Mutation of NOTCH1 is selected within normal esophageal tissues, yet leads to selective epistasis suppressive of further evolution into cancer

Kira A. Glasmacher[1] (https://orcid.org/0009-0004-6300-3812),
Vincent L. Cannataro[1] (https://orcid.org/0000-0002-6364-7747),
Jeffrey D. Mandell[3] (https://orcid.org/0000-0002-3839-2543),
Mia Jackson[2] (https://orcid.org/0009-0005-4111-6098),
J. Nic Fisk[2,3,4] (https://orcid.org/0000-0002-1940-393X),
Jeffrey P. Townsend[2,3,5†] (https://orcid.org/0000-0002-9890-3907)

[1]Department of Biology, Emmanuel College, Boston, MA,
[2]Department of Biostatistics, Yale School of Public Health, New Haven, CT,
[3]Program in Computational Biology and Bioinformatics, Yale University, New Haven, CT,
[4]Department of Cell and Molecular Biology, University of Rhode Island, Kingston, RI, and
[5]Department of Ecology and Evolutionary Biology, Yale University, New Haven, CT,

[†]To whom correspondence should be addressed: Jeffrey Townsend (203 737-7042), Department of Biostatistics, Yale School of Public Health, 135 College Street, New Haven, CT 06510.

The authors declare no potential conflicts of interest.

Running Title: NOTCH1 early selection and epistasis in esophageal tissues

Abstract

Sequencing of tissues from histologically normal esophagus, among other organs, has revealed that normal tissues harbor somatic variants that are also found in cancers arising from the same tissue types. Our understanding of how somatic mutations commonly found in normal tissue can contribute to tumorigenesis is limited: common somatic mutations may or may not confer phenotypes compatible with oncogenesis. However, the strength of selection for somatic variants that appear in both normal and cancer tissues can be quantified in each context using evolutionary modeling approaches. We studied the evolutionary trajectory from normal esophageal tissue to esophageal squamous-cell carcinoma (ESCC) by analysis of 2171 sequenced samples from previous studies on normal esophageal epithelium and ESCC to reveal the stepwise contributions of somatic mutations to increased cellular division and survival. We also analyzed pairwise selective epistasis between somatically mutated genes that may lead to stepwise substitution patterns. We found that *NOTCH1* substitutions are highly selected along the trajectory from embryogenesis to adult normal esophageal tissue, explaining their high prevalence in the tissue. In contrast, there is little to no positive selection for *NOTCH1* mutations along the trajectory from adult normal tissue to ESCC, suggesting that *NOTCH1* substitutions do not drive tumorigenesis. Furthermore, mutations in NOTCH1 exhibit antagonistic epistasis with well-known cancer drivers including TP53, reducing selection for progressive mutations in tumorigenesis. This antagonistic epistasis likely corresponds with a low likelihood of tumor progression in the presence of *NOTCH1* mutations in the esophagus.

Introduction

Somatic cells acquire mutations as we age, and as we are exposed to mutagens. Many of these mutations have little to no effect on the rates that cells proliferate and survive; others can lead to benign or malignant growths via increasing cell division and survival (1). The effects of such mutations on the development of cancer have been characterized on the levels of the genome (2), the gene (3), and even the nucleotide or amino acid (4). Sequencing of histologically precancerous tissues has revealed that precancerous tissues share some genetic variants originally detected at high prevalence within cancers originating from those tissues—for example in atypical hyperplasia, which is a precursor of uterine endometrioid carcinoma (5). However, construction of trajectories from precancerous to cancerous tissue cannot be comprehensively performed solely on the basis of relative substitution frequency because these frequencies are the product of the rate at which mutations arise and also the propensity these mutations have to reach fixation in evolving populations of cancer cells (i.e. the intensity of selection—quantifiable as scaled selection coefficients). These mutation rates differ along the trajectory of tumorigenesis and must be accounted for prior to calculating contribution of variants to increases of cellular division and survival influencing selection. Accordingly, calculation of the strength of selection early in oncogenesis can be performed by stepwise deconvolution of the influence of mutation rate on substitution prevalence in histologically normal tissues and their associated cancers.

Historically, there have been technical difficulties in detecting mutations within normal tissues due to low frequencies of genetic alterations and high normal tissue stem-lineage heterogeneity. New techniques, including deep sequencing of small biopsies containing clones

(6–9) and sequencing elements from microscopically distinct clonal units (10,11), have enabled detection of substitutions that are present within a small fraction of cells. Some of those studies, like many more typical bulk tumor sequencing studies, have relied on variant prevalence among samples (or among tumors) to draw inference about the degree to which genes contribute to tumorigenesis. However, prevalence estimates fall short of quantifying the contribution of somatic mutations to the tumor formation process. Differences in the prevalence of somatic mutations in tumors can arise in specific genes due to quantitative differences in underlying mutation rate that are not related to biological functions associated with oncogenesis, and statistical models for driver detection have striven to account for these underlying differences across the genome by martialing both gene-level synonymous mutation rates and gene-level covariates (12). Accordingly, technical advancements have been developed that estimate the strength of selection throughout the genome—enabling us to quantify selection on both the gene and genome-level. Gene-level techniques have used synonymous substitutions to quantify an underlying "neutral" mutation rate, examining their frequency relative to nonsynonymous substitutions to detect drivers (12) and quantify gene-level selection (3,8). Indeed, recent gene-level studies have identified recurrent substitutions in cancer-associated genes within physiologically normal skin, esophageal, colorectal, and endometrial epithelial cells (6,9,10,13), and clonal hematopoiesis in blood (14,15).

Application of methods that detect somatic selection in the skin epidermis have indicated that the distribution of most substitutions detected in normal tissue are consistent with neutral evolution (1,13). These methods also revealed positive selection for mutations in genes that have been previously associated with tumorigenesis (e.g. *NOTCH1*, *TP53*). However, the inference of

positive selection of putative cancer driver mutations in normal tissues does not directly imply that these mutations contribute to tumorigenesis. For example, some variants may be prevalent in cancer simply because they are positively selected for in healthy cell lineages, some of which eventually become cancerous via unrelated driver events. Epistatic interactions altering selective impact of other mutations along an evolutionary trajectory may be common (16), affecting the importance of mutations at one stage of development compared to another. Understanding the potentially dynamic roles of mutations in healthy cell lineages and in tumorigenesis has tremendous potential to inform therapy and open new avenues for drug discovery (17–19).

