Journal of Human Genetics (2016) 61, 627–632

www.nature.com/jhg

ORIGINAL ARTICLE

Evaluation of the evenness score in next-generation sequencing

Konrad Oexle

The evenness score (*E*) in next-generation sequencing (NGS) quantifies the homogeneity in coverage of the NGS targets. Here I clarify the mathematical description of *E*, which is 1 minus the integral from 0 to 1 over the cumulative distribution function *F*(*x*) of the normalized coverage *x*, where normalization means division by the mean, and derive a computationally more efficient formula; that is, 1 minus the integral from 0 to 1 over the probability density distribution *f*(*x*) times 1–*x*. An analogous formula for empirical coverage data is provided as well as fast R command line scripts. This new formula allows for a general comparison of *E* with the coefficient of variation (= standard deviation σ of normalized data) which is the conventional measure of the relative width of a distribution. For symmetrical distributions, including the Gaussian, *E* can be predicted closely as $1-\sigma^2/2 \geqslant E \geqslant 1-\sigma/2$ with $\sigma \leqslant 1$ owing to normalization and symmetry. In case of the log-normal distribution as a typical representative of positively skewed biological data, the analysis yields $E \approx \exp(-\sigma^*/2)$ with $\sigma^{*2} = \ln(\sigma^2+1)$ up to large σ ($\leqslant 3$), and $E \approx 1-F(\exp(-1))$ for very large σ ($\geqslant 2.5$). In the latter kind of rather uneven coverage, *E* can provide direct information on the fraction of well-covered targets that is not immediately delivered by the normalized σ. Otherwise, *E* does not appear to have major advantages over σ or over a simple score $\exp(-\sigma)$ based on it. Actually, $\exp(-\sigma)$ exploits a much larger part of its range for the evaluation of realistic NGS outputs.
Journal of Human Genetics (2016) **61,** 627–632; doi:10.1038/jhg.2016.21; published online 14 April 2016

INTRODUCTION

Next-generation sequencing (NGS) techniques use random ('shotgun') sequencing of the template DNA in order to cover all 'targets' with a sufficient number of sequencing reads, that is, to reach a sufficient 'coverage'. Accordingly, NGS always involves at least the fluctuation of a Poisson process. The distribution of the coverage thus cannot be entirely even but must have a variance that is at least as large as the mean (as in case of a Poisson distribution). On top of that lower bound, real NGS distributions show overdispersion and have variances that are substantially larger than their means. Overdispersion is due to various factors, including copy number variability of the template DNA, for instance, or pre-NGS manipulations, such as selective capturing of template DNA.

To assess the degree of inhomogeneity of NGS coverage quantitatively, Mokry *et al.*[1] elaborated on a consideration of Gnirke *et al.*[2] and introduced the 'evenness score' *E*. This score has found its way into the NGS field. Very recently, for instance, Lelieveld *et al.*[3] applied *E* in their comparison of exome sequencing and whole-genome sequencing. Here I derive a computationally more efficient formula for the calculation of *E*. Then I use that formula in a general analysis, producing simple but close approximations. The latter allow for a comparison of the evenness score with conventional descriptors of the relative width of a distribution, such as the coefficient of variation.

MATERIAL & METHODS AND RESULTS

The evenness score *E*

Mokry *et al.*[1] developed the evenness score as a tool to describe the dispersion of the coverage around the average coverage C_{ave}. Their idea is intuitive and proved useful for their study, but it has not been extensively characterized mathematically. Here I show that the explanation and derivation of the evenness score can be simplified significantly, leading to a more efficient computation of the score and to general insights into its relationship with more traditional statistical measures. In order to simplify the explanation, it has to be reproduced in the following paragraph. Readers asking for an immediate intuitive understanding of the evenness score are referred to Figure 1 and may then proceed directly to Equation (3).

Mokry *et al.*[1] stated

$$E = 100\% \sum_{i=1}^{C_{ave}} \frac{M_i}{C_{ave} N_{TP}} \quad (1)$$

where M_i 'is defined as number of targeted positions with at least coverage C_i, C_{ave} is defined as the average coverage through all targeted positions and N_{TP} is defined as the total number of targeted positions.' The introduction of the term C_i in this definition is an unnecessary complication as the instruction of the sum in Equation (1) obviously implies that $C_i = i$. Thus, as Mokry *et al.*[1] stated, *E* equals 1 (= 100%)

Center for Cardiovascular Genetics and Gene Diagnostics, Foundation for People with Rare Diseases, Schlieren-Zurich, Switzerland
Correspondence: Professor K Oexle, Center for Cardiovascular Genetics and Gene Diagnostics, Foundation for People with Rare Diseases, Wagistrasse 25, Schlieren-Zurich CH-8952, Switzerland.
E-mail: oexlekonrad@netscape.net
Received 5 November 2015; revised 21 February 2016; accepted 21 February 2016; published online 14 April 2016

