SUPPLEMENTARY DISCUSSION

Characterization of cancer mutations

Mutations in cancers have been conventionally identified through cytogenetic and molecular techniques[1], later supplanted with sequencing of specific cancer types[2-4], or candidate genes[5-7]. A number of national efforts are underway to comprehensively characterize the genomic alterations in cancer including The Cancer Genome Atlas Project (TCGA, http://cancergenome.nih.gov/index.asp). More recently, high throughput 'next generation sequencing' methods have been used for enumeration of genome-wide aberrations in cancers[8,9]. While considerable effort has been vested in discovering base change mutations (and SNPs) in cancers[3,4,10,11], 'gene-fusions' have not been systematically investigated thus far. Part of the reason is that solid tumors pick up many non-specific aberrations during tumor evolution, making it difficult to distinguish causal/driver aberrations from secondary/insignificant mutations. We believe that the problem of non-specific genetic aberrations is mitigated by sequencing the transcriptome, which restricts the enquiry to 'expressed sequences', thus enriching the data for potentially 'functional' mutations. Not surprisingly, the recent gene fusions discovered in prostate and lung cancer were found through transcriptome[12,13] and proteome[14] analyses. We therefore considered employing massively parallel transcriptome sequencing to discover chimeric transcripts, representing functional gene fusions.

Integration of short and long read sequencing for chimera discovery reduces false positives

Experimental validation revealed that the integration of long and short read technologies decreases false positive candidates, thereby providing a higher fidelity set of *bona fide* fusion candidates. As shown in Supplementary Figure 3, we assessed the expression of 16 chimeras nominated by the long read sequencing platform but lacked short reads spanning the predicted fusion boundary. The fact that all of these chimeras failed to produce a signal, yet 67% (**Supplementary Table 4**) of our nominated chimeras from the integrated approach were experimentally validated confirms that integrating both sequencing platforms effectively reduces false positive nominations.

Complex intra-chromosomal rearrangements in VCaP

One of the striking observations from our experimentally validated VCaP chimeras was the identification of a complex intra-chromosomal rearrangement involving *HJURP*. The fact that both exon 8 and 9 of *HJURP* fuse to different genes suggests a breakpoint resides within the intron (**Fig. 2b**). Both of these gene fusions were confirmed by qRT-PCR in VCaP and VCaP-Met, and were found to be absent in other samples tested. This complex intra-chromosomal rearrangement was also confirmed by FISH analysis. *HJURP* has been shown to be associated with genomic instability and immortality in cancer cells[15], while *INPP4A* encodes one of the enzymes involved in phosphatidylinositol signaling pathways and *EIF4E2* is a eukaryotic translation initiation factor (http://smd.stanford.edu/cgi-bin/source/sourceSearch)[2].

LNCaP chimeras

Interestingly, our top gene fusion nomination in LNCaP cells involved the fusion of *MIPOL1-DGKB*. This gene fusion may represent a harbinger of *ETV1* cryptic rearrangement, a putative driver mutation in the LNCaP prostate cancer cell line. Moreover, we observed the LNCaP cells harbor multiple fusions, similar to our observations in VCaP. One of the validated examples is the fusion between exon 7 of *MRPS10* from chromosome 6 with exon 7 of *HPR* of chromosome 16 (**Supplementary Fig. 14a**). *MRPS10-HPR* was confirmed by FISH and validated by qRT-PCR in LNCaP, but not observed in VCaP, VCaP-Met, RWPE, PREC, or Met 2 (**Supplementary Fig. 14b-e**). Overall, the detection of multiple gene fusions in individual cancer samples argues for a more widespread and more nuanced role of gene fusions in cancer.

Transcriptome sequencing reveals read-through chimeras

One of the advantages of high-throughput sequencing is that it provides an unbiased view of the transcriptome. This enabled us to identify not only inter-chromosomal gene fusions and intra-chromosomal gene fusions involving non-adjacent genes, but also chimeras between adjacent genes, or read-throughs. For instance, a chimera between exon 2 of *C19orf25* with an intron of the neighboring gene *APC2* in LNCaP cells (**Supplementary Fig. 6a**). Experimental validation demonstrated a lower expression level of *C19orf25-APC2(intron)* than observed for gene fusions and weak expression in multiple cell lines suggesting they are more broadly expressed. A similar pattern was observed for *WDR55-DND1* (**Supplementary Fig. 6b**), *MBTPS2-YY2* (**Supplementary Fig. 6c**), and *ZNF649-ZNF577* (**Supplementary Fig. 6d)**.

Array CGH analysis of gene fusions

Many studies utilize genomic information for mining gene fusion candidates[8,16]. Therefore, we were interested in determining whether transcriptome data detects chimeras that would not be apparent from genomic DNA analysis. To do so, we integrated unbalanced genomic copy number change data from array comparative genomic hybridization of matched samples and monitored genomic aberrations within our gene fusion candidates (**Supplementary Methods**). This revealed breakpoints in genes involved in two gene fusion candidates, *USP10-ZDHHC7*, and *MIPOL1-DGKB* (**Supplementary Table 4**). More specifically, we observed a homozygous deletion spanning the region between *USP10-ZDHHC7* in VCaP cells as well as in the parental metastatic prostate cancer tissue from which VCaP is derived (VCaP-Met) but not in the normal prostate cell line RWPE (**Supplementary Fig. 15**). Taken together, this suggests that a deletion coupled with a complex rearrangement may have led to the *USP10-ZDHHC7* fusion. qRT-PCR based evaluation confirmed this fusion to be specific to VCaP and its parental tissue, VCaP-Met, and not in LNCaP, RWPE, PREC, or metastatic prostate cancer tissue (Met 2) (**Fig. 2a**). In LNCaP cells, for the *MIPOL1-DGKB* fusion a breakpoint was found only in *DGKB* but not in *MIPOL1*. Furthermore, absence of breakpoints in all other fusion chimeras examined suggests that the majority of fusion gene candidates identified by sequencing would not have been discovered by mining genomic

copy number aberration data. Moreover, while only a subset of genomic rearrangements potentially represent functional gene fusions, most chimeric transcripts signify productive fusions, with likely roles in the biology of cells they are found in.

MATERIALS AND METHODS

Samples and cell lines

The benign immortalized prostate cell line RWPE and the prostate cancer cell line LNCaP was obtained from the American Type Culture Collection. Primary benign prostatic epithelial cells (PrEC) were obtained from Cambrex Bio Science. The prostate cancer cell line MDA-PCa 2B was provided by E. Keller. The prostate cancer cell line 22-RV1 was provided by J. Macoska. VCaP was derived from a vertebral metastasis from a patient with hormone-refractory metastatic prostate cancer[17], and was provided by Ken Pienta.

Androgen stimulation experiment was carried out with LNCaP and VCaP cells grown in charcoal-stripped serum containing media for 24 h, before treatment with 1% ethanol or 1 nM of methyltrienolone (R1881, NEN Life Science Products) dissolved in ethanol, for 24 and 48 h. Total RNA was isolated with RNeasy mini 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[18], University of Michigan Prostate Cancer Specialized Program of Research Excellence Tissue Core. All samples were collected with informed consent of the patients and prior approval of the institutional review board.

454 FLX Sequencing

PolyA+ RNA was purified from 50µg total RNA using two rounds of selection on oligo-dT containing paramagnetic beads using Dynabeads mRNA Purification Kit (Dynal Biotech, Oslo, Norway), according to the manufacturer's instructions. 200 ng mRNA was fragmented at 82°C in Fragmentation Buffer (40 mM Tris-Acetate, 100 mM Potassium Acetate, 31.5 mM Magnesium Acetate, pH 8.1) for 2 minutes. First strand cDNA library was prepared using Superscript II (Invitrogen) according to standard protocols and directional adaptors were ligated to the cDNA ends for clonal amplification and sequencing on the Genome Sequencer FLX. The 5'-end Adaptor A has a 5' overhang of 5 nucleotides and the 3'-end Adaptor B has a 3' overhang of 6 random nucleotides:

```
5'-NANNACTGATGGCGCGAGGGAGGC-3′
        GACTACCGCGCTCCCTCCG-5′

5'-biotin-GCCTTGCCAGCCCGCTCAGNNNNNN-P-3‘
     3′-CGGAACGGTCGGGCGAGTC
```

The adaptor ligation reaction was carried out in Quick Ligase Buffer (New England Biolabs, Ipswich, MA) containing 1.67 µM of the Adaptor A, 6.67 µM of the Adaptor B and 2000 units of T4 DNA Ligase (New England Biolabs, Ipswich, MA) at 37°C for 2 hours. Adapted library was recovered with 0.05% Sera-Mag30 streptavidin beads (Seradyn Inc, Indianapolis, IN) according to manufacturer's instructions. Finally, the sscDNA library was purified twice with RNAClean (Agencourt, Beverly, MA) as per the manufacturer's directions except the

amount of beads was reduced to 1.6X the volume of the sample. The purified sscDNA library was analyzed on an RNA 6000 Pico chip on a 2100 Bioanalyzer (Agilent Technologies, Santa Clara, CA) to confirm a size distribution between 450 to 750 nucleotides, and quantified with Quant-iT Ribogreen RNA Assay Kit (Invitrogen Corporation, Carlsbad, CA) on a Synergy HT (Bio-Tek Instruments Inc, Winooski, VT) instrument following the manufacturer's instructions. The library was PCR amplified with 2 μM each of Primer A (5'-GCC TCC CTC GCG CCA-3') and Primer B (5'-GCC TTG CCA GCC CGC-3'), 400 μM dNTPs, 1X Advantage 2 buffer and 1 μl of Advantage 2 polymerase mix (Clontech, Mountain View, CA). The amplification reaction was performed at: 96°C for 4 min; 94°C for 30 sec, 64°C for 30 sec, repeating steps 2 and 3 for a total of 20 cycles, followed by 68°C for 3 minutes. The samples were purified using AMPure beads and diluted to a final working concentration of 200,000 molecules per μl. Emulsion beads for sequencing were generated using Sequencing emPCR Kit II and Kit III and sequencing was carried out using 600,000 beads.

Normalization by Subtraction

mRNA from the prostate cancer cell line VCaP was hybridized with the subtractor cell line LNCaP 1st-strand cDNA immobilised on magnetic beads (Dynabeads, Invitrogen), according to the manufacturers instructions. Transcripts common to both the cells were captured and removed by magnetic separation of bead-bound subtractor cDNA and the subtracted VCaP mRNA left in the supernatant was recovered by precipitation and used for generating sequencing library as described. Efficiency of normalization was assessed by qRT-PCR assay of levels of select transcripts in the sample before and after the subtraction (data not shown).

Illumina Genome Analyzer Sequencing

200ng mRNA was fragmented at 70°C for 5 min in a Fragmentation buffer (Ambion), and converted to first strand cDNA using Superscript III (Invitrogen), followed by second strand cDNA synthesis using E coli DNA pol I (Invitrogen). The double stranded cDNA library was further processed by Illumina Genomic DNA Sample Prep kit, and it involved end repair using T4 DNA polymerase, Klenow DNA polymerase, and T4 Polynucleotide kinase followed by a single <A> base addition using Klenow 3' to 5' exo$^-$ polymerase, and was ligated with Illumina's adaptor oligo mix using T4 DNA ligase. Adaptor ligated library was size selected by separating on a 4% agarose gel and cutting out the library smear at 200bp (+/- 25bp). The library was PCR amplified by Phu polymerase (Stratagene), and purified by Qiaquick PCR purification kit (Qiagen). The library was quantified with Quant-iT Picogreen dsDNA Assay Kit (Invitrogen Corporation, Carlsbad, CA) on a Modulus™ Single Tube Luminometer (Turner Biosystems, Sunnyvale, CA) following the manufacturer's instructions. 10nM library was used to prepare flowcells with approximately 30,000 clusters per lane.

Sequence datasets

Human genome build 18 (hg18) was used as a reference genome. All UCSC and Refseq transcripts were downloaded from the UCSC genome browser (http://genome.ucsc.edu/)[19].

Sequences of previously identified *TMPRSS2-ERGa* fusion transcript (Genbank accession: DQ204772) and *BCR-ABL1* fusion transcript (Genbank accession: M30829) were used for reference.

Short read chimera discovery

Short reads that do not completely align to the human genome, Refseq genes, mitochondrial, ribosomal, or contaminant sequences are categorized as non-mapping. For many chimeras we expect that there will be a larger portion mapping to a fusion partner (major alignment), and smaller portion aligning to the second partner (minor alignment). Our approach is therefore divided into two phases in which we focus on first identifying the major alignment and then performing a more exhaustive approach for identifying the minor alignment. In the first phase all non-mapping reads are aligned against all exons of Refseq genes using Vmatch, a pattern matching program[20]. Only reads that have an alignment of 12 or more nucleotides to an exon boundary are kept as potential chimeras. In the second phase, the non-mapping portion of the remaining reads are then mapped to all possible exon boundaries using a Perl script that utilizes regular expressions to detect alignments of as few as six nucleotides. Only those short reads that show partial alignment to exon boundaries of two separate genes are categorized as chimeras. It is possible to have a chimera that has 28 nucleotides aligning to gene x and 8 nucleotides that align to gene y and z because the 8-mer does not provide enough sequence resolution to distinguish between gene y and gene z. Therefore we would categorize this as two individual chimeras. If a sequence forms more than five chimeras it is discarded because it is ambiguous. To minimize false positives, we require that a predicted gene fusion event has at least two supporting chimeras.

Long and short read integrated chimera discovery

All 454 reads are aligned against the human Refseq collection using BLAT, a rapid mRNA/DNA alignment tool[21]. Using a Perl script, the BLAT output files were parsed to detect potential chimeric reads. A read is categorized as completely aligning if it shows greater than 90% alignment to a known Refseq transcript. These are then discarded as they almost completely align and therefore are not characteristic of a chimera. From the remaining reads, we want to query for reads having partial alignment, with minimal overlap, to two Refseq transcripts representing putative chimeras. To accomplish this, we iterate the all possible BLAT alignments for a putative chimera, extracting only those partial alignments that have no more than a six nucleotide, or two codon, overlap. This step reduces false positive chimeras introduced by repetitive regions, large gene families, and conserved domains. Additionally, while our approach tolerates overlap between the partial alignments, it filters those having more than ten or more nucleotides between the partial alignments.

The short reads (36 nucleotides) generated from the Illumina platform are parsed by aligning them against the Refseq database and the human genome using Eland, an alignment tool for short reads. Reads that align completely or fail quality control are removed leaving only the "non-mapping" reads; a rich source for chimeras. These non-mapping short reads are subsequently aligned against all putative long read chimeras (obtained as described above) using Vmatch[20], a pattern matching program. A Perl script is used to parse the Vmatch output

to extract only those reads that span the fusion boundary by at least three nucleotides on each side. Following this integration, the remaining putative chimeras are categorized as inter- or intra-chromosomal chimeras based on whether the partial alignments are located on different or the same chromosomes, respectively. Those intra-chromosomal chimeras that have partial alignments to adjacent genes are believed to be the product of co-transcription of adjacent genes coupled with intergenic splicing (CoTIS)[22], alternatively known as read-throughs. The remaining intra-chromosomal and all inter-chromosomal chimeras are considered candidate gene fusions.

One additional source of false positive chimeras could be an unknown transcript that is not in Refseq. Due to its absence in the Refseq database, the corresponding long read would not be able to show a complete alignment, but instead show partial hits. Subsequently, short reads spanning this transcript would naturally validate the artificially produced fusion boundary. Therefore, to remove these candidates, we aligned all of the chimeras against the human genome using BLAT. If the long read had greater than 90% alignment to one genomic location, it is considered a novel transcript rather than a chimeric read. The remaining chimeras are given a score which is calculated by multiplying the long read coverage spanning the fusion boundary against the short read coverage spanning the fusion boundary.

Coverage analysis

Transcript coverage for every gene locus was calculated from the total number of passing filter reads that mapped, via ELAND, to exons. The total count of these reads was multiplied by the read length and divided by the longest transcript isoform of the gene as determined by the sum of all exon lengths as defined in the UCSC knownGene table (Mar. 2006 assembly). Nucleotide coverage was determined by enumerating the total reads, based on ELAND mappings, at every nucleotide position within a non-redundant set of exons from all possible UCSC transcript isoforms.

Array CGH analysis

Oligonucleotide comparative genomic hybridization is a high-resolution method to detect unbalanced copy number changes at whole genome level. Competitive hybridization of differentially labeled tumor and reference DNA to oligonucleotide printed in an array format (Agilent Technologies, USA) and analysis of fluorescent intensity for each probe will detect the copy number changes in the tumor sample relative to normal reference genome. We identified genomic breakpoints at regions with a change in copy number level of at least one copy (log ratio ± 0.5) for gains and losses involving more than one probe representing each genomic interval as detected by the aberration detection method (ADM) in CGH analytics algorithm.

Real Time PCR validation

Quantitative PCR (QPCR) was performed using Power SYBR Green Mastermix (Applied Biosystems, Foster City, CA) on an Applied Biosystems Step One Plus Real Time PCR System as described[23]. All oligonucleotide primers were synthesized by Integrated DNA

Technologies (Coralville, IA) and are listed in **Table S8**. *GAPDH*[24], primer was as described. All assays were performed in duplicate or triplicate and results were plotted as average fold change relative to *GAPDH*.

