Supplemental *Materials and Methods* for:

"Collections of simultaneously altered genes as biomarkers of cancer cell drug response"

David L. Masica & Rachel Karchin

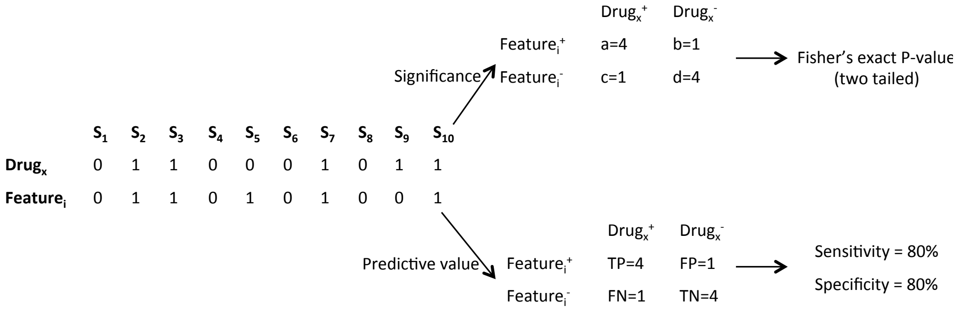
The algorithm presented here, now called MOCA (Multivariate Organization of Combinatorial Alterations), is an improvement on a method we developed for finding correlated alterations in genomics data, including the ability to correlate genomic alteration with phenotype (e.g., drug response or progression-free survival)(1). MOCA has no built-in limitations on the number of data types it can handle in a single execution, nor are there limits on the types of data (e.g., methylation, mutation, drug response) MOCA can handle. Algorithmic flexibility with respect to data type is possible because all data types are ultimately converted to binary representations (see below), and therefore there are no restrictions concerning the comparison of continuous data types (e.g., expression or methylation) with binary data types such as mutation.

MOCA takes, as input, feature-by-sample data matrices; for this study, features were genes, tissues, or drugs, and samples were cell lines. For any data type that is not binary, MOCA discretizes the data by converting every row to a feature-specific vector of Z-scores, and applying a cutoff to define each element in that vector as altered or not; in this study, a Z-score > 0.8 was used to define positive cases (overexpressed, drug sensitive, etc.), and a Z-score < -0.8 was used to define negative cases (underexpressed, copy number deleted, etc.).

Data:

For this work we utilized all data types available as part of the Cancer Cell Line Encyclopedia (CCLE; http://www.broadinstitute.org/ccle/home), this included: 18,107 genes with expression data, 23,124 genes with CNA (copy-number alteration) data, 1,644 genes with mutation data, and 24 drugs with drug response data. There were 416 cell lines common to all data types, which were used in this study. Because MOCA discretizes all continuous data types, there are twice as many features as there are genes for CNA and expression data (e.g., one over- and one under-expression matrix). We used mutation data with three distinct representations: 1) 1,644 gene-specific mutations. 2) 51,829 mutation-specific mutation features (e.g., *TP53* H193R is a unique feature). 3) 346 drug-response-optimized mutation features (see below). The 416 cancer cell lines comprised 18 distinct tissues, which were also features in all calculations. Thus, a total of 136,323 features were considered in this study. For all calculations, except for creation of drug-response-optimized mutation features (see below), feature vectors with fewer than three "true" ("1") values were filtered; for creation of drug-response-optimized mutation features this filter was not applied. Of the formats available for the drug response data, we selected IC_{50} values, which we converted to $-\log_{10}(IC_{50})$ values prior to calculation.

Calculations Performed by MOCA:



	S_1	S_2	S_3	S_4	S_5	S_6	S_7	S_8	S_9	S_{10}
$Drug_x$	0	1	1	0	0	0	1	0	1	1
$Feature_i$	0	1	1	0	1	0	1	0	0	1

Significance

	$Drug_x^+$	$Drug_x^-$
$Feature_i^+$	a=4	b=1
$Feature_i^-$	c=1	d=4

→ Fisher's exact P-value (two tailed)

Predictive value

	$Drug_x^+$	$Drug_x^-$
$Feature_i^+$	TP=4	FP=1
$Feature_i^-$	FN=1	TN=4

→ Sensitivity = 80%
Specificity = 80%

Figure 1S: Example of P-Value, statistical sensitivity, and specificity calculations used by MOCA.

