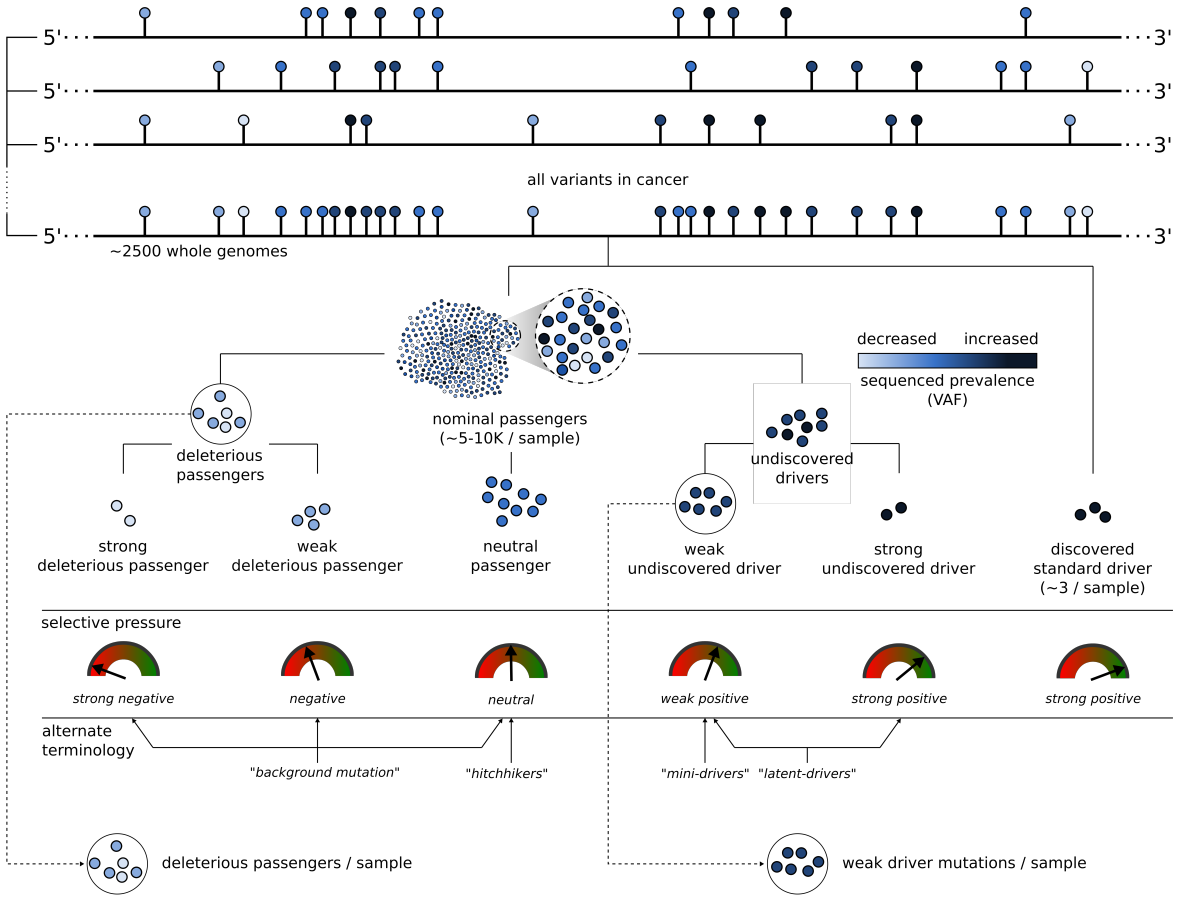
SC response presentation

Terminologies in the paper



1. Majority of these terminologies are borrowed from literature and we cite references accordingly.

2. All SNVs were classified as drivers or “nominal passengers”.
   - List of driver SNVs were obtained from the patient-centric paper which includes driver mutations curated from literature and PCAWG driver discovery effort

3. In this work, nominal passengers with higher functional impact were termed as “ impactful nominal passengers”.
   - Alternatively we can use impactful non-driver variants to avoid any confusion.

Additional analyses beyond background model & additive effects

1. Look at overall burdening of different genomic elements (TF binding)
   - Not using any randomization.
   - Provide mechanistic insight.

2. Simple enrichment analysis of nominal passengers in key genes
   - Uses sanger simulation

3. Correlation between nominal passengers' molecular impact and their prevalence.
   - Not using any randomization

4. Burdening analysis of early and late subclones
   - Not using any randomization

Sensitivity analysis for additive variance model

Sensitivity Analysis I – additive variance analysis of double randomization set

- To address potential issue related to overfitting of the model
- Created two randomized sets(using sanger simulation) for six cancer cohort.
- Calculated additive variance

Breast	CNS	Kidney	Ovary	Pancreas	Prostate
1e-6%	2e-6%	2e-6%	2e-6%	1e-6%	1e-6%

Almost 0% additive variance on double random set suggest no overfitting of the model.

Sensitivity analysis for additive variance model

Sensitivity Analysis II – evaluating the influence of window size

1. Performed two distinct set of randomization by varying permutation window length (50kb & 100kb).

2. Calculated the additive variance for different cohorts.

3. Similar additive variance scores, though slightly higher values for the 100kb window length randomization.
   - 100kb model picks up slightly higher additive variance probably due to influence of covariates.

	50kb model	100kb model
Breast	0.5105	0.5147
CNS	0.1991	0.2014
Kidney	0.5072	0.6409
Ovary	0.6485	0.6426
Prostate	0.3296	0.3326
Pan-cancer	**0.4390**	**0.4664**

Custom randomization to generate covariate corrected randomized dataset

- Following conditions were applied to generate new randomized dataset

1. Tri-nucleotide context of permuted location was same as original mutation

2. Permuted mutation lie on the same chromosome as original mutation

3. genome was divided into 10kn non-overlapping bins. Average of multiple co-variates were computed for each bin.