Quantitative PCR for *SLC45A3-ELK4* was carried out by Taqman assay method using fusion specific primers and Probe #7 of Universal Probe Library (UPL), Human (Roche) as the internal oligonucleotide, according to manufacturer's instructions. *PGK1* was used as housekeeping control gene for UPL based Taqman assay (Roche), as per manufacturer's instructions. HMBS (Applied Biosystems, Taqman assay Hs00609297_m1) was used as housekeeping gene control for Taqman assays according to standard protocols (Applied Biosystems).

Fluorescence in situ hybridization (FISH)

FISH hybridizations were performed on VCaP, LNCaP, and FFPE tumor and normal tissues. BAC clones were selected from UCSC genome browser. Following colony purification midi prep DNA was prepared using QiagenTips-100 (Qiagen, USA). DNA was labeled by nick translation labeling with biotin-16-dUTP and digoxigenin-11-dUTP (Roche, USA). Probe DNA was precipitated and dissolved in hybridization mixture containing 50% formamide, 2XSSC, 10% dextran sulphate, and 1% Denhardts solution. About 200ng 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, Germany).

Affymetrix Genome-Wide Human SNP Array 6.0

1 μg each of genomic DNA samples was sent to Affymetrix service centers (Center for Molecular Medicine, Grand Rapid, MI and Vanderbilt Affymetrix Genotyping Core, Nashville, TN) for genomic level analysis of 15 samples on the Genome-Wide Human SNP Array 6.0. Copy number analysis was conducted using the Affymetrix Genotyping Console software and visualizations were generated by the Genotyping Console (GTC) browser.

Supplementary Table 1. Summary of normalized and non-normalized VCaP 454 libraries

Sample	Normalized VCaP	Non-normalized VCaP
Subtracted	Yes	No
Total Reads	575985	551780
Average length	218.9	226.5
Genes*	2687	2857
Reads / Gene	214.36	193.14
Chimeras	118	428
Reads / chimera	4881.3	1289.3

* A read must be a best hit to a gene with greater than 90% alignment

Supplementary Table 2. Top long read chimera candidates. The following list highlights the top VCaP chimeras identified using solely 454 technology. Only those chimeras that had more than one sequence confirmed a fusion boundary are shown in this list. Chimeras highlighted in yellow were confirmed by short read technology and experimentally validated. Chimeras highlighted in blue were found by long read technology but lacked short reads spanning the predicted fusion boundary and failed experimental validation.

454 Reads	Gene 1	Chromosomal location	Description	Gene 2	Chromosomal location	Description
8	EFTUD2	chr17:40283805-40332289	U5 snRNP-specific protein	NDUFB2	chr7:140042950-140052915	NADH dehydrogenase (ubiquinone) 1 beta
7	C9orf152	chr9:112001667-112009735	hypothetical protein LOC401546	SBK1	chr16:28211341-28242671	SH3-binding domain kinase 1
6	HNRPC	chr14:20749432-20772192	heterogeneous nuclear ribonucleoprotein C	P4HB	chr17:77394323-77411833	prolyl 4-hydroxylase, beta subunit precursor
5	EIF2AK1	chr7:6029989-6065302	heme-regulated initiation factor 2-alpha kinase	JTV1	chr7:6015408-6029991	JTV1
4	LPHN1	chr19:14119549-14177997	latrophilin 1 isoform 1 precursor	SUZ12	chr17:27288185-27352162	joined to JAZF1
3	RPS14	chr5:149803985-149809512	ribosomal protein S14	PPIA	chr7:44802766-44809241	peptidylprolyl isomerase A
3	RPL13A	chr19:54682677-54687376	ribosomal protein L13a	CANX	chr5:179058536-179091245	calnexin precursor
3	NGRN	chr15:88609899-88616447	neugrin, neurite outgrowth associated	TMEM87A	chr15:40290018-40353047	transmembrane protein 87A
3	MIA3	chr1:220858067-220907974	melanoma inhibitory activity family, member 3	C1orf80	chr1:220,907,978-220,952,487	hypothetical protein LOC64853
3	MIA2	chr14:38772876-38792326	melanoma inhibitory activity 2	CTAGE5	chr14:38806079-38890148	CTAGE family, member 5
3	CNOT1	chr16:57134733-57221251	CCR4-NOT transcription complex, subunit 1	SETD6	chr16:57106928-57111053	SET domain containing 6
3	C14orf130	chr14:92743170-92765314	hypothetical protein LOC55148	HSP90B1	chr12:102848319-102865833	heat shock protein 90kDa beta, member 1
3	C12orf26	chr12:81276455-81397076	hypothetical protein LOC84190	CCDC59	chr12:81270220-81276330	coiled-coil domain containing 59
3	ARHGEF12	chr11:119713156-119865855	Rho guanine nucleotide exchange factor (GEF) 12	SCD	chr10:102096762-102114578	stearoyl-CoA desaturase
3	ARHGEF12	chr11:119713156-119865855	Rho guanine nucleotide exchange factor (GEF) 12	PLAA	chr9:26894518-26925207	phospholipase A2-activating protein
3	AMD1	chr6:111302680-111323606	S-adenosylmethionine decarboxylase 1	UBE1DC1	chr3:133861836-133879312	ubiquitin-activating enzyme E1-domain containing
2	ZNF667	chr19:61643018-61680555	zinc finger protein 667	PARD6B	chr20:48781488-48803685	PAR-6 beta
2	ZNF649	chr19:57084300-57100059	zinc finger protein 649	ZNF577	chr19:57066362-57083009	zinc finger protein 577
2	YWHAZ	chr8:102000090-102033447	tyrosine 3/tryptophan 5 -monooxygenase	CTNND1	chr11:57285810-57343228	catenin, delta 1
2	XTP7	chr19:50407719-50429309	protein 7 transactivated by hepatitis B virus X	FLJ21062	chr7:89712461-89777622	hypothetical protein LOC79846
2	USP10	chr16:83291056-83371028	ubiquitin specific protease 10	ZDHHC7	chr16:83565573-83602642	zinc finger, DHHC-type containing 7
2	TYMS	chr18:647604-663499	thymidylate synthetase	ENOSF1	chr18:664001-702676	enolase superfamily member 1 isoform rTS beta
2	TTC6	chr14:37334265-37381247	tetratricopeptide repeat domain 6	CTSC	chr11:87666408-87710589	cathepsin C isoform a preproprotein
2	TSEN34	chr19:59386009-59389338	tRNA splicing endonuclease 34 homolog	C17orf79	chr17:27203012-27210369	hypothetical protein LOC55352
2	TLOC1	chr3:171167305-171193497	translocation protein 1	LAMC1	chr1:181259176-181381350	laminin, gamma 1 precursor
2	TIA1	chr2:70290080-70329283	TIA1 protein	DIRC2	chr3:123996597-124081451	disrupted in renal carcinoma 2
2	TERF2IP	chr16:74239185-74248829	telomeric repeat binding factor 2, interacting	AKAP12	chr6:151688516-151719601	A-kinase anchor protein 12
2	SMC4L1	chr3:161600124-161614999	SMC4 structural maintenance of chromosomes	SAR1B	chr5:133970018-133996426	SAR1a gene homolog 2
2	SLCO1A2	chr12:21313094-21379099	organic anion transporting polypeptide A isoform	CTNND1	chr11:57285810-57343228	catenin, delta 1
2	SKIV2L2	chr5:54639373-54756562	superkiller viralicidic activity 2-like 2	MARCKSL1	chr1:32572027-32574410	MARCKS-like 1
2	SF1	chr11:64288654-64302817	splicing factor 1	TIAL1	chr10:121322968-121346531	TIA-1 related protein
2	SET	chr9:130485755-130498496	SET translocation (myeloid leukemia-associated)	RAC1	chr7:6380651-6410123	ras-related C3 botulinum toxin substrate 1
2	RSBN1	chr1:114105977-114156593	round spermatid basic protein 1	RPL7	chr8:74365428-74368423	ribosomal protein L7
2	RPS17	chr15:81002559-81006263	ribosomal protein S17	LOC402057	chr22:30765435-30765974	similar to 40S ribosomal protein S17

2	RPL9	chr4:39132140-39136602	ribosomal protein L9	DHRS7	chr14:59681252-59701857	dehydrogenase/reductase (SDR family) member 7
2	RAB4B	chr19:45976011-45994687	ras-related GTP-binding protein 4b	EGLN2	chr19:45998021-46006177	EGL nine (C.elegans) homolog 2
2	POLR2K	chr8:101232015-101235406	DNA directed RNA polymerase II polypeptide K	SFPQ	chr1:35421788-35431322	splicing factor proline/glutamine rich
2	PLEKHN1	chr1:891740-900347	pleckstrin homology domain containing, family N	GNG7	chr19:2462218-2653746	guanine nucleotide binding protein (G protein)
2	PLEKHN1	chr1:891740-900347	pleckstrin homology domain containing, family N	C16orf57	chr16:56592806-56613023	hypothetical protein LOC79650
2	PIGH	chr14:67125776-67136770	phosphatidylinositol glycan anchor biosynthesis,	PRDX1	chr1:45749294-45760196	peroxiredoxin 1
2	PCBP2	chr12:52132153-52161213	poly(rC) binding protein 2	KLHDC6	chr3:129124592-129189204	kelch domain containing 6
2	OTUD5	chrX:48664432-48699837	OTU domain containing 5	HSP90AB1	chr6:44322827-44329592	heat shock 90kDa protein 1, beta
2	NUDT4	chr12:92295832-92321155	nudix-type motif 4	SLC41A1	chr1:204024844-204048784	solute carrier family 41 member 1
2	NPM1	chr5:170747403-170770493	nucleophosmin 1	CLEC2D	chr12:9713576-9743306	osteoclast inhibitory lectin isoform
2	MDP-1	chr14:23752998-23755081	magnesium-dependent phosphatase 1	CHMP4A	chr14:23748627-23753025	chromatin modifying protein 4A
2	MBTPS2	chrX:21767675-21810794	membrane-bound transcription factor peptidase,	YY2	chrX:21784524-21785642	YY2 transcription factor
2	LOC402057	chr22:30765435-30765974	similar to 40S ribosomal protein S17	RPS17	chr15:81002559-81006263	ribosomal protein S17
2	LHX3	chr9:138227917-138234825	LIM homeobox protein 3	CKAP2	chr13:51927496-51948764	cytoskeleton associated protein 2
2	LEPR	chr1:65658906-65875410	leptin receptor	NEK5	chr13:51536901-51601215	NIMA (never in mitosis gene a)-related kinase 5
2	INPP4A	chr2:98427845-98564716	inositol polyphosphate-4-phosphatase, type 1	DKFZp762E1312	chr2:234410225-234427951	HJURP
2	HSP90B1	chr12:102848319-102865833	heat shock protein 90kDa beta, member 1	SERF2	chr15:41871769-41875579	small EDRK-rich factor 2
2	HNRPK	chr9:85773645-85783187	heterogeneous nuclear ribonucleoprotein K	ASAH1	chr8:17958205-17986159	N-acylsphingosine amidohydrolase 1
2	GBF1	chr10:103995299-104132639	golgi-specific brefeldin A resistance factor 1	ACTL6A	chr3:180763402-180788887	actin-like 6A
2	FLJ20273	chr4:40134545-40211349	hypothetical protein LOC54502	HSP90AA1	chr14:101616828-101623265	heat shock protein 90kDa alpha (cytosolic)
2	FARSLB	chr2:223144406-223229071	phenylalanyl-tRNA synthetase, beta subunit	TRIM61	chr4:166095048-166118268	tripartite motif-containing 61
2	EPB41L4B	chr9:111041833-111122842	erythrocyte membrane protein band 4.1 like 4B	TXNRD1	chr12:103204857-103268192	thioredoxin reductase 1
2	ENAH	chr1:223741157-223907468	enabled homolog	ASAH1	chr8:17958205-17986159	N-acylsphingosine amidohydrolase 1
2	EIF4G2	chr11:10775169-10787158	eukaryotic translation initiation factor 4	PTP4A2	chr1:32146380-32176575	protein tyrosine phosphatase type IVA, member 2
2	DNM1L	chr12:32723404-32789851	dynamin 1-like protein	KLK2	chr19:56068501-56075635	kallikrein 2
2	DNAJA5	chr5:34965455-34994826	DnaJ homology subfamily A member 5	MYST3	chr8:41906154-42028662	MYST histone acetyltransferase
2	DDX56	chr7:44571928-44580662	DEAD (Asp-Glu-Ala-Asp) box polypeptide 56	RPL37A	chr2:217071765-217074433	ribosomal protein L37a
2	CS	chr12:54951750-54980442	citrate synthase precursor	RAI14	chr5:34691597-34868474	retinoic acid induced 14
2	CARM1	chr19:10883860-10894448	coactivator-associated arginine	YIPF2	chr19:10894444-10900357	Yip1 domain family, member 2
2	C9orf152	chr9:112001667-112009735	hypothetical protein LOC401546	BASP1	chr5:17270750-17329943	brain abundant, membrane attached signal protein
2	C1orf80	chr1:220,907,978-220,952,487	hypothetical protein LOC64853	MIA3	chr1:220884287-220907974	melanoma inhibitory activity family, member 3
2	C19orf42	chr19:16617959-16631968	hypothetical protein LOC79086	MST150	chr5:150138060-150156491	putative small membrane protein NID67
2	C14orf2	chr14:103448378-103457656	6.8 kDa mitochondrial proteolipid	RPS24	chr10:79463580-79470479	ribosomal protein S24
2	BHLHB9	chrX:101862310-101894025	basic helix-loop-helix domain containing, class	RPL7	chr8:74365428-74368423	ribosomal protein L7
2	BCOR	chrX:39795443-39841663	BCL-6 interacting corepressor	ZDHHC9	chrX:128766594-128805554	zinc finger, DHHC domain containing 9
2	ANKRD21	chr21:13,904,369-13,935,777	pote protein	POTE8	chr8:43266742-43337485	
2	ABCC9	chr12:21841591-21980895	ATP-binding cassette, sub-family C, member 9	EEF1G	chr11:62083649-62098036	eukaryotic translation elongation factor 1

Supplementary Table 3. Illumina sequence summary statistics

Sample	K562		VCaP		LNCaP		RWPE		VCaP-Met		Met 3		Met 4	
Total reads (millions)	66.9		76.4		57.3		71.9		14		35		9.24	
Pass filter (millions)*	38.3	57.25%	40.3	52.75%	35.3	61.61%	44.8	62.31%	9.6	68.57%	16.4	46.86%	5.51	59.64%
Non-mapping reads (millions)**	2.08	5.43%	12.69	31.49%	1.59	4.50%	1.77	3.95%	0.4	4.17%	1.1	6.71%	0.31	5.63%
Redundantly mapping reads (millions)**	1.42	3.71%	1.08	2.68%	1.23	3.48%	1.74	3.88%	0.71	7.40%	1.32	8.05%	0.45	8.17%
Best hit uniquely maps (millions)**	19.86	51.85%	15.48	38.41%	19.34	54.79%	26.13	58.33%	7.36	76.67%	12.59	76.77%	4.3	78.04%
Mitochondrial reads (millions)**	1.89	4.93%	1.72	4.27%	3.19	9.04%	2.8	6.25%	0.81	8.44%	0.81	4.94%	0.37	6.72%
Ribosomal reads (millions)**	13.09	34.18%	9.35	23.20%	10	28.33%	12.34	27.54%	0.31	3.23%	0.62	3.78%	0.09	1.64%

* Percentage of total reads

** Percentage of reads passing filter

	UCSC transcripts per sample						
Average Transcript Coverage	K562	VCaP	LNCaP	RWPE	VCaP-Met	Met 3	Met 4
1x	13584	9382	7905	7586	15004	12310	13495
2-10x	5550	9979	9809	9182	6079	8866	6308
11-100x	549	2793	4400	5965	574	1708	746
101-1000x	33	223	301	374	97	182	95
> 1000x	9	40	26	29	8	8	9

	Nucleotide frequency per sample						
Nucleotide coverage	K562	VCaP	LNCaP	RWPE	VCaP-Met	Met 3	Met 4
1x	16890720	16354691	14044350	12626171	17318415	19561549	17677022
2-10x	19840906	34200185	40047417	37775216	18868883	30855788	20365786
11-100x	2283451	13651323	20893997	28851663	2710638	7389198	2933772
101-1000x	123205	844799	1240273	2295184	199455	438471	229998
1001-10000x	133639	86440	64267	82156	18987	17925	18509
10001-100000x	267	3063	66890	5425	0	494	0
> 100000x	0	0	266	0	0	0	0

Supplementary Table 4. Chimera nominations from transcriptome sequencing

Rank	Library	5' Gene	3' Gene	# of Reads		Score*	Validated
				Illumina	454		
1	VCaP	*ZNF649*	*ZNF577*	14	2	28	Yes
2	VCaP	*TMPRSS2*	*ERG*	21	1	21	Yes
3	VCaP	*INPP4A*	*HJURP*	8	2	16	Yes
4	VCaP	*USP10*	*ZDHHC7*	6	2	12	Yes
5	VCaP	*HJURP*	*EIF4E2*	8	1	8	Yes
6	RWPE	*WDR55*	*DND1*	7	1	7	Yes
7	LNCaP	*MIPOL1*	*DGKB*	5	1	5	Yes
8	LNCaP	*PPIA*	*RPS14*	1	3	3	No
9	VCaP	*RBM14*	*RBM4*	2	1	2	No
10	LNCaP	*C19orf25*	*APC2*	2	1	2	Yes
11	VCaP	*FLJ46838*	*SALF*	2	1	2	No
12	VCaP	*SLTM*	*ZNF621*	2	1	2	No
13	LNCaP	*MDP-1*	*CHMP4A*	1	2	2	No
14	LNCaP	*MBTPS2*	*YY2*	1	2	2	Yes
15	LNCaP	*MRPS10*	*HPR*	1	1	1	Yes

* Score = 454 read count x Illumina read count

Supplementary Table 5. Gene fusion candidates with previously reported copy number variations (CNVs) reported in the Database of Genomic Variants (http://projects.tcag.ca/variation/).

