



Chimeric transcript discovery by paired-end transcriptome sequencing

Christopher A. Maher^{a,b}, Nallasivam Palanisamy^{a,b}, John C. Brenner^{a,b}, Xuhong Cao^{a,c}, Shanker Kalyana-Sundaram^{a,b}, Shujun Luo^d, Irina Khrebtukova^d, Terrence R. Barrette^{a,b}, Catherine Grasso^{a,b}, Jindan Yu^{a,b}, Robert J. Lonigro^{a,b}, Gary Schroth^d, Chandan Kumar-Sinha^{a,b}, and Arul M. Chinnaiyan^{a,b,c,e,f,1}

^aMichigan Center for Translational Pathology, Ann Arbor, MI 48109; Departments of ^bPathology and ^cUrology, University of Michigan, Ann Arbor, MI 48109; ^dHoward Hughes Medical Institute and ^eComprehensive Cancer Center, University of Michigan Medical School, Ann Arbor, MI 48109; and ^fIllumina Inc., 25861 Industrial Boulevard, Hayward, CA 94545

Communicated by David Ginsburg, University of Michigan Medical School, Ann Arbor, MI, May 4, 2009 (received for review March 16, 2009)

Recurrent gene fusions are a prevalent class of mutations arising from the juxtaposition of 2 distinct regions, which can generate novel functional transcripts that could serve as valuable therapeutic targets in cancer. Therefore, we aim to establish a sensitive, high-throughput methodology to comprehensively catalog functional gene fusions in cancer by evaluating a paired-end transcriptome sequencing strategy. Not only did a paired-end approach provide a greater dynamic range in comparison with single read based approaches, but it clearly distinguished the high-level “driving” gene fusions, such as *BCR-ABL1* and *TPRSS2-ERG*, from potential lower level “passenger” gene fusions. Also, the comprehensiveness of a paired-end approach enabled the discovery of 12 previously undescribed gene fusions in 4 commonly used cell lines that eluded previous approaches. Using the paired-end transcriptome sequencing approach, we observed read-through mRNA chimeras, tissue-type restricted chimeras, converging transcripts, diverging transcripts, and overlapping mRNA transcripts. Last, we successfully used paired-end transcriptome sequencing to detect previously undescribed ETS gene fusions in prostate tumors. Together, this study establishes a highly specific and sensitive approach for accurately and comprehensively cataloging chimeras within a sample using paired-end transcriptome sequencing.

bioinformatics | gene fusions | prostate cancer | breast cancer | RNA-Seq

One of the most common classes of genetic alterations is gene fusions, resulting from chromosomal rearrangements (1). Intriguingly, >80% of all known gene fusions are attributed to leukemias, lymphomas, and bone and soft tissue sarcomas that account for only 10% of all human cancers. In contrast, common epithelial cancers, which account for 80% of cancer-related deaths, can only be attributed to 10% of known recurrent gene fusions (2–4). However, the recent discovery of a recurrent gene fusion, *TPRSS2-ERG*, in a majority of prostate cancers (5, 6), and *EML4-ALK* in non-small-cell lung cancer (NSCLC) (7), has expanded the realm of gene fusions as an oncogenic mechanism in common solid cancers. Also, the restricted expression of gene fusions to cancer cells makes them desirable therapeutic targets. One successful example is imatinib mesylate, or Gleevec, that targets *BCR-ABL1* in chronic myeloid leukemia (CML) (8–10). Therefore, the identification of novel gene fusions in a broad range of cancers is of enormous therapeutic significance.

The lack of known gene fusions in epithelial cancers has been attributed to their clonal heterogeneity and to the technical limitations of cytogenetic analysis, spectral karyotyping, FISH, and microarray-based comparative genomic hybridization (aCGH). Not surprisingly, *TPRSS2-ERG* was discovered by circumventing these limitations through bioinformatics analysis of gene expression data to nominate genes with marked overexpression, or outliers, a signature of a fusion event (6). Building on this success, more recent strategies have adopted unbiased high-throughput approaches, with increased resolution, for genome-wide detection of chromosomal rearrangements in cancer involving BAC end sequencing (11), fosmid paired-end sequences (12), serial analysis of gene expression

(SAGE)-like sequencing (13), and next-generation DNA sequencing (14). Despite unveiling many novel genomic rearrangements, solid tumors accumulate multiple nonspecific aberrations throughout tumor progression; thus, making causal and driver aberrations indistinguishable from secondary and insignificant mutations, respectively.

The deep unbiased view of a cancer cell enabled by massively parallel transcriptome sequencing has greatly facilitated gene fusion discovery. As shown in our previous work, integrating long and short read transcriptome sequencing technologies was an effective approach for enriching “expressed” fusion transcripts (15). However, despite the success of this methodology, it required substantial overhead to leverage 2 sequencing platforms. Therefore, in this study, we adopted a single platform paired-end strategy to comprehensively elucidate novel chimeric events in cancer transcriptomes. Not only was using this single platform more economical, but it allowed us to more comprehensively map chimeric mRNA, hone in on driver gene fusion products due to its quantitative nature, and observe rare classes of transcripts that were overlapping, diverging, or converging.

Results

Chimera Discovery via Paired-End Transcriptome Sequencing. Here, we employ transcriptome sequencing to restrict chimera nominations to “expressed sequences,” thus, enriching for potentially functional mutations. To evaluate massively parallel paired-end transcriptome sequencing to identify novel gene fusions, we generated cDNA libraries from the prostate cancer cell line VCaP, CML cell line K562, universal human reference total RNA (UHR; Stratagene), and human brain reference (HBR) total RNA (Ambion). Using the Illumina Genome Analyzer II, we generated 16.9 million VCaP, 20.7 million K562, 25.5 million UHR, and 23.6 million HBR transcriptome mate pairs (2×50 nt). The mate pairs were mapped against the transcriptome and categorized as (i) mapping to same gene, (ii) mapping to different genes (chimera candidates), (iii) nonmapping, (iv) mitochondrial, (v) quality control, or (vi) ribosomal (Table S1). Overall, the chimera candidates represent a minor fraction of the mate pairs, comprising $\approx 1\%$ of the reads for each sample.

We believe that a paired-end strategy offers multiple advantages over single read based approaches such as alleviating the reliance on sequencing the reads traversing the fusion junction, increased coverage provided by sequencing reads from the ends of a tran-

Author contributions: C.A.M. and A.M.C. designed research; C.A.M., N.P., J.C.B., X.C., S.L., I.K., T.R.B., R.J.L., G.S., C.K.-S., and A.M.C. performed research; C.A.M., S.L., I.K., R.J.L., and G.S. contributed new reagents/analytic tools; C.A.M., N.P., J.C.B., S.K.-S., C.G., J.Y., R.J.L., G.S., C.K.-S., and A.M.C. analyzed data; and C.A.M., N.P., X.C., C.K.-S., and A.M.C. wrote the paper.

The authors declare no conflict of interest.

Freely available online through the PNAS open access option.

¹To whom correspondence should be addressed. E-mail: arul@umich.edu.

This article contains supporting information online at www.pnas.org/cgi/content/full/0904720106/DCSupplemental.

scribed fragment, and the ability to resolve ambiguous mappings (Fig. S1). Therefore, to nominate chimeras, we leveraged each of these aspects in our bioinformatics analysis. We focused on both mate pairs encompassing and/or spanning the fusion junction by analyzing 2 main categories of sequence reads: chimera candidates and nonmapping (Fig. S24). The resulting chimera candidates from the nonmapping category that span the fusion boundary were merged with the chimeras found to encompass the fusion boundary revealing 119, 144, 205, and 294 chimeras in VCaP, K562, HBR, and UHR, respectively.

Comparison of a Paired-End Strategy Against Existing Single Read Approaches. To assess the merit of adopting a paired-end transcriptome approach, we compared the results against existing single read approaches. Although current RNA sequencing (RNA-Seq) studies have been using 36-nt single reads (16, 17), we increased the likelihood of spanning a fusion junction by generating 100-nt long single reads using the Illumina Genome Analyzer II. Also, we chose this length because it would facilitate a more comparable amount of sequencing time as required for sequencing both 50-nt mate pairs. In total, we generated 7.0, 59.4, and 53.0 million 100-nt transcriptome reads for VCaP, UHR, and HBR, respectively, for comparison against paired-end transcriptome reads from matched samples.

Because the UHR is a mixture of cancer cell lines, we expected to find numerous previously identified gene fusions. Therefore, we first assessed the depth of coverage of a paired-end approach against long single reads by directly comparing the normalized frequency of sequence reads supporting 4 previously identified gene fusions [*TMPRSS2-ERG* (5, 6), *BCR-ABL1* (18), *BCAS4-BCAS3* (19), and *ARFGF2-SULF2* (20)]. As shown in Fig. 1A, we observed a marked enrichment of paired-end reads compared with long single reads for each of these well characterized gene fusions.

We observed that *TMPRSS2-ERG* had a >10-fold enrichment between paired-end and single read approaches. The schematic representation in Fig. 1B indicates the distribution of reads confirming the *TMPRSS2-ERG* gene fusion from both paired-end and single read sequencing. As expected, the longer reads improve the number of reads spanning known gene fusions. For example, had we sequenced a single 36-mer (shown in red text), 11 of the 17 chimeras, shown in the bottom portion of the long single reads, would not have spanned the gene fusion boundary, but instead, would have terminated before the junction and, therefore, only aligned to *TMPRSS2*. However, despite the improved results only 17 chimeric reads were generated from 7.0 million long single read sequences. In contrast, paired-end sequencing resulted in 552 reads supporting the *TMPRSS2-ERG* gene fusion from ≈ 17 million sequences.

Because we are using sequence based evidence to nominate a chimera, we hypothesized that the approach providing the maximum nucleotide coverage is more likely to capture a fusion junction. We calculated an *in silico* insert size for each sample using mate pairs aligning to the same gene, and found the mean insert size of ≈ 200 nt. Then, we compared the total coverage from single reads (coverage is equivalent to the total number of pass filter reads against the read length) with the paired-end approach (coverage is equivalent to the sum of the insert size with the length of each read) (Fig. S2B). Overall, we observed an average coverage of 848.7 and 757.3 MB using single read technology, compared with 2,553.3 and 2,363 MB from paired-end in UHR and HBR, respectively. This increase in ≈ 3 -fold coverage in the paired-end samples compared with the long read approach, per lane, could explain the increased dynamic range we observed using a paired-end strategy.

Next we wanted to identify chimeras common to both strategies. The long read approach nominated 1,375 and 1,228 chimeras, whereas with a paired-end strategy, we only nominated 225 and 144 chimeras in UHR and HBR, respectively. As shown in the Venn diagram (Fig. 1C), there were 32 and 31 candidates common to both

technologies for UHR and HBR, respectively. Within the common UHR chimeric candidates, we observed previously identified gene fusions *BCAS4-BCAS3*, *BCR-ABL1*, *ARFGF2-SULF2*, and *RPS6KB1-TMEM49* (13). The remaining chimeras, nominated by both approaches, represent a high fidelity set. Therefore, to further assess whether a paired-end strategy has an increased dynamic range, we compared the ratio of normalized mate pair reads against single reads for the remaining chimeras common to both technologies. We observed that 93.5 and 93.9% of UHR and HBR candidates, respectively, had a higher ratio of normalized mate pair reads to single reads (Table S2), confirming the increased dynamic range offered by a paired-end strategy. We hypothesize that the greater number of nominated candidates specific to the long read approach represents an enrichment of false positives, as observed when using the 454 long read technology (15, 21).

Paired-End Approach Reveals Novel Gene Fusions. We were interested in determining whether the paired-end libraries could detect novel gene fusions. Among the top chimeras nominated from VCaP, HBR, UHR, and K562, many were already known, including *TMPRSS2-ERG*, *BCAS4-BCAS3*, *BCR-ABL1*, *USP10-ZDHHCH7*, and *ARFGF2-SULF2*. Also ranking among these well known gene fusions in UHR was a fusion on chromosome 13 between *GAS6* and *RAS43* (Fig. S34 and Table S2). The fact that *GAS6-RAS43* ranked higher than *BCR-ABL1* suggests that it may be a driving fusion in one of the cancer cell lines in the RNA pool.