Significance of drug-feature correlations is assessed by populating a two-by-two contingency table with the binary drug and feature vectors and calculating Fisher's exact, two-tailed P-value. In the worked example (Figure 1S), a vector element of "1" indicates either response to $Drug_x$ or alteration in genetic $Feature_i$, for samples S_1-S_{10}. All P-values are corrected using the Benjamini and Hochberg false discovery rate (FDR).

Using the same two-by-two contingency table, MOCA computes the statistical sensitivity and specificity of $Feature_i$ for response to $Drug_x$ (Figure 1S). For example, if a particular sample *S* is sensitive to $Drug_x$ and altered in $Feature_i$, that is considered a true positive (TP). Similarly, the coincidence of response to $Drug_x$ and wild-type $Feature_i$ is a false negative (FN). Samples not drug sensitive, but altered in the feature under comparison, constitute a false positive (FP). And, samples not drug sensitive and not altered in the feature under consideration represent a true negative (TN). From these values, MOCA computes the statistical sensitivity (not to be confused with drug sensitivity) as TP/(TP + FN) and the specificity as TN/(TN + FP).

Figure 2S is a worked example of how MOCA applies the union, intersection, and difference Boolean set operations to combine multiple genomic features. The union operation is equivalent to an `or` statement. An element in a binary feature vector representing the union of $Feature_i$ and $Feature_j$ will be "1" if $Feature_i$ `or` $Feature_j$ are altered in the corresponding sample, otherwise the element is "0" (Figure 2SA). The intersection operation is equivalent to an `and` statement. An element in a binary feature vector representing the intersection of *Feature*$_i$ and *Feature*$_j$ will be "1" if *Feature*$_i$ `and` *Feature*$_j$ are altered in the corresponding sample, otherwise the element is "0" (Figure 2SB). The difference operation is equivalent to a `not` statement. An element in a binary feature vector representing the difference of *Feature*$_i$ and *Feature*$_j$ will be "1" if *Feature*$_i$, but `not` *Feature*$_j$ is altered in the corresponding sample; otherwise, that element is "0" (Figure 2SC). There is no limit of the number of features that can be combined into a single feature using any of these three operations.

A

	S_1	S_2	S_3	S_4	S_5	S_6	S_7	S_8	S_9	S_{10}
Feature$_i$	0	1	1	0	0	0	1	0	1	1
Feature$_j$	0	1	1	0	1	0	1	0	0	1
Union	0	1	1	0	1	0	1	0	1	1

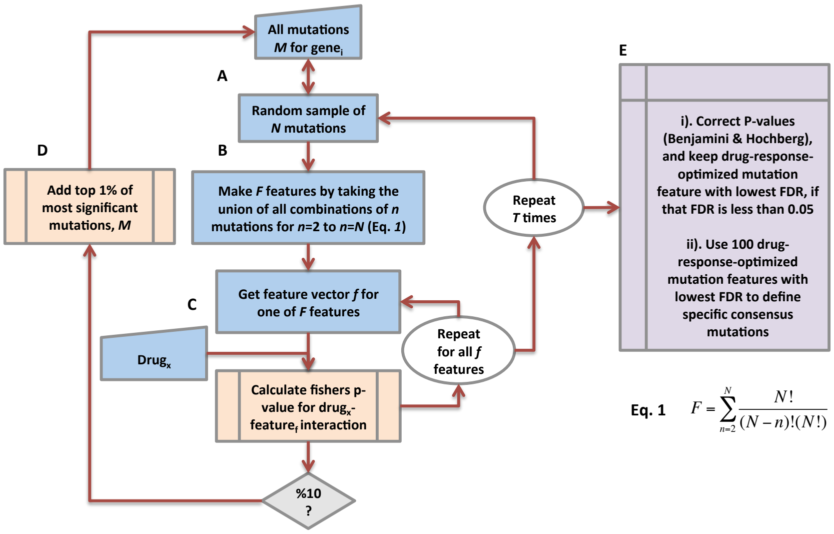
B

	S_1	S_2	S_3	S_4	S_5	S_6	S_7	S_8	S_9	S_{10}
Feature$_i$	0	1	1	0	0	0	1	0	1	1
Feature$_j$	0	1	1	0	1	0	1	0	0	1
Intersection	0	1	1	0	0	0	1	0	0	1

C

	S_1	S_2	S_3	S_4	S_5	S_6	S_7	S_8	S_9	S_{10}
Feature$_i$	0	1	1	0	0	0	1	0	1	1
Feature$_j$	0	1	1	0	1	0	1	0	0	1
Difference	0	0	0	0	0	0	0	0	1	0

Figure 2S: Combining genomic features using the union, intersection, and difference Boolean set operations.



