Searching for Gene Sets with Mutually Exclusive Mutations

Paul Ginzberg[*] Federico Giorgi[†] Andrea Califano[†]

Abstract
Cancer cells evolve through random somatic mutations. "Beneficial" mutations which disrupt key pathways (e.g. cell cycle regulation) are subject to natural selection. Multiple mutations may lead to the same "beneficial" effect, in which case there is no selective advantage to having more than one of these mutations. Hence we are interested in finding sets of genes whose mutations are approximately mutually exclusive (anti-co-occurring) within the TCGA Pancancer dataset. In principle, finding the best set is NP Hard. Nevertheless, we will show how a new Mutation anti-co-OCcurrence Algorithm (MOCA) provides an effective greedy search and testing algorithm with guaranteed control of the familywise error rate or false discovery rate, by combining some under-appreciated ideas from frequentist hypothesis testing. These ideas include: (a) A novel exact conditional test for the tendency of multiple sets to have a large/small union/intersection, which generalises Fisher's exact test of 2x2 tables. (b) Randomised hypothesis tests for discrete distributions. (c) Stouffer's method for combining p-values. (d) Weighted multiple hypothesis testing. A new approach to setting a-priori weights which generates additional implicit hypothesis tests is suggested, and allows us to preserve almost all statistical power when testing pairs despite introducing a combinatorially large number of additional hypotheses.

Key Words: hypothesis testing, Fisher's exact test, cancer, genetics, co-occurrence, exclusivity

1. Introduction

Modern sequencing technology provides a wealth of genetic, genomic and metabolomic data, and using this data to help map out the complex interactions between genes, and between genes and cancer progression is an ongoing effort. This paper describes an algorithm for this based on detecting exclusivity patterns between somatic mutations.

It is widely believed that cancer progression is linked to the gradual accumulation of somatic mutations in tumour cells, and in particular of mutations which disrupt the functioning of certain key pathways such as cell cycle regulation and DNA repair (Hanahan and Weinberg, 2011). These pathway-disrupting mutations allow the cells which acquire them to multiply faster, thus conferring a selective advantage relative to the surrounding cells, and hence these mutations will be more common in the population of cancer cells than what the background mutation rate on its own would predict. Passenger mutations which do not affect cancer progression will on the other hand occur only at the background rate. This underlies analyses such as Kandoth et al. (2013) which identify genes relevant to cancer through their individual somatic mutation rates

The functioning of each pathway is complex, involving a large number of genes, some of which perform gene regulation functions. Hence each pathway may be disrupted by different somatic mutations in different patients. Although the first mutation to disrupt a given pathway offers a selective advantage, we expect that the accumulation of additional mutations in that pathway will typically not. Hence cancer cells in which multiple mutations disrupt one pathway will be rarer than if such mutations were statistically independent. In an idealised scenario, one would expect that the set of somatic mutations in each patient

[*] Imperial College London, 180 Queen's Gate, London SW7 2AZ, UK
[†] Columbia University, 1130 St. Nicholas Avenue, New York, NY 10032, USA

will consist of exactly one mutation disrupting each of the pathways. Across patients, the mutations affecting a given pathway would then be perfectly mutually exclusive.

Reality is of course more complex than the simplified description given above. For example certain mutations may disrupt multiple pathways (e.g. TP53), and certain pairs of mutation exhibit synthetic lethality (McCarthy, 2011; Li et al., 2014). Because of various reasons including background mutation rates, possible residual selective advantages in having more than one mutation per pathway, sequencing errors, and the existence of additional pathway disruption mechanisms, some patients may have multiple mutations from genes in a given pathway, and some may have none. A tendency for mutations in some of the genes affecting a given pathway to co-occur less than one would expect by chance nevertheless will remain, and can be detected using statistical techniques. Various analyses and interpretations of both co-occurrence and anti-co-occurrence patterns in somatic mutations have been suggested, mostly based on testing pairs of genes (Yeang et al., 2008; Cui, 2010; Gu et al., 2013; Wang et al., 2011; Gu et al., 2010). In this paper we are interested only in detecting anti-co-occurrence, i.e. appproximate mutual exclusivity between genes, and we wish to consider not just pairs but also larger sets of genes.

Multiple methods have been suggested to search for sets of anti-co-occurring genes (Leiserson et al., 2015, 2013; Vandin et al., 2012; Babur et al., 2015; Constantinescu et al., 2015). The Dendrix algorithm of Vandin et al. (2012) showed that once the tendency of driver mutations to belong to mutually exclusive sets is taken into account, even relatively rare driver mutations can be identified 'de novo' from somatic mutation data, i.e. without any prior pathway information. Dendrix searches for sets of genes with high overall coverage and low overlap. This approach has been further developed in the Multi-Dendrix algorithm (Leiserson et al., 2013) which finds multiple gene sets simultaneously and the group's current state of the art CoMEt (formerly Dendrix++) (Leiserson et al., 2015) which uses a statistical test of exclusivity to score sets. See Babur et al. (2015) for a review and comparison of methods.

In Section 2 we introduce some basic notation. Section 3 describes how we compute the statistical significance of the anti-co-occurrence of a set of genes/alterations by generalising Fisher's exact test from pairs of alterations to sets of alterations, and applying a randomised Stouffer's method to combine information from multiple tumour types. Section 4 introduces a novel a weighted Bonferroni correction scheme to control the family-wise error rate whilst maintaining high power for small sets. Section 5 describes a greedy algorithm to generate a shortlist of candidate gene sets which will be tested according to Section 3. Finally Section 6 shows the pattern of anti-co-occurrence detected by our method on a combined dataset of 5807 tumour samples from 22 tumour types. R code implementing our method and the full list of 654 statistically significant gene sets are available at `https://github.com/PaulGinzberg/MOCA/`.

2. Setup and Notation

For each patient in the data we identify for each gene the presence or absence of three possible alterations: whether there are any functional SNVs (single nucleotide variations) or CNVs (copy number variations), which may be AMPs (amplifications) or DELs (deletions). This produces an $m \times n$ binary matrix $\mathbf{A} = (a_{ij})$ where each patient or sample corresponds to a column, and each gene corresponds to (up to) three rows, one for each of the three types of alterations considered (SNV, AMP, DEL). Hence $a_{ij} = 1$ means "alteration i is present/mutated in sample j". For the purpose of understanding the statistical methods, one may conflate genes with alterations, i.e. assume that each row in $\mathbf{A}$ corresponds to a different gene. After pre-processing, which is detailed in Section 6.1, our data contains $m = 1418$ alterations and a total of $n = 5807$ samples. These 5807 samples are split across 22 different types of cancer, as described in Table 2.

