Supplemental Material

A Hardness Proofs

Theorem 6. *The Maximum Coverage Exclusive Submatrix Problem is NP-hard.*

Proof. Given a mutation matrix A and an integer $k > 0$, the Maximum Coverage Exclusive Submatrix Problem requires to find the $m \times k$ column submatrix $\hat{M}$ with the largest number of non-zero rows. We prove it is NP-Hard by reduction from the Maximum Weight Independent Set Problem. We consider the Maximum Weight Independent Set Problem with positive integer weights, that is again NP-Hard, since an algorithm for the case of positive integers weights can be use to find a solution to the Maximum Cardinality Independents Set Problem.

In the Maximum Weight Independent Set problem we are given a graph $G = (V, E)$, a weight function $w : V \to \mathbb{N}^+$, and a value $k > 0$, and are asked for an independent set of size k with maximum weight. An independent set is a set $I \subset V$ of vertices such that there is no edge between the vertices of I, i.e. $\forall u, v \in I, u \neq v : (u, v) \notin E$.

Given an instance for the Maximum Weight Independent Set problem we build an instance of the Maximum Coverage Exclusive Submatrix Problem as follows. The mutation matrix A has one column for each vertex $v \in V$. Let $\delta(v)$ be the degree of $v \in V$ in G, and $\Delta = \max_{v \in V} \delta(v)$. We define the set of rows of the mutation matrix as: $\mathcal{S} = \{s_e : e \in E\} \cup (\cup_{v \in V} \mathcal{S}_v)$ where $\mathcal{S}_v = \left\{s_v^{(1)}, s_v^{(2)}, \ldots, s_v^{(\Delta - \delta(v) + w(v))}\right\}$. We define $A_{s,v} = 1$ if $s = s_e = s_{(u,v)}$ with $e \in E$ or if $s \in \mathcal{S}_v$, and $A_{s,v} = 0$ otherwise. All these operations can be performed in polynomial time.

Note that: (i) for any two columns $u, v \in V$, $\Gamma(u) \cap \Gamma(v) \neq \emptyset$ if and only if $(u, v) \in E$. (ii) $\forall v \in V$, $|\Gamma(v)| = \Delta + w(v)$ (i.e, the number of rows in which column v is 1 is equal to $\Delta + w(v)$);

Now consider a set $M = \{v_1, \ldots, v_k\}$ of k columns. From (i) we have that column submatrix induced by M is exclusive if and only if M is an independent set of size k in G. Now, if M is an exclusive matrix, the number of non-zero rows in it is equal to $\sum_{i=1}^{k} |\Gamma(v_i)|$. From (ii) above $\sum_{i=1}^{k} |\Gamma(v_i)| = k\Delta + \sum_{i=1}^{k} w(v_i)$. Since k and Δ are fixed, the exclusive column submatrix induced by M maximizes $\sum_{i=1}^{k} |\Gamma(v_i)|$ if and only if M is the independent set of size that maximizes $\sum_{i=1}^{k} w(v_i)$, i.e. M is the maximum weight independent set in G. □

Theorem 7. *The Maximum Weight Submatrix Problem is NP-Hard.*

Proof. The proof is by reduction from the Independent Set Problem, a well known NP-Hard problem (Hochbaum, 1997). In the Independent Set Problem we are given a graph $G = (V, E)$ and a value k, and we ask if there is an independent set of size k in G. An independent set for G is a set of vertices $I \subseteq V$ such that there is no edge among the vertices of I, i.e. for all pairs $u, v \in I, u \neq v$: $(u, v) \notin E$.

Given an instance of the Independent Set Problem, we build a mutation matrix representing the instance of our problem as follows. We consider a column for each vertex $v \in V$. Let $\delta(v)$ be the degree of v in G, and define $\Delta = \max \delta(v)$. We define the set of rows of the mutation matrix as: $\mathcal{S} = \{s_e : e \in E\} \cup (\cup_{v \in V} \mathcal{S}_v)$ where $\mathcal{S}_v = \left\{s_v^{(1)}, s_v^{(2)}, \ldots, s_v^{(\Delta - \delta(v))}\right\}$. We define $A_{s,v} = 1$ if $s = s_e = s_{(u,v)}$ with $e \in E$ or if $s \in \mathcal{S}_v$, and $A_{s,v} = 0$ otherwise. All these operations can be performed in polynomial time.

Note that: (i) $\forall v \in V$, $|\Gamma(v)| = \Delta$ (i.e, the number of non-zeros entries in column v is Δ); (ii) for any two vertices $u, g \in V$, $\Gamma(u) \cap \Gamma(v) \neq \emptyset$ if and only if $(u, v) \in E$.

