Published in final edited form as:

Genes Chromosomes Cancer. 2012 October ; 51(10): 933–948. doi:10.1002/gcc.21977.

Exclusion of the 750-kb Genetically Unstable Region at Xq27 as a Candidate Locus for Prostate Malignancy in HPCX1-linked Families

Natalay Kouprina^{1,*}, Nicholas CO Lee¹, Adam Pavlicek², Alexander Samoshkin¹, Jung-Hyun Kim¹, Hee-Sheung Lee¹, Sudhir Varma¹, William C Reinhold¹, John Otstot³, Greg Solomon³, Sean Davis⁴, Paul S. Meltzer⁴, Johanna Schleutker^{5,6}, and Vladimir Larionov¹

¹Laboratory of Molecular Pharmacology, NCI, NIH, Bethesda, Maryland, USA ²Computational Biology, Oncology Research Unit, Pfizer Global Research and Development, La Jolla, CA, USA ³Laboratory of Molecular Carcinogenesis, NIEHS, NIH, Research Triangle Park, North Carolina, USA ⁴Cancer Genetics Branch, NCI, NIH, Bethesda, MD, USA ⁵Department of Medical Biochemistry and Genetics, University of Turku, Finland ⁶Institute of Biomedical Technology/BioMediTech, University of Tampere and Fimlab Laboratories, Tampere, Finland

Abstract

Several linkage studies provided evidence for the presence of the hereditary prostate cancer locus, *HPCX1*, at Xq27-q28. The strongest linkage peak of prostate cancer overlies a variable region of ~750 kb at Xq27 enriched by segmental duplications (SDs), suggesting that the predisposition to prostate cancer may be a genomic disorder caused by recombinational interaction between SDs. The large size of SDs and their sequence similarity make it difficult to examine this region for possible rearrangements using standard methods. To overcome this problem, direct isolation of a set of genomic segments by *in vivo* recombination in yeast (a TAR cloning technique) was used to perform a mutational analysis of the 750 kb region in X-linked families. We did not detect disease-specific rearrangements within this region. In addition, transcriptome and computational analyses were performed to search for non-annotated genes within the Xq27 region, which may be associated with genetic predisposition to prostate cancer. Two candidate genes were identified, one of which is a novel gene termed *SPANXL* that represents a highly diverged member of the *SPANX* gene family, and the previously described *CDR1* gene that is expressed at a high level in both normal and malignant prostate cells, and mapped 210 kb of upstream the *SPANX* gene cluster. No disease-specific alterations were identified in these genes. To summarize, our results exclude the 750-kb genetically unstable region at Xq27 as a candidate locus for prostate malignancy. Adjacent regions appear to be the most likely candidates to identify the elusive *HPCX1* locus.

Keywords

Xq27; hereditary prostate cancer; HPCX1; SPANXL; CDR1; TAR cloning

*Correspondence to: Natalay Kouprina, Laboratory of Molecular Pharmacology, NCI, NIH, Bethesda, Maryland, USA. kouprinn@mail.nih.gov.

AUTHOR CONTRIBUTIONS

Conceived and designed the experiments: NK, AP, SV, SD, GS, VL
Performed the experiments: NK, NCOL, AP, AS, JHK, HSL, SV, JO, GS, SD
Analyzed the data: NK, WCR, PSM, JS, VL
Wrote the paper: NK, AP, VL

INTRODUCTION

Currently, prostate cancer is the most frequent cancer among men in many developing countries and the second leading cause of cancer-related deaths in men in the United States (Verhage and Kiemeny 2003; Rubin and De Marzo 2004; Schaid 2004). Based on 2010 estimates, in the United States alone over 200,000 new cases of prostate cancer are diagnosed and more than 30,000 men die from this disease annually. Evidence that genetics may play a critical role in the development of prostate cancer is based on a variety of studies, including case-control, cohort, twin and family-based studies. In 1956, Morganti and co-authors suggested a strong familial predisposition for prostate cancer (Morgani et al. 1956). The search for prostate cancer susceptibility genes by linkage studies offered hope that finding such genes would help in early diagnosis and that in turn would reduce the rate of mortality among prostate cancer patients. Work over the past two decades, using genome-wide scans in prostate cancer families, has identified 40 risk candidate regions for prostate cancer susceptibility loci on different chromosomes, which indicates great heterogeneity for this disease (Varghese and Easton 2010; Colloca and Venturino 2011). One of the first loci found by linkage analysis of 360 hereditary prostate cancer families is HPCX1 on chromosome Xq27-q28, estimated to be responsible for approximately 16% of hereditary prostate cancer cases (Xu et al. 1998). More recently, the prostate cancer linkage at this locus was confirmed by an independent study in a Finnish population of patients who fit the criteria for hereditary prostate carcinoma (Schleutker et al. 2000) and by others (Lange et al. 1999; Neuhausen et al. 1999; Stephan et al. 2002; Xu et al. 2003; Bochum et al. 2002; Brown et al. 2004; Gillanders et al. 2004; Farnham et al. 2005; Yaspan et al. 2008). A genome-wide scan for prostate cancer susceptibility genes in Finnish X-linked families provided the evidence that the strongest linkage peak of prostate cancer overlies a region of ~750 kb at Xq27 (Baffoe-Bonnie et al. 2005). This region contains the cluster of five *SPANX* (Sperm Protein Associated with the Nucleus on the X chromosome) gene family members, *SPANXA1*, *A2*, *B*, *C*, and *D*. These genes encode sperm proteins associated with the nucleus, whose expression is restricted to the normal testis, a few nongametogenic tissues, and certain tumors (Zendman et al. 1999; Zendman et al. 2003; Westbrook et al. 2000; Westbrook et al. 2001; Westbrook et al. 2004; Goydos et al. 2001; Wang et al. 2003; Kouprina et al. 2004; Kouprina et al. 2007a), and *LDOC1*, which has been shown to be down-regulated in some cancers (Nagasaki et al. 2003). The *SPANX* genes are embedded into large inverted SDs with a high level of similarity that represent approximately one third of the region, predicting an increased level of genomic instability. The non-overlapping deletions in two patients with prostate cancer that coincided closely with the interval under close scrutiny for the hereditary prostate cancer locus Xq27-q28 and structural chromosomal abnormalities such as duplications and interchromosomal translocations at Xq27 have been described (Kibal et al. 2003; Solomon et al. 2004; Zhu et al. 2011; Vitek et al. 2012).

During the past several years, we have studied the hypothesis for the presence of a novel mutation or pre-existing polymorphism in one of the *SPANX* genes or in *LDOC1* that may cause predisposition to prostate cancer. Our analysis revealed an extensively complex and dynamic organization of the *SPANX* genes. A variable number of *SPANX* gene copies and frequent homology-based sequence transfer events between *SPANX* gene family members were both detected. These were presumably caused by recombinational interactions between SDs harboring the *SPANX* genes (Kouprina et al. 2005). However, none of the sequence variations in the coding regions of these genes, as well as *LDOC1*, were associated with susceptibility to prostate cancer. Thus, our work excluded these genes as candidates for hereditary prostate cancer susceptibility.

In this study, we propose that the predisposition to prostate cancer in X-linked families may be a “genomic disorder” (Stankiewicz and Lupski 2006; Stankiewicz and Lupski 2010) caused by a genomic rearrangement(s) at Xq27, affecting the expression of one or more nearby genes, or it may be caused by a mutation in an unknown (not yet annotated) gene located within the candidate region.

MATERIALS AND METHODS

Human Subjects/DNA Samples

A detailed description of the study samples is presented elsewhere (Xu et al. 1998; Xu et al. 2003; Gillanders et al. 2004; Schleutker et al. 2000). DNA was isolated from transformed lymphoblast cell lines. 24 families with prostate cancer were used for analysis, among which 17 were families with X-linked prostate cancer and 7 were families with “non-male-to-male” (NMM) disease inheritance. 12 families were represented by brother pairs (Table 1). In addition, 27 unaffected controls that included blood samples from healthy individuals were analyzed, including 14 individuals with benign prostate hyperplasia (BPH) (Pir-samples) and 7 anonymous healthy males (M-samples) (University of Tampere and Tampere University Hospital, Finland). DNA was isolated using a Puregene DNA isolation kit (Gentra Systems, Inc., Minneapolis). 40 random samples of genomic DNA from normal male individuals (Caucasians) used as eligible controls were purchased from Coriell Institute for Medical Research. For transcriptome analysis a panel of the NCI-60 cell lines derived from nine tissue-of-origin types of cancer was used (Shoemaker 2006).