For ease of comparison, we will follow a notation similar to Leiserson et al. (2013). $M \subseteq \{1, \ldots, m\}$ will denote a set of alterations of size $|M|$. The coverage $\Gamma(i)$ of alteration i is the number of samples for which this alteration is present $\Gamma(i) = |\{j : a_{ij} = 1\}| = \sum_{j=1}^{n} a_{ij}$. Let $\mathbf{C} = (\Gamma(1), \ldots, \Gamma(m))$ denote the vector of these marginal coverages. Alteration set M is considered to be present/mutated in sample j if at least one of the alterations from M is present in sample j. Hence the coverage of M is defined as

$$\Gamma(M) = \left| \left\{ j : \sum_{g \in M} a_{gj} \geq 1 \right\} \right|. \tag{1}$$

M is perfectly mutually exclusive iff no sample has more than one alteration from M, i.e. iff $\Gamma(M) = \sum_{g \in M} \Gamma(g)$. The overlap of an alteration set M is the number of additional mutations compared to the perfect mutual exclusivity case $\omega(M) = \sum_{g \in M} \Gamma(g) - \Gamma(M)$.

3. Testing for group-wise anti-co-occurrence

3.1 An exact test for whether the intersection/union of multiple sets is smaller/larger than expected by chance

Given an alteration set $M = \{g_1, \ldots, g_{|M|}\}$ we wish to test the following hypotheses:

H_0 : The alterations $a_{g_1 j}, \ldots, a_{g_{|M|} j}$ occur independently.

H_1 : The alterations $a_{g_1 j}, \ldots, a_{g_{|M|} j}$ tend to co-occur less frequently than if they were independent, i.e. for $i, \ell \in M$ $\mathrm{P}(a_{ij} = 1 \cap a_{\ell j} = 1) < \mathrm{P}(a_{ij} = 1)\,\mathrm{P}(a_{\ell j} = 1)$.

When $|M| = 2$ the most commonly used test for this problem is a one-sided Fisher's exact test for the 2×2 contingency table of the events $a_{g_1 j} = 1$ and $a_{g_2 j} = 1$. The test rejects H_0 for large values of the test statistic $\Gamma(M)$. The exact conditional null distribution of $\Gamma(M) - \Gamma(g_1)$ given $(\Gamma(g_1), \Gamma(g_2))$ (or equivalently given $\mathbf{C}$) is Hypergeometric:

$$\Gamma(M) \mid (\Gamma(g_1), \Gamma(g_2)) \sim \Gamma(g_1) + \mathrm{Hypergeom}(n, n - \Gamma(g_1), \Gamma(g_2)), \tag{2}$$

where the hypergeometric distribution has probability mass function

$$f_{\mathrm{Hypergeom}(n,k,r)}(x) = \frac{\binom{k}{x}\binom{n-k}{r-x}}{\binom{n}{r}},$$

over the support $\max(0, k + r - n) \leq x \leq \min(k, r)$.

Let us now consider the case of groupwise testing, i.e. $|M| > 2$. One approach which has been suggested in the literature is do define fully parametric generative models of the mutations under both the null and alternative hypothesis and then perform an asymptotic likelihood ratio test (Szczurek and Beerenwinkel, 2014; Constantinescu et al., 2015). This approach relies on the validity of the underlying generative models. We will instead generalise Fisher's exact test by deriving the conditional null distribution of $\Gamma(M)$ given $\mathbf{C}$. The use of the test statistic $\Gamma(M)$ for $|M| \geq 2$ had already been suggested by Ciriello et al. (2012). However they relied on Monte Carlo permutation testing to compute p-values, an approach which is computationally prohibitive when the p-values are very small. We will provide an algorithm for fast computation of *exact* p-values for the test statistic $\Gamma(M)$.

Assume H_0. We will now derive the conditional null distribution of $\Gamma(M)$ given $\Gamma(g_1), \ldots, \Gamma(g_{|M|})$, (or equivalently given $\mathbf{C}$) by iterated convolution. For $s = 1, \ldots, |M|$ let $M^{(s)} = \{g_1, \ldots, g_s\} \subseteq M$, so that $M^{(|M|)} = M$. If $1 < s \leq |M|$ then

$$\Gamma(M^{(s)}) \mid (\Gamma(M^{(s-1)}), \Gamma(g_s)) \sim \Gamma(M^{(s-1)}) + \mathrm{Hypergeom}(n, n - \Gamma(M^{(s-1)}), \Gamma(g_s)). \tag{3}$$

Using (3) and the fact that under H_0 $\Gamma(M^{(s)})$ is conditionally independent of $\mathbf{C}$ given $(\Gamma(M^{(s-1)}), \Gamma(g_s))$, we can compute the conditional probability mass function $f_{\Gamma(M)|\mathbf{C}}(x)$ iteratively as

$$f_{\Gamma(M^{(s)})|\mathbf{C}}(x) = \sum_y f_{\mathrm{Hypergeom}(n,n-y,\Gamma(g_s))}(x - y) f_{\Gamma(M^{(s-1)})|\mathbf{C}}(y). \tag{4}$$

The p-value for our test is then

$$p_M = \sum_{x \geq \Gamma(M)} f_{\Gamma(M^{(|M|)})|\mathbf{C}}(x). \tag{5}$$

Note that computing this p-value only requires evaluating the probability mass function for values of x in the right tail at each iteration. In particular, if M is perfectly mutually exclusive then each sum (5), (4) contains only one term.

Because our test is based on $\Gamma(M)$, it has a simple and intuitive interpretation: "Given $|M|$ subsets of $\{1, \ldots, n\}$ of sizes $\Gamma(g_1), \ldots, \Gamma(g_{|M|})$, is the size $\Gamma(M)$ of their union larger than would be expected for independently drawn random subsets of those sizes?". If we wished to detect co-occurrence rather than anti-co-occurrence could also perform the opposite one-sided test and reject for small values of $\Gamma(M)$. Also, because $n - \Gamma(M)$ is the size of the intersection of $|M|$ subsets of $\{1, \ldots, n\}$ of sizes $n - \Gamma(g_1), \ldots, n - \Gamma(g_{|M|})$, we can apply our test to intersections instead of unions. We expect that our generalisation of Fisher's exact test will have applications beyond biostatistics.

Because it requires $|M| - 2$ "convolutions" (plus two sums), the computational complexity of our iterative approach is linear in $|M|$. Leiserson et al. (2015) perform exact groupwise testing but uses as a test statistic the number of samples for which exactly one alteration from M is present $T(M) = \left|\left\{j : \sum_{g \in M} a_{gj} = 1\right\}\right|$. For $|M| = 2$ both tests are equivalent and reduce to Fisher's exact test. The two tests are also equivalent when there is perfect mutual exclusivity. For $|M| > 2$ and imperfect mutual exclusivity however the computational complexity of the enumeration technique employed by Leiserson et al. (2015) grows exponentially with $|M|$, and becomes impractical for large sample sizes except in the extreme tail of the distribution. In practice the two tests will tend to give similar answers, although the statistic $T(M)$ has the advantage of being more robust to the presence of hypermutated phenotypes.

