Supporting Information

Demichelis et al. 10.1073/pnas.1117405109

SI Methods

Tyrol Prostate-Specific Antigen Screening Cohort. ***Cohort description.*** Tyrol Prostate-Specific Antigen (PSA) Screening Cohort is a cohort of men tested with PSA for early detection and treatment of prostate cancer. Population-based PSA testing in asymptomatic men started in 1993 and intended to evaluate the utility of intensive PSA screening in the reduction of prostate cancer-specific death. Regular annual PSA testing for the early detection and treatment of prostate cancer was offered free of charge to men aged 45–75 (and men aged 40–75 in case of a family history of prostate cancer) in Tyrol, a federal state of Austria (1–4). More than 86% of men in this age group participated in this program, resulting in increased detection rates for locally confined prostate cancer (starting from 30% to around 85%) (1–3, 5). PSA serum levels measurements were used as the only determinant to detect men at risk for prostate cancer and recommend further urological examination. Initially, age-related total PSA levels according to Oesterling et al. (6) were used. In 1995, PSA reference levels were cut by half (bisected age-related reference ranges) and PSA above 1.25 ng/mL became the lowest cutoff level (7). The age-related normal reference ranges then were 0–1.25 ng/mL for 40–49 y, 0–1.75 ng/mL for 50–59 y, 0–2.25 ng/mL for 60–69 y, and 0–3.25 ng/mL for 70–79 y free of PSA (fPSA) measurements were added in 1995. The f/tPSA ratio was used to triage a PSA gray zone with the age-related cutoff value as the lower limit and 3.25 ng/mL as the upper limit. Subsequent urological examination and prostate biopsy was recommended with a serum PSA level of $\geq$ 3.25 ng/mL, with a gray zone PSA value wherein further urological examination was recommended in case of f/tPSA of $\leq$ 18%; in case of a gray zone PSA and f/tPSA > 18%, the interval to the next PSA measurement was reduced to half a year. Further urological evaluation consisted of a digital rectal examination, ultrasonography, and transrectal ultrasound-guided prostate needle biopsies. Initially sextant biopsies were used; since 1995, 10 systematic biopsies were taken, and since 2000, 10 systematic cores and up to 5 additional cores were taken from hypervascular prostate areas identified by microbubble-enhanced color Doppler ultrasound (8). Biopsies were processed using a flat embedding technique (9) and analyzed by experienced genitourinary pathologists at the University of Innsbruck (Austria). About one third of the prostate cancer cases were detected in men with a PSA level of 2.0–4.0 ng/mL (10).

Cases were defined as men with biopsy-confirmed prostate cancer. Controls were defined as men with a benign prostate biopsy result and no cancer diagnosis in available follow-up data. Clinical data including histopathology of biopsies and radical prostatectomy specimens, PSA, and fPSA serum levels were retrieved from the Prostate Cancer Database of the Tyrol Early Prostate Cancer Detection Program. All of the samples and data have been maintained in a central location since 1993. Moreover, the Tyrol population has a low emigration rate from the region, making follow-up protocols that are currently in place more valuable over time. The demographics of trial men included in this study are presented in Dataset S1, Table S1. The study and the use of anonymized clinical data and archived DNA samples for the study were approved by the Ethics Committee of the Innsbruck Medical University and the Institutional Review Board of Weill Cornell Medical College.

Tyrol sample collection and DNA preparation. Samples were obtained from the Prostate Biorepository of Innsbruck that has been established in association with the Tyrol Early Prostate Cancer Detection Program since 1993. Blood samples of participants were collected after informed consent and stored at −80 °C until use. Ficoll purified peripheral blood mononuclear cells were used for DNA extraction (11). Peripheral blood mononuclear cells were diluted in 400 μL of chilled PBS, equilibrated to room temperature and processed using the QIAamp 96 DNA Blood kit (Qiagen).

After applying strict DNA and data-quality filters, 1,903 unrelated individuals of Caucasian origin (867 cases and 1,036 controls) were included in the study cohort. The clinical study and the use of anonymized data and DNA samples for the study were approved by the Ethics Committee of the Innsbruck Medical University and the Institutional Review Board of Weill Cornell Medical College.

Early Detection Research Network PSA Screening Cohorts. ***Cohort description.*** The Early Detection Research Network (EDRN) (http://edrn.nci.nih.gov/) is charged with the discovery, development, and validation of biomarkers related to neoplastic disease. This EDRN Prostate Cancer Clinical Validation Center includes the three institutions of Beth Israel Deaconess Medical Center ("Harvard," Harvard University, Cambridge, MA), the University of Michigan ("Michigan," Ann Arbor, MI), and Weill Cornell Medical College ("Cornell," New York, NY), prospectively enrolling men at risk for prostate cancer in three catchment areas in the United States: Boston, MA, Southeast Michigan, and New York, NY, respectively. Using a common research protocol, based on the eligibility criteria described below, men are enrolled and consented for biomarker development studies. All men underwent prostate needle biopsy and were followed on this protocol based on local clinical standards. For this cohort, cases are defined as men diagnosed with prostate cancer and controls are men who have undergone prostate needle biopsy without any detectable prostate cancer and no prior history of prostate cancer. The eligibility criteria for the EDRN Prostate Biopsy Cohort include: (*i*) Male over age 40; (*ii*) Patient scheduled for prostate biopsy for any of the following reasons: (*a*) PSA > 2.5 ng/mL, (*b*) rising PSA (>0.5 ng/mL/yr), (*c*) lower PSA value with other risk factors for prostate cancer (e.g., family history), (*d*) abnormal digital rectal examination, (*e*) percent fPSA <15%; (*iii*) No prior history of prostate cancer or prostate biopsy; (*iv*) Prostate biopsy with at least 10 cores taken in a laterally directed fashion; (*v*) Blood collected before prostate biopsy; (*vi*) Prostate biopsy pathology report available. Individual ethnicity information was collected at time of patient enrollment. The demographics of men enrolled in this prospective cohort are presented in Dataset S1, Table S5.

EDRN sample collection and DNA preparations. A total of 994 subject samples were collected from the three EDRN institutions: 399 from Michigan, 380 from Harvard, and 215 from Cornell. Collected samples comprised of pre-extracted DNA (227 from Michigan) or 100–200 μL aliquot of blood cellular-EDTA samples prepared according to EDRN standard operating procedures.

