Published online 29 July 2016

Nucleic Acids Research, 2016, Vol. 44, No. 19 ***e150***
doi: 10.1093/nar/gkw674

FocalScan: Scanning for altered genes in cancer based on coordinated DNA and RNA change

Joakim Karlsson and Erik Larsson*

Department of Medical Biochemistry and Cell Biology, Institute of Biomedicine, The Sahlgrenska Academy, University of Gothenburg, SE-405 30 Gothenburg, Sweden

Received March 15, 2016; Revised June 22, 2016; Accepted July 17, 2016

ABSTRACT

Somatic genomic copy-number alterations can lead to transcriptional activation or inactivation of tumor driver or suppressor genes, contributing to the malignant properties of cancer cells. Selection for such events may manifest as recurrent amplifications or deletions of size-limited (focal) regions. While methods have been developed to identify such focal regions, finding the exact targeted genes remains a challenge. Algorithms are also available that integrate copy number and RNA expression data, to aid in identifying individual targeted genes, but specificity is lacking. Here, we describe FocalScan, a tool designed to simultaneously uncover patterns of focal copy number alteration and coordinated expression change, thus combining both principles. The method outputs a ranking of tentative cancer drivers or suppressors. FocalScan works with RNA-seq data, and unlike other tools it can scan the genome unaided by a gene annotation, enabling identification of novel putatively functional elements including lncRNAs. Application on a breast cancer data set suggests considerably better performance than other DNA/RNA integration tools.

INTRODUCTION

Tumors develop due to acquisition of somatic genomic changes that alter the activity or function of cancer driver genes. Identification of genes affected by such changes can thus improve our understanding of oncogenesis and aid in the development of novel therapies (1). A complicating factor is that most somatic alterations in tumors are nonfunctional passenger events that do not confer selective advantages to tumor cells.

One important mechanism by which genes are altered during tumor development is through copy number aberrations, i.e. amplifications or deletions of genomic regions. Frequently, these aberrations can affect entire chromosome arms, but it may also be that events spanning shorter regions recur at roughly the same position in multiple independent tumor samples ('focal regions'). Such patterns indicate selection for altered expression of genes that may drive oncogenesis (oncogenes) or hinder cancer growth (tumor suppressors). Tools have therefore been developed to find these frequently altered regions (2,3). However, the focal regions identified by existing tools often span a large number of genes, and many times it is not clear which gene is the target, that is, mediates the selective advantage provided by a particular alteration. While it seems intuitive that the gene closest to the most recurrently altered position in a region should be the main target, this is not always the case (4).

In order for a copy number change to confer a selective advantage, the expression level of some particular target gene also needs to be altered. Genes that, in a recurrently altered region, fail to show consistent expression changes in relation to copy number changes, e.g. due to lack of expression in some samples, are thus less likely to be the drivers upon which selective forces are acting in this region. Therefore, it is attractive to integrate expression and copy number data when searching for driver candidates. A common way of doing this is through the calculation of a correlation coefficient between changes in copy number and RNA (5,6). This favors either a linear (in the case of the Pearson coefficient) or a general increasing/decreasing relationship (for instance the Spearman coefficient) between the two types of alterations. Thus, if a gene is to be deemed of any potential importance, it is required to be consistently overexpressed when amplified, or underexpressed when deleted. Other methods have been developed that examine alternate measures of association (7,8).

Some of the methods that integrate copy number and expression data have recently been compared using both simulated and real data (9,10). Unfortunately, performance has been found to be lacking, especially with regards to specificity. A likely contributor to this lack of specificity is that, while driver genes are expected to show coordinated DNA and RNA change, such correlations can be expected also for many non-causal genes, and positive correlation is thus not sufficient to conclude a functional contribution. Additionally, unlike DNA-only tools such as GISTIC (4), the width (focality) of copy number changes is not taken into

*To whom correspondence should be addressed. Tel: +46 31 786 6942; Fax: +46 31 41 61 08; Email: erik.larsson@gu.se

account by current integrative methods. Furthermore, the degree of recurrence across multiple patients is not always considered. It is also conceivable that the driver of an aberration can be an un-annotated gene, for instance producing a long non-coding RNA (lncRNA). There is thus a lack of integrative approaches that can leverage the full possibilities of modern transcriptome sequencing technologies for inquiries into copy number altered regions.

Here, we propose a strategy for uncovering focally copy number altered loci that simultaneously show coordinated changes in expression, thus drawing strength from both strategies. Rather than a two-step approach, the method relies on an integrative metric that rewards focality, recurrency and coordinated RNA change. This metric can be used at the level of genes or in an annotation-independent manner. We apply this tool, named FocalScan, to a large breast cancer data set from The Cancer Genome Atlas (TCGA), and compare its performance to some existing programs for copy number/expression integration.

MATERIALS AND METHODS

A score that rewards recurrent coordinated focal copy number and expression changes

Positive selection acting on a specific cancer driver gene may lead to copy number alterations that recur in many independent tumor samples. For these changes to be functional, they should also be associated with a consistent effect on the expression of this gene. In addition, focally altered regions are more informative about likely driver genes than arm-length aberrations. To find genes that fulfill these three criteria, a scoring metric was constructed that rewards recurrent coordinated focal copy number and gene expression changes. This metric is applied either to genes or small genomics tiles/segments.

