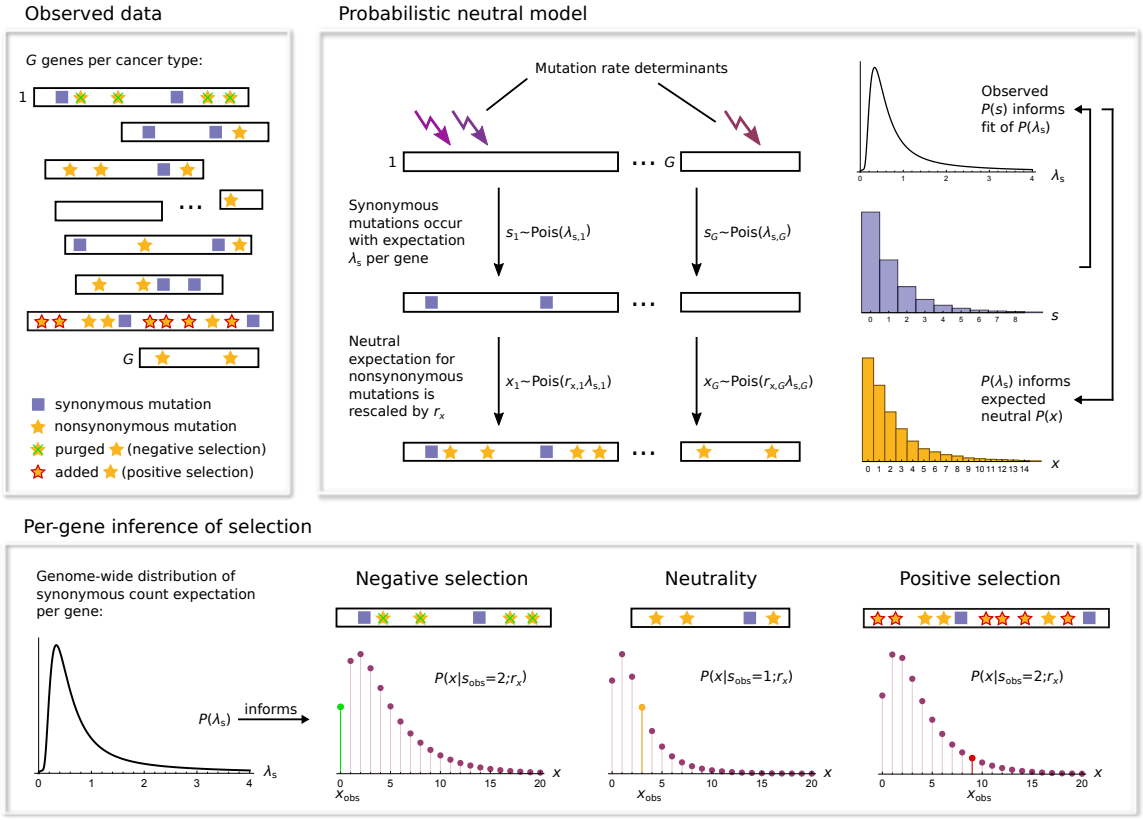
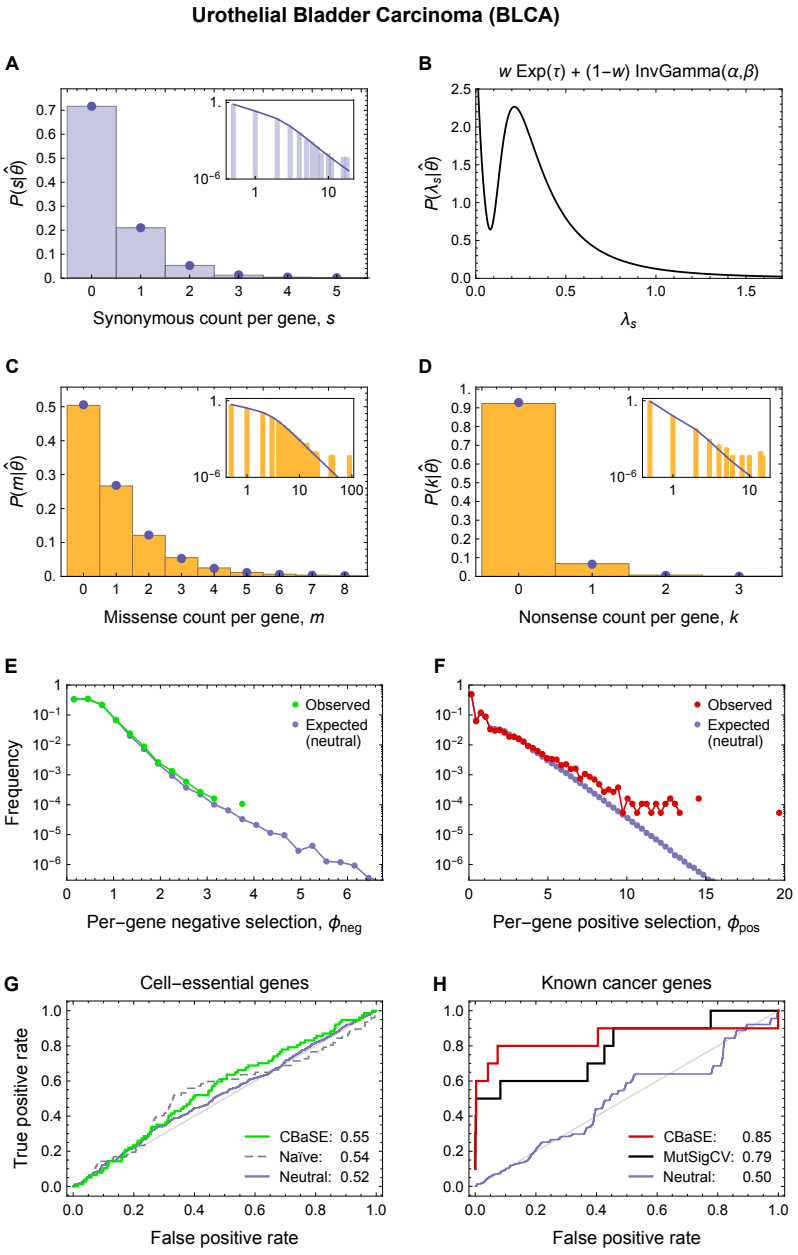


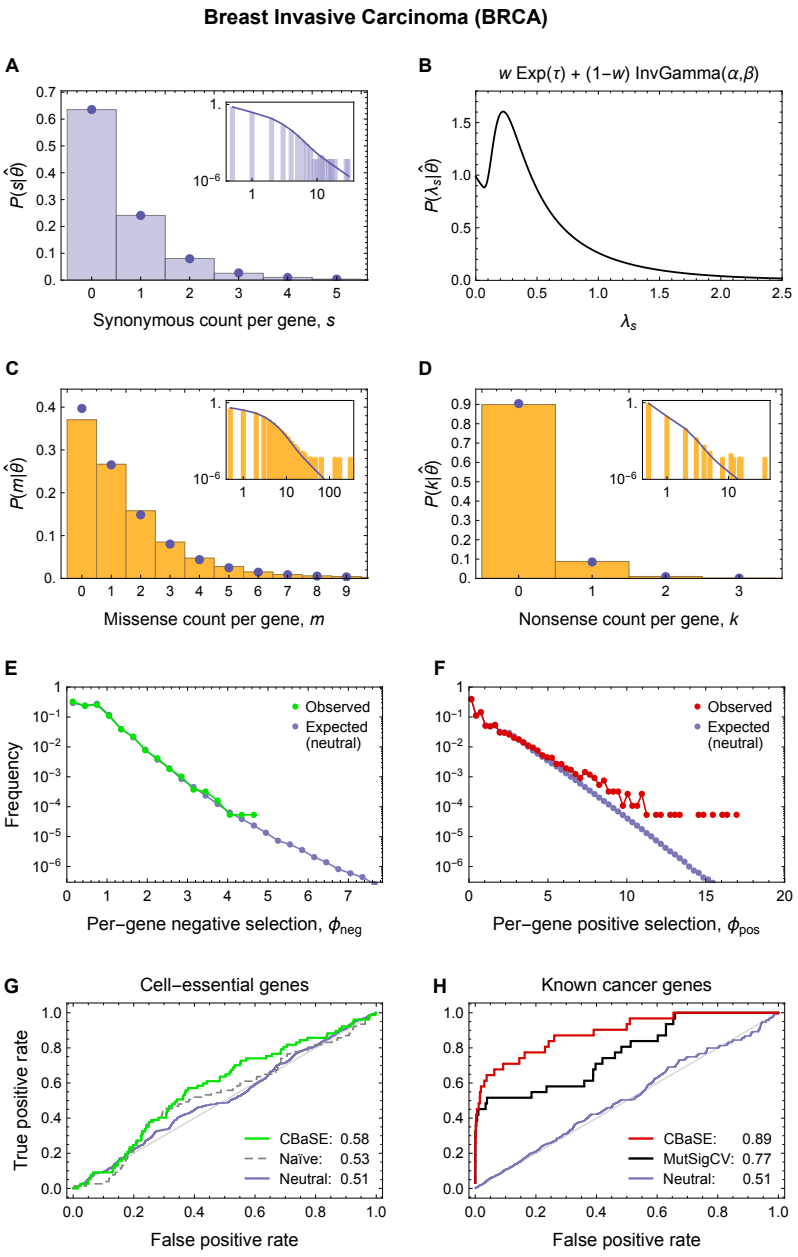
Supplementary Figure 1: Schematic illustration of the method. Top left: For each cancer type, the utilized data set consists of the cumulative count (over patients) of synonymous (s) and nonsynonymous (x) mutations per gene. Positive selection is recognized as genes showing excessive nonsynonymous variation ("added"), while negative selection manifests as nonsynonymous mutation paucity ("purged"), relative to the neutral expectation. Top right: Background mutation rate varies along the genome and determines λ_s, the expectation value of the number of synonymous mutations on a given gene, $s \sim \text{Pois}(\lambda_s)$. A hierarchical model fit to the observed genome-wide distribution of synonymous counts (blue histogram) allows to estimate the parameters of $P(\lambda_s)$. In turn, this fit informs the genome-wide expected distribution of nonsynonymous counts under neutral evolution (yellow bars; depicted as blue dots and lines in **Figures 1C,D** and **Supplementary Figures 2-26C,D**). On each gene with synonymous expectation λ_s, the nonsynonymous count expectation is $r_x\lambda_s$, where r_x denotes the ratio of nonsynonymous and synonymous target size, which takes into account the cancer type-specific mutational signature and the gene sequence composition. Bottom: Using the inferred $P(\lambda_s)$ and the observed synonymous count, s_{obs}, on each gene the conditional probability for the expected nonsynonymous count under neutrality is derived: $P(x|s_{\text{obs}}; r_x)$ (purple bars). From this, p-values for the observed number of missense ($x = m_{\text{obs}}$) and nonsense ($x = k_{\text{obs}}$) mutations are computed. These correspond to the probabilities in the tails, on the small-x end for negative selection and on the large-x end for positive selection. These p-values are then combined into the meta-statistics ϕ_{neg} and ϕ_{pos}, respectively measuring negative and positive selection strength. For each gene, the statistical significance of ϕ_{neg} and ϕ_{pos} is determined from comparison to simulated distributions under neutrality, controlling the false-discovery rate at q_{neg} and q_{pos}, respectively.

Nature Genetics: doi:10.1038/ng.3987



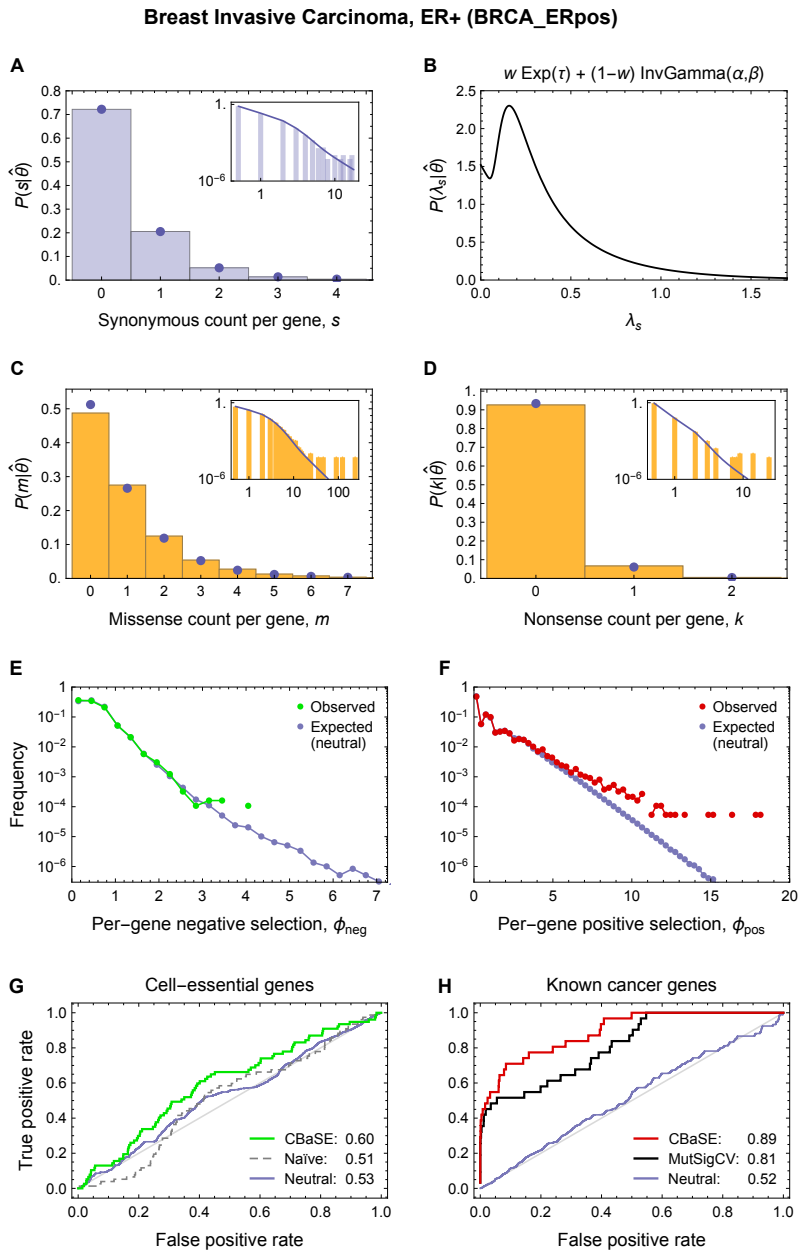
Supplementary Figure 2: BLCA. (A),(C),(D) Observed distributions of mutation counts per gene (histograms), and expected distributions under neutral evolution (blue dots and blue lines) with underlying $P(\lambda_s;\boldsymbol{\theta}) = w\,\mathrm{Exp}(\lambda_s;\tau) + (1-w)\,\mathrm{InvGamma}(\lambda_s;\alpha,\beta)$, shown in (B). Parameter estimates $\hat{\boldsymbol{\theta}}$ are $\hat{\alpha} = 2.86$, $\hat{\beta} = 0.84$, $\hat{\tau} = 24.10$, $\hat{w} = 0.14$; sum of squared deviations in (A) is 2e-6. (E),(F) Histogram of the meta-statistic ϕ in real data (green, red) and from simulation under the neutral model (blue), bin width 0.3. Genes (ϕ_{pos}) not shown in (F): *RB1* (20.2), *KMT2D* (21.0), *KDM6A* (23.1), *ARID1A* (35.4), *TP53* (49.9). (G) ROC of 77 predicted cell-essential genes from Wang et al.[2] for genes sorted by q_{neg} (green), $p_{\mathrm{AUC}} =$ 1e-1. Dashed gray line shows the ROC from ranking by hypomutation ($p_{\mathrm{AUC}} > 0.05$). (H) ROC of the gene ranking based on q_{pos}, using as true positives 11 CGC genes causally implicated in BLCA (red), $p_{\mathrm{AUC}} =$ 2e-5. Black line shows the ROC for MutSigCV[1] predictions. Blue lines in (G) and (H) show neutral ROC expectations from simulation. AUCs are given in the insets.

Nature Genetics: doi:10.1038/ng.3987



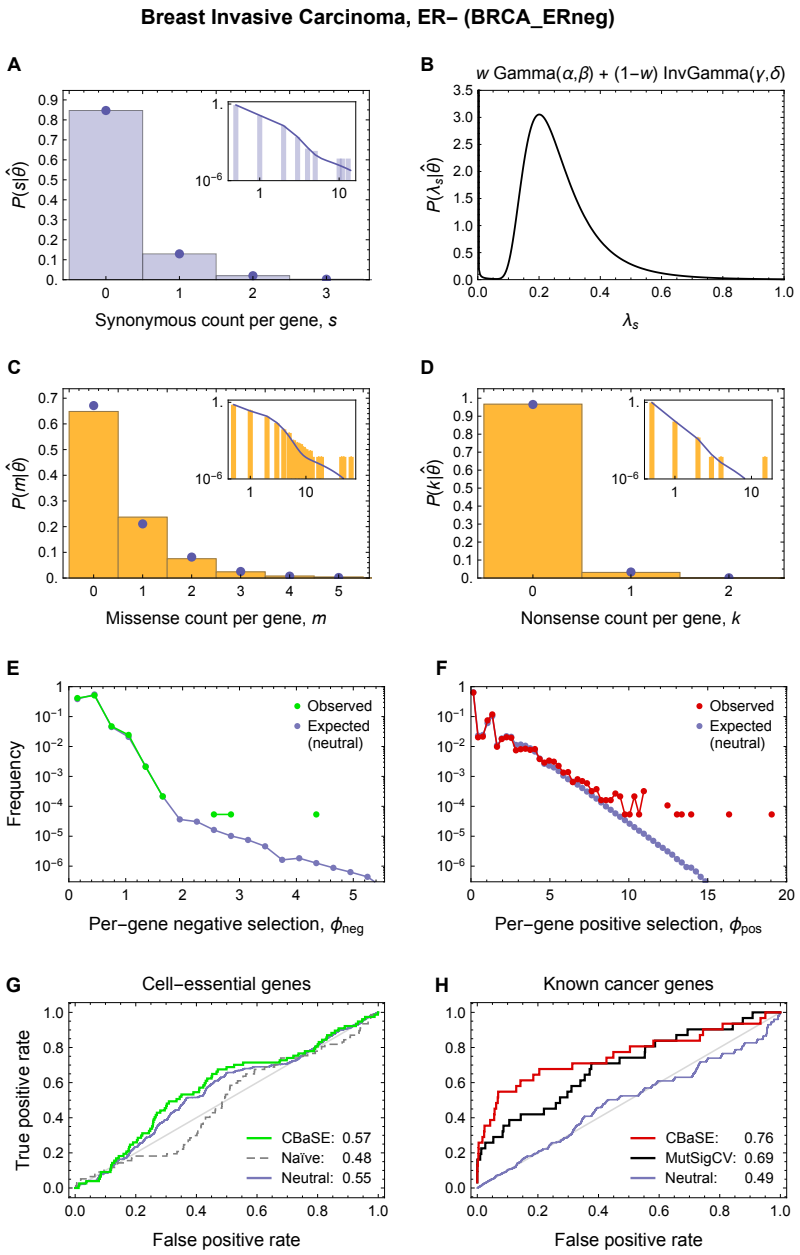
Supplementary Figure 3: BRCA. (A),(C),(D) Observed distributions of mutation counts per gene (histograms), and expected distributions under neutral evolution (blue dots and blue lines) with underlying $P(\lambda_s;\boldsymbol{\theta}) = w\,\mathrm{Exp}(\lambda_s;\tau) + (1-w)\,\mathrm{InvGamma}(\lambda_s;\alpha,\beta)$, shown in (B). Parameter estimates $\hat{\boldsymbol{\theta}}$ are $\hat{\alpha} = 2.38$, $\hat{\beta} = 0.82$, $\hat{\tau} = 1.83$, $\hat{w} = 0.54$; sum of squared deviations in (A) is 3e-6. (E),(F) Histogram of the meta-statistic ϕ in real data (green, red) and from simulation under the neutral model (blue), bin width 0.3. Genes (ϕ_{pos}) not shown in (F): *MAP2K4* (23.0), *NCOR1* (23.2), *CDH1* (29.1), *MAP3K1* (30.3), *KMT2C* (32.5), *PIK3CA* ($p_m^{\mathrm{pos}} = 0$), *TP53* ($p_m^{\mathrm{pos}} = 0$). (G) ROC of 77 predicted cell-essential genes from Wang et al.[2] for genes sorted by q_{neg} (green), $p_{\mathrm{AUC}} = 5\text{e-}2$. Dashed gray line shows the ROC from ranking by hypomutation ($p_{\mathrm{AUC}} > 0.05$). (H) ROC of the gene ranking based on q_{pos}, using as true positives 31 CGC genes causally implicated in BRCA (red), $p_{\mathrm{AUC}} = 2\text{e-}13$. Black line shows the ROC for MutSigCV[1] predictions. Blue lines in (G) and (H) show neutral ROC expectations from simulation. AUCs are given in the insets.

