LETTERS

Transcriptome sequencing to detect gene fusions in cancer

Christopher A. Maher[1,3]*, Chandan Kumar-Sinha[1,3]*, Xuhong Cao[1,2], Shanker Kalyana-Sundaram[1,3], Bo Han[1,3], Xiaojun Jing[1,3], Lee Sam[1,3], Terrence Barrette[1,3], Nallasivam Palanisamy[1,3] & Arul M. Chinnaiyan[1,2,3,4,5]

Recurrent gene fusions, typically associated with haematological malignancies and rare bone and soft-tissue tumours[1], have recently been described in common solid tumours[2–9]. Here we use an integrative analysis of high-throughput long- and short-read transcriptome sequencing of cancer cells to discover novel gene fusions. As a proof of concept, we successfully used integrative transcriptome sequencing to 're-discover' the *BCR–ABL1* (ref. 10) gene fusion in a chronic myelogenous leukaemia cell line and the *TMPRSS2–ERG*[2,3] gene fusion in a prostate cancer cell line and tissues. Additionally, we nominated, and experimentally validated, novel gene fusions resulting in chimaeric transcripts in cancer cell lines and tumours. Taken together, this study establishes a robust pipeline for the discovery of novel gene chimaeras using high-throughput sequencing, opening up an important class of cancer-related mutations for comprehensive characterization.

Characterization of specific genomic aberrations in cancers has led to the identification of several successful therapeutic targets, such as BCR–ABL1, PDGFR, ERBB2 and EGFR[11–14]. Therefore a major goal in cancer research is to identify causal genetic aberrations. Gene fusions resulting from chromosomal rearrangements in cancer are believed to define the most prevalent category of 'cancer genes'[15]. Typically, an aberrant juxtaposition of two genes may encode a fusion protein (for example, *BCR–ABL1*), or the regulatory elements of one gene may drive the aberrant expression of an oncogene (for example, *TMPRSS2–ERG*). Although gene fusions have been widely described in rare haematological malignancies and sarcomas[1], the recent discovery of recurrent gene fusions in prostate[2,4] and lung cancers[5–9] points to their role in common solid tumours as well. Considering their prevalence and common characteristics across cancer types, gene fusions may be regarded as a distinct class of 'mutations', with a causal role in carcinogenesis. In addition, being strictly confined to cancer cells, they represent ideal diagnostic markers and rational therapeutic targets.

As a proof of concept, we performed whole transcriptome sequencing of the chronic myelogenous leukaemia cell line K562, which harbours the classical gene fusion, *BCR–ABL1* (ref. 16). Using the Illumina Genome Analyser, we generated 66.9 million reads of 36 nucleotides in length and screened them for the presence of reads showing partial alignment to exon boundaries from two different genes. Although this approach was able to detect *BCR–ABL1*, it was one among a set of 111 other chimaeras (with at least two reads). Thus, in a *de novo* discovery mode, it would be difficult to pinpoint the *BCR–ABL1* fusion in the background of the other putative chimaeras. However, when we used the known fusion junction of *BCR–ABL1* (GenBank number M30829) as the reference sequence, we detected 19 chimaeric reads (Supplementary Fig. 1). Thus, we considered an integrative approach for chimaera detection, using short-read sequencing technology for obtaining deep sequence data and long-read technology (Roche 454 sequencing platform) to provide reference sequences for mapping candidate fusion genes.

An important concern in transcriptome sequencing was whether we could detect chimaeric transcripts in the background of highly abundant housekeeping genes (that is, would complementary DNA (cDNA) normalization be required). To address this, we compared sequences from normalized and non-normalized cDNA libraries of the prostate cancer cell line VCaP, which harbours the gene fusion *TMPRSS2–ERG* (Supplementary Table 1). Overall, the normalized library showed an approximately 3.6-fold reduction in the total number of chimaeras nominated. Furthermore, although we expected the normalized library would enrich for the *TMPRSS2–ERG* gene fusion, it failed to reveal any *TMPRSS2–ERG* chimaeras, which suggested that we would not benefit from normalization in our analyses.

To assess the feasibility of using massively parallel transcriptome sequencing to identify novel gene fusions, we generated non-normalized cDNA libraries from the prostate cancer cell lines VCaP and LNCaP, and a benign immortalized prostate cell line RWPE. As a first step, using the Roche 454 platform, we generated 551,912 VCaP, 244,984 LNCaP and 826,624 RWPE transcriptome sequence reads, averaging 229.4 nucleotides. These were categorized as completely aligning, partly aligning or non-mapping to the human reference database (Fig. 1a). Sequence reads that showed partial alignments to two genes (Supplementary Methods) were nominated as first-pass candidate chimaeras. This yielded 428 VCaP, 247 LNCaP and 83 RWPE candidates. Admittedly, many of these chimaeric sequences could be a result of *trans*-splicing[17] or co-transcription of adjacent genes coupled with intergenic splicing[18], or simply an artefact of the sequencing protocol. Surprisingly, among the 428 VCaP candidates, only one read spanned the *TMPRSS2–ERG* fusion junction using the long-read sequencing platform (Supplementary Table 2).

Next, using the Illumina Genome Analyser we obtained over 50 million short-transcriptome sequence reads from VCaP, LNCaP and RWPE cDNA libraries (Supplementary Table 3). Focusing initially on VCaP cells, we identified the *TMPRSS2–ERG* fusion as one among 57 candidates, many of them likely false positives. To overcome the problem of false positives, lack of depth in long reads, and difficulty in mapping partly aligning short reads, we considered integrating the long- and short-read sequence data. Following this strategy, we found the single long-read chimaeric sequence spanning *TMPRSS2–ERG* junction from VCaP transcriptome sequence, buttressed by 21 short reads (Fig. 1b), was one of only eight chimaeras nominated, overall. Thus, using the integrative approach, the total number of false candidates was reduced and the proportion of

[1]Michigan Center for Translational Pathology, [2]Howard Hughes Medical Institute, [3]Department of Pathology, [4]Department of Urology, [5]Comprehensive Cancer Center, University of Michigan Medical School, Ann Arbor, Michigan 48109, USA.
*These authors contributed equally to this work.

