Published online 27 May 2011 Nucleic Acids Research, 2011, Vol. 39, No. 15 **e100**
doi:10.1093/nar/gkr362

A novel bioinformatics pipeline for identification and characterization of fusion transcripts in breast cancer and normal cell lines

Yan W. Asmann[1], Asif Hossain[1], Brian M. Necela[2], Sumit Middha[1], Krishna R. Kalari[2], Zhifu Sun[1], High-Seng Chai[1], David W. Williamson[3], Derek Radisky[2], Gary P. Schroth[3], Jean-Pierre A. Kocher[1], Edith A. Perez[4] and E. Aubrey Thompson[2,*]

[1]Division of Biomedical Statistics and Informatics, Department of Health Sciences Research, Mayo Clinic College of Medicine, Rochester, MN, [2]Department of Cancer Biology, Mayo Clinic Comprehensive Cancer Center, Jacksonville, FL, [3]Illumina Inc., Hayward, CA and [4]Department of Medicine, Mayo Clinic, Jacksonville, FL, USA

Received February 9, 2011; Revised April 25, 2011; Accepted April 27, 2011

ABSTRACT

SnowShoes-FTD, developed for fusion transcript detection in paired-end mRNA-Seq data, employs multiple steps of false positive filtering to nominate fusion transcripts with near 100% confidence. Unique features include: (i) identification of multiple fusion isoforms from two gene partners; (ii) prediction of genomic rearrangements; (iii) identification of exon fusion boundaries; (iv) generation of a 5′–3′ fusion spanning sequence for PCR validation; and (v) prediction of the protein sequences, including frame shift and amino acid insertions. We applied SnowShoes-FTD to identify 50 fusion candidates in 22 breast cancer and 9 non-transformed cell lines. Five additional fusion candidates with two isoforms were confirmed. In all, 30 of 55 fusion candidates had in-frame protein products. No fusion transcripts were detected in non-transformed cells. Consideration of the possible functions of a subset of predicted fusion proteins suggests several potentially important functions in transformation, including a possible new mechanism for overexpression of ERBB2 in a HER-positive cell line. The source code of SnowShoes-FTD is provided in two formats: one configured to run on the Sun Grid Engine for parallelization, and the other formatted to run on a single LINUX node. Executables in PERL are available for download from our web site: http://mayoresearch.mayo.edu/mayo/research/biostat/stand-alone-packages.cfm.

INTRODUCTION

Gene fusion events resulting from inversions, interstitial deletion or translocations represent one of the most common types of genomic rearrangement (1). So far, the majority of fusion genes have been identified in leukemias, lymphomas and sarcomas. Recently, the discovery of *TMPRSS2-ERG* fusions in prostate cancer (2) and *EML4-ALK* fusion in non-small-cell lung tumors (3) suggests that gene fusion events may as well occur with a relatively high frequency in solid tumors, leading to the generation of novel fusion proteins with unique oncogenic properties. Since these fusion gene products are mainly restricted to tumor cells, they constitute potentially useful diagnostic and therapeutic targets. For example, the BRC-ABL1 fusion gene has been a diagnostic marker for chronic myelogenous leukemia (CML), as well as the drug target of Imatinib (Gleevec) in cells that harbor the *BRC-ABL1* fusion gene. In addition, the identification of fusion gene products in solid tumors may yield new insight into the etiology of certain tumors. The prostate cancer-specific *TMPRSS2-ERG* fusion events place growth regulatory genes under the influence of an androgen-regulated promoter, giving rise to a novel oncogene that has the potential to amplify normal androgen-dependent growth (2).

Identification and validation of fusion genes or their products in solid tumors have been challenging, largely due to the technical limitations inherent in techniques such as comparative genomic hybridization, fluorescent *in situ* hybridization, cytogenetic analysis and spectral karyotyping. However, the combination of bioinformatics approaches for the identification of fusion candidates followed by reverse-transcriptase PCR validation of the

*To whom correspondence should be addressed. Tel: +904 953 6226; Fax: +904 953 6233; Email: thompson.aubrey@mayo.edu fusion events has been successful with both microarray (2) and massive parallel RNA sequencing (RNA-Seq) data (4–7). Specially, the recent advances in RNA-Seq using next-generation sequencers have opened new horizons on the identification of expressed fusion transcripts. Four very recent publications substantiated the power of this approach. The use of long sequencing reads (median length of ~250 bases) generated by the Roche 454 sequencer identified 9 chimeric mRNAs in the HCC1954 breast cancer cell line in which all fusion transcripts were subsequently verified by Sanger re-sequencing of genomic DNA (8). Zhao *et al.* used an earlier version of the Roche 454 sequencer to generate reads of median length 88 nt from a primary breast cancer sample (9) and identified 6 putative gene fusion events, one of which involved *UBR4* (ubiquitin protein ligase E3 component n-recognin 4, chr1) and *GLB1* (beta-galactosidase-like protein, chr3). UBR4, also known as RB1-associated p600, is a cellular target of human papilloma virus E7 oncoprotein and is involved in anchorage-independent growth and transformation (10). Two studies from the Chinnaiyan group at the University of Michigan took advantage of the higher throughput, lower cost, short reads from the Illumina Genome Analyzer (IGA). The first of the two studies initially used the long read capacity of the Roche 454 sequencer to generate a reference library which was then interrogated using short reads (36-base) from the IGA (5). With this approach, the study 're-discovered' the *BCR-ABL1* and *TMPRSS2-ERG* in CML and prostate cancer cell lines, respectively. In addition, the authors identified and validated a number of novel fusion gene products in prostate cancer cells including several heretofore unknown ETS gene fusion products. Although the initial Chinnaiyan paper represents a landmark in the use of next-generation sequencers to identify fusion gene products, the authors noted several drawbacks related to the approaches used in the study. The first and most obvious is the overhead associated with the need to analyze each sample with two different platforms (Roche 454 and IGA). Also, the short read length gave rise to a high false discovery rate (FDR), which was not entirely eliminated by the application of a number of informatics filters. A later paper from the same group described the use of paired-end sequencing to identify fusion gene products in tumor cell lines (6). Briefly, this approach involves the generation of RNA-Seq libraries which contain different adaptors on the 5′- and 3′-ends of the cDNA fragments. The IGA was then used to carry out sequencing from each end of 16–25 million cDNA fragments (50-base from each end). Reads whose two ends mapped to transcripts from two different genes (fusion encompassing reads) are indicative of fusion events, which are supported by the reads whose one end mapped to one of the fusion partners and the other end spanning the fusion junction point (fusion junction spanning reads). In the same paper, the paired end deep sequence analysis was extended to MCF7 cells to identify fusion gene products. The analysis identified *BCAS4-BCAS3* and *ARFGEF2-SULF2,* which had previously been described in MCF7 cells (11), as well as several novel fusion transcripts. Two recent publications have extended the use of paired-end sequencing data to identify fusion gene transcripts in cell lines, including the analysis of fusion gene products in a small group of breast cancer cells (4,12).

All the bioinformatics methods for fusion discovery in transcriptome sequencing data described in the recent publications incorporated steps of false positive filtering and identified novel fusion transcripts in various types of solid tumors or cancer cell lines. However, only Sboner *et al.* (7) made their bioinformatics tool publicly available. In addition, all published bioinformatics algorithms stop at the nomination of fusion candidates, while the follow-up analyses including fusion validation and studying of the structural details of the fusion transcripts still require a substantial amount of manual and laborious efforts. For example, the design of PCR primers for fusion validations needs to take into account many factors including the original orientations of the two partner genes on the chromosome(s), the orientation of the fusion transcript, the 5′ to 3′ order of the two partners in the fusion product, as well as the knowledge of the mapping orientations of the read pairs. We developed a robust bioinformatics pipeline for the identification of fusion transcripts in paired-end RNA sequencing (RNA-Seq) data, which is part of Mayo Clinic's Next-Generation sequencing data analysis tool suite, SnowShoes. We named this pipeline SnowShoes-FTD (Fusion Transcript Detection). The analytical power of this pipeline lies in a very low false detection rate (approaching 0%), thereby overcoming the major problem that we and others have encountered in confirmation of candidate fusion transcripts. Moreover, SnowShoes-FTD incorporates several subroutines to generate template regions for PCR primer design, which facilitates quick PCR validations, as well as the amino acid sequences of the putative in-frame fusion gene products, which facilitates predictions concerning the functional significance of the fusion events. In addition, SnowShoes-FTD provides likely fusion mechanisms (translocation, inversion and deletion), strand orientation of fusion transcripts and identifies the mutations at the fusion junction points for in-frame fusions. SnowShoes-FTD is publicly available for download at web site: http://mayoresearch.mayo.edu/mayo/research/biostat/stand-alone-packages.cfm.