Array Comparative Genome Hybridization

Genomic DNA was isolated from the cell lines developed from affected and unaffected individuals by standard methods. High-resolution tiling arrays were designed and produced by Nimblegen (Madison, WI). The arrays had approximately 385,000 probes. Data were visualized using the Affymetrix Integrated Genome Browser (available from Affymetrix, Inc., Santa Clara, CA)

Construction of TAR Vectors and Cloning by *in vivo* Transformation-Associated Recombination in Yeast

Isolation of the genomic fragments from Xq27 was carried out by recombinational cloning (TAR cloning) (Fig. 2) (Kouprina and Larionov 2006; Kouprina and Larionov 2008). Nine TAR circularizing vectors, TAR1-TAR10, were constructed using the basic vector pVC604 (Kouprina and Larionov 2008). These vectors contain unique targeting sequences (or hooks) approximately 100–300 bp in size, corresponding to the 5' and 3' ends of the targeted region inserted into the polylinker of pVC604. The targeting sequences were designed based on the available information (hg18). The sizes and corresponding positions of the hooks in the human genome are described in supplementary Table S1. The TAR vectors were linearized with an endonuclease (the recognition site of which is located between the hooks) before TAR cloning. For transformations, the highly transformable *S. cerevisiae* strain VL6–48 (MAT α , his3- Δ 200, trp1- Δ 1, ura3-52, lys2, ade2-101, met14) that has *HIS3* deleted was used. For TAR cloning, agarose plugs (~60 μ l) containing approximately 1–2 μ g of high molecular weight human DNA from the cell lines of patients and healthy brothers were prepared, mixed with a linearized TAR vector (1 μ g), and presented to freshly prepared yeast spheroplasts. Yeast transformants were selected on synthetic complete medium plates lacking histidine. Six to ten transformation experiments were carried out for each construct. The yield of transformants per 1–2 μ g of human DNA using 1 μ g of vector and 5×10^8 spheroplasts varied between 10 and 50. To identify positive clones, the transformants were combined into pools and examined by PCR with the diagnostic primers for the unique fragment sequences not present in the vector (supplementary Table S1). One to five positive

pools were usually identified. The PCR products were sequenced to verify that they matched the expected sequences. Individual clones containing the expected region were found in positive pools and used for further analysis. Some TAR-cloned regions were further characterized by additional overlapping PCR using specific pairs of primers (supplementary Table S2).

Physical Characterization of YAC/TAR Clones

Chromosomal-size DNAs from yeast transformants prepared in agarose DNA plugs were digested by NotI, separated by Clamped Homogeneous Electrical Field (CHEF) gel electrophoresis, blotted, and hybridized with total human genomic DNA as previously described (Kouprina and Larionov 2006; Kouprina and Larionov 2008 and references therein).

Analysis of Chromosomal Regions at Xq27 Using a Set of Overlapping PCR Reactions

A set of primers that amplify 1–6 kb overlapping fragments was designed (supplementary Table S2). DNA sequencing of the ends of each fragment was carried out on a PE-Applied Biosystem 3100 Automated Capillary DNA Sequencer.

Analysis of the 12-kb Tandem Repeat in the *SPANXB* Segmental Duplication

The DOTTUP program from the EMBOSS package (<http://emboss.sourceforge.net/>; Rice et al. 2000) with the word size set to 30 bp was used to detect *SPANXB* duplication in the AL451048.12 contig. The exact duplication breakpoints were obtained using the MAFFT alignment program (<http://align.bmr.kyushuu.ac.jp/mafft/software/>; Katoh et al. 2005). Programs in the EMBOSS suite (Rice et al. 2000) were used for the analysis of breakpoint sequences. Short sequence repeats were identified by WORDMATCH, palindromes were identified by EINVERTED, and FUZZNUC was used for the detection of motifs associated with genome rearrangements.

Characterization of the Xq27 Region by Restriction Analysis

Chromosomal-size genomic DNA was isolated from the cell lines developed from patients and unaffected controls, digested with either BamHI, KpnI, or PmeI, CHEF gel electrophoresis-separated, transferred to a nylon membrane and blot-hybridized with the specific probe. Positions of the restriction sites at Xq27 are shown in supplementary Table S3. The probe was designed from the sequence of the *SPANX* genes with homology to all *SPANX*-containing fragments (supplementary Table S3).

Determination of Location of Consistent Gene Expression in the NCI-60

To look for the regions with consistent, robust transcription, we used expression data for the NCI-60 from five different arrays (Agilent 44K, Affymetrix U95, Affymetrix U133, Affymetrix U133 Plus 2.0, Affymetrix HuEx). Probes mapping to the specific locations were ordered according to their position in the human genome. We assumed that if transcription were not taking place at a location the probes would be uncorrelated and related to background interference. Thus, a high correlation between neighboring probes across the 60 cell lines was taken to be indicative of consistent transcription. The p-value of the Pearson correlation coefficient was calculated as a measure of statistical significance. These p-values were adjusted for multiple comparisons (across all of the neighboring probe pair correlations) using the False Discovery Rate (FDR) method (Benjamini and Hochberg 1995).

Expression of *SPANXL* and a Hypothetical Long Non-coding RNA (lncRNA)

Transcripts of the *SPANXL* gene were checked by RT-PCR with 22 normal human tissues (Ambion Inc., Austin, TX), i.e., brain, thymus, lung, lymph node, adrenal, skeletal muscle, heart, liver, breast, ovary, placenta, spleen, stomach, colon, proximal colon, pancreas, kidney, testicle, cervix, small intestine, prostate, and uterus as well as in cancer cell lines (two cell lines, Colo205 and PC3, from the NCI-60 panel). Two pairs of primers were designed to amplify genomic and spliced variants of the *SPANXL* gene as 678-bp and 489-bp fragments, respectively (supplementary Table S4). RT-PCR analysis of the hypothetical lncRNA covered by two ESTs, BM811380 and DA247186, was carried out by two pairs of primers designed from EST sequences (supplementary Table S4). cDNA was made from 1 µg of total RNA isolated from normal and cancer cells using the Superscript first strand system kit (Invitrogen, Carlsbad, CA, USA) and priming with oligo dT per their standard protocol. RT-PCR was performed using 1 µl of cDNA in a 50 µl reaction volume. Standard reaction conditions were 94°C 5 min, (94°C 1 min, 55°C 1 min, 72°C 1 min × 35 cycles), 72°C 7 min, 4°C hold. The PCR products were cloned into a TA vector. Clones were sequenced in each direction on a PE-Applied Biosystem 3100 Automated Capillary DNA Sequencer.

Mutational Analysis of the *SPANXL* and *CDR1* Genes

The promoter, coding and intronic regions, and 3' noncoding sequences of *SPANXL* and *CDR1* genes were PCR-amplified from affected and unaffected individuals using gene-specific primers (supplementary Table S4). The PCR fragments were cloned into a TA vector. Clones were sequenced in each direction on a PE-Applied Biosystem 3100 Automated Capillary DNA Sequencer. Sequences of different alleles were aligned with MAVID [Bray and Pachter 2004; <http://baboon.math.berkeley.edu/mavid/>].

Sequence Analysis

Database searches were performed using versions of the BLAST program appropriate for different types of sequence comparisons: BLASTN for nucleotide sequences, BLASTP for protein sequences, and TBLASTN for searching a nucleotide database translated in six frames with a protein query (Altschul et al. 1990; Atschul et al. 1997). DNA sequences with *SPANX* locus promoters were aligned with MAFFT (Katoh and Toh 2010). *SPANX* proteins were aligned using Dialign v2.2 (Morgenstern 2004) and conservation profiles were obtained using GeneDoc (Nicholas et al. 1997). The *SPANX* proteins phylogenetic tree was obtained by the neighbor-joining method using the Poisson distance in Seaview (Gouy et al. 2010).

RESULTS

Analysis of Genomic Deletions/Amplifications within the *HPCX1* Mapped Region by CGH

It is well documented that SDs mediate ectopic interaction of loci that can result in chromosomal rearrangements such as duplications, deletions, and inversions (Bailey et al. 2002; Sebat et al. 2004; Stankiewicz and Lupski 2006; Stankiewicz and Lupski 2010; Mefford and Eichler 2009). Thus, a high density of SDs within the *HPCX1* region suggests that the predisposition to prostate cancer in some *HPCX1* families may be a result of genomic rearrangements mediated by SDs.

To survey the landscape of DNA copy number alterations in X-linked prostate cancer, we profiled patients from three families with the strongest linkage to Xq27-q28, i.e., 232, 236 and 001 (Table 1). Oligonucleotide-base microarray analysis was performed for the region that spans ~6.5 Mb (from 138.5–145.0 Mb; hg18). Among five samples (three patients and two healthy controls), the only region of decreased copy number noted was between the

genes *MAGEC1* and *MAGEC2* in the control sample KS extending from 140.89 Mb to 140.94 Mb (Fig. 1). Such common polymorphism (variation_52943) has been previously described in the human population (Shaikh et al. 2009). Thus, CGH analysis did not reveal any deletions/amplifications at Xq27-q28 specific for the X-linked prostate cancer patients.