Genomic DNA was extracted from the blood cellular-EDTA samples in a high-throughput fashion using the QIAamp 96 DNA Blood Kit (Qiagen). Each sample aliquot was manually resuspended to 200 μL total volume with chilled phosphate buffer saline (pH 7.0). After equilibration to room temperature, the aliquots were processed according to the QIAamp 96 DNA Blood purification protocol and DNAs were eluted with 100 μL of nuclease-free water.

All DNAs were evaluated by NanoDrop spectrophotometer (NanoDrop, Thermo Scientific) and gel electrophoresis (2%

agarose). Of the 994 samples collected, 972 (394 Michigan, 363 Harvard, and 215 Cornell) passed quality control with at least 500 ng of DNA quantity; 22 failed extraction and 1 was a duplicate. For TaqMan Real-Time Quantitative PCR, each DNA sample was diluted to 10 ng/μL with nuclease-free water. A total of 800 individuals were Caucasian (based on self-declaration) and were included in the validation study.

Affymetrix Genome-Wide Human SNP 6.0 Profiling and Data Preprocessing. Blood DNA from the Tyrol cohort samples was profiled using the Affymetrix 6.0 Whole Genome SNP Array platform (Affymetrix) according to manufacturer's instructions. Raw data were preprocessed as in Oldridge et al. (12). Data quality control included call rate, and intra/interchip variability. Preprocessed data were segmented (13). Principal component analysis was performed (14), including HapMap Phase II sample genotypes to assess population stratification within the Tyrol cohort. Individuals who largely diverged from the CEU cluster were excluded from primary analysis before any downstream analysis. To check for relatedness within the Tyrol cohort, we applied an intersample genetic distance test (15): 0.38 is the average distance between unrelated individuals. A cluster spanning the range 0.19–0.28 revealed first-degree relatives (e.g., fraternal brothers or father/son pairs) consistent with CEU HapMap trio distances and a singleton around zero revealed two identical twins. One individual for each related pair was retained in the downstream analysis.

Human Prostate Samples. Localized tumors ($n = 50$) and benign tissues ($n = 10$) were collected at time of radical prostatectomy (16). De-identified frozen metastatic sample were obtained as part of an ongoing clinical study on metastatic prostate cancer (17). All cases were reviewed by the study pathologist (M.A.R.). RNAseq expression level quantification was performed as in refs. 17 and 16; transcript levels are quantified in reads per kilobase of exon model per million mapped reads (18). Analysis for transcript of interest versus copy number states from the same individuals was performed as in Banerjee et al. (19).

Risk SNP Selection and Association Analysis. The initial list of 57 published prostate cancer risk SNPs was compiled based upon four genome-wide association studies (GWAS) (20–23), two follow-up GWAS (24, 25), two replication studies (26, 27), one study reporting on a new region of 8q24 (28), and one fine-mapping study (29). SNPs reported from two earlier GWAS studies (30, 31) are included in Zheng et al. (27), which we consider in our compiled list. For SNPs that have been reported in multiple studies, we gave preference to the results of the most recent or comprehensive study. Dataset S1, Table S2 includes the list of studies (annotated by PubMedID) and reports the association information from the original study for each SNP. For the SNPs that are not represented on the Affymetrix platform ($n = 36$), tag SNPs were identified by maximum SNP-SNP pair-wise r^2 value from the HapMap Consortium annotation of pair-wise linkage disequilibrium statistics in a Utah population (CEU) with northern European ancestry (http://hapmap.ncbi.nlm.nih.gov/downloads/ld_data/2009-04_rel27/). After exclusion of one redundant tag SNP, a total of 56 SNPs were queried in the Tyrol cohort for association. Tag SNP alleles corresponding to the risk allele were called considered genotype data across 165 CEU HapMap individuals (ftp://ftp.ncbi.nlm.nih.gov/hapmap/genotypes/2009-01_phaseIII/hapmap_format/polymorphic). All SNPs passed the following criteria: minor allele frequency ≥ 1%; P value for the test for Hardy–Weinberg equilibrium greater or equal to 1e-6, and SNP-call rate ≥ 98% (genome-wide 872,055 SNPs passed the criteria). SNP association analysis was performed using PLINK (32). We tested the selected SNPs for association with prostate cancer risk and with aggressive prostate cancer risk (no covariates of age or age and PSA) according to allelic, genotypic, dominant, and recessive logistic models in the Tyrol cohort. We considered a SNP to be concordantly associated so long as at least one of the above test P value was ≤ 0.05 and the risk allele identified within the Tyrol cohort was consistent with the reported risk allele (otherwise inconsistent). Three SNPs had insufficient information to assess risk allele concordance (one because of ambiguous annotation in original study and two because of lack of phase information for the tagSNP).

Copy Number Variant Detection Approach Comparison. Detection of copy number variant (CNV) is not as refined as the detection of SNPs. A recent comprehensive assessment of array-based platforms and CNV detection algorithms evaluated high variability between calling algorithms (33), showing that different analytic tools applied to the same exact raw data yield CNV calls with less than 50% concordance. The authors suggest that using multiple detection approach consensus minimizes the number of false discoveries, but significantly reduces sensitivity. Because of the current lack of a gold standard for CNVs, disease-association studies require CNV assessment by independent techniques (e.g., quantitative PCR, DNA-sequencing). We recently implemented an algorithm called Identification of germline Changes in Copy Number, IgC2N, (19) that takes advantage of the polymorphic copy number signal across the entire sample set under study, similar to the work from Beroukhim et al. (34), and helps improve upon CNV detection. This approach minimizes false discovery for CNVs by taking into account multiple samples. In our previous work, we estimated the expected power of the approach by in silico simulation (as function of sample size, variant size, and minor allele frequency), characterized HapMap III samples for CNVs using Affymetrix 6.0 data and validated newly detected variants by CGH data from a 42 million probe dataset (35). To further benchmark the CNV detection algorithm and genotyping approach in the current study, we quantitatively compared its performance with the data from Conrad et al. (35), verified copy number states using high coverage DNA sequencing data (36, 37), and performed quantitative PCR on a selected set of CNVs. The comparison sample set included 477 HapMap samples. The percentages of CNVs detected by Banerjee et al. (19) and included in Conrad et al. (35) and vice versa were 61.2% and 66.2%, respectively, for CNVs covered by at least 10 probes, and 77.7% and 70.3% with at least 20 probes. The genotype concordance was estimated on average equal to 82.6% (SD = 1.8%) (Fig. S5). These results are in line with the across-platform and across-detection approaches recently reported by Pinto et al. (33). Importantly, our detection approach differs from the majority of germ line CNV detection approaches in that it works across samples, and therefore its performance strictly depends on the study sample size, as we quantified through the original power simulation study. With the intent to ultimately query a comprehensive and well-characterized set of CNVs, we combined the variants detected by the across-sample approach from Banerjee (19) with additional variants from Conrad et al. (35) for a total number of 2,611 CNVs.