MATERIALS AND METHODS

Breast cell lines

Twenty-two breast cancer cell lines and one non-tumorigenic breast epithelial cell line (MCF10A) were obtained from the American Type Culture Collection (ATCC) (Table 1). All cell lines were thawed and expanded to allow for the isolation of total RNA from low passage cells, which should exhibit minimal deviation from the ATCC type reference cells. Eight primary human mammary epithelial cell (HMEC) cultures were established from biopsies of Mayo Clinic patients undergoing evaluation of suspected breast lesions (Table 1). All the biopsy samples from which the cell lines were derived

Table 1. Sample information of the 31 breast cell lines

Sample number	Sample ID	Sequencing location	Sample description	Flow cell lane	Run number
1	BT-474	Illumina Hayward, CA, USA	Cancer Cell Line	1	Run #1
2	MCF10A		Non-Tumorigenic	2	
3	BT-20		Cancer Cell Line	3	
4	MCF7		Cancer Cell Line	4	
5	MDA-MB-468		Cancer Cell Line	6	
6	T47D		Cancer Cell Line	7	
7	ZR-75-1		Cancer Cell Line	8	
8	HCC1937	Mayo Clinic Sequencing Core	Cancer Cell Line	1	Run #2
9	HCC1954		Cancer Cell Line	2	
10	HCC2218		Cancer Cell Line	3	
11	HCC1599		Cancer Cell Line	4	
12	HCC1395		Cancer Cell Line	5	
13	BT549		Cancer Cell Line	6	
14	Hs578T		Cancer Cell Line	7	
15	MDA-MB-175V-II		Cancer Cell Line	8	
16	MDA-MB-361	Mayo Clinic Sequencing Core	Cancer Cell Line	1	Run #3
17	MDA-MB-436		Cancer Cell Line	2	
18	MDA-MB-453		Cancer Cell Line	3	
19	SK-BR-3		Cancer Cell Line	4	
20	UACC812		Cancer Cell Line	5	
21	HCC1187		Cancer Cell Line	6	
22	HCC1428		Cancer Cell Line	7	
23	HCC1806		Cancer Cell Line	8	
24	DHF 168	Illumina Hayward, CA, USA	Normal HMEC*	1	Run #4
25	BSO19B		Normal HMEC	2	
26	BSO28		Normal HMEC	3	
27	BSO29		Normal HMEC	4	
28	BSO30		Normal HMEC	5	
29	BSO32N		Normal HMEC	6	
30	BSO36		Normal HMEC	7	
31	BSO37		Normal HMEC	8	

HMEC, human mammalian epithelial cells primarily cultured from benign breast biopsy samples.

were assessed as benign, and in the following discussion we will operationally define these cells as ‘normal’.

RNA preparation and sequencing

Total RNA extraction was performed using Exiqon's miRCURY RNA Isolation Kit. One microgram of total RNA was used for the sequencing library preparation, which was modified from conventional Illumina mRNA-Seq protocols to facilitate paired-end RNA sequence analysis (13). The cDNA fragments were amplified by PCR and sequenced at both ends for 50 bases (50-bp-end sequencing) using the IGA IIx. Sequencing was carried out at the Illumina assay development facility at Hayward, CA, USA and at the Mayo Clinic Advanced Genomic Technology Center at Rochester, MN, USA. The FASTQ read files for each sample were used for further analysis.

Construction of exhaustive one-directional exon junction database

The exon–exon boundary database was generated using the exon and gene definition files downloaded from UCSC Table Browser (table: refFlat; track: RefSeq Genes; group: Genes and Gene Prediction Tracks) in reference to human genome build 36 (hg18). Among 35 983 total transcripts in the refFlat file, 765 transcripts with alternative haplotypes and 1482 transcripts with multiple/redundant genomic locations were removed. Based on the exon boundaries of all transcripts defined in the curated refFlat file, all possible one-directional combinations of exon–exon boundary sequences for the sequencing length of 50 bases were generated to ensure that no reads will map to more than one junction using an in-house developed algorithm (S. Middha, N. Hossain, Y.W. Asmann, unpublished data). The curated refFlat file and its future updated versions in reference to both genome builds 36 and 37, as well as the FASTA files of exon–exon boundary sequences for different sequencing lengths (50-, 75- and 100-base) can be downloaded from our web site: http://mayoresearch.mayo.edu/mayo/research/biostat/stand-alone-packages.cfm.

The analytic workflow for fusion detection

As shown in Figure 1, the SnowShoes-FTD tool consists of (i) read alignments to both reference genome and exon junction database; (ii) annotation of aligned read pairs to identify potential fusion candidates; (iii) filtering of false positive candidates; (iv) generation of a continuous sequence region spanning fusion junction points for PCR primer design for experimental validation; (v) prediction of fusion mechanism; and (vi) prediction of the in-frame versus out of frame fusion products and generation of the predicted protein sequences of the in-frame fusion products based on known transcripts of the two

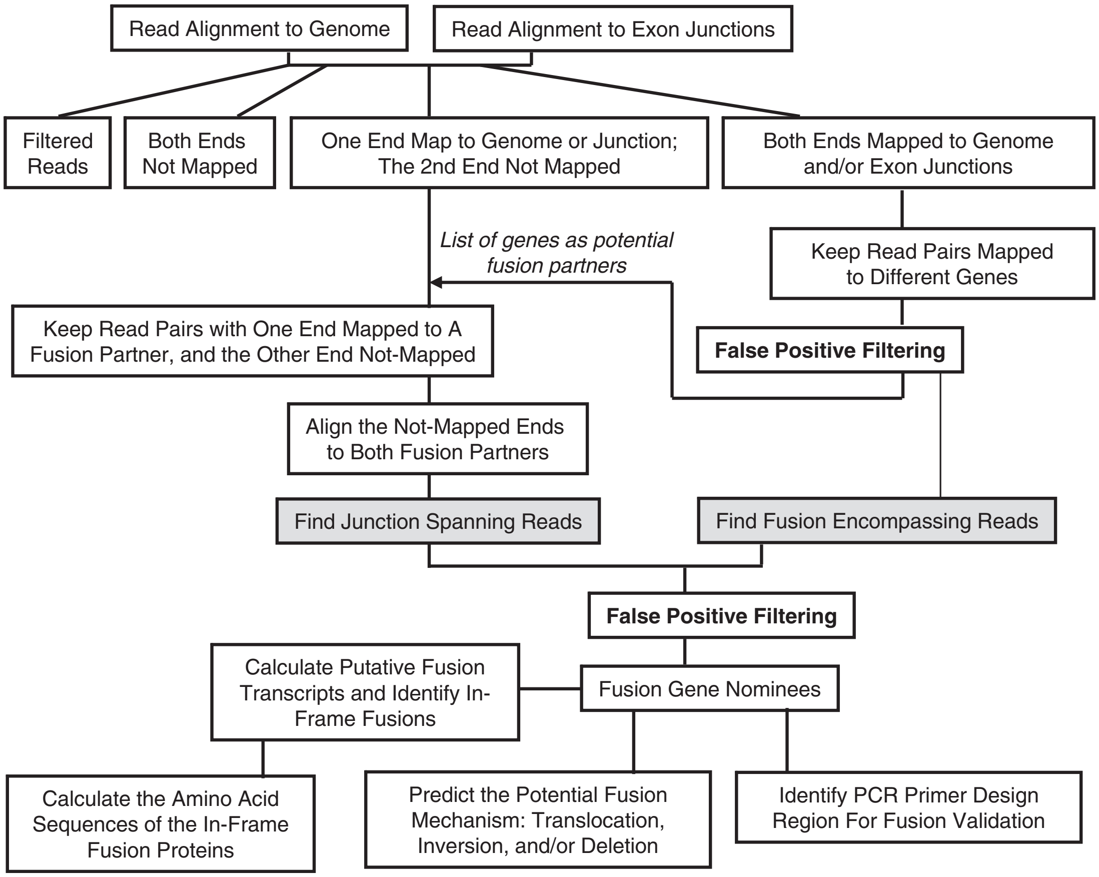
Read Alignment to Genome
Read Alignment to Exon Junctions
Filtered Reads
Both Ends Not Mapped
One End Map to Genome or Junction; The 2nd End Not Mapped
Both Ends Mapped to Genome and/or Exon Junctions
List of genes as potential fusion partners
Keep Read Pairs Mapped to Different Genes
Keep Read Pairs with One End Mapped to A Fusion Partner, and the Other End Not-Mapped
False Positive Filtering
Align the Not-Mapped Ends to Both Fusion Partners
Find Junction Spanning Reads
Find Fusion Encompassing Reads
False Positive Filtering
Calculate Putative Fusion Transcripts and Identify In-Frame Fusions
Fusion Gene Nominees
Calculate the Amino Acid Sequences of the In-Frame Fusion Proteins
Predict the Potential Fusion Mechanism: Translocation, Inversion, and/or Deletion
Identify PCR Primer Design Region For Fusion Validation

Figure 1. The work flow of the fusion detection algorithm implemented in SnowShoes-FTD.

partner genes. In addition, the tool filters out reads mapped with poor quality as described above.

Read alignment and filtering for fusion detection. The two ends of RNA-Seq reads were aligned to both the Human Reference genome build 36 (hg18) and exon junctions using Burrows-Wheeler Alignment (BWA) (14) with a seed length of 32 allowing 4% of maximum edit distance. The BWA aligned reads are stored in the Sequence Alignment/Map (SAM) format (15). The pair of SAM files from the alignment of two ends of the same sample were sorted according to read IDs using SAMtools (15). The reads with neither end mapped to genome or exon junctions are not informative and were filtered out. If the Phred-scaled Mapping Quality Score (MAPQ) of either end was <20, the end pair is considered low quality and was excluded from further analysis. Note that this will also filter out read pairs with either or both ends mapped to multiple locations since BWA assigns a MAPQ of zero to such reads.

