SUPPLEMENTARY MATERIAL FOR THE PAPER "A HIERARCHICAL BAYESIAN MODEL FOR INFERENCE OF COPY NUMBER VARIANTS AND THEIR ASSOCIATION TO GENE EXPRESSION"

BY ALBERTO CASSESE[*], MICHELE GUINDANI[†] MAHLET G. TADESSE[‡]
FRANCESCO FALCIANI[§] AND MARINA VANNUCCI[*]

Rice University[*], *MD Anderson Cancer Center*[†], *Georgetown University*[‡] *and University of Liverpool*[§]

MCMC steps. Here we describe our MCMC algorithm in detail. During the update of $\boldsymbol{R}$ (and of $\boldsymbol{\xi}$) we first select a list of gene expression, as rows of $\boldsymbol{R}$ (and a list of samples, as elements of a randomly selected column of ξ) and then update and accept/reject their individual values. For this, we first sample from a geometric distribution with probability p_R (or p_ξ) and add the result to the index of the last selected gene expression (sample). If the resulting index is greater than G (or n), then we discard the new value and stop, otherwise we add the new position to the list of selected gene expression (samples) and draw a new value from the geometric distribution. For the first draw, we simply consider the result as the position to be updated. The updates on η_j, σ_j, for $j = 1, \ldots, 4$, and the transition matrix $\mathbf{A}$ follow Guha et al. (2008), though applied to all samples simultaneously. We recall below equations involved in the MCMC steps:

$$\begin{aligned} \pi(r_{gm}|r_{g(m-1)}, r_{g(m+1)}, \boldsymbol{\xi}) &= \gamma_m \tfrac{\Gamma(e+f)\Gamma(e+r_{gm})\Gamma(f+1-r_{gm})}{\Gamma(e+f+1)\Gamma(e)\Gamma(f)} \\ &+ \textstyle\sum_{j=1}^{2} \omega_m^{(j)} I_{\{r_{gm}=r_{g(m+(-1)^j)}\}}. \end{aligned} \tag{1}$$

$$f(\boldsymbol{Y}_g|\boldsymbol{\xi}, \boldsymbol{R}) = \frac{(2\pi)^{-\frac{n}{2}}(\frac{c_\mu}{c_\mu+n})^{\frac{1}{2}}(c_\beta)^{\frac{k_g}{2}}\Gamma(\frac{n+\delta}{2})(\frac{d}{2})^{\frac{\delta}{2}}}{|\boldsymbol{U}_g|^{\frac{1}{2}}\Gamma(\frac{\delta}{2})(\frac{d+q_g}{2})^{(\frac{n+\delta}{2})}}, \tag{2}$$

Updating R. We give details on how to calculate the probability $\pi(\boldsymbol{R}|\boldsymbol{\xi})$ when updating $\boldsymbol{R}$:

$$\pi(\boldsymbol{R}|\boldsymbol{\xi}) = \prod_{g=1}^{G} \pi(r_{g1}|r_{g2}, \boldsymbol{\xi})\pi(r_{gM}|r_{g(M-1)}, \boldsymbol{\xi}) \prod_{m=2}^{M-1} \pi(r_{gm}|r_{g(m-1)}, r_{g(m+1)}, \boldsymbol{\xi}).$$

When calculating the ratio $\frac{\pi(\boldsymbol{R}^{new}|\boldsymbol{\xi})}{\pi(\boldsymbol{R}^{old}|\boldsymbol{\xi})}$ we need to consider only those quantities whose values change when a single element of $\boldsymbol{R}$ is updated. What follows is the description of the different scenarios that could occur when applying our MCMC update.

