Supplementary Text, Supplementary Figures S1 and S2, and Supplementary Table S1: The Supplementary Text, Figures and Table provide further details on three issues related to mutant-allele tumor heterogeneity (MATH): 1, how MATH incorporates copy-number alteration (CNA) and corrects for normal DNA in a tumor; 2, alternate ways to handle CNA and genomic instability; 3, how CNA and patterns of mutation sharing among cell popoulations affect the center and the width of the distribution of mutant-allele fractions among loci.

Section 1. MATH, ploidy, and purity.

MATH implicitly includes CNA in its measure of intratumor heterogeneity, through the influence of CNA on mutant-allele fractions. As a ratio of the width to the center of the distribution of mutant-allele fractions, MATH corrects for normal cells ("impurity") present along with cancer cells in a tumor sample.

Copy-number alteration (CNA). To see how MATH incorporates CNA, consider the formula for the mutant-allele fraction at an autosomal locus in a heterogenous tumor with N genetically distinct cell populations. If m_{ij} is the number of mutant copies of locus i per cell in population j, c_{ij} is the corresponding total number of copies (mutant + reference) of the locus per cell in population j, and p_j is the fraction of all cells that are members of population j, then the mutant-allele fraction f_i at locus i is:

(Eq. 1)
$$f_i = \sum_{j=1}^{N} m_{ij} p_j \bigg/ \sum_{j=1}^{N} c_{ij} p_j = \frac{1}{2a_i} \sum_{j=1}^{N} m_{ij} p_j \ ,$$

where the sum is over all cell populations and a_i is the amplification of locus i (ratio to diploid) in the sample. Each mutant-allele fraction used to calculate MATH thus incorporates CNA, with the number of mutant copies in each cell scaled by the overall amplification of the locus in the tumor. (Alternatively, if there is information on locus amplification, each mutant-allele fraction could be multiplied by its amplification a_i, with a result equal to ½ of the mean mutant-copy number of the locus per cell. This approach is considered in Section 2.)

Tumor purity. MATH is a ratio of the width to the center of the distribution of mutant-allele fractions among tumor-specific mutated loci (100 * MAD/median). If the same correction for normal cells in a tumor appears in both the numerator and the denominator, the correction cancels in the ratio so that MATH in the whole tumor is the same as MATH in just the cancer cells. Correction for normal cells motivated use of this ratio rather than the width of the distribution alone as a measure of intratumor genetic heterogeneity.

A first-order correction for normal cells is to correct for cell numbers. If population N is normal cells, the multiplicative correction factor for the "impurity" provided by normal cells is $1/(1 - p_N)$ for each cancer-cell-population fraction and thus for all mutant-allele fractions (Eq. 1). This correction for cell numbers is identical for all loci and cancels in the calculation of MATH.

This correction for normal-cell numbers is also the correction for normal-cell DNA at loci without CNA. With a median of 92 mutated loci per tumor in the head and neck squamous cell carcinomas (HNSCC) analyzed by Stransky et al (1), most mutated loci were expected to be passenger rather than driver mutations and thus not expected to be subject to direct selection for genomic gain or loss. Consistent with this expectation, among 55 HNSCC with CNA data in Supplementary Table 11 of Stransky et al (1), more than 90% of mutated loci had amplifications within ± 0.5 $\log_2$ units of normal copy number. Thus for most loci the correction for normal-cell numbers is close to the correction for normal-cell DNA (see below). With MAD and median as robust measures of distribution width and center, not greatly influenced by small numbers of individual loci, their ratio will be predominantly determined by the large numbers of loci having minimal CNA. Thus their ratio makes MATH insensitive to the presence of normal cells.

A more detailed correction for normal tissue would be to correct each locus for its own normal-cell DNA. With population *N* taken as normal cells (population fraction p_N; $m_{iN} = 0$ and $c_{iN} = 2$ for all autosomal loci), multiplying the mutant-allele fraction of locus *i* by $a_i / (a_i - p_N)$ corrects for normal DNA, providing a cancer-DNA-specific mutant-allele fraction:

(Eq. 2)
$$\frac{a_i}{a_i - p_N} f_i = \frac{1}{2(a_i - p_N)} \sum_{j=1}^{N-1} m_{ij} p_j = \sum_{j=1}^{N-1} m_{ij} p_j \Big/ \sum_{j=1}^{N-1} c_{ij} p_j \ .$$