Nature Genetics: doi:10.1038/ng.3987



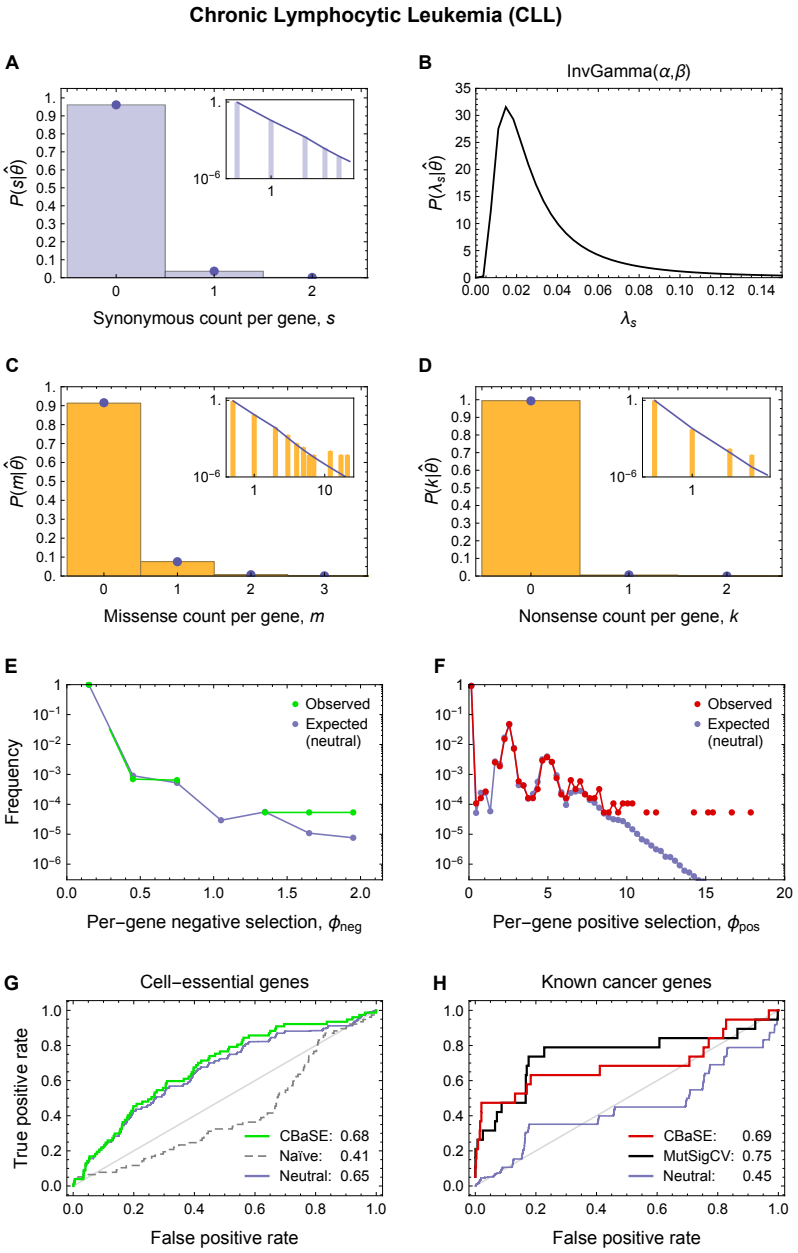
Supplementary Figure 4: BRCA (ER-positive). (A),(C),(D) Observed distributions of mutation counts per gene (histograms), and expected distributions under neutral evolution (blue dots and blue lines) with underlying $P(\lambda_s;\boldsymbol{\theta}) = w\,\mathrm{Exp}(\lambda_s;\tau) + (1-w)\,\mathrm{InvGamma}(\lambda_s;\alpha,\beta)$, shown in (B). Parameter estimates $\hat{\boldsymbol{\theta}}$ are $\hat{\alpha} = 2.36$, $\hat{\beta} = 0.58$, $\hat{\tau} = 2.82$, $\hat{w} = 0.53$; sum of squared deviations in (A) is 2e-7. (E),(F) Histogram of the meta-statistic ϕ in real data (green, red) and from simulation under the neutral model (blue), bin width 0.3. Genes (ϕ_{pos}) not shown in (F): *MAP2K4* (20.7), *CDH1* (27.1), *KMT2C* (31.0), *MAP3K1* (38.3), *TP53* (83.6), *PIK3CA* ($p_m^{\mathrm{pos}} = 0$). (G) ROC of 77 predicted cell-essential genes from Wang et al.[2] for genes sorted by q_{neg} (green), $p_{\mathrm{AUC}} = 4\text{e-}3$. Dashed gray line shows the ROC from ranking by hypomutation ($p_{\mathrm{AUC}} > 0.05$). (H) ROC of the gene ranking based on q_{pos}, using as true positives 31 CGC genes causally implicated in BRCA (red), $p_{\mathrm{AUC}} = 4\text{e-}13$. Black line shows the ROC for MutSigCV[1] predictions. Blue lines in (G) and (H) show neutral ROC expectations from simulation. AUCs are given in the insets.

Nature Genetics: doi:10.1038/ng.3987



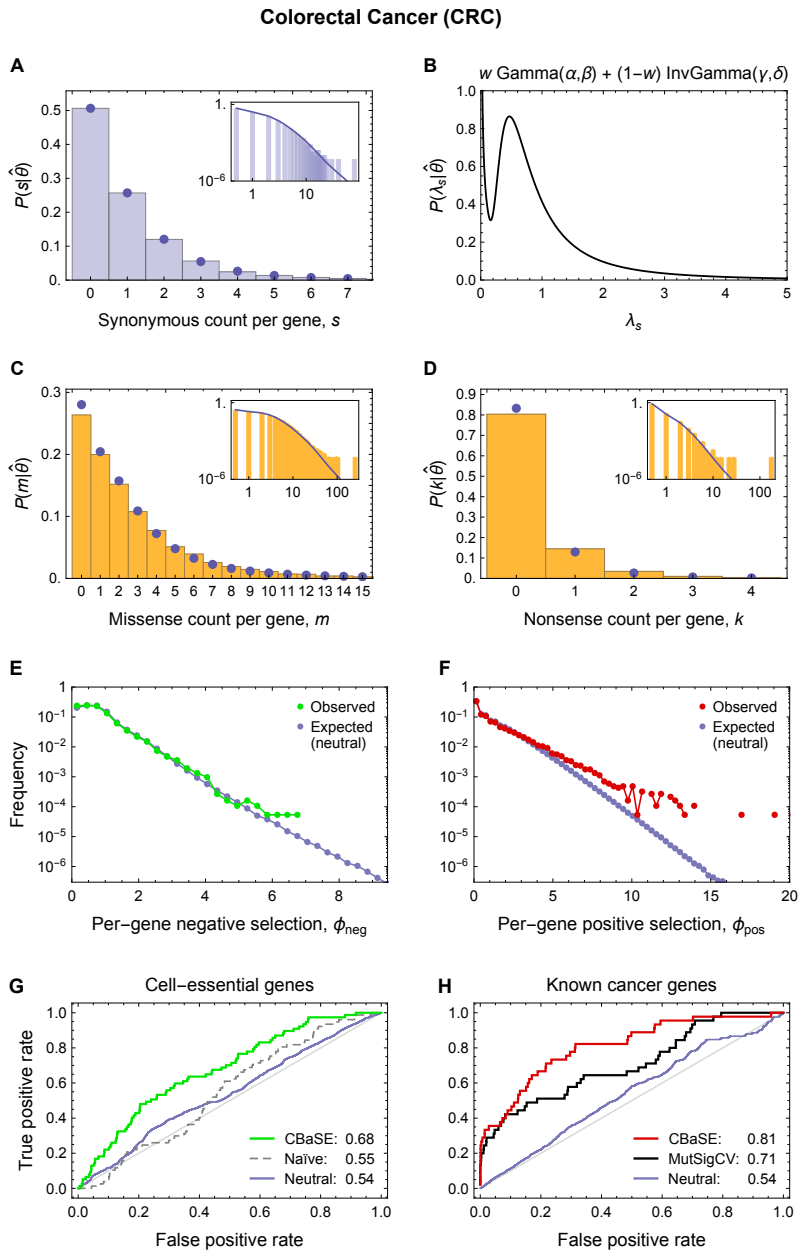
Supplementary Figure 5: BRCA (ER-negative). (A),(C),(D) Observed distributions of mutation counts per gene (histograms), and expected distributions under neutral evolution (blue dots and blue lines) with underlying $P(\lambda_s;\boldsymbol{\theta}) = w\,\mathrm{Gamma}(\lambda_s;\alpha,\beta) + (1-w)\,\mathrm{InvGamma}(\lambda_s;\gamma,\delta)$, shown in (B). Parameter estimates $\hat{\boldsymbol{\theta}}$ are $\hat{\alpha} = 0.002$, $\hat{\beta} = 6.19$, $\hat{\gamma} = 6.57$, $\hat{\delta} = 1.52$, $\hat{w} = 0.35$; sum of squared deviations in (A) is 5e-9. (E),(F) Histogram of the meta-statistic ϕ in real data (green, red) and from simulation under the neutral model (blue), bin width 0.3. Gene (ϕ_{pos}) not shown in (F): *TP53* (65.3). (G) ROC of 77 predicted cell-essential genes from Wang et al.[2] for genes sorted by q_{neg} (green), $p_{\mathrm{AUC}} = 2\text{e-}1$. Dashed gray line shows the ROC from ranking by hypomutation ($p_{\mathrm{AUC}} > 0.05$). (H) ROC of the gene ranking based on q_{pos}, using as true positives 31 CGC genes causally implicated in BRCA (red), $p_{\mathrm{AUC}} = 4\text{e-}9$. Black line shows the ROC for MutSigCV[1] predictions. Blue lines in (G) and (H) show neutral ROC expectations from simulation. AUCs are given in the insets.

Nature Genetics: doi:10.1038/ng.3987



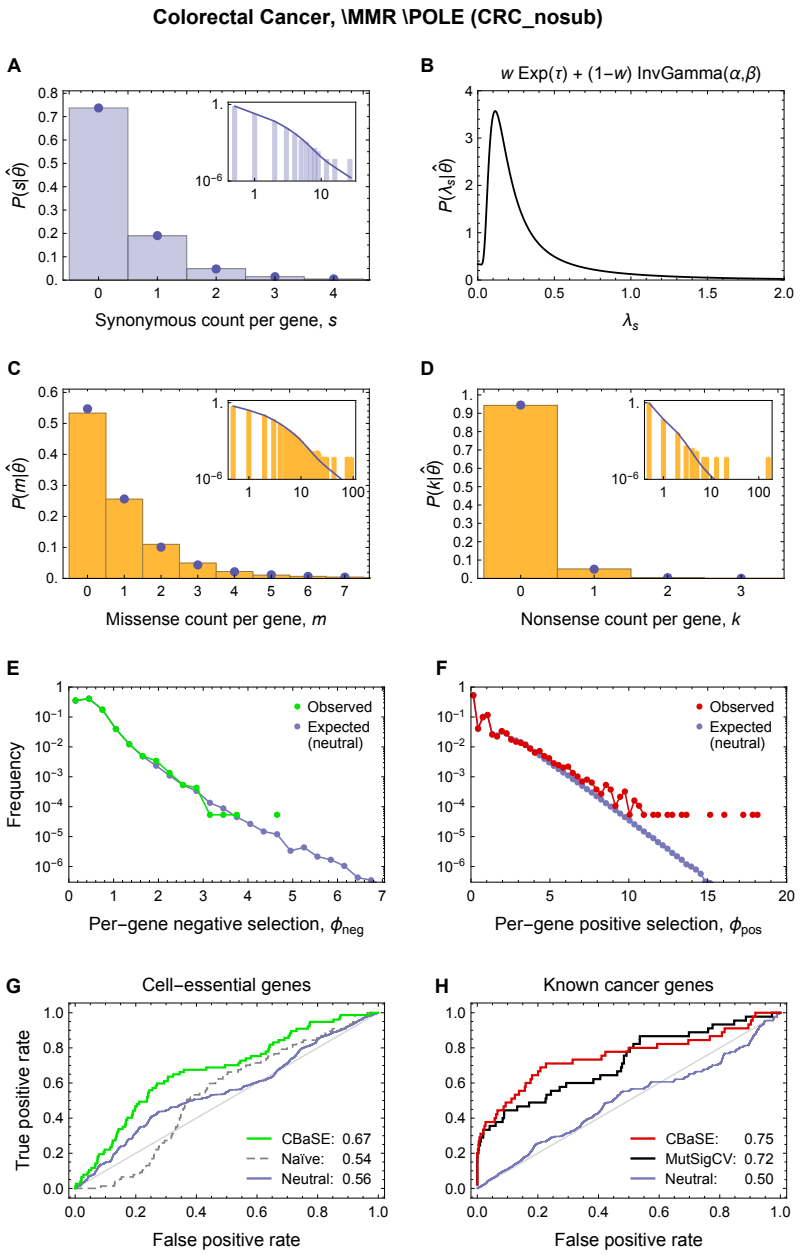
Supplementary Figure 6: CLL. (A),(C),(D) Observed distributions of mutation counts per gene (histograms), and expected distributions under neutral evolution (blue dots and blue lines) with underlying $P(\lambda_s;\boldsymbol{\theta}) = \text{InvGamma}(\lambda_s;\alpha,\beta)$, shown in (B). Parameter estimates $\hat{\boldsymbol{\theta}}$ are $\hat{\alpha} = 2.09$, $\hat{\beta} = 0.045$; sum of squared deviations in (A) is 1e-7. (E),(F) Histogram of the meta-statistic ϕ in real data (green, red) and from simulation under the neutral model (blue), bin width 0.3. Gene (ϕ_{pos}) not shown in (F): *TP53* (22.3). (G) ROC of 77 predicted cell-essential genes from Wang et al.[2] for genes sorted by q_{neg} (green), $p_{\text{AUC}} =$ 1e-1. Dashed gray line shows the ROC from ranking by hypomutation ($p_{\text{AUC}} > 0.05$). (H) ROC of the gene ranking based on q_{pos}, using as true positives 23 CGC genes causally implicated in CLL (red), $p_{\text{AUC}} =$ 1e-6. Black line shows the ROC for MutSigCV[1] predictions. Blue lines in (G) and (H) show neutral ROC expectations from simulation. AUCs are given in the insets.

Nature Genetics: doi:10.1038/ng.3987



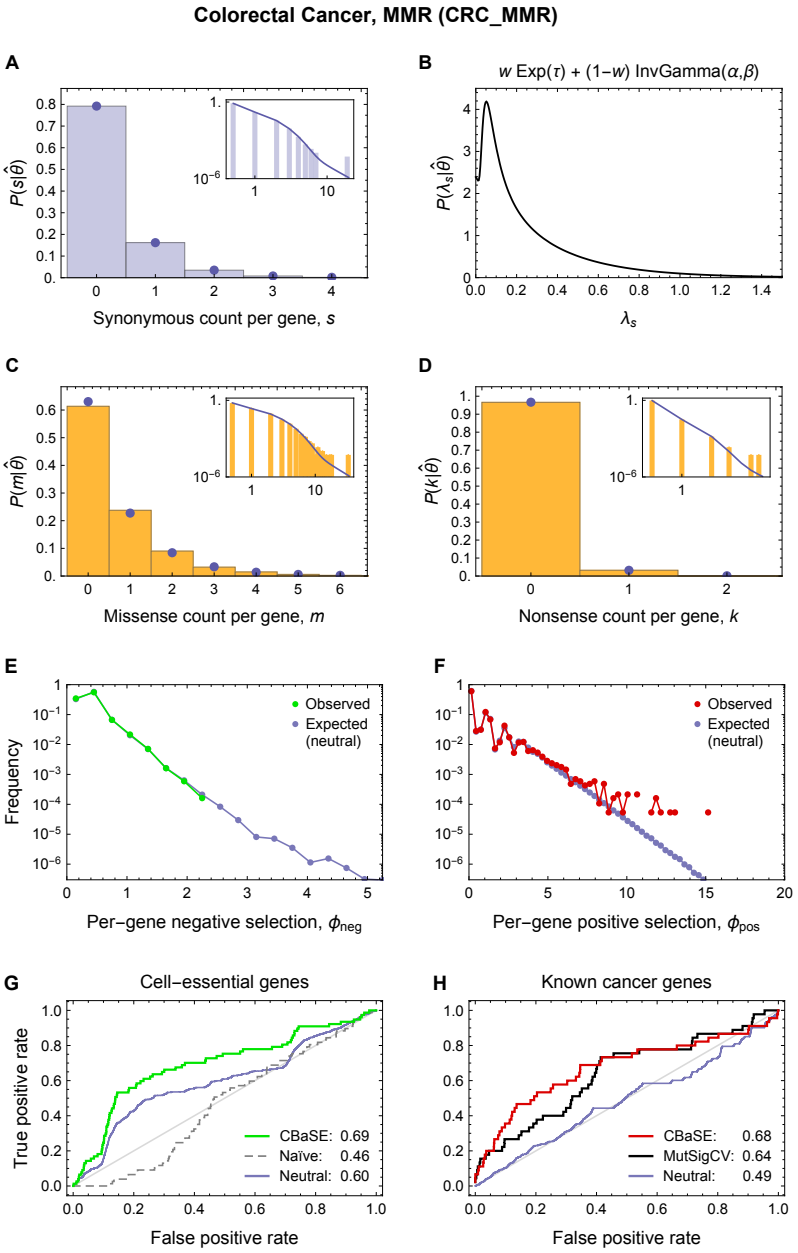
Supplementary Figure 7: CRC. (A),(C),(D) Observed distributions of mutation counts per gene (histograms), and expected distributions under neutral evolution (blue dots and blue lines) with underlying $P(\lambda_s;\boldsymbol{\theta}) = w\,\mathrm{Gamma}(\lambda_s;\alpha,\beta)+(1-w)\,\mathrm{InvGamma}(\lambda_s;\gamma,\delta)$, shown in (B). Parameter estimates $\hat{\boldsymbol{\theta}}$ are $\hat{\alpha} = 0.35$, $\hat{\beta} = 2.38$, $\hat{\gamma} = 2.58$, $\hat{\delta} = 1.74$, $\hat{w} = 0.32$; sum of squared deviations in (A) is 7e-6. (E),(F) Histogram of the meta-statistic ϕ in real data (green, red) and from simulation under the neutral model (blue), bin width 0.3. Genes (ϕ_{pos}) not shown in (F): *SMAD4* (20.1), *ARID1A* (23.1), *KRAS* (39.3), *TP53* (60.2), *APC* ($p_k^{\mathrm{pos}} = 0$). (G) ROC of 77 predicted cell-essential genes from Wang et al.[2] for genes sorted by q_{neg} (green), $p_{\mathrm{AUC}} = 4\text{e-}5$. Dashed gray line shows the ROC from ranking by hypomutation ($p_{\mathrm{AUC}} > 0.05$). (H) ROC of the gene ranking based on q_{pos}, using as true positives 45 CGC genes causally implicated in CRC (red), $p_{\mathrm{AUC}} = 8\text{e-}10$. Black line shows the ROC for MutSigCV[1] predictions. Blue lines in (G) and (H) show neutral ROC expectations from simulation. AUCs are given in the insets.

Nature Genetics: doi:10.1038/ng.3987



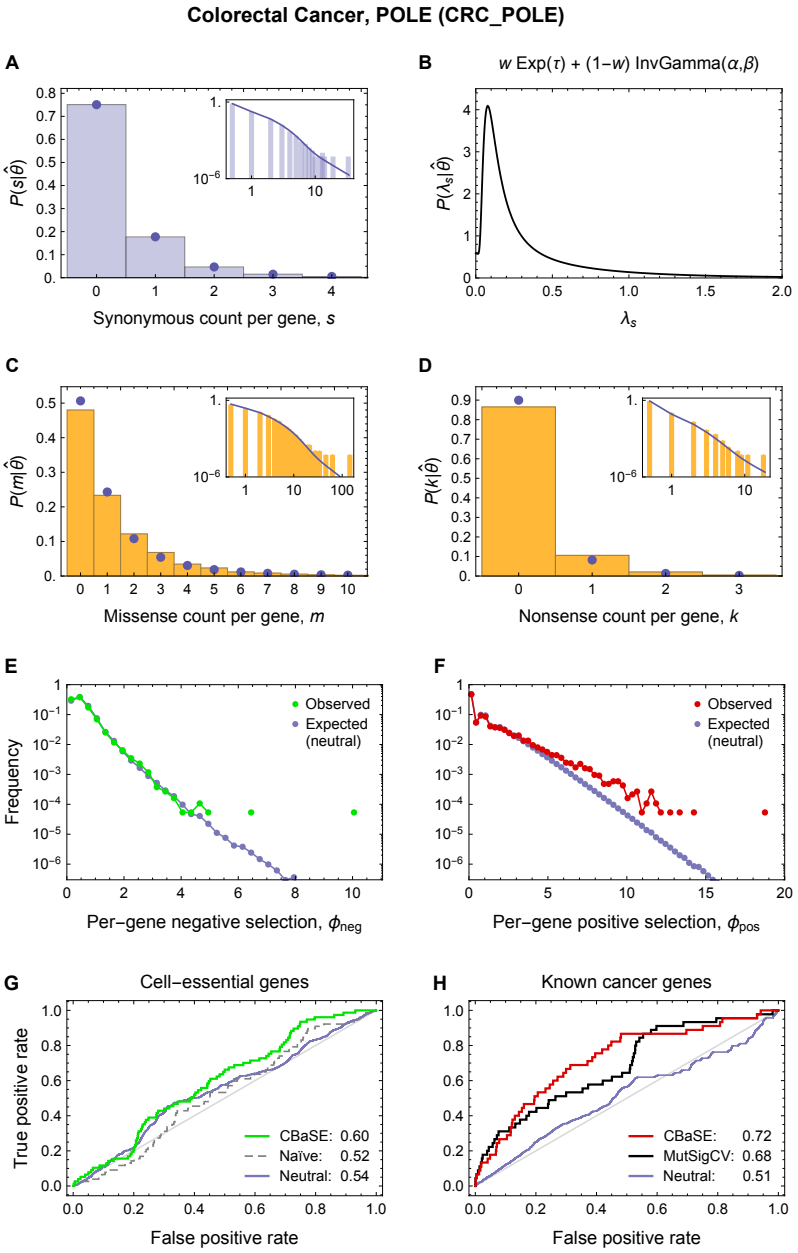
Supplementary Figure 8: CRC (without MMR or POLE subtypes). (A),(C),(D) Observed distributions of mutation counts per gene (histograms), and expected distributions under neutral evolution (blue dots and blue lines) with underlying $P(\lambda_s;\boldsymbol{\theta}) = w\,\mathrm{Exp}(\lambda_s;\tau) + (1-w)\,\mathrm{InvGamma}(\lambda_s;\alpha,\beta)$, shown in (B). Parameter estimates $\hat{\boldsymbol{\theta}}$ are $\hat{\alpha} = 2.23$, $\hat{\beta} = 0.37$, $\hat{\tau} = 1.52$, $\hat{w} = 0.22$; sum of squared deviations in (A) is 2e-6. (E),(F) Histogram of the meta-statistic ϕ in real data (green, red) and from simulation under the neutral model (blue), bin width 0.3. Genes (ϕ_{pos}) not shown in (F): *PIK3CA* (20.6), *TCF7L2* (26.1), *ARID1A* (27.3), *AMER1* (30.0), *KRAS* (34.2), *APC* (40.6), *TP53* (61.5). (G) ROC of 77 predicted cell-essential genes from Wang et al.[2] for genes sorted by q_{neg} (green), $p_{\mathrm{AUC}} = 1\mathrm{e}{-4}$. Dashed gray line shows the ROC from ranking by hypomutation ($p_{\mathrm{AUC}} > 0.05$). (H) ROC of the gene ranking based on q_{pos}, using as true positives 45 CGC genes causally implicated in CRC (red), $p_{\mathrm{AUC}} = 3\mathrm{e}{-7}$. Black line shows the ROC for MutSigCV[1] predictions. Blue lines in (G) and (H) show neutral ROC expectations from simulation. AUCs are given in the insets.

