Genes Genom (2015) 37:525–535
DOI 10.1007/s13258-015-0281-6

RESEARCH ARTICLE

Effects of omics data combinations on in silico tumor-normal tissue classification

Ho-Sik Seok[1] · Seung Hwan Seok[2] · Jaebum Kim[1]

Received: 17 February 2015 / Accepted: 27 March 2015 / Published online: 1 April 2015

Abstract A disease can be characterized by various attributes such as genomic, epigenetic, and transcriptomic features beyond physiological symptoms. The accumulation of vast datasets allows us to investigate the relative effectiveness of each omics data and their combinations for in silico analysis of diseases. Here, we employed a classification method with the well-established measure of information gain for the computational analysis of the effect of the aggregation of omics data, especially for the task of in silico classification of tumor-normal samples for bladder urothelial carcinoma and kidney renal papillary cell carcinoma. We observed that the combination of multi-omics data such as copy number variation, DNA methylation, RNA-Seq, and somatic mutations have beneficial effects. The quantitative analysis using information gain and various measures for classification-performance showed that the combination of multiple omics data improved the performance in general. The qualitative analysis referring previous researches also confirmed the relevance of genes with higher information gain to target diseases. Our results report that the combination of multiple omics data is beneficial and the information gain which focuses on the distribution of attributes across target domains could be useful as an indicator of the effect of each omics data on tumor-normal sample classification.

Keywords Multiomics data · Copy number variation · DNA methylation · RNA-Seq · Somatic mutations · The cancer genome atlas (TCGA)

Introduction

Most molecular studies on diseases have focused on just one or two data types as an attempt to show more distinct differences among target samples. However, recent efforts to produce massive data such as the TCGA (The Cancer Genome Atlas) project (Cancer Genome Atlas Research Network 2008) allowed researchers to analyze comprehensive molecular landscapes of diseases. Now various windows such as copy number variation (CNV), DNA methylation, RNA-Seq or somatic mutations are available to understand the molecular landscape of a target disease.

CNVs are structurally variant regions where copy number differences have been observed between two or more genomes (Feuk et al. 2006). CNV has received considerable interests because it is one of the important sources of genetic variation causing phenotype diversity (Henrichsen et al. 2009). It is believed that CNVs can be a predominant mechanism driving gene and genome evolution (Zhang et al. 2009). Due to its prevalence, CNVs could

Ho-Sik Seok and Seung Hwan Seok have contributed equally to this work.

Electronic supplementary material The online version of this article (doi:10.1007/s13258-015-0281-6) contains supplementary material, which is available to authorized users.

✉ Jaebum Kim
jbkim@konkuk.ac.kr

Ho-Sik Seok
hosik.seok@gmail.com

Seung Hwan Seok
stone1018@naver.com

[1] Department of Animal Biotechnology, Konkuk University, Seoul 143-701, Korea

[2] Department of Chemical and Biomolecular Engineering, Yonsei University, Seoul 120-749, Korea

drive significant intraspecific genetic variation (Henrichsen et al. 2009; She et al. 2008; Shlien and Malkin 2009). For example, consistent increase in the frequency of rare CNVs was reported among breast cancer cases (Pylkas et al. 2012). It was found that CNVs are specific to cancer types and reproducible from cell to cell (Ni et al. 2013). Because human populations show extensive polymorphism (both insertions and deletions) in the number of copies of chromosomal segments (Hastings et al. 2009), CNVs have the potential for understanding underlying factors in human diseases.

DNA methylation provides stability and diversity to the cellular phenotype through chromatin marks affecting local transcriptional potential. Methylation of DNA cytosine residues at the carbon 5 position in CpG dinucleotides is a common epigenetic mark in many eukaryotes (Laird 2010). In particular, aberrant promoter hypermethylation associated with inappropriate gene silencing affects virtually every step in tumor progression (Jones and Baylin 2002). Especially, CpG island methylation plays an important role in transcriptional regulation. Between 5 and 10 % of normally unmethylated CpG promoter islands become abnormally methylated in various cancer genomes (Dawson and Kouzarides 2012). In breast cancer transcriptionally repressed genes become aberrantly methylated, and the affected genes can be used to distinguish breast tumors of epithelial and mesenchymal lineage (Sproul et al. 2011). Four DNA methylation-based subgroups of colorectal cancer were identified using cluster analyses (Hinoue et al. 2012). Significant promoter hypermethylation in at least 50 % of CpG sites in two genes, *ABHD9* and *HOXD3*, was found in tumors from recurring patients compared with those without recurrence (Stott-Miller et al. 2014). Changes in DNA methylation patterns play a critical role in development, differentiation and diseases such as multiple sclerosis, diabetes, schizophrenia, aging, and multiple forms of cancer (Bibikova et al. 2011).

RNA-Seq is an approach for transcriptome profiling based on deep-sequencing technologies. RNA-Seq produces a genome-scale transcription map consisting of both the transcriptional structure and/or level of expression for each gene (Wang et al. 2009). Several RNA-Seq-based technologies, such as improvements in transcription start site mapping strand-specific measurements and small RNA characterization, have allowed more complete observation of RNA transcripts (Ozsolak and Milos 2011). The splicing signatures of the subtypes of breast cancer were revealed using RNA-Seq (Eswaran et al. 2013). RNA-Seq was also utilized to define the subsets of pancreatic circulating tumor cells (Ting et al. 2014). From the comprehensive landscape of the transcriptome profiles of prostate cancer in the Chinese population, it was reported that there exists wide diversity in gene fusions, long noncoding RNAs, alternative splicing and somatic mutations (Ren et al. 2012).

Although genetic mutations causing human disease can be inherited from one's parents, most mutations that cause cancer as well as other diseases arise somatically (Poduri et al. 2013; Stratton 2011). In addition to diseases, the somatic mutation theory of aging posits that the accumulation of mutations in the genetic material of somatic cells as a function of time results in a decrease in cellular function (Kennedy et al. 2012). With the widespread use of next-generation sequencing technologies, high-throughput mutation profiling identifies frequent somatic mutations in cancers. It was observed that 14.4 % of gastric cancer patients harboring mutations (Lee et al. 2012). The integrated analysis based on 27 cancer types illustrated that the variation in mutation frequency can be partly explained by cancer types, and the mutation spectra also vary across cancer types (Watson et al. 2013). The fact that examining the patterns of somatic mutations is not enough to decipher individual mutational signatures that are operative in each sample (Alexandrov and Stratton 2014) indicates the need for multiplatform-based approach for cancer analysis.