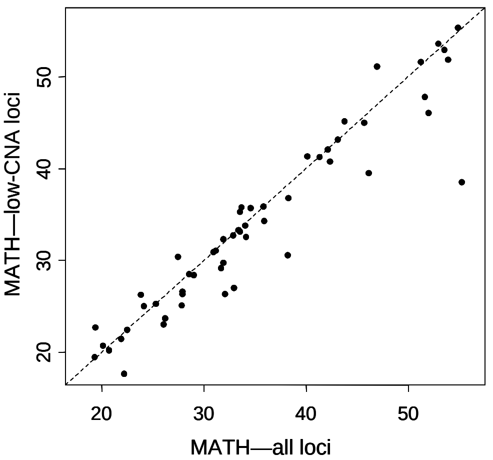
The correction for normal-cell DNA at most individual loci is very close to the general correction factor $1/(1-p_N)$ for normal-cell numbers. Based on the CNA data provided by Stransky et al, at a typical 20% normal-cell admixture the correction for normal-cell DNA would be within 10% of the correction for normal-cell numbers for 92% of loci. Even at the maximum acceptable 30% normal cells, the 2 corrections agree within 20% for 94% of loci.

To put the small differences between the 2 types of correction for normal tissue into perspective, note that binomial sampling error in determining mutant-allele fractions is generally 10% to 30%. At 100 sequence reads per locus, typical in these data, the coefficient of variation (CV) for mutant-allele fractions arising from binomial sampling of mutant versus reference alleles is 10%, 20% or 30% at mutant-allele fractions of 0.5, 0.2, or 0.1, respectively. The percentage difference between the 2 types of correction for normal tissue at a locus is almost always less than the percentage CV in measuring its mutant-allele fraction—at 20% normal tissue, that is the case for over 96% of loci in the 55 HNSCC with CNA data.

Section 2. Alternate ways to handle CNA; other measures of heterogeneity and genomic instability

CNA is included implicitly in MATH, because the observed mutant-allele fraction of each locus involves the ratio of the number of mutated copies per cell to the overall amplification of the locus (Eq. 1). This raised the issue of whether directly including available information on CNA might provide an alternate way to obtain an index of intratumor heterogeneity.

We addressed this possibility in two ways, using the subset of 55 HNSCC having locus-specific CNA data. First, we performed MATH calculations both before and after removing loci having high CNA (more than ± 0.5 $\log_2$ units away from normal copy number). The loci omitted were thus those having corrections for normal-cell DNA that were farthest from the correction for normal-cell number, as discussed in Section 1.



Supplementary Fig. S1. MATH after removing loci beyond ±0.5 $\log_2$ units of normal copy number, versus MATH for all loci. Dashed line is line of identity.

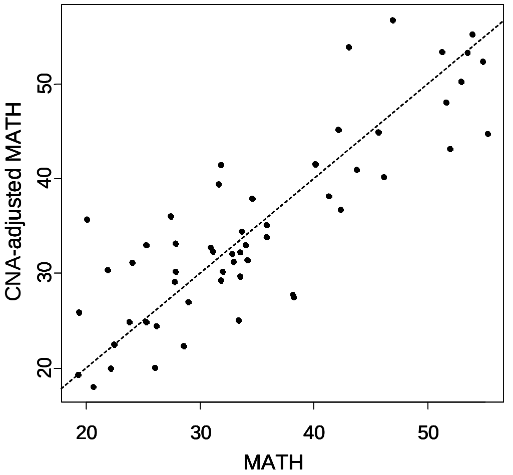
As shown in Supplementary Fig. S1, most MATH values based only on loci with low CNA were close to the values calculated for all mutated loci, typically within the range of resampling SDs of MATH values (Fig. 1D, main text). The tumors showing the larger discrepancies had the larger numbers of mutated loci outside these CNA limits.

Second, we instead used all loci but first multiplied the mutant-allele fraction of each locus by its amplification a_i to provide a CNA-adjusted mutant-allele fraction, and calculated

100 * (MAD/median) for the distribution of CNA-adjusted mutant-allele fractions for each tumor. Each CNA-adjusted mutant-allele fraction is then ½ of the average number of mutant copies of the locus per cell (cf. Eq. 1). In this case, the adjustment for normal tissue is $1/(1-p_N)$ for all loci, and the correction for normal tissue in the ratio MAD/median is exact.

These CNA-adjusted MATH values were similar to MATH values based directly on mutant-allele fractions (Supplementary Fig. S2). Thus neither of these two "corrections" of MATH for CNA (omitting high-CNA loci; adjusting for local amplification) has a major effect on MATH values, at least for these combinations of CNA, normal tissue, and mutant-allele fractions.