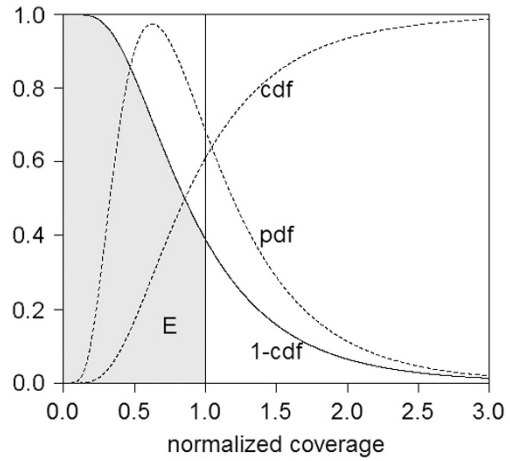

Figure 1 Explanation of the evenness score E (shaded area) according to Mokry *et al.*[1] (also see their Figure 2b). If the coverage is homogeneous, its pdf (probability density function) is narrow and close to the mean of the normalized coverage at $x=1$. Then the shaded area approximates a size of 1 or 100%. E is calculated as $\int_0^1 1-F(x)dx = 1-\int_0^1 F(x)dx = 1-\int_0^1(1-x)f(x)dx$ (see derivation of Equation (4)) where $f(x)$ is the pdf and $F(x)$ is the cumulative distribution function (cdf).

in case of completely uniform coverage of all targeted positions at a level of C_{ave} because in this case $M_i/N_{TP}=1$ for all $i \leqslant C_{ave}$ yielding $\sum_{i=1}^{Cave} \frac{M_i}{C_{ave}N_{TP}} = \sum_{i=1}^{Cave} \frac{1}{C_{ave}} = 1$. (Mokry *et al.*[1] used the letter P_i instead of M_i in Equation (1) which is avoided here as P usually indicates a probability or relative frequency. Only after dividing M_i by N_{TP}, a probability results, that is, $P(coverage \geqslant i) = M_i/N_{TP}$.)

Mokry *et al.*[1] also provided a version of Equation (1) for the continuous case, that is, '$E = 100\% \int_0^1 F(i)di$, …where $F(i)$ is the fraction of positions with normalized coverage of at least $C(i)/C_{ave}$', with 'normalization' meaning division by the mean. Again, the definition is a little complicated as the reader needs to figure out that $C(i)/C_{ave}=i$. Moreover, the use of the letters i and F in this formula is unfavorable as i has been applied in Equation (1) already, although with a different meaning (!), and F usually relates to the left-sided cumulative distribution function (from $-\infty$ to x, see https://en.wikipedia.org/wiki/Cumulative_distribution_function). Therefore, I prefer to write

$$E = \int_0^1 G(x)dx \tag{2}$$

with $x \approx i/C_{ave}$ and $G(x) \approx M_i/N_{TP}$, where i is defined as in Equation (1), and $G(x)$ is the fraction of positions with normalized coverage of at least x. In Equation (2), I omitted the factor '100%' as it equals 1 anyway. With increasing C_{ave}, the residual difference between the discrete and the continuous version of E declines. Mokry *et al.*[1] used the continuous version for a visual explanation of the evenness score (see their 'Figure 2' and Figure 1 of the present paper). This explanation implies that $G(x)$ (that is, '$F(i)$' in terms of Mokry *et al.*[1]) is the complement of the cumulative distribution function (cdf) of the normalized coverage. Hence $G(x)=1–F(x)$, because $1–G(x)$ equals the fraction of positions with normalized coverage of at most x, that is, the cdf for which I use the common descriptor $F(x)$ here. In case of a very even NGS result, almost all target positions have a coverage close to the mean, so that the probability density distribution (pdf) is restricted to the vicinity of the mean, the cdf is close to 0 for $x<1$, and the evenness score E approximates $1-\int_0^1 0di = 1$ or 100%. Conversely, a coverage that is uneven with $F(x)>0$ for $x<1$ results in $E<1$ (that is, $<100\%$).

Thus, except for the expression in percentage, the evenness score E of Mokry *et al.*[1] is given by

$$E = \int_0^1 1-F(x)dx \tag{3}$$

where $F(x)$ is the cdf of the normalized coverage $x=$coverage/mean coverage.

With $f(x)$ being the related pdf where $F(x) = \int_{-\infty}^{x} f(t)dt$, that is, $F(x) = \int_0^x f(t)dt$, as there is no negative coverage ($f(t)=0$ for $t<0$), a rather convenient expression can be derived from Equation (3) using integration by parts: As $\int_0^1 F(x)dx + \int_0^1 xf(x)dx = [xF(x)]_0^1 = F(1) = \int_0^1 f(x)dx$, Equation (3) can be written as $E = 1-\int_0^1 f(x)dx + \int_0^1 xf(x)dx$. Hence,

$$E = 1 - \int_0^1 (1-x)f(x)dx \tag{4}$$

For the discrete case with $x \approx i/C_{ave}$ and $f(x) \approx n_i/N_{TP}$, the analogous formula is

$$E = 1 - \sum_{i=0}^{C_{ave}} \frac{(C_{ave}-i)n_i}{C_{ave}N_{TP}} \tag{5a}$$

where N_{TP} and C_{ave} are defined as in Equation (1) as the total number of targeted positions and the average coverage, respectively, while n_i is the number of targets that are covered with exactly i reads (that is, i is the non-normalized coverage). Equation (5a) can be transformed to

$$E = 1 - \frac{1}{N_{TP}} \sum_{1 \leqslant j \leqslant N_{TP}, C(j) \leqslant C_{ave}} \left(1 - \frac{C(j)}{C_{ave}}\right) \tag{5b}$$

