DOI: 10.1534/genetics.111.127480

Germline Fitness-Based Scoring of Cancer Mutations

Andrej Fischer,*,† Chris Greenman* and Ville Mustonen*,[1]

*Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SA, United Kingdom and †Institut für Theoretische Physik, Universität zu Köln, 50937 Köln, Germany

Manuscript received November 8, 2010
Accepted for publication February 28, 2011

ABSTRACT

A key goal in cancer research is to find the genomic alterations that underlie malignant cells. Genomics has proved successful in identifying somatic variants at a large scale. However, it has become evident that a typical cancer exhibits a heterogenous mutation pattern across samples. Cases where the same alteration is observed repeatedly seem to be the exception rather than the norm. Thus, pinpointing the key alterations (driver mutations) from a background of variations with no direct causal link to cancer (passenger mutations) is difficult. Here we analyze somatic missense mutations from cancer samples and their healthy tissue counterparts (germline mutations) from the viewpoint of germline fitness. We calibrate a scoring system from protein domain alignments to score mutations and their target loci. We show first that this score predicts to a good degree the rate of polymorphism of the observed germline variation. The scoring is then applied to somatic mutations. We show that candidate cancer genes prone to copy number loss harbor mutations with germline fitness effects that are significantly more deleterious than expected by chance. This suggests that missense mutations play a driving role in tumor suppressor genes. Furthermore, these mutations fall preferably onto loci in sequence neighborhoods that are high scoring in terms of germline fitness. In contrast, for somatic mutations in candidate onco genes we do not observe a statistically significant effect. These results help to inform how to exploit germline fitness predictions in discovering new genes and mutations responsible for cancer.

CANCER is a genetic disease whose progression has for a long time been discussed in terms of Darwinian evolution where malignant cells have a fitness advantage over normal cells (see, *e.g.*, Merlo *et al.* 2006). This evolution is a complex stochastic process where the major evolutionary forces, mutation, genetic drift, and selection all contribute to the observed evolutionary changes, making their individual roles difficult to disentangle. However, it is precisely this decomposition that will be critical when we attempt to understand functional consequences of somatic mutations. Only very recently have these theoretical considerations found their way to data analyses at nucleotide resolution on a large scale (Yang *et al.* 2003; Greenman *et al.* 2006, 2007; Sjöblom *et al.* 2006). This transition is driven by the increased technological ability to sequence cancer and healthy tissue samples from patients. Genomics has proved powerful in finding somatic variants in cancer, yet the emerging picture is complex with a typical cancer showing a heterogenous mutation pattern across samples (Stratton *et al.* 2009). Specifically, in addition to the standard model where a cancer gene is frequently mutated at a specific location in a gene such as the V600E mutation in BRAF (Davies *et al.* 2002), there is increasing evidence that there are many driving mutations in genes occurring at a very low prevalence (Greenman *et al.* 2007; Carter *et al.* 2009). This complicates the statistical challenges of distinguishing causal driver mutations from a large number of passenger mutations that are not directly contributing to the cancer phenotype of the cells.

Methods to find driver mutations are usually classified into two main categories: mutation frequency-based analysis and bioinformatic predictions of functional effects of amino acid changes. These methods have recently been reviewed in great detail (Lee *et al.* 2009a; Torkamani *et al.* 2009). In short, frequency-based methods try to exploit the fact that mutations under positive Darwinian selection fix in the population with a higher rate than neutral or deleterious mutations. In its simplest form this involves comparing the rates of substitutions in a postulated neutral class, *e.g.*, synonymous mutations, to somatic missense mutations. The most obvious subtlety in these approaches lies in identifying a truly sound neutral model (Rubin and Green 2009). These approaches generally assume a fixed background mutation rate across the genome, an assumption that is known to be inaccurate for homozygous deletions (Bignell *et al.*

Supporting information is available online at http://www.genetics.org/cgi/content/full/genetics.111.127480/DC1.

Available freely online through the author-supported open access option.

[1]*Corresponding author:* Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SA, United Kingdom. E-mail: vm5@sanger.ac.uk

2010). Fortunately, many of the complications in the inference of positive selection have been considered extensively within the context of organismal evolution. However, the powerful utilization of both substitution and polymorphism data in conjunction still awaits implementation in the case of cancer due to the lack of somatic polymorphism data, *i.e.*, frequencies of variants in a cancer cell population. Such data could be obtained from sequencing populations of single cells from tumors. In the future, frequency-based methods in cancer mutation analysis have the potential to draw from the vast experience accumulated in the field of evolutionary genetics (Fay and Wu 2000; Kreitman 2000; Nielsen 2005; Eyre-Walker 2006).

In the second main class, bioinformatic methods use a combination of measures based on conservation, biophysical, and structural considerations to distinguish causal functional variants from passenger mutations. These methods are now commonly used to assess the potential functional consequences of discovered somatic variants in genome-wide scans. Depending on the specifics of training data and the availability of functional information, they can perform strongly (Kaminker *et al.* 2007; Torkamani and Schork 2008; Carter *et al.* 2009; Lee *et al.* 2009a). Many of the bioinformatic approaches to predict the functional effects of missense mutations were initially developed for germline variation and have been applied to better understand common and rare disease variants as well as evolution (Ng and Henikoff 2001; Ramensky *et al.* 2002; Bromberg and Rost 2007; Kryukov *et al.* 2007). With a focus on germline variation, these methods have recently been reviewed in Jordan *et al.* (2010). These approaches have seen some adaptation to somatic variation (Radivojac *et al.* 2008; Carter *et al.* 2009; Kan *et al.* 2010).

In this article, we approach the scoring of cancer missense mutations from the perspective of their germline fitness effects (germline fitness quantifies reproductive success of the organism and natural selection acts via fitness differences). Throughout the study, we carefully disentangle the predicted effects of the mutations from their bare occurrence rates, *e.g.*, by comparing to null models conditioned to the same number of missense mutations. In this sense, our work is not frequency based. However, we do rely on evolutionary theory—applied to germline variation—to develop our scoring system. On the other hand, the approach does not fall directly into the category of bioinformatic-based methods in that we do not seek to use all possible functional information available to find candidate driver mutations. Instead, we want to know what role, if any, germline fitness plays in cancer.

We focus on a cancer mutation data set of human kinases from Greenman *et al.* (2007). With this study we have full control over what was sequenced, which variants were seen, and which were not, in both germline and cancer samples. These are critical aspects for our purpose. More generally, kinases are an ideal test bed to develop methods for understanding putative disease mutations because abnormally functioning kinases are a major cause of human disease, an extensive body of functional information on kinases exists, and they are a large family and thus well suited for statistical analysis (Manning *et al.* 2002; Lahiry *et al.* 2010).

We start by calibrating a scoring system using Pfam (Finn *et al.* 2010) domain alignments, which allows us to assign a score to both missense mutations and target loci that they fall onto. We then discuss under what conditions can such a multiple alignment-based score be related to the germline fitness of the variant, using considerations from diffusion theory; see, *e.g.*, Kimura (1964). We show first that for the germline variants in our data set the score predicts the overall polymorphism rate and is thus consistent with it being an estimate of germline fitness. The scoring is then applied to germline and somatic mutations at the level of loci and genes. It will become clear that comparing germline and somatic variants directly is not appropriate for addressing our question: rather both should be individually contrasted to a null model where mutations are random with respect to the scores. Such a choice of a null model has also been shown to be important for calibrating classifiers for finding driver mutations (Carter *et al.* 2009).