All mutations
M for gene$_i$
A
Random sample of
N mutations
B
Make F features by taking the union of all combinations of n mutations for n=2 to n=N (Eq. 1)
C
Get feature vector f for one of F features
Drug$_x$
Calculate fishers p-value for drug$_x$-feature$_f$ interaction
%10
?
Repeat
for all f
features
Repeat
T times
D
Add top 1% of most significant mutations, M
E
i). Correct P-values (Benjamini & Hochberg), and keep drug-response-optimized mutation feature with lowest FDR, if that FDR is less than 0.05
ii). Use 100 drug-response-optimized mutation features with lowest FDR to define specific consensus mutations
Eq. 1 $F = \sum_{n=2}^{N} \frac{N!}{(N-n)!(N!)}$

Figure S3: Algorithmic flowchart for the creation of a drug-response-optimized mutation feature.

Creation of Drug-Response-Optimized Mutation Features:

For a specific drug-gene combination, creation of a drug-response-optimized mutation feature begins by taking a random collection of *N* mutation-specific mutation features (Figure S3A). Each unique combination of these *N* mutation-specific mutation features is combined into a distinct feature using the union operation (Figure S3B). Next, Fisher's exact P-value is calculated for the interaction of each one of these distinct features and the corresponding drug (Figure S3C). This process, of taking *N* mutations and comparing the union of all possible constituent combinations with drug response, is repeated *T* times; every 10th time, mutations belonging to the most significant 1% of combined features were appended to the list of the mutation-specific mutation features (Figure S3D). Therefore, as the algorithm progresses, the feature combinations become increasingly optimized for correlation with drug response. Finally, all P-values are converted to Benjamini & Hochberg FDRs, and the drug-response-optimized mutation feature with the lowest FDR is used for subsequent MOCA calculations, provided that FDR is < 0.05 (Figure S3E). If no drug-response-optimized mutation feature has an FDR < 0.05, that drug-gene combination is rejected entirely. For analysis purposes, we also consider the frequency that each specific mutation occurs in the 100 most significant optimized mutation features (see for instance, Figure 2A and 2B in the main manuscript); we define these as the *consensus mutations*, throughout.

For this work, *N* was set to 10, resulting in 1,013 comparisons per random sampling (see Figure 3S, *Eq. 1*). *T* scales proportional to the number of mutation-specific mutation features (*X*) available for a specific gene, using the equation *T* = 10*(*X* – 10) *for X < 110, else T = 1000*. This choice resulted in reasonable computational efficiency (see below) and the ability to recover the same drug-response-optimized mutation features over multiple trials (i.e., algorithmic convergence).

To filter genes that could be optimized below the FDR threshold by chance, a modified version of the above-described process was initially performed on randomized drug data. In Figure S3, *T* was set to 25, and every 5th time mutations belonging to the most significant 1% of combined features were appended to the list of the mutation-specific mutation features. This permutation protocol was repeated 100 times for every drug-gene combination; for each of these 100 trials, the drug under comparison was permuted to a new configuration ($Drug_x$ in Figure S3). During the 100 permutation trials, if any of the resulting drug-response-optimized mutation features achieved an FDR < 0.05, that drug-gene combination was removed from the real (i.e., non-randomized) data. This process was designed to be conservative, and of the potential 39,456 drug-response-optimized mutation features (i.e., 1,644 genes for each of 24 drugs), only 346 (< 1%) were significant and considered for subsequent analysis. The entire process for creating drug-response-optimized mutation features took ~4.5 hours per drug, on a single processor core. We include all 346 drug-response-optimized mutation features and corresponding significance, statistical sensitivity, and specificity in the accompanying supplemental data file.

Pairwise Calculations:

Next, we used MOCA to compute every pairwise drug-feature interaction. Of the 869,136 expression feature-drug interactions, 37,466 (4.0%) had an FDR < 0.05. Only ~0.03% (321) of the potential 1,109,952 CNA-drug correlations had an FDR < 0.05. Four (0.01%) of the possible 39,456 interactions between gene-specific mutation features and drug response were significant. Of the 1,243,896 interactions of drug response with mutation-specific mutation features, only six (< 0.0005%) were significant. And, 7.2% (31) of the potential 432 drug-tissue pairs were significantly correlated with drug response. MOCA can compare ~2,000 interactions per, and a single processor core. The accompanying supplemental data includes all significant pairwise interactions.