Copy number change is defined here as the $\log_2$-transformed ratio between the copy number of a genomic position in the sample and in the diploid scenario (two copies), as indicated in segmented copy number profiles required as input data. Segmented copy number levels are remapped by FocalScan onto annotation features (genes or genomic tiles, see below). Expression change of a gene/tile in a given sample is defined relative to the median expression of all samples that are diploid. Diploid samples are, by default, defined as those with absolute copy number change less than 0.1 at the position of the particular gene/tile being examined (changes less than 0.1 are often artifactual in nature (4)). The resulting expression change ratio is then $\log_2$-transformed to obtain a measure comparable to that of the copy number change. Prior to calculating the ratio, a pseudo value is added to avoid division by zero. In order for the pseudo value to be scale/unit independent, it is by default scaled in relation to the median level of all non-zero expression values in the data set. A scale factor can be given to further adjust this value relative to the median. We found that a scale factor of 10 gave the best performance on average across different cancer data sets, which was therefore selected as the default. Expression read counts supplied by the user are, by default, normalized using the median of the top 5% most highly expressed genes or genomic tiles, to compensate for cases where the reads of a few genes dominate the data. Other options also exist, such as library size normalization. It is also possible to use pre-normalized gene expression values when performing a gene level analysis.

A score is calculated, for each gene/tile, as the dot product between the two vectors that contain the above described $\log_2$ ratios for all samples (Figure 1A). To account for the focality of each copy number change, the copy number data is first subjected to a filtering procedure. This procedure acts as a high-pass filter by subtracting long copy number alterations. For a given genomic position, the amplitude is compared to those of two nearby positions, one upstream and one downstream, a fix genomic distance away, using the formula depicted in Figure 1B. This fixed distance is termed 'window size' and essentially acts as a length-based cutoff to distinguish long from focal events (by default 10 MB is used, although this parameter is user adjustable).

FocalScan allows the use of two separate analysis strategies: gene-based (Figure 1C, left) and tile-based (Figure 1C, right). The tile-based analysis scores the genome independently of any reference annotation. It can thus be used to discover potentially interesting changes also in intergenic regions (for instance, relating to novel lncRNAs). For this purpose, by default the genome is divided into 1000 bp overlapping tiles. A score is then calculated for each tile, as described above. The gene-based analysis instead calculates a score for each gene, as defined in a user provided genome reference annotation.

Positive scores indicate focal amplifications that are associated with increased expression or focal deletions that are associated with decreased expression. In some cases, negative scores can be found, suggesting an anti-coordinated relationship: amplification events associated with decreases in expression levels or deletions associated with increased expression. Large score values indicate that the coordinated changes are highly recurrent and/or that focal copy number aberrations have a large effect on expression.

As input, the tool expects gene expression read counts and segmented copy number data derived from high-density SNP/copy number arrays or genomic sequencing. Read mapping, gene/tile expression quantification and copy number segmentation are steps handled by external tools, although BED files are provided to simplify expression quantification using, e.g. bedtools (11). Several tools are available for copy number segmentation, including CNV Workshop (12) (arrays) and cn.MOPS (13) (sequencing).

Peak detection algorithm

Driver genes in copy number altered regions are assumed by FocalScan to show consistent coordination between copy number status and gene expression levels, but we can also expect a general trend of positive correlation between copy number status and gene expression. For this reason, drivers and surrounding passengers will form clusters/peaks of coordinated alteration. To select candidates on a genome-wide scale, a peak detection and gene prioritization algorithm was implemented. The construction of such an algorithm presented a number of challenges: First, peak widths cannot be assumed to be constant across the genome; Second,