The TCGA project (Cancer Genome Atlas Research Network 2008) offers various types of data for a number of cancers. Currently, the analyzed data of 23 types of cancers are available without limitation and nine types of cancers are available under publication limitations. For each cancer type, a user can download clinical data, images microsatellite instability, DNA sequencing, miRNA sequencing, protein expression, mRNA sequencing, total RNA sequencing array-based expression, DNA methylation and copy number data. The publication of the TCGA data enabled multiplatform-based analysis of various cancers.

There have been active researches to understand a disease from various dimensions. The Cancer Genome Atlas Research Network produced a catalogue of molecular aberrations causing ovarian cancer (Cancer Genome Atlas Research Network 2011). The landscape of somatic genomic alterations of chromophobe renal cell carcinomas was produced based on multidimensional and comprehensive characterization of the molecular basis of a target disease (Davis et al. 2014). Hoadley and colleagues identified 11 "integrated subtypes" from 12 tumor types, which were consistent with the histological classification. Among the cases, approximately 10 % were reclassified based on the multiple assay platforms with significantly increased accuracy in the prediction of clinical outcomes by the newly defined integrated subtypes (Hoadley et al. 2014). However, these researches did not consider the effectiveness of each molecular assay platform and their combinations for explaining the difference between tumor and normal tissues. In this paper, we analyzed bladder urothelial carcinoma (BLCA) and kidney renal papillary cell carcinoma (KIRP) from the viewpoint of the effectiveness of aberrations in CNV, DNA methylation, RNASeq version 2 (RNASeqV2) which is similar to RNA-Seq in terms of the employment of sequencing data but uses a different set of algorithms for determining expression levels, and somatic mutations (SNPs, insertions, and deletions of DNA bases) for in silico classification. We measured the effectiveness in terms of quantitative and qualitative performance by using information gain (Mitchell 1997), and observed the difference in classification performance by CNV, DNA methylation, RNASeqV2, somatic mutations, and their combinations.

Materials and methods

Data preparation

In this study, we used samples generated by The Cancer Genome Atlas (http://cancergenome.nih.gov/). Among the various types of data, we selected CNV (SNP array), DNA methylation, RNASeqV2, and somatic mutations. These four data types represent characteristics of the selected cancer types in terms of genomics, epigenomics, and transcriptomics. Especially for this study, we selected BLCA, and KIRP as target diseases, and downloaded 28 cases of the BLCA data (14 tissue samples for each tumor and normal case) and 42 cases of the KIRP data (21 tissue samples for each tumor and normal case). All of the 70 cases of samples were composed of observations on all of CNV, DNA methylation, RNASeqV2, and somatic mutations (Table 1). Our analyses were based on relevant genes with the above data types, which were identified by using the following pre-processing steps. Here we followed approaches available in literature.

The TCGA consortium performed CNV calling and provided segment mean values for all detected CNVs. Especially the segment mean value is computed as $\log_2(\text{observed intensity}/\text{reference intensity})$ and represents the extent of copy number change. We regarded CNVs with segment mean value of greater than 0.2 as *amplifications* and less than -0.2 as *deletions*, which was decided based on a previous study (Laddha et al. 2014).

To increase the likelihood of identifying differentially methylated genes, we only considered genes that met the following criteria. We first extracted a set of differentially methylated CpG sites (DMC) in a promoter region (size of 1.5 Kbp) of each gene, and then searched for differentially methylated genes based on the state of their DMCs. The degree of methylation of each probe (target CpG site) is represented using a β value. The β value is a continuous variable between 0 and 1, with β values approaching 1 (or 0) indicating complete methylation (or non-methylation) (Kim et al. 2012). This β value is used to determine a hypermethylated or hypomethylated DMC. In the case of normal tissues, the β values of ≥ 0.7 and ≤ 0.3 were used as a threshold for hyper and hypomethylation respectively. In the case of tumor tissue, high methylation values were rarely observed due to heterogeneous mix of cell types in each sample. Thus β value of 0.3 was used as a threshold for distinguishing hyper or hypomethylated states (Sproul et al. 2011). Finally, at least three quarters of multiple DMCs in the same promoter region should have the same direction of methylation and at least one DMC have at least 0.35 mean methylation difference between tumor and normal phenotypes to determine aberrant DNA methylation of a gene.

In order to find differentially expressed (DE) genes using RNASeqV2 data, we employed the *R* package EBSeq (Leng et al. 2013), which can compute the fold change (FC) value. We classified genes as DE genes when $|FC| > 2$. The DE genes were further partitioned into up- or down-regulated genes based on the FC values ($FC > 2$: up-regulated, $FC < -2$: down-regulated) (Guo et al. 2013). From the somatic mutation data, we collected genes that contain single nucleotide polymorphisms (SNPs), insertions, and/or deletions.

The above data is summarized in Table 1. There exist distinct difference between normal samples and tumor samples from both of the BLCA and KIPP data sets. In terms of the CNV data, tumor samples from the BLCA data set have more than 17.73 times and 4.58 times amplifications and deletions compared to normal samples. In contrast, the KIRP data set shows smaller differences (7.69 and 2.21 times, respectively). On average, 76.71 genes are hypermethylated in the case of tumor samples in the BLCA data set. Although more genes in normal samples from the BLCA data set is hypermethylated than genes in tumor samples of the KIRP data set, 13 times as many genes were hypermethyalted in tumor samples compared to normal samples in the case of the KIRP data set. In the RNASeqV2 data, tumor samples show fewer up-regulated genes and more down-regulated genes in both of the BLCA and KIRP data sets. The difference is less distinct in the somatic mutation data. However it is clear that tumor samples show more SNPs, insertions, and deletions per samples in both of the BLCA and KIRP data sets.