Annotation of aligned reads. After filtering, the reads remaining in the SAM files were categorized into five groups: (i) reads with both ends mapped to genome locations; (ii) reads with both ends mapped to exon junctions; (iii) reads with one end mapped to the genome and the other mapped to exons; (iv) reads with one end mapped to the genome and the other end not mapped; (v) reads with one end mapped to exon junctions and the other not mapped. All mapped ends were annotated using the genes and exons defined in the curated refFlat file. For a read to be annotated as being mapped to a gene, we required that either the start or the end of the read be mapped within the boundaries of an exon of that gene. If a read aligned to both genome and an exon junction, the annotation from the exon junction alignment takes precedence.

False positive filtering. There are two steps of filtering to minimize the false fusion rate that has plagued nomination of fusion gene candidates. The first filtering step is performed on the reads pairs that are annotated to two different genes, also known as fusion encompassing reads. This begins with the filtering of fusion candidates with significant sequence similarities between the two fusion partners (Please refer to the SnowShoes-FTD user manual for details). In addition, a gene distance filter is implemented to exclude fusions formed by two genes that are within M kb of each other on the reference genome, in order to eliminate chimeric transcripts that might arise from overlapping genes or transcriptional read through of adjacent genes. Furthermore, the fusion candidates with less than N fusion encompassing reads are filtered out. The second filtering step focuses on the fusion candidates with supporting evidences of both fusion encompassing read pairs and fusion junction spanning reads.

The mapping orientations of the end pairs are compared to the orientations of the two fusion partner genes on the genome, and the fusion candidates with inconsistent mapping orientations between end pairs are filtered out. Also, the algorithm requires at least X unique fusion junction spanning reads and no more than Y fusion junction points per fusion candidate. Note that these thresholds (M, N, X and Y) are user defined.

Prediction of the fusion mechanism. If a fusion product is formed by two partner genes from two different chromosomes, a translocation will be listed as the mechanism of fusion. The translocation event can be accompanied by inversion of the two partner genes that have the opposite strand orientations. When the two partner genes are located on the same chromosome, the mechanism of the fusion could be translocation alone, inversion alone and inversion and translocation concurrently. These three scenarios are determined based on the strand orientations and the relative chromosomal positions of the two partners. However, when an intra-chromosomal fusion arises without altering the relative orders of the two partners with the same strand orientation, the fusion can be the consequence of a translocation or an interstitial deletion.

Prediction of the fusion protein product. Prediction of the fusion protein sequences was carried out using all of the known transcripts of the two fusion partner genes as defined in the refFlat file. As shown in Figure 3, we first identify the two exons from each of the two fusion partner genes that aligned to the fusion spanning reads (fusion boundary exons). Next, among all know transcripts of the two fusion partner genes, we identify the transcripts containing the boundary exons and generated a list of putative fusion transcripts. Each of the putative fusion transcripts was then translated into predicted amino acid sequence and each of the putative fusion proteins characterized as whether it is in frame. In addition, the fusion products are categorized as: (i) coding region to coding region fusion which will result in in-frame fusion product, a frame shift for the 3′ gene or an in-frame fusion with a single amino acid mutation at the fusion junction point. The single amino acid mutation will be listed in the SnowShoes-FTD output; (ii) 5′-UTR to coding region fusion in which the promoter of the 5′-gene fused in front of a coding region of the 3′-gene; (iii) 5′-UTR to 3′-UTR fusion in which coding regions from both partner genes are fused out; (iv) 3′-UTR to 3′-UTR fusion in which the 5′-gene are intact but the coding region of the 3′-gene is fused out; (v) 5′-UTR to 5′-UTR fusion in which the promoter of the 5′-gene will potentially drive the expression of 3′-gene as the consequence of the fusion; (vi) 3′-UTR to 5′-UTR or coding region fusion in which the stop codon of the 5′-gene will terminate the translation of any coding regions of the 3′-gene; (vii) coding region to 5′-UTR fusion in which the sequence between the coding region of the 5′-gene and the start codon of the 3′-gene may result in an insertion of single or multiple amino acids that are listed in the output file; (viii) the coding region to 3′-UTR fusion which may result in the shortening of the 5′-gene with or without the addition of foreign amino acids.

Nucleotide sequences spanning fusion junction points for PCR primer design. The chromosomal orientations of the two fusion partners, the mapping orientations of the two ends from fusion encompassing read pairs, as well as the sequence and orientation of the fusion junction spanning read(s) are used to report a template region for PCR primer design in order to quickly validate the fusion candidates with RT–PCR. From 5′–3′, the template region consists of the exon region from partner A from the start of the exon to the fusion junction point, a '||' sign that signifies the fusion junction point and the exon region from partner B from the start of the fusion junction point to the end of the exon. Since the orientation of the primer template region does not necessarily define directionality (5′ to 3′) of the fusion transcript, it is necessary to use double-stranded cDNAs as the template for PCR validation.

PCR and Sanger sequencing validations of fusion candidates

Double-stranded cDNA were synthesized using the total RNAs from each of the 31 cell lines. It should be noted that, to minimize potential artifacts that might arise during library construction, different cDNA libraries were constructed and used for sequencing and for PCR validation. PCR primers were designed using the template regions recommended by SnowShoes-FTD. The 5′ and 3′ primers were complementary to the template regions that represent the two fusion partners, respectively. The fusion transcript is considered validated if we detected a PCR product of the predicted size. The PCR bands from randomly selected fusion transcripts were sequenced using Sanger sequencing to further confirm the nucleotide sequence of the predicted fusion junctions.

Quantification of gene and exon expression levels

The gene expression levels were calculated as the sum of the individual exon read counts and exon junction read counts. The expression levels of genes and exons were normalized using the total aligned reads from the sample and the length of the exon or gene (Reads per kilo bases per million, RPKM).

RESULTS

Flexibility of the choice of sequence alignment tools

Currently, there are several sequencing platforms and multiple sequence alignment algorithms designed for next-generation sequencing of transcriptome. The SnowShoes-FTD can work with raw or post-alignment files of different platforms. When FASTQ files obtained from IGA or HiSeq sequencers are provided as input, the users can choose BWA or Bowtie (16) for alignment. We also accept post-alignment files (BAM) for both genome and exon junction alignments from a different sequencing platforms including Life Technologies' SOLiD sequencer.

Since we prefer the exon junction database generated by SnowShoes-FTD over other publically available junction databases, the user will need to align the reads to the exon junctions provided by SnowShoes-FTD if BAM files are provided as input files. The results reported in the current manuscript were obtained using FASTQ as input files and BWA as the aligner.

User-defined parameters for SnowShoes-FTD

The following parameters are user-defined for detection of fusion transcripts using SnowShoes-FTD: (i) the minimum number of fusion encompassing reads (default value: 10); (ii) the minimum number of unique fusion junction spanning reads (must be ≥ 1 with a default set to 2); (iii) the minimum distance between the two fusion partner genes if both are located on the same chromosome (default value: 100 kb); (iv) the maximum number of fusion isoforms allowed between two fusion partners (default value: 2); and (v) whether the fusion transcripts feature junction points at exon boundaries (default = Yes). The default values of the parameters were chosen to minimize false positive rate. For example, the minimum number of unique fusion junction spanning reads was set to 2 by default to avoid the false detection of fusion junction spanning reads arising form the PCR artifacts, which may give multiple junction spanning reads that are identical in alignment positions. In addition, the limit of the maximum fusion isoforms between two partner genes is based on the hypothesis that if there are too many fusion isoforms between two partners, the fusion event would appear to be existing by random fusion events without obvious biological significances.

List of reference files available

A list of reference files is available for download in preparation for the fusion transcript detection using SnowShoes-FTD: (i) the one-directional exhaustive exon–exon junction database generated for read-lengths 50-, 75- and 100-bases. This is provided in the FASTA format; and (ii) the curated gene and exon definition files (refFlat files) from both genome builds 36 and 37. The gene and exon definition files will be updated periodically. All reference files can be obtained from the SnowShoes web site: http://mayoresearch.mayo.edu/mayo/research/biostat/stand-alone-packages.cfm.

Detection of fusion transcripts in 31 breast cell lines

We applied the SnowShoes-FTD tool to the 50-bp-end RNA-Seq data from 22 breast cancer cell lines, 1 established non-tumorigenic breast cell line (MCF10A) and 8 primary HMEC cultures (Table 1). The fusion transcript candidates of these 31 breast cell lines were nominated using the default parameter values based on genome build 36 (hg18). As shown in Supplementary Table S1, read pairs sequenced per sample total to 18–33 millions, among which: 45–58% have both ends mapped to the genome; 3–5% have both ends mapped to exon junctions; 11–18% with one end mapped to the genome and the other mapped to exon junctions; 5–15% with one end mapped to the genome and the other not mapped; and 1–2% with one end mapped to exon junctions and the other not mapped. In addition, there are 2–9% of the read pairs with neither ends mapped to the genome or exon junctions. In all, 11–20% of the reads were filtered out due to low mapping quality and/or redundant mapping.

We nominated 55 fusion transcript candidates (Table 2 and Supplementary Data S2). Fifty of these have unique isoforms while the rest have two isoforms. As shown in Figure 2a, all 50 fusion transcripts with a single fusion isoform were validated as evidenced by the generation of PCR products of the predicted sizes. We randomly picked several fusion transcripts for further validation using Sanger sequencing of the PCR bands. All PCR products were confirmed by Sanger sequencing with the observation that the predicted DNA sequence conformed to the actual DNA sequence of the PCR product (data not shown). All isoforms were similarly validated for the five fusion candidates with two isoforms (Figure 2b). The sequences of the primers used in PCR validations can be found in Supplementary Data S4, which includes the primers for the alternative isoforms of the five fusion candidates with two isoforms each.