Target Selection:

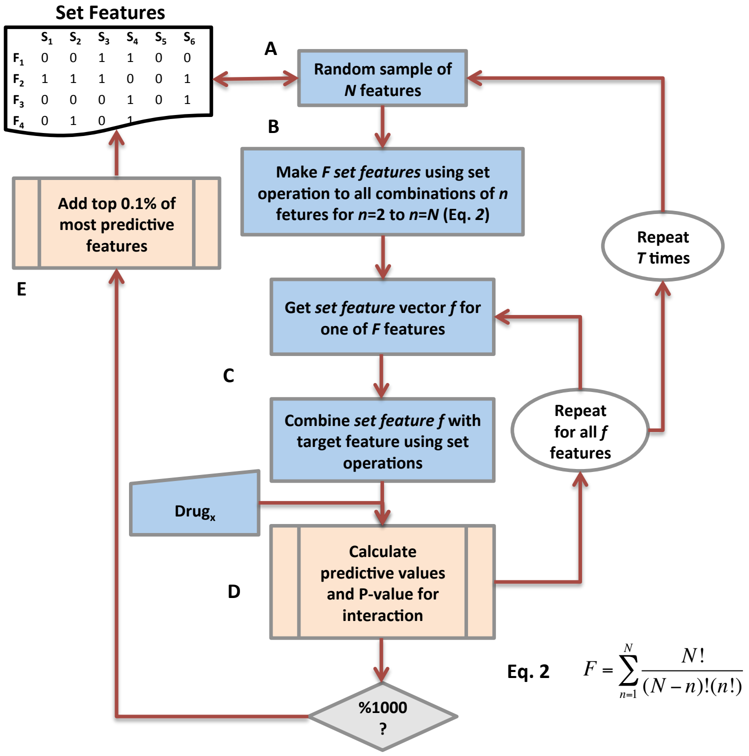
Next, we selected a single *target* feature of drug response for each drug. A feature was considered a target of drug response if it met the following two conditions: 1) the feature is the target, or part of the pathway, that the drug was designed to inhibit. 2) The feature was significantly correlated with response to that drug in the previous pairwise comparisons. Three drugs were not considered for this step because data for the target fusion gene was not publicly available at the time of this study. Of the remaining 21 drugs, 11 drugs had targets that met conditions 1 and 2 (see Table S1).

Drug	**Target**
AEW451	IGF1^{Exp+}
AZD6244	NRAS^{OptMut}
Erlotinib	EGFR^{Exp+}
Lapatinib	ERBB2^{Exp+}
LBW242	NFKB2^{OptMut}
Nutlin-3	MDM2^{Exp+}
Panobinostat	HDAC1^{OptMut}
PD-0325901	KRAS^{OptMut}
PLX4720	BRAF^{V600E}
RAF265	BRAF^{V600E}
ZD-6474	EGFR^{OptMut}

Table S1: Target features for each of 11 drugs. Superscripts delineate the alteration type used for the *Target* feature of response for the corresponding *Drug* (*Exp+* is overexpression, *OptMut* is a drug-response-optimized mutation feature (see above and Figure S3), and *V600E* is a specific BRAF amino-acid substitution).

Creation of Set Features:

Next, we used MOCA to make lists of features that, when combined with a target feature using a set operation, increased the significance of the interaction by two orders of magnitude, relative to the target alone. For instance, the erlotinib-EGFR^{Exp+} (see Table S1) interaction has a Fisher's exact P-value of $3x10^{-7}$. Therefore, if the interaction of erlotinib with the union of EGFR^{Exp+} and feature *f* had a P-Value < $3x10^{-9}$, feature *f* was added to the list of union features for the erlotinib-EGFR^{Exp+} interaction. This process was carried out for every feature for each interaction in Table 1S, considering the union, intersection, and difference operation for every comparison. Importantly, this calculation was not restricted to features determined significant in the previous pairwise calculations; it is this choice that allowed our method to capture important interactions that would be missed if pairwise significance was required to construct many-gene features of drug sensitivity. This step took ~30 minutes for each of 24 drugs, on a single processor core.