Another observation was that there were 2 candidates among the top 10 found in both UHR and K562. This observation was intriguing, because hematological malignancies are not considered to have multiple gene fusion events. In addition to *BCR-ABL1*, we were able to detect a previously undescribed interchromosomal gene fusion between exon 23 of *NUP214* located at chromosome 9q34.13 with exon 2 of *XKR3* located at chromosome 22q11.1. Both of these genes reside on chromosome 22 and 9 in close proximity to *BCR* and *ABL1*, respectively (Fig. S3B). We confirmed the presence of *NUP214-XKR3* in K562 cells using qRT-PCR, but were unable to detect it across an additional 5 CML cell lines tested (SUP-B15, MEG-01, KU812, GDM-1, and Kasumi-4) (Fig. S3C). These results suggest that *NUP214-XKR3* is a “private” fusion that originated from additional complex rearrangements after the translocation that generated *BCR-ABL1* and a focal amplification of both gene regions.

Although we were able to detect *BCR-ABL1* and *NUP214-XKR3* in both UHR and K562, there was a marked reduction in the mate pairs supporting these fusions in UHR. Although a diluted signal is expected, because UHR is pooled samples, it provides evidence that pooling samples can serve as a useful approach for nominating top expressing chimeras, and potentially enrich for “driver” chimeras.

Previously Undescribed Prostate Gene Fusions. Our previous work using integrative transcriptome sequencing to detect gene fusions in cancer revealed multiple gene fusions, demonstrating the complexity of the prostate transcriptomes of VCaP and LNCaP (15). Here, we exploit the comprehensiveness of a paired-end strategy on the same cell lines to reveal novel chimeras. In the circular plot shown in Fig. S44, we displayed all experimentally validated paired-end chimeras in the larger red circle. We found that all of the previously discovered chimeras in VCaP and LNCaP comprised a subset of the paired-end candidates, as displayed in the inner black circle.

As expected, *TMPRSS2-ERG* was the top VCaP candidate. In addition to “rediscovering” the *USP10-ZDHHCH7*, *HJURP-INPP4A*, and *EIF4E2-HJURP* gene fusions, a paired-end approach revealed several previously undescribed gene fusions in VCaP. One such example was an interchromosomal gene fusion between *ZDHHCH7*, on chromosome 16, with *ABCB9*, residing on chromosome 12, that was validated by qRT-PCR (Fig. S3D). Interestingly, the 5' partner, *ZDHHCH7*, had previously been validated as a complex intrachro-

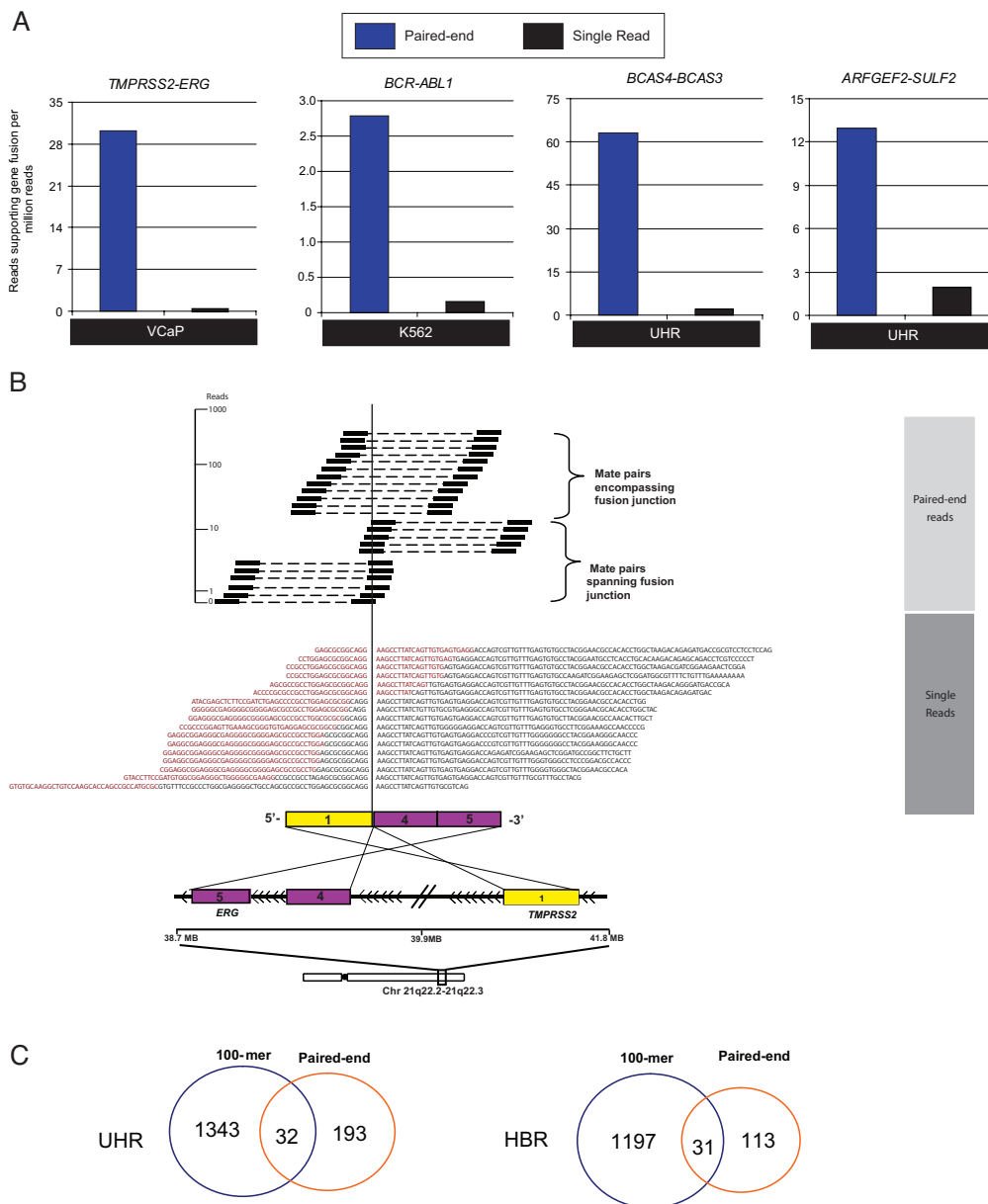


Fig. 1. Dynamic range and sensitivity of the paired-end transcriptome analysis relative to single read approaches. (A) Comparison of paired-end (blue) and long single transcriptome reads (black) supporting known gene fusions *TPRSS2-ERG*, *BCR-ABL1*, *BCAS4-BCAS3*, and *ARFGEF2-SULF2*. (B) Schematic representation of *TPRSS2-ERG* in VCaP, comparing mate pairs with long single transcriptome reads. (Upper) Frequency of mate pairs, shown in log scale, are divided based on whether they encompass or span the fusion boundary; (Lower) 100-mer single transcriptome reads spanning *TPRSS2-ERG* fusion boundary. First 36 nt are highlighted in red. (C) Venn diagram of chimera nominations from both a paired-end (orange) and long single read (blue) strategy for UHR and HBR.

mosomal gene fusion with *USP10* (15). Both fusions have mate pairs aligning to the same exon of *ZDHHC7* (15), suggesting that their breakpoints are in adjacent introns (Fig. S3D).

Another previously undescribed VCaP interchromosomal gene fusion that we discovered was between exon 2 of *TIA1*, residing on chromosome 2, with exon 3 of *DIRC2*, or disrupted in renal carcinoma 2, located on chromosome 3. *TIA1-DIRC2* was validated by qRT-PCR and FISH (Fig. S5). In total, we confirmed an

additional 4 VCaP and 2 LNCaP chimeras (Fig. S6). Overall, these fusions demonstrate that paired-end transcriptome sequencing can nominate candidates that have eluded previous techniques, including other massively parallel transcriptome sequencing approaches.

Distinguishing Causal Gene Fusions from Secondary Mutations. We were next interested in determining whether the dynamic range provided by paired-end sequencing can distinguish known high-

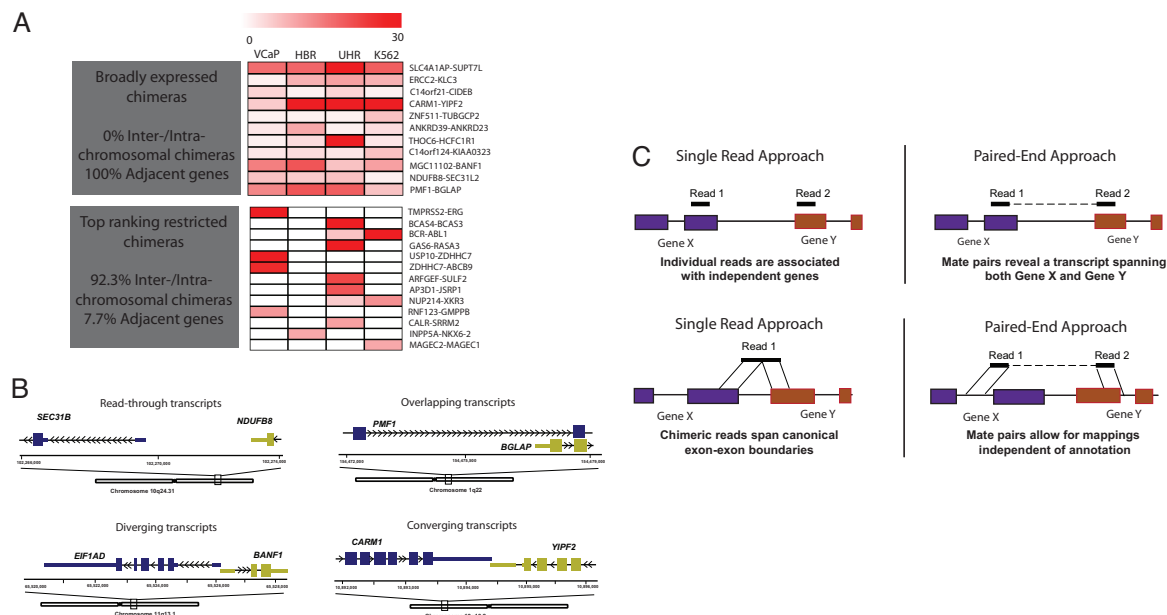


Fig. 2. RNA-based chimeras. (A) Heatmaps showing the normalized number of reads supporting each read-through chimera across samples ranging from 0 (white) to 30 (red). (Upper) The heatmap highlights broadly expressed chimeras in UHR, HBR, VCaP, and K562. (Lower) The heatmap highlights the expression of the top ranking restricted gene fusions that are enriched with interchromosomal and intrachromosomal rearrangements. (B) Illustrative examples classifying RNA-based chimeras into (i) read-throughs, (ii) converging transcripts, (iii) diverging transcripts, and (iv) overlapping transcripts. (C Upper) Paired-end approach links reads from independent genes as belonging to the same transcriptional unit (Right), whereas a single read approach would assign these reads to independent genes (Left). (Lower) The single read approach requires that a chimera span the fusion junction (Left), whereas a paired-end approach can link mate pairs independent of gene annotation (Right).

level “driving” gene fusions, such as known recurrent gene fusions *BCR-ABL1* and *TPRSS2-ERG*, from lower level “passenger” fusions. Therefore, we plotted the normalized mate pair coverage at the fusion boundary for all experimentally validated gene fusions for the 2 cell lines that we sequenced harboring recurrent gene fusions, VCaP and K562. As shown in Fig. S4B, we observed that both driver fusions, *TPRSS2-ERG* and *BCR-ABL1*, show the highest expression among the validated chimeras in VCaP and K562, respectively. This observation suggests a paired-end nomination strategy for selecting putative driver gene fusions among private nonspecific gene fusions that lack detectable levels of expression across a panel of samples (15).