where the condition '$1 \leqslant j \leqslant N_{TP}$, $C(j) \leqslant C_{ave}$' guarantees that the index of the summation runs through all target positions j whose coverage $C(j)$ is not larger than the average coverage. As each of these positions occurs exactly once, Equation (5b) does not have a weighing factor comparable to n_i in Equation (5a) where i denotes coverage level instead of position. For a direct derivation of Equations (5a and 5b), from Equation (1), see Supplementary Material A and B and Supplementary Figure S1. As C_{ave} is usually not an integer, there might be a small deviation between the values calculated by Equation (1) and Equations (5a and 5b). The deviation is small for $C_{ave}>10$ but may be considerable if C_{ave} is of order $\leqslant 1$. The difference vanishes if C_{ave} is rounded to the next integer. With Equations (5a and 5b), the computation time to calculate E is a linear function of the number of NGS reads as each read is addressed only once in the summation over $(C_{ave}–i)n_i$, whereas Equation (1) requires computational time that increases as a quadratic function of the number of reads as each M_i represents a summation itself. For theoretical considerations (see below), Equations (4, 5a and 5b) are also more useful than Equations (1–3) because for various distributions, including the Gaussian normal distribution, $F(x)$ cannot be provided in closed form. Concerning computational efficiency, the situation is then similar to the discrete version, as Equation (4) involves only one numerical integration, while the calculation according to Equation (3) implies the numerical integration of numerical integrations.

Commands to calculate E according to Equations (4, 5a and 5b) on the R command line or as parts of R programs are as follows (see

Supplementary Material B for a detailed explanation): In case of empirical (that is, discrete) data, let D be a vector that contains the data as a sequence of numbers representing the coverage of each of the targets. If this sequence is the column k of a table T, use the command '$D = T[,k]$' to produce D. Then implementation of Equation (5b) in R yields E by the command line script

```
C=round(mean(D)); D2=D[D<=C]; E=1-(length(D2)-sum(D2)/C)/length(D); E
```

where C_{ave} is rounded to the next integer (which only has a substantial effect for data whose non-normalized C_{ave} is very small; that is, of order $\leqslant 1$). This command also works after normalization of the data. For operations with a theoretical distribution $f(x)$ of a continuous normalized random variable, Equation (4) can be implemented in R as

```
E=integrate(function(x){(1-x)*f(x)}, lower=0, upper=1)$value; E
```

where 'f(x)' has to be replaced by the specific pdf.

In the following, I derive approximations of the evenness score E, especially in terms of the distribution parameter σ. I also consider alternative scores such as $e^{-\sigma}$, which is restricted to the interval between 0 and 1 by definition and thus qualifies for scoring in percentage.

E and σ in case of a symmetrical pdf

As coverage always is positive, with $f(x) = 0$ for $x < 0$, and because normalization implies $\mu = 1$, the variance is given by $\sigma^2 = \int_{-\infty}^{\infty} f(x)(x-\mu)^2 dx = \int_0^{\infty} f(x)(x-1)^2 dx$. With both $f(x) \geqslant 0$ and $(x-1)^2 \geqslant 0$, the variance is $\sigma^2 = k_2 \int_0^1 f(x)(1-x)^2 dx$ where $k_2 > 1$. As $(1-x)^2 \leqslant 1-x$ for $0 \leqslant x \leqslant 1$, Equation (4) implies $\sigma^2/k_2 \leqslant 1-E$, which yields an upper limit of E. In analogy to k_2, a constant k_0 can be defined with $1 = k_0 \int_0^1 f(x)dx$, where $k_0 > 1$ since $\int_0^{\mu} f(x)dx < 1$ because $f(x)$ is a pdf. To derive the lower limit of E, apply Jensen's inequality for convex functions such as $(x-1)^2$ (see Supplementary Material C) yielding $\int_0^1 (1-x)^2 k_0 f(x)dx \geqslant (\int_0^1 (1-x) k_0 f(x)dx)^2$.

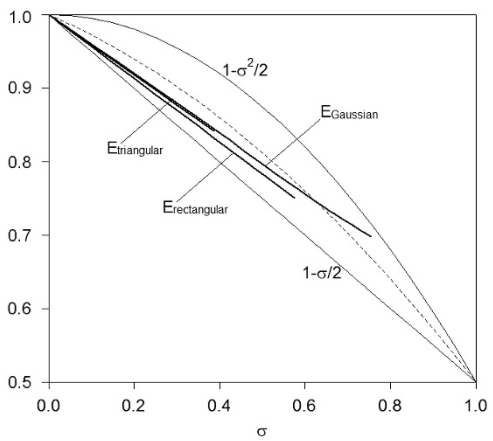

Figure 2 Evenness scores (E) of some symmetrical distributions as functions of the coefficient of variation, that is, of the standard deviation (σ) after normalization by division by the mean. Note that E is well predicted by the average (dashed line) of the upper ($1-\sigma^2/2$) and lower limits ($1-\sigma/2$) as given by inequality(7). Because of the normalization, the base length of the triangular and the rectangular distribution cannot be > 2, so that the maximal σ is $1/\sqrt{6} = 0.408$ and $1/\sqrt{3} = 0.577$, respectively. The Gaussian normal distribution necessarily is truncated at 0, which inflicts increasing skewness with increasing σ. Interestingly, the normalized left-truncated Gaussian distribution also has a maximal σ, which equals $\sqrt{\pi/2-1} = 0.756$ (see Supplementary Material F). As the figure shows, the E score of the distribution even then is well predicted by inequality(7).