- Adding/deleting:
    - If the selected element is not either the first or last CGH probe, three elements change their values (say, for example, that element r_{gm} is selected): $\pi(r_{gm}|r_{g(m-1)}, r_{g(m+1)}, \boldsymbol{\xi})$, $\pi(r_{g(m-1)}|r_{g(m-2)}, r_{gm}, \boldsymbol{\xi})$ and $\pi(r_{g(m+1)}|r_{gm}, r_{g(m+2)}, \boldsymbol{\xi})$.
    - If the selected element is either CGH probe 1 or M, only two quantities change their values:
        - $\pi(r_{g1}|r_{g2}, \boldsymbol{\xi})$ or $\pi(r_{gM}|r_{g(M-1)}, \boldsymbol{\xi})$;
        - $\pi(r_{g2}|r_{g1}, r_{g3}, \boldsymbol{\xi})$ or $\pi(r_{g(M-1)}|r_{g(M-2)}, r_{gM}, \boldsymbol{\xi})$.
- Swapping:
    - Swap between adjacent elements; four quantities change their values (say, for example, that r_{gm} get swapped with $r_{g(m-1)}$):
        - $\pi(r_{g(m-2)}|r_{g(m-3)}, r_{g(m-1)}, \boldsymbol{\xi})$;
        - $\pi(r_{g(m-1)}|r_{g(m-2)}, r_{gm}, \boldsymbol{\xi})$;
        - $\pi(r_{gm}|r_{g(m-1)}, r_{g(m+1)}, \boldsymbol{\xi})$;
        - $\pi(r_{g(m+1)}|r_{gm}, r_{g(m+2)}, \boldsymbol{\xi})$.
    - Swap between "quasi-adjacent" elements, i.e., two elements that are two CGH probes positions apart. Five quantities get involved (say, for example, that r_{gm} get swapped with $r_{g(m-2)}$):
        - $\pi(r_{g(m-3)}|r_{g(m-4)}, r_{g(m-2)}, \boldsymbol{\xi})$;
        - $\pi(r_{g(m-2)}|r_{g(m-3)}, r_{g(m-1)}, \boldsymbol{\xi})$;
        - $\pi(r_{g(m-1)}|r_{g(m-2)}, r_{gm}, \boldsymbol{\xi})$;
        - $\pi(r_{gm}|r_{g(m-1)}, r_{g(m+1)}, \boldsymbol{\xi})$;
        - $\pi(r_{g(m+1)}|r_{gm}, r_{g(m+2)}, \boldsymbol{\xi})$.
    - Swap between all other elements are just an Add and a Delete step.

Note that if the swap involves either CGH probe 1 or M then these quantities reduce by one. Equation (1) is used to calculate all quantities involved in the steps above.

Updating $\boldsymbol{\xi}$. With this update, when calculating the probability $\pi(\boldsymbol{R}|\boldsymbol{\xi})$ we need to look for changes in the values of $\boldsymbol{\gamma}$, $\boldsymbol{\omega}^{(1)}$ and $\boldsymbol{\omega}^{(2)}$. Suppose we change the value of the m-th element, then:

- We need to recalculate $\frac{1}{n}\sum_{i=1}^{n} I_{\{\xi_{im}=\xi_{i(m-1)}\}}$ and $\frac{1}{n}\sum_{i=1}^{n} I_{\{\xi_{im}=\xi_{i(m+1)}\}}$;

- These quantities result in changes in the values of γ_m, $\omega_m^{(1)}$, $\omega_m^{(2)}$, γ_{m-1}, $\omega_{m-1}^{(1)}$, $\omega_{m-1}^{(2)}$, γ_{m+1}, $\omega_{m+1}^{(1)}$, $\omega_{m+1}^{(2)}$;
- We apply equation (1) to calculate the new values of $\pi(r_{gm}|r_{g(m-1)}, r_{g(m+1)}, \boldsymbol{\xi})$, $\pi(r_{g(m-1)}|r_{g(m-2)}, r_{gm}, \boldsymbol{\xi})$ and $\pi(r_{g(m+1)}|r_{gm}, r_{g(m+2)}, \boldsymbol{\xi})$.