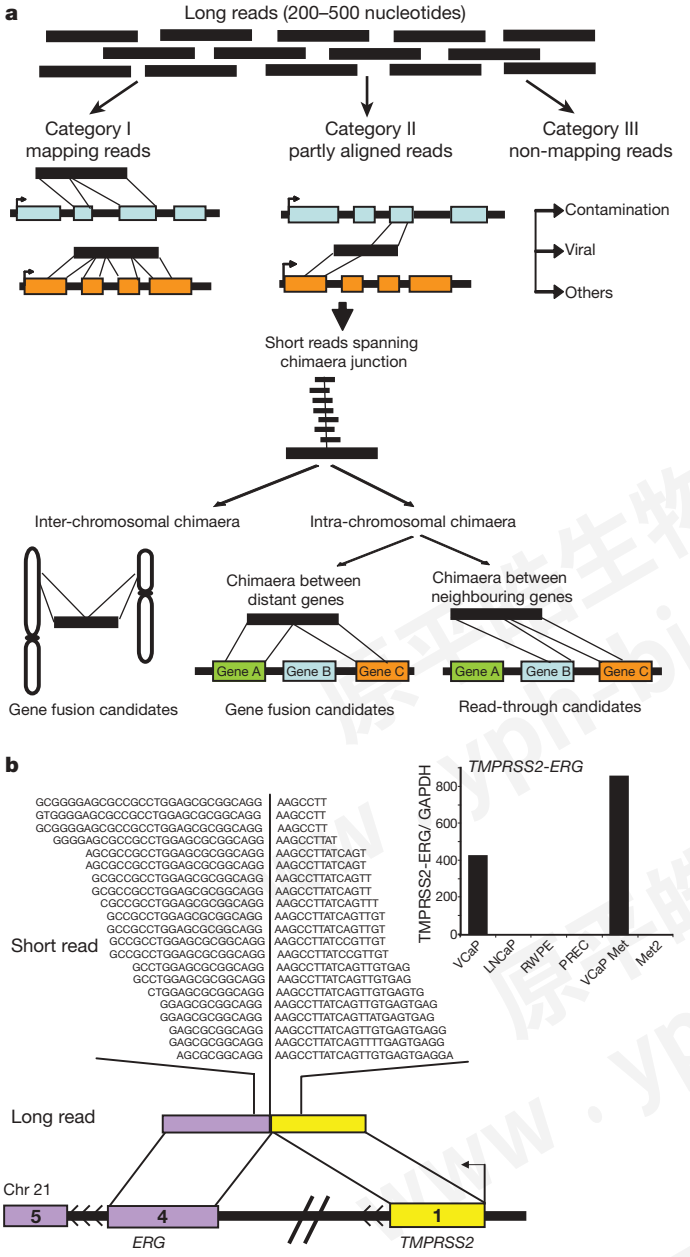

Figure 1 | Using massively parallel sequencing to discover chimaeric transcripts in cancer. a, Schema representing our use of transcriptome sequencing to identify chimaeric transcripts. 'Long-read' sequences compared with the reference database are classified as 'mapping', 'partly aligned' and 'non-mapping' reads. Partly aligning reads are considered putative chimaeras and are categorized as inter- or intra-chromosomal chimaeras. Integration with short-read sequence data are used for shortlisting candidate chimaeras and assessing the depth of coverage spanning the fusion junction. **b**, 'Re-discovery' of *TMPRSS2–ERG* fusion on chromosome (chr) 21. Short reads (Illumina) are overlaid on the corresponding long-read (454) represented by coloured bars. Sequences spanning the fusion junction are indicated by the partition in the short reads. Chromosomal context of the fusion genes is represented by coloured bars punctuated with black lines. Inset displays histogram of qRT–PCR validation of the *TMPRSS2–ERG* transcript.

experimentally validated candidates increased dramatically (Supplementary Fig. 2). Extending the integrative analysis to LNCaP and RWPE sequences provided a total of 15 chimaeric transcripts, of which ten could be experimentally confirmed (Supplementary Table 4). To ensure that the integration strategy filtered out only false positives and not valid chimaeras, we tested a panel of 16 long-read chimaera candidates that were eliminated upon integration and found that none of them confirmed a fusion transcript by quantitative real-time PCR (qRT–PCR; Supplementary Fig. 3).

To leverage the collective coverage provided by the two sequencing platforms systematically, and to prioritize the candidates, we formulated a scoring function obtained by multiplying the number of chimaeric reads derived from either method (Supplementary Table 4). Further, we categorized these chimaeras as intra-or inter-chromosomal, based on their location on the same or different chromosomes, respectively. The latter represent bona fide gene fusions as do intra-chromosomal chimaeras aligning to non-adjacent transcripts; intra-chromosomal chimaeras between neighbouring genes are classified as read-throughs. Remarkably, *TMPRSS2–ERG* was our top ranking gene fusion sequence, second only to a read-through chimaera *ZNF577–ZNF649*.

In addition to *TMPRSS2–ERG*, we identified several new gene fusions in VCaP. One such fusion was between exon 1 of *USP10*, with exon 3 of *ZDHHC7*, both genes located on chromosome 16, approximately 200 kilobases (kb) apart, in opposite orientation (Fig. 2a and Supplementary Discussion). Furthermore, two separate fusions involving the gene *HJURP* on chromosome 2 were identified. A fusion between exon 2 of *EIF4E2* with exon 8 of *HJURP* generated the fusion transcript *EIF4E2–HJURP* and a fusion between exon 9 of *HJURP* with exon 25 of *INPP4A* yielded *HJURP–INPP4A* (Fig. 2b and Supplementary Fig. 4).

