

Timing and heterogeneity of mutations associated with drug resistance in metastatic cancers

Ivana Bozic^{a,1} and Martin A. Nowak^{a,b,1}

^aProgram for Evolutionary Dynamics, Department of Mathematics, and ^bDepartment of Organismic and Evolutionary Biology, Harvard University, Cambridge, MA 02138

Edited by Herbert Levine, Rice University, Houston, TX, and approved October 8, 2014 (received for review June 28, 2014)

Targeted therapies provide an exciting new approach to combat human cancer. The immediate effect is a dramatic reduction in disease burden, but in most cases, the tumor returns as a consequence of resistance. Various mechanisms for the evolution of resistance have been implicated, including mutation of target genes and activation of other drivers. There is increasing evidence that the reason for failure of many targeted treatments is a small preexisting subpopulation of resistant cells; however, little is known about the genetic composition of this resistant subpopulation. Using the novel approach of ordering the resistant subclones according to their time of appearance, here we describe the full spectrum of resistance mutations present in a metastatic lesion. We calculate the expected and median number of cells in each resistant subclone. Surprisingly, the ratio of the medians of successive resistant clones is independent of any parameter in our model; for example, the median of the second clone divided by the median of the first is $\sqrt{2} - 1$. We find that most radiographically detectable lesions harbor at least 10 resistant subclones. Our predictions are in agreement with clinical data on the relative sizes of resistant subclones obtained from liquid biopsies of colorectal cancer patients treated with epidermal growth factor receptor (EGFR) blockade. Our theory quantifies the genetic heterogeneity of resistance that exists before treatment and provides information to design treatment strategies that aim to control resistance.

cancer | drug resistance | heterogeneity | mathematical biology

Acquired resistance to treatment is a major impediment to successful eradication of cancer. Patients presenting with early-stage cancers can often be cured surgically, but patients with metastatic disease must be treated with systemic therapies (1). Traditional treatments such as chemotherapy and radiation that exploit the enhanced sensitivity of cancer cells to DNA damage have serious side effects and, although curative in some cases, often fail due to intrinsic or resistance acquired during treatment. Targeted therapies, a new class of drugs, inhibit specific molecules implicated in tumor development and are typically less harmful to normal cells compared with chemotherapy and radiation (2–5). In the case of many targeted treatments, patients initially have a dramatic response (6, 7), only to be followed by a regrowth of most of their lesions several months later (8–10). Acquired resistance is often a consequence of genetic alterations (usually point mutations) in the drug target itself or in other genes (10–14).

Recently, mathematical modeling and clinical data were used to show that acquired resistance to an epidermal growth factor receptor (EGFR) inhibitor panitumumab in metastatic colorectal cancer patients is a *fait accompli*, because typical detectable metastatic lesions are expected to contain hundreds of cells resistant to the drug before the start of treatment (10). These cells would then expand during treatment, repopulate the tumor, and cause treatment failure. Similar conclusions should hold for targeted treatments of other solid cancers (15). Successful treatment requires drugs that are effective against the preexisting resistant subpopulation and must take into account the (possible) heterogeneity of resistance mutations present in the patient's lesions. In this article we use mathematical

modeling to investigate the heterogeneity of drug-resistant mutations in patients with metastatic cancers.

First mathematical investigations of the evolution of resistance to cancer therapy were concerned with calculating the probability that cells resistant to chemotherapy are present in a tumor of a certain size (16). Later studies expanded these results to include the effects of a fitness advantage or disadvantage provided by resistance mutations (17, 18), multiple mutations needed to achieve resistance to several drugs (15, 19–21), and density limitations caused by geometric constraints (22). These studies used generalizations of the famous Luria–Delbrück model for accumulation of resistant cells in exponentially growing bacterial populations (23). Probability distribution for the number of resistant cells in a population of a certain size in the fully stochastic formulation of the Luria–Delbrück model was recently calculated in the large population size limit (24, 25). The focus of above studies was describing the total number of all resistant cells, rather than the composition of the resistant population (26).

Results

We model the growth of a metastatic lesion as a branching process (27) that starts from a single cell (the founder cell of the metastasis) that is sensitive to treatment. Sensitive cells divide with rate b and die with rate d . The net growth rate of sensitive cells is $r = b - d$. During division, one of the daughter cells receives a resistance mutation with probability u . Resistant mutations can be neutral in the absence of treatment, which means they have the same birth and death rates as sensitive cells, and we initially focus on this case. We also expand our theory to the more general case where resistant cells are nonneutral, which means they have birth and death rates b_R and d_R , respectively. If

Significance

Metastatic dissemination to surgically inaccessible sites is the major cause of death in cancer patients. Targeted therapies, often initially effective against metastatic disease, invariably fail due to resistance. We use mathematical modeling to study heterogeneity of resistance to treatment and describe for the first time, to our knowledge, the entire ensemble of resistant subclones present in metastatic lesions. We show that radiographically detectable metastatic lesions harbor multiple resistant subclones of comparable size and compare our predictions to clinical data on resistance-associated mutations in colorectal cancer patients. Our model provides important information for the development of second-line treatments that aim to inhibit known resistance mutations.

