SUPPLEMENTARY INFORMATION

USP2a-mediated modulation of microRNAs
in prostate cancer enhances c-Myc oncogenic activity

Barbara Benassi, Richard Flavin, Luigi Marchionni, Silvio Zanata, Yunfeng Pan, Dipanjan Chowdhury, Marina Marani, Sabrina Strano, Paola Muti, Giovanni Blandino, Massimo Loda

The Supplementary information provided in this document accounts for:

- **Supplementary Data:** 12 Figures consecutively numbered from 1 to 12 and 11 Tables consecutively numbered from 1 to 11.

- **Supplementary Methods:** the detailed description of Affymetrix gene expression analysis and evaluation of the prostate mRNA signature in distinct cancer invasion models, including data sources, annotation, pre-processing, and statistical analysis.

- **Supplementary References:** the list of literature citations unique to the Supplementary Information.

Raw gene expression data and all required MIAME (Brazma et al., 2001) information is permanently hosted in the NCBI Gene Expression Omnibus (GEO) database (Wheeler et al., 2007) with the following series accession **GSE17466** at:

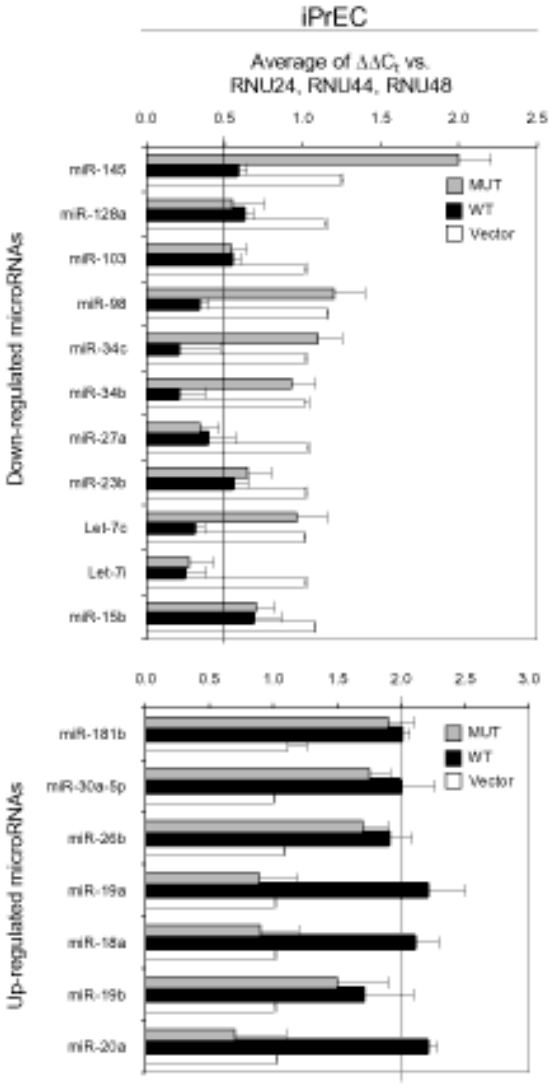
http://www.ncbi.nlm.nih.gov/geo/.

Complete Supplementary Figures and Tables for microarray pre-processing, differentially expressed genes and enriched functional gene sets, linked to external genomic annotation databases are available at:

http://astor.som.jhmi.edu/~marchion/labML/benassi.html

SUPPLEMENTARY DATA

Supplementary Fig. 1

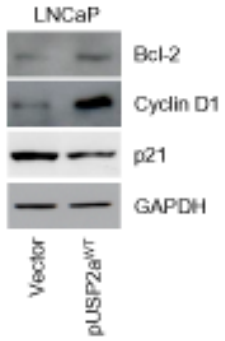


Supplementary Fig. 2

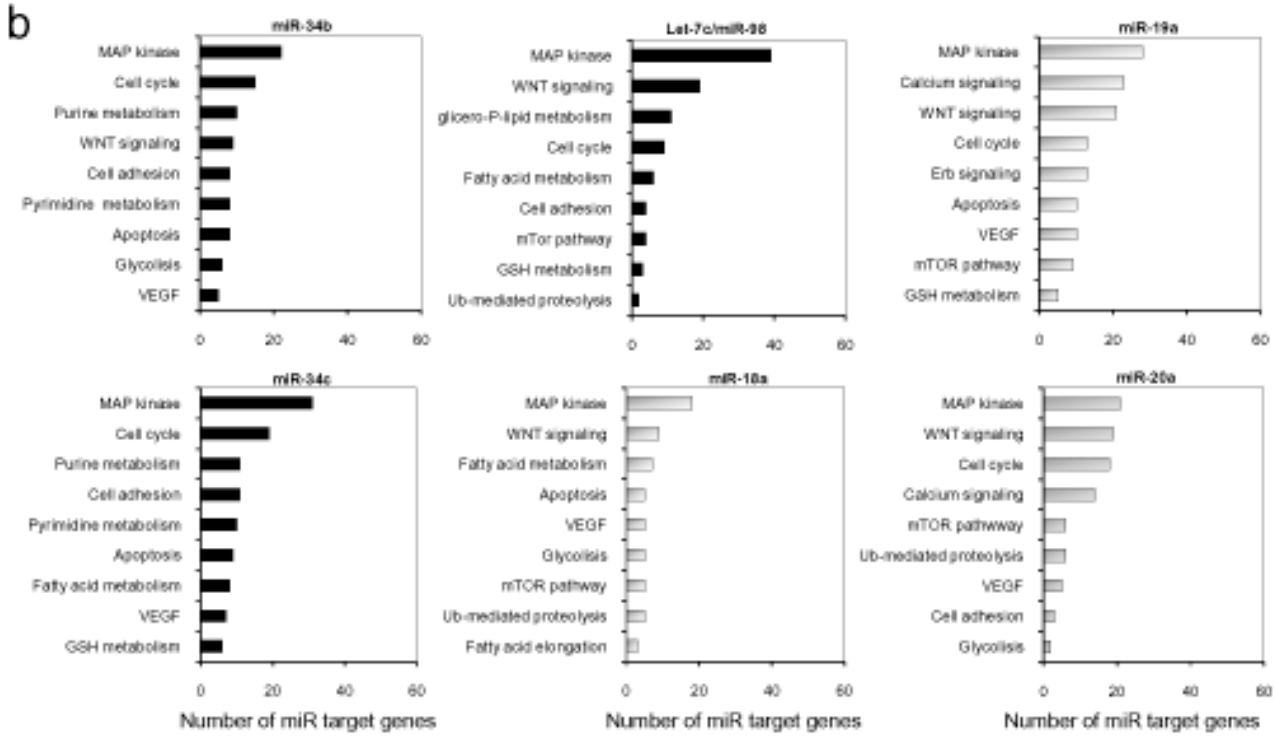
a

microRNAs and relative predicted target genes in wtUSP2a-overexpressing cells

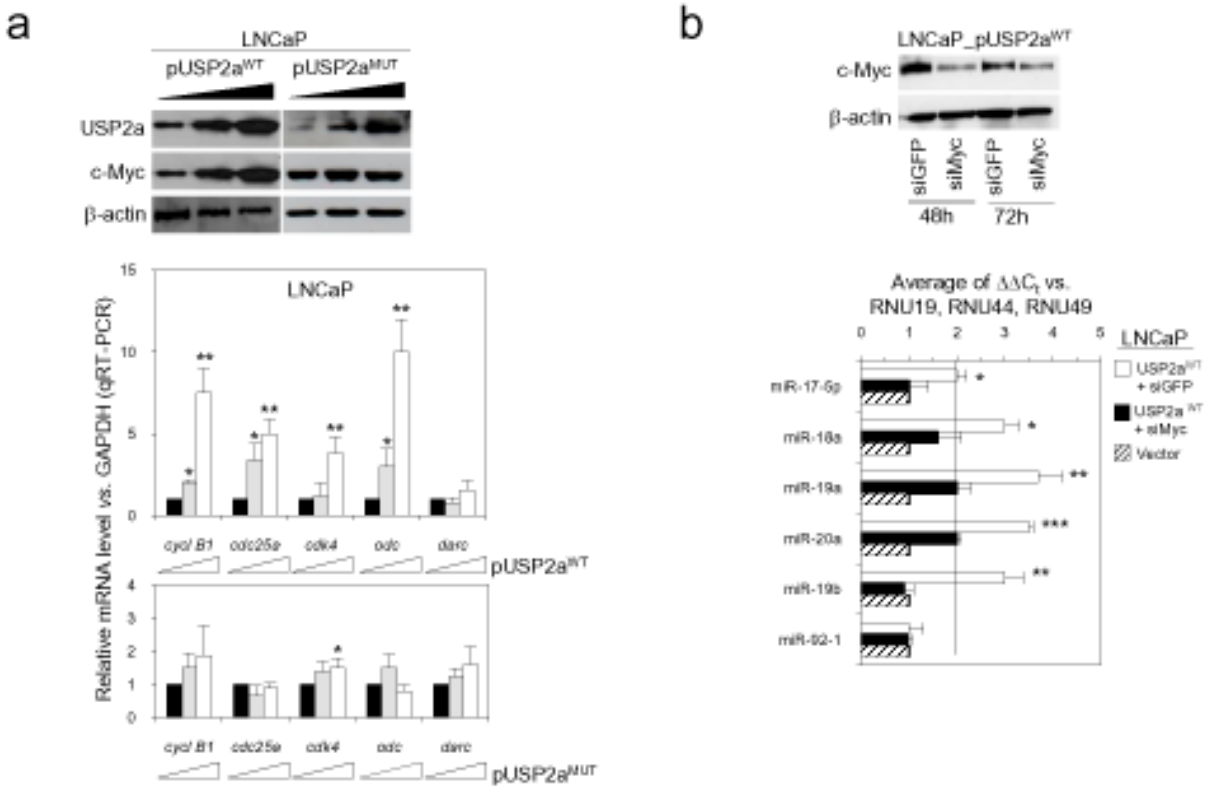
Down-regulated	*Predicted increased expression*
hsa-miR-34b	c-myb c-myc met
hsa-miR-34c	bcl-2 cdc25a cyclin D1 c-myc met ras VEGF wnt-1
hsa-let-7c, hsa-miR-98	cdc25a cyclin D1 c-myc CROP ras vav3 wnt-1
Up-regulated	*Predicted decreased expression*
hsa-miR-18a	Runx1
hsa-miR-19a	PTEN
hsa-miR-20a	CDKN1A (p21) Rb1