Among the 55 fusion candidates, 30 were in-frame (Table 2 and Supplementary Data S3). We define a fusion product as ‘in frame’ when there was no frame shift in the 3′-gene, regardless whether there is single amino acid mutation or single/multiple amino acid insertion at the fusion junction point. The fusion junction point mutations are also listed in the Supplementary Data S3. In addition, the list of fusion transcripts as the result of exhaustive combinations of all transcripts from two partner genes may contain identical fusion products if the differences between the transcripts from the same partner are ‘fused out’. For example, as shown in Figure 3d, the fusion transcript of A1-B4 is identical to that of A1-B1, and the fusion transcript of A2-B4 is identical to that of A2-B1. These identical fusion proteins are flagged in the SnowShoes output file (Supplementary Data S3).

Fusion genes identified in MCF7 cancer cell line

Fusion gene products in the MCF7 cell line had been previously described using a paired-end sequencing protocol, so we compared the list of fusion transcripts identified in MCF7 cancer cell line using SnowShoes-FTD with those identified by Maher *et al.* (6). The SnowShoes-FTD identified and validated five novel fusion transcripts that were not reported by Maher *et al.*: ADAMTS19-SLC27A6, ATXN7L3-FAM171A2, GCN1L1-MSI1, MYH9-EIF3D and RPS6KB1-DIAPH3. In addition, there were five fusion genes identified by Maher *et al.* that were not detected by SnowShoes-FTD: ARHGAP19-DRG1, BC017255-TMEM49, PAPOLA-AK7, AHCYL1-RAD51C and FCHOL-MYO9B. We found that (i) BC017255 is no longer in the RefSeq RNA database; (ii) the distance between PAPOLA-AK7 is 65 kb which is smaller than the default setting of 100 kb.

Table 2. List of fusion transcripts identified

Fusion transcript	Mechanism	Type	In frame	Strand	Total read pairs	Between exon boundaries	No. of fusion isoforms	Sample ID
LIMA1 → USP22	T	inter-chr	Yes	−	16	Yes	1	BT-20
ACACA → STAC2	T	intra-chr	Yes	−−	72	Yes	1	BT-474
FAM102A → CIZ1	T	intra-chr		−	31	Yes	2	BT-474
GLB1 → CMTM7	I	intra-chr	Yes	−	13	Yes	1	BT-474
MED1 → STXBP4	I and T	intra-chr	Yes	−	54	Yes	1	BT-474
PIP4K2B → RAD51C	I and T	intra-chr		−	15	Yes	1	BT-474
RAB22A → MYO9B	T	inter-chr		+	16	Yes	1	BT-474
RPS6KB1 → SNF8	I and T	intra-chr	Yes	+	162	Yes	1	BT-474
STARD3 → DOK5	T	inter-chr		+	21	Yes	1	BT-474
TRPC4AP → MRPL45	I and T	inter-chr	Yes	−	27	Yes	1	BT-474
ZMYND8 → CEP250	I	intra-chr		−	189	Yes	2	BT-474
CTAGE5 → SIP1	T	intra-chr		+	64	Yes	1	HCC1187
MLL5 → LHFPL3	T	intra-chr		+	23	Yes	1	HCC1187
PUM1 → TRERF1	T	inter-chr		−	58	Yes	1	HCC1187
SEC22B → NOTCH2	I and T	intra-chr		+	22	Yes	1	HCC1187
EIF3K → CYP39A1	I and T	inter-chr	Yes	+	91	Yes	1	HCC1395
RAB7A → LRCH3	D or T	intra-chr		+	14	Yes	1	HCC1395
RNF187 → OBSCN	T	intra-chr		+	11	Yes	1	HCC1428
SLC37A1 → ABCG1	T	intra-chr	Yes	+	20	Yes	1	HCC1428
CYTH1 → PRPSAP1	D or T	intra-chr	Yes	−	33	Yes	1	HCC1599
EXOC7 → CYTH1	T	intra-chr	Yes	−	20	Yes	1	HCC1599
BRE → DPYSL5	T	intra-chr	Yes	+	13	Yes	1	HCC1806
CD151 → DRD4	T	intra-chr		+	11	Yes	1	HCC1806
LDLRAD3 → TCP11L1	T	intra-chr		+	25	Yes	1	HCC1806
RFT1 → UQCRC2	I and T	inter-chr	Yes	−	102	Yes	1	HCC1806
TAX1BP1 → AHCY	I and T	inter-chr	Yes	+	54	Yes	1	HCC1806
NFIA → EHF	T	inter-chr	Yes	+	18	Yes	1	HCC1937
GSDMC → PVT1	I	intra-chr		−	23	Yes	1	HCC1954
INTS1 → PRKAR1B	D or T	intra-chr	Yes	−	24	Yes	1	HCC1954
PHF20L1 → SAMD12	I and T	intra-chr	Yes	+	106	Yes	1	HCC1954
STRADB → NOP58	D or T	intra-chr	Yes	+	10	Yes	1	HCC1954
POLDIP2 → BRIP1	T	intra-chr		−	13	Yes	1	HCC2218
ADAMTS19 → SLC27A6	T	intra-chr		+	30	Yes	1	MCF7
ARFGEF2 → SULF2	I and T	intra-chr	Yes	+	421	Yes	1	MCF7
ATXN7L3 → FAM171A2	T	intra-chr		−	10	Yes	1	MCF7
BCAS4 → BCAS3	T	inter-chr		+	1697	Yes	1	MCF7
GCN1L1 → MSI1	T	intra-chr	Yes	−	25	Yes	1	MCF7
MYH9 → EIF3D	T	intra-chr	Yes	−	16	Yes	1	MCF7
RPS6KB1 → DIAPH3	I and T	inter-chr		+	25	Yes	1	MCF7
SULF2 → PRICKLE2	T	inter-chr		−	26	Yes	1	MCF7
ODZ4 → NRG1	I and T	inter-chr	Yes	−	12	Yes	1	MDA-MB-175V-II
BRIP1 → TMEM49	I	intra-chr		−	28	Yes	1	MDA-MB-361
SUPT4H1 → CCDC46	T	intra-chr		−	17	Yes	1	MDA-MB-361
TMEM104 → CDK12	T	intra-chr	Yes	+	10	Yes	2	MDA-MB-361
RIMS2 → ATP6V1C1	T	intra-chr	Yes	+	11	Yes	1	MDA-MB-436
TIAL1 → C10orf119	T	intra-chr		−	12	Yes	1	MDA-MB-436
MECP2 → TMLHE	T	intra-chr		−	29	Yes	1	MDA-MB-453
ARID1A → MAST2	D or T	intra-chr	Yes	+	18	Yes	1	MDA-MB-468
UBR5 → SLC25A32	T	intra-chr		−	28	Yes	1	MDA-MB-468
KLHDC2 → SNTB1	I and T	inter-chr	Yes	+	25	Yes	1	SK-BR-3
ARID1A → WDTC1	D or T	intra-chr	Yes	+	23	Yes	1	UACC812
HDGF → S100A10	D or T	intra-chr	Yes	−	154	Yes	1	UACC812
PPP1R12B → SNX27	T	intra-chr	Yes	+	45	Yes	1	UACC812
SRGAP2 → PRPF3	T	intra-chr	Yes	+	22	Yes	2	UACC812
WIPF2 → ERBB2	T	intra-chr	Yes	+	66	Yes	2	UACC812

In the fusion mechanism column: T, translocation; I, inversion; D, interstitial deletion; Intra-chr, intra-chromosomal fusion; Inter-chr, inter-chromosomal fusion. The fusion transcripts are named as the 5′ gene → 3′ gene. For example, LIMA1 → USP22 is a fusion transcript formed between two partner genes, LIMA1 and USP22, in which LIMA1 is the 5′ gene and USP22 is the 3′ gene.