Quantitative PCR Validation for CNVs. Index samples were selected using HapMap samples Affymetrix array data and 1000 Genomes Project sequence data (37) (http://www.1000genomes.org) was used for determining CNV breakpoints for efficient primer design (38). Genomic profiles were visualized using the Integrative Genomics Viewer (http://www.broadinstitute.org/igv/), to determine the approximate breakpoints from read depth.

The quantitative PCR assays developed for the prostate cancer risk CNVs at 15q21.3 and 12q21.31 were used on the Early Detection Research Network (EDRN) cohort. Genomic DNA from HapMap individuals used as reference controls was obtained from Coriell Cell Repositories (http://ccr.coriell.org/). TaqMan RT-

qPCR was carried out using 7900HT Fast Real-Time PCR System (Applied Biosystems). Following the company-provided TaqMan CN Assays protocol for 384-well format, each sample was prepared in triplicate and run at least two rounds to ensure reproducibility. The EDRN subject samples were split into nine 384-well plates, with each containing HapMap reference controls (Coriell Institute for Medical Research) and no more than 120 subject samples. TaqMan Copy Number Reference Assay RNase P was used as the endogenous reference. Copy number data were called using CopyCaller v1.0 software (Applied Biosystems). Custom assay for15q21.3 probe [TCCTGAGTGCCAAAGTCC], forward primer [CTCCAAAGGCAGACTACCAAGAC], and reverse primer CGATGGCGATTTTCTCTGAAGAGTA]. Custom assay for12q21.31 probe [CCCTTCTTTGTCTCTTTTGATCTTT], forward primer [CTGTTGCATTGATCCCTTTACCATT], reverse primer [AAATAAATAAATAAATAAAGCAAGGGTTGAAACAA].

Statistical Consideration. The 238 CNVs were tested for association with prostate cancer using allelic test. Each copy number state is considered as a different category; the *P* value reported is the lowest among the different categories and the corresponding OR is reported. The logistic regression models for each CNV were repeated by adjusting for age and preoperative PSA level (categorized as <2, 2–4, >4–10, and > 10 ng/mL). Multiple-hypothesis testing correction was evaluated by calculating the false-discovery rate (39) on the *P* values of the Wald test for the coefficients of each logistic regression model. If any CN class had less than 1% frequency either in cases or controls, association tests for that CNV was not evaluated as the asymptotic approximation of the Wald's *z*-statistic might not be valid (40). For the CNVs queried in the EDRN cohort, we ran the tests mentioned above for prostate cancer risk and aggressive prostate cancer risk versus controls adjusting for age and PSA density. The *P* value for differential expression of MGAT4C between localized and metastatic human prostate samples was evaluated based on 10,000 permutations. The tool DAVID was used of gene set enrichment analysis (41).

Hi-C Approach. Hi-C is a unique approach that generalizes a previous target-specific approach referred to as chromosome conformation capture assay (3C) (42) to enable detection of long-range DNA interactions (43, 44). Hi-C was carried out according to the method published by Lieberman-Aiden et al. (43). Briefly, 5×10^7 RWPE1 cells were cross-linked, split into five separate aliquots (equivalent of 10^7 cells per aliquot), and lysed in lysis buffer [500 μL 10 mM Tris-HCl pH 8.0, 10 mM NaCl, 0.2% Igepal CA-630 and protease inhibitors (SigmaFast; Sigma)]. DNA was then digested with 400 Units of HindIII (New England Biolabs) overnight. One aliquot was prepared to generate a 3C library by ligation of the restricted DNA ends using T4 DNA ligase (Invitrogen) and the other four aliquots were processed to generate Hi-C libraries. For this process, restricted DNA ends were filled-in using Klenow (New England Biolabs) in the presence of biotin-14-dCTP. Quality of the fill-in and biotin-14-dCTP incorporation was assessed using PCR followed be NheI digestion, as described in Lieberman-Aiden et al. (43). Blunt-end ligation was performed and DNA was purified following proteinase K digestion using phenol extraction and ethanol precipitation. Following biotin removal from unligated DNA ends, the DNA was sheared (BioRuptor) and 300-bp to 600-bp fragments were gel-purified. Following end-repair, biotinylated DNA was purified using DynabeadsMyOne Streptavidin C1 Beads (Invitrogen). The resulting DNA was subjected to massively parallel paired-end DNA sequencing on the Illumina Genome Analyzer II platform to identify genome-wide long-range interactions using our in-house software. In brief, the paired-end reads were aligned to the reference human genome NCBI36 (hg18) using the BWA aligner (45). Reads mapping ambiguously to multiple locations on the genome were discarded. We further filtered out clonal reads caused by PCR artifacts and retained reads with consistent expected placement relative to HindIII enzyme digestion sites. Inter- and intrachromosomal interactions were extracted, visualized, and analyzed using the University of California at Santa Cruz Genome Browser (46) and in-house software.

RWPE1 ChIP-seq Experiments. ChIP-seq data from 50 million RWPE1-ERG cells were generated as previously described (47). Chromatin samples were immunoprecipitated with 5 μg of rabbit anti-Jun (Santa Cruz; sc-1694 X) antibody. Following extensive washing the DNA was eluted using 100 mM NaHCO3 and 1% SDS and the cross-links were reversed using 300 mM NaCl at 65 C° for 16 h. The eluted DNA was purified using Qiagen PCR Qiaquick kit following the manufacturer's protocol. Direct sequencing of the Input DNA and ChIP library was performed using Illumina Genome Analyzer according to standard manufacturer's procedures.

Functional Studies. ***Cell cultures and reagents.*** Human Prostate cell lines RWPE1, VCaP, LnCaP (Clone FGC) were purchased from American Type Culture Collection (ATCC); RWPE1: CRL-11609; LnCaP: CRL-1740; VCaP: CRL-2876. Cell lines were maintained in either DMEM or RPMI 1640 supplemented with 10% (vol/vol) FBS and penicillin/streptomycin according to manufacturer's instructions. RWPE1 cells were maintained in Keratinocyte serum-free medium (K-SFM; Invitrogen) supplemented with K-SFM kit (Invitrogen), and 1% penicillin/streptomycin (Life Technologies). pCMV6-XL5 vector expressing *α-1,3-mannosyl-glycoprotein 4-β-N-acetylglucosaminyltransferase C* (*MGAT4C*) were obtained from Origene.

