



Addendum to the  
*Cancer Sequencing Service*  
*Data File Formats*

## Addition of Fisher's Exact Test-derived Score (CGA\_SOMFE) to *somaticVcfBeta*

August 2012

In addition to the somatic score described in the *Cancer Sequencing Service Data File Formats* document, we now provide a score derived from a statistical test—Fisher's Exact Test (FET)—that states the likelihood that the variant is somatic given the observed read counts in the tumor and normal samples. The score is transformed from the fractional  $p$ -value result of the test to a Phred-scaled integer. This score is present only for variants with the "SS=Somatic" tag, and is named "CGA\_SOMFE."

Fisher's Exact Test is used to determine if there are non-random associations between two categorical variables. For a sampling of DNB sequencing reads at a given variant locus, the variables being tested are sample type (tumor or normal) and allele (reference or variant). In the case of a somatic call, a lower  $p$ -value indicates that the DNB counts are less likely to be random or more likely to represent a somatic variant. Alternatively, a higher  $p$ -value would correspond to a germline variant or maybe a false positive call in the tumor, where the DNB counts supporting reference and variant seem to be randomly assigned to the tumor or normal.

Consider this example of tumor and normal DNB counts at a somatic variant locus, represented by the following 2x2 contingency table:

	Reference	Variant	Total
Tumor	16	12	28
Normal	14	4	18
Total	30	16	46

Given each cell in the table, we can determine the expected count of DNBs. For example, the expected number of variant supporting DNBs in the tumor is equal to the total number of tumor DNBs (28) multiplied by the observed frequency of the variant allele (16/46). The observed count of tumor DNBs supporting the variant allele is greater than what is expected:

$$28 \times 16 / 46 = 9.74$$

Complete Genomics data is for Research Use Only and not for use in the treatment or diagnosis of any human subject. Information, descriptions and specifications in this publication are subject to change without notice.

The cells can be represented above with  $a$ ,  $b$ ,  $c$ , and  $d$  where  $n$  represents the total number of reads.

	Reference	Variant	Total
Tumor	$a$	$b$	$a + b$
Normal	$c$	$d$	$c + d$
Total	$a + c$	$b + d$	$n = a + b + c + d$

Fisher showed that the probability of obtaining any such set of values was given by the hypergeometric distribution:

$$p = \frac{\binom{a+b}{a} \binom{c+d}{c}}{\binom{n}{a+c}} = \frac{(a+b)! (c+d)! (a+c)! (b+d)!}{a! b! c! d! n!}$$

The observed count (12) of DNBs supporting the variant is greater than the expected count (9.74), the one-sided  $p$ -value is calculated by determining the set of possibilities that are at least as extreme as the observed and summing up the hypergeometric probabilities, holding the marginal totals ( $a + b$ ,  $a + c$ ,  $b + d$ , and  $c + d$ ) constant.

In our example, the one-sided  $p$ -value that an equal or more extreme count of DNBs supporting the variant in the tumor are selected at random is:

$$p(12) + p(13) + p(14) + p(15) + p(16) = 0.1316$$

We then convert this  $p$ -value to a Phred-like score, CGA\_SOMFE, rounding to the nearest integer, using the formula:

$$\text{CGA\_SOMFE} = -10 * \log_{10}(p - \text{value}) = -10 * \log_{10}(0.1316) = 9$$