Title	An integrative bioinformatic approach for identifying subtypes and subtype-specific drivers in cancer
Author(s)	Chen, P; Hung, YS; Fan, Y; Wong, STC
Citation	The 2012 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB'12), San Diego, CA., 9-12 May 2012. In IEEE CIBCB Proceedings, 2012, p. 169-176
Issued Date	2012
URL	http://hdl.handle.net/10722/165157
Rights	IEEE Symposium on Computational Intelligence and Bioinformatics and Computational Biology Proceedings. Copyright © IEEE.

An Integrative Bioinformatics Approach for Identifying Subtypes and Subtype-specific Drivers in Cancer

Peikai Chen and Y. S. Hung
The University of Hong Kong
Hong Kong, China
Emails: {pkchen,yshung}@eee.hku.hk

Yubo Fan and Stephen T.-C. Wong
The Methodist Hospital Research Institute
Houston, TX, USA 77030
Emails: {yfan,stwong}@tmhs.org

***Abstract*—Cancer is a complex disease and within a cancer, subtypes of patients with distinct behaviors often exist. The subtypes might have been caused by different hits, such as copy number aberrations (CNAs) and point mutations, on different pathways/cells-of-origin in a common tissue/organ. Identifying the subtypes with subtype-specific drivers, i.e., hits, is key to the understanding of cancer and development of novel treatments.**

Here, we report the development of an integrative method to identify the subtypes of cancer. Specifically, we consider CNAs and their impact on gene expressions. Based on these relations, we propose an iterative approach that alternates between kernel based gene expression clustering and gene signature selection. We applied the method to datasets of the pediatric cancer medulloblastoma (MB). The consensus number of clusters quickly converges to three; and for each of these three subtypes, the signature detection also converges to a consistent set of a few hundred highly functionally related genes.

For each of the subtypes, we correlate its signature with the set of within-subtype recurrent CNA-affected genes for identifying drivers. The top-ranked driver candidates are found to be enriched with known pathways in certain subtypes of MB as well as containing novel genes that might reveal new understandings for other subtypes.

I. Introduction

Cancer is initiated and driven by aberrant genetic events (also known as *hits*), such as copy number aberrations (CNAs). While not all hits cause cancer, those that do are called the drivers of cancer. Further, quite a few cancers, such as breast cancer [1], glioblastoma [2] and medulloblastoma [3], are confirmed to contain subtypes. The subtybes of a cancer often demonstrate inter-subtype difference and within-subtype homogeneity in molecular profiles (e.g., gene expressions and CNAs) and clinical outcomes (e.g., drug responses and survival rates). Different subtypes may have arisen because of different mechanisms, such as the hits on different pathways and/or different cells-of-origin [4] within the same tissue/organ. Classifying the patients of a cancer into appropriate subtypes is the key to uncover the drivers of these mechanisms.

There is a large body of literature dedicated to the development of supervised [5], semi-supervised [6] or unsupervised [7] approaches for the discovery of classes within a cancer dataset. However, there are major differences between class discovery and cancer subtyping:

- The levels of data for analysis are different. In class discovery, typically one level of data (e.g., gene expression data, copy numbers/sequencing data or clinical data) is used, whereas in subtyping, a combination of these datasets may need to be considered. The reason is that the establishment of a cancer subtype requires evidences at various levels of behaviors.
- The validation criteria are different. In class discovery, breadth-first cross-dataset validation is a golden measure to test the robustness of a method and the patterns it identifies, whereas in subtyping both depth-first consistency validation among the various levels of information and cross-dataset validation are important.
- The causal models are different. In class discovery, normally no conceptual models are built and the training and testing datasets are assumed to be naturally occurring and identically independent. In subtyping, since multiple-level datasets are used, conceptual models about the relations of these datasets are often necessary and should soundly reflect the underlying operations of biological systems.

Particularly, one type of genetic events, CNAs, is widely found in the cancer genomes [8] and they occur at the DNA level, which is on the upper stream of gene expressions as dictated by the central dogma of biology. CNAs are also found to be positively correlated with the raw expressions of affected genes [3]. In some cancer, such as medulloblastoma, the CNA patterns are also found to be subtype-dependent [9]. This raises an important question: *how do CNAs affect subtyping results?*

There is a possibility that the clustering of gene expressions is merely the consequences of subtype-dependent CNA patterns, instead of the result of a number differentially expressed genes (DEGs). If this is the case, clustering by the CNA patterns may be a more fundamental way of performing cancer subtyping. And the DEGs thus detected may only represent the mechanic responses of CNAs but otherwise contain little biological functions that may help trace the tumorigenic drivers/pathways.

If, on the other hand, clustering of gene expressions is more dependent on the set of DEGs than on the CNA-affected genes (which may or may not be DEGs), then it means that: (1) these

978-1-4673-1191-5/12/$31.00

DEGs are functional; (2) most of the CNAs are passenger events, although there could be a small number of tumorigenic drivers; (3) the DEGs may be caused by the tumorigenic CNAs, or any other hits, and (4) the impact of CNAs on gene expressions need to be removed in order to study the gene-gene interactions.

Towards this end, we propose an integrative approach to perform subtyping and driver-identification based on the study of the CNA-expression relation. This paper is organized as below. Section II explores the 'egg-and-chicken' relation between CNAs and gene expressions, and how it affects the clustering-based subtyping result. Based on these findings, Sections III and IV discuss the gene-signature based iterative subtyping approach and the PCA based driver identification, respectively. In Section V, we apply the algorithm to datasets of a cancer, medulloblastoma, and compare our findings with other methods. Section VI discusses the results of the algorithms.

II. A Preliminary Study of the CNA-expression Relation and Its Impact on Clustering

A total number of 75 medulloblastoma cases with matched gene expression and copy number data (by SNP arrays) were obtained from Taylor *et al.* [10] (GEO: GSE21166 and GSE14437). Another 11 gene expressions of normal cerebellum were also obtained from Cho *et al* [9] (GEO: GSE19399).

