The age incidence of chronic myeloid leukemia can be explained by a one-mutation model

Franziska Michor*†, Yoh Iwasa‡§, and Martin A. Nowak§

§Program for Evolutionary Dynamics, Departments of Mathematics and Organismic and Evolutionary Biology, and *Society of Fellows, Harvard University, Cambridge, MA 02138; and ‡Department of Biology, Kyushu University, Fukuoka 812-8581, Japan

Chronic myeloid leukemia (CML) is associated with the Philadelphia chromosome, which arises by a reciprocal translocation between chromosomes 9 and 22 and harbors the BCR-ABL fusion oncogene. It is unknown whether any other mutations are needed for the chronic phase of the disease. The CML incidence increases as a function of age with an exponent of ≈3. A slope of 3 could indicate that there are two mutations, in addition to the Philadelphia translocation, that have not yet been discovered. In this work, we explore an alternative hypothesis: We study a model of cancer initiation requiring only a single mutation. A mutated cell has a net reproductive advantage over normal cells and, therefore, might give rise to clonal expansion. The cancer is detected with a probability that is proportional to the size of the mutated cell clone. This model has three waiting times: (*i*) the time until a mutated cell is produced, (*ii*) the time of clonal expansion, and (*iii*) the time until the clone is detected. Surprisingly, this simple process can give rise to cancer incidence curves with exponents up to 3. Therefore, the CML incidence data are consistent with the hypothesis that the Philadelphia translocation alone is sufficient to cause chronic phase CML.

cancer incidence | mathematical biology | stochastic dynamics

Chronic myeloid leukemia (CML) is a malignant clonal disorder of the hematopoietic system that leads to increased numbers of myelocytes, erythrocytes, and thrombocytes in peripheral blood (1, 2). The molecular hallmark of CML is the Philadelphia chromosome, first described as shortened chromosome 22 in 1960 (3) and then as a reciprocal t(9;22) translocation in 1973 (4). This abnormal chromosome is found in cells from the myeloid, erythroid, megakaryocytic, and B lymphoid lineages, indicating the presence of a "cancer stem cell" that is capable of producing several types of differentiated cancer cells (2). The Philadelphia chromosome is present in 95% of patients. The remaining 5% have complex or variant translocations that have the same result: fusion of the breakpoint cluster region (BCR) gene on chromosome 22 to the Ableson leukemia virus (ABL) gene on chromosome 9. The chimeric oncogene BCR-ABL encodes a constitutively active cytoplasmic tyrosine kinase. This protein activates growth and differentiation pathways in hematopoietic cells; its effectors include RAS, RAF, MYC, JUN, STAT, and phosphatidylinositol 3-kinase (1). BCR-ABL initiates a process to transform hematopoietic cells such that their growth and survival become independent of cytokines (5).

The question of whether BCR-ABL is necessary and sufficient to cause CML has been addressed in different ways. Several laboratories could reproduce a CML-like disease in mice expressing the BCR-ABL oncogene alone (6–8) or in combination with v-abl (9). Exposure to ionizing irradiation was shown to increase the risk of acquiring CML, but disease usually develops only after a prolonged latent period (10); this finding could indicate that several mutations are necessary for CML, but it also could mean that the Philadelphia-positive cell clone has a slow rate of expansion. Finally, ≈30% of healthy individuals express the BCR-ABL oncogene at very low levels (11, 12). This observation, however, does not prove that additional genetic changes are necessary for CML because it can alternatively be explained by the expression of BCR-ABL in non-self-renewing differentiated cells rather than hematopoietic stem cells. In the latter case, the continuous production of healthy hematopoietic cells will eventually replace the Philadelphia-positive cell clone and the disease will be "washed out" of the system (13). Hence, the experimental evidence of how many mutations are necessary to cause CML is inconclusive.