Interestingly, based on whole transcriptome sequencing, the highest-ranked LNCaP gene fusion was between exon 11 of *MIPOL1* on chromosome 14 with the last exon of *DGKB* on chromosome 7; this was confirmed by qRT–PCR and fluorescence *in situ* hybridization (FISH) (Fig. 3 and Supplementary Fig. 5). We recently demonstrated that overexpression of *ETV1*, a member of the oncogenic erythroblast transformation-specific (ETS) transcription factor family, has a role in tumour progression in LNCaP cells[3]. The mechanism of *ETV1* overexpression was attributed to a cryptic insertion of approximately 280 kb encompassing the *ETV1* gene into an intronic region of *MIPOL1*. Thus, although our previous study suggested that *ETV1* was rearranged without evidence of an *ETV1* fusion transcript, here we show the generation of a surrogate fusion of *MIPOL1* to *DGKB*, which appears to be indicative of an *ETV1* chromosomal aberration.

In addition to gene fusions, we also identified several transcript chimaeras between neighbouring genes, referred to as read-through events. Overall, the read-through events appear to be more broadly expressed across both malignant and benign samples whereas the gene fusions were cancer-cell specific (Supplementary Fig. 6 and Supplementary Discussion).

Next, we attempted to extend this methodology to tumour samples that represent the malignant cells often admixed with benign epithelia, stromal, lymphocytic and vascular cells. To accomplish this we conducted transcriptome sequencing of two *TMPRSS2–ERG* gene-fusion-positive metastatic prostate cancer tissues, VCaP-Met (from which the VCaP cell line is derived) and Met 3, and one *ERG* negative metastatic prostate tissue, Met 4. Interestingly, in addition to the *TMPRSS2–ERG* fusion sequences detected in both VCaP-Met and Met 3 tissues, three novel gene fusions were identified (Supplementary Fig. 7a). One chimaeric transcript from Met 3 involves exon 9 of *STRN4* with exon 2 of *GPSN2* (Supplementary Fig. 7b). *GPSN2* belongs to the steroid 5-α-reductase family, the enzyme that converts testosterone to dihydrotestosterone, the key hormone that mediates androgen response in prostate tissues. Dihydrotestosterone is known to be highly expressed in prostate cancer, and is a therapeutic target[19]. Dihydrotestosterone, like its synthetic analogue R1881, has been shown to induce *TMPRSS2–ERG* expression as well as prostate-specific antigen (PSA)[2]. Additionally, we found exon 10 of *RC3H2* fused to exon 20 of *RGS3* in the VCaP-Met (and VCaP cells) (Supplementary Fig. 7c). Another novel gene fusion was between exon 1 of *LMAN2* and exon 2 of *AP3S1* (Supplementary Fig. 7d).