Gene	Copy Number Gain or Loss	Position	Reference(s)	Pubmed ID
BCR	Gain	chr22:21,973,192..22,250,412	(Perry et al., 2008)	18304495
	Gain	chr22:21,983,119..22,145,329	(Perry et al., 2008)	18304495
ABL1	Loss	chr9:132,580,357..132,584,940	(de Smith et al., 2007)	17666407
	-	chr9:132,727,387..133,018,382	(Redon et al., 2006)	17122850
MPRS10				
HPR	Loss	chr16:70,645,832..70,665,594	(Wang et al., 2007)	17921354
	Loss	chr16:70,647,656..70,669,810	(Perry et al., 2008)	18304495
	Loss	chr16:70,643,022..70,667,434	(Perry et al., 2008)	18304495
	Loss	chr16:70,655,469..70,675,458	(Kidd et al., 2008)	18451855
	Loss	chr16:70,666,462..70,671,389	(Kidd et al., 2008)	18451855
	-	chr16:70,533,845..70,831,848	(Redon et al., 2006)	17122850
MIPOL1				
DGKB	Loss	chr7:13405467..14521413	(Sebat et al., 2004)	15273396
	Loss	chr7:14477509..14477509	(Levy et al., 2007)	17803354
	Gain	chr7:14261667..14265007	(Perry et al., 2008)	18304495
	Loss	chr7:14179956..14189289	(Perry et al., 2008)	18304495
TMPRSS2				
ERG	Loss	chr21:38835017..38838626	(Wang et al., 2007)	17921354
	Loss	chr21:38839050..38877931	(Pinto et al., 2007)	17911159
USP10				
ZDHHC7				
STRN4	Loss	chr19:51,898,389..52,062,144	(Wong et al., 2007)	17160897
GPSN2				
LMAN2	Loss	chr5:175467278..177401618	(Korbel et al., 2008)	17901297
	Loss	chr5:176669915..176829018	(Jakobsson et al., 2008)	18288195
	-	chr5:176550923..176735050	(Redon et al., 2006)	17122850
AP3S1				
EIF4E2				
HJURP				
INPP4A				
RC3H2				
RGS3				
ZNF649	Loss	chr19:56951383..57296437	(Redon et al., 2006)	17122850
	Gain	chr19:56910738..57296437	(Zogopoulos et al., 2007)	17638019
	-	chr19:57037904..57225638	(Redon et al., 2006)	17122850
	Gain	chr19:56919728..57322747	(Pinto et al., 2007)	17911159
ZNF577	Loss	chr19:56951383..57296437	(Redon et al., 2006)	17122850
	Gain	chr19:56910738..57296437	(Zogopoulos et al., 2007)	17638019
	-	chr19:57037904..57225638	(Redon et al., 2006)	17122850
	Gain	chr19:56919728..57322747	(Pinto et al., 2007)	17911159
MTBPS2				
YY2				
C19orf25	Loss	chr19:1395419..1426391	(Wang et al., 2007)	17921354
	Loss	chr19:1426391..1477613	(Jakobsson et al., 2008)	18288195
	-	chr19:902641..1495933	(Redon et al., 2006)	17122850
	Loss	chr19:1,331,125..1,495,933	(Wong et al., 2007)	17160897

	-	chr19:1271267..1950204	(de Smith et al., 2007)	17666407
APC2	Loss	chr19:1395419..1426391	(Wang et al., 2007)	17921354
	Loss	chr19:1426391..1477613	(Jakobsson et al., 2008)	18288195
	-	chr19:902641..1495933	(Redon et al., 2006)	17122850
	Loss	chr19:1,331,125..1,495,933	(Wong et al., 2007)	17160897
	-	chr19:1271267..1950204	(de Smith et al., 2007)	17666407
SLC45A3				
ELK4				
WDR55				
DND1				

Supplementary Table 6. aCGH analysis of VCaP, LNCaP, and RWPE nominated chimeras from integrative approach.

		aCGH				aCGH		
Library	5' partner	vCaP	LnCaP	RWPE	3' partner	vCaP	LnCaP	RWPE
VCaP	ZNF649	one copy gain	no change	no change	ZNF577	one copy gain	no change	no change
VCaP	TMPRSS2	1.2 copy gain	no change	no change	ERG	1.2 copy gain	no change	no change
VCaP	INPP4A	no change	no change	no change	HJURP	no change	no change	no change
VCaP	USP10	breakpoint	no change	1.5 copy gain	ZDHHC7	breakpoint	no change	1.5 copy gain
VCaP	HJURP	no change	no change	no change	EIF4E2	no change	no change	no change
RWPE	WDR55	no change	no change	one copy gain	DND1	no change	no change	one copy gain
LNCaP	MIPOL1	no change	no change	no change	DGKB	breakpoint	no change	no change
LNCaP	PPIA	no change	no change	no change	RPS14	no change	no change	one copy gain
VCaP	RBM14	gain- 4copies	no change	no change	RBM4	gain-4 copies	no change	no change
LNCaP	C19orf25	no change	no change	no change	APC2	no change	no change	no change
VCaP	FLJ46838	no change	no change	no change	SALF	no change	no change	no change
VCaP	SLTM	no change	no change	no change	ZNF621	no change	no change	no change
LNCaP	MDP-1	no change	no change	no change	CHMP4A	no change	no change	no change
LNCaP	MBTPS2	one copy gain	no change	no change	YY2	one copy gain	no change	no change
LNCaP	MRPS10	no change	no change	no change	HPR	no change	no change	no change

Supplementary Table 7. Overall summary of validated chimeras. In-frame chimeras are denoted with an asterik.

Chimera	Chimera Class	Location	5' Gene	Location	3' Gene	Validated in	Validated by
BCR-ABL1	Class I: Translocation	22q11.23	BCR, breakpoint cluster region	9q34.1	ABL1, c-abl oncogene 1, receptor tyrosine kinase	K562	Short read, qRT-PCR
MRPS10-HPR	Class I: Translocation	6p21.1	MRPS10, mitochondrial ribosomal protein S10	16q22.1	HPR, haptoglobin-related protein	LNCaP	Long read, Short read, qRT-PCR, FISH
MIPOL1-DGKB	Class II: Inter-chromosomal complex	14q13.3-q21.1	MIPOL1, mirror-image polydactyly 1	7p21.2	DGKB, diacylglycerol kinase, beta 90kDa	LNCaP	Long read, Short read, qRT-PCR. FISH
*TMPRSS2-ERG**	Class III: Interstitial Deletion	21q22.3	TMPRSS2, transmembrane protease, serine 2	21q22.3	ERG, v-ets erythroblastosis virus E26 oncogene homolog (avian)	VCaP, VCaP-Met	Long read, Short read, qRT-PCR, FISH
*USP10-ZDHHC7**	Class III: Interstitial Deletion	16q24.1	USP10, ubiquitin specific peptidase 10	16q24.1	ZDHHC7, zinc finger, DHHC-type containing 7	VCaP, VCaP-Met	Long read, Short read, qRT-PCR, aCGH
*STRN4-GPSN2**	Class IV: Intra-chromosomal complex	19q13.2	STRN4, striatin, calmodulin binding protein 4	19p13.12	GPSN2, glycoprotein, synaptic 2	Met-3	Short read, qRT-PCR
LMAN2-AP3S1	Class IV: Intra-chromosomal complex	5q35.3	LMAN2 lectin, mannose-binding 2	5q22	AP3S1, adaptor-related protein complex 3, sigma 1	VCaP, VCaP-Met	Short read, qRT-PCR
*HJURP-EIF4E2**	Class IV: Intra-chromosomal complex	2q37.1	HJURP, Holliday junction recognition protein	2q37.1	EIF4E2, eukaryotic translation initiation factor 4E family member 2	VCaP, VCaP-Met	Long read, Short read, qRT-PCR, FISH
*INPP4A-HJURP**	Class II: Intra-chromosomal complex	2q11.2	INPP4A, inositol polyphosphate-4-phosphatase, type I	2q37.1	HJURP, Holliday junction recognition protein	VCaP, VCaP-Met	Long read, Short read, qRT-PCR, FISH
RC3H2-RGS3	Class IV: Intra-chromosomal complex	9q34	RC3H2, ring finger and CCCH-type zinc finger domains 2	9q32	RGS3, regulator of G-protein signaling 3	VCaP, VCaP-Met	Short read, qRT-PCR
ZNF649-ZNF577	Class V: Read-through	19q13.33	ZNF649, zinc finger protein 649	19q13.33	ZNF577, zinc finger protein 577	VCaP, VCaP-Met	Long read, Short read, qRT-PCR
*MBTPS2-YY2**	Class V: Read-through	Xp22.1-p22.2	MBTPS2, membrane-bound transcription factor peptidase, site 2	Xp22.2-p22.1	YY2, YY2 transcription factor	VCaP, LNCaP, VCaP-Met	Long read, Short read, qRT-PCR
C19ORF25-APC2	Class V: Read-through	19p13.3	C19ORF25, chromosome 19 open reading frame 25	19p13.3	APC2, adenomatosis polyposis coli 2	LNCaP	Long read, Short read, qRT-PCR
WDR55-DND1	Class V: Read-through	5q31.3	WDR55, WD repeat domain 55	5q31.3	DND1, dead end homolog 1 (zebrafish)	RWPE	Long read, Short read, qRT-PCR
*SLC45A3-ELK4**	Class V: Read-through	1q32.1	SLC45A3, Solute carrier family 45 member 3	1q32.1	ELK4, ETS domain-containing protein	*Met-4*	Short read, qRT-PCR

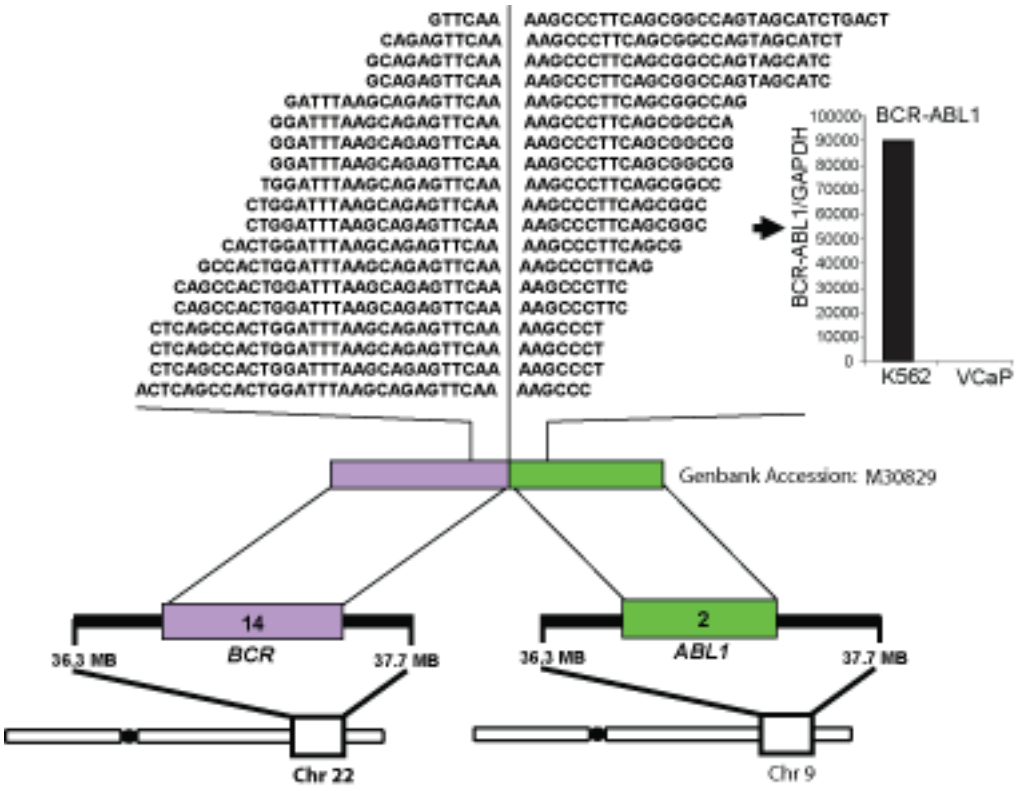
Supplementary Table 8. Primer sequences used for confirming fusion genes by qRT-PCR.

Fusion Gene	Primer Sequence (5'-3')
ARHGEF12-SCD-F	GCTAAGGAAAGGGTGGGATG
ARHGEF12-SCD-R	TTGTGTTTGTTCATAATAAAAAGTGAA
BCR-ABL(b3a2)-F	GAGTCTCCGGGGCTCTATGG
BCR-ABL(b3a2)-F	GCCGCTGAAGGGCTTTTGAA
DNM1L-KLK2-F	GGATCCTCCCCTTCTTTCTG
DNM1L-KLK2-R	CAAAACTTGCTAGTTACTGCCTACC
EFTUD2-NDUFB2-F	CCCAGCACCTCTTCTGAGTC
EFTUD2-NDUFB2-R	AGAGAGGGGTGTAGGCATCA
EGLN2-RAB4B-F	GGATTGTCAACGTGCCCTAC
EGLN2-RAB4B-R	GAGCTAGACCCGGAGAGGAT
EIF4A2-SPDEF-F	GTGCACGAACTGGTAGACGA
EIF4A2-SPDEF-R	GGCAGAAAGCAACACAACCT
LMAN2-AP3S1-F	ACTGACGGCAACAGTGAACA
LMAN2-AP3S1-R	TGGAAAGTCTCCCTGATGATTT
MDS1-EVI1-F	ATGCAACAAGGTTGTGCTGA
MDS1-EVI1-R	CAAACCTGAAAGACCCCAGT
MIA2-CTAGE5-F	AGCCGACTCCTAACCGATCT
MIA2-CTAGE5-R	TGAATTCTGCATTTTCACCAA
MIPOL1-DGKB-F	CAGAGCGAGCAAATATGGAA
MIPOL1-DGKB-R	CTTGCTTCGGTTTCTTGTCC
NDRG1-SF3B5-F	CAAAAACGAGACGCCAAATC
NDRG1-SF3B5-R	CAAAAACAAGACGCGTAGCA
PDCL2-CLOCK-F	GAAGCGGTTACAGGAATGGA
PDCL2-CLOCK-R	TTCTGAGCTCCAGCAGCTTT
PRKAR1A-HEXIM1-F	GAACTGAGCAGAGCAGAGCA
PRKAR1A-HEXIM1-R	CATTTGGCATTAACAAAGATCAA
RBM14-RBM4-F	GTGTGACGTGGTGAAAGGTG
RBM14-RBM4-R	AAATGGGCAGGAGAGGAAAG
RC3H2-RGS3-F	GCTAATGGTCAGAATGCTGCT
RC3H2-RGS3-R	CTTCTTCTGCTCCTGCGAGT
SLC35A3-HIAT1-F	GCTGTCAATAGTCCCCAAGC
SLC35A3-HIAT1-R	GGATTTGCAACCTCTTTATCG
SMAD5-IDH1-F	TTTGGGGATAAGGGAAAAGG
SMAD5-IDH1-R	GCTTTGCTCTGTGGGCTAAC
STRN4-GPSN2-F	CTGGGGGACTTGGCAGAT
STRN4-GPSN2-R	TCCAAGAAACACAGCTTCTCC
TEAD1-ASCC3L1-F	GGCTCAGGTTGTGGTAGAGG
TEAD1-ASCC3L1-R	TTGAGCCTGTCCTGGAACTT
TMPRSS2-ERG-F	GGAGTAGGCGCGAGCTAAG
TMPRSS2-ERG-R	GTCCATAGTCGCTGGAGGAG
USP10-ZDHHC7-F	CGGAGTCCCAATGAAACG
USP10-ZDHHC7-R	GAGGAGGAGGACGATGAAGA
ZNF577-ZNF649-F	CCTTCCCAGAAGTGGTGGT
ZNF577-ZNF649-R	CACACGGGAGAGAGACCCTA
MRPS10-HPR-F	GATTCTTGGGCTTCCCACAT
MRPS10-HPR-R	CAAAGACACAATTAGAACAGTTACCA
SLC45A3-ELK4-F	GCAGATCCTGCCCTACACAC
SLC45A3-ELK4-R	AGCTGAAGAAGGAACTGCCA

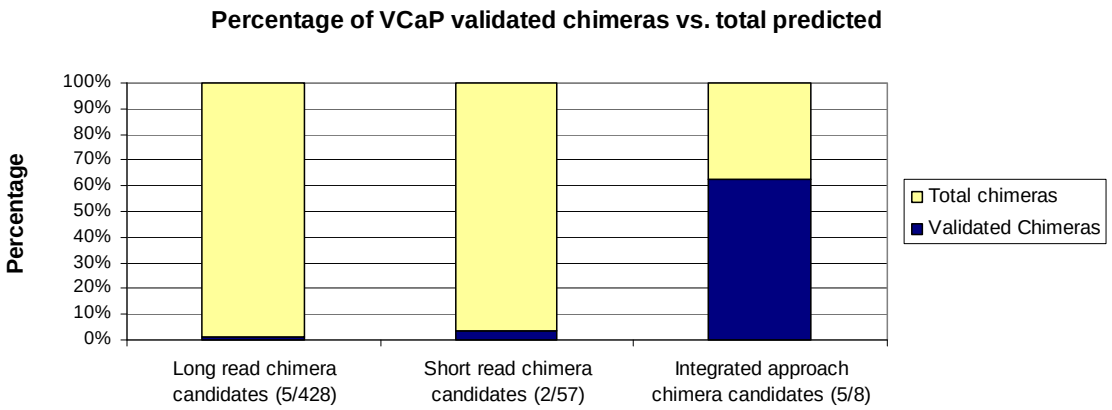
Supplementary Table 9. Sequences of chimeric transcripts, with GenBank accession numbers, reported in this manuscript. Fusion junction is denoted by '*'.