Equation (2) is then used to calculate $f(\boldsymbol{Y}|\boldsymbol{\xi}^{new}, \boldsymbol{R})$ and $f(\boldsymbol{Y}|\boldsymbol{\xi}^{old}, R)$, while $f(x_{im}|\xi_{im})$ is simply the density of a $N(\mu_{\xi_{im}}, \sigma^2_{\xi_{im}})$, calculated in the current values of $\mu_{\xi_{im}}$ and $\sigma^2_{\xi_{im}}$.

Next, we focus on the ratio:

$$\frac{\pi(\boldsymbol{\xi}^{new}|\boldsymbol{\xi}^{old}, \boldsymbol{A})q(\boldsymbol{\xi}^{old}|\boldsymbol{\xi}^{new})}{\pi(\boldsymbol{\xi}^{old}|\boldsymbol{\xi}^{old}, \boldsymbol{A})q(\boldsymbol{\xi}^{new}|\boldsymbol{\xi}^{old})},$$

that can be factorized as

$$\prod_{i=1}^{n} \frac{\pi(\xi_{im}^{new}|\xi_{i(m-1)}^{old}, \xi_{i(m+1)}^{old}, \boldsymbol{A})q(\xi_{im}^{old}|\xi_{im}^{new})}{\pi(\xi_{im}^{old}|\xi_{i(m-1)}^{old}, \xi_{i(m+1)}^{old}, \boldsymbol{A})q(\xi_{im}^{new}|\xi_{im}^{old})}.$$

The ratio of interest can be evaluated as $\frac{\pi(\xi_{i(m+1)}^{old}|\xi_{im}^{new}, \boldsymbol{A})}{\pi(\xi_{i(m+1)}^{old}|\xi_{im}^{old}, \boldsymbol{A})}$, when $m \neq M$, and simply as 1 when $m = M$, by noting that $q(\xi_{im}^{new}|\xi_{im}^{old}) = \pi(\xi_{im}^{new}|\xi_{i(m-1)}^{old}, \boldsymbol{A})$, $\frac{\pi(\xi_{im}^{new}|\xi_{i(m-1)}^{old}, \xi_{i(m+1)}^{old}, \boldsymbol{A})}{\pi(\xi_{im}^{old}|\xi_{i(m-1)}^{old}, \xi_{i(m+1)}^{old}, \boldsymbol{A})} = \frac{\pi(\xi_{i(m+1)}^{old}|\xi_{im}^{new}, \boldsymbol{A})\pi(\xi_{im}^{new}|\xi_{i(m-1)}^{old}, \boldsymbol{A})}{\pi(\xi_{i(m+1)}^{old}|\xi_{im}^{old}, \boldsymbol{A})\pi(\xi_{im}^{old}|\xi_{i(m-1)}^{old}, \boldsymbol{A})}$, and considering that we update a single sample, sample i in our example.

Updating η. Let $j = \{1, 2, 3, 4\}$ be the label for the four different states, δ_j be the center of the truncated normal distributions in the prior specification of η_j, n_j be the number of CGH in state j, $\bar{X}_j$ the mean of X's over those CGH probes that are in state j and $\mathbf{I}_j$ denote the support of η_j. Specifically

$$n_j = \sum_{m=1}^{M}\sum_{i=1}^{n} \mathbf{I}_{\{\xi_{im}=j\}}, \quad \bar{X}_j = \frac{1}{n_j}\sum_{m=1}^{M}\sum_{i=1}^{n} X_{im}\mathbf{I}_{\{\xi_{im}=j\}}.$$

The posterior probability for η is:

$$\pi(\eta_j|X, rest) \sim N(\nu_j, (\theta_j^2)^{-1})\mathbf{I}_j$$

where $\theta_j = \tau_j^{-2} + n_j\sigma_j^{-2}$ and $\nu_j = \theta_j^{-2}(\delta_j\tau_j^{-2} + \bar{X}_j n_j \sigma_j^{-2})$.