The multistep theory of carcinogenesis was inspired by the observation that the cancer incidence increases as a higher order function of age. In the 1950s, Nordling, Armitage, and Doll (14, 15) pointed out that the cancer incidence data could be explained by an increasing somatic mutation rate with age, or by the requirement of several events to cause cancer (16). In 1957, a simple deterministic model of two mutations was shown to produce incidence curves with slope six if the intermediate mutants lead to clonal expansion (17). Later, Moolgavkar *et al.* (16, 18, 19) introduced modeling of incidence curves based on multistate branching processes including cell proliferation and death.

Based on these insights, many mathematical models have been developed that investigate the number of mutations needed to cause particular kinds of cancer (19–22). A model fitted to the age-specific incidence of colorectal cancer proposes that two rare events and one high-frequency event suffice to initiate clonal expansion of the mutated stem cell, and only one more mutation is needed to give rise to an adenomatous polyp (19). The age-specific incidence of CML was used to calibrate a multistage model and predicts that three stem cell mutations are necessary for the chronic phase of the disease (20). These models neglect the population size and structure of the cells that are prone to accumulating mutations. However, the population genetics of susceptible cells must be considered if meaningful conclusions shall be drawn. For example, depending on the effective population size as compared with the mutation rate, it can take two, one, or even zero rate-limiting hits to accumulate two genetic alterations (23–26). The number of rate-limiting hits also can be larger than the actual number of mutations because of time delays due to clonal expansion (17). Therefore, the number of mutations necessary to cause a particular type of cancer cannot simply be read off the age-specific incidence data, and mathematical analyses must take into account population size and structure.

In the following, we offer a population genetics analysis of the dynamics of cancer initiation and its epidemiological consequences. Our model is motivated by CML but does not exclusively apply to this disease; rather, it is a general approach to studying many different types of cancer. We use CML as a specific example to show that a single mutation model can generate cancer incidence curves with an exponent of up to 3.
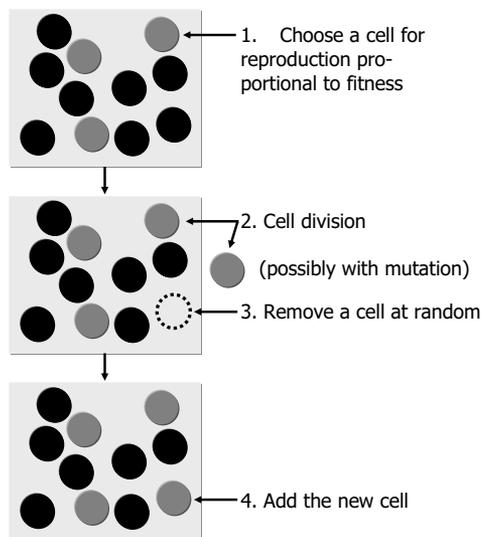
MEDICAL SCIENCES

Fig. 1. The Moran process. Initially, all N cells are wild type and have a relative growth rate of 1. At each time step, a cell is chosen for reproduction proportional to fitness, and its daughter cell replaces another randomly chosen cell. The total number of cells remains strictly constant. Wild-type cells give rise to mutated cells at rate u per cell division. Mutated cells have a relative growth rate r. Cells divide every τ days.

Results and Discussion

Assume that a population of N (hematopoietic stem) cells proliferate according to the Moran process (27, 28): Initially, all cells are wild type. Cells divide every τ days. At each time step, a cell is chosen for reproduction proportional to fitness. The newly produced daughter cell replaces another randomly chosen cell (Fig. 1). Hence, the population size remains strictly constant. A wild-type cell gives rise to a mutated cell with probability u per cell division. Back mutation is neglected. A mutated cell has relative fitness r: if $r = 1$, then the mutation is neutral and the cell has the same growth rate as a wild-type cell; if $r > 1$, the mutant is advantageous, and if $r < 1$, the mutant is disadvantageous as compared with wild-type cells. We assume that the probability to diagnose the disease is linearly proportional to the number of mutated cells in the population. If there are x_1 mutant cells, the rate of detection is qx_1. We perform exact numerical simulations of this process and compare the results with analytic approximations.

