Cancer Informatics

Libertas Academica
FREEDOM TO RESEARCH

ORIGINAL RESEARCH

Evaluating Translocation Gene Fusions by SNP Array Data

Hong Liu[1], Asher Zilberstein[2], Pascal Pannier[3], Frederic Fleche[3], Christopher Arendt[4], Christoph Lengauer[3] and Chang S. Hahn[5]

[1]Lead Generation to Candidate Realization, Sanofi, Route 202-206, Bridgewater, NJ 08807 USA. [2]Oncology, Sanofi, Route 202-206, Bridgewater, NJ 08807 USA. [3]Oncology, Sanofi, 13 quai Jules Guesde, Vitry-Sur-Seine, Vitry 94403 Cedex, France. [4]Immuno-Inflammation TSU, Sanofi, Route 202-206, Bridgewater, NJ 08807 USA. [5]Fibrosis and Wound Repair TSU, Sanofi, Route 202-206, Bridgewater, NJ 08807 USA.
Corresponding author email: hong.liu@sanofi-aventis.com

Abstract: Somatic cell genetic alterations are a hallmark of tumor development and progression. Although various technologies have been developed and utilized to identify genetic aberrations, identifying genetic translocations at the chromosomal level is still a challenging task. High density SNP microarrays are useful to measure DNA copy number variation (CNV) across the genome. Utilizing SNP array data of cancer cell lines and patient samples, we evaluated the CNV and copy number breakpoints for several known fusion genes implicated in tumorigenesis. This analysis demonstrated the potential utility of SNP array data for the prediction of genetic aberrations via translocations based on identifying copy number breakpoints within the target genes. Genome-wide analysis was also performed to identify genes harboring copy number breakpoints across 820 cancer cell lines. Candidate oncogenes were identified that are linked to potential translocations in specific cancer cell lines.

Keywords: copy number variation, copy number breakpoint, SNP array, translocation

Cancer Informatics 2012:11 15–27

doi: 10.4137/CIN.S8026

This article is available from http://www.la-press.com.

Background

One of the major goals in cancer research is to identify causal genetic aberrations. This has led to the identification of several successful therapeutic targets, including BCR-ABL fusion, EGFR amplification/mutation, and HER2 amplification. Among different types of genetic aberrations, chromosomal rearrangements creating oncogenic gene fusions are the hallmark of many haematopoietic malignancies as well as rare bone and soft-tissue tumors.[1,2] Recent reports suggest that many solid tumors also contain gene fusions that confer tumorigenic potential. Multiple methods were developed for identifying gene fusions caused by chromosomal translocations. Traditional methods involve cytogenetic analysis followed by fluorescent in situ hybridization analysis. Recent high-throughput data platforms also provide opportunities to discover novel fusions. For example, expression data based analysis such as Cancer Outlier Profile Analysis (COPA)[3] identified novel fusions, including TMPRSS2-ERG and TMPRSS2-ETV1 in prostate cancer.[4] Deep sequencing of cDNA libraries led to the discovery of the EML4-ALK fusion in non-small-cell lung cancer (NSCLC),[1] and integrative analysis of high-throughput long- and short-read transcriptome sequencing identified several gene fusions in prostate cancer cell lines.[5] However, since these methods are based on information from RNA transcripts, the fusion identified might be due to alternative splicing, or transcript level re-arrangement. Identifying translocations at the chromosomal level remains a challenging task.

Chromosomal translocations, by definition, alter genomic sequences, and may generate fusion proteins or dysregulate gene expression. Chromosomal translocations elicit DNA repair processes, which involve mis-repair of double-strand ends. Cloning genomic junctions in various chromosome translocations in leukemia shows that there are deletions, duplications, and insertions at the breakpoints in many translocations.[6] The sizes of deletions and duplications range from a few bp to a few hundred bp. Many fusion genes are also reported to have multiple copies. These will result in copy number variation (CNV) between segments retained in the fusion gene and its neighboring genomic sequences. High density SNP arrays are useful tools not only to study SNP-based genetic linkage, but also to detect DNA CNV across the whole genome. The current Affymetrix SNP array 6.0 contains 1.8 million markers for genetic variation, and has a median inter-marker distance of less than 700 bases. During the last few years, extensive efforts have been dedicated to SNP array profiling on tumor samples and cell lines. For example, the Sanger Institute has profiled over 800 cancer cell lines on the Affymetrix SNP array 6.0. We postulate that if a deletion or duplication exists at the breakpoint/junction site of a gene fusion event during tumorigenesis, or if a fusion gene is amplified to have multiple copies, then there is likely to be a copy number breakpoint juxtaposed to the genomic junction of the fusion gene, which can be detected by high-resolution SNP arrays.

We validated this hypothesis by analyzing three well-known genetic fusions in cancer: BCR-ABL1, TMPRSS2-ERG, and EML4-ALK using high density SNP array data from related patient samples and cancer cell lines. We examined whether there is a copy number breakpoint near or at the junction of each fusion gene. Two aspects were evaluated: (i) whether a deletion or amplification in copy number could be detected, and (ii) the distance relative to the specific fusion junction. Based on the information from the analysis of the validation set, we developed a search tool to identify additional cancer cell lines that may contain interesting fusion genes, by examining the SNP array data of Sanger's 820 cancer cell lines. Finally, we evaluated the distribution of gene-linked copy number breakpoints in cancer cell lines. Based on these efforts, we provide a framework for evaluating and possibly detecting translocation mutations in different cancers utilizing SNP array data.

Results

Analyzing known gene fusions utilizing SNP array data

Case 1: BCR-ABL

Examination of CML cell lines with reported Philadelphia chromosome translocations

The BCR-ABL gene fusion, also referred to as the Philadelphia chromosome or Philadelphia translocation, is the best known chromosomal abnormality resulting from a reciprocal translocation between chromosome 9 and 22. The fusion contains 5′end sequences from BCR and 3′end sequences from ABL1, which contains the kinase domain. The chimeric BCR-ABL protein has constitutively elevated tyrosine phosphokinase activity. This abnormal enzymatic activation is critical to the oncogenic potential of BCR-ABL.[7] It is found in 95% of the subjects with CML, in 25%–30% of ALL patients, and occasionally in AML patients.[7]

Among the 820 cell lines that the Sanger Institute profiled, there are eight cell lines that are of CML origin with reported Philadelphia chromosomes: MEG01, K562, BV173, EM2, LAMA84, NALM1, CMLT-1, and KU812. We analyzed the SNP array data of these cell lines and observed that five cell lines (MEG01, K562, BV173, EM2, and LAMA84) contain copy number breakpoints in both the BCR and ABL1 genes (Fig. 1). There are three known breakpoint cluster regions in BCR. The majority of breakpoints in CML patients have been reported to occur between BCR exons 12 to 16. The second breakpoint cluster region, mainly in ALL, lies within an intron between BCR exons 1 and 2, while the third one is located downstream of BCR exon 19.[8] In the five CML cell lines in which copy number breakpoints were observed, the breakpoints in BCR all reside between exons 12 to 16, while the breakpoints in ABL1 vary somewhat, but the majority reside within the first intron. As for the CNV, in BCR, the copy number at its 3´end is lower than that of its 5´end, while in ABL1, the copy number at its 3´end is always higher than that of its 5´end, thus favoring presence of BCR-ABL fusion protein. As for the other three CML cell lines (NALM1, CMLT-1, and KU812), breakpoints were detected in either BCR, ABL1, or in the neighboring regions. For example, in CMLT-1 cells, a copy number breakpoint was found in ABL1, while in KU812 a micro-deletion was found in ABL1. Of note, CNV were found in the neighbor genes of BCR in all three cell lines. In conclusion, although the number of examples is limited, the copy numbers for sequences that retain in the fusion gene are either remained as normal or increased, but are never decreased; while the copy numbers for sequences that are not in the fusion gene are either lost or remain as normal, but are never increased.