A kernel-based spectral clustering [11] was applied to the top 5% genes with largest variance. Fig. 1A shows the clustering result. Gap-statistic [12] (Fig. 1B) estimates the most appropriate number of clusters to be 3. The copy number landscapes by SNP arrays for these 75 samples are shown in Fig. 3.

To study the global impact of copy numbers on affected genes, we estimated the copy numbers of individual genes for each of the core samples. The copy number for a gene is estimated based on the average copy number states of all SNP probesets falling within a certain neighborhood ($\leq$ 30kbp, say) of that gene. Fig. 1C shows the boxplot of raw expressions (in $\log_2$ scale) with respect to their copy number states. An analysis of variance (ANOVA) indicates that there is very significant linear correlations (F-statistic: 336.87 and p-value $< 2.2 \times 10^{-16}$). Since the raw expressions represent a superimposed effects of multiple factors, e.g., cross-binding, functional regulation and copy number responses etc., we subtract the gene expression values of a gene by a background signal. This background signal is estimated based on this gene's average expression in the normal tissue (e.g., the normal cerebellum). The raw signal with background subtracted is called relative signal. By doing so, it is hoped that the cross-binding effects would be reduced and the copy number effects enhanced. Fig. 1D shows the boxplot of relative expressions (in $\log_2$ scale) with respect to their copy number states. An ANOVA reveals that there is a even more significant linear correlation (F-statistic: 1884.8, same degree of freedom; p-value $< 2.2 \times 10^{-16}$).

To further study the effects of copy numbers on gene expressions, the relative expressions are plotted w.r.t. to their

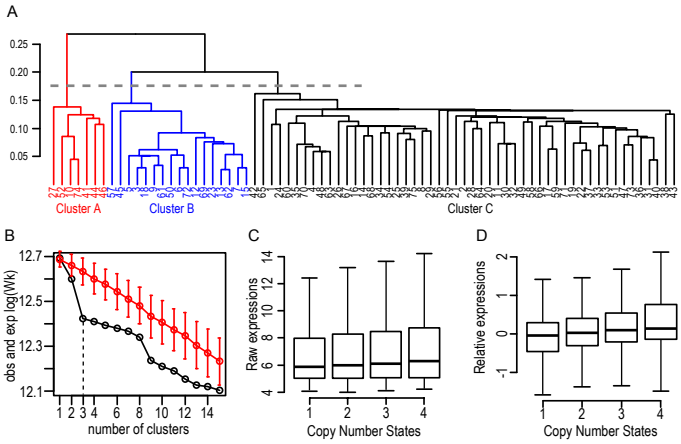

Fig. 1. A. A dendrogram showing the spectral clustering on the core samples. B. The gap-statistic indicates the first elbow at 3, i.e., the most appropriate number of clusters determined to be 3. The red error bars indicate the null intervals. The gray dashed line shown in A refers to the splitting of the dendrogram into three clusters. C. Boxplot of global relation between raw expressions and corresponding copy number states. D. Boxplot of relative expressions against copy number states.

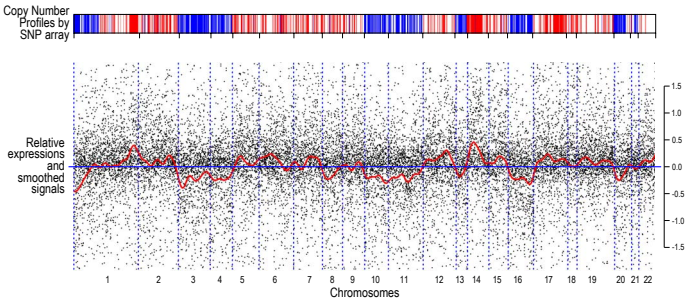

Fig. 2. The copy number profiles of a specific example (ID:73) by SNP arrays and this patient's position-dependent relative expressions. In the upper panel, red lines correspond with copy number gains and blue lines correspond with copy number losses. In the lower panel, the dots are the relative expression values, and the red curve is the smoothed signal.

chromosomal positions. A Nadaraya-Watson algorithm [13] with a Gaussian kernel (with a window width of 300 points) was used to smooth the relative expressions. Fig. 2 shows the result. It can be seen that there is strong consistency between the copy number states and the relative expressions in regions with CNAs. Given such significant responses of expressions due to CNAs, it might be reasonable to assume that most of the affected expressions may be only mechanic responses but otherwise not involved in the cancerous process.

Note that the copy number landscapes in Fig. 3 seem to suggest that the copy number features are dependent on the clusters. For example, Cluster A seems to be characterized with Chr6 deletions while the Cluster C seem to be characterized with CNAs on Chr17. An immediate question is whether the dendrogram in Fig. 1A is a consequence of the CNA patterns in Fig. 3, rather than the cause of it. For convenience, the former situation is denoted as H_0, while the latter is denoted as H_1.

As discussed in the introduction, the answer to this question

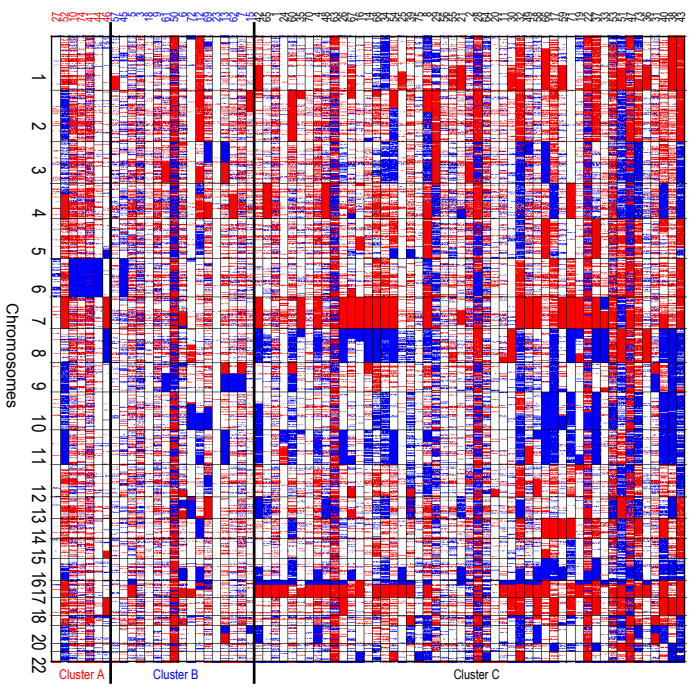

Fig. 3. The copy number profiles of the 75 core samples by SNP arrays. Red, copy number gains; blue, copy number losses.

