Zhang *et al. BMC Bioinformatics* (2016) 17:222
DOI 10.1186/s12859-016-1085-7

BMC Bioinformatics

METHODOLOGY ARTICLE Open Access

Identification of recurrent focal copy number variations and their putative targeted driver genes in ovarian cancer

Liangcai Zhang[1,4,5], Ying Yuan[2,4], Karen H. Lu[3] and Li Zhang[1*]

Abstract

Background: Genomic regions with recurrent DNA copy number variations (CNVs) are generally believed to encode oncogenes and tumor suppressor genes (TSGs) that drive cancer growth. However, it remains a challenge to delineate the key cancer driver genes from the regions encoding a large number of genes.

Results: In this study, we developed a new approach to CNV analysis based on spectral decomposition of CNV profiles into focal CNVs and broad CNVs. We performed an analysis of CNV data of 587 serous ovarian cancer samples on multiple platforms. We identified a number of novel focal regions, such as focal gain of ESR1, focal loss of LSAMP, prognostic site at 3q26.2 and losses of sub-telomere regions in multiple chromosomes. Furthermore, we performed network modularity analysis to examine the relationships among genes encoded in the focal CNV regions. Our results also showed that the recurrent focal gains were significantly associated with the known oncogenes and recurrent losses associated with TSGs and the CNVs had a greater effect on the mRNA expression of the driver genes than that of the non-driver genes.

Conclusions: Our results demonstrate that spectral decomposition of CNV profiles offers a new way of understanding the role of CNVs in cancer.

Keywords: Copy number variations, Oncogenes and tumor suppressor genes, Focal and broad CNV, Cancer

Background

DNA aberrations in cancer can take many different forms, ranging from mutations, translocations, inversions and copy number variations (CNV) [1, 2]. The scope of the aberrations can range from single nucleotides to whole chromosomes. Recent analyses of CNVs in various types of cancers showed that their scope primarily exhibits two modes: it is either focal, which is limited to a small fraction of a chromosome, or very broad, which extends to a large fraction of a chromosome arm [3, 4]. Mechanistically, it was found that the focal CNVs (*f*CNVs) occur due to errors in DNA repair and the broad CNVs (*b*CNV) occur due to incorrect segregation of chromosomes during mitosis [1, 5, 6]. Regions of frequent *f*CNV are particularly important in cancer studies because they are believed to encode key genes driving cancer growth [7]. Many of the known oncogenes, such as ERBB2, EGFR and CCND1, are frequently amplified [8–12] and many of the known tumor suppressor genes (TSGs), such as CDKN2A, PTEN, NF1 and RB1, are frequently depleted in various types of cancers [13–15]. Thus, it is generally believed that recurrent focal gains are associated with oncogenes_ENREF_3 and focal losses associated with TSGs [16].

However, it remains to be a challenge to delineate the targeted oncogenes and TSGs from the recurrent CNVs [17–19]. For example, from an analysis of over 3000 cancer genomes, Beroukhim et al. identified 150 focal regions [4] that were supposed to be the hotspots of cancer driver genes, but only less than 25 % of the regions contained known oncogenes or TSGs. To explain the phenomena, it was proposed that some CNVs may rise from inherently fragile sites and gene poor regions. Solimini et al. proposed a 'gene island' theory [16]: genes that stimulate/inhibit tumor growth may distribute very unevenly across the genome. Such genes are not classical oncogenes or

* Correspondence: lzhangli@mdanderson.org
[1] Department of Bioinformatics and Computational Biology, The University of Texas MD Anderson Cancer Center, 1400 Pressler St, Unit 1410, Houston, TX 77401, USA
Full list of author information is available at the end of the article

TSGs as they only have minor effects on tumor growth individually. However, the minor effects collectively can make a big difference through evolution of the cancer cells and produce the patterns of frequent gains/losses as we have observed. Sharon J. Diskin et al. [20] developed a statistical model called Significance Testing for Aberrant Copy number (STAC) to evaluate the randomness of the distribution of CNVs in tumors. STAC uses *p*-values to prioritize regions for down-stream analysis.

To better understand how CNVs are related to cancer driver genes, we analyzed CNVs in serous ovarian cancer, which is a type of cancer that contains relatively more CNVs than others. The dataset contained samples collected from 587 patients assayed on three different microarray platforms (Additional file 1: Table S1). Both primary tumors and adjacent normal tissues were assayed. An analysis of the dataset was reported previously by the TCGA consortium [21]. A statistical algorithm called GISTIC was used in several studies [3, 22, 23] to identify the regions that were highly enriched with copy number gains/losses. Several well-known cancer driver genes, such as RB1, PTEN and NF1, were suggested to be the targets of the regions. However, the targeted cancer genes of 70 (62 %) of the 113 regions are not known. The results reported by the TCGA consortium were largely consistent with earlier analyses of ovarian cancer, which used smaller sample sizes and obtained CNVs at lower resolution [24–29].

We developed a new method aimed at identifying the focal CNVs that drive cancer progression (See in Fig. 1a). The method is designed to decompose a CNV profile into a focal profile and a broad profile based on the hypothesis that a genomic region may undergo multiple modes of aberrations during cancer progression. We viewed a CNV profile as a spectrum, in which the high frequency component corresponded to *f*CNVs and the low frequency component corresponded to *b*CNVs. Figure 1b illustrates the method schematically. To obtain the *b*CNV profile, we used the running median smoothing algorithm, which was originally proposed by Tukey [30]_ENREF_2 to smoothen time series data that follow a piecewise-constant model. The algorithm works by taking the median value in a scanning window over the CNV profile (See details in method section). The focal profile is computed from the difference between the CNV profile and the *b*CNV profile. This method is appealing because it preserves the shape of abrupt change points and it is robust for data that follow piecewise

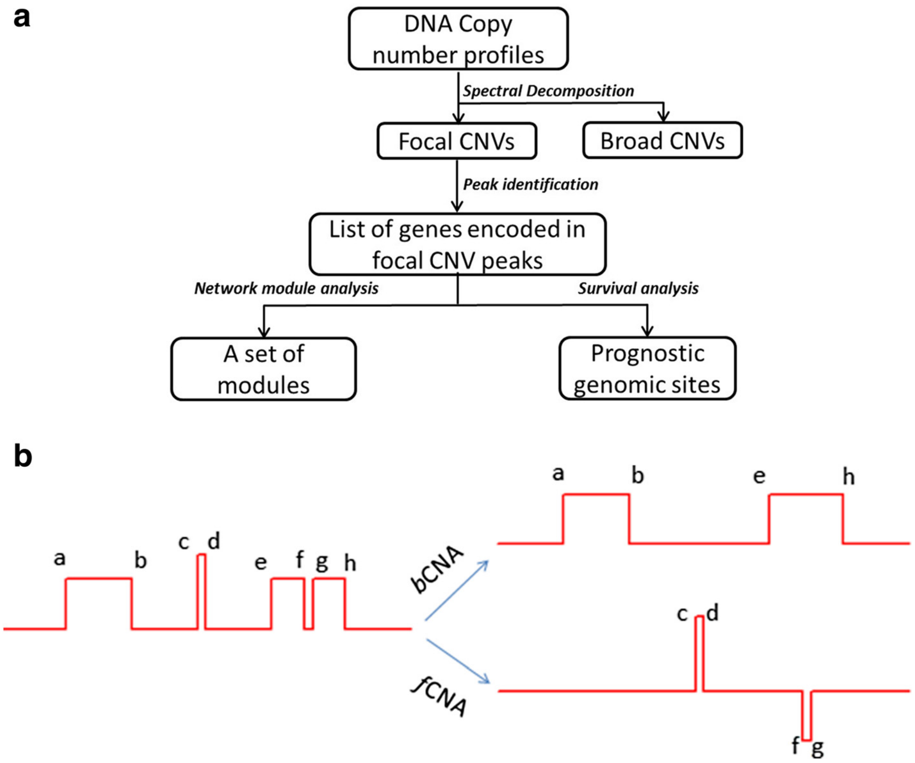

Fig. 1 Analysis framework. **a** The flowchart of our analysis. **b** Schematic illustration of spectral decomposition of CNV profile. The red line on the left shows a CNV profile with vertical position representing the copy number and horizontal position representing chromosomal location. The letters in the figure mark the change points. The red lines on the left show the decomposed profiles, the *b*CNV on the top, and *f*CNV on the bottom. The segments $\bar{ab}$ and $\bar{eh}$ are considered broad gains. $\bar{cd}$ is a focal gain and $\bar{fg}$ is focal loss

constant model. As shown in Fig. 1, the magnitude of broad CNV segment $\overline{eh}$ is not affected by focal CNV segment $\overline{fg}$ as long as $\overline{fg}$ span less than the half size of the scanning window. Using Fourier or wavelet transformations, it is also possible to separate high frequency from low frequency components from a CNV profile [31–35], but those transformations are not robust, nor do they preserve the shape of abrupt change points.