In order to evaluate CNV in the coding regions, especially in the oncogenes, we searched the copy

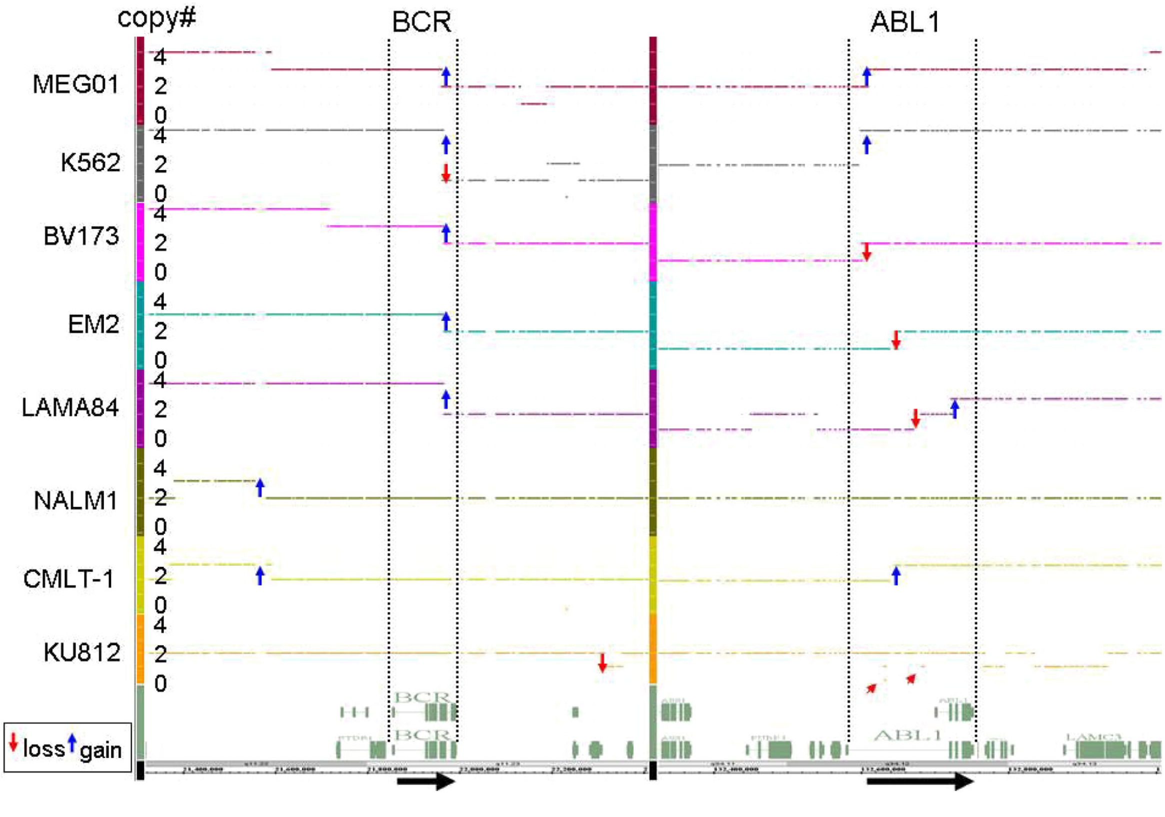

Figure 1. Genomic level CNV analysis of BCR and ABL1 genes using Affymetrix SNP 6.0 array data for eight CML cell lines. Copy number states are divided into the following categories: 0-homozygous deletion; 1-heterozygous deletion; 2-normal diploid; 3-single copy gain; and 4-multiple copy gain. Arrows highlight both amplified (blue) and deleted (red) genomic segments.
Notes: Black arrow indicates the direction of the transcript. Affymetrix Genotyping Console software was used for this analysis.

number breakpoint resides within the transcript of the oncogene in these eight CML cell lines. As shown in supplemental Table 1, ABL1 contains copy number breakpoints in six out of the eight CML cell lines. The frequency (75%) is the highest compared to other oncogenes.

Searching for other cancer cell types with potential BCR-ABL1 fusions

We asked whether copy number breakpoints occur within BCR and ABL1 in cancer cells of non-hematopoietic origins. By analyzing the Sanger 820 cell line panel, we identified two cell lines that contain breakpoints in both BCR and ABL1 (Fig. 2). The breakpoint within BCR in NCI-H747, a colon cancer cell line, was similar to that of the other CML cell lines, residing between exons 12 to 16. However, in NCI-H1581, a lung cancer cell line, the breakpoint was between exons 1 and 2, similar to that of ALL. In NCI-H747, the CNV was consistent with what was observed in the other CML cell lines: namely, the copy number at 3′end is lower than that of 5′end for BCR, and the copy number at 3′end is higher than that of 5′end for ABL1. However, in NCI-H1581, the CNV in ABL1 is different, with the 3′end (containing the kinase domain) at lower copy number relative to the 5′end. Whether these cell lines contain a functional BCR-ABL fusion is yet to be determined.

To evaluate the random chance to identify copy number breakpoints in both BCR and ABL1 genes in a given cell line, simulations were performed by assigning the information of all amplification or loss segments, such as the segment length, copy number status (gain or loss) from either NCI-H1581 or NCI-H747, on a hypothetic genome. For each cell line, the simulation was performed 100,000 times, and no breakpoint was found in both BCR and ABL1 in a single case. This suggests that the chance to observe copy number breakpoints in both BCR and ABL1 in the two cell lines is unlikely to occur by chance alone.

Case 2: TMPRSS2-ERG

Examination of prostate cancer samples and cell lines

Recent findings, through a bioinformatics approach (COPA),[3] revealed that prostate cancers frequently over-express the ETS family transcription factors ERG and ETV1 as a result of chromosomal rearrangements that lead to the fusion of the 5′end of the androgen-regulated serine protease TMPRSS2 (21q22.2) to the 3′end of either ERG (21q22.3) or ETV1 (7p21.3).[4] The consequence is the aberrant androgen receptor-driven expression of the potential oncogenes ERG or ETV1. The TMPRSS2-ERG fusion is present at high frequency in moderate to poorly differentiated prostate cancers (35/86, 40.7%),[9] in contrast to TMPRSS2-ETV1 which is the product of a rare fusion event.