Suppose for example that $g_1, g_2 \in M$ are respectively amplification and deletion of the same gene. Then it does not make sense to treat the anti-co-occurrence between g_1 and g_2 as evidence of any interesting biological relationship, since the two alterations are perfectly mutually exclusive by definition. Hence, whenever multiple alterations of a given gene are present in an alteration set, we will condition on the total coverage of these multiple alterations, and not just the coverage of each alteration. This is equivalent to treating $\{g_1, g_2\}$ as a single alteration of size $\Gamma(\{g_1, g_2\})$ in our test.[1]

[1] The only difference is that when performing our multiple hypothesis testing correction, the set size will be the original the number of alterations, not the number of genes.

3.2 Combining p-values across cancer types

The mutation frequencies of certain genes is different in different cancer types, and indeed some mutations may be specific to only one type or subtype of cancer. In a dataset with multiple cancer types, this effect can on its own induce anti-co-occurrence between functionally unrelated alterations simply because the cancer types in which they are common are different. To avoid this effect, when applying our method we will compute p-values on each cancer type separately, and then combine these p-values with Stouffer's method. Two issues must however be addressed. First, what weights to use in Stouffer's method, and second how to handle effects caused by the fact that the distribution of our test statistic $\Gamma(M)$ (and hence of our p-values) is discrete.

Stouffer's method consists in computing the combined p-value

$$p_{\text{Stouffer}} = \Phi\left(\frac{\sum_\tau v_\tau \Phi^{-1}(p_\tau)}{\sqrt{\sum_\tau v_\tau^2}}\right),$$

where Φ is the standard normal CDF, v_τ are (unnormalised) weights, and p_τ are the p-values from the independent tests being combined.

The asymptotically optimal weights to use in the case where $|M| = 2$ (and the effect size is the same for all τ) are well known, and are the inverse asymptotic standard deviation of the empirical log-odds ratio (which is a variance-stabilised parameter estimator) under the null. The (unnormalised) weight for each tumour type τ in this case is

$$\begin{aligned} v_{\{g_1,g_2\},\tau} = \Bigg(& \frac{n_\tau}{\Gamma_\tau(g_1)\,\Gamma_\tau(g_2)} + \frac{n_\tau}{\Gamma_\tau(g_1)(n - \Gamma_\tau(g_2))} \\ & + \frac{n_\tau}{(n - \Gamma_\tau(g_1))\,\Gamma_\tau(g_2)} + \frac{n_\tau}{(n - \Gamma_\tau(g_1))(n - \Gamma_\tau(g_2))} \Bigg)^{-\frac{1}{2}}, \end{aligned}$$

where n_τ is the number of samples with tumour type τ and the coverages Γ_τ are computed using only those samples with tumour type τ.

Finding a nice closed-form formula for the more general case $|M| > 2$ is non-trivial since it requires us to formulate some appropriate parametric alternative distribution generalising the non-central hypergeometric distribution, and we opt instead for the following heuristic weight v_M, based on assuming that the power of the groupwise test can be approximated as the power of combining all pairwise tests between genes in the group, as though they were independent:

$$v_{M,\tau} = \left(\sum_{k=1}^{|M|-1} \sum_{\ell=k+1}^{|M|} v^2_{\{g_k,g_\ell\},\tau}\right)^{\frac{1}{2}}.$$

The fact that these weights are based on approximate power does not make our testing procedure approximate, although it may cause the overall power of our combined test to be lower than it would be with optimal choices of weights. Although for each M our weights depend on the data, they are valid because they only use information in $\mathbf{C}$, and the tests being performed are all conditional on fixed $\mathbf{C}$.

In the case of exact tests based on continuous test statistics, the null distribution of each p_τ will be $\text{Uniform}(0, 1)$, and so will the null distribution of p_{Stouffer}. This is no longer true with discrete tests. If we apply Stouffer's method naively, then we will optain a p-value satisfying $\text{P}(p_{\text{Stouffer}} \leq \alpha) \leq \alpha$, but it may be catastrophically conservative. Indeed, if even just one of the tests being combined returns a p-value of 1, then the overall p-value is 1.

Kincaid (1962) considers various approaches for solving this problem when Fisher's method for combining p-values is used. However, their remarks are also valid for Stouffer's method. One of the simpler suggested approaches is a simplified version of Lancaster's procedure, which is equivalent to using mid-p-values $p_{\mathcal{T}}^{\mathrm{mid}} = p_{\mathcal{T}} - 0.5(p_{\mathcal{T}} - p_{\mathcal{T}}^{-})$, where $p_{\mathcal{T}}^{-} = \mathrm{P}_0(\Gamma(M) > \gamma(M))$, instead of p-values $p_{\mathcal{T}} = \mathrm{P}_0(\Gamma(M) \geq \gamma(M))$. The logic behind this is that the distribution of the mid-p-value is better approximated by the uniform distribution than that of the p-value. Because of its simplicity we will use this approximate approach to combine p-values from Fisher's exact test used internally in our greedy algorithm for generating candidate gene sets (See Section 5). However, because this approximate approach does not guarantee $\mathrm{P}(p_{\mathrm{Stouffer}} \leq \alpha) \leq \alpha$, (and does not produce a mid-p-value) we will use instead the Pearson approach (Kincaid, 1962) for combining the final p-values: This latter approach consists in using independently randomised p-values

$$p_{\mathcal{T}}^{\mathrm{rand}} = p_{\mathcal{T}} - (p_{\mathcal{T}} - p_{\mathcal{T}}^{-}) \cdot \mathrm{Uniform}(0, 1)$$

instead of the original p-values $p_{\mathcal{T}}$. Combining randomised p-values will never increases the number of false negatives compared to naively combining non-randomised p-values, but still controls the false positive rate so that $\mathrm{P}(p_{\mathrm{Stouffer}} \leq \alpha) = \alpha$. For large sample sizes, (and finite log-odds-ratio) this randomised p_{Stouffer} is restricted to lie in a small interval below the best (but computationally prohibitive) non-randomised p-value, so the variability introduced by randomisation becomes negligible.