Methods

To identify focal copy number variations and their putative cancer genes, our method consists of the following parts, which elaborate detailed steps of data collection from the TCGA data portal, CNV probe-level data denoising and decomposition, identification of focal gains and losses, identification of peak regions and further downstream functional analysis (Network module and survival analysis).

The computational time for the CNV decomposition process is about 10–12 h on the high performance cluster (six nodes with 24 cores per node, Linux operating system); while for the rest procedures the time is less than 1 h. The proposed algorithm can be easily achieved using the detailed steps in method section, and the source R code is available at https://drive.google.com/drive/folders/0B6Q6G-z3ELntWllEd29IOVpyYzA.

Data source

Copy number and mRNA expression data were downloaded from TCGA Data Portal (https://tcga-data.nci.nih.gov/tcga/dataAccessMatrix.htm) before September, 2012. Sample information of the 587 patients, progression free survival data, were summerized in Additional file 1: Table S1. Version hg18, Human Build 36.1 were used for annotating the genomic coordinates.

CNV data processing

Let x be the vector of logarithm-transformed ($base = 2$) probe signals ordered according to the chromosomal positions in a sample. x is called a copy number profile. First, the profile is normalized using

$$x_{normalized} = x - mode(x) \tag{1}$$

where $mode(x)$ is mode value of x, *i.e.*, the peak position of density distribution funciton of x.

Second, the profile is smoothed to reduce noise,

$$x_{smoothed} = runmed(x_{normalized}, w) \tag{2}$$

where *runmed* is the running median smoothing function with a scanning window of w, which was chosen to be 51, 51, 99 on the Agilent array, Illumina array, and Affymetrix array, respectively. The total number of probes are 9.6×10^5, 1.19×10^6, 1.87×10^6, for the Agilent, Illumina, and Affymetrix arrays, respectively. Consequently, the chosen window sizes corresponded to approximately the same size on the chromosomes, circa 120 kb. The smoothing algorithm treated CNV events that spaned less than half of the window size as noise and ignored them. Thus, gain/loss fragments that were smaller than 59 kb were omitted in the subseqeunt analysis.

Third, because there is little change in the smoothed profile between neighboring probed sites, we chose to represent the smoothed profile at a lower resolution to reduce data dimension, *i.e.*,

$$x_{reduced} = x_{smoothed}[1,\ 1 + w/3, 1 + 2w/3, \ldots,\ n] \tag{3}$$

where n is length of $x_{smoothed}$.

To obtain the broad CNV profile and focal CNV profile, we used

$$x_{bcnv} = runmed(x_{reduced}, wb) \tag{4}$$

$$x_{fcnv} = x_{reduced} - x_{bcnv} \tag{5}$$

where wb was chosen to be 641, 641, 793 on the Agilent, Illumina, and Affymetrix arrays, respectively. The window sizes correspond to approxmiately 32 Mb on the chromosomes, which means that CNV segments longer than 16 Mb are treated as broad changes and CNV segments shorter than 16 Mb are treated as focal changes. The results obtained in this study were not very sensitive to the choice of wb. Additional file 2: Figure S1 showed that the results were similar for wb between 30 and 40 Mb.

Identification of focal gains or losses

At any genomic locus, the copy number has three states: gain, neutral, loss, which was determined as follows:

$$x_{fcnv} \begin{cases} > \ 3\varepsilon : & gain; \\ > -3\varepsilon \ and \ < 3\varepsilon : & neutral; \\ < \ -3\varepsilon : & loss \end{cases} \tag{6}$$

where ε is the estimated error of x_{fcnv}.

Identification of peak positions and their confidence intervals

To identify peak positions of focal CNV distribution, we used a scanning window of 41 x_{fcnv} sites to search for local maximums. Peaks with maximum value less than 8 were ignored.

To estimate the confidence intervals of the peak positions, a bootstrapping method [36, 37] was used. Boostrap samples were constructed using random sampling with replacement from the 587 focal CNV profiles. 500 sets of bootstrap samples were created, each set containing 587 profiles. For each set of the profiles, peak positions were identified. Because the number of such peak positions from bootstrap samples may be different from number of the original peaks, it is not possible to pair-up the two kinds of peaks one-to-one. To identify a new peak position for each orginal peak, we used the nearest peak position in the bootstrapping set with regard to each original peak to represent bootstrapped peak position. From the 500 sets of bootstrap samples, 500 sets of peak positions were obtained. The top 2.5 percentile and bottom 2.5 percentile of the bootstrapped peak positions were used as the estimates of 95 % confidence interval of each original peak.

Identification of copy number polymorphism (CNP) from normal tissues

To identify CNPs that rise from germline mutation, we used focal CNV profiles of adjacent normal tissue samples. Because some of the samples appeared to be outliers as they contained a large number of CNPs, we suspected that they have been contaminated by tumor cells or the data had poor quality and decided to exclude them. We computed the number of CNPs for each sample as number of sites with $|f\text{CNV}| > 5\boldsymbol{\varepsilon}$, where $\boldsymbol{\varepsilon}$ is the estimated error. Five samples on the Agilent platform with largest number of CNPs were determined as outliers and were excluded. Similarly, 61 and 48 normal samples were excluded from Affymetrix and Illumina arrays, respectively. After the removal of outliers, 240, 512 and 466 samples were included for CNP identification on the Agilent, Illumina, and Affymetrix arrays, respectively.

Putative cancer driver genes based on *f*CNV and literature search

We created a list of cancer related genes from combining the cancer genes lists in OMIM database (http://omim.org/), Cancer Gene database of MSKCC [38], Sanger cancer gene census [39], TAG (http://www.binfo.ncku.edu.tw/TAG/) and TSGene databases (http://bioinfo.mc.vanderbilt.edu/TSGene/). The list contains 11,595 gene symbols.

An overlapping gene set of 200 genes was obtained between focal region genes and this cancer gene list.

We found 1245 *mRNA* transcripts were located in the 42 focal regions that were identified from our analyses. 200 out of the 1245 were contained in the list of cancer related genes.

For each of the 200 genes, we examined the references listed on GeneCards (http://www.genecards.com). A gene was defined as an oncogene if it has been reported that over-expression or activation mutation of this gene can cause cancer, promote cancer growth or metastasis. A gene was defined as a TSG if knockdown or deleterious mutation of this gene can cause cancer, promote cancer growth or metastasis.