Updating σ^2. Let $j = \{1, 2, 3, 4\}$ be the label for the four different states, and $\mathbf{I}_j$ denote the support of σ_j^2, the posterior probability for σ is:

$$\pi(\sigma_j^2|X, rest) \sim IG(b_j + \frac{n_j}{2}, l_j + \frac{V_j}{2})\mathbf{I}_j$$

where $V_j = \sum_{m=1}^{M} \sum_{i=1}^{n} (X_{im} - \mu_j)^2 \mathbf{I}_{\{\xi_{im}=j\}}$.

Updating A. Let's focus on a single row of the transition matrix $\mathbf{A}$, then the distribution of the states arises from a multinomial distribution (except for the first element of each sample), and the prior distribution of any row of the matrix is $Dir(\phi_1, \phi_2, \phi_3, \phi_4)$. We follow Guha et al. (2008) and generate a proposal $\boldsymbol{A}^{new}$ from the distribution $a_h|rest \sim Dir(\phi_1 + o_{h1}, \phi_2 + o_{h2}, \phi_3 + o_{h3}, \phi_4 + o_{h4})$, ignoring the marginal distribution of state $\boldsymbol{\xi_1}$. We then accept the proposal with probability $\min[1, \prod_{i=1}^{n} \frac{\pi_{A^{new}}(\xi_{i1})}{\pi_{A^{old}}(\xi_{i1})}]$, where π_A denotes the stationary distribution of the transition matrix $\boldsymbol{A}$.

Additional results for the simulations. For simulated scenario 1 ($\sigma_\epsilon = .1$) with the independent prior, the empirical transition matrix corresponding to the simulated data and the estimated transition matrix were, respectively,

$$\begin{bmatrix} 0.3513 & 0.6183 & 0.0216 & 0.0088 \\ 0.1101 & 0.7762 & 0.1118 & 0.0019 \\ 0.0089 & 0.6209 & 0.3267 & 0.0436 \\ 0 & 0.6000 & 0.0586 & 0.3414 \end{bmatrix} \qquad \begin{bmatrix} 0.3264 & 0.6417 & 0.0217 & 0.0102 \\ 0.1027 & 0.7826 & 0.1038 & 0.0109 \\ 0.0079 & 0.6202 & 0.3023 & 0.0695 \\ 0.0008 & 0.6022 & 0.0461 & 0.3508 \end{bmatrix}$$

with the empirical transition matrix obtained by counting the number of changes from state i to state j that occur between adjacent positions in the true matrix $\boldsymbol{\xi}$. We note that in our simulation the data generating mechanism for $\boldsymbol{\xi}$ is based on randomly selecting L columns, with some stretches of adjacent columns, therefore violating the stationarity assumption of the HMM chain. The results above, jointly with those shown in Section 4.2, suggest that our estimates are robust even in cases where the stationary assumption of the HMM is violated.

For simulated scenario 2, we generated the data by fixing the error variance σ_ϵ to a same value for every gene g, even though our proposed model does allow each gene to have its own variance. This is not restrictive, as, with real data, one can always perform the analysis on standardized data, i.e., with $\sigma_\epsilon = 1$. We did however perform an additional simulation where we generated the data using standard deviations $\sigma_{\epsilon g}$ that vary with g. These were chosen by randomly selecting 100 genes from the set used in the case study and calculating their raw s.d.'s. These values were constrained to be in the range $[.1, .5]$, to facilitate the comparison with the simulation settings reported in the paper. Using an FDR threshold of .05, our model with a dependent prior resulted in specificity $= .99996$ and sensitivity $= .95$ for $\alpha = 10$ and specificity $= .99998$ and sensitivity $= .85$ for $\alpha = 50$, in line with the results from previous simulations. Inference on the HMM parameters was also comparable to the other simulated settings.

Additional results for the case study. Below we present a table that contains functions of mutations linked to the target pathways identified in our analysis (see Section 5 of the paper).

