Transfer Learning Across Cancers on DNA Copy Number Variation Analysis

Huanan Zhang
Department of Computer Science
and Engineering
University of Minnesota Twin Cities
Minneapolis, Minnesota 55455-0213
Email: huanan@cs.umn.edu

Ze Tian
Microsoft Corporation
Redmond, Washington 98052
Email: zetian@microsoft.com

Rui Kuang
Department of Computer Science
and Engineering
University of Minnesota Twin Cities
Minneapolis, Minnesota 55455-0213
Email: kuang@cs.umn.edu

Abstract—**DNA copy number variations (CNVs) are prevalent in all types of tumors. It is still a challenge to study how CNVs play a role in driving tumorgenic mechanisms that are either universal or specific in different cancer types. To address the problem, we introduce a transfer learning framework to discover common CNVs shared across different tumor types as well as CNVs specific to each tumor type from genome-wide CNV data measured by arrayCGH and SNP genotyping array. The proposed model, namely Transfer Learning with Fused Lasso (TLFL), detects latent CNV components from multiple CNV datasets of different tumor types to distinguish the CNVs that are common across the datasets and those that are specific in each dataset. Both the common and type-specific CNVs are detected as latent components in matrix factorization coupled with fused lasso on adjacent CNV probe features. TLFL considers the common latent components underlying the multiple datasets to transfer knowledge across different tumor types. In simulations and experiments on real cancer CNV datasets, TLFL detected better latent components that can be used as features to improve classification of patient samples in each individual dataset compared with the model without the knowledge transfer. In cross-dataset analysis on bladder cancer and cross-domain analysis on breast cancer and ovarian cancer, TLFL also learned latent CNV components that are both predictive of tumor stages and correlate with known cancer genes.**

I. Introduction

Normally there are two copies of each gene in the human genome located on paired DNAs in a chromosome. Large scale DNA alternations such as insertions or deletions could lead to copy number gain or loss of the genes, which are called DNA copy number variations (CNVs). CNVs have been found extremely common in human cancer genome [1], [2] and it is believed that CNVs play significant roles in cancer [3], [4]. New technologies such as array-based comparative genomic hybridization (arrayCGH) [5], [6] and SNP genotyping arrays, are now available to measure genome-wide CNVs in high resolution at a population scale for characterizing CNV patterns in cancer samples [7]. Identification and systematic analysis of CNVs can provide important insights into the cellular defects that are cancer causative and suggest potential therapeutic strategies.

Most previous computational research work focused on developing models for identifying individual CNV events from CNV samples of a single cancer type. [8] studied 17 cancer types with at least 40 samples in each cancer type and reported that about 80% somatic copy number alternations found in one cancer type can also be found in pooled analysis excluding that cancer type. The detected regions in the pooled analysis were also found in other cancer types that are better localized. These common and type-specific CNVs can potentially reveal unknown cancer mechanisms in the light of cross-cancer-type analysis. However, currently there is no unified mathematical model to simultaneously detect the CNV events common or specific to multiple cancer types from CNV array datasets.

In this paper, we propose a Transfer Learning with Fused Lasso model (TLFL) to detect latent CNV components from CNV datasets of multiple cancer types, in which each cancer type can be regarded as one domain in transfer learning. Common latent CNV components are used as a bridge to transfer knowledge among different cancer domains along with the domain-specific components for each cancer type to explain the observed CNV datasets. To represent the pattern of CNV events, fused lasso is applied on each latent CNV component to preserve the sparsity and block structure. By using alternating optimization to solve the TLFL model, common latent features and domain specific features could be detected from multiple domains. Compared with a baseline method without knowledge transfer, TLFL is more robust and identifies more accurate latent CNV components in simulations and experiments on real arrayCGH CNV datasets and SNP genotyping array datasets.

II. Related Work

DNA CNVs tend to occur in continuous blocks of various sizes and thus, the adjacent probe features are more likely to be associated in the same CNV event. Previously, several models, such as change-point detection [9], [10], hidden Markov models [11], [12] and Gaussian models [13], [14] have been applied to address the challenge. More recently, fused lasso model [15] which introduces ℓ_1 norm constraint to encourage sparse change points and fused CNV features, has been found to be effective in discovering more interpretable CNV events [16]. A fused lasso latent feature model, FLLat [17] was proposed to take full advantage of any shared information among samples. The model assumes each CNV sample is a linear combination of a few latent CNV components. By factorizing the arrayCGH data matrix into the product of a coefficient matrix and a latent feature matrix, FLLat is able to detect underlying CNV events and discern specific relationships between samples. [18] proposed a latent fused-lasso feature method to use prior knowledge to learn group specific CNVs. Other multiple sample analysis methods which are powerful to identify frequent individual CNVs [19], [20], [21], [22], are neither designed to identify CNV components nor capture the heterogeneity of samples. None of the previous methods was specially designed as a unified mathematical formulation to discover CNV events from multiple datasets across different cancer types.

Transfer learning uses common knowledge or structures among different domains to enhance multiple learning tasks [23], [24]. Recently, a lot of research work on transfer learning has been published for various learning problems such as Co-Clustering based Classification [25], Label Propagation [26], Collaborative Dual-PLSA [27] and Matrix Tri-Factorization based Classification [28]. The paradigm of transfer learning also fits the learning tasks of finding CNV components across cancer types since datasets of the same or similar cancer types presumably bear the same or similar pathogenic cause. However, to the best of our knowledge, no transfer learning method has been designed for latent fused-lasso component discovery.

III. Method

Figure 1 is an outline of the TLFL model. In the Figure, each of the three cancer CNV datasets is factorized into a product of a coefficient matrix and k latent components. In each set of the k components, τ components are shared across the three datasets and the remain $k-\tau$ components are specific to each dataset. The framework assumes that the CNV features are measured on the same set of probe locations sampled from a chromosome. Each component is learned with fused lasso on the adjacent probe features to enforce a shape of step function to mimic true CNV signals. In the following, we first describe the optimization formulation of the model and then introduce an alternating optimization algorithm to minimize the cost function. Strategies for selecting hyper-parameters and initialization are also suggested for the empirical practice of the algorithm.

A. Transfer Learning Framework

The notations are given in Table I. Given δ datasets measured from the same m probe locations, each dataset X_d contains n_d samples from one cancer domain. The objective is to recover k latent components $[\hat{U}, U_d]$ to reconstruct each dataset X_d with the minimal loss of information, where U_d are $k-\tau$ latent components specific to dataset X_d and $\hat{U}$ are τ common components shared by all the datasets. V_d is the corresponding coefficient matrix of $[\hat{U}, U_d]$ for reconstructing X_d. Specifically, the TLFL model assumes that each sample in X_d can be represented as a linear combination of k latent components as follows,