The issue still remained whether either of these "corrected" versions of MATH, or other measures of genomic instability or diversity, might perform better than MATH as a candidate biomarker. In the 55-case subset with CNA data, we looked at relations of MATH and 5 other potential measures of genomic diversity or instabilty to 3 clinically important HNSCC variables: disruptive TP53 mutations (versus all other TP53 status), HPV status (in wild-type TP53 cases), and pack-years (among HPV-negative cigarette smokers, taking disruptive TP53 into account). For each tumor, measures considered were: MATH as calculated in the main text; the number of mutated loci (a measure of overall mutation rate); number of genomic segments showing substantial CNA (segments longer than 1000 base pairs beyond ±0.5 $\log_2$ units from normal copy number); mean-square CNA per base (estimate of overall genomic copy-number diversity); MATH restricted to loci with low CNA (Supplementary Fig. S1); and "CNA-adjusted" MATH, based on mutant-allele fractions multiplied by locus amplification (Supplementary Fig. S2).



Supplementary Fig. S2. CNA-adjusted MATH, based on product of mutant-allele fraction and amplification for each locus, versus MATH. Dashed line is line of identity.

Supplementary Table 1. Relations of measures of genomic instability or intratumor heterogeneity to clinically important HNSCC variables, in cases having CNA data for mutated loci.

Variable examined	Number of cases	Test	*p*-value for relation of measure to variable					
			No. of mutated loci	CNA, segment numbers	CNA, mean-square	MATH, low CNA loci	CNA-adjusted MATH	MATH (main text)
Disruptive TP53 mutation	55	Wilcoxon rank-sum	0.26	0.13	0.002	0.056	0.020	0.024
HPV status in wild-type TP53	13 HPV- 7 HPV+	Wilcoxon rank-sum	0.20	0.59	0.94	0.037	0.024	0.024
Pack-years in HPV-negative cigarette smokers	41	*t*-test for pack-year coefficient in bivariate model with disruptive TP53	0.46*	0.40*	0.58*	0.015	0.062	0.049

*These 3 measures were log transformed for the bivariate model.

As shown in Supplementary Table 1, neither mutation rate nor the number of segments with substantial CNA was significantly related to these clinical variables in these 55 cases. Overall genomic copy-number variability (mean-square CNA) was related to disruptive TP53 mutation but not to HPV status or to pack-years. As might be expected from Supplementary Figs. S1 and S2, both types of "correction" of MATH for CNA provided relations to these 3 clinical variables close to those seen for MATH based solely on mutant-allele fractions; each of the "corrected" versions slightly missed significance at $p < 0.05$ with respect to one of the clinical variables.

For these data, none of the other measures performed better overall with respect to the 3 clinical variables than did MATH, as calculated in the main text directly from mutant-allele fractions (with its implicit inclusion of CNA). MATH calculated in this way does not require separate analysis of CNA or imputation of CNA from numbers of sequence reads, so it provides the most straightforward way for now to assess, from NGS results, a type of intratumor heterogeneity that appears to be clinically significant in HNSCC. Incorporation of information on CNA should be re-assessed in future work on the relations of MATH to outcome in HNSCC and other types of cancer, in larger data sets.

Section 3. The center and the width of a distribution of mutant-allele fractions.

To illustrate the principles of how mutation patterns among cell populations can affect the center and the width of a distribution of mutant-allele fractions, we first examine the mean and the standard deviation (SD) of the distribution, to take advantage of their analytical simplicity. We then explain why we used robust measures (median and MAD) for application to NGS results.

Mean mutant-allele fraction. Based on Eq. 1, the mean mutant-allele fraction over all *L* mutated loci in N cell populations, $\mathrm{mean}(f)$, is:

(Eq. 3)
$$\mathrm{mean}(f) = \frac{\sum_{i=1}^{L} \frac{1}{a_i} \sum_{j=1}^{N} m_{ij} p_j}{2L}.$$

The numerator is related to the average number of mutations per cell, with each locus *i* scaled by its overall amplification a_i. The denominator is 2 times the total number of mutated loci among all cancer cell populations.