Validation of the Known Genomic Inversions at Xq27 in Prostate Cancer Patients

In the recent paper by Kidd et al. (2008) the authors identified 217 polymorphic inversions in the human genome by fingerprint and DNA sequence analyses. They found a twofold enrichment of inversions mapping to clustered regions of the X chromosome, consistent with theoretical predictions of increased inversion content based on unusual inverted repeat structures. Two inversions (_37352 and _37253) were mapped to Xq27, the most complex region enriched by SDs (positions of breakpoints are presented in Table 2). To check whether inversion _37352 is present in prostate cancer patients, we used two pairs of primers, F40-F1/R-443-R1 and F5387-F2/R53879-R2 (supplementary Table S4), designed from the non-rearranged genomic sequence, i.e., diagnostic for lack of inversion. We analyzed all members from 24 Finnish families (Table 1), 46 healthy Finnish and 40 Caucasian random males (total 119 individuals). The PCR worked for all DNA samples. The amplified products were sequenced all matching the predicted genomic sequences which indicate the absence of inversion _37352 in the analyzed samples. In inversion _37253, only one of the breakpoints (right) can be checked by PCR because another breakpoint (left) is localized within the SD common for five *SPANX* genes preventing its study by PCR. Absence of the right breakpoint was proven using a pair of diagnostic primers, In353-F/In353-R (supplementary Table S4). Absence of the left breakpoint of inversion _37253 was proven using TAR cloning. The predicted size genomic fragments (42 kb) with no evidence of genomic rearrangements were isolated from affected and unaffected members by TAR8 vector (see below and Fig. 3c). Thus, we concluded that these described inversions are a rare event and not a cause of hereditary prostate cancer in X-linked prostate cancer patients. However, this does not exclude the presence of other inversions in this region induced by large inverted SDs.

Analysis of Genomic Rearrangements within the *HPCX1* Region by TAR Cloning

Before searching for rearrangements, we analyzed the sequence at Xq27 for potential fragile sites. Fragile sites are common targets for chromosome rearrangements (Arlt et al. 2006). Fragile site instability is well documented and includes gaps and breaks on metaphase chromosomes, translocation and deletions breakpoints, and sister chromosome exchanges. For example, one of the breakpoints described involved in translocations occurred at the center of a near-perfect palindromic AT-rich repeat (NPPR) that is 595 bp long (Gotter et al. 2007). Fragile sites may increase chromosome instability caused by SDs. We found two NPPR repeats at Xq27 (Fig. 3a; Table 2). The length of the largest NPPR repeat (#1) is 261 bp. We developed primers (supplementary Table S4) and amplified and sequenced this repeat in the three families with the strongest linkage to Xq27 (001, 232, 236). Sequence analysis did not reveal significant differences between affected and unaffected brothers. We found that in patient 232-001 six ATs were deleted. In patient 001-002 four nucleotides containing "G" (TGTA) were deleted. In patient 236-006 the sequence corresponds to the expected genome sequence (supplementary Fig. S1). These variants correspond to a common polymorphism in population.

To determine whether the Xq27 region enriched by SDs contains inversions or other genomic rearrangements, TAR cloning was used for direct isolation of the genomic segments. A scheme of TAR cloning applied for one of the isolated genomic segments is shown in Figure 2. We considered the most likely recombination event to be an inversion due to an interaction between the largest inverted SDs, SD-B (119 kb) and SD-A1/A1 (113

kb) (Fig. 3b), and thus concentrated on the isolation of overlapping genomic fragments that would completely cover the potential breakpoint on one side of the inversion (Fig. 3a and 3b). For these experiments, TAR vectors were constructed to isolate the fragments of 40–140 kb in size covering the region from 139,713,026 to 140,293,973 (581 kb) and from 140,492,674 to 40,534,123 (42 kb) (hg18) (Fig. 3a; Table 3). Three families with the strongest linkage to the *HPCX1* locus (232, 236, and 001) (Table 1) were chosen for cloning. Segments were TAR-cloned from affected and unaffected brothers. To verify the size of the cloned fragments, the TAR/YAC-isolates were linearized by a unique endonuclease, CHEF-gel separated, and blot-hybridized with the human DNA probe. For all TAR clones except TAR5, the size of the inserts corresponded to the predicted size, based on the genome sequence information (Fig. 3c). A physical analysis of the TAR5 clones showed variability in the size of the inserts, ranging from 54 kb to 114 kb, due to a variable number of a 12-kb repeat within SD-B that is not linked to prostate malignancy as described earlier (Kouprina et al. 2005). The mechanism of a 12-kb repeat amplification is described in supplementary Figure S2. Each TAR isolate was checked by a set of diagnostic primers for the 3' and 5' ends and middle of the insert (supplementary Table S1). We found only one polymorphic marker in the 001 family in the region cloned by the TAR1 vector using the diagnostic primers Diag 713-F/Diag 713-R (supplementary Table S1). It was a 34 bp deletion of the LTR/ERV1-MaLRs sequence (gcttctgtacttgctgtactactataaaca).

In addition, we carried out PCR analysis of the 88 kb (PCR1), 53 kb (PCR2), and 84 kb (PCR3) genomic regions (Fig. 3a; Table 3) from the *HPCX1* locus using a set of primers that amplify 1–6 kb overlapping fragments (supplementary Table S2). The positions of the covered regions are presented in Table 3. Note that PCR1 covers TAR4 and partially overlaps with TAR2, TAR3, and TAR5 isolates. PCR3 overlaps with the TAR10 clone. All predicted overlapping fragments were amplified, sequenced and verified. The total length of genomic DNA covered by TAR cloning and PCR is ~ 623 kb. The results of TAR and PCR analyses were supported by Southern-blot hybridization (supplementary Fig. S3; Table S3). Thus, these analyses do not support the presence of any rearrangements or inversions within the region analyzed.

Search for Novel Genes in the Critical Region

The NCI-60 cell lines derived from nine tissue-of-origin types of cancer have been characterized in multiple manners, including transcript expression (Shoemaker 2006). Therefore, to search for new, non-annotated regions/genes with consistent, robust transcription at Xq27, we used expression data accumulated for NCI-60 from five different microarray platforms (see Materials and Methods). Briefly, we assumed that if transcription did not take place at a location, the probes would be measuring noise and thus would be mostly un-correlated. Thus, a high correlation between neighboring probes across the 60 cancer cell lines was taken to be indicative of consistent transcription. Figure 4a shows the correlation between neighboring pairs of probes plotted along the midpoint between the probes in each pair for the region that spans ~ 1 Mb. Correlations with FDR<0.05 are marked by red points. As seen in the Figure 4a, no transcripts except those for *LDOC1*, *CDR1* genes and two ESTs (BM811380 and DA247186) were detected.

A recent publication on whole transcriptome sequencing in prostate cancer samples (Prensner et al. 2011) enabled us to re-analyze an expression profile in the 750 kb critical region and beyond. We downloaded the BAM files from 21 prostate cancer cell lines with sequence reads aligned to the human genome provided by the authors (GSE25183). The BAM files were merged, and the mapped read coverage was inspected in the Integrative Genomics Viewer (Robinson et al. 2011). The results were similar to those obtained with NCI-60 cell lines. Two potential non-annotated transcripts were found from two short regions at the 5' and 3' ends of *CDR1*. The first region was on the 5' end of *CDR1* at the

approximate positions chrX:139,674,300-139,688,550 (hg18) and overlaps with the 3' end of two ESTs (BM811380 and DA247186) mentioned above (Fig. 4b). Because no significant similarity to known genes was found when translated Blastx searches were used, these transcripts (overlapping ~63 kb) likely represent a long non-coding RNA with an unknown function. The second apparent expressed region from the 3' end of *CDR1* (chrX:140,060,370-140,060,750) shares a strong similarity to the 60S ribosomal protein *L36a* locus and represents a partial pseudogene. Given the presence of two mismatches from the chromosome X sequence in all reads, these reads likely represent a misalignment of sequences from another genomic locus to the *L36a* pseudogene at Xq27. No other novel transcripts were found within or close to the critical association region at Xq27.

Thus, computational analysis revealed only one protein-encoding gene within the candidate region, *CDR1*, that is expressed in normal and cancer prostate cells. Therefore, this gene may be considered as a candidate gene for prostate malignancy.

Mutational Analysis of the *CDR1* Gene in Prostate Cancer Patients

The human *CDR1* gene is found as a single copy and encodes a protein with a distinctive tandemly repeated structure. It is predominantly expressed in normal neuroectodermal tissues and in certain malignant tumors (Rettig et al. 1987; Dropcho et al. 1987). The *CDR1* gene, 1,368 bp in size, was mapped to the region Xq26-q27.2 (Furneaux et al. 1990). The length of the open reading frame (ORF) is 801 bp.