In *Methods*, we show that the probability of detecting the disease before time t is given by

$$P(t) = \int_0^t \left[1 - \left(1 + \frac{e^{cz} - 1}{N(1 - 1/r)} \right)^{-qN/c} \right] e^{-b(t-z)} b \, dz. \quad [\mathbf{1}]$$

Here, $b = Nu(1 - 1/r)/\tau$ and $c = (r - 1)/\tau$. The stochastic process is characterized by three waiting times: (*i*) the waiting time until production of the first successful mutated cell is given by $1/b$; (*ii*) the time for clonal expansion of its lineage is given by $\ln[N(1 - 1/r)]/c$, and (*iii*) the time until detection of the disease is given by $1/qN$. This stochastic process can give rise to incidence curves with exponents up to 3. Define t_0 as the age at which we are interested in the slope of the incidence; for example, $t_0 = 50$ years. The factor b is always much less than the inverse of t_0 because otherwise, almost everybody would be diagnosed with cancer. Patients are diagnosed with a probability proportional to their number of leukemic stem cells, and if this probability is small, then the average time to detection might be >100 years. If $qN \gg 1/t_0$ and $c \gg 1/t_0$, then the incidence increases with an exponent of 1. In this case, the waiting time is dominated by the time it takes to produce the first successful mutant, and the time needed for clonal expansion and
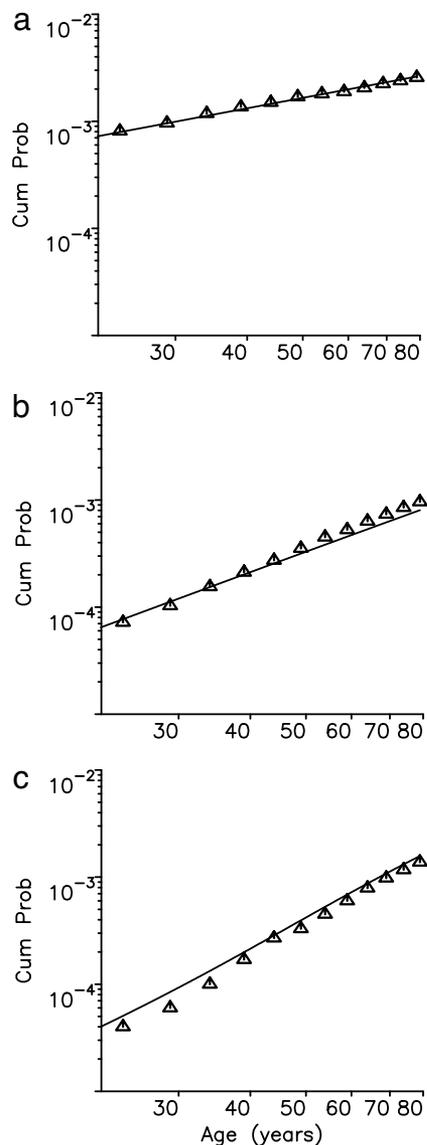
Fig. 2. Simulation and theory. The figure shows the results of the exact stochastic computer simulation (triangles) and Eq. **1** (lines). Parameter values are mutation rate $u = 10^{-9}$, population size $N = 1,000$, relative fitness of mutated cells $r = 1.1$, mean cell generation time $\tau = 1$ day, and probability of detection $q/N = \alpha = 0.1$ for an exponent of 1 (*a*); $u = 10^{-5}$, $N = 2,000$, $r = 1.001$, $\tau = 100$, and $\alpha = 2 \times 10^{-5}$ for an exponent of 2 (*b*); and $u = 10^{-6}$, $N = 2,000$, $r = 1.01$, $\tau = 40$, and $\alpha = 2 \times 10^{-5}$ for an exponent of 3 (*c*).

detection is negligible and, hence, does not contribute to the slope. If $qN \ll 1/t_0$ and $c \gg 1/t_0$, then the slope is $2 + \alpha$, where α is between 0 and 1. Finally, if $qN \ll 1/t_0$ and $c \ll 1/t_0$, then the slope can take a value of 3. This contribution is only important if the time for clonal expansion is sufficiently long, i.e., if the mutated cell clone has a small fitness advantage. Therefore, a simple one-mutation model can give rise to incidence curves with an exponent of up to 3. Fig. 2 shows the exact stochastic simulation and Eq. **1** shows exponents 1–3.

