E-05
15	Hepatocellular carcinoma with HCV	0.47%	0.93%	2.28E-07	1.08E-06
16	Lung adenocarcinoma (nonsmokers)	0.28%	0.56%	5.02E-08	2.79E-07
17	Lung adenocarcinoma (smokers)	0.11%	0.21%	2.64E-09	1.02E-08
19	Melanoma	1.37%	1.37%	5.92E-06	5.92E-06
20	Osteosarcoma	0.06%	0.12%	5.28E-10	3.17E-09
21	Osteosarcoma of the arms	0.07%	0.15%	9.18E-10	5.77E-09
22	Osteosarcoma of the head	0.08%	0.16%	1.22E-09	7.76E-09
23	Osteosarcoma of the legs	0.06%	0.11%	3.66E-10	1.76E-09
24	Osteosarcoma of the pelvis	0.06%	0.12%	5.40E-10	3.31E-09
26	Pancreatic ductal adenocarcinoma	0.76%	1.51%	1.01E-06	6.45E-06

27	Pancreatic endocrine (islet cell) carcinoma	0.95%	1.87%	1.94E-06	1.24E-05
30	Thyroid papillary/follicular carcinoma	0.06%	0.12%	4.95E-10	1.09E-09
31	Thyroid medullary carcinoma	0.09%	0.18%	1.52E-09	4.43E-09

The two tables above demonstrate same conclusions as in Tables 11 and 12.

Tables 15 and 16 below provide results with intrinsic mutation rate $\mathbf{10^{-7}}$ and mutation enlargement factor 1.

Table 15: Estimated excess mutation rate, total risk and intrinsic risk for selected cancer types with two settings of clonal expansion (as an effect of mutation M_5) factor $\gamma^{(1)} = 1$ and $\gamma^{(2)} = 2$.
We use intrinsic mutation rate 10^{-7} and mutation enlargement factor 1.
R_{obs} is the observed risk, $\hat{u}_{\text{exc}}^{(1)}, \hat{u}_{\text{exc}}^{(2)}$ represent estimated excess rate with two settings respectively. Similarly, $\hat{R}_{\text{total}}$ and $\hat{R}_{\text{int}}$ represent estimated total risk (from mutation rate $u_{\text{int}} + \hat{u}_{\text{exc}}$), and intrinsic risk (from mutation rate u_{int}).
Note:
***1. In cases where $\hat{R}_{\text{total}}$ is much higher than R_{obs} with originally estimated $\hat{u}_{\text{exc}}$, we take $\hat{u}_{\text{exc}}/2$ instead.**
***2. This is a special case where observed risk is 1.0, in which our algorithm for estimation $\hat{u}_{\text{exc}}$ does not apply.**
***3. In cases where $\hat{R}_{\text{total}}$ is much higher than R_{obs} with originally estimated $\hat{u}_{\text{exc}}$; and even with $\hat{u}_{\text{exc}}/2$ taken as excess rate, $\hat{R}_{\text{total}}$ is still ~ 2 times larger than R_{obs}, we mark these particular cancer types/settings and use the intrinsic contribution values just as a reference but not solid evidence for making conclusions.**
***4. In cases where $\hat{R}_{\text{int}}$ is already bigger than R_{obs}, which usually occurs for tissues with very long stem lineages, we could have a negative $\hat{u}_{\text{exc}}$. This indicates only a smaller intrinsic mutation rate could be reasonable and therefore we ignore the interpretations for this situation.**

id	cancer type	R_{obs}	$\hat{R}_{\text{total}}^{(1)}$	$\hat{R}_{\text{total}}^{(2)}$	$\hat{u}_{\text{exc}}^{(1)}$	$\hat{u}_{\text{exc}}^{(2)}$	$\hat{R}_{\text{int}}^{(1)}$	$\hat{R}_{\text{int}}^{(2)}$
1	Acute myeloid leukemia	4.10E-03	3.42E-03*1	4.10E-03*1	7.57E-08	7.57E-08	5.12E-04	8.90E-04
2	Basal cell carcinoma	3.00E-01	2.92E-01	1.67E-01*1	4.40E-07	2.20E-07	2.22E-03	5.29E-03
3	Chronic lymphocytic leukemia	5.20E-03	4.15E-03*1	4.89E-03*1	8.61E-08	8.61E-08	5.12E-04	8.90E-04
4	Colorectal adenocarcinoma	4.80E-02	5.04E-02	2.49E-01	-3.18E-09*4	-3.18E-09*4	5.54E-02	2.72E-01
5	Colorectal adenocarcinoma with FAP	1.00E+00*2	1.00E+00	1.00E+00	1.00E+00	1.00E+00	5.54E-02	2.72E-01
7	Duodenum adenocarcinoma	3.00E-04	3.12E-04	6.23E-04*3	1.11E-07	5.54E-08	3.26E-05	1.47E-04
10	Gallbladder non papillary adenocarcinoma	2.80E-03	2.91E-03	3.17E-03*1	1.61E-05	8.05E-06	6.42E-10	1.50E-09
14	Hepatocellular carcinoma	7.10E-03	6.13E-03	2.64E-03*1	8.87E-07	4.43E-07	7.04E-06	1.60E-05
15	Hepatocellular carcinoma with HCV	7.10E-02	5.90E-02	2.48E-02*1	2.05E-06	1.02E-06	7.04E-06	1.60E-05
16	Lung adenocarcinoma (nonsmokers)	4.50E-03	1.61E-03	1.20E-03*1	3.46E-06	1.73E-06	4.32E-08	7.86E-08
17	Lung adenocarcinoma (smokers)	8.10E-02	2.67E-02	3.46E-02*1	9.35E-06	4.67E-06	4.32E-08	7.86E-08
19	Melanoma	2.03E-02	1.87E-02	1.87E-02	6.29E-07	6.29E-07	5.17E-05	5.17E-05

20	Osteosarcoma	3.50E-04	2.45E-04	1.50E-03[*1]	1.61E-05	8.04E-06	4.83E-11	1.33E-10
21	Osteosarcoma of the arms	4.00E-05	2.98E-05	1.77E-04[*3]	1.34E-05	6.68E-06	1.02E-11	2.85E-11
22	Osteosarcoma of the head	3.02E-05	2.21E-05	1.24E-04[*3]	1.22E-05	6.08E-06	1.02E-11	2.85E-11
23	Osteosarcoma of the legs	2.20E-04	1.66E-04	1.11E-03[*3]	1.80E-05	9.02E-06	2.22E-11	6.17E-11
24	Osteosarcoma of the pelvis	3.00E-05	2.43E-05	1.63E-04[*3]	1.60E-05	7.98E-06	4.63E-12	1.31E-11
26	Pancreatic ductal adenocarcinoma	1.36E-02	1.14E-02	4.19E-03[*3]	1.21E-06	6.07E-07	5.68E-06	1.26E-05
27	Pancreatic endocrine (islet cell) carcinoma	1.94E-04	1.83E-04	6.99E-05[*1]	9.57E-07	4.78E-07	1.59E-07	3.57E-07
30	Thyroid papillary/follicular carcinoma	1.03E-02	5.08E-03	3.76E-03[*1]	1.61E-05	8.07E-06	1.42E-09	2.47E-09
31	Thyroid medullary carcinoma	3.24E-04	1.94E-04	1.20E-04[*1]	1.12E-05	5.59E-06	1.46E-10	2.56E-10

Table 16: Estimated intrinsic contribution percentage for selected cancer types with two settings of mutation rate enlargement (as an effect of mutation M_4) factor $\alpha^{(1)} = 1$ and $\alpha^{(2)} = 2$.
We use intrinsic mutation rate 10^{-7} and mutation enlargement factor 1.
"$\hat{C}_{int}$ all cells" gives the estimated intrinsic contribution based on the ratio of number of acquired mutations among all cells due to u_{int} vs. $(u_{int} + u_{ext})$, while "$\hat{C}_{int}$ cancer cells" gives the same quantity considering only cancer cells, i.e. those cells acquired sufficient types of driver mutations for cancer onset.

id	cancer type	$\hat{C}_{int}^{(1)}$ all cells	$\hat{C}_{int}^{(2)}$ all cells	$\hat{C}_{int}^{(1)}$ cancer cells	$\hat{C}_{int}^{(2)}$ cancer cells
1	Acute myeloid leukemia	56.92%	56.92%	12.17%	12.27%
2	Basal cell carcinoma	18.50%	31.21%	0.63%	2.70%
3	Chronic lymphocytic leukemia	53.74%	53.74%	9.79%	9.87%
4	Colorectal adenocarcinoma	103.28%	103.29%	110.48%	113.41%
5	Colorectal adenocarcinoma with FAP	0.02%	0.68%	0.00%	0.00%
7	Duodenum adenocarcinoma	47.41%	64.30%	10.30%	18.78%
10	Gallbladder non papillary adenocarcinoma	0.62%	1.22%	0.00%	0.00%
14	Hepatocellular carcinoma	10.13%	18.40%	0.10%	0.52%
15	Hepatocellular carcinoma with HCV	4.65%	8.89%	0.01%	0.05%
16	Lung adenocarcinoma (nonsmokers)	2.81%	5.47%	0.00%	0.01%
17	Lung adenocarcinoma (smokers)	1.06%	2.09%	0.00%	0.00%
19	Melanoma	13.72%	13.72%	0.26%	0.26%
20	Osteosarcoma	0.62%	1.23%	0.00%	0.00%
21	Osteosarcoma of the arms	0.74%	1.47%	0.00%	0.00%
22	Osteosarcoma of the head	0.82%	1.62%	0.00%	0.00%
23	Osteosarcoma of the legs	0.55%	1.10%	0.00%	0.00%
24	Osteosarcoma of the pelvis	0.62%	1.24%	0.00%	0.00%
26	Pancreatic ductal adenocarcinoma	7.61%	14.14%	0.04%	0.27%
27	Pancreatic endocrine (islet cell) carcinoma	9.46%	17.28%	0.08%	0.49%
30	Thyroid papillary/follicular carcinoma	0.62%	1.22%	0.00%	0.00%
31	Thyroid medullary carcinoma	0.89%	1.76%	0.00%	0.00%

We ignore the results for cancer types 4 and 5 due to the reasons stated at Table 15, then we can see that the maximum $\hat{C}_{\text{int}}$ from only cancer cells is ~18%.

Tables 17 and 18 below provide results with intrinsic mutation rate $\mathbf{10^{-7}}$ and mutation enlargement factor 2.

Table 17: Estimated excess mutation rate, total risk and intrinsic risk for selected cancer types with two settings of clonal expansion (as an effect of mutation M_5) factor $\gamma^{(1)} = 1$ and $\gamma^{(2)} = 2$.
We use intrinsic mutation rate 10^{-7} and mutation enlargement factor 2.
R_{obs} is the observed risk, $\hat{u}_{\text{exc}}^{(1)}, \hat{u}_{\text{exc}}^{(2)}$ represent estimated excess rate with two settings respectively. Similarly, $\hat{R}_{\text{total}}$ and $\hat{R}_{\text{int}}$ represent estimated total risk (from mutation rate $u_{\text{int}} + \hat{u}_{\text{exc}}$), and intrinsic risk (from mutation rate u_{int}).
Note:
***1. In cases where $\hat{R}_{\text{total}}$ is much higher than R_{obs} with originally estimated $\hat{u}_{\text{exc}}$, we take $\hat{u}_{\text{ext}}/2$ instead.**
***2. This is a special case where observed risk is 1.0, in which our algorithm for estimation $\hat{u}_{\text{exc}}$ does not apply.**
***3. In cases where $\hat{R}_{\text{total}}$ is much higher than R_{obs} with originally estimated $\hat{u}_{\text{exc}}$; and even with $\hat{u}_{\text{exc}}/2$ taken as excess rate, $\hat{R}_{\text{total}}$ is still ~ 2 times larger than R_{obs}, we mark these particular cancer types/settings and use the intrinsic contribution values just as a reference but not solid evidence for making conclusions.**
***4. In cases where $\hat{R}_{\text{int}}$ is already bigger than R_{obs}, which usually occurs for tissues with very long stem lineages, we could have a negative $\hat{u}_{\text{ext}}$. This indicates only a smaller intrinsic mutation rate could be reasonable and therefore we ignore the interpretations for this situation.**

id	cancer type	R_{obs}	$\hat{R}_{\text{total}}^{(1)}$	$\hat{R}_{\text{total}}^{(2)}$	$\hat{u}_{\text{exc}}^{(1)}$	$\hat{u}_{\text{exc}}^{(2)}$	$\hat{R}_{\text{int}}^{(1)}$	$\hat{R}_{\text{int}}^{(2)}$
1	Acute myeloid leukemia	4.10E-03	5.97E-03*1	7.11E-03*1	7.57E-08	7.57E-08	1.32E-03	2.27E-03
2	Basal cell carcinoma	3.00E-01	3.38E-01	2.07E-01*1	4.40E-07	2.20E-07	5.13E-03	1.06E-02
3	Chronic lymphocytic leukemia	5.20E-03	7.02E-03*1	8.24E-03*1	8.61E-08	8.61E-08	1.32E-03	2.27E-03
4	Colorectal adenocarcinoma	4.80E-02	1.19E-01	4.11E-01	-3.18E-09*4	-3.18E-09*4	1.27E-01	4.37E-01
5	Colorectal adenocarcinoma with FAP	1.00E+00*2	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.27E-01	4.37E-01
7	Duodenum adenocarcinoma	3.00E-04	4.86E-04	9.54E-04*3	1.11E-07	5.54E-08	7.67E-05	2.88E-04
10	Gallbladder non papillary adenocarcinoma	2.80E-03	2.93E-03	3.20E-03*1	1.61E-05	8.05E-06	1.48E-09	3.08E-09
14	Hepatocellular carcinoma	7.10E-03	6.74E-03	3.05E-03*1	8.87E-07	4.43E-07	1.62E-05	3.21E-05
15	Hepatocellular carcinoma with HCV	7.10E-02	6.16E-02	2.66E-02*1	2.05E-06	1.02E-06	1.62E-05	3.21E-05
16	Lung adenocarcinoma (nonsmokers)	4.50E-03	1.65E-03	1.27E-03*1	3.46E-06	1.73E-06	9.80E-08	1.73E-07
17	Lung adenocarcinoma (smokers)	8.10E-02	2.69E-02	3.53E-02*1	9.35E-06	4.67E-06	9.80E-08	1.73E-07
19	Melanoma	2.03E-02	2.13E-02	2.13E-02	6.29E-07	6.29E-07	1.19E-04	1.19E-04
20	Osteosarcoma	3.50E-04	2.46E-04	1.52E-03*3	1.61E-05	8.04E-06	1.10E-10	3.16E-10
21	Osteosarcoma of the arms	4.00E-05	3.01E-05	1.80E-04*3	1.34E-05	6.68E-06	2.32E-11	6.76E-11

22	Osteosarcoma of the head	3.02E-05	2.23E-05	1.26E-04[*3]	1.22E-05	6.08E-06	2.32E-11	6.76E-11
23	Osteosarcoma of the legs	2.20E-04	1.67E-04	1.13E-03[*3]	1.80E-05	9.02E-06	5.08E-11	1.47E-10
24	Osteosarcoma of the pelvis	3.00E-05	2.45E-05	1.65E-04[*3]	1.60E-05	7.98E-06	1.06E-11	3.10E-11
26	Pancreatic ductal adenocarcinoma	1.36E-02	1.22E-02	4.67E-03[*1]	1.21E-06	6.07E-07	1.31E-05	2.52E-05
27	Pancreatic endocrine (islet cell) carcinoma	1.94E-04	2.01E-04	7.98E-05[*1]	9.57E-07	4.78E-07	3.69E-07	7.18E-07
30	Thyroid papillary/follicular carcinoma	1.03E-02	5.11E-03	3.80E-03[*1]	1.61E-05	8.07E-06	3.24E-09	5.26E-09
31	Thyroid medullary carcinoma	3.24E-04	1.96E-04	1.22E-04[*1]	1.12E-05	5.59E-06	3.33E-10	5.45E-10

Table 18: Estimated intrinsic contribution percentage for selected cancer types with two settings of mutation rate enlargement (as an effect of mutation M_4) factor $\alpha^{(1)} = 1$ and $\alpha^{(2)} = 2$.
We use intrinsic mutation rate 10^{-7} and mutation enlargement factor 2.
"$\hat{C}_{int}$ all cells" gives the estimated intrinsic contribution based on the ratio of number of acquired mutations among all cells due to u_{int} vs. $(u_{int} + u_{ext})$, while "$\hat{C}_{int}$ cancer cells" gives the same quantity considering only cancer cells, i.e. those cells acquired sufficient types of driver mutations for cancer onset.

id	cancer type	$\widehat{C}_{int}^{(1)}$ all cells	$\widehat{C}_{int}^{(2)}$ all cells	$\widehat{C}_{int}^{(1)}$ cancer cells	$\widehat{C}_{int}^{(2)}$ cancer cells
1	Acute myeloid leukemia	56.92%	56.92%	19.03%	19.12%
2	Basal cell carcinoma	18.50%	31.21%	1.22%	4.29%
3	Chronic lymphocytic leukemia	53.74%	53.74%	15.87%	15.95%
4	Colorectal adenocarcinoma	103.28%	103.29%	107.90%	109.96%
5	Colorectal adenocarcinoma with FAP	0.02%	0.68%	0.00%	0.00%
7	Duodenum adenocarcinoma	47.41%	64.30%	15.54%	26.09%
10	Gallbladder non papillary adenocarcinoma	0.62%	1.22%	0.00%	0.00%
14	Hepatocellular carcinoma	10.13%	18.40%	0.22%	0.92%
15	Hepatocellular carcinoma with HCV	4.65%	8.89%	0.02%	0.09%
16	Lung adenocarcinoma (nonsmokers)	2.81%	5.47%	0.00%	0.03%
17	Lung adenocarcinoma (smokers)	1.06%	2.09%	0.00%	0.00%
19	Melanoma	13.72%	13.72%	0.53%	0.53%
20	Osteosarcoma	0.62%	1.23%	0.00%	0.00%
21	Osteosarcoma of the arms	0.74%	1.47%	0.00%	0.00%
22	Osteosarcoma of the head	0.82%	1.62%	0.00%	0.00%
23	Osteosarcoma of the legs	0.55%	1.10%	0.00%	0.00%
24	Osteosarcoma of the pelvis	0.62%	1.24%	0.00%	0.00%
26	Pancreatic ductal adenocarcinoma	7.61%	14.14%	0.09%	0.48%
27	Pancreatic endocrine (islet cell) carcinoma	9.46%	17.28%	0.18%	0.87%
30	Thyroid papillary/follicular carcinoma	0.62%	1.22%	0.00%	0.00%
31	Thyroid medullary carcinoma	0.89%	1.76%	0.00%	0.00%

We can see that except cancer types 4 and 5, the largest intrinsic contribution computed from only cancer cells is 26.09% with Duodenum adenocarcinoma. However, according to Table 17, the excess rate was overestimated; we take the range [26.09%, 64.30%] as the intrinsic contribution under the condition of clonal expansion.

5.2 Extensive Study with 18 Tissue Types

In this section, we extend our study in Section 5.1 to 18 selected tissues listed in Table 3.2. In estimating intrinsic contribution, we use a more time consuming yet more accurate algorithm, binary search, to compute the total rate/excess rate. We will demonstrate that under various parameter settings, intrinsic factor has very limited contribution to cancer risk for most tissues and therefore the majority of observed risk is due to non-intrinsic factors.