***In vitro expression of* MGAT4C.** The prostate cells expressing *MGAT4C* were generated using FuGENE-6 transfection (Roche) and were selected using G418 selection (1 mg/mL). For LnCaP cells, electroporation (Nucleofection) was used to transfect the plasmids according to the manufacturer's instructions (Amaxa Biosystems). With this method, we achieved about 80% transfection efficiency as monitored by GFP expression. Briefly, 10^7 cells at 70% confluency were transfected with either 5 μg of empty vector (pCMV6-XL5) or appropriate vectors. The cells were initially plated in a six-well plate on poly-D-lysine–coated plates. After 24 h, the cells were transferred to regular tissue culture plates. The effects of *MGAT4C* overexpression or down-regulation were measured after 72 h by using quantitative RT-PCR. For proliferation assays, cells were seeded in 96-well tissue culture plates (1×10^4 per well) and proliferation was assessed at 0, 24, 48, and 72 h, by performing WST-1 assay (Roche). The absorbance at 450 nm was recorded according to the manufacturer's instructions. Values from four wells were obtained for each treatment.

RNA interference. For siRNA transfection, LNCaP (1×10^4 per well) and VCaP (1×10^4 per well) cells were seeded on 96-well tissue culture plates. After 24 h, cells were transfected with 100 nM *MGAT4C* SMARTpoolsiRNA (L-020586), or non-targeting (NT) siRNAs (ON-TARGETplus; Thermo Scientific) using Lipofectamine 2000 (Invitrogen). At 0, 24, 48, and 72 h, growth was assessed by performing WST-1 assay (Roche). Values from four wells were obtained for each treatment. The efficacy of the siRNA knockdown was assessed in several independent experiments by quantitative RT-PCR, and the optimal amount of siRNA used for transfection was determined to be 100 nM.

RT-PCR. The MGAT4C primer sequences used in these experiments are: Sense, ACACAGGAACACCAGATTGCCATC, and Antisense, TGCCCGCAAAGCCATGACTTG. RNA was extracted using the TRIzol reagent (Invitrogen), followed by DNase treatment (DNA-free kit; Applied Biosystems) according to the

manufacturer's instructions. Quantitative RT-PCR was performed using the ABI 7900HT Fast Real-Time PCR System (Applied Biosystems) following the manufacturer's RNA-to-CT one-step protocol. Each target gene was run in triplicate, and expression levels relative to the housekeeping gene *HMBS* were determined by relative quantitation using the comparative CT method ($2^{-\Delta\Delta CT}$).

Migration assay. For the migration assay, 1×10^6 RWPE1 and VCaP cells transfected with control vector and *MGAT4C* expression plasmid were resuspended in RPMI-1640 medium containing 1% FBS and placed into Transwell inserts (8-mm pore size) (Chemicon). The inserts were placed into wells containing RPMI, 10% FBS and epidermal growth factor (10 ng/mL). After 24 h, the cell suspension in the insert was removed by pipetting. Migrated cells on the insert were dislodged in cell detachment solution followed by lysis using Lysis buffer/Dye Solution following the manufacturer's instructions. The migration was quantitated by measuring fluorescence at 480/520 nm. A set of inserts was fixed and stained with Crystal violet 0.5% for 30 min for visualization. The migrated cells were visualized and imaged in five microscopic fields (at 20× objective magnification) per filter.