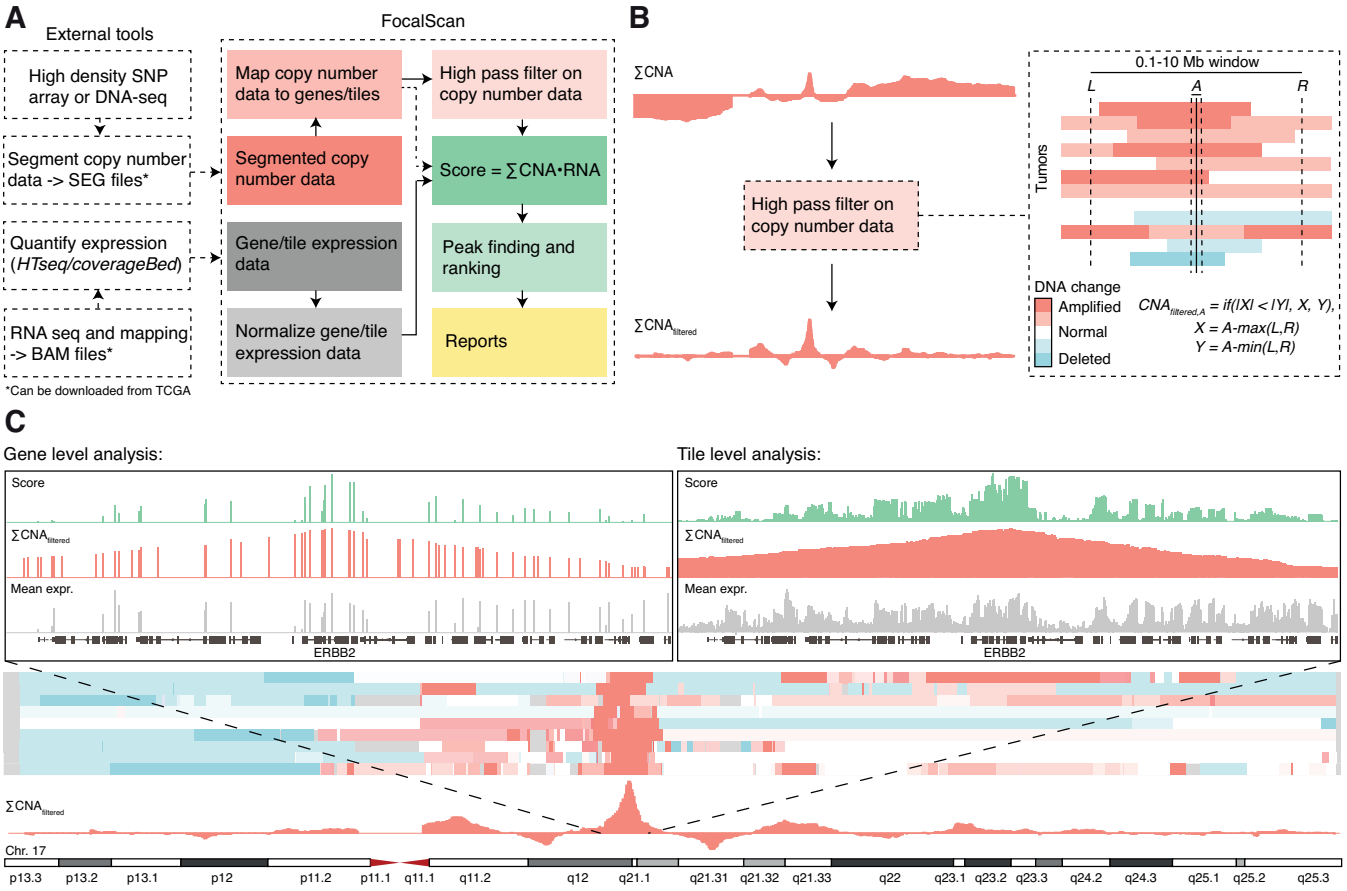

Figure 1. Illustration of the FocalScan method and output. (**A**) Overview of the method workflow. (**B**) Illustration of the 'high-pass' filter employed to filter out arm-length events. For each genomic position and sample, the copy number amplitude with minimal absolute value at two positions, a fix distance apart, is subtracted from the copy number amplitude at the current position. (**C**) Visualization of score tracks spanning the *ERBB2* locus, resulting from gene level analysis (left) and tile-level analysis (right). Red indicates amplification, blue deletion. CNA, copy number amplitude (change relative diploid).

coordination scores are highly variable from one position to another; Third, there are multiple scale levels on which peaks may be defined: depending on the desired sensitivity, a given event may constitute either an independent peak or a sub-peak to a stronger nearby driver.

To deal with these challenges, an algorithm was developed that considers the genomic score pattern on different 'scale levels' (taking some inspiration from another recent multi-scale approach (14)). The lowest, most granular, scale level defines a position as a peak if it has a larger score than its two neighbors. The next (higher) scale level uses the peaks defined on the previous level as input and again defines a peak if it is larger than its two neighbors, etc. Thus, on each successive level, sub-peaks merge into larger ones until only one peak dominates each chromosome (the 'maximal' scale). The strongest candidates persist across multiple scales and no prior assumptions regarding peak widths have to be made. An appropriate scale that picks out the most dominant genes/tiles across the genome can then be selected relative to the maximal one. As default, a scale corresponding to 70% of max is used. A lower one can be selected for increased sensitivity or a higher one for a more conservative result. Amplifications and deletions are analyzed separately this way and compiled into the final ranked list of candidates. The solution is illustrated in Supplementary Figure S1, where peaks detected at the 70% scale level (Supplementary Figure S1A) are contrasted with those detected on the more sensitive 60% level (Supplementary Figure S1B).

Data retrieval, pre-processing and method comparison

RNA-seq reads for each cancer type, aligned to the hg19 reference, were downloaded in BAM format from the online repository of TCGA (Cancer Genomics Hub: https://cghub.ucsc.edu). The reads were further binned to either genes or tiles to quantify expression levels. For gene-level expression quantification, low quality alignments were first filtered out with samtools and the flag '-q 1'. Then, htseq-count (15) was used with the parameters '-m intersection-strict' (a read is mapped to the intersect of all features spanning any read position) and '-s no' (do not assume a stranded sequencing protocol). As reference annotation, GENCODE v17 (16) was used. For tile-level expression quantification, the bedtools coverageBed program (11) was used with '-split', '-counts' and a tiled genome as reference. Segmented (with circular binary segmentation) copy number data filtered for germline copy number alterations,

originally measured on Affymetrix SNP6 arrays, were also downloaded from TCGA.

The performance of FocalScan was compared with three other tools, pint (17), edira (18), DR-Correlate (5). In order to make the data compatible with these, some additional pre-processing steps were required. For pint, the documentation stated that the data should be approximately Gaussian. As such, gene read counts were library size ('RPM') normalized and $\log_2$-transformed. A pseudo value of one was added to avoid issues with log transforming zero (values based on median expression, as with FocalScan, were also evaluated, without much difference in performance). Copy number levels were given in $\log_2$-ratio format. For DR-Correlate, the documentation did not give any specific instructions for pre-processing. The data were therefore library size normalized and $\log_2$ transformed (adding a pseudo value equal to median non-zero expression). Similarly, copy number data were given on log scale (non-transformed data and other pseudo values were also evaluated, without noticeable performance differences). For edira, a specific function exists that takes ratio format data. Therefore, gene expression ratios were calculated relative to the median of diploid samples as described above for FocalScan. Again a pseudo value was added to avoid division by zero (median expression across the data set, only considering genes with reads). Copy number data were provided on $\log_2$ scale. In addition, since not all of the tools could handle missing values in the copy number data, genes with more than 10% missing values were removed and remaining cases were set to zero. Ranked gene lists were then calculated, aided by the comparison framework developed by Lahti *et al*. (9). Default parameters were used for all methods, including FocalScan.

RESULTS

A tool for uncovering recurrent focal DNA alterations coordinated with expression

We implemented a computational tool (FocalScan) that combines DNA copy number and RNA expression data from tumors for the purpose of identifying candidate cancer drivers, which aims to address shortcomings in existing methodology. FocalScan rewards genes or genomic regions where many tumors show coordinated changes in DNA copy number and RNA expression, and where DNA changes at the same time are focal/narrow (Figure 1A; Materials and Methods).

Briefly, to account for the size/focality of affected regions, a filter is applied that subtracts arm-length size events to favor focal copy number changes using a windowing approach (essentially a 'high pass' filter, Figure 1B). Next, in order to detect changes in DNA and RNA amplitude that are coordinated as well as recurrent, FocalScan next calculates the dot product between copy number changes (relative the diploid state) and expression level changes (relative a reference set of samples that are diploid at a given position) across all included tumors. This score rewards events that are recurrent (more terms are added to the sum) and show large equally directed changes in both quantities (larger terms are added), while at the same time being focal at the DNA level (surviving the high pass filter). In summary, for a given locus, a high score indicates a highly recurrent focal copy number event where copy number gain is associated with prominent RNA induction or, conversely, copy number loss is associated with strong RNA reduction.