Set Features

	S_1	S_2	S_3	S_4	S_5	S_6
F_1	0	0	1	1	0	0
F_2	1	1	1	0	0	1
F_3	0	0	0	1	0	1
F_4	0	1	0	1		

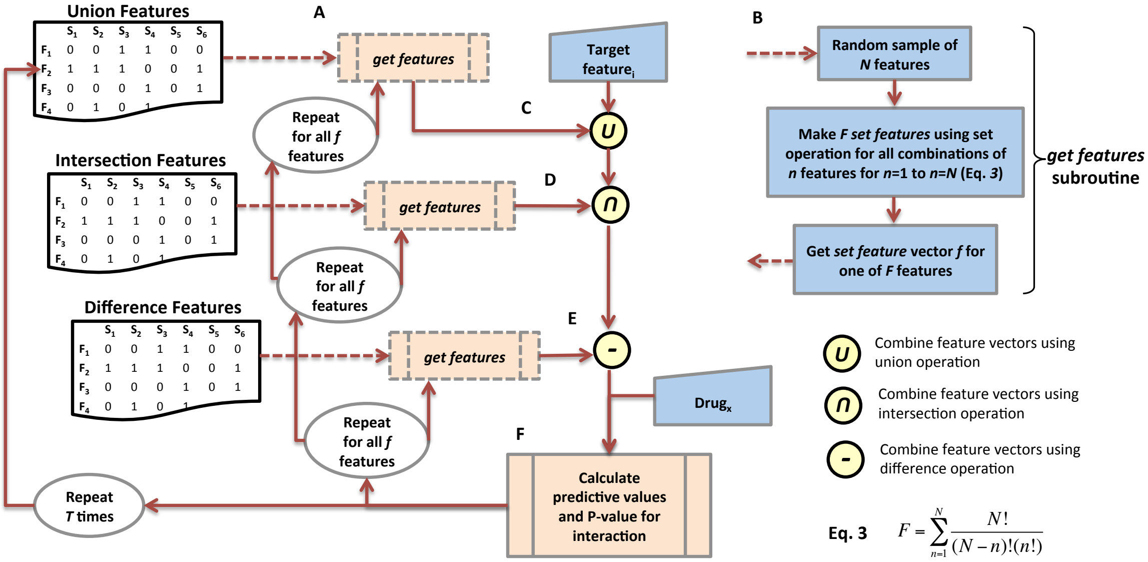
A
Random sample of *N* features
B
Make *F set features* using set operation to all combinations of *n* fetures for *n*=2 to *n*=*N* (Eq. *2*)
Repeat *T* times
Add top 0.1% of most predictive features
E
Get *set feature* vector *f* for one of *F* features
C
Combine *set feature f* with target feature using set operations
Repeat for all *f* features
Drug$_x$
Calculate predictive values and P-value for interaction
D
%1000 ?
Eq. 2

$$F = \sum_{n=1}^{N} \frac{N!}{(N-n)!(n!)}$$

Figure S4: Enriching the set-feature lists for features with the highest correlation with drug response.

Then, we used an optimization protocol to enrich the *set-feature* lists created in the previous step with features that, when combined, further improved correlation with drug response. This process begins by selecting a random collection of *N* features from one of three set-feature lists (i.e., union, intersection, or difference), for a particular target-drug interaction (Figure S4A). Each unique combination of these features is combined into a distinct feature using the set operation relevant to the set list being enriched (Figure S4B). Next, each of these distinct features is combined with the target feature using the relevant set operation (Figure S4C), and the Fisher's exact P-value is calculated for the interaction of that combined feature and the corresponding drug (Figure S4D). This process, of taking *N* features, combining all possible feature combinations with the target feature, and comparing the combined feature with drug response is repeated *T* times; every 1000th time, features belonging to the most correlated 0.1% of combined features were appended to the corresponding set-interaction list (Figure S4E). Therefore, as the algorithm progresses, a given set-interaction list becomes increasingly populated with features that combine to optimize correlation with drug response. For this work, *T* was set to 10,000 and *N* was 5, resulting in 3.1×10^5 combinations tested, per set operation, per target feature-drug interaction. Additionally, we enriched set-feature lists for features that contributed to correlations with high statistical sensitivity, specificity, or the sum of both.