We evaluated whether SNP data could be reveal insights about the mechanisms involved in these fusion events in prostate cancer. To this end, we analyzed the SNP array data of 20 paired prostate cancer

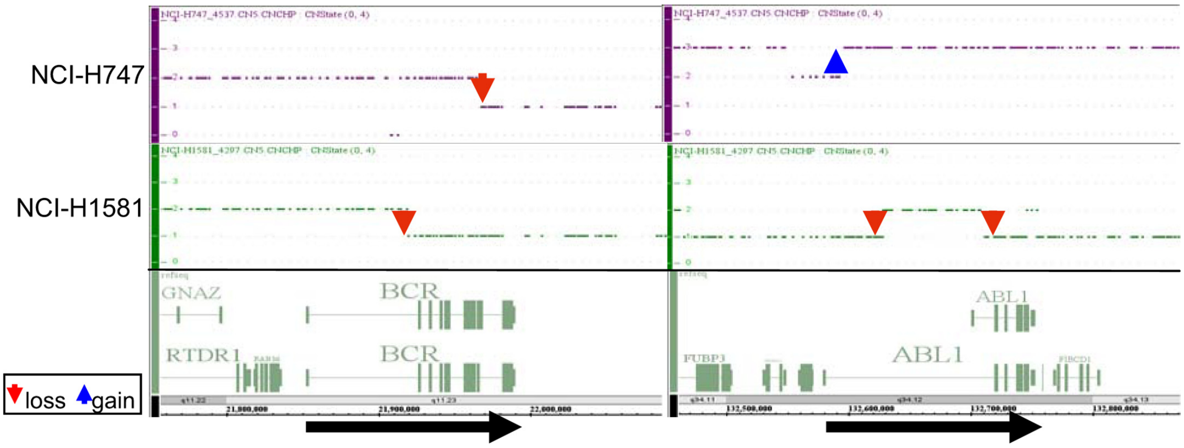

Figure 2. Affymetrix Genotyping Console Browser view of BCR and ABL1 genes for two non-CML cell lines that possess copy number breakpoints in both BCR and ABL1 genes. Copy number states are divided into the following categories: 0-homozygous deletion; 1-heterozygous deletion; 2-normal diploid; 3-single copy gain; and 4-multiple copy gain.

Notes: Arrows highlight both amplified (blue) and deleted (red) genomic segments. Black arrows indicate the direction of the transcript.

samples with matched normal samples from GEO. Both TMPRSS2 and ERG are on the same chromosome, chromosome 21, and reside on the negative DNA strand with an invervening distance of 2.7 Mb. As shown in Figure 3, three out of 20 samples (15%) contain a segment deletion between TMPRSS2 and ERG. Although the positions of the breakpoints at both TMPRSS2 and ERG are different in the three samples, the final fusions retain the coding sequence of ERG linked to the 5' regulatory sequences of TMPRSS2. In addition, we found that ETV1 was amplified in three of the 10 prostate samples (data not shown).

In order to evaluate CNV in the coding regions, especially in the oncogenes, we searched the copy number breakpoint resides within the transcript of the oncogene in these 20 prostate cancer samples. As shown in supplemental Table 2, ERG contains copy number breakpoints in three out of the 20 prostate cancer samples. The frequency (15%) is the highest compared to other oncogenes.

The TMPRSS2-ERG fusion is known to exist in the prostate cell line VCaP.[5] However, we were unable to locate publicly available SNP array data for this cell line. Instead, we investigated prostate cell lines profiled by Sanger Institute. Among 820 cancer cell lines, five prostate cancer cell lines are represented: 22RV, BPH-1, DU-145, LNCaP, PC-3. None of these lines possess segmental deletions between TMPRSS2 and ERG, nor is the copy number of ETV1 amplified (data not shown).

Searching for other cancer cell types with the TMPRSS2-ERG fusion

In order to assess whether the TMPRSS2-ERG fusion is unique to prostate cancers, a search was performed to identify other cell line containing segmental deletions between ERG and TMPRSS2. Specifically, we queried for a deletion segment with one end anchored either within TMPRSS2 or between TMPRSS2 and its neighbor RIPK4, and the other end anchored either within the 5´end of ERG or between ERG and its neighbor gene ETS2. This type of deletion has the potential to create a fusion gene that utilizes the promoter of TMPRSS2 and links to the coding of ERG. However, among the 820 cell lines that the

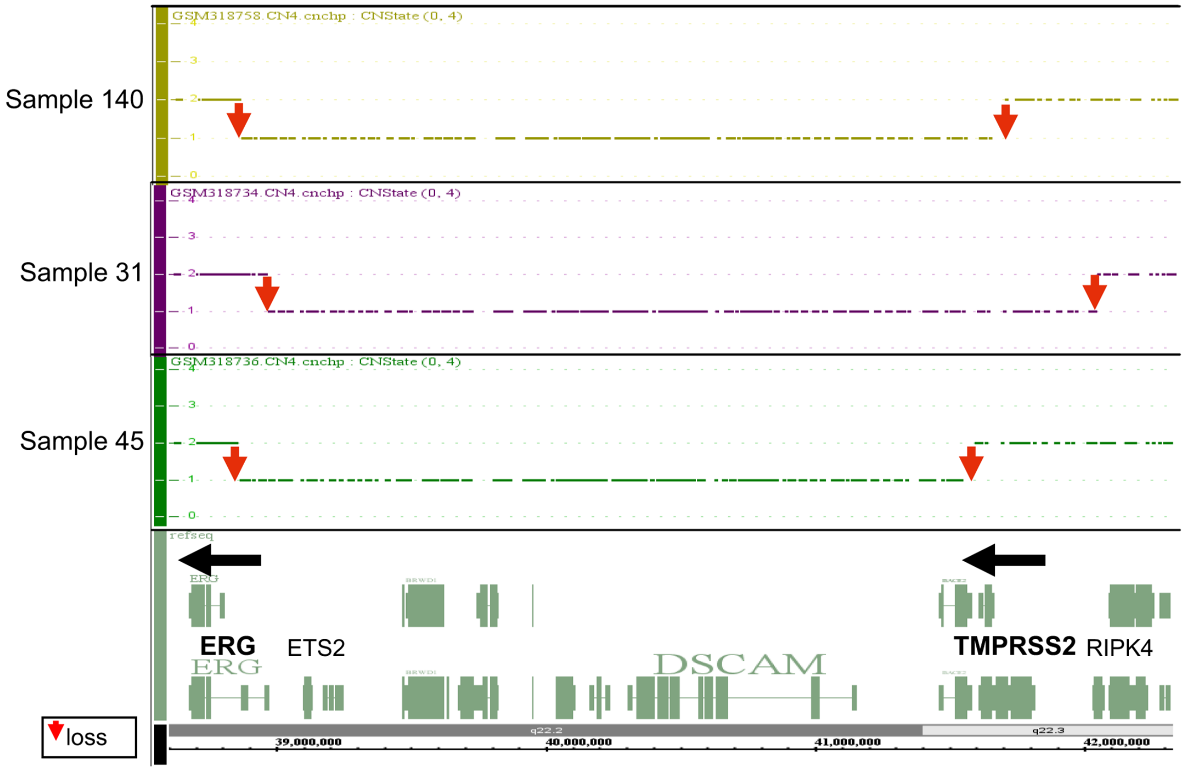

Figure 3. Affymetrix Genotyping Console Browser view of the segment deletion between TMPRSS2 and ERG in prostate cancer samples. Copy number states are divided into the following categories: 0-homozygous deletion; 1-heterozygous deletion; 2-normal diploid; 3-single copy gain; and 4-multiple copy gain.
Notes: Red arrows demarcate deleted genomic segments, while black arrows designate the directions of the two transcripts.