Survival analysis

We used Cox proportional hazard model to search for genomic loci where *f*CNVs or *b*CNV are correlated with

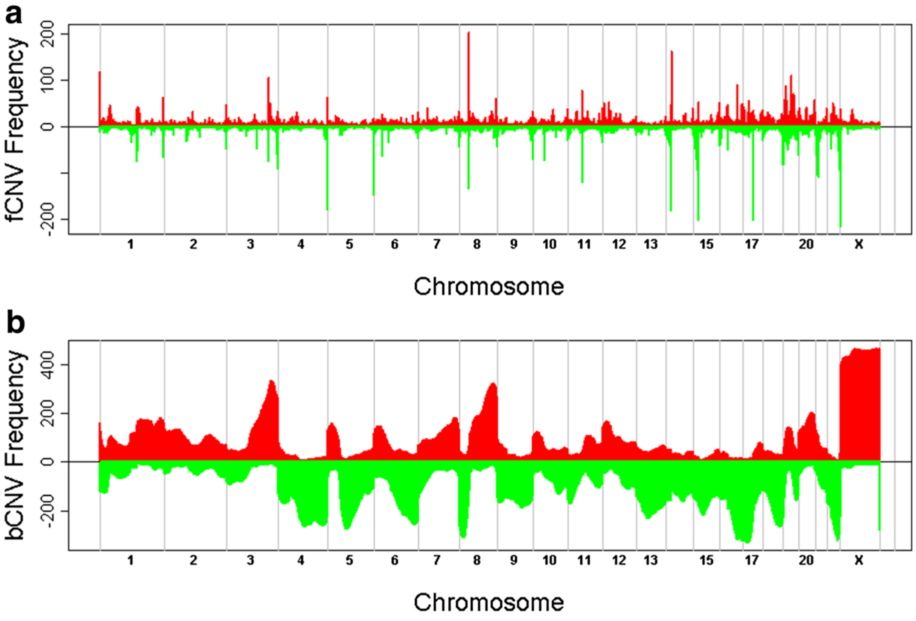

Fig. 2 Distribution of focal and broad CNVs in 587 ovarian cancer samples. **a** *f*CNV; (**b**) *b*CNV. The *x*-axis shows the chromosomes. Data regarding the Y-chromosome were removed. The *y*-axis shows the numbers of samples with gains as positive numbers in red and numbers of samples with losses as negative numbers in green. Numerical data of the distributions can be found in Additional file 1: Table S2

progression free survival. At each the j^{th} genomic site, we constructed two models:

$$M_{f,bcnv}[j] = coxph\left(Surv(time,\ status)^{\sim} x_{fcnv}[j] + x_{bcnv}[j]\right) \tag{7}$$

$$M_{cnv}[j] \quad = coxph(Surv(time,\ status)^{\sim} x_{cnv}[j]) \tag{8}$$

where x_{fcnv} is *f*CNV, and x_{bcnv} is the *b*CNV, and $x_{cnv} = x_{fcnv} + x_{bcnv}$. *coxph* is an *R* function that constructs an object of the cox proportional model [40]. *Surv*() creates a survival data object with censored *time* and *status*. Model A used the decomposed CNV profiles. Model B used CNV profiles without the decomposition.

Cox model for screening of prognostic sites based on mRNA Expression data

For each gene *i*, a single-variate Cox model is set to be:

$$M_i \ = \ coxph\left(Surv(time,\ status)^{\sim} g_i\right)$$

where g_i is the logarithm transformed mRNA expression levels of the gene and *coxph* is an *R* function to construct an object of the cox proportional survival model. *Surv*() creates a survival data object with censored *time* and *status*.

Network/module analysis on genes in *f*CNV regions

We used the NetBox server (http://cbio.mskcc.org/netbox), which is pre-loaded with a Human Interaction Network (HIN) derived from curated literature. HIN contains 9264 nodes and 68111 edges. The genes in the focal regions identified from our analyses were mapped to HIN to identify the core modules. The parameters used in the NetBox analyses were as follows: shortest path threshold = 2; *p*-value threshold = 0.05; number of global trials = 1000; number of local trials = 100.

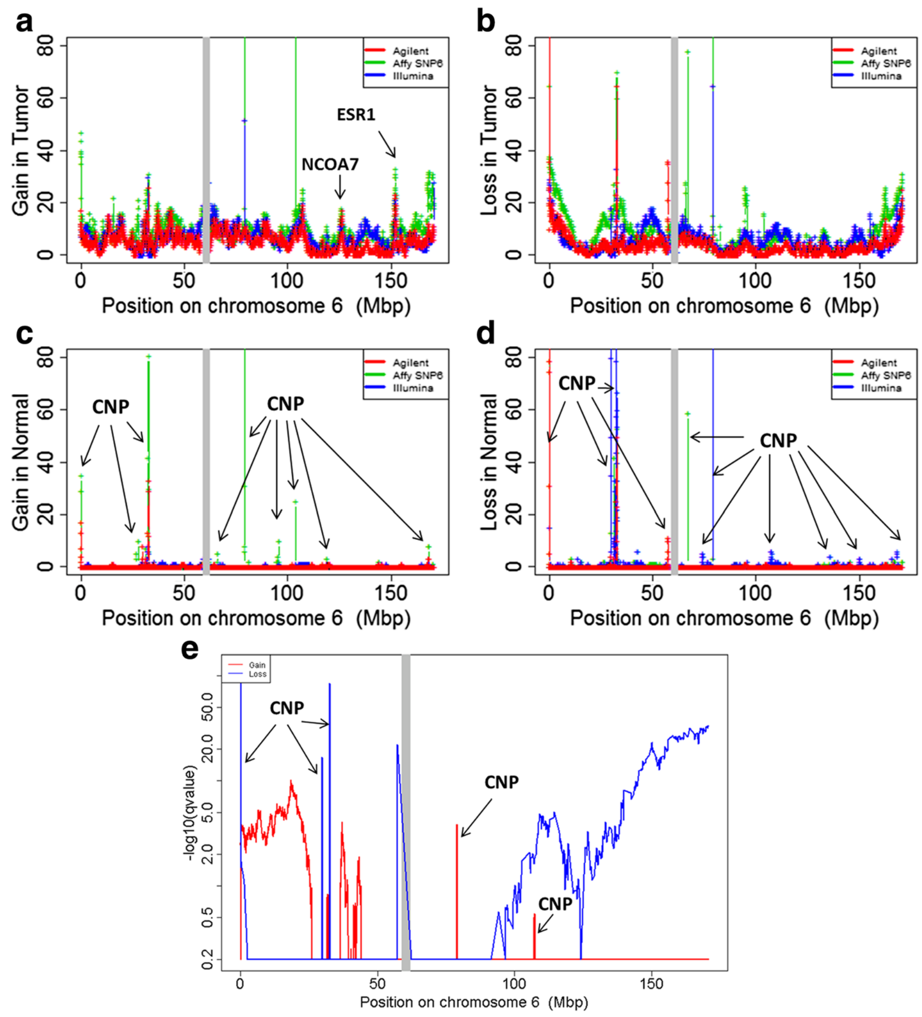

Fig. 3 Focal gains and losses on chromosome 6. **a** Focal gains in tumor samples; (**b**) focal losses in tumor samples; (**c**) focal gains in normal samples, CNP means the copy number polymorphism in normal population; (**d**) focal losses in normal samples; (**e**) GISTIC results

Assessment of sensitivity and specificity based on computer simulated data

To generate simulated CNV profile, we conducted the simulation as follows: we set the dimension of a CNV profile having $L = 1.6 \times 10^4$ probes for $N = 500$ patients. The measurement error of the probe signals is set to follow normal distribution with a standard deviation $\varepsilon = 1$. Each patient has a normal CNV profile and a tumor CNV profile. A fraction of the patients (f^*N) have an amplicon in their tumor CNV profile. The height (h) and the width (represented as number of probes n) of the amplicon are set to be variable. We then apply the decomposition algorithm to identify status of gain/loss at the center of the CNV profiles. False positive rates, true positive rates, true negative rates, and false negative rates were calculated according to nominal truth. The significance p-value is calculated based on the test of whether the number of focal gains is significantly greater than the number of focal losses in the tumors at a particular genomic locus.

