Paper E updates

07/24/2017

PCAWG driver and functional interpretation group

Aim and deliverables for the functional impact paper

Decipher overall functional burdening in cancer genomes in the PCAWG project.

- Avg cancer has ~5 drivers & thousands of nominal passenger mutations. What is the cumulative effect of nominal passengers in progression of cancer ?

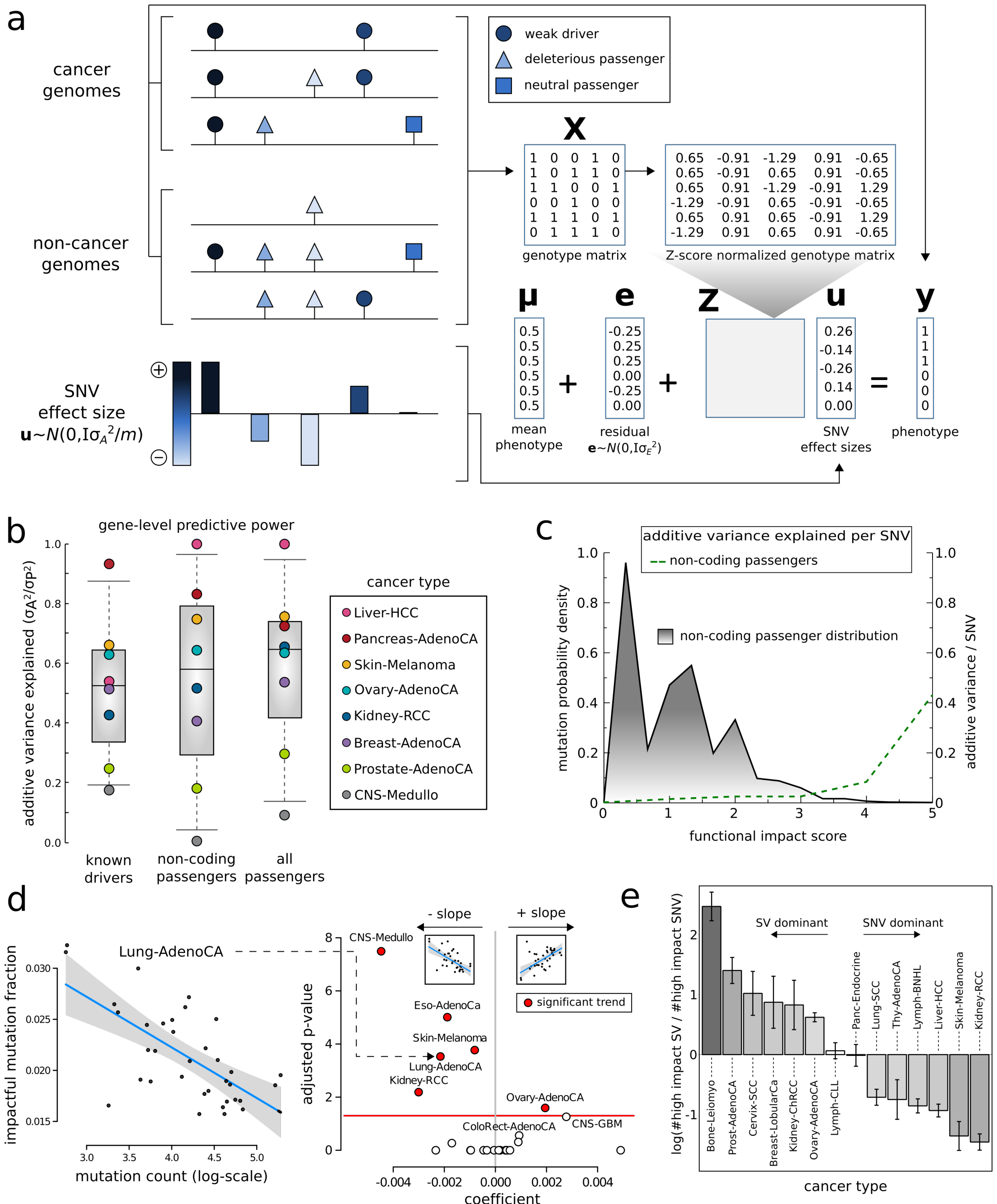
- Look at additive effect of nominal passengers and their overall functional impact in different PCAWG cohorts.
  - ➢ This work will provide **comprehensive functional annotations across all of pcawg** (FunSeq & aloft score)
  - ➢ Framework to evaluate structural variation impact score