Previously Undescribed Breast Cancer Gene Fusions. Our ability to detect previously undescribed prostate gene fusions in VCaP and LNCaP demonstrated the comprehensiveness of paired-end transcriptome sequencing compared with an integrated approach, using short and long transcriptome reads. Therefore, we extended our paired-end analysis by using breast cancer cell line MCF-7, which has been mined for fusions using numerous approaches such as expressed sequence tags (ESTs) (22), array CGH (23), single nucleotide polymorphism arrays (24), gene expression arrays (25), end sequence profiling (20, 26), and paired-end diTag (PET) (13).

A histogram (Fig. S4C) of the top ranking MCF-7 candidates highlights *BCAS4-BCAS3* and *ARFGEF-SULF2* as the top 2 ranking candidates, whereas other previously reported candidates, such as *SULF2-PRICKLE*, *DEPDC1B-ELÖVL7*, *RPS6KB1-TMEM49*, and *CXorf15-SYAP1*, were interspersed among a comprehensive list of previously undescribed putative chimeras. To confirm that these previously undescribed nominations were not false positives, we experimentally validated 2 interchromosomal and 3 intrachromosomal candidates using qRT-PCR (Fig. S6). Overall, not only was

a paired-end approach able to detect gene fusions that have eluded numerous existing technologies, it has revealed 5 previously undescribed mutations in breast cancer.

RNA-Based Chimeras. Although many of the inter and intrachromosomal rearrangements that we nominated were found within a single sample, we observed many chimeric events shared across samples. We identified 11 chimeric events common to UHR, VCaP, K562, and HBR (Table S3). Via heatmap representation (Fig. 2A) of the normalized frequency of mate pairs supporting each chimeric event, we can observe these events are broadly transcribed in contrast to the top restricted chimeric events. Also, we found that 100% of the broadly expressed chimeras resided adjacent to one another on the genome, whereas only 7.7% of the restricted candidates were neighboring genes. This discrepancy can be explained by the enrichment of inter and intrachromosomal rearrangements in the restricted set.

Unlike, previously characterized restricted read-throughs, such as *SLC45A3-ELK4* (15), which are found adjacent to one another, but in the same orientation, we found that the majority of the broadly expressed chimera candidates resided adjacent to one another in different orientations. Therefore, we have categorized these events as (i) read-throughs, adjacent genes in the same orientation, (ii) diverging genes, adjacent genes in opposite orientation whose 5' ends are in close proximity, (iii) convergent genes, adjacent genes in opposite orientation whose 3' ends are in close proximity, and (iv) overlapping genes, adjacent genes who share common exons (Fig. 2B). Based on this classification, we found 1 read-through, 2 convergent genes, 6 divergent genes, and 2 overlapping genes. Also, we found that ~81.8% of these chimeras had at least 1 supporting EST, providing independent confirmation of the event (Table S3). In contrast to paired-end, single read ap-

proaches would likely miss these instances as each mate would have aligned to their respective genes based on the current annotations (Fig. 2C). Also, these instances may represent extensions of a transcriptional unit, which would not be detectable by a single read approach that identifies chimeric reads that span exon boundaries of independent genes. Overall, we believe that many of these broadly expressed RNA chimeras represent instances where mate pairs are revealing previously undescribed annotation for a transcriptional unit.

Previously Undescribed ETS Gene Fusions in Clinically Localized Prostate Cancer. Given the high prevalence of gene fusions involving ETS oncogenic transcription factor family members in prostate tumors, we applied paired-end transcriptome sequencing for gene fusion discovery in prostate tumors lacking previously reported ETS fusions. For 2 prostate tumors, aT52 and aT64, we generated 6.2 and 7.4 million transcriptome mate pairs, respectively. In aT64, we found that *HERPUD1*, residing on chromosome 16, juxtaposed in front of exon 4 of *ERG* (Fig. 3A), which was validated by qRT-PCR (Fig. S6) and FISH (Fig. 3B), thus identifying a third 5' fusion partner for *ERG*, after *TMPRSS2* (6) and *SLC45A3* (27), and presumably, *HERPUD1* also mediates the overexpression of *ERG* in a subset of prostate cancer patients. Also, just as *TMPRSS2* and *SLC45A3* have been shown to be androgen regulated by qRT-PCR (5), we found *HERPUD1* expression, via RNA-Seq, to be responsive to androgen treatment (Fig. S7). Also, ChIP-Seq analysis revealed androgen binding at the 5' end of *HERPUD1* (Fig. S7).

Also, in the second prostate tumor sample (aT52), we discovered an interchromosomal gene fusion between the 5' end of a prostate cDNA clone, *AX747630* (*FLJ35294*), residing on chromosome 17, with exon 4 of *ETV1*, located on chromosome 7 (Fig. 3C), which was validated via qRT-PCR (Fig. S6) and FISH (Fig. 3D). Interestingly, this fusion has previously been reported in an independent sample found by a fluorescence in situ hybridization screen (27); thus, demonstrating that it is recurrent in a subset of prostate cancer patients. As previously reported, gene expression via RNA-Seq confirmed that *AX747630* is an androgen-inducible gene (Fig. S7). Also, ChIP-Seq revealed androgen occupancy at the 5' end of *AX747630* (Fig. S7).

Discussion

This study demonstrates the effectiveness of paired-end massively parallel transcriptome sequencing for fusion gene discovery. By using a paired-end approach, we were able to rediscover known gene fusions, comprehensively discover previously undescribed gene fusions, and hone in on causal gene fusions. The ability to detect 12 previously undescribed gene fusions in 4 commonly used cell lines that eluded any previous efforts conveys the superior sensitivity of a paired-end RNA-Seq strategy compared with existing approaches. Also, it suggests that we may be able to unveil previously undescribed chimeric events in previously characterized samples believed to be devoid of any known driver gene fusions as exemplified by the discovery of previously undescribed ETS gene fusions in 2 clinically localized prostate tumor samples that lacked known driver gene fusions.

By analyzing the transcriptome at unprecedented depth, we have revealed numerous gene fusions, demonstrating the prevalence of a relatively under-represented class of mutations. However, one of the major goals remains to discover recurrent gene fusions and to distinguish them from secondary, nonspecific chimeras. Although quantifying expression levels is not proof of whether a gene fusion is a driver or passenger, because a low-level gene fusion could still be causative, it still of major significance that a paired-end strategy clearly distinguished known high-level driving gene fusions, such as *BCR-ABL1* and *TMPRSS2-ERG*, from potential lower level passenger chimeras. Overall, these fusions serve as a model for employing a paired-end nomination strategy for prioritizing leads

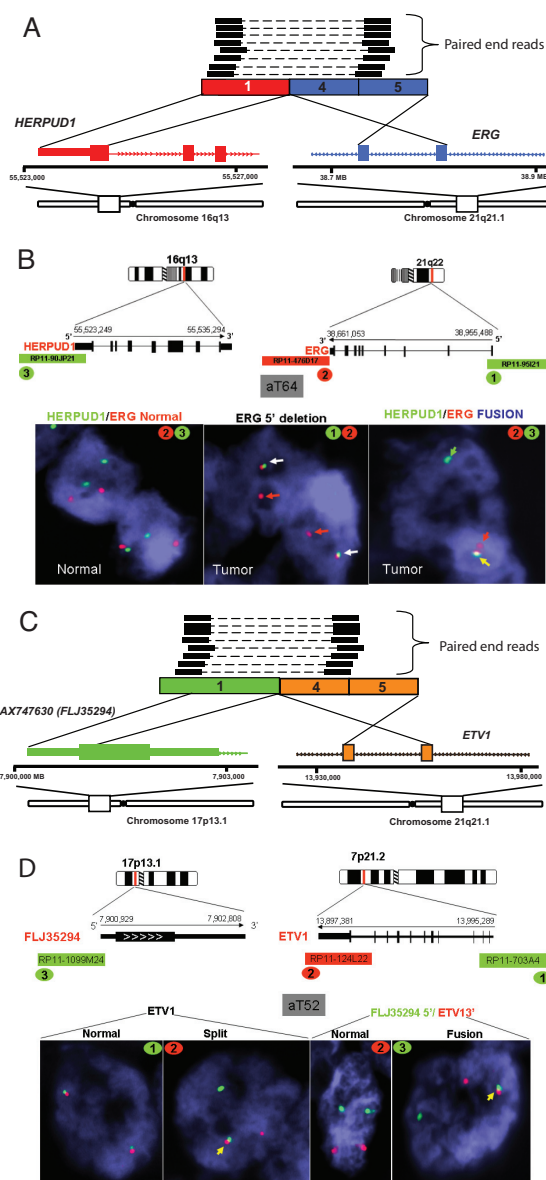


Fig. 3. Discovery of previously undescribed ETS gene fusions in localized prostate cancer. (A) Schematic representation of the interchromosomal gene fusion between exon 1 of *HERPUD1* (red), residing on chromosome 16, with exon 4 of *ERG* (blue), located on chromosome 21. (B) Schematic representation showing genomic organization of *HERPUD1* and *ERG* genes. Horizontal red and green bars indicate the location of BAC clones. (Lower) FISH analysis using BAC clones showing *HERPUD1* and *ERG* in a normal tissue (Left), deletion of the *ERG* 5' region in tumor (Center), and *HERPUD1-ERG* fusion in a tumor sample (Right). (C) Schematic representation of the interchromosomal gene fusion between *FLJ35294* (green), residing on chromosome 17, with exon 4 of *ETV1* (orange) located on chromosome 21. (D Upper) Schematic representation of the genomic organization of *FLJ35294* and *ETV1* genes. (Lower) FISH analysis using BAC clones showing split of *ETV1* in tumor sample (Left) and the colocalization of *FLJ35294* and *ETV1* in a tumor sample (Right).

likely to be high-level driving gene fusions, which would subsequently undergo further functional and experimental evaluation.

One of the major advantages of using a transcriptome approach is that it enables us to identify rearrangements that are not detectable at the DNA level. For example, conventional cytogenetic methods would miss gene fusions produced by paracentric inversions, or sub microscopic events, such as *GAS6-RAS43*. Also, transcriptome sequencing can unveil RNA chimeras, lacking DNA aberrations, as demonstrated by the discovery of a recurrent, prostate specific, read-through of *SLC45A3* with *ELK4* in prostate cancers. Further classification of RNA based events using paired-end sequencing revealed numerous broadly expressed chimeras between adjacent genes. Although these events were not necessarily read-throughs events, because they typically had different orientations, we believe they represent extensions of transcriptional units beyond their annotated boundaries. Unlike single read based approaches, which require chimeras to span exon boundaries of independent genes, we were able to detect these events using paired-end sequencing, which could have significant impact for improving how we annotate transcriptional units.

Overall, we have demonstrated the advantages of employing a paired-end transcriptome strategy for chimera discovery, established a methodology for mining chimeras, and extensively catalogued chimeras in a prostate and hematological cancer models. We believe that the sensitivity of this approach will be of broad impact and significance for revealing novel causative gene fusions in various cancers while revealing additional private gene fusions that may contribute to tumorigenesis or cooperate with driver gene fusions.

Methods

Paired-End Gene Fusion Discovery Pipeline. Mate pair transcriptome reads were mapped to the human genome (hg18) and Refseq transcripts, allowing up to 2 mismatches, using Efficient Alignment of Nucleotide Databases (ELAND) pair within the Illumina Genome Analyzer Pipeline software. Illumina export output files were parsed to categorize passing filter mate pairs as (i) mapping to the same transcript, (ii) ribosomal, (iii) mitochondrial, (iv) quality control, (v) chimera candidates, and (vi) nonmapping. Chimera candidates and nonmapping categories were used for gene fusion discovery. For the chimera candidates category, the following criteria were used: (i) mate pairs must be of high mapping quality (best unique match across genome), (ii) best unique mate pairs do not have a more logical alternative combination (i.e., best mate pairs suggest an interchromosomal rearrangement, whereas the second best mapping for a mate reveals the pair have an alignment within the expected insert size), (iii) the sum of the distances between the most 5' and 3' mate on both partners of the gene fusion must be <500 nt, and (iv) mate pairs supporting a chimera must be nonredundant.