The age-specific incidence data for CML were obtained from the Surveillance, Epidemiology, and End Results (SEER) registry, which covers ≈10% of the U.S. population (www.seer.cancer.gov). In total, 5,256 CML cases were observed between 1973 and 2002, and these cases were recorded in age classes of five years (Table 1, columns 1 and 2). To calculate the probability to get CML per year from this data, some adjustments are made. First, the cases re-

Table 1. CML incidence data

Age classes, years	Cases in SEER (1973–2002)	Cases per year	U.S. standard population	Adjusted cumulative probability
1–4	20	17	15,191,619	0.0000055
5–9	22	18	19,919,840	0.0000102
10–14	27	23	20,056,779	0.0000159
15–19	50	42	19,819,518	0.0000264
20–24	92	77	18,257,225	0.0000475
25–29	147	123	17,722,067	0.0000823
30–34	189	158	19,511,370	0.0001228
35–39	225	188	22,179,956	0.0001653
40–44	241	202	22,479,229	0.0002103
45–49	281	235	19,805,793	0.0002698
50–54	351	294	17,224,359	0.0003552
55–59	436	365	13,307,234	0.0004925
60–64	508	425	10,654,272	0.0006922
65–69	569	476	9,409,940	0.0009455
70–74	611	511	8,725,574	0.0012388
75–79	635	532	7,414,559	0.0015973
80–84	465	389	4,900,234	0.0019944
85+	379	317	4,259,173	0.0023666

Shown are the number of CML cases in the SEER registry (1973–2002) for age classes of five years (columns 1 and 2), cases per year (column 3), and the U.S. Census data from 2000 (column 4) that is used to calculate the cumulative probability per year to be diagnosed with CML until a given age class (column 5). Further explanations are given in the text.

corded in SEER have to be normalized to account for the number of cases per year. In 2000, there were $\approx$4,400 CML cases (29), hence, each SEER entry is multiplied with 4,400/5,256 (Table 1, column 3). This operation gives the number of CML cases per year per age class. To obtain the probability to get CML per age class per year, the number of cases has to be normalized by the number of susceptible people. Hence, we divide the cases by the U.S. Census data from 2000 (Table 1, column 4) and get probabilities p_i to be diagnosed with CML per year of age. Finally, these probabilities p_i are used to calculate the probabilities P_k to be diagnosed with CML anytime before age k (measured in years). We have

$$P_k = 1 - \prod_{i=1}^{k} (1 - p_i)$$

(Table 1, column 5). The resulting incidence curve is a nearly straight line on a doubly logarithmic plot with slope 2.86.

We compare these data with the direct computer simulation and with Eq. **1** and find that our simple one-mutation model can be fit to the CML incidence curve (Fig. 3). Hence, the Philadelphia chromosome alone might be sufficient to cause chronic phase CML.

In this work, we have shown that a simple one-mutation model can give rise to cancer incidence curves with an exponent up to 3. We assume that the susceptible cells proliferate according to a stochastic process. The probability of diagnosis of the disease increases linearly with the number of mutated cells. We compare exact simulations of this process with an analytical formula for the cumulative probability of being diagnosed until a certain age. The stochastic process is characterized by three waiting times: the time until the first successful mutant is produced, the time of clonal expansion, and the time until detection of the disease. If the time until production of the first successful mutant is the only dominant-waiting time, then the resulting incidence data increase with exponent 1. Depending on the magnitude of the other two waiting times, epidemiological data with exponents up to 3 can be obtained. These results, together with previous work on the dynamics of tumor

Fig. 3. CML incidence. The figure shows the numerical simulation of Eq. **1** (line) and the adjusted cumulative CML incidence data from Table 1, column 5 (circles). Parameter values are $u = 3 \times 10^{-8}$, $N = 10^5$, $r = 1.01$, $\tau = 60$, and $\alpha = 10^{-3}$ (a) and 5×10^{-10}, $N = 10^6$, $r = 1.02$, $\tau = 100$, and $\alpha = 0.1$ (b).

suppressor gene inactivation (23–26), suggest that the incidence curves of cancer cannot be used to estimate the number of mutations necessary to cause the disease. Population size, mutation rates, cell division times, and fitness values all interact to determine the number of rate-limiting hits a mutation causes, and this number can be larger or smaller than the actual number of mutations. Thus, a meaningful mathematical analysis of cancer incidence must take the population genetics of the susceptible cells into account.