1. Bartsch G, et al.; Tyrol Prostate Cancer Screening Group (2008) Tyrol Prostate Cancer Demonstration Project: Early detection, treatment, outcome, incidence and mortality. *BJU Int* 101:809–816.
2. Bartsch G, et al.; Tyrol Prostate Cancer Screening Group (2001) Prostate cancer mortality after introduction of prostate-specific antigen mass screening in the Federal State of Tyrol, Austria. *Urology* 58:417–424.
3. Horninger W, et al. (2005) Screening for prostate cancer: Updated experience from the Tyrol study. *Can J Urol* 12(Suppl 1):7–13; discussion 92–93.
4. Reissigl A, Horninger W, Fink K, Klocker H, Bartsch G (1997) Prostate carcinoma screening in the county of Tyrol, Austria: Experience and results. *Cancer* 80:1818–1829.
5. Oberaigner W, et al. (2006) Reduction of prostate cancer mortality in Tyrol, Austria, after introduction of prostate-specific antigen testing. *Am J Epidemiol* 164:376–384.
6. Oesterling JE, Martin SK, Bergstralh EJ, Lowe FC (1993) The use of prostate-specific antigen in staging patients with newly diagnosed prostate cancer. *JAMA* 269:57–60.
7. Bartsch G, et al. (2008) Tyrol Prostate Cancer Screening Group. Tyrol Prostate Cancer Demonstration Project: Early detection, treatment, outcome, incidence and mortality. *BJU Int* 101:809–816.
8. Mitterberger M, et al. (2007) A prospective randomized trial comparing contrast-enhanced targeted versus systematic ultrasound guided biopsies: Impact on prostate cancer detection. *Prostate* 67:1537–1542.
9. Rogatsch H, et al. (2000) Diagnostic effect of an improved preembedding method of prostate needle biopsy specimens. *Hum Pathol* 31:1102–1107.
10. Pelzer AE, et al. (2005) Detection rates and biologic significance of prostate cancer with PSA less than 4.0 ng/mL: Observation and clinical implications from Tyrol screening project. *Urology* 66:1029–1033.
11. Setlur SR, et al. (2010) Genetic variation of genes involved in dihydrotestosterone metabolism and the risk of prostate cancer. *Cancer Epidemiol Biomarkers Prev* 19: 229–239.
12. Oldridge DA, Banerjee S, Setlur SR, Sboner A, Demichelis F (2010) Optimizing copy number variation analysis using genome-wide short sequence oligonucleotide arrays. *Nucleic Acids Res* 38:3275–3286.
13. Olshen AB, Venkatraman ES, Lucito R, Wigler M (2004) Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics* 5:557–572.
14. Price AL, et al. (2006) Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* 38:904–909.
15. Demichelis F, et al. (2008) SNP panel identification assay (SPIA): A genetic-based assay for the identification of cell lines. *Nucleic Acids Res* 36:2446–2456.
16. Pflueger D, et al. (2011) Discovery of non-ETS gene fusions in human prostate cancer using next-generation RNA sequencing. *Genome Res* 21:56–67.
17. Beltran H, et al. (2011) Molecular characterization of neuroendocrine prostate cancer and identification of new drug targets. *Cancer Discov* 1:487–495.
18. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods* 5:621–628.
19. Banerjee S, et al. (2011) A computational framework discovers new copy number variants with functional importance. *PLoS ONE* 6:e17539.
20. Duggan D, et al. (2007) Two genome-wide association studies of aggressive prostate cancer implicate putative prostate tumor suppressor gene DAB2IP. *J Natl Cancer Inst* 99:1836–1844.
21. Eeles RA, et al.; UK Genetic Prostate Cancer Study Collaborators; British Association of Urological Surgeons' Section of Oncology; UK ProtecT Study Collaborators (2008) Multiple newly identified loci associated with prostate cancer susceptibility. *Nat Genet* 40:316–321.
22. Gudmundsson J, et al. (2009) Genome-wide association and replication studies identify four variants associated with prostate cancer susceptibility. *Nat Genet* 41:1122–1126.
23. Gudmundsson J, et al. (2008) Common sequence variants on 2p15 and Xp11.22 confer susceptibility to prostate cancer. *Nat Genet* 40:281–283.
24. Eeles RA, et al.; UK Genetic Prostate Cancer Study Collaborators/British Association of Urological Surgeons' Section of Oncology; UK ProtecT Study Collaborators; PRACTICAL Consortium (2009) Identification of seven new prostate cancer susceptibility loci through a genome-wide association study. *Nat Genet* 41:1116–1121.
25. Thomas G, et al. (2008) Multiple loci identified in a genome-wide association study of prostate cancer. *Nat Genet* 40:310–315.
26. Ghoussaini M, et al.; UK Genetic Prostate Cancer Study Collaborators/British Association of Urological Surgeons' Section of Oncology; UK ProtecT Study Collaborators (2008) Multiple loci with different cancer specificities within the 8q24 gene desert. *J Natl Cancer Inst* 100:962–966.
27. Zheng SL, et al. (2008) Cumulative association of five genetic variants with prostate cancer. *N Engl J Med* 358:910–919.
28. Al Olama AA, et al.; UK Genetic Prostate Cancer Study Collaborators/British Association of Urological Surgeons' Section of Oncology; UK Prostate testing for cancer and Treatment study (ProtecT Study) Collaborators (2009) Multiple loci on 8q24 associated with prostate cancer susceptibility. *Nat Genet* 41:1058–1060.
29. Sun J, et al. (2008) Evidence for two independent prostate cancer risk-associated loci in the HNF1B gene at 17q12. *Nat Genet* 40:1153–1155.
30. Freedman ML, et al. (2006) Admixture mapping identifies 8q24 as a prostate cancer risk locus in African-American men. *Proc Natl Acad Sci USA* 103:14068–14073.
31. Haiman CA, et al. (2007) Multiple regions within 8q24 independently affect risk for prostate cancer. *Nat Genet* 39:638–644.
32. Purcell S, et al. (2007) PLINK: A tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* 81:559–575.
33. Pinto D, et al. (2011) Comprehensive assessment of array-based platforms and calling algorithms for detection of copy number variants. *Nat Biotechnol* 29:512–520.
34. Beroukhim R, et al. (2007) Assessing the significance of chromosomal aberrations in cancer: methodology and application to glioma. *Proc Natl Acad Sci USA* 104: 20007–20012.
35. Conrad DF, et al.; Wellcome Trust Case Control Consortium (2010) Origins and functional impact of copy number variation in the human genome. *Nature* 464: 704–712.
36. Berger MF, et al. (2011) The genomic complexity of primary human prostate cancer. *Nature* 470:214–220.
37. Durbin RM, et al.; 1000 Genomes Project Consortium (2010) A map of human genome variation from population-scale sequencing. *Nature* 467:1061–1073.
38. Mills RE, et al.; 1000 Genomes Project (2011) Mapping copy number variation by population-scale genome sequencing. *Nature* 470:59–65.
39. Benjamini Y, Hochberg Y (1995) Controlling the false discover rate: A practical and powerful approach to multiple testing. *J R Stat Soc, B* 57:289–300.
40. Yates D, Moore D, McCabe G (1999) *The Practice of Statistics* (W.H. Freeman, NewYork), 1st Ed.
41. Huang W, Sherman BT, Lempicki RA (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc* 4:44–57.
42. Dekker J, Rippe K, Dekker M, Kleckner N (2002) Capturing chromosome conformation. *Science* 295:1306–1311.
43. Lieberman-Aiden E, et al. (2009) Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* 326:289–293.
44. van Berkum NL, et al. Hi-C: a method to study the three-dimensional architecture of genomes. *J Vis Exp* 39:1869.
45. Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25:1754–1760.
46. Kent WJ, et al. (2002) The human genome browser at UCSC. *Genome Res* 12: 996–1006.
47. Rickman DS, et al. (2010) ERG cooperates with androgen receptor in regulating trefoil factor 3 in prostate cancer disease progression. *Neoplasia* 12:1031–1040.

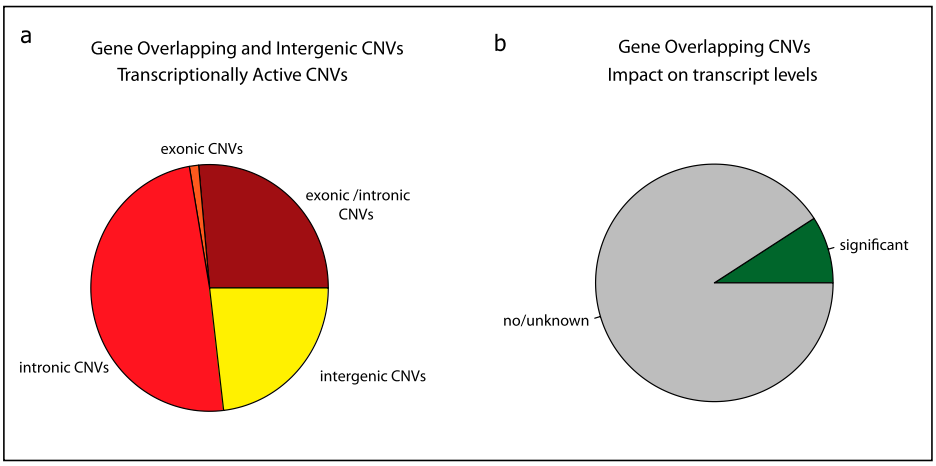


Fig. S1. Genomic characteristics of CNVs selected for prostate cancer risk association analysis. (*A*) Proportions of CNVs mapping to gene coding regions and to intergenic functionally active regions, and (*B*) proportions of gene coding variants with defined copy number state impact on corresponding transcript levels.