SCORING SYSTEM FOR CANCER VARIATION

Scoring of mutations: The scoring system that we use for mutations should reflect the impact—in a beneficial or a deleterious sense—of a genomic alteration. Each gene consists of functional subunits—the protein domains—that represent the most conserved part of the gene. Well-organized information about these domains is available in the Pfam database (Finn *et al.* 2010). Here we use the Pfam-A seed alignment of each domain as a basis for its scoring system (Moses and Durbin 2009). The composition of each column in these alignments is the result of evolution and it will usually differ markedly from a neutral distribution. We denote the counts of amino acids in the alignment column i by $n_i(a)$, where $a \in \mathcal{A}$ (all amino acids), and compare this observed distribution to a prior (expected) null distribution $p_0(a)$, derived for instance from overall genomic frequencies of amino acids. Taking the log odds ratio of the functional distribution and the null gives a so-called position-specific score [or position weight matrix (PWM) score (Durbin *et al.* 1998)],

$$s_i(a) = \log\frac{\tilde{q}_i(a)}{p_0(a)} = \log\frac{n_i(a) + p_0(a)}{(N_i + 1)p_0(a)}, \qquad (1)$$

where N_i is the total number of residues in the column. The above construct of the observed q-distribution is

regularized using pseudocounts proportional to p_0 to account for nonobserved residues in the finite sample (Lawrence *et al.* 1993; Henikoff and Henikoff 1996). The two extreme cases are columns that are highly conserved—where the most prevalent letter receives a large positive score and all others large negative ones—and columns that are highly variable and close to neutral—where all letters receive scores close to zero. For a given mutation away from the reference, we can now record the score difference between the final and the initial residue:

$$\begin{aligned}\Delta s_i &= s_i(a) - s_i(a_{\text{ref}}) \\ &= \log \frac{\tilde{q}_i(a)}{\tilde{q}_i(a_{\text{ref}})} - \log \frac{p_0(a)}{p_0(a_{\text{ref}})}.\end{aligned} \tag{2}$$

It can be shown that assuming a Dirichlet prior for the frequency vector q with parameters p_0 before observing the counts n, the probability of the score Δs has a maximum *a posteriori* (MAP) value of Equation 2. For mutations where the final amino acid was not observed ($n(a) = 0$), the posterior distribution is strongly skewed and neither the MAP nor, *e.g.*, the posterior mean value is representative. We proceed using the MAP value as a conservative estimate for Δs. Using the mean instead does not change our results significantly. We show next using population genetic theory that this score difference is closely related to the germline fitness difference caused by the mutation.

Linking scores Δs to germline fitness: Consider a population of N individuals evolving under genetic drift, mutation, and selection. Every individual has either allele a or b, with mutation rates $\mu_{a\to b}$ and $\mu_{b\to a}$, and fitness values f_a and f_b. Let us denote the fraction of alleles a in the population by x. For eukaryotic evolution, the mutation rates are usually very small so we can consider this generic model in the limit $N\mu << 1$. It is well known (for a review see Rouzine *et al.* 2001) that in this case the population is mostly monomorphic with infrequent periods of polymorphism and substitution events between the two alleles. The model allows for a description of the time evolution of the probability density of x, $P(x, t)$ (including boundaries) with a diffusion equation of the form

$$\begin{aligned}&\partial_t P(x,t) \\ &= \partial_x \left[\frac{1}{2N}\partial_x x(1-x) - \sigma_0 x(1-x) + \mu_{a\to b}x - \mu_{b\to a}(1-x)\right] \\ &\quad \times P(x,t),\end{aligned} \tag{3}$$

where $\sigma_0 = f_a - f_b$ is the selective advantage (disadvantage if $\sigma_0 < 0$) of allele a with respect to allele b. We can then solve for the equilibrium density of the process,

$$P(x) = Z^{-1}(1-x)^{-1+2N\mu_{a\to b}} x^{-1+2N\mu_{b\to a}} e^{2N\sigma_0 x}, \tag{4}$$

where Z normalizes the distribution. In what follows, we denote the scaled selection coefficient as $2N\sigma_0 = \sigma$. It is clear from Equation 4 that for systems with $N\mu << 1$ the density assumes a "U-shape" with most of the probability concentrated at the boundaries. The rates of substitution from monomorphic (all individuals carry either allele b or a) populations can be evaluated from Equation 3 by solving the corresponding backward equation (see, *e.g.*, Kimura 1964) for appropriate boundary conditions, yielding

$$u_{b\to a}(\sigma) = \frac{\mu_{b\to a}\sigma}{1 - e^{-\sigma}},\ u_{a\to b}(\sigma) = \frac{-\mu_{a\to b}\sigma}{1 - e^{+\sigma}}. \tag{5}$$

The substitution rate $u(\sigma)$ depends on the fitness difference σ so that deleterious mutations are suppressed while beneficial mutations fix with enhanced rates. We can put the polymorphic states aside for a moment and evaluate the fixed-state probabilities by solving the two-state substitution dynamics with rates from Equation 5 [neglecting terms $O(N\mu)$]:

$$q(a) = \frac{\mu_{b\to a}}{\mu_{b\to a} + \mu_{a\to b}e^{-\sigma}}, \quad q(b) = \frac{\mu_{a\to b}e^{-\sigma}}{\mu_{b\to a} + \mu_{a\to b}e^{-\sigma}}. \tag{6}$$

It then follows that the ratio of the probabilities at the two fixed states is

$$\frac{q(a)}{q(b)} = \frac{\mu_{b\to a}}{\mu_{a\to b}}e^{\sigma}. \tag{7}$$

In fact, we can understand each alignment column in our scoring systems as a finite sample from such a distribution,

$$s_i(a) = \log \frac{\tilde{q}_i(a)}{p_0(a)} \tag{8}$$

with

$$\tilde{q}_i(a) = \frac{n_i(a) + p_0(a)}{N_i + 1} \xrightarrow{N_i \gg 1} q_i(a), \tag{9}$$

where $p_0(a)$ is the frequency of letter a in some background sequence that we are comparing the q sequence to. If we take $p_0(a)$ to correspond to a neutral evolutionary model with $\sigma = 0$, $\mu_{a\to b}$, $\mu_{b\to a}$, we note that

$$\begin{aligned}\Delta s = s(a) - s(b) &= \log \frac{q(a)}{p_0(a)} - \log \frac{q(b)}{p_0(b)} \\ &= \log \frac{q(a)}{q(b)} - \log \frac{p_0(a)}{p_0(b)} \\ &= \sigma + \log \frac{\mu_{b\to a}}{\mu_{a\to b}} - \log \frac{\mu_{b\to a}}{\mu_{a\to b}} \\ &= \sigma.\end{aligned} \tag{10}$$

In other words, the score difference Δs equals the scaled fitness difference σ. The picture above is also easily generalized to cover loci with >2 alleles (4 for nucleotides and 20 for amino acids), as long as detailed balance (see, *e.g.*, Gardiner 2009) holds for the neutral process and selection is given by a static fitness landscape. Such approaches that equate observed frequency differences between functional and neutral classes of sequences to evolutionary fitness have been exploited in many systems: in the contexts of codon usage bias (Bulmer 1991), amino acid evolution (Halpern and Bruno 1998), binding site evolution (Berg *et al.* 2004; Moses *et al.* 2004; Mustonen and Lässig 2005; Doniger and Fay 2007), and human germline polymorphism analysis (Moses and Durbin 2009).