Sanger Institute profiled, none contains such deletion. This is suggestive that the TMPRSS2-ERG fusion is restricted to specific subtypes of prostate cancers.

Case 3: EML4-ALK

Examination of lung cancer cell lines

The EML4-ALK fusion was identified in a NSCLC sample by full-length cDNA cloning. It was also detected in other lung cancers with a frequency of 9.1% (3 out of 33).[1] To further evaluate the utility of SNP array data for identifying potential translocations, both EML4 and ALK were examined for potential copy number breakpoints in Sanger's 140 lung cancer cell lines. Among these 140 lung cancer cell lines, six (4%) were found to carry breakpoints for both EML4 and ALK. As shown in Figure 4, most of the copy number breakpoints in EML4 are caused by the amplification of the 5′end of the gene, while breakpoints in ALK are closer to its 3′end. This is consistent with the junctional sequence described for EML4-ALK, in which the 5′end sequence of EML4 is linked to the 3′end sequence of ALK, where the kinase domain resides. A literature search revealed that, among these six lung cancer cell lines, NCIH2228 is reported to contain EML4-ALK fusion that links EML4 exon 6 to ALK exon 20.[10]

In order to evaluate CNV in the coding regions, especially in the oncogenes, we searched the copy number breakpoint resides within the transcript of the oncogene in these 140 lung cancer cell lines. As shown in supplemental Table 3, ALK contains copy number breakpoints in 22 out of the 140 lung cancer cell lines. Its frequency (15.7%) is among the top ranked but not the highest. Among the ten oncogenes that have higher copy number breakpoint frequencies than that of ALK, five of them: PVT1 (t(2;8) and t(8;22) in Burkitt lymphoma)[11] AKT3 (t(1;13)(q44;q32) in microcephaly and agenesis of the corpus callosum),[12] VAV2 (t(1;9)(p36.32;q34.2) in bilateral exotropia, left ptosis)[13] ABL2 (t(1;12)(q25;p13) in AML),[14] and NTRK3 (t(12;15)(p13;q25) in salivary gland tumors and AML),[15,16] were reported to be involved in reciprocal translocation. It will be of interest to validate whether these translocations are also present in the lung cancer cell lines.

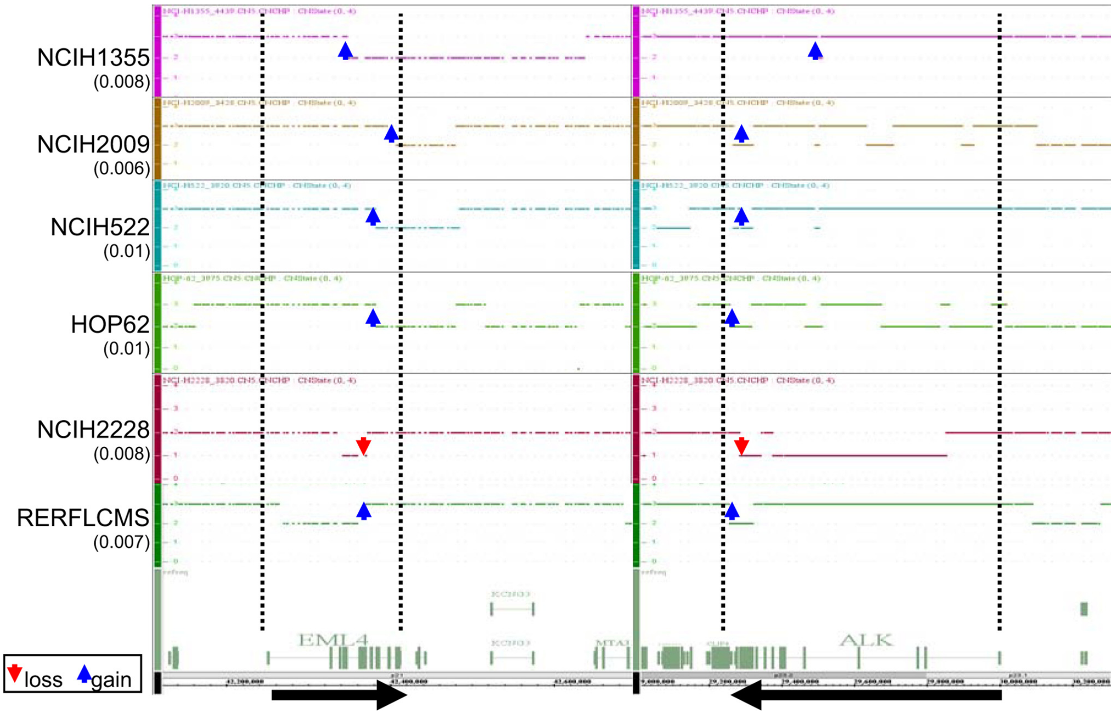

Figure 4. Affymetrix Genotyping Console Browser view of EML4 and ALK genes in six lung cancer cell lines that contain copy number breakpoints in both genes. Copy number states are divided into the following categories: 0-homozygous deletion; 1-heterozygous deletion; 2-normal diploid; 3-single copy gain; and 4-multiple copy gain.

Notes: Arrows highlight both amplified (blue) and deleted (red) genomic segments. Black arrows indicate the directions of the transcripts. The numbers in brackets indicate the *P* values.

