Identifying Novel Expressed Gene Fusions in MCF-7 Cell Line Using Next Generation Sequencing

Onur Sakarya[1], Fiona C Hyland[1], Yulei Wang[1], Heinz Breu[1], Paolo Vatta[1], Liviu Popescu[1], Chieh-Yuan Li[1], Matthew Muller[1], Alec Wong[1], Nriti Garg[1], Sowmi Utiremur[1], Zheng Zhang[1], John P Bodeau[1], Robert C Nutter[1], Mark Mooney[1], Milan Radovich[2], Penn Whitley[3], Kelli Bramlett[3], Francisco de la Vega[1], and Asim Siddiqui[1]

[1] Life Technologies, Foster City, CA, USA, 94404 and [3] Ambion, Austin, TX, USA 78744
[2] Indiana University School of Medicine, Indianapolis, IN, USA 46202

ABSTRACT

High throughput RNA sequencing (RNA-Seq) enables detection and quantification of novel transcripts including gene fusions. Gene fusions are potential chemotherapeutic sites such as in the case of BCR/ABL and imatinib. We sequenced the breast cancer cell line MCF-7 using paired-end RNA-Seq protocol with the SOLiD™ 4 System. By using a new gene fusion detection algorithm called SASR (Suffix Array Single Read splice detection) implemented in the BioScope™ software, we called forty gene fusions and validated twenty five of them to be expressed specifically in MCF-7 including four novel inter-chromosomal events. We report most MCF-7 fusion breakpoints on the 5' gene had likely occurred at the early introns (median 23% of gene size) while no bias was observed for the 3' fusion genes' breakpoints. Additionally, we ran TaqMan® assays for select gene fusions on a set of cancer cell lines, and on forty-eight clinical normal and breast tumor samples.

INTRODUCTION

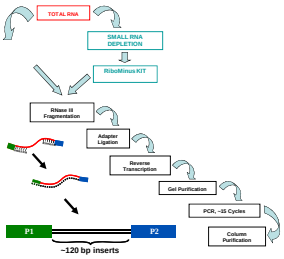
Chromosome aberrations, especially gene fusions, are implicated in the initiation of tumorigenesis. Various gene fusions are important diagnostic and prognostic indicators in leukemia, sarcomas, and other solid tumors.
The high throughput of massively parallel sequencers (up to 1 billion mapped reads on a single run of the SOLiD System) enables genome-wide hypothesis-free detection of gene fusions. The availability of DNA barcoded paired-end reads facilitates concurrent sequencing of many samples, reducing per-sample costs.

MATERIALS AND METHODS

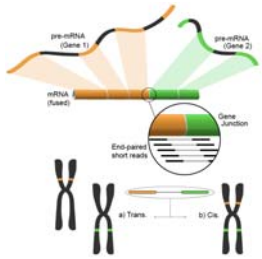
We sequenced one slide of MCF-7 and two slides of UHR (Universal Human Reference) and HBR (Human Brain) with various insert sizes. We used DNA barcoded samples with paired-end SOLiD System sequencing. We predicted 40 fusions in MCF-7 of which 36 were validated by TaqMan® assays and 25 were specific to MCF-7. Gene fusions were called with the BioScope software.

We describe a new suffix array single read (SASR) based intron splicing detection algorithm which allows detection of fusion breakpoints to single base and is not prone to homology based miscalls. In conjunction with the paired-end approach, we devise an integrated, evidence based system to construct splicing graphs and detect fusion transcripts.

RNA-Seq Workflow

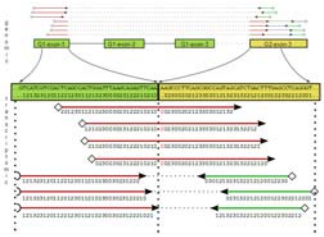


Gene Fusions



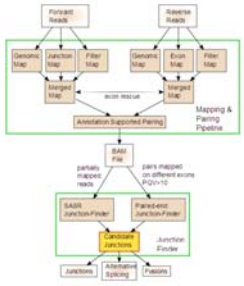
ALGORITHM OVERVIEW

Figure 1. Fusion Detection in color space



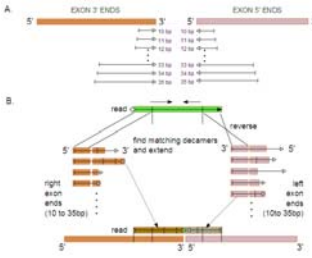
Single read and paired-end evidence for a hypothetical fusion splicing junction is illustrated. Four SOLiD System color space reads that span (split single read), and three reads that bridge (paired-end) the junction are shown. Top chart shows the bird view of the genomic alignments detected for seven pairs of reads between the two exons.