Information gain and decision tree

A gene with a specific state of CNV, DNA methylation, RNASeqV2, or somatic mutations could be regarded as an attribute capable of explaining associated tumor or normal tissue samples. A proper quantitative measure of the worth of an attribute is *information gain* (Mitchell 1997), which can be used as an indicator for the quantitative

Table 1 Data summary

	BLCA[a]		KIRP[b]	
	Normal	Tumor	Normal	Tumor
Number of samples	14	14	21	21
Total size of data (MB)	657.67	658.84	987.81	984.39
CNV				
Averaged number of amplifications per sample	3.57	63.29	1.95	15.00
Averaged number of deletions per sample	14.57	66.79	17.14	37.86
DNA methylation				
Averaged number of hypermethylated genes per sample	16.50	76.71	0.67	9.05
Averaged number of hypomethylated genes per sample	40.71	9.93	4.81	1.00
RNASeqV2				
Averaged number of up-expressed genes per sample	670.71	431.36	790.57	421.24
Averaged number of down-expressed genes per sample	428.64	667.71	411.43	782.43
Somatic mutations				
Averaged number of SNPs per sample	265.93	474.29	67.38	75.52
Averaged number of insertions per sample	2.86	5.36	2.90	3.14
Averaged number of deletions per sample	6.50	11.86	7.90	8.86

[a] Bladder urothelial carcinoma

[b] Kidney renal papillary cell carcinoma

importance of the gene for the tumor or normal tissue samples. Information gain is defined based on the Shannon's entropy formula if the target domain can take on c different classes,

$$Entropy(S) \equiv \sum_{i=1}^{c} -p_i \log_2 p_i \tag{1}$$

where S is the set of samples (the BLCA or KIRP data sets in our case), p_i is the proportion of S belonging to class i (normal or tumor in our case). *Entropy* characterizes the purity of a collection of samples. Entropy is interpreted as the minimum number of bits of information needed to encode the classification of an arbitrary member of S. Thus, one can quantify the effectiveness of an attribute based on the expected reduction in entropy and information gain is a measure for the expected reduction in entropy. Information gain, $Gain(S, A)$ of attribute A relative to a set of samples S is defined as,

$$Gain(S,A) \equiv Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{S} Entorpy(S_v) \tag{2}$$

where $Values(A)$ is the set of all possible values for attribute A, and S_v is the subset of S for which attribute A has value v. In our case, A denotes genes and $Values(A)$ can have different values based on which omic data type is in consideration. By using the pre-processing steps described in the previous section, we defined the following values for each different omics data type:

- CNV: amplification, deletion, and non-variant
- DNA methylation: hypermethylated, hypomethylated, and non-methylated
- RNASeqV2: up-regulated, down-regulated, and normal
- Somatic mutations: SNPs, insertions, deletions, and no mutation

For example, if we are testing the effect of CNV, the values of A can be "amplification", "deletion", and "non-variant." If we are interested in the joint effect of CNV and RNASeqV2, nine different values are possible from the three values of each CNV and RNASeqV2.

Because information gain is the measure of the expected entropy-reduction of an attribute A, information gain is utilized to build a decision tree (Mitchell 1997; Quinlan 1993). The decision tree is composed of nodes specifying a test of an attribute, and an instance is classified by sorting down the decision tree from the root to some leaf nodes. Information gain is used to select the best attribute for current decision node. In this paper, we utilized the implementation of a decision tree in WEKA (Hall et al. 2009), a public data mining software, for measuring the contribution of CNV, DNA methylation, RNASeqV2, somatic mutations, and their combinations for distinguishing tumor-normal tissue samples.

Performance measure

To evaluate the effect of omic data and their combinations on tumor and normal tissue classification, we used two widely used measures, precision and recall, in the field of data mining. In terms of true positive (TP), false negative (FN), and false positive (FP), precision and recall are defined as:

$$\text{Precision} = \frac{TP}{TP + FP} \quad (3)$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad (4)$$

Results and discussion

Information gain and classification performance

Information theory-based approach was already introduced for transcriptomes, where gene diversity, the specialization of transcriptomes and gene specificity were defined using Shannon's entropy (Martinez and Reyes-Valdes 2008). In this paper, we adopted similar approach to analyze the effectiveness of CNV, DNA methylation, RNASeqV2, somatic mutations and their combinations for explaining tumor-normal tissues. Specifically, for two target diseases, BLCA, and KIRP, we collected total 70 samples containing CNV, DNA methylation, RNASeqV2, and somatic mutation data with information whether each sample is normal or tumor ("Materials and methods"). We created total ten different evaluation cases based on the number and type of used omics data: CNV only (C), DNA methylation only (M), RNASeqV2 only (R), somatic mutations only (S), CNV and DNA methylation (CM), CNV and RNASeqV2 (CR), CNV and somatic mutations (CS), DNA methylation and RNASeqV2 (MR), DNA methylation and somatic mutations (MS), and RNASeqV2 and somatic mutations (RS). For each different evaluation case, we computed the information gain of each gene and re-classified the samples based on a decision tree by only using the assigned omic data, and the classification accuracy was measured by using precision and recall ("Materials and methods"). Here the combinations with more than two omics data types could not be used because we could not find genes that contain more than two omic data signatures due to our stringent rules ("Materials and methods"). We used the average of the information gain across all genes as the information gain of each evaluation case.

Table 2 summarizes the important findings in this study. In general the aggregation of omics data clearly increased information gain compared with results from single omics data. For example, in the case of the BLCA data set, the information gain of the combinations of two omics data types, except for MS, was higher than or equal to each one of single omics data. A similar pattern was also observed in the KIRP data set. In addition, the aggregation of multiple omics data is beneficial especially for omics data with lower discriminating performance. In the BLCA data set, the precision and recall of CR, CS, MR, MS, and RS are better than those values of R and S only cases. The KIRP data set also showed a similar pattern. For every measure, the aggregation of multiple omics data resulted in better performances in Table 2, which confirms the benefits of the aggregation of multiple omics data. Figure 1 reports correlation between precision/recall and the average information gain of each evaluation cases ("Materials and methods"). In general, the precision/recall and the information gain have positive correlations in both of the BLCA and KIRP data sets. For instance, the Pearson correlation coefficients between precision and information gain were 0.62 and 0.71 for BLCA and KIRP respectively, and 0.65 and 0.74 for BLCA and KIRP respectively in the comparison of recall and information gain. Overall these results confirmed the suitability of the considered platforms or their combinations for in silico classification of tumor and normal tissue samples.