Searching for additional cancer cell lines with EML4-ALK fusions

We further searched for cell lines carrying breakpoints for both EML4 and ALK by surveying the entire 820 Sanger cell lines, and identify an additional 18 cell lines that carry copy number breakpoints similar to those of the six lung cell lines. As shown in Table 1, these cell lines are from different cancer origins, including breast, skin, stomach, and colon. Calculated *P* values suggest that it is highly unlikely that copy number breakpoints within both EML4 and ALK in there cell lines occur by random chance. This is in agreement with a report indicating that the EML4-ALK fusion is not only found in lung cancer samples, but also in other cancers such as breast and colorectal.[17] It is worth noting that this fusion sequence was detected by RT-PCR assay in the colon cell line SW1417.[17]

Identifying oncogenes with copy number breakpoints

A proto-oncogene can become an oncogene as a consequence of a relatively small modification such as mutations or increased expression. Chromosomal rearragement can lead to the increased gene expression, or the expression of a constitutively active hybrid protein[18] In order to identify cell lines carrying similar breakpoint as ABL1, we searched the CNV in the oncogenes that were defined by Affymetrix chip annotation across Sanger's 820 cancer cell lines according to the following criteria: (1) The copy number breakpoint resides within the transcript of the oncogene; and (2) the copy number for the coding sequence is either normal (n = 2) or amplified, but not deleted. Among the 205 oncogenes annotated by Affymetrix, 26% (54) oncogenes contain copy number breakpoints in at least one cancer cell line (Supplemental Table 4). This is substantially higher (*P* value = 0.001) than the genome-wide search, where we found that 17% (2,945 of out total 17,609) of the genes contain copy number breakpoint in at least one cancer cell line.

As a comparison, we further evaluated the copy number breakpoints of oncogenes and all genes in normal cell lines, which were converted from the blood samples of healthy donors collected by the International HapMap project. Based on the limited 38 HapMap normal cell line data retrieved from GEO, none of the oncogenes contains copy number breakpoints, while only 0.1% (19 out of total 17,609) of the genes in the genome contain CNV in at least one normal cell line. There is no significant difference between oncogenes versus all genes (*P* value = 0.638) with regard to the frequency of copy number breakpoints in these normal cell lines.

Table 1. Additional Sanger's non-lung cancer cell lines that contain breakpoints in both EML4 and ALK genes.

Cell line	Primary tissue	*P* value
Saos-2	bone	0.002
MDA-MB-468	breast	0.017
UACC-893	breast	0.028
COLO-824	breast	0.004
KNS-42	brain	0.053
DoTc2-4510	cervix	0.042
DEL	haematopoietic	0.011
LAMA-84	haematopoietic	0.03
SW403	colon	0.006
SW1417	colon	0.016
OVCAR-5	ovary	0.011
PANC-08-13	pancreas	0.013
A101D	skin	0.009
CHL-1	skin	0.006
SK-MEL-5	skin	0.004
SCH	stomach	0.008
MKN28	stomach	0.002
639-V	urinary	0.009

Note: *P* value indicates the random chance for the cell line to contain breakpoints in both EML4 and ALK.

Genome-Wide evaluation of gene-linked copy number breakpoints in cancer cell lines

We sought to go beyond our initial analyses of oncogenes and broadly evaluate copy number breakpoints across all genes for all 820 of the Sanger cancer cell lines. A gene was considered to contain a copy number breakpoint if the CNV was present within its transcript, regardless of the location or whether it was associated with amplification or deletion. The results of this analysis reveal that the mean value of the number of genes with copy number breakpoints in Sanger's cancer lines is 369, which is much higher than the mean of 25 obtained for HapMap normal cell lines ($P <= 4.054e^{-10}$). When the numbers of genes containing copy number breakpoints were plotted against cancers of various tissue origins, a significant variability was observed amongst cancer origins. On one end of the spectrum, liver cancers and mesotheliomas possess medians of greater than 600 genes containing copy number breakpoints. In contrast, hematopoietic cancers exhibited a median of 70 genes containing copy number breakpoints (Fig. 5).

Discussion

In this study, we have utilized SNP array data and evaluated three well known fusion genes that are associated with tumorigenesis for the presence of copy number breakpoints within genomic sequences.

For BCR-ABL1, we observed breakpoints in both BCR and ABL1 in 5 out of 8 CML cell lines. This suggests the potential limitation of detecting translocations by using SNP array data, which may be attributed to a number of factors. First, the SNP array data can only detect gene fusions with underlying chromosomal changes, and not copy neutral alterations, such as balanced translocations with no gain or loss of genetic information. Second, the lack of deletion/duplication signatures at junctional sites in certain cell lines may due to the resolution of SNP array data if the deletion/duplication segment is relatively small. It should be noted that for the other three CML cell lines, breakpoints were present in either BCR, ABL1, or in the neighboring genes. Additional searches identified two non-leukemia cancer cell lines that also contain copy number breakpoints in BCR and ABL1. It will be of interest to confirm whether these two cell lines do contain BCR-ABL1 fusions since BCR-ABL1 has only been reported in leukemias to date.

For TMPRSS2-ERG, based on the 20 prostate patient sample sets, the SNP array data reveal that some of the fusions are likely results of genomic sequence deletion. As for other Ets family members, such as ETV1, the altered expression may be due to different genetic mutations, such as gene amplification. Alternatively, the frequency of TMPRSS2-ETV1 translocation fusions may be much lower than for TMPRSS2-ERG. In this limited data set, the samples that contain amplified ETV1 and fusion ERG were mutually exclusive, which implies that