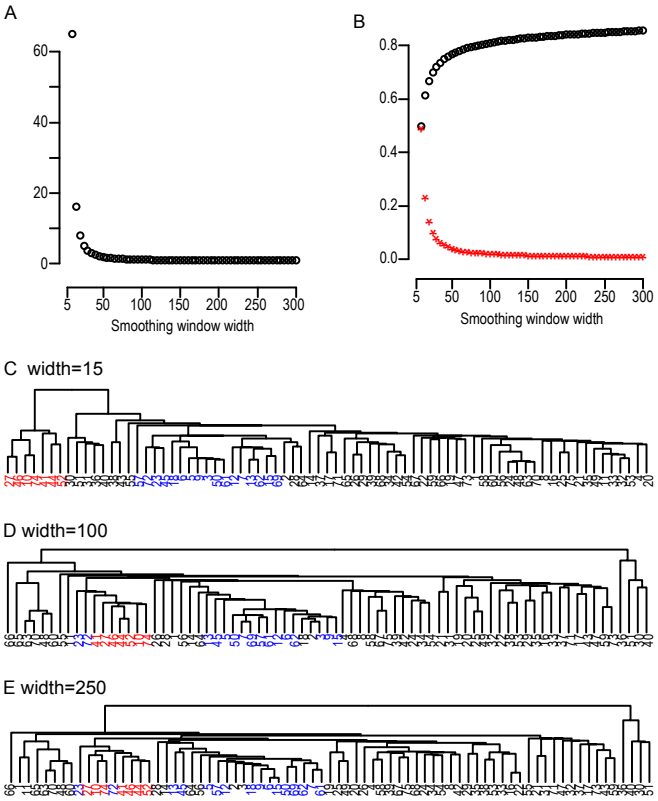

Fig. 4. The clustering stability metrics as smoothing window width varies and corresponding impact on clustering. A. M_1^s from $s = 2$ to $s = 61$, as the window width increases by 5 in each step, from 5 to 300. B. Black, M_2^s; red, M_3^s. C-E. The dendrograms under different smoothing parameters.

is important because it dictates which level of data is more fundamental for performing subtyping, and hence determine the success or failure of a subtyping method. To answer this question, note the smoothed signal in Fig. 2. If H_0 holds, instead of using the relative expressions (note that either raw or relative expression will produce the same dendrogram as Fig. 1A), the position-dependent smoothed signal of it, might produce a similar dendrogram. Furthermore, as the smoothing window's width increases, say from a few points to a few hundred points, the smoothed signal would represent more of the CNA responses than other functional effects. As a result, the clustering of samples would remain largely stable, i.e., the relative positions among the samples shall remain stable as the smoothing enhances. To test this, we devise a metric to measure the relative position of samples.

Given N samples, and a distance matrix $D \in \mathbb{R}^{N\times N}$, where $D^{i,j}$ denotes the distance between samples i and j. Suppose we vary the smoothing window width small enough such that $D^{i,j}$ also varies only by a small amount $\Delta^{i,j}$ in two consecutive steps. The average change of distance from step s to (s+1), i.e., $E\{|D_{s+1}^{i,j} - D_s^{i,j}|,\ \forall\, i,j\} \triangleq E\{|\Delta_s|\}$, can be used to indicate the degree of clustering stability. To broaden the scope, two more metrics are devised. For simplicity, they are denoted as:

$$\begin{cases} M_1^s = E\{|\Delta_s|\} \\ M_2^s = E\{|D_{s+1}^{i,j} - D_1^{i,j}|/D_1^{i,j}\} \\ M_3^s = E\{|D_{s+1}^{i,j} - D_s^{i,j}|/D_s^{i,j}\} \end{cases} \quad (1)$$

where $D_1^{i,j}$ refers to the between-sample distance before smoothing. The same clustering as before was applied to the smoothed signals. Fig. 4A-B show these metrics as the smoothing window varies.

It can be seen that when smoothing window is close to 0 (0 indicates the un-smoothed data), there are large changes of relative positions. But as the window width increases, its impact on the relative position reduces significantly. This is confirmed by the resulting dendrograms in Fig. 4C-E. Note that a slight smoothing in Fig. 4C quickly distort the cluster boundaries as compared with Fig. 1A. But as the smoothing enhances, the dendrograms remain quite stable. Note that throughout the procedure, even though the cluster boundaries are no longer consistent with that by the un-smoothed data, the local positions seem to remain stable. Particularly, the Cluster A samples (red) remain close to each other even when the smoothing is strong. So are the Cluster B samples (blue). This means that the mechanic responses of CNAs alone have a major impact on the clustering, and hence subtyping result.

Nevertheless, the smoothing almost completely obscures the cluster boundaries, which means the high frequency components of expressions might have played a more important role in establishing the subtypes. Since a gene is regulated by an unknown number of factors, including CNAs, and the smoothing used here is position-dependent, impact of other non-position-dependent regulators (e.g., point mutations) will be eliminated. But these factors, together with some of the

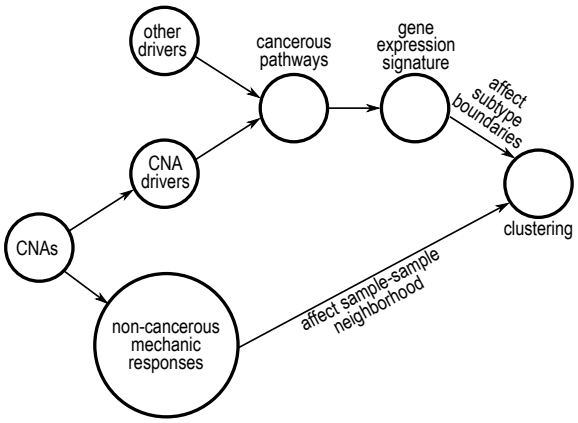

Fig. 5. The role of CNAs in cancer, subtypes and clustering of expression data.