Nature Genetics: doi:10.1038/ng.3987



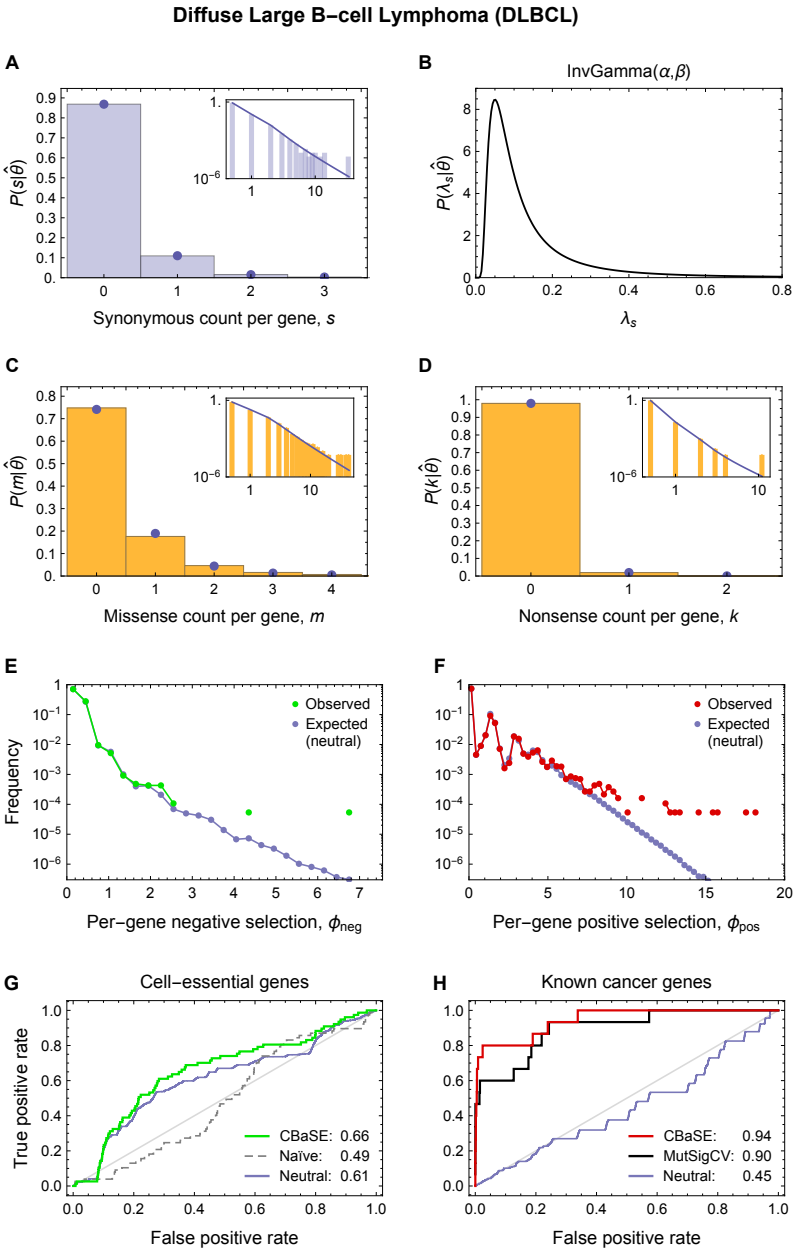
Supplementary Figure 9: CRC (mismatch repair deficient). (A),(C),(D) Observed distributions of mutation counts per gene (histograms), and expected distributions under neutral evolution (blue dots and blue lines) with underlying $P(\lambda_s;\boldsymbol{\theta}) = w\,\mathrm{Exp}(\lambda_s;\tau)+(1-w)\,\mathrm{InvGamma}(\lambda_s;\alpha,\beta)$, shown in (B). Parameter estimates $\hat{\boldsymbol{\theta}}$ are $\hat{\alpha} = 1.97$, $\hat{\beta} = 0.16$, $\hat{\tau} = 3.26$, $\hat{w} = 0.73$; sum of squared deviations in (A) is 5e-8. (E),(F) Histogram of the meta-statistic ϕ in real data (green, red) and from simulation under the neutral model (blue), bin width 0.3. Gene (ϕ_{pos}) not shown in (F): *APC* (23.4). (G) ROC of 77 predicted cell-essential genes from Wang et al.[2] for genes sorted by q_{neg} (green), $p_{\mathrm{AUC}} = 2\text{e-}3$. Dashed gray line shows the ROC from ranking by hypomutation ($p_{\mathrm{AUC}} > 0.05$). (H) ROC of the gene ranking based on q_{pos}, using as true positives 45 CGC genes causally implicated in CRC (red), $p_{\mathrm{AUC}} = 2\text{e-}5$. Black line shows the ROC for MutSigCV[1] predictions. Blue lines in (G) and (H) show neutral ROC expectations from simulation. AUCs are given in the insets.

Nature Genetics: doi:10.1038/ng.3987



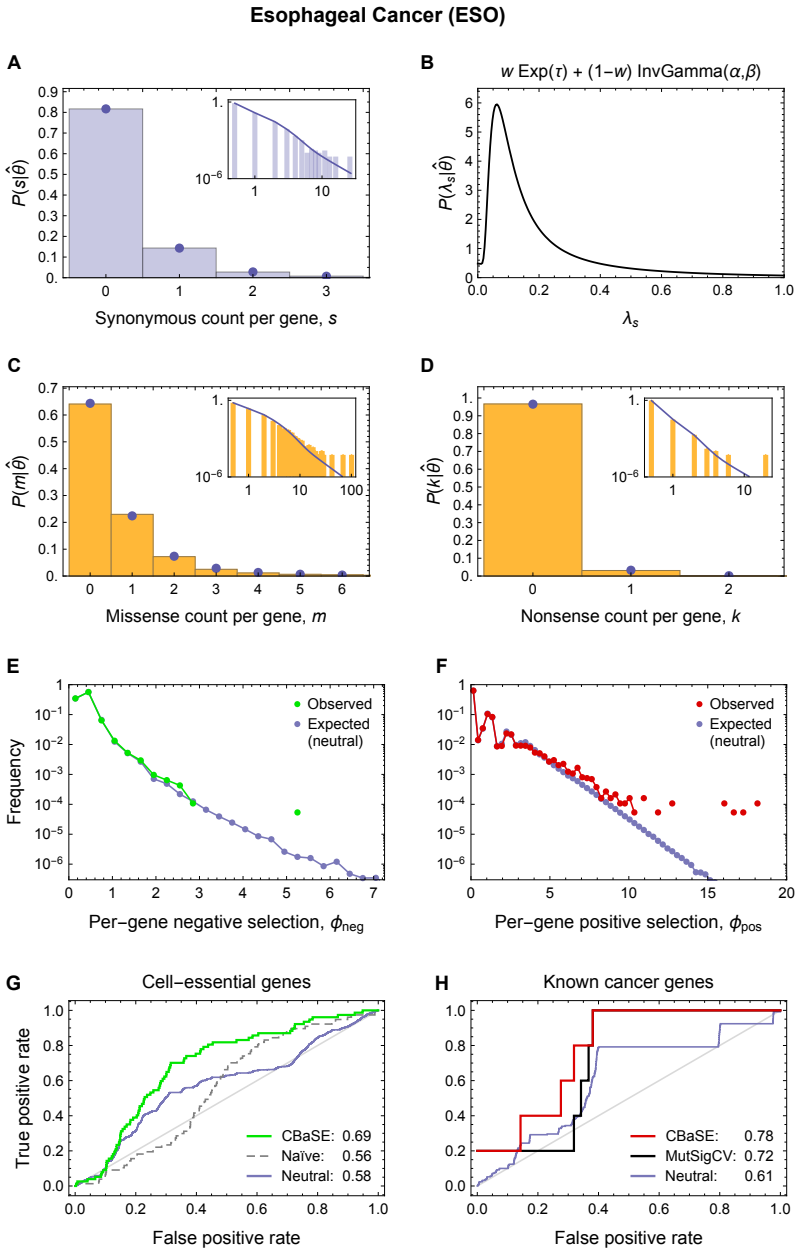
Supplementary Figure 10: CRC (polymerase ϵ deficient). (A),(C),(D) Observed distributions of mutation counts per gene (histograms), and expected distributions under neutral evolution (blue dots and blue lines) with underlying $P(\lambda_s;\boldsymbol{\theta}) = w\,\mathrm{Exp}(\lambda_s;\tau) + (1-w)\,\mathrm{InvGamma}(\lambda_s;\alpha,\beta)$, shown in (B). Parameter estimates $\hat{\boldsymbol{\theta}}$ are $\hat{\alpha} = 1.87$, $\hat{\beta} = 0.23$, $\hat{\tau} = 1.72$, $\hat{w} = 0.34$; sum of squared deviations in (A) is 4e-7. (E),(F) Histogram of the meta-statistic ϕ in real data (green, red) and from simulation under the neutral model (blue), bin width 0.3. (G) ROC of 77 predicted cell-essential genes from Wang et al.[2] for genes sorted by q_{neg} (green), $p_{\mathrm{AUC}} = 6\text{e-}2$. Dashed gray line shows the ROC from ranking by hypomutation ($p_{\mathrm{AUC}} > 0.05$). (H) ROC of the gene ranking based on q_{pos}, using as true positives 45 CGC genes causally implicated in CRC (red), $p_{\mathrm{AUC}} = 9\text{e-}6$. Black line shows the ROC for MutSigCV[1] predictions. Blue lines in (G) and (H) show neutral ROC expectations from simulation. AUCs are given in the insets.

Nature Genetics: doi:10.1038/ng.3987



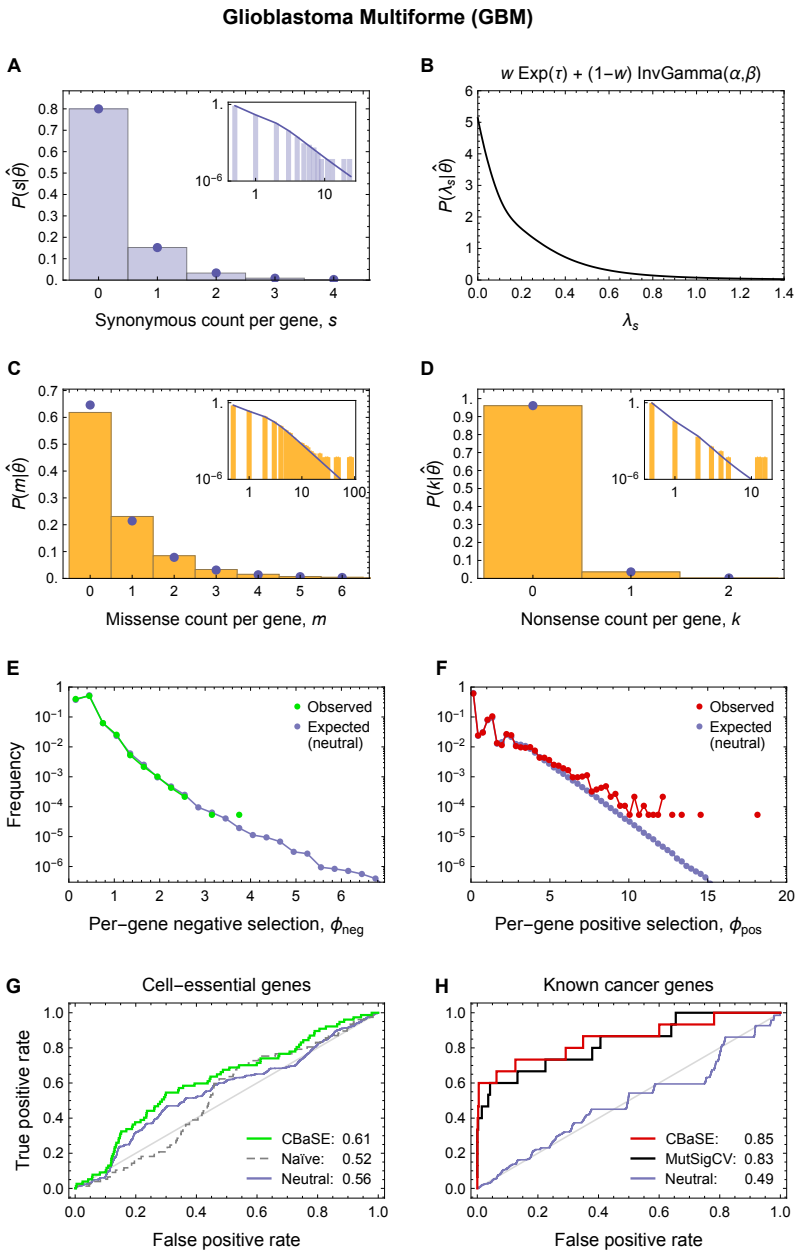
Supplementary Figure 11: DLBCL. (A),(C),(D) Observed distributions of mutation counts per gene (histograms), and expected distributions under neutral evolution (blue dots and blue lines) with underlying $P(\lambda_s;\boldsymbol{\theta}) = \text{InvGamma}(\lambda_s;\alpha,\beta)$, shown in (B). Parameter estimates $\hat{\boldsymbol{\theta}}$ are $\hat{\alpha} = 1.84$, $\hat{\beta} = 0.142$; sum of squared deviations in (A) is 7e-7. (E),(F) Histogram of the meta-statistic ϕ in real data (green, red) and from simulation under the neutral model (blue), bin width 0.3. Gene (ϕ_{pos}) not shown in (F): *KMT2D* (33.2). (G) ROC of 77 predicted cell-essential genes from Wang et al.[2] for genes sorted by q_{neg} (green), $p_{\text{AUC}} = 4\text{e-}2$. Dashed gray line shows the ROC from ranking by hypomutation ($p_{\text{AUC}} > 0.05$). (H) ROC of the gene ranking based on q_{pos}, using as true positives 15 CGC genes causally implicated in DLBCL (red), $p_{\text{AUC}} = 5\text{e-}12$. Black line shows the ROC for MutSigCV[1] predictions. Blue lines in (G) and (H) show neutral ROC expectations from simulation. AUCs are given in the insets.

Nature Genetics: doi:10.1038/ng.3987



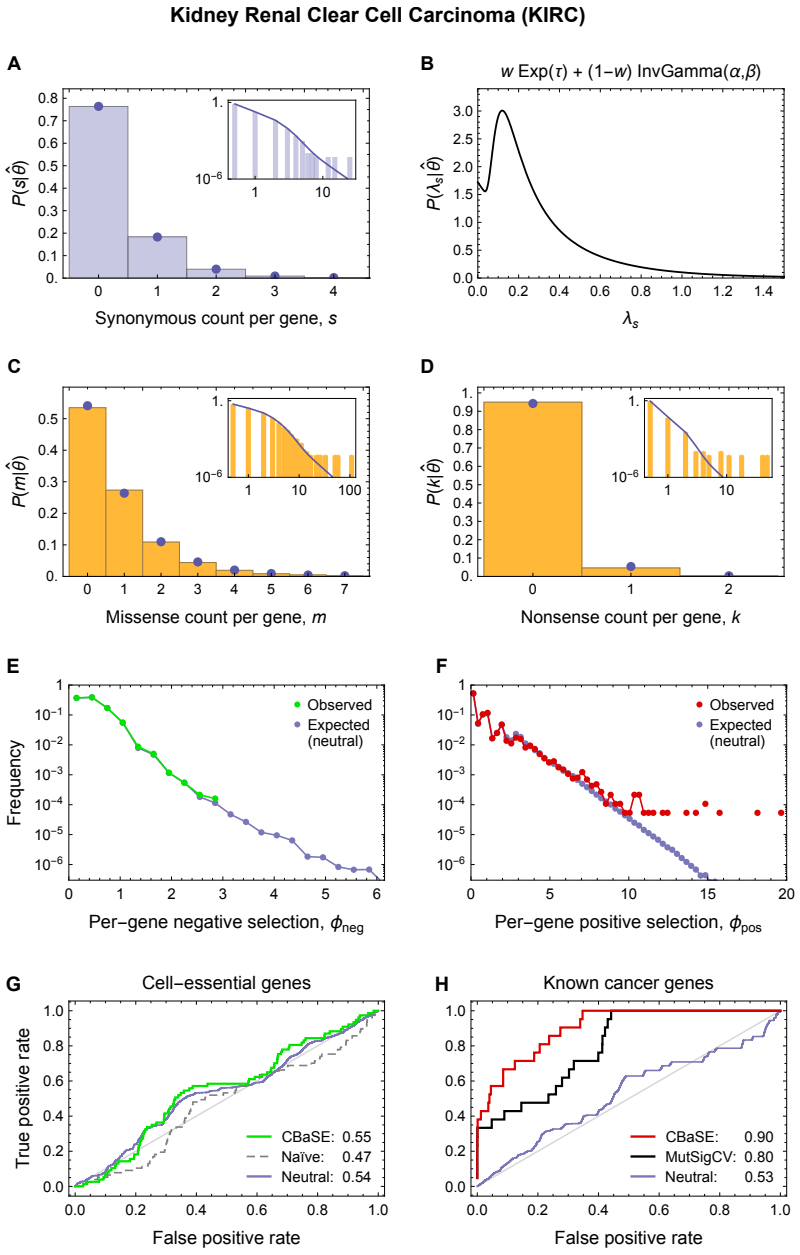
Supplementary Figure 12: ESO. (A),(C),(D) Observed distributions of mutation counts per gene (histograms), and expected distributions under neutral evolution (blue dots and blue lines) with underlying $P(\lambda_s;\boldsymbol{\theta}) = w\,\mathrm{Exp}(\lambda_s;\tau) + (1-w)\,\mathrm{InvGamma}(\lambda_s;\alpha,\beta)$, shown in (B). Parameter estimates $\hat{\boldsymbol{\theta}}$ are $\hat{\alpha} = 1.89$, $\hat{\beta} = 0.18$, $\hat{\tau} = 2.37$, $\hat{w} = 0.20$; sum of squared deviations in (A) is 8e-7. (E),(F) Histogram of the meta-statistic ϕ in real data (green, red) and from simulation under the neutral model (blue), bin width 0.3. Gene (ϕ_{pos}) not shown in (F): *TP53* (62.9). (G) ROC of 77 predicted cell-essential genes from Wang et al.[2] for genes sorted by q_{neg} (green), $p_{\mathrm{AUC}} = 6\text{e-}5$. Dashed gray line shows the ROC from ranking by hypomutation ($p_{\mathrm{AUC}} = 0.05$). (H) ROC of the gene ranking based on q_{pos}, using as true positives 5 CGC genes causally implicated in ESO (red), $p_{\mathrm{AUC}} = 4\text{e-}2$. Black line shows the ROC for MutSigCV[1] predictions. Blue lines in (G) and (H) show neutral ROC expectations from simulation. AUCs are given in the insets.

Nature Genetics: doi:10.1038/ng.3987



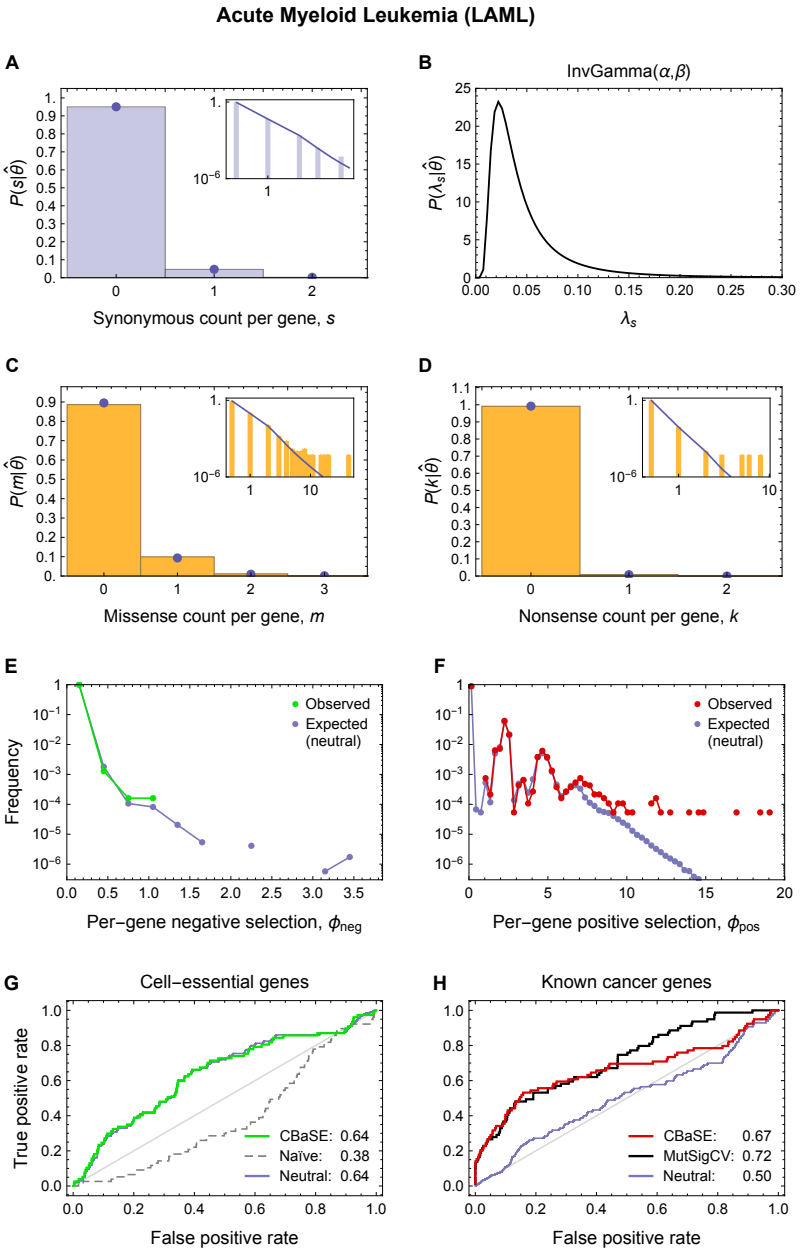
Supplementary Figure 13: GBM. (A),(C),(D) Observed distributions of mutation counts per gene (histograms), and expected distributions under neutral evolution (blue dots and blue lines) with underlying $P(\lambda_s;\boldsymbol{\theta}) = w\,\mathrm{Exp}(\lambda_s;\tau) + (1-w)\,\mathrm{InvGamma}(\lambda_s;\alpha,\beta)$, shown in (B). Parameter estimates $\hat{\boldsymbol{\theta}}$ are $\hat{\alpha} = 2.66$, $\hat{\beta} = 1.10$, $\hat{\tau} = 6.84$, $\hat{w} = 0.75$; sum of squared deviations in (A) is 3e-7. (E),(F) Histogram of the meta-statistic ϕ in real data (green, red) and from simulation under the neutral model (blue), bin width 0.3. Genes (ϕ_{pos}) not shown in (F): *PIK3CA* (20.8), *RB1* (21.8), *EGFR* (26.4), *NF1* (30.8), *TP53* (44.2), *PTEN* (57.1). (G) ROC of 77 predicted cell-essential genes from Wang et al.[2] for genes sorted by q_{neg} (green), $p_{\mathrm{AUC}} = 4\text{e-}2$. Dashed gray line shows the ROC from ranking by hypomutation ($p_{\mathrm{AUC}} > 0.05$). (H) ROC of the gene ranking based on q_{pos}, using as true positives 16 CGC genes causally implicated in GBM (red), $p_{\mathrm{AUC}} = 3\text{e-}7$. Black line shows the ROC for MutSigCV[1] predictions. Blue lines in (G) and (H) show neutral ROC expectations from simulation. AUCs are given in the insets.