b

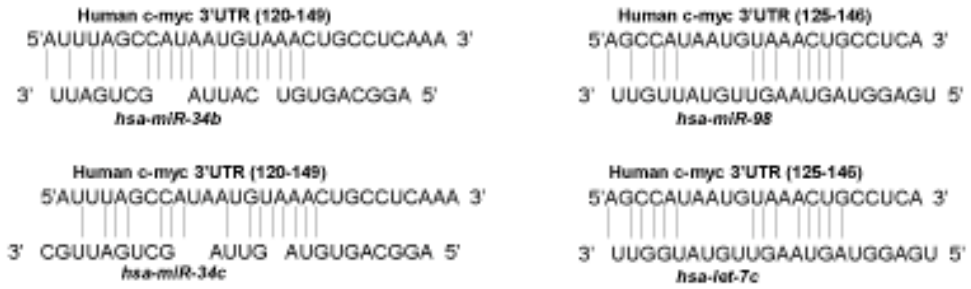


Supplementary Fig. 3

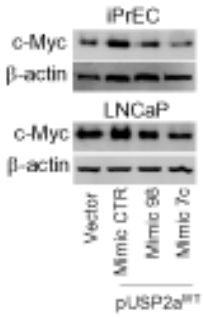


Supplementary Fig. 4

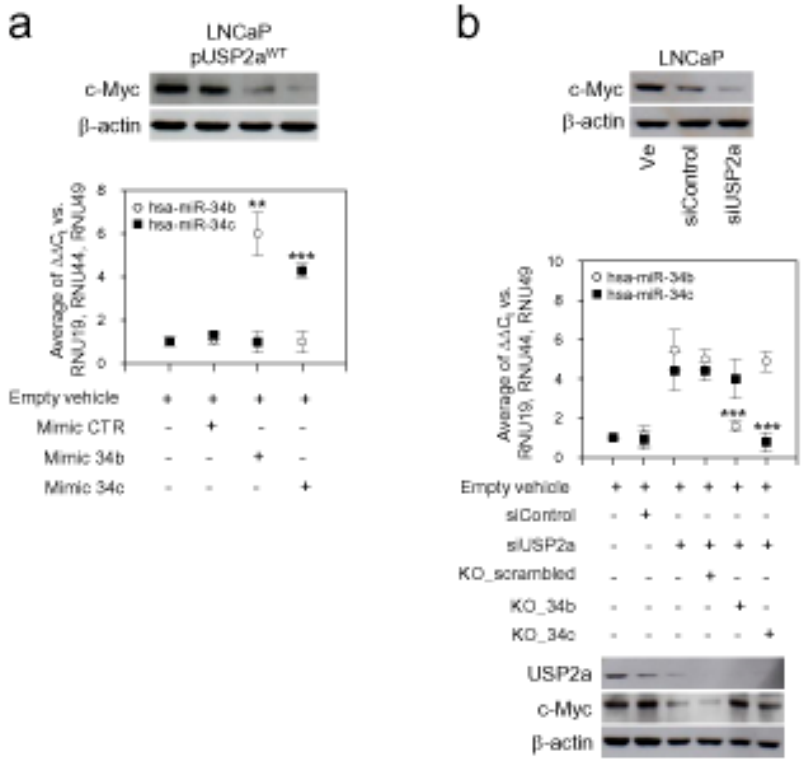
a



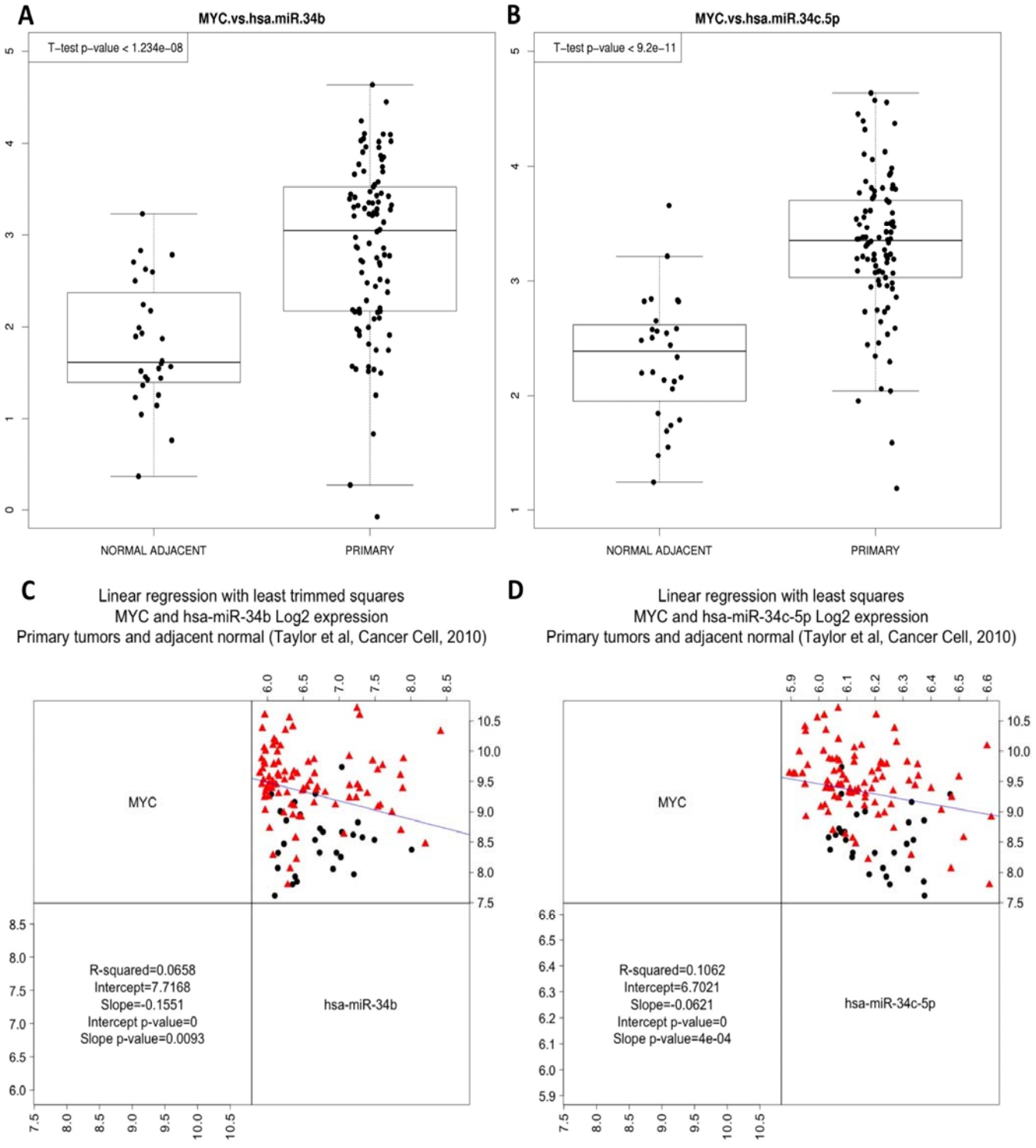
b



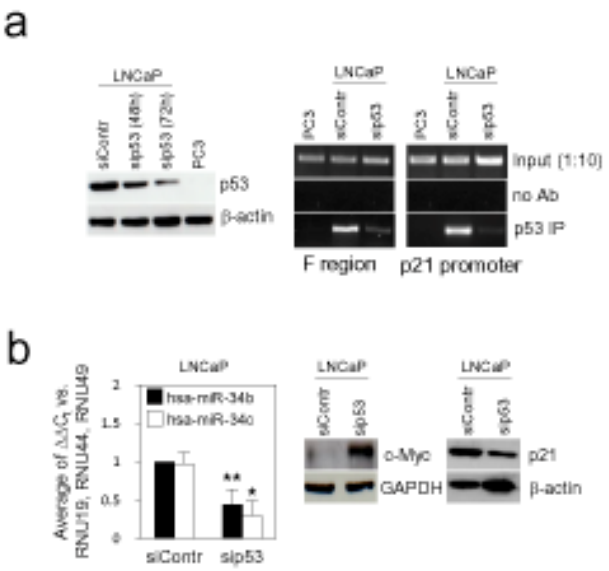
Supplementary Fig. 5



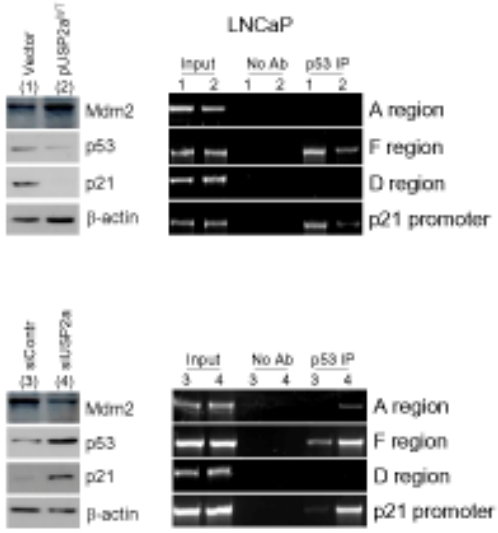
Supplementary Fig. 6



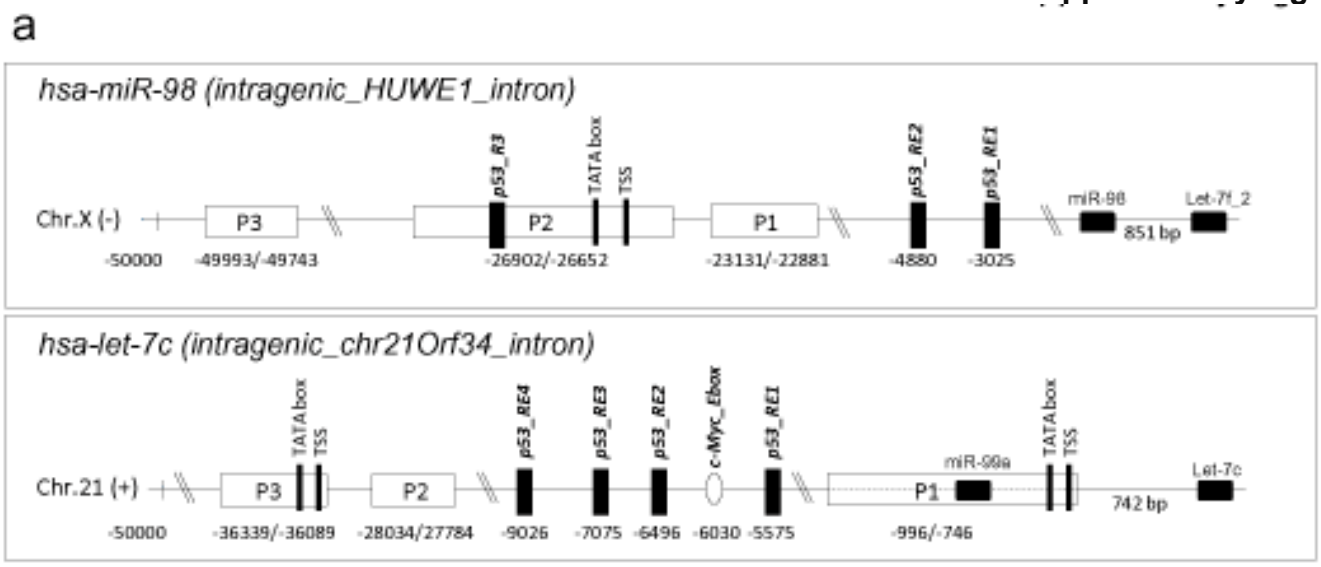
Supplementary Fig. 7

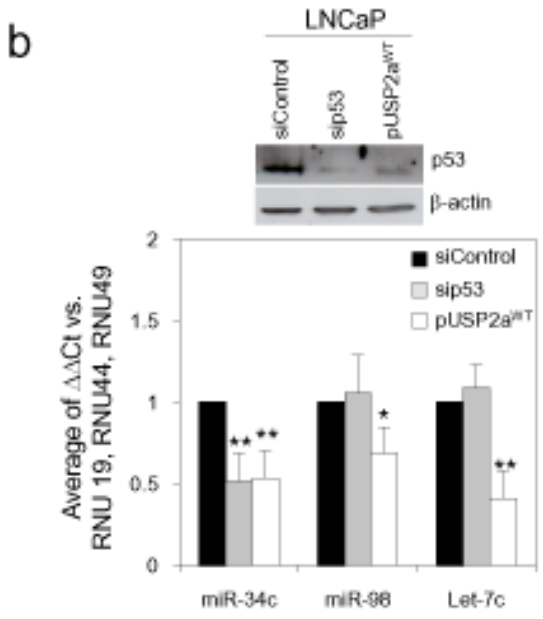


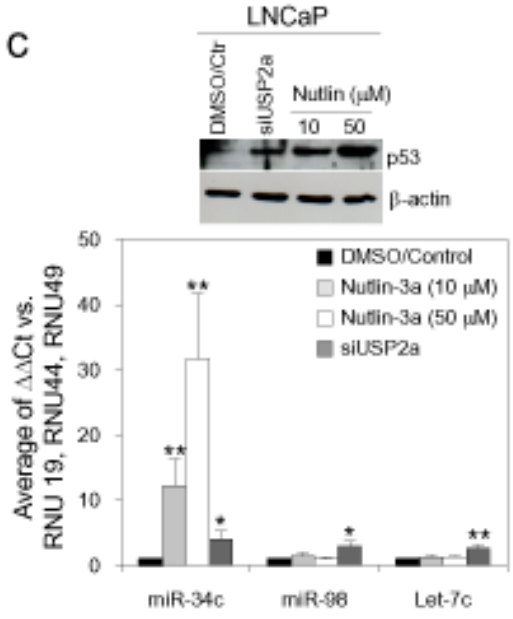