Now consider a set $M = \{v_1, \ldots, v_k\}$ of k columns. From (i) we have $\sum_{i=1}^{k} |\Gamma(v_k)| = k\Delta$, and $|\Gamma(M)| \leq k\Delta$. From (ii) we have that $|\Gamma(M)| = k\Delta$ if and only if $\{v_1, v_2, \ldots, v_k\}$ is an independent set of G, thus $W(M) = k\Delta$ if and only if $\{v_1, v_2, \ldots, v_k\}$ is an independent set of G. If we can solve the The Maximum Weight Genes Set Problem on A, we can then solve the Independent Set Problem on G. □

B Analysis of Greedy Algorithm

In this section we give a proof of Theorem 3. We need to prove that: (i) the first step of the greedy algorithm is correct, i.e. the pair $M = \{g_1, g_2\}$ of columns that maximizes $W(M)$ is a subset of $\hat{M}$; (ii) the loop in step 2 is correct, i.e. all the subsets of size ℓ built in the loop are subsets of $\hat{M}$.

To prove (i) and (ii), we need to lower bound the weight of the subsets that we build in the loop of step 2, assuming they are the correct ones.

Lemma 1. *Let M be a subset of $\hat{M}$ with $|M| = \ell, 0 \leq \ell < k$. If $W(M) \geq \frac{\ell}{k}\hat{M}$, then there exists a gene $g \in \hat{M} \setminus M$ such that $W(M \cup \{g\}) \geq \frac{\ell+1}{k}W(\hat{M})$.*

Proof. The proof is by contradiction. Let assume that for a given $M \subset \hat{M}$ with $|M| = \ell < k$, there exists no gene $g \in \hat{M} \setminus M$ such that $W(M \cup \{g\}) \geq \frac{\ell+1}{k}W(\hat{M})$.

Let $\hat{M} \setminus M = \{g_1, \ldots, g_{k-\ell}\}$, and $M_i = M \cup \{g_1, \ldots, g_i\}, i \leq k - \ell$. Since $W(M_{k-\ell}) \geq W(\hat{M})$ (in particular, $W(M_{k-\ell}) = W(\hat{M})$), there exists i such that $W(M_i) < \frac{\ell+i}{k}W(\hat{M})$ and $W(M_{\ell+1}) \geq \frac{\ell+i+1}{k}W(\hat{M})$. Let i^* be the minimum such i, $M' = M_{i^*-1}$and $g^* = g_{i^*}$.

We have

$$\begin{aligned} W(M' \cup \{g^*\}) &= 2|\Gamma(M' \cup \{g^*\})| - \sum_{g \in M' \cup \{g^*\}} |\Gamma(g^*)| \\ &= 2|\Gamma(M')| - \sum_{g \in M'} |\Gamma(g)| + 2|\Gamma(g^*) \cap (\mathcal{S} \setminus \Gamma(M'))| - |\Gamma(g^*)| \\ &= W(M') + \Delta(M', g^*) \end{aligned}$$

where $\Delta(M', g^*) = 2|\Gamma(g^*) \cap (\mathcal{S} \setminus \Gamma(M'))| - |\Gamma(g^*)|$.

Now, since $W(M' \cup \{g^*\}) = W(M') + \Delta(M', g^*) \geq \frac{\ell+i+1}{k}W(\hat{M})$ and $W(M') < \frac{\ell+i}{k}W(\hat{M})$, we have

$$\Delta(M', g^*) \geq \frac{\ell + i + 1}{k}W(\hat{M}) - \frac{\ell + i}{k}W(\hat{M}) \geq \frac{1}{k}W(\hat{M}).$$

Moreover, since $M \subseteq M'$, we have $\Gamma(M) \subseteq \Gamma(M')$, that implies $\Gamma(g^*) \cap (\mathcal{S} \setminus \Gamma(M')) \subseteq \Gamma(g^*) \cap (\mathcal{S} \setminus \Gamma(M))$.

Thus

$$\Delta(M, g^*) \geq \Delta(M', g^*) \geq \frac{1}{k}W(\hat{M}),$$

that implies

$$W(M \cup \{g^*\}) = W(M) + \Delta(M, g^*) \geq \frac{\ell}{k}\hat{M} + \frac{1}{k}W(\hat{M}) \geq \frac{\ell + 1}{k}W(\hat{M}),$$

that is a contradiction. □

We now prove that if the number of patients is large enough, (i) holds. Note that since in Gene Independence Model the frequency of mutation of the genes in $\hat{M}$ and the frequency of mutation of genes not in $\hat{M}$ can be the same, the most frequent genes are not guaranteed to be in $\hat{M}$. Instead we prove that if the the number of patients is large enough, the greedy algorithm, that checks sets of size 2, correctly identifies a subset of $\hat{M}$ of size 2.