In its simplest form (as was presented above), the theoretical basis for these considerations can be traced back to Kimura's solution of the one-locus, two-alleles model (Kimura 1955). The picture has also been extended to traits under selection defined over larger functional units than single nucleotides or amino acids (Berg *et al.* 2004).

Assumptions underlying the link between scores Δs and germline fitness: When we link scoring of an alignment column to the evolutionary model discussed above, we should keep in mind the following assumptions:

- Alleles on loci forming the alignment columns are understood to be the result of *independent* draws from the *same* underlying distribution. For evolutionary dynamics this means that the σ's are fixed for every column individually, the neutral mutation process defined by the μ's is shared between the columns, and the sequences have diverged beyond the relevant correlation times and can thus be considered independent.
- For sequences of length L, the scoring also assumes that the sequence probability is factorizable, *i.e.*, $Q = \prod_{i=1}^{L} q_i$, and thus there are no correlations across the loci caused by, for example, genomic linkage or epistatic fitness interactions. The probabilities q_i can of course vary as a function of the sequence position. In evolutionary theory these assumptions can be expressed as infinite recombination and additive fitness contributions across loci.
- The above derivation for the two-allele model is applicable to protein alignments if detailed balance holds for the neutral evolutionary process at each locus (column). This is a necessary condition to equate the ratio of any two amino acid frequencies to their substitution rates.

It is clear that these assumptions are never fully met in the systems that we study. Fortunately, we have a way of testing the sensibility of the scoring scheme with germline polymorphism data—a piece of information not used in the derivation. In particular, we can ask: What is the probability to find a polymorphism of effect σ given that we have sampled m individuals? The diffusion equation can be analyzed for the so-called forward spectrum of the polymorphism frequency (Sawyer and Hartl 1992). Up to first order in $N\mu$ we get the polymorphism density to be

$$P_p(\sigma, m) = 2N\mu \sum_{k=1}^{m-1} \frac{m}{k(m-k)} \frac{\left(e^{\sigma} - F_1(k, m, \sigma)\right)}{2(e^{\sigma} - 1)}, \tag{11}$$

where F_1 is the hypergeometric function and terms in the sum with $k = 0$, m would correspond to monomorphic samples. Importantly, this function includes the mutation rate only as a prefactor. Later we use Equation 11 to predict the germline polymorphism rate of occurrence as a function of our score Δs.

Scoring of target loci: On top of the score Δs that we assign to every mutation with Equation 2, we incorporate the local neighborhoods of the mutation target sites into the scoring. This is done by evaluating the mean germline fitness per locus of the subsequence consisting of $l_w = 2w + 1$ amino acids centered around the mutation site i (a set denoted by w_i):

$$S_i^w = \frac{1}{l_w} \sum_{a_j \in w_i} s_j(a_j). \tag{12}$$

In contrast to the mutation score, the locus score does not depend on the particular mutation but only on its location. It derives its information from several loci and gives a scale for how evolutionarily important the target locus and its surroundings are. We can then weight every mutation score by its neighborhood score in the process of scoring: $f(S^w)\Delta s$. Both scores are illustrated in Figure 1.

Defining genomic observables: We use the scores Δs and S^w to analyze the cancer, *i.e.*, somatic and germline variation at the level of individual loci and genes. All observables and distributions at locus (gene) level have a superindex l (g). The kinome sequence consisting of N^l loci belonging to an individual k is denoted by $\mathbf{a}_k = \{a_{1k}, \ldots, a_{ik}, \ldots, a_{N^l k}\}$ and the score of a mutation with respect to the reference genome ($a_{\mathrm{ref}} \to a$) at a genomic locus i by $\Delta s_i(a) \equiv s_i(a) - s_i(a_{\mathrm{ref}})$. We can thus define the effect per locus i,

$$\Delta s_i^l = \sum_{k=1}^{m} \Delta s_i(a_{ik}), \tag{13}$$

where m is the number of sample genomes and a_{ik} denotes the amino acid at locus i in individual k. The projection (summing) over all samples at this early stage of the analysis is necessary due to the relative scarcity of the mutation data at hand.

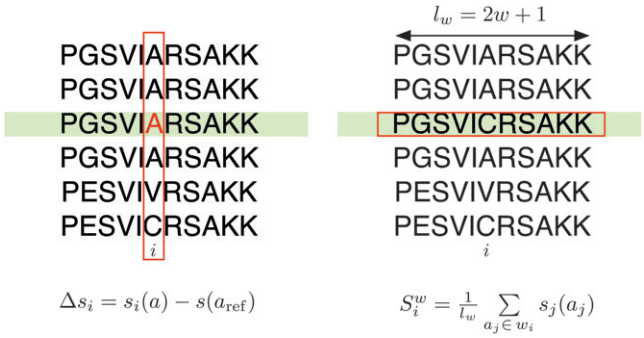

FIGURE 1.—Mutation and locus scores. An example alignment window is shown that illustrates the scoring system described in the text. We want to score a mutation $a_{\mathrm{ref}} = \mathrm{C} \rightarrow \mathrm{A} = a$ (colored with red) in position i. First, we evaluate the difference in the position-specific score between the final and the initial states, as defined by the alignment column (left panel, vertical red box). Second, we evaluate a score for the target locus onto which the mutation falls by summing up the scores of the amino acids within a window w_i (right panel, horizontal red box). Locus score information is derived from several loci and gives a scale for how evolutionarily important the target locus and its surroundings are.

Similarly we define the effect per locus in gene j by summing over all mutations in that gene and scaling by its opportunity size l_j^g (approximately proportional to the total length of all domains in it):

$$\Delta s_j^g = \frac{1}{l_j^g} \sum_{i \in \mathrm{gene}_j} \Delta s_i^l. \tag{14}$$

Analogously, we also define locus scores S^w and weighted scores $\exp(S^w)\Delta s$ (the rationale behind the nonlinear weighting function is discussed later). Finally, to expose the effect of mutation counts alone we also define a count score per locus,

$$c_i^l = \sum_{k=1}^{m} c(a_{ik}), \tag{15}$$

with $c = 1$ if $a_{ik} \neq a_{i,\mathrm{ref}}$ and 0 otherwise. This count score can also be applied at the gene level as defined above.

Defining a null ensemble: Before we can analyze the sets of germline and somatic mutations we need to define a null model. Our initial assumption is that all mutations are random with respect to our scoring system. To test this hypothesis, we construct *in silico* all possible missense point mutations away from the reference sequence of the kinases $\mathbf{a}_{\mathrm{ref}}$. This set is called *mutational opportunity space* and denoted by $\mathcal{M}$. Using a null based on all genes would not be appropriate as the frequencies of domains are not homogenous over the genes. This heterogeneity is a smaller problem for data sets with a large number of genes; however, a data set-specific null becomes increasingly more important for smaller gene sets.