Supplementary Fig. 8

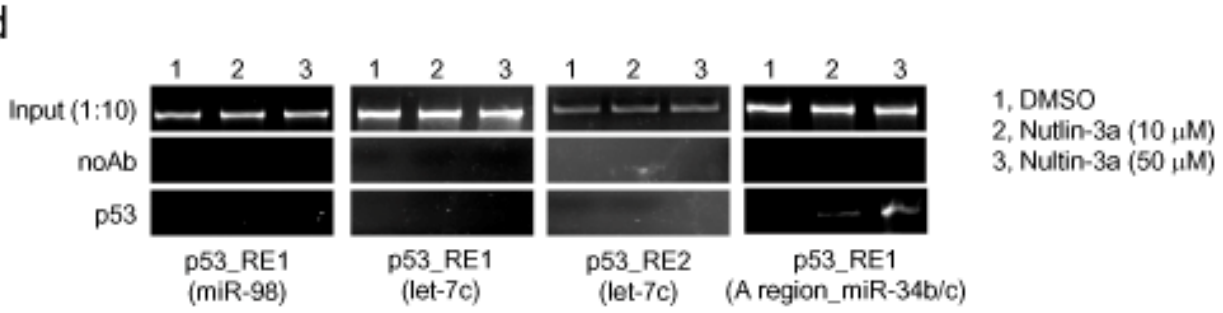


Supplementary Fig. 9

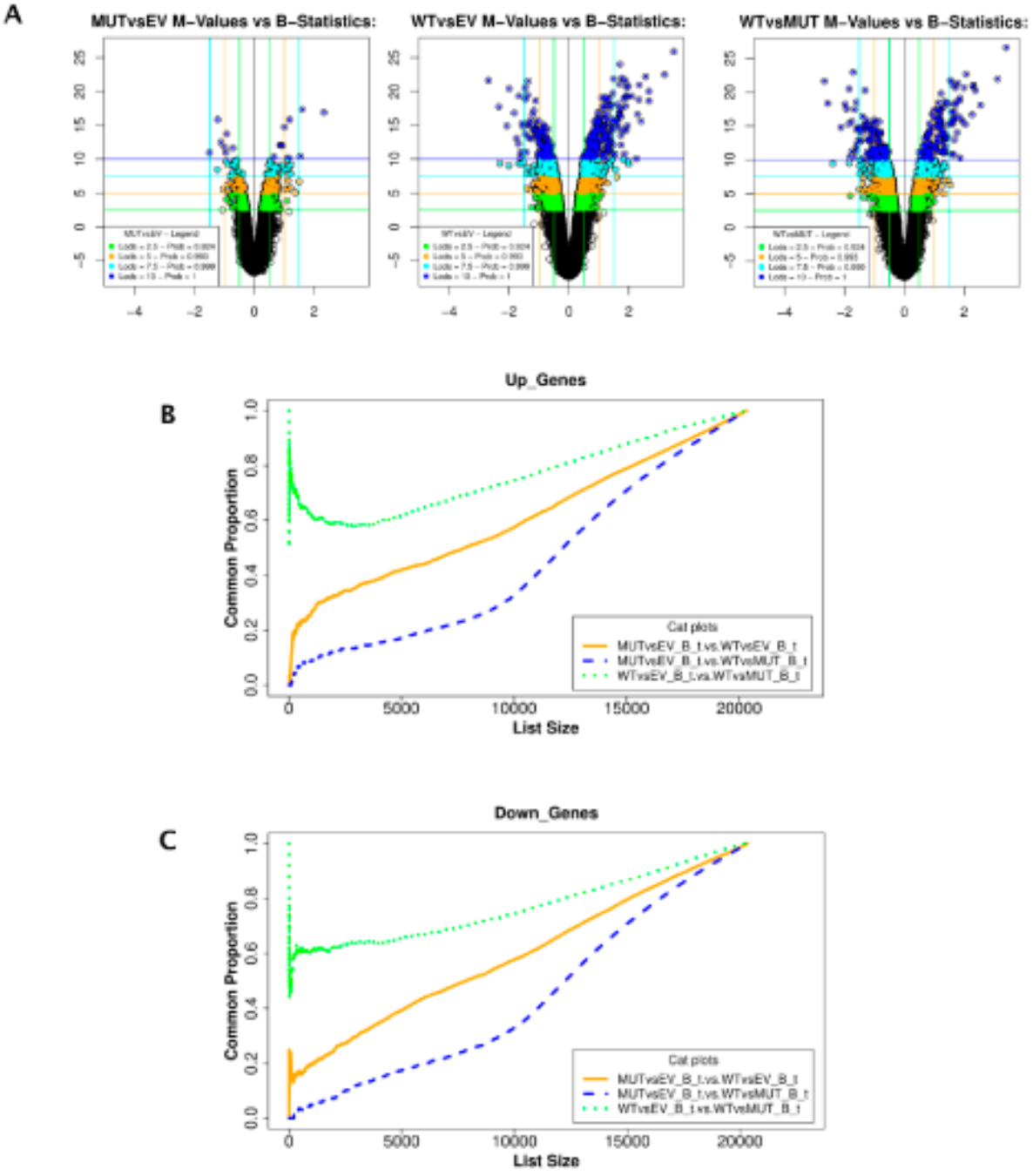




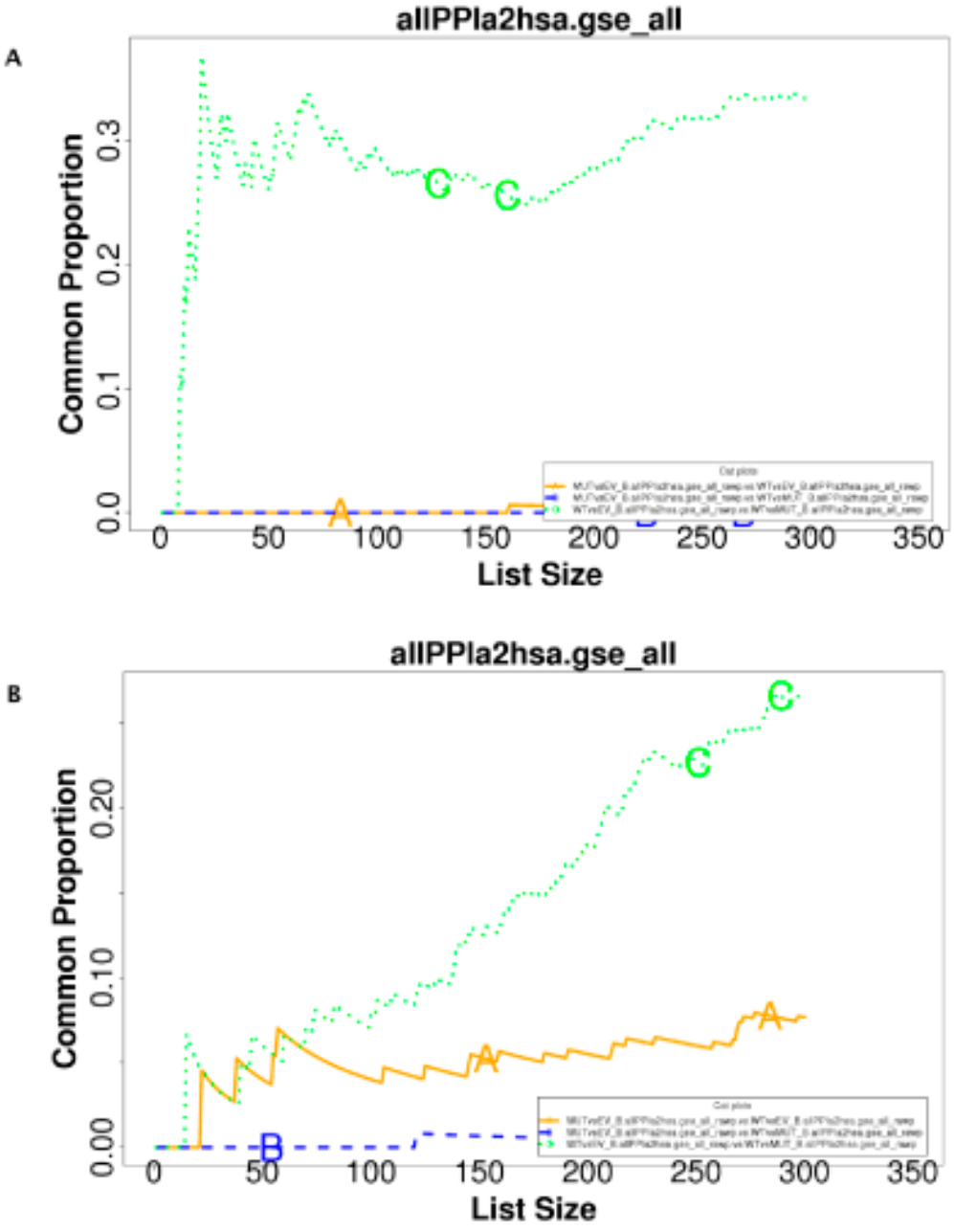




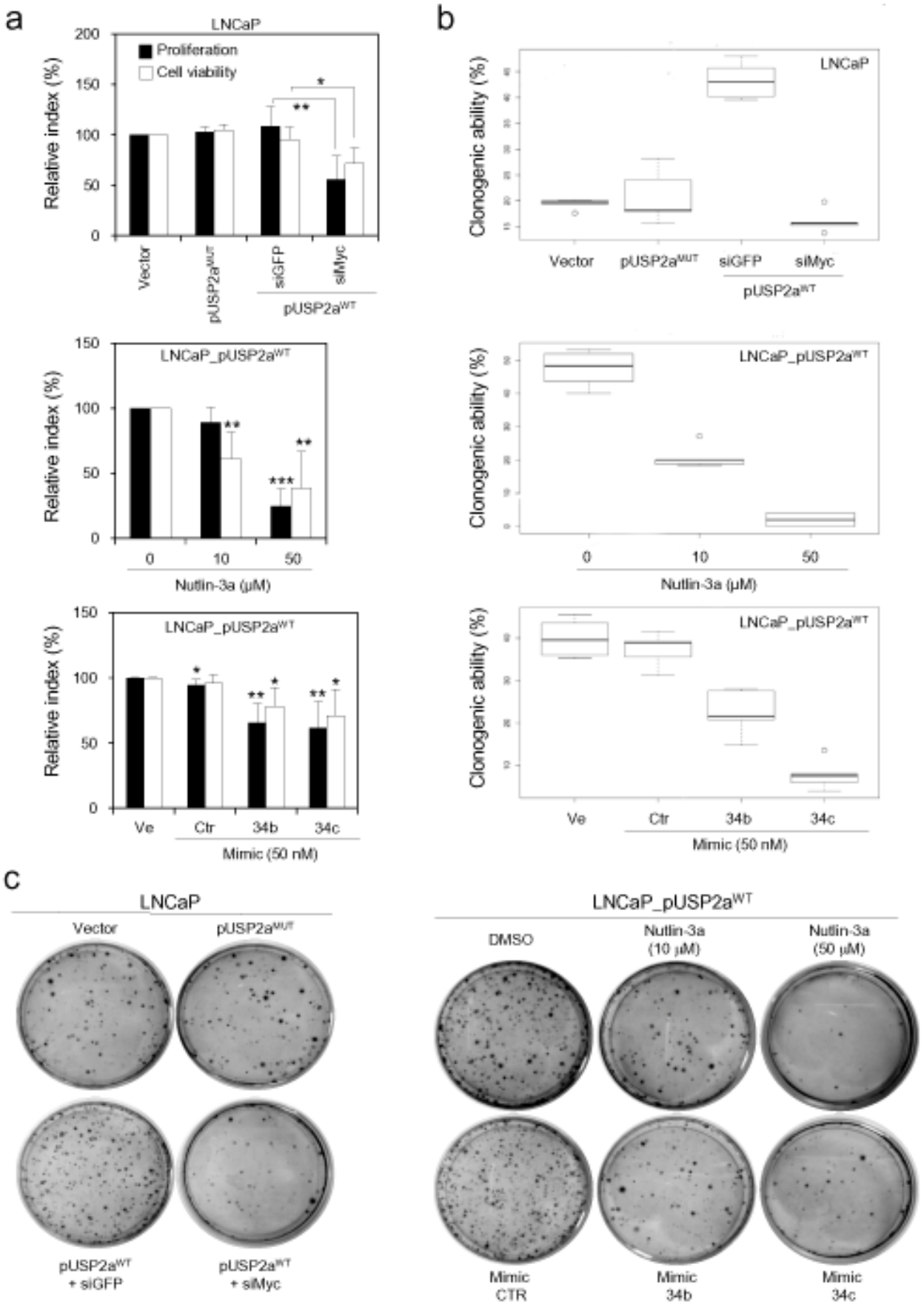
Supplementary Fig. 10



Supplementary Fig. 11



Supplementary Fig. 12



Supplementary Figures legends.

Supplementary Figure 1. Evaluation of miRs expression performed by TaqMan qRT-PCR assay in the control (Vector), USP2a^{MUT} and USP2a^{WT} iPrEC cells.