Lemma 2. *Let P be the pair of genes in $\hat{M}$ with the highest weight $W(P)$, and let $W(\hat{M}) = rm$. Define the event $E =$"there exists a pair $\mathcal{R} \not\subseteq \hat{M}$ of genes such that $W(\mathcal{R}) \geq W(P)$". If*

$$m \geq \frac{(2 + \varepsilon)}{2[2r/k - 2(p_U - p_L^2)]^2} \log n$$

then

$$\mathbf{Pr}[E] \leq n^{-\varepsilon}.$$

Proof. Consider a pair of genes $\mathcal{R} = \{g_i, g_j\} \not\subseteq \hat{M}$. We can rewrite

$$W(\mathcal{R}) = X = \sum_{i=1}^{m} \in X_i,$$

where X_i is the random variable that counts the "contribution" of patient $s_i \in \mathcal{S}$ to $W(\mathcal{R})$. Note that for all i we have $X_i \in \{0; 1\}$, since when only one gene in $\mathcal{R}$ is mutated the contribution of X_i is 1, while when none or both genes in $\mathcal{R}$ are mutated in S the contribution of X_i is 0.

Let p_i, p_j be the frequency of mutation of g_i, g_j respectively. The expectation of $W(\mathcal{R})$ is

$$\mathbf{E}[W(\mathcal{R})] = \mathbf{E}[X] = m(p_i(1-p_j) + p_j(1-p_i)) \leq |\mathcal{S}|(2p_U - 2p_L^2).$$

Since the X_i are independent random variables, we can use the Chernoff bound we can then derive the probability that a particular set $\mathcal{R}$ has $W(\mathcal{R}) \geq \frac{2r}{k}|\mathcal{S}|$:

$$\begin{aligned}
\mathbf{Pr}\left[W(\mathcal{R}) \geq \frac{2r}{k}m\right] &= \mathbf{Pr}\left[W(\mathcal{R}) - \mathbf{E}[W(\mathcal{R})] \geq \frac{2r}{k}m - \mathbf{E}[W(\mathcal{R})]\right] \\
&\leq \mathbf{Pr}\left[W(\mathcal{R}) - \mathbf{E}[W(\mathcal{R})] \geq \frac{2r}{k}|\mathcal{S}| - m(2p_U - 2p_L^2)\right] \\
&\leq e^{-\frac{2m^2\left(\frac{2r}{k} - (2p_U - 2p_L^2)\right)^2}{m}}.
\end{aligned}$$

Now, since $m \geq \frac{(2+\varepsilon)}{2[2r/k - 2(p_U - p_L^2)]^2} \log n$, we have:

$$\begin{aligned}
\mathbf{Pr}\left[W(\mathcal{R}) \geq \frac{2r}{k}m\right] &\leq e^{-(2+\varepsilon)\ln n} \\
&\leq n^{-(2+\varepsilon)}.
\end{aligned}$$

The lemma follows by applying a union bound on all the possible pairs $\mathcal{P}$. □

Assume that at each step the greedy algorithm chooses a gene in $\hat{M}$. With $\hat{M}_\ell$ we denote the subset of size of ℓ genes of $\hat{M}$ obtained from the procedure above. Note that this defines an order on the genes of $\hat{M}$: in particular, we denote with g_ℓ the gene in $\hat{M}$ added to $\hat{M}_\ell$ in order to obtain $\hat{M}_{\ell+1}$. We now find a lower bound to the number of patients required to guarantee that with high probability there does not exist an iteration of the step 2 of the Greedy Algorithm in which a gene not in $\hat{M}$ is chosen.

Lemma 3. *Let* $g_\ell^* = \arg\max_g W(M_\ell \cup \{g\})$. *If*

$$m \geq \frac{(2+\varepsilon)\log n}{2\left(\frac{r(1-d)}{k} - p_U + \frac{4rp_L}{k}\right)^2},$$

then

$$\mathbf{Pr}[\exists \ell : g_\ell^* \notin \hat{M}] \leq n^{-\varepsilon}.$$

Proof. We assume that the subset built at each step is $\hat{M}_i \subseteq \hat{M}$. Since at the end the theorem will hold, this assumption will be proven correct.

Consider the set $\hat{M}_i$, that is mutated in the set $\Gamma(\hat{M}_i)$. By the assumptions of the Independence Gene Model, its weight is bounded by $W(\hat{M}_i) \leq \frac{i+d}{k} W(\hat{M})$. Now consider a gene $g_j \notin \hat{M}$, with mutation frequency $p_j \in [p_L; p_U]$. If we now add g_j to $\hat{M}_i$, we have that g_j contributes 1 to $W(\hat{M}_i \cup \{g_j\})$ for each patient in $\mathcal{S} \setminus \Gamma(\hat{M}_i)$ in which it is mutated, and -1 for each patient in $\Gamma(\hat{M}_i)$ in which it is mutated.