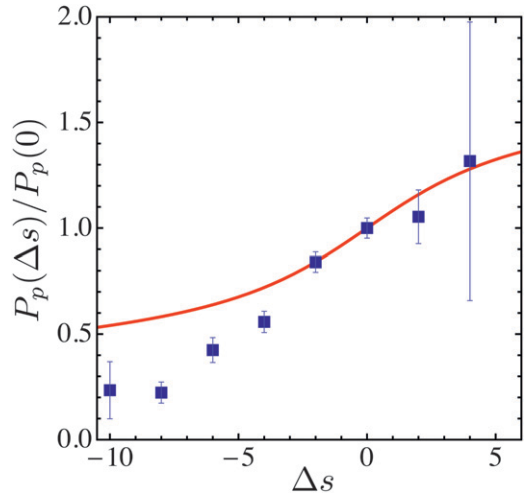

FIGURE 2.—Germline polymorphism density. $P_p(\Delta s)$ is shown in units of $P_p(0)$: blue squares are data and the red line is the theory curve from Equation 11 with $m = 210$. Predicted polymorphism density is proportional to the values measured from the germline variation, somewhat underestimating the reduction of strongly deleterious mutations (error bars evaluated with $N_{\mathrm{counts}} \pm \sqrt{N_{\mathrm{counts}}}$).

We then draw synthetic sets of mutations from $\mathcal{M}$ that resemble the original sets (germline or somatic) in their essential characteristics; *i.e.*, they have the same total number of missense mutations and the same biases in the different mutation channels (mutations C:G > T:A occur more frequently in both mutation sets than would be expected by chance) (GREENMAN *et al.* 2007). Such biases appear already at the level of the neutral process, *e.g.*, transition–transversion bias.

MATERIALS AND METHODS

Data set: We work with human cancer mutation data as given by GREENMAN *et al.* (2007). The data consist of the following:

i. A reference kinase gene set in nucleotide space with introns removed. We have translated the sequences to amino acid space and this set of genes forms our reference kinome $\mathbf{a}_{\mathrm{ref}}$.
ii. A list of somatic mutations, *i.e.*, exclusive to 210 cancer samples. We consider only missense mutations.
iii. A list of germline variants from these patients, which we have polarized using chimpanzee as an outgroup to determine their ancestral alleles. For 142 germline variants the chimpanzee allele did not decide the ancestral allele, because it had a third option, or the amino acid was missing, or we could not identify it by our blast search against chimpanzee refseq sequences (PRUITT *et al.* 2007). Extrapolating from the set that we could polarize unambiguously we estimate that within the no-call set there are <10 variants for which using the $\mathbf{a}_{\mathrm{ref}}$ as the ancestral state results in an error. We therefore decided to include these no-call variants nevertheless, but note that leaving them out altogether does not change our results.
iv. A list of candidate cancer genes selected from Supplementary Table 4c in KAN *et al.* (2010) with the condition that copy number loss and gain frequencies are available. We

TABLE 1

Germline variation in kinases

		Germline	
Score	Level	*P*-value	Effect size
Δs	Locus	$<10^{-5}$	0.61
	Gene	$<10^{-5}$	0.52
$e^{S^{10}}\Delta s$	Locus	$<10^{-5}$	0.56
	Gene	$<10^{-5}$	0.49

TABLE 2

Genomic observables for candidate tumor suppressor genes

		Somatic		Germline	
Score	Level	*P*-value	Effect size	*P*-value	Effect size
c	Locus	0.004	1.39	NS	—
	Gene	0.02	1.33	NS	—
Δs	Locus	0.003	1.69	0.00006	0.55
	Gene	0.002	1.80	0.00002	0.50
$e^{S^{10}}\Delta s$	Locus	0.0007	1.98	0.00007	0.49
	Gene	0.0003	2.10	0.00002	0.45

use the copy number information to assign these genes to candidate tumor suppressor and onco gene categories. The decision criterion is as follows: if the rate of loss is greater than the rate of gain, we call the gene a candidate tumor suppressor gene; otherwise the gene is labeled as a candidate onco gene. This criterion is based on Figure 3.a in Kan *et al.* (2010), which shows that such a classification is a sensible first-order estimate for these genes. We do not claim that all these genes are tumor suppressor or onco genes, only that these lists should be enriched with real tumor suppressor and onco genes. The data set is summarized in Tables 4 and 5.

Evaluating other alignment-based scores: The HMMER (v3.0) program provides the hmmsearch utility that searches a set of sequences against a single hmm profile, giving *E*-values and bit scores for each. Using the Pfam-A seed profiles of each domain, we obtained these observables for all variant sequences that are one missense mutation away from the reference, *i.e.*, mutational opportunity $\mathcal{M}$. We then used the difference in bit score between variant and reference sequence, converted to natural logarithms (see Clifford *et al.* 2004). The fundamental similarity to Equation 2 should be clear:

$$\frac{\Delta s_{\mathrm{HMM}}}{\log 2} = \log_2 \frac{\mathrm{P}\left(\mathrm{seq}_{\mathrm{var}} \in \mathrm{HMM}\right)}{\mathrm{P}\left(\mathrm{seq}_{\mathrm{ref}} \in \mathrm{HMM}\right)} - \log_2 \frac{\mathrm{P}\left(\mathrm{seq}_{\mathrm{var}} \in \mathrm{Null}\right)}{\mathrm{P}\left(\mathrm{seq}_{\mathrm{ref}} \in \mathrm{Null}\right)}. \quad (16)$$

For the SIFT and B-SIFT scores, we installed the latest version of SIFT (v4.0.3 together with BLIMPS v3.8) locally and likewise produced SIFT scores for all mutational opportunity (for SIFT see Ng and Henikoff 2003). We used cutoffs of SIFT <0.05 and B-SIFT >0.5 to call a variant deleterious or beneficial, respectively. To make the comparison more definite, we based the SIFT predictions on the same domain alignments that are used to infer Δs. However, more generally, an essential part of the SIFT procedure is to find homolog sequences in a database like UniProt/TrEMBL via PSI-BLAST, which naturally results in different alignments and different sets of scorable mutations.

RESULTS AND DISCUSSION

Germline mutations: It is clear that the assumptions under which we can equate the scores Δs to germline fitness are not fully satisfied within the data set at hand. However, the fact that the calibration of the scores does not use information about the germline polymorphism gives us an opportunity to predict their behavior as a function of the score. This is done in Figure 2, which shows the polymorphism density as a function of the predicted fitness effect of the variant.

We can evaluate this density analytically (Equation 11) and this prediction is consistent with the score Δs being a measurement of germline fitness. The agreement is surprisingly good given the simplicity of the model—albeit our score is clearly underestimating the real fitness cost of the big-effect mutations. Such an application of Pfam domain alignments has been performed at a genome-wide scale by Moses and Durbin (2009), who investigated polymorphism frequency spectra, rates of substitutions, and so-called MK ratios (after the McDonald–Kreitman test). Our results for germline variation are consistent with their findings except one notable difference: their theoretical prediction of substitution rate underestimates strongly deleterious substitution rates and overestimates the rate for beneficial substitutions (see their Figure 5C). This means that for substitutions at least, the effect of selection is somewhat overestimated. The difference may stem from the fact that our data set contains only kinases, which is a quite homogenous group of genes. Nevertheless, it is clear that even when applied at a whole-genome level, the predictions that follow from such a scoring provide results consistent with it being an estimate of germline fitness to a good degree. We note that several studies have shown a correlation between germline polymorphism allele frequencies and different conservation-based scores (see discussion in Jordan *et al.* 2010). However, as these results are strongly dependent on the quality of the multiple alignment in the sense discussed earlier (*Assumptions underlying the link between scores Δs and germline fitness*), applying such methods to somatic mutations should be done in conjunction with a corresponding germline variation analysis to address the validity of the scoring.