We choose the set of intrinsic mutation rate to be $\mathbf{mr} = \{1.0\times10^{-8}, 1.1\times10^{-8}, 2.5\times10^{-8}, 1.0\times10^{-7}\}$. Note that 10^{-7} is an intentionally chosen large mutation rate which is above the range used in [34]. As before, we choose the factor of mutation rate enlargement to be $\mathbf{emr} = \{1.0, 2.0\}$ and clonal expansion factor $\mathbf{fd} = \{1.0, 2.0\}$. Different from our preliminary experiments in Section 5.1, we don't apply the heuristic regulation on clonal expansion, meaning that the clonal expansion effect will continue through the entire lifespan. In this way, we obtain an extreme-case upper bound of intrinsic risk under clonal expansion. By default, the required mutation hits for stem, progenitor and terminal cells are $(M_3, M_4, M_5), (M_2, M_3, M_4, M_5)$ and $(M_1, M_2, M_3, M_4, M_5)$, respectively. We represent the hits in our parameters as $mt_S = 00111, mt_P = 01111$ and $mt_T = 11111$. For some tissues, we also use $mt_S = 01111, mt_P = 11111$ and $mt_T = 11111$, which makes it harder for cancer onset.

5.2.1 Lifetime Risk and Intrinsic Contribution

As in previous experiments, we study the lifetime intrinsic risks and compare them to observed risks, under various parameters. In addition, we provide the estimated intrinsic contribution percentages and compare them to those reported in Tomasetti et al. [41]. We first focus on comparing different intrinsic mutation rates, then different mutation effect factors, and clonal expansion. For certain tissues, we also study the change of required mutation hits for cancer onset. Eventually we will summarize all parameter settings into a comprehensive result.

Figures 23 and 24 below illustrate the lifetime intrinsic risk computed from Extended Risk Model under different intrinsic rate. We can see the comparison of intrinsic risk and observed risk.

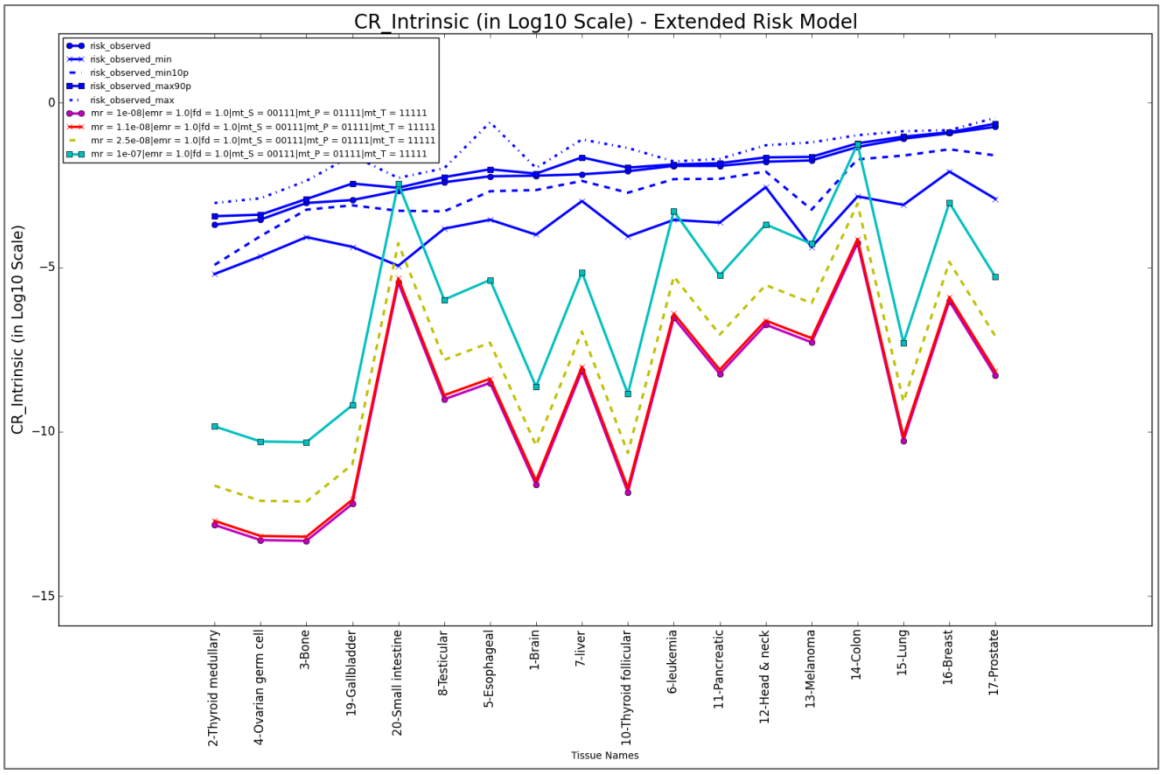


Figure 23: Lifetime intrinsic risk (log10 scale) computed from Extended Risk Model and statistics of NPCR observed risk in U.S. Tissue id/names are given below horizontal axis and the tissues are sorted in ascending order of "risk_observed", the average risk in U.S. Intrinsic mutation rate is selected to be {1e-08, 1.1e-08, 2.5e-08, 1e-07};

no mutation effects and clonal expansion were applied here ($emr = 1.0$ and $fd = 1.0$); also default required mutation hits were used ($mt_S = 00111, mt_P = 01111$ and $mt_T = 11111$).

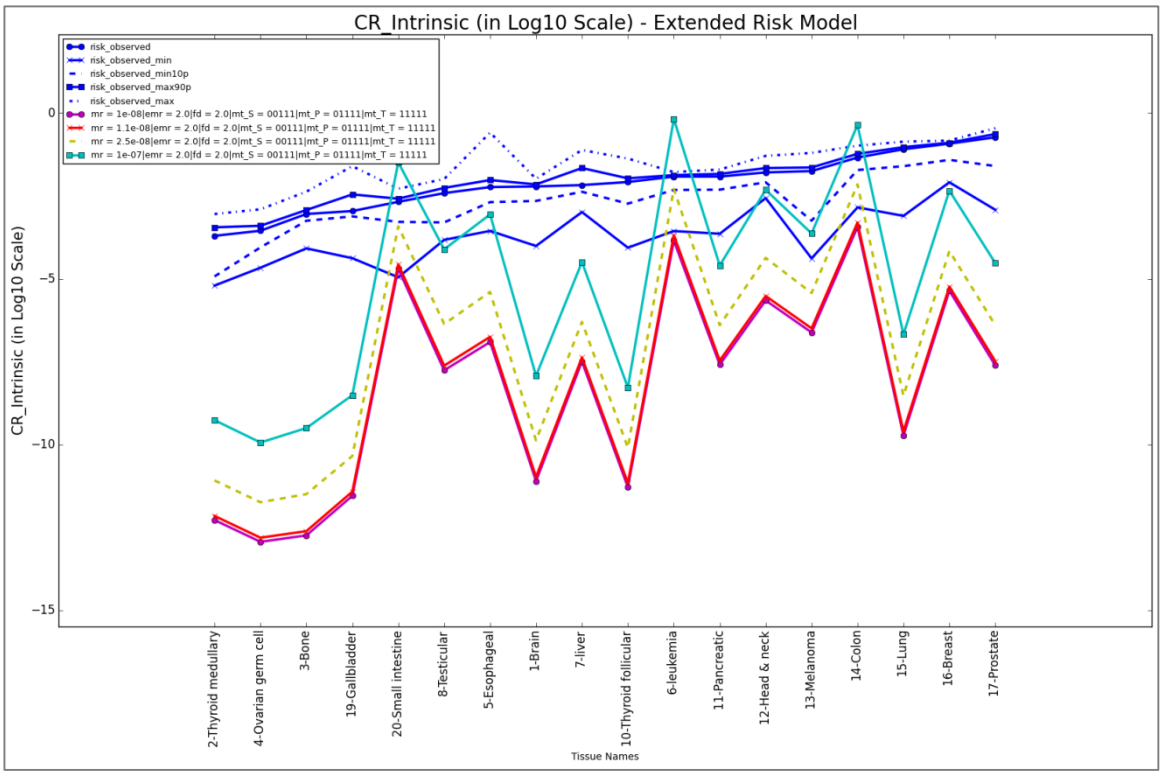


Figure 24: Lifetime intrinsic risk (log10 scale) computed from Extended Risk Model and statistics of NPCR observed risk in U.S. Tissue id/names are given below horizontal axis and the tissues are sorted in ascending order of "risk_observed", the average risk in U.S. Intrinsic mutation rate is selected to be {1e-08, 1.1e-08, 2.5e-08, 1e-07}. Mutation effects and clonal expansion were applied here with factors $emr = 2.0$ and $fd = 2.0$; also default required mutation hits were used ($mt_S = 00111, mt_P = 01111$ and $mt_T = 11111$).

We can see from Figures 23 and 24 that except small intestine, leukemia and colon, the intrinsic risks from all other tissues are far below the average observed risk under all conditions. For large mutation rate 10^{-7}, small intestine, leukemia and colon all give intrinsic risks larger than observed risks under mutation effects and clonal expansion. We will later focus on these three tissues for further sensitivity analysis. Aside from the comparison of risk magnitude, the overall trend of intrinsic risk is very different from that of observed risk across these 18 tissues, which indicates significant variations coming from non-intrinsic factors.

To see a quantitative analysis on how much intrinsic/non-intrinsic factors contributes to total risk, Figures 25 and 26 below illustrate the intrinsic contribution percentage estimated according to the binary search algorithm described in 4.6.

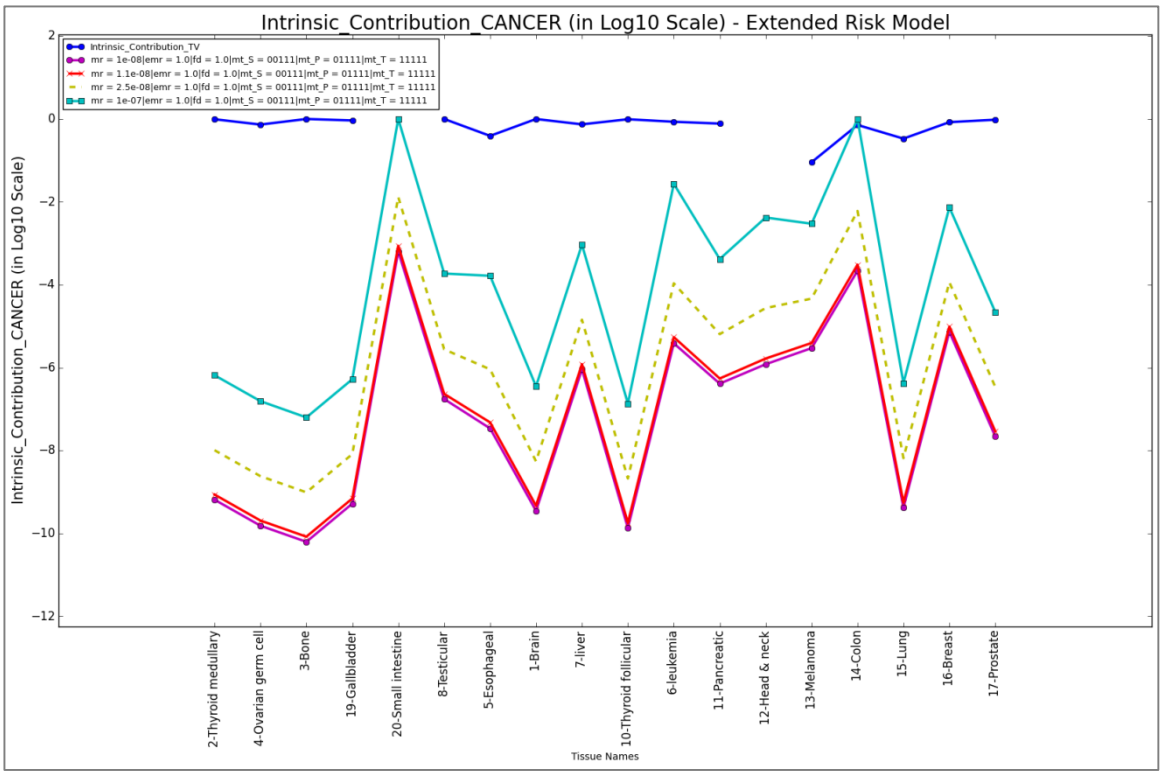


Figure 25: Lifetime intrinsic contribution (log10 scale) computed from Extended Risk Model and intrinsic contribution percentage, "Intrinsic_Contribution_TV", reported by Tomasetti, Li and Vogelstein [41] except Small Intestine and Head & Neck. Tissue id/names are given below horizontal axis and the tissues are sorted in ascending order of "risk_observed", the average risk in U.S. Intrinsic mutation rate is selected to be {1e-08, 1.1e-08, 2.5e-08, 1e-07}. No mutation effects and clonal expansion were applied here ($emr = 1.0$ and $fd = 1.0$); also default required mutation hits were used ($mt_S = 00111, mt_P = 01111$ and $mt_T = 11111$). Note that in rare cases where computed intrinsic contribution is greater than 1.0, (there are more mutations acquired due to intrinsic rate than that due to estimated total rate), the intrinsic contribution was set to 1.0 (0.0 in log10 scale).

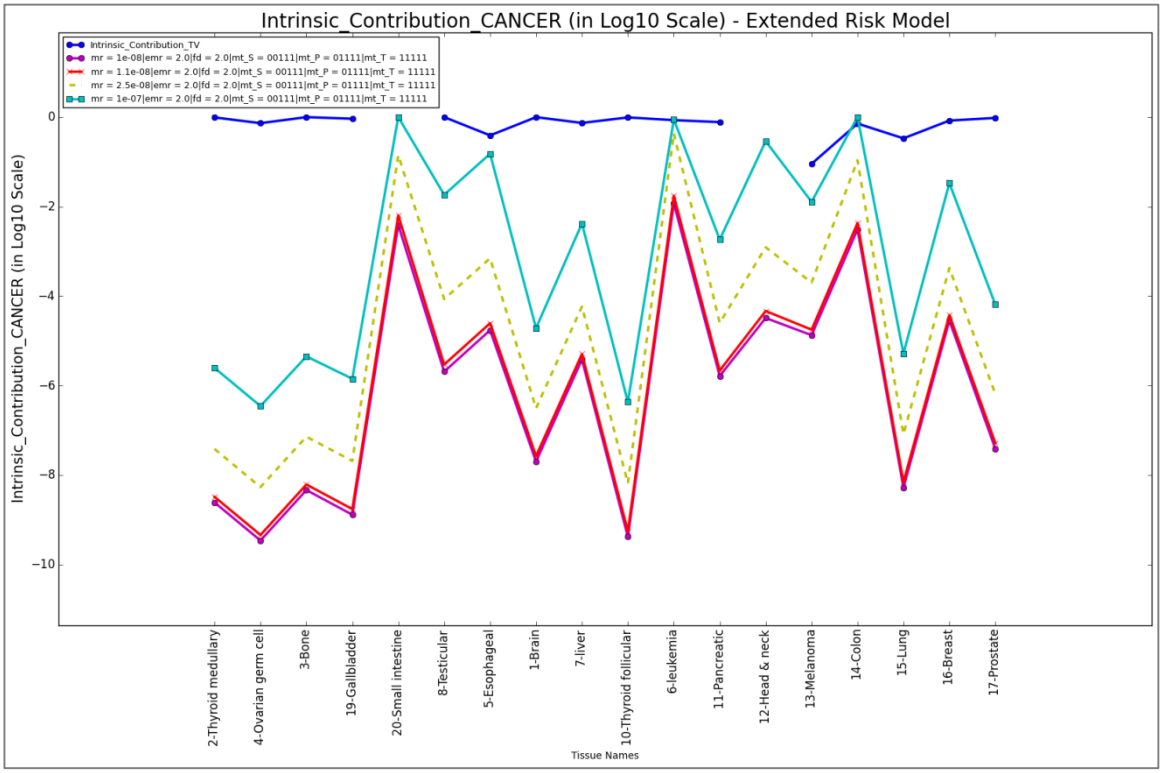


Figure 26: Lifetime intrinsic contribution (log10 scale) computed from Extended Risk Model and intrinsic contribution percentage, "Intrinsic_Contribution_TV", reported by Tomasetti, Li and Vogelstein [41] except Small Intestine and Head & Neck. Tissue id/names are given below horizontal axis and the tissues are sorted in ascending order of "risk_observed", the average risk in U.S. Intrinsic mutation rate is selected to be {1e-08, 1.1e-08, 2.5e-08, 1e-07}. Mutation effects and clonal expansion were applied here with factors $emr = 2.0$ and $fd = 2.0$; also default required mutation hits were used ($mt_S = 00111, mt_P = 01111$ and $mt_T = 11111$). Note that in rare cases where computed intrinsic contribution is greater than 1.0, (there are more mutations acquired due to intrinsic rate than that due to estimated total rate), the intrinsic contribution was set to 1.0 (0.0 in log10 scale).

From Figures 25, 26, the intrinsic contribution percentages for most tissues are much smaller than the reported values in Tomasetti et al. [41], if any. With large mutation rate 10^{-7} (especially with clonal expansion), small intestine, leukemia and colon have unreasonably high intrinsic contribution percentages.

From Figures 23 to 26, we can conclude that for most tissues, non-intrinsic factors contribute to most of the total cancer risk.

Small intestine, leukemia and colon have relatively long stem and progenitor lineages, leading to unreasonably large intrinsic risk and intrinsic contribution under large mutation rate. We now take required mutation hits into considerations, under mutation

rate 10^{-7}. Figures 27 and 28 provide the results for small intestine, leukemia and colon at different mutation hits required for cancer onset.

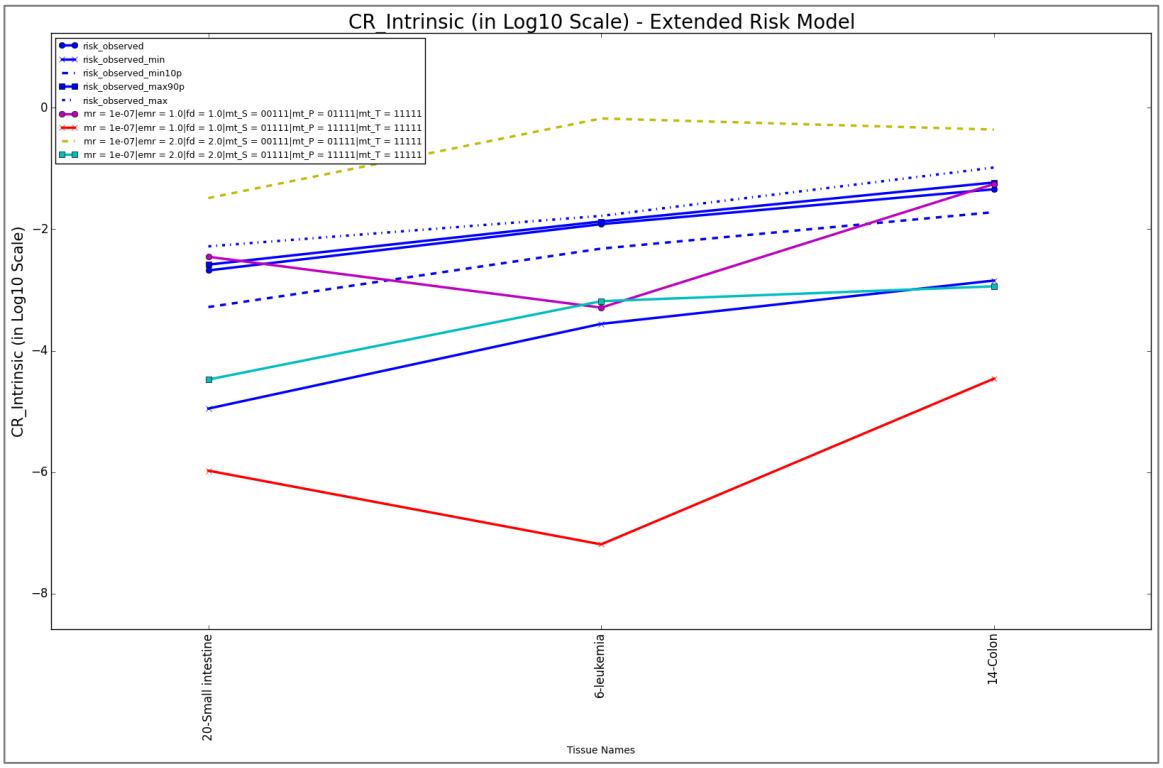


Figure 27: Lifetime intrinsic risk (log10 scale) computed from Extended Risk Model and statistics of NPCR observed risk in U.S. Tissue id/names are given below horizontal axis and the tissues are sorted in ascending order of "risk_observed", the average risk in U.S. Intrinsic mutation rate is selected to be {1e-07}. Mutation effects and clonal expansion were applied here with factors ranging from $emr = \{1.0, 2.0\}$ and $fd = \{1.0, 2.0\}$; also we consider different required mutation hits: $(mt_S = 00111, mt_P = 01111$ and $mt_T = 11111)$ vs. $(mt_S = 01111, mt_P = 11111$ and $mt_T = 11111)$.