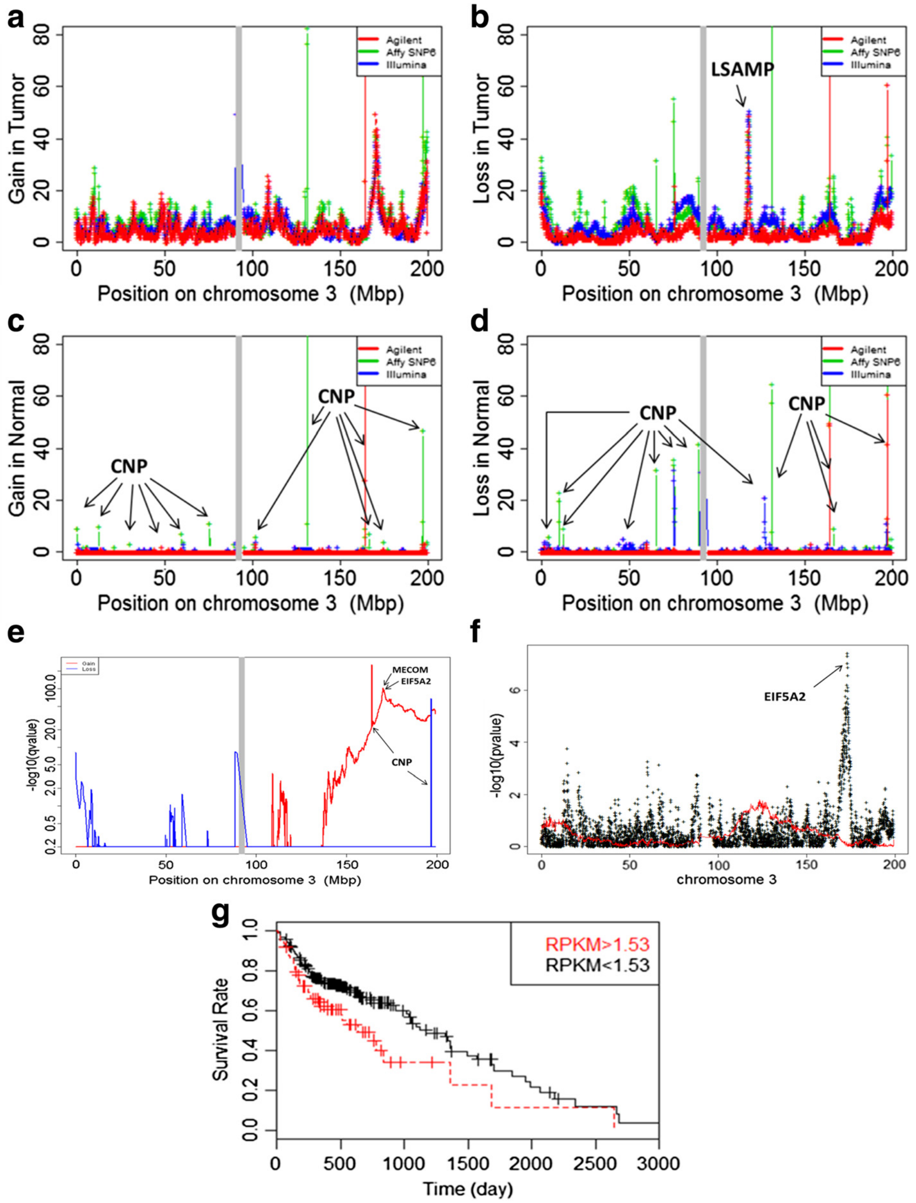

Fig. 4 Focal gains and losses on chromosome 3. **a** focal gains in tumor samples; (**b**) focal losses in tumor samples; (**c**) focal gains in normal samples, CNP means the copy number polymorphism in normal population; (**d**) focal losses in normal samples; (**e**) GISTIC results; (**f**) PFS ~ *f*CNV + *b*CNV, the black points represent the prognostic power of *f*CNVs, while the red represent that of *b*CNVs; (**g**) Figure of KM-plot for EIF5A2. We used two Cox model analyses. Model 1: model: PFS ~ focal. Wald test *p*-value = 4.23E-08. Model 2: the samples were partitioned into two groups the upper quartile (RPKM > 1.53) and the rest (RPKM < 1.53). *P*-value = 0.006808, model: PFS ~ Group

Results and discussion

Using the decomposition method, we performed the analysis of focal copy number identification for the ovarian cancer data from the TCGA data portal. We found focal regions contain putative cancer driver genes, and have significant co-occurrences with tumor suppressors and oncogenes. Also, we investigated the relationship between gene expression alterations and their copy number variations in focal copy number regions. Finally, we checked if focal copy number variations play a role in patients' survival. Detailed results are elaborated in the following (or in Additional file 2: Supplementary information).

Identification of focal regions containing putative cancer drivers

Figure 1a shows a flowchart of our analysis process. First, we obtained the distributions of focal changes and broad changes across the human genome (Fig. 2 and Additional file 1: Table S2). We evaluated the distributions for tumor and normal tissue samples separately on each of the three microarray platforms (Additional file 2: Figure S2). Then we identified the peak positions of the *f*CNV distribution and calculated the confidence interval (CI) of the peak positions using bootstrap samples (See in methods). We searched for focal regions that met the following criteria: (1) Peak height ≥ 8; (2) The 95 % CI of the peak position is less than 1 million base pairs. (3) Less than 4 gains or losses were found in the normal tissue samples within the 95 % CI. This criterion filtered out regions that are polymorphic in the healthy population; (4) The number of focal gains must be significantly different from the number of losses in the same region.

We identified 42 focal regions that met our criteria, which are composed of 26 focal gain regions and 16 focal loss regions (Additional file 1: Table S3). These regions collectively encode 1245 transcripts (809 gene symbols). Based on literature search, we found 47 oncogenes in the focal gain regions and 15 TSGs in the focal loss regions, which we regarded as putative driver genes (Additional file 1: Table S3).

We compared results with previous analysis of ovarian cancer [21] using GISTIC, which found 50 focal regions with peak-width less than 1 Mb. 25 of the 50 had overlaps with the focal regions in our analysis. Additionally, we compared our results with a report of GISTIC analysis of multiple cancer types [41], which found 75 focal regions with peak-width less than 1 Mb, and 29 of them overlapped with our focal regions.

We noticed a number of features that had not been reported in the previously. Our analyses unveiled recurrent focal gains in the midst of broad losses. For example, 6q loss is common in many cancers (Fig. 2b). Surprisingly, we found a focal gain region around 6q25.1. The focal gains in tumors were observed on all the three microarray platforms (Fig. 3a) but not in normal tissues (Fig. 3c, d). The region encodes an oncogene ESR1 known in many cancers including ovarian cancer [42–45]. The focal gain in the tumor samples was not identified using GISTIC 2.0 (Additional file 2: Figure S3), presumably because the gain of ESR1 can only be identified after CNV decomposition.

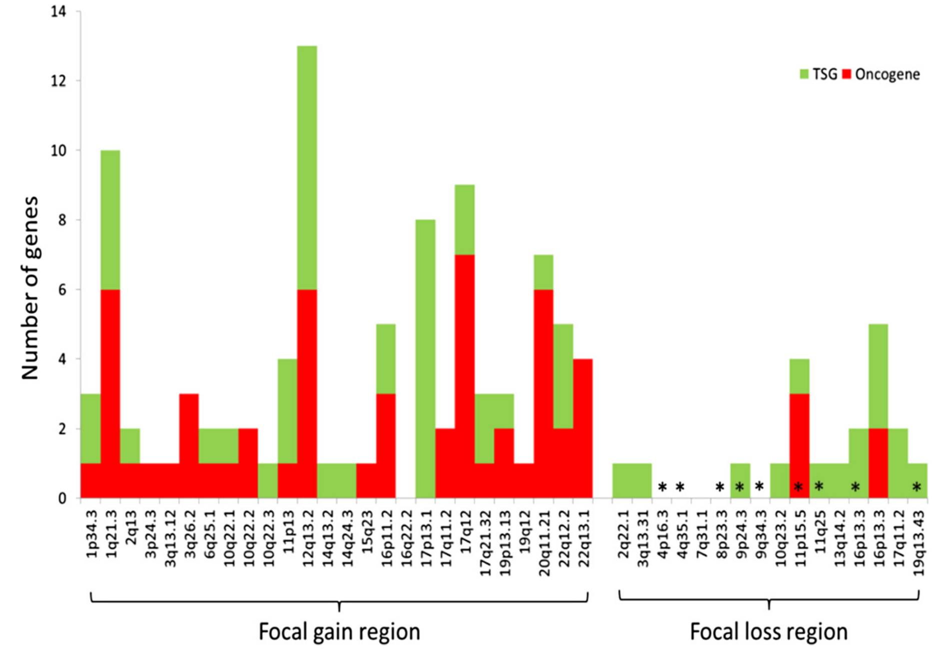

Fig. 5 Known cancer genes (Oncogenes and TSGs) in regions of recurrent focal CNVs

Similarly, LSAMP located at 3q26 is an example of focal losses in the midst of broad gains (Figs. 2b and 4). The focal losses in tumors were observed on all the three microarray platforms (Fig. 4a) but not in normal tissues (Fig. 4c, d). LSAMP was identified as a candidate tumor suppressor and focal deletion LSAMP were reported in other cancers [46–50]. Expression of LSAMP was also shown to be associated with osteosarcoma progression [50]. GISTIC analysis of LSAMP region identified no significant deletions (Fig. 4e).