In addition, no fusion junction spanning reads were observed to support this fusion. Therefore, this fusion transcript would only have been detected with a different distance threshold and by reducing the default for fusion spanning reads to 0; (iii) there are no junction spanning reads in our data set for AHCYL1-RAD51C although we did find 10 fusion encompassing reads supporting the existence of this fusion transcript; and (iv) there was only one fusion junction spanning read for FCHOL-MYO9B and the default setting for SnowShoes-FTD is 'at least two unique junction spanning reads'. On the other hand, we found no evidence in support of an ARHGAP19-DRG1

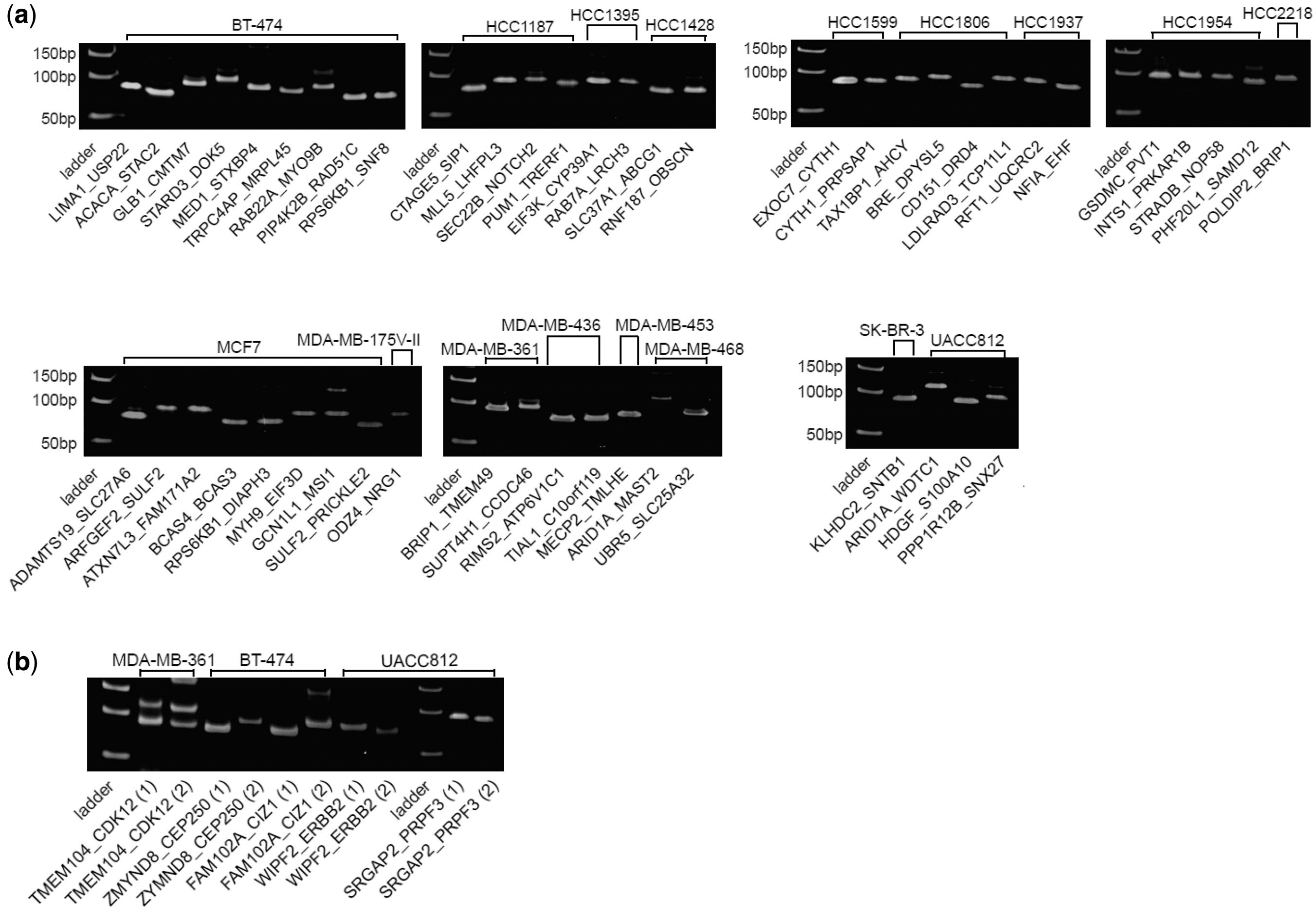

Figure 2. PCR validation of candidate fusion products. The PCR primers were designed using the template sequences generated by SnowShoes-FTD. The double-stranded cDNA libraries were constructed using total RNAs from each of the cell lines. The primer sequences and the expected PCR product sizes for each of the fusion candidates were detailed in Supplementary Data S4. (**a**) The PCR products from 50 fusion candidates with unique isoforms. The fusion candidates were grouped by the cell lines in which the fusion candidates were discovered. (**b**) The PCR products from five fusion candidates with two fusion isoforms each. Note that there are multiple PCR bands in the lanes for CDK12-TMEM104, and the lowest bands were those from the fusion product.

fusion, as the alignment file (SAM file) did not contain any read pairs that mapped to both of these genes. When we performed RT-PCR using the PCR primers provided by Maher *et al.* (Supplementary Data S5), the results also supported the existence of the fusion products BC017255-TMEM49, PAPOLA-AK7, AHCYL1-RAD51C and FCHOL-MYO9B, while no PCR product was observed for the ARHGAP19-DRG1 fusion. Thus, 4 out 5 'known' fusion transcripts that were not identified by SnowShoes-FTD can be explained by differences in the RefSeq database used for the analyses or by the choice of parameter settings for our various filtering steps. The ARHGAP19-DRG1 fusion transcript reported by Maher *et al.* does not appear to be expressed in the MCF7 cells that we have obtained from ATCC.

As our analyses were in preparation for submission, Edgren *et al.* (12) reported on the detection of fusion transcripts in four breast cancer cell lines, including MCF7 in which three fusion transcripts were validated. We detected eight fusion transcripts in MCF, including two of the three reported by Edgren *et al.* (BCAS4_BCAS3 and ARFGEF2_SULF2, both previously identified). Overall, our results correspond reasonably well with those reported by Edgren *et al.*

Pathway analysis of genes involved in fusion transcripts in breast cancer cell lines

There were a total of 105 fusion partner genes from the 55 fusion candidates, among which 58 genes formed in-frame fusion transcripts of 30 chimeric RNAs. We performed pathway and regulatory network analyses of these 58 genes using MetaCore (GeneGo Inc., San Diego, CA, USA). Two pathways that are enriched among these 58 genes: the non-genomic action of androgen receptor and ligand-independent activation of ESR1 and ESR2. Three GeneGo process network were significantly enriched: androgen receptor signaling cross-talk, ESR1-nuclear pathway and FGF/ERBB signaling. This observation suggests that fusion transcripts may have functional significance in signal transduction in breast cancer cells.