In addition to mining mate pairs encompassing a fusion boundary, the non-mapping category was mined for mate pairs that had 1 read mapping to a gene, whereas its corresponding read fails to align, because it spans the fusion boundary. First, the annotated transcript that the "mapping" mate pair aligned against was extracted, because this transcript represents one of the potential partners involved in the gene fusion. The "nonmapping" mate pair was then aligned against all of the exon boundaries of the known gene partner to identify a perfect partial alignment. A partial alignment confirms that the nonmapping mate pair maps to our expected gene partner while revealing the portion of the nonmapping mate pair, or overhang, aligning to the unknown partner. The overhang is then aligned against the exon boundaries of all known transcripts to identify the fusion partner. This process is done using a Perl script that extracts all possible University of California Santa Cruz (UCSC) and Refseq exon boundaries looking for a single perfect best hit.

Mate pairs spanning the fusion boundary are merged with mate pairs encompassing the fusion boundary. At least 2 independent mate pairs are required to support a chimera nomination, which can be achieved by (i) 2 or more nonredundant mate pairs spanning the fusion boundary, (ii) 2 or more nonredundant mate pairs encompassing a fusion boundary, or (iii) 1 or more mate pairs encompassing a fusion boundary and 1 or more mate pairs spanning the fusion boundary. All chimera nominations were normalized based on the cumulative number of mate pairs encompassing or spanning the fusion junction per million mate pairs passing filter.

RNA Chimera Analysis. Chimeras found from UHR, HBR, VCaP, and K562 were grouped based on whether they showed expression in all samples, "broadly expressed," or a single sample, "restricted expression." Because UHR is comprised of K562, chimeras found in only these 2 samples were also considered as restricted. Heatmap visualization was conducted by using TIGR's MultiExperiment Viewer (TMEV) version 4.0 (www.tm4.org).

Additional Details. Additional details can be found in [SI Text](#).

ACKNOWLEDGMENTS. We thank Lu Zhang, Eric Vermaas, Victor Quijano, and Juying Yan for assistance with sequencing, Shawn Baker and Steffen Durinck for helpful discussions, Rohit Mehra and Javed Siddiqui for collecting tissue samples, and Bo Han and Kalpana Ramnarayanan for technical assistance. C.A.M. was supported by a National Institutes of Health (NIH) Ruth L. Kirschstein postdoctoral training grant, and currently derives support from the American Association of Cancer Research Amgen Fellowship in Clinical/Translational Research and the Canary Foundation and American Cancer Society Early Detection Postdoctoral Fellowship. J.Y. was supported by NIH Grant 1K99CA129565-01A1 and Department of Defense (DOD) Grant PC080665. A.M.C. was supported in part by the NIH (Prostate SPOR P50CA69568, R01 R01CA132874), the DOD (BC075023, W81XWH-08-0110), the Early Detection Research Network (U01 CA111275), a Burroughs Wellcome Foundation Award in Clinical Translational Research, a Doris Duke Charitable Foundation Distinguished Clinical Investigator Award, and the Howard Hughes Medical Institute. This work was also supported by National Center for Integrative Biomedical Informatics Grant U54 DA021519.

- Futreal PA, et al. (2004) A census of human cancer genes. *Nat Rev* 4:177-183.
- Kumar-Sinha C, Tomlins SA, Chinnaiyan AM (2008) Recurrent gene fusions in prostate cancer. *Nat Rev* 8:497-511.
- Mitelman F, Johansson B, Mertens F (2004) Fusion genes and rearranged genes as a linear function of chromosome aberrations in cancer. *Nat Genet* 36:331-334.
- Mitelman F, Mertens F, Johansson B (2005) Prevalence estimates of recurrent balanced cytogenetic aberrations and gene fusions in unselected patients with neoplastic disorders. *Gene Chromosome Canc* 43:350-366.
- Tomlins SA, et al. (2007) Distinct classes of chromosomal rearrangements create oncogenic ETS gene fusions in prostate cancer. *Nature* 448:595-599.
- Tomlins SA, et al. (2005) Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer. *Science* 310:644-648.
- Soda M, et al. (2007) Identification of the transforming EML4-ALK fusion gene in non-small-cell lung cancer. *Nature* 448:561-566.
- Druker BJ, et al. (2006) Five-year follow-up of patients receiving imatinib for chronic myeloid leukemia. *New Engl J Med* 355:2408-2417.
- Druker BJ, et al. (1996) Effects of a selective inhibitor of the Abl tyrosine kinase on the growth of Bcr-Abl positive cells. *Nat Med* 2:561-566.
- Kantarjian H, et al. (2002) Hematologic and cytogenetic responses to imatinib mesylate in chronic myelogenous leukemia. *New Engl J Med* 346:645-652.
- Volik S, et al. (2003) End-sequence profiling: Sequence-based analysis of aberrant genomes. *Proc Natl Acad Sci USA* 100:7696-7701.
- Tuzun E, et al. (2005) Fine-scale structural variation of the human genome. *Nat Genet* 37:727-732.
- Ruan Y, et al. (2007) Fusion transcripts and transcribed retrotransposed loci discovered through comprehensive transcriptome analysis using Paired-End diTags (PETs). *Genome Res* 17:828-838.
- Campbell PJ, et al. (2008) Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. *Nat Genet* 40:722-729.
- Maher CA, et al. (2009) Transcriptome sequencing to detect gene fusions in cancer. *Nature* 458:97-101.
- Marioni JC, Mason CE, Mane SM, Stephens M, Gilad Y (2008) RNA-seq: An assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res* 18:1509-1517.
- Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods* 5:621-628.
- Shtivelman E, Lifshitz B, Gale RP, Canaani E (1985) Fused transcript of abl and bcr genes in chronic myelogenous leukaemia. *Nature* 315:550-554.
- Barlund M, et al. (2002) Cloning of BCAS3 (17q23) and BCAS4 (20q13) genes that undergo amplification, overexpression, and fusion in breast cancer. *Gene Chromosome Canc* 35:311-317.
- Hampton OA, 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 Res* 19:167-177.
- Zhao Q, et al. (2009) Transcriptome-guided characterization of genomic rearrangements in a breast cancer cell line. *Proc Natl Acad Sci USA* 106:1886-1891.
- Hahn Y, et al. (2004) Finding fusion genes resulting from chromosome rearrangement by analyzing the expressed sequence databases. *Proc Natl Acad Sci USA* 101:13257-13261.
- Shadeo A, Lam WL (2006) Comprehensive copy number profiles of breast cancer cell model genomes. *Breast Cancer Res* 8:R9.
- Huang J, et al. (2004) Whole genome DNA copy number changes identified by high density oligonucleotide arrays. *Hum Genom* 1:287-299.
- Neve RM, et al. (2006) A collection of breast cancer cell lines for the study of functionally distinct cancer subtypes. *Cancer Cell* 10:515-527.
- Volik S, et al. (2006) Decoding the fine-scale structure of a breast cancer genome and transcriptome. *Genome Res* 16:394-404.
- Han B, et al. (2008) A fluorescence in situ hybridization screen for E26 transformation-specific aberrations: Identification of DDX5-ETV4 fusion protein in prostate cancer. *Cancer Res* 68:7629-7637.

Supporting Information

Maher et al. 10.1073/pnas.0904720106

SI Text

Results. Effectiveness of paired-end filtering steps. The chimera candidates, comprised of mate pairs that align to different genes, were subjected to a series of filters incorporating insert size, duplicate reads, and ambiguous mappings to reduce potential false positives. To confirm the effectiveness of the filters, we tested 12 candidates that did not pass the filters, and found that they all failed qRT-PCR validation. This suggests these filters are removing false positive nominations.

Paracentric inversion generates novel universal human reference (UHR) gene fusion, GAS6-RASA3. We were intrigued to observe the gene fusion between *GAS6* and *RASA3* residing on chromosome 13. The fact that *GAS6-RASA3* ranked higher than *BCR-ABL1* suggests that it may be a driving fusion in one of the cancer cell lines in the RNA pool. *GAS6* is a gamma-carboxyglutamic acid (Gla)-containing protein believed to stimulate cell proliferation. It resides ≈ 200 MB, in opposite orientation and separated by *FAM70B*, from *RASA3* suggesting that this fusion gene is generated by a small paracentric inversion. *RASA3*, is a member of the GAP1 family of GTPase-activating proteins. Overall, *GAS6-RASA3* may represent one of many novel gene fusions that could shed light into the tumorigenesis of one of the anonymous cancer cell lines within the UHR pool.

Novel interchromosomal VCaP gene fusions, TIA1-DIRC2. One novel VCaP interchromosomal gene fusion found by a paired-end strategy that eluding our integrative transcriptome sequencing approach was between exon 2 of *TIA1*, residing on chromosome 2, with exon 3 of *DIRC2*, or disrupted in renal carcinoma 2, located on chromosome 3. *TIA1-DIRC2* was validated by qRT-PCR and FISH (Fig. S5). The splicing regulator, *TIA1*, is a member of a RNA-binding protein family that has nucleolytic activity against cytotoxic lymphocyte (CTL) target cells and could have a role inducing apoptosis. Also, the disruption of *DIRC2* has been associated with haplo-insufficiency, which could provide a possible mechanism for tumor growth in renal cell carcinoma (1).

Discussion. We attribute the comprehensiveness of a paired-end strategy for gene fusion discovery to the increased coverage provided by sequencing reads from both ends of a fragment, the ability to resolve ambiguous mappings, thus, maximizing the information from the sequences generated, and the lack of reliance on having to span the fusion junction. In comparison, single read approaches using short reads (36 nt) are limited not only by requiring it to span the fusion junction, but with enough sequence on each side to confidently identify the fusion partners. Although long transcriptome reads are highly desirable to provide sequence specificity when aligning to a reference genome, a 454 based approach is limited by the depth of coverage. Therefore, it is not surprising that many of the novel paired-end gene fusions, such as *TIA1-DIRC2* or *ZDHHC7-ABCB9*, eluded an integrative transcriptome sequencing approach. However, to circumvent this issue, we unveiled one of the first long single read (100 nt) runs generated by the Illumina platform. Despite offering a deeper coverage of the transcriptome, compared with previous long single read approaches such as expressed sequence tags (ESTs) or 454 long reads, we still observed an increased dynamic range by paired-end sequencing. Also, our results demonstrate that, despite the slightly longer time, it takes to generate 2×50 -nt paired-end over 100-nt transcriptome reads, the paired-end data resulted in 3-fold greater nucleotide coverage. Overall, for comparable resources of generating long single

reads, paired-end sequencing provides a more comprehensive catalog of gene fusions within a given sample.

Materials and Methods. Samples and cell lines. VCaP cell line was derived from a vertebral metastasis from a patient with hormone-refractory metastatic prostate cancer (2), and was provided by Ken Pienta (University of Michigan, Ann Arbor, MI). LNCaP or VCaP cells were starved in phenol red free media supplemented with charcoal-dextran filtered FBS and 5% penicillin/streptomycin for 48 h before the addition of 1 nM synthetic androgen (R1881) as indicated. RNA was then isolated using the microRNAeasy kit (Qiagen) according to the manufacturer's instructions. Prostate tissues were obtained from the radical prostatectomy series at the University of Michigan and from the Rapid Autopsy Program (3), University of Michigan Prostate Cancer Specialized Program of Research Excellence (SPORE) Tissue Core. All samples were collected with informed consent of the patients and prior approval of the institutional review board. K562, SUP-B15, MEG-01, KU812, GDM-1, and Kasumi-4 cell lines were obtained from American Type Culture Collection (ATCC). UHR was obtained from Strategene. Human brain RNA (HBR) was obtained from Ambion.

Sequence datasets. Human genome build 18 (hg18) was used as a reference genome. All Refseq and University of California Santa Cruz (UCSC) transcripts were downloaded from the UCSC genome browser (<http://genome.ucsc.edu/>). Sequences of previously identified *TMPRSS2-ERG* fusion transcript (GenBank accession no. DQ204772) and *BCR-ABL1* fusion transcript (GenBank accession no. M30829) were used for reference. Previously validated prostate gene fusion chimaeras were extracted using GenBank accession nos. FJ423742–FJ423755.