Relationship between gains/losses and oncogenes/TSGs

Although it is widely believed that oncogenes are associated with the gains and TSGs associated with losses, the associations have not been evaluated quantitatively in previous studies. We have tested this association using our results. Figure 5 showed the numbers of focal regions that contained known cancer driver genes in the 42 focal regions identified in our analysis. Most of the regions contained oncogenes as well as TSGs. However, the oncogene-to-TSG ratio in the focal gain regions was different from that in the focal loss region: The ratio is 22:14 in the focal gain regions and 2:11 in the focal loss regions. The ratios differ by 8.6-fold, which is statistically significant (p-value = 8.30×10^{-3}, Fisher's exact test.)

Third, there were 9 regions of focal losses located at the ends of chromosomes (4p, 4q, 8p, 9p, 9q, 11p, 11q, 16p and 19p, in Fig. 5), which may be related to telomere loss in serous ovarian cancer [51–53]. In four of the focal regions, no known oncogenes or TSGs were found.

Network modularity analysis of genes encoded in the focal regions

To explore the underlying relationship among the 1245 genes encoded in the 42 focal regions, we performed a

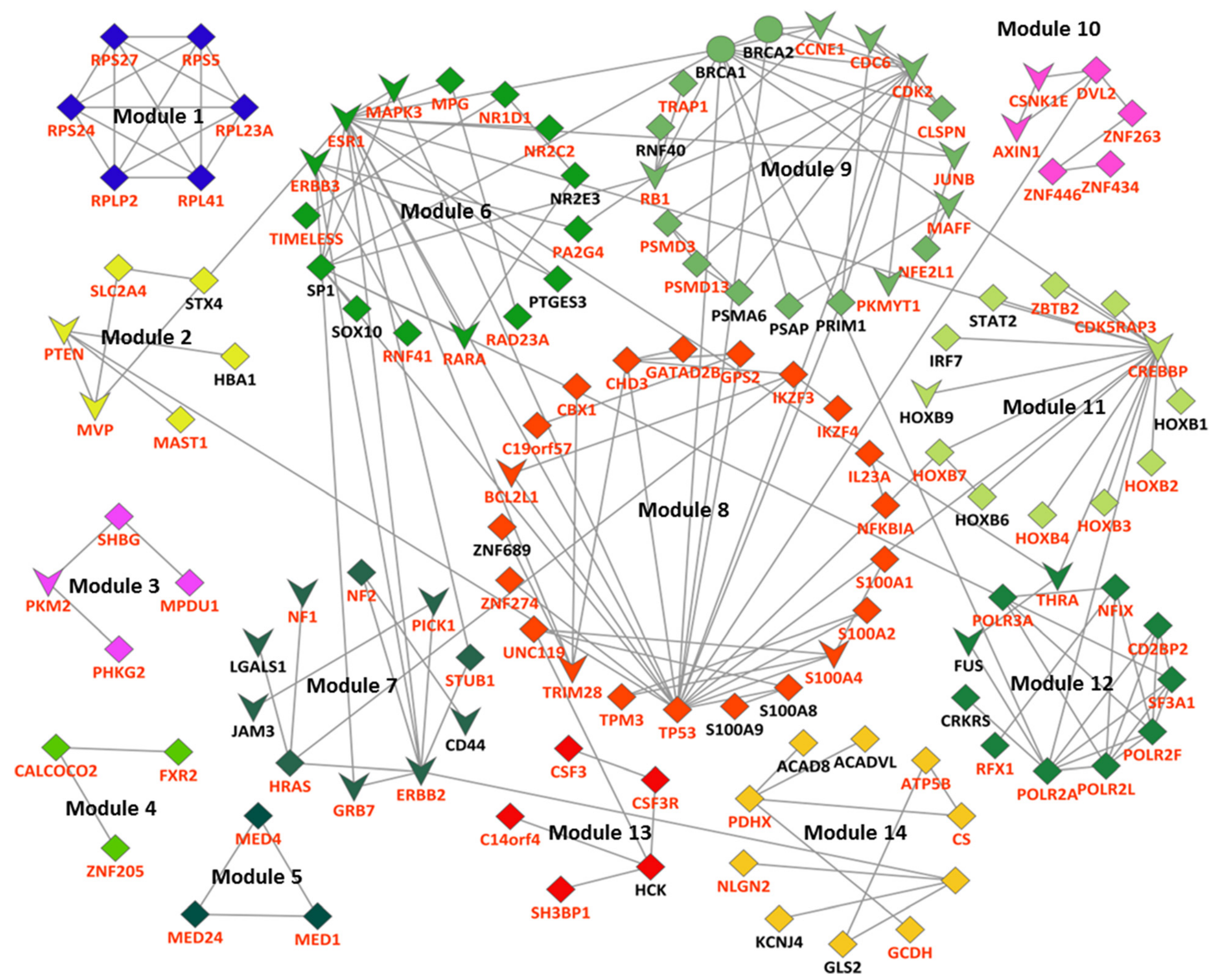

Fig. 6 Network modules of cancer drivers in ovarian cancer. There are 130 nodes and 188 edges in this network. Nodes with the label in red have their *f*CNV changes responsive to gene expression. The 'diamond/V' node represents the gene is located within 42 focal regions. The 'V' node represents the current gene is a known-cancer driver based on our curated literature search

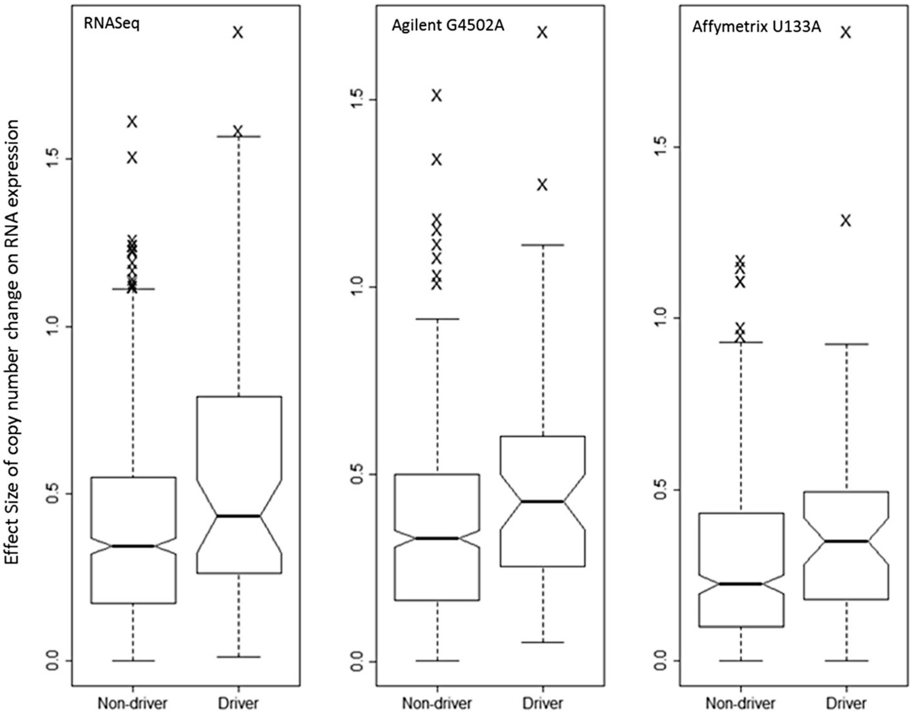

Fig. 7 Effects of copy number change on RNA expression for cancer drivers and non-cancer drivers. The vertical axis shows effect size of copy number change on RNA expression as boxplots. In the focal gain regions, each data point used in the boxplots were computed from the average log-fold change of mRNA expression of a gene between tumor samples with focal gain and tumor samples without focal gain. In the focal loss regions, each data point used in the boxplots were computed from the average log-fold change of mRNA expression of a gene between tumor samples without focal loss and tumor samples with focal loss. The boxplots were generated from data on three different platforms: RNAseq, Agilent 244 K and Affymetrix U133A, corresponding to the panels on the left, middle and right, respectively. The differences in the effect sizes are significant (*p*-values =0.012, 0.0059, 0.020, for the left, middle and right panels, respectively)