Nature Genetics: doi:10.1038/ng.3987



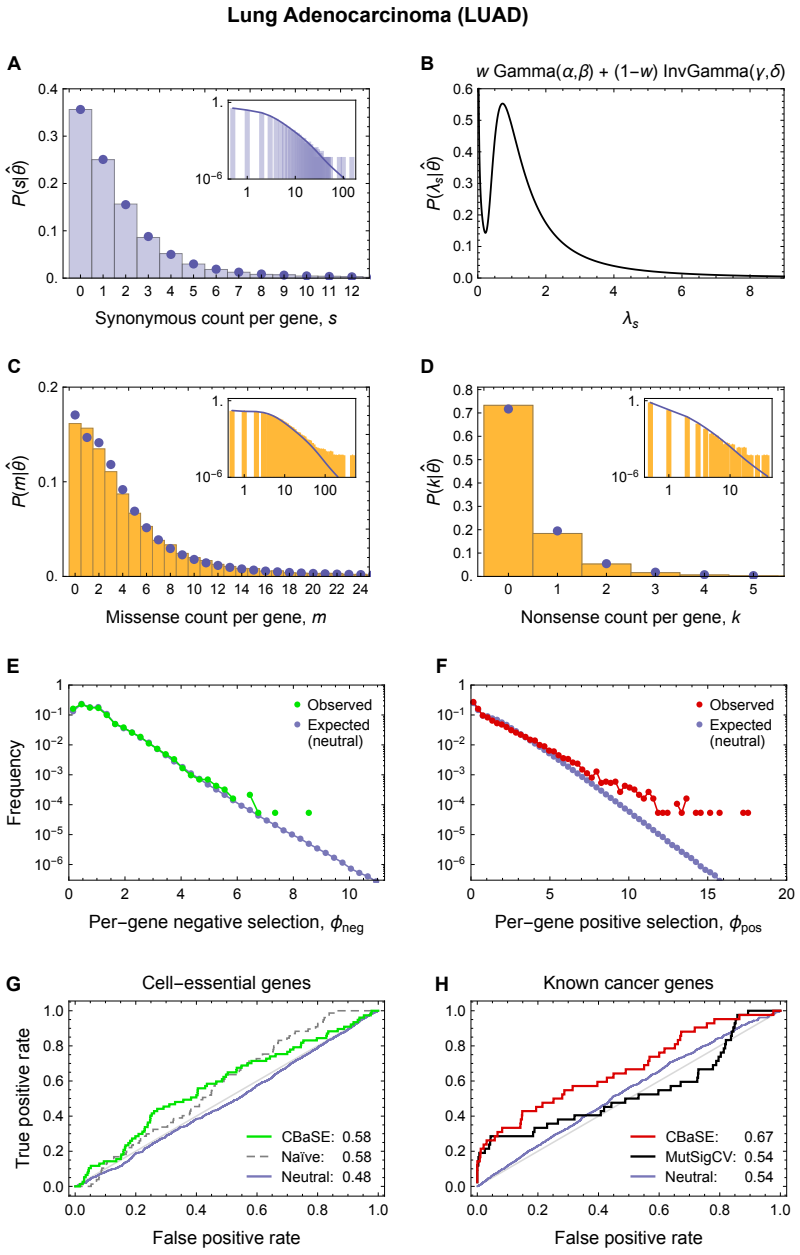
Supplementary Figure 14: KIRC. (A),(C),(D) Observed distributions of mutation counts per gene (histograms), and expected distributions under neutral evolution (blue dots and blue lines) with underlying $P(\lambda_s;\boldsymbol{\theta}) = w\,\mathrm{Exp}(\lambda_s;\tau) + (1-w)\,\mathrm{InvGamma}(\lambda_s;\alpha,\beta)$, shown in (B). Parameter estimates $\hat{\boldsymbol{\theta}}$ are $\hat{\alpha} = 2.46$, $\hat{\beta} = 0.44$, $\hat{\tau} = 3.18$, $\hat{w} = 0.54$; sum of squared deviations in (A) is 4e-7. (E),(F) Histogram of the meta-statistic ϕ in real data (green, red) and from simulation under the neutral model (blue), bin width 0.3. Genes (ϕ_{pos}) not shown in (F): *KDM5C* (27.2), *BAP1* (40.7), *SETD2* (47.4), *PBRM1* ($p_k^{\mathrm{pos}} = 0$), *VHL* ($p_k^{\mathrm{pos}} = 0$). (G) ROC of 77 predicted cell-essential genes from Wang et al.[2] for genes sorted by q_{neg} (green), $p_{\mathrm{AUC}} = 4\text{e-}1$. Dashed gray line shows the ROC from ranking by hypomutation ($p_{\mathrm{AUC}} > 0.05$). (H) ROC of the gene ranking based on q_{pos}, using as true positives 23 CGC genes causally implicated in KIRC (red), $p_{\mathrm{AUC}} = 2\text{e-}8$. Black line shows the ROC for MutSigCV[1] predictions. Blue lines in (G) and (H) show neutral ROC expectations from simulation. AUCs are given in the insets.

Nature Genetics: doi:10.1038/ng.3987



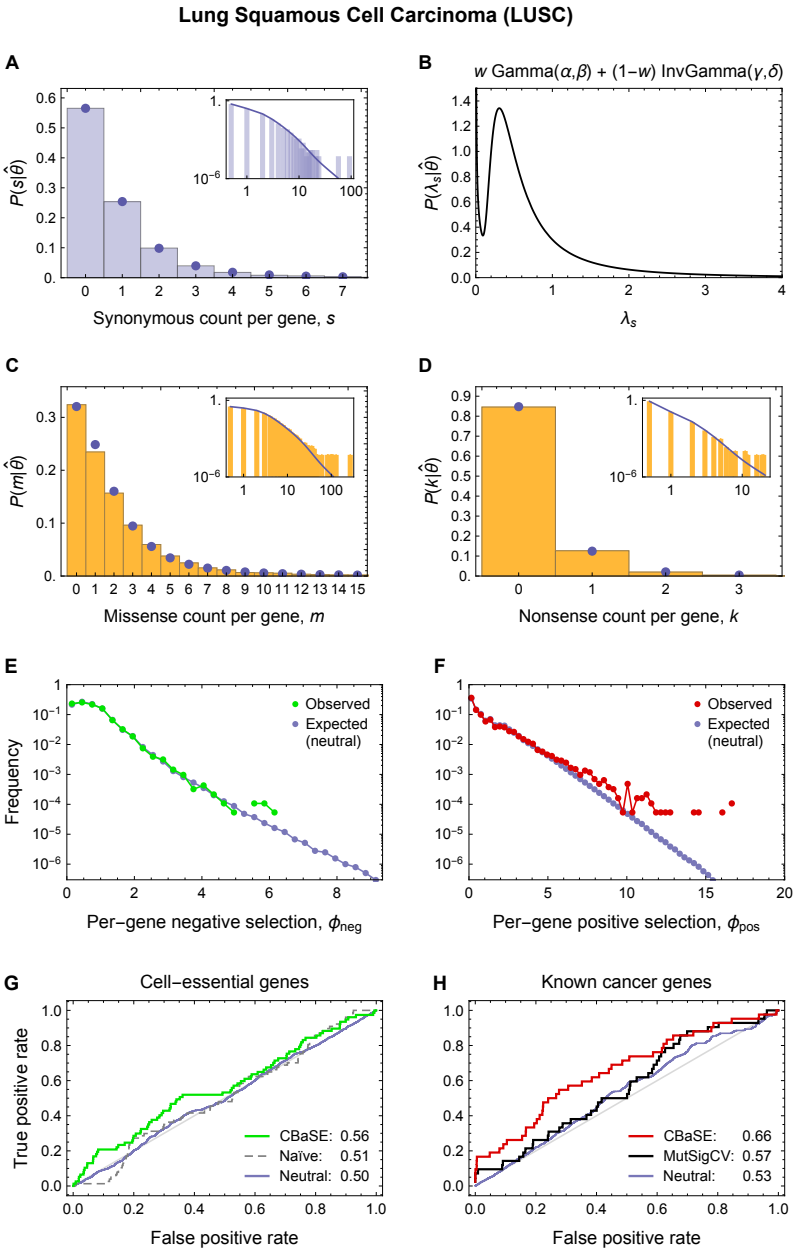
Supplementary Figure 15: LAML. (A),(C),(D) Observed distributions of mutation counts per gene (histograms), and expected distributions under neutral evolution (blue dots and blue lines) with underlying $P(\lambda_s;\boldsymbol{\theta}) = \text{InvGamma}(\lambda_s;\alpha,\beta)$, shown in (B). Parameter estimates $\hat{\boldsymbol{\theta}}$ are $\hat{\alpha} = 2.44$, $\hat{\beta} = 0.08$; sum of squared deviations in (A) is 1e-8. (E),(F) Histogram of the meta-statistic ϕ in real data (green, red) and from simulation under the neutral model (blue), bin width 0.3. Genes (ϕ_{pos}) not shown in (F): *IDH2* (20.7), *TET2* (34.2), *RUNX1* (37.5), *DNMT3A* (57.3). (G) ROC of 77 predicted cell-essential genes from Wang et al.[2] for genes sorted by q_{neg} (green), $p_{\text{AUC}} =$ 5e-1. Dashed gray line shows the ROC from ranking by hypomutation ($p_{\text{AUC}} > 0.05$). (H) ROC of the gene ranking based on q_{pos}, using as true positives 83 CGC genes causally implicated in LAML (red), $p_{\text{AUC}} =$ 6e-16. Black line shows the ROC for MutSigCV[1] predictions. Blue lines in (G) and (H) show neutral ROC expectations from simulation. AUCs are given in the insets.

Nature Genetics: doi:10.1038/ng.3987



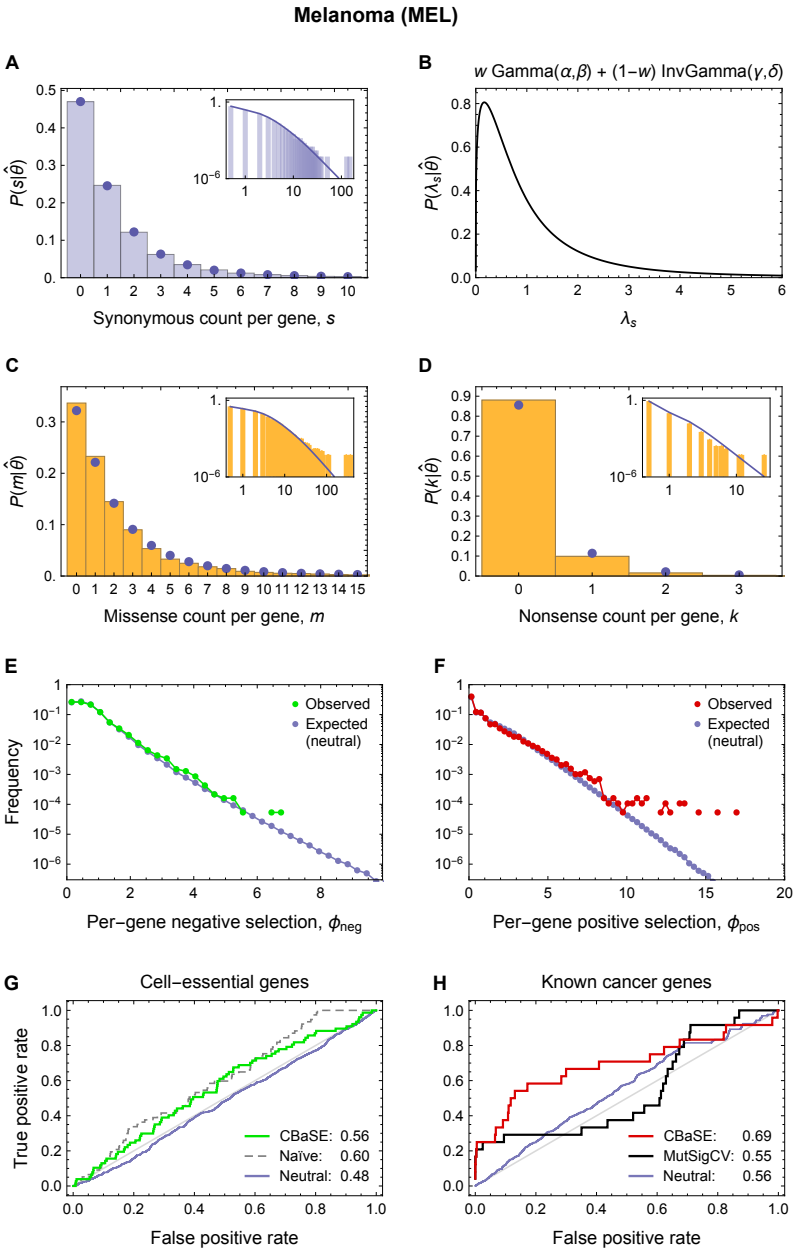
Supplementary Figure 16: LUAD. (A),(C),(D) Observed distributions of mutation counts per gene (histograms), and expected distributions under neutral evolution (blue dots and blue lines) with underlying $P(\lambda_s;\boldsymbol{\theta}) = w\,\mathrm{Gamma}(\lambda_s;\alpha,\beta)+(1-w)\,\mathrm{InvGamma}(\lambda_s;\gamma,\delta)$, shown in (B). Parameter estimates $\hat{\boldsymbol{\theta}}$ are $\hat{\alpha} = 0.31$, $\hat{\beta} = 5.74$, $\hat{\gamma} = 2.30$, $\hat{\delta} = 2.47$, $\hat{w} = 0.24$; sum of squared deviations in (A) is 1e-5. (E),(F) Histogram of the meta-statistic ϕ in real data (green, red) and from simulation under the neutral model (blue), bin width 0.3. Genes (ϕ_{pos}) not shown in (F): *SMARCA4* (23.5.), *ARID1A* (24.9), *KRAS* (29.8), *KEAP1* (35.6), *STK11* (37.9), *TP53* ($p_k^{\mathrm{pos}} = 0$). (G) ROC of 77 predicted cell-essential genes from Wang et al.[2] for genes sorted by q_{neg} (green), $p_{\mathrm{AUC}} = 9\text{e-}4$. Dashed gray line shows the ROC from ranking by hypomutation ($p_{\mathrm{AUC}} = 1\text{e-}2$). (H) ROC of the gene ranking based on q_{pos}, using as true positives 44 CGC genes causally implicated in LUAD (red), $p_{\mathrm{AUC}} = 3\text{e-}4$. Black line shows the ROC for MutSigCV[1] predictions. Blue lines in (G) and (H) show neutral ROC expectations from simulation. AUCs are given in the insets.

Nature Genetics: doi:10.1038/ng.3987



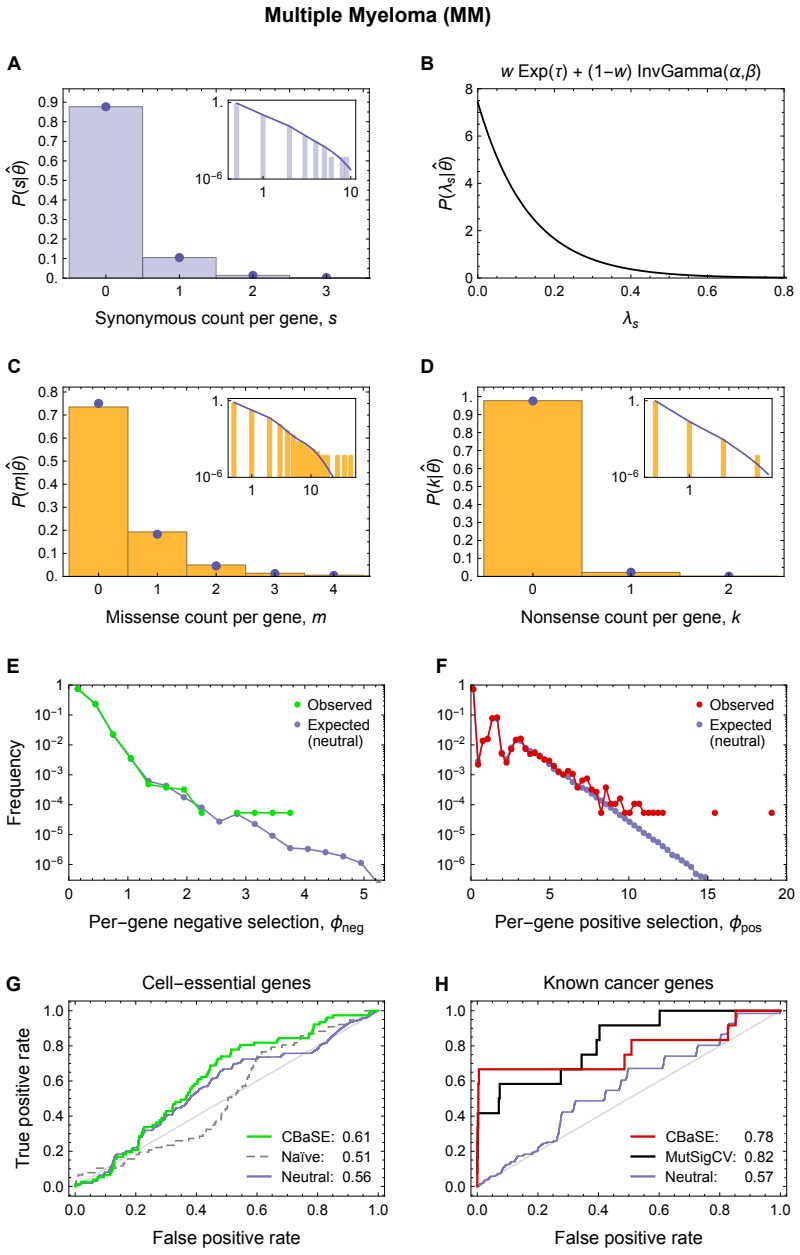
Supplementary Figure 17: LUSC. (A),(C),(D) Observed distributions of mutation counts per gene (histograms), and expected distributions under neutral evolution (blue dots and blue lines) with underlying $P(\lambda_s;\boldsymbol{\theta}) = w\,\mathrm{Gamma}(\lambda_s;\alpha,\beta)+(1-w)\,\mathrm{InvGamma}(\lambda_s;\gamma,\delta)$, shown in (B). Parameter estimates $\hat{\boldsymbol{\theta}}$ are $\hat{\alpha}=0.41$, $\hat{\beta}=2.21$, $\hat{\gamma}=2.33$, $\hat{\delta}=1.05$, $\hat{w}=0.24$; sum of squared deviations in (A) is 3e-6. (E),(F) Histogram of the meta-statistic ϕ in real data (green, red) and from simulation under the neutral model (blue), bin width 0.3. Genes (ϕ_{pos}) not shown in (F): *KMT2D* (20.9), *TP53* (48.3). (G) ROC of 77 predicted cell-essential genes from Wang et al.[2] for genes sorted by q_{neg} (green), p_{AUC} = 3e-2. Dashed gray line shows the ROC from ranking by hypomutation ($p_{\mathrm{AUC}} > 0.05$). (H) ROC of the gene ranking based on q_{pos}, using as true positives 44 CGC genes causally implicated in LUSC (red), p_{AUC} = 2e-3. Black line shows the ROC for MutSigCV[1] predictions. Blue lines in (G) and (H) show neutral ROC expectations from simulation. AUCs are given in the insets.

Nature Genetics: doi:10.1038/ng.3987



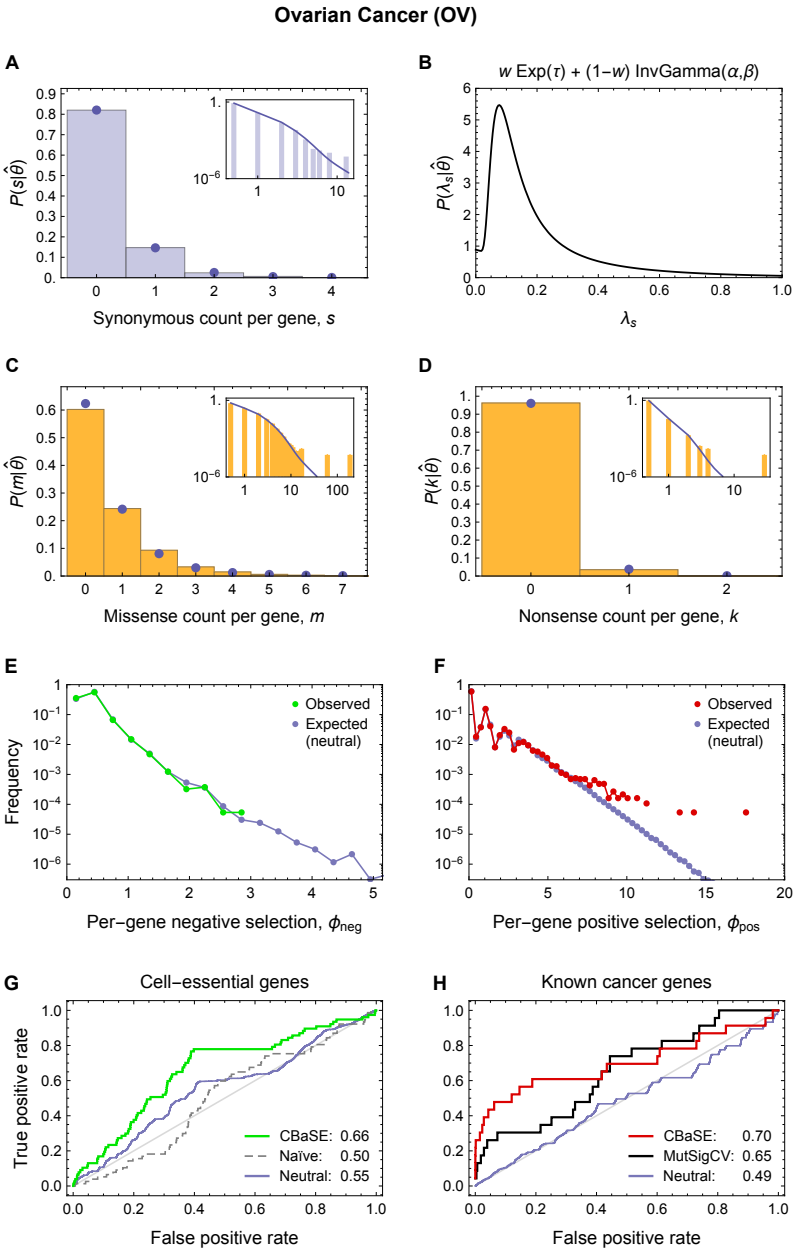
Supplementary Figure 18: MEL. (A),(C),(D) Observed distributions of mutation counts per gene (histograms), and expected distributions under neutral evolution (blue dots and blue lines) with underlying $P(\lambda_s;\boldsymbol{\theta}) = w\,\mathrm{Gamma}(\lambda_s;\alpha,\beta)+(1-w)\,\mathrm{InvGamma}(\lambda_s;\gamma,\delta)$, shown in (B). Parameter estimates $\hat{\boldsymbol{\theta}}$ are $\hat{\alpha} = 1.26$, $\hat{\beta} = 0.65$, $\hat{\gamma} = 2.91$, $\hat{\delta} = 9.15$, $\hat{w} = 0.87$; sum of squared deviations in (A) is 5e-6. (E),(F) Histogram of the meta-statistic ϕ in real data (green, red) and from simulation under the neutral model (blue), bin width 0.3. Genes (ϕ_{pos}) not shown in (F): *BRAF* (20.8), *TP53* (21.9). (G) ROC of 77 predicted cell-essential genes from Wang et al.[2] for genes sorted by q_{neg} (green), p_{AUC} = 5e-3. Dashed gray line shows the ROC from ranking by hypomutation (p_{AUC} = 1e-3). (H) ROC of the gene ranking based on q_{pos}, using as true positives 25 CGC genes causally implicated in MEL (red), p_{AUC} = 1e-2. Black line shows the ROC for MutSigCV[1] predictions. Blue lines in (G) and (H) show neutral ROC expectations from simulation. AUCs are given in the insets.