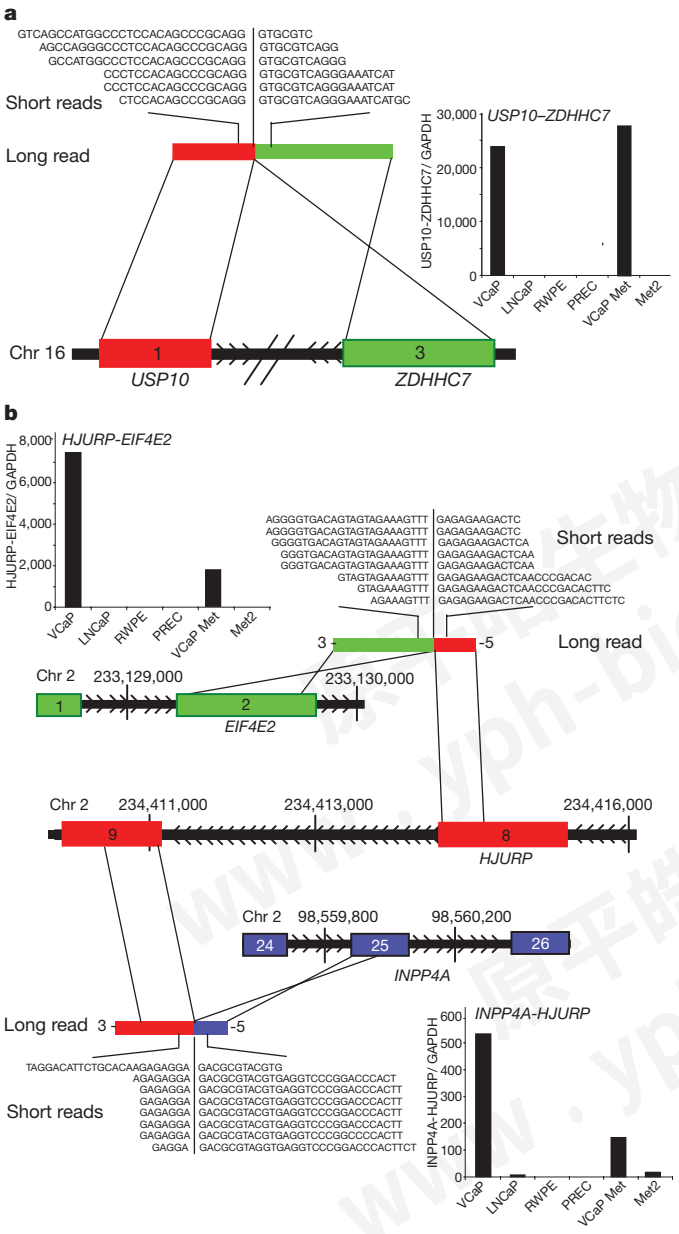

Figure 2 | Representative gene fusions characterized in the prostate cancer cell line VCaP. a, Schematic of *USP10–ZDHHC7* fusion on chromosome 16. Exon 1 of *USP10* (red) is fused with exon 3 of *ZDHHC7* (green), located on the same chromosome in opposite orientation. Inset displays histogram of qRT–PCR validation of *USP10–ZDHHC7* transcript. **b**, Schematic of a complex intra-chromosomal rearrangement leading to two gene fusions involving *HJURP* on chromosome 2. Exon 8 of *HJURP* (red) is fused with exon 2 of *EIF4E2* (green) to form *HJURP–EIF4E2*. Exon 25 of *INPP4A* (blue) is fused with exon 9 of *HJURP* (red) to form *INPP4A–HJURP*. Insets display histograms of qRT–PCR validation of *HJURP–EIF4E2* and *INPP4A–HJURP* transcripts.

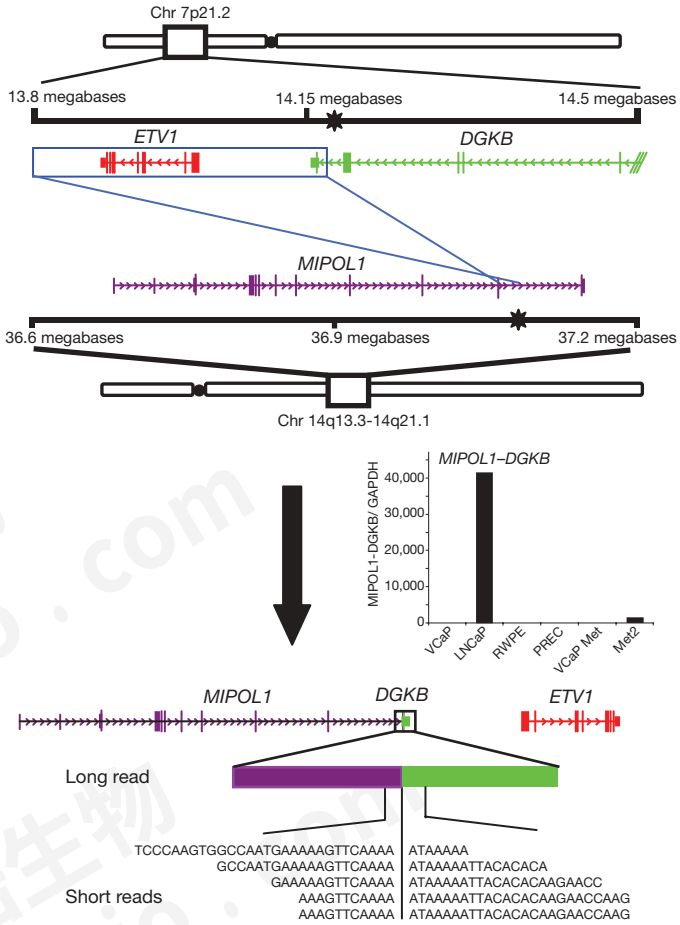

Figure 3 | Schematic of *MIPOL1–DGKB* gene fusion in the prostate cancer cell line LNCaP. *MIPOL1–DGKB* is an inter-chromosomal gene fusion accompanying the cryptic insertion of *ETV1* locus (red) on chromosome 7 into the *MIPOL1* (purple) intron on chromosome 14. Previously determined genomic breakpoints (black stars) are shown in *DGKB* and *MIPOL1*. An insertion event results in the inversion of the 3′ end of *DGKB* and *ETV1* into the *MIPOL1* intron between exons 10 and 11. Inset displays histogram of qRT–PCR validation of the *MIPOL1–DGKB* transcript.

Interestingly, one read-through chimaera, *SLC45A3–ELK4*, between the fourth exon of *SLC45A3* with exon 2 of *ELK4*, a member of the ETS transcription factor family, was identified in metastatic prostate cancer, Met 4, and the LNCaP cell line, suggesting recurrence (Fig. 4a, upper panel). Taqman qRT–PCR assay for this fusion performed in a panel of cell lines revealed a high level of expression in LNCaP cells and much lower levels in other prostate cancer cell lines including 22Rv1, VCaP and MDA-PCA-2B. Benign prostate epithelial cells, PREC and RWPE, and non-prostate cell lines including breast, melanoma, lung, chronic myeloid leukaemia (CML) and pancreatic cancer cell lines were negative for this fusion (Fig. 4a, middle panel).

SLC45A3 has been reported to be fused to *ETV1* in a prostate cancer sample[3]; notably, it is a prostate-specific, androgen-responsive gene. Interestingly, the fusion transcript *SLC45A3–ELK4* was also found to be induced by the synthetic androgen R1881 (Fig. 4a, middle panel, inset). Further, we interrogated a panel of prostate tissues for this fusion, and found it expressed in seven out of 20 metastatic prostate cancer tissues examined (Fig. 4a, lower panel). Interestingly, six of those seven positive cases have been identified as negative for ETS genes *ERG*, *ETV1*, *ETV4* and *ETV5* in our previous work, based on a FISH screen[20]. One *TMPRSS2–ETV1*-positive metastatic prostate cancer sample was also found to be positive for *SLC45A3–ELK4* (similar to LNCaP, which is also *ETV1* positive[3]). Unlike the previous ETS gene fusions identified, *SLC45A3–ELK4* is a read-through event between adjacent genes and does not harbour detectable alterations at the DNA level by FISH (Supplementary Fig. 8), array comparative genomic hybridization (aCGH) (data not shown) or high-density single nucleotide polymorphism (SNP) arrays (Supplementary Fig. 9). As LNCaP and Met 4 harbour genomic aberrations of *ETV1*, and express high levels of the *SLC45A3–ELK4* chimaeric transcript, this suggests that *ETV1* and *ELK4* may cooperate to drive prostate carcinogenesis in those tumours. To our knowledge, *SLC45A3–ELK4* may represent the first description of a recurrent RNA chimaeric transcript specific to cancer that does not have a detectable DNA aberration. Overall, *SLC45A3–ELK4* appears to be the only recurrent chimaeric transcript identified in our transcriptome sequencing