4. A novel weighted multiple hypothesis testing scheme

It is appropriate to correct the alteration set p-values for multiple hypothesis testing. Because the procedure for generating candidate alteration sets described in Section 5 uses information from the data (other than $\mathbf{C}$), the multiple hypothess testing correction must also take into account all tests which could have been performed, even when the number of candidate alteration sets actually tested is limited. The total number of alteration sets which could have been tested is $\sum_{k=2}^{k_{\max}} \binom{m}{k}$, where $k_{\max}$ denotes the maximum allowed alteration set size. If we apply a standard unweighted multiple hypothesis testing correction, e.g. a Bonferroni correction, then the corrected p-values will depend strongly on the choice of $k_{\max}$. If we are agnostic a-priori about the size of alteration sets and use $k_{\max} = m$, then the number of alteration sets which could have been tested is $2^m - m - 1 > 10^{426}$. This is so large that there is little hope of any statistical significance after such a standard unweighted multiple hypothesis testing correction.

Since the main motivation behind performing groupwise testing is that the groupwise test can be much more powerful than pairwise tests, we must ensure that the loss of power caused by introducing additional hypotheses with $|M| > 2$ does not overwhelm the gains. We propose to use a *weighted* multiple hypothesis testing correction scheme which upweights smaller alteration sets at the expense of larger ones, and which is based on the argument that the smallest p-values after weighting should typically correspond to the most "interesting" alteration sets. In practice our proposed approach leads to only a small increase in the multiple hypothesis correction applied to p_M when $|M| = 2$, despite the large number of additional hypotheses.

For simplicity, we will assume that there is a single cancer type in this section. Define the conditional null probability measure $\mathrm{P}_0(\bullet) = \mathrm{P}(\bullet|\mathbf{C}, H_{\emptyset})$ corresponding to the global null hypothesis $H_{\emptyset}$ that the rows of $\mathbf{A}$ are independent. Then the p-value obtained when applying our test to alteration set M is $p_M = \mathrm{P}_0(\Gamma(M) \geq \gamma(M))$, where $\gamma(M)$ is the observed value of the test statistic $\Gamma(M)$.

Let $0 < \alpha \leq 1$ (which need not be equal to the significance level used for selecting statistically significant alteration sets) and define the weights

$$w_M = \frac{\left(\sum_{\ell=2}^{k_{\max}} \binom{m}{\ell}\right) \prod_{k=2}^{|M|} \left(1 - (1-\alpha)^{\frac{1}{m-k+1}}\right)}{\sum_{\ell=2}^{k_{\max}} \binom{m}{\ell} \prod_{k=2}^{\ell} \left(1 - (1-\alpha)^{\frac{1}{m-k+1}}\right)}, \tag{6}$$

which are normalised to have an average value of 1.

Lemma 1.

$$\mathrm{P}_0\left(\exists g \notin M : w_{M\cup\{g\}}^{-1} p_{M\cup\{g\}} < w_M^{-1} p_M \middle| \Gamma(M)\right) \leq \alpha$$

Proof. $\Gamma(M)$ is a random variable for each M, and let us first consider arbitrary fixed data leading to an observed value $\gamma(M)$.

$$\begin{aligned} p_{M\cup\{g\}} &= \sum_{k\geq 0} \mathrm{P}_0\left(\Gamma(M\cup\{g\}) \geq \gamma(M\cup\{g\}) | \Gamma(M) = k\right) \mathrm{P}_0(\Gamma(M) = k) \\ &\geq \sum_{k\geq\gamma(M)} \mathrm{P}_0\left(\Gamma(M\cup\{g\}) \geq \gamma(M\cup\{g\}) | \Gamma(M) = k\right) \mathrm{P}_0(\Gamma(M) = k) \\ &\geq \mathrm{P}_0\left(\Gamma(M\cup\{g\}) \geq \gamma(M\cup\{g\}) | \Gamma(M) = \gamma(M)\right) \sum_{k\geq\gamma(M)} \mathrm{P}_0(\Gamma(M) = k) \\ &= \mathrm{P}_0\left(\Gamma(M\cup\{g\}) \geq \gamma(M\cup\{g\}) | \Gamma(M) = \gamma(M)\right) \cdot p_M \\ &= p_{M\cup\{g\}|M} \cdot p_M, \end{aligned}$$

where $p_{M\cup\{g\}|M} = \mathrm{P}_0\left(\Gamma(M\cup\{g\}) \geq \gamma(M\cup\{g\}) | \Gamma(M) = \gamma(M)\right)$ denotes the p-value obtained by a standard one-sided Fisher's exact test of independence between the events "having a mutation for alteration g" and "having at least one mutation amongst the alterations in set M".

Let us now treat the data, and hence $p_{M\cup\{g\}}$ and $p_{M\cup\{g\}|M}$, as random variables. Note that for any $g \notin M$, $|M\cup\{g\}| = |M| + 1$ and $\frac{w_{M\cup\{g\}}}{w_M} = 1 - (1-\alpha)^{\frac{1}{m-|M|}}$. Also note that that because $p_{M\cup\{g\}|M}$ is a p-value for fixed $\Gamma(M)$, $\mathrm{P}_0\left(p_{M\cup\{g\}|M} \leq \alpha \middle| \Gamma(M)\right) \leq \alpha$.

$$\begin{aligned} &1 - \mathrm{P}_0\left(\exists g \notin M : w_{M\cup\{g\}}^{-1} p_{M\cup\{g\}} < w_M^{-1} p_M \middle| \Gamma(M)\right) \\ &= \prod_{g\notin M} \mathrm{P}_0\left(w_{M\cup\{g\}}^{-1} p_{M\cup\{g\}} \geq w_M^{-1} p_M \middle| \Gamma(M)\right) \\ &\geq \prod_{g\notin M} \mathrm{P}_0\left(p_{M\cup\{g\}|M} \geq \frac{w_{M\cup\{g\}}}{w_M} \middle| \Gamma(M)\right) \\ &\geq \prod_{g\notin M} \left(1 - \frac{w_{M\cup\{g\}}}{w_M}\right) \\ &= 1 - \alpha \end{aligned}$$

□

The inequality $p_{M\cup\{g\}} \geq p_{M\cup\{g\}|M} \cdot p_M$ becomes an equality when $M\cup\{g\}$ has perfect mutual exclusivity; and the null distribution of the discrete p-value $p_{M\cup\{g\}|M}$ converges to a $\mathrm{Uniform}(0,1)$ for large sample sizes. Hence the inequalities in the proof of Lemma 1 are "tight" for small α, and any weighting scheme with weights proportional to

	$k_{\max} = 2$	$k_{\max} = 3$	$k_{\max} = 4$	$k_{\max} \geq 5$
$\lvert M\rvert = 2$	$\mathbf{1.004653 \cdot 10^6}$	$1.021830 \cdot 10^6$	$1.022050 \cdot 10^6$	$1.022053 \cdot 10^6$
$\lvert M\rvert = 3$	∞	$2.820910 \cdot 10^{10}$	$2.821518 \cdot 10^{10}$	$2.821524 \cdot 10^{10}$
$\lvert M\rvert = 4$	∞	∞	$7.783707 \cdot 10^{14}$	$7.783725 \cdot 10^{14}$
$\lvert M\rvert = 5$	∞	∞	∞	$2.145775 \cdot 10^{19}$

Table 1: Values for the weighted Bonferroni Correction (7) where $m = 1418$ and $\alpha = 0.05$. The case of unweighted pairwise tests is in bold.