Nature Genetics: doi:10.1038/ng.3987



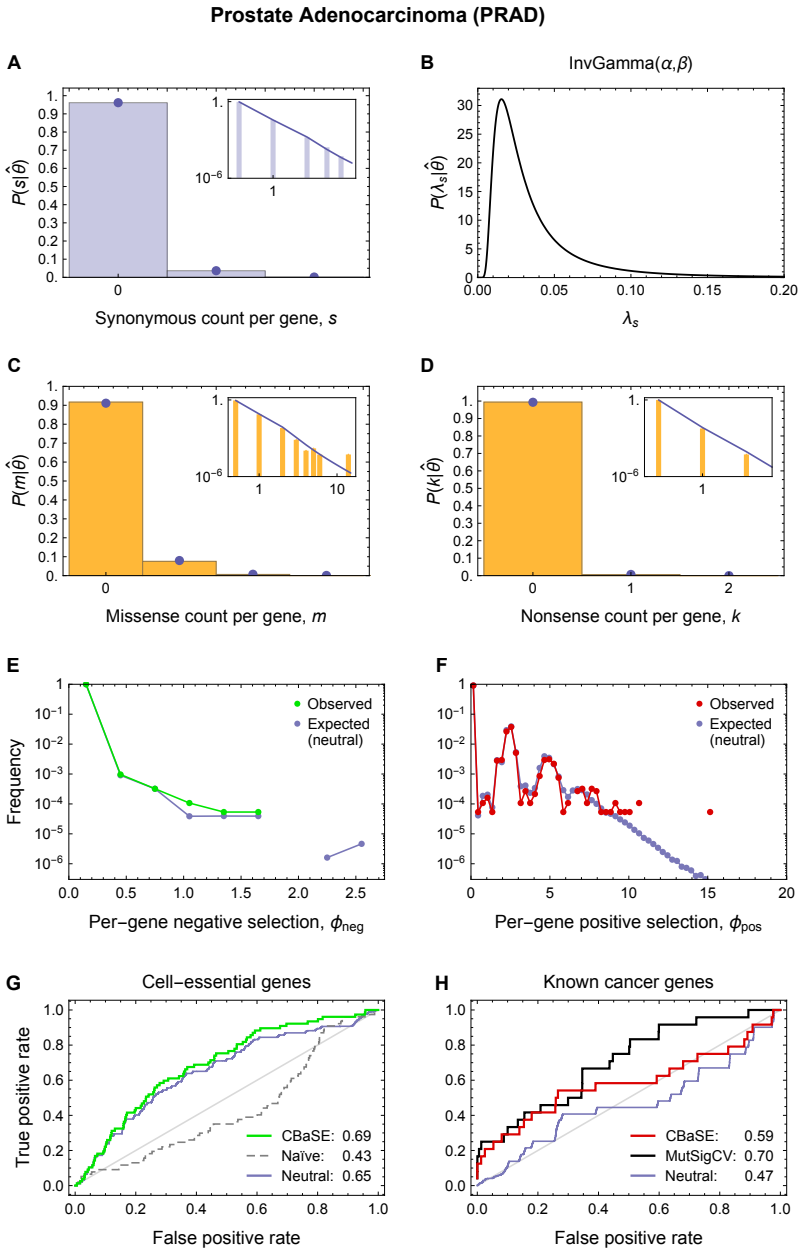
Supplementary Figure 19: MM. (A),(C),(D) Observed distributions of mutation counts per gene (histograms), and expected distributions under neutral evolution (blue dots and blue lines) with underlying $P(\lambda_s;\boldsymbol{\theta}) = w\,\mathrm{Exp}(\lambda_s;\tau) + (1-w)\,\mathrm{InvGamma}(\lambda_s;\alpha,\beta)$, shown in (B). Parameter estimates $\hat{\boldsymbol{\theta}}$ are $\hat{\alpha} = 8.85$, $\hat{\beta} = 16.39$, $\hat{\tau} = 7.48$, $\hat{w} = 0.993$; sum of squared deviations in (A) is 9e-8. (E),(F) Histogram of the meta-statistic ϕ in real data (green, red) and from simulation under the neutral model (blue), bin width 0.3. Genes (ϕ_{pos}) not shown in (F): *TRAF3* (21.9), *NRAS* (25.3), *KRAS* (28.3). (G) ROC of 77 predicted cell-essential genes from Wang et al.[2] for genes sorted by q_{neg} (green), $p_{\mathrm{AUC}} = 2\text{e-}2$. Dashed gray line shows the ROC from ranking by hypomutation ($p_{\mathrm{AUC}} > 0.05$). (H) ROC of the gene ranking based on q_{pos}, using as true positives 13 CGC genes causally implicated in MM (red), $p_{\mathrm{AUC}} = 6\text{e-}4$. Black line shows the ROC for MutSigCV[1] predictions. Blue lines in (G) and (H) show neutral ROC expectations from simulation. AUCs are given in the insets.

Nature Genetics: doi:10.1038/ng.3987



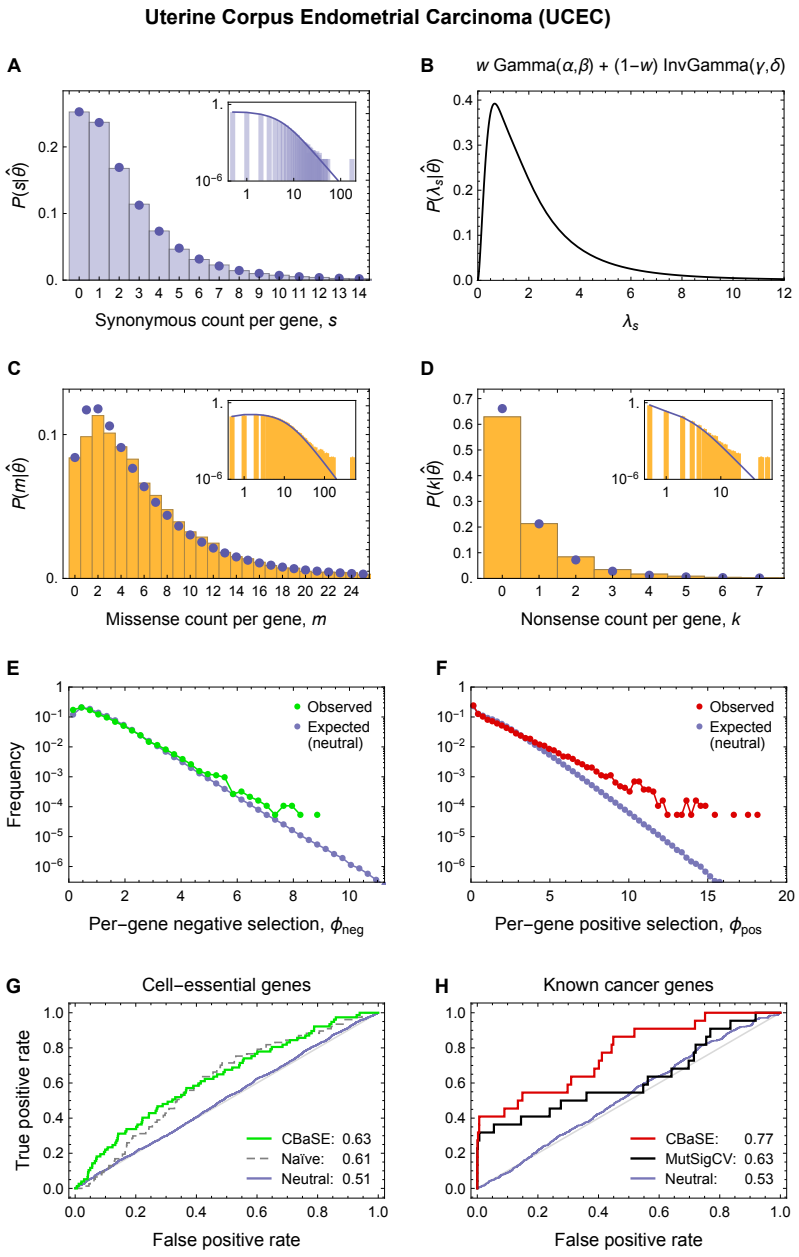
Supplementary Figure 20: OV. (A),(C),(D) Observed distributions of mutation counts per gene (histograms), and expected distributions under neutral evolution (blue dots and blue lines) with underlying $P(\lambda_s;\boldsymbol{\theta}) = w\,\mathrm{Exp}(\lambda_s;\tau) + (1-w)\,\mathrm{InvGamma}(\lambda_s;\alpha,\beta)$, shown in (B). Parameter estimates $\hat{\boldsymbol{\theta}}$ are $\hat{\alpha} = 2.49$, $\hat{\beta} = 0.27$, $\hat{\tau} = 3.05$, $\hat{w} = 0.29$; sum of squared deviations in (A) is 4e-6. (E),(F) Histogram of the meta-statistic ϕ in real data (green, red) and from simulation under the neutral model (blue), bin width 0.3. Gene (ϕ_{pos}) not shown in (F): *TP53* ($p_m^{\mathrm{pos}} = 0$). (G) ROC of 77 predicted cell-essential genes from Wang et al.[2] for genes sorted by q_{neg} (green), $p_{\mathrm{AUC}} = 6\text{e-}4$. Dashed gray line shows the ROC from ranking by hypomutation ($p_{\mathrm{AUC}} > 0.05$). (H) ROC of the gene ranking based on q_{pos}, using as true positives 24 CGC genes causally implicated in OV (red), $p_{\mathrm{AUC}} = 2\text{e-}4$. Black line shows the ROC for MutSigCV[1] predictions. Blue lines in (G) and (H) show neutral ROC expectations from simulation. AUCs are given in the insets.

Nature Genetics: doi:10.1038/ng.3987



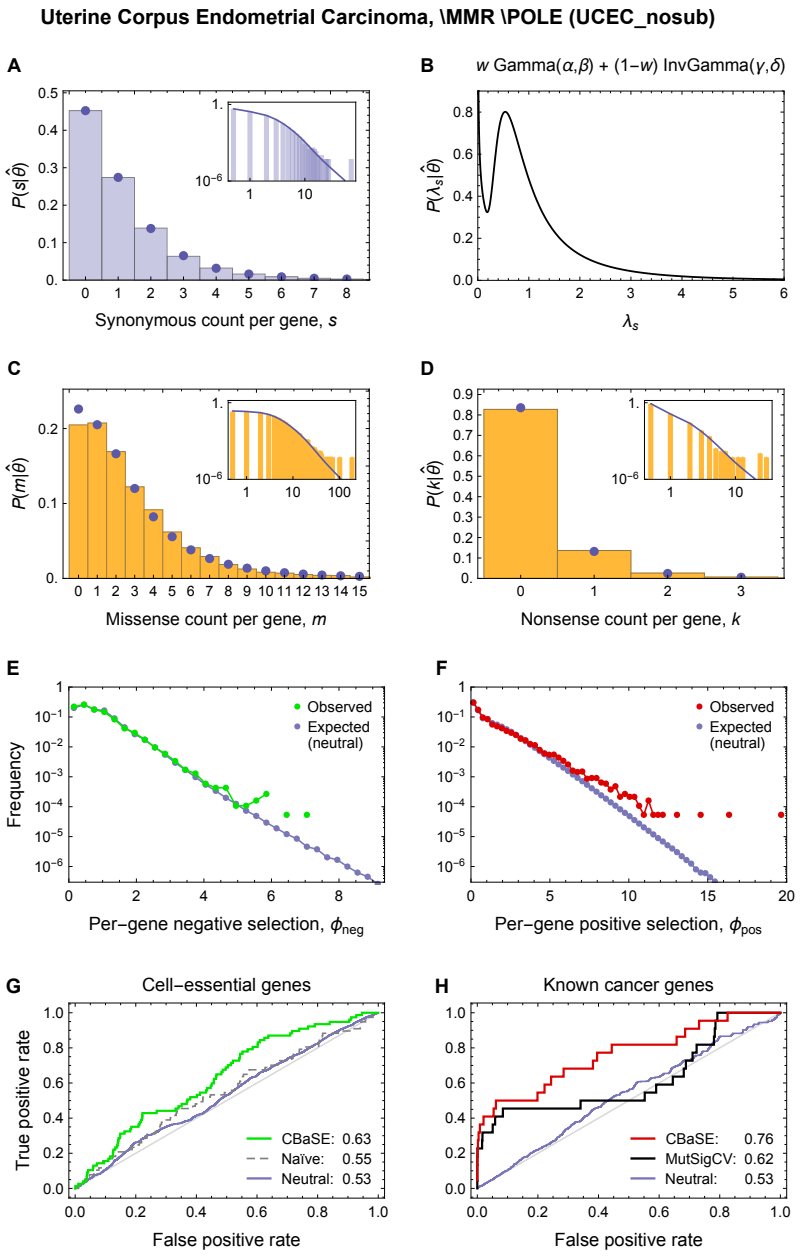
Supplementary Figure 21: PRAD. (A),(C),(D) Observed distributions of mutation counts per gene (histograms), and expected distributions under neutral evolution (blue dots and blue lines) with underlying $P(\lambda_s; \boldsymbol{\theta}) = \text{InvGamma}(\lambda_s; \alpha, \beta)$, shown in (B). Parameter estimates $\hat{\boldsymbol{\theta}}$ are $\hat{\alpha} = 2.22$, $\hat{\beta} = 0.05$; sum of squared deviations in (A) is 8e-9. (E),(F) Histogram of the meta-statistic ϕ in real data (green, red) and from simulation under the neutral model (blue), bin width 0.3. (G) ROC of 77 predicted cell-essential genes from Wang et al.[2] for genes sorted by q_{neg} (green), $p_{\text{AUC}} = 6\text{e-}2$. Dashed gray line shows the ROC from ranking by hypomutation ($p_{\text{AUC}} > 0.05$). (H) ROC of the gene ranking based on q_{pos}, using as true positives 26 CGC genes causally implicated in PRAD (red), $p_{\text{AUC}} = 8\text{e-}4$. Black line shows the ROC for MutSigCV[1] predictions. Blue lines in (G) and (H) show neutral ROC expectations from simulation. AUCs are given in the insets.

Nature Genetics: doi:10.1038/ng.3987



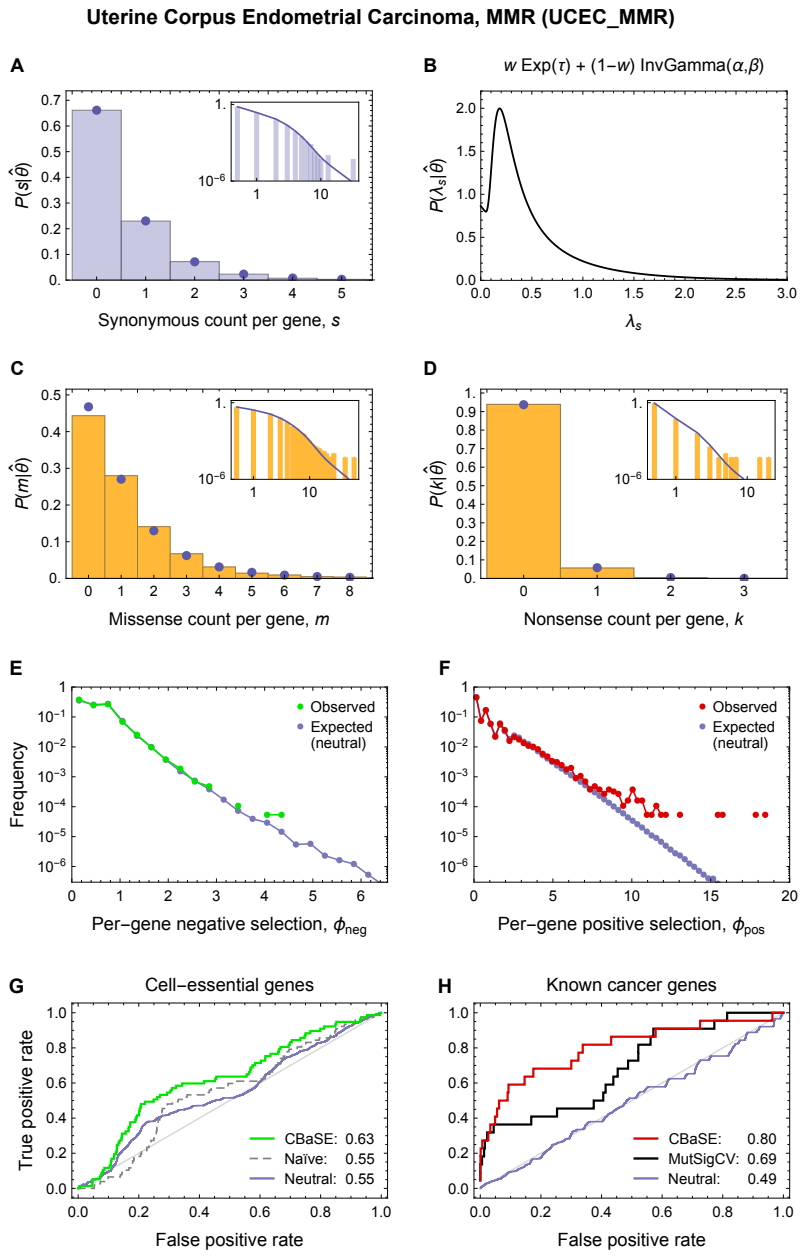
Supplementary Figure 22: UCEC. (A),(C),(D) Observed distributions of mutation counts per gene (histograms), and expected distributions under neutral evolution (blue dots and blue lines) with underlying $P(\lambda_s;\boldsymbol{\theta}) = w\,\mathrm{Gamma}(\lambda_s;\alpha,\beta)+(1-w)\,\mathrm{InvGamma}(\lambda_s;\gamma,\delta)$, shown in (B). Parameter estimates $\hat{\boldsymbol{\theta}}$ are $\hat{\alpha} = 2.76$, $\hat{\beta} = 0.36$, $\hat{\gamma} = 3.27$, $\hat{\delta} = 8.67$, $\hat{w} = 0.49$; sum of squared deviations in (A) is 1e-5. (E),(F) Histogram of the meta-statistic ϕ in real data (green, red) and from simulation under the neutral model (blue), bin width 0.3. Genes (ϕ_{pos}) not shown in (F): *CCDC147* (21.5), *ARHGAP35* (22.2), *PIK3R1* (24.2), *TP53* (29.1), *PIK3CA* (32.0), *CTCF* (37.2), *PTEN* ($p_k^{\mathrm{pos}} = 0$), *ARID1A* ($p_k^{\mathrm{pos}} = 0$). (G) ROC of 77 predicted cell-essential genes from Wang et al.[2] for genes sorted by q_{neg} (green), p_{AUC} = 2e-5. Dashed gray line shows the ROC from ranking by hypomutation (p_{AUC} = 5e-4). (H) ROC of the gene ranking based on q_{pos}, using as true positives 24 CGC genes causally implicated in UCEC (red), p_{AUC} = 2e-5. Black line shows the ROC for MutSigCV[1] predictions. Blue lines in (G) and (H) show neutral ROC expectations from simulation. AUCs are given in the insets.

Nature Genetics: doi:10.1038/ng.3987



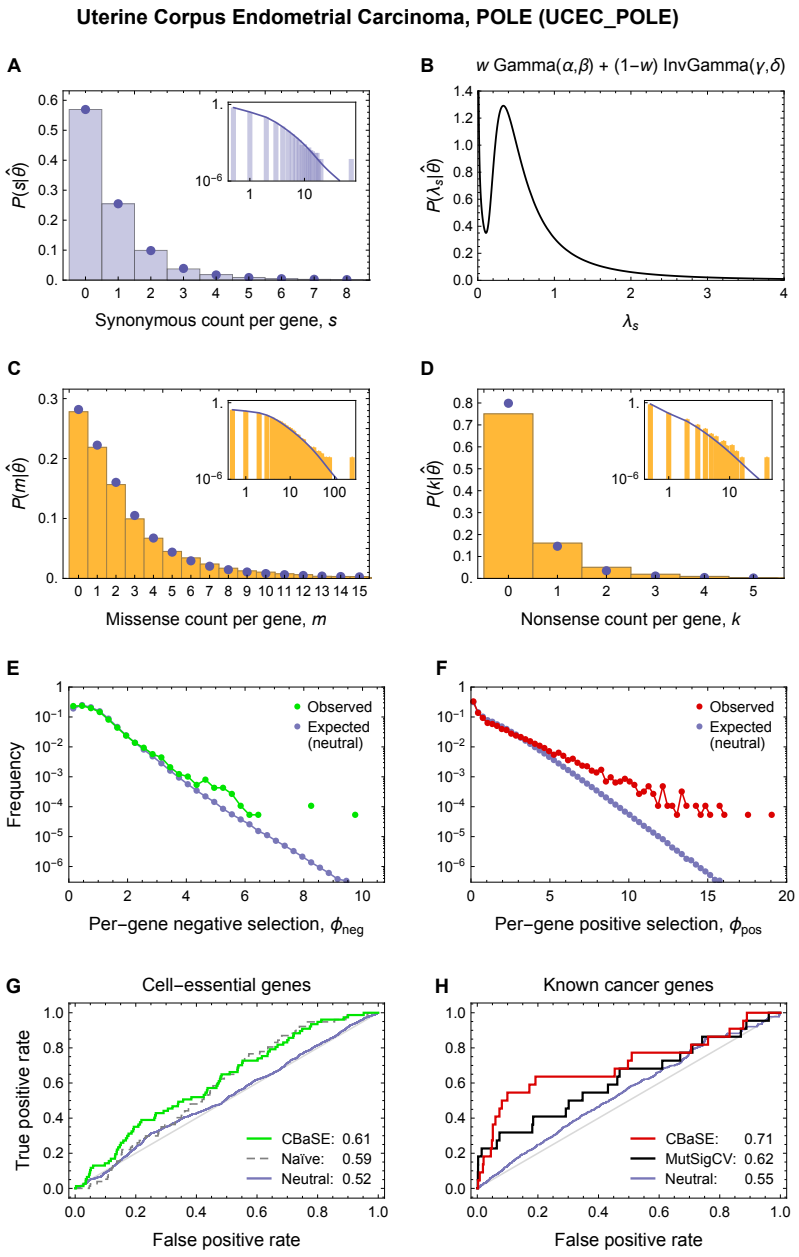
Supplementary Figure 23: UCEC (without MMR or POLE subtypes). (A),(C),(D) Observed distributions of mutation counts per gene (histograms), and expected distributions under neutral evolution (blue dots and blue lines) with underlying $P(\lambda_s;\boldsymbol{\theta}) = w\,\mathrm{Gamma}(\lambda_s;\alpha,\beta)+(1-w)\,\mathrm{InvGamma}(\lambda_s;\gamma,\delta)$, shown in (B). Parameter estimates $\hat{\boldsymbol{\theta}}$ are $\hat{\alpha}=0.63$, $\hat{\beta}=1.58$, $\hat{\gamma}=2.75$, $\hat{\delta}=2.12$, $\hat{w}=0.35$; sum of squared deviations in (A) is 5e-6. (E),(F) Histogram of the meta-statistic ϕ in real data (green, red) and from simulation under the neutral model (blue), bin width 0.3. Genes (ϕ_{pos}) not shown in (F): *PIK3CA* (27.4), *TP53* (27.7), *CTCF* (30.5), *ARID1A* (37.8), *PTEN* (55.8). (G) ROC of 77 predicted cell-essential genes from Wang et al.[2] for genes sorted by q_{neg} (green), $p_{\mathrm{AUC}}=3\text{e-}4$. Dashed gray line shows the ROC from ranking by hypomutation ($p_{\mathrm{AUC}}>0.05$). (H) ROC of the gene ranking based on q_{pos}, using as true positives 24 CGC genes causally implicated in UCEC (red), $p_{\mathrm{AUC}}=3\text{e-}4$. Black line shows the ROC for MutSigCV[1] predictions. Blue lines in (G) and (H) show neutral ROC expectations from simulation. AUCs are given in the insets.