- Mutational burden observed in various genomic sub-systems (coding & noncoding) in different PCAWG cohorts.
  - ➢ Correlation of passenger burdening with downstream gene expression changes

- Decipher the the differential passenger burdening in various cohorts (how it relates to mechanism)
  - ➢ Relate to different Signature, sub-clonality & other clinical information

- Role of weak drivers and deleterious passengers in cancer progression
  - ➢ Conservative estimate of the weak driver and deleterious passengers frequencies

Additive variance and overall molecular functional impact



Presence of few key variants and large numbers of passengers is analogous to genome-wide association studies (GWAS) that implicated a handful of variants that significantly influence complex traits.

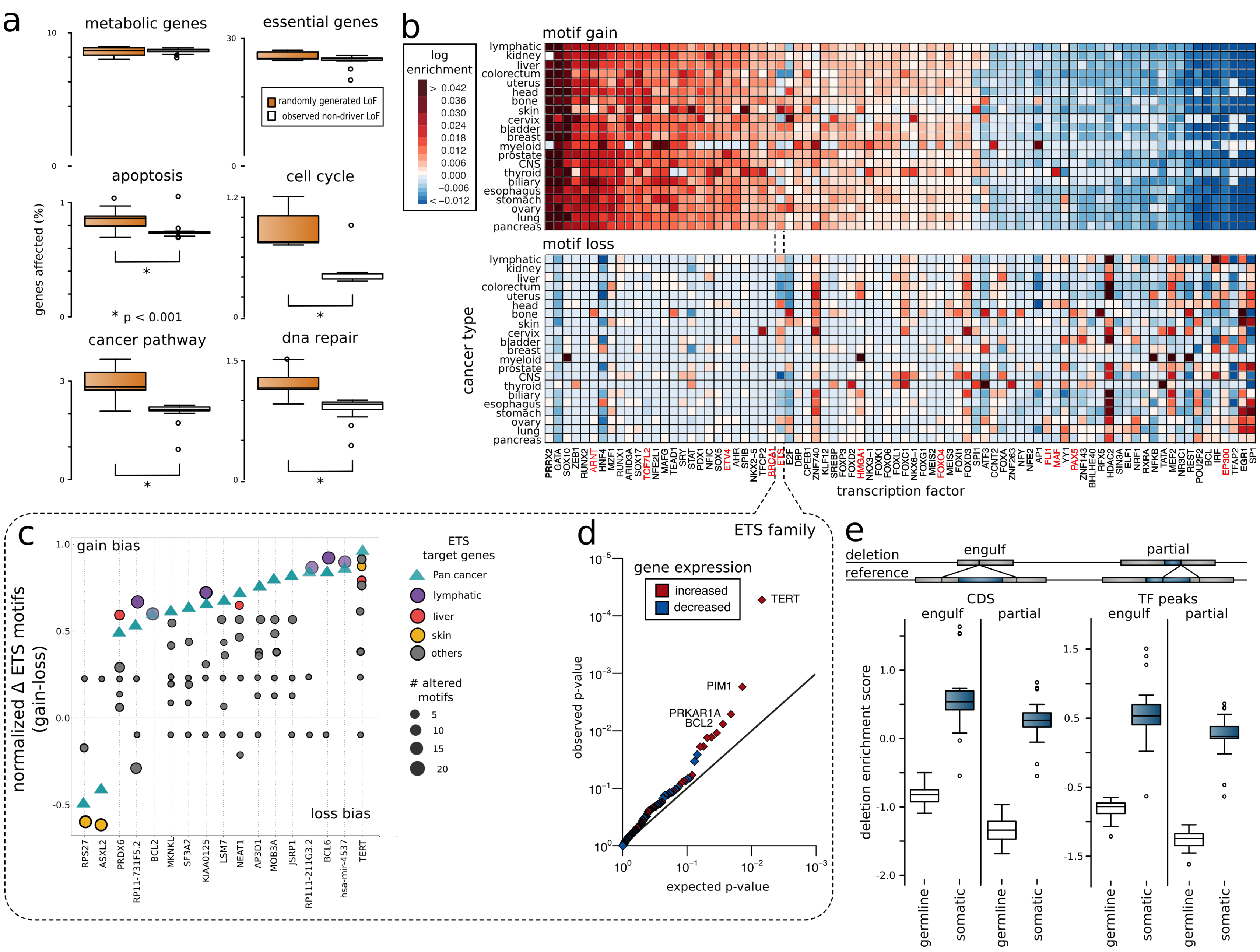
We apply an additive effects model to quantify the relative size of these aggregated effects of nominal passengers in relation to known drivers.