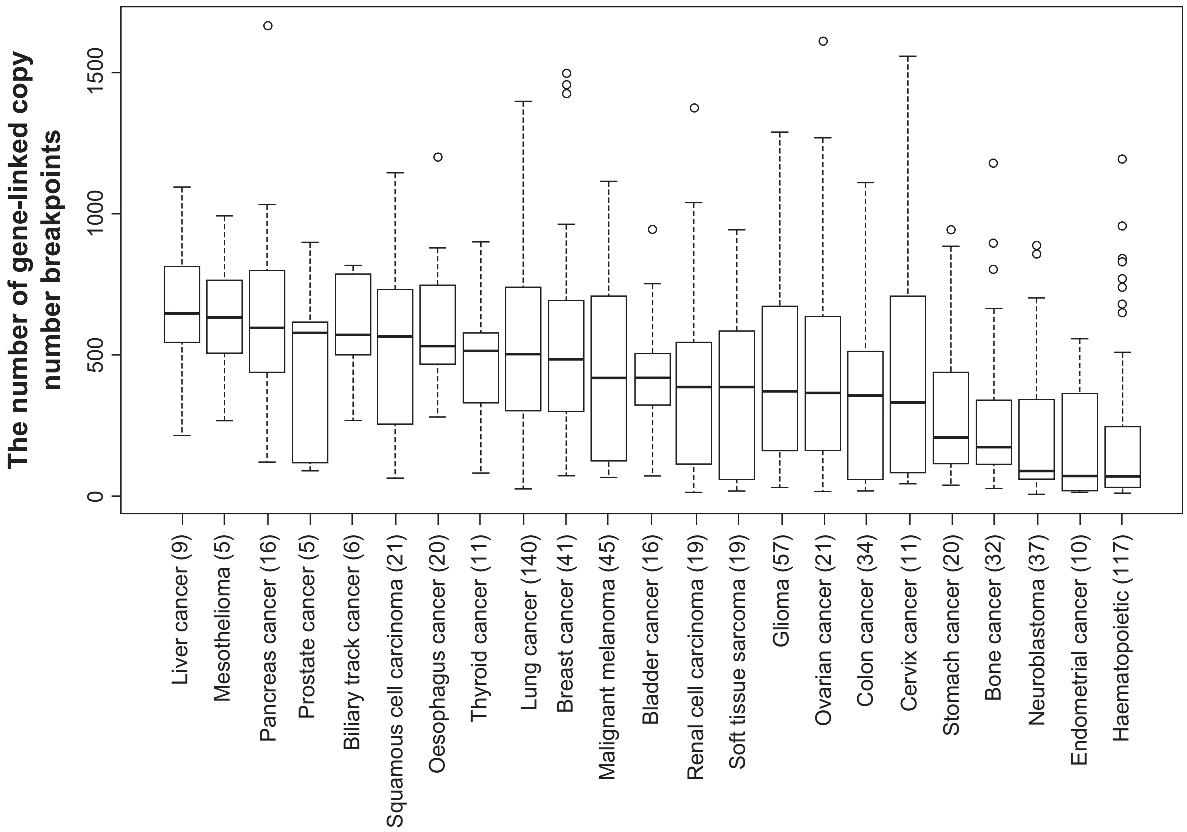

Figure 5. Boxplot view of the numbers of gene-linked copy number breakpoint in 820 cancer cell lines from different cancer origins.
Notes: The numbers in brackets indicate the number of cell lines mapped to each cancer type. Cancer types with less than 5 cell lines are not included. Y axis: the number of gene-linked copy number breakpoints. Median: bolded line inside the box; 25 percentile: top line of the box; 75 percentile: bottom line of the box; Maximum (excluding outliers): top bar outside of the box; Minimum (excluding outliers): lowest bar outside of the box; Outlier: O.

the over-expression of one of the Ets genes may contribute to initiation or progression of prostate cancer. Genome wide analysis indicates that none of the 820 Sanger cancer cell lines contain a segment deletion between TMPRSS2 and ERG similar to what we observed in the primary prostate cancer samples. There are two possibilities: (1) with only five prostate cancer cell lines in the panel, the number is too small to represent the wide genetic diversity of prostate cancers; (2) TMPRSS2-ERG is unique to prostate cancer, since TMPRSS2 expression is almost exclusively in the prostate with some expression in the GI track. In addition, the fusion is regulated by androgen as TMPRSS2 expression is androgen dependent[19,20] which provides it a growth advantage only in prostate and not in other tissues. However, we can not rule out other fusion transcript mechanisms, such as transcript read-through, which can not be captured by SNP array data at the DNA level. This is all the more possible given that TMPRSS2 and ERG are closely located on the same chromosome, and that read-through fusion transcripts were identified in prostate cancer.[5]

For the EML4-ALK fusion, we have identified 6 lung cancer cell lines, and 18 cell lines from other cancer origins with copy number breakpoints. This is in agreement with a report by Lin et al, where a RT-PCR assay was used to examine a panel of 124 cell lines from breast cancer, colorectal cancer, and NSCLC for fusion transcripts of EML4-ALK. With this panel, 9 cell lines, including a known positive control (H2228), were identified to harbor the EML4-ALK fusion. Since the detailed information of the 124 cell lines was not available in the publication, we were not able to compare them with our copy number analysis. However, based on some of the positive and negative cell lines listed in the paper, we were able to evaluate nine cell lines that were common. Among them, two cell lines (H2228 and SW1417) are positive, and five cell lines (T47D, CAL120, HCT116, H1299, and H838) are negative supported by both methods. There are two cell lines (H460 and H1975), which were found to contain fusion transcripts, but do not show copy number breakpoints by SNP array data. This again suggests the potential limitation of detecting translocations by using SNP array data as discussed above.

In the above evaluations, we also assessed all copy number breakpoints in the coding regions, especially in the oncogenes. The frequencies of the copy number breakpoint for ABL1 (75%), ERG (15%) and ALK (15.7%) are high in the eight CML cell lines, 20 prostate cancer samples, and 140 lung cancer cell lines, respectively. However, the frequency for ERG may be underestimated since the SNP array data of the prostate cancer samples were generated from Affymetrix 500 K array Set, whose probe coverage is less than one third of that of the Affymetrix array 6.0 used for the Sanger panel. In the CML cell lines and prostate cancer samples, the frequencies of copy number breakpoint for ABL1 and ERG, respectively, were ranked top compared to other oncogenes. In the lung cancer cell lines, there are several oncogenes whose copy number breakpoint frequencies are higher than that of ALK. Some of them were reported to be involved in reciprocal translocation in other cancers or developmental diseases. Since the overall frequency of EML4-ALK fusion is low (9.1%) in lung cancers, other mechanisms including other fusions, may possibly be involved in these lung cancer cell lines. It will be of interest to further experimentally validate the potential translocation candidates, which will help us to discover novel fusion and provide a better understanding of the false positive rate. For the potential utility of applying SNP array data to detect novel translocation fusion, we suggest that the focus should be on the oncogenes with high frequency of copy number breakpoints in the studied samples. The prediction can be powerful if the fusion gene is present at high frequency in the studied samples. In conclusion, analyzing SNP array data for copy number breakpoint frequencies of oncogenes could provide us a potential translocation candidate list to be further followed up experimentally.

Genome-wide copy number breakpoint analysis was also performed in the 820 Sanger cell lines. It had identified genes that contain copy number breakpoints in Sanger cancer cell lines, but not in the 38 Hapmap normal cell lines. We evaluated the top ten genes that are highly linked with copy number breakpoints only in the cancer cell lines (Supplemental Table 5). The genomic sizes of these genes tend to be very large, which may increase the chance for them to link with a copy number breakpoint. Nevertheless, it is interesting to note that five out of the ten genes: MACROD2[21] FHIT,[22] CNTNAP2,[23] MAGI2,[24] LRP1B[25] were reported that were involved in genomic sequence re-arrangement/translocation.

Many cancer cell lines have extensive in vitro passage histories and therefore may accumulate additional mutations associated with extended laboratory propagation. Cancer cell lines, many of which have defective DNA repair machinery, are particularly susceptible to genetic alterations that confer growth or survival advantages under conventional tissue culture conditions. In order to evaluate whether cancer lines contain greater frequencies of translocation mutations, we assessed the gene-linked copy number breakpoints in primary cancer samples. First, we examined the number of genes containing copy number breakpoints in the 20 paired primary prostate cancer samples. The average value of 74 was significantly lower (t-test with P value $< = 0.0001$) than the average value (= 462) of the five prostate cancer cell lines in the Sanger set. However, the number for the 20 primary prostate cancer samples may be underestimated due to the low density of SNP array used. Next, we examined a public data set from primary gastrointestinal stromal tumors (GIST) samples that were profiled on the same one million SNP array. In the 25 GIST samples, the average number for the gene linked copy number breakpoint is 239, and it is lower but not significantly different from the average value (= 310) of the 20 stomach cancer cell lines in Sanger's set ($P > 0.01$). In conclusion, based on our initial analysis, there is no significant difference of gene linked copy number breakpoints between primary cancer samples and related cancer cell line.