Two primers, CDR-F/CDR-R, (supplementary Table S4) were used for mutational analysis of the *CDR1* gene in 24 Finnish families (Table 1). In three families with the strongest linkage to Xq27 (232, 236, and 001) the R-224-C mutation of the CDR1 protein was found (Table 4). However, there is no strong correlation between affected and unaffected members of the families. For example, in the family 232 both affected and unaffected brothers have this mutation. Analysis of the 1094 genomes (1000 Genomes Project Consortium 2010; http://browser.1000genomes.org/Homo_sapiens/Variation/Individual) where the Finnish population is represented by 93 samples revealed 12 samples with the R-224-C mutation (KG_X_139865862). From these 12 samples, three are from Finnish individuals. We also analyzed 27 samples of Finnish random male individuals and found only a known polymorphic mutation I-19-V (supplementary Table S5). Therefore, a frequency of the R-224-C mutation is higher in Finns compared to random population (3.2% vs 0.89%), suggesting that this mutation is a common polymorphism in the Finns. To summarize, the *CDR1* gene is unlikely a candidate for prostate malignancy.

Discovery and Mutational Analysis of the *SPANXL* Gene in Prostate Cancer Patients

Given some limitation of transcriptome analyses (a possible lack of probes for a non-annotated gene(s) in microarrays or a gene's low expression level), we searched the GenBank database for EST sequences from Xq27-q28. Two ESTs in dbEST (AX748173 and BC042039) were found. Analysis of the corresponding genomic sequence revealed a novel gene consisting of two exons separated by a short intron. The positions of the coding regions of the gene are presented in Table 2. The length of the first exon is 387 bp. The length of the second exon is 156 bp. The length of intron is 189 bp (Fig. 5b). Alignment of the novel protein with SPANXA/D and SPANXN proteins showed conservation of key residues between all SPANX proteins and relatedness of the novel locus to *SPANXN* genes (Fig. 5d). A phylogenetic tree confirmed that the novel protein is distantly related to the SPANXN family (supplementary Fig. S4), and we thus termed it SPANXL for 'SPANX-like'. *SPANXL* shares promoter sequence conservation with other *SPANX* family members (supplementary Fig. S5), indicating that conservative selection is operating on the promoter sequences and possibly shared co-regulation of all *SPANX* genes. However, the *SPANXL*

gene contains an insertion of unknown origin that contains an alternative initiation AUG codon confirmed by RT-PCR. As a result, the SPANXL protein contains an additional 97 amino acids at the C-terminus that are unrelated to the other SPANX proteins (Fig. 5d). The sequence of the *SPANXL* gene was deposited into GenBank under accession number JN874479. RT-PCR analysis showed that the gene is expressed in normal tissues and some cancer cell lines (see Materials and Methods). A 489-bp band of the expected size confirmed by sequencing was detected (Fig. 5c). Note that transcription of *SPANXL* is relatively low, similar to that observed for other *SPANX* genes, which explains the lack of signal on microarray.

A mutational analysis of the *SPANXL* gene in X-linked prostate cancer families and 46 random Finnish males was performed using the primers Dog-F1/Dog-R1 (supplementary Table S4) that amplify a 1,024 bp sequence of the gene (122 bp upstream of the ATG codon and 170 bp downstream of the STOP codon in exon 2). Three identical mutations were found in random Finnish males [A to G in position 160 of exon 1 (I-83-V), a synonymous A to G mutation in position 180 of exon 1, and A to G 53 bp downstream of the STOP codon] while no mutations were found within the members of 25 families. Thus, the analysis of *SPANXL* does not support this gene as a susceptibility gene for hereditary prostate cancer.

Mutational Analysis of the microRNA hsa-mir-320d-2 in Prostate Cancer Patients

In the past decade, microRNAs (miRNAs) have been uncovered as key regulators of fundamental cellular processes in diverse organisms, including gene expression at the post-transcriptional level (Van Wynsberghe et al. 2011). Dysregulation/mutations of these microRNAs have been found to have relevance to tumorigenesis and also to neurological, cardiovascular, developmental, and other diseases (Esteller 2009). Therefore, we examined the Xq27-q28 region for the known microRNAs but did not find any. In 2009, Ren and co-authors described *Mus musculus* microRNA, Mir-320 (MI0000704), that they have shown to be involved in the regulation of cardiac ischemia/reperfusion injury by targeting heat-shock protein 20 (Ren et al. 2009), and mapped it to chromosome X. Human BLAT revealed a microRNA 48 bp in length, hsa-mir-320d-2 (MIR320D2), at positions 139,836,003 - 139,836,050 on chromosome X (MI000819) (Table 2) that has 87% homology to the mouse Mir-320 (Fig. 5e).

To check for the presence of any mutations in the hsa-mir-320d-2 DNA sequence in prostate cancer patients from 25 families (Table 1), we developed a pair of primers, F20980-F/ SpanR-R (supplementary Table S4), that amplify a 3,569 bp fragment. A 741 bp sequence reading by the primer F20980-F (~394 bp downstream of the promoter and ~130 bp upstream of hsa-mir-320d-2) did not reveal any nucleotide substitution except for one nucleotide C to T transition at position 139,836,181 that is located ~130 bp upstream of hsa-mir-320d-2. However, the mutation in this microRNA gene does not cause predisposition to prostate cancer.

DISCUSSION

Genetic linkage studies identified a locus for susceptibility to prostate cancer called *HPCX1* at Xq27-q28 (Xu et al. 1998). The linkage peak of prostate cancer overlies a region of ~750 kb at Xq27 containing the cluster of nine *SPANX* genes encoding cancer-testis (CT) antigens and the *LDOC1* gene which were excluded as candidates for the susceptibility gene by our previous studies (Kouprina et al. 2005; Kouprina et al. 2007b). Because of the presence of large SDs with a high level of sequence homology at Xq27-q28, we proposed that the predisposition to prostate cancer in X-linked families may be a “genomic disorder” (Stankiewicz and Lupski 2006; Stankiewicz and Lupski 2010) caused by a genomic rearrangement(s) affecting the expression of one or more nearby genes. Searching for

rearrangements within a structurally unstable genomic region is a challenging task. The standard methods are unable to overcome all uncertainties related to SDs, polymorphism, and copy number variations.

In this work, we took an advantage of transformation-associated recombination (TAR) cloning technology that allows direct isolation of genomic fragments and gene clusters with a length up to 250 kb from human genomes (Kouprina and Larionov 2006; Kouprina and Larionov 2008). Specifically, genomic fragments were isolated from affected and unaffected brothers of X-linked families. A physical analysis of the TAR-isolated genomic segments corresponding to the *SPANX* gene cluster and overlapping regions did not reveal any detectable rearrangements. In addition, the regions not having SDs were checked by a set of overlapping PCR reactions and also did not reveal any rearrangements. The results of CGH and Southern-blot hybridization analyses were in agreement with the TAR cloning data. Thus, our data do not support the hypothesis that hereditary prostate cancer at Xq27 is a “genomic disorder” caused by instability of this region.

A failure to detect any specific genomic changes within the candidate region raised two other possibilities: i) the presence of a gene(s) that has escaped annotation but mutations in which may cause prostate cancer and ii) the possibility that the susceptibility gene is located outside of the analyzed chromosomal region, which is quite possible because of the difficulty in converting the genetic map into a physical map of the precise chromosomal position for each gene. To address these questions, we analyzed the transcription pattern of the candidate region in a set of 60 different cancer cell lines (NCI-60) representing nine different types of cancer. We have also analyzed RNA-Seq data available for samples of sporadic prostate cancer and normal controls (Prensner et al. 2011). Neither analysis revealed unknown transcribed regions within the 750 kb region analyzed. However, search of the GenBank database for EST sequences from Xq27-q28 resulted in the discovery of a novel *SPANX*-like gene, *SPANXL*, located between the *SPANXA1/2* and *SPANXD* genes. A mutational analysis did not reveal any specific changes of this gene in the X-linked families. We have also excluded a hypothetical microRNA, hsa-mir-320d-2, located within this region as a prostate cancer gene candidate. Unexpectedly, transcriptome analysis of the candidate region in prostate samples revealed a high level of expression of the *CDR1* gene, which is mapped close to the *SPANX* gene cluster. This gene is still poorly characterized, and nothing is known about its function. Antibodies directed against the CDR1 protein have been identified in some patients with paraneoplastic cerebellar degeneration, which occurs in association with small cell carcinoma of the lung, neoplasms of the breast and ovary, and Hodgkin disease (Chen et al. 1990). DNA sequence analysis of the *CDR1* gene in affected and unaffected individuals revealed two alleles of the gene, including a nonsynonymous deleterious mutation, R-224-C. However, none of these genetic variants is associated with susceptibility to X-linked prostate cancer.

It is worth noting that in addition to HPCX1, other disease-associated loci were mapped to Xq27-q28, including the *TGCT* locus linked to hereditary testicular cancer and the susceptibility loci for dyslexia, autism and migraine (Lutke et al. 2004; de Kovel et al 2004; Slavotinek et al. 2005; Vincent et al. 2005; Maher et al. 2012). Therefore, the study of newly identified coding regions in a series of these disorders would be warranted.