Effect of multi-omics data combinations

The information gain-based multi-omics data analysis produced mixed results. As described in the previous section, the aggregation of two omics data resulted in higher information gain in almost all the cases for both of the BLCA and KIRP data sets. However, the improvement in precision/recall was less clear although all average performance of multi-omics data sets outperformed the performance of single-omics datasets. In terms of precision/recall, the combinations of multiple-omics datasets produced improvement for the omics data with lower information gain. The most striking example was somatic mutations. The precision/recall performance of somatic mutations were 0.14/0.14 and 0.18/0.19 for the BLCA and KIRP datasets, respectively. These low precision/recall were improved to 0.91/0.89 and 0.93/0.93 for BLCA and KIRP respectively through the combination with DNA methylation. The low information gain due to the lack of discriminating process to identify significantly mutated genes was overcome with the combination of other omics-data. For somatic mutations, the improvement in precision/recall were also observed in the cases of CS and RS in both of BLCA and KIRP. For the remaining cases, the combination of multiple omics data tends to be beneficial for the omics data with lower information gain and not very useful for the omics data with higher information gain in general. The example cases are CM, MR, and MS for BLCA, and CM, CS, MR, MS and RS for KIRP. There were cases where the combination of multiple omics data produced better precision/recall than each single-omics data set. In

Table 2 Information gain and classification accuracy

	BLCA[a]			KIRP[b]		
	IG−mean (IG−SD)	Precision	Recall	IG−mean (IG−SD)	Precision	Recall
C	0.20 (0.09)	0.79	0.75	0.16 (0.08)	0.87	0.86
M	0.52 (0.29)	1.00	1.00	0.60 (0.22)	1.00	1.00
R	0.43 (0.26)	0.71	0.68	0.14 (0.12)	0.65	0.64
S	0.00 (0.01)	0.14	0.14	0.00 (0.00)	0.18	0.19
CM	0.63 (0.25)	0.79	0.79	0.67 (0.18)	0.91	0.91
CR	0.54 (0.22)	0.73	0.71	0.28 (0.12)	0.89	0.88
CS	0.20 (0.09)	0.81	0.79	0.17 (0.09)	0.86	0.81
MR	0.82 (0.20)	0.82	0.82	0.80 (0.13)	0.96	0.95
MS	0.51 (0.25)	0.91	0.89	0.60 (0.11)	0.93	0.93
RS	0.44 (0.26)	0.82	0.82	0.16 (0.13)	0.60	0.60
Average performance						
One-platform	0.19 (0.09)	0.75	0.71	0.16 (0.08)	0.84	0.83
Multi-omics	0.34 (0.15)	0.79	0.76	0.25 (0.11)	0.87	0.85

In the first column "C", "M", "R", "S", "CM", "CR", "CS", "MR", "MS", and "RS" mean "CNV only", "DNA methylation only", "RNASeqV2 only", "somatic mutations only", "CNV and DNA methylation", "CNV and RNASeqV2", "CNV and somatic mutations", "DNA methylation and RNA-SeqV2", "DNA methylation and somatic mutations", and "RNASeqV2 and somatic mutations", respectively

In the second row, "IG−mean" and "IG−SD" denote "Information gain−mean" and "Information gain−standard deviation", respectively

[a] Bladder urothelial carcinoma

[b] Kidney renal papillary cell carcinoma

BLCA, the combinations of CS and RS were beneficial for each one of single omics data. In KIRP, CR combinations produced better precision/recall than single omics data sets. Basically, the information gain represents the imbalance in the distribution of attributes across samples. Therefore, the pairing with an attribute with higher information gain is beneficial only for an attribute with lower information gain. This is because the pairing induces more skewed distribution of attributes. For the same reason, the pairing is not beneficial for an attribute with higher information gain.

An important benefit of the multi-omics data-based approach is that it can prevent a "large" platform (with a large number of features) from dominating a solution (Hoadley et al. 2014). Information gain-based approach produced similar effect. An omics-category with lower information gain suffered from the too balanced distribution of attributes among tumor-normal tissue samples. The aggregation with the imbalanced attribute such as DNA methylation induced in more skewed distributions and resulted the improvements in the in silico classification-performance.

Role of genes with higher information gain

For the qualitative analysis of the genes with higher information gain, we investigated the role of genes with higher information gain in the development of target diseases or tumorigenesis in general. For BLCA, we selected 44 genes from the ten omics data-combinations. We investigated 36 genes for KIRP from eight omics data-combinations. Those selected genes were first checked using the CaGe (Park et al. 2012) and GeneCards (http://www.genecards.org) database (Rebhan et al. 1997). A gene covered by one of two databases is regarded as target diseases or tumorigenesis-related genes, and its role was summarized. Genes not covered by two databases were also checked through literature searches. The list of relevant genes is shown in Table 3. We also investigated pathways related with the higher information gain genes by using the NCI-Nature Pathway Interaction Database (Schaefer et al. 2009). The results are listed in supplementary Table 1.

In the case of BLCA, 32 genes were relevant to tumorigenesis in general. Among the relevant genes, *SLC6A6* was associated with higher information gain in terms of CNV. It was found that *SLC6A6* is important for the maintenance of side population cells and their cancer stem cell properties. It was suggested that *SLC6A6* signaling is a significant player in the survival and maintenance of cancer stem cell population and its capacity for tumor initiation, starvation tolerance and multidrug resistance (Yasunaga and Matsumura 2014). *ZNF154* was another informative gene in terms of DNA methylation. *ZNF154* encodes a protein belonging to the zinc finger Kruppel family of