Author contributions: I.B. designed research; I.B. and M.A.N. performed research; I.B. and M.A.N. analyzed data; and I.B. and M.A.N. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

¹To whom correspondence may be addressed. Email: ibozic@math.harvard.edu or martin_nowak@harvard.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1412075111/-DCSupplemental.

$c = (b_R - d_R)/(b - d) > 1$, then resistance mutations are advantageous before treatment; if $c < 1$, they are deleterious.

A resistant cell may appear in the population and be lost due to stochastic drift or it can establish a resistant subclone. We number the resistant subclones that survive stochastic drift by the order of appearance (Fig. 1A). A reasonable assumption for the number of point mutations that can provide resistance to a targeted drug is on the order of 100 (10, 28). Thus, the different resistant subclones will typically contain different resistance mutations, especially if we only focus on the largest ones.

We calculate the number and sizes of resistant subclones in a metastatic lesion containing M cells. Typical radiographically detectable lesions are ~ 1 cm in diameter and contain $\sim 10^9$ cells. The mutation rate, u , leading to resistance is the product of the point mutation rate μ , which is on the order of $\sim 10^{-9}$ per base pair per cell division, and the number of point mutations that can confer resistance, which is ~ 100 . In our analysis we will

assume a large M and small u limit and mostly focus on the case when $Mu \gg 1$.

Tumor sizes at which successful resistant mutants are produced can be viewed as a Poisson process on $[0, M]$ with rate u (SI Text) (10, 17). The number of successful mutant lineages is thus Poisson distributed with mean $\lambda = Mu$. If M_k is the number of cancer cells in the lesion when the k th mutant appeared, which survived stochastic drift (Fig. 1A), then $M_{k+1} - M_k$ is exponentially distributed with mean $1/u$. Therefore, we expect that the k th clone appeared when the total population size was $M_k \sim k/u$ and that roughly the size of the first clone is k times the size of the k th clone. The probability that exactly k clones are present in the population of size M is $\lambda^k e^{-\lambda}/k!$.

Counting new successful resistant clones in the order of appearance, we calculate the probability distribution for the number of cells in the k th resistant clone. In particular, if $k \ll Mu$, the cumulative distribution function for the number of resistant cells in the k th clone simplifies to

$$F_k(y) \approx 1 - \left(\frac{Mu}{Mu + y - dy/b} \right)^k. \quad [1]$$

The excellent agreement between Formula 1 and exact computer simulations of the stochastic process is shown in Fig. 1B.

The mean number of cells in the k th resistant clone is $E(Y_1) \approx [bMu/r][\log(r/bu) - 1]$ and $E(Y_k) \approx bMu/[r(k-1)]$ for $k \geq 2$. The median for the number of cells in the k th subclone is given by

$$\text{Med}(Y_k) \approx \frac{bMu}{r} (2^{1/k} - 1). \quad [2]$$

Interestingly, the ratio of the means of the two subclones k and j is $(j-1)/(k-1)$ for $k, j > 1$. The ratio of their medians is

$$\frac{\text{Med}(Y_k)}{\text{Med}(Y_j)} = \frac{2^{1/k} - 1}{2^{1/j} - 1}. \quad [3]$$

These ratios are independent of any parameters of the process. In particular, the ratio of the medians of the first and second clone is $\sqrt{2} - 1$, which implies that they have comparable size (same order of magnitude).

Liquid biopsy data were used to obtain estimates for the birth and death rates of cells in metastatic lesions and the number of point mutations providing resistance to the EGFR inhibitor panitumumab in colorectal cancer (10). The resulting parameter values ($b=0.25$ and $d=0.181$ per day, point mutation rate $\mu = 10^{-9}$ per base pair per replication, and 42 point mutations conferring resistance) can be used to calculate the mean and median sizes of the resistant subclones in a metastatic lesion containing $M = 10^9$ cells. The mean numbers of cells in the first, second, and third appearing resistant clone are $E(Y_1) \approx 2237$, $E(Y_2) \approx 152$, and $E(Y_3) \approx 76$, respectively. However, the mean for Y_1 , the size of the first resistant clone, is heavily influenced by the realizations of the stochastic process in which the first resistance mutation appeared early and is not a good summary of the probability distribution for Y_1 . Namely, the realizations in which the number of cells in the first clone is greater than the mean (2,237) account for less than 7% of all cases. The median number of cells in the first resistant clone [$\text{Med}(Y_1)$] for the above parameters is 152, whereas the medians for Y_2 and Y_3 are 63 and 40, respectively.