Our genomic observables evaluated for the germline mutations recapitulate the discussion above: scores for germline variants have significantly lower deleterious effects than the mutations in the corresponding null model have. The effect sizes for germline mutation scores are mostly ~50% of the expectation under the null. These results are summarized in Table 1 and are contrasted to somatic variation, discussed next.

Somatic mutations: The global pattern of somatic variation in the kinase set is not significantly different

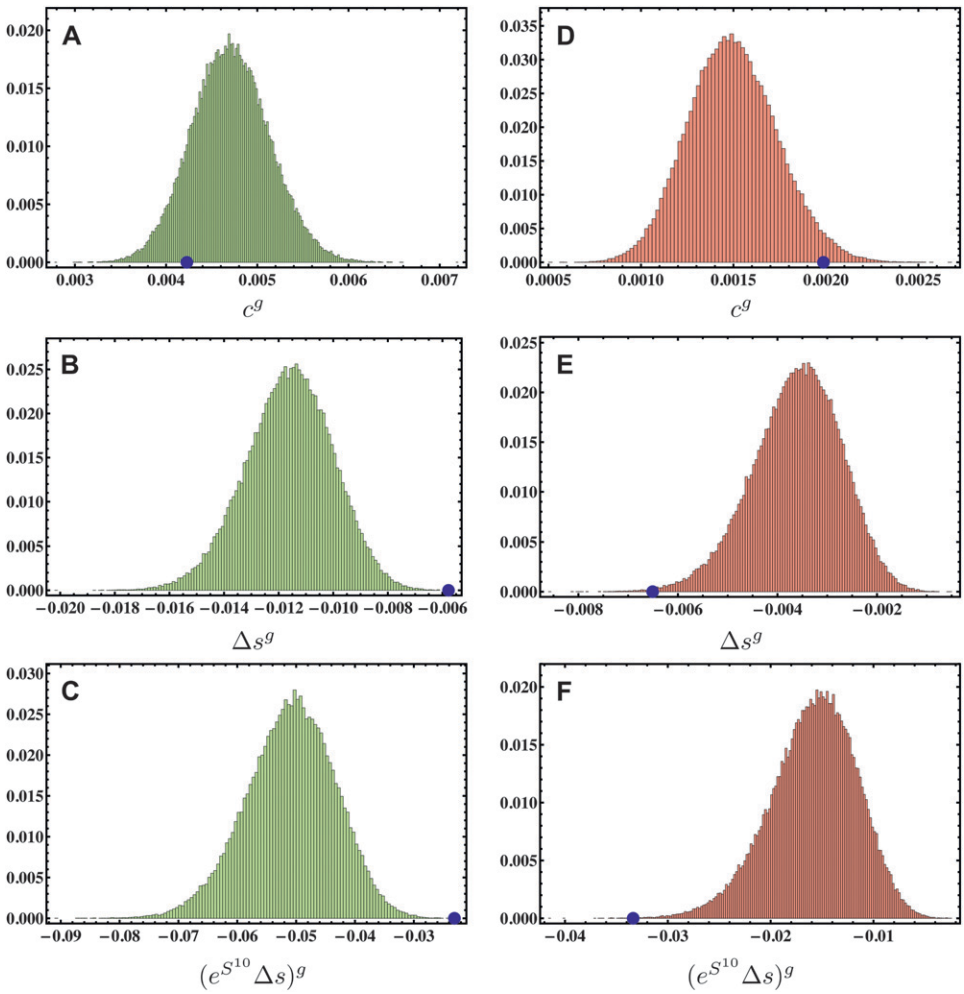

FIGURE 3.—Gene-level observables averaged over candidate tumor suppressor genes. Histograms denote the obtained averages in 10^5 synthetic sets (null model) and blue dots denote the values in the data. (A–C) Germline mutations (green). (A) Count scores c^g show no significant effect. In contrast, scores for germline mutations are less deleterious than for mutations in the null. (B) Δs^g, *P*-value $= 2 \times 10^{-5}$, effect size 0.50 (evaluated as data value divided by the mean of synthetic sets). (C) $(e^{S^{10}}\Delta s)^g$, *P*-value $= 2 \times 10^{-5}$, effect size 0.45. (D–F) Somatic mutations (red). (D) Count scores c^g, *P*-value $= 0.02$, effect size 1.33. There is a surplus of counts over what would be expected within the null. Furthermore, germline fitness scores for somatic mutations are more deleterious than for mutations in the null. (E) Δs^g, *P*-value $= 0.002$, effect size 1.80. (F) $(e^{S^{10}}\Delta s)^g$, *P*-value $= 0.0003$, effect size 2.10.

from random for the locus- and gene-level observables studied here. This is surprising even though a majority of somatic mutations are expected to be passengers (STRATTON *et al.* 2009); it is interesting to contemplate how large a genetic load passenger mutations impose on the cells. To assess whether somatic mutations in cancer genes are different we use a list of hand-curated candidate cancer genes from KAN *et al.* (2010) to delineate the data set (see MATERIALS AND METHODS). As distinct classes of cancer genes are known to contribute to tumorigenesis in a different manner, we further divide these genes into candidate tumor suppressor and onco genes, using information on copy number loss and gain frequencies (see MATERIALS AND METHODS for the precise criterion).

Somatic mutations in candidate tumor suppressor genes: In this gene set the impact of somatic mutations is on average more deleterious than the impact of random mutations from the null (see Table 2 and Figure 3). Furthermore, they fall preferably onto loci residing in neighborhoods with high overall germline fitness per allele. We see this pattern for both locus- and gene-level observables but the improvement with respect to count scores alone is stronger at the gene level where we see almost a two orders of magnitude drop in the *P*-values and a considerable increase in effect sizes of 1.33 → 2.10 (see Table 2). This implies that using the germline scores at the gene level integrates the underlying biological signal coherently and thus enhances the differences between the somatic variants and the null model. In Figure 3 we show results for this set for both germline and somatic mutations.

Furthermore, we calculated the correlation between the copy number loss frequency (data from KAN *et al.* 2010) and the combined score at the gene level. Considering only genes that have a nontrivial score (*i.e.*, only genes with somatic mutations), we observed a modest correlation of −0.4 with a *P*-value of 0.01 (32 genes in the sample; the mutation scores for genes are given in supporting information, File S1.). This correlation seems to arise largely from a few outliers, which suggests that the scores work for a subset of genes. This scatter may stem from the fact that in the candidate tumor suppressor gene set we expect only an enrichment of real tumor suppressor genes.

Among all kinases with respect to their somatic mutation scores, MAP2K4 stands out with a *P*-value of 0.015 for the mutation score and 0.041 for the

TABLE 3

Results for s_{SIFT} and Δs_{HMM} scores

		Germline all		Somatic candidate tumor suppressor		Germline candidate tumor suppressor	
Score	Level	*P*-value	Effect size	*P*-value	Effect size	*P*-value	Effect size
s_{SIFT}	Locus	$<10^{-5}$	0.61	0.02	1.61	0.0001	0.46
	Gene	$<10^{-5}$	0.55	0.02	1.73	0.00008	0.43
Δs_{HMM}	Locus	$<10^{-5}$	0.51	0.006	1.75	$<10^{-5}$	0.43
	Gene	$<10^{-5}$	0.45	0.003	1.91	$<10^{-5}$	0.39

combined score at gene level (after Bonferroni correction for multiple testing, with the number of kinase genes being 518; the mutation scores for genes are given in supporting information, File S1). The germline scores thus predict MAP2K4 as a cancer gene (without the information from the copy number rates or its membership in the curated candidate cancer gene list). For comparison we note that for the count score this is not the case (*P*-value = 0.36), and thus the additional information from the germline scores proves decisive. This is an example of how germline fitness scoring can be exploited in finding individual cancer genes. Interestingly, experimental results suggest a dominant negative role for MAP2K4 (Kan *et al.* 2010). The important role of MAP2K4 in oncogenesis is reviewed in Whitmarsh and Davis (2007).