Notably, if there are *N* cell populations, at least one population must have a cell fraction no greater than 1/*N*. In particular, heterozygous mutations (without CNA) specific to the smallest population will have mutant-allele fractions no greater than 1/(2*N*). Increasing numbers of cell populations thus can tend to shift the mean of the distribution toward lower values, although details depend on the specific patterns of mutation sharing among populations, locus amplification, and cell-population fractions. Even if all *N* populations are similar in size and do not share mutations, so that the width of the distribution is small (see below), the center of the distribution of mutant-allele fractions will thus tend to be lower than for a homogeneous tumor and the ratio of width to center will be higher.

SD of mutant-allele fractions. Mutation sharing among cell populations and differences among cell-population fractions in a tumor increase the SD of the distribution of mutant-allele fractions among loci. Use matrix notation for the (column) vector of mutant-allele fractions **F** formed from the individual locus values, f_i (Eq. 1):

(Eq. 4)
$$\mathbf{F} = \frac{1}{2}\mathbf{Diag(1/a)MP}$$

where **P** is the (N x 1) vector of cell-population fractions, **M** is the (L x N) mutation-number matrix (m_{ij} =mutant copies of locus i per cell in population j), and **Diag(1/a)** is a diagonal (L x L) matrix with reciprocals of the locus amplifications along the diagonal. Then the variance (square of the SD) of mutant-allele fractions among loci is:

(Eq. 5) $$\mathrm{var}(f) = \frac{\sum_{i=1}^{L} f_i^2}{L} - (\mathrm{mean}(f))^2 = \frac{\mathbf{F}^{\mathbf{T}}\mathbf{F}}{L} - (\mathrm{mean}(f))^2 = \frac{\mathbf{P}^{\mathbf{T}}(\mathbf{M}^{\mathbf{T}}\mathbf{Diag(1/a^2)M})\mathbf{P}}{4L} - (\mathrm{mean}(f))^2$$

(the superscript ^{T} represents the transpose). The matrix product $\mathbf{M^T Diag(1/a^2)M}$ is an (N x N) matrix that represents the pattern of mutation sharing among the N cell populations. Element j,k of $\mathbf{M^T Diag(1/a^2)M}$ is a weighted sum of mutations shared between cell populations j and k, with locus i weighted by $(m_{ij}\, m_{ik})/(a_i)^2$. With heterozygous mutations and without CNA, element j,k of $\mathbf{M^T Diag(1/a^2)M}$ is simply the total number of mutations shared by populations j and k.

Thus for a tumor having a given mean mutant-allele fraction $\mathrm{mean}(f)$, the distribution of mutant-allele fractions among loci can be wide due to mutation sharing among cell populations (non-zero off-diagonal elements of $\mathbf{M^T Diag(1/a^2)M}$) or variation among cell-population fractions (even in the unlikely event that no populations share any mutations). Insofar as a larger number of cell populations lowers $\mathrm{mean}(f)$, the SD is also increased.

Robust measures of center and width. We used robust measures of the center and the width of each tumor's distribution of mutant-allele fractions, the median and the median absolute deviation (MAD), rather than the mean and the SD. We made this choice to minimize the influence of the small numbers of mutated loci that had very high mutant-allele fractions. About 5% of loci in the data of Stransky et al (1) had mutant-allele fractions greater than ½, versus a median mutant-allele fraction of 0.21 and a mean of 0.25. Many loci with such high mutant-allele fractions represent mutations that are present in almost all cells of a tumor with CNA favoring the mutant allele. Among the 55 HNSCC with CNA data, over 20% of these high-mutant-allele loci had copy numbers beyond ± 0.5 $\log_2$ units of normal, with correspondingly high differences between the corrections for normal-cell number and for normal-cell DNA (Section 1). Such loci widen the distribution of mutant-allele fractions even for a homogeneous tumor. Furthermore, the root-mean-square calculation for SD would highly weight these few loci with high mutant-allele fractions, potentially masking heterogeneity arising from small cell populations.

The MAD, in contrast, is based on the half of loci closest to the median mutant-allele fraction, so the exact values both of the loci with the highest mutant-allele fractions and of the loci with the lowest fractions (where binomial sampling error of mutant-allele fractions is greatest) do not matter. Corrections for normal-cell numbers appear identically in MAD and median values, canceling in their ratio. The MAD and median, and their ratio used to calculate MATH, thus incorporate information about the existence of loci having high or low mutant-allele fractions, without being unduly influenced by the specific values of the outlier loci or the presence of normal cells in a tumor.

Reference for Supplementary Material

1. Stransky N, Egloff AM, Tward AD, Kostic AD, Cibulskis K, Sivachenko A, et al. The mutational landscape of head and neck squamous cell carcinoma. Science 2011; 333*:* 1157-60.