In SI Text, we calculate the probability distribution for the ratio of resistant clone sizes Y_1/Y_k and show that it is also independent of the parameters of the process. Even though the first appearing clone is expected to be the largest, followed by the second clone and so on, we show that this ordering is often

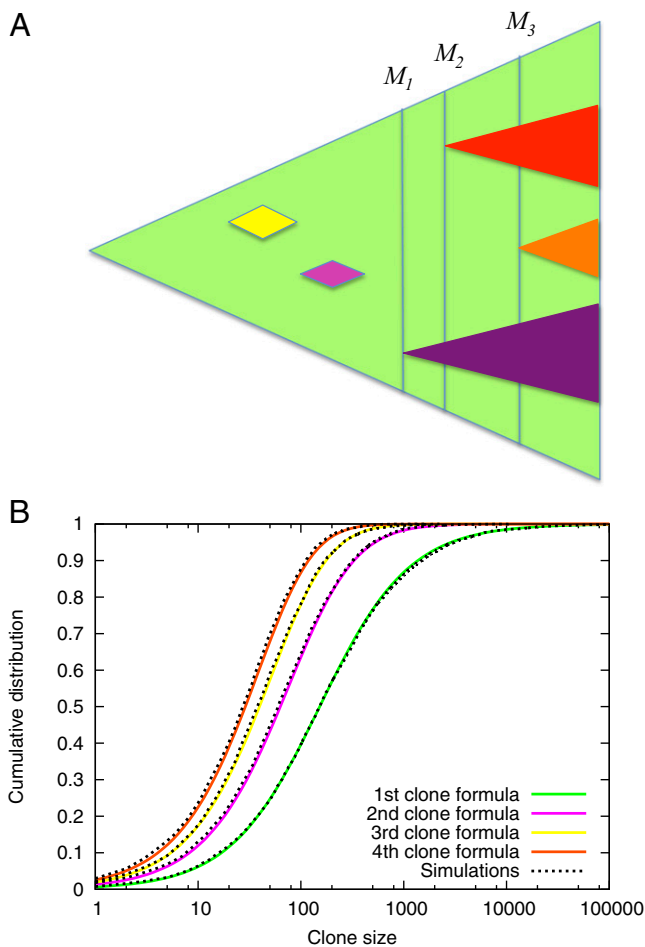


Fig. 1. Evolution of resistance in a metastatic lesion. (A) As the lesion (green) grows from one cell to detectable size, new resistant subclones appear. Some of them are lost to stochastic drift (yellow and pink), while others survive (purple, red and orange triangle). Instead of looking at the time of appearance of new clones, our approach takes into account the total size of the lesion when the resistance mutation first occurred. (B) Agreement between computer simulations and formula (1) for the cumulative distribution function for the number of cells in the first four resistant clones. The first subclone contains 10 or fewer cells with probability 0.06, between 10 and 100 cells with probability 0.34, between 100 and 1000 cells with probability 0.47 and more than 1000 cells with probability 0.13. The second subclone contains more than 100 cells with probability 0.36. Parameters $b=0.25$, $d=0.181$, $M=10^9$, $u=42 \cdot 10^{-9}$.

violated. In 31% of lesions, the first successful subclone is smaller than the second one; on the other hand, in 24% of lesions the first subclone is at least 10 times larger than the second one.

Fig. 2 shows different realizations of the stochastic process of evolution of resistance in metastatic lesions containing 10^8 and 10^9 cancer cells. The same parameters were used to generate all lesions. The size of each subclone is shown (in number of cells), and the subclones are ordered by their time of appearance. In lesion L1, the first three subclones are the largest, and each have around 100 cells. Lesion L5 contains only two subclones, whereas L6 contains seven subclones, but none has more than 10 cells. In each lesion of total size 10^9 cells, there are more than 10 resistant subclones. In L7, the two largest subclones contain 1,500 and 460 cells. In L8, there are five subclones of about 100 cells.

In Table 1, we show clinical data for the number of circulating tumor DNA (ctDNA) fragments harboring mutations in five genes associated with resistance to anti-EGFR treatment in 18 colorectal cancer patients who developed more than one mutation in those genes (29). These mutations were not detectable in patients' serum before therapy, but became detectable during the course of anti-EGFR treatment. The number of ctDNA fragments correlates with the number of tumor cells harboring that mutation: it was previously estimated (using the tumor burdens and pretreatment ctDNA levels measured in patients who had KRAS mutations in their tumors before therapy) that one mutant DNA fragment per milliliter of serum corresponds to 44 million mutant cells in the patient's tumor (10). Thus, the ratios of the resistant clone sizes can be obtained from the ratios of the numbers of ctDNA fragments harboring resistance-associated mutations. These data provide a unique opportunity to test our theory and compare the relative sizes of resistant clones inferred from the data with those predicted using our model. Assuming that resistance-associated mutations with higher ctDNA counts appeared before those with lower ctDNA counts, we find excellent agreement between the data and our model predictions. For example, the median ratio of the sizes of the first two resistant clones inferred from clinical data (29) is 2.21, whereas our model predicts 2.51. The median ratio of the sizes of the first and third clones from clinical data are 4.3, and our model predicts 4.12 (Table 1). This comparison is parameter free, as we showed

that the ratio of resistant clone sizes is independent of parameters.