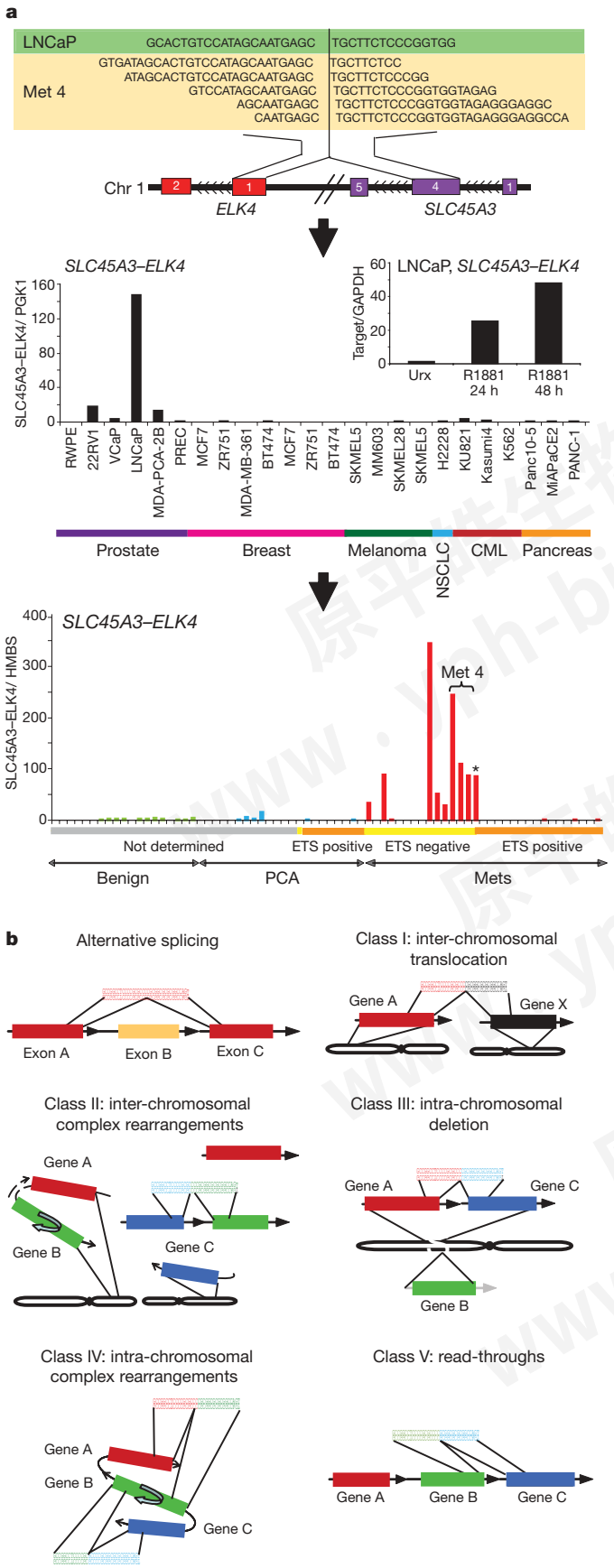

Figure 4 | Discovery of the recurrent *SLC45A3–ELK4* chimaera in prostate cancer and a general classification system for chimaeric transcripts in cancer. **a**, Upper panel, schematic of the *SLC45A3–ELK4* chimaera located on chromosome 1. Middle panel, qRT–PCR validation of *SLC45A3–ELK4* transcript in a panel of cell lines. Inset, histogram of qRT–PCR assessment of the *SLC45A3–ELK4* transcript in LNCaP cells treated with R1881. Lower panel, histogram of qRT–PCR validation in a panel of prostate tissues: benign adjacent prostate, localized prostate cancer (PCA) and metastatic prostate cancer (Mets). ETS family gene rearrangement status (by FISH) indicated by horizontal coloured bars below graph. Grey, not determined; yellow, ETS negative; orange, ETS positive. Horizontal bracket indicates three different metastatic tissues from the same patient (Met 4). Asterisk denotes an ETV1-positive sample. **b**, Chimaera classification schema (described in the text).

study, as other gene fusions tested in a panel of prostate cancer samples appeared to be restricted to the sample in which they were identified (at least in the limited number of samples we analysed) and thus may represent rare or private mutations (Supplementary Fig. 10).

Next, we tested if the novel gene fusions identified in this study represent acquired somatic mutations or simply germline variations. Based on qPCR (Supplementary Fig. 11) and FISH (Supplementary Figs 12 and 13) assessment of a representative set of fusion genes on patient-matched germline tissues, we found the chimaeras restricted to the cancer tissues. Further, we interrogated the 29 genes involved in our gene fusions in the Database of Genomic Variants (http://projects.tcag.ca/variation/) and found only eight of them with previously reported copy number variations (Supplementary Table 5). However, our matched aCGH data did not reveal any copy number variation in those genes (Supplementary Table 6), suggesting that our samples did not harbour copy number variations common to the human population.