CNAs, are both possible drivers of cancer. Note that some samples in Fig. 3 do not have any CNAs at all, but may belong to separate clusters. A reasonable theory is that, the CNAs within a sample do harbor some tumorigenic drivers, which activate certain pathways in certain cells-of-origin [14], and results in subtypes. And CNAs may not be the only possible factor to activate these pathways. The samples in Fig. 3 without obvious CNAs may contain other aberrations that hit the same pathways as other samples in the cluster. From this study, it appears that neither H_0 nor H_1 is accuate to describe the relations among CNAs, gene expressions and subtypes. Instead, the relations are summarized in Fig. 5.

Fig. 5 indicates that gene signature, i.e., subtype-specific DEGs might be more directly related to the subtype mechanisms. And gene signature is more important in determining the class boundaries, whereas CNA-induced responses may only affect the localized relation of a sample, i.e., what is in the neighborhood of a sample. Therefore, performing cancer subtyping by signature selection may be closer to unveil the common mechanism for a subtype. The next section elaborates on a gene signature based subtyping algorithm.

III. The Subtyping Algorithm

As described above, gene expression signature underpins the biological processes, and is usually enriched with known canonical pathways, which may help improve the interpretability of each subtype. We therefore propose to integrate signature detection into the subtyping process. A general framework is proposed as below:

I. Given an expression dataset $X = \{X_{i,j} | i = 1..N, j = 1..J\}$ with N samples and J genes, divide X into two mutually-exclusive and equally-sized datasets, X_1 and X_2, such that X=[X_1 X_2].
II. Perform clustering on X_1 and determine the number (say, K) of clusters in X_1, dividing the samples into K subgroups.
III. Use the trained class labels of X_1 by Step II to predict the class label for each sample in X_2.
IV. Detect subtype gene signature Φ_k for subtype k of X_2 based on the predicted class labels in Step III.
V. Let $\Phi = \Phi_1 \cup \Phi_2 ... \cup \Phi_K$. Update X_2 with $\hat{X}_2$, where each row in $\hat{X}_2$ corresponds to a signature gene in Φ.
VI. Perform Step II on $\hat{X}_2$ and use the training results to obtain a updated $\hat{X}_1$ and updated signature Φ.
VII. Repeat Steps II to VI, till certain convergence criterion, (e.g. stability of the signature genes) is reached.

The above framework assumes that there are no noisy and/or outlier samples in the dataset, which may not be true. Noisy and/or outlier samples tend to have huge impacts on the clustering results. A strategy to handle this problem is to detect such samples and retain them from the training samples.

Besides outliers, other issues including the clustering, signature detection, training, testing and convergence shall be discussed in the follows.

A. Outlier Detection

Most clustering algorithms are extremely sensitive to outliers, which may lie in-between clusters or lie beyond all clusters. Here, the outliers are identified in two steps. First, a one-class outlier detection is performed using the Mahalanobis distance method [15]. This step efficiently removes samples that are far from any clusters. These samples might have been incorrectly diagnosed or labeled. Second, samples that lie in-between clusters are detected by consensus clustering [16]. This category of outliers might have been due to noisy measurements but not mis-labeling.

In the first step, a sample that has large Mahalanobis distance tends to be an outlier and its squared value can be approximated by a χ^2_p distribution, where p is the dimension of the data. A major challenge here is that there tend to be more genes than samples, making the covariance matrix rank-deficient and hard to determine. To handle this, usually a singular value decomposition (SVD) is performed and only the first few principal components are selected. In the second step, in-between cluster outliers are prone to misclassification after repetitive re-sampling. This can be easily spotted by the consensus clustering method. Outliers detected by both methods are retained for subsequent validation and the remaining sample are used for training the subtype patterns.

B. A Consensus Algorithm for Clustering and Determining the Number of Clusters

There is a large body of literature in unsupervised clustering. But two major issues in gene expression subtyping post challenge on the stability of the clustering results. First, samples within a subtype may not be multivariate Gaussian in the gene expression space, but may be of rather irregular geometries. Second, the densities vary from subtype to subtype, with some samples lying in-between clusters.

An efficient method to handle the within-cluster and between-cluster variability by consensus clustering was proposed by Monti *et al.* [16]. The algorithm works by checking–under a number of clusters K enumerating from 2 to certain upper limit–how stable two samples are within the same cluster

after a large number of re-samplings. The degree of how much two samples i and j stay in the same cluster is measured by the entry of a consensus matrix, $\mathcal{M}(i,j)$. Since the matrix $\mathcal{M}$ is similar with the distance matrix in hierarchical clustering, a meta-algorithm to perform clustering based on $\mathcal{M}$ is used to determine the *consensus clusters*. While this algorithm has the advantage of uncovering cluster patterns in the face of sample variability, it also has the advantages that the number of samples needs to pre-specified and enumerated, and that the re-sampling scheme needs to be carefully devised.

Here, we propose a modified consensus clustering, where the exhaustive enumeration is avoided, the re-sampling scheme is optimized and the clustering metric is topology-based. Specifically, given a training dataset X_1 with n samples, the following steps are taken:

a. Perform sampling with replacement from X_1 by H times to obtain $X_1^{(1)}, ..., X_1^{(H)}$. As a result of this scheme, some samples may be duplicated. For each re-sampled dataset, $X_1^{(h)}$, only the set of unique samples, denoted as $\tilde{X}_1^{(h)}$, is used for clustering. However, the number of duplicates for each sample in $\tilde{X}_1^{(h)}$ is recorded for later use.
b. The spectral clustering [11] used in Section II is applied to each dataset $\tilde{X}_1^{(h)}$. Instead of traversing from $K = 2$ to K equals a certain number, the most appropriate number of clusters within $\tilde{X}_1^{(h)}$ is determined by the gap statistic method [12].
c. If two samples i and j (i, j=1,...,n) both appear in $\tilde{X}_1^{(h)}$ and they belong to the same cluster, then the corresponding entry of a h-dependent matrix $\mathcal{M}^{(h)}(i,j)$ can be evaluated with $\mathcal{M}^{(h)}(i,j) = c_i c_j$. Here, c_i and c_j are the numbers of repeats for each sample in $\tilde{X}_1^{(h)}$ in Step a., respectively. If i and j are not in the same cluster, or not simultaneously selected by re-sampling, $\mathcal{M}^{(h)}(i,j) = 0$.
d. The consensus matrix is evaluated as: $\mathcal{M}(i,j) = \sum_h \mathcal{M}^{(h)}(i,j)$.