Table 1: Functions of mutations linked to the target pathways identified in our analysis (see Section 5 of the paper). Information extracted from the web based resource GeneCards (*http://www.genecards.org*).

Official gene symbol	**Name**	**Function**
MTERFD1	MTERF domain containing 1	Mitochondrial transcription termination factor Binds promoter DNA and regulates initiation of transcription. Required for normal mitochondrial transcription, and for normal assembly of mitochondrial respiratory complexes. Required for normal mitochondrial function
PTK2B	PTK2B protein tyrosine kinase 2 beta	Related adhesion focal tyrosine kinase. Non-receptor protein-tyrosine kinase that regulates reorganization of the actin cytoskeleton, cell polarization, cell migration, adhesion, spreading and bone remodeling. Plays a role in the regulation of the humoral immune response, and is required for normal levels of marginal B-cells in the spleen and normal migration of splenic B-cells. Required for normal macrophage polarization and migration towards sites of inflammation. Regulates cytoskeleton rearrangement and cell spreading in T-cells, and contributes to the regulation of T-cell responses. Promotes osteoclastic bone resorption
DEFA5	defensin, alpha 5, Paneth cell-specific	Has antimicrobial activity against Gram-negative and Gram-positive bacteria. Defensins are thought to kill microbes by permeabilizing their plasma membrane

Continued on next page

Table 1 – continued from previous page

Official gene symbol	Name	Function
NPM2	nucleophosmin/ nucleoplasmin, 2	Core histones chaperone involved in chromatin reprogramming, specially during fertilization and early embryonic development
LRP12	low density lipoprotein-related protein 12	This gene encodes a member of the low-density lipoprotein receptor related protein family. The product of this gene is a transmembrane protein that is differentially expressed in many cancer cells. Alternate splicing results in multiple transcript variants
PPP1R3B	protein phosphatase 1, regulatory (inhibitor) subunit 3B	This gene encodes the catalytic subunit of the serine/theonine phosphatase, protein phosphatase-1. The encoded protein is expressed in liver and skeletal muscle tissue and may be involved in regulating glycogen synthesis in these tissues. This gene may be a involved in type 2 diabetes and maturity-onset diabetes of the young. Alternate splicing results in multiple transcript variants that encode the same protein
MTUS1	mitochondrial tumor suppressor 1	This gene encodes a protein which contains a C-terminal domain able to interact with the angiotension II (AT2) receptor and a large coiled-coil region allowing dimerization. Multiple alternatively spliced transcript variants encoding different isoforms have been found for this gene. One of the transcript variants has been shown to encode a mitochondrial protein that acts as a tumor suppressor and partcipates in AT2 signaling pathways. Other variants may encode nuclear or transmembrane proteins but it has not been determined whether they also participate in AT2 signaling pathways

Continued on next page

Table 1 – continued from previous page

Official gene symbol	Name	Function
NUDCD1	NudC domain containing 1	Chronic myelogenous leukemia tumor antigen 66Isoform 1 is the dominant immunogenic isoform and is capable of eliciting a humoral response in individuals with a variety of solid tumors. Expression of isoform 1 in a wide variety of malignancies as well as the presence of an immunogenic epitope suggest that it may be a suitable target for antigen-specific immunotherapy
RIMS2	regulating synaptic membrane exocytosis 2	Rab effector involved in exocytosis. May act as scaffold protein regulating synaptic membrane exocytosis protein 2
OTUD6B	OTU domain containing 6B	Deubiquitinating enzymes (DUBs; see MIM 603478) are proteases that specifically cleave ubiquitin (MIM 191339) linkages, negating the action of ubiquitin ligases. DUBA5 belongs to a DUB subfamily characterized by an ovarian tumor (OTU) domain
RP1	retinitis pigmentosa 1 (autosomal dominant)	Microtubule-associated protein regulating the stability and length of the microtubule-based axoneme of photoreceptors. Required for the differentiation of photoreceptor cells, it plays a role in the organization of the outer segment of rod and cone photoreceptors ensuring the correct orientation and higher order stacking of outer segment disks along the photoreceptor axoneme
LPL	lipoprotein lipase	Lipoprotein lipase (LPL), like LIPG, is a vascular lipase, however it is not synthesized in endothelial cells. It is anchored to the capillary endothelium by proteoglycans and catalyzes the hydrolysis of triglycerides to release free fatty acids into the circulation. LPL therefore initiates the processing of triglyceride-rich lipoproteins such as chylomicrons and VLDL.
CSMD1	CUB and Sushi multiple domains 1	CSMD1 is a novel multiple domain complement-regulatory protein highly expressed in the central nervous system and epithelial tissues.