Since our analysis does not support the 750 kb repeat-rich polymorphic region as a susceptibility locus for prostate cancer, the gene responsible for hereditary prostate cancer remains unknown. Therefore, the search should be continued outside of this region. One candidate gene for prostate malignancy, *MAGEC1*, was recently proposed, based on the analysis of the *HPCX1* region with brother pairs from X-linked families using a recently developed NMD microarray technology (Mattila et al. 2011). The gene encodes a CT

antigen and is located ~200 kb away from the *SPANX* gene cluster (Lucas et al. 1998). Our transcriptome and computational analyses revealed a ~63 kb transcribed region covered by two ESTs, BM811380 and DA247186. Using pairs of specific primers (supplementary Table S4), this long transcript was found to be spliced and detectable in normal tissues (Fig. 5f), suggesting that it may represent a new long non-coding RNA involved in multiple regulation processes (Wang, Chang 2011). This hypothesis is further reinforced as the region corresponding to this long non-coding RNA is partially conserved and transcribed in mouse (RIKEN cDNA C230004F18 gene). Other known genes in the regions flanking the *SPANX* gene cluster have to also be examined. Given the incomplete annotation of genes in the human genome, a search for new transcribed regions would be warranted. The identification of the gene responsible for *HPCX1* might lead to promising diagnostic tests and therapies.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This study was supported by the Intramural Research Program of the NIH NCI Center for Cancer Research (V.L.). We thank Ms. Nina Kouprina for expert editing of this manuscript.

REFERENCES

- Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic Local Alignment Search Tool. *J Mol Biol.* 1990; 215:403–410. [PubMed: 2231712]
- Altschul SF, Madden TL, Schaffer AA, Zhang JH, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 1997; 25:3389–3402. [PubMed: 9254694]
- Arlt MF, Durkin SG, Ragland RL, Glover TW. Common fragile sites as targets for chromosome rearrangements. *DNA Repair.* 2006; 5:1126–1135. [PubMed: 16807141]
- Baffoe-Bonnie AB, Smith JB, Stephan DA, Schleutker J, Carpten JD, Kainu T, Gillanders EM, Matikainen M, Teslovich TM, Tammela T, Sood R, Balshem AM, Scarborough SD, Xu JF, Isaacs WB, Trent JM, Kallioniemi OP, Bailey-Wilson JE. A major locus for hereditary prostate cancer in Finland: localization by linkage disequilibrium of a haplotype in the *HPCX* region. *Hum Genet.* 2005; 117:307–316. [PubMed: 15906096]
- Bailey JA, Gu ZP, Clark RA, Reinert K, Samonte RV, Schwartz S, Adams MD, Myers EW, Li PW, Eichler EE. Recent segmental duplications in the human genome. *Science.* 2002; 297:1003–1007. [PubMed: 12169732]
- Benjamini Y, Hochberg Y. Controlling the False Discovery Rate - a Practical and Powerful Approach to Multiple Testing. *J Roy Stat Soc B Met.* 1995; 57:289–300.
- Bochum S, Paiss T, Vogel W, Herkommer K, Hautmann R, Haeussler J. Confirmation of the prostate cancer susceptibility locus *HPCX* in a set of 104 German prostate cancer families. *Prostate.* 2002; 52:12–19. [PubMed: 11992616]
- Bray N, Pachter L. MAVID: Constrained ancestral alignment of multiple sequences. *Genome Res.* 2004; 14:693–699. [PubMed: 15060012]
- Brown WM, Lange EM, Chen H, Zheng SL, Chang B, Wiley KE, Isaacs SD, Walsh PC, Isaacs WB, Xu J, Cooney KA. Hereditary prostate cancer in African American families: linkage analysis using markers that map to five candidate susceptibility loci. *Brit J Cancer.* 2004; 90:510–514. [PubMed: 14735201]
- Chen YT, Rettig WJ, Yenamandra AK, Kozak CA, Chaganti RSK, Posner JB, Old LJ. Cerebellar degeneration-related antigen - a highly conserved neuroectodermal marker mapped to chromosomes-X in human and mouse. *Proc Natl Acad Sci USA.* 1990; 87:3077–3081. [PubMed: 2326268]

- Colloca G, Venturino A. The evolving role of familial history for prostate cancer. *Acta Oncol.* 2011; 50:14–24. [PubMed: 20874046]
- de Kovel CG, Hol FA, Heister JG, Willemen JJ, Sandkuijl LA, Franke B, Padberg GW. Genomewide scan identifies susceptibility locus for dyslexia on Xq27 in an extended Dutch family. *J Med Genet.* 2004; 41:652–657. [PubMed: 15342694]
- Dropcho EJ, Chen YT, Posner JB, Old LJ. Cloning of a brain protein identified by autoantibodies from a patient with paraneoplastic cerebellar degeneration. *Proc Natl Acad Sci U S A.* 1987; 84:4552–4556. [PubMed: 3474619]
- Esteller M. Non-coding RNAs in human disease. *Nat Rev Genet.* 2011; 12:861–874. [PubMed: 22094949]
- Farnham JM, Camp NJ, Swensen J, Tavtigian SV, Albright LAC. Confirmation of the *HPCX* prostate cancer predisposition locus in large Utah prostate cancer pedigrees. *Hum Genet.* 2005; 116:179–185. [PubMed: 15592687]
- Furneaux HM, Wong E, Posner JB. Isolation of cDNA clones encoding the major Yo paraneoplastic antigen. *Neurology.* 1990; 40:166.
- Gillanders EM, Xu JF, Chang BL, Lange EM, Wiklund F, Bailey-Wilson JE, Baffoe-Bonnie A, Jones M, Gildea D, Riedesel E, Albertus J, Isaacs SD, Wiley KE, Mohai CE, Matikainen MP, Tammela TLJ, Zheng SL, Brown WM, Rokman A, Carpten JD, Meyers DA, Walsh PC, Schleutker J, Gronberg H, Cooney KA, Isaacs WB, Trent JM. Combined genome-wide scan for prostate cancer susceptibility genes. *J Natl Cancer I.* 2004; 96:1240–1247.
- Gotter AL, Nimmakayalu MA, Jalali GR, Hacker AM, Vorstman J, Duffy DC, Medne L, Emanuel BS. A palindrome-driven complex rearrangement of 22q11.2 and 8q24.1 elucidated using novel technologies. *Genome Res.* 2007; 17:470–481. [PubMed: 17351131]
- Gouy M, Guindon S, Gascuel O. SeaView version 4: A multiplatform graphical user interface for sequence alignment and phylogenetic tree building. *Mol Biol Evol.* 2010; 27:221–224. [PubMed: 19854763]
- Goydos JS, Patel M, Shih WC. NY-ESO-1 and CTp11 expression may correlate with stage of progression in melanoma. *J Surg Res.* 2001; 98:76–80. [PubMed: 11397121]
- Katoh K, Kuma K, Toh H, Miyata T. MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res.* 2005; 33:511–518. [PubMed: 15661851]
- Katoh K, Toh H. Parallelization of the MAFFT multiple sequence alignment program. *Bioinformatics.* 2010; 26:1899–1900. [PubMed: 20427515]
- Kibel AS, Faith DA, Bova GS, Isaacs WB. Xq27–28 deletions in prostate carcinoma. *Genes Chromosomes Can.* 2003; 37:381–388.
- Kidd JM, Cooper GM, Donahue WF, Hayden HS, Sampas N, Graves T, Hansen N, Teague B, Alkan C, Antonacci F, Haugen E, Zerr T, Yamada NA, Tsang P, Newman TL, Tuzun E, Cheng Z, Ebling HM, Tusneem N, David R, Gillett W, Phelps KA, Weaver M, Saranga D, Brand A, Tao W, Gustafson E, McKernan K, Chen L, Malig M, Smith JD, Korn JM, McCarroll SA, Altshuler DA, Peiffer DA, Dorschner M, Stamatoyannopoulos J, Schwartz D, Nickerson DA, Mullikin JC, Wilson RK, Bruhn L, Olson MV, Kaul R, Smith DR, Eichler EE. Mapping and sequencing of structural variation from eight human genomes. *Nature.* 2008; 453:56–64. [PubMed: 18451855]
- Kouprina N, Larionov V. Innovation - TAR cloning: insights into gene function, long-range haplotypes and genome structure and evolution. *Nat Rev Genet.* 2006; 7:805–812. [PubMed: 16983376]
- Kouprina N, Larionov V. Selective isolation of genomic loci from complex genomes by transformation-associated recombination cloning in the yeast *Saccharomyces cerevisiae*. *Nat Protoc.* 2008; 3:371–377. [PubMed: 18323808]
- Kouprina N, Mullokandov M, Rogozin IB, Collins NK, Solomon G, Otstot J, Risinger JI, Koonin EV, Barrett JC, Larionov V. The *SPANX* gene family of cancer/testis-specific antigens: Rapid evolution and amplification in African great apes and hominids. *Proc Natl Acad Sci USA.* 2004; 101:3077–3082. [PubMed: 14973187]
- Kouprina N, Noskov VN, Pavlicek A, Collins NK, Bortz PDS, Ottolenghi C, Loukinov D, Goldsmith P, Risinger JI, Kim JH, Westbrook VA, Solomon G, Sounders H, Herr JC, Jurka J, Lobanekov V,