$w_M \cdot (1+\epsilon)^{|M|}$, $\epsilon > 0$ will in general no longer satisfy Lemma 1. In this sense, the weighting scheme (1) is the flattest (the one least penalising large sets) which satisfies Lemma 1.

When using a weighting scheme which does not satisfy Lemma 1 for $\alpha = 0.5$ (e.g. uniform weights), we run the risk that most of the highly significant aletration sets will contain passenger mutations. Hence such weighting schemes arguably favour larger alteration sets unfairly over smaller ones. In our analysis we will set $\alpha = 0.05$.

We will use a weighted Bonferroni correction. Because the weights (6) are normalised so that the average weight is 1 this controls the familywise error rate (Roeder and Wasserman, 2009, Lemma 2.1). The same weighting scheme could be used with a weighted Benjamini-Hochberg procedure to control the false discovery rate (Genovese et al., 2006, Theorem 1). The weighted Bonferroni correction consists in multiplying the uncorrected p-value p_M corresponding to an alteration set M by

$$\frac{\sum_{\ell=2}^{k_{\max}} \binom{m}{\ell}}{w_M} = \frac{\sum_{\ell=2}^{k_{\max}} \binom{m}{\ell} \prod_{k=2}^{\ell} \left(1 - (1-\alpha)^{\frac{1}{m-k+1}}\right)}{\prod_{k=2}^{|M|} \left(1 - (1-\alpha)^{\frac{1}{m-k+1}}\right)}. \tag{7}$$

For small α, large m, and large $k_{\max}$, (7) can be well approximated by

$$(\mathrm{e}^{\alpha} - 1 - \alpha)\alpha^{-|M|}|M|!\binom{m}{|M|}.$$

It is clear from Table 1 that our weighted Bonferroni correction depends weakly on $k_{\max}$, and indeed when compared to the standard approach of only testing pairs of alterations ($k_{\max} = 2$), it causes only a small increase in the multiple hypothesis correction applied to pairwise tests, even when setting $k_{\max} = m$. The results in Section 6.2 show that despite almost all of the weight being focused on the smallest sets ($|M| = 2$), in most cases larger sets ($3 \leq |M| \leq 6$) still obtain smaller weighted p-values than any of the pairs that they contain.

5. A greedy algorithm for generating candidate sets

Because even for moderate $k_{\max}$ the number of possible alteration sets to consider $\sum_{k=2}^{k_{\max}} \binom{m}{k}$ is too large, we cannot exhaustively compute all p_M, and must select a limited number of candidate alteration sets to test. We set the maximum alteration set size to $k_{\max} = 10$. The number of alteration sets satisfying $2 \leq |M| \leq 10$ is $\sum_{k=2}^{10} \binom{1418}{k} \approx 8.84 \cdot 10^{24}$.

Constantinescu et al. (2015) generate candidate sets by considering the cliques in a graph where the edges indicate statistically significant pairwise anti-co-occurrence. Leiserson et al. (2015) explores the space of possible sets of sets of alterations through an MCMC algorithm which swaps out one alteration at a time. As a simpler alternative to MCMC, Vandin et al. (2012) also considers a greedy approach to finding the alteration set which maximises their test statistic $W(M) = \Gamma(M) - \omega(M)$, and show that for large sample sizes the greedy approach succeeds with high probability.

The greedy algorithm for generating candidate alteration sets is based on two insights: Firstly, each alteration can also be considered as an alteration set of size 1. Secondly, Fisher's exact test can be used to test for anti-co-occurrence between alteration sets, i.e. (2) remains valid if $M = g_1 \cup g_2$ where g_1 and g_2 are alteration sets rather than single alterations. The proposed algorithm takes as inputs the binary alteration matrix $\mathbf{A}$ and the number of alteration sets one wishes to generate maxIter (which we will set to 5000). It then generates at each iteration a new alteration set by taking the union of the two existing alteration sets which are the most significantly anti-co-occurring. Pairs of alteration sets whose union is an existing alteration set or is larger than $k_{\max}$ are ignored. To ensure that the space of small alteration sets is properly explored, the iterations are split into $k_{\max} - 1$ equal-sized epochs and the maximum allowed size of new alteration sets is increased by 1 from 2 to $k_{\max}$ between epochs.

As described in Section 3.2, within this greedy algorithm significance is measured by performing Fisher's exact tests independently on each of the tumour types and combining the mid p-values of these tests. An advantage of using mid-p-values over p-values here is that the ordering of mid-p-values is more appropriate for our purposes: If g_1 and g_2 are perfectly mutually exclusive, but their p-value is close to 1 because of discreteness, then they are arguably more significantly anti-co-occurring than two alterations with a large overlap. The mid-p-value of perfectly mutually exclusive alterations will always be ≤ 0.5. The fact that the tests used in the greedy algorithm may be liberal is not an issue since the purpose of this step is only to heuristically search for candidate alteration sets on which accurate groupwise testing will be performed.

All subsets of the candidate alteration sets are then added to the set of candidate alteration sets.

Finally, the exact p-value $p_\mathcal{T}$ for each candidate alteration set is computed for each tumour type as described in Section 3.1, and these are randomised and combined as described in Section 3.2 to produce an overall p-value $p_M = p_M^{\text{rand}}$. The Bonferroni correction (7) with $\alpha = 0.05$ is then applied, and Bonferroni-corrected p-values greater than $\alpha = 0.05$ are discarded as non-significant. The ordering of Bonferroni-corrected p-values is the same as the ordering of weighted p-values, and in light of Lemma 1 we also discard an alteration set as non-significant if any of its subsets (or supersets) has a smaller Bonferroni-corrected p-value.

6. Application to the TCGA Pancancer dataset

6.1 Pre-processing

Somatic SNP and CNV data is obtained for 22 tumour types as described in Table 2. All but one of the tumour type datasets (and more than 99% of samples) was collected by TCGA. The total sample size across all 22 datasets is $n = 5807$.