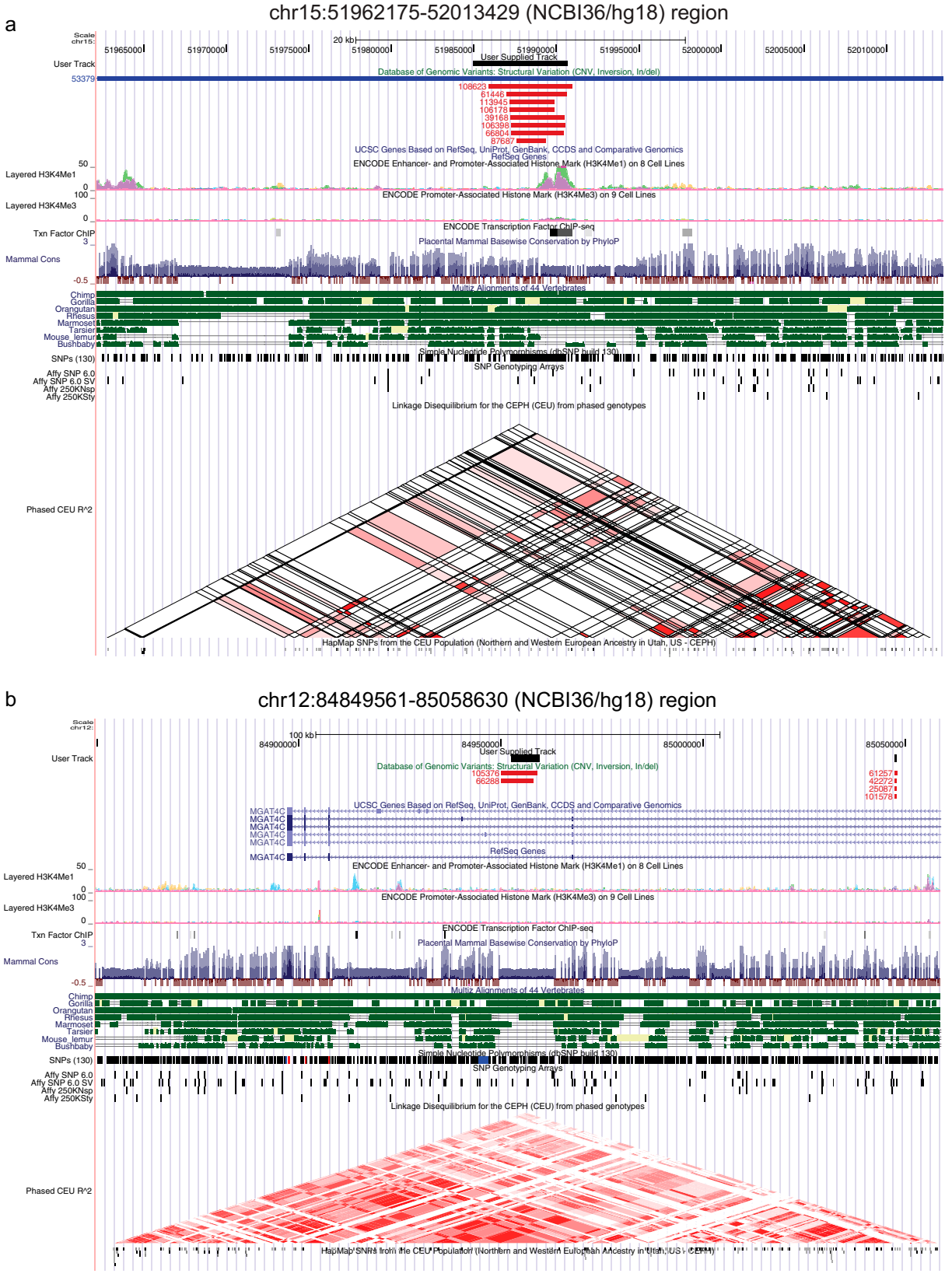


Fig. S2. Genome browser view of (*A*) 15q21.3 and (*B*) 12q21.31 variant regions including linkage disequilibrium (LD) data as evaluated from phased genotypes of HapMap II CEU individuals. Each diagonal represents a different SNP with each diamond representing a pair-wise comparison between two SNPs. Shades are used to indicate LD between the pair of SNPs, with darker shades indicating stronger LD. LD was specifically evaluated for each CNV and nearby SNPs (200-kb window) using Haploview. The analysis revealed the best r^2 equal to 0.32 (rs2725627) and to 0.52 (rs12321159) for 15q21.3 and 12q21.31 variants, respectively.

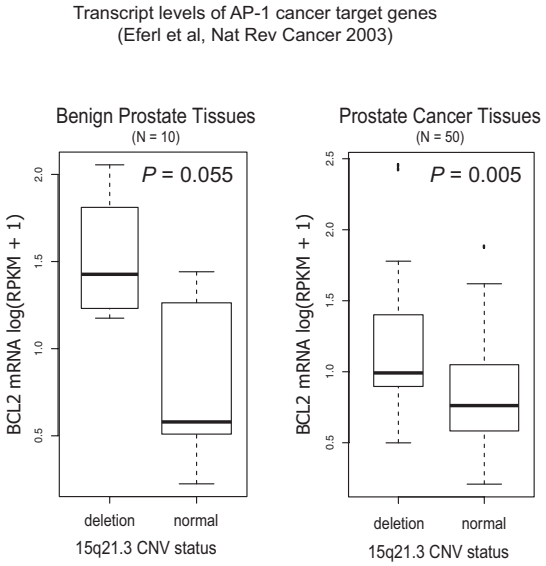


Fig. S3. Transcript levels of activator protein 1 (AP-1) cancer-related gene targets with respect to 15q21.3 locus. Prostate tissue data (benign, *Left*; tumor, *Right*) are visualized using boxplots for BCL2 mRNA. Higher levels of transcript are observed in the deletion groups consistently in benign and tumor prostate tissue.