>*TMPRSS2-ERG* FJ423744
GGAGTAGGCGCGAGCTAAGCAGGAGGCGGAGGCGGAGGCGGAGGGCGAGGGGCGGGGAGC
GCCGCCTGGAGCGCGGCAG*GAAGCCTTATCAGTTGTGAGTGAGGACCAGTCGTTGTTTGA
GTGTGCCTACGGAACGCCACACCTGGCTAAGACAGAGATGACCGCGTCCTCCTCCAGCGA
CTATGGACAGACTTCCAAGATGAGCCCACGCGTCCCTCAGCAGGATTGGCTGTCT
>*INPP4A-HJURP* FJ423742
AGGTCTCAAGAATCAAAAACAAAACAAAAATACAAACAGAGAGCAAGTGGGAAGATAAAT
AACACTCCGAAATAACCTAGCTACACACTTTTAGTTTCCAATTTTTCTTAGCATGAAATC
ACTTTTCTCTTCCATCCTGTAAGACGTGTTCTCTCCT*CTGCGCATGCACTCCAGGGCCTG
GGTGAAGACCTGCGGGGCCATGCCATGCTCGTGTTGCAGGATCAGGCACTGCTCCAGTGT
CACCG
>*ZNF649-ZNF577* FJ423743
GGGGCTAGCAACTCTAGTATGTTTTCTCTCTTCTGTCTATTCTGGGCCTTCCCAGAAGTG
GTGGTCAGGTATCATCTCAGGTCAAGCTACCACTGGAAATGATGATCTTCCCCAGCCTGG
AAGCTCCTTCTTCCATTACTGAAAATGTCTTGTTCCTATAGGCCAGAAC*ACTCATCACAG
CCATAGGGTCTCTCTCCCGTGTGAGTTCTGTGATGTACAATGAGCATTG
>*USP10-ZDHHC7* FJ423745
ACGCGGGGGAAGCAGCGTGAGCAGCCGGAGGATCGCGGAGTCCCAATGAAACGGGCAGCC
ATGGCCCTCCACAGCCCGCAG*GGTGCGTCAGGGAAATCATGCAGCCATCAGGACACAGGC
TCCGGGACGTCGAGCACCATCCTCTCCTGGCTGAAAATGACAACTATGACTCTTCATCGT
CCTCCTCCTCCGAGGCTGACGTGGCTGACCGGGTCTGGTTCATCCGTGACGG
>*HJURP-EIF4E2* FJ423746
CGATTCTTGTCTCGTTCCGTTTTTTCCTTCTCACCATCTTTCTGTGTGCTGTTTTCTTCA
TTCTGATCATGGTCCCCACTGTCATCATCTTTCAAA*CTCTCTTCTGAGTTGGGCTGTGAA
GAGCTGCCCTGGTCTCCCGGTCTGACGGTGTTGTCCACCCCATCTGAGGCACCCAGGGAA
TTGCCCTGGCGTCCGGAGCCCGTGGGTTCTGATAGCCTGGGTCTTTTTGCAGGGAACTGA
TGGT
>*MIPOL1-DGKB* FJ423747
ACAGAGAGAACATTGTTTCCATCACTCAACAACAAAATGAGGAACTGGCTACTCAACTGC
AACAAGCTCTGACAGAGCGAGCAAATATGGAATTACAACTTCAACATGCCAGAGAGGCCT
CCCAAGTGGCCAATGAAAAAGTTCAAAA*ATAAAAATTACACACAAGAACCAAGCCCCAAT
GCTGATGGGCCCGCCTCCAAAAACCGGTTTATTCTGCTCCCTCGTCAAAAGGACAAGAAA
CCGAAGCAAGGAATAA
>*MRPS10-HPR* FJ423748
GTCACTGGGTTTGCCGGATTCTTGGGCTTCCCACATA*TTTCTTCTTTTTCTTCTGATAGT
GTTTCCCAGATTGGCTCCTTGATGTGTTCTGGTAACTGTTCTAATTGTGTCTTTGTTACT
TCCATGGCAACCCCTTCAGGTAAGTTTCA
>*WDR55-DND1* FJ423749
CGCAAAAAAAAGGGAGGACCACTGCGGGCTCTGAGCAGCAAGACTTGGAGCACCGATGAC
TTCTTCGCAGGACTGAGGGAAGAGGGAGAAGACTCCATGGCTCAGGAAGAAAAGGAGGAG
ACTGGGGATGACAATGACTGAAGGAATGAATTGAATCTTGAGACGGGTCCTCACCAGGGT
GCCTGTGGAGAAAGAATGGAGTCACTGTTTAACCATGGTACCTGCCTCAGCCCCAGCAGA
CCACAGGAGGTTCGG
>*C19orf25-APC2 (Intron)* FJ423750
GAATCGGAAGTGGCTGCGTCGTCGACGCTGGGCTTTCGGGTCCCGCGCCCAGAGATGGGC
TCCAAGGCAAAGAAGCGCGTGCTGCTGCCCCACCCGCCCAGCGCCCCCCACGGGTGGAGC
AGATCCTGGAGGATGTGCGGGGTGCGCCGGCAGAGGATCCAGTGTTCACCATCCTGGCCC
CGGAAG*GCTGGAGTGCAGTGGCGAGATCTCGACTCACTGCAGGCTCCGACTCCCCAGTTC
AAGCGATT
>*MBTPS2-YY2* FJ423751
TTGGGATTTTTCTCTTCATTATTTATCCCGGAGCATTTGTTGATCTGTTCACCACTCATT
TGCAACTTATATCGCCAGTCCAGCAGCAAGGATATTTTGTGCAG*CCATGGCCTCCAACGA
AGATTTCTCCATCACACAAGACCTGGAGATCCCGGCAGATATTGTGGAGCTCCACGACAT

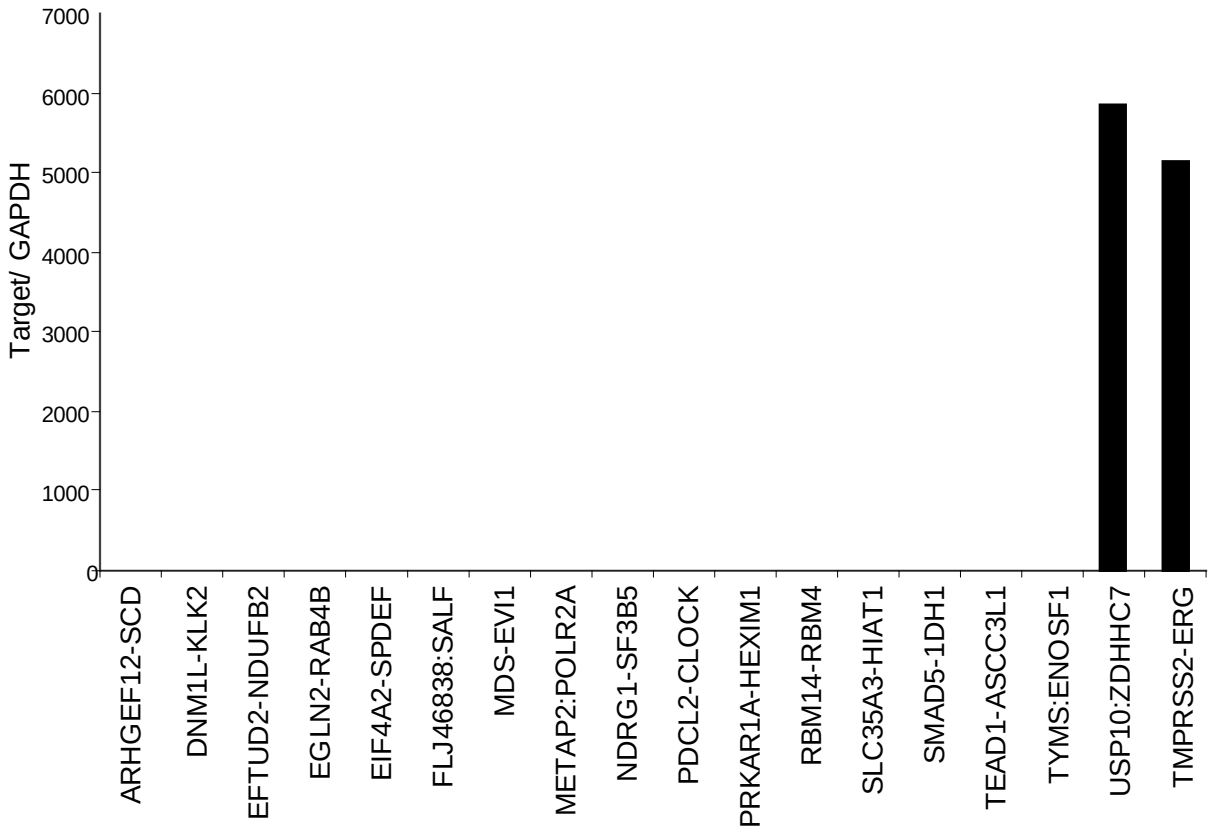
CAATGTGGAGCCCCTTCCTATGGAGGACATTCCGACGGAAAGCGTCCAGTACG
>*STRN4-GPSN2* FJ423752
CTGGGGGACTTGGCAGATCTCACCGTCACCAACGACAACGACCTCAGCTGCGAT*GTGGA
GATTCTGGACGCAAAGACAAGGGAGAAGCTGTGTTTCTTGGA
>*LMAN2-AP3S1* FJ423753
ACTGACGGCAACAGTGAACATCTCAAGCGGGAGCATTCGCTCATTAAGCCCTACCAAG*A
GTGAAGATACACAACAGCAAATCATCAGGGAGACTTTCCA
>*RC3H2-RGS3* FJ423754
GCTAATGGTCAGAATGCTGCTGGGCCCTCTGCAGATTCTGTAACTGAAAA*AAGGCAGAG
TGCTTATTCACTTTGGAAGCGCACTCGCAGGAGCAGAAGAAG
>*SLC45A3-ELK4* FJ423755
GCTGAAGAAGGAACTGCCACAGGGTGATAGCACTGTCCATAGCAATGAG*CTGCTTCTCC
CGGTGGTAGAGGGAGGCCAGTGTGTAGGGGAGG



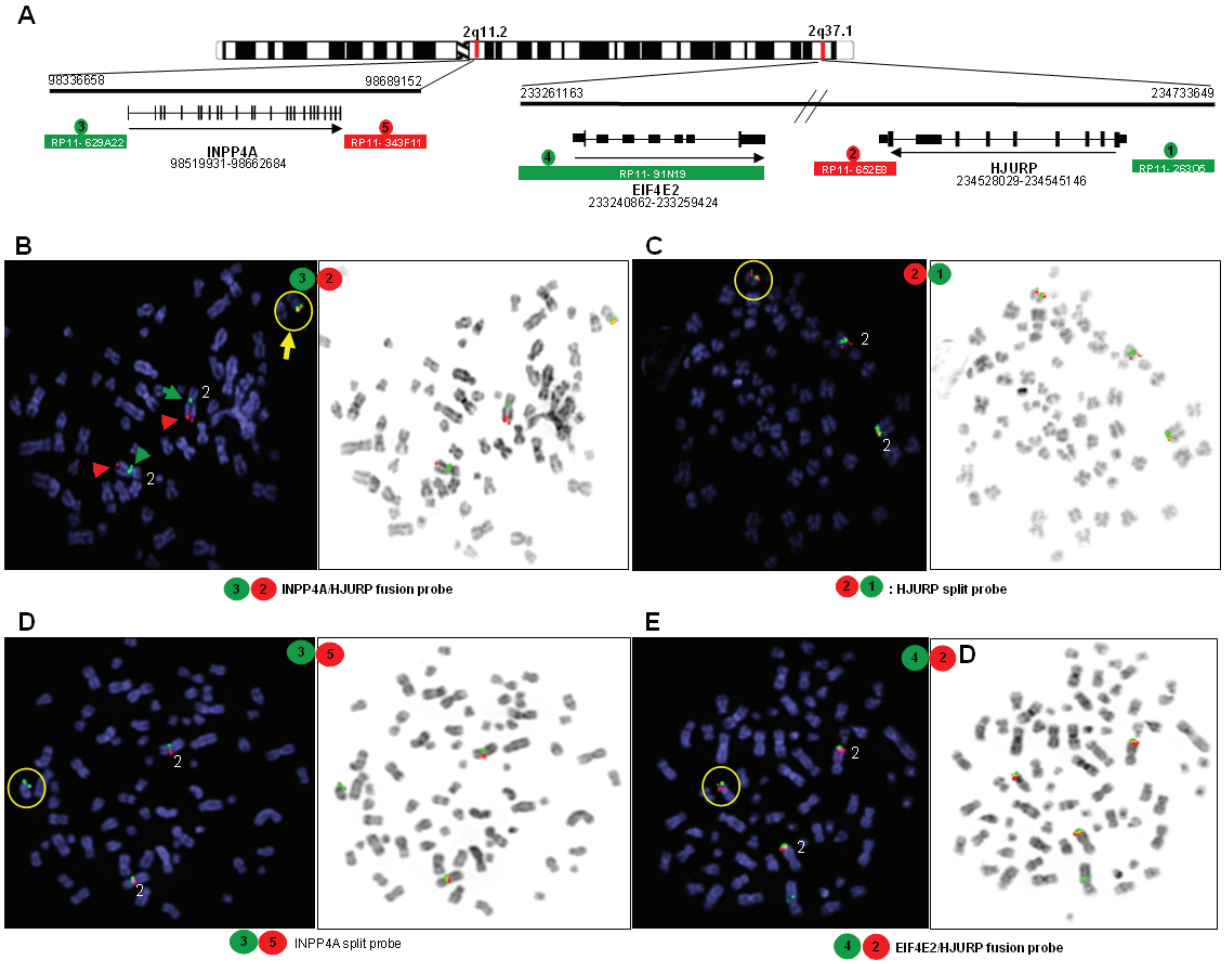
Supplemental Figure 1: "Re-discovery" of the *BCR-ABL1* gene fusion using massively parallel sequencing of the transcriptome in the chronic myelogenous leukemia cell line K652. a, The schematic bar represents the reference sequence *BCR-ABL1* (Genbank Accession No. M30829). *BCR* (purple), located on chromosome 22, fused with *ABL1* (green) on chromosome 9. Short read sequencing obtained from the Illumina platform are aligned above the bar. The inset represents qRT-PCR validation of the expression of *BCR-ABL1* fusion gene in K562 cells.



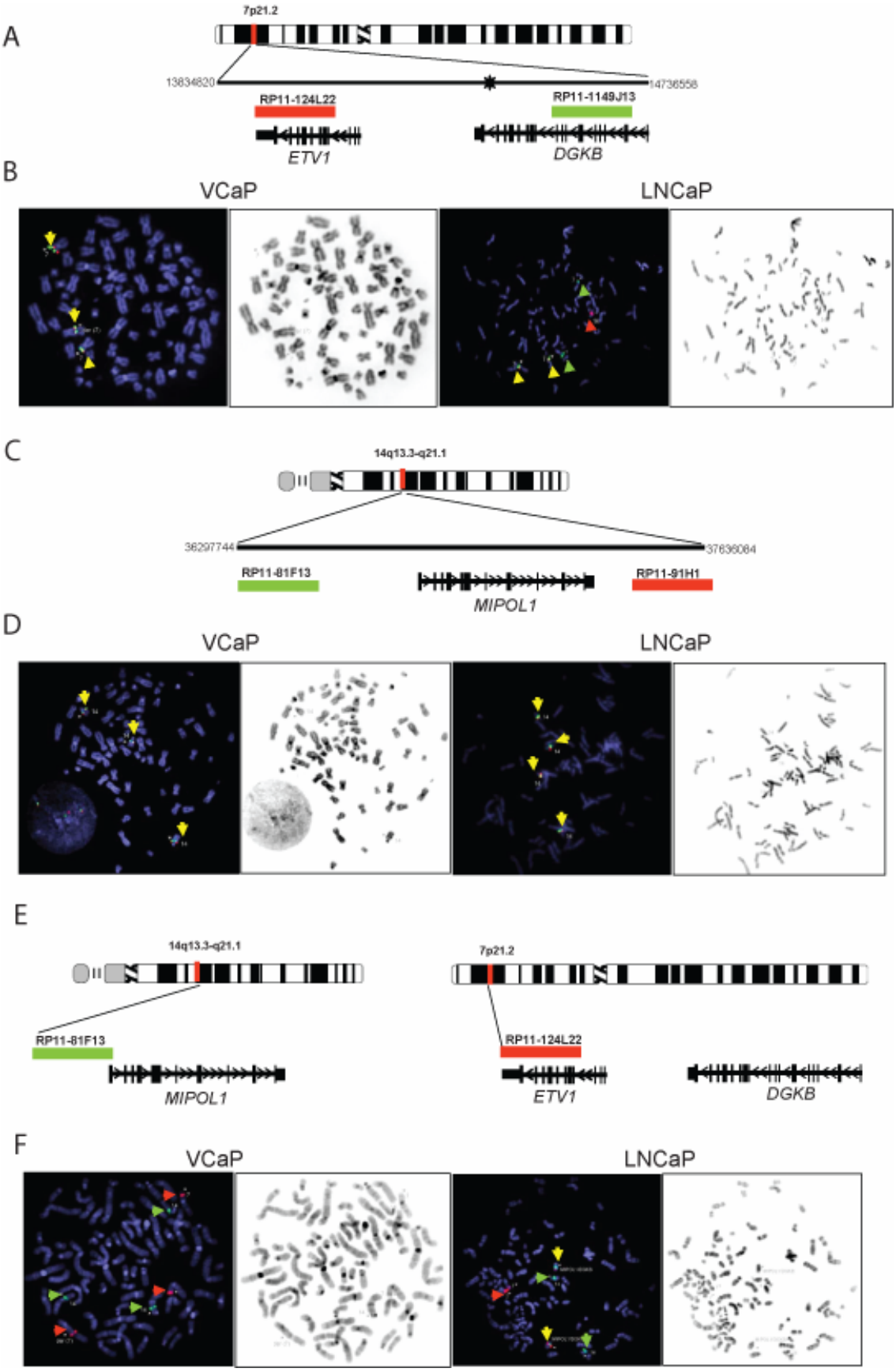
Supplemental Figure 2: Histogram of predicted VCaP validated chimeras compared to total number of computationally predicted chimeras based on long read technology, short read technology, and an integrative approach.