In the above steps, if $\mathcal{M}(i,j) = 0$, it means i and j never appear to be under the same cluster. Otherwise, the larger the value $\mathcal{M}(i,j)$, the more likely i and j actually belong to the same cluster. Further, suppose the average of the most appropriate number of clusters for all H re-samplings is K, presumably this is also consensus number of clusters for the original dataset X_1. Hierarchical clustering can then be performed on $1/\mathcal{M}(i,j)$ and the resulting dendrogram is cut with K clusters. As a result, each sample in the training dataset is now assigned a trained class label $k \in \{1, ..., K\}$.

C. Gene Signature Detection

Step III. of the framework proposes to predict the class labels of the testing set X_2 based on the trained results of X_1. To implement this step, we use the adaptive boosting (AdaBoost) algorithm [17], [18], which was shown to outperform linear classifiers (e.g., SVM) in the case of irregular sample spaces common in biomedical data [19]. Specifically, a real number two-class AdaBoost is used. To predict samples in X_2 that are in cluster k, all samples with trained label k in X_1 are regarded as the $\{-1\}$ class, while all other samples (which may include non-cancer controls) in X_1 are regarded as the class $\{+1\}$. The two classes of samples are then used in training AdaBoost. Each sample in X_2 is predicted with a real value $\eta \in (-1, +1)$, indicating how much likely it is in the k cluster. The closer η is to -1, the more likely it is in the k cluster. In the case a sample is predicted to be in more than one cluster, it is further predicted to be the cluster for which η is smallest. This step assign a testing class label $k \in \{1, ..., K\}$ for each sample in X_2.

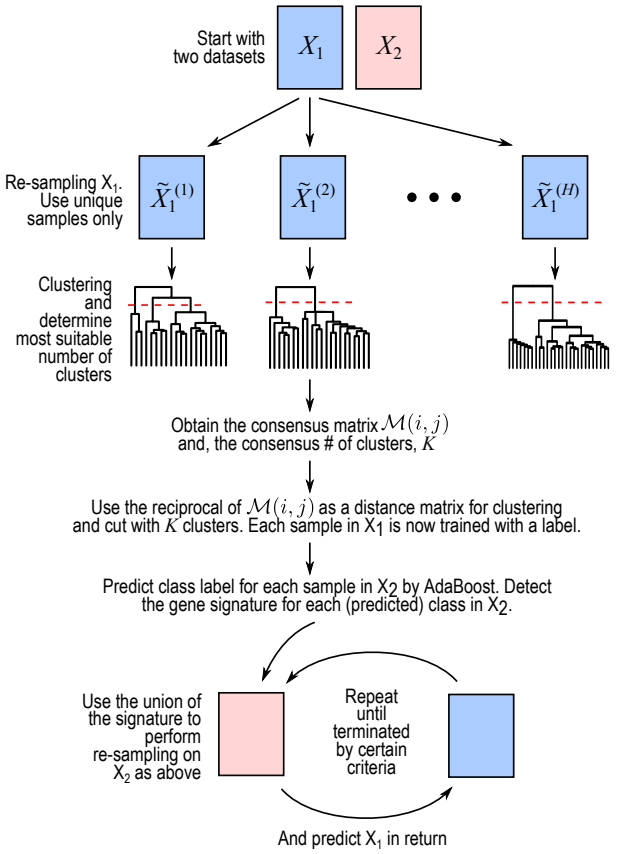

Fig. 6. Overview of the integrative subtyping algorithm.

For each predicted cluster in X_2, the SAM method [20] is used to obtain a list of differentially-expressed genes (DEGs) between this cluster and the normal (non-cancer) samples. Consequently, a set of unique DEGs, or signature genes, Φ_k is obtained for cluster k. A reduced dataset of X_2 based on the union of all Φ_k is then used in return for the consensus clustering training. The trained class labels are then used to predict the class labels for samples in X_1 and so on so forth. This procedure continues till the set of signature genes between the last two steps are stable enough, say the percentage of change is less than 5%.

An overview of the above algorithms is shown on Fig. 6.

IV. The Subtype-specific Driver

IDENTIFICATION ALGORITHM

According to the preliminary analysis in Section II, even within a subtype, there could be different driving events. Here, we focus on the dominant subtype-specific events, i.e., CNAs. CNAs often occur on a broad region in the cancer genome and presumably harbor far more genes than necessary to trigger cancer. As a result of the efficient responses (cf. Fig. 2) from CNAs, most of which may be mechanic but otherwise non-cancerous, the search for candidate CNAs that might be responsible for the cancerous process becomes challenging (cf. Fig. 5). Nevertheless, the subtype-specific DEGs, i.e., gene signature, offer a retrospective clue for subtype-specific driver identification.

A CNA-affected gene can activate the cancerous process through its CNA-induced aberrant expressions. Therefore, its expression may likely be correlated with the signature genes, which are believed to be the consequences of subtype-specific processes. To identify the drivers, for a subtype k and its signature S_k ($|S_k| \triangleq M$), and L candidate CNA genes, an M-by-L matrix Z_k of pair-wise Pearson correlation coefficients between the signature genes and the expressions of the CNA genes is computed. We may assume that a row-wise zero-meaned operation has been applied to Z_k. A PCA approach can next be applied to determine the correlation of each CNA with the set S_k, as follows:

a. Perform an SVD: $Z_k = U\Sigma V^T$, and then project Z_k onto the first principle vector u_1 of U, to give $z1 = u_1^T Z_k$. The individual entry of z_1 represents the overall correlation of each CNA gene with the set of signature genes.
b. For a candidate CNA gene $l \in \{1, .., L\}$, the more positive z_1^l is, the more gene l is positively correlated with S_k, and vice versa. Therefore, the values of z_1 provide a ranking for the candidate CNA genes.
c. The empirical confidence interval (CI) of z_1^l can be obtained by bootstrapping. If the CI of z_1^l at significance level α does not include 0, gene l is determined to be a significant regulator.