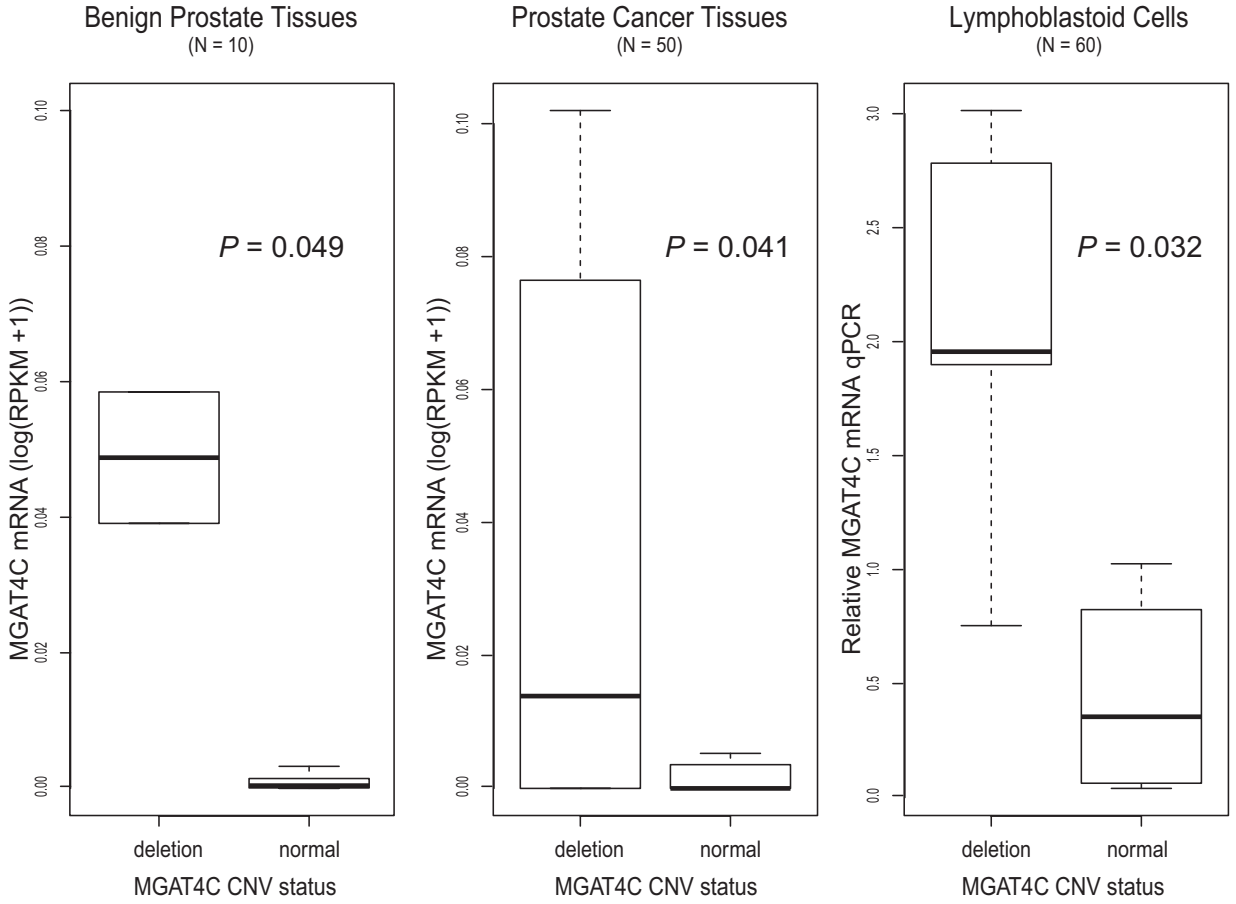


Fig. S4. MGAT4C transcript levels versus 12q21.31 CNV normal state and deletion. Prostate tissue data (benign, *Left*; tumor, *Center*) and lymphoblastoid cells data (*Right*) (1) are visualized using boxplots. Higher levels of transcript are observed in the deletion groups.

1. Montgomery SB, et al. (2010) Transcriptome genetics using second generation sequencing in a Caucasian population. *Nature* 464:773–777.

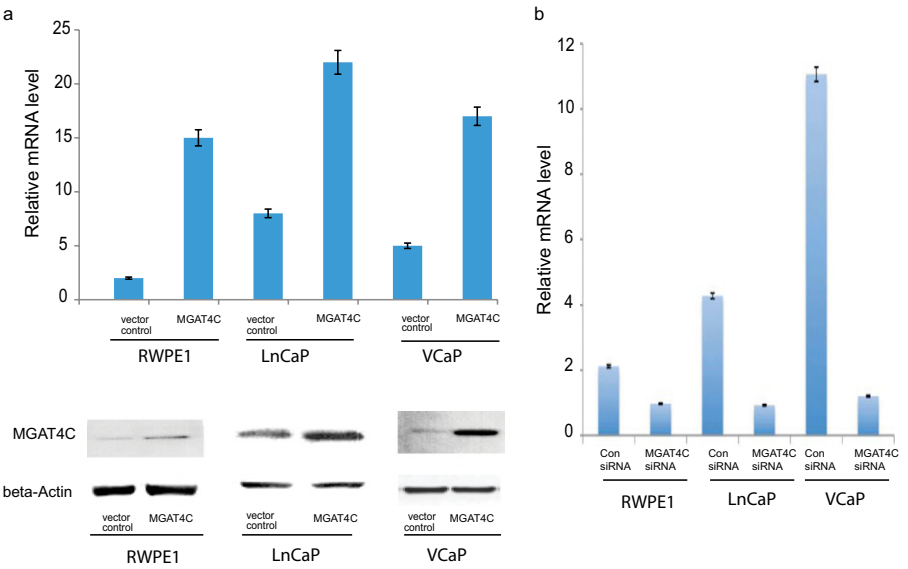


Fig. S5. Characterization of MGAT4C in prostate cell lines. (*A*) Transcript and protein expression of MGAT4C in RWPE1, LNCaP, and VCaP (aggressive prostate cancer) cell lines alone and following overexpression of MGAT4C. (*B*) Expression of MGAT4C in RWPE1, LNCaP, and VCaP cells treated with nontargeting control siRNAs or siRNAs against MGAT4C.