Selection of genes: Although the proposed algorithm can be used in theory to detect genes which are relevant to cancer purely through their mutation pattern relative to other genes, this will not be attempted in this paper, and we will focus instead only on the search for relationships between genes. Hence we will restrict our analysis to genes which have already been identified as potentially relevant to cancer because of their high mutation rate, or through other means. Specifically, we include in our analysis the 127 genes which were identified as significantly mutated in (Kandoth et al., 2013, Supplementary Table 4) and the 487 genes listed by the Catalogue Of Somatic Mutations in Cancer (COSMIC) cancer gene census (Futreal et al., 2004, Supplementary Table S1). Combining these two sources

Tumour Type	Acronym	Samples	Publication
Acute myeloid leukemia	AML	193	(TCGA, 2013b)
Urothelial bladder cancer	BLCA	129	(TCGA, 2014b)
Breast cancer	BRCA	975	(TCGA, 2012c)
Colon adenocarcinoma	COAD	216	(TCGA, 2012b)
Glioblastoma multiforme	GBM	281	(TCGA, 2008)
Head & neck squamous cell carcinoma	HNSC	302	(TCGA, 2015a)
Clear cell kidney carcinoma	KIRC	292	(TCGA, 2013a)
Papillary kidney carcinoma	KIRP	161	(TCGA, 2016)
Lower grade glioma	LGG	286	(TCGA, 2015b)
Liver hepatocellular carcinoma	LIHC	187	
Lung adenocarcinoma	LUAD	466	(TCGA, 2014c)
Lung squamous cell carcinoma	LUSC	178	(TCGA, 2012a)
Ovarian serous cystadenocarcinoma	OV	139	(TCGA, 2011)
Pancreatic ductal adenocarcinoma	PAAD	168	
Prostate cancer (castration-resistant)	CRPC	58	(Grasso et al., 2012)
Prostate adenocarcinoma	PRAD	199	(TCGA, 2015d)
Rectal adenocarcinoma	READ	81	(TCGA, 2012b)
Sarcoma	SARC	257	
Cutaneous melanoma	SKCM	364	(TCGA, 2015c)
Stomach adenocarcinoma	STAD	229	(TCGA, 2014a)
Papillary thyroid carcinoma	THCA	403	(TCGA, 2014d)
Uterine corpus endometrial carcinoma	UCEC	243	(TCGA, 2013c)

Table 2: Number of samples for each tumour type. All data is from The Cancer Genome Atlas Research Network, with the exception of the additional prostate cancer dataset from Grasso et al.

yields a whitelist of 547 genes. Because there are three possible alteration types (SNV, AMP, DEL) for each gene, the whitelist corresponds to 1641 possible alterations.

Alterations which are mutated in exactly the same set of samples (i.e. identical rows in $\mathbf{A}$) are merged into one since in terms of their anti-co-occurrence they are mathematically indistinguishable. This step can reduce the computational requirements and can also aid interpretation. None of these merged alterations appear in the significant sets.

The number of alterations considered is reduced further by eliminating those alterations which are too rare for there to be a realistic chance of measuring statistically significant anti-co-occurrence after a Bonferroni correction is applied. The minimum number of mutations required for inclusion is set to be greater than $\log_2$(Number of alterations remaining$-$ 1). Based on this heuristic, all alterations with $\Gamma(g) < 11$ were removed. This leaves us with a final binary mutation matrix $\mathbf{A}$ describing $m = 1418$ alterations across $n = 5807$ samples.

Missing values in $\mathbf{A}$ are set to 0 (i.e. the alteration is assumed to be absent).

6.2 Results

After a weighted Bonferroni correction with $\alpha = 0.05$ as the significance level, We find 654 statistically significant sets of anti-co-occurring alterations (6 pairs, 147 triplets, 222 quadruplets, 261 quintuplets and 18 sextuplets). The 10 most statistically significant sets are given in Table 3. These alteration sets contain a total of 125 different alterations (116 different genes). We can think of each alteration set as a fully connected graph with alterations as vertices. The union of these 654 graphs contains 125 vertices and 969 edges. This graph is displayed in Figure 1. The rarest alteration out of these 125 is TNFAIP3(D) with 35 mutations (0.6%).

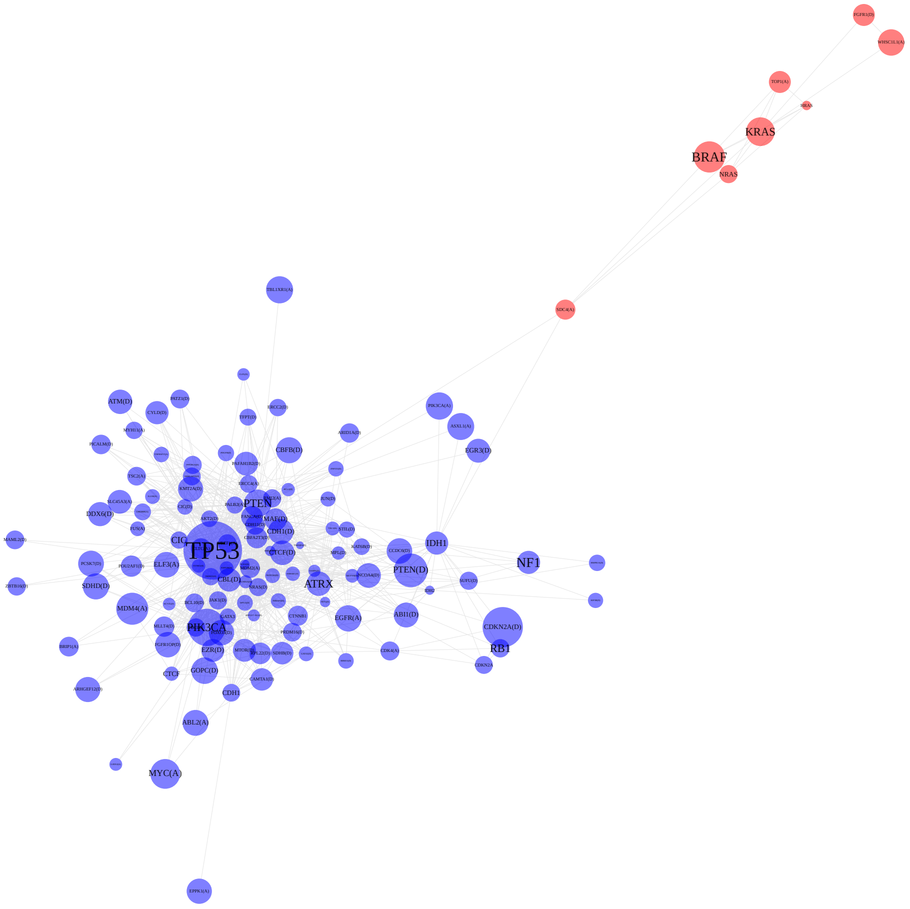

Figure 1: The graph obtained from the union of the 654 statistically significant ($\alpha = 0.05$) alteration sets obtained with our approach. The size of nodes is proportional to the coverage of the alteration. The color of nodes indicates the two communities (detected by label propagation).