Paired-end transcriptome sequencing using Illumina Genome Analyzer II. Messenger RNA (1 μ g) was fragmented at 70 °C for 2 min in a fragmentation buffer (Ambion) and converted to single-stranded cDNA using SuperScript II reverse transcriptase (Invitrogen), followed by second-strand cDNA synthesis using *Escherichia coli* DNA polymerase I (Invitrogen). The double-stranded cDNA was further processed by Illumina mRNA-sequencing Prep kit. Briefly, double-stranded cDNA was end repaired by using T4 DNA polymerase and T4 polynucleotide kinase, monoadenylated using a Klenow DNA polymerase I (3' to 5' exonuclease activity), and ligated with adaptor oligo mix (Illumina) using T4 DNA ligase. The adaptor-ligated cDNA library was then fractionated on a 4% agarose gel, and a smear corresponding to ≈ 300 nt was excised, purified, and PCR amplified (15 cycles) by Pfu polymerase (Stratagene). The PCR product was again size selected on a 4% agarose gel by cutting out the library smear at 300 base pairs. The library was then purified with the Qiaquick Minelute PCR Purification Kit (Qiagen) and quantified with the Agilent DNA 1000 kit on the Agilent 2100 Bioanalyzer following the manufacturer's instructions. Library (10 nM) was used to prepare flowcells with $\approx 100,000$ – $130,000$ clusters per lane for analysis on the Illumina Genome Analyzer II.

Long transcriptome read gene fusion discovery. All 100-nt passing filter transcriptome reads generated from the Illumina sequencing platform were processed similar to the method described for detecting chimeras from 454 reads (4). All chimera nominations were normalized based on the total number reads spanning the fusion junction per million reads passing filter.

Comparison of single transcriptome reads with paired-end approach. As the 100-nt single transcriptome reads were aligned against only

Refseq transcripts to identify chimeras spanning exon–exon boundaries, only those paired-end chimera nominations that had supporting evidence of an exon–exon fusion junction were used for comparison.

RNA chimera classification. Chimeras between adjacent genes were categorized based on their orientation to one another and whether they are overlapping. The categories are (i) read-throughs, adjacent genes in the same orientation, (ii) diverging genes, adjacent genes in opposite orientation whose 5' sites are in close proximity, (iii) convergent genes, adjacent genes whose 3' ends are in close proximity, and (iv) overlapping genes, adjacent genes who share common exons. Genes are defined as overlapping if they have even 1 nt overlapping.

Real-time PCR validation. Quantitative PCR was performed using Power SYBR Green Mastermix (Applied Biosystems) on an Applied Biosystems Step One Plus Real Time PCR System as described (5). All oligonucleotide primers were synthesized by Integrated DNA Technologies. *GAPDH* (6) primer was as described. All assays were performed in duplicate or triplicate, and results were plotted as average fold change relative to *GAPDH*.

FISH. FISH hybridizations were performed on VCaP and prostate tumor samples. BAC clones were selected from the UCSC genome browser. After colony purification, midi prep DNA was prepared using QiagenTips-100 (Qiagen). DNA was labeled by nick translation labeling with biotin-16-dUTP and digoxigenin-11-dUTP (Roche). Probe DNA was precipitated and dissolved in hybridization mixture containing 50% formamide, 2× SSC, 10% dextran sulfate, and 1% Denhardt's solution. Approximately 200 ng of labeled probes was hybridized to normal human chromosomes to confirm the map position of each BAC clone. FISH signals were obtained using anti digoxigenin-fluorescein and alexa fluor594 conjugate for green and red colors, respectively. Fluorescence images were captured using a high resolution CCD camera controlled by ISIS image processing software (Metasystems).

ChIP-Seq analysis. ChIP from the cultured cells was carried out as previously described Yu et al. (7), using antibodies against AR (no. 06-680; Millipore), ERG (no. sc354; Santa Cruz), and rabbit IgG (no. sc-2027; Santa Cruz). ChIP samples were prepared for sequencing using the Genomic DNA sample prep kit (Illumina) following manufacturers protocols. The raw sequencing image data were analyzed by the Illumina analysis pipeline, aligned to the unmasked human reference genome (NCBI v36, hg18) using the ELAND software (Illumina) to generate sequence reads of 25–32 bps. These short reads were subsequently analyzed using HPeak (<http://www.sph.umich.edu/csg/qin/HPeak/>). Statistically significant peaks, representing binding regions, were exported into wiggle files for visualization in the UCSC genome browser.

Calculating gene expression from RNA-Seq data. Transcriptome reads were trimmed to 32 nt by removing the first 2 bases and sufficient bases from the end necessary to yield a 32 mer. The 32-mer reads were aligned to the human genome plus 54-mer splice junctions generated by concatenating 28 bases from the end of the 5' and 3' splicing partner. This ensures that reads that map to the splice junction overlap the splice junction by 4 bases (8). The reads were aligned using Bowtie and allowing up to 2 bases of mismatch. Reads that did not yield a unique best hit, were discarded. Gene expression was calculated by first summing the coverage over all of the positions included in any isoform of the gene that is included in the UCSC mRNA dataset and then dividing by the number of positions included in the sum to yield the average coverage for the gene (9). Next, the average coverage was normalized by the number of reads mapping to the human genome in the sample and then multiplied by 1 million to yield a gene expression value in reads per kilobase million (RPKM). **Establishment of mate-pair filtering steps.** The criteria for filtering mate pairs encompassing a fusion boundary were selected for the following reasons. First, because our initial chimera candidates were derived from mappings against known transcripts, it is likely they have multiple alignments to the genome that do not correspond to an annotated transcript. Therefore, a mate pair was discarded if either of the mates failed to have a single unique best hit against the genome. If the mate pair does reveal single best hits, we iterate through secondary mappings to ensure none of those reveal a mate pair combination that is in agreement with the expected insert size as this represents a more logical event. In addition to having a secondary hit residing approximately the insert size away on the same transcript, we also filter candidates within 50,000 kb on the genome, presuming this alignment does not overlap a different gene. For the remaining candidates, we established a filter that leverages the insert size between the mate pairs. We expect that if multiple mate pairs were to support the same fusion event, their mappings will aggregate within the region flanking the fusion junction. We calculated an *in silico* insert size for each sample using mate pairs aligning to the same gene and found the mean size of ≈200 nt. Therefore, we expect that if 2 mate pairs were both encompassing the same breakpoint, the furthest apart that they could reside from one another would have to be nearly equivalent to the insert size. Next, we observed that some candidates had identical mate pair reads that were in close proximity on the flow cell. These duplicates were likely an artifact of the analysis pipeline and resulted in the overrepresentation of a subset of chimeras. To circumvent this, for each chimera candidate, we generated a nonredundant set of mate pairs supporting the predicted fusion event. Last, we required that a chimera have a minimum of 2 nonredundant mate pairs, unless there was supporting evidence of a mate pair spanning the fusion junction, to increase our confidence in the nominated event.

1. Bodmer D, et al. (2002) Disruption of a novel MFS transporter gene, DIRC2, by a familial renal cell carcinoma-associated t(2;3)(q35;q21). *Hum Mol Genet* 11:641–649.
2. Korenchuk S, et al. (2001) VCaP, a cell-based model system of human prostate cancer. *In Vivo* 15:163–168.
3. Rubin MA, et al. (2000) Rapid ("warm") autopsy study for procurement of metastatic prostate cancer. *Clin Cancer Res* 6:1038–1045.
4. Maher CA, et al. (2009) Transcriptome sequencing to detect gene fusions in cancer. *Nature* 458:97–101.
5. Tomlins SA, et al. (2007) Distinct classes of chromosomal rearrangements create oncogenic ETS gene fusions in prostate cancer. *Nature* 448:595–599.

6. Vandesompele J, et al. (2002) Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes. *Genome Biol* 3:34.
7. Yu J, et al. (2007) Integrative genomics analysis reveals silencing of beta-adrenergic signaling by polycomb in prostate cancer. *Cancer Cell* 12:419–431.
8. Wang ET, et al. (2008) Alternative isoform regulation in human tissue transcriptomes. *Nature* 456:470–476.
9. Sultan M, et al. (2008) A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. *Science* 321:956–960.

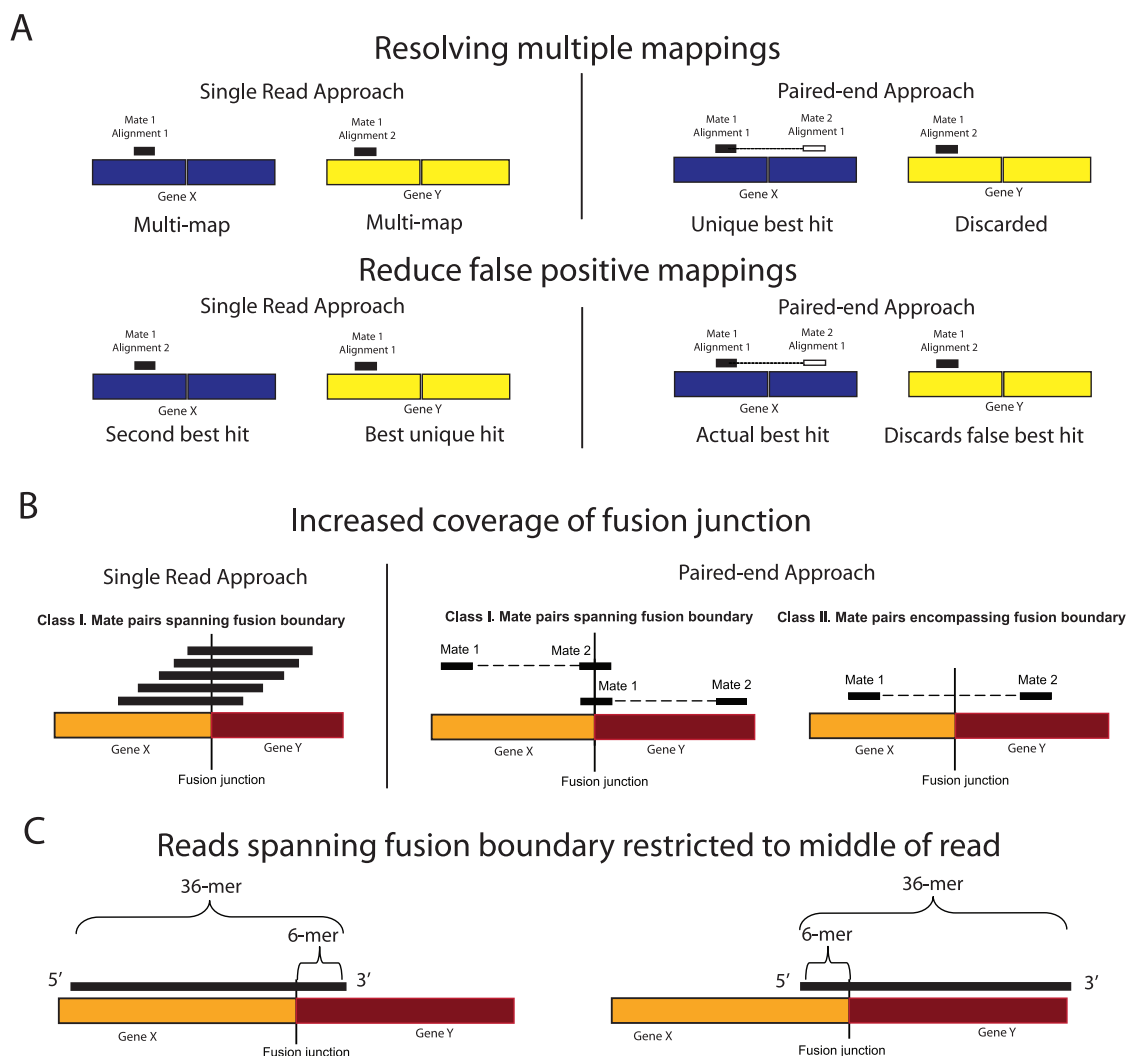
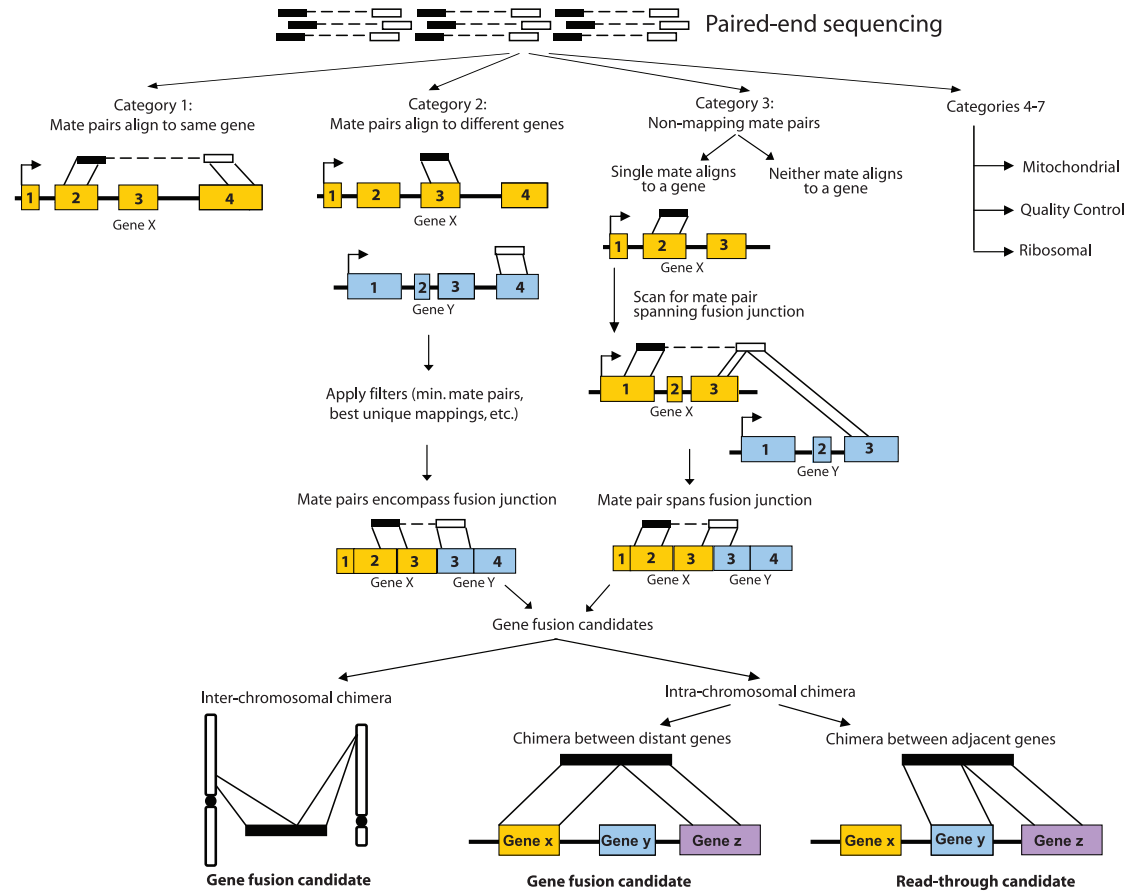