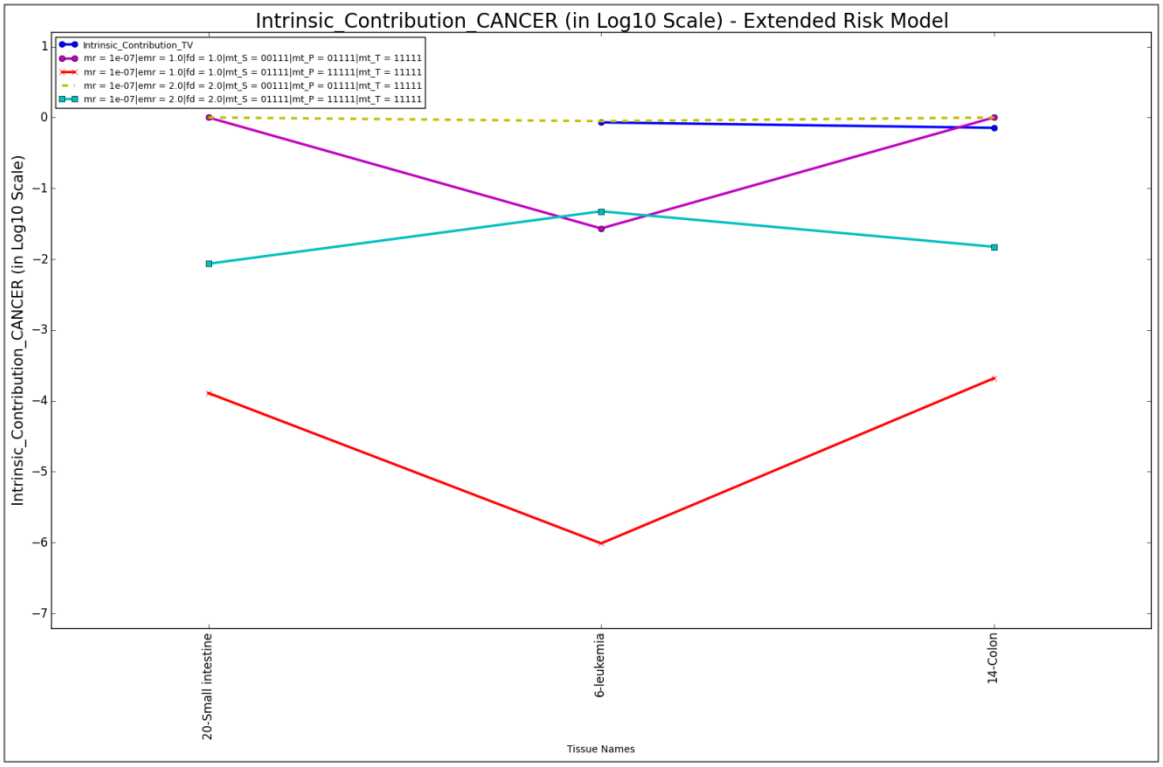


Figure 28: Lifetime intrinsic contribution (log10 scale) computed from Extended Risk Model and intrinsic contribution percentage, "Intrinsic_Contribution_TV", reported by Tomasetti, Li and Vogelstein [41] except Small Intestine and Head & Neck. Tissue id/names are given below horizontal axis and the tissues are sorted in ascending order of "risk_observed", the average risk in U.S. Intrinsic mutation rate is selected to be {1e-07}. Mutation effects and clonal expansion were applied here with factors ranging from $emr = \{1.0, 2.0\}$ and $fd = \{1.0, 2.0\}$; also we consider different required mutation hits: ($mt_S = 00111, mt_P = 01111$ and $mt_T = 11111$) vs. ($mt_S = 01111, mt_P = 11111$ and $mt_T = 11111$). Note that in cases where computed intrinsic contribution is greater than 1.0, (there are more mutations acquired due to intrinsic rate than that due to estimated total rate), the intrinsic contribution was set to 1.0 (0.0 in log10 scale).

We can see that by increasing the number of required hits for cancer onset by 1 on stem and progenitor cells, we can significantly reduce intrinsic risk and intrinsic contribution. Even under large mutation rate 10^{-7}, we can see that intrinsic factor only contributes a small portion of total risk with $mt_S = 01111, mt_P = 11111$ and $mt_T = 11111$.

Figures 29 and 30 below plot together the results with all parameter settings. Note that for tissues other than small intestine, leukemia and colon, we use the default required mutation hits of $mt_S = 00111, mt_P = 01111$ and $mt_T = 11111$.

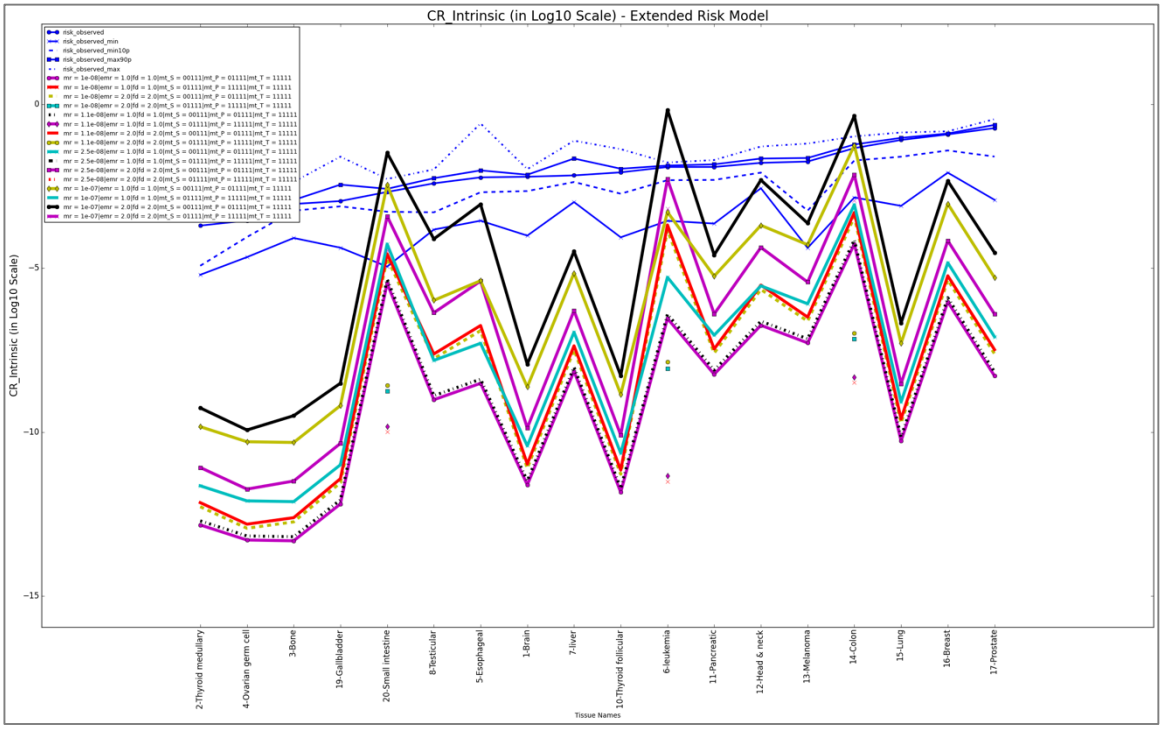


Figure 29: Lifetime intrinsic risk (log10 scale) computed from Extended Risk Model and statistics of NPCR observed risk in U.S. Tissue id/names are given below horizontal axis and the tissues are sorted in ascending order of "risk_observed", the average risk in U.S. Intrinsic mutation rate is selected to be {1e-08, 1.1e-08, 2.5e-08, 1e-07}; mutation effects and clonal expansion factors are selected from $emr = \{1.0, 2.0\}$ and $fd = \{1.0, 2.0\}$; the default required mutation hits are ($mt_S = 00111, mt_P = 01111$ and $mt_T = 11111$); for small intestine, leukemia and colon, we add an additional set of mutation hits ($mt_S = 01111, mt_P = 11111$ and $mt_T = 11111$).

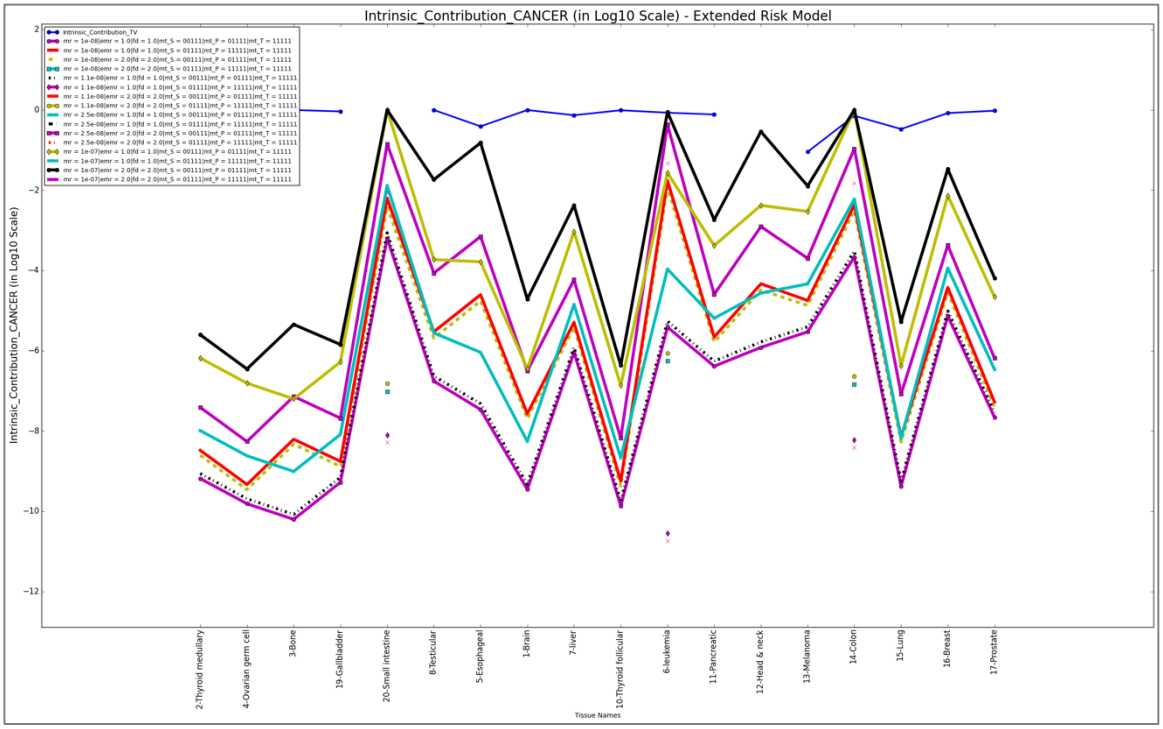


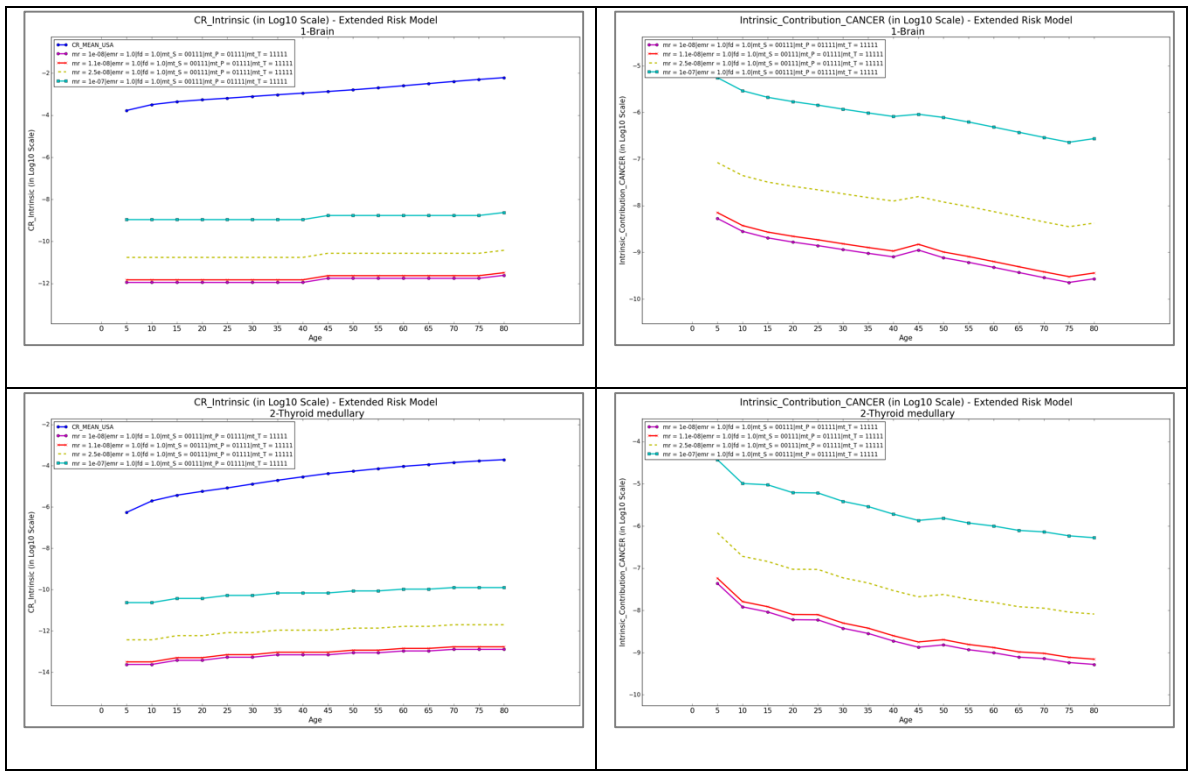
Figure 30: Lifetime intrinsic contribution (log10 scale) computed from Extended Risk Model and intrinsic contribution percentage, "Intrinsic_Contribution_TV", reported by Tomasetti, Li and Vogelstein [41] except Small Intestine and Head & Neck. Tissue id/names are given below horizontal axis and the tissues are sorted in ascending order of "risk_observed", the average risk in U.S. Intrinsic mutation rate is selected to be {1e-08, 1.1e-08, 2.5e-08, 1e-07}; mutation effects and clonal expansion factors are selected from $emr = \{1.0, 2.0\}$ and $fd = \{1.0, 2.0\}$; the default required mutation hits are ($mt_S = 00111, mt_P = 01111$ and $mt_T = 11111$); for small intestine, leukemia and colon, we add an additional set of mutation hits ($mt_S = 01111, mt_P = 11111$ and $mt_T = 11111$). Note that in cases where computed intrinsic contribution is greater than 1.0, (there are more mutations acquired due to intrinsic rate than that due to estimated total rate), the intrinsic contribution was set to 1.0 (0.0 in log10 scale).

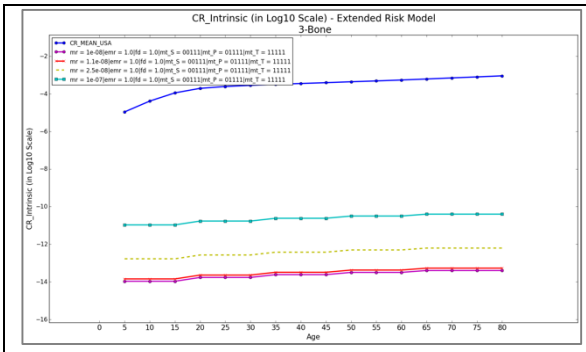
We see that for most tissues and parameter settings, intrinsic rate only contributes a small portion of total risk.

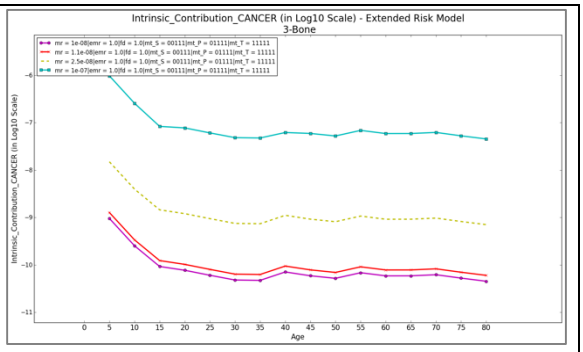
5.2.2 Age Dependent Risk and Intrinsic Contribution

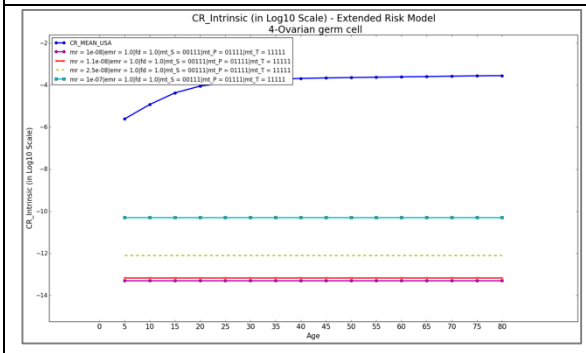
NPCR provides the average observed risk (in U.S.) at each of 5 years within an 80-year lifespan for the 18 tissues. We described the algorithm to compute age-dependent intrinsic risk in Section 4.4.4 which can be easily extended to the estimation of intrinsic contribution, with algorithms in Section 4.6. We use selected tissues to show longitudinal patterns of how observed risk, intrinsic risk, and intrinsic contribution changes with age, as shown below in Tables 19 and 20.

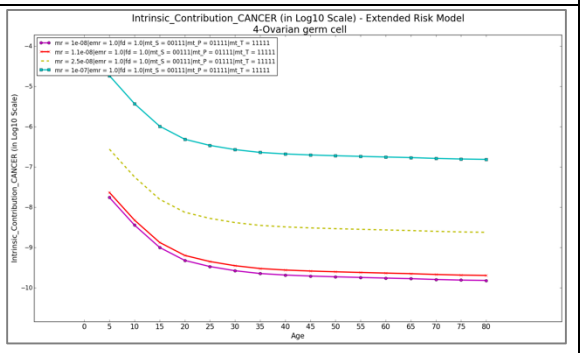
Table 19: Age dependent observed risk, intrinsic risk, and estimated intrinsic contribution at each of 5 years within an 80-year lifespan. Left side figures plot the average observed risk ("CR_MEAN_USA") and intrinsic risk under selected parameter settings; right side figures plot the estimated intrinsic contribution for the same tissue. The intrinsic contribution in above figures for age x, represents cumulative average intrinsic contribution percentage from age 0 to age x. All figures are in log10 scale. Intrinsic mutation rate is selected to be {1e-08, 1.1e-08, 2.5e-08, 1e-07}. No mutation effects and clonal expansion were applied here ($emr = 1.0$ and $fd = 1.0$); also default required mutation hits were used ($mt_S = 00111, mt_P = 01111$ and $mt_T = 11111$). Note that in rare cases where computed intrinsic contribution is greater than 1.0, (there are more mutations acquired due to intrinsic rate than that due to estimated total rate), the intrinsic contribution was set to 1.0 (0.0 in log10 scale).