The increased power from using groupwise testing is evident if we compare with the results from a pairwise analysis. If we compute pairwise p-values based on Fisher's exact test (combined over all cancer types),[2] then after applying the Bonferroni correction $\frac{m(m-1)}{2}$, only 82 pairs of alterations are significant at the level $\alpha = 0.05$, between 67 unique alterations (63 genes). The rarest alteration out of these 67 is TLX1(D) with 59 mutations (1%). Only one of the pairs detected by the pairwise approach is missing an edge in the graph obtained with the groupwise approach, namely (EIF4A2(A),PTEN), and this is because EIF4A2(A) is not in any of the sets detected by the groupwise approach. If we take the graph union of these pairs, displayed in Figure 2, there are no cliques of size ≥ 3, i.e. the pairwise approach fails to detect any groups of size ≥ 3. The graph does however split into 7 connected components.

[2] In the pairwise analysis we used mid-p-values instead of randomised p-values in Stouffer's method.

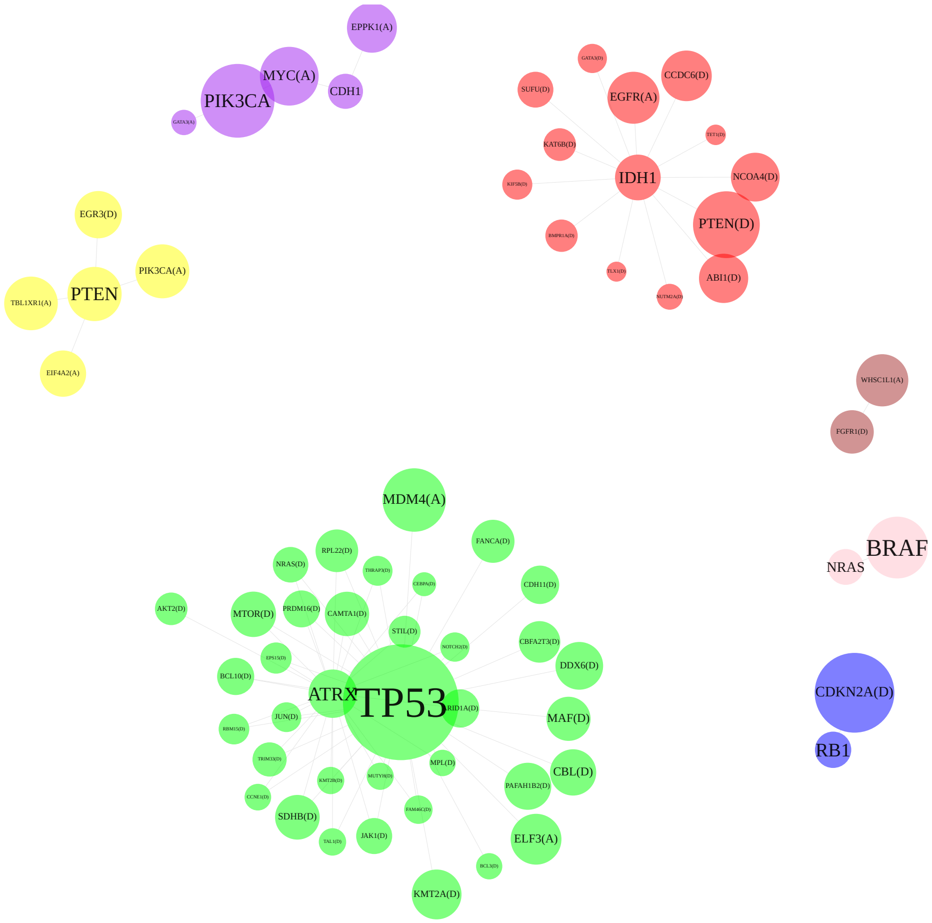

Figure 2: The graph obtained from the statistically significant pairs ($\alpha = 0.05$) of anti-co-occurring alterations. The size of nodes is proportional to the coverage of the alteration. The color of nodes is based on the connected components.

	$\frac{1}{n}\Gamma(M)$	p-value (uncorrected)	p-value (Weighted Bonferroni)	Alteration Set
1	25.6%	$1.22 \cdot 10^{-53}$	$5.20 \cdot 10^{-39}$	TOP1(A), BRAF, HRAS, KRAS, NRAS
2	24.0%	$8.11 \cdot 10^{-45}$	$1.22 \cdot 10^{-32}$	SDC4(A), BRAF, KRAS, NRAS
3	42.8%	$3.77 \cdot 10^{-27}$	$1.61 \cdot 10^{-17}$	ARID1A(D), PTEN, TP53
4	37.2%	$2.04 \cdot 10^{-26}$	$8.70 \cdot 10^{-17}$	STIL(D), HMGA2(A), TP53
5	9.2%	$3.84 \cdot 10^{-26}$	$1.64 \cdot 10^{-16}$	TLX1(D), RBM15(D), ATRX
6	36.3%	$4.78 \cdot 10^{-26}$	$2.04 \cdot 10^{-16}$	MUTYH(D), TNFAIP3(D), TP53
7	48.1%	$5.73 \cdot 10^{-31}$	$2.45 \cdot 10^{-16}$	STIL(D), TP53, MDM2(A), PTEN, CTCF(D)
8	36.8%	$6.84 \cdot 10^{-26}$	$2.92 \cdot 10^{-16}$	RBM15(D), TNFAIP3(D), TP53
9	13.9%	$8.82 \cdot 10^{-26}$	$3.76 \cdot 10^{-16}$	FAM46C(D), NCOA4(D), ATRX
10	45.3%	$4.26 \cdot 10^{-28}$	$6.43 \cdot 10^{-16}$	JUN(D), FANCA(D), TP53, PTEN

Table 3: The top 10 most significant alteration sets obtained with our method after Bonferroni correction. (A) indicates an amplification, (D) indicates a deletion. The full list is available at `https://github.com/PaulGinzberg/MOCA/blob/master/example-output/pancancer-test2-significantv2.txt`

7. Acknowledgements

This research was supported in part by EPSRC Mathematics Platform Grant EP/I019111/1. The empirical results in Section 6 are based upon data generated by the TCGA Research Network: `http://cancergenome.nih.gov/`.

References

Babur, z., Gönen, M., Aksoy, B. A., Schultz, N., Ciriello, G., Sander, C., and Demir, E. (2015), "Systematic identification of cancer driving signaling pathways based on mutual exclusivity of genomic alterations," *Genome Biology*, 16, 45.

Ciriello, G., Cerami, E., Sander, C., and Schultz, N. (2012), "Mutual exclusivity analysis identifies oncogenic network modules," *Genome Research*, 22, 398–406.

Constantinescu, S., Szczurek, E., Mohammadi, P., Rahnenführer, J., and Beerenwinkel, N. (2015), "TiMEx: a waiting time model for mutually exclusive cancer alterations," *Bioinformatics*, btv400.