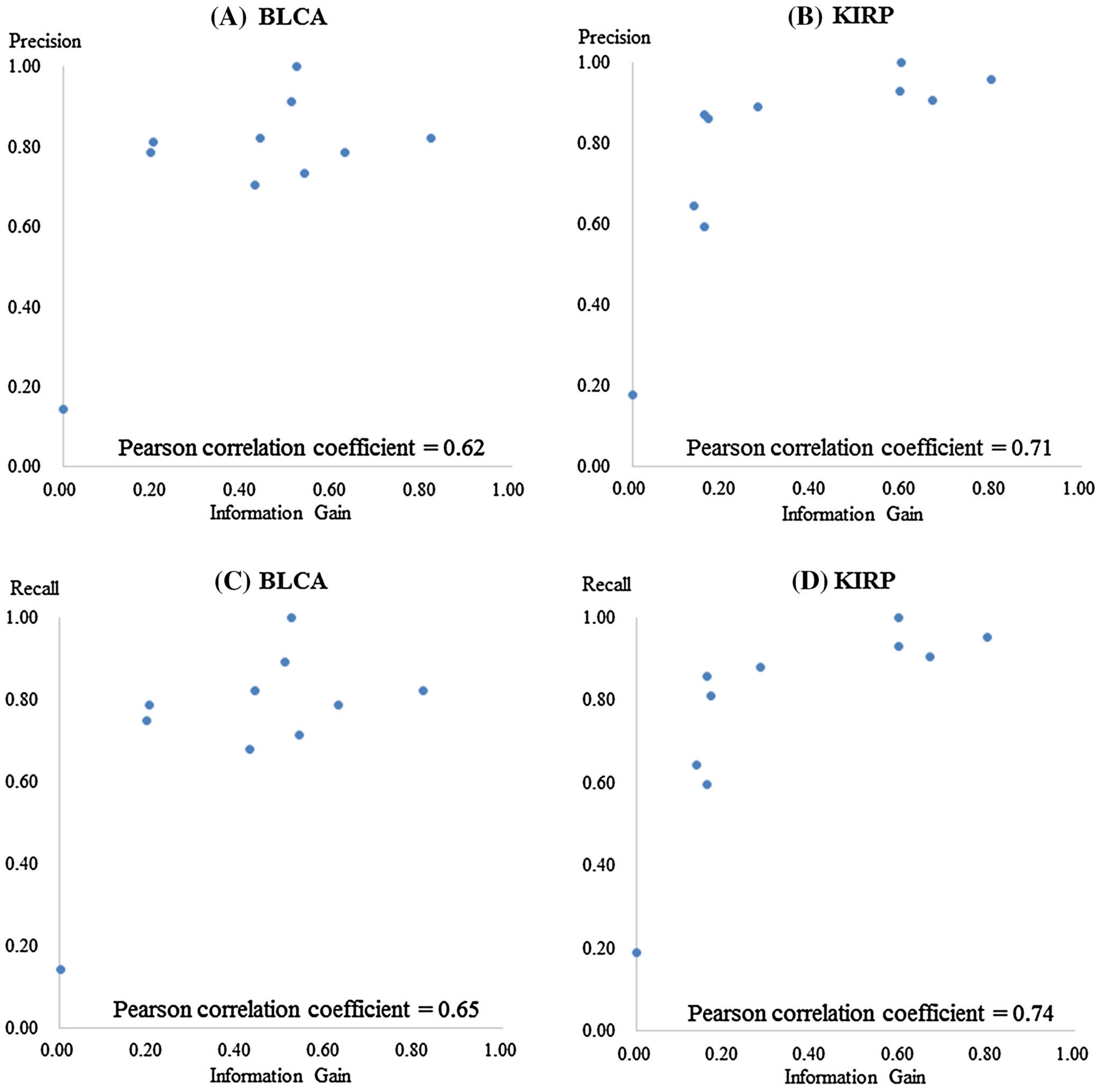

Fig. 1 Correlation between information gain and classification accuracy

transcriptional regulators, whose members are deemed to function in normal/abnormal cell growth and differentiation. The methylation of *ZNF154* was validated as a tumor marker gene (Reinert et al. 2011) and it was shown that it is possible to detect a concomitant tumor recurrence with a single marker *ZNF154* (Reinert et al. 2012). *RBP7* was highly informative in terms of DNA methylation. *RBP7* (also named *CRBP-IV*) is the member of the cellular retinal-binding protein family and it was shown that transcription silencing of this gene by aberrant methylation is involved in the tumorigenesis of human cancers (Kwong et al. 2005). The possibility was raised that the mutation of *RBP7* impaired retinoid's function in breast cancer cells where tamoxifen-induced ZR-75-1 cell death requires intact retinoid signaling (Zarubin et al. 2005). *PI16* was associated with higher information gain in the case of RNASeqV2 and the combination of RNASeqV2 and somatic mutations. It was suggested that *PI16* would be a tumor suppressor and a metastasis enhancer because the cell lines ectopically expressing *PI16* display a net increase in the rate of secondary lesion (Crawford et al. 2008). *MTUS1* was selected considering the combination of CNV

Table 3 The list of genes with higher information gain and their relevance to tumorigenesis in general

Target cancer	Omics combination	Gene name
BLCA[a]	CNV only	***SLC6A6***, *DEFB109P1*, *DEFB109P1B*, *DEFB108P4*, *DEFB109*
	DNA methylation only	***CD8A***, ***ZNF154***, ***RBP7***, ***MIR663A***, ***ZSCAN18***
	RNASeqV2 only	***CLEC3B***, ***PI16***, ***PYGM***, ***ADH1B***, ***XPNPEP2***
	Somatic mutations only	***KIAA0100***, ***ARID1A***, ***MUC16***, ***ELF3***, *ANKRD36*
	CNV and DNA methylation	***CD8A***, ***ZNF154***, ***CMTM2***, ***BOLL***, ***DRD4***
	CNV and RNASeqV2	***PCP4***, ***MYRIP***, ***FAM107A***, *SORCS1*, *VIT*
	CNV and somatic mutations	***MTUS1***, ***RAB11FIP1***, *SLC7A2*, *KCNU1*, *ERICH1-AS1*
	DNA methylation and RNASeqV2	***MMP23B***, ***CDO1***, ***ZNF154***, ***ZIC5***, *NKAPL*
	DNA methylation and somatic mutations	***PCDHA6***, ***CMTM2***, ***PRDM14***, ***FOXG1***, *NKAPL*
	RNASeqV2 and somatic mutations	***KRT24***, ***PI16***, ***ITIH5***, ***ZNF695***, *MYOC*
KIRP[b]	CNV only	***SDK1***, *AC004160.4*, *AC004538.3*, *THSD7A*, *AC073109.2*
	DNA methylation only	***JDP2***, ***IFITM10***, ***KLHDC7B***, ***BNC1***, ***COL9A2***
	RNASeqV2 only	***TMEM207***, ***CALB1***, ***TYRP1***, ***MOGAT2***, ***MT1G***
	Somatic mutations only	***VCP***, ***BCL11B***, *LINC00971*, *AC004381.6*, *CIDECP*
	CNV and DNA methylation	***COL9A2***, ***IFITM10***, ***BNC1***, ***CA3***, ***DRD4***
	CNV and RNASeqV2	***AQP2***, ***AKR1B10***, ***PTPRO***, ***EPO***, *WEE2*
	CNV and somatic mutations	***TTYH3***, ***ACTB***, *HEATR2*, *PMS2CL*, *C7orf26*
	RNASeqV2 and somatic mutations	***AQP2***, ***CDH10***, *SLC7A13*, *UPP2*, *SLC12A3*

Bold face genes denote carcinogenesis-relevant genes verified through literature search

For KIRP, we ignored two omics-combinations where insufficient number of data instances was produced

[a] Bladder urothelial carcinoma

[b] Kidney renal papillary cell carcinoma

and somatic mutations. A hypothesis was suggested that *MTUS1* is involved in the regulation of tumor progression in various malignant diseases including human colon carcinoma. *MTUS1* is down-regulated in undifferentiated tumor cell lines and inhibits tumor cell proliferation after recombinant over-expression (Zuern et al. 2010).