Our understanding of the differential evolutionary trajectories in humans that lead to benign clonal expansions instead of malignant cancers is limited. A conventional understanding of clonal evolution and selection suggests that driver alterations would occur more frequently as the progression from benign to premalignant to malignant neoplasms takes place. However, this assumption has been challenged. Some puzzling examples that could relate to such a phenomenon include the facts that BRAF V600E driver mutations have been reported in 70–88% of benign nevi compared to 40–45% of melanomas, and HER2 expression is more commonly found in ductal carcinoma in situ than in invasive mammary cancers (20). Some variants could even confer a degree of cancer resistance, reducing selection for other cancer drivers. For example, Higa & DeGregori (21) hypothesized that *NOTCH1* mutant cells may have a lower likelihood of accumulating mutations that drive malignant trajectories in the esophagus because they have reached distinct adaptive fitness peaks associated with clonal expansions of the normal esophageal epithelium. Specifically, mutations that result in a reduction or loss of NOTCH1 function are thought to increase cell fitness: nonsense and essential splice mutations are under

especially strong positive selection in normal esophageal epithelia (22,23). The result of strong positive selection of *NOTCH1* mutants in normal tissue may be a reduced competitive advantage in tumor cells (24). Indeed, a recent study in mice demonstrated that mutation of *NOTCH1* in normal esophageal tissue caused clonal expansion, but impaired tumor growth (25). Quantification of selection in tumor and normal tissue of the esophagus under an evolutionary model that accounts for changes in the rate at which mutations occur across the steps of tumorigenesis could clarify the role of mutations in *NOTCH1* and other genes in clonal expansion within normal tissue and also oncogenesis.

We investigated the differential selective effects of variants within normal esophageal tissue and esophageal squamous-cell carcinoma (ESCC) at the nucleotide level. We quantified mutation rates and scaled selection coefficients on ESCC-associated genes within normal esophageal and ESCC tissue. To reveal interactions between putative cancer-causing genes, we analyzed ordered, pairwise epistatic effects of genes associated with ESCC, including *NOTCH1* and *TP53,* on precursor cell selection in normal tissue and malignant tumors during tumorigenesis.

Methods

Data acquisition and preprocessing

Targeted sequencing data from 137 ESCC samples were obtained from cBioPortal for Cancer Genomics (UCLA). Additionally, we obtained targeted sequencing data from 142 physiologically normal and 66 tumor samples of esophageal epithelia from Yokoyama et al. (7) and targeted sequences from 832 normal samples reported in Martincorena et al. (8). In

Martincorena et al. (8), multiple biopsies were taken from the same donor. To prevent the possibility of counting the same instance of fixed mutations as independent evolutionary events, only the sequences that were considered independent from each other and that were not merged into single clones based on shared mutations and spatial information were used for this analysis. Similarly, we analyzed variants in sequences taken from the same donor in data from Yokoyama et al. (7) and only kept samples that had no variants in common whenever they were taken from biopsies that were physically close together. In addition to those data, Yuan et al. (26) provided whole-exome sequencing data from 42 normal and 10 tumor samples. Within the Yuan et al. (26) data, seven samples that were classified as normal but exhibited severe dysplasia or carcinoma in-situ were excluded from the analysis: they potentially have clonal relation to a neighboring tumor, sharing molecular variants identical with the primary tumor from that patient. We obtained whole-exome sequencing data from an additional 25 ESCC samples sequenced by (27). Lastly, whole-exome and whole-genome sequencing data from 1503 samples were obtained from an integrated cohort of ESCC data including 17 different datasets from TCGA, ICGC, and other published studies (28). The deduplication of samples described above resulted in $n = 2171$ samples classified as $n = 430$ that were physiologically normal tissue and $n = 1741$ that were malignant. All samples used in this analysis, their sourcing from their original study, and their normal/tumor designation are listed within **Supplementary Table 1**.

ESCC-associated genes were compiled based on previous studies on the mutational landscape of normal esophageal epithelium and ESCC (29–35). To focus our analysis on genes with variants affecting substantial numbers of patients, we performed our analysis on genes from this list that exhibited substitutions in 25 or more samples in the combined dataset. For the

analyses of selection, all nucleotide variants located one or two positions apart in the same tumor sample were removed to ensure that the single-nucleotide variants (SNVs) we analyzed were reflective of mutation rates estimated for SNVs and not possible multi-nucleotide events. Reanalysis of the data without exclusion of these possible multi-nucleotide events did not result in any substantial differences to our conclusions.

Data availability and reproducibility statement

All data are publicly available and all analyses are reproducible via a freely available pipeline located here: https://github.com/Cannataro-Lab/ESCC_step_epistasis.

Stepwise mutation rates

Under a stepwise model of tumorigenesis, tissues accumulate mutations and progress from embryogenesis to adult normal tissue and in some cases to ESCC. Dividing continuous chronological time into discrete pathology-determined steps, mutations can occur and substitutions may fix early (during the mutation accumulation from embryogenesis to adult normal tissue; Step 0→1) or late (from adult normal tissue to ESCC; Step 1→2). During these discrete steps, genes may be subject to differing mutation rates. Since synonymous mutations can be presumed to be neutral (1,3), and neutral mutations fix in a population at a rate equal to the rate of mutation (36), synonymous substitution rates are informative of mutation rates across all sites. To estimate gene-specific proportions of early and late mutations, we executed dNdScv (3), using all whole-exome and whole-genome data and tissue-specific covariates (4), separately for normal and ESCC samples. For each cohort, dNdScv provided expected numbers of synonymous

substitutions for each gene under its covariate-informed model (*exp_syn_cv*). We divided this output by the number of samples in respective cohorts, then by the number of synonymous sites per gene to yield expected rates of mutation per nucleotide for each gene *g*, $E_{0\rightarrow1,g}$ and $E_{0\rightarrow2,g}$, for each of the pathologically determined tissue sets. The proportion of mutations accumulated in the trajectory from embryogenesis to adult normal tissue for each gene was assessed as $p_{0\rightarrow1,g} = E_{0\rightarrow1,g}/E_{0\rightarrow2,g}$, and the proportion accumulated in the trajectory from adult normal tissue to ESCC was $p_{1\rightarrow2,g} = 1 - p_{0\rightarrow1,g}$.

We used cancereffectsizeR (37 v. 2.7.0) to calculate sample-specific mutation rates summed over sets of genomic sites of interest within the ESCC data set and to fit our model of stepwise selection. For each tumor *i* and a defined set of sites in a gene *g,* cancereffectsizeR calculated a mutation rate $\mu_{0\rightarrow2,g,i}$ summed over sites of interest, informed by rates of synonymous gene substitution ($E_{0\rightarrow2,g}$) and by sample-specific patterns of mutation obtained via mutational signature analysis (4). Then we define the mutation rate for the defined set of sites as $\mu_{0\rightarrow1,g,i} = p_{0\rightarrow1,g} \times \mu_{0\rightarrow2,g,i}$ in the initial step, and the mutation rate during the second step as $\mu_{1\rightarrow2,g,i} = p_{1\rightarrow2,g} \times \mu_{0\rightarrow2,g,i}$. For oncogenes *PIK3CA* and *NFE2L2,* we defined the set of sites of interest to be those that are putatively selected—including any recurrent variant in the data set; for tumor-suppressor genes *FAT1*, *NOTCH1*, *NOTCH2*, *TP53*, *FBXW7*, and *RB1,* we included recurrent variants as well as all nonsense/stop-loss variants (37). The classification of genes as oncogenes or tumor suppressors followed Chakravarty et al. (38). All mutation rates and scaled selection coefficients reported here are calculated over these defined sites of interest.