Our mathematical results describe the relative sizes of resistant clones ordered by age, whereas the experimental data in Table 1 are ordered by size, which serves as a proxy for age, because exact clonal age is unknown. We quantify the extent to which this difference in clonal ordering by size vs. age influences our statistics using exact computer simulations (Table 1). In the relevant parameter regime of large lesion size, M , and small mutation rate, u , with $Mu \gg 1$, the results are largely independent of parameters (median ratios of clone sizes vary by <10% for different parameter combinations). We show simulation results for median ratios of clone sizes when clones are ordered by size for typical parameter values (10). As we see in Table 1, the ordering of experimental data by size does not significantly change the results of our analysis.

We can generalize our approach to the case when resistance mutations are not neutral, but provide a fitness effect already before treatment (formulas shown in *SI Text*). In Table 2, we compare the predicted medians for the first five resistant clones in a metastatic lesion containing $M = 10^9$ cells when resistance is deleterious, neutral, or advantageous. We see from Table 2 that even if resistant cells are only 10% as fit as sensitive cells, they will still be present in typical lesions. The average number of resistant cells produced until the lesion reaches size M is Mu/s . Here $s = 1 - d/b$ is the survival probability of sensitive cells, which is the probability that the lineage of a single sensitive cell will not die out. For typical parameter values (i.e., those used in Table 2), the number of resistant cells produced by sensitive cells in a single lesion is ~ 150 . Resistant cells that are 10% as fit as sensitive cells have a survival probability of $\sim 4\%$; so on average, six of them will form surviving clones. The effect that mutations can cause treatment failure, although they have high fitness cost is a consequence of the high number of resistant mutants produced by billion(s) of sensitive cells in a lesion and the specific properties of the branching process, namely the independence of lineages.

Discussion

In this paper we describe the heterogeneity of mutations providing resistance to cancer therapy that can be found in any one

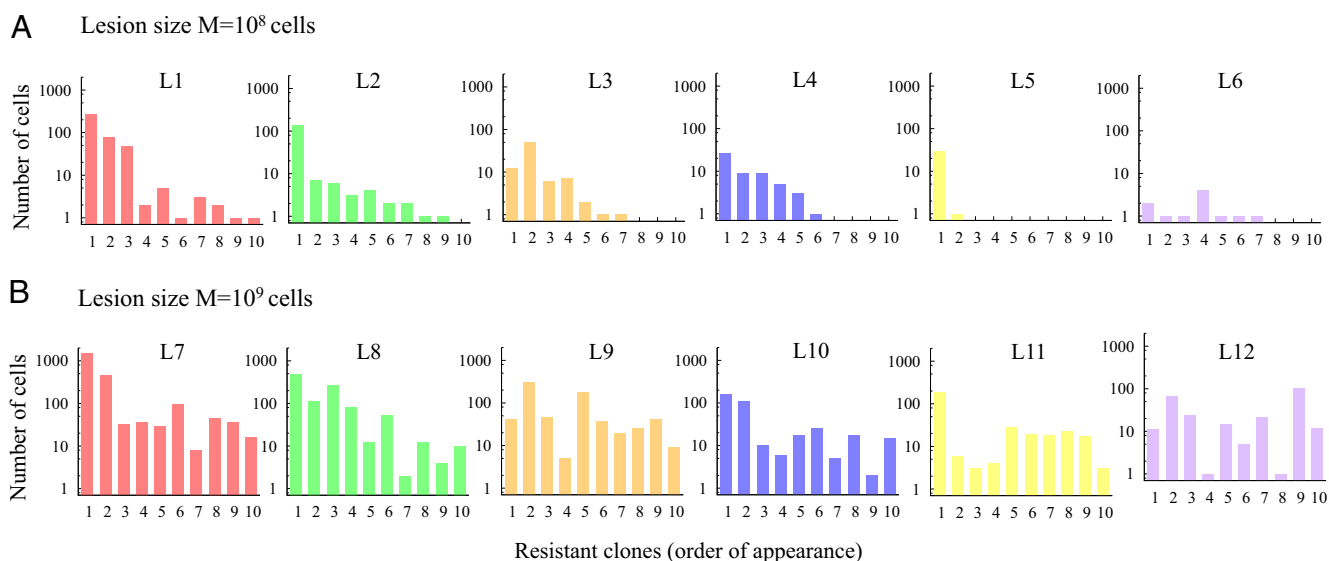


Fig. 2. Resistant subclones in metastatic lesions. Different realizations of the same stochastic process are shown in each panel. (A) Six lesions of size 10^8 and (B) six lesions of size 10^9 cells. The first ten resistant clones are shown, which survived until time of detection. They are ordered according to their time of appearance. Parameter values for all simulations: $b = 0.25$, $d = 0.181$, $u = 42 \cdot 10^{-9}$.

providing resistance to BCR-ABL (fusion of breakpoint cluster region gene and Abelson murine leukemia viral oncogene homolog 1) inhibitor imatinib were used to calculate the number of resistant clones and the expected number of resistant cells with a particular resistance mutation at the time of diagnosis of chronic myeloid leukemia. The authors found that at most one resistant clone is expected to be present, as the total number of CML stem cells at diagnosis is estimated to be approximately $M \sim 100,000$ cells and is much smaller than the billions of cells typically present in a single detectable lesion of a solid tumor. In this paper, we use a different mathematical technique and the novel approach of ordering the resistant clones according to their time of appearance, which allows us for the first time, to our knowledge, to describe the full spectrum of resistance mutations present in a lesion.