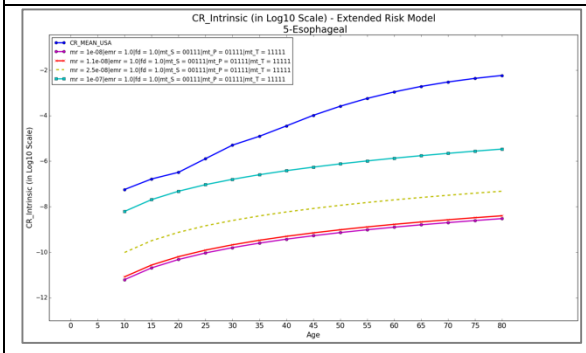


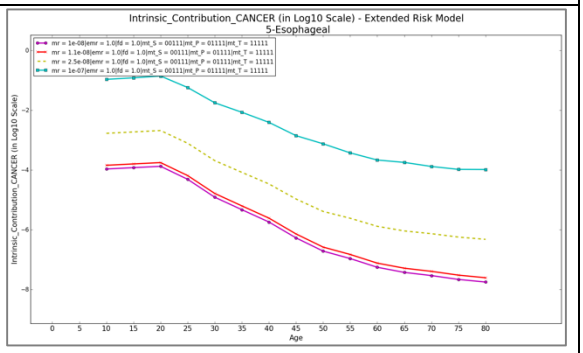

CR_Intrinsic (in Log10 Scale) - Extended Risk Model
3-Bone
CR_Intrinsic (in Log10 Scale)
Age

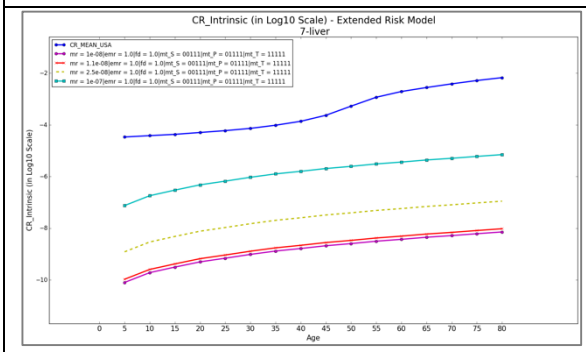

Intrinsic_Contribution_CANCER (in Log10 Scale) - Extended Risk Model
3-Bone
Intrinsic_Contribution_CANCER (in Log10 Scale)
Age

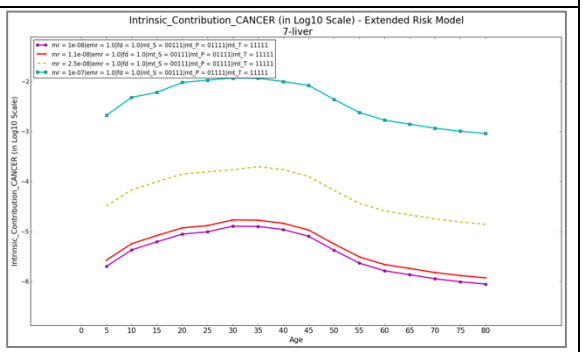

CR_Intrinsic (in Log10 Scale) - Extended Risk Model
4-Ovarian germ cell
CR_Intrinsic (in Log10 Scale)
Age

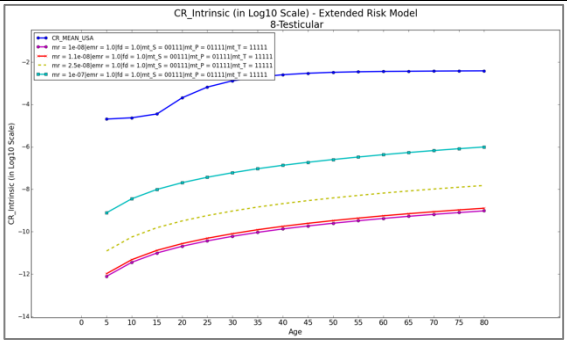

Intrinsic_Contribution_CANCER (in Log10 Scale) - Extended Risk Model
4-Ovarian germ cell
Intrinsic_Contribution_CANCER (in Log10 Scale)
Age

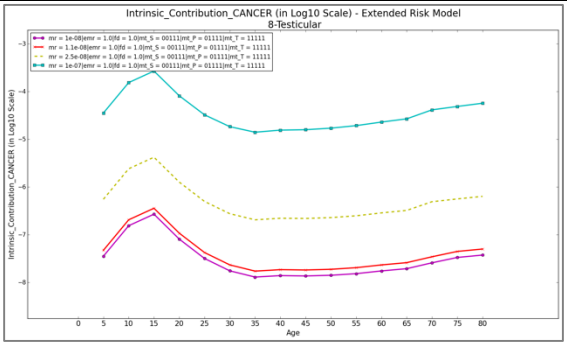

CR_Intrinsic (in Log10 Scale) - Extended Risk Model
5-Esophageal
CR_Intrinsic (in Log10 Scale)
Age

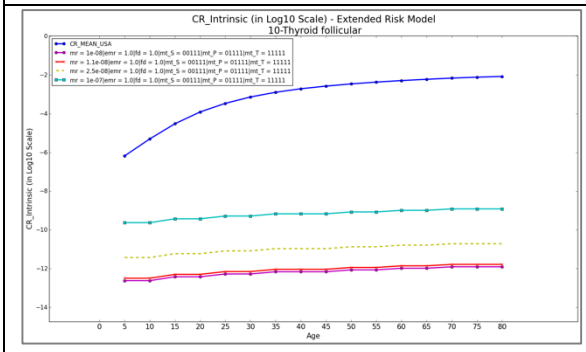

Intrinsic_Contribution_CANCER (in Log10 Scale) - Extended Risk Model
5-Esophageal
Intrinsic_Contribution_CANCER (in Log10 Scale)
Age

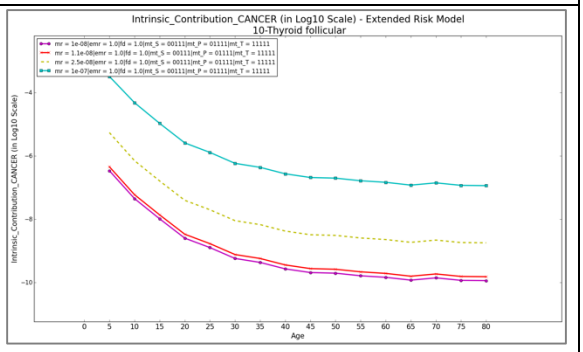

CR_Intrinsic (in Log10 Scale) - Extended Risk Model
7-liver
CR_Intrinsic (in Log10 Scale)
Age

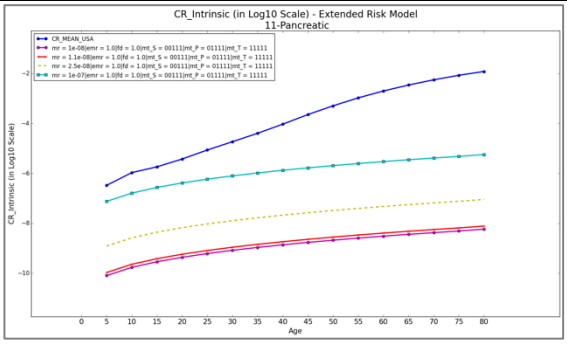

Intrinsic_Contribution_CANCER (in Log10 Scale) - Extended Risk Model
7-liver
Intrinsic_Contribution_CANCER (in Log10 Scale)
Age

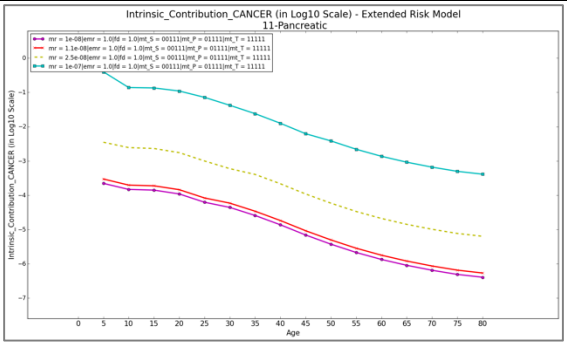

CR_Intrinsic (in Log10 Scale) - Extended Risk Model
8-Testicular
CR_MEAN_USA
mr = 1e-08|emr = 1.0|fd = 1.0|mt_S = 00111|mt_P = 01111|mt_T = 11111
mr = 1.1e-08|emr = 1.0|fd = 1.0|mt_S = 00111|mt_P = 01111|mt_T = 11111
mr = 2.5e-08|emr = 1.0|fd = 1.0|mt_S = 00111|mt_P = 01111|mt_T = 11111
mr = 1e-07|emr = 1.0|fd = 1.0|mt_S = 00111|mt_P = 01111|mt_T = 11111
CR_Intrinsic (in Log10 Scale)
Age

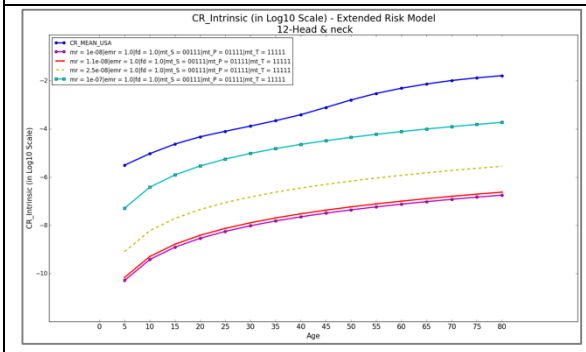

Intrinsic_Contribution_CANCER (in Log10 Scale) - Extended Risk Model
8-Testicular
mr = 1e-08|emr = 1.0|fd = 1.0|mt_S = 00111|mt_P = 01111|mt_T = 11111
mr = 1.1e-08|emr = 1.0|fd = 1.0|mt_S = 00111|mt_P = 01111|mt_T = 11111
mr = 2.5e-08|emr = 1.0|fd = 1.0|mt_S = 00111|mt_P = 01111|mt_T = 11111
mr = 1e-07|emr = 1.0|fd = 1.0|mt_S = 00111|mt_P = 01111|mt_T = 11111
Intrinsic_Contribution_CANCER (in Log10 Scale)
Age

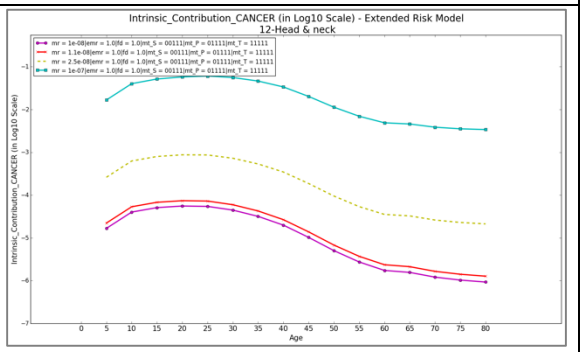

CR_Intrinsic (in Log10 Scale) - Extended Risk Model
10-Thyroid follicular
CR_MEAN_USA
mr = 1e-08|emr = 1.0|fd = 1.0|mt_S = 00111|mt_P = 01111|mt_T = 11111
mr = 1.1e-08|emr = 1.0|fd = 1.0|mt_S = 00111|mt_P = 01111|mt_T = 11111
mr = 2.5e-08|emr = 1.0|fd = 1.0|mt_S = 00111|mt_P = 01111|mt_T = 11111
mr = 1e-07|emr = 1.0|fd = 1.0|mt_S = 00111|mt_P = 01111|mt_T = 11111
CR_Intrinsic (in Log10 Scale)
Age

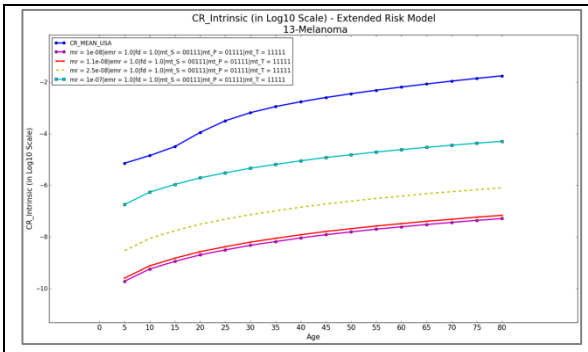

Intrinsic_Contribution_CANCER (in Log10 Scale) - Extended Risk Model
10-Thyroid follicular
mr = 1e-08|emr = 1.0|fd = 1.0|mt_S = 00111|mt_P = 01111|mt_T = 11111
mr = 1.1e-08|emr = 1.0|fd = 1.0|mt_S = 00111|mt_P = 01111|mt_T = 11111
mr = 2.5e-08|emr = 1.0|fd = 1.0|mt_S = 00111|mt_P = 01111|mt_T = 11111
mr = 1e-07|emr = 1.0|fd = 1.0|mt_S = 00111|mt_P = 01111|mt_T = 11111
Intrinsic_Contribution_CANCER (in Log10 Scale)
Age

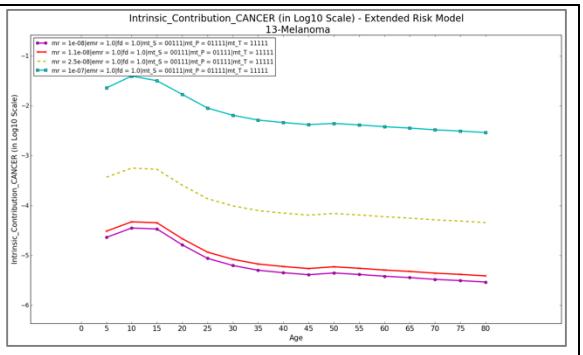

CR_Intrinsic (in Log10 Scale) - Extended Risk Model
11-Pancreatic
CR_MEAN_USA
mr = 1e-08|emr = 1.0|fd = 1.0|mt_S = 00111|mt_P = 01111|mt_T = 11111
mr = 1.1e-08|emr = 1.0|fd = 1.0|mt_S = 00111|mt_P = 01111|mt_T = 11111
mr = 2.5e-08|emr = 1.0|fd = 1.0|mt_S = 00111|mt_P = 01111|mt_T = 11111
mr = 1e-07|emr = 1.0|fd = 1.0|mt_S = 00111|mt_P = 01111|mt_T = 11111
CR_Intrinsic (in Log10 Scale)
Age

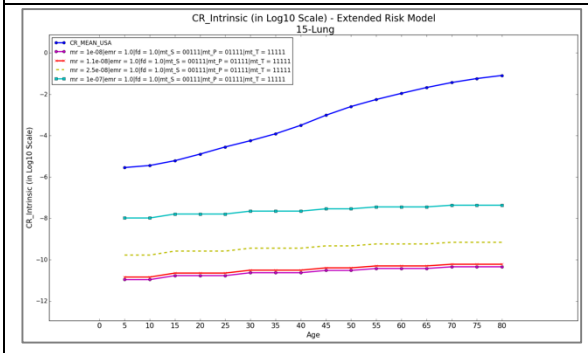

Intrinsic_Contribution_CANCER (in Log10 Scale) - Extended Risk Model
11-Pancreatic
mr = 1e-08|emr = 1.0|fd = 1.0|mt_S = 00111|mt_P = 01111|mt_T = 11111
mr = 1.1e-08|emr = 1.0|fd = 1.0|mt_S = 00111|mt_P = 01111|mt_T = 11111
mr = 2.5e-08|emr = 1.0|fd = 1.0|mt_S = 00111|mt_P = 01111|mt_T = 11111
mr = 1e-07|emr = 1.0|fd = 1.0|mt_S = 00111|mt_P = 01111|mt_T = 11111
Intrinsic_Contribution_CANCER (in Log10 Scale)
Age

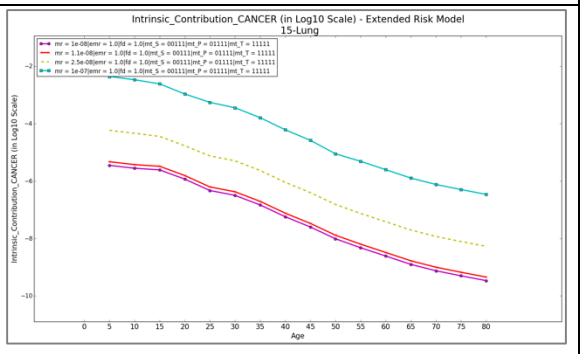

CR_Intrinsic (in Log10 Scale) - Extended Risk Model
12-Head & neck
CR_MEAN_USA
mr = 1e-08|emr = 1.0|fd = 1.0|mt_S = 00111|mt_P = 01111|mt_T = 11111
mr = 1.1e-08|emr = 1.0|fd = 1.0|mt_S = 00111|mt_P = 01111|mt_T = 11111
mr = 2.5e-08|emr = 1.0|fd = 1.0|mt_S = 00111|mt_P = 01111|mt_T = 11111
mr = 1e-07|emr = 1.0|fd = 1.0|mt_S = 00111|mt_P = 01111|mt_T = 11111
CR_Intrinsic (in Log10 Scale)
Age

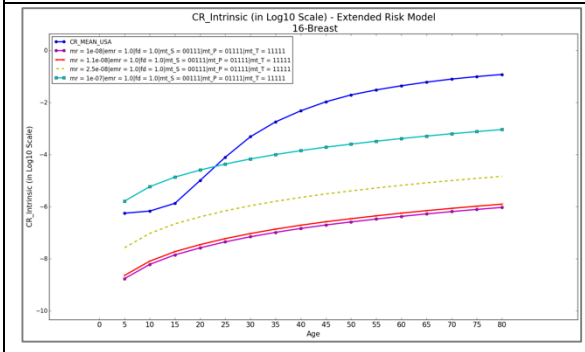

Intrinsic_Contribution_CANCER (in Log10 Scale) - Extended Risk Model
12-Head & neck
mr = 1e-08|emr = 1.0|fd = 1.0|mt_S = 00111|mt_P = 01111|mt_T = 11111
mr = 1.1e-08|emr = 1.0|fd = 1.0|mt_S = 00111|mt_P = 01111|mt_T = 11111
mr = 2.5e-08|emr = 1.0|fd = 1.0|mt_S = 00111|mt_P = 01111|mt_T = 11111
mr = 1e-07|emr = 1.0|fd = 1.0|mt_S = 00111|mt_P = 01111|mt_T = 11111
Intrinsic_Contribution_CANCER (in Log10 Scale)
Age

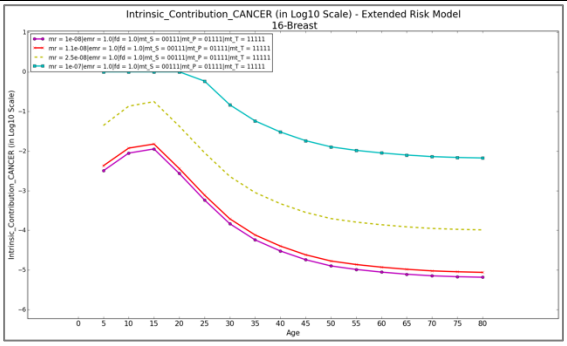

CR_Intrinsic (in Log10 Scale) - Extended Risk Model
13-Melanoma
CR_Intrinsic (in Log10 Scale)
Age

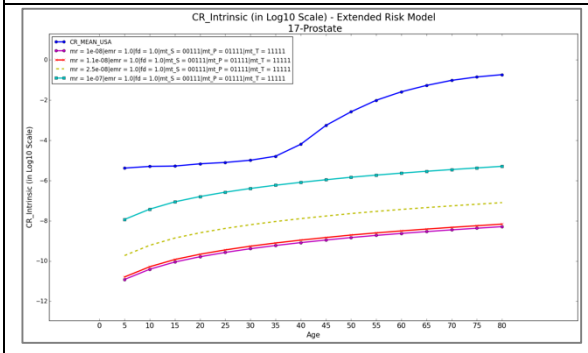

Intrinsic_Contribution_CANCER (in Log10 Scale) - Extended Risk Model
13-Melanoma
Intrinsic_Contribution_CANCER (in Log10 Scale)
Age