Stepwise scaled selection calculations

In Mandell et al. (37), likelihood for the strength of selection γ was calculated for any single cohort of tumors by maximizing

$$\mathcal{L}(\gamma|\mu_1,\ldots,\mu_n,\ldots,\mu_N) = \prod_{i=1}^{n} \mathrm{e}^{-\mu_i\gamma} \times \prod_{i=n+1}^{N} 1-\mathrm{e}^{-\mu_i\gamma}, \quad \textbf{(1)}$$

where μ_i, $1 \leq i \leq N$, is the mutation rate of a defined set of sites in sample i, and where $n < N$ are defined such that any variant in the defined set of sites is absent in n samples and is present in samples $n+1\ldots N$, and where the subscript *g* is suppressed in Eq. 1 and later equations for clarity of presentation

To estimate scaled selection coefficients along a multi-step sequential model of tumorigenesis, we extended **Eq. 1** to differentiate two steps—evolution from embryogenesis to adult normal tissue (Step 0→1) and evolution from adult normal tissue to tumor tissue (Step 1→2):

$$\begin{aligned}\mathcal{L}(\gamma_{0\to1},\gamma_{1\to2}|\mu_{0\to1,i}\forall i \in 1\ldots M+N, \mu_{1\to2,i}\forall i \in 1\ldots N) = \\ \prod_{i=1}^{n}(e^{-\mu_{0\to1,i}\times\gamma_{0\to1}} \times e^{-\mu_{1\to2,i}\times\gamma_{1\to2}})\times \\ \prod_{i=n+1}^{N}(1-e^{-\mu_{0\to1,i}\times\gamma_{0\to1}} \times e^{-\mu_{1\to2,i}\times\gamma_{1\to2}})\times \\ \prod_{i=N+1}^{N+m} e^{-\mu_{0\to1,i}\times\gamma_{0\to1}}\times \\ \prod_{i=N+m+1}^{N+M}(1-e^{-\mu_{0\to1,i}\times\gamma_{0\to1}})\end{aligned} \quad ,\textbf{(2)}$$

for a dataset with M normal tissue samples, of which $M - m$ exhibit a substitution and m do not, and where there are N tumor tissue samples in total, of which $N - n$ exhibit a substitution and n do not. The normal tissue samples are identified by index $i = 1 + N, \ldots, m + N, \ldots, M + N$ and the tumor tissue samples are also identified by index $i = 1, \ldots, n, \ldots, N$. The first product term in **Eq. 2** represents tumor samples where all variants in the defined set of sites are absent, the second product term represents tumor samples where any such variant is present, the third product represents normal samples where all such variants are absent, and the fourth product represents normal samples with any such variant present.

Eq. 2 can further be generalized for any number of steps S:

$$\mathcal{L}(\gamma_{0\to1}, \ldots, \gamma_{S-1\to S} | \mu_{s-1\to s,i} \forall i \in 1 \cdots \sum_{a=s}^{S} N_a) =$$
$$\prod_{h=1}^{S} \left(\prod_{i=(\sum_{a=h+1}^{S} N_a)+1}^{(\sum_{a=h+1}^{S} N_a)+n_h} \left(\prod_{j=1}^{h} e^{-\mu_{h-1\to h,i} \times \gamma_{j-1\to j}} \right) \times \prod_{i=(\sum_{a=h+1}^{S} N_a)+n_h+1}^{\sum_{a=h}^{S} N_a} \left(1 - \prod_{j=1}^{h} e^{-\mu_{h-1\to h,i} \times \gamma_{j-1\to j}} \right) \right), \quad \textbf{(3)}$$

where $s = 1, \ldots, S$ denotes the step of progression toward cancer, and S corresponds to the last step present in the data. For each step s, there are N_s samples: n_s without a variant and $N_s - n_s$ with a variant in the defined set of sites. The product index *h* specifies the step of a sample, *i* specifies the sample index, and *j* functions as an index that breaks down all *h* steps of a sample's trajectory into the distinct stepwise processes. For each step, the samples are indexed by

$$i_s = 1 + \sum_{a=s+1}^{S} N_a, \ldots, n_s + \sum_{a=s+1}^{S} N_a, \ldots, N_s + \sum_{a=s+1}^{S} N_a$$

resulting in

$$i = 1, \ldots, \sum_{a=1}^{S} N_s$$

across all steps.

Mutation rates for epistatic selection

For our analysis of pairwise epistatic selection, we employed a model that does not distinguish between normal and ESCC tissue states. We calculated sample-specific site mutation rates within a standard cancereffectsizeR workflow (37), without partitioning the mutation rate as in the stepwise selection model.

Epistatic scaled selection calculation

Under a model of pairwise epistasis, we allowed selection for each defined set of sites in each of a pair of genes to differ depending on the variant status in the defined set of sites of the other gene. We calculated pairwise epistasis using all samples as in Alfaro-Murillo and Townsend (39) and Mandell et al. (37). Under a no-epistasis model of selection, the Poisson rate for a substitution (acquired in a cell then fixed throughout a cell population) that is designated A in tumor i can be estimated as $\lambda_{A_i} = \mu_{A_i}\gamma_A$ (40), where μ_{A_i} is the sample-specific neutral mutation rate and γ_A is the scaled selection coefficient for mutation A. Without epistatic effects, mutations are acquired independently, so the probability $P_i(AB)$ of a tumor having two substitutions designated A and B is $(1 - e^{-\lambda_{A_i}})(1 - e^{-\lambda_{B_i}})$, and the expected number of tumors with both substitutions is $\sum P_i(AB)$. Under a model of pairwise epistatic selection,

each tumor has four estimated rates of acquisition for each pair of variants: λ_{A_0} and λ_{B_0} for mutations A and B in the absence of the other mutation, and λ_{A_B} and λ_{B_A} for acquisition of A and B, respectively, in the presence of the other substitution (for brevity, omitting the tumor index i arising from tumor-specific μ_{A_i}.) These rates incorporate sample-specific mutation rates and epistatic scaled selection coefficients. The tumor-specific probability of observing both substitutions is

$$1 - e^{-(\lambda_{A_0}+\lambda_{B_0})} - \left(\frac{\lambda_{A_0}}{\lambda_{A_0}+\lambda_{B_0}-\lambda_{B_A}}\right)\left(e^{-\lambda_{B_A}} - e^{-(\lambda_{A_0}+\lambda_{B_0})}\right) - \left(\frac{\lambda_{B_0}}{\lambda_{A_0}+\lambda_{B_0}-\lambda_{A_B}}\right)\left(e^{-\lambda_{A_B}} - e^{-(\lambda_{A_0}+\lambda_{B_0})}\right)$$

(39), and the expected number of tumors with both substitutions is the sum of tumor probabilities across the cohort.