network modularity analysis using NetBox [54]. 809 of 1245 transcripts can be mapped into the background network stored in NetBox. The NetBox analysis found a sub-network of 14 modules consists of 130 genes and 188 edges (Fig. 6). This sub-network has an overlap of 31 with the 62 putative drivers. This level of overlap is statistically significant (p-value = 6.45×10^{-5}, Fisher's exact test), which suggests that the driver genes are more "connected" than other genes.

Association between focal CNVs and genes expression

To further explore the functional role of focal CNVs, we assessed association between focal CNVs and gene expression. We found that the focal changes have a greater effect on putative drivers (Fig. 7). Using gene expression data obtained on the RNAseq platform, the mean fold change in the gene expression in response to focal gain/loss is 0.54 for cancer driver genes and 0.39 for other genes, respectively. On the Agilent 244 K platform, the corresponding values were 0.48 and 0.36. And on the Affymetrix U133A platform, the values were 0.40 and 0.29.

Association between focal CNVs and patient survival

We searched for genomic loci where CNVs were associated with patient survival. We used Cox proportional

Table 1 Prognostic loci of ovarian cancer progression. The (Wald test) *p* values were obtained from Cox proportional hazard model with *f*CNV and *b*CNV data on the Affymetrix array platform as covariates

Chr	Region(Mb)	*p*-value	Representative genes
1	239.017–239.540	2.7E-04	KMO
3	171.107–174.249	1.0E-05	ECT2, EIF5A2, CLDN11, MYNN, LRRC31, EVI1
4	19.677–20.283	8.2E-06	
12	22.873–23.388	6.5E-04	KIAA0528
18	24.633–24.729	1.7E-04	
X	32.237–32.418	5.8E-06	

hazard model with the disease progression free survival (PFS) data of the patients. At each genomic site, we used *f*CNV and *b*CNV as two covariates to evaluate their association with PFS. We obtained similar results are similar on all of the three microarray platforms (Additional file 2 Figure S4 and Additional file 1: Table S4). The most significant loci for prognosis are listed in Table 1. Additional file 2: Figure S5 showed the *p*-values of the *f*CNV and *b*CNV in the Cox models. One of the most significant site was located at 3q26.2 (genome coordinate: Ch3:171–174 Mb). We found that mRNA expression of the genes encoded in the region is also significantly associated with PFS (Fig. 7). Interestingly, EIF5A2, which is one of the genes in the region, was found to be prognostic in ovarian cancer [55], colorectal carcinoma [56] and urothelial carcinoma [57]. Several other genes also had expression values significantly associated with PFS (p-value < 0.001, Wald test). These associations lend further support of the genomic with prognostic value.

Test of *f*CNV detection algorithm with computer simulated data

To evaluate the performance of the algorithm that we developed for *f*CNV detection, we applied the algorithm into the computer-generated data. The simulated data were generated following a very simple scenario: The profile contains only one amplicon at the center (See Method section for details). Detection of the amplicon depends on the amplitude (height h) of the amplicon over the noise level (ε), and the width of the amplicon (n, number of probes). The false positive rate approaches to 0 when $h/\varepsilon > 0.9$ and $n > 30$ (Additional file 2: Figure S6.1 and S6.2). In addition, since one of the prerequisites of oncogenes is that the number of focal gains ought to be significantly greater than the number of focal losses, it follows that the frequency of patients with the amplicon cannot be too low for the positive/correct identification. We found that f should be greater than 0.02 such that we can have a p-value <0.05 when 500 patients were tested (Additional file 2: Figure S6.3). These results suggested that we may not be able to identify the cancer driver genes if the frequency is less than 2 % of the patients in our study, or the focal CNVs are too short (e.g., with fewer than 30 probes covering the aberrant genomic region).

Conclusions

In this study, we developed a new approach to CNV analysis based on spectral decomposition CNV profile that separates focal CNV from broad CNVs. Using this approach, we performed an analysis of 587 serous ovarian cancer samples and found significant focal regions that are likely to contain cancer drivers. These regions have partial overlaps with regions that had been reported in previous analyses, but significant differences were also noted. Our results yielded a list of interesting findings, such as focal gains around ESR1, focal loss around LSAMP, prognostic site at 3q26.2 and sub-telomeric losses. 29 of the 42 focal regions from our analysis overlapped with the focal regions reported by previous pan-cancer analysis using GISTIC, which suggests that our results are in general agreements with previous analyses but also offered new focal regions of interest, which demand further investigations.

We also formerly tested the association between gain/loss and oncogene/TSG. Our results confirmed that the recurrent focal gains were significantly associated with the known oncogenes and recurrent losses associated with TSGs and the CNVs had a greater effect on the mRNA expression of the driver genes than that of the non-driver genes. Our results also showed focal CNVs had greater effects on expression of cancer driver genes than that of the non-driver genes. Our study demonstrated that spectral decomposition of CNVs offers a powerful new way of understanding the role of CNVs in cancer.

Additional files

Additional file 1: Table S1. Sample information of 587 ovarian patients and normal tissue samples with Copy number variation and mRNA data on different platforms. **Table S2.** Distributions of focal gains/losses in primary tumors and adjacent normal tissue separately on each of the three microarray platforms. **Table S3.** List of genes encoded in the recurrent focal event regions. **Table S4.** Prognostic results of ovarian cancer progression using *f*CNV on three platforms. (XLSX 20403 kb)

Additional file 2: The additional file contains all the details of additional/supplementary figures (**Figure S1**–**S6**) and tables (**Table S1**–**S4**). (DOCX 1574 kb)

Abbreviations

Affymetrix SNP 6.0, affymetrix genome-wide human SNP array 6.0; Affymetrix U133A, affymetrix human genome HTS U133A 2.0; Agilent 244 K, agilent 244 K custom human gene expression G4502A-07–3; Agilent G4447A 1 M, agilent SurePrint G3 human CGH microarray kit 1×1M; bCNV, broad copy number variation; CI, confidence interval; CNP, copy number polymorphism; fCNV, focal copy number variation; GISTIC, genomic identification of significant targets in cancer; Illumima 1 M-Duo, Illumina Human1M-Duo BeadChip; Mb, million base pair; PFS, progression free survival; RNAseq, whole transcriptome shotgun sequencing; RPKM, reads per kilobase per million of mapped reads; SNP, single nucleotide polymorphism; TCGA, the cancer genome atlas; TSG, tumor suppressor gene

Acknowledgements

This work was supported in part by the U.S. National Cancer Institute for the MD Anderson TCGA Genome Data Analysis Center through grant numbers CA143883 and CA083639 and the Bioinformatics Shared Resource through grant P30 CA016672. This work was also partially supported by the National Nature Science Foundation of China through grant number 31301040. This work was also partially supported by Award Number P50 CA098258 from the National Cancer Institute.