Nominal passengers predicted a large fraction of the variance (64.5% median), a significant fraction of which remained even when coding variants were excluded (57.9%)

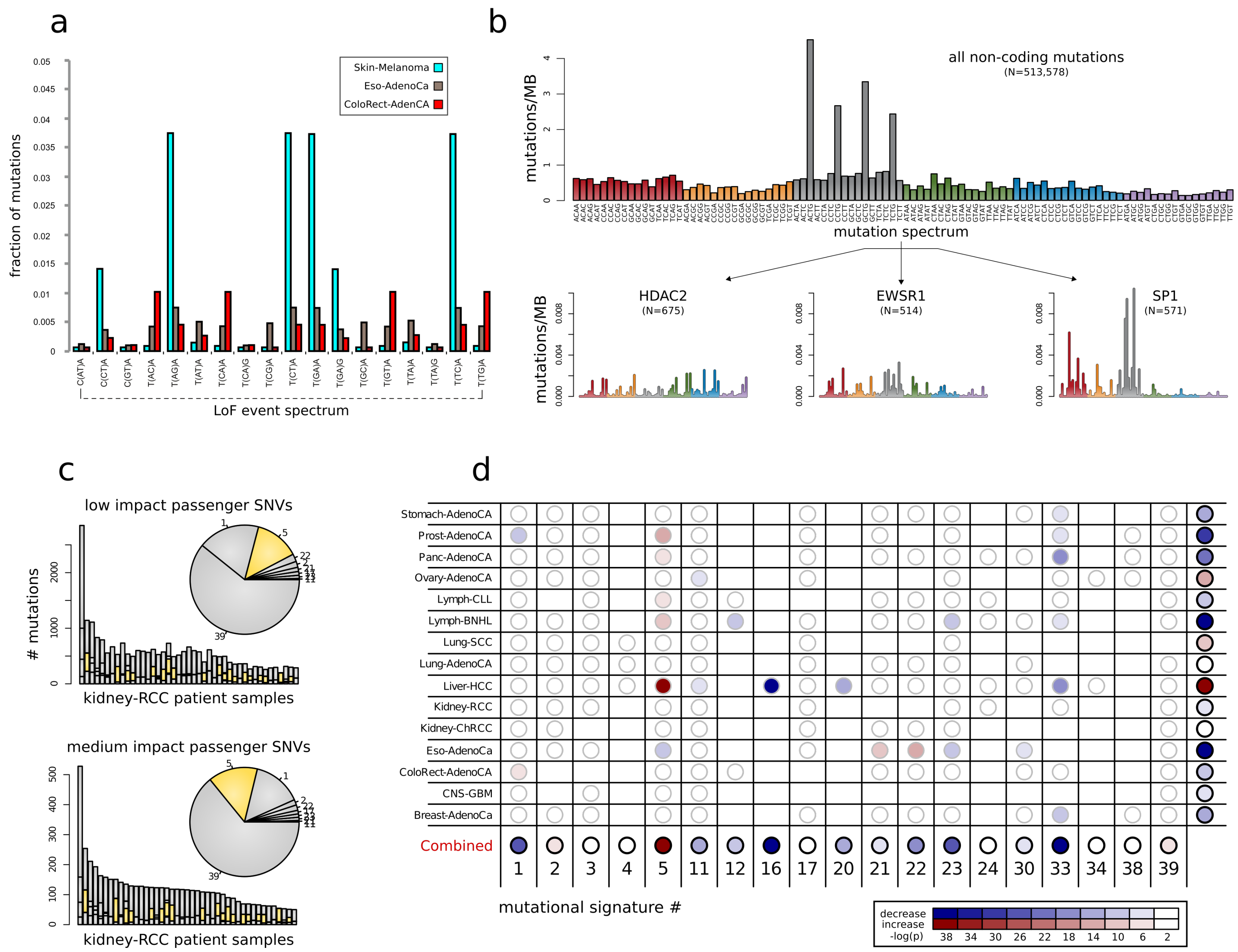
Large number of SNVs leads to decrease in the fraction of impactful passengers.