Our study is challenging the conventional view of the evolution of resistance in cancer. For every therapy that is opposed by multiple potential resistance mutations, which is the case for every targeted drug developed thus far, we can expect multiple resistant clones of comparable size in every lesion. Our theory provides a precise quantification of the relative sizes of those resistant subclones. The heterogeneity of resistance mutations is further amplified when taking into account multiple

metastatic lesions in a patient. This information is pertinent to the development of second line treatments that aim to inhibit known resistance mutations.

Materials and Methods

Model. We model the growth and evolution of a metastatic lesion as a continuous time multitype branching process (34). The growth of a lesion is initiated by a single cell sensitive to the drug. Sensitive cells produce a resistant cell at each division with probability u and each resistant cell produced by sensitive cells starts a new resistant type.

Analysis. In our analysis, we use the approximation that resistant cells produced by sensitive cells appear as a Poisson process on the number of sensitive cells (17). For more details and derivations of our results, please see *SI Text*.

Simulations. We perform Monte Carlo simulations of the multitype branching process using the Gillespie algorithm (35). Between 5,000 and 10,000 surviving runs are used for each parameter combination.

ACKNOWLEDGMENTS. We thank Bert Vogelstein for critical reading of the manuscript and Rick Durrett for discussion during the conception of this work. We are grateful for the support from Foundational Questions in Evolutionary Biology Grant RFP-12-17 and the John Templeton Foundation.

- Vogelstein B, et al. (2013) Cancer genome landscapes. *Science* 339(6127):1546–1558.
- Sawyers C (2004) Targeted cancer therapy. *Nature* 432(7015):294–297.
- Michor F, et al. (2005) Dynamics of chronic myeloid leukaemia. *Nature* 435(7046):1267–1270.
- Gerber DE, Minna JD (2010) ALK inhibition for non-small cell lung cancer: From discovery to therapy in record time. *Cancer Cell* 18(6):548–551.
- Komarova NL, Wodarz D (2013) *Targeted Cancer Treatment In Silico: Small Molecule Inhibitors and Oncolytic Viruses* (Springer, New York).
- Chapman PB, et al.; BRIM-3 Study Group (2011) Improved survival with vemurafenib in melanoma with BRAF V600E mutation. *N Engl J Med* 364(26):2507–2516.
- Maemondo M, et al.; North-East Japan Study Group (2010) Gefitinib or chemotherapy for non-small-cell lung cancer with mutated EGFR. *N Engl J Med* 362(25):2380–2388.
- Katayama R, et al. (2011) Therapeutic strategies to overcome crizotinib resistance in non-small cell lung cancers harboring the fusion oncogene EML4-ALK. *Proc Natl Acad Sci USA* 108(18):7535–7540.
- Sosman JA, et al. (2012) Survival in BRAF V600-mutant advanced melanoma treated with vemurafenib. *N Engl J Med* 366(8):707–714.
- Diaz LA, Jr, et al. (2012) The molecular evolution of acquired resistance to targeted EGFR blockade in colorectal cancers. *Nature* 486(7404):537–540.
- Pao W, et al. (2005) Acquired resistance of lung adenocarcinomas to gefitinib or erlotinib is associated with a second mutation in the EGFR kinase domain. *PLoS Med* 2(3):e73.
- Antonescu CR, et al. (2005) Acquired resistance to imatinib in gastrointestinal stromal tumor occurs through secondary gene mutation. *Clin Cancer Res* 11(11):4182–4190.
- O'Hare T, Eide CA, Deininger MV (2007) Bcr-Abl kinase domain mutations, drug resistance, and the road to a cure for chronic myeloid leukemia. *Blood* 110(7):2242–2249.
- Misale S, et al. (2012) Emergence of KRAS mutations and acquired resistance to anti-EGFR therapy in colorectal cancer. *Nature* 486(7404):532–536.
- Bozic I, et al. (2013) Evolutionary dynamics of cancer in response to targeted combination therapy. *eLife* 2:e00747.
- Coldman AJ, Goldie JH (1983) A model for the resistance of tumor cells to cancer chemotherapeutic agents. *Math Biosci* 65(2):291–307.
- Iwasa Y, Nowak MA, Michor F (2006) Evolution of resistance during clonal expansion. *Genetics* 172(4):2557–2566.
- Durrett R, Moseley S (2010) Evolution of resistance and progression to disease during clonal expansion of cancer. *Theor Popul Biol* 77(1):42–48.
- Komarova NL, Wodarz D (2005) Drug resistance in cancer: Principles of emergence and prevention. *Proc Natl Acad Sci USA* 102(27):9714–9719.
- Komarova N (2006) Stochastic modeling of drug resistance in cancer. *J Theor Biol* 239(3):351–366.
- Haeno H, Iwasa Y, Michor F (2007) The evolution of two mutations during clonal expansion. *Genetics* 177(4):2209–2221.
- Bozic I, Allen B, Nowak MA (2012) Dynamics of targeted cancer therapy. *Trends Mol Med* 18(6):311–316.
- Luria SE, Delbrück M (1943) Mutations of bacteria from virus sensitivity to virus resistance. *Genetics* 28(6):491–511.
- Kessler DA, Levine H (2013) Large population solution of the stochastic Luria-Delbrück evolution model. *Proc Natl Acad Sci USA* 110(29):11682–11687.
- Kessler DA, Levine H (2014) Scaling solution in the large population limit of the general asymmetric stochastic Luria-Delbrück evolution process. arXiv:1404.2407.
- Foo J, Michor F (2014) Evolution of acquired resistance to anti-cancer therapy. *J Theor Biol* 355:10–20.
- Bailey NTJ (1964) *The Elements of Stochastic Processes With Applications to the Natural Sciences* (Wiley, New York).
- Leder K, et al. (2011) Fitness conferred by BCR-ABL kinase domain mutations determines the risk of pre-existing resistance in chronic myeloid leukemia. *PLoS ONE* 6(11):e27682.
- Bettgowda C, et al. (2014) Detection of circulating tumor DNA in early- and late-stage human malignancies. *Sci Transl Med* 6(224):224ra24.
- O'Connell N (1993) Yule process approximation for the skeleton of a branching process. *J Appl Probab* 30(3):725–729.
- Manrubia SC, Zanette DH (2002) At the boundary between biological and cultural evolution: The origin of surname distributions. *J Theor Biol* 216(4):461–477.
- Maruvka YE, Shnerb NM, Kessler DA (2010) Universal features of surname distribution in a subsample of a growing population. *J Theor Biol* 262(2):245–256.
- Dewanji A, Luebeck EG, Moolgavkar SH (2005) A generalized Luria-Delbrück model. *Math Biosci* 197(2):140–152.
- Athreya KB, Ney PE (1972) *Branching Processes* (Springer-Verlag, Berlin).
- Gillespie DT (1977) Exact stochastic simulation of coupled chemical reactions. *J Phys Chem* 81(25):2340–2361.