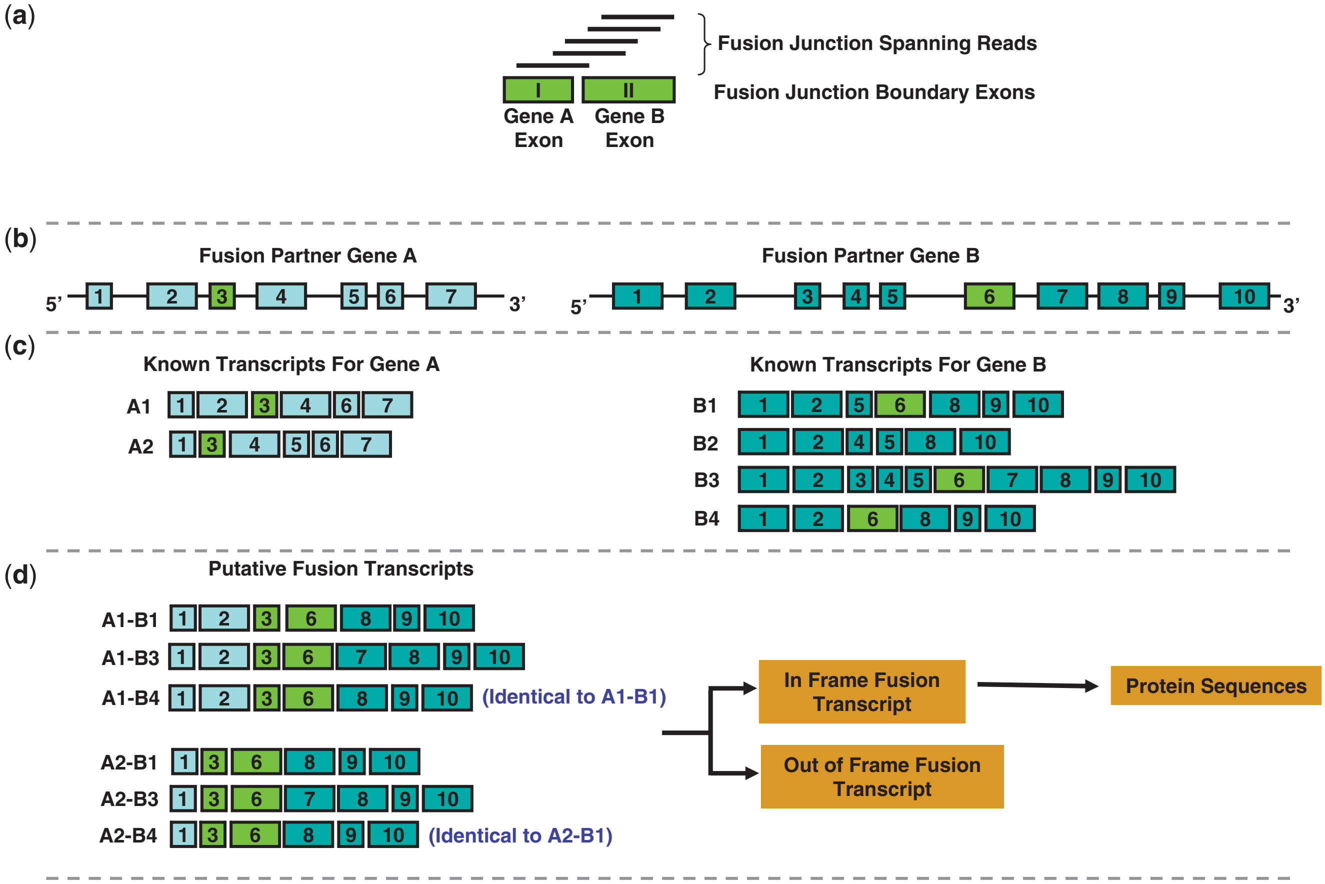

Figure 3. The identification of in-frame fusion transcripts and their predicted protein sequences. (**a**) Staring from the fusion junction spanning reads that aligned to both fusion partner genes, the two junction boundary exons from fusion partner genes A and B are identified; (**b**) obtaining the IDs and sequences of all exons belonging to the two fusion partner genes A and B based on the curated refFlat file. In this example, Gene A has 7 exons with the third exon as the fusion boundary exon, and gene B has 10 exons with the sixth exon as the fusion boundary exon; (**c**) obtaining all known transcripts for the two fusion partner genes. Gene A has two known transcripts (A1 and A2) both of which contain the fusion boundary exon. Gene B has 4 known transcripts (B1 → B4) and three of which (B1, B3 and B4) contain the fusion boundary exon. (**d**) Generating the list of exhaustive fusion transcripts using the known transcripts containing the fusion boundary exons. There are six possible fusion transcripts: A1-B1, A1-B3, A1-B4, A2-B1, A2-B3 and A2-B4. Note that because the differences between the transcripts B1 and B4 are 'fused out', the fusion transcript of A1-B1 is identical to that of A1-B4. Similarly, A2-B1 is identical to A2-B4. The fusion transcripts that cause frame shift in gene B are defined as 'out of frame', and the ones that did not cause any frame shift will be defined as 'in frame' fusions. Each of the in-frame fusions will be translated into amino acid sequences of the fusion proteins.

Structural analysis of fusion transcripts suggests a preponderance of 'promoter swap' mutations, one of which may represent a novel mechanism for ERBB2 overexpression

The analytical power of the SnowShoes-FTD pipeline lies in part in the very low false detection rate and in very large part in the downstream features that predict the structure of the hypothetical fusion transcripts and the amino acid sequence of the resultant translation products. Such analyses indicate that the nature of the fusion transcripts that were detected in breast cancer cells is strikingly non-random, as evidenced by the fact that 23 of the 60 confirmed chimeric transcripts result from fusion of exon 1 of the 5′/upstream partners to the 3′/downstream partners. The most probable cause of such chimeric RNAs is a genomic rearrangement (12) that results in juxtaposition of a promoter that potentially alters the level of expression and/or the regulation of the downstream partner in response to changes in the cellular environment. In addition, all the fusion transcripts that we have reported and validated map precisely to exon/exon junctions between the upstream and downstream fusion partners, suggesting that such transcripts are processed. There were only five additional fusion transcripts in which the fusion junction points are in the middle of exons (detected with different parameter settings for SnowShoes-FTD, data not shown). About half of the fusion events are in frame and therefore predicted to encode fusion proteins. The preponderance of such events in our samples suggests that some of the fusion transcripts may convey a growth advantage, such that transcript enrichment results from selection. For example, MDA-MB-468 cells express an ARID1A_MAST2 fusion transcript (Figure 4a) that might result from translocation without inversion of the ARID1A promoter (1p36.11) to the more centromeric MAST2 locus (1p34.1) Alternatively, this fusion transcript might result from interstitial deletion of those portions of

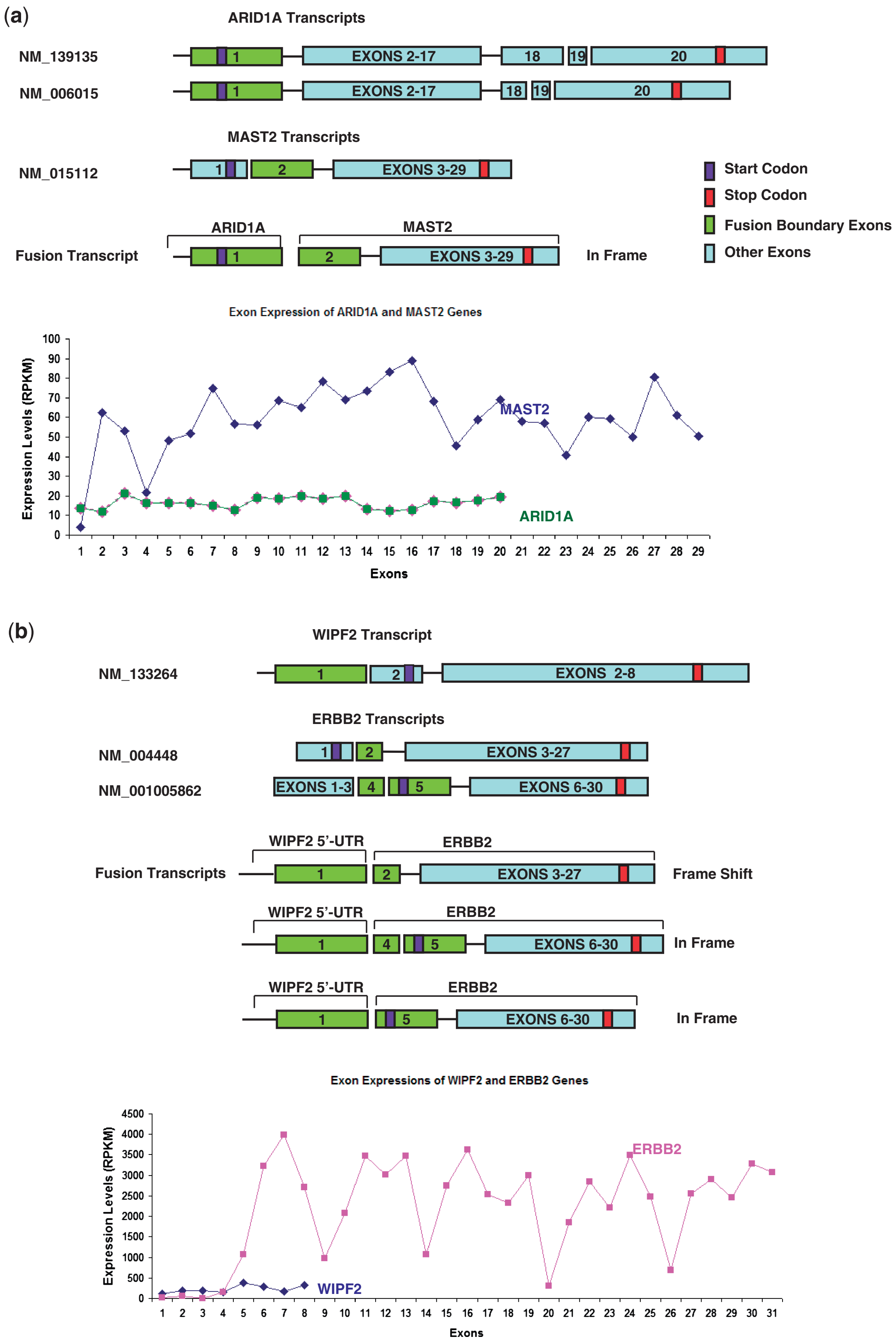

Figure 4. Detailed description of ARID1A_MAST2 (**a**) and WIPF2_ERBB2 (**b**) fusion transcripts. Using the process described in Figure 3, SnowShoes-FTD uses the RNA sequence of all known transcripts of the fusion partners to predict the sequence of all potential in-frame and out of frame fusion transcripts. Abundance of individual exons for each of the fusion partners, normalized to total exon abundance, was extracted from the mRNA-Seq data.

chromosome 1 that intervene between exon 1 of ARID1A (coordinates 26896618) and exon 2 of MAST2 (coordinates 46062691). Juxtaposition of the ARID1A promoter would place control of MAST2 which is downstream of the RB1 pathway, as evidenced by the preponderance of E2F sites in the ARID1A promoter and by the observation that ARID1A is regulated in a cell cycle-dependent manner (17). SnowShoes-FTD predicts that in-frame fusion between ARID1A exon 1 and MAST2 exon 3 will give rise to a chimeric transcript with a predicted open reading frame of 2118 amino acids. The N-terminal 378 amino acids of this hypothetical fusion protein are derived from ARID1A and appear to contain no known or predicted functional domain. Conversely, the C-terminal 1740 amino acids are derived from MAST2 and contain the protein kinase, AGC kinase and PDZ domains of the parental protein. It is likely that this fusion protein has serine/threonine kinase activity. Whether loss of the N-terminal 58 amino acids from MAST2, insertion of the 378 amino acid N-terminus of ARID1A or aberrant expression of MAST2 driven from the ARID1A promoter conveys novel oncogenic potential remains to be determined. However, MAST2 plays an important role in the regulation of NFκB activity downstream of TRAF6 (18), and a SNP within the MAST2 locus has been implicated in breast cancer risk (19). We examined the exon level expressions of the fusion transcript. As shown in Figure 4a, exon 1 expression of MAST2 was significantly lower than the other exons (exon 2–29), which might be due to the fact the exon 1 was fused out. However, there were no obvious expression differences between the exons of the ARID1A gene.