Based on the gene fusions we have characterized (Supplementary Table 7), we propose a chimaera classification system (Fig. 4b). Inter-chromosomal translocation (Class I) involves fusion between two genes on different chromosomes (for example, *BCR–ABL1*). Inter-chromosomal complex rearrangements (Class II) occur when two genes from different chromosomes fuse together while a third gene follows along and becomes activated (*MIPOL1–DGKB*). Intra-chromosomal deletion (Class III) results when deletion of a genomic region fuses the flanking genes (*TMPRSS2–ERG*). Intra-chromosomal complex rearrangements (Class IV) involve a breakpoint in one gene fusing with multiple regions (*HJURP–EIF4E2* and *INPP4–HJURP*). Read-through chimaeras (Class V) include chimaeric transcripts between neighbouring genes (*ZNF649–ZNF577*).

Overall, transcriptome sequencing was found to be a powerful tool for detecting gene fusions, exemplified by our ability to detect multiple gene fusions in cancer cell lines and tissues. One important limitation is in cases where the proximal partner contributes only the regulatory sequence to the fusion and no transcript sequence (for example IgH–Myc in Burkitt's lymphoma). Although it has been known that gene fusion events can play a causative role in cancer, the current study has demonstrated that a particular cancer cell line or tissue can harbour multiple gene fusions, many of which are likely not recurrent. Although it is unclear whether these private gene fusions play a role in malignant transformation, they could potentially cooperate with the driver mutation/gene fusions. Like the cataloguing of point mutations associated with cancer[21–27], it will be important to catalogue and investigate the function of the multiple gene fusions present in a single cancer. The discovery of the chimaeric transcript *SLC45A3–ELK4* underscores that a refinement of next-generation sequencing technologies and attendant analytical tools may well unravel the full scope of these 'dangerous liaisons' in carcinogenesis.

METHODS SUMMARY

Long-read sequencing was conducted using 454 FLX Sequencing, whereas short-read sequencing was performed on the Illumina Genome Analyzer. qPCR for fusion candidates was performed using indicated oligonucleotide primers (Supplementary Table 8). Interphase FISH was performed in cell lines and tissues using bacterial artificial chromosome probes (Supplementary Figs 4b–e, 5b, d, f, 8, 12, 13 and 14c, e). Oligonucleotide aCGH was performed using Agilent arrays, and copy number analysis was conducted in CGH Analytics. Affymetrix Genome-wide Human SNP Array 6.0 was processed using the Affymetrix

Genotyping Console. Prostate tissues were obtained from the radical prostatectomy series at the University of Michigan and from the Rapid Autopsy Program, University of Michigan Specialized Program of Research Excellence (SPORE) in prostate cancer.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 21 July; accepted 10 November 2008.
Published online 11 January 2009.

1. Mitelman, F., Johansson, B. & Mertens, F. Fusion genes and rearranged genes as a linear function of chromosome aberrations in cancer. *Nature Genet.* **36,** 331–334 (2004).
2. Tomlins, S. A. *et al.* Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer. *Science* **310,** 644–648 (2005).
3. Tomlins, S. A. *et al.* Distinct classes of chromosomal rearrangements create oncogenic ETS gene fusions in prostate cancer. *Nature* **448,** 595–599 (2007).
4. Kumar-Sinha, C., Tomlins, S. A. & Chinnaiyan, A. M. Recurrent gene fusions in prostate cancer. *Nature Rev. Cancer* **8,** 497–511 (2008).
5. Choi, Y. L. *et al.* Identification of novel isoforms of the EML4-ALK transforming gene in non-small cell lung cancer. *Cancer Res.* **68,** 4971–4976 (2008).
6. Koivunen, J. P. *et al.* EML4-ALK fusion gene and efficacy of an ALK kinase inhibitor in lung cancer. *Clin. Cancer Res.* **14,** 4275–4283 (2008).
7. Perner, S. *et al.* EML4-ALK fusion lung cancer: a rare acquired event. *Neoplasia* **10,** 298–302 (2008).
8. Rikova, K. *et al.* Global survey of phosphotyrosine signaling identifies oncogenic kinases in lung cancer. *Cell* **131,** 1190–1203 (2007).
9. Soda, M. *et al.* Identification of the transforming EML4-ALK fusion gene in non-small-cell lung cancer. *Nature* **448,** 561–566 (2007).
10. Rowley, J. D. Chromosome translocations: dangerous liaisons revisited. *Nature Rev. Cancer* **1,** 245–250 (2001).
11. Lynch, T. J. *et al.* Activating mutations in the epidermal growth factor receptor underlying responsiveness of non-small-cell lung cancer to gefitinib. *N. Engl. J. Med.* **350,** 2129–239 (2004).
12. Slamon, D. J. *et al.* Use of chemotherapy plus a monoclonal antibody against HER2 for metastatic breast cancer that overexpresses HER2. *N. Engl. J. Med.* **344,** 783–792 (2001).
13. Demetri, G. D. *et al.* Efficacy and safety of imatinib mesylate in advanced gastrointestinal stromal tumors. *N. Engl. J. Med.* **347,** 472–480 (2002).
14. Druker, B. J. *et al.* Five-year follow-up of patients receiving imatinib for chronic myeloid leukemia. *N. Engl. J. Med.* **355,** 2408–2417 (2006).
15. Futreal, P. A. *et al.* A census of human cancer genes. *Nature Rev. Cancer* **4,** 177–183 (2004).
16. Shtivelman, E., Lifshitz, B., Gale, R. P. & Canaani, E. Fused transcript of *abl* and *bcr* genes in chronic myelogenous leukaemia. *Nature* **315,** 550–554 (1985).
17. Takahara, T., Tasic, B., Maniatis, T., Akanuma, H. & Yanagisawa, S. Delay in synthesis of the 3′ splice site promotes trans-splicing of the preceding 5′ splice site. *Mol. Cell* **18,** 245–251 (2005).
18. Communi, D., Suarez-Huerta, N., Dussossoy, D., Savi, P. & Boeynaems, J. M. Cotranscription and intergenic splicing of human P2Y11 and SSF1 genes. *J. Biol. Chem.* **276,** 16561–16566 (2001).
19. Gleave, M. *et al.* The effects of the dual 5α-reductase inhibitor dutasteride on localized prostate cancer – results from a 4-month pre-radical prostatectomy study. *Prostate* **66,** 1674–1685 (2006).
20. Han, B. *et al.* 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 (2008).
21. Barber, T. D., Vogelstein, B., Kinzler, K. W. & Velculescu, V. E. Somatic mutations of EGFR in colorectal cancers and glioblastomas. *N. Engl. J. Med.* **351,** 2883 (2004).
22. Cheung, V. G. *et al.* Integration of cytogenetic landmarks into the draft sequence of the human genome. *Nature* **409,** 953–958 (2001).
23. Greenman, C. *et al.* Patterns of somatic mutation in human cancer genomes. *Nature* **446,** 153–158 (2007).
24. Stephens, P. *et al.* A screen of the complete protein kinase gene family identifies diverse patterns of somatic mutations in human breast cancer. *Nature Genet.* **37,** 590–592 (2005).
25. Strausberg, R. L., Buetow, K. H., Emmert-Buck, M. R. & Klausner, R. D. The cancer genome anatomy project: building an annotated gene index. *Trends Genet.* **16,** 103–106 (2000).
26. Weir, B. A. *et al.* Characterizing the cancer genome in lung adenocarcinoma. *Nature* **450,** 893–898 (2007).
27. Wood, L. D. *et al.* The genomic landscapes of human breast and colorectal cancers. *Science* **318,** 1108–1113 (2007).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank Illumina and 454 for technical support, R. Mehra and J. Siddiqui for providing tissue samples, Y. Gong, S. Shankar, X. Wang and A. Menon for technical assistance, J. Yu for help with the Illumina Genome Analyzer, and R. J. Lonigro for discussions. C.A.M. was supported by a National Institutes of Health Ruth L. Kirschstein post-doctoral training grant, and currently derives support from the American Association of Cancer Research Amgen Fellowship in Clinical/Translational Research, the Canary Foundation and American Cancer Society Early Detection Postdoctoral Fellowship. This work was supported in part by the National Institutes of Health (to A.M.C.), the Department of Defense (to A.M.C.), the Early Detection Research Network (to A.M.C.), and NCIBI (grant number U54 DA 021519).