FocalScan is designed to be used with RNA-seq data and, similar to existing tools, can score and assess individual genes with the help of a user provided gene annotation file (Figure 1C, left). However, it also gives the option to evaluate genomic regions unaided by a gene annotation, achieved by dividing chromosomes into partially overlapping genomic tiles (by default 1000 bp). This feature makes it possible to search for novel, non-annotated, transcribed elements, such as long noncoding RNAs, with potential roles in cancer development (Figure 1C, right). The results are presented in the form of a ranked list with statistics for each locus examined (gene or tile). In addition, a peak detection algorithm is employed to produce a reduced list of candidate driver loci, to account for genomic regions where multiple signals cluster closely together (see Materials and Methods). Results are also written to .wig files for visualization with compatible genome browsers including IGV (19).

FocalScan discovers known cancer genes and performs favorably compared to alternative methods

In order to test our method, we applied FocalScan to a set of 971 breast invasive carcinoma (BRCA) samples from TCGA (20). The data were pre-processed as described in Materials and Methods. To compare our results to other integrative methods, we identified additional tools that were applicable on segmented copy number data from high-density SNP arrays and RNA-seq data, and that did not require a separate data set of normal tissue samples. The additional methods were required to have available documentation, be open source and to execute without errors upon following the provided instructions. Three tools were found to be suitable: 'edira' (18) (which uses a modified correlation coefficient), 'pint (SimCCA)' (17) (based on similarity constrained canonical correlation analysis) and 'DR-Correlate' (5) (Pearson or Spearman correlation, or a *t*-test based method; Pearson correlation was used in this study). Although other methods are available, their relative performances have been studied previously (9,10,21).

We next assessed the degree of enrichment of known cancer genes among top ranked hits nominated by the different tools, using a list of known cancer drivers. The list was compiled from the databases Cancer Gene Census (CGC) (22) and intOGen (23). While all methods had enrichments higher than what would be expected by chance, we found that FocalScan performed considerably better than the reference methods based on this metric (Figure 2A). A total of 70% of the top 10, 35% of the top 20 genes, and 25% of the top 100 genes overlapped with known cancer genes, to be compared with 20%, 10% and 6%, respectively, for pint, which scored best among the reference tools. The proportion known cancer genes discovered by FocalScan among the top 100 genes was significantly higher (Fisher test; $p = 0.016$ versus pint, $p = 6.6 \times 10^{-5}$ versus edira and $p = 6.5 \times 10^{-4}$ versus DR-Correlate) than for the other three methods.

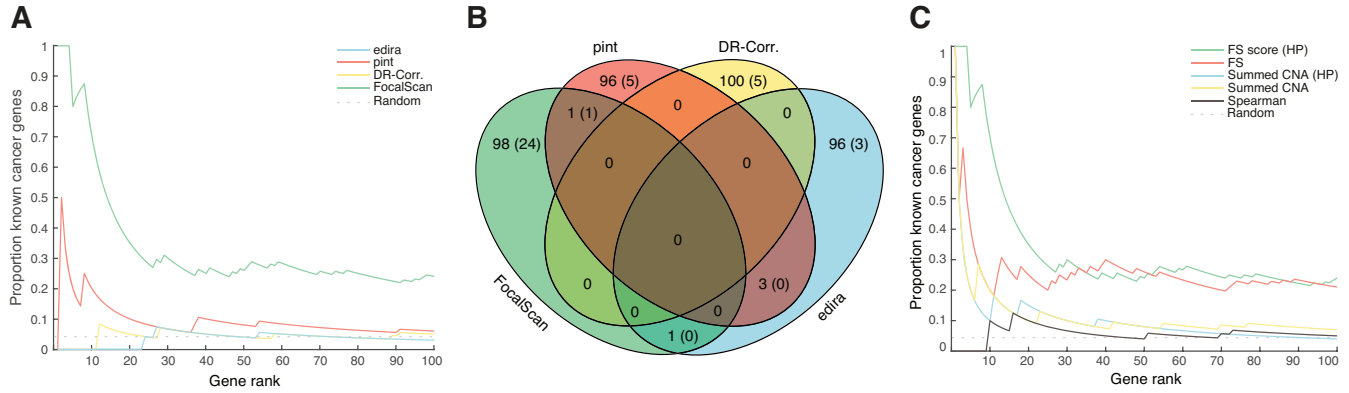

Figure 2. Performance comparison results. (**A**) Cumulative proportion known cancer genes across the ranked gene lists obtained using FocalScan, pint, DR-Correlate and edira. 'Random' indicates the proportion of cancer genes expected by chance. (B) Overlaps among the top 100 genes found by each method, with known cancer genes indicated in parenthesis. (**C**) Cumulative proportion of known cancer genes obtained with rankings based on FocalScan score ('FS') with ('HP') and without high-pass filtered copy numbers, as well as peaks detected based on summed copy number amplitudes alone (with and without filter) and Spearman correlation between filtered copy number and expression changes. CNA, copy number amplitude (change relative diploid).

Among the top 100 ranked genes found by FocalScan, only two (*WHSC1L1*, *ADIPOR1*) were identified by more than one method (Figure 2B). Of these, *WHSC1L1* is a known driver. The only genes co-discovered exclusively by the other methods were *VPS4A*, *DDX19A* and *COG9,* none of which are known drivers. Table 1 lists the top ten ranked cancer gene candidates suggested by FocalScan. Seven of these, *ERBB2*, *CCND1*, *WHSC1L1*, *EGFR*, *IGF1R*, *FGFR2* and *CCNE1* are cancer genes according to either CGC or intOGen. Notably, these ten genes were outliers not only in terms of enrichment, but also in terms of FocalScan score, which dropped sharply beyond this rank (Supplementary Figure S2). Additionally, we compared performance on data from 20 other cancer types characterized by TCGA and found that FocalScan offered improved performance in the vast majority of cancers (Supplementary Figure S3), with the exception of adrenocortical carcinoma (ACC) and thyroid carcinoma (THCA).