Fig. S1. Paired-end improvements over single-read approach. (A) Paired-end approach resolves ambiguous mappings. (Upper) The single-read approach (Left) displays a single read, or "mate 1," with identical matches to gene X (blue) and gene Y (yellow), thus resulting in this read being classified as having multiple mappings. The paired-end approach (Right) displays the same read as the single-read approach aligning to gene X and gene Y. However, the corresponding mate pair, or "mate 2," aligns with the expected insert size to gene X, but not gene Y. Therefore, using a paired-end approach, we obtain a unique best hit to gene X that would have been ambiguous using a single-read approach. (Lower) Mate 1 shows a best unique hit to gene Y, and a second best hit to gene X, based on single-read approach (Left). However, the second mate, using paired-end (Right), reveals a best unique hit to gene X, revealing the actual best hit. (B) Paired-end sequencing increases coverage spanning fusion junction. Although a single-read approach can detect gene fusions solely by spanning the fusion junction (Left), a paired-end approach can detect a chimera if a mate pairs spans the fusion junction or if the mate pairs encompass the fusion junction (Right), thus providing more opportunity for chimera discovery. (C) Limitation of single-read spanning fusion junction. A 36-mer spanning a chimera typically necessitates a minimum of 6-nt (depending on the stringency) to align to each partner of the fusion event. Therefore, for a given 36-mer, there are only 24 possible alignments that could capture the fusion junction.

A



B

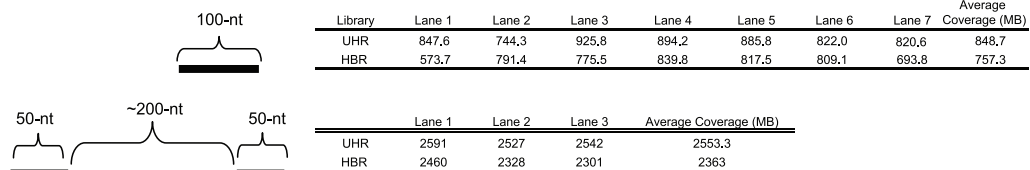


Fig. S2. Employing paired-end transcriptome sequencing for chimera discovery. (A) Schematic representation of bioinformatics methodology for using paired-end transcriptome sequencing to identify chimeric transcripts. Sample cDNA libraries are sequenced, using the Illumina Genome Analyzer II, and subsequently aligned against the human transcriptome. The mate pairs are classified into the following categories (i) mate pairs align to same gene, (ii) mate pairs align to different genes (chimera candidates), (iii) nonmapping, (iv) mitochondrial, (v) ribosomal, and (vi) quality control. Chimeras found in category 2 undergo additional processing to reduce false positives. The nonmapping mate pairs are further classified based on whether (i) they both fail to map to a gene or (ii) only a single mate read fails to align to a gene. The latter group represents candidate mate pairs in which one mate spans the fusion boundary; thus eluding any alignment to an annotated gene. These nonmapping reads are, therefore, mined for potential chimeras. Next, we merge the candidates where one mate spans the fusion boundary with our mate pairs encompassing the fusion boundary (from category 2) and subsequently classify them as intrachromosomal or interchromosomal, based on their location on the same or different chromosomes, respectively. (B) Coverage statistics for UHR and HBR paired-end and long transcriptome read approaches distributed by lane.

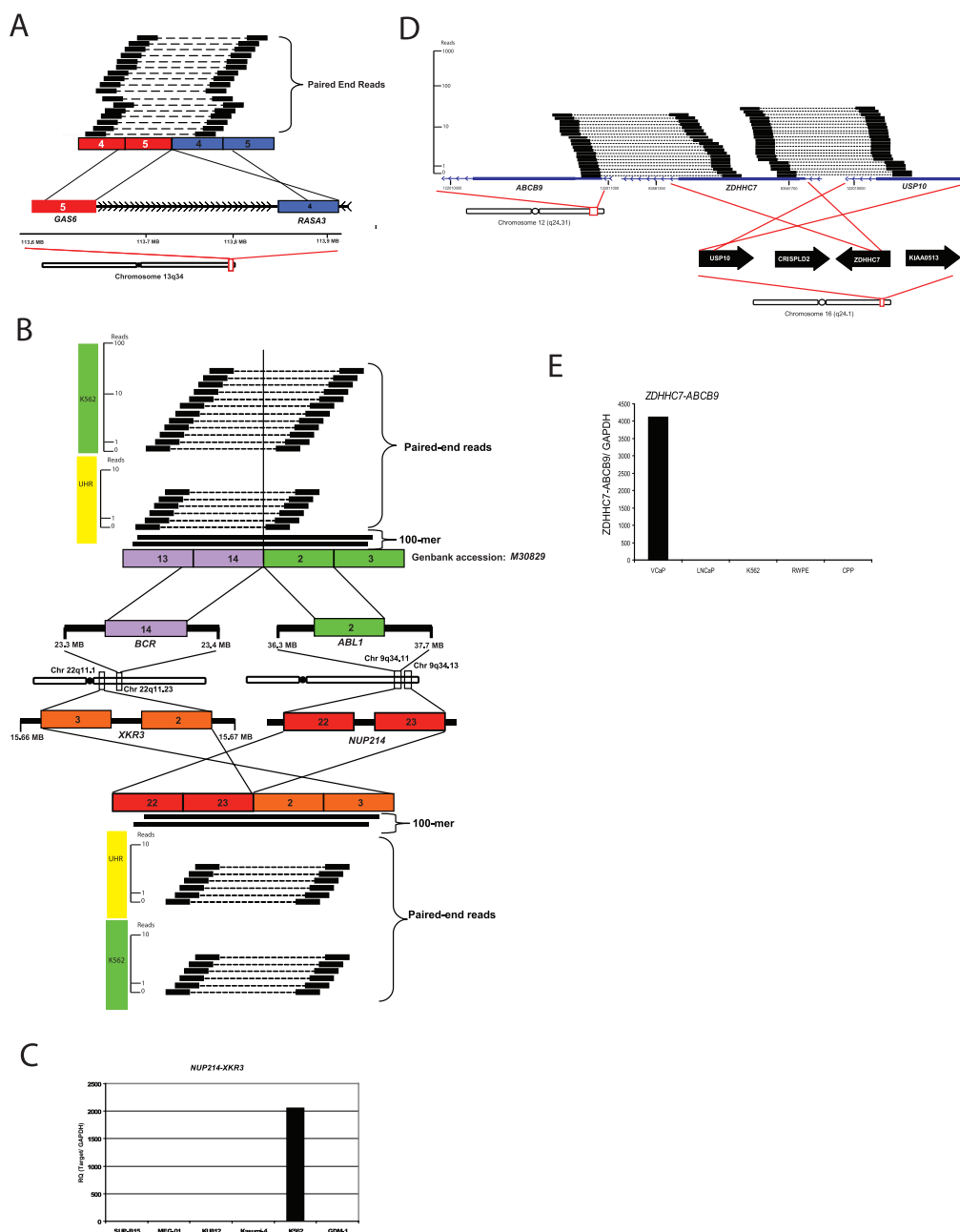


Fig. S3. Novel paired-end schematics and experimental validation. (A) Schematic representation of the UHR paracentric inversion on chromosome 13q34 generating the gene fusion between exon 5 of *GAS6* (red) and exon 4 of *RASA3* (blue). (B) Novel hematological gene fusion *NUP214-XKR3*. Schematic representation of *BCR-ABL1* and *NUP214-XKR3* interchromosomal gene fusions between chromosomes 9 and 22. Representative distributions of mate pairs and long single reads are shown on log scale for both UHR (yellow) and K562 (green). (C) Histogram of qRT-PCR validation of the *NUP214-XKR3* transcript across chronic myeloid leukemia cell lines. (D) Novel complex interchromosomal rearrangement *ZDHHC7-ABC9*. Schematic representation of the intrachromosomal rearrangement of *USP10-ZDHHC7* and the interchromosomal gene fusion, *ZDHHC7-ABC9*. (E) Histogram of qRT-PCR validation of the *ZDHHC7-ABC9* transcript.