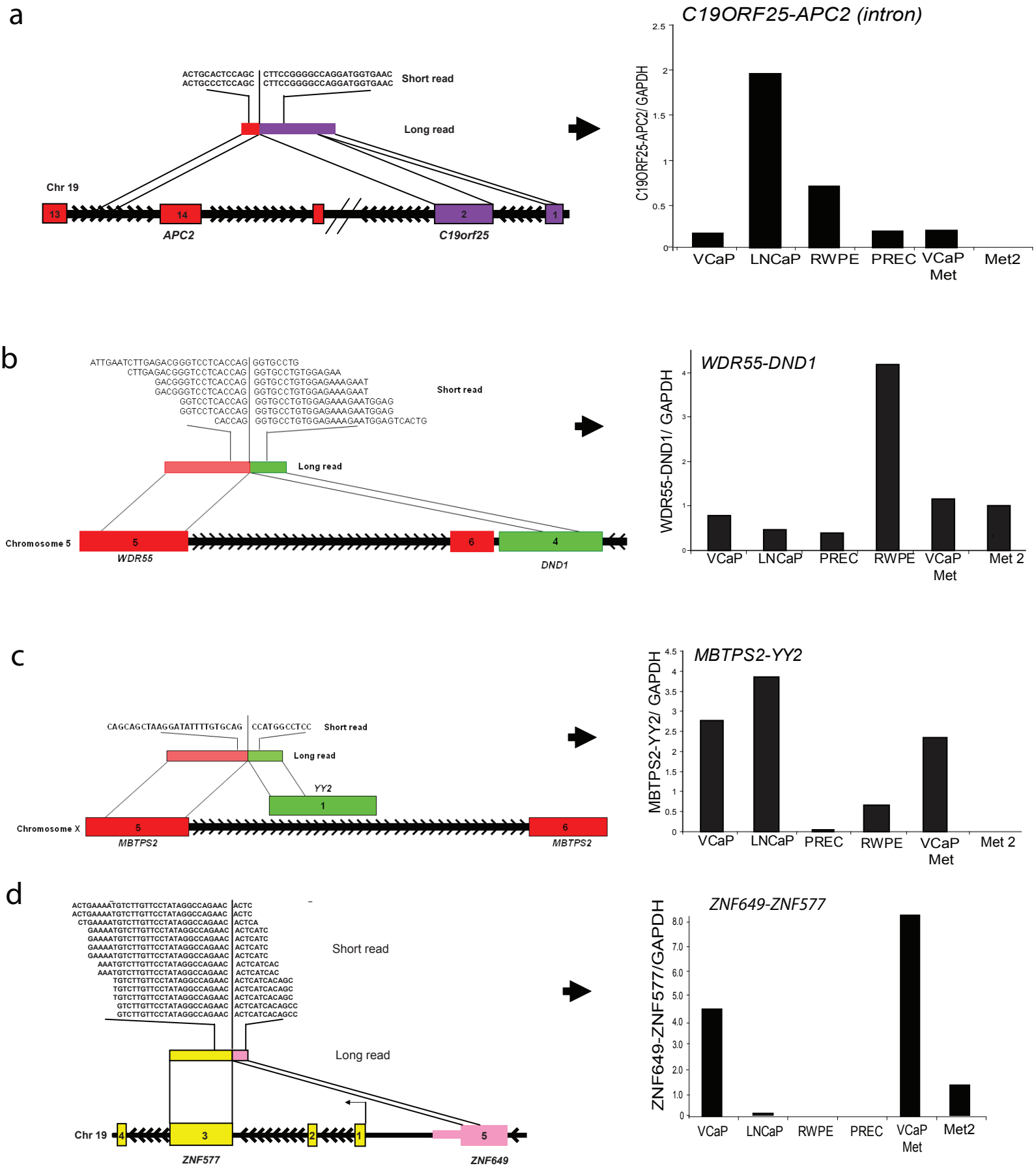
Supplemental Figure 3: Fusion-chimeras nominated by long read sequences that failed validation by qRT-PCR. *TMPRSS2-ERG* and *USP10-ZDHHC7* were the only two chimeras validated in this set of eighteen candidates in VCaP cells.



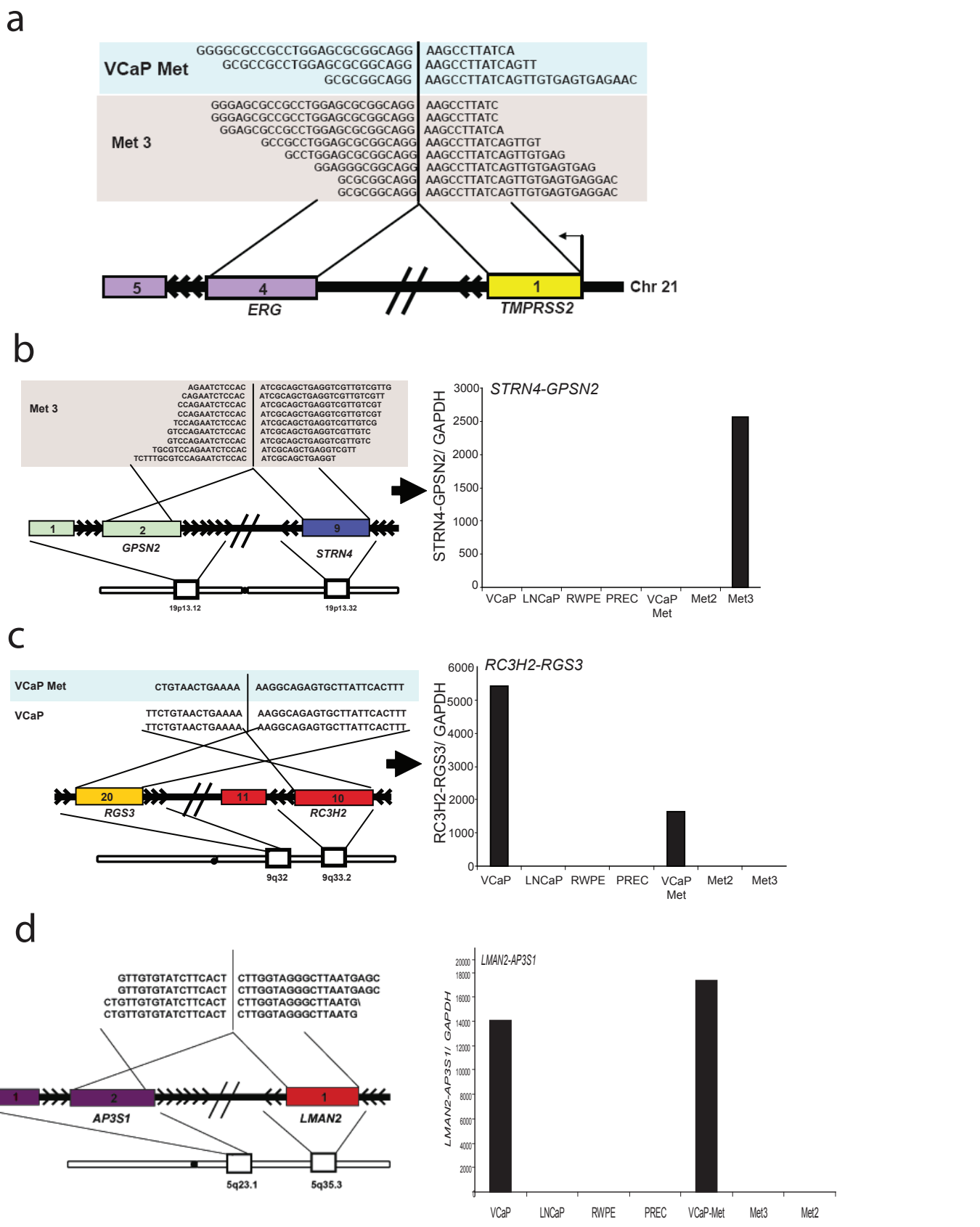
Supplemental Figure 4: FISH analysis of the chromosomal rearrangements at 2q11 and 2q37, involving *INPP4A*, *EIF4E2* and *HJURP* genes. a, Schematic showing genomic organization of *INPP4A*, *EIF4E2* and *HJURP* genes. Horizontal red and green bars indicate the location of BAC clones. **b,** FISH analysis using BAC clones 2 and 3 showing the fusion of *INPP4A* and *HJURP* genes on a marker chromosome (yellow circle). Green and red arrow indicate the hybridization of 5'*INPP4A* probe at 2q11 and 3'*HJURP* probe at 2q37, respectively, on two copies of normal chromosome 2. **c,** Hybridization of *HJURP* probe to two normal copies of chromosome 2 and on the marker chromosome (yellow circle) suggest a breakpoint between *EIF4E2* and *HJURP* genes resulting in translocation of 3' end of chromosome 2q onto the marker chromosome. **d,** Hybridization of probes 2 and 4 onto two normal chromosome 2, marker chromosome (yellow circle) and a split green signal on the derivate chromosome 2 (confirming a breakpoint within probes 2 and 4 resulting in an insertion into the marker chromosome. **e,** Rearrangement of *INPP4A* gene confirmed by the presence of probe 3 on the marker chromosomes (yellow circle) in addition to the co-localizing signal on two copies of normal chromosome 2.


A
7p21.2
13834820
14736558
RP11-124L22
RP11-1149J13
ETV1
DGKB
B
VCaP
LNCaP
C
14q13.3-q21.1
36297744
37639084
RP11-81F13
RP11-91H1
MIPOL1
D
VCaP
LNCaP
E
14q13.3-q21.1
7p21.2
RP11-81F13
RP11-124L22
MIPOL1
ETV1
DGKB
F
VCaP
LNCaP

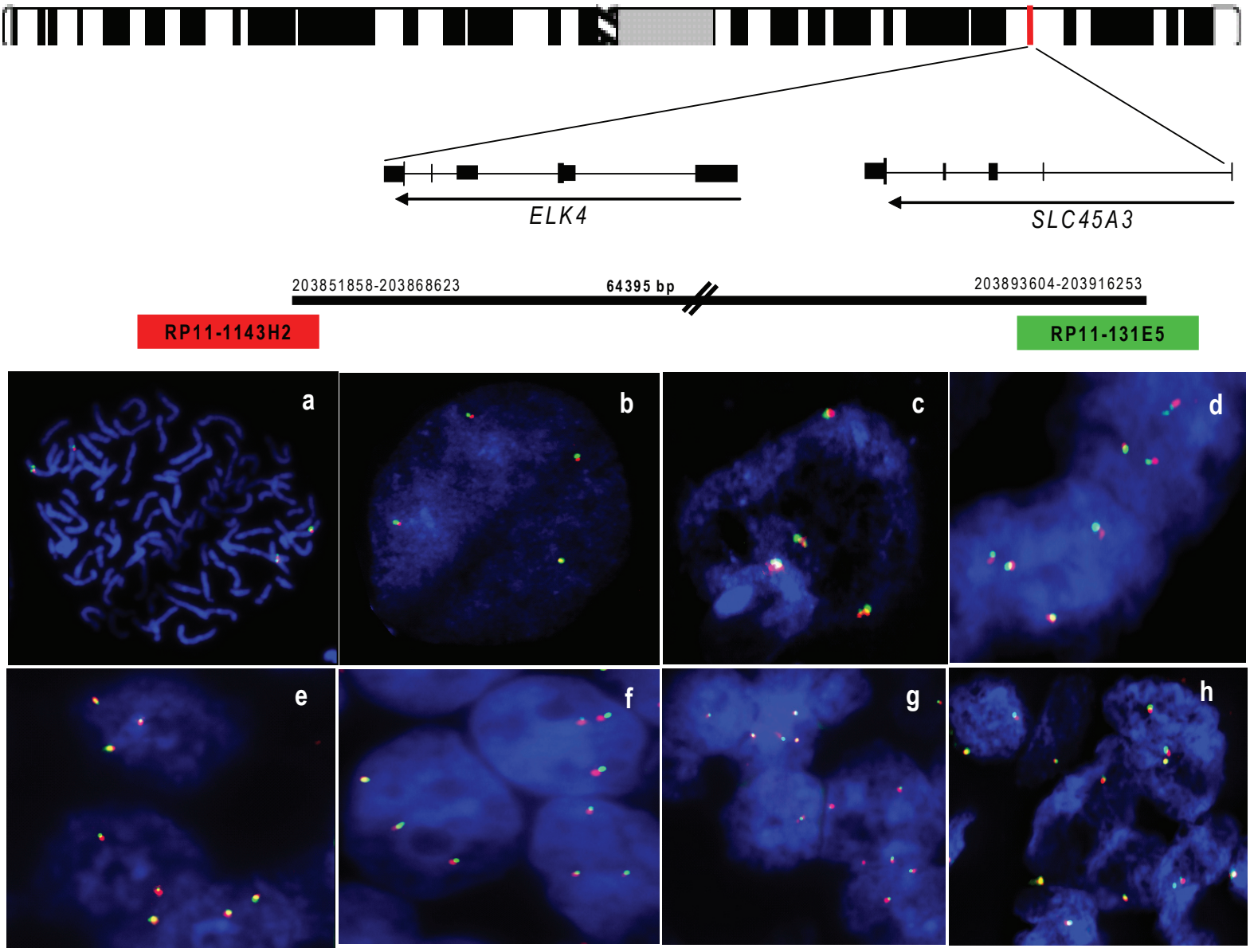
Supplemental Figure 5: FISH analysis of the chromosomal rearrangements involving *MIPOL1*, *DGKB*, and *ETV1*. a, Schematic of the genomic organization of *ETV1* and *DGKB* locus on chromosome 7p21.2. Gene orientation is indicated by arrows. Previously identified genomic breakpoint in *DGKB* is marked with a star. FISH analysis was performed using BAC clones on VCaP and LNCaP cells. Probe locations encompassing both *ETV1* and *DGKB* are indicated with horizontal red and green bars, respectively. Genomic coordinates indicate the region spanning the two BAC clones. **b,** Co-localized signals (normal) are indicated by yellow arrows and red and green arrowheads indicate the split signal. Split signals are observed only in LNCaP cells. The rearranged signal (red) was observed on chromosome 14. **c,** Schematic diagram showing genomic organization of *MIPOL1* locus on chromosome 14q13.3-q21.1, **d,** FISH analysis did not reveal split signals in LNCaP or VCaP cells. **e,** Genomic organization of *MIPOL1*, *ETV1*, and *DGKB* gene locus on chromosomes 7p21.2 and 14q13.3-q21.1, respectively. BAC clones from the 3' end of *ETV1* (red bar) and the 5'end of *MIPOL1* (green bar) were employed to detect co-localization of *ETV1* downstream of *MIPOL1*. **f,** FISH analysis shows co-localizat ion of in LNCaP but not VCaP cells.