Funding

This work was supported in part by the U.S. National Cancer Institute for the MD Anderson TCGA Genome Data Analysis Center through grant numbers

CA143883 and CA083639 and the Bioinformatics Shared Resource through grant P30 CA016672. This work was also partially supported by the National Nature Science Foundation of China through grant number 31301040. This work was also partially supported by Award Number P50 CA098258 from the National Cancer Institute. Non-financial competing interests are included.

Availability of data and materials

The code for our analysis could be found via the link below: https://drive.google.com/drive/folders/0B6Q6G-z3ELntWlIEd29IOVpyYzA.

Authors' contributions

LcZ and LZ wrote the main manuscript text and YY provided additional ideas. LcZ and YY developed analytical methods and tools. LcZ, YY and KHL prepared figures and tables. KHL performed the pathway analysis and interpretation of results. All authors reviewed the manuscript. All authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Consent for publication

Not applicable.

Ethics approval and consent to participate

Not applicable.

Author details

[1]Department of Bioinformatics and Computational Biology, The University of Texas MD Anderson Cancer Center, 1400 Pressler St, Unit 1410, Houston, TX 77401, USA. [2]Department of Biostatistics, The University of Texas MD Anderson Cancer Center, 1400 Pressler St, Unit 1410, Houston, TX 77401, USA. [3]Department of Gynecologic Oncology and Reproductive Medicine, The University of Texas MD Anderson Cancer Center, Houston, TX, USA. [4]Department of Statistics, Rice University, Houston, TX, USA. [5]Department of Biophysics, College of Bioinformatics Sciences and Technology, Harbin Medical University, Harbin, China.

Received: 4 December 2015 Accepted: 14 May 2016
Published online: 26 May 2016

References

1. Hastings PJ, Lupski JR, Rosenberg SM, Ira G. Mechanisms of change in gene copy number. Nat Rev Genet. 2009;10(8):551–64.
2. Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. Cell. 2011;144(5):646–74.
3. Mermel CH, Schumacher SE, Hill B, Meyerson ML, Beroukhim R, Getz G. GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. Genome Biol. 2011;12(4):R41.
4. Beroukhim R, Mermel CH, Porter D, Wei G, Raychaudhuri S, Donovan J, Barretina J, Boehm JS, Dobson J, Urashima M et al. The landscape of somatic copy-number alteration across human cancers. Nature. 2010; 463(7283):899–905.
5. van Gent DC, Hoeijmakers JH, Kanaar R. Chromosomal stability and the DNA double-stranded break connection. Nat Rev Genet. 2001;2(3):196–206.
6. Pihan GA, Purohit A, Wallace J, Knecht H, Woda B, Quesenberry P, Doxsey SJ. Centrosome defects and genetic instability in malignant tumors. Cancer Res. 1998;58(17):3974–85.
7. Guichard C, Amaddeo G, Imbeaud S, Ladeiro Y, Pelletier L, Maad IB, Calderaro J, Bioulac-Sage P, Letexier M, Degos F, et al. Integrated analysis of somatic mutations and focal copy-number changes identifies key genes and pathways in hepatocellular carcinoma. Nat Genet. 2012;44(6):694–8.
8. Kallioniemi OP, Kallioniemi A, Kurisu W, Thor A, Chen LC, Smith HS, Waldman FM, Pinkel D, Gray JW. ERBB2 amplification in breast cancer analyzed by fluorescence in situ hybridization. Proc Natl Acad Sci U S A. 1992;89(12):5321–5.
9. Smith JS, Tachibana I, Passe SM, Huntley BK, Borell TJ, Iturria N, O'Fallon JR, Schaefer PL, Scheithauer BW, James CD, et al. PTEN mutation, EGFR amplification, and outcome in patients with anaplastic astrocytoma and glioblastoma multiforme. J Natl Cancer Inst. 2001;93(16):1246–56.
10. Albertson DG. Gene amplification in cancer. Trends Genet. 2006;22(8):447–55.
11. Callender T, el-Naggar AK, Lee MS, Frankenthaler R, Luna MA, Batsakis JG. PRAD-1 (CCND1)/cyclin D1 oncogene amplification in primary head and neck squamous cell carcinoma. Cancer. 1994;74(1):152–8.
12. Bhargava R, Gerald WL, Li AR, Pan Q, Lal P, Ladanyi M, Chen B. EGFR gene amplification in breast cancer: correlation with epidermal growth factor receptor mRNA and protein expression and HER-2 status and absence of EGFR-activating mutations. Mod Pathol. 2005;18(8):1027–33.
13. Illei PB, Rusch VW, Zakowski MF, Ladanyi M. Homozygous deletion of CDKN2A and codeletion of the methylthioadenosine phosphorylase gene in the majority of pleural mesotheliomas. Clin Cancer Res. 2003;9(6):2108–13.
14. Li C, Larsson C, Futreal A, Lancaster J, Phelan C, Aspenblad U, Sundelin B, Liu Y, Ekman P, Auer G et al. Identification of two distinct deleted regions on chromosome 13 in prostate cancer. Oncogene. 1998;16(4):481–7.
15. Meuwissen R, Linn SC, Linnoila RI, Zevenhoven J, Mooi WJ, Berns A. Induction of small cell lung cancer by somatic inactivation of both Trp53 and Rb1 in a conditional mouse model. Cancer Cell. 2003;4(3):181–9.
16. Solimini NL, Xu Q, Mermel CH, Liang AC, Schlabach MR, Luo J, Burrows AE, Anselmo AN, Bredemeyer AL, Li MZ, et al. Recurrent hemizygous deletions in cancers may optimize proliferative potential. Science. 2012;337(6090):104–9.
17. Wiedemeyer R, Brennan C, Heffernan TP, Xiao Y, Mahoney J, Protopopov A, Zheng H, Bignell G, Furnari F, Cavenee WK et al. Feedback circuit among INK4 tumor suppressors constrains human glioblastoma development. Cancer Cell. 2008;13(4):355–64.
18. The Cancer Genome Atlas Research Network. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature. 2008; 455(7216):1061–8.
19. Chitale D, Gong Y, Taylor BS, Broderick S, Brennan C, Somwar R, Golas B, Wang L, Motoi N, Szoke J et al. An integrated genomic analysis of lung cancer reveals loss of DUSP4 in EGFR-mutant tumors. Oncogene. 2009; 28(31):2773–83.
20. Diskin SJ, Eck T, Greshock J, Mosse YP, Naylor T, Stoeckert Jr CJ, Weber BL, Maris JM, Grant GR. STAC: a method for testing the significance of DNA copy number aberrations across multiple array-CGH experiments. Genome Res. 2006;16(9):1149–58.
21. Cancer Genome Atlas Research N. Integrated genomic analyses of ovarian carcinoma. Nature. 2011;474(7353):609–15.
22. Yuan X, Yu G, Hou X, Shih Ie M, Clarke R, Zhang J, Hoffman EP, Wang RR, Zhang Z, Wang Y. Genome-wide identification of significant aberrations in cancer genome. BMC Genomics. 2012;13:342.
23. Engler DA, Gupta S, Growdon WB, Drapkin RI, Nitta M, Sergent PA, Allred SF, Gross J, Deavers MT, Kuo WL et al. Genome wide DNA copy number analysis of serous type ovarian carcinomas identifies genetic markers predictive of clinical outcome. PLoS One. 2012;7(2):e30996.
24. Davis SJ, Sheppard KE, Pearson RB, Campbell IG, Gorringe KL, Simpson KJ. Functional analysis of genes in regions commonly amplified in high-grade serous and endometrioid ovarian cancer. Clin Cancer Res. 2013;19(6):1411–21.
25. Wang ZGC, Birkbak NJ, Culhane AC, Drapkin R, Fatima A, Tian RY, Schwede M, Alsop K, Daniels KE, Piao HY, et al. Profiles of genomic instability in high-grade serous ovarian cancer predict treatment outcome. Clin Cancer Res. 2012;18(20):5806–15.
26. Gorringe KL, George J, Anglesio MS, Ramakrishna M, Etemadmoghadam D, Cowin P, Sridhar A, Williams LH, Boyle SE, Yanaihara N, et al. Copy number analysis identifies novel interactions between genomic loci in ovarian cancer. PLoS One. 2010;5(9):e11408.
27. Bignell GR, Greenman CD, Davies H, Butler AP, Edkins S, Andrews JM, Buck G, Chen LN, Beare D, Latimer C, et al. Signatures of mutation and selection in the cancer genome. Nature. 2010;463(7283):893–8.
28. Etemadmoghadam D, de Fazio A, Beroukhim R, Mermel C, George J, Getz G, Tothill R, Okamoto A, Raeder MB, Harnett P, et al. Integrated genome-wide DNA copy number and expression analysis identifies distinct mechanisms of primary chemoresistance in ovarian carcinomas. Clin Cancer Res. 2009; 15(4):1417–27.
29. Gorringe KL, Ramakrishna M, Williams LH, Sridhar A, Boyle SE, Bearfoot JL, Li J, Anglesio MS, Campbell IG. Are there any more ovarian tumor suppressor genes? A new perspective using ultra high-resolution copy number and loss of heterozygosity analysis. Genes Chromosomes Cancer. 2009;48(10): 931–42.
30. Tukey JW. Exploratory data analysis. In: Reading: Addison-Wesley Publishing Co.; 1977.
31. Hur Y, Lee H. Wavelet-based identification of DNA focal genomic aberrations from single nucleotide polymorphism arrays. BMC Bioinformatics. 2011;12:146.