The most provocative chimeric transcript that we have detected involves fusion of the WIPF2 and ERBB2 RNAs. Two isoforms of the fusion were predicted and validated. These chimeric transcripts are expressed in UACC812 cells, which were derived from a HER2+ tumor (20). The WIPF2 locus (also known as WIRE) is located at chr17q21.2 and is transcribed toward the telomere. ERBB2 is located at chr17q11.2, centromeric to WIPF2. Like WIPF2, ERBB2 is transcribed toward the telomere. It is therefore probable that this fusion transcript arises as a result of translocation without inversion of the WIPF2 promoter to give rise to two in-frame transcripts in which the 5′-untranslated region of WIPF2 is fused to one of several 5′-untranslated exons of ERBB2 (Figure 4b). The genomic structure of this hypothetical translocation remains to be verified, but the net result of such an event would be to place ERBB2 expression under control of a promoter that appears, from analysis of potential transcription factor binding sites in the WIPF2 5′ flanking region, to be susceptible to regulation by NFκB, NOTCH and MYC signaling. It is tempting to speculate that this hypothetical promoter swap may account, at least in part, for the observation that ERBB2 transcripts account for about 12 632 tags per million total tags, as determined from our mRNA-Seq data, which translates to about 1.3% of the total polyA+ mRNA pool in UACC812 cells. The observation that there is a dramatic increase in ERBB2 exon expression at the fusion junction (Figure 4b) is consistent with this hypothesis.

SnowShoes-FTD predicted two WIPF2_ERBB2 fusion junctions which were verified in UACC812 cells: WIPF2 chromosomal coordinates 35 629 270 fused to ERBB2 coordinates 35 104 766 or 35 116 768. The latter coordinates fall within the coding sequence of one of the RefSeq variants of ERBB2 mRNA (exon 2 of NM_004448) and would introduce a frame shift mutation in that variant (Figure 4b). However, two of the three predicted fusion sequences (composed of exon 1 of WIPF2 NM_133264 fused to exon 4 or 5 of ERBB2 NM_001005862) would produce transcripts that encode full-length ERBB2 protein (Figure 4b). Unfortunately, it is not possible at this time to determine the sequence of full-length transcripts from mRNA-Seq data. Consequently, it will be necessary to clone and sequence longer cDNA fragments that correspond to the first few hundred nucleotides of the fusion transcript in order to determine which of the hypothetical transcripts are expressed. When we examined the exon expression levels of ERBB2, exons 1–4 are substantially less abundant than downstream exons, suggesting that the transcript with the first 4 exons of ERBB2 fused out might be the more plausible fusion product.

Several cautionary notes warrant emphasis in this discussion of ERBB2 fusion transcripts. We are inclined to believe that such transcripts arise as a result of genomic rearrangement, although our data do not exclude alternative mechanisms. The relative position of the two transcription units on chromosome 17 precludes read through transcription as a potential mechanism to generate the WIPF2_ERBB2 (but not the ARID1A_MAST2) transcripts. However, it is formally possible that some of these chimeric transcripts may arise by trans-splicing events. Genomic sequencing of the relevant loci will be required to rigorously define the mechanism that gives rise to any of these potentially important fusion transcripts. It will also be necessary to determine if there is amplification of the WIPF2-ERBB2 fusion gene in UACC812 cells, if these cells retain intact ERBB2 alleles and if these are also amplified. Until these questions are resolved, it will not be possible to say with any conviction that overexpression of ERBB2 in these cells results from a promoter swap mutation. Nevertheless, we are intrigued by the notion that our analyses may have identified a novel mechanism that accounts for ERBB2 overexpression in HER2+ breast cancer.

DISCUSSION

It has been our experience, as well as that of other investigators, that the predominant analytical issue in detection of fusion transcripts from paired-end mRNA-Seq data is high false detection rate. Different approaches have been proposed to deal with this issue. Wang *et al.* (21) proposed that the fusion events in a certain type of tumor/disease are more likely to arise from genes with similar functions; therefore, one can prioritize fusion candidates based on pathway and/or gene ontology analyses. Others have filtered out the false fusion candidates with homology between fusion partners, with adjacent/overlap partners and/or without the presence of multiple unique junction spanning reads (6,7,12). The SnowShoes-FTD implements similar false positive filtering steps, as well as additional filters including the requirement of the consistencies between the mapping orientations of the end pairs and the two fusion partners, and capping the maximum number of fusion junction points between two fusion partners. Implementation of these features leads to the identification of a subset of high-confidence candidates, as evidenced by our ability to confirm 100% of the putative fusion transcripts described in this study.

The results reported in this manuscript were generated using a set of default parameter values that were defined to minimize the false positive rate. The sensitivity and false positive rate are influenced by all four user-defined parameters. It is very difficult to provide a matrix of the sensitivity versus FDR using different settings of the parameters. For example, by not requiring fusion junction spanning reads (parameter ii), the number of fusion candidates increased from 8 to 50 in MCF7 cells. In addition, if we set parameters to accept candidates with no fusion junction spanning reads, SnowShoes-FTD cannot generate a PCR primer template for quick RT-PCR validation, which makes the validation of the additional fusion candidates difficult. Since the PCR validation is required for accurately calculating the sensitivity and FDR, we chose to give the users the flexibility of adjusting the parameters rather than dictating the input parameters by providing a sensitivity/FDR matrix.

A comparison of our results with those reported for MCF7 in two previous studies reveals a high degree of correspondence, but emphasizes the fact that a significant factor in the outcome of all analyses of this sort resides in the user-defined abundance thresholds for selecting candidates for validation. This observation raises a significant consideration about depth of sequencing. Most of these chimeric transcripts are of low abundance, and the fusion junction point is a small ‘target’. All published fusion gene analysis of breast cancer cells to date have been done with relatively low depth of sequencing, which introduces a certain level of chance that one will, in a given sequencing run, be able to detect two or more tiled fusion spanning reads. We posit that a complete profile of fusion transcripts will likely require a greater depth of sequence analysis than has achieved to date. A more comprehensive profile should emerge from transcriptome sequencing with the newer Illumina Hi-Seq 2000 platform, which routinely yields >100M tags/lane.

Additional sources of variation may include intrinsic differences in the cell lines used in the analyses. It is well known that tumor cell lines exhibit genomic drift, which, in combination with different growth conditions in different laboratories, may result in significant divergence between cell lines maintained in different laboratories. Since genomic instability is probably the driving force behind generation of fusion transcripts, it is plausible that the cells with the greatest genomic instability will not only have the greatest number of fusion transcripts, but are also likely to be more variable in this respect from laboratory to laboratory. Cognizant of this possibility, we elected to use only cell lines that were recently thawed from ATCC reference cultures and maintained in culture for no longer than necessary to isolate RNA for library construction and sequence analysis. Implicit in this line of reasoning is the idea that many of the fusion transcripts that we and others have identified in cell lines may have arisen after isolation of the cell lines and may not reflect the fusion transcript profile that was obtained in the primary tumor from which these cells were initially isolated. If this hypothesis is correct, we would expect to see fewer fusion transcripts in primary tumor samples. Our preliminary analyses suggest that this proposition is likely correct (data not shown).

There are many distinctive features of SnowShoes-FTD compared with other pipelines. Some of these features reflect our decision at the outset to make this pipeline flexible as to input format and user-defined filter parameters, as user friendly as possible, and freely available to any investigator who wishes to use it. In developing SnowShoes-FTD, we were motivated by the concept that downstream analysis of candidate fusion transcripts would require not only a low FDR, but also a set of tools that would make testable predictions about fusion transcript structure and function. For this reason, we developed and incorporated tools that output a number of key features of each fusion transcript. At the validation stage, the most useful of these features is the predicted nucleotide sequence at the fusion junction. This information facilitates rapid design of PCR primers for verification of candidate transcripts. In addition, knowledge of the fusion junction sequence permits relative straightforward output of a number of key structural features of the chimeric RNA, including relationship to known exons and genomic coordinates of the fusion junction. These features are included in the output, making it a rather simple matter to scan the data for chimeric RNAs that may result from either promoter or 3′-UTR swaps. Knowledge of the fusion junction sequence is also critical for predicting frame shifts or substitution mutations that might attend fusion events. We also designed the analysis to provide predicted transcript structures that correspond to all known variants of both the upstream and downstream partners. This information is particularly critical in the case of upstream partners that exhibit alternative promoter utilization (different exon 1 coordinates), as well as evaluation of the relationship between the fusion junction and the translational start site of downstream partners (namely, WIPF2_ERBB2). This output defines the scope of alternative forms of the fusion transcript and, in cases of multiple forms of the downstream partner, may identify critical questions that can only be resolved by conventional cDNA cloning and sequencing.

One of the features of fusion transcripts identified here is that the structure of these chimeric RNAs is strikingly non-random. We see a preponderance of transcripts that may arise due to promoter swap mutations as well as a significant number of transcripts in which the C-terminal sequence of the upstream partner as well as the nucleotide sequence of the 3′-UTR has been altered. Similar results were reported by Edgren *et al.* (12). Clearly changing the promoter or the 3′-UTR may have considerable regulatory significance. Altering either the N- or C-terminal amino acid sequence of the fusion protein may affect function in a variety of ways. Based on these considerations, we posit that the non-random nature of the events that we have observed is likely to be due to selection for altered function leading to growth advantage. This prediction must be tested, of course, and experiments to that end are in progress. But based upon what is known about fusion gene products in hematopoietic malignancy, we believe that it is most likely that breast cancer fusion transcripts arise from genomic rearrangements (rather than read through transcription or trans-splicing), that these events are themselves largely random, that some of these rearrangements give rise to fusion transcripts with novel oncogenic properties and that cells that express these RNAs are selected for growth. Analysis of array comparative genomic hybridization data by Edgren *et al.* (12) is consistent with this hypothesis.