In the case of KIRP, 22 genes were regarded as relevant ones. For example, *JDP2* reported higher information gain in DNA methylation. The role of *JDP2* is prominent in the regulation of the differentiation and proliferation of cells. The overexpression of *JDP2* inhibits the retinoic acid-dependent differentiation of embryonic carcinoma F9 cells (Jin et al. 2002). In mice, the overexpression of *JDP2* induces arrest of the cell cycle. The absence of *JDP2* decreases the expression of both $\mathrm{p}^{16\mathrm{Ink4a}}$ and $\mathrm{p19}^{\mathrm{Arf}}$, which inhibits progression of the cell cycle. It is supposed that *JDP2* not only inhibits the transformation of cells but also plays a role in the induction senescence. These two functions imply that *JDP2* might act as an inhibitor of tumor formation (Nakade et al. 2009). *TMEM207* achieved distinctively higher information gain considering RNASeqV2. *TMEM207* facilitates tumor invasion possibly through binding to WWOX (WW domain-containing oxidoreductase), a molecule plays an important role in the regulation of a wide variety of cellular functions such as protein degradation, transcription, and RNA splicing. Human *TMEM207* was found to be overexpressed in many aggressive gastric signet-ring cell carcinomas and *TMEM207* expression is relatively restricted to the kidney physiologically (Kito et al. 2014). *VCP* was associated with higher information gain considering somatic mutations. *VCP* regulates various cellular processes such as chromatin decondensation, homotypic membrane fusion, and ubiquitin-dependent protein degradation by the proteasome. Interference of proteasome inhibitors with the ubiquitin proteasome pathway leads to the accumulation of proteins engaged in cell cycle progression, which ultimately put a halt to cancer cell division and induce apoptosis (Rastogi and Mishra 2012; Tresse et al. 2010). *BCL11B* was a highly informative gene in the domain of somatic mutations. The *BCL11B* gene is responsible for the regulation of the apoptotic process and cell proliferation. *BCL11B* has recently been identified as a tumor suppressor gene. In particular, *BCL11B* is known as a haplo-insufficient tumor suppressor, the absence of *BCL11B* resulted in vulnerability to DNA replication stress and damage, and down-regulation of *BCL11B* gene by siRNA (small interfering RNA) led to growth inhibition and apoptosis in a human T-ALL cell line (Huang et al. 2012). *EPO* was associated with higher information gain considering the combination of CNV and RNVSeqV2. Tumor necrosis factor-alpha (TNF-alpha) selectively kills tumor cells in vitro and in vivo. It was shown that *EPO* could be used to prevent TNF-alpha-induced erythroid suppression (Johnson et al. 1990) (Supplementary Material for the details of other relevant genes).

Information gain-based analysis of putative significantly mutated genes

Genomic variant causing ‘gain of function’ or ‘loss of function’ plays a key role in cancer diagnostics and targeted therapy (Krishnan and Ng 2012). Therefore, the identification of potential cancer drivers is another role of cancer genomics. The Cancer Genome Atlas Research Network analyzed urothelial bladder carcinoma. As a result, 29 significantly mutated genes and 20 genes with statistically significant focal copy number changes were identified (The Cancer Genome Atlas Research Network 2014). For 29 significantly mutated genes identified in (The Cancer Genome Atlas Research Network 2014), we collected their information gain from our experiment results and found that the average information gain was 0.254. The maximum information gain is 0.705 of *RHOB* and the minimum information gain is 0.0367 of *FOXA1*. Because information gain is quantified by the distribution of attributes in tumor and normal tissue samples, if an attribute is evenly observed in tumor and normal tissue samples then the attribute would have low information gain. As a result, the 29 identified genes reported relatively low average information gain with high standard deviation of 0.146. Among 20 genes with statistically significant focal copy number changes, the average information gain is 0.294. The maximum information gain is 0.610 of *CCNE1* and the minimum information gain is 0.0728 of *CCND1*. The most differentially regulated 48 genes in renal cell carcinoma were identified (Beleut et al. 2012). For the identified 48 genes, we observed information gain of genes excluding genes no longer serviced by NCBI (http://www.ncbi.nlm.nih.gov/). For the remaining genes, the average information gain is 0.175. The maximum information gain is 0.759 of *PTPRO* and the minimum information gain is 0.0198 of *SMARCA4*. The complete list of significant genes and their corresponding information gain values are provided in Supplementary Tables 2 and 3.

Conclusions

In this paper, we investigated the potential of information gain for analysis of biomedical datasets generated from multiple platforms. The quantitative analysis based on the concept of information gain showed that the utilization of multiple-omics data is beneficial for in silico classification of tumor-normal instances. Furthermore, the experimental results reported that the classification power of each omics data and their combinations are very distinct. The qualitative analysis based on previous researches also verified the usefulness of the concept of information gain. We verified the relevance of genes with higher information gain to tumorigenesis through literature search.