TABLE I
Notations

Notation	Description
δ	# of domains
n_d	# of samples in domain $d \in [1, \delta]$
m	# of CNV features
k	total # of components in one domain
τ	# of common components
X_d	data matrix of domain d, size $m \times n_d$
$\hat{U}$	matrix of common components, size $m \times \tau$
U_d	domain-specific components of domain d, size $m \times (k-\tau)$
V_d	coefficient matrix of domain d, size $k \times n_d$

$$X_d = [\hat{U}, U_d]V_d.$$

To obtain the k latent components $[\hat{U}, U_d]$ and coefficient matrix V_d that best reconstruct X_d , the objective function minimizes the reconstruction error of all the datasets by a sum of the squared loss across the datasets,

$$\sum_{d=1}^{\delta} ||X_d - [\hat{U}, U_d]V_d||_F^2.$$

To capture the spatial relation in the CNV probe features, each latent component (a column in $[\hat{U}, U_d]$) is constrained by a fused lasso. Specifically, the cost function for the common components in $\hat{U}$ is defined as,

$$\begin{aligned} &g(\hat{U}, \lambda_C, \gamma_C) \\ &= \lambda_C \sum_{j=1}^{\tau}\sum_{i=1}^{m} |\hat{U}_{(i,j)}| + \gamma_C \sum_{j=1}^{\tau}\sum_{l=2}^{m} |\hat{U}_{(l,j)} - \hat{U}_{(l-1,j)}|, \end{aligned} \tag{1}$$

where λ_C and $\gamma_C \in \mathbb{R}$ are parameters to weight the penalties and the lasso penalty is introduced to obtain sparse CNV events in the components. Similarly, the cost function for each domain-specific latent component is

$$\begin{aligned} &g(U_d, \lambda_d, \gamma_d) \\ &= \lambda_d \sum_{j=1}^{k-\tau}\sum_{i=1}^{m} |U_{d(i,j)}| + \gamma_d \sum_{j=1}^{k-\tau}\sum_{l=2}^{m} |U_{d(l,j)} - U_{d(l-1,j)}|, \end{aligned} \tag{2}$$

where λ_d and $\gamma_d \in \mathbb{R}$ are also parameters to weight the penalties. Here, λ_C, γ_C, λ_d and γ_d for $d = 1, 2, \ldots, \delta$ are hyper-parameters to be tuned (see section III-C).

Given all the cost terms introduced above, the complete objective function is defined as

$$\begin{aligned} \mathcal{L} = &\sum_{d=1}^{\delta}(\frac{1}{2}||X_d - [\hat{U}, U_d]V_d||_F^2 \\ &+ g(U_d, \lambda_d, \gamma_d)) + g(\hat{U}, \lambda_C, \gamma_C) \\ &s.t. \end{aligned} \tag{3}$$

$$V_d \geq \mathbf{0} \text{ and } V_{d(i,:)}V_{d(i,:)}^T = 1 \text{ for } i = 1, 2, \ldots, k,$$

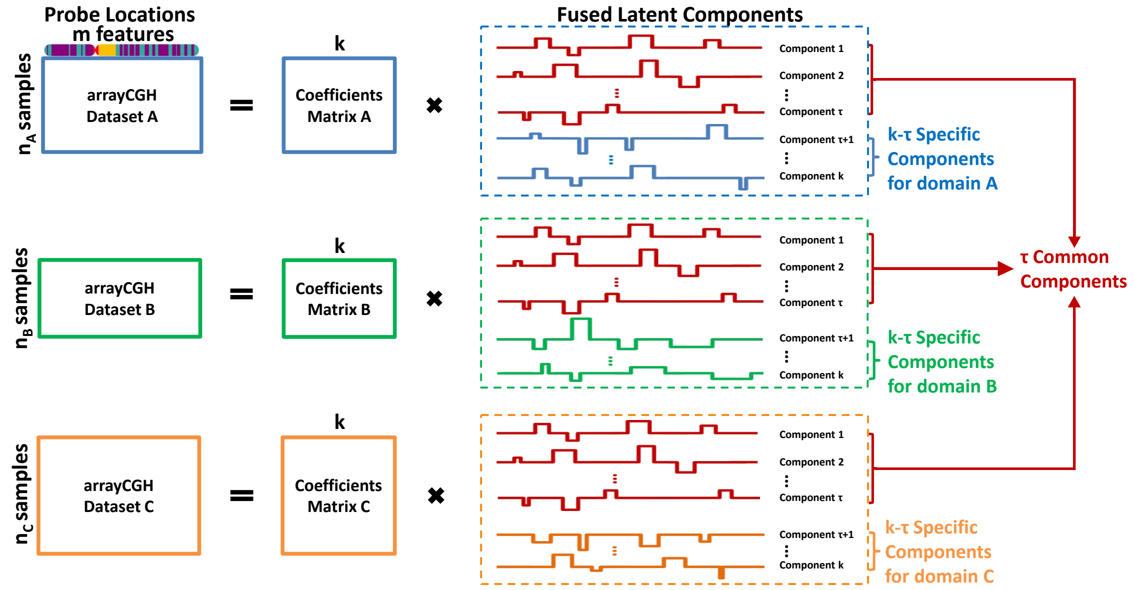


Fig. 1. Outline of TLFL model. ArrayCGH or SNP genotyping array datasets from three domains are decomposed into coefficient matrices and matrices of k latent components. The probe locations are identical in all three datasets (m features) while the number of samples (n_A, n_B and n_C) can be different. The red latent components are τ common components shared in the three domains, and the remaining components in the same color of each dataset are $k - \tau$ domain specific components. For better visualization, matrices in this figure are transpose from equations.

where $V_d \geq 0$ denotes the condition that each element in V_d is nonnegative and $V_{d(i,:)}$ is the ith row of V_d. This cost function combines the reconstruction errors with the lasso and fused lasso terms weighted by λ_C, γ_C, λ_d and γ_d for $d = 1, 2, \ldots, \delta$. The nonnegative constraints on V_d only allow positive coefficients to combine latent components which might contain both amplification (positive) and deletion (negative) events. Each row in every V_d is also normalized across the samples such that the learned latent components are scaled to be comparable with each other [17]. The normalization also encourages even contributions from every latent component features to prevent being dominated by a few. Those considerations are meant to improve the interpretability of both the coefficients and the components.

B. Alternating Optimization

The optimization problem in eqn 3 can be solved by alternating updates to the variables $\hat{U}$, U_d and V_d iteratively. Specifically, we solve subproblems on only one group of variables by fixing the other two and alternate through the three groups of variables in each iteration. The alternating procedure is repeated until convergence. The detailed TLFL algorithm is described in Algorithm 1. Below we outline the solution to each subproblem to solve for $\hat{U}$, U_d and V_d, respectively.

1) Updating coefficient matrix V_d: When $\hat{U}$ and U_d are fixed, eqn 3 is only a function on V_d simplified as

$$\begin{aligned} &\arg\min_{V_d} ||X_d - [\hat{U}, U_d]V_d||_F^2 \\ &s.t. \\ &V_d \geq \mathbf{0} \text{ and } V_{d(i,:)}V_{d(i,:)}^T = 1 \text{ for } i = 1, 2, \ldots, k. \end{aligned} \tag{4}$$

For each column $X_{d(:,j)}$, we can solve a nonnegative least-square problem to obtain a solution for $V_{d(:,j)}$.

$$\begin{aligned} &\arg\min_{V_{d(:,j)}} ||X_{d(:,j)} - [\hat{U}, U_d]V_{d(:,j)}||_F^2 \\ &s.t. \\ &V_{d(:,j)} \geq \mathbf{0}. \end{aligned} \tag{5}$$

Then V_d can be normalized as $V_{d(i,:)}V_{d(i,:)}^T = 1$ for $i = 1, 2, \ldots, k$.

2) Updating domain-specific components U_d: When $\hat{U}$ and V_d are fixed, eqn 3 is only a function on U_d simplified as

$$\begin{aligned} &\frac{1}{2}||X_d - [\hat{U}, U_d]V_d||_F^2 + g(U_d, \gamma_d, \lambda_d) \\ =&\frac{1}{2}||\dot{X}_d - U_d V_{d(\tau+1:k,:)}||_F^2 + g(U_d, \gamma_d, \lambda_d), \end{aligned} \tag{6}$$

where residue $\dot{X}_d$ is defined as

$$\dot{X}_d \equiv X_d - \hat{U}V_{d(1:\tau,:)}.$$

This problem is equivalent to the general fused lasso problem, which can be solved by the SLEP package [29].