Creation of Optimized, Many-Gene Features:



Union Features

	S_1	S_2	S_3	S_4	S_5	S_6
F_1	0	0	1	1	0	0
F_2	1	1	1	0	0	1
F_3	0	0	0	1	0	1
F_4	0	1	0	1		

Intersection Features

	S_1	S_2	S_3	S_4	S_5	S_6
F_1	0	0	1	1	0	0
F_2	1	1	1	0	0	1
F_3	0	0	0	1	0	1
F_4	0	1	0	1		

Difference Features

	S_1	S_2	S_3	S_4	S_5	S_6
F_1	0	0	1	1	0	0
F_2	1	1	1	0	0	1
F_3	0	0	0	1	0	1
F_4	0	1	0	1		

A
get features
Repeat for all *f* features
get features
Repeat for all *f* features
get features
Repeat for all *f* features
Target feature$_i$
C
U
D
∩
E
-
Drug$_x$
F
Calculate predictive values and P-value for interaction
Repeat *T* times
B
Random sample of *N* features
Make *F set features* using set operation for all combinations of *n* features for *n*=1 to *n*=*N* (Eq. *3*)
Get *set feature* vector *f* for one of *F* features
get features subroutine
U Combine feature vectors using union operation
∩ Combine feature vectors using intersection operation
- Combine feature vectors using difference operation
Eq. 3

$$F = \sum_{n=1}^{N} \frac{N!}{(N-n)!(n!)}$$

Figure S5: Combining the enriched set-feature lists into many-gene features highly correlated with drug response.

Next, we create many-gene features by combining features from each enriched set-feature list, for each target feature-drug interaction (Figure S5). First, MOCA selects a random sample of *N* features from the enriched union-feature list (Figure S5A). Every possible combination of these *N* features is combined using the union operation (Figure S5B); one at a time, the target feature is added to one of these combined features, using the union operation, and passed along to the next step (Figure S5C). Next, this process is repeated for the intersection-feature list, and one at a time, the combined intersection features are added to the previous union feature, using the intersection operation, and passed along to the next loop of the algorithm (Figure S5D). Then, the same process is repeated for the difference-feature list, and one at a time, the combined difference features are added to the previous intersection feature, using the difference operation (Figure S5E), and the P-value, statistical sensitivity, and specificity of the resulting many-gene feature is calculated with respect the corresponding drug (Figure S5F). Because the three lists (i.e., union, intersection, and difference) are sampled within nested `for` loops, every possible combination of all selected features is tested to find those with highest statistical sensitivity and specificity. For this study, *N* was set to 5, 4, and 3 for sampling from the union, intersection, and difference lists, respectively. Therefore, for each of 1,00 executions of *T*, 6,000 combinations were tested, resulting in a total of ~$6.0\text{x}10^5$ many-gene features tested for each drug. As in previous steps, all correlations were subject to multiple testing correction, and were required to have an FDR < 0.05 for subsequent analysis (step not shown in Figure 5S). See the last three rows of Table 1, in the main text, for an example of the output from this protocol. The combined processes illustrated in Figure's S4 and S5 took ~15 minutes, for each of 11 drugs, on a single processor core.

Finally, we assessed the potential utility of many-gene markers for blind prediction of drug response. The 416 CCLE cell lines were randomly divided (using the python random number generator) into training and testing datasets of 80% and 20%, respectively. Testing and training datasets were visually inspected to assure each contained a representative distribution of tissue types. First, training data was used to select single- and multi-feature predictors of drug response using the training data. Features were only selected for subsequent testing if their Benjamini and Hochberg FDR-corrected Fisher's P-value was less than 0.05. Many-gene features were derived as above, and it was therefore additionally required that gene-target-drug interactions had P-values at least two orders of magnitude more significant than the corresponding drug-target interaction alone. Biomarkers selected during the training phase were then tested using the testing data, and the resulting predictive value was ranked as the sum of statistical sensitivity and specificity.

Other Details:

MOCA is written in python. Fisher's exact test is computed using the fast *fisher* module for python (v0.1.4; http://pypi.python.org/pypi/fisher/). Benjamini and Hochberg FDRs are computed using the R *p.adjust* module, which interfaces with MOCA via RPy2.

Structural models were rendered with PyMol (2). PDB accession codes for crystal structures used in the structural models are as follows: 3BBT for lapatinib-bound ERBB4; 1XKK for lapatinib-bound EGFR; 1M17 for erlotinib-bound EGFR; 1RV1 for nutlin-2-bound MDM2; 3V3B for p53-bound MDM2; 3FE7 for p53-bound MDM4. Heatmaps were rendered using the *heatmap2* module in R.

Calculations were performed either on a Dell workstation with a Quad-Core Xeon processor or an IBM iDataPlex cluster with 500 Quad Core Intel Xeon E5472 processors (2,000 processor cores).

1. **Masica DL & Karchin R (2011) Correlation of Somatic Mutation and Expression Identifies Genes Important in Human Glioblastoma Progression and Survival. *Cancer Research* 71(13):4550-4561.**
2. **DeLano WL (2002) The PyMOL molecular graphics system.**