4. Clustering approach was applied to identify relevant bins, where a given mutation can be permuted satisfying condition1 and condition2.

5. Following co-variate corrections were considered for generating the background mutations:
   a) Replication timing
   b) Chromatin accessibility
   c) GC content
   d) Penta-nucleotide context for Liver and Melanoma cohorts

Updated randomization based additive variance analysis

	Drivers	Coding	Promoters	Non-coding	Total
Breast	0.5132	0.0080	0.0113	0.0423	0.5749
CNS	0.1738	0.0057	0.0117	0	0.1912
Kidney	0.426	0.0225	0.0042	0.0325	0.4852
Liver	0.5256	0.0158	0.0518	0.1229	0.7163
Ovary	0.5622	0.0073	0.0514	0.1661	0.7870
Pancreas	0.9312	0.0391	0	0	0.9703
Prostate	0.248	0	0	0.1524	0.4004
Skin	0.6043	0.0219	0	0	0.6262
Pan-cancer	**0.4981**	**0.0150**	**0.0163**	**0.0645**	**0.5939**

We updated our random effects model to quantify contribution of each category of variants in total additive variance.

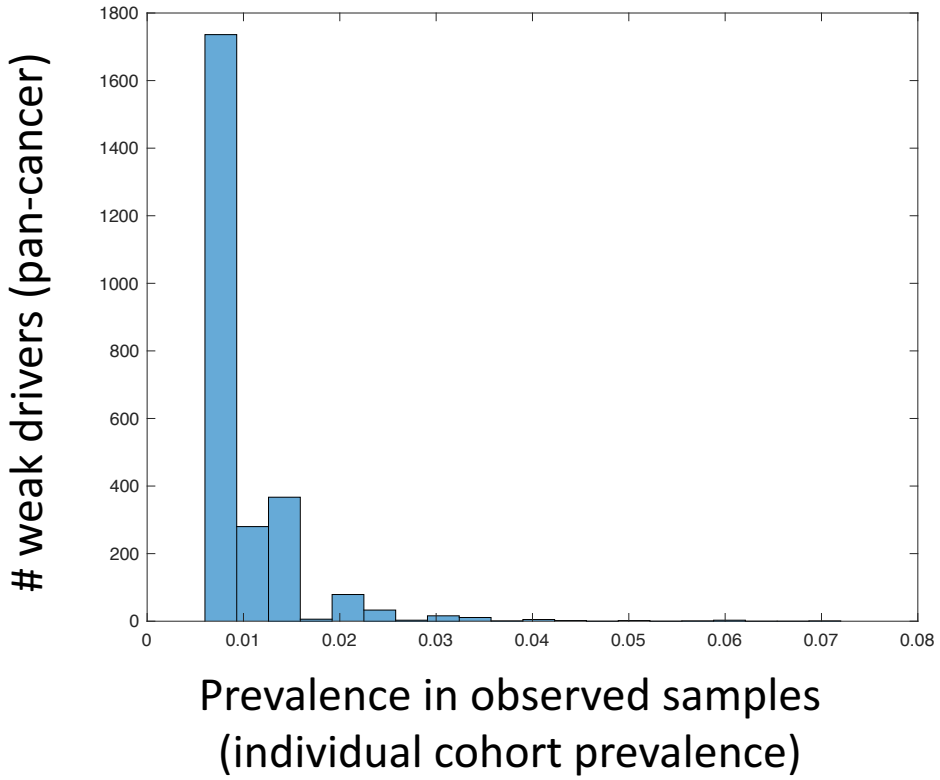
Extra Slides

Nested model:

$$y_j = \mu + z_j^{\mathrm{dr}} u^{\mathrm{dr}} + \sum_i z_{ij}^{\mathrm{cd}} u_i^{\mathrm{cd}} + \sum_i z_{ij}^{\mathrm{prm}} u_i^{\mathrm{prm}} + \sum_i z_{ij}^{\mathrm{ncd}} u_i^{\mathrm{ncd}} + e_j$$

$$u_i^{\mathrm{dr}} \sim N(0, \sigma_{\mathrm{dr}}^2) \qquad u_i^{\mathrm{prm}} \sim N(0, \sigma_{\mathrm{prm}}^2)$$
$$u_i^{\mathrm{cov}} \sim N(0, \sigma_{\mathrm{cov}}^2) \qquad u_i^{\mathrm{ncd}} \sim N(0, \sigma_{\mathrm{ncd}}^2)$$
$$u_i^{\mathrm{cd}} \sim N(0, \sigma_{\mathrm{cd}}^2) \qquad e_j \sim N(0, \sigma_{\mathrm{e}}^2)$$

Prevalence analysis



Intersections with PCAWG candidate drivers:

Element level: FDR<0.25, **174** (p=7e-24); FDR<0.1, **115** (p=3e-17); 0.1<FDR<0.25, **59** (p=4e-8)
CDS only: FDR<0.25, **52** (1e-5); FDR<0.1, **33** (4e-3); 0.1<FDR<0.25, **19** (p=2e-4)

Estimation updates with full filtering

(Lower estimates)

	Breast	CNS	Kidney	Liver	Ovary	Pancreas	Prostate	Skin	Cross-cohort av.
WD	5.1031	0.2695	4.2606	2.0356	1.6364	3.2554	1.5808	5.022	2.8954
DP removed	5.4485	1.3121	4.2042	4.2492	1.8818	7.013	3.6414	19.5275	5.9097
DP retained	1.9871	0.1507	1.4177	0.6718	0.1733	1.1749	1.2109	4.2134	1.375