Based on this formulation, a general test for epistatic selective effects can be performed by comparing expected mutation counts under the epistatic model to a no-epistasis model. In the no-epistasis model, selection for each set of sites is assumed to be independent of the mutation status of the other set of sites; effectively, scaled selection coefficients for each set of sites are constrained to be equivalent before and after mutation in the other set of sites. We performed likelihood ratio tests on the fitted epistatic and no-epistasis models to identify significant epistatic effects ($P < .05$). This test for epistasis is substantially different from typical cancer data science tests of observed co-occurrence (e.g. 41,42): our model-based test discounts for commonalities of underlying mutation rates for some driver gene pairs that are driven by common underlying mutational processes (40). We further substantiate epistatic relationships by reporting changes of scaled selection coefficients arising from epistasis wherein confidence intervals under the best-fit pairwise epistatic model are non-overlapping with the confidence intervals under a best-fit model with no epistasis.

Selective epistatic relationships were referred to as synergistic when an epistatic model predicted co-substitution patterns statistically significantly better than a null model of no epistasis, and when there were more co-occurrences expected in the epistatic model than the model without epistasis. Selective epistatic relationships were referred to as antagonistic when the best-fit pairwise epistatic model predicted co-substitution patterns statistically significantly better than a null model of no epistasis, and when there were fewer co-occurring samples expected in the pairwise epistatic model than were expected in the best-fit model without epistasis.

Results

Cell-lineage mutation rates differ among normal and tumor tissues

The mean of the count of substitutions across all whole-exome sequenced ESCC samples was greater than the mean of the substitution count across all whole-exome sequenced normal samples ($P < 2.2 \times 10^{-16}$). For each individual gene in this analysis, the per-nucleotide somatic mutation rate for the evolutionary trajectory from embryogenesis to adult normal tissue ($E_{0\rightarrow 1,g}$) was lower than it was for the evolutionary trajectory from embryogenesis to tumor tissue ($E_{0\rightarrow 2,g}$; e.g. **Fig. 1**). Overall, gene-level mutation rates of tumor samples were 38 times larger than the mutation rates of normal samples. The distribution of the ratios was skewed, exhibiting a number of substantially higher ratios. Among ESCC-associated genes, the one exhibiting the largest increase in mutation rate was *NOTCH1*, with a mutation rate in tumors 116 times higher than the mutation rate in normal samples.

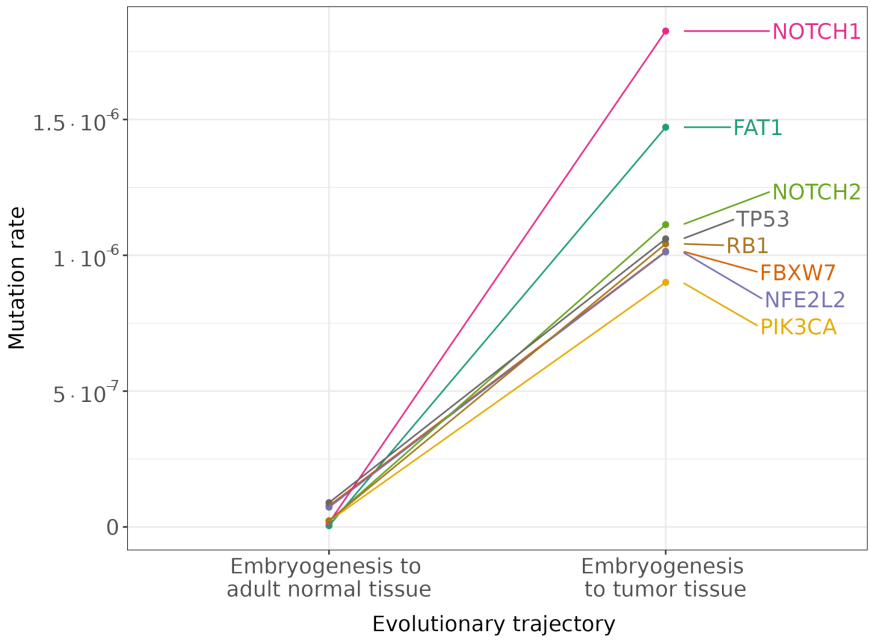

Figure 1. Expected gene-specific nucleotide mutation rates along the evolutionary trajectory from embryogenesis to adult normal tissue and from embryogenesis to esophageal carcinoma, for a selection of ESCC-associated genes.

NOTCH1, FAT1, and NOTCH2 exhibit high scaled selection coefficients within normal tissue, but are not selected within cancers

In the esophagus we detected positive selection for mutations of nonsense, readthrough, and recurrent missense variants in numerous genes. Selection varies between evolutionary trajectories from embryogenesis to adult normal tissue and from adult normal tissue to esophageal squamous-cell carcinoma (**Fig. 2**). In adult normal esophageal tissue, *NOTCH1* substitutions were present in 58.1% of the sequences of samples. In contrast, they were only present in 6.2% of the sequences of tumor samples. *NOTCH2* mutants were present in 7.2 % of the sequences of normal samples, whereas they were present in less than 0.2% of the sequences of tumor samples. The estimated mutation rate for *NOTCH1* within normal tissue cell lineages is low (1.57×10^{-8} in the span from embryogenesis to biopsy) relative to tumor tissue (1.83×10^{-6} in the span encompassing embryogenesis to tumor resection). In contrast to this low mutation rate, a higher prevalence of *NOTCH1* mutations in normal tissue is observed relative to tumor

tissue, a result that is consistent with a previous report from a smaller cohort (8). *FAT1* exhibits similar trends in selection in normal and tumor tissue with a low estimated mutation rate in normal tissue (4.64×10^{-9} from embryogenesis to biopsy) and a higher estimated mutation rate in tumor tissue (1.47×10^{-6} from embryogenesis to resection). The prevalence of *FAT1* was also higher in adult normal tissues (10.9%) relative to tumor samples (4.3%).