Nature Genetics: doi:10.1038/ng.3987



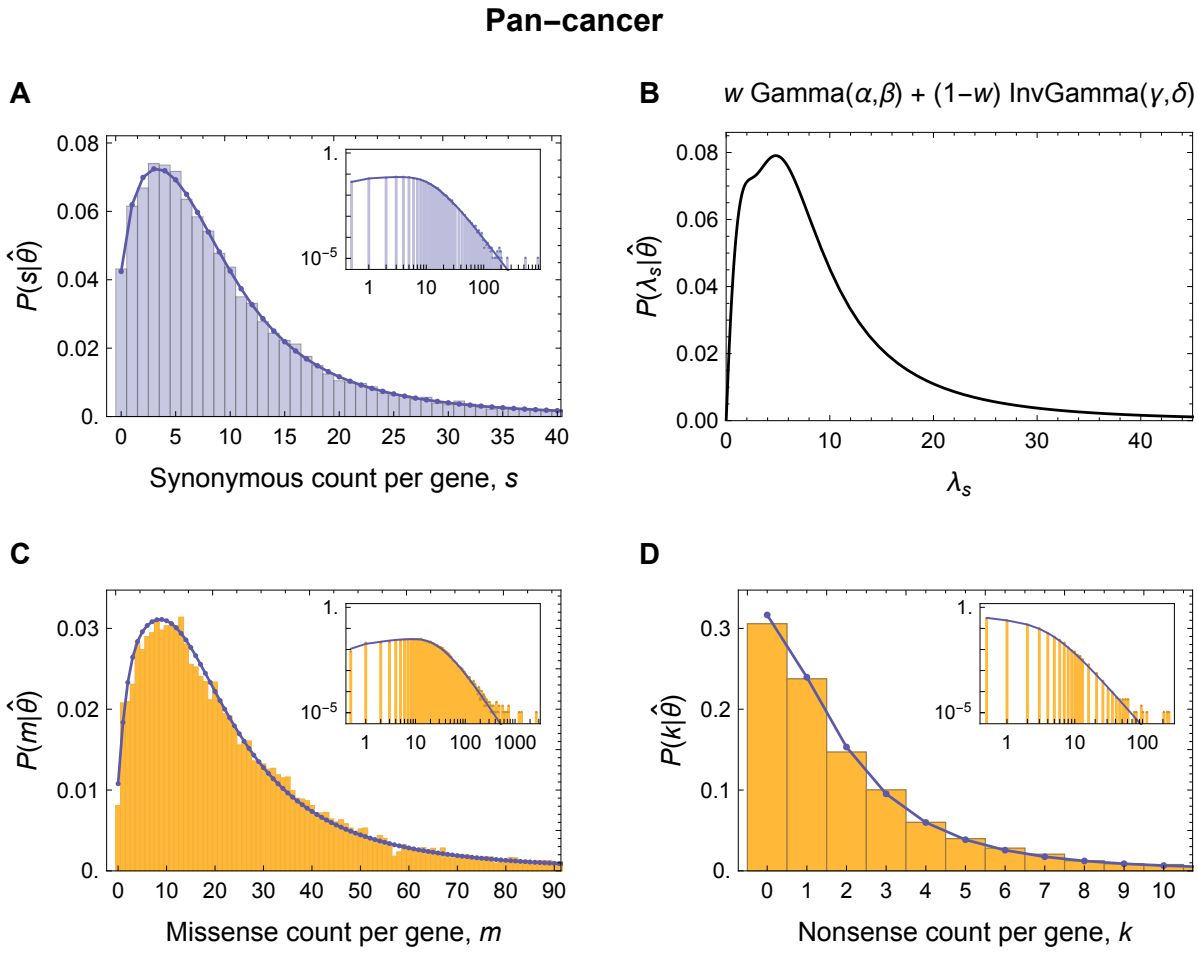
Supplementary Figure 24: UCEC (mismatch repair deficient). (A),(C),(D) Observed distributions of mutation counts per gene (histograms), and expected distributions under neutral evolution (blue dots and blue lines) with underlying $P(\lambda_s;\boldsymbol{\theta}) = w\,\mathrm{Exp}(\lambda_s;\tau)+(1-w)\,\mathrm{InvGamma}(\lambda_s;\alpha,\beta)$, shown in (B). Parameter estimates $\hat{\boldsymbol{\theta}}$ are $\hat{\alpha} = 2.53$, $\hat{\beta} = 0.68$, $\hat{\tau} = 1.73$, $\hat{w} = 0.50$; sum of squared deviations in (A) is 1e-6. (E),(F) Histogram of the meta-statistic ϕ in real data (green, red) and from simulation under the neutral model (blue), bin width 0.3. Genes (ϕ_{pos}) not shown in (F): *PIK3CA* (23.7), *ARID1A* (34.6), *PTEN* ($p_k^{\mathrm{pos}} = 0$). (G) ROC of 77 predicted cell-essential genes from Wang et al.[2] for genes sorted by q_{neg} (green), $p_{\mathrm{AUC}} = 3\text{e-}3$. Dashed gray line shows the ROC from ranking by hypomutation ($p_{\mathrm{AUC}} > 0.05$). (H) ROC of the gene ranking based on q_{pos}, using as true positives 24 CGC genes causally implicated in UCEC (red), $p_{\mathrm{AUC}} = 3\text{e-}4$. Black line shows the ROC for MutSigCV[1] predictions. Blue lines in (G) and (H) show neutral ROC expectations from simulation. AUCs are given in the insets.

Nature Genetics: doi:10.1038/ng.3987



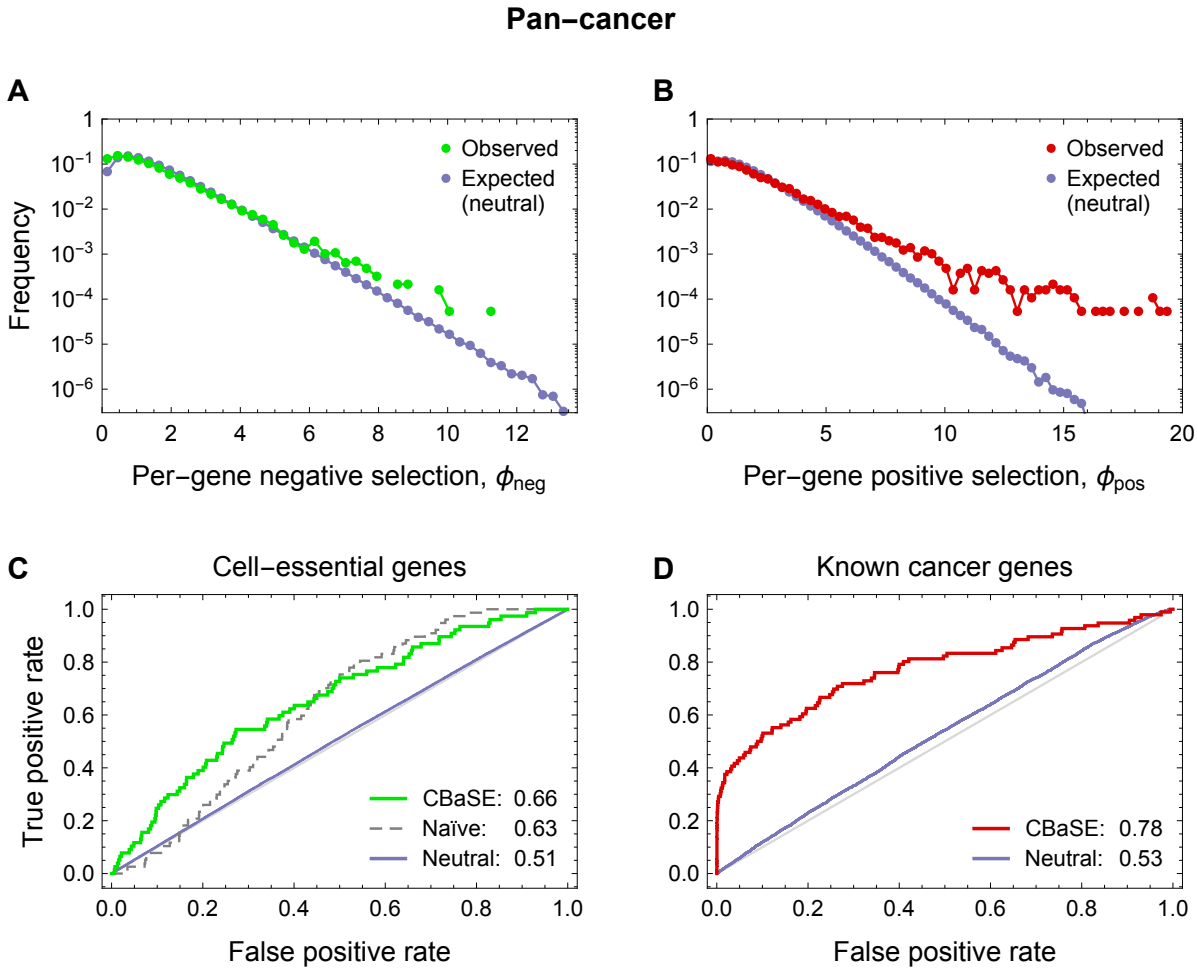
Supplementary Figure 25: UCEC (polymerase ϵ deficient). (A),(C),(D) Observed distributions of mutation counts per gene (histograms), and expected distributions under neutral evolution (blue dots and blue lines) with underlying $P(\lambda_s;\boldsymbol{\theta}) = w\,\mathrm{Gamma}(\lambda_s;\alpha,\beta) + (1-w)\,\mathrm{InvGamma}(\lambda_s;\gamma,\delta)$, shown in (B). Parameter estimates $\hat{\boldsymbol{\theta}}$ are $\hat{\alpha} = 0.39$, $\hat{\beta} = 2.14$, $\hat{\gamma} = 2.62$, $\hat{\delta} = 1.24$, $\hat{w} = 0.27$; sum of squared deviations in (A) is 4e-6. (E),(F) Histogram of the meta-statistic ϕ in real data (green, red) and from simulation under the neutral model (blue), bin width 0.3. Genes (ϕ_{pos}) not shown in (F): *CCDC147* (23.8), *ARID1A* (26.5). (G) ROC of 77 predicted cell-essential genes from Wang et al.[2] for genes sorted by q_{neg} (green), $p_{\mathrm{AUC}} = 7\text{e-}4$. Dashed gray line shows the ROC from ranking by hypomutation ($p_{\mathrm{AUC}} = 5\text{e-}3$). (H) ROC of the gene ranking based on q_{pos}, using as true positives 24 CGC genes causally implicated in UCEC (red), $p_{\mathrm{AUC}} = 7\text{e-}3$. Black line shows the ROC for MutSigCV[1] predictions. Blue lines in (G) and (H) show neutral ROC expectations from simulation. AUCs are given in the insets.

Nature Genetics: doi:10.1038/ng.3987



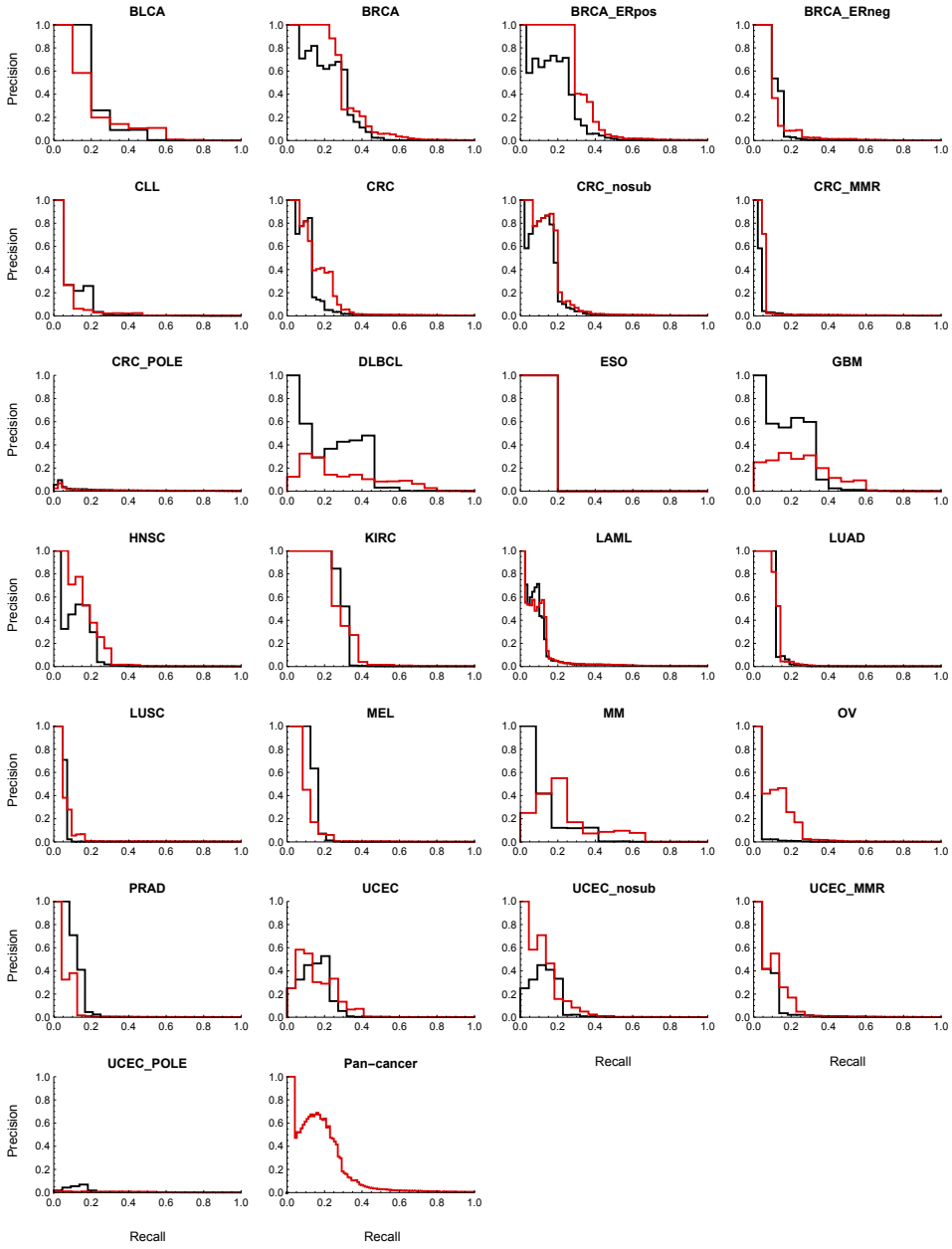
Supplementary Figure 26: Pan-cancer analysis. (A) Observed distribution of synonymous counts per gene summed over all cancer types (histogram), and fitted model distribution (blue dots joined by blue line) with underlying $P(\lambda_s;\boldsymbol{\theta}) = w\,\text{Gamma}(\lambda_s;\alpha,\beta)+(1-w)\,\text{InvGamma}(\lambda_s;\gamma,\delta)$, shown in (B). Parameter estimates $\hat{\boldsymbol{\theta}}$ are $\hat{\alpha} = 2.08$, $\hat{\beta} = 1.91$, $\hat{\gamma} = 2.68$, $\hat{\delta} = 25.51$, $\hat{w} = 0.38$; sum of squared deviations in (A) is 4e-5. (C),(D) Observation (histograms) and expectation under neutral evolution (blue dots joined by blue line) for the distribution of the per-gene number of missense and nonsense mutations, respectively, summed over cancer types. Large panels show the distributions on a linear scale, while the insets show the same distributions on a log-log-scale to visualize the large-count regime (blue line showing the neutral expectation). While the two genes with the highest counts in the synonymous category ($s > 300$; *MUC16*, *TTN*) are captured by the neutral model, the nonsynonymous categories enrich for known cancer genes in the large-count regime: genes with $m > 1000$ are *TP53*, *MUC16* and *TTN*, while genes with $k > 100$ are *ARID1A*, *APC*, *TTN*, and *TP53*. Since the selection signal depends on the relative size of s and x, neither *MUC16* ($q_{\text{pos}} = 1$) nor *TTN* ($q_{\text{pos}} = 1$) are significantly mutated (see **Supplementary Table 2**).

Nature Genetics: doi:10.1038/ng.3987



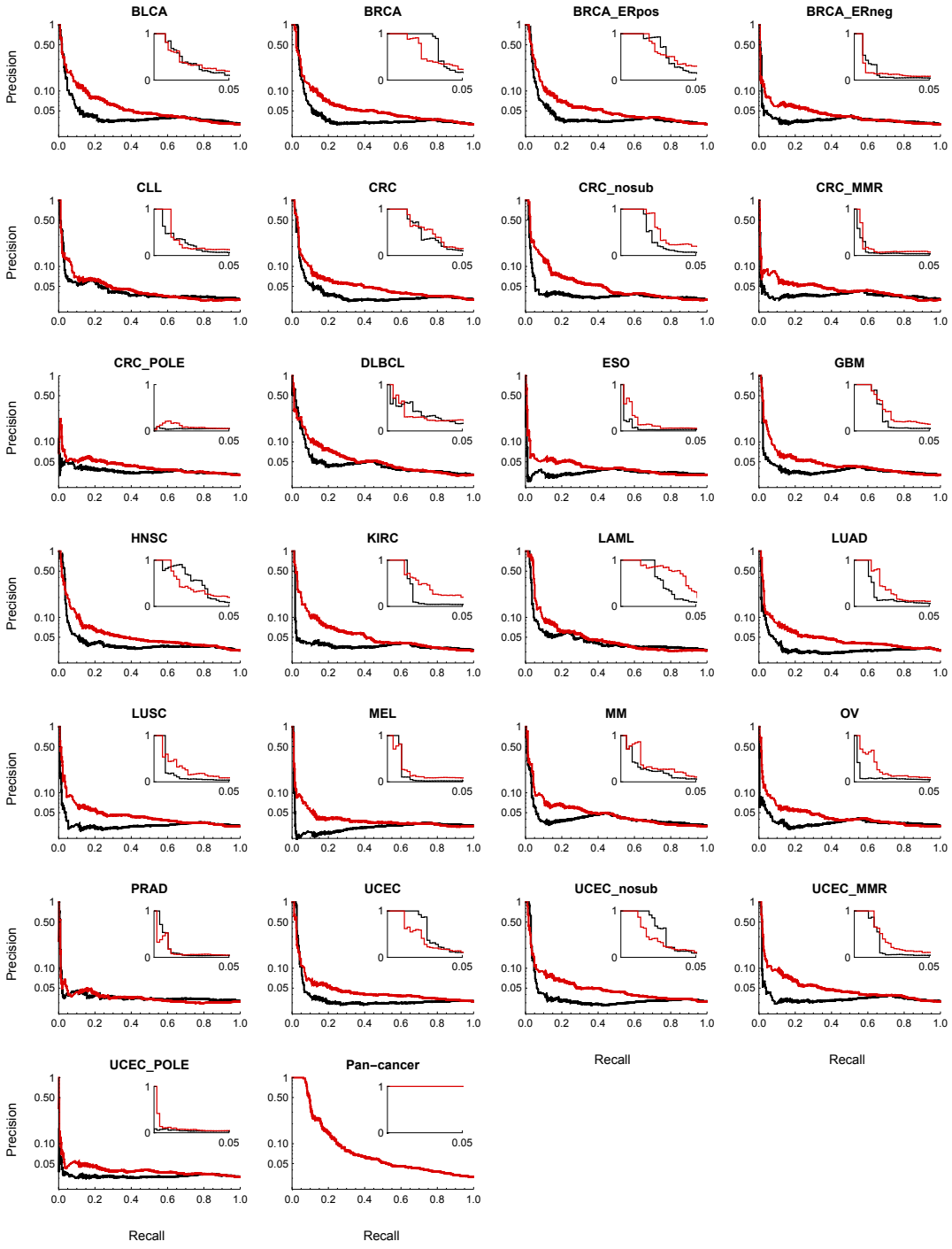
Supplementary Figure 27: Pan-cancer analysis. (A) Histogram of the meta-statistic for negative selection, ϕ_{neg}, in real data (green) and from simulation under the neutral model (blue). (B) Same as (A) for ϕ_{pos} (real data in red); both distributions shown for bins of width 0.3. Genes (ϕ_{pos}) not shown in (B): *SMARCA4* (21.0), *SMAD4* (22.1), *HLA-A* (22.4), *NFE2L2* (23.7), *MAP3K1* (25.3), *BAP1* (26.5), *BRAF* (26.7), *CTCF* (28.5), *ARID2* (29.3), *NOTCH1* (30.3), *NRAS* (30.6), *PIK3R1* (31.8), *CASP8* (32.7), *FBXW7* (33.0), *STK11* (33.5), *KEAP1* (35.9), *ARID1A* (36.9), *SETD2* (38.0), *ARHGAP35* (38.8), *PBRM1* (39.4), *PIK3CA* ($p_m^{\text{pos}} = 0$), *KRAS* ($p_m^{\text{pos}} = 0$), *KMT2D* ($p_k^{\text{pos}} = 0$), *APC* ($p_k^{\text{pos}} = 0$), *NF1* ($p_k^{\text{pos}} = 0$), *RB1* ($p_k^{\text{pos}} = 0$), *FAT1* ($p_k^{\text{pos}} = 0$), *VHL* ($p_k^{\text{pos}} = 0$), *TP53* ($p_m^{\text{pos}} = 0$), *PTEN* ($p_k^{\text{pos}} = 0$). (C) ROC of 77 predicted cell-essential genes from Wang et al.[2] for genes sorted by q_{neg} (green), $p_{\text{AUC}} = 5\text{e-}6$. Dashed gray line shows the ROC from ranking by hypomutation in all mutational categories ($p_{\text{AUC}} = 5\text{e-}5$). (D) ROC of the gene ranking based on significance level q_{pos}, using as true positives 100 genes described by the CGC as causally implicated in at least two of the 17 cancer types, $p_{\text{AUC}} = 0$. MutSigCV results for the pan-cancer data set could not be obtained, hence the comparison cannot be shown. Blue lines in (C) and (D) show neutral ROC expectations from simulation. AUCs are given in the insets.