Algorithm 1 TLFL

```
Input: $\{X_d\}_{d=1}^{\delta}$, $k$, $\tau$, $\{\gamma_d\}_{d=1}^{\delta}$,$\{\lambda_d\}_{d=1}^{\delta}$, $\gamma_C$, $\lambda_C$
Output: $\hat{U}$, $\{U_d\}_{d=1}^{\delta}$, $\{V_d\}_{d=1}^{\delta}$
initialize $\hat{U}$, $\{U_d\}_{d=1}^{\delta}$
repeat
  for $d = 1, \ldots, \delta$ do
    for $j = 1, \ldots, n_d$ do
      solve $\arg\min_{V_{d(:,j)}} ||X_{d(:,j)} - [\hat{U}, U_d] V_{d(:,j)}||_F^2$
      s.t. $V_{d(:,j)} \geq \mathbf{0}$ (eqn 5)
    end for
    normalize $V_d$ s.t. $V_{d(i,:)} \times V_{d(i,:)}^T = 1$ for $i = 1, 2, \ldots, k$
    $\dot{X}_d = X_d - \hat{U} V_{d(1:\tau,:)}$
    solve $\arg\min_{U_d} (\frac{1}{2}||\dot{X}_d - U_d V_{d(\tau+1:k,:)}||_F^2 + g(U_d, \gamma_d, \lambda_d))$ (eqn 6)
  end for
  for $d = 1, \ldots, \delta$ do
    $\ddot{X}_d = X_d - U_d V_{d(\tau+1:k,:)}$
  end for
  $X_{all} = [\ddot{X}_1, \ddot{X}_2, \ldots, \ddot{X}_\delta]$
  $V_{all} = [V_{1(1:\tau,:)}, V_{2(1:\tau,:)}, \ldots, V_{\delta(1:\tau,:)}]$
  solve $\arg\min_{\hat{U}} (\frac{1}{2}||X_{all} - \hat{U} V_{all}||_F^2 + g(\hat{U}, \gamma_C, \lambda_C))$ (eqn 7)
until $\hat{U}$, $\{U_d\}_{d=1}^{\delta}$, $\{V_d\}_{d=1}^{\delta}$ converge
```

3) Updating common components $\hat{U}$:

3) Updating common components $\hat{U}$: When U_d and V_d are fixed, eqn 3 is only a function on $\hat{U}$ simplified as

$$\begin{aligned} &\sum_{d=1}^{\delta} (\frac{1}{2}||X_d - [\hat{U}, U_d] V_d||_F^2) + g(\hat{U}, \gamma_C, \lambda_C) \\ =&\frac{1}{2}||X_{all} - \hat{U} V_{all}||_F^2 + g(\hat{U}, \gamma_C, \lambda_C), \end{aligned} \tag{7}$$

where we define

$$\begin{aligned} \ddot{X}_d &\equiv X_d - U_d V_{d(\tau+1:k,:)}, \\ X_{all} &\equiv [\ddot{X}_1, \ddot{X}_2, \ldots, \ddot{X}_\delta], \\ V_{all} &\equiv [V_{1(1:\tau,:)}, V_{2(1:\tau,:)}, \ldots, V_{\delta(1:\tau,:)}]. \end{aligned}$$

Similarly, this problem is also equivalent to the general fused lasso problem, which can be solved by the SLEP package.

C. Initialization and Hyper-parameter Selection

Since eqn 3 is not convex, alternating updates in TLFL do not guarantee a global optimal solution. The local optimal solution heavily relies on proper initialization of $\hat{U}$ and U_d. We adopt a simple strategy to choose the initialization. We use Principle Component Analysis (PCA) on pooled data $[X_1, X_2, \ldots, X_\delta]$ to select top τ components as the initialization of common components $\hat{U}$. For domain specific components, PCA is applied on each domain data separately to select the top k components for each domain. Then, the top τ components of the k components of each domain that are most similar to the initialization of $\hat{U}$ are removed. The similarity is measured by the absolute correlation coefficients. For each domain, the remaining $k - \tau$ components are used as the initialization of domain specific components U_d.

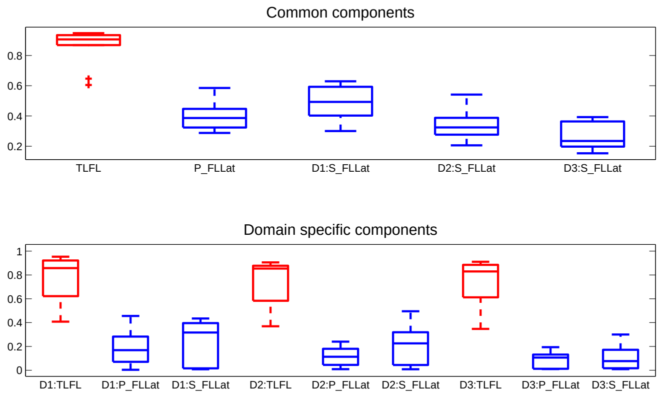


Fig. 2. Performance of latent component detection by TLFL, pool FLLat (P_FLLat) and split FLLat (S_FLLat) section IV-A. The box-plots are computed from 10 random experiments. D1, D2 and D3 denote the three domains.

The number of latent component k was chosen as the number of principle components that can explain $\alpha \in [0, 1]$ variation of the arrayCGH or SNP genotyping array datasets. For multiple domains, the calculated k could vary among the datasets. We simply choose the maximal as a global k to explain at least α variance in each dataset. A user also needs to select a parameter $\beta \equiv \tau/k$ to control the ratio between common component number τ and total component number k. For similar datasets such as datasets of the same or closely related cancer types, β should be chosen larger while for datasets from different cancer types, β should be chosen smaller. Presumably, β could be determined by a user's perception of the similarity across the domains.

Parameters λ_C, γ_C, λ_d and γ_d are chosen by the same Bayesian Information Criterion (BIC) introduced in [17]. BIC controls both model complexity and training error to avoid overfitting. For each domain, λ_d and γ_d are selected with dataset X_d and k components. λ_C and γ_C are selected with the combined dataset $[X_1, X_2, \ldots, X_\delta]$ and $\tau + \delta * (k - \tau)$ components. Note that we could apply BIC to the complete model in eqn 3 to jointly select λ_C, γ_C, λ_d and γ_d. However, jointly choosing four parameters is not scalable even on datasets of moderate size. Thus, we divided the estimation into smaller BIC problems as described above.

IV. Simulation

In the section, we generated artificial datasets to test TLFL model in three measurements: 1) performance of recovering latent components; 2) performance of detecting hidden sample group structures in coefficient matrix for classification and clustering; and 3) convergence and robustness under different noise levels and ratios between common and domain-specific components. The synthetic datasets are constructed as $X_d = [\hat{U}, U_d] * V_d + \Xi$, where latent component matrix $[\hat{U}, U_d]$ and coefficient matrix V_d are either predefined or