It was suggested that common chromatin structures at breakpoint cluster regions may lead to chromosomal translocations found in chronic and acute leukemias.[26] These chromatin structural elements include topo II and DNase I cleavage sites, scaffold attachment regions (SARs) and lowest free energy level sites. Such elements have previously been shown to co-localize with genomic breakpoints of chromosome translocation in leukemia, suggesting their potential involvement in non-homologous recombination.[26,27] SARs, topo II and DNase I sites were found to be dispersed throughout the genome every 60–100 kb. It is of interest whether these chromatin structural elements play a role in the progression of other cancer types. Incorporating the genomic position of those sites, together with the copy number breakpoints, may help in better defining the translocation events and delineating the underlying mechanisms.

Recent reports in the *New England Journal of Medicine* demonstrated that the ALK inhibitor crizotinib yielded a stunning overall response rate of 55% and an estimated 6 month, progression-free survival rate of 72% in NSCLC patients.[28,29] Remarkably, it took less than three years from the finding of rearrangement of ALK in NSCLC to translate this into a clinic therapy, with the repositioning of an existing ALK inhibitor in development.[30] This prompts the use of genome-wide analysis to identify new gene mutations, which will not only help us to identify potential novel targets for specific cancers in the context of personalized medicine, but also facilitate the repositioning of drugs either already on the market or in development for new purposes. Next-generation sequencing is likely powerful enough to detect different kinds of gene mutations, including new variant and fusion genes. However, generating the data, as well as performing correct gene mapping and alignment, are not trivial tasks and will take time to be practiced in the patient population. Our analysis reveals the utility of SNP array data that can serve as a complementary data set and assist in the identification of genetic translocations associated with the tumorigenesis process. High density SNP array profiles have been generated extensively in the past few years. Querying GEO pubic database (http://www.ncbi.nlm.nih.gov/geo/) alone for Affymetrix and Illumina 1 million SNP array profiles have identified ~4000 samples, and most of them are cancer related. For The Cancer Genome Atlas project at NCI, the goal is to provide comprehensive genomic characterization and sequencing to the research community on at least 3,000 new cancer cases by the fall of 2011, which includes high density SNP array profiling. The vast amount data provide us a great resource and opportunity to evaluate the known fusions: for their prevalence in a specific cancer type, as well as the presence and prevalence across different cancers. This type of data also has the potential utility to predict novel fusions for oncogenes and to prioritize the fusion candidates to be validated experimentally. Recently shown by Berger et al, integrative analysis of transcriptomic and SNP array data indicates that there are clear changes in DNA copy number, with well defined breakpoints evident within both partner genes in six out of eight melanoma fusions.[31] This further supports the utility of SNP array data in fusion gene identification. In conclusion, thousands of publicly available high-density SNP array datasets have provided us useful information regarding genes with aberrant copy numbers. However, less attention has been paid to those genes containing copy number breakpoints, or at the border of a CNV. Our evaluation indicates that genes with copy number breakpoints are represented at much higher level in cancer versus normal cells. These genes are worthy of further studies aimed at elucidating the functional consequences of these somatic alterations with respect to tumor progression.

Methods

Gene position on the chromosome

For each gene, the transcript start and end positions, as well as the coding start and end positions, were retrieved from the hg18 human genome assembly RefGene table. In the case of genes with multiple splice transcripts, the transcript start position was based on the first transcript start site, and the transcript end position was based on the last transcript end site on the chromosome. Genes were excluded from analysis if any of the following criteria were met: (1) transcripts annotations were on different chromosomes, or on un-assembled segments; (2) genes were on either the X or Y chromosomes; (3) the transcript sequence length was less than 0. Based on the hg18 RefGene table, 19,642 genes meet the above criteria, and among them, 224 genes were annotated as oncogenes based on Affymetrix "HG-U133_Plus_2.na27.annot.txt" annotation table. For genes with multiple splice transcripts, the start position for the coding sequences was based on the last start site, and the end position was based on the first end site on the chromosome, provided that the resulting coding sequence length was larger than 0. There are total 17,609 genes in the genome that meet these criteria, and 205 genes were annotated as oncogenes.

SNP array data analysis

SNP array data were analyzed using Affymetrix Genotyping Console 3.0.2 and Birdseed v2 genotype algorithm. All of the arrays passed quality control requirements, with contrast QC and MAPD values within boundaries. If there were no paired samples, samples were normalized against default Affymetrix normal samples. For the copy number analysis, we used regional GC correction and required 5 markers to be found within the changed region and the size of the region to be at least 100 kb. Genotyping Console Browser (Affymetrix) was used to illustrate copy number changes detected.

SNP array data of 820 cell lines, kindly provided by Sanger Institute, were profiled on the Affymetrix SNP Array 6.0 set.

SNP array data of 38 unique HapMap normal cell lines downloaded from GEO (datasets GSE15096 and GSE17359), were profiled on the Affymetrix SNP Array 6.0 set.

SNP array data of 20 paired prostate tumor samples with matched normal samples were downloaded from GEO (GSE12702), and were profiled on the Affymetrix Mapping 500 K Array set. Samples were normalized against the matched normal samples.

SNP array data of 25 GIST cancer samples were downloaded from GEO (dataset GSE20709) and were profiled on the Affymetrix SNP Array 6.0 set.

Searching for genes containing copy number breakpoint

For each sample, a segment reporting file was exported from Affymetrix Genotyping Console 3.0.2, which contains CNV segment information, including copy number state, chromosome location, start position, and end position. The information was utilized to search for copy number breakpoint in an interested gene.

Simulation

In order to evaluate the probability of finding, by random chance, a start/end point of a CNV segment that falls inside a gene of interest in a specific cell line, a simulation was performed as following: 1. For a defined cell line, extract the information of all CNV segments, such as size, copy number status (gain or loss). 2. Randomly assign the CNV segments, with the defined size, gain/loss status, to a hypothetical human genome, avoiding generating CNV segment overlaps. 3. Examine whether there is a start/end point of a CNV segment that falls inside the gene(s) of interested. 4. Repeat steps 1–3 for 100,000 iterations. 5. Calculate *P* values based on the number of iterations required to identify a CNV inside the gene(s) of interest based on the above criteria, per 100,000 iterations.

Acknowledgements

Thanks to Dr. Jack Pollard for helpful discussions and suggestions.

Disclosures

Author(s) have provided signed confirmations to the publisher of their compliance with all applicable legal and ethical obligations in respect to declaration of conflicts of interest, funding, authorship and contributorship, and compliance with ethical requirements in respect to treatment of human and animal test subjects. If this article contains identifiable human subject(s) author(s) were required to supply signed patient consent prior to publication. Author(s) have confirmed that the published article is unique and not under consideration nor published by any other publication and that they have consent to reproduce any copyrighted material. The peer reviewers declared no conflicts of interest.