With Equation (4), this is equivalent to $k_0\sigma^2/k_2 \geqslant (k_0(1-E))^2$. Hence,

$$1 - \frac{\sigma^2}{k_2} \geqslant E \geqslant 1 - \frac{\sigma}{\sqrt{k_0 k_2}}, \text{ with } k_0 > 1, k_2 > 1 \qquad (6)$$

The constants k_0 and k_2 depend on the form of the distribution; k_0 is associated with the relation of median m and mean $\mu = 1$. If $1 > m$, then $\int_0^1 f(x)dx > 0.5$, and $k_0 < 2$. In case of symmetrical pdfs, $m = \mu = 1$ and $k_0 = 2$. Moreover, symmetry implies $\sigma^2 = \int_0^2 f(x)(1-x)^2 dx = \int_0^2 f(x)|1-x|^2 dx = 2\int_0^1 f(x)(1-x)^2 dx$, and therefore, $k_2 = 2$. Hence,

$$1 - \sigma^2/2 \geqslant E \geqslant 1 - \sigma/2 \quad \text{for symmetrical pdfs} \qquad (7)$$

(see Figure 2 for some examples). Inequality(7) makes sense only if the normalized standard deviation σ, which equals $(\int_0^2 f(x)(x-1)^2 dx)^{1/2}$ for symmetrical pdfs, ranges between 0 and 1. Indeed, this can be shown using the extreme types of symmetrical pdfs: If $f(x) \to 0$ for $x \neq 1$, we get $\sigma \to 0$, whereas if $f(x)$ has a U-form, with $f(x) \to 0$ for $x(x-2) \neq 0$, thus maximizing the distance of the random variable from the mean, we have $\sigma \to 1$ as $(1-x)^2$ equals either $(1-0)^2$ or $(1-2)^2$. For these two extremes, E is precisely determined by inequality(7), being 1 and 0.5, respectively. The relative error in estimating E by inequality (7), that is, by the mean of the limits $(1-\sigma/2)$ and $(1-\sigma^2/2)$, must be smaller than half of their difference divided by the lower limit, $0.5(\sigma/2-\sigma^2/2)/(1-\sigma/2)$. The maximum of the latter term is found at $\sigma = 2-\sqrt{2} \approx 0.59$ and is only 0.086. Analyzing realistic distributions (see below) yields relative errors even much smaller than that.

Among the pdfs that are symmetrical and unimodal (for example, bell-shaped), the pdf with the maximal σ is realized by an approximate rectangular distribution over the interval [0, 2] with $f(x) = 0.5$ for $0 \leqslant x \leqslant 2$, and $f(x) = 0$ otherwise. A simple calculation yields $\sigma = \sqrt{1/3} = 0.58$, $1-\sigma/2 = 0.71$, $1-\sigma^2/2 = 0.83$, $E = 0.75$, a relative error in estimating E by inequality (7) of $((1-\sigma/2+1-\sigma^2/2)/2-E)/E = 0.03$, and $e^{-\sigma} = 0.56$. More so than the rectangular, the triangular distribution might serve as a semi-realistic but still analytically treatable model of a symmetrical and unimodal pdf. For a triangular pdf with its base on the interval $[1-b, 1+b]$, $b \leqslant 1$, and, consequently, peak height of $1/b$, one gets $\sigma = b/\sqrt{6} \leqslant 0.41$, $e^{-\sigma} \geqslant 0.66$, $1-\sigma/2 \geqslant 0.80$ and $E = 1-b/6 = 1-\sigma/\sqrt{6} \geqslant 0.83$. Again, E can be predicted rather well by inequality(7) with a relative error of < 0.028. Of note, the range of E, that is, the interval [0.83, 1], is only half as large as the ranges of σ or $e^{-\sigma}$. Even more realistic, of course, than a triangular pdf is the assumption of a Gaussian normal distribution. The latter is reasonably symmetrical as long as the standard deviation is small with $\sigma \ll \mu$ (see Figure 2 and Supplementary Material F for the effect of truncation at $x = 0$). If the coverage results from a random production of reads as in a Poisson process, its distribution is approximately Gaussian with a variance before normalization that is as large as the mean coverage. Assuming a mean coverage of 100 before normalization, the standard deviation after normalization then is $\sqrt{100}/100 = 0.1$. Numerical integration using R (see Supplementary Material B) yields $E(\sigma)$ as 0.96, 0.92 and 0.84 for σ being 0.1, 0.2, and 0.41, respectively, which is almost the same as in case of the triangular distribution (Figure 2).

One might think that $E \geqslant 1-\sigma/2$ (see inequality(7)) also applies to all positively skewed normalized distributions, that is, normalized pdfs with positive third moment. However, this may not necessarily be the case. Defining $k_n = \int_0^{\infty} f(x)|1-x|^n dx / \int_0^1 f(x)|1-x|^n dx$ with $n \in \{0,1,2,3,\ldots\}$ we get k_0 and k_2 according to their definitions in the derivation of inequality(6), $k_1 = 2$ owing to the definition of the mean μ, which equals 1 after normalization and $k_3 > 2$ in case of positive skewness. For $E \geqslant 1-\sigma/2$ to be true, the product $k_0 k_2$ needs to be > 4 (see inequality(6)). Proofs in that matter are not trivial. For 'positively

slanted' distributions[4] (that is, pdfs for which $f(\mu+x)$–$f(\mu-x)$ is not identically zero and changes sign in $x>0$ at most once and from negative to positive, which include the Pearson family and the log-normal distribution) it can be derived, using the reasoning of MacGillivray,[5] that $k_2>2$ and $2>k_0>1$ (not shown). However, this is not very helpful. For the log-normal distribution, better approximations are derived in the following section.