However, this research also revealed the weakness of the information gain-based approach. Basically, the concept of information gain employed in this research utilizes the distribution of attributes across classes or categories of a target disease. As a result, our research was able to find genes whose expression pattern is biased to tumor or normal samples but it was unable to find “significantly” mutated genes as in (Beleut et al. 2012; The Cancer Genome Atlas Research Network 2014). This weakness was expressed during quantitative analysis. Although, it was confirmed that the majority of genes with higher information gain are tumorigenesis-relevant genes, the computed information gain of significantly mutated genes or genes with significant focal copy numbers reported relatively lower information gain. These findings suggest a number of useful guides for future researches. Firstly, the information gain-based approach would be useful for limiting candidates for detailed analysis. For example, DNA methylation is superior to other omics data for in silico classification of tumor-normal samples. Therefore, DNA methylation focused approach would produce tumor-intensive expressed genes. Secondly, a novel method is needed to reflect prior knowledge. Current approach is unable to recommend significantly mutated genes without incorporating prior knowledge. Finally, a novel approach is needed to utilize the unbalanced composition of biomedical data sets. For the same target cancer, the omics composition of each data instance from TCGA is very different. As a result, data instances lacking observations on the target omics data should be ignored in this study. In future works, we will focus on developing a novel method capable of utilizing data instances with different composition of data sources.

Acknowledgments This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT & Future Planning (2013R1A1A2008183). The results shown here are in whole based upon data generated by the TCGA Research Network: http://cancergenome.nih.gov/.

Conflict of interest The authors declare that they have no competing interests.

References

Alexandrov LB, Stratton MR (2014) Mutational signatures: the patterns of somatic mutations hidden in cancer genomes. Curr Opin Genet Dev 24:52–60

Beleut M, Zimmermann P, Baudis M, Bruni N, Buhlmann P, Laule O, Luu VD, Gruissem W, Schraml P, Moch H (2012) Integrative genome-wide expression profiling identifies three distinct molecular subgroups of renal cell carcinoma with different patient outcome. BMC Cancer 12:310

Bibikova M, Barnes B, Tsan C, Ho V, Klotzle B, Le JM, Delano D, Zhang L, Schroth GP, Gunderson KL et al (2011) High density DNA methylation array with single CpG site resolution. Genomics 98:288–295

Cancer Genome Atlas Research Network (2008) Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature 455:1061–1068

Cancer Genome Atlas Research Network (2011) Integrated genomic analyses of ovarian carcinoma. Nature 474:609–615

Crawford NPS, Walker RC, Lukes L, Officewala JS, Williams RW, Hunter KW (2008) The Diasporin pathway: a tumor progression-related transcriptional network that predicts breast cancer survival. Clin Exp Metastasis 25:357–369

Davis CF, Ricketts CJ, Wang M, Yang L, Cherniack AD, Shen H, Buhay C, Kang H, Kim SC, Fahey CC et al (2014) The somatic genomic landscape of chromophobe renal cell carcinoma. Cancer Cell 26:319–330

Dawson MA, Kouzarides T (2012) Cancer epigenetics: from mechanism to therapy. Cell 150:12–27

Eswaran J, Horvath A, Godbole S, Reddy SD, Mudvari P, Ohshiro K, Cyanam D, Nair S, Fuqua SA, Polyak K et al (2013) RNA sequencing of cancer reveals novel splicing alterations. Sci Rep 3:1689

Feuk L, Carson AR, Scherer SW (2006) Structural variation in the human genome. Nat Rev Genet 7:85–97

Guo Y, Sheng QH, Li J, Ye F, Samuels DC, Shyr Y (2013) Large scale comparison of gene expression levels by microarrays and RNAseq using TCGA data. PLoS One 8:e71462

Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH (2009) The WEKA data mining software: an update. SIGKDD Explor 11:10–18

Hastings PJ, Lupski JR, Rosenberg SM, Ira G (2009) Mechanisms of change in gene copy number. Nat Rev Genet 10:551–564

Henrichsen CN, Chaignat E, Reymond A (2009) Copy number variants, diseases and gene expression. Hum Mol Genet 18:R1–R8

Hinoue T, Weisenberger DJ, Lange CP, Shen H, Byun HM, Van Den Berg D, Malik S, Pan F, Noushmehr H, van Dijk CM et al (2012) Genome-scale analysis of aberrant DNA methylation in colorectal cancer. Genome Res 22:271–282

Hoadley KA, Yau C, Wolf DM, Cherniack AD, Tamborero D, Ng S, Leiserson MD, Niu B, McLellan MD, Uzunangelov V et al (2014) Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin. Cell 158:929–944

Huang X, Du X, Li Y (2012) The role of BCL11B in hematological malignancy. Exp Hematol Oncol 1:22

Jin C, Li H, Murata T, Sun K, Horikoshi M, Chiu R, Yokoyama KK (2002) JDP2, a repressor of AP-1, recruits a histone deacetylase 3 complex to inhibit the retinoic acid-induced differentiation of F9 cells. Mol Cell Biol 22:4815–4826

Johnson CS, Cook CA, Furmanski P (1990) In vivo suppression of erythropoiesis by tumor necrosis factor-alpha (TNF-alpha): reversal with exogenous erythropoietin (EPO). Exp Hematol 18:109–113

Jones PA, Baylin SB (2002) The fundamental role of epigenetic events in cancer. Nat Rev Genet 3:415–428

Kennedy SR, Loeb LA, Herr AJ (2012) Somatic mutations in aging, cancer and neurodegeneration. Mech Ageing Dev 133:118–126

Kim JW, Kim ST, Turner AR, Young T, Smith S, Liu WN, Lindberg J, Egevad L, Gronberg H, Isaacs WB et al (2012) Identification of new differentially methylated genes that have potential functional consequences in prostate cancer. PLoS One 7:e48455

Kito Y, Saigo C, Atsushi K, Mutsuo F, Tamotsu T (2014) Transgenic mouse model of cutaneous adnexal tumors. Dis Model Mech 7:1379–1383

Krishnan VG, Ng PC (2012) Predicting cancer drivers: are we there yet? Genome Med 4:88

Kwong J, Lo KW, Chow LSN, To KF, Choy KW, Chan FL, Mok SC, Huang DP (2005) Epigenetic silencing of cellular retinol-binding proteins in nasopharyngeal carcinoma. Neoplasia 7:67–74

Laddha SV, Ganesan S, Chan CS, White E (2014) Mutational landscape of the essential autophagy gene BECN1 in human cancers. Mol Cancer Res 12:485–490

Laird PW (2010) Principles and challenges of genome-wide DNA methylation analysis. Nat Rev Genet 11:191–203

Lee J, van Hummelen P, Go C, Palescandolo E, Jang J, Park HY, Kang SY, Park JO, Kang WK, MacConaill L et al (2012) High-throughput mutation profiling identifies frequent somatic mutations in advanced gastric adenocarcinoma. PLoS One 7:e38892

