Supplementary Materials

Online methods
Figures S1, S2, S3, S4, S5, S6, S7, S8, S9, S10, S11
Tables S1, S2, S3, S4, S5, S6, S7, S8, S9, S10, S11, S12
References (supplementary)

Online methods

SEER data

Surveillance, Epidemiology, and End Results (SEER) data were downloaded from http://seer.cancer.gov/data/.

Tumor / normal exon sequencing, DNA copy number, and RNA-sequencing data

Genomic data was obtained from the TCGA data portal (https://tcga-data.nci.nih.gov/tcga/) and from additional Broad Institute datasets[39,50]. Informed consent was obtained from all subjects by local institutional review boards and consents were reviewed by the sequencing centers, per TCGA guidelines. To visualize gene expression by mutation status for males and females, TCGA data were downloaded and visualized using cBioPortal (www.cbioportal.org)[51,52].

Mutation classification

Tumor-associated DNA variants were called "truncating" if they resulted in nonsense, insertion, deletion, translation start site, or nonstop mutations. Truncating and likely functional missense (LOF) included all truncating mutations as well as missense mutations that occurred at the same site in at least 3 patients (n=419 across the pan-cancer dataset, of which 9 were male-predominant) or had a Cons 46-vertebrates score (http://ucscbrowser.genap.ca/cgi-bin/hgTrackUi?db=hg19&g=cons46way)[53] higher than 0.9 indicating high conservation across species (n=3083, of which 124 were male-predominant).

Permutation analysis

For each set of events, the probability of observing the number of events seen in males was determined by a series of permutations. Within each tumor type, the probability that any given gene alteration would occur in a male was assumed to be directly related to the number of mutations across all males, divided by the total number of mutations across all samples. This controlled for the number of males in a set as well as the relative mutation rate on chrX in males vs females in the set. For instance, if 60% of coding mutations on chrX in the dataset for a tumor type were in males, then for any given chrX gene, assuming there is no selective advantage for males vs. females, we should see 60% of the events occurring in males. For each gene, the total number of events was counted and a random set of permutations was constructed, in which each event was randomly assigned a 'male' or 'female' value based on this calculated probability. For the pan-cancer analyses, the probability of an event being in a male was calculated on a per tumor type basis, and events were counted and permuted within each tumor type. Once the number of males in the permutation was calculated, it was compared against the observed data to see if the number of males randomly generated in this manner was greater than or equal to that observed in the data. For each gene, 1 million permutations were done, and the fraction of permutations with a greater than or equal number of male events was reported as the P value. Multiple hypothesis correction was then performed on the complete set of genes that were affected by the set of events examined, and genes with FDR <0.1 (by Benjamini-Hochberg false discovery rate correction) were reported as significant. For permutations on truncating and LOF mutations, this probability was calculated based on the coding mutation count of chrX. For permutations on copy-loss events, the frequency of copy-loss events on chrX was used. For the combined LOF/CN loss permutations, the sum of these two was used.

Log likelihood ratio test

Assuming that the only factor differentiating the male mutation rate of a gene and the female mutation rate of a gene is the difference in the background mutation rate on chrX (i.e. assuming there is no selective bias in males for mutating a gene), then the probability of a male having a mutation in a gene should be directly related to the probability of a female having a mutation in a gene, corrected for the number of copies of chrX in females (n=2) and males (n=1):

$$p_{female} = r * p_{male}$$

Where r is the F/M ratio of coding mutations across chrX, i.e. if females in a set have twice the number of coding mutations on chrX as males, it can be expected that any given gene on X should be twice as likely to be mutated in a female patient vs. a male patient. For each test, the actual F/M ratio of coding mutations on chrX in the analysis set was calculated and used. Because there is a direct relationship between the probability of a male mutation in this model and a female mutation, we can express the likelihood that the observed data is consistent with this model using a single value of p:

$$L_0 = max((p_{male})^m(1-p_{male})^{M-m}(r*p_{male})^f(1-r*p_{male})^{F-f})$$

Where *M* and *F* are the total number of male and female patients, respectively, and *m* and *f* are the number of mutated males and females, respectively. The alternative hypothesis is that there are independent factors affecting males and females, in which case two values of p are needed to calculate the likelihood that the data fits the model:

$$L_1 = max((p_{male})^m(1-p_{male})^{M-m}(p_{female})^f(1-p_{female})^{F-f})$$

Which can be maximized by using the observed mutation counts:

$$L_1 = (m/M)^m(M-m/M)^{M-m}(f/F)^f(F-f/F)^{F-f}$$

The log-likelihood ratio (*LLR*) is calculated simply by taking the log of the ratio of these two numbers:

$$LLR = log(L_1/L_0)$$

Which, using Wilks's theorem, was converted to a P value for each gene. The Benjamini-Hochberg procedure for controlling the False Discovery Rate (FDR) was then applied, and genes with FDR <0.1 were reported as significant.