To investigate how the properties unique to FocalScan contributed to the result, we compared known cancer gene enrichment results from FocalScan to those obtained using the following four alternative scoring metrics: Spearman correlation instead of FocalScan score (dot product), absolute sum of copy number amplitudes, absolute sum of copy number amplitudes with focality filter and FocalScan score without focality filter. The comparison was done both with (Figure 2C) and without (Supplementary Figure S4A) peak detection on the resulting genome-wide score peaks. We found that both the focality filter and the dot product approach contributed to elevating the performance of FocalScan above the alternatives. The window size of the focality filter, a tunable parameter in FocalScan, was not found to have a large effect on performance (Supplementary Figure S4B). In general, different peak detection scale levels (see Materials and Methods) also performed similarly in terms of known cancer gene enrichment among the top ranked genes (Supplementary Figure S4C).

While CGC/intOGen enrichment should be useful as an indirect measure of overall performance (24), it should be noted that true positive cancer driver genes may be missing in these databases. Among the top 10 genes nominated by FocalScan, we note that *TRAF4* (rank 5), while absent in CGC/intOGen, has been shown to drive breast cancer metastasis (25) and was initially identified as overexpressed and amplified in breast cancer (26). Likewise, *PHGDH* (rank 8) is a known target of focal amplification in breast and other cancers (27). Taken together, we find that the list of top candidates in breast cancer is strongly enriched for known cancer genes, supporting the usefulness of FocalScan for nominating likely drivers.

Transcribed intergenic regions associated with genomic aberrations

Having shown that the gene-centric approach gives sensible rankings of known cancer genes, we next applied the tile-based method to the same breast cancer data set in order to identify putative unannotated altered genes. The top ranking genomic peaks detected using this approach largely overlapped with the top candidates from the gene-centric analysis (Supplementary Table S1). Cases of disparity could be due to, for instance, uncertainty of read origin when two genes overlap. Multiple intergenic peaks were additionally detected (Supplementary Table S1, rows without gene names), including a strong signal in between *ETV6* (involved in tumorigenic rearrangements in several types of cancer (28–30)) and the neighboring *BCL2L14* (originally thought to have pro-apoptotic activity (31), which has since been refuted (32)) on chromosome 12 (Figure 3A). Raw read data from the tumors, as well as tissues in the Human BodyMap 2.0 compendium (accessed via Ensemble (33)), revealed an approximately 2.5 kb mono-exonic region with elevated read coverage (Figure 3A). Expression was higher in focally amplified tumors than in those having broad amplification of this locus (Figure 3B, $p = 1.27 \times 10^{-10}$ with a Wilcoxon rank-sum test). A pan-cancer comparison showed that these focally amplified breast cancer tumors accounted for the highest expression levels of this RNA observed across 19 TCGA cancer types, although the median level was higher in some other cancers (Figure 3C).

Table 1. Top ranked genes in breast cancer samples from TCGA

Rank	Gene symbol	Score	Sum CNA_{HP}	# Amp.	# Del.	ρ	*P*
1	ERBB2[a]	943.9	263	142	5	0.312	2.36E-23
2	CCND1[a]	446.9	220	197	11	0.370	7.77E-33
3	WHSC1L1[a]	441.0	202	170	39	0.327	1.10E-25
4	EGFR[a]	129.6	21	16	2	0.025	0.444
5	TRAF4	118.4	81	108	15	0.210	4.75E-11
6	IGF1R[a]	116.8	27	38	15	0.128	6.14E-5
7	FGFR2[a]	96.9	25	24	5	0.018	0.574
8	CCNE1[a]	84.5	27	37	22	0.085	0.008
9	PHGDH	80.5	21	37	16	0.032	0.314
10	AIM1	68.8	24	31	3	0.063	0.050

[a]Known cancer gene (Cancer Gene Census and/or intOGen).
'Sum CNA_{HP}': Summed filtered copy number amplitide (change relative to diploid). '# Amp.': Number of samples with amplification. '#Del.': Number of samples with deletion. ρ: Spearman correlation coefficient. *P*: *P*-value of Spearman correlation.

Translating the sequence of the expressed region using ExPASy (34) revealed no open reading frames longer than 125 amino acids. A Pfam (35) search showed no matches to known protein domains, further supporting that the transcript is not protein coding, although it could theoretically code for a small peptide. Additionally, the region appears conserved across several mammalian species, suggesting it may have functional relevance (Figure 3A). H3K27Ac modifications in the region as revealed by ENCODE data (36) could indicate enhancer activity, possibly implying a role as an enhancer RNA (Figure 3A).

Additional signals in unannotated regions included a peak in the long non-coding RNA *PVT1* (Supplementary Figure S5A), known to be co-amplified with, and to up-regulate the protein levels of, the nearby *MYC* oncogene (37). Notably, this transcribed region was intronic and did not overlap with annotated *PVT1* exons. An absence of spliced reads at this position indicates that it could be a novel transcript, rather than an unannotated *PVT1* exon. Active transcription and open chromatin structure was further supported by flanking DNase I hypersensitive regions and overlapping H3K27Ac peaks (36). Additional focally amplified intergenic elements of potential interest are shown in Supplementary Figure S5B–D, on chromosome 8 in BRCA, chromosome 12 in glioblastoma multiforme (GBM) and chromosome 3 in head and neck squamous cell carcinoma (HNSC). The results show that FocalScan can nominate candidate non-annotated transcripts with amplification and expression patterns characteristic of cancer drivers, which may serve as starting points for functional studies.