Supporting Information

Bozic and Nowak 10.1073/pnas.1412075111

SI Text

The Model. We model the growth of a metastatic lesion as a branching process (1) that starts from a single cell sensitive to treatment. Sensitive cells divide with rate b and die with rate d . The net growth rate of sensitive cells is $r = b - d$. During division, one of the daughter cells receives a resistance mutation with probability u . Resistant cells have birth and death rates b_R and d_R . Resistant mutations can be neutral in the absence of treatment, which means they have the same birth and death rates as sensitive cells, and we initially focus on this case. Alternatively, if $c = (b_R - d_R)/(b - d) > 1$, then resistance mutations are advantageous before treatment; if $c < 1$, they are deleterious. We assume that mutation rate u is small, final lesion size M is large, $Mu \gg 1$, and we are mostly interested in the behavior of the early surviving clones. Furthermore, since mutation rate u is small, we assume that the size of the resistant population is much smaller than the size of the sensitive population, and approximate the size of the sensitive population with the size of the lesion.

Rate of Production of Mutants. We use the result (2) that the collection of tumor sizes at which resistance mutations are produced can be viewed (approximated) as a homogeneous Poisson process on $[1, M]$ with intensity $u/(1 - d/b)$. The reasoning follows from the fact that the average total number of resistance mutations produced when there are exactly x sensitive cells in the population is given by

$$R_x = \frac{bux}{1 - d/b} \int_0^\infty f_x(t) dt, \quad [\text{S1}]$$

where $f_x(t)$ is the probability that there are exactly x sensitive cells after time t and the factor $1 - d/b$ comes from only looking at lineages in which the tumor population did not go extinct. In the small mutation rate u limit, one can neglect the production of mutants to calculate $f_x(t)$ as pertaining to a single type branching process on the sensitive cells. Even though it seems that Iwasa et al. (2) were not aware of it, $f_x(t)$ was derived by Bailey (1)

$$f_x(t) = (1 - \alpha)(1 - \beta)^{\beta^{x-1}}, \quad [\text{S2}]$$

with $\alpha = (de^{rt} - d)/(be^{rt} - d)$ and $\beta = (be^{rt} - b)/(be^{rt} - d)$. Plugging in the expression for f_x into Eq. 1 leads to the rate at which mutants are produced when there are x sensitive cells

$$R_x = u/(1 - d/b). \quad [\text{S3}]$$

A more intuitive way to prove this result is as follows: when the population contains exactly x sensitive cells, the probability that they will produce a mutant before going to $x - 1$ or $x + 1$ sensitive cells is $bu/(b + d)$. The (average) number of occurrences of exactly x cells in the process is one plus the (average) number of returns of a biased random walk with $p = b/(b + d)$. Multiplying the probability of producing a mutant cell while at state x with the number of occurrences of that state leads to $R_x = u/(1 - d/b)$.