Cui, Q. (2010), "A Network of Cancer Genes with Co-Occurring and Anti-Co-Occurring Mutations," *PLoS ONE*, 5, e13180.

Futreal, P. A., Coin, L., Marshall, M., Down, T., Hubbard, T., Wooster, R., Rahman, N., and Stratton, M. R. (2004), "A census of human cancer genes," *Nature Reviews Cancer*, 4, 177–183.

Genovese, C. R., Roeder, K., and Wasserman, L. (2006), "False discovery control with p-value weighting," *Biometrika*, 93, 509–524.

Grasso, C. S., Wu, Y.-M., Robinson, D. R., Cao, X., Dhanasekaran, S. M., Khan, A. P., Quist, M. J., Jing, X., Lonigro, R. J., Brenner, J. C., Asangani, I. A., Ateeq, B., Chun, S. Y., Siddiqui, J., Sam, L., Anstett, M., Mehra, R., Prensner, J. R., Palanisamy, N., Ryslik, G. A., Vandin, F., Raphael, B. J., Kunju, L. P., Rhodes, D. R., Pienta, K. J., Chinnaiyan, A. M., and Tomlins, S. A. (2012), "The mutational landscape of lethal castration-resistant prostate cancer," *Nature*, 487, 239–243.

Gu, Y., Wang, H., Qin, Y., Zhang, Y., Zhao, W., Qi, L., Zhang, Y., Wang, C., and Guo, Z. (2013), "Network analysis of genomic alteration profiles reveals co-altered functional modules and driver genes for glioblastoma," *Molecular BioSystems*, 9, 467–477.

Gu, Y., Yang, D., Zou, J., Ma, W., Wu, R., Zhao, W., Zhang, Y., Xiao, H., Gong, X., Zhang, M., Zhu, J., and Guo, Z. (2010), "Systematic Interpretation of Comutated Genes in Large-Scale Cancer Mutation Profiles," *Molecular Cancer Therapeutics*, 9, 2186–2195.

Hanahan, D. and Weinberg, R. A. (2011), "Hallmarks of Cancer: The Next Generation," *Cell*, 144, 646–674.

Kandoth, C., McLellan, M. D., Vandin, F., Ye, K., Niu, B., Lu, C., Xie, M., Zhang, Q., McMichael, J. F., Wyczalkowski, M. A., Leiserson, M. D. M., Miller, C. A., Welch, J. S., Walter, M. J., Wendl, M. C., Ley, T. J., Wilson, R. K., Raphael, B. J., and Ding, L. (2013), "Mutational landscape and significance across 12 major cancer types," *Nature*, 502, 333–339.

Kincaid, W. M. (1962), "The Combination of Tests Based on Discrete Distributions," *Journal of the American Statistical Association*, 57, 10–19.

Leiserson, M. D., Wu, H.-T., Vandin, F., and Raphael, B. J. (2015), "CoMEt: a statistical approach to identify combinations of mutually exclusive alterations in cancer," *Genome Biology*, 16, 160.

Leiserson, M. D. M., Blokh, D., Sharan, R., and Raphael, B. J. (2013), "Simultaneous Identification of Multiple Driver Pathways in Cancer," *PLoS Comput Biol*, 9, e1003054.

Li, X.-j., Mishra, S. K., Wu, M., Zhang, F., and Zheng, J. (2014), "Syn-Lethality: An Integrative Knowledge Base of Synthetic Lethality towards Discovery of Selective Anticancer Therapies," *BioMed Research International*, 2014, e196034.

McCarthy, N. (2011), "Systems biology: Lethal weaknesses," *Nature Reviews Cancer*, 11, 538–539.

Roeder, K. and Wasserman, L. (2009), "Genome-Wide Significance Levels and Weighted Hypothesis Testing," *Statistical science : a review journal of the Institute of Mathematical Statistics*, 24, 398–413.

Szczurek, E. and Beerenwinkel, N. (2014), "Modeling Mutual Exclusivity of Cancer Mutations," *PLoS Comput Biol*, 10, e1003503.

TCGA (2008), "Comprehensive genomic characterization defines human glioblastoma genes and core pathways," *Nature*, 455, 1061–1068.

— (2011), "Integrated genomic analyses of ovarian carcinoma," *Nature*, 474, 609–615.

— (2012a), "Comprehensive genomic characterization of squamous cell lung cancers," *Nature*, 489, 519–525.

— (2012b), "Comprehensive molecular characterization of human colon and rectal cancer," *Nature*, 487, 330–337.

— (2012c), "Comprehensive molecular portraits of human breast tumours," *Nature*, 490, 61–70.

— (2013a), "Comprehensive molecular characterization of clear cell renal cell carcinoma," *Nature*, 499, 43–49.

— (2013b), "Genomic and Epigenomic Landscapes of Adult De Novo Acute Myeloid Leukemia," *New England Journal of Medicine*, 368, 2059–2074.

— (2013c), "Integrated genomic characterization of endometrial carcinoma," *Nature*, 497, 67–73.

— (2014a), "Comprehensive molecular characterization of gastric adenocarcinoma," *Nature*, 513, 202–209.

— (2014b), "Comprehensive molecular characterization of urothelial bladder carcinoma," *Nature*, 507, 315–322.

— (2014c), "Comprehensive molecular profiling of lung adenocarcinoma," *Nature*, 511, 543–550.

— (2014d), “Integrated Genomic Characterization of Papillary Thyroid Carcinoma,” *Cell*, 159, 676–690.

— (2015a), “Comprehensive genomic characterization of head and neck squamous cell carcinomas,” *Nature*, 517, 576–582.

— (2015b), “Comprehensive, Integrative Genomic Analysis of Diffuse Lower-Grade Gliomas,” *New England Journal of Medicine*, 372, 2481–2498.

— (2015c), “Genomic Classification of Cutaneous Melanoma,” *Cell*, 161, 1681–1696.

— (2015d), “The Molecular Taxonomy of Primary Prostate Cancer,” *Cell*, 163, 1011–1025.

— (2016), “Comprehensive Molecular Characterization of Papillary Renal-Cell Carcinoma,” *New England Journal of Medicine*, 374, 135–145.

Vandin, F., Upfal, E., and Raphael, B. J. (2012), “De novo discovery of mutated driver pathways in cancer,” *Genome Research*, 22, 375–385.

Wang, J., Zhang, Y., Shen, X., Zhu, J., Zhang, L., Zou, J., and Guo, Z. (2011), “Finding co-mutated genes and candidate cancer genes in cancer genomes by stratified false discovery rate control,” *Molecular BioSystems*, 7, 1158–1166.

Yeang, C.-H., McCormick, F., and Levine, A. (2008), “Combinatorial patterns of somatic gene mutations in cancer,” *The FASEB Journal*, 22, 2605–2622.