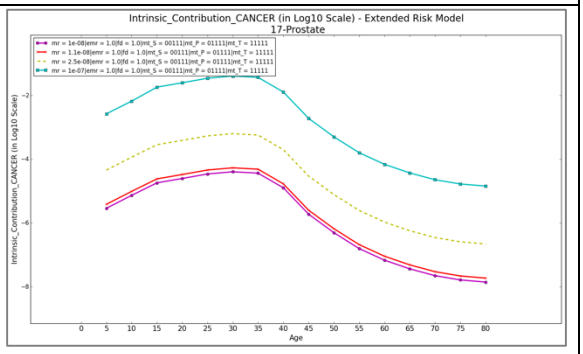

CR_Intrinsic (in Log10 Scale) - Extended Risk Model
15-Lung
CR_Intrinsic (in Log10 Scale)
Age


Intrinsic_Contribution_CANCER (in Log10 Scale) - Extended Risk Model
15-Lung
Intrinsic_Contribution_CANCER (in Log10 Scale)
Age


CR_Intrinsic (in Log10 Scale) - Extended Risk Model
16-Breast
CR_Intrinsic (in Log10 Scale)
Age


Intrinsic_Contribution_CANCER (in Log10 Scale) - Extended Risk Model
16-Breast
Intrinsic_Contribution_CANCER (in Log10 Scale)
Age


CR_Intrinsic (in Log10 Scale) - Extended Risk Model
17-Prostate
CR_Intrinsic (in Log10 Scale)
Age


Intrinsic_Contribution_CANCER (in Log10 Scale) - Extended Risk Model
17-Prostate
Intrinsic_Contribution_CANCER (in Log10 Scale)
Age

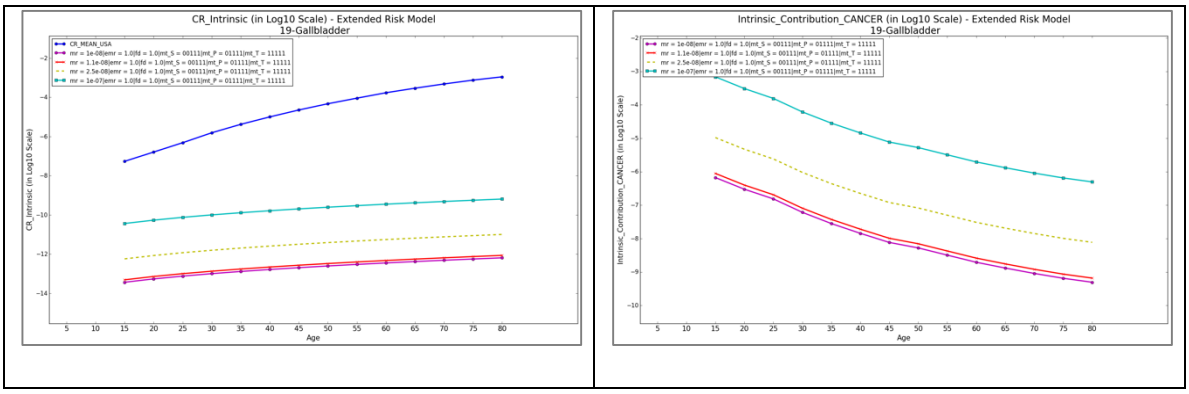
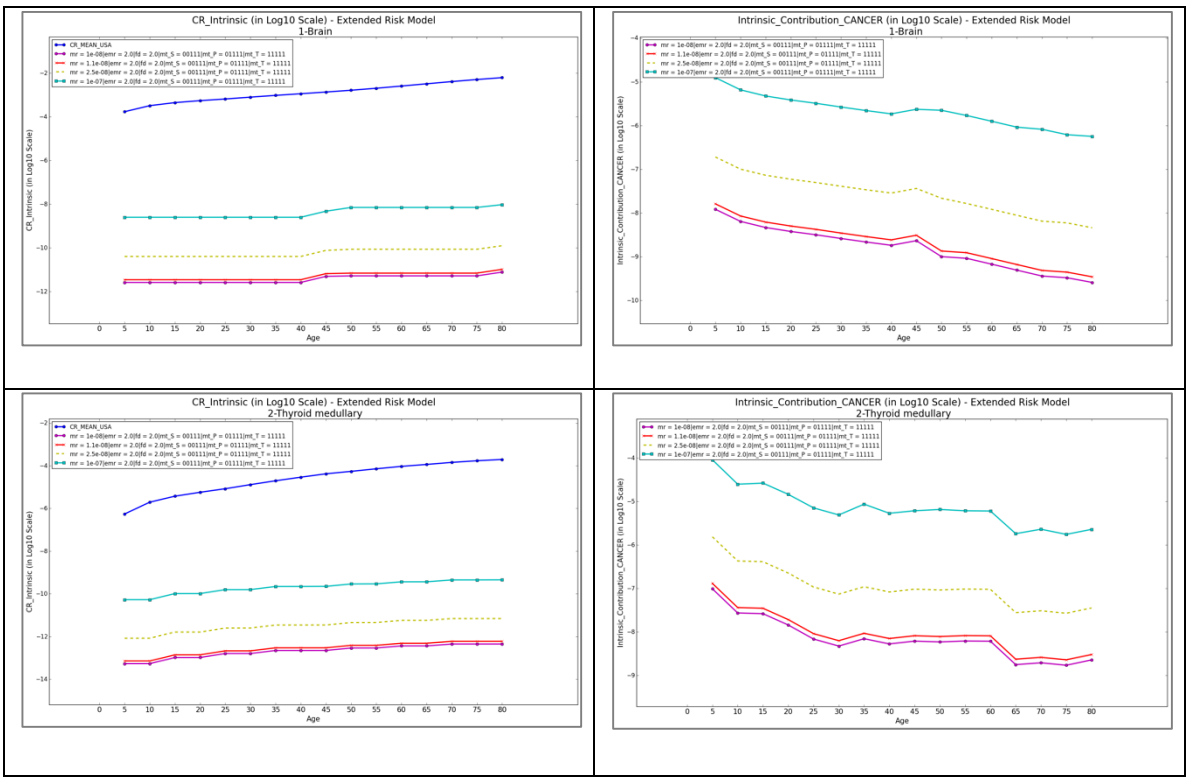
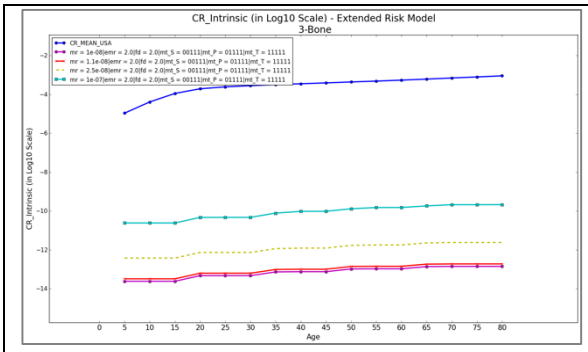
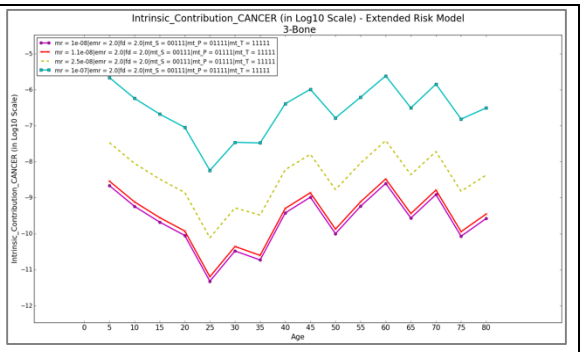
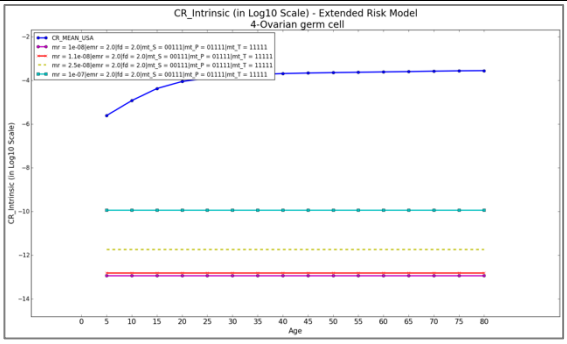


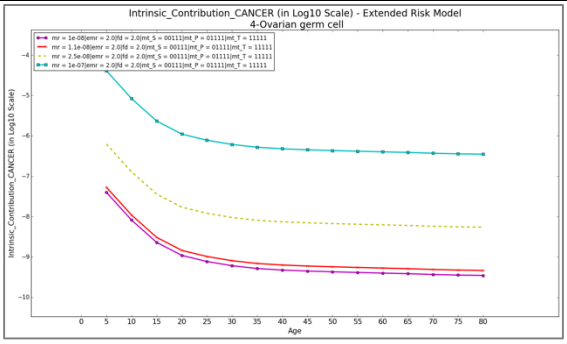
Table 20: Age dependent observed risk, intrinsic risk, and estimated intrinsic contribution at each of 5 years within an 80-year lifespan. Left side figures plot the average observed risk ("CR_MEAN_USA") and intrinsic risk under selected parameter settings; right side figures plot the estimated intrinsic contribution for the same tissue. The intrinsic contribution in above figures for age x, represents cumulative average intrinsic contribution percentage from age 0 to age x. All figures are in log10 scale. Intrinsic mutation rate is selected to be {1e-08, 1.1e-08, 2.5e-08, 1e-07}. Mutation effects and clonal expansion were applied here with factors $emr = 2.0$ and $fd = 2.0$; also default required mutation hits were used ($mt_S = 00111, mt_P = 01111$ and $mt_T = 11111$). Note that in rare cases where computed intrinsic contribution is greater than 1.0, (there are more mutations acquired due to intrinsic rate than that due to estimated total rate), the intrinsic contribution was set to 1.0 (0.0 in log10 scale).

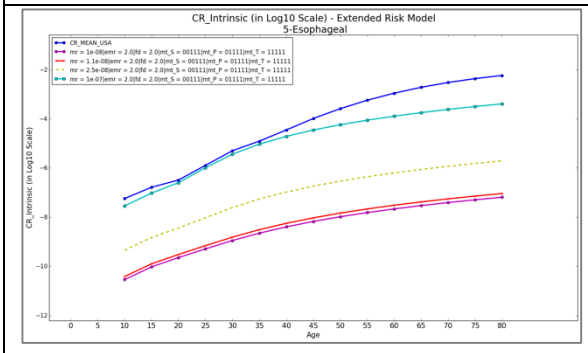


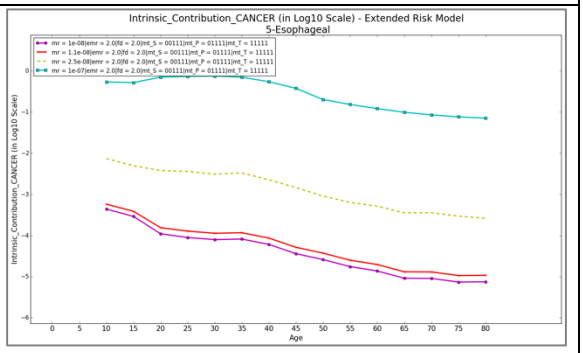

CR_Intrinsic (in Log10 Scale) - Extended Risk Model
3-Bone
CR_Intrinsic (in Log10 Scale)
Age

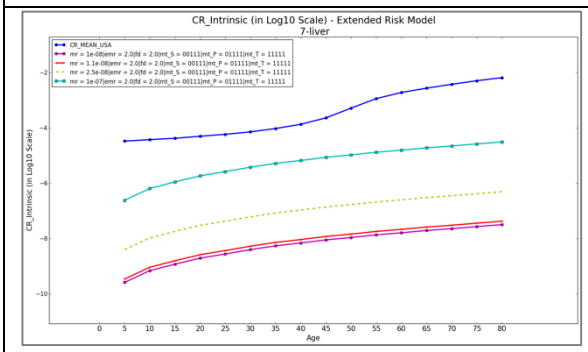

Intrinsic_Contribution_CANCER (in Log10 Scale) - Extended Risk Model
3-Bone
Intrinsic_Contribution_CANCER (in Log10 Scale)
Age

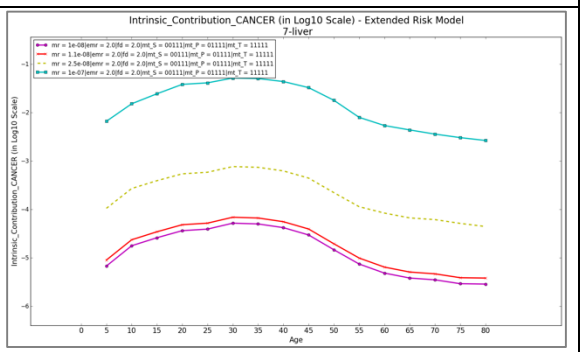

CR_Intrinsic (in Log10 Scale) - Extended Risk Model
4-Ovarian germ cell
CR_Intrinsic (in Log10 Scale)
Age

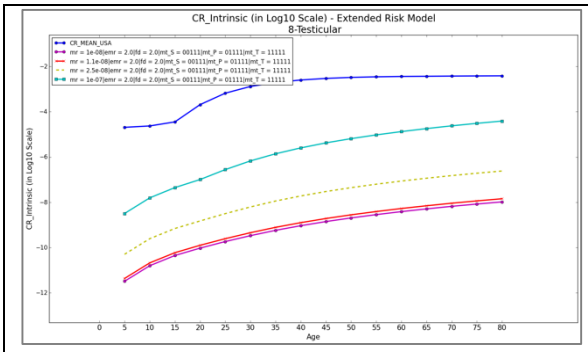

Intrinsic_Contribution_CANCER (in Log10 Scale) - Extended Risk Model
4-Ovarian germ cell
Intrinsic_Contribution_CANCER (in Log10 Scale)
Age

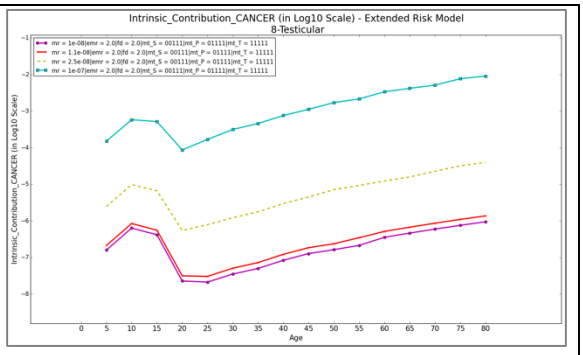

CR_Intrinsic (in Log10 Scale) - Extended Risk Model
5-Esophageal
CR_Intrinsic (in Log10 Scale)
Age

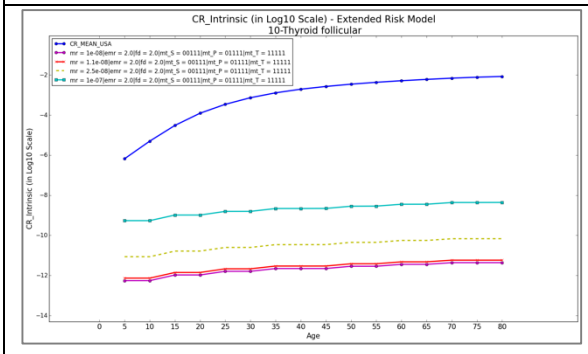

Intrinsic_Contribution_CANCER (in Log10 Scale) - Extended Risk Model
5-Esophageal
Intrinsic_Contribution_CANCER (in Log10 Scale)
Age

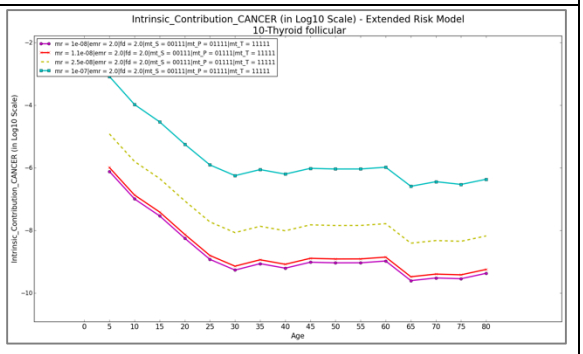

CR_Intrinsic (in Log10 Scale) - Extended Risk Model
7-liver
CR_Intrinsic (in Log10 Scale)
Age

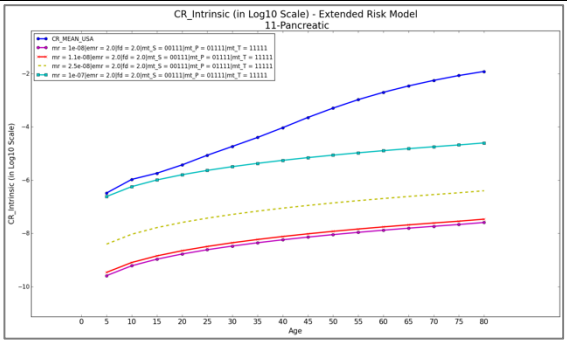

Intrinsic_Contribution_CANCER (in Log10 Scale) - Extended Risk Model
7-liver
Intrinsic_Contribution_CANCER (in Log10 Scale)
Age

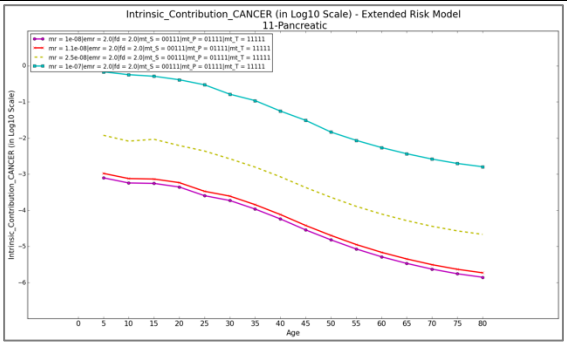

CR_Intrinsic (in Log10 Scale) - Extended Risk Model
8-Testicular
CR_MEAN_USA
mr = 1e-08|emr = 2.0|fd = 2.0|mt_S = 00111|mt_P = 01111|mt_T = 11111
mr = 1.1e-08|emr = 2.0|fd = 2.0|mt_S = 00111|mt_P = 01111|mt_T = 11111
mr = 2.5e-08|emr = 2.0|fd = 2.0|mt_S = 00111|mt_P = 01111|mt_T = 11111
mr = 1e-07|emr = 2.0|fd = 2.0|mt_S = 00111|mt_P = 01111|mt_T = 11111
CR_Intrinsic (in Log10 Scale)
Age

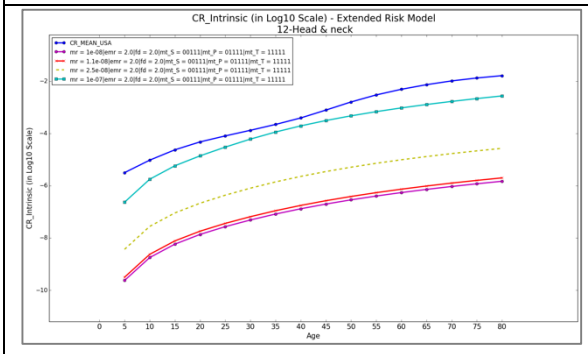

Intrinsic_Contribution_CANCER (in Log10 Scale) - Extended Risk Model
8-Testicular
mr = 1e-08|emr = 2.0|fd = 2.0|mt_S = 00111|mt_P = 01111|mt_T = 11111
mr = 1.1e-08|emr = 2.0|fd = 2.0|mt_S = 00111|mt_P = 01111|mt_T = 11111
mr = 2.5e-08|emr = 2.0|fd = 2.0|mt_S = 00111|mt_P = 01111|mt_T = 11111
mr = 1e-07|emr = 2.0|fd = 2.0|mt_S = 00111|mt_P = 01111|mt_T = 11111
Intrinsic_Contribution_CANCER (in Log10 Scale)
Age