Several investigations have used conservation-based measures to analyze cancer mutations: Talavera *et al.* (2010) report putative driver mutations to be enriched in conserved positions when compared to passenger mutations, Izarzugaza *et al.* (2009) find that predefined driver mutations are closer to regions important for function and conserved residues, Dixit *et al.* (2009) find driver mutations falling onto locations with slightly higher conservation signals than passenger mutations, and Mort *et al.* (2010) show that cancer mutations have relative enrichment of deleterious effects when compared to a neutral polymorphism set. While our results are in general consistent with these studies, at the same time, we make the case for the role that germline fitness plays in cancer substantially sharper. This clarity stems from the following reasons in particular. First, we use an explicit population genetic-based model to derive our scoring system, which allows us to state under which conditions we can expect the score to be germline fitness. Second, we show by predicting to a good degree the germline polymorphism rate that our scoring is performing well on its primary task. Third, we have a null model that allows us to analyze the germline and somatic mutations separately. Thus, effects of somatic mutations are not misinterpreted by comparing them to the germline variants that can be under selective pressures themselves. Finally, we focus on a data set where we have precise knowledge on the variation seen and not seen—aspects important for the calibration of the null model.

Somatic mutations in candidate onco genes: Mutations in candidate onco genes do not differ from the null model in a statistically significant way for the observables reported here. This clear contrast to mutations in candidate tumor suppressor genes may point to activating mutations being more heterogenous in conferring their cancerous effect, or to the germline fitness playing no (or a smaller) role for them, or to our simple classification criterion being a definition not accurate enough for onco genes. However, we note that Lee *et al.* (2009b) put forward a proposition that a subset of activating mutations may have positive scores, *i.e.*, that they are germline beneficial, and they provide structural evidence for predictions that such a criterion produces. They define a relative score B-SIFT = SIFT(mutant) − SIFT(wild type) (B for bidirectional; for SIFT see Ng and Henikoff 2003), which in principle can capture both deleterious and beneficial substitutions, similar to our Δs. In our candidate onco genes, we neither find any B-SIFT beneficial mutations (those with B-SIFT > 0.5) nor see a surplus of positive scores Δs. An explanation, also given by Lee *et al.* (2009b), could be that a majority of *functionally* activating mutations in cancer are still germline *deleterious*, in a sense that they are not observed in healthy cells.

TABLE 4

Number of (available) mutations in the different categories

	Opportunity (average) (10^5)			Somatic			Germline		
	All	T. supp.	Onco	All	Tumor suppressor	Onco	All	Tumor suppressor	Onco
Total	29.37	3.63	3.68	620	100	83	2423	277	264
Scored	14.26	1.78	1.87	324	56	49	1018	125	102

TABLE 5

Mutational biases

Channel	Opportunity (%)	Somatic (%)	Germline (%)
A:T > T:A	17	7	5
A:T > C:G	19	3	5
A:T > G:C	16	10	21
C:G > G:C	19	13	11
C:G > A:T	16	10	9
C:G > T:A	13	57	49

Comparison to other scores extracted from alignments: There are several ways the effects of mutations can be scored given an alignment. Our focus here is to explore the relation between germline and somatic fitness, and thus the score Δs defined in Equation 2 is particularly appealing for it allows a direct and intuitive connection to population genetic theory as discussed earlier. However, it is also of interest to see how Δs compares to other alignment scores in terms of its ability to separate the germline and somatic variation from the null. To that extent we evaluated the widely used SIFT (s_{SIFT}) (Ng and Henikoff 2003) and HMMER3 scores (Δs_{HMM}) (Clifford *et al.* 2004) (http://hmmer.org) for the domain alignments in the data set (see materials and methods). For germline variation, both s_{SIFT} and Δs_{HMM} scores give equivalent results to that of Δs. For somatic mutations in candidate tumor suppressor genes we observe that s_{SIFT} shows lower performance in separating the data and the null than Δs_{HMM} and Δs, which give very similar results (see Table 3).

We further correlated the scores Δs_{HMM} and Δs and record a correlation of 0.71 for germline variants and 0.80 for somatic variants. The similarity in terms of results and the high correlation between these two measures is not unexpected. In bioinformatic terms, they both score a mutation via its predicted change to the probability of that residue conforming to the particular alignment. We note, however, that HMMER3 is underpinned by a more sophisticated (and more complicated) probability model for the sequences. Thus, linking its predictions to an explicit population genetic model may not be feasible. In summary, while Δs_{HMM} performs bioinformatically similarly to Δs for the studied data, the latter has an advantage if we wish to utilize population genetic theory either as a basis for interpretation of the results or to develop scoring systems going beyond individual mutations (*e.g.*, by forming locus scores S^w).

Different levels of integration: Our scoring system is built on domain alignments that are used to extract germline fitness. Above we have used the scores at the level of loci and genes and we have seen that integration of germline fitness scores at the gene level enhances our ability to differentiate between the somatic and null sets. Ideally, we wish to perform this integration also at the level of domains. Indeed, we note that there is an effect in the number of somatic mutations that can be scored, *i.e.*, fall onto Pfam domains. This set is significantly larger than expected if they were falling randomly onto the kinases genes (P-value $= 3.1 \times 10^{-2}$). This result seems to hold more generally: see enrichment analysis of Pfam domain mutations across multiple cancer data sets in Li *et al.* (2009) and clustering analysis of so-called mutation hotspots in protein domains (Yue *et al.* 2010). However, this effect is conditioned out in our analysis as our null model has precisely the same number of scorable missense mutations as the somatic set does. Therefore any direct comparison to our gene-level scores would be biased. Once genome-wide cancer sequences arrive in numbers, the question of the optimal level of integration should be addressed also at the pathway level (for pathway considerations see, *e.g.*, Ding *et al.* 2008; Kan *et al.* 2010).

Relationship between somatic $F_{\mathrm{som.}}$ and germline fitness $F_{\mathrm{germ.}}$: We note that there is no *a priori* reason why *germline fitness* should reflect the *somatic fitness* and evolution at all. However, on the basis of the results presented above it seems clear that within the class of candidate tumor suppressor genes there is a definite statistical relationship between predicted germline fitness effects and the acquired somatic mutations. In this study we used for the neighborhood scores a window size $l_w = 21$ ($w = 10$), which we selected on the basis of an autocorrelation analysis of the information content of the domain alignments. However, finding the relevant length scale for cancer variation systematically is left to future work. At this time point we note that for our main result, *i.e.*, that the germline fitness score is relevant for tumor suppressor genes, it is sufficient to use only mutation scores without a window score weight factor, but this weakens the effect as can be seen from Figure 3 and Table 2. Here we chose a nonlinear weighting function for two reasons: first, to underline the fact that such a nonlinearity would affect the analysis since the order of taking the averages would matter, *i.e.*, $F_{\mathrm{som.}}(\langle F_{\mathrm{germ.}}\rangle) \neq \langle F_{\mathrm{som.}}(F_{\mathrm{germ.}})\rangle$, where $\langle \dots \rangle$ denotes ensemble averages [given that the substitution rate of a mutation in a population is by itself a strongly nonlinear function of fitness effect (see Equation 5), it does not seem unreasonable to anticipate such a behavior]; and second, to highlight our general ignorance about the specifics of the relationship [our ansatz for $F_{\mathrm{som.}}(F_{\mathrm{germ.}})$ with a sign reversal (*i.e.*, germline deleterious roughly equals somatic beneficial) in conjunction with a neighborhood-based weight relies implicitly on the relationship being monotonic]. This is likely not going to hold for many strongly deleterious germline variants that in essence kill the cell. It has been proposed that the alleles underlying complex traits have more subtle effects on disease risk and are hence more likely to include variants that affect the gene function modestly (Hirschhorn and Daly 2005). Indeed, common disease mutations in the kinases have been shown to fall

preferably onto regions where they have *moderate* effect on function (Torkamani *et al.* 2008). It will be of considerable interest to try to quantify both the cancer-specific length scale (if it exists) and the functional form of the fitness relation in future work.