The evenness score of the log-normal distribution

Measurements on biological entities usually are positive with a maximum at $x>0$ and a tail towards higher values. As such, their distributions resemble a log-normal distribution (see Limpert *et al.*[6] for a review). This type of distribution has been found in a great variety of cases, including gene expression,[7] telomere length,[8] neuronal activity,[9] fecundity[10] or time-to-event duration (for example, incubation time) of infectious and other diseases,[11,12] for instance, although log-normal genesis (multiplicative interaction of many random effects) cannot always be demonstrated perfectly. The pdf of the coverage in NGS may also have log-normal appearance (Figure 4): The rolling circle technique of Complete Genomics or the use of selective capturing of targets as in exome sequencing produce such distributions, whereas whole-genome sequencing with the Illumina technique results in rather symmetrical distributions.[13,14] Therefore, I examined the evenness score $E(\sigma)$ of the log-normal distribution (see Figure 3).

The log-normal distribution[15] is the density of a variable whose logarithm $\ln(x)$ has a Gaussian normal distribution $No(\mu^\star, \sigma^\star)$. Being the first derivative of the cdf with $\partial\ln(x)/\partial x = x^{-1}$, the log-normal pdf thus is

$$f(x) = \frac{1}{x\sigma^*\sqrt{2\pi}} e^{\frac{-(\ln(x)-\mu^*)^2}{2\sigma^{*2}}} \tag{8}$$

where $\mu^\star$ and $\sigma^\star$ now are form parameters only that relate to mean and variance of x as $\mu = e^{(\mu^*+\sigma^{*2}/2)}$ and $\sigma^2 = \mu^2\left(e^{(\sigma^{*2})}-1\right)$, respectively.[16] Normalization (division by the mean) conserves the log-normal form of a distribution, since $\ln(x/\mu) = \ln(x)$–$\ln(\mu)$ implies that $\ln(x/\mu)$ has the Gaussian normal distribution $No(\mu^*$–$\ln(\mu),\sigma^*)$ if the distribution of $\ln(x)$ is $No(\mu^\star,\sigma^\star)$. For normalized coverage with $\mu=1$, the relation of $\sigma^\star$ and σ simplifies to

$$\sigma^{*2} = \ln\left(\sigma^2+1\right) = -2\mu^*, \text{if}\, \mu = 1 \tag{9}$$

The log-normal distribution is increasingly skewed with increasing σ (see Figure 3a), whereas it approximates a Gaussian normal distribution $No(\mu, \sigma)$ if $\sigma\to 0$ (see Supplementary Material D for a proof of the latter tendency). In case of small σ, the evenness score of a normalized log-normal distribution thus can be estimated by inequality(7) (see Figures 3b and 4b). First-order approximation of Equation (9) in the vicinity of $\ln(1)$ yields $\sigma^{\star 2} = \ln(1+\sigma^2) \approx \sigma^2$, that is, $\sigma \approx \sigma^\star$, so that inequality(7) translates to $1-\sigma^{\star 2}/2 \geqslant E \geqslant 1-\sigma^\star 2$ for $\sigma \approx \sigma^\star \to 0$. With first-order approximation in the vicinity of e^0 as $e^{0+\Delta t} \approx 1+\Delta t$, this results in $e^{-\sigma^{*2}/2} \geqslant E \geqslant e^{-\sigma^*/2}$. Figures 3b and 4 and Supplementary Figure S2 show that $E \approx e^{-\sigma^*/2}$ also holds beyond the region of small σ. At $\sigma=1.3$ where $\sigma^\star=1$, the values of $E=0.62$ and $e^{-1/2}=0.61$ still are almost identical. Indeed, the approximation $e^{-\sigma^*/2}$ is valid up to $\sigma=3$ (that is, $\sigma^\star=1.5$), with a maximal absolute error of 0.02,

$$E \approx e^{\frac{-\sigma^*}{2}} \text{ for normalized log-normal distribution with } \sigma \leqslant 3 \tag{10}$$

To derive an approximation for even larger σ, use $y=\ln(x)$, which, by definition, has a Gaussian normal distribution $No(\mu^\star,\sigma^\star)$. Substituting x by y, that is, $F_{LogNo}(x)$ by $F_{No}(y)$ in Equation (3), yields $E = 1 - \int_{-\infty}^{0} F_{No}(y)e^y dy$, taking into consideration that $dx = \partial x/\partial y dy = e^y dy$. The value of $\int_{-\infty}^{0} F_{No}(y)e^y dy$ is determined by the region close to the origin as the factor e^y is approaching 0 for negative values of y beyond that region. For large σ (and, therefore, large $\sigma^\star$), $F_{No}(y)$ approximates its maximum ($=1$) in that region so that its graph becomes flat and rather linear, because its mean $\mu^\star$ moves away from the origin with the square of its standard deviation, $\mu^\star = \sigma^{\star 2}/2$, according to Equation (9). Hence, for large σ, $F_{No}(y)$ can be replaced by a low-grade Taylor series approximation. The Taylor series can be expressed as $F_{No}(0) + \sum_{n=1}^{\infty}(\partial^{n-1} f_{No}(0)/\partial x^{n-1}) y^n/n!$, considering that f_{No} is the first derivative of F_{No}. With $n! = \int_0^\infty y^n e^{-y}\, dy = \int_{-\infty}^{0} (-1)^n y^n e^y dy$, one gets $E = 1 - F_{No}(0) + \sum_{n=0}^{\infty}(\partial^n f_{No}\ (0)/\partial x^n)(-1)^n$ (see Supplementary Material E for a