32. Hsu L, Self SG, Grove D, Randolph T, Wang K, Delrow JJ, Loo L, Porter P. Denoising array-based comparative genomic hybridization data using wavelets. Biostatistics. 2005;6(2):211–26.
33. Pique-Regi R, Tsau ES, Ortega A, Seeger R, Asgharzadeh S. Wavelet footprints and sparse Bayesian learning for DNA copy number change analysis. Int Conf Acoust Spee. 2007;353–356.
34. Wang YH, Wang SL, Zinn AR. rSWTi: a robust stationary wavelet denoising method for array CGH data. In: Proceedings of the 7th Ieee International Symposium on Bioinformatics and Bioengineering, Vols I and Ii. 2007. p. 1066–70.
35. Alqallaf AK, Teak AH. DNA copy number detection and sigma filter. IEEE Int Workshop Genomic Signal Proc Stat. 2007;2007:66–9.
36. Haugen TB. [Bootstrapping–popular method without a Norwegian name]. Tidsskr Nor Laegeforen. 2012;132(12-13):1497.
37. Moore JH. Bootstrapping, permutation testing and the method of surrogate data. Phys Med Biol. 1999;44(6):L11–2.
38. Higgins ME, Claremont M, Major JE, Sander C, Lash AE. CancerGenes: a gene selection resource for cancer genome projects. Nucleic Acids Res. 2007; 35(Database issue):D721–6.
39. Futreal PA, Coin L, Marshall M, Down T, Hubbard T, Wooster R, Rahman N, Stratton MR. A census of human cancer genes. Nat Rev Cancer. 2004;4(3): 177–83.
40. Bradburn MJ, Clark TG, Love SB, Altman DG. Survival analysis part II: multivariate data analysis–an introduction to concepts and methods. Br J Cancer. 2003;89(3):431–6.
41. Zack TI, Schumacher SE, Carter SL, Cherniack AD, Saksena G, Tabak B, Lawrence MS, Zhsng CZ, Wala J, Mermel CH, et al. Pan-cancer patterns of somatic copy number alteration. Nat Genet. 2013;45(10):1134–40.
42. Elangovan S, Ramachandran S, Venkatesan N, Ananth S, Gnana-Prakasam JP, Martin PM, Browning DD, Schoenlein PV, Prasad PD, Ganapathy V, et al. SIRT1 is essential for oncogenic signaling by estrogen/estrogen receptor alpha in breast cancer. Cancer Res. 2011;71(21):6654–64.
43. Moelans CB, Monsuur HN, de Pinth JH, Radersma RD, de Weger RA, van Diest PJ. ESR1 amplification is rare in breast cancer and is associated with high grade and high proliferation: a multiplex ligation-dependent probe amplification study. Cell Oncol. 2010.
44. Vincent-Salomon A, Raynal V, Lucchesi C, Gruel N, Delattre O. ESR1 gene amplification in breast cancer: a common phenomenon? Nat Genet. 2008; 40(7):809. author reply 810-802.
45. McIntyre MH, Kantoff PW, Stampfer MJ, Mucci LA, Parslow D, Li H, Gaziano JM, Abe M, Ma J. Prostate cancer risk and ESR1 TA, ESR2 CA repeat polymorphisms. Cancer Epidemiol Biomark Prev. 2007;16(11):2233–6.
46. Kresse SH, Ohnstad HO, Paulsen EB, Bjerkehagen B, Szuhai K, Serra M, Schaefer KL, Myklebost O, Meza-Zepeda LA. LSAMP, a novel candidate tumor suppressor gene in human osteosarcomas, identified by array comparative genomic hybridization. Genes Chromosomes Cancer. 2009; 48(8):679–93.
47. Kuhn MW, Radtke I, Bullinger L, Goorha S, Cheng J, Edelmann J, Gohlke J, Su X, Paschka P, Pounds S, et al. High-resolution genomic profiling of adult and pediatric core-binding factor acute myeloid leukemia reveals new recurrent genomic alterations. Blood. 2012;119(10):e67–75.
48. Dereli-Oz A, Versini G, Halazonetis TD. Studies of genomic copy number changes in human cancers reveal signatures of DNA replication stress. Mol Oncol. 2011;5(4):308–14.
49. Arlt MF, Mulle JG, Schaibley VM, Ragland RL, Durkin SG, Warren ST, Glover TW. Replication stress induces genome-wide copy number changes in human cells that resemble polymorphic and pathogenic variants. Am J Hum Genet. 2009;84(3):339–50.
50. Yen CC, Chen WM, Chen TH, Chen WY, Chen PC, Chiou HJ, Hung GY, Wu HT, Wei CJ, Shiau CY et al. Identification of chromosomal aberrations associated with disease progression and a novel 3q13.31 deletion involving LSAMP gene in osteosarcoma. Int J Oncol. 2009;35(4):775–88.
51. Counter CM, Gupta J, Harley CB, Leber B, Bacchetti S. Telomerase activity in normal leukocytes and in hematologic malignancies. Blood. 1995;85(9): 2315–20.
52. McPherson JP, Hande MP, Poonepalli A, Lemmers B, Zablocki E, Migon E, Shehabeldin A, Porras A, Karaskova J, Vukovic B et al. A role for Brca1 in chromosome end maintenance. Hum Mol Genet. 2006;15(6):831–8.
53. Mirabello L, Garcia-Closas M, Cawthon R, Lissowska J, Brinton LA, Peplonska B, Sherman ME, Savage SA. Leukocyte telomere length in a population-based case-control study of ovarian cancer: a pilot study. Cancer Causes Control. 2010;21(1):77–82.
54. Cerami E, Demir E, Schultz N, Taylor BS, Sander C. Automated network analysis identifies core pathways in glioblastoma. PLoS One. 2010;5(2):e8918.
55. Yang GF, Xie D, Liu JH, Luo JH, Li LJ, Hua WF, Wu HM, Kung HF, Zeng YX, Guan XY. Expression and amplification of eIF-5A2 in human epithelial ovarian tumors and overexpression of EIF-5A2 is a new independent predictor of outcome in patients with ovarian carcinoma. Gynecol Oncol. 2009;112(2):314–8.
56. Xie D, Ma NF, Pan ZZ, Wu HX, Liu YD, Wu GQ, Kung HF, Guan XY. Overexpression of EIF-5A2 is associated with metastasis of human colorectal carcinoma. Hum Pathol. 2008;39(1):80–6.
57. Chen W, Luo JH, Hua WF, Zhou FJ, Lin MC, Kung HF, Zeng YX, Guan XY, Xie D. Overexpression of EIF-5A2 is an independent predictor of outcome in patients of urothelial carcinoma of the bladder treated with radical cystectomy. Cancer Epidemiol Biomark Prev. 2009;18(2):400–8.