The set of candidate CNA genes can be obtained by finding the set of significantly recurrent CNAs in subtype k via GISTIC [21] using the SNP arrays matching to the samples in k.

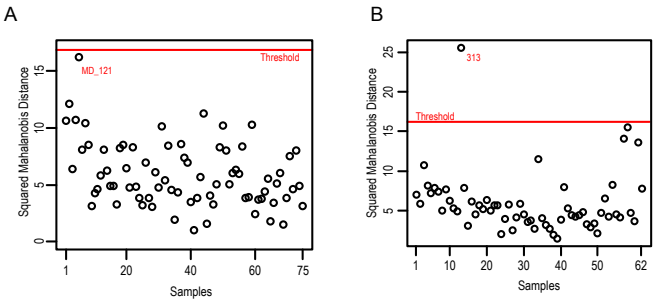

Fig. 7. Outlier detection. A, outlier detection by the Mahalanobis distance method in $\mathcal{D}_1$, and; B. in $\mathcal{D}_2$.

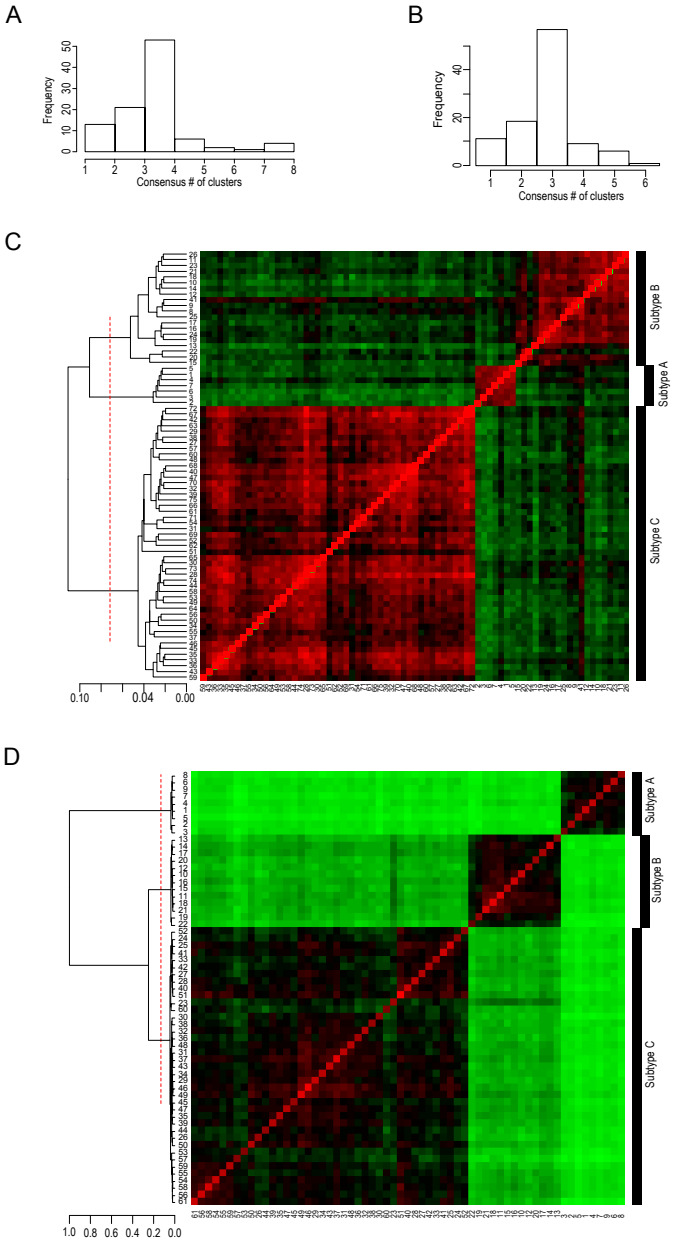

Fig. 8. A and B. The histograms for the consensus number of clusters for the two datasets. C and D. The converged consensus matrices in heatmaps for the two datasets, $\mathcal{D}_1$ (C), and $\mathcal{D}_2$ (D), respectively. Red color indicates two samples are repetitively predicted to be in the same cluster, while green color indicates they are rarely so.

V. RESULTS AND DISCUSSION

To implement the proposed approach, we used two published and publicly available medulloblastoma datasets. The first dataset consists of 75 samples by Taylor's group [10] (GEO: GSE21166 and GSE14437). The second dataset consists of 62 medulloblastoma samples from Kool *et al.* [3] (GEO: GSE10327). For convenience, the two datasets are referred to as $\mathcal{D}_1$ and $\mathcal{D}_2$, respectively. Only probesets common to both $\mathcal{D}_1$ and $\mathcal{D}_2$ are used.

TABLE I
PATHWAY ANALYSIS FOR EACH SUBTYPE BY ITS SIGNATURE

Pathways (p-value)	Genes
Subtype A	
NF-κB Signaling (5.89×10^{-5})	TLR1, BMP4, TRD@, KRAS, IRAK1, PIK3R3, MAP3K7, TGFA, PDGFRA, TDP2, LTBR, TRA@, TNFRSF11B
Wnt/β-catenin Signaling (6.60×10^{-5})	FZD10, PPP2R5D, WNT16, WIF1, MAP3K7, GNAO1, CD44, FZD6, DKK4, DKK2, LEF1, DKK1, WNT11
Subtype B	
Axonal Guidance Signaling (2.28×10^{-4})	GLI2, ARHGEF7, SEMA6A, PTCH1, PLXND1, BMP5, MICAL1, NTRK2, NTRK3, CXCL12, NGFR, PDGFD, GLI1, PRKD1, WNT5A, UNC5C
Hedgehog signaling (5.57×10^{-3})	PTCH1, GLI1, GLI2, WNT5A, CSNK1E, BMP5, GAS1, ATOH1
Subtype C	
PPAR Signaling (6.9×10^{-4})	NR2F1, PDGFA, PDGFRA, NCOR2, AIP
Protein Kinase A Signaling (2.4×10^{-3})	NR2F1, PDGFA, PDGFRA, NCOR2, AIP
p53 Signaling (3.7×10^{-3})	GADD45G, C12orf5, TP53BP2, SERPINE2
Glucocorticoid Receptor Signaling (8.9×10^{-3})	POU2F1, TAF5, PRL, SGK1, MAPK10, NCOR2, SMARCA4