- Schlessinger D, Larionov V. Evolutionary diversification of *SPANX-N*sperm Protein gene structure and expression. *PloS One*. 2007a; 2:E359. [PubMed: 17406683]
- Kouprina N, Noskov VN, Solomon G, Otstot J, Isaacs W, Xu J, Schleutker J, Larionov V. Mutational analysis of *SPANX* genes in families with X-linked prostate cancer. *Prostate*. 2007; 67:820–828. [PubMed: 17373721]
- Kouprina N, Pavlicek A, Noskov VN, Solomon G, Otstot J, Isaacs W, Carpten JD, Trent JM, Schleutker J, Barrett JC, Jurka J, Larionov V. Dynamic structure of the *SPANX* gene cluster mapped to the prostate cancer susceptibility locus *HPCX* at Xq27. *Genome Res*. 2005; 15:1477–1486. [PubMed: 16251457]
- Lange EM, Chen H, Brierley K, Perrone EE, Bock CH, Gillanders E, Ray ME, Cooney KA. Linkage analysis of 153 prostate cancer families over a 30-cM region containing the putative susceptibility locus *HPCX*. *Clin Cancer Res*. 1999; 5:4013–4020. [PubMed: 10632333]
- Lucas S, De Smet C, Arden KC, Viars CS, Lethe B, Lurquin C, Boon T. Identification of a new *MAGE* gene with tumor-specific expression by representational difference analysis. *Cancer Res*. 1998; 58:743–752. [PubMed: 9485030]
- Lutke-Holzlik MF, Rapley EA, Hoekstra HJ, Sleijfer DT, Nolte IM, Sijmons RH. Genetic predisposition to testicular germ-cell tumours. *Lancet Oncol*. 2004; 5:363–371. [PubMed: 15172357]
- Maher BH, Kerr M, Cox HC, MacMillan JC, Brimage PJ, Esposito T, Gianfrancesco F, Haupt LM, Nyholt DR, Lea RA, Griffiths LR. Confirmation that Xq27 and Xq28 are susceptibility loci for migraine in independent pedigrees and a case-control cohort. *Neurogenetics*. 2012; 13:97–101. [PubMed: 22294494]
- Mattila H, Schindler M, Isotalo J, Ikonen T, Vihinen M, Oja H, Tammela TLJ, Wahlfors T, Schleutker J. NMD and microRNA expression profiling of the *HPCX1* locus reveal *MAGEC1* as a candidate prostate cancer predisposition gene. *BMC Cancer*. 2011; 11:327. [PubMed: 21810217]
- Mefford HC, Eichler EE. Duplication hotspots, rare genomic disorders, and common disease. *Curr Opin Genet Dev*. 2009; 19:196–204. [PubMed: 19477115]
- Morganti G, Gianferrari L, Cresseri A. Recherches clinico-statistiques et genetiques sur les neoplasies de la prostate. *Acta Genet*. 1956; 6:304–305. [PubMed: 13410498]
- Morgenstern B. DIALIGN: multiple DNA and protein sequence alignment at BiBiServ. *Nucleic Acids Res*. 2004; 32(Web Server issue):W33–W36. [PubMed: 15215344]
- Nagasaki K, Schem C, von Kaisenberg C, Biallek M, Rosel F, Jonat W, Maass N. Leucine-zipper protein, *LDOC1*, inhibits NF-kappaB activation and sensitizes pancreatic cancer cells to apoptosis. *Int J Cancer*. 2003; 105:454–458. [PubMed: 12712434]
- Neuhausen SL, Farnham JM, Kort E, Tavtigian SV, Skolnick MH, Cannon-Albright LA. Prostate cancer susceptibility locus *HPC1* in Utah high-risk pedigrees. *Hum Mol Genet*. 1999; 8:2437–2442. [PubMed: 10556291]
- Nicholas KB, Nicholas HB Jr, Deerfield DW II. GeneDoc: Analysis and visualization of genetic variation. *EMBNEW.NEWS*. 1997; 4:14.
- Prensner JR, Iyer MK, Balbin OA, Dhanasekaran SM, Cao Q, Brenner JC, Laxman B, Asangani IA, Grasso CS, Kominsky HD, Cao XH, Jing XJ, Wang XJ, Siddiqui J, Wei JT, Robinson D, Iyer HK, Palanisamy N, Maher CA, Chinnaiyan AM. Transcriptome sequencing across a prostate cancer cohort identifies *PCAT-1*, an unannotated lincRNA implicated in disease progression. *Nat Biotechnol*. 2011; 29:742–749. [PubMed: 21804560]
- Ren XP, Wu JH, Wang XH, Sartor MA, Qian J, Jones K, Nicolaou P, Pritchard TJ, Fan GC. MicroRNA-320 Is Involved in the Regulation of Cardiac Ischemia/Reperfusion Injury by Targeting Heat-Shock Protein 20. *Circulation*. 2009; 119:2357–2366. [PubMed: 19380620]
- Rettig WJ, Chen YT, Yenamandra AK, Chaganti RSK, Dropcho EJ, Posner JB, Old LJ. Assignment of the cerebellar degeneration-related autoantigen CDR to human chromosome, X. Abstract. *Cytogenet. Cell Genet*. 1987; 46:681.
- Rice P, Longden I, Bleasby A. EMBOSS: The European molecular biology open software suite. *Trends Genet*. 2000; 16:276–277. [PubMed: 10827456]
- Robinson JT, Thorvaldsdottir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP. Integrative genomics viewer. *Nat Biotechnol*. 2011; 29:24–26. [PubMed: 21221095]

- Rubin MA, De Marzo AM. Molecular genetics of human prostate cancer. *Mod Pathol*. 2004; 17:380–388. [PubMed: 14752525]
- Schaid DJ. The complex genetic epidemiology of prostate cancer. *Hum Mol Genet* 13 Spec No. 2004; 1:103–121.
- Schleutker J, Matikainen M, Smith J, Koivisto P, Baffoe-Bonnie A, Kainu T, Gillanders E, Sankila R, Pukkala E, Carpten J, Stephan D, Tammela T, Brownstein M, Bailey-Wilson J, Trent J, Kallioniemi OP. A genetic epidemiological study of hereditary prostate cancer. HPC in Finland: frequent HPCX linkage in families with late-onset disease. *Clin Cancer Res*. 2000; 6:4810–4815. [PubMed: 11156239]
- Sebat J, Lakshmi B, Troge J, Alexander J, Young J, Lundin P, Maner S, Massa H, Walker M, Chi MY, Navin N, Lucito R, Healy J, Hicks J, Ye K, Reiner A, Gilliam TC, Trask B, Patterson N, Zetterberg A, Wigler M. Large-scale copy number polymorphism in the human genome. *Science*. 2004; 305:525–528. [PubMed: 15273396]
- Shaikh TH, Gai XW, Perin JC, Glessner JT, Xie HB, Murphy K, O'Hara R, Casalunovo T, Conlin LK, D'Arcy M, Frackelton EC, Geiger EA, Haldeman-Englert C, Imielinski M, Kim CE, Medne L, Annaiah K, Bradfield JP, Dabaghyan E, Eckert A, Onyiah CC, Ostapenko S, Otieno FG, Santa E, Shaner JL, Skraban R, Smith RM, Elia J, Goldmuntz E, Spinner NB, Zackai EH, Chiavacci RM, Grundmeier R, Rappaport EF, Grant SFA, White PS, Hakonarson H. High-resolution mapping and analysis of copy number variations in the human genome: A data resource for clinical and research applications. *Genome Res*. 2009; 19:1682–1690. [PubMed: 19592680]
- Shoemaker RH. The NCI60 human tumour cell line anticancer drug screen. *Nat Rev Cancer*. 2009; 6:813–823. [PubMed: 16990858]
- Slavotinek A, Lee SS, Hamilton SP. A family with X-linked anophthalmia: Exclusion of *SOX3* as a candidate gene. *Am J Med Genet*. 2005; 138A:89–94. [PubMed: 16114045]
- Solomon NM, Ross SA, Morgan T, Belsky JL, Hol FA, Karnes PS, Hopwood NJ, Myers SE, Tan AS, Warne GL, Forrest SM, Thomas PQ. Array comparative genomic hybridisation analysis of boys with X linked hypopituitarism identifies a 3.9 Mb duplicated critical region at Xq27 containing *SOX3*. *J Med Genet*. 2004; 41:669–678. [PubMed: 15342697]
- Stankiewicz P, Lupski JR. The genomic basis of disease, mechanisms and assays for genomic disorders. *Genome Dyn*. 2006; 1:1–16. [PubMed: 18724050]
- Stankiewicz P, Lupski JR. Structural variation in the human genome and its role in disease. *Annu Rev Med*. 2010; 61:437–455. [PubMed: 20059347]
- Stephan DA, Howell GR, Teslovich TM, Coffey AJ, Smith L, Bailey-Wilson JE, Malechek L, Gildea D, Smith JR, Gillanders EM, Schleutker J, Hu P, Steingruber HE, Dhami P, Robbins CM, Makalowska I, Carpten JD, Sood R, Mumm S, Reinbold R, Bonner TI, Baffoe-Bonnie A, Bubendorf L, Heiskanen M, Kallioniemi OP, Baxevas AD, Joseph SS, Zucchi I, Burk RD, Isaacs W, Ross MT, Trent JM. Physical and transcript map of the hereditary prostate cancer region at Xq27. *Genomics*. 2002; 79:41–50. [PubMed: 11827456]
- Van Wynsberghe PM, Chan SP, Slack FJ, Pasquinelli AE. Analysis of microRNA expression and function. *Methods Cell Biol*. 2011; 106:219–252. [PubMed: 22118279]
- Varghese JS, Easton DF. Genome-wide association studies in common cancers--what have we learnt? *Curr Opin Genet Dev*. 2010; 20:201–209. [PubMed: 20418093]
- Verhage BA, Kiemeny LA. Genetic susceptibility to prostate cancer: a review. *Fam Cancer*. 2003; 2:57–67. [PubMed: 14574167]
- Vincent JB, Melmer G, Bolton PF, Hodgkinson S, Holmes D, Curtis D, Gurling HM. Genetic linkage analysis of the X chromosome in autism, with emphasis on the fragile X region. *Psychiatr Genet*. 2005; 15:83–90. [PubMed: 15900222]
- Vitek WS, Pagidas K, Gu G, Pepperell JR, Simpson JL, Tantravahi U, Plante BJ. Xq;autosome translocation in POF: Xq27.2 deletion resulting in haploinsufficiency for *SPANX*. *J Assist Reprod Genet*. 2012; 29:63–66. [PubMed: 22071885]
- Wang KC, Chang HY. Molecular mechanisms of long noncoding RNAs. *Mol Cell*. 2011; 43:904–914. [PubMed: 21925379]