Since g_j is mutated with probability p_j in a patient, we have

$$\mathbf{E}[W(\hat{M}_i \cup \{g_j\})] = W(\hat{M}_i) + p_j(m - |\Gamma(\hat{M}_i)|) - p_j|\Gamma(\hat{M}_i)| = W(\hat{M}_i) + p_j m - 2p_j|\Gamma(\hat{M}_i)|).$$

Now, given the assumptions on $W(\hat{M}_i)$, $W(\hat{M})$, p_j, and since $|\Gamma(\hat{M}_i)| \geq W(\hat{M}_i)$ we have

$$\mathbf{E}[W(\hat{M}_i \cup \{g_j\})] \leq \frac{i+d}{k} rm + p_U m - 2p_L \frac{ir}{k} m \leq m\left(\frac{i+d}{k} r + p_U - 2p_L \frac{2r}{k}\right).$$

(In the last inequality we use $i \geq 2$.)

Since for each patient s the absolute value of the contribution to $W(\hat{M}_i \cup \{g_j\})$ of g_j in s is bounded by 1, we can use the Chernoff-Hoeffding bound to compute the probability that $W(\hat{M}_i \cup \{g_j\}) \geq \frac{i+1}{k} W(\hat{M})$:

$$\begin{aligned}
\mathbf{Pr}\left[W(\hat{M}_i \cup \{g_j\}) \geq \frac{i+1}{k} W(\hat{M})\right] &\leq e^{-2\frac{\left(\frac{i+1}{k} rm - m\left(\frac{i+d}{k} r + p_U - 2p_L \frac{2r}{k}\right)\right)^2}{m}} \\
&\leq e^{-2m\left(\frac{r(1-d)}{k} - p_U + \frac{4rp_L}{k}\right)^2}.
\end{aligned}$$

Now, since $m \geq \frac{(2+\varepsilon)\log n}{2\left(\frac{r(1-d)}{k} - p_U + \frac{4rp_L}{k}\right)^2}$, we have

$$\mathbf{Pr}\left[W(\hat{M}_i \cup \{g_j\}) \geq \frac{i+1}{k} W(\hat{M})\right] \leq n^{-(2+\varepsilon)}.$$

The total number of pairs $(\hat{M}_i, g_j)$ we have to consider is bounded by $nk \leq n^2$, since there are n genes $g_j \notin \hat{M}$, and there are k sets in $\hat{M}_i$, since $\hat{M}$ contains k genes. The lemma follows by union bound. □

Theorem 3 follows from Lemma 2 and Lemma 3.

C Proof of MCMC Convergence

Our analysis applies the following simple version of path coupling adapted to our setting (see (Bubley and Dyer, 1997) and (Mitzenmacher and Upfal, 2005)):

Theorem 8. *Let $\phi_t = |M_t - M'_t|$, and assume that for some constant $0 < \beta < 1$, $E[\phi_{t+1} |\ \phi_t = 1] \leq \beta$, then the mixing time*

$$\tau(\epsilon) \leq \frac{k \log(k\epsilon^{-1})}{1-\beta}.$$

Using the above, we prove the following convergence result for our chain.

Theorem 9. *The MCMC is rapidly mixing for some $c > 0$.*

Proof. Let $D = \max_{g \in \mathcal{G}} |\Gamma(g)|$. Assume first that in the first chain $v = y$. The probability that the first chain performs the switch is

$$\frac{e^{cW(M-\{y\}+\{w\})}}{e^{cW(M)}} \geq e^{-c(\Gamma(y)+\Gamma(w))} \geq e^{-2cD}.$$

Similarly the probability that the second chain performs the switch is $\geq e^{-2cD}$. Since $v = y$ with probability $1/k$ we have

$$Pr(\phi_{t+1} = 0 \mid \phi_t = 1) \geq \frac{1}{k} e^{-4cD} \geq \frac{1}{k} - \frac{4cD}{k}$$

for $4cD < 1$. Next assume that $v \in M \cap M'$. We need to upper bound the probability that exactly one of the chains perform the switch (otherwise $\phi_{t+1} = \phi_t$). The probability that exactly one chain performs a switch is given by

$$Q = \left| \min[1, e^{cW(M-\{v\}+\{w\})-cW(M)}] - \min[1, e^{cW(M'-\{v\}+\{w\})-cW(M')}] \right|.$$

Let $\tilde{\Gamma}_M(v) = \Gamma(v) - \Gamma(M - \{v\})$, i.e. patients where v is altered but no other gene in M is altered. Clearly

$$|(W(M - \{v\} + \{w\}) - W(M)) - (W(M' - \{v\} + \{w\}) - W(M'))| \leq$$

$$2(|\tilde{\Gamma}_M(v) \cap \Gamma(z)| + |\tilde{\Gamma}_{M'}(v) \cap \Gamma(y)| + |\Gamma(w) \cap \tilde{\Gamma}_M(y)| + |\Gamma(w) \cap \tilde{\Gamma}_{M'}(z)|).$$