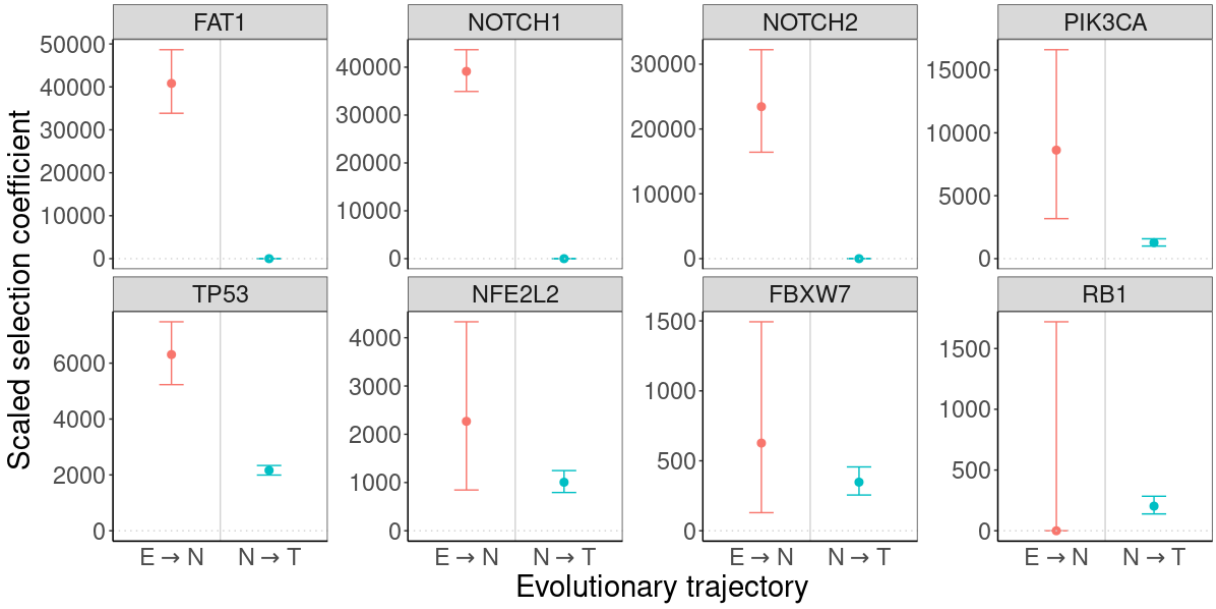

Figure 2. Scaled selection coefficients of mutations in known drivers *FAT1*, *NOTCH1*, *NOTCH2*, *PIK3CA*, *TP53, NFE2L2, FBXW7*, and *RB1* in evolutionary trajectories from embryogenesis to adult normal tissue (E → N) and adult normal tissue to primary esophageal squamous cell tumor (N → T) with 95% confidence intervals.

There is no increase in the relative prevalence of *NOTCH1* substitutions in the trajectory from adult normal to tumor tissue compared to the trajectory from embryogenesis to normal tissue. Such an increase in the relative prevalence of *NOTCH1* substitutions would be expected based on the increase in mutation rate if selection were operating similarly in the tumorigenic

context as in the normal tissue context. Indeed, the scaled selection coefficient of NOTCH1 variants was much higher in normal tissue (3.9×10^4) than in primary ESCC tissue, where the scaled selection coefficient was less than one (i.e. negatively selected against fixation). Mutations of *NOTCH1* are not alone in their stronger selection in normal tissue—a similar result obtains regarding the scaled selection coefficients on *NOTCH2* and *FAT1* variants: much greater selection in normal tissue, and much lower (often negligibly low) selection in tumor tissue.

Selection on mutations of *TP53*, *NFE2L2*, *PIK3CA*, and *FBXW7* contrasted: at mutated sites in these genes, the scaled selection coefficient was also higher in normal tissue than in tumor tissue. However, in these genes, variants were still selected during malignant growth of tumor tissue.

Selection on some mutations of cancer-driving genes depends on the presence or absence of other cancer driver substitutions

Quantification of selection on specific driver variants can be complicated by interactions with other cancer-driving genetic variants. To understand these interactions, we measured the epistatic effects of various genetic combinations. The epistatic interactions were calculated across the evolutionary trajectory from embryogenesis to tumor resection by analyzing data gathered across both steps of cancer evolution.

Statistically significant pairwise selective epistasis was inferred between mutations of *NOTCH1* and *TP53*, as well as mutations of *NOTCH1* or *TP53* paired with mutations of other cancer-driving genes (**Fig. 3**). Mutations of NOTCH1 exhibited significant pairwise epistatic relationships with mutations of NOTCH2 ($P < 0.001$), TP53 ($P < 0.001$), RB1 ($P < 0.01$) and

FAT1 ($P < 0.01$; **Fig. 3A**). The calculated mutation rates of *TP53* and *NOTCH1* variants in esophageal tissues would lead us to expect a higher frequency of co-occurrence than is observed: These genes exhibit antagonistic selective epistasis in the context of oncogenesis. The scaled selection coefficient for mutations of TP53 decreases substantially on a background of NOTCH1 substitution (0.001, 95% CI 0–434) compared to wild-type (2210, CI 2080–2345). Similarly, mutations of NOTCH1 and RB1 exhibit antagonistic epistasis: driver-site mutation of RB1 results in a substantial decrease of selection on driver-site mutations of NOTCH1 (0, CI 0–282; versus 347, CI 310–387) and driver-site mutation of NOTCH1 results in a decrease of selection on driver-site mutations of RB1 (0, CI 0–139; versus 175, CI 120–245). We also detected significant antagonistic epistasis between TP53 and NOTCH2. Substitution in TP53 resulted in a significant decrease of selection on NOTCH2 mutants (73, CI 0–367; versus 644, CI 444–897).

Given the calculated mutation rates of NOTCH1 and NOTCH2, we would anticipate finding fewer tumors with co-substitutions in a model of no epistasis versus a model of epistasis. Observed substitution patterns match our model of epistasis, revealing synergistic epistasis. There is a significant increase in selection on mutations of NOTCH2 with a NOTCH1-substituted background (2713, CI 1523–4413) compared to wild-type (302, CI 188–454). In an additional example of asymmetrical synergistic epistasis, we also detect significant synergistic epistasis between NOTCH1 and FAT1, wherein selection on a mutation of NOTCH1 significantly increases on a FAT1-substituted background (1399, CI 945–1999; versus 317, CI 281–356). We also detected significant synergistic epistasis between TP53 and NFE2L2. A substitution in TP53 resulted in a significant increase in selection on a NFE2L2 mutation (2382, CI 1894–2950; versus 333, CI 224–472).