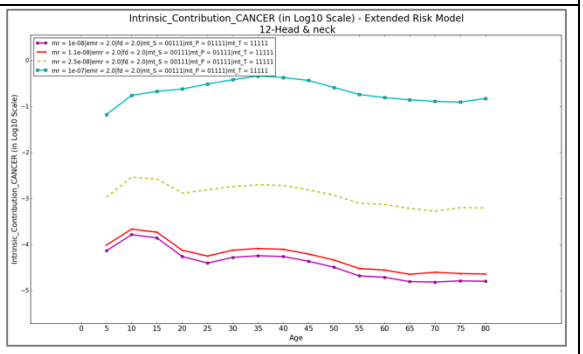

CR_Intrinsic (in Log10 Scale) - Extended Risk Model
10-Thyroid follicular
CR_MEAN_USA
mr = 1e-08|emr = 2.0|fd = 2.0|mt_S = 00111|mt_P = 01111|mt_T = 11111
mr = 1.1e-08|emr = 2.0|fd = 2.0|mt_S = 00111|mt_P = 01111|mt_T = 11111
mr = 2.5e-08|emr = 2.0|fd = 2.0|mt_S = 00111|mt_P = 01111|mt_T = 11111
mr = 1e-07|emr = 2.0|fd = 2.0|mt_S = 00111|mt_P = 01111|mt_T = 11111
CR_Intrinsic (in Log10 Scale)
Age

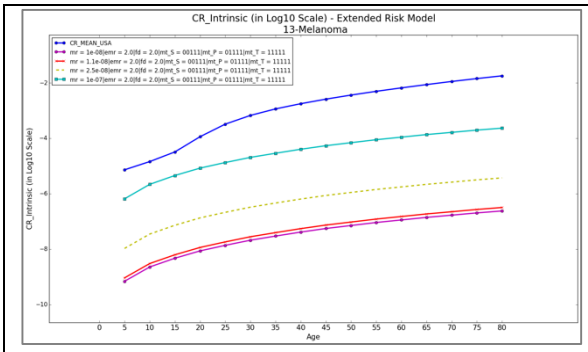

Intrinsic_Contribution_CANCER (in Log10 Scale) - Extended Risk Model
10-Thyroid follicular
mr = 1e-08|emr = 2.0|fd = 2.0|mt_S = 00111|mt_P = 01111|mt_T = 11111
mr = 1.1e-08|emr = 2.0|fd = 2.0|mt_S = 00111|mt_P = 01111|mt_T = 11111
mr = 2.5e-08|emr = 2.0|fd = 2.0|mt_S = 00111|mt_P = 01111|mt_T = 11111
mr = 1e-07|emr = 2.0|fd = 2.0|mt_S = 00111|mt_P = 01111|mt_T = 11111
Intrinsic_Contribution_CANCER (in Log10 Scale)
Age

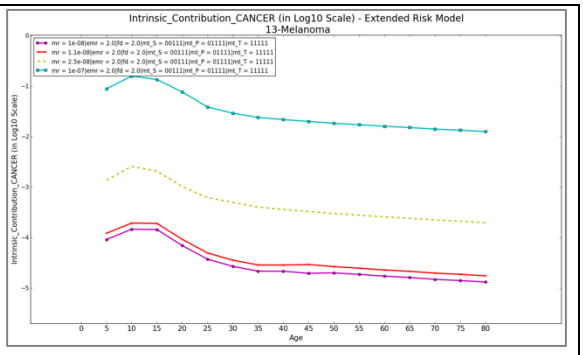

CR_Intrinsic (in Log10 Scale) - Extended Risk Model
11-Pancreatic
CR_MEAN_USA
mr = 1e-08|emr = 2.0|fd = 2.0|mt_S = 00111|mt_P = 01111|mt_T = 11111
mr = 1.1e-08|emr = 2.0|fd = 2.0|mt_S = 00111|mt_P = 01111|mt_T = 11111
mr = 2.5e-08|emr = 2.0|fd = 2.0|mt_S = 00111|mt_P = 01111|mt_T = 11111
mr = 1e-07|emr = 2.0|fd = 2.0|mt_S = 00111|mt_P = 01111|mt_T = 11111
CR_Intrinsic (in Log10 Scale)
Age

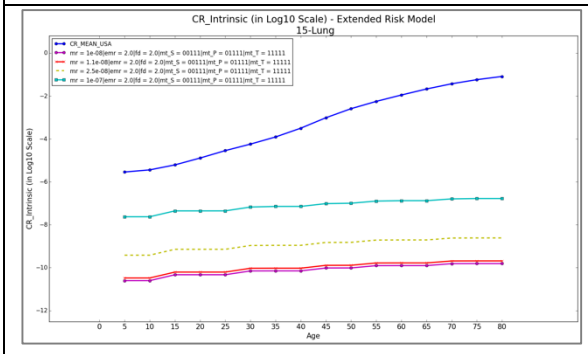

Intrinsic_Contribution_CANCER (in Log10 Scale) - Extended Risk Model
11-Pancreatic
mr = 1e-08|emr = 2.0|fd = 2.0|mt_S = 00111|mt_P = 01111|mt_T = 11111
mr = 1.1e-08|emr = 2.0|fd = 2.0|mt_S = 00111|mt_P = 01111|mt_T = 11111
mr = 2.5e-08|emr = 2.0|fd = 2.0|mt_S = 00111|mt_P = 01111|mt_T = 11111
mr = 1e-07|emr = 2.0|fd = 2.0|mt_S = 00111|mt_P = 01111|mt_T = 11111
Intrinsic_Contribution_CANCER (in Log10 Scale)
Age

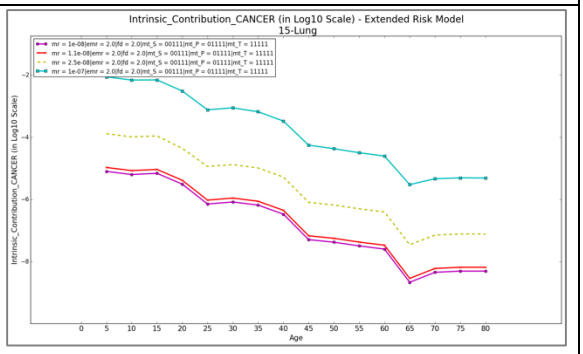

CR_Intrinsic (in Log10 Scale) - Extended Risk Model
12-Head & neck
CR_MEAN_USA
mr = 1e-08|emr = 2.0|fd = 2.0|mt_S = 00111|mt_P = 01111|mt_T = 11111
mr = 1.1e-08|emr = 2.0|fd = 2.0|mt_S = 00111|mt_P = 01111|mt_T = 11111
mr = 2.5e-08|emr = 2.0|fd = 2.0|mt_S = 00111|mt_P = 01111|mt_T = 11111
mr = 1e-07|emr = 2.0|fd = 2.0|mt_S = 00111|mt_P = 01111|mt_T = 11111
CR_Intrinsic (in Log10 Scale)
Age

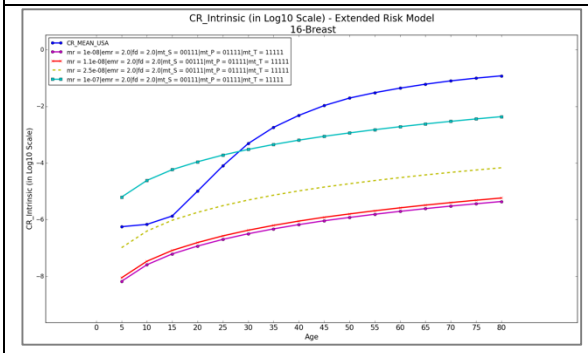

Intrinsic_Contribution_CANCER (in Log10 Scale) - Extended Risk Model
12-Head & neck
mr = 1e-08|emr = 2.0|fd = 2.0|mt_S = 00111|mt_P = 01111|mt_T = 11111
mr = 1.1e-08|emr = 2.0|fd = 2.0|mt_S = 00111|mt_P = 01111|mt_T = 11111
mr = 2.5e-08|emr = 2.0|fd = 2.0|mt_S = 00111|mt_P = 01111|mt_T = 11111
mr = 1e-07|emr = 2.0|fd = 2.0|mt_S = 00111|mt_P = 01111|mt_T = 11111
Intrinsic_Contribution_CANCER (in Log10 Scale)
Age

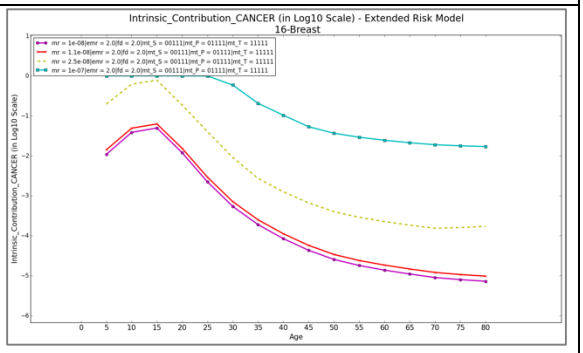

CR_Intrinsic (in Log10 Scale) - Extended Risk Model
13-Melanoma
CR_MEAN_USA
mr = 1e-08|emr = 2.0|fd = 2.0|mt_S = 00111|mt_P = 01111|mt_T = 11111
mr = 1.1e-08|emr = 2.0|fd = 2.0|mt_S = 00111|mt_P = 01111|mt_T = 11111
mr = 2.5e-08|emr = 2.0|fd = 2.0|mt_S = 00111|mt_P = 01111|mt_T = 11111
mr = 1e-07|emr = 2.0|fd = 2.0|mt_S = 00111|mt_P = 01111|mt_T = 11111
CR_Intrinsic (in Log10 Scale)
Age

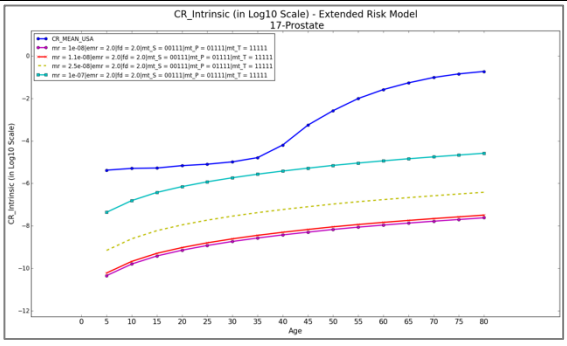

Intrinsic_Contribution_CANCER (in Log10 Scale) - Extended Risk Model
13-Melanoma
mr = 1e-08|emr = 2.0|fd = 2.0|mt_S = 00111|mt_P = 01111|mt_T = 11111
mr = 1.1e-08|emr = 2.0|fd = 2.0|mt_S = 00111|mt_P = 01111|mt_T = 11111
mr = 2.5e-08|emr = 2.0|fd = 2.0|mt_S = 00111|mt_P = 01111|mt_T = 11111
mr = 1e-07|emr = 2.0|fd = 2.0|mt_S = 00111|mt_P = 01111|mt_T = 11111
Intrinsic_Contribution_CANCER (in Log10 Scale)
Age

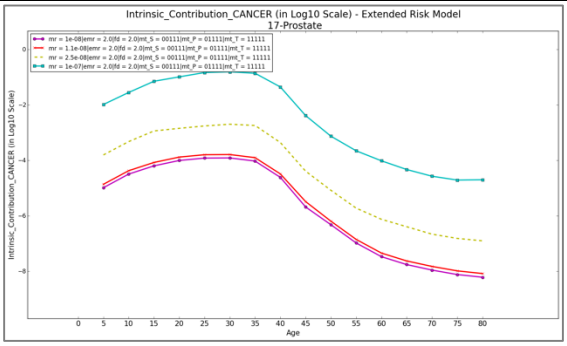

CR_Intrinsic (in Log10 Scale) - Extended Risk Model
15-Lung
CR_MEAN_USA
mr = 1e-08|emr = 2.0|fd = 2.0|mt_S = 00111|mt_P = 01111|mt_T = 11111
mr = 1.1e-08|emr = 2.0|fd = 2.0|mt_S = 00111|mt_P = 01111|mt_T = 11111
mr = 2.5e-08|emr = 2.0|fd = 2.0|mt_S = 00111|mt_P = 01111|mt_T = 11111
mr = 1e-07|emr = 2.0|fd = 2.0|mt_S = 00111|mt_P = 01111|mt_T = 11111
CR_Intrinsic (in Log10 Scale)
Age

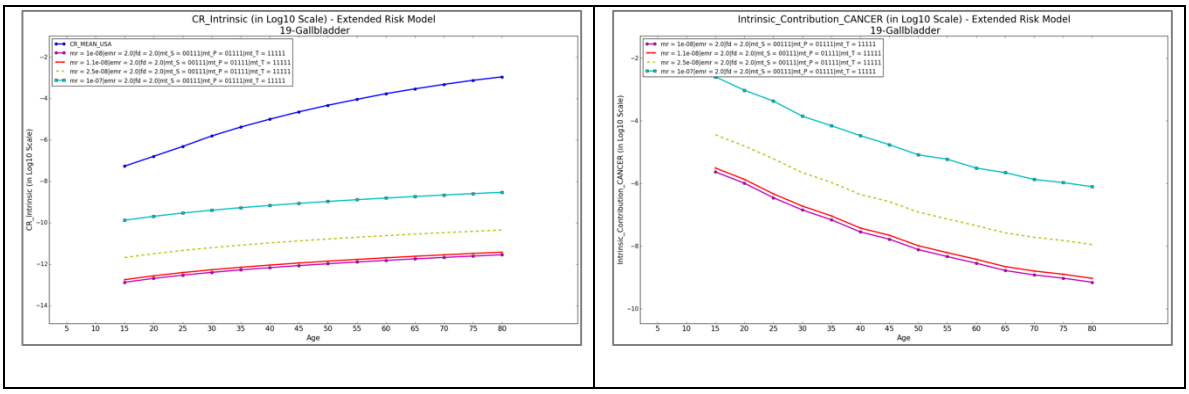

Intrinsic_Contribution_CANCER (in Log10 Scale) - Extended Risk Model
15-Lung
mr = 1e-08|emr = 2.0|fd = 2.0|mt_S = 00111|mt_P = 01111|mt_T = 11111
mr = 1.1e-08|emr = 2.0|fd = 2.0|mt_S = 00111|mt_P = 01111|mt_T = 11111
mr = 2.5e-08|emr = 2.0|fd = 2.0|mt_S = 00111|mt_P = 01111|mt_T = 11111
mr = 1e-07|emr = 2.0|fd = 2.0|mt_S = 00111|mt_P = 01111|mt_T = 11111
Intrinsic_Contribution_CANCER (in Log10 Scale)
Age


CR_Intrinsic (in Log10 Scale) - Extended Risk Model
16-Breast
CR_MEAN_USA
mr = 1e-08|emr = 2.0|fd = 2.0|mt_S = 00111|mt_P = 01111|mt_T = 11111
mr = 1.1e-08|emr = 2.0|fd = 2.0|mt_S = 00111|mt_P = 01111|mt_T = 11111
mr = 2.5e-08|emr = 2.0|fd = 2.0|mt_S = 00111|mt_P = 01111|mt_T = 11111
mr = 1e-07|emr = 2.0|fd = 2.0|mt_S = 00111|mt_P = 01111|mt_T = 11111
CR_Intrinsic (in Log10 Scale)
Age


Intrinsic_Contribution_CANCER (in Log10 Scale) - Extended Risk Model
16-Breast
mr = 1e-08|emr = 2.0|fd = 2.0|mt_S = 00111|mt_P = 01111|mt_T = 11111
mr = 1.1e-08|emr = 2.0|fd = 2.0|mt_S = 00111|mt_P = 01111|mt_T = 11111
mr = 2.5e-08|emr = 2.0|fd = 2.0|mt_S = 00111|mt_P = 01111|mt_T = 11111
mr = 1e-07|emr = 2.0|fd = 2.0|mt_S = 00111|mt_P = 01111|mt_T = 11111
Intrinsic_Contribution_CANCER (in Log10 Scale)
Age


CR_Intrinsic (in Log10 Scale) - Extended Risk Model
17-Prostate
CR_MEAN_USA
mr = 1e-08|emr = 2.0|fd = 2.0|mt_S = 00111|mt_P = 01111|mt_T = 11111
mr = 1.1e-08|emr = 2.0|fd = 2.0|mt_S = 00111|mt_P = 01111|mt_T = 11111
mr = 2.5e-08|emr = 2.0|fd = 2.0|mt_S = 00111|mt_P = 01111|mt_T = 11111
mr = 1e-07|emr = 2.0|fd = 2.0|mt_S = 00111|mt_P = 01111|mt_T = 11111
CR_Intrinsic (in Log10 Scale)
Age


Intrinsic_Contribution_CANCER (in Log10 Scale) - Extended Risk Model
17-Prostate
mr = 1e-08|emr = 2.0|fd = 2.0|mt_S = 00111|mt_P = 01111|mt_T = 11111
mr = 1.1e-08|emr = 2.0|fd = 2.0|mt_S = 00111|mt_P = 01111|mt_T = 11111
mr = 2.5e-08|emr = 2.0|fd = 2.0|mt_S = 00111|mt_P = 01111|mt_T = 11111
mr = 1e-07|emr = 2.0|fd = 2.0|mt_S = 00111|mt_P = 01111|mt_T = 11111
Intrinsic_Contribution_CANCER (in Log10 Scale)
Age



It is obvious that both observed risk and intrinsic risk are increasing with age for all tissues. For age x, cancer risk and intrinsic contribution above represents cumulative risk and average contribution percentage from age 0 to age x.

Among selected tissues except breast, intrinsic cancer risk under all conditions is much smaller than the observed risk at any age. For break cancer, its intrinsic risk exceeds the observed one with aggressive mutation rate and clonal expansion, before the age of 30; however, the intrinsic risk falls far below the observed curve after 30. More importantly, the intrinsic risk presents a greatly different trend than observed risk curve, for all tissues. This demonstrates significant contribution from non-intrinsic factors.

For most tissues, the intrinsic contribution percentages have an overall decreasing trend, indicating increasing importance of non-intrinsic factors to cancer onset as one gets older. Some tissues, especially with clonal expansion effects, present non-monotonous intrinsic contribution patterns. For example, the intrinsic contribution for prostate cancer was seen decreasing after age 30.

In general, our results demonstrate significant contribution to cancer risk from non-intrinsic factors in lifetime, and the percentage of intrinsic/non-intrinsic contributions varies with age.

5.2.3 Comparing Different Models

In this section, we compare original stem cell model, intermediate model and extended risk model. We will see that they yield close intrinsic risk values under the same parameter configuration. Only the extended risk model has the capability to incorporate mutation effects and clonal expansion, which is the main reason why the experiment results above were all based on the extended risk model. Figures 31 to 34 below compare Extended Risk Model, Original Stem Cell Model and Intermediate Model on their computed intrinsic risk under different mutation rates without mutation effects and clonal expansion.