DISCUSSION

Integrating copy number and expression data has the potential to reduce the search space when identifying targets of recurrent amplifications and deletions in cancer. FocalScan is a tool implemented to perform such a joint analysis, thereby aiding the search for functional alterations across multiple tumor samples in both annotated and non-annotated DNA regions. According to our assessments, FocalScan performs favorably compared to other integrative tools when evaluated on RNA-seq data and segmented copy number data from TCGA. The results suggest that top ranked genes should be useful candidates to evaluate as potential cancer drivers. The tool is primarily intended be used as a standalone application (as demonstrated here), but could also be used in conjunction with, e.g. GISTIC, to get further clues about genes in already established focal regions.

The overlap between the genes identified by the different tools was limited, likely due to a generally low sensitivity offered by available copy number/expression integrative methods (9,10). In contrast to FocalScan, enrichment of known cancer genes was low, and well-known breast cancer genes such as *ERBB2* and *CCND1*, were absent. There was also a lack of consensus among these tools for genes not discovered by FocalScan. While one may consider combining the results from multiple tools to increase sensitivity, our results suggest that this could lead to reduced specificity.

FocalScan can be applied in an annotation-independent manner, enabling discovery of non-annotated intergenic regions showing patterns of coordinated expression and focal copy number change. This is made possible by the non-gene-centric nature of RNA-seq. We demonstrate this functionality through application on breast cancer data, highlighting one such region on chromosome 12. While experimental validation would be needed to establish a functional role, this demonstrates the ability to identify putatively functional loci that would be undetectable in gene-centric analyses.

Challenges still remain in accurately separating main drivers, co-drivers and passengers within a given focally altered region. FocalScan employs a peak detection step designed to select the highest-scoring gene in a cluster of signals, assuming that each score peak only has one gene under selection. However, it is also possible that multiple genes (co-drivers) contribute in a given region, and one may therefore still wish to investigate other nearby highly scoring genes. In cases of uncertainty, FocalScan can output auxiliary statistics such as mean expression levels and correlation for each gene to enable additional filtering of results.

In conclusion, we describe a novel bioinformatics method, FocalScan, designed to suggest candidate driver genes that are targets of focal copy number alteration in cancer. This is accomplished by integrating copy-number and gene expression data across multiple tumor samples. The tool is capable of fully utilizing the potential of RNA-seq data to inquire into changes in intergenic regions, aiding the search for novel transcribed elements of cancer relevance.

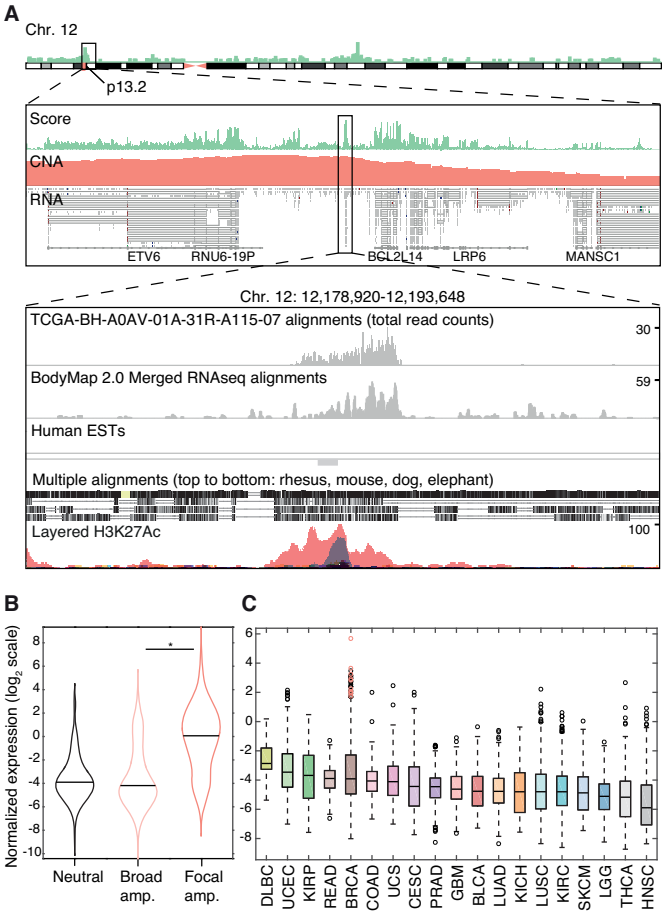

Figure 3. Characteristics of the region spanning a putative novel lncRNA. (**A**) From top to bottom: FocalScan scores for chromosome 12; tile-based scores across the indicated zoomed in region; copy number amplitudes across 971 BRCA tumors; individual RNA-seq reads for TCGA-BH-A0AV-01A-31R-A115-07; read counts in TCGA-BH-A0AV-01A-31R-A115-07 across the region chr12:12 178 920–12 193 648; Human BodyMap 2.0 coverage; human expressed sequence tags (ESTs); conservation scores for rhesus, mouse, dog and elephant; enrichment of H3K27ac modifications from ChIP-seq data. (**B**) Expression of chr12:12 184 830–12 187 738 in samples without copy number alteration (absolute copy number amplitude < 0.1), broad amplifications (copy number amplitude > 0.1 and copy number segment longer than 1 Mbp) and focal (shorter than 1 Mbp) amplifications. Horizontal black lines indicate median. '*' indicates a significant difference according to a Wilcoxon rank-sum test ($P = 1.27 \times 10^{-10}$). Expression was normalized using the percentile method, as described in 'Materials and methods'. (**C**) Normalized expression of this region in 19 cancer types (names abbreviated according to TCGA). Red circles indicate focally amplified samples in BRCA. CNA, copy number amplitude (change relative diploid).