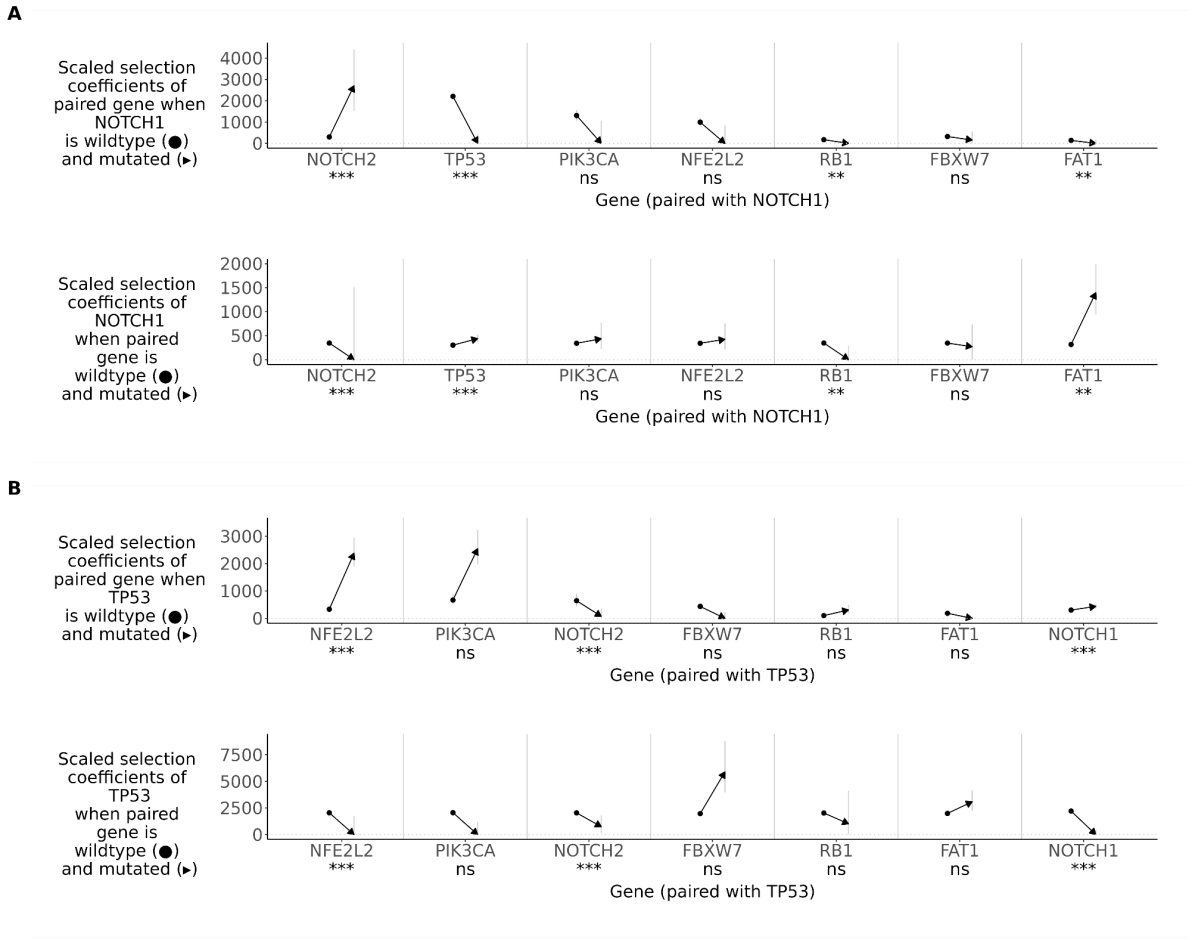

Figure 3. Selective epistasis of mutations in NOTCH1 and TP53 with other genes. (**A**) Context-specific (epistatic) scaled selection coefficients for mutations in *NOTCH1* and other known drivers of esophageal carcinoma *TP53, PIK3CA, NFE2L2, NOTCH2, RB1, FBXW7,* and *FAT1.*
(**B**) Context-specific (epistatic) scaled selection coefficients for mutations in *TP53* and other known drivers of esophageal carcinoma *PIK3CA, NFE2L2, NOTCH2, FBXW7, RB1, FAT1,* and *NOTCH1*.

Discussion

Here we have developed an approach to analyze sequential steps of an evolutionary trajectory toward cancer. Using sequence data obtained from 430 normal esophageal tissue

samples and 1741 primary esophageal squamous-cell carcinoma samples, we quantified distinct scaled selection coefficients for compound sets of variants within genes that rise to high variant allele frequency along the trajectory from embryogenesis to adult normal tissue and along the subsequent trajectory from adult normal tissue to tumor tissue. Our analysis showed that somatic single-nucleotide mutations are differentially selected for during sequential steps of esophageal cancer development. Consistent with past studies that have detected driver variants at higher rates in premalignant conditions compared to tumors, our analysis of NOTCH1 mutations demonstrates that they are strongly selected for in normal esophageal tissue, driving clonal expansions, but not selected for in the trajectory from normal tissue to tumor tissue. Furthermore, we applied a model of pairwise epistasis to calculate the effects of prior mutations on gene-level scaled selection coefficients, demonstrating the selective consequences of previously hypothesized gene-gene interactions between *TP53* and *NOTCH1* (21,43). *TP53* mutations are well-known drivers of esophageal cancer and are frequently found to be mutated within normal tissues and tumors, whereas *NOTCH1* mutations are prevalent in clonal expansions found in human normal esophageal epithelia (7,8) and have been shown to drive clonal expansion in a mouse model of normal esophageal epithelia (25). *NOTCH1* substitutions have been frequently detected in normal adult esophageal epithelia but substantially less frequently detected in ESCC. Here we have shown that this observation is not a result of the difference in mutation rates of normal and tumor samples, but rather a difference in selection, causing *NOTCH1* substitutions to fix at higher rates in normal esophageal tissue. According to (20), the exact mechanism responsible for the paradoxical decrease in the occurrence of driver alterations during malignant progression remains unknown. We have demonstrated that epistatic relationships of a driver gene

like *NOTCH1* with other cancer-associated genes such as *TP53* resolves this paradox. We showed that the mutation rate of *NOTCH1* in adult normal tissue is low, yet the observed prevalence of *NOTCH1* mutations in adult normal tissue is high. We conclude that there is strong selection for *NOTCH1* mutations in normal tissue. Conversely, we show that *NOTCH1* exhibits a high mutation rate in tumor tissue, yet the prevalence of *NOTCH1* mutations in tumors is lower than in normal samples. Accordingly, new *NOTCH1* mutations in tumors do not contribute to increased cellular division and survival in esophageal squamous-cell carcinomas.