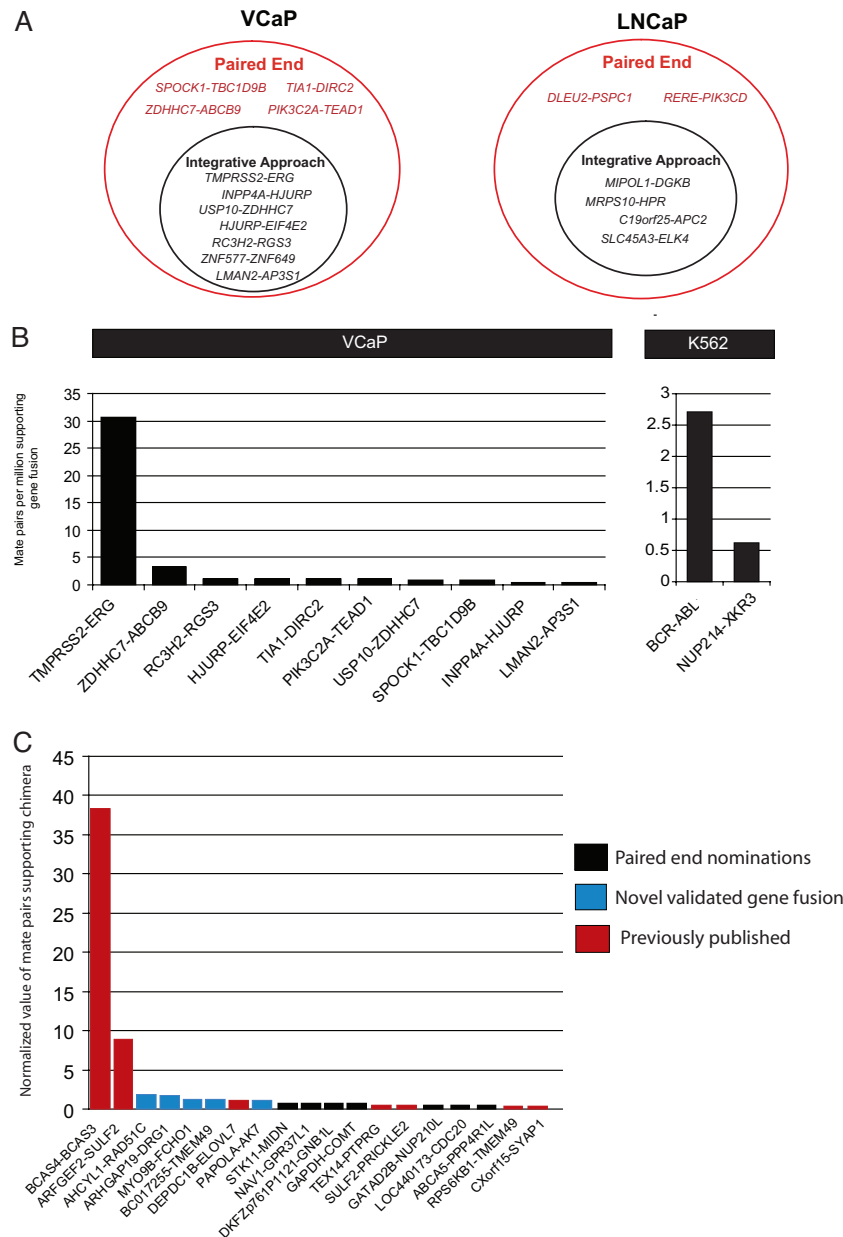
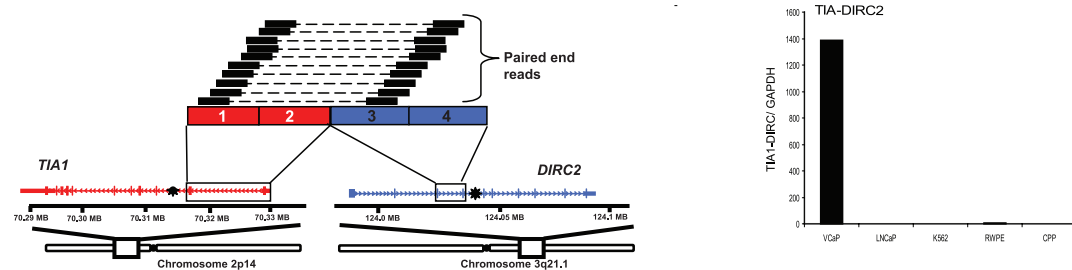


Fig. S4. Comprehensiveness of paired-end transcriptome analysis. (A) Venn diagram to highlight the overlap between paired-end gene fusion discovery and the previously reported integrated approach applied to VCaP (Left) and LNCaP (Right). Larger circle (red) encompasses all experimentally validated chimeras nominated by paired-end sequencing. The inner circle (black) demonstrates that all previously validated chimeras, previously reported by the integrated approach, are a subset of the paired-end nominations. (B) Histogram of the experimentally validated chimeras in VCaP and K562 highlighting the distinction between known recurrent gene fusions *TMPRSS2-ERG* and *BCR-ABL1* from secondary gene fusions within their respective cell lines. (C) Comprehensive detection of chimeras in MCF-7 using paired-end transcriptome sequencing. Histogram highlights the normalized frequency of mate pairs supporting each predicted chimera. Previously reported gene fusions are shown in red, novel experimentally validated chimeras are shown in blue, and putative chimeras that have not yet been experimentally tested are shown in black.

A



B

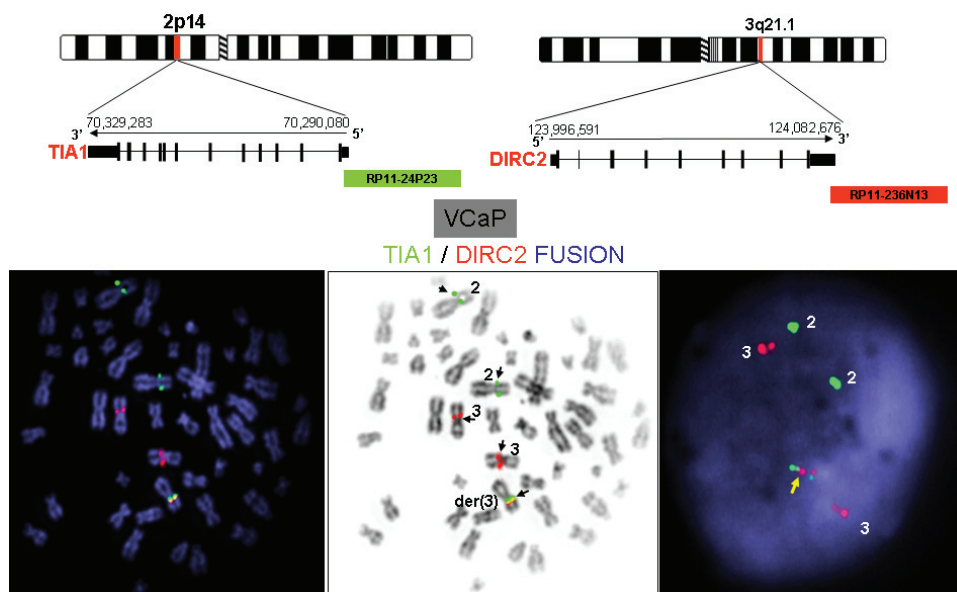


Fig. 55. Validation of novel VCaP interchromosomal gene fusion *TIA1-DIRC2*. (A) Schematic representation of the VCaP interchromosomal gene fusion between *TIA1* (red) residing on chromosome 2 with *DIRC2* (blue) located on chromosome 3. Inset displays histogram of qRT-PCR validation of the *TIA1-DIRC2* transcript. (B) Schematic representation showing genomic organization of *TIA1* and *DIRC2* genes. Horizontal red and green bars indicate the location of BAC clones (Upper). FISH analysis using BAC clones showing the fusion of *TIA1* and *DIRC2* genes on a marker chromosome (yellow) (Lower).

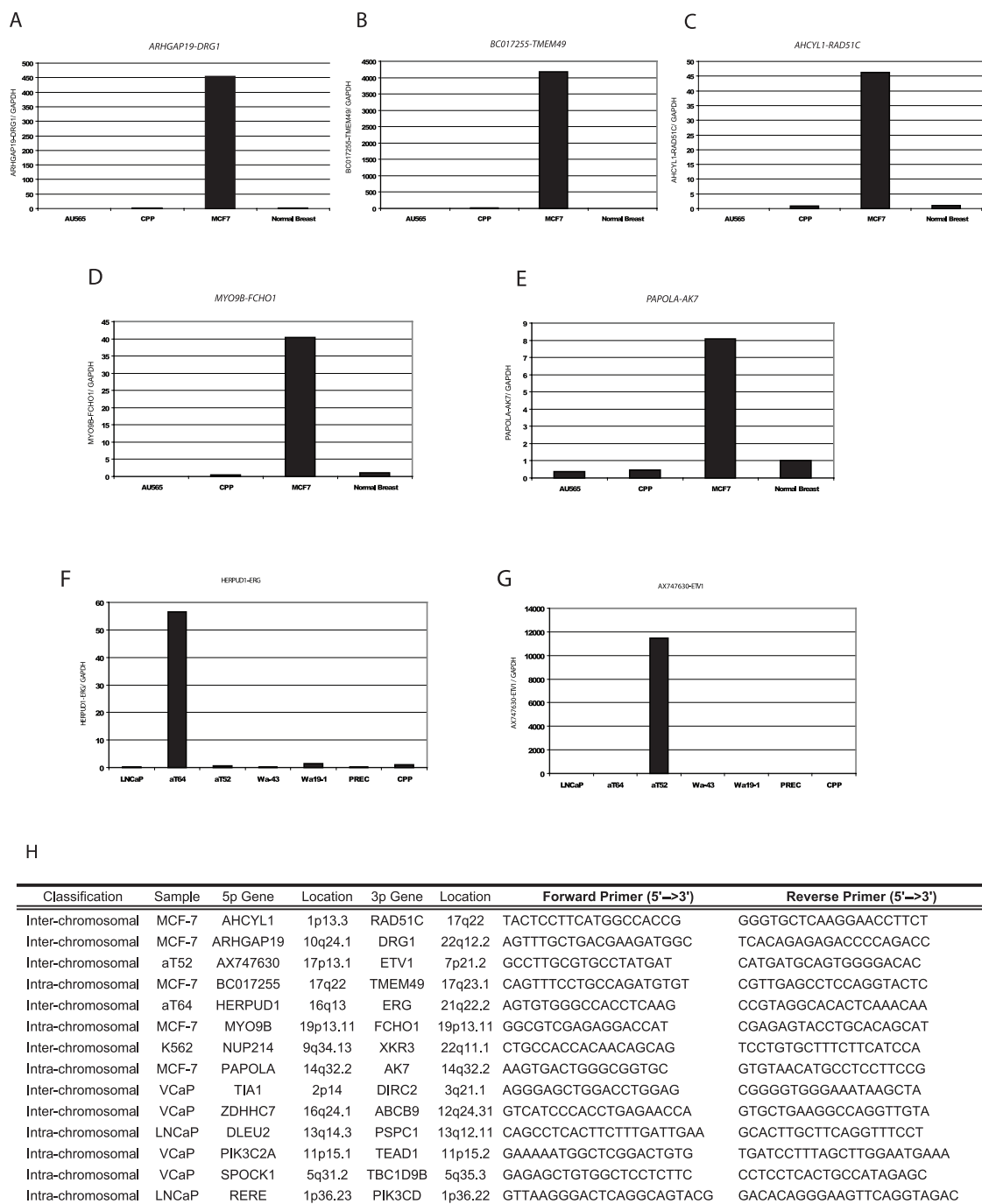


Fig. S6. Experimental validation of novel chimeras. Quantitative RT-PCR validation of novel paired end nominations (A) *ARHGAP19-DRG1*, (B) *BC017255-TMEM49*, (C) *AHCYL1-RAD51C*, (D) *MYO9B-FCHO1*, and (E) *PAPOLA-AK7* in MCF-7. Validation of prostate tumor chimeras includes (F) *HERPUD1-ERG* in aT64 and (G) *AX747630-ETV1* in aT52. (H) Overall summary of novel validated chimeras.