- Wang ZQ, Zhang YN, Liu HC, Salati E, Chiriva-Internati M, Lim SH. Gene expression and immunologic consequence of *SPAN-Xb* in myeloma and other hematologic malignancies. *Blood*. 2003; 101:955–960. [PubMed: 12393489]
- Westbrook VA, Diekman AB, Klotz KL, Khole VV, von Kap-Herr C, Golden WL, Eddy RL, Shows TB, Stoler MH, Lee CYG, Flickinger CJ, Herr JC. Spermatid-specific expression of the novel X-linked gene product SPAN-X localized to the nucleus of human spermatozoa. *Biol Reprod*. 2000; 63:469–481. [PubMed: 10906052]
- Westbrook VA, Diekman AB, Naaby-Hansen S, Coonrod SA, Klotz KL, Thomas TS, Norton EJ, Flickinger CJ, Herr JC. Differential nuclear localization of the cancer/testis-associated protein, SPAN-X/CTp11, in transfected cells and in 50% of human spermatozoa. *Biol Reprod*. 2001; 64:345–358. [PubMed: 11133693]
- Westbrook VA, Schoppee PD, Diekman AB, Klotz KL, Allietta M, Hogan KT, Slingluff CL, Patterson JW, Frierson HF, Irvin WP, Flickinger CJ, Coppola MA, Herr JC. Genomic organization, incidence, and localization of the SPAN-X family of cancer-testis antigens in melanoma tumors and cell lines. *Clin Cancer Res*. 2004; 10:101–112. [PubMed: 14734458]
- Xu J, Gillanders EM, Isaacs SD, Chang BL, Wiley KE, Zheng SL, Jones M, Gildea D, Riedesel E, Albertus J, Freas-Lutz D, Markey C, Meyers DA, Walsh PC, Trent JM, Isaacs WB. Genome-wide scan for prostate cancer susceptibility genes in the Johns Hopkins hereditary prostate cancer families. *Prostate*. 2003; 57:320–325. [PubMed: 14601028]
- Xu J, Meyers D, Freije D, Isaacs S, Wiley K, Nusskern D, Ewing C, Wilkens E, Bujnovszky P, Bova GS, Walsh P, Isaacs W, Schleutker J, Matikainen M, Tammela T, Visakorpi T, Kallioniemi OP, Berry R, Schaid D, French A, McDonnell S, Schroeder J, Blute M, Thibodeau S, Gronberg H, Emanuelsson M, Damber JE, Bergh A, Jonsson BA, Smith J, Bailey-Wilson J, Carpten J, Stephan D, Gillanders E, Amundson I, Kainu T, Freas-Lutz D, Baffoe-Bonnie A, Van Aucken A, Sood R, Collins F, Brownstein M, Trent J. Evidence for a prostate cancer susceptibility locus on the X chromosome. *Nat Genet*. 1998; 20:175–179. [PubMed: 9771711]
- Yaspan BL, McReynolds KM, Elmore JB, Breyer JP, Bradley KM, Smith JR. A haplotype at chromosome Xq27.2 confers susceptibility to prostate cancer. *Hum Genet*. 2008; 123:379–386. [PubMed: 18350320]
- Zendman AJW, Cornelissen IMHA, Weidle UH, Ruiter DJ, van Muijen GNP. CTp11, a novel member of the family of human cancer/testis antigens. *Cancer Res*. 1999; 59:6223–6229. [PubMed: 10626816]
- Zendman AJ, Zschocke J, van Kraats AA, de Wit NJ, Kurpisz M, Weidle UH, Ruiter DJ, Weiss EH, van Muijen GN. The human *SPANX* multigene family: genomic organization, alignment and expression in male germ cells and tumor cell lines. *Gene*. 2003; 309:125–133. [PubMed: 12758128]
- Zhu HW, Shang DD, Sun M, Choi SJ, Liu Q, Hao JJ, Figuera LE, Zhang F, Choy KW, Ao Y, Liu Y, Zhang XL, Yue FZ, Wang MR, Jin L, Patel PI, Jing T, Zhang X. X-Linked Congenital Hypertrichosis Syndrome Is Associated with Interchromosomal Insertions Mediated by a Human-Specific Palindrome near *SOX3*. *Am J Hum Genet*. 2011; 88:819–826. [PubMed: 21636067]

Figure 1.

Microarray comparative genomic hybridization. Three patients (001-002, 236-006, 232-001) from the families with the strongest linkage to Xq27-q28 and two healthy controls (KS and K1) were chosen for analysis. Positions correspond to 138.5-145.0 Mb in NCBI human genome build 36.

Figure 2.

Selective isolation of the TAR10 genomic region by transformation-associated recombination in yeast (TAR cloning). **(a)** To clone the genomic fragment, the TAR10 circularizing vector, which contains 5' and 3' sequences of the targeted region on chromosome X, was constructed using the basic vector pVC604 which contains an yeast centromeric locus (*CEN6*) and an yeast selectable marker (*HIS3*). Hook 1 and hook 2 correspond to 196-bp XhoI-BamHI and 177-bp BamHI-XbaI fragments of the 5' and 3' targeted genomic regions. The hook sequences were designed based on the available information (hg18) and correspond to positions 140,180,833-140,181,028 and

140,293,797-140,293,973 on the chromosome X sequence. Size of the targeted genomic fragment is ~113.1 kb. The TAR10 vector was linearized with BamHI (the site is located between the hooks) before transformation to yield a molecule bounded by the targeting region sequences. Recombination between the hooks in the vector and the homologous sequences of the fragment of interest results in isolation of the desired fragment as a circular DNA molecule capable of stable propagation in yeast cells. Arrows indicate the orientation of the targeting sequences in the TAR vector relative to the genome sequence. **(b)** To identify TAR10-positive clones, transformants were combined into pools and examined by PCR with diagnostic primers for the sequences that are unique to the desired insert. Red arrows indicate positive pools (2, 9, 16, 20). The PCR products were sequenced and found to match the expected TAR10 sequences. Individual clones containing the TAR10 region were found in each TAR10-positive pool. **(c)** The integrity of the genomic insert was confirmed by PCR using additional pairs of diagnostic primers.

Figure 3.

Analysis of genomic rearrangements within the *HPCX1* region by TAR cloning and PCR.