As a specific example, we have studied the age-specific incidence data of CML, which was obtained from the SEER registry. We calculate the cumulative probability of developing CML from the SEER cases and find that the incidence has a slope of 2.86 on a doubly logarithmic plane. We fit this slope with the one-mutation model and identify appropriate parameter values. Therefore, the hypothesis that the Philadelphia chromosome alone is sufficient to cause chronic phase CML is consistent with the observed incidence curve (16, 30). However, further experimental efforts must be made to firmly establish whether BCR-ABL is sufficient to cause chronic phase CML.

Methods

We consider a population of N cells by following the Moran process with a mean cell generation time of τ. Initially all cells are wild type. A mutant cell is produced with probability u per cell division and has a relative fitness r. We assume that $r > 1$. Diagnosis of the

disease occurs as a stochastic event described as a continuous-time Markovian transition model. The probability of detection increases proportionally to the number of mutant cells: If there are x_1 mutant cells, the rate of detection is qx_1. Let $P(t)$ be the probability that the cancer is detected before time t; it is the cumulative incidence of cancer. The incidence curves of many cancers (apart from childhood cancers) are approximately straight lines on a doubly logarithmic plot.

The stochastic process can be calculated as follows: at a random time, a mutation occurs that eventually will become fixed in the population. The appearance of the first successful mutant is a random event occurring at rate $b = Nu_1(1 - 1/r)/\tau$. This rate is the product of the expected number of mutants produced per unit of time and the probability of nonextinction of the cell lineage. An advantageous mutation may go extinct because of the stochasticity caused by the small population size, but once the mutant cell number increases to sufficiently many, the lineage grows deterministically, and extinction can be neglected. We approximate the clonal expansion of mutant cells by the logistic equation

$$\frac{dx}{da} = cx(1 - x), \quad \text{with } x(0) = 1/N. \qquad [2]$$

Here, a is the time since the mutation occurred, $x(a)$ is the frequency of mutants in the population at that time where $0 < x(a) < 1$, and $c = (r - 1)/\tau$. This equation implies that the frequency of mutants increases as $x(a) = 1/(1 + (N - 1)\exp[-ca])$.

The probability of cancer detection per unit time is given by qNx. Hence, the cumulative risk of cancer detection at or before time t is given by

$$qN\int_0^t x(a)da,$$

if the mutant arises at time $t = 0$ and its lineage grows until time t after the deterministic growth rule given by Eq. 2. The time of appearance of the first successful mutant, m, is given by a negative exponential distribution, $b = Nu_1(1 - 1/r)/\tau$. This stochasticity is much greater than the stochasticity caused by the clonal expansion; the latter is approximated by Eq. 2. The derivation of the following formula neglects the stochasticity related to clonal expansion.

The risk of detection at or before time t is given by

$$qN\int_0^{t-m} x(a)da.$$

There is a small probability that the first successful mutant arises exceptionally early (small m); for this unlucky patient, the risk is much higher than for the average person, and it is those unlucky patients that will be diagnosed first. Therefore, we have to average the risk with respect to m. The probability of detection for a particular patient who has a successful mutation event at m is given by

$$1 - \exp\left[-qN\int_0^{t-m} x(a)da \right], \qquad [3]$$

and the averaged risk of diagnosis is given by

$$P(t) = \int_0^t \left\{ 1 - \exp\left[-qN\int_0^{t-m} \frac{da}{1 + (N - 1)\exp[-ca]} \right] \right\}$$
$$\cdot e^{-bm}b\,dm. \qquad [4]$$