Copy number determination by SNP array

Probe-level signal intensities from Affymetrix SNP6 .CEL files for tumor samples across different cancer types were combined, calibrated, normalized, and segmented in uniform fashion as previously described[50]. Markers identified as having recurrent germline copy number variations using normal samples were excluded. For samples for which ABSOLUTE[54] purity/ploidy calls were available, copy numbers were scaled by a factor inversely proportional to the purity estimate to remove the effects of admixed normal cells, as previously described[50]. Copy number profiles were deconstructed into underlying somatic events using GISTIC[55]; only focal events with length less than 1 megabase were used to limit normalization bias between males and females on chrX. Copy number for a gene was determined by using the most extreme copy number among the markers spanning the gene. The threshold for calling somatic copy number events was chosen by considering the distribution of copy number across both male and female cohorts (Supplementary Figure 3). Because normalization decouples chrX overall copy number from that of the autosomes, retention or loss of the entire chrX was called by applying a threshold of 1.6 to a robust average of the un-normalized, calibrated signal across the X chromosome (Supplementary Figure 3). The maximum and minimum 0.2% probe-level signals were excluded.

Power analysis

The power calculation was performed using a binomial model similarly as previously described[39]. The power to detect a male-biased mutation can be determined by first calculating the maximum number of mutations in males, m_{max}, that would be considered non-significant in the null model, based on an inverse binomial cumulative distribution function, given the desired bound on the false discovery rate (<0.1), the total number of mutations in a gene in the entire cohort, n (a function of the fraction of patients with a mutation and the total number of patients), and the fraction of all events that occur in males (as opposed to in females) in the cohort, which determines the probability that a mutation will be in a male tumor, p_0. The power is then determined by the probability of discovering at least that many mutations in male patients under the alternate hypothesis, which adjusts p_0 by the hypothesized increased relative risk of a mutation occurring in a male tumor, R_{signal}, the fold change compared to the null hypothesis.

$$R_{background} = p_0/(1- p_0)$$

$$R_{total} = R_{background} * R_{signal}$$

$$p_{alt} = 1 / (1 + 1/R_{total})$$

To generate figures, we calculated the number of patients (N) required for power to detect 80% of genes with a statistically significant male bias, given hypothetical values of R_{signal}, n, and p_0.

Y copy loss determination

We calculated chrY coverage at a per-megabase level for paired tumor and normal sequencing data, and then normalized to total coverage of the exome. A region of chrY between bases 29000001-58000000 was omitted due to poor coverage. An additional region, 13000001-14000000, was removed due to high incidence of misalignment. A biphasic pattern of copy number quantitation was observed in male samples (Supplementary Figure 6), with a distinct population of tumors demonstrating significant median copy loss compared to the paired normal sample. ChrY loss was therefore assigned to those samples with <25% in overall coverage of chrY in the tumor compared to the paired normal sample.

Allele-specific expression analysis

Exome sequencing and RNA-seq data from somatic and germline samples from GBM, LGG, HNSC, KIRC, LUSC, and LUAD sets from TCGA, and brain, lung, and whole blood sample sets from GTEx were analyzed. For each, allele-specific RNA-seq pileup counts were called at heterozygous germline sites and tumor sites from cancers (determined by exome sequencing) from the respective analysis. Duplicate, non-primary, soft clipped and Phred quality 0 reads were not included in the pileup count. For sites where at least 20 reads (tumor samples) or 8 reads (normal samples) were detected by RNA-seq, the count of the less frequent allele was divided by the total allele count to obtain the minor allele fraction. Data are represented as the average minor allele fraction for all sites in the indicated gene, or as the average of all sites in all genes in the 'non-escape' group.

Analysis of GTEx male vs female expression data

We obtained RNA-Seq expression data from the GTEx project as gene reads per sample by accessing the project's pipeline (http://www.gtexportal.org/home/). For each tissue, we normalized the signal across samples. The whole blood samples used were restricted to those collected ante mortem. There were 85 whole blood samples (25 female and 60 male) and 360 brain samples (124 female and 233 male) that passed quality control and fit the selection criteria. We tested for bimodality of *ATRX* expression in the brain in two ways. First, we fit a Gaussian mixture model to the empirical densities using the R package "mixtools" (http://cran.r-project.org/web/packages/mixtools/), function normalmixEM with parameters: k=2, epsilon = 1e-08, maxit = 1000, maxrestarts=20, and assessed the distance between resulting means. Second, we tested for bimodality with a likelihood ratio test for bimodality in two-component mixtures. The method contrasts the likelihood of the data obtained under restricted and unrestricted maximum likelihood fits of mixture of normal distributions. Under the assumption of equal variance, bimodality solely depends on the mixture component weight and the ratio of the distance between means and the variance. With unequal variance, it is also determined by the ratio of variances.