(a) A schematic representation of the genomic fragments isolated by TAR cloning and checked by PCR analysis. The entire segment is drawn to scale. The scheme shows the positions of *SPANX*, *LDOC1* and *CDR1* genes (red). (b) A scheme illustrating recombinational interaction between two largest segmental duplications (SD-B 119kb in size and SD-A1/A2 113 kb in size) potentially leading to inversion. Blue dotted arrows indicate possible breakpoints. The green line in (a) and (b) corresponds to the continuous region (~581 kb) analyzed by TAR cloning and PCR. (c) A physical characterization of the TAR

isolates by Southern-blot hybridization. TAR3 (81 kb): lanes 1–4 correspond to affected brothers (232-001, 236-006, 001-002) (lanes 3 and 4 are two independent TAR clones for 001-002); lanes 5-8 correspond to unaffected controls (232-005, 236-003, 001-003) (lanes 7 and 8 are two independent TAR clones for 001-003). TAR4 (67 kb): lanes 1–3 correspond to three patients (232-001, 236-006, 001-002); lanes 4–6 correspond to three unaffected brothers (232-005, 236-003, 001-003). TAR5 (63 kb): lanes 1–3 correspond to the 001-002 patient (three independent TAR clones containing one copy of a 12-kb repeat). TAR7 (84 kb): lanes 1–3 correspond to three patients (232-001, 236-006, 001-002); lanes 4–6 correspond to three unaffected brothers (232-005, 236-003, 001-003). TAR8 (42 kb): lanes 1–2 correspond to two patients (232-001, 236-006); lanes 3–4 correspond to unaffected controls (232-005, 236-003).

Figure 4.

Analysis of the transcriptional pattern in the critical region. **(a)** Transcriptional analysis of NCI-60 cells. Correlation between neighboring pairs of probes plotted against the midpoint location between the probes in each pair. The correlation cutoff corresponding to a False Discovery Rate (FDR) of 0.1 is plotted a horizontal line. Correlations with FDR < 0.05 are marked by red points. **(b)** Analysis of transcriptome sequencing in prostate samples. The BAM files from 21 prostate cancer cell lines deposited by Prensner et al. (2011) in the Gene Expression Omnibus (GSE25183) were merged and visualized in the Integrative Genomics Viewer (Robinson et al. 2011). The region on the 5' end of *CDR1* (a blue box at the bottom)

in approximate positions chrX:139,674,300-139,688,550 (shown as a red line at the bottom) contains many RNA-Seq reads (gray vertical bars). The region also overlaps with the 3' end of two ESTs, BM811380 and DA247186 (a green line), and could represent a novel gene at Xq26-q27.2.

Figure 5.

Analysis of the novel genes located at Xq27. **(a)** Location of all genes identified at the Xq27 region is shown. **(b)** A structure of the *SPANXL* gene. Exons 1 and 2 are marked in green; intron is marked in orange. I (iso-leucine) in position 83 of *SPANXL* is replaced by V (valine) in some samples. **(c)** RT-PCR analysis of the *SPANXL* gene. Oligonucleotides were designed within exons 1 and 2 to amplify a putative transcript. Using Dog-F2/Dog-R2 primers (supplementary Table S4) specific to the gene, the *SPANXL* expression was analyzed in a panel of normal tissues. A 489-bp band of expected size was detected in testis and kidney. Lane 1-Kidney; lane 2-Testicle; lane 3-Cervix, lane 4-Small Intestine; lane 5-

Prostate. Blue arrow indicates to the 489-bp bands that were isolated from the gel and sequenced. M-GeneRuler™DNA Ladder Mix. **(d)** Multiple alignment of SPANXL with SPANXA/D and SPANXN families. The plot highlights color-coded physiochemical properties. The conserved positions are shown as a consensus at the bottom (uppercase means identity in all proteins; lowercase means high conservation). **(e)** Hypothetical human microRNA homologous to mouse Mir-320. **(f)** RT-PCR analysis of a novel long non-coding RNA (lncRNA) covered by two ESTs, BM811380 and DA247186. Using two pairs of specific primers (EST-F1/ EST-R33 and EST-F2/ EST-R44), the long transcript was found to be spliced and detectable in cervix and testis. Lanes 1, 3 - Cervix; lanes 2, 4 - Testis. The 260-bp and 489-bp bands were isolated from the gel and sequenced. M - GeneRuler™DNA Ladder Mix.

Table 1

Finnish families used in this study

Patients	X-linked	Healthy brothers	
232-002/232-001	the strongest	HLOD > 0.4	232-005/232-004
236-005/236-006	the strongest	HLOD > 0.4	236-003
001-004/001-002	the strongest	HLOD > 0.4	001-003/001-001
015-001	the strongest	HLOD > 0.4	015-006
102-001	the strongest	HLOD > 0.4	
037-002/037-001	suggestive	HLOD < 0.4	
051-002	suggestive	HLOD < 0.4	
084-002/084-001	suggestive	HLOD < 0.4	084-003
241-003/241-002	suggestive	HLOD < 0.4	241-009
145-004	suggestive	HLOD < 0.4	145-001
028-001	suggestive	HLOD < 0.4	028-002/028-003
021-002	suggestive	HLOD < 0.4	021-003
045-003/045-001	suggestive	HLOD < 0.4	
125-006	suggestive	HLOD < 0.4	
101-101	suggestive	HLOD < 0.4	
113-004	suggestive	HLOD < 0.4	
064-006	suggestive	HLOD < 0.4	
420-001/420-002	NMM *		
248-006	NMM		248-001
311-003	NMM		
138-007	NMM		
083-002	NMM		
043-001	NMM		043-002
257-003/257-002	NMM		

* NMM - no-male-to-male inheritance

Table 2

Positions of features analyzed at the Xq27 region

Name	Positions in genome (hg18)	Size
Inversion1 (in GenBank)	139,874,095 – 140,538,795	-
Inversion 2 (in GenBank)	140,516,337 – 140,551,160	
AT-rich site 1	139,845,431 – 139,845,684	254 bp
AT-rich site 2	139,839,878 – 139,839,949	72 bp
<i>SPANXL</i> gene	140,541,566 – 140,542,646	1080 bp
<i>CDRI</i> gene	139,693,029 – 139,694,396	1368 bp
has-mir-320d-2	139,836,003 – 139,836,050	48 bp
hypothetical lncRNA	139,619,611 – 139,682,508	62,898 bp

Table 3

Positions of genomic regions covered by overlapping PCRs and TAR cloning

PCR/TAR	Positions (hg18)	Expected size
PCR1	139,816,257 – 139,900,471	88 kb
PCR2	140,044,723 – 140,097,869	53 kb
PCR3	140,181,947 – 140,267,284	85 kb
TAR1	139,713,026 – 139,771,894	59 kb
TAR2	139,771,633 – 139,837,585	66 kb
TAR3	139,835,037 – 139,915,604	81 kb
TAR4	139,835,037 – 139,904,731	67 kb
TAR5	139,885,566 – 139,948,402	63 kb *
TAR6	139,904,589 – 140,045,002	140 kb
TAR7	140,097,851 – 140,181,147	84 kb
TAR8	140,492,674 – 140,534,123	42 kb
TAR10	140,180,833 – 140,293,973	113 kb

* The size of the genomic insert in the case when a 12-kb repeat is present as one copy.

Table 4

Mutational analysis of the *CDRI* gene in Finnish families

Families	Healthy/Affected	Residue	Alleles	Residue change	X-linked
232-					the strongest
232-001	Affected	224	C/T	R to C	
232-004	Healthy	224	C/T	R to C	
232-005	Healthy	no mutations			
001-					the strongest
001-002	Affected	224	C/T	R to C	
001-003	Healthy	no mutations			
236-					the strongest
236-002	Affected	224	C/T	R to C	
236-005	Affected	no mutations			
236-006	Affected	no mutations			
236-003	Healthy	no mutations			
102-					the strongest
102-001	Affected	no mutations			
037-					suggestive
037-002	Affected	no mutations			
037-001	Affected	no mutations			
084-					suggestive
084-002	Affected	no mutations			
084-003	Healthy	no mutations			
241-					suggestive
241-003	Affected	no mutations			
241-009	Healthy	no mutations			
045-					suggestive
045-001	Affected	no mutations			
045-003	Affected	no mutations			
145-					suggestive
145-004	Affected	no mutations			
145-001	Healthy	no mutations			

Families	Healthy/Affected	Residue	Alleles	Residue change	X-linked
028-					suggestive
028-001	Affected	no mutations			
028-003	Healthy	no mutations			
028-002	Healthy	no mutations			
125-006	Affected	no mutations			suggestive
101-					suggestive
101-101	Affected	no mutations			
101-004	Healthy	no mutations			
113-004	Affected	no mutations			suggestive
064-006	Affected	no mutations			suggestive
420-					NMM
420-002	Affected	no mutations			
420-001	Affected	no mutations			
257-					NMM
257-002	Affected	no mutations			
257-003	Affected	no mutations			
248-					NMM
248-006	Affected	no mutations			
248-001	Healthy	no mutations			
311-003	Affected	no mutations			NMM
138-007	Affected	no mutations			NMM
083-002	Affected	no mutations			NMM
043-					NMM
043-001	Affected	no mutations			
043-002	Healthy	no mutations			
242-002	Affected	no mutations			