Thus, (using $1 - e^{-x} \leq x$ for $x < 1$, and $c \leq 1/12D$)

$$Q \leq 2c(|\tilde{\Gamma}_M(v) \cap \Gamma(z)| + |\tilde{\Gamma}'_M(v) \cap \Gamma(y)| + |\Gamma(w) \cap \tilde{\Gamma}_M(y)| + |\Gamma(w) \cap \tilde{\Gamma}_{M'}(z)|).$$

Summing over all choices of w and $v \in M \cap M'$ we compute that the probability that exactly one of the chains performs a switch is bounded by $\frac{k-1}{k}\left(\frac{4cD}{k-1} + \frac{4cD\bar{D}}{m}\right)$, where $\bar{D} = \frac{\sum_{g \in \mathcal{G}} |\Gamma(g)|}{n}$. Setting $c < \min\left\{\frac{m}{8kD\bar{D}}, \frac{1}{16D}\right\}$ we have

$$E[\phi_{t+1} \mid \phi_t = 1] \leq 1 + 2\frac{k-1}{k}\left(\frac{2cD}{k-1} + \frac{2cD\bar{D}}{m}\right) - \frac{1}{k} + \frac{4cD}{k} \leq \frac{1}{2}.$$

Thus, with this value of c the mixing time satisfies $\tau(\epsilon) \leq 2k \log(k\epsilon^{-1})$. □

How good is the sampling process with this value of c in sampling sets of significant gene mutations? Assume that (after removing the genes that appear in most patients) we have $Dk = O(m)$, and that we have two sets M and M' such that $W(M) \geq \gamma W(M')$ for some $\gamma < 1$. In that case the sampling procedure samples M with frequency that is at least $e^{\Omega(k)}$ larger than the frequency that it samples M'.

D MCMC results

D.1 Results for Lung adenocarcinoma

Table 2 reports the results obtained on the lung data with $k = 2$ and $k = 3$. Table 3 reports the results for $k = 2$ obtained after removing the set (*EGFR, KRAS, STK11*). In all tables k is the size of set, M is the set of genes, $\tilde{\pi}(M)$ is the frequency of the set in the sample obtained with the MCMC, and $W(M)$ is the weight of the set.

k	$\tilde{\pi}(M)$	$W(M)$	M
2	99.9%	90	*EGFR KRAS*
3	8.2%	96	*STK11 EGFR KRAS*
	3.2%	94	*PRKCG EGFR KRAS*
	2%	93	*NRAS EGFR KRAS*
	2%	93	*PFTK1 EGFR KRAS*
	1.8%	94	*EPHB1 EGFR KRAS*
	1.7%	93	*MAP3K3 EGFR KRAS*
	1.6%	92	*VAV2 EGFR KRAS*
	1.5%	92	*ERBB4 EGFR KRAS*
	1.4%	92	*YES1 EGFR KRAS*
	1.3%	92	*TSC1 EGFR KRAS*
	1.3%	93	*NTRK3 EGFR KRAS*
	1.2%	93	*NF1 EGFR KRAS*
	1.1%	92	*PTK2 EGFR KRAS*
	1.1%	92	*FES EGFR KRAS*
	1.1%	91	*KLF6 EGFR KRAS*
	1.1%	92	*PAK6 EGFR KRAS*
	1.1%	91	*AURKB EGFR KRAS*
	1.1%	92	*EPHA3 EGFR KRAS*
	1%	92	*TP73L EGFR KRAS*
	1%	91	*TERT EGFR KRAS*
	1%	91	MG *EGFR KRAS*
	1%	92	*CYSLTR2 EGFR KRAS*
	1%	90	*FLT4 EGFR KRAS*

Table 2: Results for Lung mutation data with $k = 2$ and $k = 3$. M is the set of genes, $\tilde{\pi}(M)$ is the frequency of the set in the sample obtained with the MCMC, and $W(M)$ is the weight of the set. MG is a metagene containing: *ACVR2B, MAP2K5, RPS6KA6, EPHA2, FOXO3, STK3, CDC2L2, KSR2, CCNT2, FBXW7*.

k	$\tilde{\pi}(M)$	$W(M)$	M
2	56%	75	*ATM TP53*
	1%	67	*PAK4 TP53*
	1%	66	*IGFR1 TP53*

Table 3: Results for Lung mutation data with $k = 2$ after removing the set (*EGFR, KRAS, STK11*) from analysis. M is the set of genes, $\tilde{\pi}(M)$ is the frequency of the set in the sample obtained with the MCMC, and $W(M)$ is the weight of the set.