Supplementary Figure 2. (**a**) The panel details strongly predicted target genes for USP2a-regulated miRNAs. The identification of transcripts potentially targeted by miRs was carried out by interrogating public databases and algorithms available on the web (Sanger miRbase, PicTar, Target Scan, miRanda, DIANA miRGen, miRNA Map). The western blot panel highlights the protein expression levels of a selected set of predicted target genes for USP2a-regulated miRNAs. (**b**) The graphs represent a gene ontology analysis (gene number and ontologic categories) of miR putative targets performed by the miRNApath algorithm (http://lgmb.fmrp.usp.br/mirnapath/). The analysis focuses on characterizing each USP2a-modulated miRNA on the basis of the biologic functions exerted by its target genes.

Supplementary Figure 3. (**a**) Analysis of USP2a and c-Myc protein expression and evaluation of transcripts levels of selected c-Myc target genes carried out in LNCaP, following transfection with increasing concentrations (0 μg/black bars, 2 μg/gray bars, 5 μg/white bars) of pUSP2a^{WT} or pUSP2a^{MUT} expression vectors. (**b**) Evaluation of c-Myc protein and c-Myc-regulated miRs expression, performed in empty vector (Vector) and in USP2a^{WT} over-expressing LNCaP transfectants silenced either for GFP or Myc expression. All data represent mean ± s.d. of at least three independent replicates. *P* values: (*) $P< 0.05$, (**) $P< 0.01$, (***) $P< 0.001$.

Supplementary Figure 4. (**a**) Schematic diagram of potential miR-34b, miR-34c, miR-98 and let-7c binding sites within the human c-myc transcript 3'UTR. (**b**) Western blot of c-Myc protein level carried out in iPrEC and USP2a^{WT}-overexpressing LNCaP transfectants undergoing treatment with control (CTR), miR-98 and let-7c synthetic molecules (50 nM).

Supplementary Figure 5. (**a**) Western blot analysis of c-Myc protein and evaluation of miRNA 34b/c expression carried out in USP2a^{WT} LNCaP transfectants, treated with control (CTR), miR34b and miR-34c synthetic molecules. (**b**) Western blot of c-Myc and USP2a protein levels and miRNA 34b/c expression performed in Control and in USP2a-silenced LNCaP cells following treatment with either control (KO_scrambled) or specific anti-miR (KO_34b, KO_34c) LNA molecules. All data represent mean ± s.d. of at least three independent replicates. *P* values: (*) $P< 0.05$, (**) $P< 0.01$, (***) $P< 0.001$.

Supplementary Figure 6. Comparison of MYC, hsa-miR-34b, and hsa-miR-34c-5p expression levels across 28 adjacent normal prostate samples, and 94 primary localized prostate cancer not previously treated with any type of neo-adjuvant therapy. An inverse correlation between MYC and the selected microRNAs expression levels was observed upon local invasion. In the top portion of the figure the box plots show the different distribution of the fold-change between MYC and hsa-miR-34b (panel A), and MYC versus hsa-miR-34c-5p (panel B), along with t-test p-value. In the bottom part of the figure the scatter plots show the negative association (blue fitted lines) between MYC and hsa-miR-34b expression (panel C), and between MYC and hsa-miR-34c-5p expression (panel D), across all 122 analyzed samples (red triangles and black circles for cancer and normal respectively). R-squared, intercept, coefficient, and the associated p-values, as obtained by linear regression analysis are reported.

Supplementary Figure 7. Western blot analysis of p53 (**a,** top left panel), c-Myc and p21 (**b,** bottom right panel) protein levels, ChIP assay (**a**, top right panel) and miRNA 34b/c expression analysis (**b**, bottom left panel) carried out in LNCaP and PC3 prostate cancer cell lines, following transient transfection with either Control (scrambled) or specific p53 silencing oligonucleotides.

Supplementary Figure 8. Western blot analysis of MdM2, p53 and p21 protein levels (left panels) and ChIP assay (right panels) carried out in LNCaP cells upon transient transfection with pCDNA3 empty vector (1), pUSP2a^{WT} (2), control silencing (3) or specific USP2a silencing oligonucleotides (4).

Supplementary Figure 9. (**a**) Schematic representation of the 50 kb upstream regions of miR-98 and let-7c on human chromosomes X and 21 respectively. TATA boxes, TSS sites and both p53 and c-Myc putative binding sites are highlighted. The analysis of predicted regulatory regions was conducted as previously described for miR-34b/c promoters. (**b**) Analysis of p53 and miR-34c, miR-98 and let-7c expression in p53-silenced and pUSP2a^{WT}-overexpressing LNCaP cells. (**c**) Analysis of p53 and miR-34c, miR-98 and let-7c expression levels carried out in USP2a-silenced and Nutlin-3a-treated LNCaP cells. (**d**) ChIP assay carried out in DMSO/Ctr and Nutlin-3a-treated LNCaP cells. Immuno-precipitation of chromatin was performed without antibody (no Ab), with specific anti-p53 and with further PCR amplification of the indicated regions on the putative miR-98, let-7c and miR-34b/c regulating regions. The selected p53 response elements were chosen on the basis of their highest predicted score (indicating the probability that region actually represents an *in vivo* p53 binding site) calculated by MatInspector software (Quandt et al., 1995, Cartharius et al., 2005). All data represent mean ± SD of at least three independent technical replicates. Asterisks represent p values: (*) $p< 0.05$, (**) $p< 0.01$, (***) $p< 0.001$.

Supplementary Figure 10. (**a**) Volcano plots showing differentially expressed genes in each pair-wise comparison among USP2a^{MUT}, USP2a^{WT} and EV cell lines. On the x-axis the $\log_2$ fold-cange, on the y-axis lods of differential expression (the corresponding probabilities are shown in the legend). (**b**) Correspondence-at-the-top (CAT) plot for the up-regulated genes. (**c**) Correspondence-at-the-top (CAT) plot for the down-regulated genes.In Panels (b) and (c) the green line corresponds to the comparison of the USP2a^{MUT} and EV cell lines, the orange line to the one between the USP2a^{WT} and EV cell lines, and the blue line to the one between the USP2a^{WT} and USP2a^{MUT} cell lines. On the y-axis are represented the proportion of gene in common, on the x -axis the rank determined by the moderated t-statistics, as resulting from our linear model analysis.

Supplementary Figure 11. (**a**) Correspondence-at-the-top (CAT) plot for Protein-Protein-Interaction (PPI) gene lists. The enrichment was determined by one-sided Wilcoxon rank-sum test (upper tail), and the corresponding p-values were used for the CAT-plot ranking. (**b**) CAT-plot for PPI gene lists. The enrichment was determined by one-sided Wilcoxon rank-sum test (lower tail), and the corresponding *p*-values were used for the CAT-plot ranking. In both Panels the green line corresponds to the comparison between the USP2a^{MUT} and EV cell lines, the orange line to the one between the USP2a^{WT} and EV cell lines, and the blue line to the one between the USP2a^{WT} and USP2a^{MUT} cell lines. On the y-axis are represented the proportion of gene sets in common, on the x -axis is reported the corresponding rank as determined by the Wilcoxon rank-sum test. Similar results were also obtained with the other investigated functional themes (i.e. Gene Ontology, KEGG, TFBS, miRNA targets).

Supplementary Figure 12. (**a**) Cell proliferation and viability assays performed in LNCaP cells and pUSP2a^{WT} clones, following transfection and treatment with Nutlin-3a and miR Mimic molecules. All data represent mean ± sd of at least five independent replicates. *P* values: (*) *P*< 0.05, (**) *P*< 0.01, (***) *P*< 0.001. (**b**) Colony formation assay carried out in LNCaP cells and pUSP2a^{WT} clones following transfection and treatment with Nutlin-3a and miR Mimic molecules, *P* <0.001. (**c**) Clonogenic assay plates of LNCaP cells and pUSP2a^{WT} clones reported in (b).

Supplementary Table 1: siRNAs sequence

siRNA	Sequence
siUSP2a (Graner et al., 2004)	r(UGCUUGUGCCCGGUUCGAC)d(TT)
siControl (siUSP2a scrambled) (Graner et al., 2004)	r(UUCUUCGAACGUGUCACGU)d(TT)
siLacZ	r(GUGACCAGCGAAUACCUGU)d(TT)
siGFP	r(AAGUUCAGCGUGUCCGGGGAG)d(TT)
siMyc	r(GCCACAGCAUACAUCCUGU)d(TT)
siMdm2_#1	r(UAACCACCUCACAGAUUCCAG)d(TT)
siMdm2_#2	r(UGGUUGCAUUGUCCAUGGC)d(TT)
sip53	r(GACUCCAGUGGUAAUCUAC)d(TT)
siControl (sip53 scrambled)	r(CUAUAACGGCGCUCGAUAU)d(TT)

Supplementary Table 2: Semiquantitative RT-PCR primers

Coding Gene	Primer sequence
hUSP2a	5'-TGC TGA GAC CCG ACA TCA CT-3' (forward) 5'-TGG GGT CTA TCC GGT AGC TA-3' (reverse)
hCyclin B	5'-TCG AGC AAC ATA CTT TGG CCA-3' (forward) 5'-GCA AAA AGC TCC TGC TGC AA-3' (reverse)
hCdc25a	5'-AGA TAG CAG TGA ACC AGG-3' (forward) 5'-ATC GGT TGT CAA GG-3' (reverse)
hCDK4	5'-CCT GGC CAG AAT CTA CAG CTA-3' (forward) 5'-ACA TCT CGA GGC CAG TCA TC-3' (reverse)
hODC	5'-AGA CCT TCG TGC AGG CAA TC-3' (forward) 5'-AGG AAA GCC ACC GCC AAT AT-3' (reverse)
hDARC	5'-CCT GTG GGC CTG GTT TAT TTT CT-3' (forward) 5'-ATT CAG GTT GAC AGG TGG GAA GA-3' (reverse)
hGAPDH	5'-GAG TCA ACG GAT TTG GTC GT-3' (forward) 5'-GAC AAG CTT CCC GTT CTC AG-3' (reverse)

Supplementary Table 3: Quantitative RT-PCR (qRT-PCR) primers

Coding Gene	Primer sequence
hUSP2a	5'-TGC CTC CAG AGG CTC TAC AT-3' (forward) 5'-CGA AGG AAC TCC TGA GCA TC-3' (reverse)
hGAPDH	5'-GAG TCA ACG GAT TTG GTC GT-3' (forward) 5'-GAC AAG CTT CCC GTT CTC AG-3' (reverse)
hALD_A	5'-CGC AGA AGG GGT CCT GGT GA-3' (forward) 5'-CAG CTC CTT CTT CTG CTC CGG GGT-3' (reverse)

Supplementary Table 4: miR-34b/34c putative promoter ChIP primers

miR-34b/34c putative promoter	Primer sequence
Amplicon A (p53_RE1)	5'-TACCTAAGAGCTCCGGCAAC -3' 5'-CTTGGCCTCTCACAGTGCTA-3'
Amplicon B	5'-TCTTCAAAGGCACTGAATTGAC-3' 5'-CTTGGCCTCTCACAGTGCTA-3'
Amplicon C	5'-CCTGAGGTCAGGAGTTCGAG-3' 5'-CACCATCTCCCACAGCTTTT-3'
Amplicon D	5'-CAAAAGCTGTGGGAGATGGT-3' 5'-TCACACCTGTAATCCCAGCA-3'
Amplicon E	5'-TGGCCAGGATGGTCTCTATC-3' 5'-CCAGGGAGACCCATGATTTA-3'
Amplicon F (p53_RE2)	5'-AAATCATGGGTCTCCCTGGT-3' 5'-TTACCTGAGCTGGATTGCTG-3'
Amplicon G	5'-TCAGTTGGGAAGAAGTGTTCG-3' 5'-CACCCCATAGTGAAGGGAAA-3'
Amplicon H	5'-ACCTGCCCCTGTTCATGTTA-3' 5'-GGGCAAAAGGGACAGTTACA-3'
Amplicon I	5'-GCTTCCAGCTGAATTTCACA-3' 5'-CCCAGAGGAGGTGAGACTTG-3'
Amplicon L	5'-CCTCCTCTGGGAACCTTCTT-3' 5'-GGCTTCCCAGGTACCTCAA-3'
Amplicon M	5'-GCACAGAGGTGCAGATGAGA-3' 5'-GACCAACCGTCCTTGGAAC-3'
Amplicon N	5'-CACAGCGCTTTCTCTCAGC-3' 5'-CCATGACCCCCAGGAGTG-3'
Amplicon O	5'-AAGGAAAAGCGAGGGGAAC-3' 5'-ACTGCCTACAAACCGAGCAC 3'

Supplementary Table 5: Human coding gene ChIP primers

Genomic region	Primer sequence
Human p21 promoter	5'-ATGTATAGGAGCGAAGGTGCAGAC-3' 5'-CCTCCTTTCTGTGCCTGAAACA-3'
Human cyclin B1 intron	5'-GAGTCTCTATCGGCTCTTATACCG3' 5'-GTCCAGTTTCCCAAGGCCAAT-3'

Supplementary Table 6: miR-98 putative promoter ChIP primers

miR-98 putative promoter	Primer sequence
p53_RE1	5'-GATGATGAGCTGCTGGATGA-3' 5'-TGATCCACCGATCCAAAAAT-3'

Supplementary Table 7: Let-7c putative promoter ChIP primers

Let-7c putative promoter	Primer sequence
p53_RE1	5'-GGAGACAGATGAAGGGATGC -3' 5'-GAAACGTCTACCTGGCAAGC -3'
p53_RE2	5'-TGAAGGACATGGCACAAAAA -3' 5'-AGCGGCACTAAAAAGAACCA -3'

SUPPLEMENTARY METHODS

1. miRNA target and molecular pathways prediction. Analysis of putative miRNA targets was carried out by union of results obtained from the following web-based databases:

1. Sanger miRbase (http://microrna.sanger.ac.uk);
2. PicTar (http://pictar.mdc-berlin.de);
3. TargetScan(http://www.targetscan.org);
4. miRanda (http://www.microrna.org/microrna/home.do),
5. DIANAmiRGen(http://www.diana.pcbi.upenn.edu/miRGen.html);
6. miRNA Map (http://mirnamap.mbc.nctu.edu.tw/).