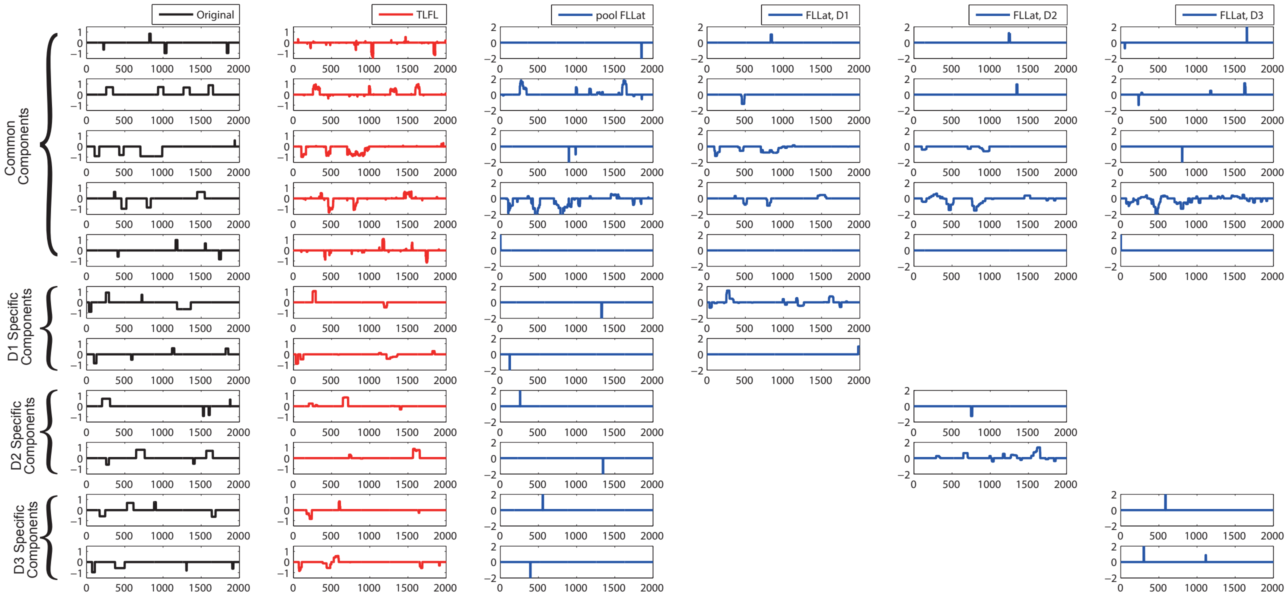


Fig. 3. Latent components detected by TLFL and FLLat are compared with the known true components. The rows represent the common components and the components specific to the domains (D1, D2 and D3). The columns from left to right represent true components, components detected by TLFL, components detected by pool FLLat, and components detected by split FLLat for D1, D2 and D3 with one column for each domain.

randomly generated, and the entries in Ξ are IID gaussian noises. In all simulations, the hyper-parameters λ and γ are selected as described in section III-C, and k and τ are assumed known. In each component in $[\hat{U}, U_d]$, 4 independent copy number gain or loss events were assumed and randomly located with magnitudes in $[-1, 1]$ over 2000 probe features. The components are not strictly orthogonal but the correlation between any two components is required to be smaller than 0.3. The entries in V_d are random nonnegative values in $[0, 1]$ and normalized as $V_{d(i,:)}V_{d(i,:)}^T = 1, i = 1, 2, \ldots, k$. We compared TLFL with FLLat [17] to show the advantage of transfer learning and discrimination of common and domain specific components. In each experiment, TLFL is applied jointly on three datasets. FLLat was applied on 1) a pooled dataset of all the domain datasets (pool FLLat) and 2) each domain dataset individually (split FLLat).

A. Recovering Latent CNV Components

Three synthetic datasets of sample size 300, 420 and 510 respectively were generated. In all the datasets, there are 7 latent components, 5 of which are common components $\hat{U}$ and 2 are domain-specific components U_d for each dataset, Note that no structure is assumed in the coefficient matrices V_d in this simulation. Gaussian noises $\Xi \sim (\mu = 0, \sigma = 0.3)$ were added. In this simulation, we focused on recovering the known latent components used to generate the synthetic datasets with added noise. The performance is measured by the average Pearson correlation coefficients of each estimated latent component with its corresponding known component. Since FLLat allows negative coefficients, some latent components were negated to obtain the best correlation coefficients with the known components. With the components were fixed, randomized coefficient matrices and noise were generated for 10 trials.

The performance of TLFL, split FLLat and pool FLLat for recovering the known components is shown in Figure 2. TLFL outperformed both split and pool FLLat in each domain under the comparison across either common components or domain-specific components. Interestingly, TLFL tends to identify more consistent common components than the FLLat models in the 10 repeats with smaller variance. Paired-sample t-test of the component correlations by TLFL and FLLat for common components, domain-specific components and all components are all significant with the largest $p - value = 4.46E - 04$, which indicates that TLFL significantly outperforms both split and pool FLLat in detecting the known latent CNV components. To illustrate the detect components, Figure 3 shows the side-by-side comparison of each component detected by FLTL, split FLLat or pool FLLat with the known component from one trial. In this example, pool FLLat failed to detect the third common component and split FLLat detected no signal correlating with the second common component in all three domains while TLFL captured all the true events accurately. In the fifth common component, both FLLat methods failed to separate the signal from the other components. Similar advantages by TLFL are also seen in the comparison of domain-specific components.

B. Sample Classification by Coefficient Matrices

Under the assumption that the latent components are underlying features describing tumor characteristics, the coefficient matrices are presumably informative for patient classification or clustering. For example, some latent features might represent CNV abberations disrupting a gene pathway in a

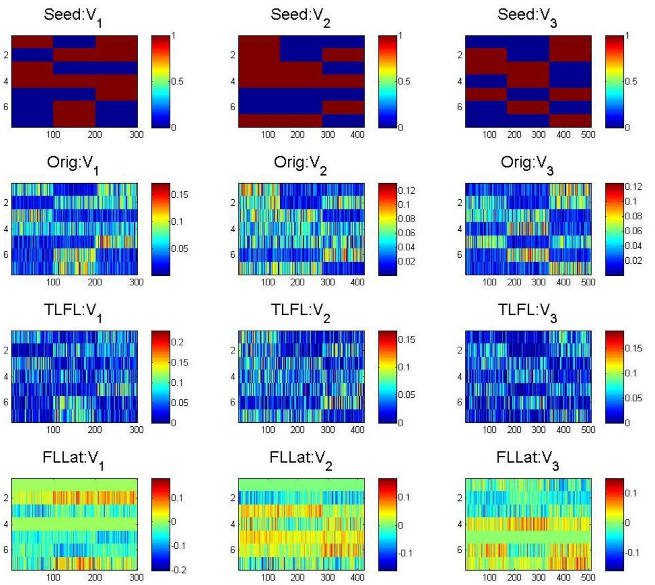


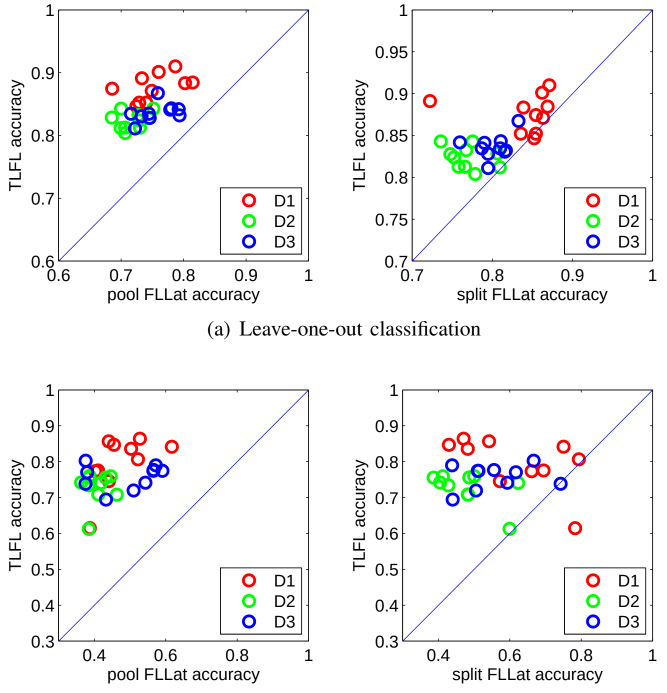
Fig. 4. Comparison of learned coefficient matrices (components by samples). The plots are shown for row 1: structured seed matrices; row 2: true coefficient matrices constructed by adding noise to the structured seed matrices; row 3: coefficient matrices learned by TLFL; and row 4: coefficient matrices learned by split FLLat. Three classes of equal sizes are assumed in each domain.

certain tumor stage, and thus samples with a large coefficient on the latent features are more likely to be associated with that particular tumor stage. Therefore, in this simulation we focused on using the learned coefficient matrices for sample classification and clustering.