D.2 Results for Glioblastoma

Table 4 reports the result obtained for Glioblastoma data with $k = 2$ and $k = 3$. Table 5 reports the results for $k = 2$ obtained after removing the set (*CDKN2B, RB1, CDK4*). Table 6 reports the results for $k = 2$ obtained after removing the sets (*CDKN2B, RB1, CDK4*) and the set (*CDKN2A, TP53*). In all tables k is the size of set, M is the set of genes, $\tilde{\pi}(M)$ is the frequency of the set in the sample obtained with the MCMC, and $W(M)$ is the weight of the set.

k	$\tilde{\pi}(M)$	$W(M)$	M
2	18.4%	54	*CYP27B1 CDKN2B*
	10.9%	53	MG_1 *CDKN2B*
	9.7%	53	*TP53 CDKN2B*
	9.6%	53	*CDKN2A TP53*
	7.2%	52	*EGFR TP53*
	5.8%	52	MG_2 *CDKN2B*
	5.4%	52	*MTAP TP53*
	4.9%	51	*OS9 CDKN2B*
	4.9%	51	*RB1 CDKN2B*
	2.6%	50	*NF1 EGFR*
	1.6%	49	*PTEN CDKN2A*
	1.6%	48	*SEC61G TP53*
	1.4%	49	*DTX3 CDKN2B*
	1.2%	48	MG_3 *CDKN2B*
	1.2%	48	*PTEN MTAP*
3	9.7%	62	*CYP27B1 RB1 CDKN2B*
	5.7%	61	MG_1 *RB1 CDKN2B*
	3.1%	60	MG_2 *RB1 CDKN2B*
	2.1%	59	*OS9 RB1 CDKN2B*
	1.4%	57	*MTAP CYP27B1 RB1*

Table 4: Results for Glioblastoma mutation data with $k = 2$ and $k = 3$. M is the set of genes, $\tilde{\pi}(M)$ is the frequency of the set in the sample obtained with the MCMC, and $W(M)$ is the weight of the set. MG_1, MG_2 and MG_3 are metagenes corresponding to the following sets: MG_1 = (*TSFM, MARCH9, TSPAN31, FAM119B, METTL1, CDK4, CENTG1*); MG_2 = (*AVIL, CTDSP2*); MG_3 = (*SLC26A10, GEFT, PIP4K2C*).

D.3 Results for Known Mutations on Multiple Cancer Type

Table 7 and reports the results obtained on the oncogenes mutations data with $k = 8$. Table 8 and reports the results obtained on the oncogenes mutations data with $k = 10$. In all tables k is the size of set, M is the

k	$\tilde{\pi}(M)$	$W(M)$	M
2	30.1%	53	*CDKN2A TP53*
	19.8%	52	*MTAP TP53*
	18.6%	52	*EGFR TP53*
	6.8%	50	*NF1 EGFR*
	3.9%	50	*PTEN CDKN2A*
	3.4%	49	*SEC61G TP53*
	3%	48	*CYP27B1 CDKN2A*
	2.6%	48	*PTEN MTAP*
	1.7%	46	MG_2 *CDKN2A*
	1.2%	47	*CYP27B1 MTAP*
	1%	47	*OS9 CDKN2A*
	1%	45	*DTX3 CDKN2A*

Table 5: Results for Glioblastoma mutation data with $k = 2$ after removing the set (*CDKN2B, RB1, CDK4*) from analysis. M is the set of genes, $\tilde{\pi}(M)$ is the frequency of the set in the sample obtained with the MCMC, and $W(M)$ is the weight of the set. See caption of Table 4 for the definition of MG_2.

k	$\tilde{\pi}(M)$	$W(M)$	M
2	44.3	50	*NF1 EGFR*
	16.9	48	*PTEN MTAP*
	9.3	47	*CYP27B1 MTAP*
	3.2	45	*AVIL MTAP*
	2.4	44	*PTEN EGFR*
	2.0	43	*IFNA21 PTEN*
	1.8	44	*OS9 MTAP*
	1.3	42	*MTAP NF1*

Table 6: Results for Glioblastoma mutation data with $k = 2$ after removing the sets (*CDKN2B, RB1, CDK4*) and (*CDKN2A,TP53*) from analysis. M is the set of genes, $\tilde{\pi}(M)$ is the frequency of the set in the sample obtained with the MCMC, and $W(M)$ is the weight of the set. See caption of Table 4 for the definition of MG_2.