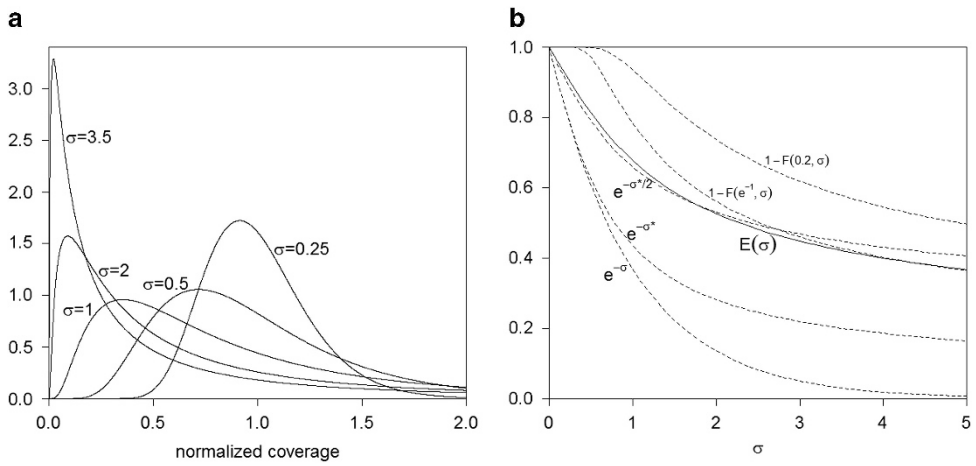

Figure 3 (**a**) Assumed log-normal probability density functions (pdfs) of the normalized coverage x for different values of the standard deviation σ. (See the main text for the relation between σ and the form parameter σ^* of the log-normal pdf.) As normalization means division by the mean here, the mean μ of x is always 1. Therefore, the coefficient of variation ($=\sigma/\mu$) equals σ. Note that for $\sigma\to 0$, the pdf approximates the form of a Gaussian normal distribution while with increasing σ the skewness also increases. (**b**) Evenness score $E(\sigma)$ and alternative scores ($e^{-\sigma}$, $e^{-\sigma^*}$,$e^{-\sigma^*/2}$ and $1-F(x,\sigma)$) for normalized log-normal distributions with varying σ. The cumulative distribution function $F(0.2,\sigma)$ quantifies the fraction or targets with a coverage of <0.2 (that is, $\leqslant 20\times$ if the mean of the non-normalized random variable is $100\times$) depending on σ. Note that $E(\sigma)$ is well approximated by $e^{-\sigma^*/2}$ up to large σ ($=3$) and by $1-F(e^{-1},\sigma)$ for very large σ ($\geqslant 2.5$).