A. Outlier Detection

Fig. 7A-B shows the result of the Mahalanobis distance method of outlier detection. Note that at 99% confidence, only $\mathcal{D}_2$ is found to contain an outlier 313. One sample in $\mathcal{D}_1$ is close to but never exceed the threshold. As a result, 75 and 61 samples are used for training and testing in $\mathcal{D}_1$ and $\mathcal{D}_2$, respectively. The total number of 136 samples were further processed by quantile normalization, to ensure that a gene has close dynamic ranges in both datasets.

B. Subtyping

To implement the subtyping framework, we used the $\mathcal{D}_1$ and $\mathcal{D}_2$ samples as the X_1 and X_2 in Section III, respectively. One hundred re-samplings (i.e., H=100) were performed for each training step. The algorithm converges extremely fast, both in terms of the number of clusters K and the stability of the gene signatures. Fig. 8 shows the histograms for the converged consensus numbers of clusters, and expected value of $K = 3$ in both datasets; and the consensus matrices as clustered and cut with $K = 3$ clusters.

It appears from Fig. 8C and D that, although both datasets demonstrate strong consensus in having three subtypes, the second dataset is much less noisy than the first one.

Table I summarizes the top enriched canonical pathways for each subtype. It is easy to see that Subtype A corresponds to the Wnt-pathway associating subtype, Subtype B to the Shh-pathway associating subtype, and Subtype C to the non-Wnt/non-Shh patients, in medulloblastoma.

C. Driver Identification

Copy number measurements via SNP arrays of $\mathcal{D}_1$ were processed and submitted to GISTIC (via genepattern.broadinstitute.org) for detection of recurrent CNAs within each subtype. The results of this pre-selection step are listed in Table II.

TABLE II
NUMBERS OF CANDIDATE CNAS WITHIN EACH SUBTYPE

	Subtype A		Subtype B		Subtype C	
	gains	losses	gains	losses	gains	losses
# genes	0	359	0	281	692	561

As described in Section IV, for each subtype, pair-wise correlations between the pre-selected CNAs and the gene signature were then computed and the PCA-based ranking method was applied to determine the significance score for each candidate regulator. Ten thousand bootstrap replications were performed to determine the empirical confidence intervals of the obtained scores. In all three subtypes, regulators with empirical p-values < 0.005 were determined to be significantly correlated with the signature genes. The results are summarized in Table III. Of note, Subtype A is only marginally enriched (not significantly) with Wnt-pathway genes, but some other highly ranked candidate genes such as NRN1 and TULP4 seem to be interpretable. While TULP4 was proposed to be a Wnt subtype tumor suppressor by [14] and its copy losses may cause the loss-of-function of tumor suppressing; NRN1 was found to be among the most down-regulated signature for a drug resistant medulloblastoma cell-line [22]. Subtype B contains significant axonal guidance genes, and this pathway is also represented in its signature (Table I). Particularly, PTCH1, which is frequently deleted in Subtype B, plays an important role in Shh-pathway signaling by inhibiting the *Smoothened*. The deletions of PTCH1 might cause the Shh pathway to be permanently turned on, which could be cancerous if it occurred during brain development. In Subtype C, where a large number of significant CNA drivers are found, it is of interest to note that even though signature of Subtype C is not enriched with Wnt genes, the CNA drivers do seem to suggest this trend. This might shed light on the disease mechanism of this group, which is much less understood compared with the other two subtypes.

VI. CONCLUSION

In this work, we have conducted a thorough investigation on the CNA-gene expression relations, based on which we formulated a theory for the CNA-induced subtype-specific cancerous process. A two-step algorithmic framework was then developed to perform gene-signature based cancer subtyping and to identify subtype-specific CNAs drivers. The algorithm was applied to datasets of medulloblastoma, and produced dataset-invariant subtyping results. The driver identification results were found to be enriched with cancer-driving pathways. This study has contributions in several aspects.

First, the CNA-gene expression study unveils the efficient responses of CNAs on gene expressions. Second, a subtyping technique that avoids the effects of the mechanic responses of CNAs by depending on the signature-based clustering is

TABLE III
CANDIDATE CNA DRIVERS WITHIN EACH SUBTYPE AND PATHWAY ANALYSIS

Pathways (p-value)	Total #	Candidate CNA Drivers
Subtype A	171	
Wnt/β-catenin Signaling (0.183)		SOX4, MAP3K7, PPP2R5D
Some top-ranked non-pathway genes		NRN1 (No.1), TULP4 (No. 18)
Subtype B	181	
Axonal Guidance Signaling (2.04×10^{-3})		AKT1, NTRK2, RGS3, CDK5, PRKCD, PTCH1, ABL1, NFATC1, PLXNB2, RAC3
Glioma Signaling (8.91×10^{-3})		IGF2, AKT1, PRKCD, ABL1
Subtype C	878	
EIF2 Signaling (9.33×10^{-7})		RPS20, EIF3H, GRB2, RPS17/RPS17L, EIF3F, RPS2, RPL30, RPL23A, RPL23, RPS23, RPLP0, EIF4E, EIF3M, RPS24, EIF4G2, RPL36A/hCG 1787519, EIF4A3, PAIP1, RPS15, RPLP2, RPL8, RPL18, RPL13A, RPL13
mTOR Signaling (3.24×10^{-5})		RPS6KB1, RPS20, PLD2, EIF3H, DDIT4, PPP2R2A, RPS17/RPS17L, EIF3F, RPS2, RAC1, RPS23, EIF4E, EIF3M, RPS24, RHOG, EIF4G2, RHOT1, EIF4A3, TSC2, RPS15, PRKCA
Wnt/β-catenin Signaling (1.14×10^{-4})		TP53, SFRP4, WNT3, FRAT1, PPP2R2A, CSNK1D, WNT16, FZD1, CDH1, SOX9, NLK, CDH5, DKK3, RARA, CD44, SFRP5, DKK4, SFRP1, FZD2

function-oriented and addresses quite a few issues in unsupervised learning of cancer datasets. Third, subtype-specific driver identification provides an efficient algorithm to relate the dysregulated pathway activities to the aberrations at the DNA level. Its capability to rank such candidates provides a way to automate the process of driver identification.