Figure 2 Paired-end Whole Transcriptome Pipeline in BioScope software



Forward and reverse reads were aligned separately, and paired using an algorithm that finds the best pair based on a probabilistic scheme. This algorithm considered mapping mismatches of individual tags and proximity of the pair in comparison to the expected insert size. After the BAM file is generated, another program is used to detect splicing junctions and gene fusions (Figure 2).

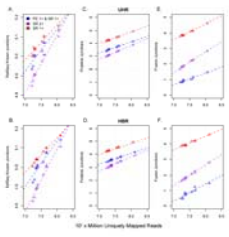
Figure 3. Suffix Array Single Read Splicing Detection



In order to discover single reads that span hypothetical fusion junctions, we constructed a data structure of 12,692,600 left- and right-end suffixes (from hg18 and UCSC gene models) that were sorted lexicographically in arrays allowing logarithmic time string comparison with mismatches (Figure 3)

Increasing coverage (X axis Figure 4) versus A,B: Known RefSeq junctions (A: UHR, B:HBR) C,D: Putative junctions (C:UHR, D:HBR), E,F: Fusion junctions (E: UHR, F: HBR). Three different evidence thresholds were compared: 1) Red line: One single read (SR) evidence required for junction call, 2) Magenta line: Two different start point ST evidence required for junction call, and 3) Blue line: One SR and one paired-end (PE) evidence required for junction call.

Figure 4. Comparison and Effects of Evidence Thresholds On Splice and Fusion Calling



RESULTS

Table 2. List of Validated MCF-7 gene fusions

5' gene-exon	Chr	3' gene-exon	Chr	Distance	MCF7	UHR	HBR	PC-3
ARFGEF2-1	20	SULF2-3	20	Inverted	20.6	24.2	40.0	39.7
SLC25A24-4	1	NBPF6-16	1	Inverted	23.9	27.9	40.0	40.0
USP31-1	16	CRYL1-4	13	Inter-chr	27.5	31.8	40.0	40.0
TBL1XR1-1	3	RGS17-2	6	Inter-chr	26.1	30.6	40.0	40.0
TAF4-1	20	BRIP1-5	17	Inter-chr	25.6	29.2	40.0	40.0
RPS6KB1-6	17	DIAPH3-30	13	Inter-chr	22.6	26.1	40.0	36.7
BCAS4-1	20	BCAS3-24	17	Inter-chr	21.3	25.3	40.0	40.0
AHCYL1-1	1	RAD51C-10	17	Inter-chr	31.0	34.8	40.0	40.0
ABCA5-4	17	PPP4R1L-4	20	Inter-chr	26.1	29.9	40.0	40.0
C16orf45-1	16	ABCC1-15	16	641567	25.3	29.2	40.0	40.0
C16orf62-8	16	IQCK-10	16	264613	26.7	30.5	40.0	40.0
CXorf15-1	X	SYAP1-2	X	-51362	29.1	32.7	40.0	40.0
MYO6-1	6	SENP6-15	6	-70841	28.4	31.9	40.0	40.0
RPS6KB1-2	17	TMEM49-11	17	-72316	24.4	28.4	40.0	39.8
SMARCA4-7	19	CARM1-2	19	-81642	29.9	33.1	40.0	40.0
POP1-2	8	MATN2-15	8	-86928	28.5	31.8	40.0	40.0
GATAD2B-1	1	NUP210L-28	1	-107321	28.3	32.4	40.0	40.0
ESR1-2	6	C6orf97-7	6	-116116	32.3	35.0	40.0	40.0
DEPDC1B-7	5	ELOVL7-8	5	-118895	25.6	29.0	39.8	40.0
ESR1-2	6	C6orf97-6	6	-128831	25.2	29.1	40.0	40.0
GCN1L1-2	12	MSI1-12	12	-157216	25.3	28.2	40.0	39.8
ATXN7L3-1	17	FAM171A2-4	17	-158568	24.8	28.3	40.0	40.0
SYTL2-1	11	PICALM-20	11	-217187	26.7	30.7	40.0	40.0
ADAMTS19-1	5	SLC27A6-10	5	-432137	26.5	31.3	40.0	40.0
ADAMTS19-2	5	SLC27A6-10	5	-433412	25.8	30.5	40.0	40.0

In MCF-7, a total of 40 gene fusions were detected with ~80 million uniquely mapped pairs of reads, of which 11 were detected again with a second sequencing run of 23 million pairs. 25 of the fusions were validated with TaqMan to be specific to MCF-7 and showed 10-fold less expression in UHR, and no expression in HBR (Table 2). Eleven of these fusions (including five inter-chromosomal) were reported by Maher et al or Hampton et al. [1-2]

MCF-7 fusions are enriched between three chromosome hot-spots (Figure 5). Graph drawn using Circos software [3].