Conclusions: There is little doubt that evolutionary theory will continue to be at the center stage in analyses of somatic mutations as the ongoing large-scale sequencing efforts generate more data (see the International Cancer Genome Consortium 2010 for data efforts and Frank 2010 for a discussion on the time evolution of somatic variation). There is a lot to be learned as we still do not have a quantitative understanding of the key evolutionary parameters for cancer such as the selective advantages of the driver mutations, albeit estimates based on modeling and known timescales of tumor progression have been put forward (Beerenwinkel *et al.* 2007; Bozic *et al.* 2010).

Here we focused on analyzing somatic variants from the perspective of their germline fitness effects. As it is not clear from the outset that germline fitness has anything to do with somatic fitness of the cells, this is an interesting theoretical question. We have shown that for candidate tumor suppressor genes there is a relation where cancer mutations are on average more germline deleterious than random mutations—practical implications of this result are obvious, *e.g.*, the prediction of MAP2K4 as a cancer gene above. Our findings point to future directions of integrating germline scores into a computational framework that uses additional information, *e.g.*, synonymous mutations and missense events that do not fall into domains. Such an integration will be important as we could assign a germline score for only slightly more than half of the missense variants seen. Another important direction would be developing similar methods to score also insertions and deletions. This will require a basic understanding of germline rates of these events that can be attained from detailed analyses of human genomes (see, *e.g.*, http://www.1000genomes.org/). Finally, we point out that calibrating germline fitness effects as was done here has been applied also to noncoding variation in the context of binding sites (Moses *et al.* 2004; Mustonen and Lässig 2005; Doniger and Fay 2007), so the framework will be applicable for cancer variation in regulatory regions as well.

We thank the editor and two anonymous reviewers for helpful comments and suggestions and Leonid Mirny for discussions. A.F. thanks the Deutsche Forschungsgemeinschaft for funding through the Bonn-Cologne Graduate School for Physics and Astronomy and the Sonderforschungsbereich/Transregio 12. V.M. acknowledges the Wellcome Trust for support under grant 091747.

LITERATURE CITED

Beerenwinkel, N., T. Antal, D. Dingli, A. Traulsen, K. W. Kinzler *et al.*, 2007 Genetic progression and the waiting time to cancer. PLoS Comput. Biol. **3:** e225.

Berg, J., S. Willmann and M. Lässig, 2004 Adaptive evolution of transcription factor binding sites. BMC Evol. Biol. **4:** 1.

Bignell, G. R., C. D. Greenman, H. Davies, A. P. Butler and S. Edkins *et al.*, 2010 Signatures of mutation and selection in the cancer genome. Nature **463:** 893–898.

Bozic, I., T. Antal, H. Ohtsuki, H. Carter, D. Kim *et al.*, 2010 Accumulation of driver and passenger mutations during tumor progression. Proc. Natl. Acad. Sci. USA **107:** 18545–18550.

Bromberg, Y., and B. Rost, 2007 SNAP: predict effect of non-synonymous polymorphisms on function. Nucleic Acids Res. **35:** 3823–3835.

Bulmer, M., 1991 The selection-mutation-drift theory of synonymous codon usage. Genetics **129:** 897–907.

Carter, H., S. Chen, L. Isik, S. Tyekucheva, V. E. Velculescu *et al.*, 2009 Cancer-specific high-throughput annotation of somatic mutations: computational prediction of driver missense mutations. Cancer Res. **69:** 6660–6667.

Clifford, R. J., M. N. Edmonson, C. Nguyen and K. H. Buetow, 2004 Large-scale analysis of non-synonymous coding region single nucleotide polymorphisms. Bioinformatics **20:** 1006–1014.

Davies, H., G. R. Bignell, C. Cox, P. Stephens, S. Edkins *et al.*, 2002 Mutations of the BRAF gene in human cancer. Nature **417:** 949–954.

Ding, L., G. Getz, D. A. Wheeler, E. R. Mardis, M. D. McLellan *et al.*, 2008 Somatic mutations affect key pathways in lung adenocarcinoma. Nature **455:** 1069–1075.

Dixit, A., L. Yi, R. Gowthaman, A. Torkamani, N. J. Schork *et al.*, 2009 Sequence and structure signatures of cancer mutation hotspots in protein kinases. PLoS ONE **4:** e7485.

Doniger, S. W., and J. C. Fay, 2007 Frequent gain and loss of functional transcription factor binding sites. PLoS Comput. Biol. **3:** e99.

Durbin, R., S. Eddy, A. Krogh and G. Mitchison, 1998 *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids.* Cambridge University Press, Cambridge, UK/London/New York.

Eyre-Walker, A., 2006 The genomic rate of adaptive evolution. Trends Ecol. Evol. (Amst.) **21:** 569–575.

Fay, J. C., and C. I. Wu, 2000 Hitchhiking under positive Darwinian selection. Genetics **155:** 1405–1413.

Finn, R. D., J. Mistry, J. Tate, P. Coggill, A. Heger *et al.*, 2010 The Pfam protein families database. Nucleic Acids Res. **38**(Database issue): D211–D222.

Frank, S. A., 2010 Evolution in health and medicine Sackler colloquium: somatic evolutionary genomics: mutations during development cause highly variable genetic mosaicism with risk of cancer and neurodegeneration. Proc. Natl. Acad. Sci. USA **107** (Suppl 1): 1725–1730.

Gardiner, C. W., 2009 *Stochastic Methods: A Handbook for the Natural and Social Sciences.* Springer, Heidelberg.

Greenman, C., R. Wooster, P. A. Futreal, M. R. Stratton and D. F. Easton, 2006 Statistical analysis of pathogenicity of somatic mutations in cancer. Genetics **173:** 2187–2198.

Greenman, C., P. Stephens, R. Smith, G. L. Dalgliesh, C. Hunter *et al.*, 2007 Patterns of somatic mutation in human cancer genomes. Nature **446:** 153–158.

Halpern, A. L., and W. J. Bruno, 1998 Evolutionary distances for protein-coding sequences: modeling site-specific residue frequencies. Mol. Biol. Evol. **15:** 910–917.

Henikoff, J. G., and S. Henikoff, 1996 Using substitution probabilities to improve position-specific scoring matrices. Bioinformatics **12:** 135–143.

Hirschhorn, J. N., and M. J. Daly, 2005 Genome-wide association studies for common diseases and complex traits. Nat. Rev. Genet. **6:** 95–108.