Similarly, three synthetic datasets of sample size 300, 420 and 510 respectively were generated with 5 common latent components and 2 domain specific components in each domain. To create patient classes (clusters), we designed coefficient matrices representing patterns of three classes (patient subgroups) in each domain as shown in Figure 4. The true coefficient matrices shown at row 2 in Figure 4 are constructed by adding gaussian noise on the structured seed matrices at row 1. The coefficient matrices were then multiplied with components similarly generated as in section IV-A and added with gaussian noises $\Xi \sim (\mu = 0, \sigma = 0.3)$ to get the synthetic datasets. With the latent components and structure seeds fixed, we repeated the simulation procedure 10 times under the gaussian noises.

The last two rows of matrices in Figure 4 show the coefficient matrices learned by TLFL and split FLLat in one trial. In this visualization, it is clear that split FLLat made mistakes in several places such as zero coefficient of the first component in domain 1 and domain 2, and the fifth component on domain 3. The overall structure of the coefficient matrices in not as distinguishable as those detected by TLFL. Since pool FLLat learned a different number features (number of rows in V_d), it is not directly comparable in Figure 4 .

To better measure the accuracy of the coefficients, classification and clustering of samples were performed on the learned



(a) Leave-one-out classification

(b) K-means clustering

Fig. 5. Classification and clustering performance on coefficient matrices learned by TLFL, pool FLLat and split FLLat. The comparisons are between the methods on the three domains (D1, D2 and D3) in 10 random trials.

coefficient matrices. The leave-one-out cross-validation with linear SVM classifier was performed for classification of the samples. K-means clustering (K=3) was applied to cluster the samples. For K-means clustering, the averages of 100 runs are reported for each domain in each trial. Figure 5 shows the comparison of the classification and clustering results by TLFL and FLLat (pool and split) by scatter plots. In both classification and clustering comparisons, almost all the cases are well above the diagonal line, i.e. TLFL performed better than FLLat by a large margin. In addition, TLFL also detected better components in this simulation (results not shown).

C. Robustness and Convergence

To understand the robustness of TLFL and FLLat under the presence of different noise level, we tested datasets with varying amount of added noise in this simulation. Three domain datasets of sizes 60, 75 and 90 respectively were generated with 5 common components and 2 domain specific components in each domain. The gaussian noises were drew from $(\mu = 0, \sigma)$ with σ ranging from 0 to 1 with 0.1 step. To test each noise level, the simulations were repeated 10 times. Figure 6 shows that the performance of component detection drops as the noise level increases for both TLFL and FLLat. TLFL performs consistently better than both pool FLLat and split FLLat when the noise level is reasonable (≤ 0.5) with the benefit of transfer learning. TLFL and FLLat performs similarly due to the extremely high noise level that almost

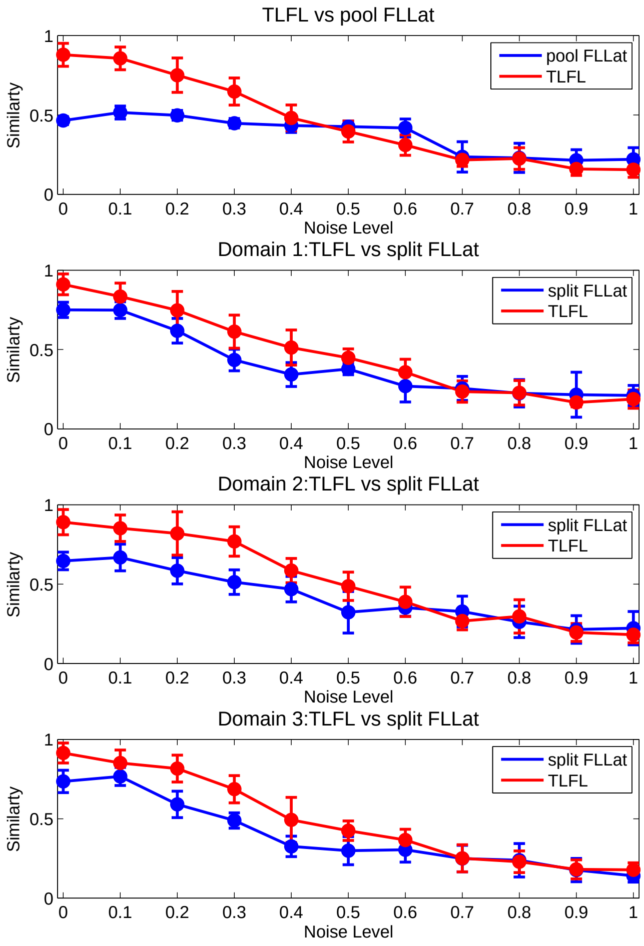


Fig. 6. Components detection performance comparison between TLFL and pool/split FLLat under different noise levels.

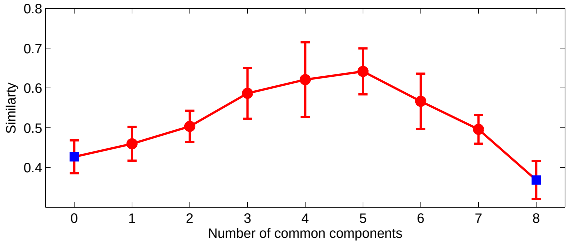


Fig. 7. Effect of varying the number of common components. The errorbars show TLFL performance under different τ with fixed datasets. Note that when $\tau = 0$, TLFL is equivalent to split FLLat and when $\tau = 8$, TLFL is equivalent to pool FLLat.

completely blurred the original signals. And at this noise level the accuracy of the learn components is very low.

In most of the real cases, the best ratio of τ and k is unknown. It is thus interesting to understand the performance of TLFL when τ varies. Intuitively, τ is directly related to how much knowledge to transfer across the different domains. The more similar the domains, the larger τ desired. In the two extremes, when $\tau = 0$ TLFL is equivalent to split FLLat, and when $\tau = k$ TLFL is equivalent to pool FLLat. We generated synthetic datasets of sample size 150, 180 and 210, each with 600 features and 8 latent components in each domain, 4 of which are common components. Similarly, we fixed the components and generated coefficient matrices randomly with gaussian noises $\Xi \sim (\mu = 0, \sigma = 0.3)$ added in 10 trials for each choice of $\tau \in [1, 2, \ldots, 7]$. The results of 10 trials is shown in Figure 7. It is clear that when $\tau = 4$ or 5, which is close to the true τ, TLFL performs the best.

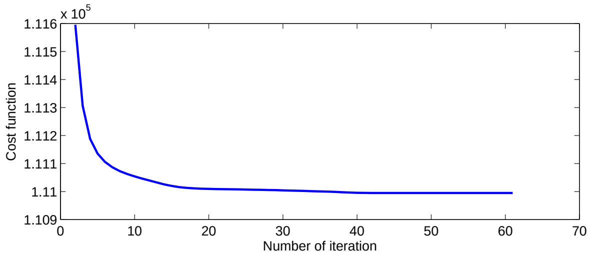


Fig. 8. Convergence of TLFL for one run from section IV-A. After around 60 iteration, components and coefficient matrices are converged.

Figure 8 shows one example of convergency in running the TLFL algorithm. TLFL convergences fast within lower tens of iterations. Most of the simulations aforementioned converged less within 100 times regardless of the sample sizes.

V. Experiments on Cancer Datasets

We performed two experiments on real cancer CNV datasets. The first experiment is a cross-dataset analysis on bladder cancer to show that TLFL can utilize information from other similar datasets to improve classification. The second experiment is a cross-domain analysis on breast cancer and ovarian cancer.