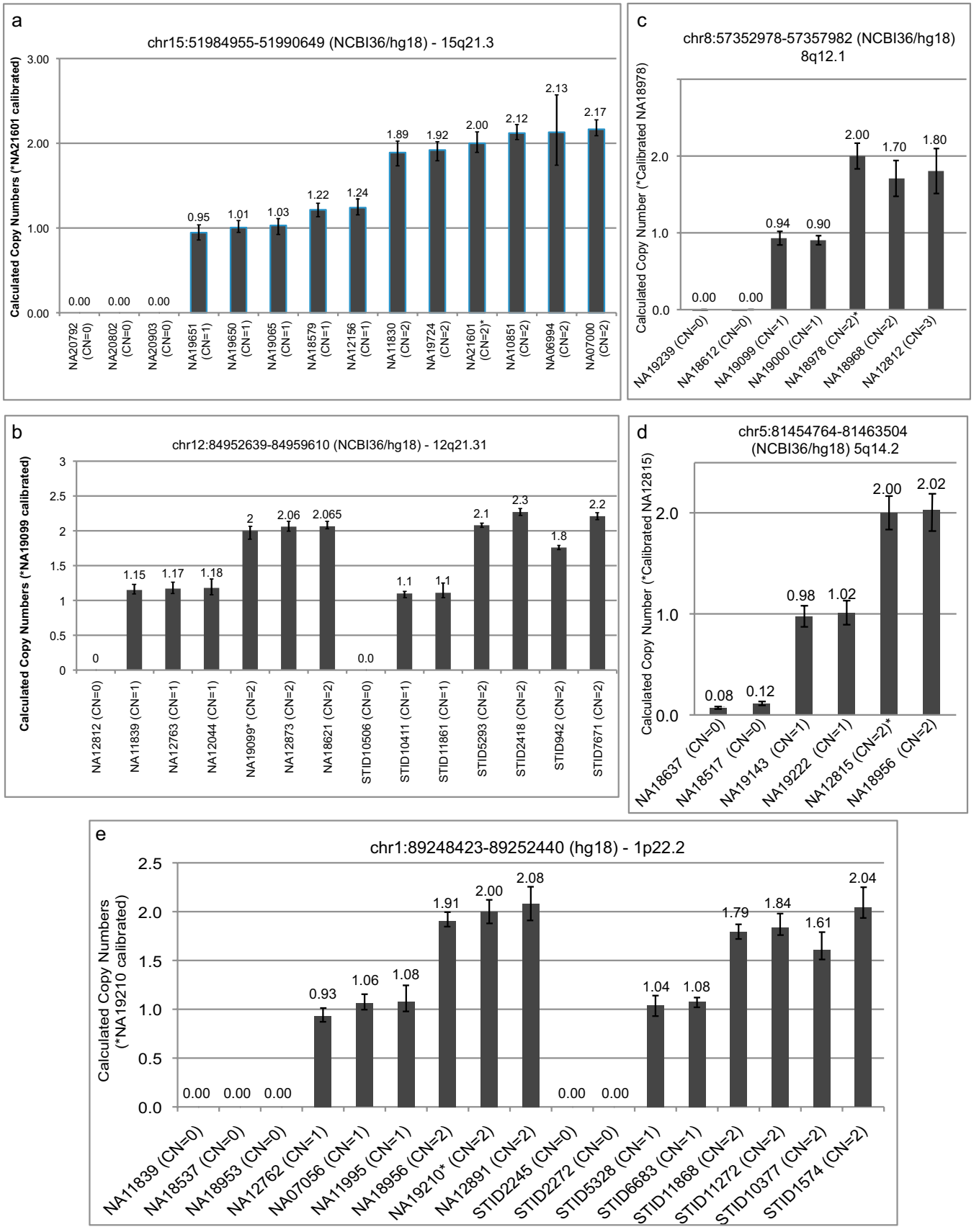
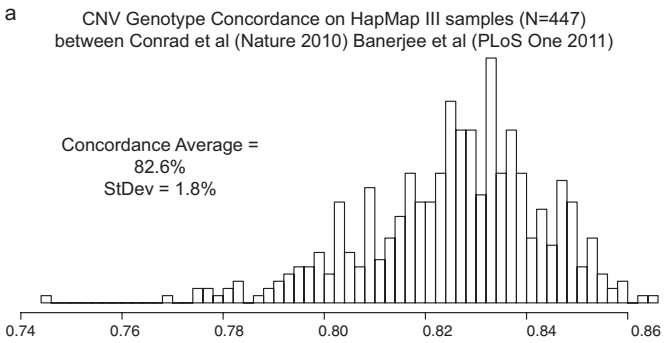


Fig. S6. (*A–E*) Quantitative PCR validation for a subset of study CNVs. Sample labels report the corresponding estimated copy number state (estimated through the study pipeline on Affymetrix data).

a



b

CNV Detection Comparison on HapMap III samples (N=447)
between Conrad et al (Nature 2010) Banerjee et al (PLoS One 2011)

Probes	Percentage of CNVs detected by Banerjee et al included in Conrad et al	Percentage of CNVs detected by Conrad et al included in Banerjee et al
> 2	31.68	56.13
> 5	47.75	62.83
> 10	61.16	66.21
> 20	72.73	70.31

Fig. S7. CNV detection and genotyping comparison between (*A*) Banerjee et al. (26) approach and (*B*) Conrad et al. (28) approach on 447 HapMap samples.

Dataset S1. Tables pertaining to the study data

Dataset S1 (XLSX)

The tables listed in the dataset are as follows: Table S1: Demographics of the Tyrol Cohort; Table S2: Replication of PCA risk SNP associations in Tyrol cohort based on previous studies; Tables S3 *A* and *B*: Summary tables of the 2,611 CNVs (*A*) and of the selected 238 CNVs (*B*); Table S4: Complete association data for the 238 biallelic low-frequency CNVs in the Tyrol cohort; Tables S5 *A*–*D*: Demographics of the EDRN cohort. All (*A*); Cornell (*B*); Harvard (*C*); Michigan (*D*); Table S6: AP-1 cancer target transcripts (1) with respect to 15q21.3 CNV; Table S7: Hi-C predicted interacting genes and miRNA for noncoding CNV on 15q21.3.

1. Eferl R, Wagner EF (2003) AP-1: A double-edged sword in tumorigenesis. *Nat Rev Cancer* 3:859–868.