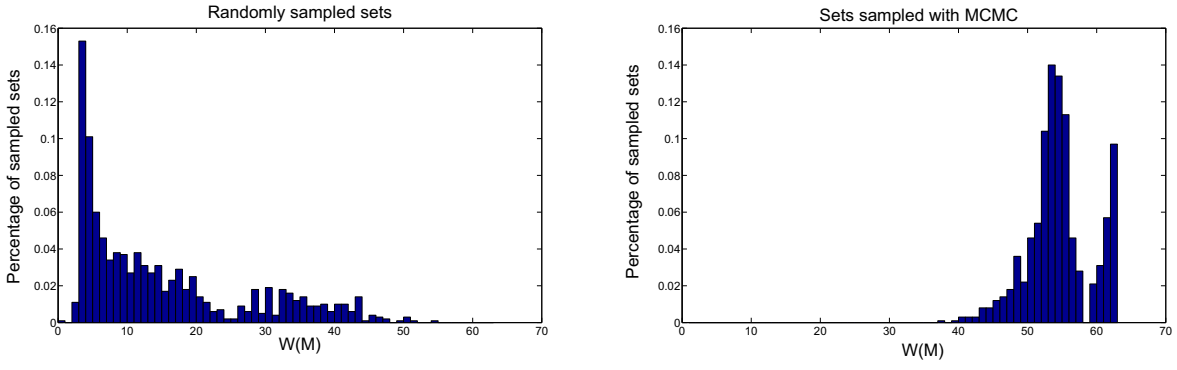
set of genes, $\tilde{\pi}(M)$ is the frequency of the set in the sample obtained with the MCMC, and $W(M)$ is the weight of the set.

D.4 Evaluation of MCMC sampling

To evaluate the MCMC sampling, we compared the distribution of weights $W(M)$ for sets M sampled by MCMC and randomly sampled sets. If the MCMC approach was merely performing a random walk among the sets of size k, the two distributions would be similar. Figure 6 shows sets sampled randomly (a) and by the MCMC approach (b) for sets of size $k = 3$ from the TCGA GBM dataset (The Cancer Genome Atlas Research Network, 2008). The two distributions are clearly different, with the value of $W(M)$ being typically much larger for the sets sampled with the MCMC than for randomly sampled sets.

k	$\tilde{\pi}(M)$	$W(M)$	M
8	13.2	265	*BRAF_600-601 EGFR_ECD EGFR_KD HRAS KRAS NRAS PIK3CA_HD PIK3CA_KD*
	6.4	264	*BRAF_600-601 EGFR_ECD EGFR_KD HRAS KIT KRAS NRAS PIK3CA_HD*
	4.6	263	*BRAF_600-601 EGFR_ECD EGFR_KD HRAS KIT KRAS NRAS PIK3CA_KD*
	4.5	263	*BRAF_600-601 EGFR_ECD EGFR_KD HRAS JAK2 KRAS NRAS PIK3CA_HD*
	4.2	263	*BRAF_600-601 EGFR_ECD EGFR_KD KIT KRAS NRAS PIK3CA_HD PIK3CA_KD*
	4.1	263	*BRAF_600-601 EGFR_ECD EGFR_KD FGFR1 HRAS KRAS NRAS PIK3CA_HD*
	3.2	262	*BRAF_600-601 EGFR_ECD EGFR_KD HRAS JAK2 KRAS NRAS PIK3CA_KD*
	2.8	262	*BRAF_600-601 EGFR_ECD EGFR_KD FGFR3 HRAS KRAS NRAS PIK3CA_HD*
	2.7	262	*BRAF_600-601 EGFR_ECD EGFR_KD FGFR1 KRAS NRAS PIK3CA_HD PIK3CA_KD*
	2.5	261	*BRAF_600-601 EGFR_ECD EGFR_KD HRAS KRAS NRAS PDGFRA PIK3CA_HD*
	2.4	262	*BRAF_600-601 EGFR_ECD EGFR_KD FGFR1 HRAS KRAS NRAS PIK3CA_KD*
	2.1	262	*BRAF_600-601 EGFR_ECD EGFR_KD ERBB2 HRAS KRAS NRAS PIK3CA_HD*
	2.1	261	*BRAF_600-601 EGFR_ECD EGFR_KD FGFR3 HRAS KRAS NRAS PIK3CA_KD*
	2.0	261	*BRAF_600-601 EGFR_ECD EGFR_KD JAK2 KIT KRAS NRAS PIK3CA_HD*
	1.9	261	*BRAF_600-601 EGFR_ECD EGFR_KD FGFR3 KRAS NRAS PIK3CA_HD PIK3CA_KD*
	1.7	261	*BRAF_600-601 EGFR_ECD EGFR_KD ERBB2 HRAS KRAS NRAS PIK3CA_KD*
	1.5	261	*BRAF_600-601 EGFR_ECD EGFR_KD FGFR1 KIT KRAS NRAS PIK3CA_HD*
	1.5	260	*BRAF_600-601 EGFR_ECD EGFR_KD KRAS NRAS PDGFRA PIK3CA_HD PIK3CA_KD*
	1.5	261	*BRAF_600-601 EGFR_ECD EGFR_KD ERBB2 KRAS NRAS PIK3CA_HD PIK3CA_KD*
	1.4	262	*BRAF_600-601 EGFR_ECD EGFR_KD JAK2 KRAS NRAS PIK3CA_HD PIK3CA_KD*
	1.3	260	*BRAF_600-601 EGFR_ECD EGFR_KD FGFR3 KIT KRAS NRAS PIK3CA_HD*
	1.1	260	*BRAF_600-601 EGFR_ECD EGFR_KD ERBB2 KIT KRAS NRAS PIK3CA_HD*

Table 7: Results for oncogenes mutations data with $k = 8$. M is the set of genes, $\tilde{\pi}(M)$ is the frequency of the set in the sample obtained with the MCMC, and $W(M)$ is the weight of the set.