We used R package "diptest" to test for multimodality (http://cran.r-project.org/web/packages/diptest/).

Calculation of percentage excess male risk

For a given gene in a specific disease, we calculated the excess male risk associated with loss-of-function mutation of a single gene on chrX. We cannot assume that the M:F ratio in our sample set is the same as the general population incidence of that disease, so we adjusted the overall M:F ratio to SEER data for each cancer in the US population. The fraction of the excess male risk in the disease attributable to a specific gene mutation was calculated as follows: (# males with the gene mutated in our dataset – (Z * # of females with the gene mutated in our dataset)) / (# males in our dataset – (Z * # of females in our dataset)); where Z = our dataset M:F ratio / SEER data M:F ratio.

Cnksr2 knockdown, western blotting, and soft agar colony assays

Mouse *Cnksr2* or control RFP shRNA constructs (oligonucleotide sequences in Supplementary Table 12) were cloned per protocol (http://www.addgene.org/tools/protocols/plko/) into a pLKO.1 vector modified to express a GFP reporter in place of the puromycin resistance gene. Lentivirus was produced and murine 3T3 cells (from ATCC, mycoplasma-free) were infected using standard protocols (Addgene) and GFP-positive cells were sorted after 7 days. Western blotting was performed as previously described[56] using antibodies recognizing phospho-ERK (Cell Signaling #4370), total ERK (Cell Signaling #9102), and tubulin (Sigma #T6074). Soft agar colony assay was performed as previously described[57]. After three weeks, random microscopy fields with an area of 3.29 mm^2 were scanned and colonies of minimum 50 um^2 in size were counted using a CellCelector (ALS, Germany).

RNA-sequencing and gene set enrichment analysis (GSEA) in *Cnksr2* knock-down cells

Total RNA was prepared using a MiRNeasy kit (Qiagen). Illumnia sequencing libraries were prepared using Illumina TruSeq Stranded mRNA sample preparation kits from 500ng of purified total RNA according to the manufacturer's protocol. The finished dsDNA libraries were quantified by Qubit fluorometer, Agilent TapeStation 2200, and RT-qPCR using the Kapa Biosystems library quantification kit according to manufacturer's protocols. Uniquely indexed libraries were pooled in equimolar ratios and sequenced on an Illumina NextSeq500 with single-end 75bp reads by the Dana-Farber Cancer Institute Molecular Biology Core Facilities. Reads were aligned to the mm9 reference genome assembly using STAR (v25.1b) (https://github.com/alexdobin/STAR). FPKM expression values were calculated using cufflinks (v2.2.1) (http://cole-trapnell-lab.github.io/cufflinks/). Spearman correlation and principal component analysis were performed using VIPER (https://bitbucket.org/cfce/viper/). GSEA (http://www.broadinstitute.org/gsea/) was performed as previously described[41,58]. Network enrichment mapping was performed using Cytoscape (http://www.cytoscape.org/).

Methods-only References

50 Zack, T. I. *et al.* Pan-cancer patterns of somatic copy number alteration. *Nature genetics* **45**, 1134-1140, doi:10.1038/ng.2760 (2013).

51 Gao, J. *et al.* Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Science signaling* **6**, pl1, doi:10.1126/scisignal.2004088 (2013).

52 Cerami, E. *et al.* The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer discovery* **2**, 401-404, doi:10.1158/2159-8290.CD-12-0095 (2012).

53 Siepel, A. *et al.* Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome research* **15**, 1034-1050, doi:10.1101/gr.3715005 (2005).

54 Carter, S. L. *et al.* Absolute quantification of somatic DNA alterations in human cancer. *Nature biotechnology* **30**, 413-421, doi:10.1038/nbt.2203 (2012).

55 Beroukhim, R. *et al.* Assessing the significance of chromosomal aberrations in cancer: methodology and application to glioma. *Proceedings of the National Academy of Sciences of the United States of America* **104**, 20007-20012, doi:10.1073/pnas.0710052104 (2007).

56 Yoda, A. *et al.* Mutations in G protein beta subunits promote transformation and kinase inhibitor resistance. *Nature medicine* **21**, 71-75, doi:10.1038/nm.3751 (2015).

57 Hammerman, P. S. *et al.* Mutations in the DDR2 kinase gene identify a novel therapeutic target in squamous cell lung cancer. *Cancer discovery* **1**, 78-89, doi:10.1158/2159-8274.CD-11-0005 (2011).

58 Lane, A. A. *et al.* Triplication of a 21q22 region contributes to B cell transformation through HMGN1 overexpression and loss of histone H3 Lys27 trimethylation. *Nature genetics* **46**, 618-623, doi:10.1038/ng.2949 (2014).