In contrast to *NOTCH1* mutations, our analyses of *TP53* mutations indicate that they have been selected for within both normal and ESCC tissues. Both *NOTCH1* and *TP53* experience high mutation rates and strong selection in normal and/or oncogenic tissue environments, leading to a straightforward expectation of proportional co-occurrence in the absence of epistatic selective effects. In contrast, co-occurrence is very low. This low co-occurrence can be explained by antagonistic selective epistasis: non-linear contributions of these two gene alterations to the total cell fitness. Having a *NOTCH1* mutant present in physiologically normal tissue lowers selection for *TP53* mutations. Consequently, tumorigenesis is suppressed because mutated *TP53* is one of the strongest-selected drivers of non-NOTCH1-mutated ESCC. NOTCH1 mutation appears to similarly suppress selective advantage and therefore tumorigenesis derived from mutations in PIK3CA, NFE2L2, RB1, FBXW7, and FAT1; selective antagonistic epistasis is statistically significant for RB1 and FAT1.

The validity of our selection estimates depends on the extent to which same-gene putatively selected sites are under shared selection and have shared epistatic relationships with sites in other genes. While we believe that this assumption reasonably approximates selective dynamics for

many cancer-associated genes, such as for wholly disabling mutations of tumor suppressors, previous analyses have already demonstrated a spectrum of selective effects among variants within the same gene (1,4,44). Future tumor sequencing studies with larger sample sizes will enable the measurement of independent and pairwise epistatic selective effects of individual variants.

On a background of NOTCH1 mutations, estimates of selection on PIK3CA, NFE2L2, and FBXW7 also decrease. However, with current sample sizes, this selective epistasis is not statistically significant ($P > 0.05$). Sequencing data for normal esophageal epithelium and other histologically normal tissue samples are scarce, currently limiting higher-dimensional quantitative analyses in early trajectories of oncogenesis. A larger dataset would likely support additional epistatic interactions (cf. 45,46). Increasingly large datasets will enable estimation of three-way, four-way, and higher-dimensional epistatic interactions, revealing the complex, mutation- and selection-driven fitness landscapes navigated by evolving cancer cells.

We have demonstrated that somatic mutations in these genes confer a change in selection. However, it would also be valuable to investigate copy-number alterations. As with single-nucleotide mutations, the importance of copy-number alterations to cancer evolution has often been assessed by tallying prevalence across a cohort. For instance, (47) have found substantial differences in copy-number alterations in ESCC compared to normal esophageal tissue. Ultimately, an increasingly comprehensive understanding of cancer evolution would arise from integrating copy-number alterations into step-specific and epistatic selective models to better reveal their dynamic roles in cancer.

The different selective advantages conferred by NOTCH1 and TP53 mutations in normal and cancerous esophageal tissue, coupled with their inferred epistatic interactions, underscore the importance of studying early cancer progression, even prior to the initiation of tumorigenesis initiation. Early cancer diagnosis and treatment can substantially improve patient survival, quality of life, and cost of treatment. However, the most promising target genes for future treatments may not be the same for precancerous and cancerous tissues. Our quantification of the stepwise and epistatic changes in selection suggests precision targets in both.

Acknowledgements

We thank Christopher Cross for helping to organize early efforts to start this project, Stephen Gaffney for arranging early data acquisition, Jose Costa for discussion of the topic, and Wes Lewis and Elizabeth Perry for helpful comments on an early draft.

References

1. Cannataro VL, Townsend JP. Neutral Theory and the Somatic Evolution of Cancer. Mol Biol Evol. 2018;35:1308–15.

2. Ostrow SL, Barshir R, DeGregori J, Yeger-Lotem E, Hershberg R. Cancer evolution is associated with pervasive positive selection on globally expressed genes. PLoS Genet. 2014;10:e1004239.

3. Martincorena I, Raine KM, Gerstung M, Dawson KJ, Haase K, Van Loo P, et al. Universal Patterns of Selection in Cancer and Somatic Tissues [Internet]. Cell. 2017. page 1029–41.e21. Available from: http://dx.doi.org/10.1016/j.cell.2017.09.042

4. Cannataro VL, Gaffney SG, Townsend JP. Effect Sizes of Somatic Mutations in Cancer. J Natl Cancer Inst. 2018;110:1171–7.

5. Li L, Yue P, Song Q, Yen T-T, Asaka S, Wang T-L, et al. Genome-wide mutation analysis in precancerous lesions of endometrial carcinoma. J Pathol. 2021;253:119–28.

6. Martincorena I, Roshan A, Gerstung M, Ellis P, Van Loo P, McLaren S, et al. Tumor evolution. High burden and pervasive positive selection of somatic mutations in normal human skin. Science. 2015;348:880–6.

7. Yokoyama A, Kakiuchi N, Yoshizato T, Nannya Y, Suzuki H, Takeuchi Y, et al. Age-related remodelling of oesophageal epithelia by mutated cancer drivers. Nature. 2019;565:312–7.

8. Martincorena I, Fowler JC, Wabik A, Lawson ARJ, Abascal F, Hall MWJ, et al. Somatic mutant clones colonize the human esophagus with age [Internet]. Science. 2018. page 911–7. Available from: http://dx.doi.org/10.1126/science.aau3879

9. Moore L, Leongamornlert D, Coorens THH, Sanders MA, Ellis P, Dentro SC, et al. The mutational landscape of normal human endometrial epithelium. Nature. 2020;580:640–6.

10. Suda K, Nakaoka H, Yoshihara K, Ishiguro T, Tamura R, Mori Y, et al. Clonal Expansion and Diversification of Cancer-Associated Mutations in Endometriosis and Normal Endometrium. Cell Rep. 2018;24:1777–89.

11. Lee-Six H, Olafsson S, Ellis P, Osborne RJ, Sanders MA, Moore L, et al. The landscape of somatic mutation in normal colorectal epithelial cells. Nature. 2019;574:532–7.

12. Lawrence MS, Stojanov P, Polak P, Kryukov GV, Cibulskis K, Sivachenko A, et al. Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature. 2013;499:214–8.

13. Martincorena I. Somatic mutation and clonal expansions in human tissues. Genome Med. 2019;11:35.

14. Jaiswal S, Fontanillas P, Flannick J, Manning A, Grauman PV, Mar BG, et al. Age-related clonal hematopoiesis associated with adverse outcomes. N Engl J Med. 2014;371:2488–98.

15. Genovese G, Kähler AK, Handsaker RE, Lindberg J, Rose SA, Bakhoum SF, et al. Clonal hematopoiesis and blood-cancer risk inferred from blood DNA sequence. N Engl J Med.

2014;371:2477–87.

16. Weinreich DM, Lan Y, Wylie CS, Heckendorn RB. Should evolutionary geneticists worry about higher-order epistasis? Curr Opin Genet Dev. 2013;23:700–7.

17. Majewski IJ. On fitness: how do mutations shape the biology of cancer? Biochem Soc Trans. 2019;47:559–69.

18. Rahal Z, Sinjab A, Wistuba II, Kadara H. Game of clones: Battles in the field of carcinogenesis. Pharmacol Ther. 2022;237:108251.