Nature Genetics: doi:10.1038/ng.3987



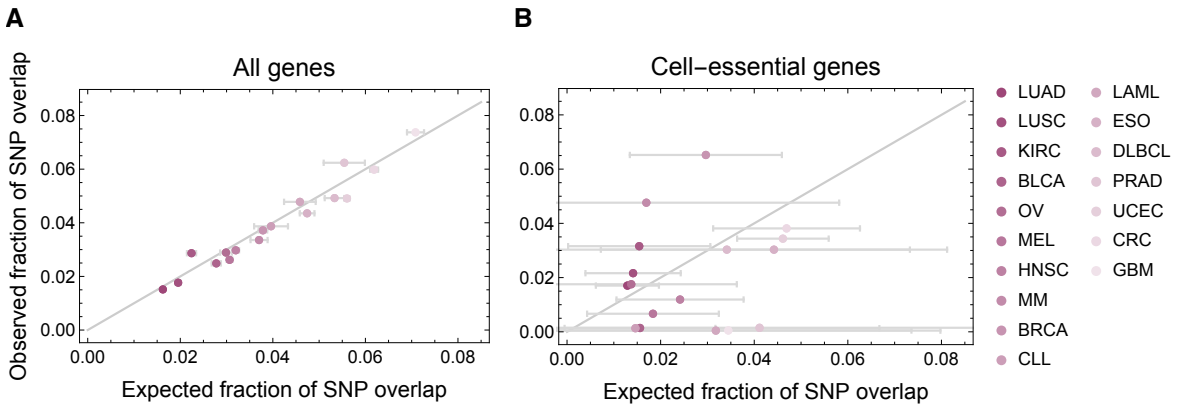
Supplementary Figure 28: Precision-recall (PR) curves using cancer type-specific driver genes. Shown are PR curves for all cancer types and the pan-cancer data set using as true positives the same cancer type-specific known driver genes used in the ROC curves from **Figure 2** and **Supplementary Figures 2-25,27**; red: CBaSE, black: MutSigCV. Depending on the distributions of output values (i.e. q-values for true driver genes and their complement, respectively), PR curves can exhibit strong deviations from the theoretical expectation.[3] Because of this and the absence of ground truth information on weak driver genes that may be operative in only a subset of tumors from a given cancer type, we also derived PR curves using the whole CGC gene set as true positives. The corresponding results are shown in **Supplementary Figure 29**.

Nature Genetics: doi:10.1038/ng.3987



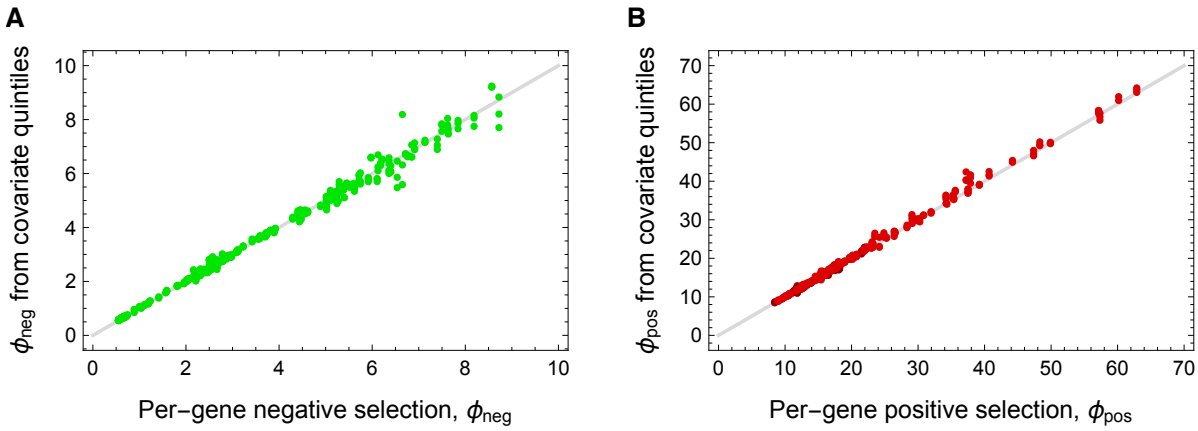
Supplementary Figure 29: Precision-recall curves using all CGC driver genes. Shown are PR curves for all cancer types and the pan-cancer data set using as true positives all 575 cancer-associated genes listed in the CGC (v80); red: CBaSE, black: MutSigCV. Note the logarithmic y-axis in the large plots (linear y-axis shown in inset). Effects on PR curves due to absence of ground truth information on all relevant driver genes in a given cancer type are expected to be alleviated relative to the more narrowly defined gene sets used in **Supplementary Figure 28**.

Nature Genetics: doi:10.1038/ng.3987



Supplementary Figure 30: Overlap with germline polymorphisms. Observed vs. random expected fraction of somatic mutation overlap with 875,804 known germline SNPs from the dbSNP database for all 17 original cancer types. Overlap was computed by trinucleotide context to account for cancer-type specific mutational context dependence, and assuming equipartition of SNPs (see **Supplementary Note**). Error bars denote s.e.m. assuming binomial sampling. (A) Overlap on all genes, (B) on 77 predicted cell-essential genes.[2] The agreement between expected and observed overlap of somatic mutations with SNPs suggests adequate filtering of germline variants from the somatic mutation data set.[4]

Nature Genetics: doi:10.1038/ng.3987



Supplementary Figure 31: Effects of heterogeneity of mutation signature. Comparison of the per-gene meta-statistic ϕ to the corresponding values obtained under separate inference of the trinucleotide context-dependent mutation matrix in gene quintiles of factors which can affect the mutation signature. We repeated the analysis for three different factors: (1) replication time, (2) expression level, and (3) chromatin state. Shown are results for the ten top-ranking genes, pooled across cancer types and factors. (A) Negative selection, ϕ_{neg}; (B) positive selection, ϕ_{pos}, bright red: genes listed in the CGC. Genes with $p_k^{\mathrm{pos}} = 0$ or $p_m^{\mathrm{pos}} = 0$ in any of the analyses were omitted, as the corresponding values $\phi_{\mathrm{pos}} \to \infty$ cannot be directly compared.

Nature Genetics: doi:10.1038/ng.3987

Supplementary Note

Contents

1 Data

1.1 Cancer data set

The Tumor Portal data set[4] (tumorportal.org) we used has the following 21 different cancer types and corresponding numbers of patients / total numbers of exonic mutations in brackets: Urothelial Bladder Carcinoma (BLCA; 99 / 29,835), Breast Invasive Carcinoma (BRCA; 890 / 45,738), Carcinoid (CARC; 54 / 1,583), Chronic Lymphocytic Leukemia (CLL; 159 / 3,021), Colorectal Cancer (CRC; 233 / 80,533), Diffuse Large B-cell Lymphoma (DLBCL; 58 / 12,154), Esophageal Cancer (ESO; 141 / 20,936), Glioblastoma Multiforme (GBM; 291 / 21,230), Head-Neck Squamous Cell Carcinoma (HNSC; 384 / 64,948), Kidney Renal Clear Cell Carcinoma (KIRC; 417 / 24,171), Acute Myeloid Leukemia (LAML; 196 / 4,089), Lung Adenocarcinoma (LUAD; 404 / 152,185), Lung Squamous Cell Carcinoma (LUSC; 177 / 66,999), Medulloblastoma (MED; 92 / 1,169), Melanoma (MEL; 118 / 85,929), Multiple Myeloma (MM; 207 / 12,976), Neuroblastoma (NB; 76 / 1,600), Ovarian Cancer (OV; 316 / 18,581), Prostate Adenocarcinoma (PRAD; 138 / 3,228), Rhabdoid Tumor (RHAB; 35 / 284), and Uterine Corpus Endometrial Carcinoma (UCEC; 248 / 210,278). We excluded cancer types with less than 2,000 exonic mutations outside likely selected genes, as these are used for the derivation of the type-specific mutational signatures (CARC, MED, NB and RHAB), leaving 17 used cancer types. The set of likely selected genes was set to comprise all 575 known cancer genes from the CGC[5] (v80) and 77 likely cell-essential genes from Wang *et al.* (2015)[2] (**Supplementary Table 13**).

1.2 Germline polymorphism

The Tumor Portal data set is based on tumor-normal subtraction of mutations and is additionally filtered for germline variation through removal of variants that occurred in a panel of normals comprising over 4,000 BAM files.[4] To further address the possibility of germline contamination we derived the expected fraction of mutations overlapping by chance with known single-nucleotide polymorphisms (SNPs), and compared it to the observed fraction. To this end, all validated human non-somatic coding SNPs submitted by the ExAC Consortium and with annotated minor allele frequency were downloaded from the dbSNP data base (ftp://ftp.ncbi.nih.gov/snp/organisms/human_9606_b150_GRCh37p13/ASN1_flat), resulting in 875,804 SNPs in total. We then computed the mean SNP density on coding regions separately for each of the 64 trinucleotide contexts, assuming a total coding sequence length of $\sim$ 33.28 Mbp (from 19,977 genes on 22 autosomes and sex chromosomes) and mean whole-exome trinucleotide occurrences. For a given context and mutated nucleotide, the expected number of somatic mutations overlapping SNPs by chance was estimated by multiplication of the mean SNP density with the total observed number of corresponding somatic mutations in the context. **Supplementary Figure 30A** shows the comparison of the expected and observed

fraction of somatic mutations overlapping with ExAC SNPs for each cancer type, and we find a high correspondence. The SNP-overlapping fraction of somatic mutations varies between 1.5 and 7.4% across cancer types, consistent with the observations from a recent study.[6] When we repeat the comparison for SNPs and somatic mutations within the constrained subset of the 77 cell-essential genes (using the same whole-exome trinucleotide occurrences), we also find good agreement between the expected and observed fraction within error bars (**Supplementary Figure 30B**). The reason why germline contamination is important for selection inference is that SNPs are affected by selection pressures at the human population level, where they evolve under largely negative selection. The resulting reduction in nonsynonymous (pN) compared to synonymous (pS) polymorphisms could contribute to the signal of negative selection if unaccounted for, or cause spurious detection of positive selection if overcorrected. The agreement between expected and observed overlap suggests that inflation (or deflation) of synonymous mutation counts relative to the nonsynonymous classes due to contamination with germline SNPs is negligible, both across the genome as well as on the cell-essential subset of genes. We carried out additional analyses of the effect of SNPs on the downstream results, which are described further below.

1.3 Cancer subtypes

Cancer subtypes were identified by their mutational signature, as defined in the COSMIC database (cancer.sanger.ac.uk/cosmic/signatures). We decomposed each sample into relative contributions from all 30 signatures using the algorithm deconstructSigs.[7] Colorectal cancer (CRC) and uterine corpus endometrial carcinoma (UCEC) were subdivided into tumors with DNA mismatch repair deficiency (MMR; sum of fractions of signatures 6, 15, 20, 26 > 0.5), those with altered function of polymerase ϵ (POLE; fraction of signature 10 > 0.5) and the remainder (nosub). There are 21 MMR samples in CRC, 64 in UCEC, while 7 CRC and 12 UCEC samples are identified as having aberrant POLE. Breast cancer (BRCA) was stratified by endocrine receptor (ER) status according to preponderance of either signature 1 (ER-positive; ERpos) or signature 3 (ER-negative; ERneg), based on associations between hormonal subtype and mutational signatures described in Nik-Zainal *et al.* (2016).[8] 649 (73%) of breast cancer samples were thus classified as ER-positive, and 241 (27%) as ER-negative. Consequently, in addition to the 17 full cancer types we consider 8 cancer subtypes. **Supplementary Table 11** shows all sample IDs associated with the different cancer subtypes.

1.4 Mutation annotation

The curated somatic single-nucleotide variant calls are a mixture of whole-genome and whole-exome tumor-normal sequencing data.[4] To ensure homogeneity of annotation, we re-annotated all called mutations using the PolyPhen-2 annotator mapSNPs, which annotates functional categories missense, nonsense, synonymous, intron, 3'-UTR and 5'-UTR. The settings were `-n`

`-x1`, corresponding to skipping of alleles without a matching reference nucleotide and including only knownCanonical transcripts out of all UCSC knownGene transcripts. All mutations that were not annotated by mapSNPs were assigned intergenic status. This resulted in a total of 3,104,600 called mutations: 542,234 missense, 46,240 nonsense (stop-gain and stop-loss), 208,822 synonymous, 830,126 intronic, 47,282 3'-UTR, 16,889 5'-UTR, and 1,413,007 intergenic. Only exonic mutations (including UTRs) were used in the downstream analyses.

1.5 Gene set

We considered all genes that have at least one coding mutation (missense m, nonsense k, synonymous s) in any of the 21 original cancer types (19,098) and which have a peptide sequence in the UCSC knownGene track that matches the DNA sequence of the UCSC reference genome, leaving 19,048 genes. Genes from the olfactory group anomalously cluster in length (around mean $L \approx 945$ bp), causing a discontinuity in the distribution of gene lengths. Because gene length is the main factor affecting the Poisson parameter we hence excluded the set of 382 olfactory receptor genes from the likelihood fit (**Supplementary Table 12**). This leaves $G =$ 18,666 genes used in the inference.

1.6 Gene meta data

Clusters of missense mutations were determined with mutation3d[9] (mutation3d.org) with preset cluster criteria. For each gene that had cluster results we used all unique clusters with p-values < 0.1, i.e. no mutation was counted in more than one cluster. We always used the protein model with the highest ModPipe Quality Score. We then matched the positions of the clustered mutations with protein domain positions obtained from Pfam (pfam.xfam.org) and computed for each unique cluster the fraction of mutations that lie within a single domain.

Gene expression levels in tumor and normal tissues were obtained from the TCGA Expression Browser (tools.stamlab.org/tcga) for the 22 cancer types BLCA, BRCA, CESC, CHOL, COAD, GBM, HNSC, KICH, KIRC, KIRP, LIHC, LUAD, LUSC, PAAD, PCPG, PRAD, READ, SARC, SKCM, THCA, THYM, and UCEC. A minimum threshold of 200 transcripts per million on average was set for both tumor and normal expression levels. Significant over-expression in a cancer tissue was reported for types with $y_\mathrm{t} - 1.57\,IQR/\sqrt{n_\mathrm{t}} > y_\mathrm{n} + 1.57\,IQR/\sqrt{n_\mathrm{n}}$, where y_t and y_n denote the median of the gene transcripts per million in the tumor and normal samples, respectively, n_t and n_n are the corresponding sample sizes, and the interquartile range is approximated by that of a normal distribution, i.e. $IQR = 1.35\,\sigma$.[10] An analogous expression follows for genes with significantly decreased expression. Constitutive expression in normal tissues for pan-cancer negatively selected genes was defined as a median of at least 200 transcripts per million across all 22 tissue types.

Nature Genetics: doi:10.1038/ng.3987

2 Probabilistic model for mutation counts under neutral evolution

2.1 Expected number of neutral mutations per gene

For a neutrally evolving nucleotide site, standard evolutionary theory predicts that the per-generation substitution rate exactly equals the mutation rate.[11] A gene with L mutable neutral sites then has an expected value $\langle x \rangle = T \sum_{i=1}^{L} \mu_i^x$ of the observed number of neutral substitutions after T generations, allowing for each site to have its own mutation rate μ_i^x. This result is independent of assumptions about genetic linkage and thus also holds in case of full linkage, like for a population of cancer cells. From a stochastic point of view, nucleotide substitutions are rare events, which lets us write the conditional probability of observing x_j neutral substitutions on a given gene in a tumor j as

$$x_j \,|\, \vec{\mu}_j^x, L, T_j \sim \text{Pois}\left(\lambda_j\right) \,, \tag{1}$$

where $\vec{\mu}_j^x = (\mu_{1j}^x, ..., \mu_{Lj}^x)$ is the vector of site mutation rates and the Poisson parameter

$$\lambda_j = T_j \sum_{i=1}^{L} \mu_{ij}^x \tag{2}$$

is just the expected number of substitutions above. We can extend the summation over all sites within the gene, Eq. 1, to the sum over all t tumors of the given cancer type, utilizing that the sum over Poisson random variables is again Poisson distributed:

$$\left(\sum_{j=1}^{t} x_j\right) \,|\, \{\vec{\mu}_j^x, L, T_j\}_t \sim \text{Pois}\left(\sum_{j=1}^{t} \lambda_j\right) \equiv \text{Pois}\left(\lambda_x\right) \,, \tag{3}$$

where in the last step we defined the expected number of neutral substitutions on the gene, λ_x.

2.2 Estimation of the distribution $P(\lambda_s|\boldsymbol{\theta})$

Assuming synonymous mutations to evolve neutrally, we aim to estimate the parameters $\boldsymbol{\theta}$ of the distribution of the expected number of synonymous mutations across the ensemble of genes, $P(\lambda_s|\boldsymbol{\theta})$. This distribution is designed to implicitly capture all factors influencing mutation rate variability across the genome and to fit the data exactly. Importantly, no explicit assumptions have to be made about mutation rate covariates, or information about them provided. We fit the observed distribution of synonymous counts s with the hierarchical model shown in Eq. 1 of the main text, allowing $P(\lambda_s|\boldsymbol{\theta})$ to assume one of the following parametric forms:

1. $\text{Gamma}(\lambda_s; \boldsymbol{\theta} = \{\alpha, \beta\}) = \frac{1}{\Gamma(\alpha)\beta^\alpha} \mathrm{e}^{-\lambda_s/\beta} \lambda_s^{\alpha-1}$

2. $\text{InvGamma}(\lambda_s; \boldsymbol{\theta} = \{\alpha, \beta\}) = \frac{\beta^\alpha}{\Gamma(\alpha)} \mathrm{e}^{-\beta/\lambda_s} \lambda_s^{-\alpha-1}$

Nature Genetics: doi:10.1038/ng.3987

3. $P_{\mathrm{E,G}}(\lambda_s;\boldsymbol{\theta}=\{\alpha,\beta,\tau,w\}) = w\,\mathrm{Exp}(\lambda_s;\tau) + (1-w)\,\mathrm{Gamma}(\lambda_s;\alpha,\beta)$

4. $P_{\mathrm{E,IG}}(\lambda_s;\boldsymbol{\theta}=\{\alpha,\beta,\tau,w\}) = w\,\mathrm{Exp}(\lambda_s;\tau) + (1-w)\,\mathrm{InvGamma}(\lambda_s;\alpha,\beta)$

5. $P_{\mathrm{G,G}}(\lambda_s;\boldsymbol{\theta}=\{\alpha,\beta,\gamma,\delta,w\}) = w\,\mathrm{Gamma}(\lambda_s;\alpha,\beta) + (1-w)\,\mathrm{Gamma}(\lambda_s;\gamma,\delta)$

6. $P_{\mathrm{G,IG}}(\lambda_s;\boldsymbol{\theta}=\{\alpha,\beta,\gamma,\delta,w\}) = w\,\mathrm{Gamma}(\lambda_s;\alpha,\beta) + (1-w)\,\mathrm{InvGamma}(\lambda_s;\gamma,\delta)$,

where models 5 and 6, respectively, represent generalizations of models 3 and 4 to account for more data complexity, as the exponential distribution is a special case of the gamma distribution (both addressing potential inflation at small λ_s). Consideration of a range of analytical models can also factor in generation time and per-gene mutation rate heterogeneity across tumors of a given cancer type.
Parameters $\boldsymbol{\theta}$ are estimated from maximizing the log-likelihood function

$$\mathcal{L} = \sum_{g=1}^{G} \log\left[P(s_{\mathrm{obs},g}|\boldsymbol{\theta})\right] , \tag{4}$$

where the sum is over all G genes with their respective observed synonymous mutation counts $s_{\mathrm{obs},g}$. Model selection uses penalization according to the Akaike information criterion, accounting for the number of fitted parameters. We find that all cancer (sub)types are best fitted with an inverse gamma component (models 2, 4, and 6), which can capture the heavy tail observed in the genome-wide distribution of synonymous mutation counts. The dimensionality of the best fit is naturally determined by the total mutation count availability, resulting in four cancer (sub)types to be best described by model 2, thirteen by model 4 and the remaining eight cancer (sub)types are best fit in combination with a gamma component, model 6 (**Figure 1** and **Supplementary Figures 2-26**).

2.3 Neutral expectation for the distribution of nonsynonymous mutation counts

The estimated distribution of Poisson parameters, $P(\lambda_s|\hat{\boldsymbol{\theta}})$, underlies the distribution of synonymous mutations, $P(s|\hat{\boldsymbol{\theta}})$ (Eq. 1 in main text). By introducing the ratio $r_x = \ell_x/\ell_s$ of the target size of mutation category x and the synonymous target size on a given gene, we can write a more general definition of the neutral model distribution, encompassing nonsynonymous mutations:

$$\begin{aligned} P(x|\hat{\boldsymbol{\theta}}; r_x) &= \int \mathrm{d}\lambda_s\, P(x|\lambda_s; r_x)\, P(\lambda_s|\hat{\boldsymbol{\theta}}) \\ &= \int \mathrm{d}\lambda_s\, \mathrm{Pois}(\lambda_s r_x)\, P(\lambda_s|\hat{\boldsymbol{\theta}})\,. \end{aligned} \tag{5}$$

Nature Genetics: doi:10.1038/ng.3987

The target size of a mutation category x (missense m, nonsense k, or synonymous s) depends on the gene sequence and the mutational signature of the cancer type. It is obtained by summing the transition probabilities to a given category over all L coding base pairs i:

$$\ell_x = \sum_{i=1}^{L} \sum_{Y \in \{\mathrm{A,C,G,T}\}} p_0(X_{i-1}X_iX_{i+1} \to X_{i-1}YX_{i+1})\, \delta_{xv_Y}\,, \tag{6}$$

where δ denotes the Kronecker delta. $p_0(X_{i-1}X_iX_{i+1} \to X_{i-1}YX_{i+1})$ is the cancer type-specific trinucleotide context-dependent mutation probability of nucleotide X_i (embedded in context $X_{i-1}.X_{i+1}$) to nucleotide Y, derived from all observed changes at exonic sequence sites outside of genes likely under selection (defined above), where $p_0 = 0$ for $Y = X_i$. We sum over all possible final states of the mutation event, where the indicator variable $v_Y = x$ for an amino acid change of type x and $v_Y \neq x$ otherwise.