AVAILABILITY

FocalScan is command-line based and implemented in MATLAB. The software and accompanying documentation is available on GitHub at https://github.com/jowkar/focalscan. It does not require any additional MATLAB toolboxes, and can be run as a standalone application without any MATLAB installation (for Mac and Linux). The software has been tested on Mac, Linux and Windows. Note, however, that the standalone application is restricted to those platforms that are capable of running the v901 MATLAB runtime. The amount of RAM required is dependent on the amount of samples and on whether a gene- or tile-level analysis is performed. A gene-level analysis with 1000 samples can be performed on a laptop with ~8 GB of memory. A corresponding tile-level analysis requires about 30 GB RAM.

To run FocalScan, input data can be given in several ways. For gene-level analysis, RNA data can be given in CSV-format. For both tile and gene-level analysis, RNA data can also be provided as read count files corresponding to each sample. For tile-level analysis, only the second option is available. Copy number data should be given in segmented ('SEG') format for both types of analyses. A script to quantify tile-level expression in included with FocalScan (requires bedtools). For gene-level analysis, a genome annotation should be given in BED format (GENCODE v17 is provided by default).

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The results published here are in whole or part based upon data generated by The Cancer Genome Atlas pilot project established by the NCI and NHGRI. Information about TCGA and the investigators and institutions who constitute the TCGA research network can be found at 'http://cancergenome.nih.gov'. The computations were in part performed on resources provided by SNIC through Uppsala Multidisciplinary Center for Advanced Computational Science (UPPMAX) under project b2012108.

FUNDING

Knut and Alice Wallenberg Foundation; Swedish Foundation for Strategic Research; Swedish Medical Research Council; Swedish Cancer Society; Åke Wiberg foundation; Lars Erik Lundberg Foundation for Research and Education.

Conflict of interest statement. None declared.

REFERENCES

1. Stratton,M.R. (2011) Exploring the genomes of cancer Cells: Progress and promise. *Science*, **331**, 1553–1558.
2. Beroukhim,R., Mermel,C.H., Porter,D., Wei,G., Raychaudhuri,S., Donovan,J., Barretina,J., Boehm,J.S., Dobson,J., Urashima,M. *et al.* (2010) The landscape of somatic copy-number alteration across human cancers. *Nature*, **463**, 899–905.
3. Klijn,C., Holstege,H., De Ridder,J., Liu,X., Reinders,M., Jonkers,J. and Wessels,L. (2008) Identification of cancer genes using a statistical framework for multiexperiment analysis of nondiscretized array CGH data. *Nucleic Acids Res.*, **36**, e13.
4. Mermel,C.H., Schumacher,S.E., Hill,B., Meyerson,M.L., Beroukhim,R. and Getz,G. (2011) GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biol.*, **12**, R41.
5. Salari,K., Tibshirani,R. and Pollack,J.R. (2009) DR-Integrator: A new analytic tool for integrating DNA copy number and gene expression data. *Bioinformatics*, **26**, 414–416.
6. Lipson,D., Ben-Dor,A., Dehan,E. and Yakhini,Z. (2004) Joint analysis of DNA copy numbers and gene expression levels. *Algorithms Bioinformatics Proc.*, **3240**, 135–146.