Pathway and structural analysis was performed using miRNApath (http://lgmb.fmrp.usp.br/mirnapath/).

2. *In silico* prediction of putative miRNA promoter regions. Genomic regions encoding for human miRNAs were obtained from the Sanger miRbase website (http://microrna.sanger.ac.uk), and loaded into the UCSC Genomic bioinformatics database (Kuhn et al., 2009, Kent et al., 2002) (http://genome.ucsc.edu/). Four kb upstream the miR34b/c genome locus, and up to 50 kb upstream the predicted miR-98 and let-7c encoding sequences, were scanned by Promoter Scan (Prestridge, 1995) (http://www-bimas.cit.nih.gov/molbio/ proscan) and MatInspector (Quandt et al., 1995, Cartharius et al., 2005) (http://www.genomatix. de/products/MatInspector) to search for putative promoter sequences (TATA boxes and transcription starting site-TSS) and binding sites for transcription factors (TFs). miR-34b and miR-34c map (within an inter-genic region) on chromosome 11q23, with miR-34b located about 500 bp upstream of miR-34c. miR-98 and let-7c are both intra-genic and map on chromosome X and chromosome 21, respectively.

3. Microarray data sets description. Briefly, raw gene expression data and all required MIAME information is permanently hosted in the NCBI Gene Expression Omnibus (GEO) database (series GSE17466). All analyses were performed with packages from R/Bioconductor (Gentleman et al., 2004; Ihaka and Gentleman, 1996). Briefly, raw gene expression data obtained on Affymetrix platforms were normalized at the probe level using the robust multi-array average (RMA) empirical stochastic model described by Irizarry (Irizarry et al., 2003). Standardization across Affymetrix DNA-chips was attained by quantile normalization (Bolstad et al., 2003). For dual-color arrays, no background subtraction was performed, and within-array “loss” and between-arrays ”scale” normalization methods were applied to log2 expression ratios (Yang et al., 2002). For all features on the array moderated t-statistics (by empirical Bayes shrinkage of standard errors), log-odds ratios of differential expression and adjusted p-values (Benjamini and Hochberg method) (Benjamini and Hochberg, 1995) were obtained after fitting a gene-wise linear model that accounted for correlation of biological replicates, group effects, samples origin, and labeling (for two-colors design). Functional Gene Sets (FGS) were obtained from a number of different genomic databases, encompassing distinct functional themes, including Gene Ontology (GO) (Ashburner et al., 2000), the Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa et al., 2004), the Molecular Signatures Database (Subramanian et al., 2005), the UCSC Genome Browser database (Kanehisa et al., 2004; Karolchik et al., 2008), the Stanford Microarray Database (SMD) (Demeter et al., 2007), and the NCBI Entrez Gene database (Wheeler et al., 2008). Detailed information on FGS construction and cross-referencing and on databases releases is reported in the Supplemental Experimental Procedures Section. Enrichment analysis was performed using a one-sided Wilcoxon rank-sum test, separately for up- and down-regulation, after ranking the genes by their signed or absolute moderated t-statistics as previously described (Daniel et al., 2009; Schaeffer et al., 2008). Multiple testing correction was performed to adjust the p-values using the Benjamini and Hochberg method (Benjamini and Hochberg, 1995). Detail methods are reported in the

Supplemental Experimental Procedures Section. Supplementary Material linked to external genomic databases are also available at: http://astor.som.jhmi.edu/~marchion/labML/benassi.html.

In details:

3.1 In the present study we analyzed differential gene expression from *de novo* generated data, as well as from data sets available from the public domain.

De novo USP2a data set**:** Using the Affymetrix Human Genome U133 Plus 2.0 Array platform (hereafter referred to as hgu133plus2), we analyzed differential gene expression in prostate cancer cell cultures transfected with wild-type USP2a (WT), mutant USP2a (MUT), and with the empty vector used as control (EV). This analysis was performed by direct comparison of the three different groups (WT, MUT, EV), using a generalized linear model approach. The fit was obtained by robust M-estimation, allowing for a small proportion of outliers, by the iterated re-weighted least squares (IWLS) method (Venables and Ripley, 2002). An empirical Bayes method was used to moderate the standard errors of the estimated log-fold changes among groups (Smyth, 2005, Smyth, 2004, Smyth et al., 2005). Such *de novo* microarray analysis was performed using 5 different biological replicates from each analyzed group, for a total of 15 microarrays, hybridized in two distinct batches of three and two biological replicates respectively.

Tomlins human prostate cancer data set (GSE6099 series**):** Raw expression data for this data set were retrieved from NCBI Gene Expression Omnibus database (Wheeler et al., 2008) (GEO, series GSE6099) and used in the present analysis. In this study (Tomlins et al., 2007) Tomlins and colleagues performed a large gene expression profiling study on Laser-Capture Microdissected (LCM) cell specimens from normal and cancerous human prostate, along sequential stages of disease progression. In this study the authors also demonstrated that in most of human prostate cancer gene expression studies available from the public domain there is a major bias caused by differential expression of transcripts of stromal origin. For this reason we selected the this large LCM data set for comparison with the gene expression signatures associated with USP2a expression (our *de novo* generated data).

3.2. Gene Annotation. Gene annotation for commercial arrays (Affymetrix) was obtained from metadata packages available from the R-Bioconductor project (Ihaka and Gentleman, 1996, Gentleman et al., 2004). The annotation package for the custom microarray platform used by Tomlins and colleagues (GSE6099 data set (Tomlins et al., 2007)) was prepared using the AnnotationDBIR-Bioconductor package, using the methods described for the annBuilderlibrary (Zhang et al., 2003). We used both IMAGE clone identifiers and GenBank accession numbers to retrieved the associated annotation information. Cross-referencing of array features across platforms was based on Entrez Gene identifiers. When cross-referencing was required, Affymetrix probe sets mapping to multiple Entrez Gene identifiers were excluded from the analysis, while multiple probe sets mapping to same Entrez Gene identifiers (redundant probe sets) were filtered by keeping the most differentially expressed ones in each data set. Below is the detailed annotation for the considered platforms used in the present study as obtained from each meta-data package.

This information includes mapping between the following entities:

- ACCNUM: feature IDs to GenBank Accession Numbers;
- ALIAS2PROBE: feature IDs to alternative Gene Symbol;
- CHRLOC: feature IDs to Chromosomal Location;
- CHRLENGTHS: length of each of the Chromosomes;
- CHR: feature IDs to Chromosomes;
- ENSEMBL2PROBE: feature IDs to Ensembl gene accession numbers;
- ENSEMBL: feature IDs to Ensembl gene accession numbers;

- ENTREZID: feature IDs to Entrez Gene;
- ENZYME: feature IDs to Enzyme Commission (EC) Numbers;
- GENENAME: feature IDs to Gene names;
- GO: feature IDs to Gene Ontology (GO);
- MAP: feature IDs and cytogenetic maps/bands;
- OMIM: feature IDs to Mendelian Inheritance in Man (MIM) identifiers;
- PATH: feature IDs to KEGG pathway identifiers;
- PMID: feature IDs to PubMed identifiers;
- REFSEQ: feature IDs to RefSeq identifiers
- SUMFUNC: feature IDs to Gene Function Summaries;
- SYMBOL: feature IDs to Gene Symbols;
- UNIGENE: feature IDs to UniGene cluster identifiers;
- CHRLENGTHS: A named vector for the length of each of the chromosomes;
- ENZYME2PROBE: Enzyme Commission Numbers to feature IDs;
- GO2ALLPROBES: Gene Ontology (GO) identifiers to all feature IDs;
- GO2PROBE: Gene Ontology (GO) identifiers to feature IDs;
- ORGANISM: the Organism for each specific microarray platform;
- PATH2PROBE: KEGG pathway identifiers to feature IDs;
- PFAM: feature IDs to Pfam identifiers;
- PMID2PROBE: PubMed identifiers to feature IDs.

For **all platforms** annotation was retrieved using the following database releases:
Additional Information about packages: DB schema: HUMANCHIP_DB
DB schema version: 1.0
Organism: Homo sapiens
Date for NCBI data: 2008-Apr2
Date for GO data: 200803
Date for KEGG data: 2008-Apr1
Date for Golden Path data: 2006-Apr14
Date for IPI data: 2008-Mar19
Date for Ensembl data: 2007-Oct24

3.3. Microarray data pre-processing. Raw data were obtained for all hybridization considered (GPR files for two color arrays and Affymetrix CEL files). All pre-processing procedures described below were performed using functions and methods implemented in the packages affy (Gautier et al., 2004) and limma (Smyth, 2005, Smyth, 2004), available though the R/Bioconductor project (Ihaka and Gentleman, 1996, Gentleman et al., 2004). We pre-processed each dataset separately, according to the type of platform involved. Pre-processing appropriateness was monitored using standard diagnostic plots (i.e. 2-D image plots, RNA degradation plots, foreground-background plots, MA-plots, and box plots).

3.4. De novo USP2a data set. Affymetrix raw data were normalized at probe-level by fitting the RMA empirical stochastic model described by Irizarry (Irizarry et al., 2003). Standardization across Affymetrix DNA-chips was attained by quantile normalization (Bolstad et al., 2003).

3.5. Tomlins human prostate cancer data set (GSE6099 series). Individual GPR files containing raw gene expression data from the study by Tomlins and colleagues (Tomlins et al., 2007) were retrieved from GEO (series GSE6099). Median intensities were normalized using the ”loess” method (Yang et al., 2002). Flagged and ”empty” microarray features were not used to compute the ”loess” smoothing and were further excluded from the subsequent analysis. No background subtraction was performed (Scharpf et al., 2007). The ”scale” method was applied to normalize M-values across arrays (Yang and Thorne, 2003, Smyth and Speed,

2003).

3.6. Differential gene expression. In all data sets considered in the present study differential gene expression was investigated using functions and methods implemented in the R/Bioconductor (Ihaka and Gentleman, 1996, Gentleman et al., 2004) package limma (Smyth, 2005, Smyth, 2004). Briefly, a fixed effects linear model was fit for each individual feature to estimate expression differences between groups of samples to be compared. When technical replicates or matched samples from the same individual were available, correlation coefficients were computed between replicates and the associated consensus correlation was added to the model (Smyth et al., 2005). An empirical Bayes approach was applied to moderate standard errors of M-values (Lonnstedt and Speed, 2002, Smyth, 2004).
Finally, for each analyzed feature moderated t-statistics, log-odds ratios of differential expression (B-statistics), raw and adjusted p-values (FDR control by the Benjamini and Hochberg method (Benjamini and Hochberg, 1995)) were obtained.