References

1. Soda M, Choi YL, et al. Identification of the transforming EML4-ALK fusion gene in non-small-cell lung cancer. *Nature*. 2007;448(7153):561–6.
2. Mertens F, Antonescu CR, et al. Translocation-related sarcomas. *Semin Onco*. 2009;l36(4):312–23.
3. MacDonald JW, Ghosh D. COPA—cancer outlier profile analysis. *Bioinformatics*. 2006;22(23):2950–1.
4. Tomlins SA, Rhodes DR, et al. Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer. *Science*. 2005;310(5748):644–8.
5. Maher CA, Kumar-Sinha C, et al. Transcriptome sequencing to detect gene fusions in cancer. *Nature*. 2009;458(7234):97–101.
6. Nickoloff JA, De Haro LP, et al. Mechanisms of leukemia translocations. *Curr Opin Hematol*. 2008;15(4):338–45.
7. Kurzrock R, Kantarjian HM, et al. Philadelphia chromosome-positive leukemias: from basic mechanisms to molecular therapeutics. *Ann Intern Med*. 2003;138(10):819–30.
8. Uphoff CC, Habig S, et al. ABL-BCR expression in BCR-ABL-positive human leukemia cell lines. *Leuk Res*. 1999;23(11):1055–60.
9. Rajput AB, Miller MA, et al. Frequency of the TMPRSS2:ERG gene fusion is increased in moderate to poorly differentiated prostate cancers. *J Clin Pathol*. 2007;60(11):1238–43.
10. Rikova K, Guo A, et al. Global survey of phosphotyrosine signaling identifies oncogenic kinases in lung cancer. *Cell*. 2007;131(6):1190–203.
11. Shtivelman E, Henglein B, et al. Identification of a human transcription unit affected by the variant chromosomal translocations 2;8 and 8;22 of Burkitt lymphoma. *Proc Natl Acad Sci U S A*. 1989;86(9):3257–60.
12. Boland E, Clayton-Smith J, et al. Mapping of deletion and translocation breakpoints in 1q44 implicates the serine/threonine kinase AKT3 in postnatal microcephaly and agenesis of the corpus callosum. *Am J Hum Genet*. 2007;81(2):292–303.
13. Gajecka M, Glotzbach CD, et al. Identification of cryptic imbalance in phenotypically normal and abnormal translocation carriers. *Eur J Hum Genet*. 2006;14(12):1255–62.
14. Iijima Y, Ito T, et al. A new ETV6/TEL partner gene, ARG (ABL-related gene or ABL2), identified in an AML-M3 cell line with a t(1;12)(q25;p13) translocation. *Blood*. 2000;95(6):2126–31.
15. Skalova A, Vanecek T, et al. Mammary analogue secretory carcinoma of salivary glands, containing the ETV6-NTRK3 fusion gene: a hitherto undescribed salivary gland tumor entity. *Am J Surg Pathol*. 2010;34(5): 599–608.
16. Eguchi M, Eguchi-Ishimae M, et al. Fusion of ETV6 to neurotrophin-3 receptor TRKC in acute myeloid leukemia with t(12;15)(p13;q25). *Blood*. 1999;93(4):1355–63.
17. Lin E, Li L, et al. Exon array profiling detects EML4-ALK fusion in breast, colorectal, and non-small cell lung cancers. *Mol Cancer Res*. 2009;7(9): 1466–76.
18. Croce CM. Oncogenes and cancer. *N Engl J Med*. 2008;358(5):502–11.
19. Bastus NC, Boyd LK, et al. Androgen-induced TMPRSS2:ERG fusion in nonmalignant prostate epithelial cells. *Cancer Res*. 2010;70(23):9544–8.
20. Mwamukonda K, Chen Y, et al. Quantitative expression of TMPRSS2 transcript in prostate tumor cells reflects TMPRSS2-ERG fusion status. *Prostate Cancer Prostatic Dis*. 2010;13(1):47–51.
21. Stephens PJ, McBride DJ, et al. Complex landscapes of somatic rearrangement in human breast cancer genomes. *Nature*. 2009;462(7276):1005–10.
22. Gemmill RM, West JD, et al. The hereditary renal cell carcinoma 3;8 translocation fuses FHIT to a patched-related gene, TRC8. *Proc Natl Acad Sci U S A*. 1998;95(16):9572–7.
23. Belloso JM, Bache I, et al. Disruption of the CNTNAP2 gene in a t(7;15) translocation family without symptoms of Gilles de la Tourette syndrome. *Eur J Hum Genet*. 2007;15(6):711–3.
24. Berger MF, Lawrence MS, et al. The genomic complexity of primary human prostate cancer. *Nature*. 2011;470(7333):214–20.
25. Moller RS, Kubart S, et al. Truncation of the Down syndrome candidate gene DYRK1A in two unrelated patients with microcephaly. *Am J Hum Genet*. 2008;82(5):1165–70.
26. Strick R, Zhang Y, et al. Common chromatin structures at breakpoint cluster regions may lead to chromosomal translocations found in chronic and acute leukemias. *Hum Genet*. 2006;119(5):479–95.
27. Zhang Y, Rowley JD. Chromatin structural elements and chromosomal translocations in leukemia. *DNA Repair (Amst)*. 2006;5(9–10):1282–97.
28. Choi YL, Soda M, et al. EML4-ALK mutations in lung cancer that confer resistance to ALK inhibitors. *N Engl J Med*. 2010;363(18):1734–9.
29. Kwak EL, Bang YJ, et al. Anaplastic lymphoma kinase inhibition in non-small-cell lung cancer. *N Engl J Med*. 2010;363(18):1693–703.
30. Gerber DE, Minna JD. ALK inhibition for non-small cell lung cancer: from discovery to therapy in record time. *Cancer Cell*. 2010;18(6):548–51.
31. Berger MF, Levin JZ, et al. Integrative analysis of the melanoma transcriptome. *Genome Res*. 2010;20(4):413–27.

Supplementary Tables

Supplemental Table 1. Oncogenes that contain copy number breakpoints in eight CML cell lines.
Notes: Breakpoint is within the transcript, regardless of the location or whether it was associated with amplification or deletion.

Supplemental Table 2. Oncogenes that contain copy number breakpoints in 20 prostate cancer samples.
Notes: Breakpoint is within the transcript, regardless of the location or whether it was associated with amplification or deletion.

Supplemental Table 3. Oncogenes that contain copy number breakpoints in 140 lung cancer cell lines.
Notes: Breakpoint is within the transcript, regardless of the location or whether it was associated with amplification or deletion.

Supplemental Table 4. Oncogenes that contain copy number breakpoints in Sanger's 820 cancer cell lines.
Notes: Breakpoint is within the transcript, and copy number of the coding region is either remained normal or amplified, but not deleted.

Supplemental Table 5. Top ten genes that contain copy number breakpoints in Sanger's 820 cancer cell lines, and not in 38 Hapmap normal cell lines.

Supplementary tables provided in 8026supplementarytables.zip