Each mutant cell survives stochastic drift with probability $1 - d/b$, so the tumor sizes at which mutations that survive stochastic drift are produced can be viewed as a Poisson process on $[1, M]$ with intensity u . Because M is large and u is small, we can replace the interval $[1, M]$ by $[0, M]$ without losing much accuracy (3).

Size of the k th Resistant Clone. Let M_k be the size of the sensitive population when the k th successful resistant mutant appears. Furthermore, let Y_k denote the size of the k th resistant sub-

population that survives stochastic drift when there are M -sensitive cells, conditioned on $M_k \leq M$. By the time the sensitive population reached size M , Y_k can be approximated by MV/M_k , where V is an exponentially distributed random variable with mean $b/(b - d)$ (4). If $F_k(y) = \Pr[Y_k \leq y]$ is the cumulative distribution of Y_k , then, expanding on the reasoning in ref. 3, we have

$$\begin{aligned} F_k(y) &\approx 1 - \Pr[MV/M_k \geq y | M_k \leq M] \\ &= 1 - \int_0^M \text{Prob.Density}[M_k = z | M_k \leq M] \times \Pr\left[V \geq \frac{yz}{M}\right] dz \\ &= 1 - \int_0^M \frac{(zu)^{k-1} e^{-zu} u}{(k-1)!} \left(1 - \sum_{l=0}^{k-1} \frac{(Mu)^l e^{-Mu}}{l!}\right)^{-1} \exp\left(-\frac{ryz}{Mb}\right) dz. \end{aligned}$$

Evaluating the integral above leads to

$$F_k(y) \approx 1 - \left(\frac{Mu}{Mu + y - dy/b}\right)^k \frac{\Gamma(k) - \Gamma(k, Mu + y - dy/b)}{\Gamma(k) - \Gamma(k, Mu)}, \quad [\text{S4}]$$

where $\Gamma(k) = (k - 1)!$ and $\Gamma(a, z) = \int_z^\infty t^{a-1} e^{-t} dt$ is the incomplete Gamma function.

The probability density function for Y_k , $f_k = F'_k$, is given by

$$f_k(y) \approx rk(bMu)^k (bMu + ry)^{-1-k} \frac{\Gamma(k) - \Gamma(k, Mu + y - dy/b)}{\Gamma(k) - \Gamma(k, Mu)}. \quad [\text{S5}]$$

In particular, for $k \ll Mu$, we have

$$F_k(y) \approx 1 - \left(\frac{Mu}{Mu + y - dy/b}\right)^k, \quad [\text{S6}]$$

and

$$f_k(y) \approx k(1 - d/b)(Mu)^k (Mu + y - dy/b)^{-1-k}. \quad [\text{S7}]$$

Comparison of Formula S6 and the cumulative distribution function for Y_k obtained from 5,000 runs of the exact computer simulation of the branching process is shown in Fig. 1B.

Calculating the expected number of cells in the k th clone using Formula S7 (integrating from 0 to M) and expanding it in the small u limit, we obtain

$$E(Y_k) \approx \frac{bMu}{r(k-1)} + O(u^2), \quad [\text{S8}]$$

for $k \geq 2$ and

$$E(Y_1) \approx \frac{bMu}{r} \left(\log \frac{r}{bu} - 1\right) + O(u^2). \quad [\text{S9}]$$

We can also obtain the median for the number of cells in the k th clone, $Y_k^{1/2}$, from the cumulative distribution function (S6)

$$Y_k^{1/2} = \frac{bMu}{r} (2^{1/k} - 1). \quad [\text{S10}]$$

Ratio of Resistant Clone Sizes. To more precisely determine the relationship between the sizes of different subclones, we next

calculate the probability distribution of Y_1/Y_k : the ratio of sizes of first and k th clone. We again use the fact that the random variable describing the size of the k th clone, Y_k , can be approximated by $V_k M/M_k$, where $V_k \sim \text{Exp}[b/(b-d)]$, and that M_k , size of the population on the arrival of the k th clone, is the sum of k exponential random variables with mean $1/u$. We have

$$\Pr[Y_1/Y_k \leq x] = \Pr[V_1 M/M_1 \leq x V_k M/M_k] = \Pr[V_1/V_k \leq x M_1/M_k].$$

We note that $Z = V_1/V_k$ is the ratio of two independent, identically distributed exponential random variables and thus its probability density function is $f_Z(z) = 1/(z+1)^2$. Similarly, $W = M_1/M_k \sim \Gamma[1, \lambda]/(\Gamma[1, \lambda] + \Gamma[k-1, \lambda]) \sim \beta[1, k-1]$ is a β -distributed random variable with probability density function $f_W(w) = (k-1)(1-w)^{k-2}$. It follows that

$$\begin{aligned} \Pr[Y_1/Y_k \leq x] &= \Pr[Z \leq xW] \\ &= \int_0^1 (k-1)(1-w)^{k-2} \int_0^{wx} \frac{1}{(z+1)^2} dz dw \\ &= \frac{x}{1+x} \left(1 - \frac{k-1}{k} {}_2F_1[1, 1, 1+k, -x] \right), \end{aligned} \quad [\text{S11}]$$

where ${}_2F_1$ is the hypergeometric function. Notably, this distribution depends only on k and not on any parameters of the process.