3.7. De novo USP2a data set. This data set accounted for:

1. Five biological replicates of wild-type (WT) USP2a transfected cells;
2. Five biological replicates of mutant (MUT) USP2a transfected cells;
3. Five biological replicates of mock-transfected cells with empty vector (EV);

For each group of samples we processed two distinct batches, since three biological replicates were obtained in a first set of experiments, while the remaining two replicates were obtained on a second time. Therefore this information was used as a covariate and added to the model. Differential gene expression was investigated for the following contrasts:

- WT versus EV, using WT as the numerator in the log-ratio;
- WT versus MUT, using WT as the numerator in the log-ratio;
- MUT versus EV, using MUT as the numerator in the log-ratio;

3.8. Tomlins human prostate cancer data set (GSE6099 series). The data set accounted for the following sample groups:

- Three RNA specimens from normal prostatic epithelial cells;
- Fifteen RNA specimens from epithelial cells dissected from normal areas adjacent to PCA;
- Two RNA specimens from atrophic epithelial cells;
- Four RNA specimens from prostatic inflammatory atrophy (PIA) epithelial cells;
- Four RNA specimens from benign prostatic hyperplasia (BPH) epithelial cells;
- Three RNA specimens from normal prostatic epithelial cells;
- Thirteen RNA specimens from epithelial cells from Prostatic Intraepithelial Neoplasia (PIN);
- Thirty-two RNA specimens from epithelial cells from PCA (10 with Gleason score 8 or higher, 8 with a Gleason score of 7, and 14 with a Gleason score of 6);
- Three RNA specimens from epithelial cells from hormone untreated PCA metastases;
- Seventeen RNA specimens from epithelial cells from hormone independent PCA metastases;
- Two RNA specimens from stromal cells dissected from normal areas adjacent to PCA;
- Three RNA specimens from stromal cells dissected from epithelial BPH specimens;
- Three RNA specimens from stromal cells dissected from epithelial BPH specimens;
- Four RNA specimens from stromal cells dissected from stromal BPH nodules;
- Three RNA specimens from stromal cells dissected from normal prostate specimens (organ donors);

Covariates for specimen groups described above, Gleason grade, and the consensus correlation within matched specimens from the same patients were added to the models. All

samples were hybridized in dual-color experiments using a common reference sample (Clontech prostate pool RNA, used in the Cy3 channel), therefore a covariate for the dye was added as well. Differential gene expression was investigated for the following contrasts:

- Neoplasia: PIN epithelial cells versus NORMAL epithelial cells;
- Cancer: PIN and PCA epithelial cells versus NORMAL epithelial cells;
- Invasion: PCA epithelial cells versus PIN epithelial cells;
- Gleason: High Gleason score PCA epithelial cells versus Low Gleason score PCA epithelial cells;
- Progression: All MET epithelial cells versus PCA epithelial cells;
- ARindependence: Hormone independent MET epithelial cells versus hormone nahive MET epithelial cells;
- EpivsStro: All epithelial cells specimens versus all stromal cells specimens;
- StroPCSvsStroNOR: Stromal cells from normal areas adjacent to PCA versus Normal stromal cells specimens;

Each contrast was constructed to have gene expression levels from the most advanced disease stage as the numerator (i.e., PCA/PIN, or MET/PCA, and so on ...).

3.9. Analysis of Functional Annotation. Analysis of Functional Annotation (AFA) has been already successfully applied (Schaeffer et al., 2008, Daniel et al., 2009, He et al., 2009), and was performed in this study to accomplish the following goals:

- Capture biological processes relevant in the investigated contrasts;
- Compare pathways and biological themes associated with gene expression programs associated with USP2a expression;

To perform the AFA we used a one-sided Wilcoxon rank-sum test to identify the biological concepts (Functional Gene Sets, FGS hereafter) associated with the phenotypes and comparisons we performed. The Wilcoxon rank-sum test computes a p-value to test the hypothesis that a FGS, defined by a functional annotation, tends to be more highly ranked in an ordered list. In the present study individual genes on the arrays were ranked by their absolute and signed moderated t-statistics, and statistical tests were performed for all the contrasts considered in the linear models described above. The use of the absolute moderated t-statistics enabled the investigation of gene set enrichment irrespective to differential gene expression direction (up- or down-regulation), while the use of the signed t-statistics enabled investigating enrichment driven by up- or down-regulated genes. All these analyses were performed using two distinct reference populations:

- All the non-redundant genes (according to Entrez Gene identifiers) present on the microarray platform;• All the non-redundant genes annotated in each specific functional theme analyzed (i.e. the genes with a GO annotation when analyzing GO, the genes with a KEGG annotation when analyzing KEGG, . . .).

For each type of functional theme the results obtained from the Wilcoxon rank-sum test by using the two reference populations above, were compared to one another for each considered contrast (i.e. WT versus MUT). This comparison allowed us to identify the Functional Gene Sets (FGS) that are enriched irrespective to the reference population used, and those for which the enrichment depended on the reference population of choice, which were therefore excluded from further analyses. Such filtering was accomplished by ordering the FGS by their p-value, comparing the relative ranks, and excluding the gene sets showing a rank-difference below the 1st percentile and above the 99th percentile of the ranks difference distribution. Overall, this approach allowed us to avoid calling significant gene sets enriched only in one of the two analyses. After the statistical tests were performed and filtered, control of false discovery rate (correction for multiple hypothesis testing) was obtained by applying the Benjamini and Hochberg method (Benjamini and Hochberg, 1995), as implemented in the multtestR/Bioconductor package.

Functional annotation was retrieved from the following sources:

- Gene Ontology Terms (GO) (Ashburner et al., 2000);
- KEGG PATHWAY sets (Kanehisa et al., 2004);
- Cytogenetic bands and chromosomes;
- Segal's Cancer Modules (Segal et al., 2005, Segal et al., 2004), module gene sets from Eran Segal's lab (described at http://ai.stanford.edu/~erans/cancer/browse_by_modules.html; retrieved from the Rob Tibshirani webpage (http://www-stat.stanford.edu/~tibs/GSA/index.html).
- PMID: PubMed gene sets, defined as groups of genes co-cited in the same published manuscripts;
- ENZYME: Enzyme Commission number (EC number) gene sets, defined as the genes sharing the same EC number;
- Functional themes from MSigDb (Subramanian et al., 2005) (Molecular Signature Database, please refer to: http://www.broad.mit.edu/gsea/msigdb/msigdb_index.html);
- FGS defined by Protein-Protein-Interaction (PPI) data available through the NCBI Entrez Gene database (Wheeler et al., 2008), which includes evidence from the Human Protein Reference Databases (HPRD) (Peri et al., 2004), BioGrid (Stark et al., 2006) and the Biomolecular Interaction Network Database (BIND) (Bader et al., 2001). PPI FGS were build using the information from all the three sources, irrespective to the technique or the type of evidence available on the physical interaction (i.e. yeast-two-hybrid, immunoprecipitation).
- FGS defined by shared miRNA target sequences, as obtained from the miRGen data base (Megraw et al., 2007) (http://www.diana.pcbi.upenn.edu/miRGen.html). Such collection of miRNA FGS accounted for gene lists resulting from the union or the intersection of miRNA targets predicted by a set of distinct prediction algorithms (PicTar (Krek et al., 2005), TargetScanS (Lewis et al., 2005), miRanda (John et al., 2004), and DIANA-microT (Maragkakis et al., 2009, Kiriakidou et al., 2004)).
- Transcription Factor Binding Site (TFBS) target gene lists defined by shared and conserved TRANSAC TFBS in their promoter regions, as obtained from the GoldenPath data base of the UCSC Genome Browser (Kuhn et al., 2009, Kent et al., 2002). Such FGS were constructed using two genomic regions around the transcription starting site (TSS) of each gene (10Kb and 30Kb), and using two distinct Z-scores for conservation (Z=1.64 and Z=2.33),corresponding to a False Discovery Rate of 5%. Details are available from the UCSC Genome Browser (http://genome.ucsc.edu/cgi-bin/hgGateway)).

Mappings between individual features of each platform to the FGS considered were based on NCBI Entrez Gene, as obtained from the R-Bioconductor metadata packages (see Annotation in the Methods section).

We also used the approach described above to ascertain whether the USP2a gene signatures where enriched along human prostate cancer progression (PCA). To this end we defined the list of genes differentially expressed in the contrast between the WT and MUT groups, excluding the genes that proved to be differentially expressed in any other comparison (WT versus EV, and MUT versus EV). We then split such list in genes that were up- or down- regulated, and used such groups of genes as FGS in Wilcoxon rank-sum tests performed using the contrasts from our linear analysis of the Tomlins data set, similarly to what we performed elsewhere (Schaeffer et al., 2008). For this analysis we considered only the genes annotated to the NCBI Entrez Gene data base, and we repeated it using lists of increasing size (which is by selecting the USP2a genes at different significance levels) with similar results.

3.10. Concordance-at-the-top plots. The " Concordance-at-the-top" plot (cat-plot) (Irizarry et al., 2005) was developed to asses the agreement between microarray results from different contrasts considered. In particular, this technique enables comparing the correspondence of two lists ranked by a predefined statistics at their top. This is accomplished as follows:

1. Ordering of the two lists according to a suitable statistics (i.e. differential gene expression, significance, probability, ...);
2. Computing the proportion of elements in common for a given list size;
3. Reiterating the two steps above increasing the list size up to all common elements;
4. Plotting the proportion of common elements against the increasing size of the the considered lists;

Since cat-plots evaluate such agreement at the top of ranked lists they are particularly useful in gene expression analysis, where only a small fraction of genes is expected to be differentially expressed over the large total number of analyzed genes.

We used cat-plots to evaluate the agreement of gene expression signatures and AFA results across the comparisons considered in the present study (i.e. among the WT-to-MUT and the WT-to-EV contrasts). For comparisons at the feature level, we ranked each gene list using the moderate t-statistics resulting from our linear model analysis, and the median gene expression by sample group. In this study we ordered the features identifiers (Affymetrix probe sets) and performed the cat-plots as follows:

- By increasing ordering using the signed moderate t-statistics, and the group log fold-change, to investigate down-regulated genes separately from up-regulated ones;
- By increasing ordering using the inverse signed moderate t-statistics, and the group log fold-change, to investigate up-regulated genes separately from down-regulated ones;
- By increasing and decreasing ordering using the median expression intensity for each group to investigate up-regulated genes separately from down-regulated ones;
- By decreasing ordering using the absolute moderate t-statistics, and the group log fold-change, to investigate up and down-regulated genes at once;

We also use cat-plots to compare the agreement among contrasts across data sets at the level of Functional Gene Sets (FGS). In this case, FGS were ordered according to raw p-values obtained from Wilcoxon rank sum tests.

3.11. Comparison of MYC, hsa-miR-34b, and hsa-miR-34c-5p expression levels. In order to further confirm the relationship between USP2a, c-Myc, and microRNA expression levels we queried the NCBI Gene Expression Omnibus database to identify datasets in which measurements were obtained for both mRNA and microRNA levels from the same patients. We have therefore identified and analyzed the two NCBI GEO series GSE21034 and GSE21036 (Taylor et al, Cancer Cell, 2010,).

Taylor human prostate cancer data set (GSE21034 and GSE21036 series)

Individual CEL and text files containing raw mRNA and microRNA expression data from the study by Taylor and colleagues (Taylor et al., 2010) were retrieved from GEO (GSE21034 and GSE21036 series), pre-processed and normalized using standard procedures. Briefly, for microRNA expression data, which was measured on the Agilent-019118 Human miRNA Microarray version 2.0 G4470B, no background subtraction was performed (Scharpf et al., 2007), and median intensities were normalized across samples by quantile normalization (Bolstad et al., 2003). Finally A 10% trimmed mean was used to summarize microRNA measurements from identical probes. Messenger RNA expression was measured using the Affymetrix Human Exon 1.0 ST Array. Raw data were normalized at probe-level by fitting the RMA empirical stochastic model described by Irizarry (Irizarry et al., 2003). Standardization across Affymetrix DNA-chips was attained by quantile normalization (Bolstad et al., 2003), and probe set level summarization was obtained at the NCBI RefSeq transcript level using the R/Bioconductor “oligo” package.

After pre-processing and normalization we compared log2 transformed expression levels of MYC, hsa-miR-34b, and hsa-miR-34c-5p across 28 adjacent normal prostate samples, and 94 primary localized prostate cancer not previously treated with any type of neo-adjuvant therapy. Since we expected to observe an inverse correlation between expression levels of MYC and the selected microRNAs upon invasion, we compared across normal and cancer samples the log2 fold-change between MYC and each microRNA (MYC versus hsa-miR-34b, and MYC versus hsa-miR-34c-5p). This analysis revealed significant differences between normal and cancerous samples (t-test p-value < 10e-7 in both comparisons, see Supplementary Figure 11, panels A and B), with higher MYC expression in cancer compare to normal, and an opposite trend for the selected microRNA. This analysis was further confirmed by negative correlations between MYC and the selected microRNA, as assessed by linear regression analysis (coefficient p-value < 0.01 and 0.001 for hsa-miR-34b and hsa-miR-34c-5p respectively, Supplementary Figure 11, panels C and D).