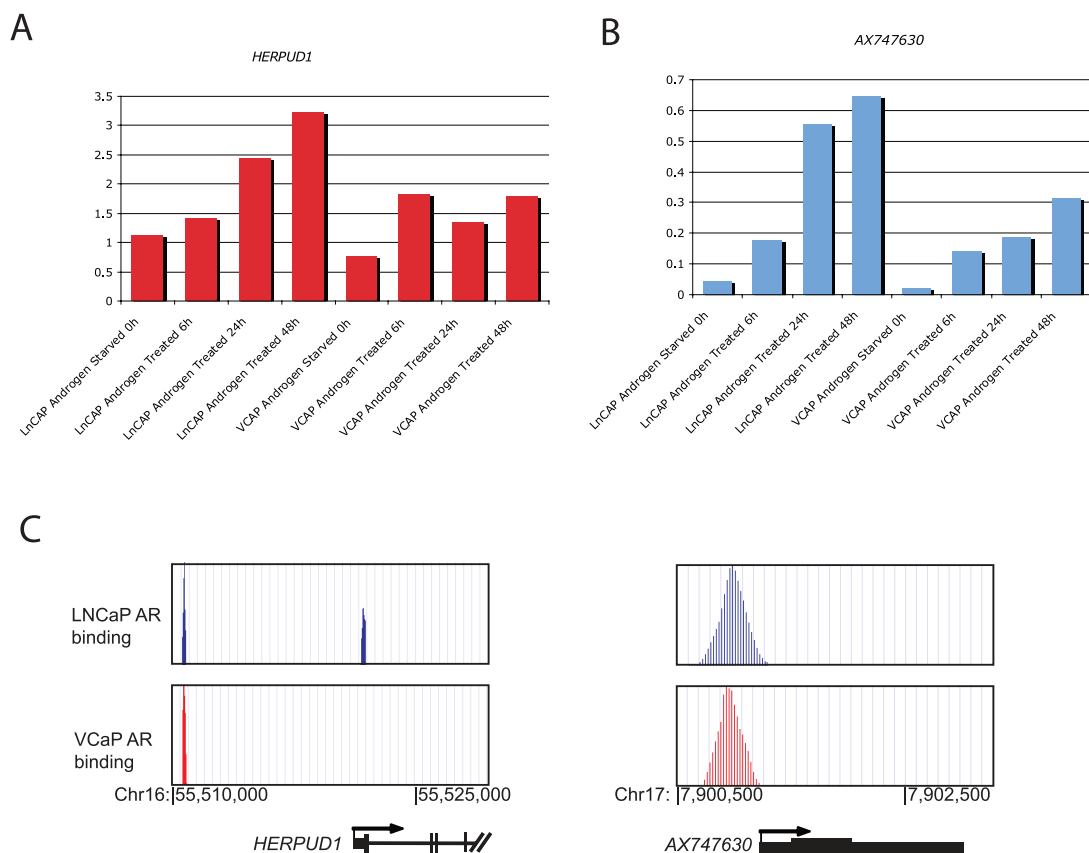


Fig. S7. RNA-Seq gene expression and androgen regulation of *HERPUD1* and *AX747630*, also referred to as *FLI35294*, in LNCaP and VCaP androgen time course. Histogram represents the normalized gene expression value of (A) *HERPUD1* and (B) *AX747630* in LNCaP and VCaP cell lines starved and treated with R1881 at 6, 24, and 48 h. Both genes appear to be androgen responsive in both cell lines. (C) ChIP-Seq binding reveals AR regulation of *HERPUD1* and *AX747630* in prostate cell lines. Schematic representation of ChIP-Seq peaks representing androgen binding near the upstream of *HERPUD1* (Left) and *AX747630* (Right) in LNCaP and VCaP.

Table S1. Paired-end summary statistics

Paired end	Lane 1	Lane 2	Lane 3	Lane 4	Total	Percentage
VCaP						
Same gene	3196295	3005894	2746073	2223161	11171423	65.8
Fusion genes	35249	31217	29465	22380	118311	0.7
Ribosomal	2509	2340	2243	1833	8925	0.1
Nonmapping	1445940	1333170	1261170	1143923	5184203	30.5
Mitochondrial	122035	114042	105123	84184	425384	2.5
Quality control	22579	18351	14427	10675	66032	0.4
	4824607	4505014	4158501	3486156	16974278	
K562						
Same gene	3774056	3758169	3737171	3505675	14775071	71.3
Fusion genes	49565	48127	47826	43390	188908	0.9
Ribosomal	184435	182938	179565	167912	714850	3.4
Nonmapping	1031211	1047680	1080464	1073374	4232729	20.4
Mitochondrial	209455	209451	208877	195094	822877	4.0
Quality control	28	19	30	37	114	0.0
	5248750	5246384	5253933	4985482	20734549	
UHR						
Same gene	6175075	6083374	5924187		18182636	71.2
Fusion genes	53671	52328	51205		157204	0.6
Ribosomal	231218	228336	221872		681426	2.7
Nonmapping	1702980	1589245	1819067		5111292	20.0
Mitochondrial	472404	463064	453238		1388706	5.4
Quality control	2646	6442	4917		14005	0.1
				Total	25535269	
HBR						
Same gene	5462892	5173109	4920238		15556239	65.8
Fusion genes	40116	37624	36344		114084	0.5
Ribosomal	157676	149884	144159		451719	1.9
Nonmapping	1776145	1678741	1884455		5339341	22.6
Mitochondrial	758180	716163	677967		2152310	9.1
Quality control	6259	4570	5336		16165	0.1
				Total	23629858	

Table S2. Chimera candidates nominated by 100-nt reads and paired-end sequencing

Sample	5p gene	5p Refseq	3p gene	3p Refseq	Paired end normalized	Long read normalized	Ratio
HBR	SEMA4D	NM.006378	LOC349236	NM.182635	0.26	0.0169	15.39
	DPH1	NM.001383	OVCA2	NM.080822	0.22	0.0169	13.02
	HSN2	NM.213655	WNK1	NM.018979	0.17	0.0169	10.06
	PRH1	NM.006250	PRR4	NM.007244	0.17	0.0169	10.06
	FBXL15	NM.024326	PSD	NM.002779	0.13	0.0169	7.7
	MGC2474	NM.023931	FLJ23436	NM.024671	0.22	0.0337	6.53
	INPP5A	NM.005539	NKX6-2	NM.177400	0.43	0.0674	6.38
	C16orf60	NM.018455	C16orf61	NM.020188	0.09	0.0169	5.33
	SNF1LK2	NM.015191	PPP2R1B	NM.181699	0.09	0.0169	5.33
	GFAP	NM.002055	TP53INP2	NM.021202	0.09	0.0169	5.33
	HLA-E	NM.005516	HLA-C	NM.002117	0.09	0.0169	5.33
	ARRDC1	NM.152285	C9orf37	NM.032937	0.09	0.0169	5.33
	TIGD7	NM.033208	ZNF263	NM.005741	0.09	0.0169	5.33
	COG1	NM.018714	FAM104A	NM.032837	0.09	0.0169	5.33
	WDR77	NM.024102	OVGP1	NM.002557	0.09	0.0169	5.33
	APEH	NM.001640	RNF123	NM.022064	0.09	0.0169	5.33
	MAPK12	NM.002969	HDAC10	NM.032019	0.09	0.0169	5.33
	BUD31	NM.003910	PTCD1	NM.015545	0.13	0.0337	3.86
	PHPT1	NM.014172	EDF1	NM.153200	0.13	0.0337	3.86
	C15orf38	NM.182616	AP3S2	NM.005829	0.22	0.0674	3.27
	APITD1	NM.199295	CORT	NM.001302	0.09	0.0337	2.68
	ABHD14A	NM.015407	ACY1	NM.000666	0.09	0.0337	2.68
	GPR27	NM.018971	EIF4E3	NM.173359	0.09	0.0337	2.68
	EIF4E3	NM.173359	GPR27	NM.018971	0.13	0.0505	2.58
	ABCB8	NM.007188	ACCN3	NM.004769	0.17	0.0674	2.53
	EIF2AK1	NM.014413	JTV1	NM.006303	0.17	0.0674	2.53
	TUBB3	NM.006086	TUBB4	NM.006087	0.09	0.0505	1.79
	NKX6-2	NM.177400	INPP5A	NM.005539	0.09	0.0505	1.79
	BPTF	NM.004459	KPNA2	NM.002266	0.22	0.1347	1.64
	CENPT	NM.025082	NUTF2	NM.005796	0.09	0.1179	0.77
	CHAD	NM.001267	FLJ20920	NM.025149	0.09	0.2694	0.34
UHR	BCAS4	NM.017843	BCAS3	NM.001099432	46.2106	2.8297	16.34
	BCR	NM.021574	ABL1	NM.007313	2.7414	0.1887	14.53
	ANP32B	NM.006401	CALR	NM.004343	2.3497	0.1887	12.46
	RPS6KB1	NM.003161	TMEM49	NM.030938	1.9581	0.1887	10.38
	DND1	NM.194249	WDR55	NM.017706	1.5665	0.1887	8.31
	FADS1	NM.013402	FADS2	NM.004265	1.5665	0.1887	8.31
	NUP214	NM.005085	XKR3	NM.175878	2.3497	0.3773	6.23
	ADCK4	NM.024876	NUMBL	NM.004756	2.3497	0.3773	6.23
	ZNF561	NM.152289	ZNF562	NM.017656	1.1749	0.1887	6.23
	VAMP8	NM.003761	VAMP5	NM.006634	1.1749	0.1887	6.23
	TUBA6	NM.032704	K-ALPHA-1	NM.006082	1.1749	0.1887	6.23
	GAS6	NM.000820	RASA3	NM.007368	13.7066	2.4524	5.59
	FLJ14640	NM.032816	PEPD	NM.000285	2.3497	0.566	4.16
	ZFP41	NM.173832	GLI4	NM.138465	1.5665	0.3773	4.16
	DNAJB7	NM.145174	LOC63929	NM.022098	0.7833	0.1887	4.16
	MCFD2	NM.139279	TTC7A	NM.020458	0.7833	0.1887	4.16
	HEXB	NM.000521	GFM2	NM.170691	0.7833	0.1887	4.16
	DGCR8	NM.022720	HTF9C	NM.182984	0.7833	0.1887	4.16
	C11orf2	NM.013265	TM7SF2	NM.003273	0.7833	0.1887	4.16
	AP4B1	NM.006594	RSBN1	NM.018364	0.7833	0.1887	4.16
	POLA2	NM.002689	CDC42EP2	NM.006779	0.7833	0.1887	4.16
	TMSB4X	NM.021109	TMSL3	NM.183049	1.9581	0.566	3.46
	ARFGEF2	NM.006420	SULF2	NM.198596	8.6156	2.6411	3.27
	PLCXD2	NM.153268	PHLDB2	NM.145753	1.1749	0.3773	3.12
	DOC2A	NM.003586	FLJ90652	NM.173618	0.7833	0.3773	2.08
	CGI-96	NM.015703	SERHL	NM.170694	0.7833	0.3773	2.08
	CRIP2	NM.001312	CRIP1	NM.001311	0.7833	0.566	1.39
	HLA-G	NM.002127	HLA-C	NM.002117	0.7833	0.7546	1.04
	ZNF276	NM.152287	C16orf7	NM.004913	0.7833	0.7546	1.04
	ACTN4	NM.004924	ACTN1	NM.001102	0.7833	0.7546	1.04
	C16orf7	NM.004913	ZNF276	NM.152287	0.7833	0.7546	1.04
	HLA-A	NM.002116	HLA-B	NM.005514	0.7833	0.9433	0.84
	FLJ14346	NM.025029	NSMASE3	NM.017751	0.7833	1.5092	0.52
	K-ALPHA-1	NM.006082	TUBA3	NM.006009	0.7833	2.2638	0.35

Table S3. Chimeras nominated in all samples (VCaP, K562, UHR, and HBR)

5p Gene	5p Refseq	3p Gene	3p Refseq	Category	EST confirmation
CARM1	NM.199141	YIPF2	NM.024029	Converging	Yes
MGC11102	NM.032325	BANF1	NM.003860	Diverging	Yes
SLC4A1AP	NM.018158	SUPT7L	NM.014860	Diverging	Yes
ERCC2	NM.000400	KLC3	NM.177417	Converging	Yes
PMF1	NM.007221	BGLAP	NM.199173	Overlapping	Yes
THOC6	NM.024339	HCFC1R1	NM.017885	Diverging	Yes
NDUFB8	NM.005004	SEC31L2	NM.015490	Read-through	Yes
ANKRD39	NM.016466	ANKRD23	NM.144994	Diverging	No
C14orf124	NM.020195	KIAA0323	NM.015299	Overlapping	Yes
C14orf21	NM.174913	CIDEB	NM.014430	Diverging	No
ZNF511	NM.145806	TUBGCP2	NM.006659	Diverging	Yes