A. Analysis Across Bladder Cancer Datasets

TLFL, split FLLat and pool FLLat were tested on two bladder cancer arrayCGH datasets: Blaveri05 [30] and Stansky06 [31]. Both datasets contain urothelial carcinomas with whole-genome tiling resolution arrayCGH and high density expression profiling. There are 98 samples in Blaveri05 dataset and 57 in Stansky06. Since the two datasets were not measured by the same resolution, we interpolated the datasets in whole genome to obtain CNV readings at the same probe positions with a resolution of 500k bps per probe. All the samples from the two arrayCGH datasets are provided with information on tumor stage. In Blaveri05 dataset, the stages are Ta, T1, T2, T3 and T4, and in Stransky06 dataset, the stages are Ta, T1a, T1b, T2, T3a, T3b, T4a and T4b. We relabel the stages into 3 classes for each dataset: Blaveri05 with stages ({Ta}, {T1, T2}, {T3, T4}) and Stransky05 with stages ({Ta},{T1a, T1b, T2},{T3a, T3b, T4a, T4b}), ordered from less severe stage to more advanced stage.

For each chromosome in the two datasets, the number of latent components was chosen as the number of principle components that could explain at least 80% variance of the data. The parameter k for a certain chromosome was then set as the larger number of principle components of the two datasets. Since both datasets are on similar bladder carcinomas, we assume a large fraction of common components. For each chromosome, we took the ratio of τ/k as 70%. Parameters $\lambda_1, \gamma_1, \lambda_2, \gamma_2, \lambda_C$ and γ_C were calculated by BIC as described in section III-C. Table II reports the leave-one-out SVM classification results of the three classes using the coefficient matrices learned by TLFL, split FLLat and pool FLLat. Among the tests on all 22 chromosomes, 11 tests of Stransky06 and 10 tests of Blaveri05 present the best classification results by TLFL than both FLLat methods (numbers with color red) while on two chromosomes of Stransky06 dataset and 7 of Blaveri05 dataset, TLFL performed worse classification than both FLLat methods (numbers with color blue). Overall improvement is observed on both datasets for the average classification results of the 22 chromosomes.

TABLE II
CLASSIFICATION OF BLADDER CANCER DATASETS.

Chr	Stransky06			Blaveri05			Average		
	TLFL	pool FLLat	split FLLat	TLFL	pool FLLat	split FLLat	TLFL	pool FLLat	split FLLat
1	0.4795	0.4444	0.4386	0.5748	0.5714	0.5782	0.5272	0.5079	0.5084
2	0.4912	0.4737	0.4737	0.6361	0.6224	0.6361	0.5636	0.5481	0.5549
3	0.5906	0.5614	0.5029	0.6429	0.6429	0.6599	0.6168	0.6021	0.5814
4	0.6608	0.5848	0.6082	0.5544	0.5578	0.5544	0.6076	0.5713	0.5813
5	0.5439	0.5088	0.5263	0.6565	0.6565	0.6429	0.6002	0.5826	0.5846
6	0.5731	0.5556	0.5556	0.5884	0.6190	0.5918	0.5808	0.5873	0.5737
7	0.5906	0.6667	0.6374	0.6633	0.6395	0.6361	0.6270	0.6531	0.6367
8	0.6199	0.6316	0.6140	0.5952	0.5986	0.5714	0.6076	0.6151	0.5927
9	0.6140	0.6082	0.5146	0.6020	0.6224	0.6156	0.6080	0.6153	0.5651
10	0.6023	0.6316	0.5322	0.5850	0.5748	0.5748	0.5937	0.6032	0.5535
11	0.6140	0.6082	0.6023	0.6088	0.6395	0.6361	0.6114	0.6238	0.6192
12	0.5848	0.5556	0.5380	0.6020	0.5748	0.5748	0.5934	0.5652	0.5564
13	0.5439	0.5205	0.5673	0.5952	0.5816	0.5952	0.5695	0.5511	0.5812
14	0.5848	0.6433	0.5789	0.5680	0.5816	0.5918	0.5764	0.6125	0.5854
15	0.4737	0.4444	0.4795	0.6293	0.6190	0.5918	0.5515	0.5317	0.5357
16	0.6433	0.6491	0.6316	0.5782	0.6122	0.5884	0.6108	0.6307	0.6100
17	0.5205	0.6257	0.5322	0.5000	0.5034	0.5646	0.5102	0.5646	0.5484
18	0.5380	0.5322	0.4971	0.6224	0.6122	0.6054	0.5802	0.5722	0.5513
19	0.5322	0.5146	0.5789	0.5850	0.6122	0.6054	0.5586	0.5634	0.5922
20	0.6550	0.6667	0.6491	0.5986	0.6020	0.5918	0.6268	0.6344	0.6205
21	0.4561	0.4795	0.4561	0.5374	0.5136	0.5238	0.4968	0.4966	0.4900
22	0.5673	0.5380	0.4678	0.5782	0.5136	0.5340	0.5727	0.5258	0.5009
ave	0.5673	0.5657	0.5447	0.5955	0.5942	0.5938	0.5814	0.5799	0.5693

B. *Analysis Across Cancer Domains*

We applied TLFL method on two related cancer types, breast cancer and ovarian cancer, to detect common CNV patterns. The two CNV datasets were downloaded from TCGA data-portal[1] SNP level 2 tangent data, generated from Affymetrix Genome-Wide Human SNP Array 6.0 platform. To label the patients for survival prediction, we chose breast cancer patient samples that had a survival time less than 5 years as the positive group and longer than 8 years as the negative group. Similarly, we chose the ovarian cancer patients with survival time less 1 year as positive samples and longer than 5 years as negative samples. With this criteria, 103 breast cancer samples (56 positive and 47 negative) and 124 ovarian cancer samples (46 positive and 78 negative) were selected. To reduce the computational load, we sampled data with 150k bp per probe resolution. Based on the genetic relevance of breast cancer and ovarian cancer described in OMIM, we focused on chromosomes 3, 8, 10, 13 and 17 in this analysis. The number of components were chosen to explain between 60%-75% of variance in each chromosome respectively. Since these are two different but related cancer types, we took a smaller ratio of τ/k as 60%.

Similarly, leave-one-out classification was performed on the coefficient matrices learned by TLFL, pool FLLat and split FLLat. The results are shown in Table III. TLFL performed similar classification to FLLat on chromosome 3 and 8 but better on the other chromosomes and overall average of both the breast cancer and ovarian cancer datasets.

To detect more focal CNV events (short CNV regions), we increased the hyper-parameter of common components γ_C and λ_C by multiplying a factor 2.5 and reran TLFL on both datasets. The common CNVs between breast cancer and ovarian cancer detected by TLFL are shown in Figure 9. Eighteen known cancer genes locate in these very focal CNV regions. thirteen among the eighteen genes (except CCDC6, FAM22A, ZMYM2 and SRSF2. GATA3 is found only related with breast cancer) were reported to play a role in both breast cancer and ovarian cancer as reported by details in Table IV. For example, deletion or hyper-methylation of tumor suppressor FHIT leads to high proliferation of both breast cancer and ovarian cancer [32], [33], [34], [35]; and BRIP1 interacts with BRCA1 and its variants are candidates of breast and ovarian cancer susceptibility [36]. The extensive literature supports that those common CNVs might play an important role in both breast and ovarian cancer.

TABLE III
CLASSIFICATION OF BREAST AND OVARIAN CANCER DATASETS.