3.12. Software. All analyses were performed using analytical packages from the R/Bioconductor project (Ihaka and Gentleman, 1996, Gentleman et al., 2004), including limma (Smyth et al., 2005), affy (Gautier et al., 2004), and multtest. Addtional functions and methods developed by Dr. Marchionni are implemented in the moreFGS and funcBox packages, which were already applied in to compare gene expression data in the prostate (Schaeffer et al., 2008), and are available at http://astor.som.jhmi.edu/~marchion/labLM/packages.html.

4. Proliferation and clonogenic assay. Cell proliferation index and cell viability were assessed by staining with trypan blue (Invitrogen) and by counting in a Thoma chamber.
For clonogenic test, LNCaP cells were seeded in quintuplicate in 60-mm Petri dishes, at 250-500 cells/plate density in complete medium. After 15 days, cells were fixed and stained with 0.5% crystal violet / 80% methanol, and counted. Colonies with a cell number > 50 were included in the count. Clonogenic ability was calculated as the average (expressed in percentage) of the ratio of grown colonies out of the total number of cells seeded in each plate.

5. Analysis of mRNA signature of distinct cancer invasion models.

5.1. Migration efficiency/velocity in Boyden assays. Upon a review of the literature (Sheng et al., 1996; Sieuwerts et al., 1997; Laniado et al., 1997; Rochefort et al., 1998; Karadag et al., 2004; Adachi et al., 2009; Bowen et al., 2009; Pan et al., 2010; Tarantola et al., 2010), we identified a number of cell lines from various cancer types displaying extreme differences in migration efficiency/velocity as assayed in Boyden assays, as reported in the following Supplementary Tables 19-21

Supplementary Table 8: Cell lines migration efficiency/velocity in Boyden assays

Cell line	Boyden	Tissue	PMID
PC3	Fast	Prostate	15199115 ; 9094978
HOP-62	Fast	Lung	19362386
HCT116	Fast	Colon	20032390
HT29	Fast	Colon	20473392 ; 19493905
MDA-MB-231	Fast	Breast	9699869 ; 9009106 ; 8876194 ; 15199115
MDA-MB-435S	Fast	Breast	9699869 ; 8876194
LNCAP	Slow	Prostate	8876194 ; 9094978
EKVX	Slow	Lung	19362386
SW480	Slow	Colon	19493905 ; 20032390
MCF7	Slow	Breast	9699869 ; 9009106
TD47	Slow	Breast	9699869

Supplementary Table 9: Affymetrix CEL files retrieved from the NCBI GEO expression database corresponding to fast (PC3) and slow (LNCAP) migrating prostate cancer cell lines in Boyden assays.

Filename	DataSet	CellLine	Tissue	Boyden
GSM133661.CEL.gz	GSE5720	PC3	Prostate	FAST
GSM86079.CEL.gz	GSE3737	PC3	Prostate	FAST
GSM86080.CEL.gz	GSE3737	PC3	Prostate	FAST
GSM86081.CEL.gz	GSE3737	PC3	Prostate	FAST
GSM86082.CEL.gz	GSE3737	PC3	Prostate	FAST
GSM310120.CEL.gz	GSE12348	PC3	Prostate	FAST
GSM63051.CEL.gz	GSE4636	LNCaP	Prostate	SLOW
GSM64845.CEL.gz	GSE4636	LNCaP	Prostate	SLOW
GSM64855.CEL.gz	GSE4636	LNCaP	Prostate	SLOW
GSM64858.CEL.gz	GSE4636	LNCaP	Prostate	SLOW
GSM64861.CEL.gz	GSE4636	LNCaP	Prostate	SLOW
GSM64864.CEL.gz	GSE4636	LNCaP	Prostate	SLOW
GSM301120.CEL.gz	GSE11914 ; GSE11915	LNCaP	Prostate	SLOW
GSM301121.CEL.gz	GSE11914 ; GSE11915	LNCaP	Prostate	SLOW
GSM310084.CEL.gz	GSE12348	LNCaP	Prostate	SLOW
GSM231996.CEL.gz	GSE9182	LNCaP	Prostate	SLOW
GSM231997.CEL.gz	GSE9182	LNCaP	Prostate	SLOW

Supplementary Table 10: Affymetrix CEL files retrieved from the NCBI GEO expression database corresponding to fast and slow migrating cancer cell lines in Boyden assays.

FileName	DataSet	CellLine	Tissue	Boyden
GSM133661.CEL.gz	GSE5720	PC3	Prostate	FAST
GSM310120.CEL.gz	GSE12348	PC3	Prostate	FAST
GSM133591.CEL.gz	GSE5720	HOP-62	Lung	FAST
GSM133592.CEL.gz	GSE5720	HCT116	Colon	FAST
GSM450448.CEL.gz	GSE18005	HCT116	Colon	FAST
GSM455560.CEL.gz	GSE18232	HCT116	Colon	FAST
GSM133562.CEL.gz	GSE5720	HT29	Colon	FAST
GSM450453.CEL.gz	GSE18005	HT29	Colon	FAST
GSM455566.CEL.gz	GSE18232	HT29	Colon	FAST
GSM133644.CEL.gz	GSE5720	MDA-MB-231	Breast	FAST
GSM421873.CEL.gz	GSE16795	MDA-MB-231	Breast	FAST
GSM133628.CEL.gz	GSE5720	MDA-MB-435S	Breast	FAST
GSM421877.CEL.gz	GSE16795	MDA-MB-435S	Breast	FAST
GSM310084.CEL.gz	GSE12348	LNCaP	Prostate	SLOW
GSM310084.CEL.gz	GSE12348	LNCaP	Prostate	SLOW
GSM133581.CEL.gz	GSE5720	EKVX	Lung	SLOW
GSM133581.CEL.gz	GSE5720	EKVX	Lung	SLOW
GSM450458.CEL.gz	GSE18005	SW480	Colon	SLOW
GSM455572.CEL.gz	GSE18232	SW480	Colon	SLOW
GSM450458.CEL.gz	GSE18005	SW480	Colon	SLOW
GSM455572.CEL.gz	GSE18232	SW480	Colon	SLOW
GSM133558.CEL.gz	GSE5720	MCF7	Breast	SLOW
GSM421869.CEL.gz	GSE16795	MCF7	Breast	SLOW
GSM133558.CEL.gz	GSE5720	MCF7	Breast	SLOW
GSM421869.CEL.gz	GSE16795	MCF7	Breast	SLOW
GSM133601.CEL.gz	GSE5720	T47D	Breast	SLOW
GSM421894.CEL.gz	GSE16795	T47D	Breast	SLOW
GSM133601.CEL.gz	GSE5720	T47D	Breast	SLOW
GSM421894.CEL.gz	GSE16795	T47D	Breast	SLOW

5.2. Three-dimensional model of prostate cancer invasion.

Supplementary Table 11: Description of prostasphere and cell culture samples analyzed in the GSE19426 series, corresponding to distinct prostate cancer growth and invasion phenotypes (Harma et al., 2010) .

GEOid	CellLine	CultureType	Origin	Time3D	BioRep	Pheno	OtherInfo
GSM483220	PC3	D2	AdenoCA	NA	rep1	D2	BoneMetastasis
GSM483221	PC3	D2	AdenoCA	NA	rep2	D2	BoneMetastasis
GSM483222	PC3	D3	AdenoCA	day8	rep1	Round	BoneMetastasis
GSM483223	PC3	D3	AdenoCA	day8	rep2	Round	BoneMetastasis
GSM483224	PC3	D3	AdenoCA	day13	rep1	Stellate	BoneMetastasis
GSM483225	PC3	D3	AdenoCA	day13	rep2	Stellate	BoneMetastasis
GSM483226	PC3	D3	AdenoCA	day15	rep1	Stellate	BoneMetastasis
GSM483227	PC3	D3	AdenoCA	day15	rep2	Stellate	BoneMetastasis
GSM483228	ALVA31	D3	AdenoCA	day11	rep1	Stellate	PC3derivative
GSM483229	ALVA31	D3	AdenoCA	day11	rep2	Stellate	PC3derivative
GSM483230	LNCaP	D3	AdenoMET	day11	rep1	Mass	LymphNodeMetastasis
GSM483231	LNCaP	D3	AdenoMET	day11	rep2	Mass	LymphNodeMetastasis
GSM483232	EP156T	D3	Immortalized	day11	rep1	Round	hTERT
GSM483233	EP156T	D3	Immortalized	day11	rep2	Round	hTERT
GSM483234	RWPE1	D3	Immortalized	day11	rep1	Branching	HPV18
GSM483235	RWPE1	D3	Immortalized	day11	rep2	Branching	HPV18
GSM483236	22rv1	D3	PCA	day11	rep1	Mass	CastrationResistant
GSM483237	22rv1	D3	PCA	day11	rep2	Mass	CastrationResistant
GSM483238	DU145	D3	AdenoMET	day11	rep1	Round	BrainMetastasis
GSM483239	DU145	D3	AdenoMET	day11	rep2	Round	BrainMetastasis
GSM483240	RWPE2/w99	D3	Transformed	day11	rep1	Stellate	HighKiRas
GSM483241	RWPE2/w99	D3	Transformed	day11	rep2	Stellate	HighKiRas
GSM483242	22rv1	D2	PCA	NA	rep1	D2	CastrationResistant
GSM483243	22rv1	D2	PCA	NA	rep2	D2	CastrationResistant
GSM483244	DU145	D2	AdenoMET	NA	rep1	D2	BrainMetastasis
GSM483245	DU145	D2	AdenoMET	NA	rep2	D2	BrainMetastasis
GSM483246	ALVA31	D2	AdenoCA	NA	rep1	D2	PC3derivative
GSM483247	ALVA31	D2	AdenoCA	NA	rep2	D2	PC3derivative
GSM483248	LNCaP	D2	AdenoMET	NA	rep1	D2	LymphNodeMetastasis
GSM483249	LNCaP	D2	AdenoMET	NA	rep2	D2	LymphNodeMetastasis
GSM483250	RWPE2/w99	D2	Transformed	NA	rep1	D2	HighKiRas
GSM483251	RWPE2/w99	D2	Transformed	NA	rep2	D2	HighKiRas
GSM483252	RWPE1	D2	Immortalized	NA	rep1	D2	HPV18
GSM483253	RWPE1	D2	Immortalized	NA	rep2	D2	HPV18
GSM483254	EP156T	D2	Immortalized	NA	rep1	D2	hTERT
GSM483255	EP156T	D2	Immortalized	NA	rep2	D2	hTERT
GSM483256	PC3	D3	AdenoCA	day4	rep1	Round	BoneMetastasis
GSM483257	PC3	D3	AdenoCA	day4	rep2	Round	BoneMetastasis
GSM483258	PC3M	D2	AdenoMET	NA	rep1	D2	HighlyMetastatic
GSM483259	PC3M	D2	AdenoMET	NA	rep2	D2	HighlyMetastatic
GSM483260	PC3M	D3	AdenoMET	day4	rep1	Round	HighlyMetastatic
GSM483261	PC3M	D3	AdenoMET	day4	rep2	Round	HighlyMetastatic
GSM483262	PC3M	D3	AdenoMET	day11	rep1	Stellate	HighlyMetastatic
GSM483263	PC3M	D3	AdenoMET	day11	rep2	Stellate	HighlyMetastatic
GSM483264	RWPE1	D3	Immortalized	day5	rep1	Branching	HPV18
GSM483265	RWPE1	D3	Immortalized	day5	rep2	Branching	HPV18
GSM483266	EP156T	D3	Immortalized	day5	rep1	Round	hTERT

GSM483267	EP156T	D3	Immortalized	day5	rep2	Round	hTERT
GSM483268	PrEC	D2	Normal	NA	rep1	D2	PrimaryCells
GSM483269	PrEC	D2	Normal	NA	rep2	D2	PrimaryCells
GSM483270	PrEC	D3	Normal	day5	rep1	Branching	PrimaryCells
GSM483271	PrEC	D3	Normal	day5	rep2	Branching	PrimaryCells
GSM483272	PrEC	D3	Normal	day12	rep1	Branching	PrimaryCells
GSM483273	PrEC	D3	Normal	day12	rep2	Branching	PrimaryCells

SUPPLEMENTARY REFERENCES

Adachi Y, Li R, Yamamoto H, Min Y, Piao W, Wang Y, et al. Insulin-like growth factor-I receptor blockade reduces the invasiveness of gastrointestinal cancers via blocking production of matrilysin. Carcinogenesis. 2009;30(8):1305-13.