7. Yang,J., Wang,X., Kim,M., Xie,Y. and Xiao,G. (2014) Detection of candidate tumor driver genes using a fully integrated Bayesian approach. *Stat. Med.*, **33**, 1784–1800.
8. Lê Cao,K.-A., González,I. and Déjean,S. (2009) integrOmics: An R package to unravel relationships between two omics datasets. *Bioinformatics*, **25**, 2855–2856.
9. Lahti,L., Schafer,M., Klein,H., Bicciato,S. and Dugas,M. (2013) Cancer gene prioritization by integrative analysis of mRNA expression and DNA copy number data: a comparative review. *Brief. Bioinform.*, **14**, 27–35.
10. Louhimo,R., Lepikhova,T., Monni,O. and Hautaniemi,S. (2012) Comparative analysis of algorithms for integration of copy number and expression data. *Nat. Methods*, **9**, 351–355.
11. Quinlan,A.R. and Hall,I.M. (2010) BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics*, **26**, 841–842.
12. Gai,X., Perin,J.C., Murphy,K., O'Hara,R., D'arcy,M., Wenocur,A., Xie,H.M., Rappaport,E.F., Shaikh,T.H. and White,P.S. (2010) CNV Workshop: an integrated platform for high-throughput copy number variation discovery and clinical diagnostics. *BMC Bioinformatics*, **11**, 74.
13. Klambauer,G., Schwarzbauer,K., Mayr,A., Clevert,D.A., Mitterecker,A., Bodenhofer,U. and Hochreiter,S. (2012) cn.MOPS: Mixture of Poissons for discovering copy number variations in next-generation sequencing data with a low false discovery rate. *Nucleic Acids Res.*, **40**, e69.
14. Knijnenburg,T.A., Ramsey,S.A., Berman,B.P., Kennedy,K.A., Smit,A.F.A., Wessels,L.F.A., Laird,P.W., Aderem,A. and Shmulevich,I. (2014) Multiscale representation of genomic signals. *Nat. Methods*, **11**, 689–694.
15. Anders,S., Pyl,P.T. and Huber,W. (2014) HTSeq–a Python framework to work with high-throughput sequencing data.*Bioinformatics*, **31**, 166–169.
16. Harrow,J., Frankish,A., Gonzalez,J.M., Tapanari,E., Diekhans,M., Kokocinski,F., Aken,B.L., Barrell,D., Zadissa,A., Searle,S. *et al.* (2012) GENCODE: The reference human genome annotation for the ENCODE project. *Genome Res.*, **22**, 1760–1774.
17. Lahti,L., Myllykangas,S., Knuutila,S. and Kaski,S. (2009) Dependency detection with similarity constraints. In: *2009 IEEE International Workshop on Machine Learning for Signal Processing*. IEEE, pp. 1–6.
18. Schäfer,M., Schwender,H., Merk,S., Haferlach,C., Ickstadt,K. and Dugas,M. (2009) Integrated analysis of copy number alterations and gene expression: a bivariate assessment of equally directed abnormalities. *Bioinformatics*, **25**, 3228–3235.
19. Thorvaldsdóttir,H., Robinson,J.T. and Mesirov,J.P. (2013) Integrative Genomics Viewer (IGV): High-performance genomics data visualization and exploration. *Brief. Bioinform.*, **14**, 178–192.
20. Koboldt,D.C., Fulton,R.S., McLellan,M.D., Schmidt,H., Kalicki-Veizer,J., McMichael,J.F., Fulton,L.L., Dooling,D.J., Ding,L., Mardis,E.R. *et al.* (2012) Comprehensive molecular portraits of human breast tumours. *Nature*, **490**, 61–70.
21. Huang,N., Shah,P.K. and Li,C. (2012) Lessons from a decade of integrating cancer copy number alterations with gene expression profiles. *Brief. Bioinform.*, **13**, 305–316.
22. Forbes,S.A., Bindal,N., Bamford,S., Cole,C., Kok,C.Y., Beare,D., Jia,M., Shepherd,R., Leung,K., Menzies,A. *et al.* (2011) COSMIC: Mining complete cancer genomes in the catalogue of somatic mutations in cancer. *Nucleic Acids Res.*, **39**, 945–950.
23. Gonzalez-Perez,A., Perez-Llamas,C., Deu-Pons,J., Tamborero,D., Schroeder,M.P., Jene-Sanz,A., Santos,A. and Lopez-Bigas,N. (2013) IntOGen-mutations identifies cancer drivers across tumor types. *Nat. Methods*, **10**, 1081–1082.
24. Van den Eynden,J., Fierro,A.C., Verbeke,L.P. and Marchal,K. (2015) SomInaClust: Detection of cancer genes based on somatic mutation patterns of inactivation and clustering. *BMC Bioinformatics*, **16**, 125.
25. Zhang,L., Zhou,F., García de Vinuesa,A., de Kruijf,E.M., Mesker,W.E., Hui,L., Drabsch,Y., Li,Y., Bauer,A., Rousseau,A. *et al.* (2013) TRAF4 promotes TGF-β receptor signaling and drives breast cancer metastasis. *Mol. Cell*, **51**, 559–572.
26. Bièche,I., Tomasetto,C., Régnier,C.H., Moog-Lutz,C., Rio,M.C. and Lidereau,R. (1996) Two distinct amplified regions at 17q11-q21 involved in human primary breast cancer. *Cancer Res.*, **56**, 3886–3890.
27. Possemato,R., Marks,K.M., Shaul,Y.D., Pacold,M.E., Kim,D., Birsoy,K., Sethumadhavan,S., Woo,H.-K., Jang,H.G., Jha,A.K. *et al.* (2011) Functional genomics reveal that the serine synthesis pathway is essential in breast cancer. *Nature*, **476**, 346–350.
28. Letessier,A., Ginestier,C., Charafe-Jauffret,E., Cervera,N., Adélaïde,J., Gelsi-Boyer,V., Ahomadegbe,J.-C., Benard,J., Jacquemier,J., Birnbaum,D. *et al.* (2005) ETV6 gene rearrangements in invasive breast carcinoma. *Genes. Chromosomes Cancer*, **44**, 103–108.
29. Li,Z., Tognon,C.E., Godinho,F.J., Yasaitis,L., Hock,H., Herschkowitz,J.I., Lannon,C.L., Cho,E., Kim,S.J., Bronson,R.T. *et al.* (2007) ETV6-NTRK3 fusion oncogene initiates breast cancer from committed mammary progenitors via activation of AP1 complex. *Cancer Cell*, **12**, 542–558.
30. De Braekeleer,E., Douet-Guilbert,N., Morel,F., Le Bris,M.J., Basinko,A. and De Braekeleer,M. (2012) ETV6 fusion genes in hematological malignancies: A review. *Leuk. Res.*, **36**, 945–961.
31. Lin,M.-L., Park,J.-H., Nishidate,T., Nakamura,Y. and Katagiri,T. (2007) Involvement of maternal embryonic leucine zipper kinase (MELK) in mammary carcinogenesis through interaction with Bcl-G, a pro-apoptotic member of the Bcl-2 family. *Breast Cancer Res.*, **9**, R17.
32. Giam,M., Okamoto,T., Mintern,J.D., Strasser,A. and Bouillet,P. (2012) Bcl-2 family member Bcl-G is not a proapoptotic protein. *Cell Death Dis.*, **3**, e404.
33. Cunningham,F., Amode,M.R., Barrell,D., Beal,K., Billis,K., Brent,S., Carvalho-Silva,D., Clapham,P., Coates,G., Fitzgerald,S. *et al.* (2014) Ensembl 2015. *Nucleic Acids Res.*, **43**, D662–D669.
34. Gasteiger,E., Gattiker,A., Hoogland,C., Ivanyi,I., Appel,R.D. and Bairoch,A. (2003) ExPASy: The proteomics server for in-depth protein knowledge and analysis. *Nucleic Acids Res.*, **31**, 3784–3788.
35. Finn,R.D., Bateman,A., Clements,J., Coggill,P., Eberhardt,R.Y., Eddy,S.R., Heger,A., Hetherington,K., Holm,L., Mistry,J. *et al.* (2014) Pfam: The protein families database. *Nucleic Acids Res.*, **42**, D222–D230.
36. Consortium,T.E.P., Dunham,I., Kundaje,A., Aldred,S.F., Collins,P.J., Davis,C. a, Doyle,F., Epstein,C.B., Frietze,S., Harrow,J. *et al.* (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57–74.
37. Tseng,Y.-Y., Moriarity,B.S., Gong,W., Akiyama,R., Tiwari,A., Kawakami,H., Ronning,P., Reuland,B., Guenther,K., Beadnell,T.C. *et al.* (2014) PVT1 dependence in cancer with MYC copy-number increase. *Nature*, **512**, 82–86.