Chr	Breast cancer			Ovarian cancer			Average		
	TLFL	pool FLLat	split FLLat	TLFL	pool FLLat	split FLLat	TLFL	pool FLLat	split FLLat
3	0.5777	0.5922	0.5971	0.5363	0.6048	0.5040	0.5570	0.5985	0.5506
8	0.4466	0.4223	0.4612	0.4234	0.4153	0.5081	0.4350	0.4188	0.4846
10	0.6553	0.5194	0.5922	0.4758	0.3992	0.4516	0.5656	0.4593	0.5219
13	0.5194	0.4951	0.4612	0.5887	0.5887	0.5847	0.5541	0.5419	0.5229
17	0.5291	0.5049	0.5194	0.5766	0.5645	0.5323	0.5529	0.5347	0.5258
ave	0.5456	0.5068	0.5262	0.5202	0.5145	0.5161	0.5329	0.5107	0.5212

VI. CONCLUSIONS

Application of transfer learning to CNV analysis across multiple cancer types is promising since CNVs are a hallmark of cancer genomes. To the best of our knowledge, TLFL is the first transfer learning method to utilize multiple cancer domains for detecting common and domain-specific CNVs as fused latent components. The transfer learning enables sharing information in datasets of different cancer domains to discover latent CNV features that can explain common and domain-specific cancer characteristics and better classify patient samples as shown in the experiments. In the recent TCGA (The Cancer Genome Atlas) initiative, more and more CNV datasets are becoming available for 21 types of cancer. It is expected that transfer learning will play an important role in the comparative analysis of the large patient cohorts

[1] https://tcga-data.nci.nih.gov/.

TABLE IV
CANCER GENES IN COMMON COMPONENTS

Gene	Association with breast cancer and ovarian cancer	Hyperlink to reference
MLH1	Loss of MLH1 plays a role in drug resistance in breast cancer; methylation of the hMLH1 promoter is possibly related to cisplatin-resistance in ovarian cancer.	Mackay, H. J., et al. Samimi, Goli, et al. Strathdee, G., et al.
FHIT	Deletion or hyper-methylation of tumor suppressor FHIT leads to high proliferation of both breast cancer and ovarian cancer.	Fullwood, P., et al. Dhillon, V.S., et al. Campiglio, M., et al. Zochbauer-Muller, S., et al.
TFRC	TFRC together with ACTB are used for breast cancer quantification; TFRC expresses differently between normal and poorly differentiated serous papillary adenocarcinoma (PD-SPA) of the ovary.	Majidzadeh-A, K., et al. Martoglio, A. M., et al.
BMPR1A	BMPR1A highly expresses in breast cancer and ovarian cancer.	Alarmo, E. L., et al. Shepherd, T. G., et al. Bowen, N. J., et al.
CCDC6	Lack of evidence	
FAM22A	Lack of evidence	
FGFR2	Four SNPs of FGFR2 are confirmed highly associated with breast cancer and FGFR2 expresses increasingly in the rare homozygotes; combining FGFR2 inhibitors with platinum-containing cytotoxic agents for the treatment of epithelial ovarian cancer may yield increased anti-tumor activity.	Hunter, D. J., et al. Meyer, K. B., et al. Cole, C., et al.
GATA3	Low GATA3 expression is associated with higher histologic grade and short survival time in breast cancer; No direct evidence to show relation between GATA3 with ovarian cancer.	Mehra, R., et al. Hoch, R. V., et al.
MYST4	MYST4 is up-regulated in ER-positive breast cancer cells and ovarian cancer cells.	Kok, M., et al. Vignati, S., et al.
PTEN	PTEN may suppress tumor cell growth and regulate tumor cell invasion and metastasis through interactions at focal adhesions in breast cancer; PTEN mutations are frequent in endometrioid ovarian tumors.	Li, J., et al. Obata, K., et al.
FAS	FAS is a reliable prognostic marker to predict DFS and OS in patients with early breast cancer; Decreased sensitivity to Fas-mediated apoptosis could contribute to ovarian tumorigenesis and may play a role in ovarian tumorigenesis.	Alo, P. L., et al. Baldwin, R. L., et al. Meinhold-Heerlein, I., et al.
RB1	RB1 is most likely involved in the development of breast cancer; Two SNPs of RB1 showed significant association with ovarian cancer risk.	Spandidos, D. A., et al. Song, H., et al.
ZMYM2	Lack of evidence	
BRCA1	The 17q-linked BRCA1 gene is identified to have influences susceptibility to breast and ovarian cancer.	Ford, D., et al. Miki, Y., et al.
BRIP1	BRIP1 interacts with BRCA1 and its variants are candidates of breast and ovarian cancer susceptibility.	Song, H., et al.
SEPT9	Increased SEPT9_v1 expression contributes to the malignant pathogenesis of some breast tumors; Experiment shows consistent and specific overexpression of both SEPT9_v1 and SEPT9_v4 transcripts in the epithelial component of ovarian tumors.	Gonzalez, M. E., et al. Scott, M., et al.
SRSF2	Lack of evidence	
YWHAE	Expression level upregulated gene YWHAE together with other 5 genes show a significant association to both disease-free and overall survival in breast cancer; YWHAE is identified from the TOV-112D ovarian cancer cell line.	Cimino, D., et al. Gagné J. P., et al.

to improve the current knowledge of cancer development and progression in the light of both common and specific cancer CNVs.

ACKNOWLEDGMENT

This work is supported by NSF grant #IIS1117153 and #IIS1149697. The results published in this paper are in part based upon data generated by The Cancer Genome Atlas established by the NCI and NHGRI. Information about TCGA and the investigators and institutions who constitute the TCGA research network can be found at http://http://cancergenome.nih.gov/. The dbGaP accession number to the specific version of the TCGA dataset is phs000178.v8.p7.

REFERENCES

[1] M. Baudis, “Genomic imbalances in 5918 malignant epithelial tumors: an explorative meta-analysis of chromosomal CGH data,” *BMC cancer*, vol. 7, no. 1, p. 226, 2007.

[2] F. Mitelman, B. Johansson, and F. Mertens, *Mitelman database of chromosome aberrations in cancer*. Cancer Genome Anatomy Project., 2007.

[3] L. Feuk, A. Carson, and S. Scherer, “Structural variation in the human genome,” *Nature Reviews Genetics*, vol. 7, no. 2, pp. 85–97, 2006.

[4] R. Redon *et al.*, “Global variation in copy number in the human genome,” *nature*, vol. 444, no. 7118, pp. 444–454, 2006.

[5] D. Pinkel *et al.*, “High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays,” *Nature genetics*, vol. 20, no. 2, pp. 207–211, 1998.

[6] D. Pinkel and D. Albertson, “Array comparative genomic hybridization

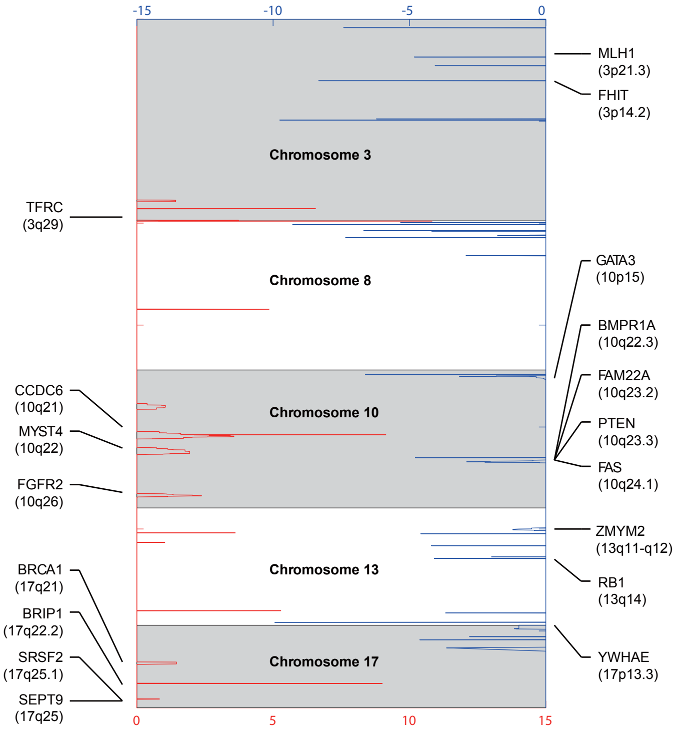


Fig. 9. Common CNV events in breast cancer and ovarian cancer with co-located cancer genes annotated. Amplification (red) and deletion (blue) CNV events are plot along the selected chromosomes.

and its applications in cancer," *Nature genetics*, vol. 37, pp. S11–S17, 2005.