International Cancer Genome Consortium, 2010 International network of cancer genome projects. Nature **464:** 993–998.

Izarzugaza, J. M. G., O. C. Redfern, C. A. Orengo and A. Valencia, 2009 Cancer-associated mutations are preferentially distributed in protein kinase functional sites. Proteins **77:** 892–903.

Jordan, D. M., V. E. Ramensky and S. R. Sunyaev, 2010 Human allelic variation: perspective from protein function, structure, and evolution. Curr. Opin. Struct. Biol. **20:** 342–350.

Kaminker, J. S., Y. Zhang, C. Watanabe and Z. Zhang, 2007 CanPredict: a computational tool for predicting cancer-associated

missense mutations. Nucleic Acids Res. **35**(Web Server issue): W595–W598.

Kan, Z., B. S. Jaiswal, J. Stinson, V. Janakiraman, D. Bhatt *et al.*, 2010 Diverse somatic mutation patterns and pathway alterations in human cancers. Nature **466:** 869–873.

Kimura, M., 1955 Stochastic processes and distribution of gene frequencies under natural selection. Cold Spring Harbor Symp. Quant. Biol. **20:** 33–53.

Kimura, M., 1964 Diffusion models in population genetics. J. Appl. Probab. **1:** 177–232.

Kreitman, M., 2000 Methods to detect selection in populations with applications to the human. Annu. Rev. Genomics Hum. Genet. **1:** 539–559.

Kryukov, G. V., L. A. Pennacchio and S. R. Sunyaev, 2007 Most rare missense alleles are deleterious in humans: implications for complex disease and association studies. Am. J. Hum. Genet. **80:** 727–739.

Lahiry, P., A. Torkamani, N. J. Schork and R. A. Hegele, 2010 Kinase mutations in human disease: interpreting genotype–phenotype relationships. Nat. Rev. Genet. **11:** 60–74.

Lawrence, C. E., S. F. Altschul, M. S. Boguski, J. S. Liu, A. F. Neuwald *et al.*, 1993 Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. Science **262:** 208–214.

Lee, W., P. Yue and Z. Zhang, 2009a Analytical methods for inferring functional effects of single base pair substitutions in human cancers. Hum. Genet. **126:** 481–498.

Lee, W., Y. Zhang, K. Mukhyala, R. A. Lazarus and Z. Zhang, 2009b Bi-directional SIFT predicts a subset of activating mutations. PLoS ONE **4:** e8311.

Li, L., K. Zhang, J. Lee, S. Cordes, D. P. Davis *et al.*, 2009 Discovering cancer genes by integrating network and functional properties. BMC Med. Genomics **2:** 61.

Manning, G., D. B. Whyte, R. Martinez, T. Hunter and S. Sudarsanam, 2002 The protein kinase complement of the human genome. Science **298:** 1912–1934.

Merlo, L. M. F., J. W. Pepper, B. J. Reid and C. C. Maley, 2006 Cancer as an evolutionary and ecological process. Nat. Rev. Cancer **6:** 924–935.

Mort, M., U. S. Evani, V. G. Krishnan, K. K. Kamati, P. H. Baenziger *et al.*, 2010 In silico functional profiling of human disease-associated and polymorphic amino acid substitutions. Hum. Mutat. **31:** 335–346.

Moses, A. M., and R. Durbin, 2009 Inferring selection on amino acid preference in protein domains. Mol. Biol. Evol. **26:** 527–536.

Moses, A. M., D. Y. Chiang, D. A. Pollard, V. N. Iyer and M. B. Eisen, 2004 MONKEY: identifying conserved transcription-factor binding sites in multiple alignments using a binding site-specific evolutionary model. Genome Biol. **5:** R98.

Mustonen, V., and M. Lässig, 2005 Evolutionary population genetics of promoters: predicting binding sites and functional phylogenies. Proc. Natl. Acad. Sci. USA **102:** 15936–15941.

Ng, P. C., and S. Henikoff, 2001 Predicting deleterious amino acid substitutions. Genome Res. **11:** 863–874.

Ng, P. C., and S. Henikoff, 2003 SIFT: predicting amino acid changes that affect protein function. Nucleic Acids Res. **31:** 3812–3814.

Nielsen, R., 2005 Molecular signatures of natural selection. Annu. Rev. Genet. **39:** 197–218.

Pruitt, K. D., T. Tatusova and D. R. Maglott, 2007 NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res. **35**(Database issue): D61–D65.

Radivojac, P., P. H. Baenziger, M. G. Kann, M. E. Mort, M. W. Hahn *et al.*, 2008 Gain and loss of phosphorylation sites in human cancer. Bioinformatics **24:** i241–i247.

Ramensky, V. E., P. Bork and S. R. Sunyaev, 2002 Human non-synonymous SNPs: server and survey. Nucleic Acids Res. **30:** 3894–3900.

Rouzine, I. M., A. Rodrigo and J. M. Coffin, 2001 Transition between stochastic evolution and deterministic evolution in the presence of selection: general theory and application to virology. Microbiol. Mol. Biol. Rev. **65:** 151–185.

Rubin, A. F., and P. Green, 2009 Mutation patterns in cancer genomes. Proc. Natl. Acad. Sci. USA **106:** 21766–21770.

Sawyer, S. A., and D. L. Hartl, 1992 Population genetics of polymorphism and divergence. Genetics **132:** 1161–1176.

Sjöblom, T., S. Jones, L. D. Wood, D. W. Parsons, J. Lin *et al.*, 2006 The consensus coding sequences of human breast and colorectal cancers. Science **314:** 268–274.

Stratton, M. R., P. J. Campbell and A. Futreal, 2009 The cancer genome. Nature **458:** 719–724.

Talavera, D., M. S. Taylor and J. M. Thornton, 2010 The (non) malignancy of cancerous amino acidic substitutions. Proteins **78:** 518–529.

Torkamani, A., and N. J. Schork, 2008 Prediction of cancer driver mutations in protein kinases. Cancer Res. **68:** 1675–1682.

Torkamani, A., N. Kannan, S. Taylor and N. Schork, 2008 Congenital disease SNPs target lineage specific structural elements in protein kinases. Proc. Natl. Acad. Sci. USA **105:** 9011–9016.

Torkamani, A., G. Verkhivker and N. J. Schork, 2009 Cancer driver mutations in protein kinase genes. Cancer Lett. **281:** 117–127.

Whitmarsh, A., and R. Davis, 2007 Role of mitogen-activated protein kinase kinase 4 in cancer. Oncogene **26:** 3172–3184.

Yang, Z., S. Ro and B. Rannala, 2003 Likelihood models of somatic mutation and codon substitution in cancer genes. Genetics **165:** 695–705.

Yue, P., W. F. Forrest, J. S. Kaminker, S. Lohr, Z. Zhang *et al.*, 2010 Inferring the functional effects of mutation through clusters of mutations in homologous proteins. Hum. Mutat. **31:** 264–271.

Communicating editor: L. M. Wahl

GENETICS

Supporting Information
http://www.genetics.org/cgi/content/full/genetics.111.127480/DC1

Germline Fitness-Based Scoring of Cancer Mutations

Andrej Fischer, Chris Greenman and Ville Mustonen

DOI: 10.1534/genetics.111.127480

File S1

Supporting Data

File S1 is available for download as a .txt file at http://www.genetics.org/cgi/content/full/genetics.111.127480/DC1.