Figure 5. MCF-7 Gene Fusion Circular Graph

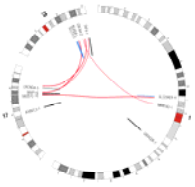
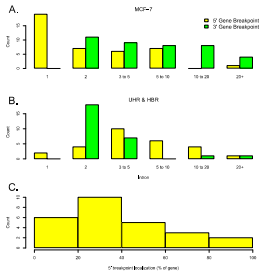
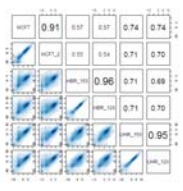


Figure 6. 5' Gene Fusion Breakpoint Bias



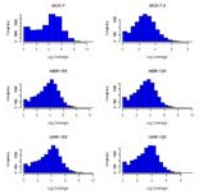
MCF-7 fusion breakpoints were enriched to early 5' introns of the gene (Figure 6 A), but the fusions of UHR and HBR were not (Figure 6 B). On average, first introns of the human genome constitute 22% of the genes and are larger than other introns. We asked whether the large intron size alone would explain the breakpoint bias to 5' introns. The breakpoints fell on average to the first 23% to 36% of the genes (Figure 6 C). In 16 out of 23 fusions, the introns that contained the putative breakpoint were larger than 10KB, which is considerably larger than the human median intron size of 1334 bp

Figure 7. Gene Expression (RPKM) scatter plots



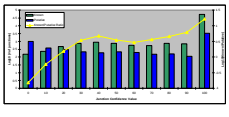
Lower panels show the scatter plot of logarithm (base 2) of gene RPKMs between the six sequencing experiments. Data points are smoothed color density representation of the scatterplot, obtained through a kernel density estimate. Upper panels show the square of the correlation (R2) between two distributions. Technical replicates show high correlation (Figure 7)

Figure 8. Histogram of gene coverage



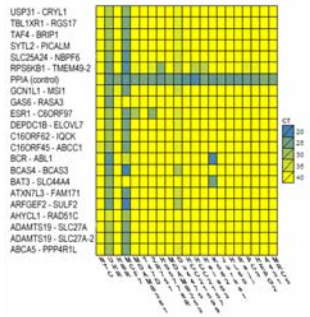
RNA-Seq enables identification of dynamic range of gene expression. Gene coverages were calculated from expressed exons of each gene with the logarithm (base 10) of total expressed reads x read length / total expressed exon length (Figure 8)

Figure 9. Junction Confidence Value



Increasing MapQV threshold improved junction specificity (Figure 9).

Figure 10. Fusion Assays Tested in 20 Cancer Cell Lines



In UHR, we identified the previously described gene fusions including BCR-ABL1, GAS6-RASA3, ARFGEF2-SULF2, NUP214-XKR3 and BAT3-SLC44A4. We prepared TaqMan assays for these three fusions as well as selected MCF-7 gene fusions and tested them in twenty cancer cell lines (Figure 10)

Two 'adjacent gene' fusions were expressed in multiple samples: ESR1-C6ORF97 and RPS6KB1-TMEM49. The fusion between the estrogen receptor alpha gene ESR1 and its neighboring gene C6ORF97 on Chr.6 was expressed in two other ER+ breast cancer cell lines in addition to MCF-7 and Du4475. This fusion may have occurred due to a rearrangement or trans-splicing. We further tested these fusions in 48 clinical normal and tumor breast cancer samples, and ESR1 fusion was found expressed in only one sample.

CONCLUSIONS

Whole transcriptome paired-end sequencing with the SOLiD system v4.0 analyzed with BioScope software allows easy, low-cost, genome-wide sensitive and specific detection of gene fusions, including novel gene fusions. This allows interrogation of large numbers of tumor samples and detection and discovery of biologically important gene fusions.

Comprehensive exon junction detection within genes suggests splice variants and facilitates prediction of alternative splicing. RNA-seq also measures gene expression, and allows quantitation of changes in gene expression between samples with large dynamic range. Strand-specific sequencing disentangles expression of proximate exons on opposite strands.

REFERENCES

1. Maher et al (2009) Transcriptome sequencing to detect gene fusions in cancer, Nature 458(7234):97-101
2. Hampton et al (2009) A sequence-level map of chromosomal breakpoints in the MCF-7 breast cancer cell line yields insights into the evolution of a cancer genome. Genome Research 19:167-177
3. Krzywinski et al (2009) Circos: an information aesthetic for comparative genomics. Genome Research 19(9): 1639-1645

ACKNOWLEDGEMENTS

We thank Benjamin Kong and Caifu Chen (Life Technologies) for designing and running the TaqMan validation of fusion transcripts. We thank Yuandan Lou, Goke Ojewole, Brijesh Krishnaswami and the software team for software integration and verification.

TRADEMARKS/LICENSING