[7] M. Stratton, P. Campbell, and P. Futreal, "The cancer genome," *Nature*, vol. 458, no. 7239, pp. 719–724, 2009.

[8] R. Beroukhim *et al.*, "The landscape of somatic copy-number alteration across human cancers," *Nature*, vol. 463, no. 7283, pp. 899–905, 2010.

[9] A. Olshen, E. Venkatraman, R. Lucito, and M. Wigler, "Circular binary segmentation for the analysis of array-based DNA copy number data," *Biostatistics*, vol. 5, no. 4, pp. 557–572, 2004.

[10] E. Venkatraman and A. Olshen, "A faster circular binary segmentation algorithm for the analysis of array CGH data," *Bioinformatics*, vol. 23, no. 6, pp. 657–663, 2007.

[11] J. Fridlyand, A. Snijders, D. Pinkel, D. Albertson, and A. Jain, "Hidden markov models approach to the analysis of array CGH data," *Journal of multivariate analysis*, vol. 90, no. 1, pp. 132–153, 2004.

[12] S. Stjernqvist, T. Rydén, M. Sköld, and J. Staaf, "Continuous-index hidden markov modelling of array CGH copy number data," *Bioinformatics*, vol. 23, no. 8, pp. 1006–1014, 2007.

[13] F. Picard, S. Robin, M. Lavielle, C. Vaisse, and J. Daudin, "A statistical approach for array CGH data analysis," *BMC bioinformatics*, vol. 6, no. 1, p. 27, 2005.

[14] P. Hupé, N. Stransky, J. Thiery, F. Radvanyi, and E. Barillot, "Analysis of array CGH data: from signal ratio to gain and loss of DNA regions," *Bioinformatics*, vol. 20, no. 18, pp. 3413–3422, 2004.

[15] R. Tibshirani and P. Wang, "Spatial smoothing and hot spot detection for CGH data using the fused lasso," *Biostatistics*, vol. 9, no. 1, pp. 18–29, 2008.

[16] F. Rapaport, E. Barillot, and J. Vert, "Classification of arrayCGH data using fused svm." *Bioinformatics*, vol. 24, no. 13, pp. i375–82, 2008.

[17] G. Nowak, T. Hastie, J. Pollack, and R. Tibshirani, "A fused lasso latent feature model for analyzing multi-sample aCGH data," *Biostatistics*, vol. 12, no. 4, pp. 776–791, 2011.

[18] Z. Tian, H. Zhang, and R. Kuang, "Sparse group selection on fused lasso components for identifying group-specific DNA copy number variations," in *Proceedings of IEEE International Conference on Data Mining*, 2012.

[19] S. Diskin, T. Eck, J. Greshock, Y. Mosse, T. Naylor, C. Stoeckert, B. Weber, J. Maris, and G. Grant, "Stac: A method for testing the significance of DNA copy number aberrations across multiple array-CGH experiments," *Genome research*, vol. 16, no. 9, pp. 1149–1158, 2006.

[20] M. Guttman and others., "Assessing the significance of conserved genomic aberrations using high resolution genomic microarrays," *PLoS Genetics*, vol. 3, no. 8, p. e143, 2007.

[21] R. Beroukhim *et al.*, "Assessing the significance of chromosomal aberrations in cancer: methodology and application to glioma," *Proceedings of the National Academy of Sciences*, vol. 104, no. 50, pp. 20 007–20 012, 2007.

[22] D. Witten, R. Tibshirani, and T. Hastie, "A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis," *Biostatistics*, vol. 10, no. 3, pp. 515–534, 2009.

[23] S. Pan and Q. Yang, "A survey on transfer learning," *Knowledge and Data Engineering, IEEE Transactions on*, vol. 22, no. 10, pp. 1345–1359, 2010.

[24] M. Long, J. Wang, G. Ding, W. Cheng, X. Zhang, and W. Wang, "Dual transfer learning," in *Proceedings of the 12th SIAM International Conference on Data Mining*, 2012.

[25] W. Dai, G. Xue, Q. Yang, and Y. Yu, "Co-clustering based classification for out-of-domain documents," in *Proceedings of the 13 th ACM SIGKDD international conference on Knowledge discovery and data mining*, vol. 12, no. 15, 2007, pp. 210–219.

[26] Z. Wang, Y. Song, and C. Zhang, "Knowledge transfer on hybrid graph," in *Proceedings of the 21st international jont conference on Artifical intelligence*. Morgan Kaufmann Publishers Inc., 2009, pp. 1291–1296.

[27] F. Zhuang, P. Luo, Z. Shen, Q. He, Y. Xiong, Z. Shi, and H. Xiong, "Collaborative dual-plsa: mining distinction and commonality across multiple domains for text classification," in *Proceedings of the 19th ACM international conference on Information and knowledge management*. ACM, 2010, pp. 359–368.

[28] F. Zhuang, P. Luo, H. Xiong, Q. He, Y. Xiong, and Z. Shi, "Exploiting associations between word clusters and document classes for cross-domain text categorization," *Statistical Analysis and Data Mining*, vol. 4, no. 1, pp. 100–114, 2011.

[29] J. Liu, S. Ji, and J. Ye, *SLEP: Sparse Learning with Efficient Projections*, Arizona State University, 2009. [Online]. Available: http://www.public.asu.edu/~jye02/Software/SLEP

[30] E. Blaveri *et al.*, "Bladder cancer stage and outcome by array-based comparative genomic hybridization," *Clinical cancer research*, vol. 11, no. 19, pp. 7012–7022, 2005.

[31] N. Stransky *et al.*, "Regional copy number–independent deregulation of transcription in cancer," *Nature genetics*, vol. 38, no. 12, pp. 1386–1396, 2006.

[32] P. Fullwood and others., "Detailed genetic and physical mapping of tumor suppressor loci on chromosome 3p in ovarian cancer." *Cancer Res*, vol. 59, no. 18, pp. 4662–4667, 1999.

[33] M. Campiglio, Y. Pekarsky, S. Menard, E. Tagliabue, S. Pilotti, and C. Croce, "FHIT loss of function in human primary breast cancer correlates with advanced stage of the disease." *Cancer Res*, vol. 59, no. 16, pp. 3866–3869, 1999.

[34] V. Dhillon, M. Shahid, and S. Husain, "Cpg methylation of the FHIT, FANCF, cyclin-D2, BRCA2 and RUNX3 genes in granulosa cell tumors (gcts) of ovarian origin." *Mol Cancer*, vol. 3, p. 33, 2004.

[35] S. Zochbauer-Muller, K. Fong, A. Maitra, S. Lam, J. Geradts, R. Ashfaq, A. Virmani, S. Milchgrub, A. Gazdar, and J. Minna, "5' cpg island methylation of the fhit gene is correlated with loss of gene expression in lung and breast cancer." *Cancer Res*, vol. 61, no. 9, pp. 3581–3585, 2001.

[36] H. Song and others., "Tagging single nucleotide polymorphisms in the BRIP1 gene and susceptibility to breast and ovarian cancer." *PLoS One*, vol. 2, no. 3, p. e268, 2007.