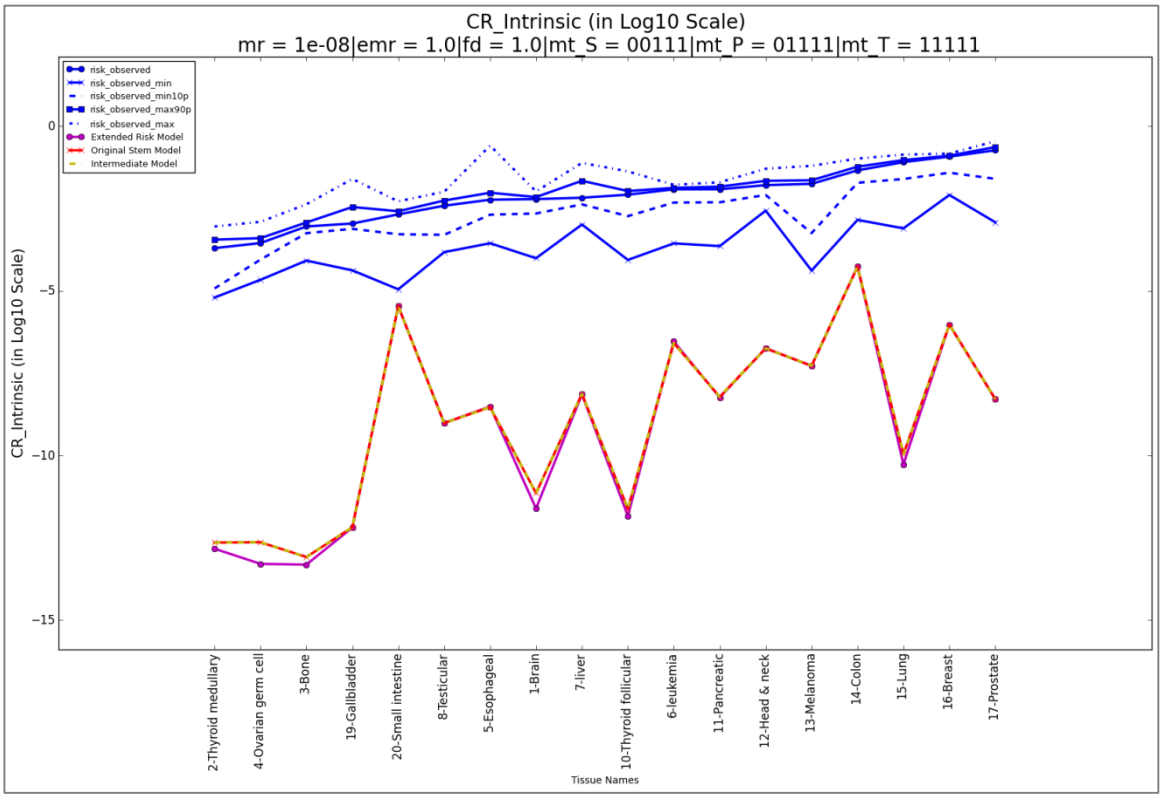


Figure 31: Comparison of Extended Risk Model, Original Stem Cell Model and Intermediate Model on computed lifetime intrinsic risk (log10 scale) computed from Extended Risk Model and statistics of NPCR observed risk in U.S. Tissue id/names are given below horizontal axis and the tissues are sorted in ascending order of "risk_observed", the average risk in U.S. Intrinsic mutation rate is selected to be {1e-08}; mutation effects and clonal expansion

factors are selected from $emr = \{1.0\}$ and $fd = \{1.0\}$; the default required mutation hits are ($mt_S = 00111, mt_P = 01111$ and $mt_T = 11111$).

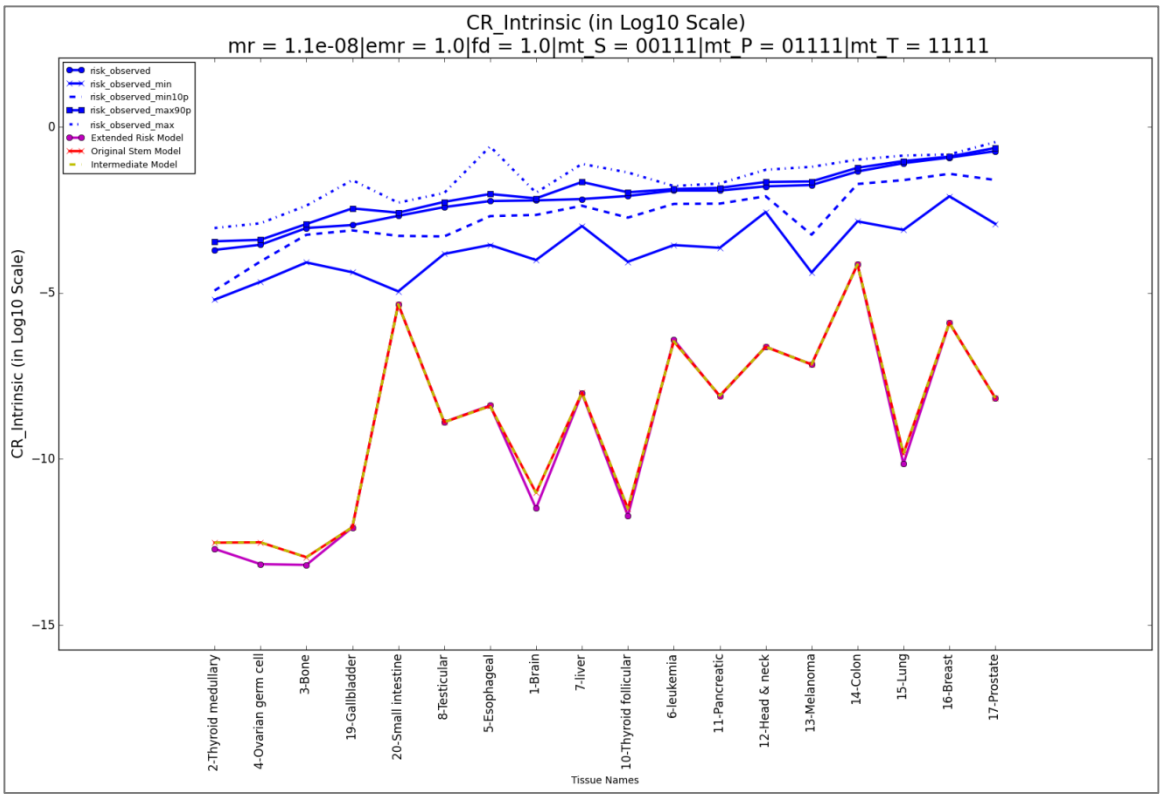


Figure 32: Comparison of Extended Risk Model, Original Stem Cell Model and Intermediate Model on computed lifetime intrinsic risk (log10 scale) computed from Extended Risk Model and statistics of NPCR observed risk in U.S. Tissue id/names are given below horizontal axis and the tissues are sorted in ascending order of "risk_observed", the average risk in U.S. Intrinsic mutation rate is selected to be {1.1e-08}; mutation effects and clonal expansion factors are selected from $emr = \{1.0\}$ and $fd = \{1.0\}$; the default required mutation hits are ($mt_S = 00111, mt_P = 01111$ and $mt_T = 11111$).

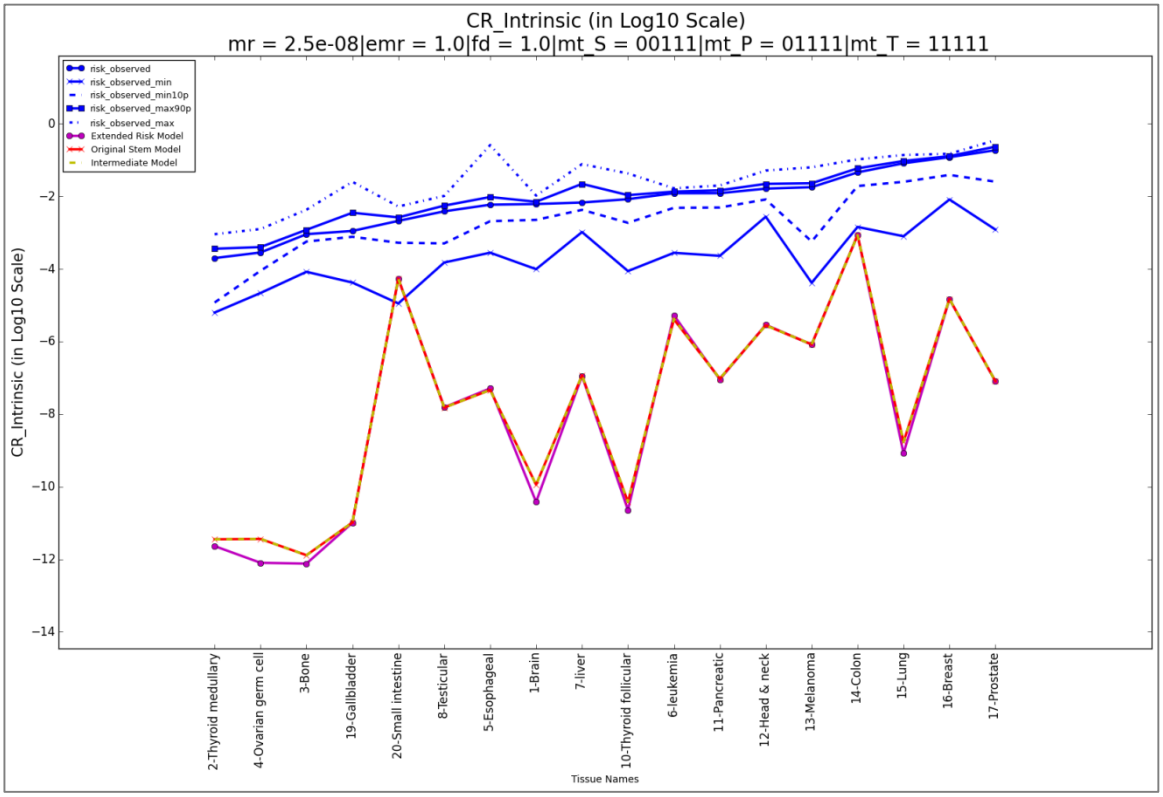


Figure 33: Comparison of Extended Risk Model, Original Stem Cell Model and Intermediate Model on computed lifetime intrinsic risk (log10 scale) computed from Extended Risk Model and statistics of NPCR observed risk in U.S. Tissue id/names are given below horizontal axis and the tissues are sorted in ascending order of "risk_observed", the average risk in U.S. Intrinsic mutation rate is selected to be {2.5e-08}; mutation effects and clonal expansion factors are selected from $emr = \{1.0\}$ and $fd = \{1.0\}$; the default required mutation hits are ($mt_S = 00111, mt_P = 01111$ and $mt_T = 11111$).

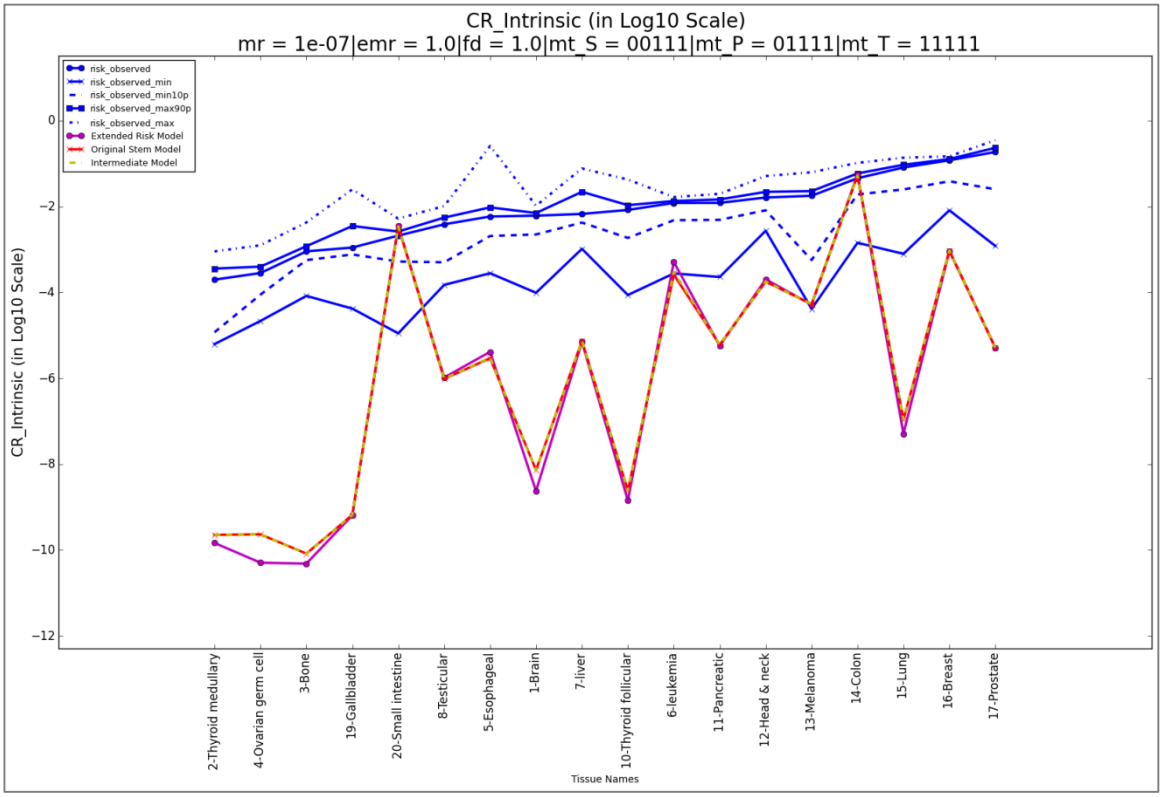


Figure 34: Comparison of Extended Risk Model, Original Stem Cell Model and Intermediate Model on computed lifetime intrinsic risk (log10 scale) computed from Extended Risk Model and statistics of NPCR observed risk in U.S. Tissue id/names are given below horizontal axis and the tissues are sorted in ascending order of "risk_observed", the average risk in U.S. Intrinsic mutation rate is selected to be {1e-07}; mutation effects and clonal expansion factors are selected from $emr = \{1.0\}$ and $fd = \{1.0\}$; the default required mutation hits are ($mt_S = 00111, mt_P = 01111$ and $mt_T = 11111$).

We can see that the three different models yielded very close intrinsic risks provided mutation effects and clonal expansion are ignored. The original stem cell model does not have progenitor branches, but it yields almost identical risk comparing to the intermediate models, which includes progenitor lineages. This is because progenitor lineages make very little contribution to cancer onset due to its short lineage length and more conservative criteria for cancer onset. In addition, the difference between original stem cell model and extended risk model can be explained by the derivations in Section 4.4. Figures 35 to 38 below compare the models when mutation effects and clonal expansion are included for extended risk model.

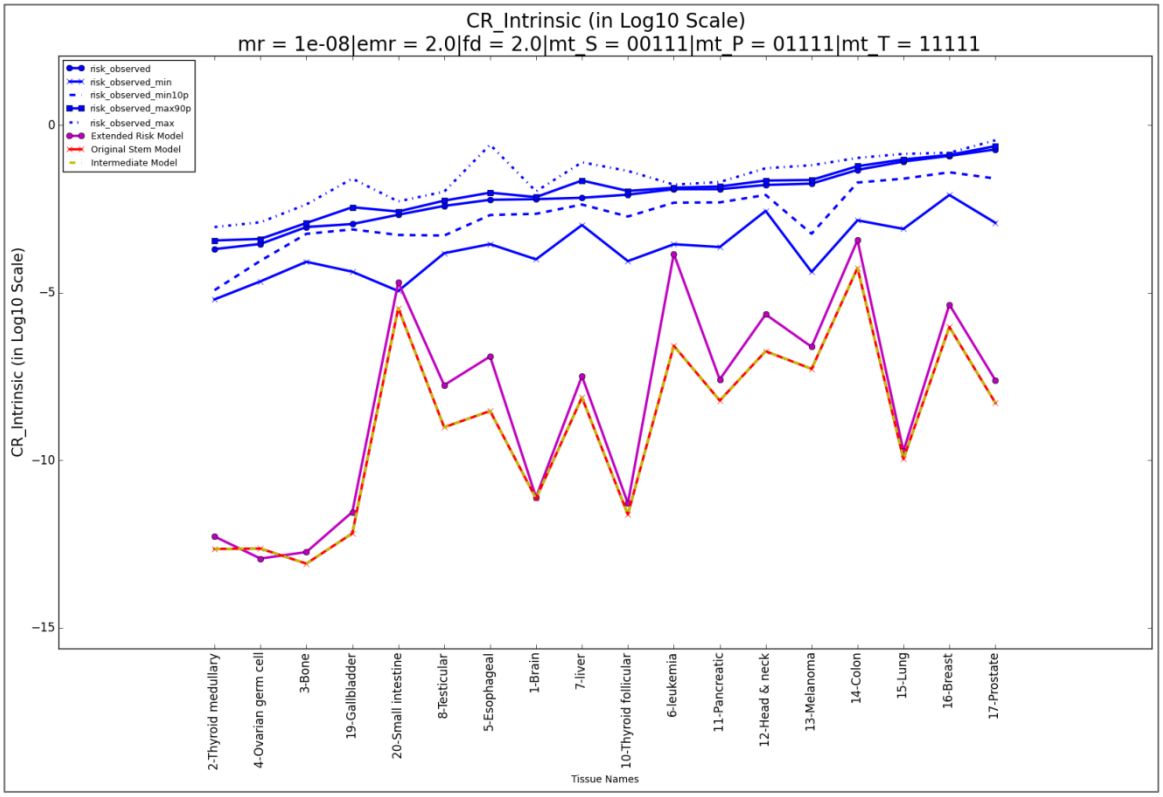


Figure 35: Comparison of Extended Risk Model, Original Stem Cell Model and Intermediate Model on computed lifetime intrinsic risk (log10 scale) computed from Extended Risk Model and statistics of NPCR observed risk in U.S. Tissue id/names are given below horizontal axis and the tissues are sorted in ascending order of "risk_observed", the average risk in U.S. Intrinsic mutation rate is selected to be {1e-08}; mutation effects and clonal expansion factors are selected from $emr = \{2.0\}$ and $fd = \{2.0\}$; the default required mutation hits are ($mt_S = 00111, mt_P = 01111$ and $mt_T = 11111$).

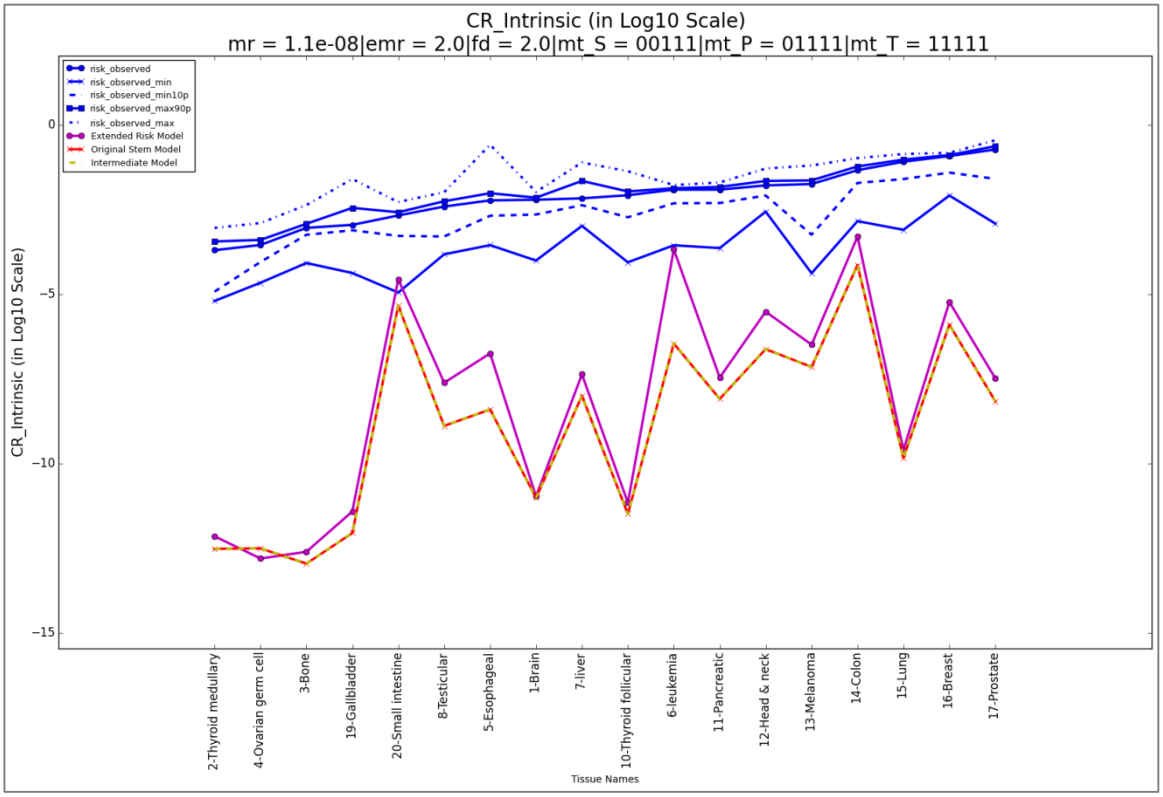


Figure 36: Comparison of Extended Risk Model, Original Stem Cell Model and Intermediate Model on computed lifetime intrinsic risk (log10 scale) computed from Extended Risk Model and statistics of NPCR observed risk in U.S. Tissue id/names are given below horizontal axis and the tissues are sorted in ascending order of "risk_observed", the average risk in U.S. Intrinsic mutation rate is selected to be {1.1e-08}; mutation effects and clonal expansion factors are selected from $emr = \{2.0\}$ and $fd = \{2.0\}$; the default required mutation hits are ($mt_S = 00111, mt_P = 01111$ and $mt_T = 11111$).