Bader, G. D., Donaldson, I., Wolting, C., Ouellette, B. F., Pawson, T., and Hogue, C. W. (2001). BIND–The Biomolecular Interaction Network Database. Nucleic Acids Res, 29(1):242–245.

Bowen KA, Doan HQ, Zhou BP, Wang Q, Zhou Y, Rychahou PG, et al. PTEN loss induces epithelial--mesenchymal transition in human colon cancer cells. Anticancer Res. 2009;29(11):4439-49. PMCID: 2932708.

Brazma, A., Hingamp, P., Quackenbush, J., Sherlock, G., Spellman, P., Stoeckert, C., Aach, J., Ansorge, W., Ball, C. A., Causton, H. C., Gaasterland, T., Glenisson, P., Holstege, F. C., Kim, I. F., Markowitz, V., Matese, J. C., Parkinson, H., Robinson, A., Sarkans, U., Schulze-Kremer, S., Stewart, J., Taylor, R., Vilo, J., and Vingron, M. (2001). Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. Nat Genet, 29(4):365–71. 1061-4036 (Print) Journal Article.

Cartharius, K., Frech, K., Grote, K., Klocke, B., Haltmeier, M., Klingenhoff, A., Frisch, M., Bayerlein, M., and Werner, T. (2005). Matinspector and beyond: promoter analysis based on transcription factor binding sites. Bioinformatics, 21(13):2933–2942.

Gautier, L., Cope, L., Bolstad, B. M., and Irizarry, R. A. (2004). affy–analysis of Affymetrix GeneChip data at the probe level. Bioinformatics, 20(3):307–15. 1367-4803 (Print) Evaluation Studies Journal Article.

Harma V, Virtanen J, Makela R, Happonen A, Mpindi JP, Knuuttila M, et al. A comprehensive panel of three-dimensional models for studies of prostate cancer growth, invasion and drug responses. PLoS One. 2010;5(5):e10431. PMCID: 2862707.

Irizarry, Rafael A, Warren, Daniel, Spencer, Forrest, Kim, Irene F, Biswal, Shyam, Frank, Bryan C, Gabrielson, Edward, Garcia, Joe G N, Geoghegan, Joel, Germino, Gregory, Griffin, Constance, Hilmer, Sara C, Hoffman, Eric, Jedlicka, Anne E, Kawasaki, Ernest, Martínez-Murillo, Francisco, Morsberger, Laura, Lee, Hannah, Petersen, David, Quackenbush, John, Scott, Alan, Wilson, Michael, Yang, Yanqin, Ye, Shui Qing, and Yu, Wayne (2005). Multiple-laboratory comparison of microarray platforms. Nat Methods, 2(5):345–350.

John, Bino, Enright, Anton J, Aravin, Alexei, Tuschl, Thomas, Sander, Chris, and Marks, Debora S (2004). Human microrna targets. PLoS Biol, 2(11):e363.

Karadag A, Ogbureke KU, Fedarko NS, Fisher LW. Bone sialoprotein, matrix metalloproteinase 2, and alpha(v)beta3 integrin in osteotropic cancer cell invasion. J Natl Cancer Inst. 2004;96(12):956-65.

Kent, W. James, Sugnet, Charles W, Furey, Terrence S, Roskin, Krishna M, Pringle, Tom H, Zahler, Alan M, and Haussler, David (2002). The human genome browser at ucsc. Genome Res, 12(6):996–1006.

Kiriakidou, Marianthi, Nelson, Peter T, Kouranov, Andrei, Fitziev, Petko, Bouyioukos, Costas, Mourelatos, Zissimos, and Hatzigeorgiou, Artemis (2004). A combined computational-experimental approach predicts human microrna targets. Genes Dev, 18(10):1165–1178.

Krek, Azra, Grün, Dominic, Poy, Matthew N, Wolf, Rachel, Rosenberg, Lauren, Epstein, Eric J, MacMenamin, Philip, da Piedade, Isabelle, Gunsalus, Kristin C, Stoffel, Markus, and Rajewsky, Nikolaus (2005). Combinatorial microrna target predictions. Nat Genet, 37(5):495–500.

Kuhn, R. M., Karolchik, D., Zweig, A. S., Wang, T., Smith, K. E., Rosenbloom, K. R., Rhead, B., Raney, B. J., Pohl, A., Pheasant, M., Meyer, L., Hsu, F., Hinrichs, A. S., Harte, R. A., Giardine, B., Fujita, P., Diekhans, M., Dreszer, T., Clawson, H., Barber, G. P., Haussler, D., and Kent, W. J. (2009). The ucsc genome browser database: update 2009. Nucleic Acids Res, 37(Database issue):D755–D761.

Laniado ME, Lalani EN, Fraser SP, Grimes JA, Bhangal G, Djamgoz MB, et al. Expression and functional analysis of voltage-activated Na+ channels in human prostate cancer cell lines and their contribution to invasion in vitro. Am J Pathol. 1997;150(4):1213-21. PMCID: 1858184.

Lewis, Benjamin P, Burge, Christopher B, and Bartel, David P (2005). Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microrna targets. Cell, 120(1):15–20.

Lonnstedt, I. and Speed, T. P. (2002). Replicated microarray data. Statistica Sinica, 12:31–46.

Maragkakis, M., Reczko, M., Simossis, V. A., Alexiou, P., Papadopoulos, G. L., Dalamagas, T., Giannopoulos, G., Goumas, G., Koukis, E., Kourtis, K., Vergoulis, T., Koziris, N., Sellis, T., Tsanakas, P., and Hatzigeorgiou, A. G. (2009). Diana-microt web server: elucidating microrna functions through target prediction. Nucleic Acids Res, 37(Web Server issue):W273–W276.

Pan SH, Chao YC, Chen HY, Hung PF, Lin PY, Lin CW, et al. Long form collapsin response mediator protein-1 (LCRMP-1) expression is associated with clinical outcome and lymph node metastasis in non-small cell lung cancer patients. Lung Cancer. 2010;67(1):93-100.

Peri, Suraj, Navarro, J. Daniel, Kristiansen, Troels Z, Amanchy, Ramars, Surendranath, Vineeth, Muthusamy, Babylakshmi, Gandhi, T. K B, Chandrika, K. N., Deshpande, Nandan, Suresh, Shubha, Rashmi, B. P., Shanker, K., Padma, N., Niranjan, Vidya, Harsha, H. C., Talreja, Naveen, Vrushabendra, B. M., Ramya, M. A., Yatish, A. J., Joy, Mary, Shivashankar, H. N., Kavitha, M. P., Menezes, Minal, Choudhury, Dipanwita Roy, Ghosh, Neelanjana, Saravana, R., Chandran, Sreenath, Mohan, Sujatha, Jonnalagadda, Chandra Kiran, Prasad, C. K., Kumar-Sinha, Chandan, Deshpande, Krishna S, and Pandey, Akhilesh (2004). Human protein reference database as a discovery resource for proteomics. Nucleic Acids Res, 32(Database issue):D497–D501.

Prestridge, D. S. (1995). Predicting pol ii promoter sequences using transcription factor binding sites. J Mol Biol, 249(5):923–932.

Quandt, K., Frech, K., Karas, H., Wingender, E., and Werner, T. (1995). Matind and matinspector: new fast and versatile tools for detection of consensus matches in nucleotide sequence data. Nucleic Acids Res, 23(23):4878–4884.

Rochefort H, Platet N, Hayashido Y, Derocq D, Lucas A, Cunat S, et al. Estrogen receptor mediated inhibition of cancer cell invasion and motility: an overview. J Steroid Biochem Mol Biol. 1998;65(1-6):163-8

Schaeffer EM, Marchionni L, Huang Z, Simons B, Blackman A, Yu W, et al. Androgen-induced programs for prostate epithelial growth and invasion arise in embryogenesis and are reactivated in cancer. Oncogene. 2008;27(57):7180-91. PMCID: 2676849.

Scharpf, Robert B, Iacobuzio-Donahue, Christine A, Sneddon, Julie B, and Parmigiani, Giovanni (2007). When should one subtract background fluorescence in 2-color microarrays? Biostatistics, 8(4):695–707.

Segal, Eran, Friedman, Nir, Kaminski, Naftali, Regev, Aviv, and Koller, Daphne (2005). From signatures to models: understanding cancer using microarrays. Nat Genet, 37 Suppl:S38–S45.

Segal, Eran, Friedman, Nir, Koller, Daphne, and Regev, Aviv (2004). A module map showing conditional activity of expression modules in cancer. Nat Genet,36(10):1090–1098.

Sheng S, Carey J, Seftor EA, Dias L, Hendrix MJ, Sager R. Maspin acts at the cell membrane to inhibit invasion and motility of mammary and prostatic cancer cells. Proc Natl Acad Sci U S A. 1996;93(21):11669-74. PMCID: 38116.

Sieuwerts AM, Klijn JG, Foekens JA. Assessment of the invasive potential of human gynecological tumor cell lines with the in vitro Boyden chamber assay: influences of the ability of cells to migrate through the filter membrane. Clin Exp Metastasis. 1997;15(1):53-62.

Smyth, G. K. (2004). Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. Statistical Applications in Genetics and Molecular Biology, 3(Article 3).

Smyth, G. K. (2005). Limma: linear models for microarray data. In Gentleman, R., Carey, R. V., Dudoit, S., Irizarry, R., and Huber, W., editors, Bioinformatics and Computational Biology Solutions using R and Bioconductor, pages 397–420. Springer, New York.

Smyth, G. K., Michaud, J., and Scott, H. S. (2005). Use of within-array replicate spots for assessing differential expression in microarray experiments. Bioinformatics, 21(9):2067–75. 1367-4803 (Print) Evaluation Studies Journal Article Validation Studies.

Smyth, G. K. and Speed, T. (2003). Normalization of cDNA microarray data. Methods, 31(4):265–73. 1046-2023 (Print) Journal Article.

Stark, Chris, Breitkreutz, Bobby-Joe, Reguly, Teresa, Boucher, Lorrie, Breitkreutz, Ashton, and Tyers, Mike (2006). BioGRID: a general repository for interaction datasets. Nucleic Acids Res, 34(Database issue):D535–D539.

Tarantola M, Marel AK, Sunnick E, Adam H, Wegener J, Janshoff A. Dynamics of human cancer cell lines monitored by electrical and acoustic fluctuation analysis. Integr Biol (Camb). 2010;2(2-3):139-50.

Venables, W. N. and Ripley, B. D. (2002). Modern Applied Statistics with S. Statistics and Computing. Springer, fourth edition. ISBN 0-387-95457-0.

Wheeler, David L, Barrett, Tanya, Benson, Dennis A, Bryant, Stephen H, Canese, Kathi, Chetvernin, Vyacheslav, Church, Deanna M, DiCuccio, Michael, Edgar, Ron, Federhen, Scott, Geer, Lewis Y, Kapustin, Yuri, Khovayko, Oleg, Landsman, David, Lipman, David J, Madden, Thomas L, Maglott, Donna R, Ostell, James, Miller, Vadim, Pruitt, Kim D, Schuler, Gregory D, Sequeira, Edwin, Sherry, Steven T, Sirotkin, Karl, Souvorov, Alexandre, Starchenko, Grigory, Tatusov, Roman L, Tatusova, Tatiana A, Wagner, Lukas, and Yaschenko, Eugene (2007). Database resources of the National Center for Biotechnology Information. Nucleic Acids Res, 35(Database issue):D5–12.

Yang, Y. H., Dudoit, S., Luu, P., Lin, D. M., Peng, V., Ngai, J., and Speed, T. P. (2002). Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. Nucleic Acids Res, 30(4):e15. 1362-4962 (Electronic) Journal Article.

Yang, Y. H. and Thorne, N. (2003). Normalization for two-color cDNA microarray data. In Goldstein, D. R., editor, Science and Statistics: A Festschrift for Terry Speed, volume 40, pages 403–41.

Zhang, J., Carey, V., and Gentleman, R. (2003). An extensible application for assembling annotation for genomic data. Bioinformatics, 19(1):155–6. 1367-4803 (Print) Journal Article.