Finally, our analysis has focused largely on fusion transcripts that are likely to be translated into fusion proteins that may exhibit altered function. It is likely that frame shift transcripts may be rapidly degraded and therefore difficult to detect at the depth of sequence analysis that we have generated. About half of the fusion events that we have detected appear to impose premature termination of translation. These may be of considerable significance if the open reading frame of the upstream partner encodes a potentially functional protein. From a translational standpoint, fusion transcripts are virtually certain to be tumor specific, and even non-functional chimeric RNAs may be useful as tumor biomarkers. Furthermore, out of frame fusion events effectively silence one allele of two different genes and might contribute to haploinsufficiency or be associated with loss of heterozygosity of one or both of the un-rearranged alleles. We have also noted that several of these putative rearrangements may affect the expression of intragenic microRNAs. For example, the miR-1204-1207 family is intragenic to the 5′-end of PVT1 (8:128,875,961-129,182,407), which is 'fused out' of the GSCMC_PVT1 chimera in HCC1954 cells. This fusion transcript probably results from an intrachromosomal inversion which likely results in deletion of the microRNA loci associated with the rearranged allele. The functional significance of fusion mutations in breast cancer remains to be determined. At this point, it is clear that we now have the analytical ability to begin to look at primary tumor samples and to determine the prevalence, origin, druggability and role of fusion transcripts in the natural history of breast cancer.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

Cell culture, RNA isolation and cDNA library construction was ably carried out by Ms Tiffany R. Baker and Ms Jennifer M. Carr (Mayo Clinic). Sequence analysis was performed under the expert supervision of Ms Shujun Lu (Illumina) and Mr Bruce Eckloff (Mayo Clinic Advanced Genomics Technology Center). Data management and discussion were provided by Drs Irina Khrebtokova and Lu Zhang (Illumina). Y.W.A. wrote the manuscript and is the author of the algorithm; E.A.P., E.A.T., J.-P.A.K. and G.P.S. provided overall guidance to the project, as well as critical input into experimental design and data interpretation. D.R. established the HMEC lines. K.R.K., A.H., S.M., Z.S. and H.-S.C. assisted in database management, development of the analytical pipeline and refinement of the computer code. D.W.W. provided essential input into construction of reference libraries. B.M.N. designed and carried out all validation protocols. Y.W.A., E.A.T., E.A.P., J.-P.A.K., K.R.K., A.H., S.M., Z.S., H.S.C. and B.M.N. declare no competing financial interests. G.P.S. and D.W.W. are full-time employees of Illumina, Inc. We declare that there was no corporate influence on experimental design, interpretation of data, decision to publish or any other aspect of these studies.

FUNDING

26.2 with Donna Foundation; Bankhead Coley program of the State of Florida (10-BG012-1); Mayo Foundation and the Mayo Clinic Comprehensive Cancer Center (CA151160); by a generous gift from James and Donna Barksdale. Funding for open access charge: Grants.

REFERENCES

1. Futreal,P.A., Coin,L., Marshall,M., Down,T., Hubbard,T., Wooster,R., Rahman,N. and Stratton,M.R. (2004) A census of human cancer genes. *Nat. Rev. Cancer*, **4**, 177–183.
2. Tomlins,S.A., Rhodes,D.R., Perner,S., Dhanasekaran,S.M., Mehra,R., Sun,X.W., Varambally,S., Cao,X., Tchinda,J., Kuefer,R. *et al.* (2005) Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer. *Science*, **310**, 644–648.
3. Soda,M., Choi,Y.L., Enomoto,M., Takada,S., Yamashita,Y., Ishikawa,S., Fujiwara,S., Watanabe,H., Kurashina,K., Hatanaka,H. *et al.* (2007) Identification of the transforming EML4-ALK fusion gene in non-small-cell lung cancer. *Nature*, **448**, 561–566.
4. Pflueger,D., Terry,S., Sboner,A., Habegger,L., Esgueva,R., Lin,P.C., Svensson,M.A., Kitabayashi,N., Moss,B.J., MacDonald,T.Y. *et al.* Discovery of non-ETS gene fusions in human prostate cancer using next-generation RNA sequencing. *Genome Res.*, **21**, 56–67.
5. Maher,C.A., Kumar-Sinha,C., Cao,X., Kalyana-Sundaram,S., Han,B., Jing,X., Sam,L., Barrette,T., Palanisamy,N. and Chinnaiyan,A.M. (2009) Transcriptome sequencing to detect gene fusions in cancer. *Nature*, **458**, 97–101.
6. Maher,C.A., Palanisamy,N., Brenner,J.C., Cao,X., Kalyana-Sundaram,S., Luo,S., Khrebtukova,I., Barrette,T.R., Grasso,C., Yu,J. *et al.* (2009) Chimeric transcript discovery by paired-end transcriptome sequencing. *Proc. Natl Acad. Sci. USA*, **106**, 12353–12358.
7. Sboner,A., Habegger,L., Pflueger,D., Terry,S., Chen,D.Z., Rozowsky,J.S., Tewari,A.K., Kitabayashi,N., Moss,B.J., Chee,M.S. *et al.* FusionSeq: a modular framework for finding gene fusions by analyzing paired-end RNA-sequencing data. *Genome Biol.*, **11**, R104.
8. Zhao,Q., Caballero,O.L., Levy,S., Stevenson,B.J., Iseli,C., de Souza,S.J., Galante,P.A., Busam,D., Leversha,M.A., Chadalavada,K. *et al.* (2009) Transcriptome-guided

characterization of genomic rearrangements in a breast cancer cell line. *Proc. Natl Acad. Sci. USA*, **106**, 1886–1891.

9. Guffanti,A., Iacono,M., Pelucchi,P., Kim,N., Solda,G., Croft,L.J., Taft,R.J., Rizzi,E., Askarian-Amiri,M., Bonnal,R.J. *et al.* (2009) A transcriptional sketch of a primary human breast cancer by 454 deep sequencing. *BMC Genomics*, **10**, 163.
10. Huh,K.W., DeMasi,J., Ogawa,H., Nakatani,Y., Howley,P.M. and Munger,K. (2005) Association of the human papillomavirus type 16 E7 oncoprotein with the 600-kDa retinoblastoma protein-associated factor, p600. *Proc. Natl Acad. Sci. USA*, **102**, 11492–11497.
11. Barlund,M., Monni,O., Weaver,J.D., Kauraniemi,P., Sauter,G., Heiskanen,M., Kallioniemi,O.P. and Kallioniemi,A. (2002) Cloning of BCAS3 (17q23) and BCAS4 (20q13) genes that undergo amplification, overexpression, and fusion in breast cancer. *Genes Chromosomes Cancer*, **35**, 311–317.
12. Edgren,H., Murumaegi,A., Kangaspeska,S., Nicorici,D., Hongisto,V., Kleivi,K., Rye,I.H., Nyberg,S., Wolf,M., Boerresen-Dale,A.L. *et al.* Identification of fusion genes in breast cancer by paired-end RNA-sequencing. *Genome Biol.*, **12**, R6.
13. Sun,Z., Asmann,Y.W., Kalari,K.R., Bot,B., Eckel-Passow,J.E., Baker,T.R., Carr,J.M., Khrebtukova,I., Luo,S., Zhang,L. *et al.* (2011) Integrated analysis of gene expression, CpG island methylation, and gene copy number in breast cancer cells by deep sequencing. *PLoS One*, **6**, e17490.
14. Li,H. and Durbin,R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, **25**, 1754–1760.
15. Li,H., Handsaker,B., Wysoker,A., Fennell,T., Ruan,J., Homer,N., Marth,G., Abecasis,G. and Durbin,R. (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.
16. Langmead,B., Trapnell,C., Pop,M. and Salzberg,S.L. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.*, **10**, R25.
17. Nagl,N.G. Jr, Wang,X., Patsialou,A., Van Scoy,M. and Moran,E. (2007) Distinct mammalian SWI/SNF chromatin remodeling complexes with opposing roles in cell-cycle control. *EMBO J.*, **26**, 752–763.
18. Xiong,H., Li,H., Chen,Y., Zhao,J. and Unkeless,J.C. (2004) Interaction of TRAF6 with MAST205 regulates NF-kappaB activation and MAST205 stability. *J. Biol. Chem.*, **279**, 43675–43683.
19. Wang,X., Fredericksen,Z.S., Vierkant,R.A., Kosel,M.L., Pankratz,V.S., Cerhan,J.R., Justenhoven,C., Brauch,H., Olson,J.E. and Couch,F.J. Association of genetic variation in mitotic kinases with breast cancer risk. *Breast Cancer Res. Treat.*, **119**, 453–462.
20. Meltzer,P., Leibovitz,A., Dalton,W., Villar,H., Kute,T., Davis,J., Nagle,R. and Trent,J. (1991) Establishment of two new cell lines derived from human breast carcinomas with HER-2/neu amplification. *Br. J. Cancer*, **63**, 727–735.
21. Wang,X.S., Prensner,J.R., Chen,G., Cao,Q., Han,B., Dhanasekaran,S.M., Ponnala,R., Cao,X., Varambally,S., Thomas,D.G. *et al.* (2009) An integrative approach to reveal driver gene fusions from paired-end sequencing data in cancer. *Nat. Biotechnol.*, **27**, 1005–1011.