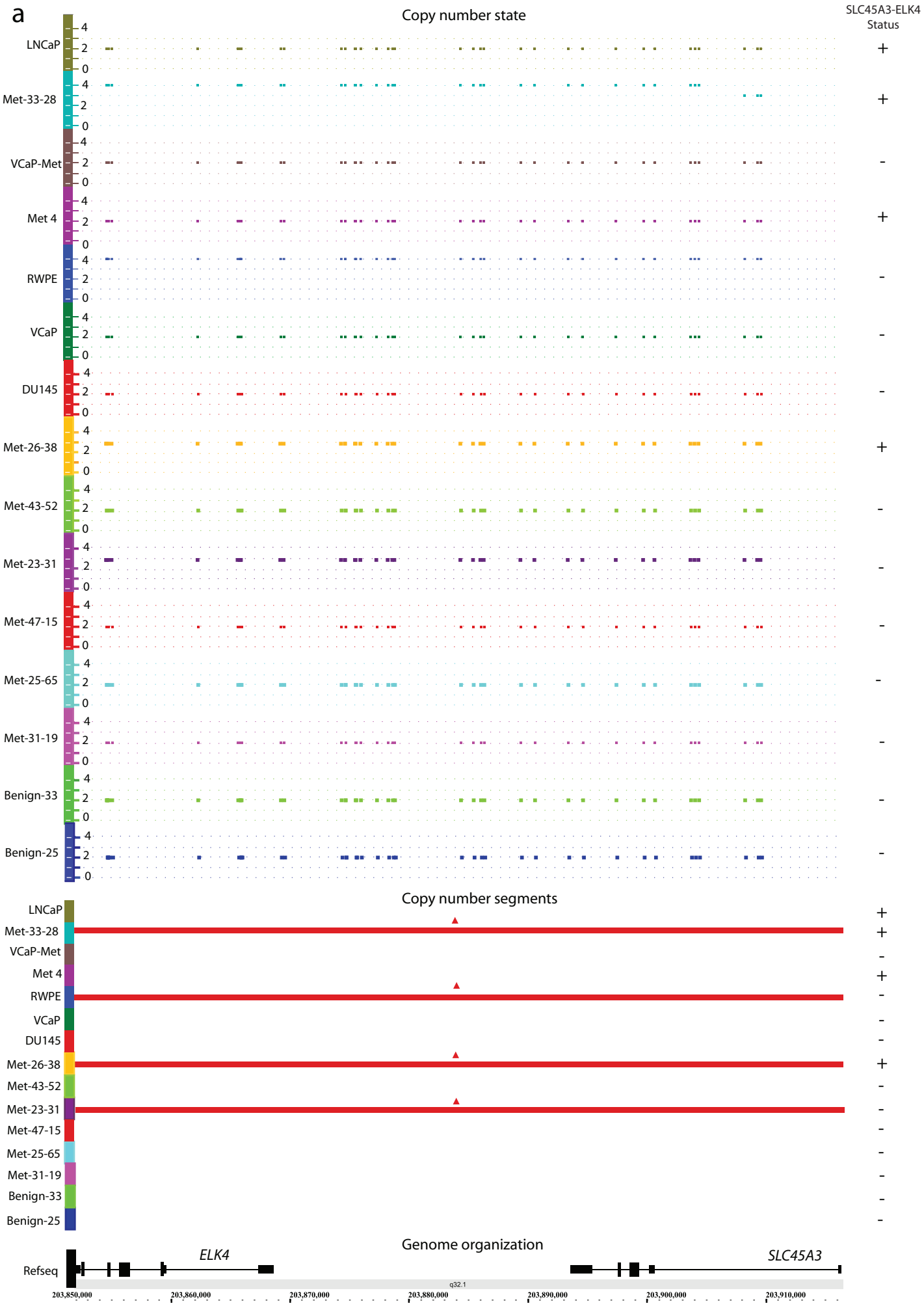
Supplemental Figure 6: Chimeric Class V, Read-through fusions. Schematics of the read-through fusions accompanied with qRT-PCR validations of the fusion transcripts in prostate cancer cell lines VCaP and LNCaP, metastatic prostate tissues VCaP-met and Met 2, and benign prostate cell lines, RWPE and PREC, **a,** *C19orf25-APC2* (intron), **b,** *WDR55-DND1*, **c,** *MBTPS2-YY2*, and **d,** *ZNF649-ZNF577*.

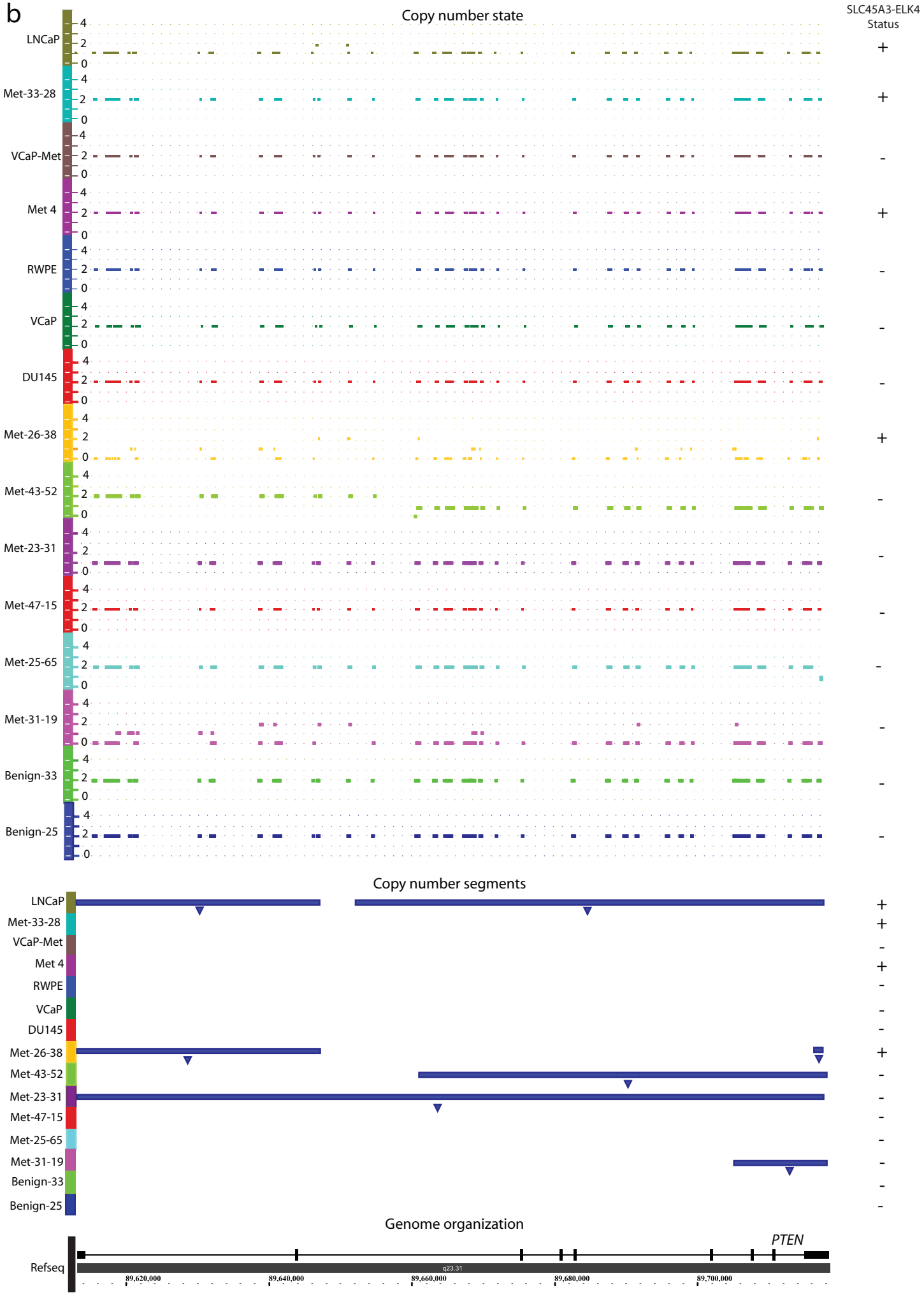


Supplemental Figure 7: Chimera candidates in prostate tissues. a, Schematic of *TMPRSS2-ERG* fusion boundary populated with short reads sequenced in both VCaP-Met (light blue) and Met 3 (gray) tissues. **b,** Schematic of the STRN4-GPSN2 fusion on chromosome 19 in the metastatic prostate cancer tissue, Met 3. The 5' portion of STRN4 (purple) is fused with exon 2 of GPSN2 (green), which resides in the opposite orientation on the same chromosome. **c,** Schematic of RC3H2-RGS3 fusion on chromosome 9 in metastatic prostate cancer tissue, VCaP-Met. The 5' portion of RC3H2 (red) is fused with exon 20 of RGS3 (yellow), which resides in the opposite orientation on the same chromosome. **d,** Schematic of the complex intra-chromosomal gene fusion between exon 1 of lectin, mannose-binding 2 (*LMAN2)* and exon 2 of adaptor-related protein complex 3, subunit 1 (*AP3S1*). qRT-PCR validation of *LMAN2-AP3S1* fusion transcript expression in prostate cancer cell line, VCaP and metastatic prostate tissue, VCaP-Met.

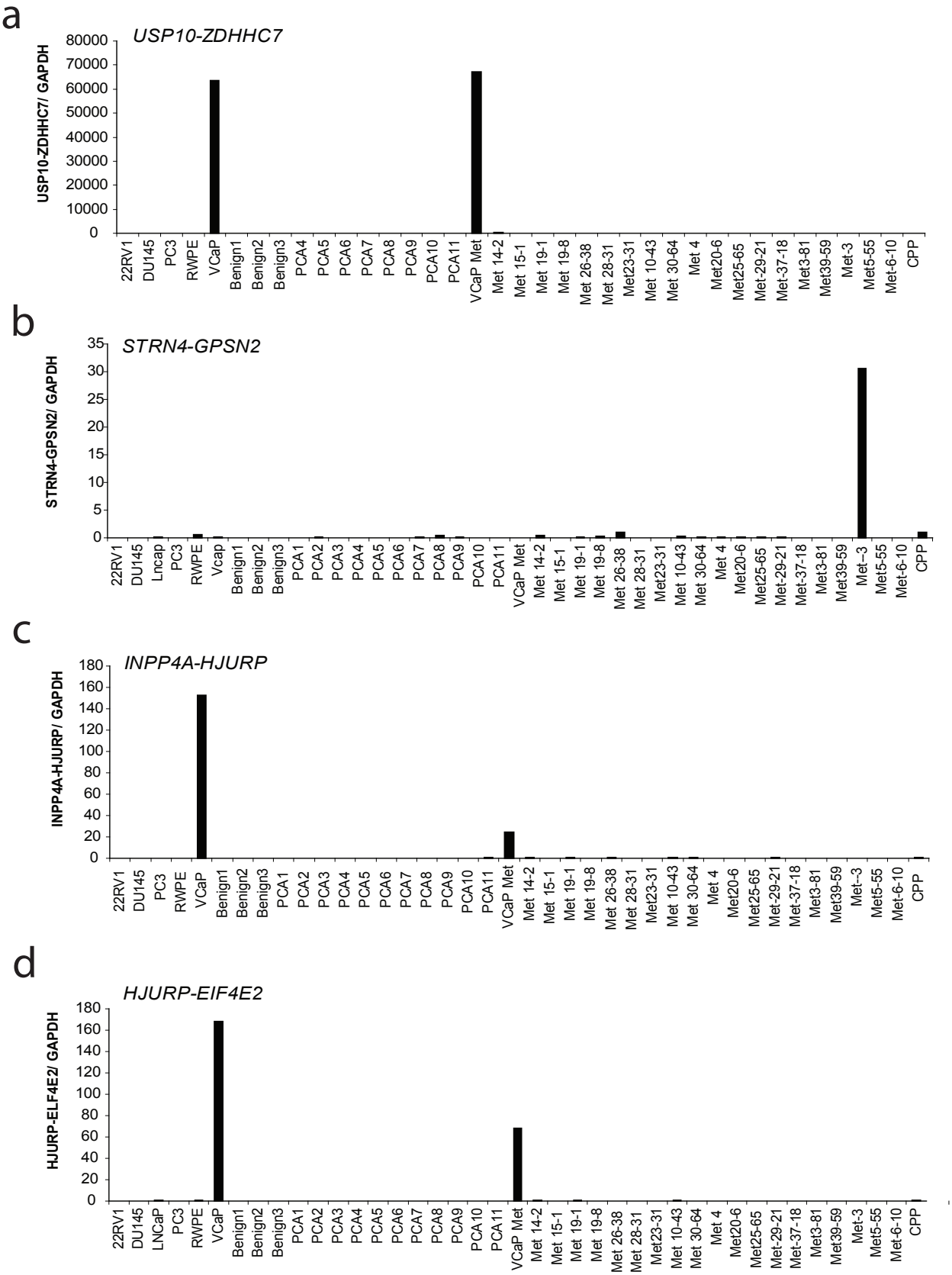


Supplemental Figure 8: Lack of rearrangement of the *SLC45A3-ELK4* locus in prostate cancers that express the *SLC45A3-ELK4* mRNA chimera. Fluorescence in situ hybridization analysis of the *ELK4* gene for rearrangement. Schematic diagram (top panel) shows the genomic organization of the *SLC45A3* and *ELK4* genes on chromosome 1q32.1. BAC clones (red and green horizontal bars with BAC clone ID) were derived from the immediately flanking 3' and 5' regions of *ELK4* and *SLC45A3* genes, respectively. Probes were hybridized on the *SLC45A3-ELK4* chimera positive cell line LNCaP (a, metaphase spread; b, interphase) , and 5 index prostate tumors that express the mRNA chimera (a, e, f, g & h). **c,** DU145 is a an *SLC45A3-ELK4* chimera negative prostate cancer cell line. All the samples show co-localization of red and green signals indicating the absence of gross rearrangement or large deletions within the genes.

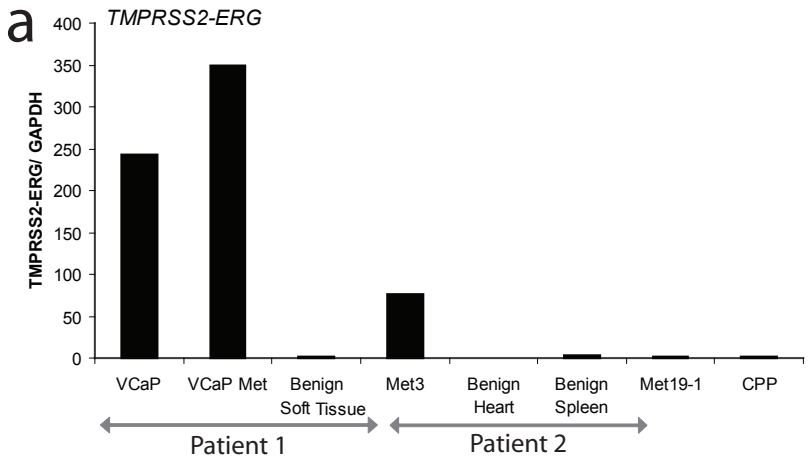

a
Copy number state
SLC45A3-ELK4
Status
LNCaP
Met-33-28
VCaP-Met
Met 4
RWPE
VCaP
DU145
Met-26-38
Met-43-52
Met-23-31
Met-47-15
Met-25-65
Met-31-19
Benign-33
Benign-25
Copy number segments
Genome organization
Refseq
ELK4
SLC45A3
q32.1
203,850,000
203,860,000
203,870,000
203,880,000
203,890,000
203,900,000
203,910,000

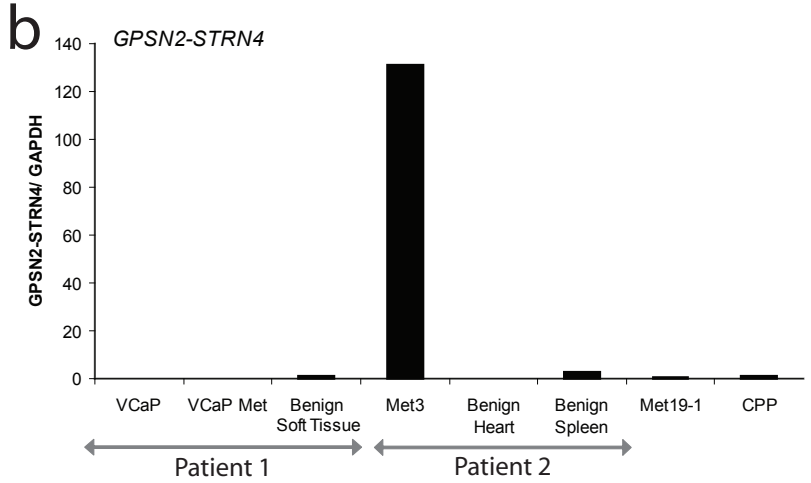

b
Copy number state
SLC45A3-ELK4 Status
LNCaP
Met-33-28
VCaP-Met
Met 4
RWPE
VCaP
DU145
Met-26-38
Met-43-52
Met-23-31
Met-47-15
Met-25-65
Met-31-19
Benign-33
Benign-25
Copy number segments
Genome organization
PTEN
Refseq
89,620,000
89,640,000
89,660,000
89,680,000
89,700,000

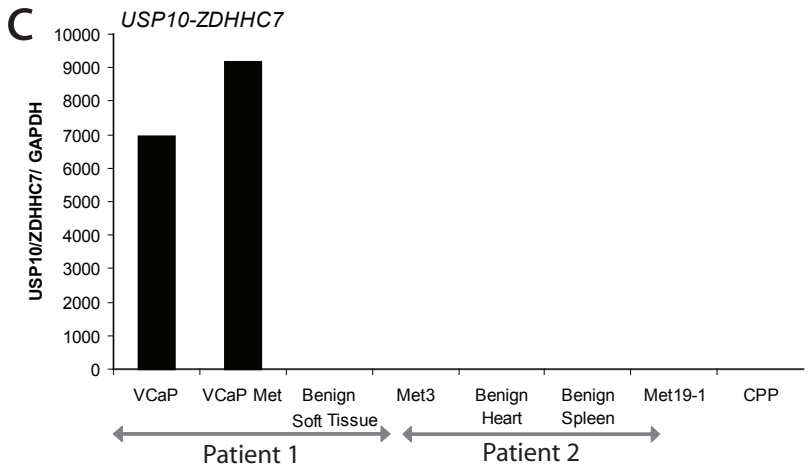
Supplemental Figure 9: Genomic level analysis, using Affymetrix SNP 6.0, of 15 samples using the Genotyping Console software. Copy number states are divided into the following categories: 0 - homozygous deletion; 1 - heterozygous deletion; 2 - normal diploid; 3 - single copy gain; and 4 - multiple copy gain. Copy number segments highlight both amplified (red) and deleted (blue) genomic segments. Genome organization shows the genomic aberrations relative to **(a)** *SLC45A3-ELK4* and **(b)** *PTEN*. For each sample, the *SLC45A3-ELK4* status is shown on the right using '+' and '-'. Overall, we did not observe a genomic deletion within the *SLC45A3-ELK4* region. As a positive control, we have demonstrated the ability of our SNP array analysis to detect the *PTEN* loss in multiple samples. While the *SLC45A3-ELK4* region lacked a deletion, multiple samples (which are not correlated with *SLC45A3-ELK4* status) show a copy number gain.