(a) **(b)**

Figure 6: Distribution of number of samples with respect to $W(M)$ for the GBM dataset with $k = 3$ for (a) randomly sampled sets and (b) sets sampled with the MCMC.

k	$\tilde{\pi}(M)$	$W(M)$	M
10	4.8%	272	*BRAF_600-601 EGFR_ECD EGFR_KD FGFR1 HRAS KIT KRAS NRAS PIK3CA_HD PIK3CA_KD*
	3.6%	272	*BRAF_600-601 EGFR_ECD EGFR_KD HRAS JAK2 KIT KRAS NRAS PIK3CA_HD PIK3CA_KD*
	3.4%	271	*BRAF_600-601 EGFR_ECD EGFR_KD ERBB2 HRAS KIT KRAS NRAS PIK3CA_HD PIK3CA_KD*
	3.4%	270	*BRAF_600-601 EGFR_ECD EGFR_KD ERBB2 FGFR1 HRAS KRAS NRAS PIK3CA_HD PIK3CA_KD*
	3.2%	271	*BRAF_600-601 EGFR_ECD EGFR_KD FGFR1 HRAS JAK2 KRAS NRAS PIK3CA_HD PIK3CA_KD*
	3.1%	271	*BRAF_600-601 EGFR_ECD EGFR_KD FGFR3 HRAS KIT KRAS NRAS PIK3CA_HD PIK3CA_KD*
	2.7%	270	*BRAF_600-601 EGFR_ECD EGFR_KD FGFR1 HRAS JAK2 KIT KRAS NRAS PIK3CA_HD*
	2.5%	270	*BRAF_600-601 EGFR_ECD EGFR_KD HRAS KIT KRAS NRAS PDGFRA PIK3CA_HD PIK3CA_KD*
	2.4%	270	*BRAF_600-601 EGFR_ECD EGFR_KD ERBB2 HRAS JAK2 KRAS NRAS PIK3CA_HD PIK3CA_KD*
	2.4%	270	*BRAF_600-601 EGFR_ECD EGFR_KD FGFR1 FGFR3 HRAS KRAS NRAS PIK3CA_HD PIK3CA_KD*
	1.6%	269	*BRAF_600-601 EGFR_ECD EGFR_KD ERBB2 HRAS JAK2 KIT KRAS NRAS PIK3CA_HD*
	1.6%	269	*BRAF_600-601 EGFR_ECD EGFR_KD FGFR1 FGFR3 HRAS KIT KRAS NRAS PIK3CA_HD*
	1.5%	270	*BRAF_600-601 EGFR_ECD EGFR_KD FGFR3 HRAS JAK2 KRAS NRAS PIK3CA_HD PIK3CA_KD*
	1.4%	269	*BRAF_600-601 EGFR_ECD EGFR_KD FGFR1 JAK2 KIT KRAS NRAS PIK3CA_HD PIK3CA_KD*
	1.4%	269	*BRAF_600-601 EGFR_ECD EGFR_KD ERBB2 FGFR1 HRAS KIT KRAS NRAS PIK3CA_HD*
	1.3%	269	*BRAF_600-601 EGFR_ECD EGFR_KD FGFR3 HRAS JAK2 KIT KRAS NRAS PIK3CA_HD*
	1.2%	269	*BRAF_600-601 EGFR_ECD EGFR_KD HRAS JAK2 KRAS NRAS PDGFRA PIK3CA_HD PIK3CA_KD*
	1.1%	268	*BRAF_600-601 EGFR_ECD EGFR_KD FGFR1 HRAS KIT KRAS NRAS PDGFRA PIK3CA_HD*
	1.1%	268	*BRAF_600-601 EGFR_ECD EGFR_KD FGFR1 FGFR3 HRAS JAK2 KRAS NRAS PIK3CA_HD*
	1.1%	268	*BRAF_600-601 EGFR_ECD EGFR_KD ERBB2 HRAS JAK2 KIT KRAS NRAS PIK3CA_KD*
	1.0%	269	*BRAF_600-601 EGFR_ECD EGFR_KD FGFR1 HRAS KRAS NRAS PDGFRA PIK3CA_HD PIK3CA_KD*

Table 8: Results for oncogenes mutations data with $k = 10$. M is the set of genes, $\tilde{\pi}(M)$ is the frequency of the set in the sample obtained with the MCMC, and $W(M)$ is the weight of the set.