Author Contributions C.A.M., C.K.-S. and A.M.C. wrote the manuscript. C.K.-S., X.C., X.J., B.H. and N.P. performed the sequencing and biochemical experiments. C.A.M., S.K.-S., L.S. and T.B. performed bioinformatics analysis.

Author Information Sequences of the gene fusion chimaeras are deposited in GenBank under accession numbers FJ423742–FJ423755. Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to A.M.C. (arul@umich.edu).

METHODS

Samples and cell lines. The benign immortalized prostate cell line RWPE and the prostate cancer cell line LNCaP were 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[28], and was provided by K. Pienta.

The androgen stimulation experiment was performed 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[29], University of Michigan Prostate Cancer Specialized Program of Research Excellence Tissue Core. All samples were collected with the informed consent of the patients and prior approval of the institutional review board.

454 FLX sequencing. PolyA+ RNA was purified from 50 μg total RNA using two rounds of selection on oligo-dT containing paramagnetic beads using Dynabeads mRNA Purification Kit (Dynal Biotech), according to the manufacturer's instructions. mRNA (200 ng) 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 min. 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 adaptor ligation reaction was performed in Quick Ligase Buffer (New England Biolabs) containing 1.67 μM of the Adaptor A, 6.67 μM of the Adaptor B and 2,000 units of T4 DNA Ligase (New England Biolabs) at 37 °C for 2 h. Adapted library was recovered with 0.05% Sera-Mag30 streptavidin beads (Seradyn Inc.) according to the manufacturer's instructions. Finally, the single-stranded cDNA (sscDNA) library was purified twice with RNAClean (Agencourt) according to the manufacturer's directions, except that the amount of beads was reduced to 1.6× the volume of the sample. The purified sscDNA library was analysed on an RNA 6000 Pico chip on a 2100 Bioanalyser (Agilent Technologies) to confirm a size distribution between 450 and 750 nucleotides, and quantified with a Quant-iT Ribogreen RNA Assay Kit (Invitrogen Corporation) on a Synergy HT (Bio-Tek Instruments Inc.) 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, 1× Advantage 2 buffer and 1 μl of Advantage 2 polymerase mix (Clontech). The amplification reaction was performed at: 96 °C for 4 min; 94 °C for 30 s, 64 °C for 30 s, repeating steps 2 and 3 for a total of 20 cycles, followed by 68 °C for 3 min. The samples were purified using AMPure beads and diluted to a final working concentration of 200,000 molecules per microlitre. Emulsion beads for sequencing were generated using Sequencing emPCR Kit II and Kit III, and sequencing was performed using 600,000 beads.

Normalization by subtraction. mRNA from the prostate cancer cell line VCaP was hybridized with the subtractor cell line LNCaP first-strand cDNA immobilized on magnetic beads (Dynabeads, Invitrogen), according to the manufacturer's instructions. Transcripts common to both the cells were captured and removed by magnetic separation of bead-bound subtractor cDNA. The subtracted VCaP mRNA left in the supernatant was recovered by precipitation and used for generating the 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. mRNA (200 ng) 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 *Escherichia coli* DNA pol I (Invitrogen). The double-stranded cDNA library was further processed by Illumina Genomic DNA Sample Prep kit. 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 200 base pairs (bp) (±25 bp). 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) on a Modulus Single Tube Luminometer (Turner Biosystems) following the manufacturer's instructions. Library (10 nM) was used to prepare flowcells with approximately 30,000 clusters per lane.