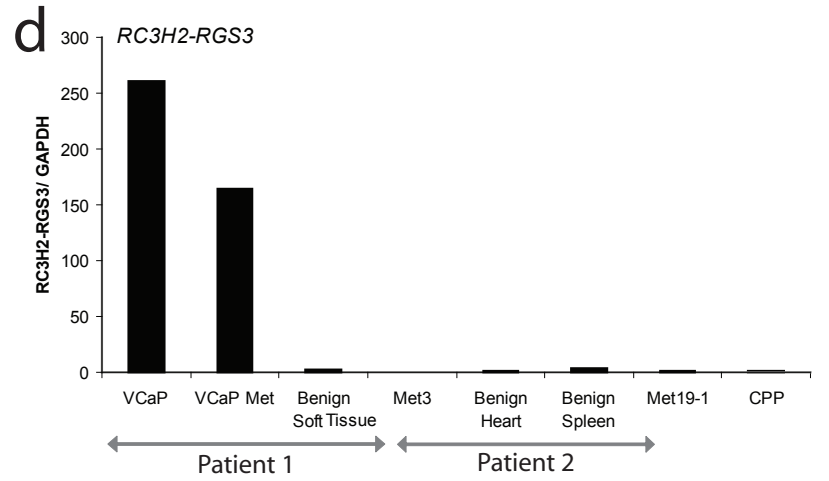


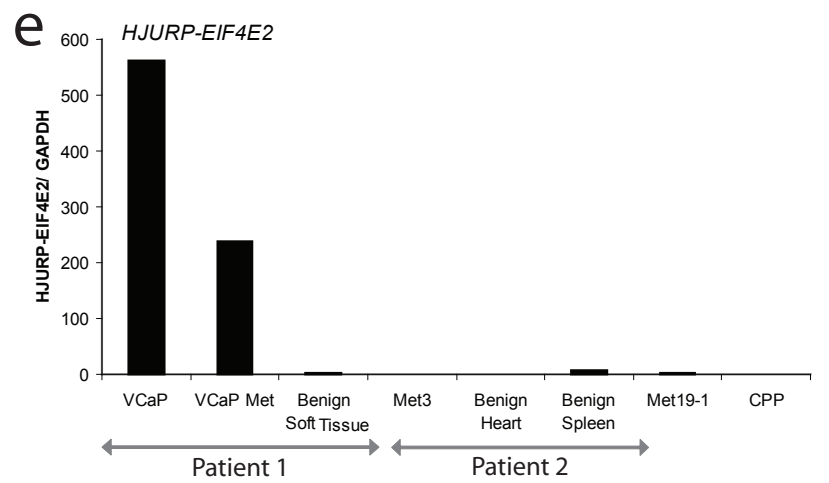
Supplemental Figure 10: qRT-PCR based survey of a panel of prostate cancer cell lines and tissues- benign, localized prostate cancer, and metastatic tissues for recurrence. *USP10-ZDHHC7* (a)*, INPP4A-HJURP* (c), and *HJURP-EIF4E2* (d) all show expression in VCaP and VCaP-Met, as expected, and were not confirmed in any other samples from the panel. (b) *STRN4-GPSN2* expression is confirmed in Met 3 but does not appear to be expressed in any other sample tested.

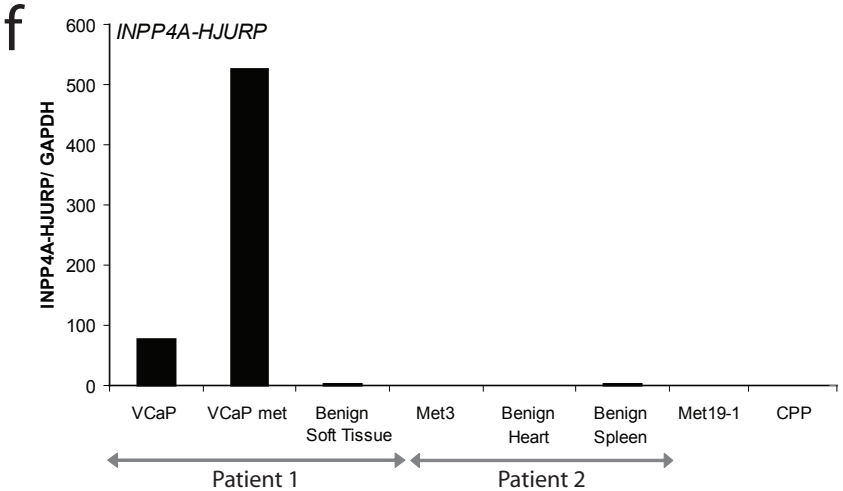

a
TMPRSS2-ERG
TMPRSS2-ERG/ GAPDH
400
350
300
250
200
150
100
50
0
VCaP
VCaP Met
Benign Soft Tissue
Met3
Benign Heart
Benign Spleen
Met19-1
CPP
Patient 1
Patient 2

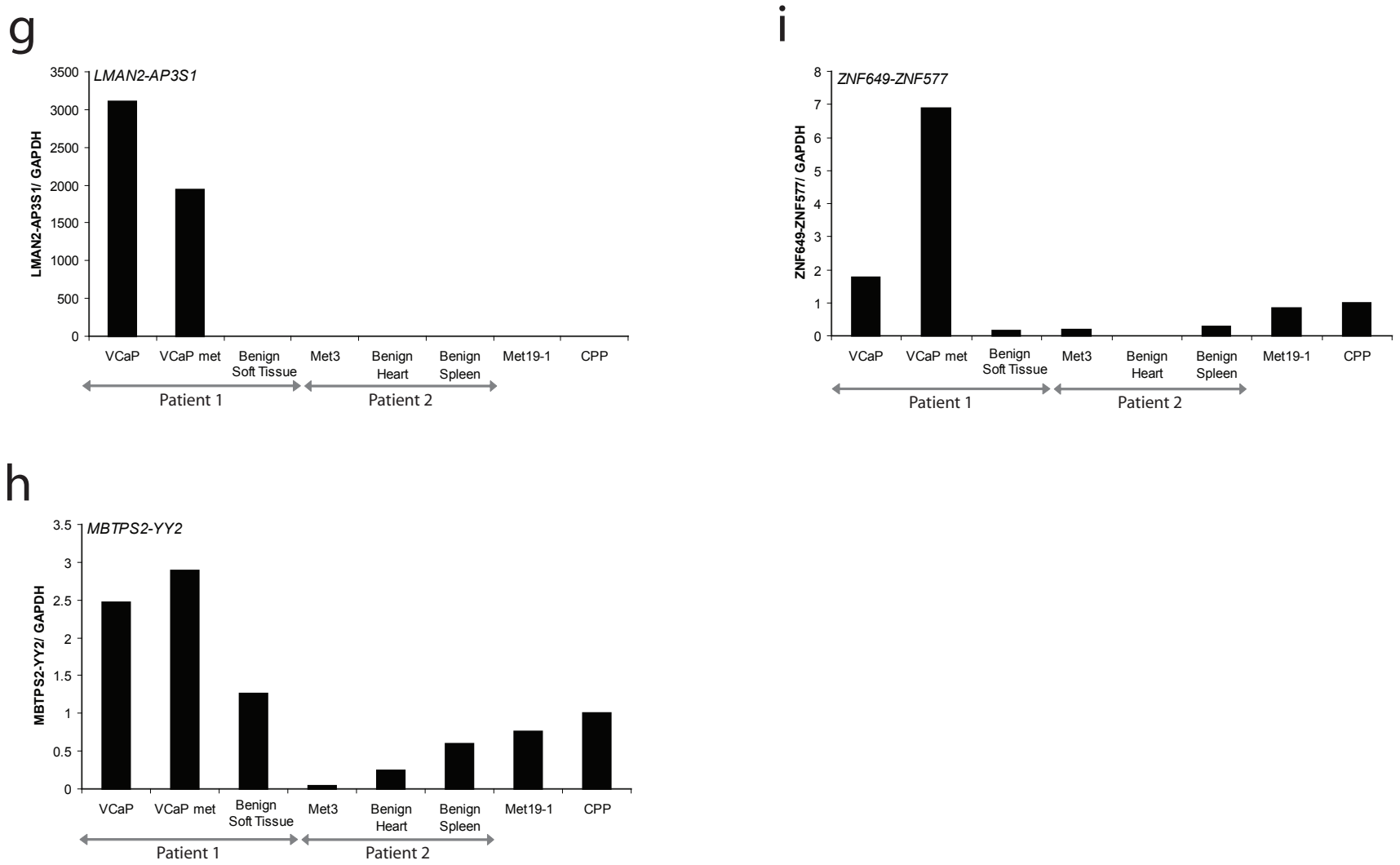

b
GPSN2-STRN4
GPSN2-STRN4/ GAPDH
140
120
100
80
60
40
20
0
VCaP
VCaP Met
Benign Soft Tissue
Met3
Benign Heart
Benign Spleen
Met19-1
CPP
Patient 1
Patient 2


c
USP10-ZDHHC7
USP10/ZDHHC7/ GAPDH
10000
9000
8000
7000
6000
5000
4000
3000
2000
1000
0
VCaP
VCaP Met
Benign Soft Tissue
Met3
Benign Heart
Benign Spleen
Met19-1
CPP
Patient 1
Patient 2

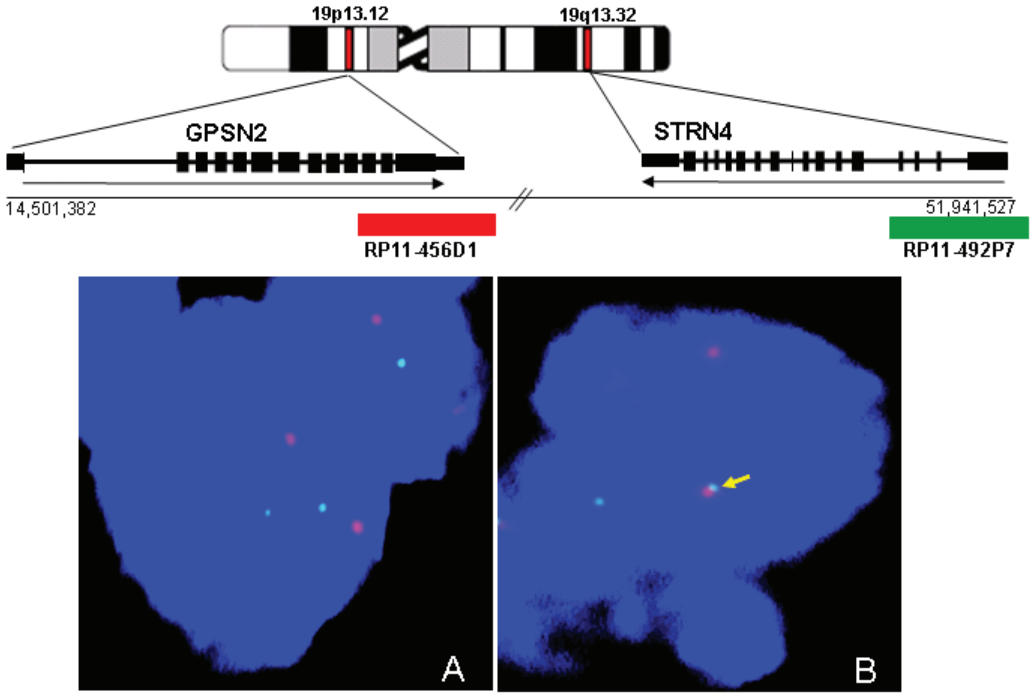

d
RC3H2-RGS3
RC3H2-RGS3/ GAPDH
300
250
200
150
100
50
0
VCaP
VCaP Met
Benign Soft Tissue
Met3
Benign Heart
Benign Spleen
Met19-1
CPP
Patient 1
Patient 2


e
HJURP-EIF4E2
HJURP-EIF4E2/ GAPDH
600
500
400
300
200
100
0
VCaP
VCaP Met
Benign Soft Tissue
Met3
Benign Heart
Benign Spleen
Met19-1
CPP
Patient 1
Patient 2

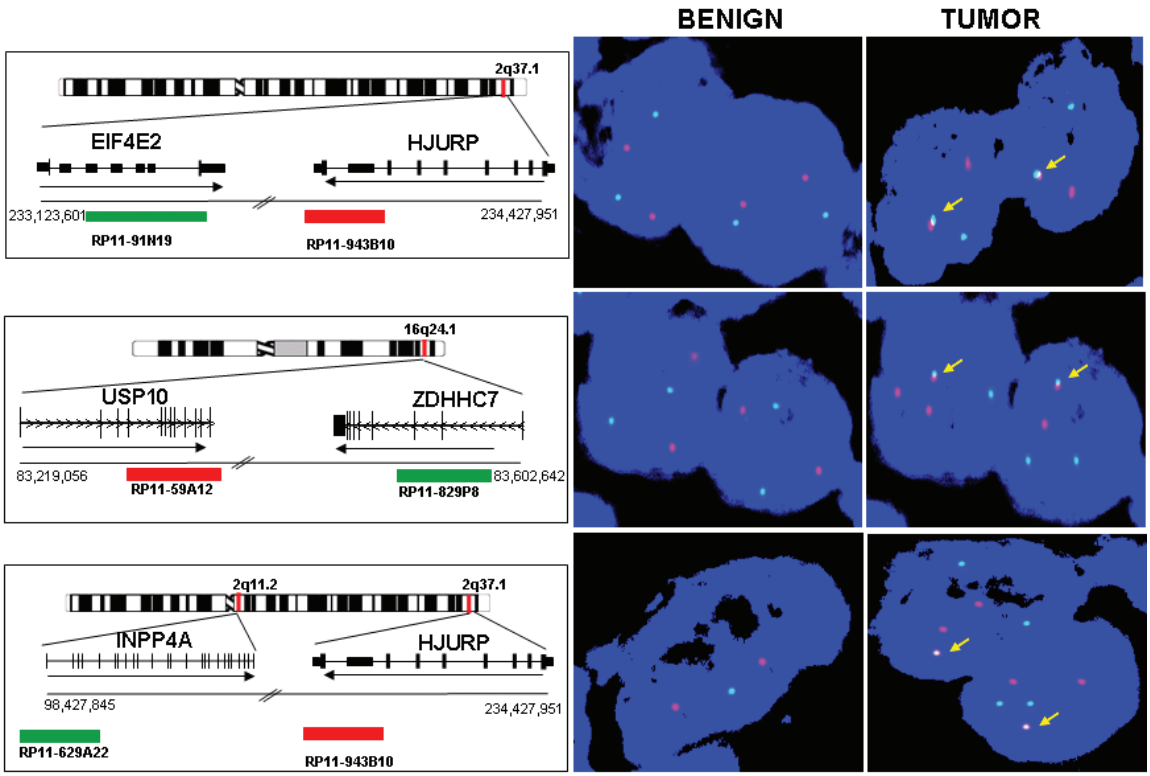

f
INPP4A-HJURP
INPP4A-HJURP/ GAPDH
600
500
400
300
200
100
0
VCaP
VCaP met
Benign Soft Tissue
Met3
Benign Heart
Benign Spleen
Met19-1
CPP
Patient 1
Patient 2



Supplemental Figure 11: qRT-PCR based confirmation of fusion transcript expression restricted to prostate cancer samples and absent in somatic tissues from the same patient. Five fusion genes, *TMPRSS2-ERG* (a), *GPSN2-STRN4* (b), *USP10-ZDHHC7* (c), *RC3H2-RGS3* (d), *HJURP-EIF4E2* (e), *INPP4A-HJURP* (f), *LMAN2-AP3S1* (g), *MBTPS2-YY2* (h), and *ZNF649-ZNF577* (i) were tested in two patients, Patient 1 and Patient 2.