Leng N, Dawson JA, Thomson JA, Ruotti V, Rissman AI, Smits BM, Haag JD, Gould MN, Stewart RM, Kendziorski C (2013) EBSeq: an empirical Bayes hierarchical model for inference in RNA-seq experiments. Bioinformatics 29:2073

Martinez O, Reyes-Valdes MH (2008) Defining diversity, specialization, and gene specificity in transcriptomes through information theory. Proc Natl Acad Sci USA 105:9709–9714

Mitchell TM (1997) Machine learning. McGraw Hill, New York

Nakade K, Pan JZ, Yamasaki T, Murata T, Wasylyk B, Yokoyama KK (2009) JDP2 (Jun dimerization protein 2)-deficient mouse embryonic fibroblasts are resistant to replicative senescence. J Biol Chem 284:10808–10817

Ni XH, Zhuo ML, Su Z, Duan JC, Gao Y, Wang ZJ, Zong CH, Bai H, Chapman AR, Zhao J et al (2013) Reproducible copy number variation patterns among single circulating tumor cells of lung cancer patients. Proc Natl Acad Sci USA 110:21083–21088

Ozsolak F, Milos PM (2011) RNA sequencing: advances, challenges and opportunities. Nat Rev Genet 12:87–98

Park YK, Kang TW, Baek SJ, Kim KI, Kim SY, Lee D, Kim YS (2012) CaGe: a web-based cancer gene annotation system for cancer genomics. Genomics Inform 10:33–39

Poduri A, Evrony GD, Cai X, Walsh CA (2013) Somatic mutation, genomic variation, and neurological disease. Science 341:1237758

Pylkas K, Vuorela M, Otsukka M, Kallioniemi A, Jukkola-Vuorinen A, Winqvist R (2012) Rare copy number variants observed in hereditary breast cancer cases disrupt genes in estrogen signaling and TP53 tumor suppression network. PLoS Genet 8:e1002734

Quinlan JR (1993) C4.5: programs for machine learning. Morgan Kaufmann Publishers Inc., San Francisco

Rastogi N, Mishra DP (2012) Therapeutic targeting of cancer cell cycle using proteasome inhibitors. Cell Div 7:26

Rebhan M, Chalifa-Caspi V, Prilusky J, Lancet D (1997) GeneCards: integrating information about genes, proteins and diseases. Trends Genet 13:163

Reinert T, Modin C, Castano FM, Lamy P, Wojdacz TK, Hansen LL, Wiuf C, Borre M, Dyrskjot L, Orntoft TF (2011) Comprehensive genome methylation analysis in bladder cancer: identification and validation of novel methylated genes and application of these as urinary tumor markers. Clin Cancer Res 17:5582–5592

Reinert T, Borre M, Christiansen A, Hermann GG, Orntoft TF, Dyrskjot L (2012) Diagnosis of bladder cancer recurrence based on urinary levels of EOMES, HOXA9, POU4F2, TWIST1, VIM, and ZNF154 hypermethylation. PLoS one 7:e46297

Ren S, Peng Z, Mao JH, Yu Y, Yin C, Gao X, Cui Z, Zhang J, Yi K, Xu W et al (2012) RNA-seq analysis of prostate cancer in the Chinese population identifies recurrent gene fusions, cancer-associated long noncoding RNAs and aberrant alternative splicings. Cell Res 22:806–821

Schaefer CF, Anthony K, Krupa S, Buchoff J, Day M, Hannay T, Buetow KH (2009) PID: the Pathway Interaction Database. Nucleic Acids Res 37:D674–D679

She X, Cheng Z, Zollner S, Church DM, Eichler EE (2008) Mouse segmental duplication and copy number variation. Nat Genet 40:909–914

Shlien A, Malkin D (2009) Copy number variations and cancer. Genome Med 1:62

Sproul D, Nestor C, Culley J, Dickson JH, Dixon JM, Harrison DJ, Meehan RR, Sims AH, Ramsahoye BH (2011) Transcriptionally repressed genes become aberrantly methylated and distinguish tumors of different lineages in breast cancer. Proc Natl Acad Sci USA 108:4364–4369

Stott-Miller M, Zhao S, Wright JL, Kolb S, Bibikova M, Klotzle B, Ostrander EA, Fan JB, Feng Z, Stanford JL (2014) Validation study of genes with hypermethylated promoter regions associated with prostate cancer recurrence. Cancer Epidemiol Biomark Prev 23:1331–1339

Stratton MR (2011) Exploring the genomes of cancer cells: progress and promise. Science 331:1553–1558

The Cancer Genome Atlas Research Network (2014) Comprehensive molecular characterization of urothelial bladder carcinoma. Nature 507:315–322

Ting DT, Wittner BS, Ligorio M, Vincent Jordan N, Shah AM, Miyamoto DT, Aceto N, Bersani F, Brannigan BW, Xega K et al (2014) Single-cell RNA sequencing identifies extracellular matrix gene expression by pancreatic circulating tumor cells. Cell Rep 8:1905–1918

Tresse E, Salomons FA, Vesa J, Bott LC, Kimonis V, Yao TP, Dantuma NP, Taylor JP (2010) VCP/p97 is essential for maturation of ubiquitin-containing autophagosomes and this function is impaired by mutations that cause IBMPFD. Autophagy 6:217–227

Wang Z, Gerstein M, Snyder M (2009) RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet 10:57–63

Watson IR, Takahashi K, Futreal PA, Chin L (2013) Emerging patterns of somatic mutations in cancer. Nat Rev Genet 14:703–718

Yasunaga M, Matsumura Y (2014) Role of SLC6A6 in promoting the survival and multidrug resistance of colorectal cancer. Sci Rep 4:4852

Zarubin T, Jing Q, New L, Han JH (2005) Identification of eight genes that are potentially involved in tamoxifen sensitivity in breast cancer cells. Cell Res 15:439–446

Zhang F, Gu W, Hurles ME, Lupski JR (2009) Copy number variation in human health, disease, and evolution. Annu Rev Genomics Hum Genet 10:451–481

Zuern C, Heimrich J, Kaufmann R, Richter KK, Settmacher U, Wanner C, Galle J, Seibold S (2010) Down-regulation of MTUS1 in human colon tumors. Oncol Rep 23:183–189