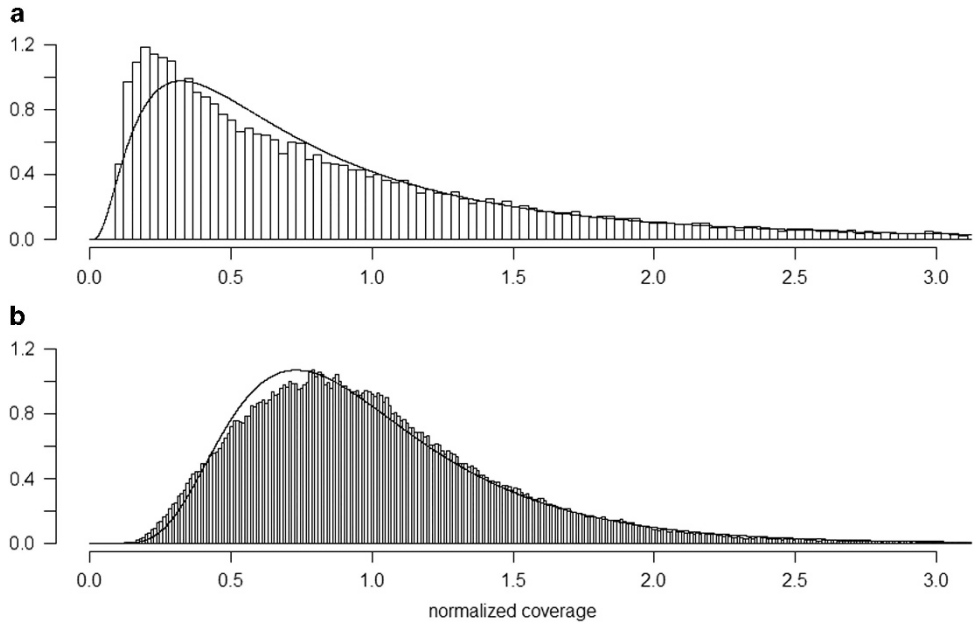

Figure 4 Fitting log-normal distributions to exome data. (**a**) Variant coverage distribution of an individual exome ('son') that can be downloaded from Glusman *et al.*[17] (**b**) Average coverage of each variant position on chromosome 18 that is covered above threshold ($>7\times$) in all 4300 European American samples of the Exome Variant Server (http://evs.gs.washington.edu/EVS, Jan 2016). For the moments and scores of the normalized coverage as discussed in the present article, distribution (**a**) yields $\sigma=1.006$, $e^{-\sigma}=0.366$, $E=0.657$, $1-\sigma/2=0.497$, $1-\sigma^2/2=0.494$, $e^{-\sigma^*/2}=0.658$ and $1-F(e^{-1})=0.724$, and (**b**) yields $\sigma=0.456$, $e^{-\sigma}=0.634$, $E=0.827$, $1-\sigma/2=0.772$, $1-\sigma^2/2=0.896$, $e^{-\sigma^*/2}=0.804$ and $1-F(e^{-1})=0.965$. Thus the evenness score E of realistic data is well approximated by $e^{-\sigma^*/2}$ as in Equation (10) while $1-F(e^{-1})$ as in Equation (11) is not sufficient yet. For (**b**) where the deviation from symmetry is relatively small, even inequality(7) yields a good approximation with $0.5(1-\sigma/2+1-\sigma^2/2)=0.834$. Moreover, panels (**a**) and (**b**) show that, for the characterization of realistic data, the score $e^{-\sigma}$ exploits a much larger part of its range than E. (See Supplementary Figure S2 for exome data that fit the log-normal distribution less perfectly while the relations of the moments and scores are still quite similar to here.)

detailed derivation). Stopping that series at $n=0$ yields $E\approx 1-(F_{No}(0)+f_{No}(0)(-1))$. The term in the brackets amounts to a first-order Taylor approximation of $F_{No}(y)$ for $y=-1$. Returning to the log-normal distribution of x with $x=e^y$ then results in $E\approx 1-F(e^{-1})$. As shown in Figure 3b, this approximation is rather good for very large values of σ. For $\sigma\geqslant 2.5$, its maximal absolute error is at most 0.02. Hence,

$$E\approx 1-F(e^{-1}) \text{ for normalized log-normal distribution with } \sigma\geqslant 2.5 \quad (11)$$

As $F(e^{-1})=F(0.38)$, which is the number of reads with a normalized coverage of at most 0.38, E here indicates the number of reads with a normalized coverage of at least 0.38. Figure 3b shows that $1-F(e^{-1})$ is quite parallel to $1-F(0.2)$; in case of a $100\times$ average coverage, which is frequently aimed for in NGS projects, $F(0.2)$ indicates the limit of $20\times$ that is usually considered as the minimal coverage necessary for reliable mutation detection. As such, for NGS projects with very large variance of the coverage, E may serve as useful and more or less direct indicator of the fraction of sufficiently covered targets.

DISCUSSION

The evenness score is used in NGS to quantify the homogeneity of target coverage with sequencing reads.[1,3] As such, it is a measure of the relative width of a distribution, with the coverage being the distributed variable. Its use can be recommended only if it has advantages compared with the coefficient of variation, which is the parameter conventionally applied for this purpose. Here I have performed that comparison. To do so, I used the evenness score in its continuous version, which assumes a normalized random variable (that is, having a mean μ of 1). Therefore, the evenness score E was compared with the standard deviation σ, as σ equals the coefficient of variation if μ equals 1.

At first, I clarified the mathematical definition of E and derived a computational more efficient version (see Equation (4)), which then was also translated to the non-normalized, discrete case of empirical coverage data (see Equations (5a and 5b)). Using this version, the calculation of E avoids double summations thus making it now about as fast as the calculation of σ. As most software applications still do not contain a built-in routine for the calculation of E, I have provided short R commands that will be easily translatable to analogous commands in other programming languages.

Besides the unconventionality of E, its definition might appear to imply another disadvantage: Since the integration $\int_0^1(1-x)f(x)dx$ in its calculation, runs only up to the mean ($=1$ owing to normalization), E might appear to be insensitive to the variable's distribution above the mean. However, this is not true because we see that $\int_0^\infty(1-x)f(x)dx=1-\mu=0$. Hence, by influencing the location of the mean (before normalization), the upper part of the distribution influences the upper end of the lower part and, thereby, the result of the integration of the normalized lower part.

More important is the outcome of the general analysis of E performed in the present paper. For any symmetrical distribution, including the Gaussian normal distribution, I showed that E can be predicted with little error by σ, that is, by $1-\sigma/2$ (see inequality(7) and Figure 2). Moreover, as some NGS methods entail positively skewed coverage data (see Figure 4, Ernani *et al.*[13] and Lam *et al.*[14]), I examined the evenness score of the log-normal distribution, which is the typical distribution of positively skewed results of biological measurements:[6] For a rather wide range of $\sigma(\leqslant 3)$, E was found to be predictable by $e^{-\sigma^*/2}$ with $\sigma^{*2}=\ln(\sigma^2+1)$ (see Equation (10) and

Figure 3b). In these cases, E also does not seem to provide much information that is not easily derivable from σ. An advantage of E was revealed only for cases with very large coefficient of variation (that is, σ of normalized data $\geqslant 2.5$), as it then satisfyingly and directly predicts the fraction of targets with sufficiently high coverage (see Equation (11) and Figure 3b), whereas this fraction cannot be easily estimated directly from σ.