Here, the frequency of the mutant cells conditional to fixation is given by the deterministic model, Eq. 2. However, some trajectories go extinct, and, hence, the deterministic model is an underestimation for the trajectory of clonal growth conditional to fixation. For a more accurate approximation, we use the initial condition $x(0) = 1/[N(1 - 1/r)]$ in Eq. 2. Hence, we have

$$P(t)$$

$$= \int_0^t \left\{ 1 - \exp\left[-qN\int_0^{t-m} \frac{da}{1 + (N(1 - 1/r) - 1)\exp[-ca]} \right] \right\}$$
$$\cdot e^{-bm}b\,dm. \qquad [5]$$

If the integral is calculated with respect to a, we have

$$P(t) = \int_0^t \left\{ 1 - \exp\left[-\frac{qN}{c} \ln\left(1 + \frac{e^{cz} - 1}{N(1 - 1/r)} \right) \right] \right\} e^{-b(t-z)}b\,dz,$$

$$[6]$$

which also can be written as Eq. 1 in *Results*.

1. Sawyers CL (1999) *N Engl J Med* 340:1330–1340.
2. Reya T, Morrison SJ, Clarke MF, Weissman IL (2001) *Nature* 414:105–111.
3. Nowell PC (1997) *Cancer Genet Cytogenet* 94:13–19.
4. Rowley JD (1973) *Nature* 243:290–293.
5. Gishizky ML, Witte ON (1992) *Science* 256:836–839.
6. Heisterkamp N, Jenster G, ten Hoeve J, Zovich D, Pattengale PK, Groffen J (1990) *Nature* 344:251–253.
7. Daley GQ, VanEtten RA, Baltimore D (1990) *Science* 247:824–830.
8. Elefanty AG, Hariharan IK, Cory S (1990) *EMBO J* 9:1069–1078.
9. Kelliher MA, McLaughlin J, Witte ON, Rosenberg N (1990) *Proc Natl Acad Sci USA* 87:6649–6653.
10. Lichtman M (1995) *Williams Hematology* (McGraw–Hill, New York).
11. Biernaux C, Loos M, Sels A, Huez G, Stryckmans P (1996) *Blood* 86:3118–3122.
12. Bose S, Deininger M, Gora-Tybor J, Goldman JM, Melo JV (1998) *Blood* 92:3362–3367.
13. Nowak MA, Michor F, Iwasa Y (2003) *Proc Natl Acad Sci USA* 100:14966–14969.
14. Nordling CO (1953) *Br J Cancer* 7:68–72.
15. Armitage P, Doll R (1954) *Br J Cancer* 8:1–12.
16. Moolgavkar SH, Knudson AG, Jr (1981) *J Natl Cancer Inst* 66:1037–1052.
17. Armitage P, Doll R (1957) *Br J Cancer* 11:161–169.
18. Moolgavkar SH, Venzon DJ (1979) *Math Biosci* 47:55–77.
19. Luebeck EG, Moolgavkar SH (2002) *Proc Natl Acad Sci USA* 99:15095–15100.
20. Vickers M (1996) *Br J Haematol* 94:1–4.
21. Frank SA (2004) *Curr Biol* 14:242–246.
22. Frank SA (2005) *Proc Natl Acad Sci USA* 102:242–246.
23. Nowak MA, Michor F, Komarova NL, Iwasa Y (2004) *Proc Natl Acad Sci USA* 101:10635–10638.
24. Iwasa Y, Michor F, Nowak MA (2004) *Genetics* 166:1571–1579.
25. Komarova NL, Sengupta A, Nowak MA (2003) *J Theor Biol* 223:433–450.
26. Michor F, Iwasa Y, Nowak MA (2004) *Nat Rev Cancer* 4:197–206.
27. Moran P (1962) *The Statistical Processes of Evolutionary Theory* (Clarendon, Oxford).
28. Nowak MA, Sasaki A, Taylor C, Fudenberg D (2004) *Nature* 428:646–650.
29. American Cancer Society (2000) *Cancer Facts and Figures 2000* (Am Cancer Soc, Atlanta).
30. Knudson AG (2001) *Nat Rev Cancer* 1:157–162.

Michor *et al.*