19. Dujon AM, Aktipis A, Alix-Panabières C, Amend SR, Boddy AM, Brown JS, et al. Identifying key questions in the ecology and evolution of cancer. Evol Appl. 2021;14:877–92.

20. Adashek JJ, Kato S, Lippman SM, Kurzrock R. The paradox of cancer genes in non-malignant conditions: implications for precision medicine. Genome Med. 2020;12:16.

21. Higa KC, DeGregori J. Decoy fitness peaks, tumor suppression, and aging. Aging Cell. 2019. page e12938.

22. Hall MWJ, Shorthouse D, Alcraft R, Jones PH, Hall BA. Mutations observed in somatic evolution reveal underlying gene mechanisms. Commun Biol. 2023;6:753.

23. Colom B, Alcolea MP, Piedrafita G, Hall MWJ, Wabik A, Dentro SC, et al. Spatial competition shapes the dynamic mutational landscape of normal esophageal epithelium. Nat Genet. 2020;52:604–14.

24. Liao G, Tang J, Bai J. Early development of esophageal squamous cell cancer: Stem cells, cellular origins and early clone evolution. Cancer Lett. 2023;555:216047.

25. Abby E, Dentro SC, Hall MWJ, Fowler JC, Ong SH, Sood R, et al. Notch1 mutations drive clonal expansion in normal esophageal epithelium but impair tumor growth. Nat Genet. 2023;55:232–45.

26. Yuan W, Liu Z, Wang Y, Liu M, Pan Y, Lei W, et al. Clonal evolution of esophageal squamous cell carcinoma from normal mucosa to primary tumor and metastases. Carcinogenesis. Oxford Academic; 2019;40:1445–51.

27. Liu X, Zhang M, Ying S, Zhang C, Lin R, Zheng J, et al. Genetic Alterations in Esophageal Tissues From Squamous Dysplasia to Carcinoma. Gastroenterology. 2017;153:166–77.

28. Li M, Zhang Z, Wang Q, Yi Y, Li B. Integrated cohort of esophageal squamous cell cancer reveals genomic features underlying clinical characteristics. Nat Commun. Nature Publishing Group; 2022;13:1–15.

29. Gao Y-B, Chen Z-L, Li J-G, Hu X-D, Shi X-J, Sun Z-M, et al. Genetic landscape of esophageal squamous cell carcinoma. Nat Genet. 2014;46:1097–102.

30. Sawada G, Niida A, Hirata H, Komatsu H, Uchi R, Shimamura T, et al. An Integrative Analysis to Identify Driver Genes in Esophageal Squamous Cell Carcinoma. PLoS One. 2015;10:e0139808.

31. Zhang Y, Zhang L, Wang R, Wang B, Hua P, Li J. LncRNA Erbb4-IR promotes esophageal

squamous cell carcinoma (ESCC) by downregulating miR-145. J Cell Biochem. 2019;120:17566–72.

32. Mangalaparthi KK, Patel K, Khan AA, Manoharan M, Karunakaran C, Murugan S, et al. Mutational Landscape of Esophageal Squamous Cell Carcinoma in an Indian Cohort. Front Oncol. 2020;10:1457.

33. Wang L, Jia Y-M, Zuo J, Wang Y-D, Fan Z-S, Feng L, et al. Gene mutations of esophageal squamous cell carcinoma based on next-generation sequencing. Chin Med J . 2021;134:708–15.

34. Zou B, Guo D, Kong P, Wang Y, Cheng X, Cui Y. Integrative Genomic Analyses of 1,145 Patient Samples Reveal New Biomarkers in Esophageal Squamous Cell Carcinoma. Front Mol Biosci. 2021;8:792779.

35. Cancer Genome Atlas Research Network, Analysis Working Group: Asan University, BC Cancer Agency, Brigham and Women's Hospital, Broad Institute, Brown University, et al. Integrated genomic characterization of oesophageal carcinoma. Nature. 2017;541:169–75.

36. Gillespie JH. Population Genetics: A Concise Guide. JHU Press; 2004.

37. Mandell JD, Cannataro VL, Townsend JP. Estimation of Neutral Mutation Rates and Quantification of Somatic Variant Selection Using cancereffectsizeR. Cancer Res. 2023;83:500–5.

38. Chakravarty D, Gao J, Phillips SM, Kundra R, Zhang H, Wang J, et al. OncoKB: A Precision Oncology Knowledge Base. JCO Precis Oncol [Internet]. 2017;2017. Available from: http://dx.doi.org/10.1200/PO.17.00011

39. Alfaro-Murillo JA, Townsend JP. Pairwise and higher-order epistatic effects among somatic cancer mutations across oncogenesis [Internet]. bioRxiv. 2022 [cited 2022 Jul 6]. page 2022.01.20.477132. Available from: https://www.biorxiv.org/content/10.1101/2022.01.20.477132v1.abstract

40. Cannataro VL, Mandell JD, Townsend JP. Attribution of Cancer Origins to Endogenous, Exogenous, and Preventable Mutational Processes. Mol Biol Evol [Internet]. 2022;39. Available from: http://dx.doi.org/10.1093/molbev/msac084

41. Zhong F, Zhu T, Pan X, Zhang Y, Yang H, Wang X, et al. Comprehensive genomic profiling of high-grade serous ovarian carcinoma from Chinese patients identifies co-occurring mutations in the Ras/Raf pathway with TP53. Cancer Med. 2019;8:3928–35.

42. Remy E, Rebouissou S, Chaouiya C, Zinovyev A, Radvanyi F, Calzone L. A Modeling Approach to Explain Mutually Exclusive and Co-Occurring Genetic Alterations in Bladder Tumorigenesis. Cancer Res. 2015;75:4042–52.

43. Kostiou V, Hall MWJ, Jones PH, Hall BA. Simulations reveal that different responses to cell crowding determine the expansion of and mutant clones in squamous epithelia. J R Soc Interface. 2021;18:20210607.

44. Kennedy MC, Lowe SW. Mutant p53: it's not all one and the same. Cell Death Differ. 2022;29:983–7.

45. Armenia J, Wankowicz SAM, Liu D, Gao J, Kundra R, Reznik E, et al. The long tail of oncogenic drivers in prostate cancer. Nat Genet [Internet]. 2018; Available from:

http://dx.doi.org/10.1038/s41588-018-0078-z

46. Cannataro VL, Townsend JP. Wagging the long tail of drivers of prostate cancer. PLoS Genet. Public Library of Science; 2019;15:e1007820.

47. Liang J, Wang Y, Cai L, Liu J, Yan J, Chen X, et al. Comparative Genomic Analysis Reveals Genetic Variations in Multiple Primary Esophageal Squamous Cell Carcinoma of Chinese Population. Front Oncol. 2022;12:868301.