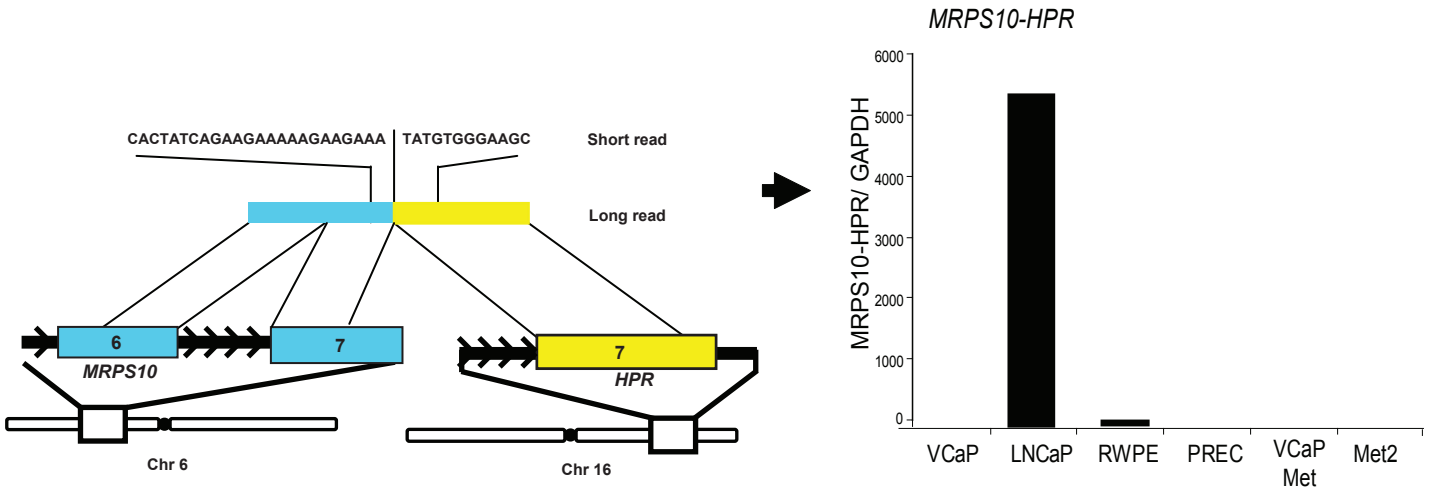


Supplemental Figure 12: FISH analysis of the chromosomal rearrangements involving *STRN4-GPSN2* gene fusion in tumor sample MET3. Top panel show the genomic organization of the *GSPN2* and *STRN4* genes located on chromosome 19. The red and green horizontal bars indicate the relative position of the BAC clones. Normal signal patterns were observed in benign sample (a) whereas a co-localizing red and green (yellow arrow) signal indicates a gene fusion in tumor sample only (b).

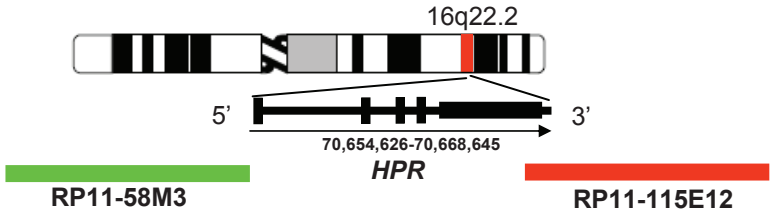


Supplemental Figure 13: FISH analysis of the chromosomal rearrangements involving *EIF4E2-HJURP*, *USP10-ZDHHC7*, and *INPP4A-HJURP* gene fusions in tumor and paired normal tissues from VCaP-Met. Schematic diagrams on the left panel show the genomic organization of the genes on their respective chromosomes. The green and red horizontal bars indicate the relative position of the BAC clones. The black arrows indicate the orientation of the genes in the chromosome. The probes were tested on both tumor and paired normal tissues. Individual red and green signals indicate the signal on the normal chromosomes. The co-localizing red and green (yellow arrows) signals indicate fusion, in tumor samples only.

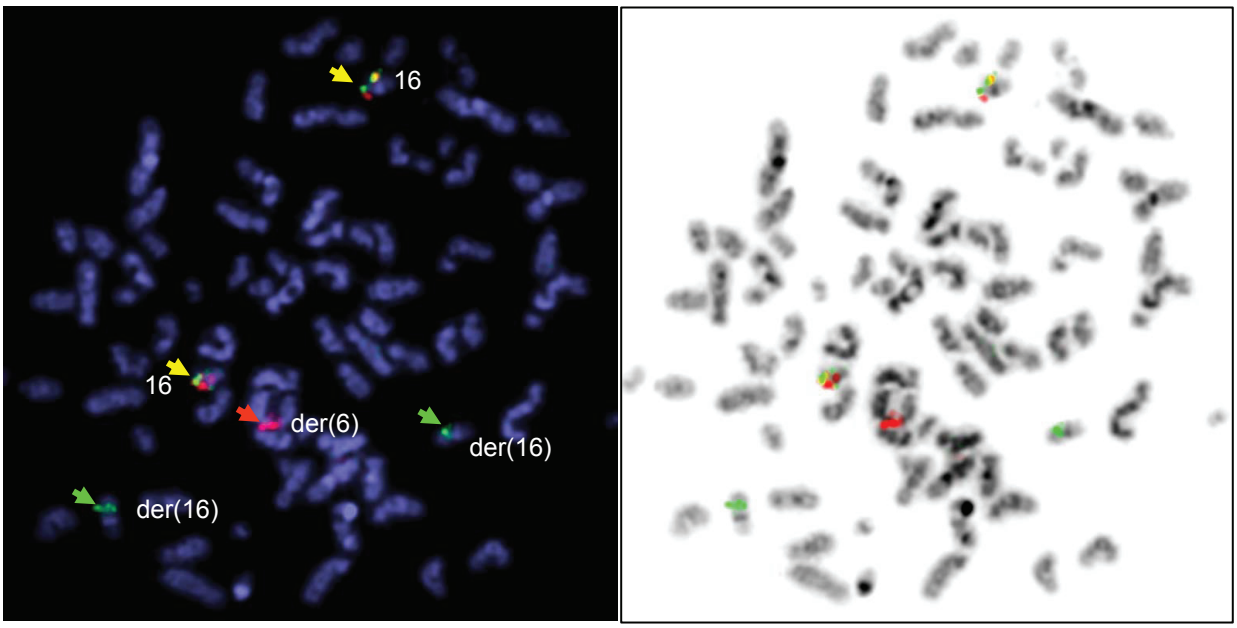
a


CACTATCAGAAGAAAAAGAAGAAA
TATGTGGGAAGC
Short read
Long read
6
7
MRPS10
7
HPR
Chr 6
Chr 16
MRPS10-HPR
MRPS10-HPR/ GAPDH
6000
5000
4000
3000
2000
1000
0
VCaP
LNCaP
RWPE
PREC
VCaP Met
Met2

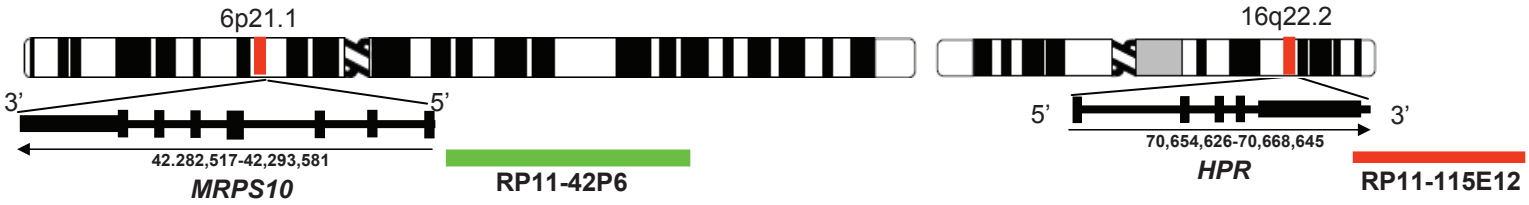
b


16q22.2
5'
3'
70,654,626-70,668,645
HPR
RP11-58M3
RP11-115E12

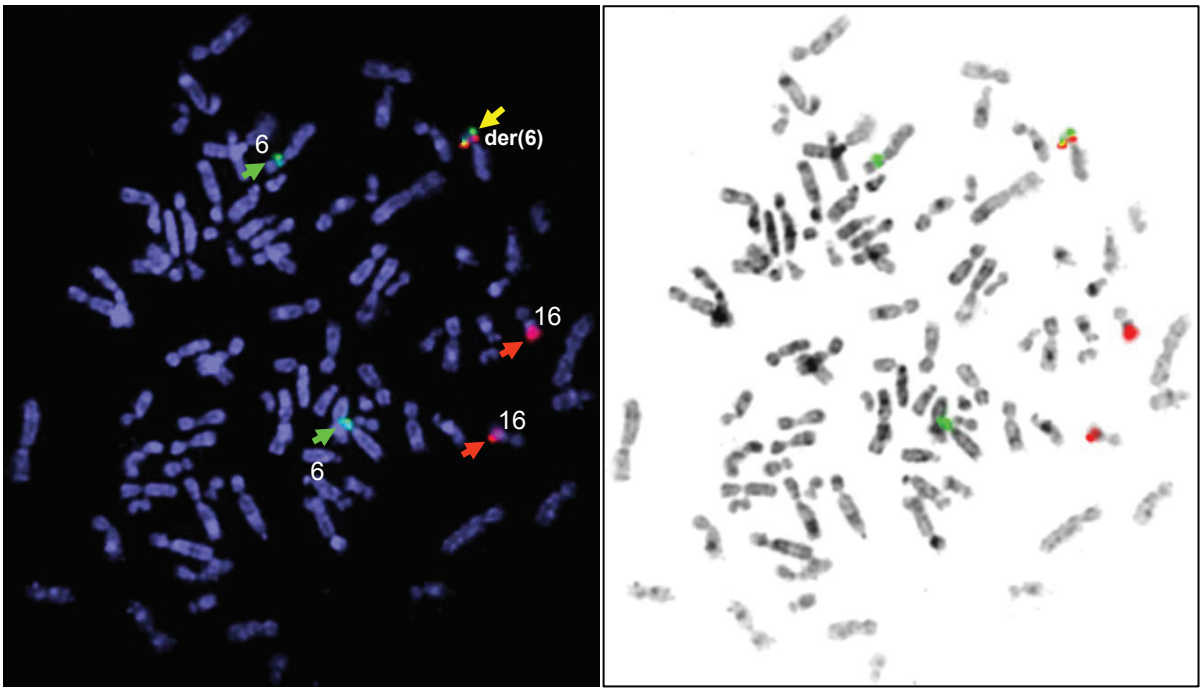
c


16
16
der(6)
der(16)
der(16)

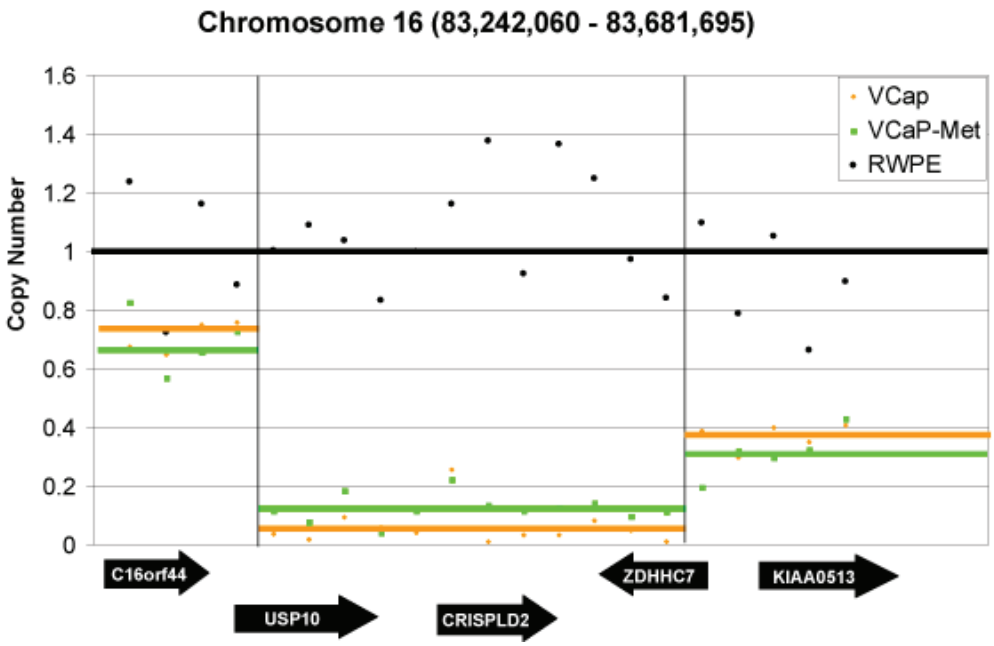
d



e



Supplemental Figure 14: FISH analysis of the chromosomal rearrangements involving *MRPS10* and *HPR*. **a,** Schematic of the MRPS10-HPR fusion. The exons 6-7 of MRPS10 (blue) located on chromosome 6 are fused with exon 7 of HPR (yellow), on chromosome 16. **b,** Schematic diagram showing the genomic organization of the *HPR* gene locus. The horizontal green and red bars indicate the approximate location of the BAC clones from the 5' and 3' end of the gene, respectively. **c,** FISH image from LNCaP cells show two copies of normal chromosome 16 (yellow arrows), two copies of derivative chromosome 16 [der(16)] (green arrows), and single red signal on derivative chromosome 6 [der(6)] confirming a rearrangement in the *HPR* gene. **d,** Schematic diagram showing the genomic organization of the *MRPS10* and *HPR* gene locus. The horizontal green and red bars indicate the approximate location of the BAC clones from the 5' and 3' end of *MRPS10* and *HPR* genes, respectively. **e,** FISH image from LNCaP cells show hybridization of *MRPS10* probe to two copies of chromosome 6 (green arrows), and red arrows indicate the hybridization of *HPR* probe to two copies of normal chromosome 16. A single co-localizing green and red signal (yellow arrow) on der(6) confirms the fusion of *MRPS10* with *HPR*.



Supplemental Figure 15: Plot of genomic aberrations on chromosome 16 located near the *USP10-ZDHHC7* fusion, as seen by array CGH. A deletion involving the two genes is observed in VCaP (orange) and the VCaP parental tissue (VCaP-Met) (green), but not in normal prostate cell line, RWPE (black).

SUPPLEMENTARY REFERENCES:

1 Mitelman F, J. B. a. M. F. E., (Cancer Genome Anatomy Project, 2008).

2 Greenman, C. *et al.*, Patterns of somatic mutation in human cancer genomes. *Nature* **446** (7132), 153 (2007).

3 Weir, B. A. *et al.*, Characterizing the cancer genome in lung adenocarcinoma. *Nature* **450** (7171), 893 (2007).

4 Wood, L. D. *et al.*, The genomic landscapes of human breast and colorectal cancers. *Science* **318** (5853), 1108 (2007).

5 Barber, T. D., B. Vogelstein, K. W. Kinzler & V. E. Velculescu, Somatic mutations of EGFR in colorectal cancers and glioblastomas. *The New England journal of medicine* **351** (27), 2883 (2004).

6 Stephens, P. *et al.*, A screen of the complete protein kinase gene family identifies diverse patterns of somatic mutations in human breast cancer. *Nature genetics* **37** (6), 590 (2005).

7 Stephens, P. *et al.*, Lung cancer: intragenic ERBB2 kinase mutations in tumours. *Nature* **431** (7008), 525 (2004).

8 Campbell, P. J. *et al.*, Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. *Nature genetics* **40** (6), 722 (2008).

9 Parsons, D. W. *et al.*, An integrated genomic analysis of human glioblastoma multiforme. *Science (New York, N.Y* **321** (5897), 1807 (2008).

10 Cheung, V. G. *et al.*, Integration of cytogenetic landmarks into the draft sequence of the human genome. *Nature* **409** (6822), 953 (2001).

11 Strausberg, R. L., K. H. Buetow, M. R. Emmert-Buck & R. D. Klausner, The cancer genome anatomy project: building an annotated gene index. *Trends Genet* **16** (3), 103 (2000).

12 Soda, M. *et al.*, Identification of the transforming EML4-ALK fusion gene in non-small-cell lung cancer. *Nature* **448** (7153), 561 (2007).

13 Tomlins, S. A. *et al.*, Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer. *Science* **310** (5748), 644 (2005).

14 Rikova, K. *et al.*, Global Survey of Phosphotyrosine Signaling Identifies Oncogenic Kinases in Lung Cancer. *Cell* **131**, 14 (2007).

15 Kato, T. *et al.*, Activation of Holliday junction recognizing protein involved in the chromosomal stability and immortality of cancer cells. *Cancer research* **67** (18), 8544 (2007).

16 Bashir, A., S. Volik, C. Collins, V. Bafna & B. J. Raphael, Evaluation of paired-end sequencing strategies for detection of genome rearrangements in cancer. *PLoS computational biology* **4** (4), e1000051 (2008).

17 Korenchuk, S. *et al.*, VCaP, a cell-based model system of human prostate cancer. *In vivo (Athens, Greece)* **15** (2), 163 (2001).

18 Rubin, M. A. *et al.*, Rapid ("warm") autopsy study for procurement of metastatic prostate cancer. *Clin Cancer Res* **6** (3), 1038 (2000).

19 Karolchik, D. *et al.*, The UCSC Table Browser data retrieval tool. *Nucleic Acids Res* **32** (Database issue), D493 (2004).

20 Abouelhoda, M. I., S. Kurtz & E. Ohlebusch, Replacing suffix trees with enhanced suffix arrays. *Journal of Discrete Algorithms* **2** (1), 53 (2004).
21 Kent, W. J., BLAT--the BLAST-like alignment tool. *Genome research* **12** (4), 656 (2002).
22 Communi, D., N. Suarez-Huerta, D. Dussossoy, P. Savi & J. M. Boeynaems, Cotranscription and intergenic splicing of human P2Y11 and SSF1 genes. *The Journal of biological chemistry* **276** (19), 16561 (2001).
23 Tomlins, S. A. *et al.*, Distinct classes of chromosomal rearrangements create oncogenic ETS gene fusions in prostate cancer. *Nature* **448** (7153), 595 (2007).
24 Vandesompele, J. *et al.*, Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes. *Genome biology* **3** (7), 34 (2002).