Sequence data sets. Human genome build 18 (hg18) was used as a reference genome. All University of California, Santa Cruz (UCSC) and Refseq transcripts were downloaded from the UCSC genome browser (http://genome.ucsc.edu/)[30]. Sequences of previously identified *TMPRSS2–ERGa* fusion transcript (GenBank accession number DQ204772) and *BCR–ABL1* fusion transcript (GenBank accession number M30829) were used for reference.

Short-read chimaera 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 chimaeras we expect that there will be a larger portion mapping to a fusion partner (major alignment), and a smaller portion aligning to the second partner (minor alignment). Our approach is therefore divided into two phases in which we focus first on identifying the major alignment, and then performing a more exhaustive approach to identify 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[31]. Only reads that have an alignment of 12 or more nucleotides to an exon boundary are kept as potential chimaeras. 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 uses 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 chimaeras. It is possible to have a chimaera that has 28 nucleotides aligning to gene *x* and eight nucleotides that align to genes *y* and *z* because the eight-base polymer does not provide enough sequence resolution to distinguish between gene *y* and gene *z*. Therefore we would categorize this as two individual chimaeras. If a sequence forms more than five chimaeras it is discarded because it is ambiguous. To minimize false positives, we require that a predicted gene fusion event has at least two supporting chimaeras.

Long- and short-read integrated chimaera discovery. All 454 reads are aligned against the human Refseq collection using BLAT, a rapid mRNA/DNA alignment tool[32]. Using a Perl script, the BLAT output files were parsed to detect potential chimaeric 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 are therefore not characteristic of a chimaera. From the remaining reads, we want to query for reads having partial alignment, with minimal overlap, to two Refseq transcripts representing putative chimaeras. To accomplish this, we iterate the all the possible BLAT alignments for a putative chimaera, extracting only those partial alignments that have no more than a six-nucleotide, or two-codon, overlap. This step reduces false-positive chimaeras introduced by repetitive regions, large gene families and conserved domains. Additionally, although 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 chimaeras. These non-mapping short reads are subsequently aligned against all putative long-read chimaeras (obtained as described above) using Vmatch[31], 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. After this integration, the remaining putative chimaeras are categorized as inter- or intra-chromosomal chimaeras based on whether the partial alignments are located on different or the same chromosomes, respectively. Those intra-chromosomal chimaeras that have partial alignments to adjacent genes are believed to be the product of co-transcription of adjacent genes coupled with intergenic splicing (CoTIS)[18], alternatively known as read-throughs. The remaining intra-chromosomal and all inter-chromosomal chimaeras are considered candidate gene fusions.

One additional source of false-positive chimaeras could be an unknown transcript that is not in Refseq. Owing 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 chimaeras against the human genome using BLAT. If the long read has greater than 90% alignment to one genomic location, it is considered a novel transcript rather than a chimaeric read. The remaining chimaeras are given a score that 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, by 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 (March 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 for detecting unbalanced copy number changes at whole-genome level. Competitive hybridization of differentially labelled tumour and reference DNA to oligonucleotide printed in an array format (Agilent Technologies), and analysis of fluorescent intensity for each probe, will detect the copy number changes in the tumour 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 $\pm$ 0.5), for gains and losses involving more than one probe representing each genomic interval, as detected by the aberration detection method (ADM) in the CGH analytics algorithm.

Real-time PCR validation. qPCR was performed using Power SYBR Green Mastermix (Applied Biosystems) on an Applied Biosystems Step One Plus Real Time PCR System as described[3]. All oligonucleotide primers were synthesized by Integrated DNA Technologies and are listed in Supplementary Table 8. *GAPDH*[33] primer was as described. All assays were performed in duplicate or triplicate, and results were plotted as average fold change relative to *GAPDH*.

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

FISH. FISH hybridizations were performed on VCaP, LNCaP and FFPE tumour and normal tissues. Bacterial artificial chromosome clones were selected from the UCSC genome browser. After colony purification, midi prep DNA was prepared using QiagenTips-100 (Qiagen). DNA was labelled by nick translation labelling 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 sulphate and 1% Denhardt's solution. About 200 ng of labelled probes was hybridized to normal human chromosomes to confirm the map position of each bacterial artificial chromosome clone. FISH signals were obtained using anti digoxigenin-fluorescein and Alexa Fluor594 conjugate for green and red colours, respectively. Fluorescence images were captured using a high-resolution CCD (charge-coupled device) camera controlled by ISIS image processing software (Metasystems).

Affymetrix Genome-Wide Human SNP Array 6.0. One microgram each of genomic DNA samples was sent to Affymetrix service centres (Center for Molecular Medicine and Vanderbilt Affymetrix Genotyping Core) 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 browser.

28. Korenchuk, S. *et al.* VCaP, a cell-based model system of human prostate cancer. *In Vivo* **15,** 163–168 (2001).
29. Rubin, M. A. *et al.* Rapid ('warm') autopsy study for procurement of metastatic prostate cancer. *Clin. Cancer Res.* **6,** 1038–1045 (2000).
30. Karolchik, D. *et al.* The UCSC Table Browser data retrieval tool. *Nucleic Acids Res.* **32** (Database issue), D493–D496 (2004).
31. Abouelhoda, M. I., Kurtz, S. & Ohlebusch, E. Replacing suffix trees with enhanced suffix arrays. *J. Discrete Algorithms* **2,** 53–86 (2004).
32. Kent, W. J. BLAT – the BLAST-like alignment tool. *Genome Res.* **12,** 656–664 (2002).
33. Vandesompele, J. *et al.* Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes. *Genome Biol.* **3,** 34–50 (2002).