We also note that there are a few limitations in current study. For example, it is assumed that relationships between the CNA-genes and the gene signature are time-invariant. This limits the candidates to those already in equilibrium states and lacks the potential to uncover those drivers that play roles dynamically. It is also desirable that in future works, other cancers with subtypes can be tested with the current method for identifying cancer drivers.

ACKNOWLEDGMENT

This work is supported by the HKU UPF and the HKU EEE department. PKC would like to acknowledge Drs. Ching C. Lau and Tsz-kwong Man of the Texas Children's Hospital for their invaluable discussions. The most intensive computations were carried out on the supercomputing cluster, Ranger, at TACC and we would like to acknowledge the computing time provided by TACC (TG-MCB110130).

REFERENCES

[1] C. M. Perou, T. Sorlie *et al.*, "Molecular portraits of human breast tumours," *Nature*, vol. 406, no. 6797, pp. 747–52, 2000.

[2] R. G. Verhaak, K. A. Hoadley *et al.*, "Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in pdgfra, idh1, egfr, and nf1," *Cancer Cell*, vol. 17, no. 1, pp. 98–110, 2010.

[3] M. Kool, J. Koster *et al.*, "Integrated genomics identifies five medulloblastoma subtypes with distinct genetic profiles, pathway signatures and clinicopathological features," *PLoS One*, vol. 3, no. 8, p. e3088, 2008.

[4] J. E. Visvader, "Cells of origin in cancer," *Nature*, vol. 469, no. 7330, pp. 314–322, 2011.

[5] T. R. Golub, D. K. Slonim *et al.*, "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring," *Science*, vol. 286, no. 5439, pp. 531–7, 1999.

[6] E. Bair and R. Tibshirani, "Semi-supervised methods to predict patient survival from gene expression data." *PLoS Biol*, vol. 2, no. 4, p. E108, Apr 2004.

[7] Y. Gao and G. Church, "Improving molecular cancer class discovery through sparse non-negative matrix factorization," *Bioinformatics*, vol. 21, no. 21, pp. 3970–5, 2005.

[8] M. R. Stratton, P. J. Campbell, and P. A. Futreal, "The cancer genome," *Nature*, vol. 458, no. 7239, pp. 719–24, 2009.

[9] Y. Cho, P. Tamayo *et al.*, "Integrative genomic analysis of medulloblastoma identifies a molecular subgroup that drives poor clinical outcome," *J Clin Oncol.*, vol. 12, no. 6, pp. 1424–1430, 2010.

[10] P. A. Northcott, A. Korshunov *et al.*, "Medulloblastoma comprises four distinct diseases," *Journal of Clinical Oncology*, vol. 12, no. 6, pp. 1408–1414, 2010.

[11] A. Y. Ng, M. I. Jordan, and Y. Weiss, "On spectral clustering: Analysis and an algorithm," in *Advances In Neural Information Processing Systems*. MIT Press, 2001, pp. 849–856.

[12] R. Tibshirani, G. Walther, and T. Hastie, "Estimating the number of clusters in a data set via the gap statistic," *J. R. Stat. Soc. Ser. B-Methodological*, vol. 63, pp. 411–423, 2001.

[13] M. G. Schimek, *Smoothing and regression : approaches, computation, and application*, ser. Wiley series in probability and statistics Applied probability and statistics section. New York: Wiley, 2000.

[14] P. Gibson, Y. Tong *et al.*, "Subtypes of medulloblastoma have distinct developmental origins," *Nature*, vol. 468, no. 7327, pp. 1095–9, 2010.

[15] A. S. Hadi, "Identifying multiple outliers in multivariate data," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 54, no. 3, pp. pp. 761–771, 1992.

[16] S. Monti, P. Tamayo *et al.*, "Consensus clustering: A resampling-based method for class discovery and visualization of gene expression microarray data," *Machine Learning*, vol. 52, no. 1-2, pp. 91–118, 2003.

[17] Y. Freund and R. E. Schapire, "Experiments with a New Boosting Algorithm," in *International Conference on Machine Learning*, 1996, pp. 148–156.

[18] M. Culp, K. Johnson, and G. Michailidis, "ada: an r package for stochastic boostong," *Journal of Statistical Software*, vol. 17, no. 2, pp. pp. 1–27, October 2006.

[19] B. Niu, Y.-D. Cai *et al.*, "Predicting protein structural class with adaboost learner." *Protein Pept Lett*, vol. 13, no. 5, pp. 489–492, 2006.

[20] V. G. Tusher, R. Tibshirani, and G. Chu, "Significance analysis of microarrays applied to the ionizing radiation response," *Proc Natl Acad Sci U S A*, vol. 98, no. 9, pp. 5116–21, 2001.

[21] R. Beroukhim, G. Getz *et al.*, "Assessing the significance of chromosomal aberrations in cancer: Methodology and application to glioma," *Proc. Nat. Acad. of Sci.*, vol. 104, no. 50, pp. 20 007–20 012, 2007.

[22] M. D. Bacolod, S. M. Lin *et al.*, "The gene expression profiles of medulloblastoma cell lines resistant to preactivated cyclophosphamide." *Curr Cancer Drug Targets*, vol. 8, no. 3, pp. 172–179, May 2008.