With this, one can express the sum over the site-specific mutation rates μ_{ij}^x to category x (for tumor j) as a product of the target size ℓ_x and the background mutation rate at the gene locus, μ_j^{bg}. Then the expected number of mutations in category x under neutrality (cf. Eq. 2) summed over tumors j becomes

$$\lambda_x = \sum_{j=1}^{t} T_j \sum_{i=1}^{L} \mu_{ij}^x = \ell_x \sum_{j=1}^{t} T_j\, \mu_j^{\mathrm{bg}}\,. \tag{7}$$

With this we can transform from the synonymous to the nonsynonymous categories without the need for further parameters, only via rescaling λ_s with the ratio of target sizes, as done in Eq. 5. In order to compare the neutral prediction of Eq. 5 to the corresponding observed nonsynonymous distributions across all genes in a given cancer type, we compute the superposition of all G probability densities from the gene set:

$$P(x|\hat{\boldsymbol{\theta}}) \equiv \frac{1}{G} \sum_{g=1}^{G} P(x|\hat{\boldsymbol{\theta}}; r_{x,g})\,, \qquad x \in \{m, k\}\,. \tag{8}$$

2.4 Nucleotide context-dependent mutation probability

We compute the expected number of mutations of a given category (m, k, s) on a gene explicitly taking into account the cancer type-specific mutational signature. Exonic mutations outside of the defined set of 575 known cancer genes and 77 likely cell-essential genes are used to construct the trinucleotide context-dependent 64×4 neutral mutation matrix with elements

$$p_0^{kl} = \frac{M_{kl}}{\omega_k \sum_{m,n} M_{mn}}\,, \tag{9}$$

which corresponds to $p_0(X_{i-1}X_iX_{i+1} \to X_{i-1}YX_{i+1})$. Here $k = 1, ..., 64$ indexes the trinucleotide context $X_{i-1}X_iX_{i+1}$ of the mutated base pair and $l = 1, ..., 4$ the target nucleotide Y.

M_{kl} represents the number of observed mutations that occurred in context k to nucleotide l, and $\sum_{m,n} M_{mn} = \text{const.}$ is the total number of mutations. We need to normalize by the probability of occurrence of any given context, ω_k, which is derived from a whole-exome count. While it is conceivable to compute p_0^{kl} separately for each tumor in the case of highly mutable cancer types, we here approximate the tumor-specific matrices with the result from pooling mutations across all tumors of a given type. Since codons consist of three nucleotides, accounting for trinucleotide context of mutations is usually sufficient. However, it has been found that the UV mutational signature has a pentameric context dependence component through the possibility of duplet and triplet mutations.[12] Since this mutational process is the primary contributor to melanoma, we compute the neutral mutation matrix for melanoma and the pan-cancer data set based on pentanucleotides instead of trinucleotides.

3 Bayesian inference of selection at the gene level

3.1 Posterior distribution of nonsynonymous counts

Apart from a quantification of the genome-wide effects of selection using Eq. 8, we are interested in an estimate of negative and positive selective pressure at the level of individual genes. We can write the per-gene posterior probability for the expected number of nonsynonymous mutations in category x, given the observed synonymous count s_{obs}, as:

$$\begin{aligned} P(x|s_{\text{obs}}; r_x) &= \int \mathrm{d}\lambda_s \, P(x, \lambda_s|s_{\text{obs}}; r_x) \\ &= \int \mathrm{d}\lambda_s \, P(x|\lambda_s; r_x) P(\lambda_s|s_{\text{obs}}) \,, \end{aligned} \tag{10}$$

where we made use of the Poisson counts s and $x \in \{m, k\}$ being independent given λ_s. The first term in this compound distribution is, as before:

$$P(x|\lambda_s; r_x) = \text{Pois}(\lambda_s r_x) \,, \tag{11}$$

and Bayes' theorem gives

$$P(\lambda_s|s_{\text{obs}}) = \frac{P(s_{\text{obs}}|\lambda_s) P(\lambda_s|\hat{\boldsymbol{\theta}})}{P(s_{\text{obs}}|\hat{\boldsymbol{\theta}})} \,. \tag{12}$$

From the posterior distribution, Eq. 10, we compute p-values for the per-gene observed numbers of missense and nonsense mutations, m_{obs} and k_{obs}, given the observed synonymous count s_{obs}. Paucity and excess of nonsynonymous mutations, respectively, are used to define negative and positive selection:

Nature Genetics: doi:10.1038/ng.3987

$$
\begin{aligned}
p_k^{\text{neg}} &= \sum_{k=0}^{k_{\text{obs}}} P(k|s_{\text{obs}}; r_k) \\
p_m^{\text{neg}} &= \sum_{m=0}^{m_{\text{obs}}} P(m|s_{\text{obs}}; r_m) \\
p_k^{\text{pos}} &= 1 - \sum_{k=0}^{k_{\text{obs}}-1} P(k|s_{\text{obs}}; r_k) \\
p_m^{\text{pos}} &= 1 - \sum_{m=0}^{m_{\text{obs}}-1} P(m|s_{\text{obs}}; r_m)\,. \qquad (13)
\end{aligned}
$$

While tumor subclonality and other factors affecting local mutation density are statistically accounted for, these p-values are designed to be effective measures of selection, not aimed at deconvolving time-dependent selection pressures that may arise at different stages of tumor evolution.

3.2 Meta-statistic ϕ

In order to combine the signals from the selective effects acting on the missense and the nonsense category in a joint measure, we then meta-analyze these p-values in a similar manner as Fisher's method:

$$
\begin{aligned}
\phi_{\text{neg}} &= -\log p_k^{\text{neg}} - \log p_m^{\text{neg}} \\
\phi_{\text{pos}} &= -\log p_k^{\text{pos}} - \log p_m^{\text{pos}}\,. \qquad (14)
\end{aligned}
$$

Hence, with increasing deviation of the observed mutation count in either category m or k from the expectation (i.e. decreasing p_m or p_k), the meta-statistic ϕ increases in value. For each gene, two deviations are analyzed: overall depletion (ϕ_{neg}) and overall excess (ϕ_{pos}).
Due to the discrete nature of the observables m and k, p-values p_m and p_k are not uniformly distributed across the range $[0, 1]$. In order to correct for multiple testing in this case, we derive the null distribution of the resulting meta-statistic ϕ from simulation, since it is not simply given by a χ^2-distribution. We simulate this null expectation under neutral evolution by generating missense and nonsense counts for each gene in each cancer type according to the conditional probabilities, Eq. 10, in 10,000 independent runs (1,000 independent runs for the pan-cancer data set).
In turn, we compute the rank ordered p-values $p_{\phi,(g)}$ and q-values $q_{\phi,(g)}$ for ϕ_{neg} and ϕ_{pos}, in

Nature Genetics: doi:10.1038/ng.3987

direct analogy to the Benjamini-Hochberg correction for the uniform case[13]:

$$q_{\phi,(g)} = \min_{j \geq g} \left(I(p_{\phi,(g)}) \frac{G}{j} \right) , \tag{15}$$

where $I(p_{\phi,(g)}) = \sum_{p_\phi=0}^{p_{\phi,(g)}} f_0(p_\phi)$, and $f_0(p_\phi)$ is the distribution of p_ϕ under the null hypothesis, taken from the simulation. The q-values defined in Eq. 15 thus represent an upper bound of the false discovery rate (FDR).[13] Note that this is equivalent to computing the FDR at q^* directly as the ratio of the fraction of genes with $\phi \geq \phi^*$ under the simulated neutral expectation and the observed fraction with $\phi \geq \phi^*$. Alternatively, we could have computed a p-value from each gene-specific simulated neutral distribution, and in turn applied Eq. 15 substituting the superposition of all gene-specific distributions of p_ϕ for f_0.

4 Genome-wide fraction of genes under selection

To estimate the overall fraction of genes under negative and positive selection, we evaluated the excess cumulative probability in the respective large-ϕ regime of the observed distributions of ϕ_{neg} and ϕ_{pos} relative to the neutral expectation from simulation (sum over histogram bins of width $\Delta\phi$). This regime is defined as the region in which the observed distribution consistently exceeds the expectation, as determined from a moving average. Note that this amounts to assigning a weight of 1 to the neutral component and that it does not consider the left tail of the true underlying distribution of the genes under selection, both of which makes this estimate of the selected fraction conservative. **Supplementary Table 3** shows the thresholds in the different cancer (sub)types that define the large-ϕ regime beyond which we computed the cumulative probabilities (cf. **Figure 2** and **Supplementary Figures 2-25,27**). In order to account for stochasticity introduced by the bin choice $\Delta\phi$, we derived the mean selected fraction from the five binning schemes $\Delta\phi \in \{0.15, 0.20, 0.25, 0.30, 0.35\}$, which is what is reported in **Figure 3** for each cancer (sub)type. We obtained a standard error of the mean of 0.017% and 0.020%, for negative and positive selection, respectively. In addition to the estimation method, limited power to detect selection, particularly of the negative kind, makes the inferred selected fractions conservative.

5 Selection prediction validation

5.1 Negative selection

We compared our negative selection predictions to likely cell-essential genes from Wang *et al.* (2015).[2] From that data set we selected all genes that have a significant CRISPR score < 0 (adjusted p-value ≤ 0.1) in any of the four human cell lines (KBM7, K562, Raji and Jiyoye) as well as a gene-trap score ≤ 0.2, leaving 77 genes as assumed true positives (**Supplementary Table 13**). CRISPR scores listed in **Supplementary Tables** denote means from all cell lines

Nature Genetics: doi:10.1038/ng.3987

with adjusted p-value ≤ 0.1. We computed significance of the AUC, i.e. area under the receiver operating characteristic (ROC) curve, and find $p_{\mathrm{AUC}} \leq 0.05$ (≤ 0.002) for 18 (10) out of the 25 cancer (sub)types as well as for the pan-cancer data set ($p_{\mathrm{AUC}} =$ 5e-6; **Supplementary Table 4**). We cross-validated the gene essentiality screen with a second set of 1,734 genes predicted to be cell-essential.[14] Here we find $p_{\mathrm{AUC}} < 0.05$ (≤ 0.002) in 14 (11) of the 25 (sub)types, and in the pan-cancer data set ($p_{\mathrm{AUC}} =$ 2e-5).

In addition, we derived results under a naïve model of ranking genes according to overall mutation paucity (i.e. across all categories m, k and s) relative to the expectation considering gene length and sequence context. Transcription-coupled repair mechanisms entail a reduced overall substitution rate on highly expressed genes[1] (see also **Supplementary Table 16**). Because of a positive correlation between gene expression and gene essentiality (cf. e.g. Tu *et al.*, 2006;[15] Blomen *et al.*, 2015[14]), an incidental enrichment with cell-essential genes in a naïve screen for global hypomutation is therefore not unexpected. We find $p_{\mathrm{AUC}} < 0.05$ (≤ 0.002) for cell-essential genes in the hypomutation screen for 6 (2) of the 25 (sub)types (**Supplementary Table 4** shows the corresponding p_{AUC}). This shows that the enrichment of the negative selection signal with cell-essential genes is not driven by global mutation paucity.

5.2 Positive selection

Positive selection estimates were assessed through prediction power of cancer type-specific causally implicated genes from the COSMIC cancer gene census[5] (**Supplementary Table 14**). For comparing the sensitivity of CBaSE to that of MutSigCV[1] (v1.2), MutSigCV was run with all exonic mutations (columns: Chromosome, Start_position, Reference_Allele, Tumor_Seq_Allele1, Tumor_Seq_Allele2, gene, patient and effect). Gene and effect (nonsilent, silent, null, noncoding) annotations were the same as used in our analysis, and we used the covariate and coverage tables provided with the online tool. The overlap between our list of 18,666 genes and the list for which MutSigCV generated predictions is 16,929, which is the total number used in the ROCs (**Supplementary Table 15**). Because of memory restrictions, MutSigCV could not produce predictions for the pan-cancer analysis (even using the maximum possible memory allocation). Hence we only plot the ROC corresponding to CBaSE predictions of positive selection in the pan-cancer data set in **Supplementary Figure 27**. MutSigCV is outperformed for 21 out of 25 cancer (sub)types by CBaSE, as measured by AUC. The fit of $P(\lambda_s|\boldsymbol{\theta})$ tracks with the amount of synonymous variation and, consistently, the four cancer types for which the predictions of cancer driver genes currently do not produce higher AUC than MutSigCV are the four with the smallest numbers of synonymous mutations (PRAD: 774, CLL: 787, LAML: 1,002, MM: 2,820). Because the amount of data is widely expected to grow, it is also to be expected that this mutation regime will quickly be exceeded for many more cancer types.

Nature Genetics: doi:10.1038/ng.3987

5.3 Computation of $p_{\mathbf{AUC}}$

When using rank-order statistics like area under the ROC curve in a setting where individual observations from the data set have differential power, it is important to note that the rank of an observation is influenced by both its signal (here: selection) as well as its power, or detectability (here: number of mutations relative to expectation; determined by the mutability of a gene and its target size, i.e. λ_x). That is, two genes under the same selection strength can be ranked far apart if only one of them has enough statistical power to be detected as selected. Likewise, two genes that evolve neutrally can also, in expectation, be ranked differently if they have different power. If one then considers a subgroup of genes that is enriched (depleted) for detectability, their relative ranks within the full set can be artificially inflated (deflated). This effect will be more pronounced in case the "signal" is hypomutation (negative selection) compared to hypermutation (positive selection). In the ROCs for cell-essential and known cancer genes, we corrected for this possibility by deriving the expected ROCs from the mean of 50 replicates of simulated neutral counts (m, k, s). This shows that for some cancer types, cell-essential genes have on average indeed a shifted detectability compared to the rest and, hence, exhibit AUCs that are divergent from the random expectation of 0.5 even in the neutral simulation. While a few ROCs for the positive selection signal also show slight deviations from the random expectation, the effect is far less pronounced. Additionally, we hypothesize that both rankings from CBaSE and MutSigCV are subject to the same dependence on power, as the underlying per-gene data are the same, ensuring direct comparability of the two AUCs. We derived the corrected significance of the area under the ROC curve of the observed data, p_{AUC}, from the variance of AUCs over the 50 replicates of neutral simulation, assuming a Gaussian error term.

Since the expectation for AUCs in the naïve hypomutation screen is 0.5, significance was in this case computed from a Mann-Whitney test using $AUC = U/(n_{\mathrm{P}} n_{\mathrm{N}})$, with n_{P} true positives and n_{N} true negatives, from which we derive the corresponding standard score $z = (U - n_{\mathrm{P}} n_{\mathrm{N}}/2)/\sqrt{(n_{\mathrm{P}} n_{\mathrm{N}}(n_{\mathrm{P}} + n_{\mathrm{N}} + 1)/12}$, and p-value, p_{AUC}.[16]

6 Model validation

6.1 Correlation of inferred mutation probability with known mutation covariates

In order to validate the correspondence between known mutation rate covariates and our per-gene predictions of mutation probability, we computed the correlations of the predicted values with gene expression, replication timing, and chromatin status as defined by HiC. For each gene, we derived the expectation $\bar{\lambda}_{\mathrm{s}}$ from the posterior distribution Eq. 12, given the observed s_{obs}. The estimator of the local mutation probability then is $\hat{\mu} = \bar{\lambda}_{\mathrm{s}}/\ell_{\mathrm{s}}$, which we compare to per-gene data from Table S5 of Lawrence *et al.* (2013)[1] (columns "expression_CCLE", "replication_time", and "HiC_compartment"). 2,452 genes that did not have a corresponding

Nature Genetics: doi:10.1038/ng.3987

covariate data entry in the table were omitted. We find for all cancer types that mutation probability $\hat{\mu}$ is negatively correlated with gene expression (-0.02 $\geq r \geq$ -0.22), positively correlated with replication time ($0.04 \leq r \leq 0.31$), and negatively correlated with HiC status (-0.03 $\geq r \geq$ -0.25). All three directions of correlation are in line with previously reported results about mutational covariates.[1,17] **Supplementary Table 16** shows Pearson's r for each covariate category and each cancer type.

6.2 Effect of SNPs on cell-essential gene enrichment

To further investigate the effects of SNPs on the detection of negative selection, we repeated the analysis for the artificial data set that is created when discarding all SNP-overlapping somatic mutations, reducing power to detect negative selection. To avoid a bias with respect to the null, we also adjusted the target sizes ℓ_x by recomputing the context-dependent mutation matrix without SNP-overlapping somatic mutations. The numbers of removed somatic mutations are: 689 (BLCA), 1610 (BRCA), 118 (CLL), 4768 (CRC), 554 (DLBCL), 789 (ESO), 1473 (GBM), 1830 (HNSC), 655 (KIRC), 180 (LAML), 2199 (LUAD), 1114 (LUSC), 2035 (MEL), 357 (MM), 523 (OV), 165 (PRAD), and 9424 (UCEC). We find that under this artificial bias against the negative selection signal, the 77 cell-essential genes are still significantly enriched with negative selection in the same set of the 17 original cancer types (as measured by $p_{\mathrm{AUC}} \leq 0.05$ of the ROC curves). The mean decrease in AUC across those cancer types is 0.005, amounting to a change of $< 1\%$ in the average AUC.

6.3 Variance in r_x

Another relevant issue concerns the stochasticity of the estimate of the ratio of nonsynonymous and synonymous target size, r_x. Two potential sources of variation are sampling variance and the effects of heterogeneity of the mutation signature, which we shall consider in turn.
To estimate sampling variance of r_x, we generated 100 bootstrapped versions of the original per-type data sets, obtained from random sampling of patients with replacement. In each bootstrap sample, we computed the mutation signature p_0^{kl} (Eq. 9) and derived per-gene values r_x, as for the real data set. We then simulated nonsynonymous mutation counts according to Eq. 10, as for the generation of the null distribution of p-values, but with $r_x \sim P_{\mathrm{bootstrap}}(r_x)$. This simulates a noisy version of a neutrally evolving set of genes, and can expose whether sampling noise alone can generate a significant signal of selection. We compared the resulting distributions of ϕ_{neg} and ϕ_{pos} again to the null distributions shown in **Figure 2** and **Supplementary Figures 2-25**. In 50 independent replicates per cancer type, we never find a significant genome-wide signal of positive or negative selection for any of the cancer types after Bonferroni correction ($p \leq 0.001$). We conclude that the measured signals of both negative and positive selection are robust under sampling variation.

Nature Genetics: doi:10.1038/ng.3987

Another potential source of variability of r_x is heterogeneity of the effective mutational signature along the genome, or "topography": If multiple mutational processes underlie a given mutational signature and their relative contributions vary as a function of genomic locus, this may cause deviations from the estimated r_x at a given locus. For example, differential relative contributions to the effective signature as a function of replication time have been reported for breast cancer.[18] We investigated the effects of mutation signature heterogeneity on the selection measures ϕ_{neg} and ϕ_{pos} by comparing the estimates based on the whole-exome signature to those based on signatures inferred separately in quintiles of (1) replication time, (2) expression level, and (3) HiC status[1] ("covariates", as described above). **Supplementary Figure 31** shows the comparison of ϕ_{neg} and ϕ_{pos} for the top 10 genes in the respective selection screen, pooled across all 17 original cancer types and across all three covariates. Overall per cancer type, we find Pearson correlation coefficients across both types of selection and all covariates in the range $\rho \in [0.96, 1.00]$ (mean: 0.997). We conclude that topography as measured here does not confound our inference of selection.

7 Comparison to breast cancer study

Going beyond covariate clustering-based algorithms to infer local mutation probabilities, a recent advance in the field was presented by Nik-Zainal *et al.* (2016),[8] who estimated per-gene mutation rates in breast cancer samples by assuming observed synonymous mutations to follow a negative binomial distribution. This amounts to assuming that per-gene mutation probabilities follow a gamma distribution. However, it is not guaranteed that the observed distribution of synonymous counts matches the negative binomial form and, hence, that the true distribution of per-gene mutation probabilities is approximated accurately by the presupposed gamma functional form. We have shown here that a probabilistic model that fits the observations exactly and adapts to the data complexity can enable addressing the broad range of variability across cancer types caused by differences in mutation processes and detection quality. Set in a Bayesian framework, we have found that this holds the potential to infer selection at higher sensitivity, including the subtle signal of negative selection.

Nature Genetics: doi:10.1038/ng.3987

References

[1] Lawrence, MS, Stojanov, P, Polak, P et al. Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* 499, 214-218 (2013).

[2] Wang, T, Birsoy, K, Hughes, NW & Krupczak, KM. Identification and characterization of essential genes in the human genome. *Science* 350, 1096-1101 (2015).

[3] Boyd, K, Eng, KH & Page, CD. Area under the precision-recall curve: Point estimates and confidence intervals. *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer Berlin Heidelberg (2013).

[4] Lawrence, MS et al. Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature* 505, 495 (2014).

[5] Futreal, PA et al. A census of human cancer genes. *Nature Reviews Cancer* 4, 177 (2004).

[6] Martincorena, I et al. Universal patterns of selection in cancer and somatic tissues. *bioRxiv* 10.1101/132324 (2017)

[7] Rosenthal, R et al. deconstructSigs: delineating mutational processes in single tumors distinguishes DNA repair deficiencies and patterns of carcinoma evolution. *Genome Biology* 17, 31 (2016).

[8] Nik-Zainal, S et al. Landscape of somatic mutations in 560 breast cancer whole-genome sequences. *Nature* 534, 47-54 (2016).

[9] Meyer, MJ et al. mutation3D: Cancer Gene Prediction Through Atomic Clustering of Coding Variants in the Structural Proteome. *Human Mutation* 37, 447-456 (2016).

[10] McGill, R, Tukey, JW & Larsen, WA. Variations of box plots. *The American Statistician* 32, 12-16 (1978).

[11] Kimura M. On the probability of fixation of mutant genes in a population. *Genetics* 47, 713 (1962).

[12] Ikehata, H & Ono, T. The mechanisms of UV mutagenesis. *Journal of Radiation Research* 52, 115-125 (2011).

[13] Benjamini, Y & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society.* Series B (Methodological) 289-300 (1995).

[14] Blomen, VA et al. Gene essentiality and synthetic lethality in haploid human cells. *Science* 350, 1092-1096 (2015).

[15] Tu, Z et al. Further understanding human disease genes by comparing with housekeeping genes and other genes. *BMC Genomics* 7, 31 (2006).

[16] Mason, SJ & Graham, NE. Areas beneath the relative operating characteristics (ROC) and relative operating levels (ROL) curves: Statistical significance and interpretation. *Quarterly Journal of the Royal Meteorological Society* 128, 2145-2166 (2002).

[17] Polak, P et al. Cell-of-origin chromatin organization shapes the mutational landscape of cancer. *Nature* 518, 360 (2015).

[18] Morganella, S, et al. The topography of mutational processes in breast cancer genomes. *Nature Communications* 7 (2016).