Some might argue that the evenness score has the advantage of being a score between 0 and 1 (0% and 100%). However, a simple score with that quality can also be devised using σ, namely as $e^{-\sigma}$, which is 1 (that is, 100%) for absolutely homogeneous coverage and approaches 0 for inhomogeneous coverage. The major difference between E and $e^{-\sigma}$ is given by the rate of approaching 0 as can be seen in Figure 3b. There, E still indicates an evenness of $0.37 = 37\%$ if $F(0.2) = 0.5$ with 50% of the targets having a coverage of at most 0.2 (that is, of at most $20\times$ if the mean is $100\times$), while $e^{-\sigma}$ is already down to a level of $e^{-5} = 0.007 = 0.7\%$. If such NGS outputs were unacceptable due to insufficient coverage of too many targets, E would not exploit its full range (0–1) for the evaluation of the acceptable NGS outputs. Indeed, the minimal E values of published NGS outputs as calculated in Mokry *et al.*,[1] Lelieveld *et al.*[3] and this present paper are still as large as 0.62, 0.68 and 0.66, respectively, while $e^{-\sigma}$ goes down to 0.37 (see Figure 4). On the other hand, if outputs with 50% of the targets having a coverage of at most 20% of the mean coverage were acceptable, E would have the advantage to preserve some of its range for their quantitative evaluation.

Dealing with log-normal distributions, it might also be worth considering the analog of the standard deviation of a Gaussian normal distribution, that is, the 'multiplicative standard deviation' $\sigma^{\star}$ as recommended by Limpert *et al.*[6] (note that the naming of the variables is different in Limpert *et al.*[6]). It is one of the two form parameters in Equations (8 and 9). In case of empirical data, it can be calculated as the standard deviation of the natural logarithm of the random variable. In Figure 3b, the score $e^{-\sigma^{\star}}$ is presented as a possible tool for the quantitative evaluation of the homogeneity of NGS outputs. It may provide a compromise between $e^{-\sigma}$ and E. However, the 'multiplicative standard deviation' does not yet seem to be in common use and the NGS community may therefore hesitate to take it into consideration.

In summary, the general evaluation presented in this paper reveals that in most circumstances the evenness score E of a NGS output can be predicted quite well by the standard deviation σ of the normalized data (that is, by the coefficient of variation σ/μ in case of non-normalized data). Only if σ is very large ($\geqslant 2.5\mu$), does E have the advantage of directly reflecting the fraction of sufficiently covered targets. The general relation between E and σ set out here should also apply to other scientific fields that develop a parameter equivalent to E for their statistics.

CONFLICT OF INTEREST

The author declares no conflict of interest.

ACKNOWLEDGEMENTS

I thank Kay E Reed for inspiring talks and critical reading.

1 Mokry, M., Feitsma, H., Nijman, I. J., de Bruijn, E., van der Zaag, P. J., Guryev, V. *et al.* Accurate SNP and mutation detection by targeted custom microarray-based genomic enrichment of short-fragment sequencing libraries. *Nucleic Acids Res.* **38**, e116 (2010).
2 Gnirke, A., Melnikov, A., Maguire, J., Rogov, P., LeProust, E. M., Brockman, W. *et al.* Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. *Nat. Biotechnol.* **27**, 182–189 (2009).
3 Lelieveld, S. H., Spielmann, M., Mundlos, S., Veltman, J. A. & Gilissen, C. Comparison of exome and genome sequencing technologies for the complete capture of protein coding regions. *Hum. Mutat.* **36**, 815–822 (2015).
4 Rösler, U. Distributions slanted to the right. *Stat. Neerl.* **49**, 83–93 (1995).
5 MacGillivray, H. L. The mean, median, mode inequality and skewness for a class of densities. *Aust. J. Stat.* **23**, 247–250 (1981).
6 Limpert, E., Stahel, W. A. & Abbt, M. Log-normal distributions across the sciences: keys and clues. *BioScience* **51**, 341–352 (2001).
7 Bengtsson, M., Ståhlberg, A., Rorsman, P. & Kubista, M. Gene expression profiling in single cells from the pancreatic islets of Langerhans reveals lognormal distribution of mRNA levels. *Genome Res.* **15**, 1388–1392 (2005).
8 Oexle, K. Telomere length distribution and Southern blot analysis. *J .Theor. Biol.* **190**, 369–377 (1998).
9 Rupasov, V. I., Lebedev, M. A., Erlichman, J. S. & Linderman, M. Neuronal variability during handwriting: lognormal distribution. *PLoS ONE* **7**, e34759 (2012).
10 Herrera, C. M. & Jovani, R. Lognormal distribution of individual lifetime fecundity: insights from a 23-year study. *Ecology* **91**, 422–430 (2010).
11 Sartwell, P. E. The distribution of incubation periods of infectious disease. *Am. J. Hyg.* **51**, 310–318 (1950).
12 Horner, R. D. Age at onset of Alzheimer's disease: clue to the relative importance of etiologic factors? *Am. J. Epidemiol.* **126**, 409–414 (1987).
13 Ernani, F. P., LeProust, E. M. & Agilent Technologies. Target enrichment for NGS. *Euro Biotech. News* **8**, 42–44 (2009).
14 Lam, H. Y., Clark, M. J., Chen, R., Chen, R., Natsoulis, G., O'Huallachain, M. *et al.* Performance comparison of whole-genome sequencing platforms. *Nat. Biotechnol.* **30**, 78–82 (2012).
15 McAlister, D. The law of the geometric mean. *Proc. R. Soc.* **29**, 367–376 (1879).
16 Rinne, H. *Taschenbuch der Statistik* 3rd edn (eds Harri Deutsch, Frankfurt a.M.) 301–305 (Germany, 2003).
17 Glusman, G., Cariaso, M., Jimenez, R., Swan, D., Greshake, B., Bhak, J. *et al.* Low budget analysis of Direct-To-Consumer genomic testing familial data version1; referees: 2 approved. *F1000Research* **1**, 3 (2012).

Supplementary Information accompanies the paper on Journal of Human Genetics website (http://www.nature.com/jhg)