For example, the probability that the ratio of sizes of the first and the second clone is smaller than x is

$$\Pr[Y_1/Y_2 \leq x] = 1 - \frac{\log(1+x)}{x}. \quad [\text{S12}]$$

In particular, the first successful subclone that appears is smaller than the second appearing subclone in 31% of cases. The probability that the first appearing subclone is twice as large or larger than the second appearing clone is 55%, and the probability that it is 10 or more times larger is 24%.

Nonneutral Resistance. In this section, we will obtain similar results for the probability distributions of clone sizes and their ratios in the case in which resistant cells have birth and death rates b_R and d_R , respectively. Tumor sizes at which resistance mutations appear can still be viewed as a homogeneous Poisson process on $[0, M]$ with intensity $u/(1-d/b)$. However, the lineages of newly produced resistance mutations will escape extinction with probability $1-d_R/b_R$, so the successful resistant subclones in this scenario will arrive with rate $u_R = u(1-d_R/b_R)/(1-d/b)$. Another change from the neutral case is that the size of the k th clone when the total population size is M can be approximated with $Y_k \sim (M/M_k)^c U$, where M_k is the population size when the k th successful resistance mutation appeared and U is an exponentially distributed random variable with mean $b_R/(b_R-d_R)$. We recall that $c = (b_R-d_R)/(b-d)$.

With the above caveats, we can write the derivation of the cumulative distribution function for the size of the k th appearing resistant subclone, Y_k , similarly as before

$$\begin{aligned} F_k(y) &\approx 1 - \Pr[(M/M_k)^c U \geq y | M_k \leq M] \\ &= 1 - \int_0^M \text{Prob.Density}[M_k = z | M_k \leq M] \times \Pr\left[U \geq \frac{yz^c}{M^c}\right] dz \\ &= 1 - \left(1 - \sum_{l=0}^{k-1} \frac{(Mu_R)^l e^{-Mu_R}}{l!} \right)^{-1} \int_0^M \frac{u_R (zu_R)^{k-1} e^{-zu_R}}{(k-1)!} \\ &\quad \times \exp\left(-\frac{r_R y z^c}{b_R M^c}\right) dz. \end{aligned}$$

[S13]

The difference is that in the nonneutral case the integral has to be evaluated numerically.

For the ratio of clone sizes Y_k/Y_1 in the nonneutral case, we have

$$\begin{aligned} \Pr[Y_1/Y_k \leq x] &= \Pr[U_1 (M/M_1)^c \leq x U_k (M/M_k)^c] \\ &= \Pr[U_1/U_k \leq x (M_1/M_k)^c] \\ &= \int_0^1 (k-1)(1-w^{1/c})^{k-2} \frac{1}{c} w^{1/c-1} \int_0^{wx} \frac{1}{(z+1)^2} dz dw \\ &= \int_0^1 \frac{k-1}{c} (1-w^{1/c})^{k-2} \frac{xw^{1/c}}{1+wx} dw. \end{aligned}$$

[S14]

We see that even when resistance is not neutral, the ratio of clone sizes depends only on the order of appearance and the relative fitness c and not on M , u , and the specific birth and death rates of cells.

Our formulas rely on the approximation $Y \sim \exp[(b_R-d_R)t]U$ for the size of a resistant clone Y , where U is an exponentially distributed random variable with mean $b_R/(b_R-d_R)$ and time t is measured from the appearance of the founder cell of the clone. This approximation assumes large t and loses accuracy for t close to 0 [e.g., it predicts that the average size of a clone at time 0 is $b_R/(b_R-d_R)$ rather than 1]. Thus, our formulas lose accuracy when successful resistant clones are produced shortly before reaching size M . Successful resistant clones appear as a Poisson process on the number of sensitive cells with rate $u_R = u(1-d_R/b_R)/(1-d/b)$ when resistance is not neutral, and the first such clone will appear when there are $\sim 1/u_R$ sensitive cells. For our approximation to hold, M must be significantly larger than $1/u_R$, which is equivalent to $Muc(b/b_R) \gg 1$. In other words, we assume that the relative fitness c is such that the expected number of successful resistant clones, Mu_R , is much larger than 1.

1. Bailey NTJ (1964) *The Elements of Stochastic Processes with Applications to the Natural Sciences* (Wiley, New York).
2. Iwasa Y, Nowak MA, Michor F (2006) Evolution of resistance during clonal expansion. *Genetics* 172(4):2557–2566.

3. Diaz LA, Jr, et al. (2012) The molecular evolution of acquired resistance to targeted EGFR blockade in colorectal cancers. *Nature* 486(7404):537–540.
4. Durrett R, Moseley S (2010) Evolution of resistance and progression to disease during clonal expansion of cancer. *Theor Popul Biol* 77(1):42–48.