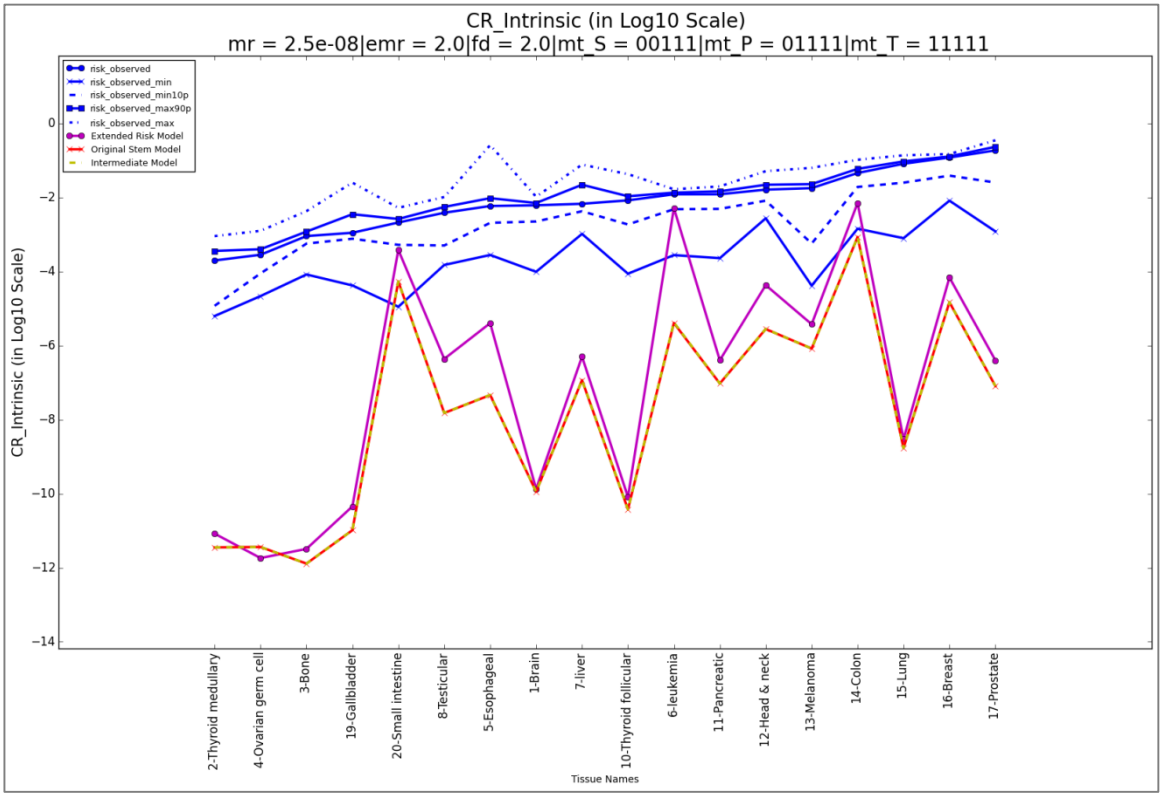


Figure 37: Comparison of Extended Risk Model, Original Stem Cell Model and Intermediate Model on computed lifetime intrinsic risk (log10 scale) computed from Extended Risk Model and statistics of NPCR observed risk in U.S. Tissue id/names are given below horizontal axis and the tissues are sorted in ascending order of "risk_observed", the average risk in U.S. Intrinsic mutation rate is selected to be {2.5e-08}; mutation effects and clonal expansion factors are selected from $emr = \{2.0\}$ and $fd = \{2.0\}$; the default required mutation hits are ($mt_S = 00111, mt_P = 01111$ and $mt_T = 11111$).

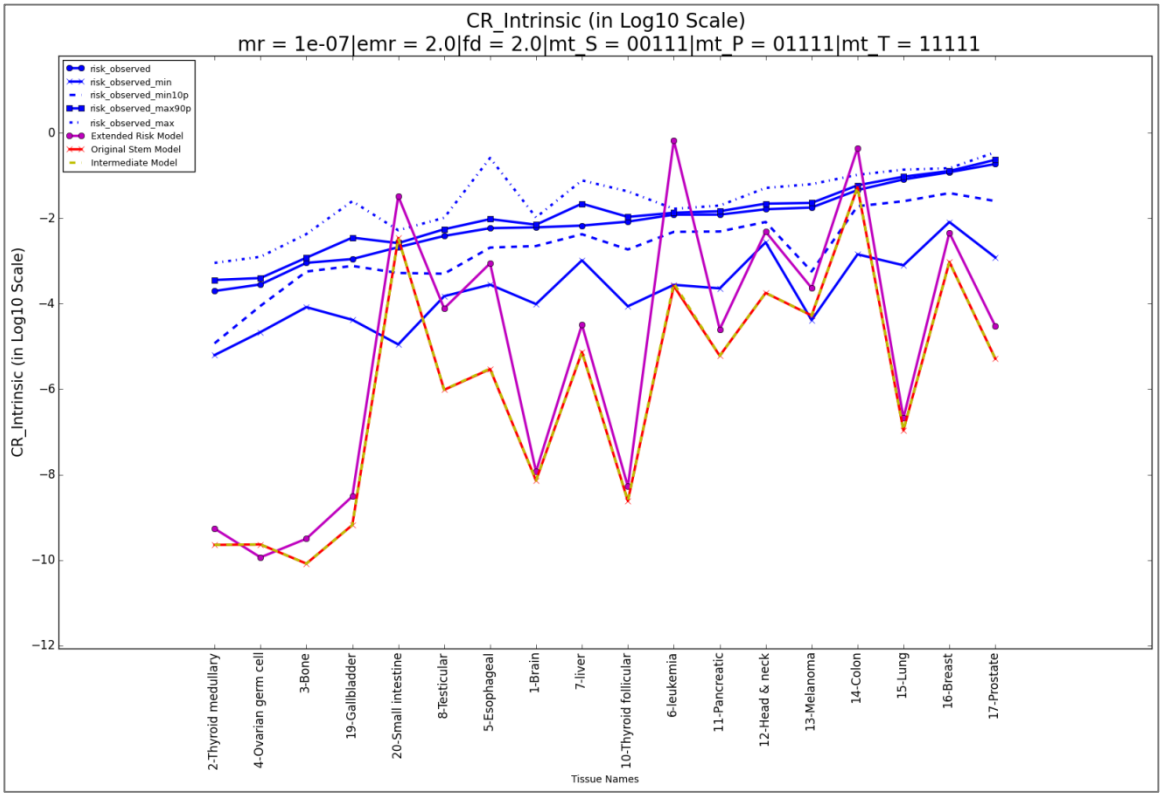


Figure 38: Comparison of Extended Risk Model, Original Stem Cell Model and Intermediate Model on computed lifetime intrinsic risk (log10 scale) computed from Extended Risk Model and statistics of NPCR observed risk in U.S. Tissue id/names are given below horizontal axis and the tissues are sorted in ascending order of "risk_observed", the average risk in U.S. Intrinsic mutation rate is selected to be {1e-07}; mutation effects and clonal expansion factors are selected from $emr = \{2.0\}$ and $fd = \{2.0\}$; the default required mutation hits are ($mt_S = 00111, mt_P = 01111$ and $mt_T = 11111$).

With mutation effects and clonal expansion, extended risk model gives significantly different results than those from other two models. However, intrinsic risk from different models still maintain most of the qualitative relations, indicating that all three models are able to capture inherent cell evolution mechanisms of each tissue that determine cancer onset and development due to intrinsic factors.

5.3 Discussion and Future Work

In this thesis, we built a comprehensive model for cell dynamics and cancer risk. Our original target was to compute the theoretical lifetime cancer risk due to intrinsic risk factors. For this purpose, we started from a very simple assumption that mutation

acquisition along stem cell divisions is the major cause for the risk of cancer [31, 34]. Our original stem cell model in Section 2.5.1 can quickly compute theoretical risk through a closed form formula derived from a discrete Markov process. The model can explain most of the cancer onset mechanisms for tissues with long stem lineages.

The original stem cell model has several limitations. First, it did not consider non-stem cell population, which accounts for more than 99% of a tissue in most cases (see Table 3). Second, it did not have an algorithm to build homeostasis within a tissue, in which the overall cell death rate approximately equals overall cell birth rate. Also, the algorithm to compute theoretical risk could give an overestimated risk, as explained in Section 4.4.

We generalized our original stem cell model to the intermediate risk model that incorporates progenitor lineages and terminal cells. We built the homeostatic condition that can be used to determine progenitor lineage length or cell cycle time. In this model different types of cells could require different number of mutations for to become cancer cells. For example, stem cells usually need 3 mutations, while progenitor cells need 4 or 5 mutations for cancer onset. We also developed a recursive algorithm for cancer risk based on a similar assumption as in the original stem cell model, that cells of the same generation within their lineage are independent in mutation acquisition. The intermediate risk model provided a way to capture the overall cell dynamics and thus its estimation of intrinsic risk is more convincing.

However, the intermediate risk model is unable to fully describe complicated process for the multi-stage cancer development. First, by using the number of acquired mutations as the state space, the intermediate model did not differentiate driver mutations that may have different effects on cell dynamics and mutation acquisition. In addition, the model

requires cells to divide at the same pace, which eludes the possibility of clonal expansion, where cells with certain mutations could divide faster. Moreover, its algorithm for computing theoretical risk could still lead to an overestimated theoretical risk.

Our extended risk model overcomes the above limitations and provides a highly flexible framework to simulate cell dynamics and cancer development. The model has an efficient algorithm to compute cancer risk and cell numbers at any given time point within an 80-year lifespan; in addition, the model incorporates mutation effects and clonal expansion. Currently the model assumes 5 driver mutations in total, but it can support any number of driver mutations. In addition, the extended risk model does not need to assume any fixed cell division patterns. It allows a general form of 6 cell division activities for both stem and progenitor cells with time varying division probabilities. More importantly, the extended risk model derives cancer risk computation based on the extended dependency structures within general cell divisions.

With the extended risk model, we analyzed the impact from different mutation effects, on the trend of cancer risk for each age within a lifetime. In addition, we evaluated the portion of cancer risk and the proportion of cancer mutations due to intrinsic risk factors alone using the metrics evaluated from the expected number of mutations.

Our analyses suggest that non-intrinsic factors are the major cause for cancer initiation for most cancer types under various conditions and parameter configurations.

As possible future work directions, the current model can be extended to provide more insight for the cancer development process. For example, we can incorporate more complicated regulation process and immune process, in which cancer cells could be

suppressed, as a simulation of cancer treatment. We can also extend the pool of driver mutations and analyze its influence on cancer risk.

Reference

[1] Ashkenazi, R., Gentry, S. N., & Jackson, T. L. (2008). Pathways to Tumorigenesis — Modeling Mutation Acquisition in Stem Cells and Their Progeny. *Neoplasia, 10*(11).

[2] Adams, P., Jasper, H., & Rudolph, K. (2015). Aging-Induced Stem Cell Mutations as Drivers for Disease and Cancer. *Cell Stem Cell, 16*(6), 601-612.

[3] Athreya, K. B., & Ney, P. (1972). *Branching process*. New York: Springer-Verlag.

[4] Barrett, J. C. (1993). Mechanisms of multistep carcinogenesis and carcinogen risk assessment. *Environmental Health Perspectives, 100*, 9-20.

[5] Bozic, I., Antal, T., Ohtsuki, H., Carter, H., Kim, D., Chen, S., . . . Nowak, M. A. (2010). Accumulation of driver and passenger mutations during tumor progression. *Proceedings of the National Academy of Sciences, 107*(43), 18545-18550.

[6] Bankhead, A., Magnuson, N. S., & Heckendorn, R. B. (2007). Cellular automaton simulation examining progenitor hierarchy structure effects on mammary ductal carcinoma in situ. *Journal of Theoretical Biology, 246*(3), 491-498.

[7] Durrett, R., & Moseley, S. (2010). Evolution of resistance and progression to disease during clonal expansion of cancer. *Theoretical Population Biology, 77*(1), 42-48.

[8] Frank, S. A., Iwasa, Y., & Nowak, M. A. (2003). Patterns of cell division and the risk of cancer. *Genetics*, *163*(4), 1527–1532.

[9] Gibbons, D. L., Byers, L. A., & Kurie, J. M. (2014). Smoking, p53 Mutation, and Lung Cancer. *Molecular Cancer Research : MCR*, *12*(1), 3–13.

[10] Grivennikov, S. I., Greten, F. R., & Karin, M. (2010). Immunity, Inflammation, and Cancer. *Cell*, *140*(6), 883–899.

[11] Gentry, S. N., & Jackson, T. L. (2013). A Mathematical Model of Cancer Stem Cell Driven Tumor Initiation: Implications of Niche Size and Loss of Homeostatic Regulatory Mechanisms. *PLoS ONE, 8*(8).

[12] Hannezo, E., Prost, J., & Joanny, J. F. (2014). Growth, homeostatic regulation and stem cell dynamics in tissues. *Journal of The Royal Society Interface, 11*(93), 20130895-20130895.

[13] Marx, V. (2014). Cancer genomes: Discerning drivers from passengers. *Nature Methods, 11*(4), 375-379.

[14] Hanahan, D., & Weinberg, R. A. (2016). The hallmarks of cancer. *Oxford Medicine Online*.

[15] Knudson, A. G. (2001). Two genetic hits (more or less) to cancer. *Nature Reviews Cancer* 1(2): 157-62.

[16] Korthauer, K. D., & Kendziorski, C. (2015). MADGiC: A model-based approach for identifying driver genes in cancer. *Bioinformatics, 31*(10), 1526-1535.

[17] Kinzler, K. W., & Vogelstein, B. (1996). Lessons from Hereditary Colorectal Cancer. *Cell, 87*(2), 159-170.

[18] Lin, H. (2002). The stem-cell niche theory: Lessons from flies. *Nature Reviews Genetics, 3*(12), 931-940.

[19] Luebeck, E. G., & Moolgavkar, S. H. (2002). Multistage carcinogenesis and the incidence of colorectal cancer. *Proceedings of the National Academy of Sciences, 99*(23), 15095-15100.

[20] Maley, C. C. (2004). Selectively Advantageous Mutations and Hitchhikers in Neoplasms: P16 Lesions Are Selected in Barrett's Esophagus. *Cancer Research, 64*(10), 3414-3427.

[21] Morrison, S. J., & Kimble, J. (2006). Asymmetric and symmetric stem-cell divisions in development and cancer. *Nature, 441*(7097), 1068-1074.

[22] Morrison, S. J., Shah, N. M., & Anderson, D. J. (1997). Regulatory Mechanisms in Stem Cell Biology. *Cell, 88*(3), 287-298.

[23] Owens, D. M., & Watt, F. M. (2003). Contribution of stem cells and differentiated cells to epidermal tumours. *Nature Reviews Cancer, 3*(6), 444-451.

[24] Mcgurk, S. (2013). The Molecular Biology of Cancer: A Bridge from Bench to Bedside – Second edition The Molecular Biology of Cancer: A Bridge from Bench to Bedside – Second edition. *Nursing Standard, 28*(1), 30-30.

[25] Simpson, A. J. (2009). Sequence-based advances in the definition of cancer-associated gene mutations. *Current Opinion in Oncology, 21*(1), 47-52.

[26] Simons, B., & Clevers, H. (2011). Strategies for Homeostatic Stem Cell Self-Renewal in Adult Tissues. *Cell, 145*(6), 851-862.

[27] Sjoblom, T., Jones, S., Wood, L. D., Parsons, D. W., Lin, J., Barber, T. D., . . . Velculescu, V. E. (2006). The Consensus Coding Sequences of Human Breast and Colorectal Cancers. *Science, 314*(5797), 268-274.

[28] Sun, X., & Yu, Q. (2015). Intra-tumor heterogeneity of cancer cells and its implications for cancer treatment. *Acta Pharmacologica Sinica, 36*(10), 1219-1227.

[29] Tomasetti, C., Marchionni, L., Nowak, M. A., Parmigiani, G., & Vogelstein, B. (2014). Only three driver gene mutations are required for the development of lung and colorectal cancers. *Proceedings of the National Academy of Sciences, 112*(1), 118-123.

[30] Jackson, A. L., & Loeb, L. A. (1998). The mutation rate and cancer. Genetics, 148(4), 1483-1490.

[31] Tomasetti, C., & Vogelstein, B. (2015). Variation in cancer risk among tissues can be explained by the number of stem cell divisions. *Science, 347*(6217), 78-81.

[32] Weinberg, R. A. (2014). *The biology of cancer*. New York: Garland Science.

[33] Waclaw, B., Bozic, I., Pittman, M. E., Hruban, R. H., Vogelstein, B., & Nowak, M. A. (2015). A spatial model predicts that dispersal and cell turnover limit intratumour heterogeneity. *Nature*, *525*(7568), 261–264.

[34] Wu, S., Powers, S., Zhu, W., & Hannun, Y. A. (2016). Substantial contribution of extrinsic risk factors to cancer development. *Nature*, *529*(7584), 43–47.

[35] Wodarz, D., & Zauber, A. G. (2015). Cancer: Risk factors and random chances. *Nature, 517*(7536), 563-564.

[36] Lifetime Risk of Developing or Dying From Cancer. (n.d.). Retrieved June 08, 2017, from https://www.cancer.org/cancer/cancer-basics/lifetime-probability-of-developing-or-dying-from-cancer.html

[37] The Editors of Encyclopædia Britannica. (n.d.). Somatic mutation. Retrieved June 08, 2017, from https://www.britannica.com/science/somatic-mutation

[38] Talk: Adult stem cell. (2017, May 20). Retrieved June 08, 2017, from https://en.wikipedia.org/wiki/Talk:Adult_stem_cell

[39] Frank, S. A. (2007). *Dynamics of cancer: Incidence, inheritance, and evolution*. Princeton, NJ: Princeton University Press.

[40] Little, M. P., & Hendry, J. H. (2017). Mathematical models of tissue stem and transit target cell divisions and the risk of radiation- or smoking-associated cancer. *PLoS Computational Biology*, *13*(2), e1005391.

[41] Tomasetti, C., Li, L., & Vogelstein, B. (2017). Stem cell divisions, somatic mutations, cancer etiology, and cancer prevention. *Science, 355*(6331), 1330-1334.