Continued on next page

Table 1 – continued from previous page

Official gene symbol	Name	Function
INTS9	integrator complex subunit 9	INTS9 is a multiprotein mediator of small nuclear RNA processing that associates with the C-terminal repeat of RNA polymerase II. It is required for Cell cycle progression but not cell growth.
RAB2A	RAB2A, member RAS oncogene family	the RAB2 protein is a resident of pre-Golgi intermediates and is required for protein transport from the endoplasmic reticulum to the Golgi complex. They found that RAB2 is essential for the maturation of pre-Golgi intermediates.
TG	thyroglobulin	hyroglobulin provides 3 things: a thyroid hormone precursor, storage of iodine, and storage of inactive thyroid hormones.
CSGALNACT1	chondroitin sulfate N-acetyl-galactosaminyl-transferase 1	TG expression was decreased in thyroid carcinomas but was normal in the other tissues. TSHR expression was normal in most tissues studied and was decreased in only some thyroid carcinomas. In thyroid cancer tissues, a positive relationship was found between the individual levels of expression of NIS, TPO, TG, and TSHR.
TPD52	tumor protein D52	D52 was expressed at significant levels in some breast carcinomas but at much lower levels in breast fibroadenomas.
ASH2L	ash2 (absent, small, or homeotic)-like (Drosophila)	in yeast, the HCF1-associated human SET1/ASH2 HMT complex possesses histone H3-K4 methylation activity, which activates transcription.
DPYS	dihydropyrimidinase	Dihydropyrimidinase (DPYS), also known as 5,6-dihydropyrimidine amidohydrolase, or DHP; (EC 3.5.2.2), is the second enzyme in the 3-step degradation pathway of uracil and thymine after the action of dihydropyramidine dehydrogenase

Continued on next page

Table 1 – continued from previous page

Official gene symbol	Name	Function
CYP7B1	cytochrome P450, family 7, subfamily B, polypeptide 1	The synthesis of primary bile acids from cholesterol occurs via 2 pathways: the classic neutral pathway involving cholesterol 7-alpha-hydroxylase (CYP7A1; 118455), and the acidic pathway involving a distinct microsomal oxysterol 7-alpha-hydroxylase (CYP7B1)

References.

S. Guha, Y. Li, and D. Neuberg. Bayesian hidden Markov modelling of array cgh data. JASA, 103: 485–497, 2008.

DEPARTMENT OF STATISTICS
RICE UNIVERSITY
HOUSTON, TEXAS 77005
USA,
E-MAIL: Alberto.Cassese@rice.edu
marina@rice.edu

DEPARTMENT OF BIOSTATISTICS
MD ANDERSON CANCER CENTER
HOUSTON, TEXAS 77030
USA,
E-MAIL: mguindani@mdanderson.org

DEPARTMENT OF MATHEMATICS AND STATISTICS
GEORGETOWN UNIVERSITY
WASHINGTON, DC 20057
USA,
E-MAIL: mgt26@georgetown.edu

CENTER OF COMPUTATIONAL BIOLOGY AND MODELLING (CCMB)
INSTITUTE OF INTEGRATIVE BIOLOGY
UNIVERSITY OF LIVERPOOL
LIVERPOOL
UK,
E-MAIL: f.falciani@liverpool.ac.uk