Certain cancer subtypes tend to harbor large number of high impact SVs, while others were more burdened with high impact SNVs

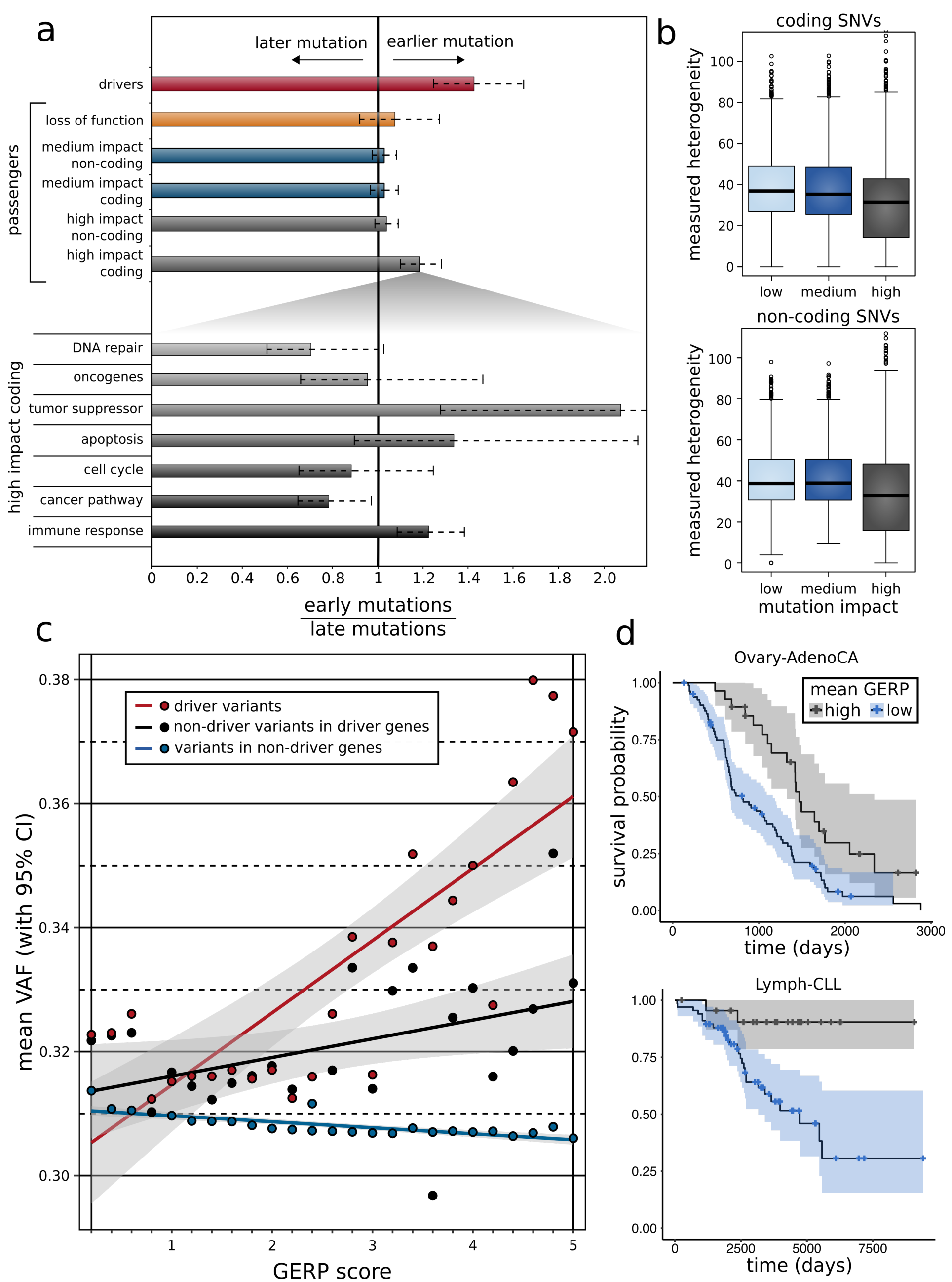
Burdening of different genomic sub-systems in cancer



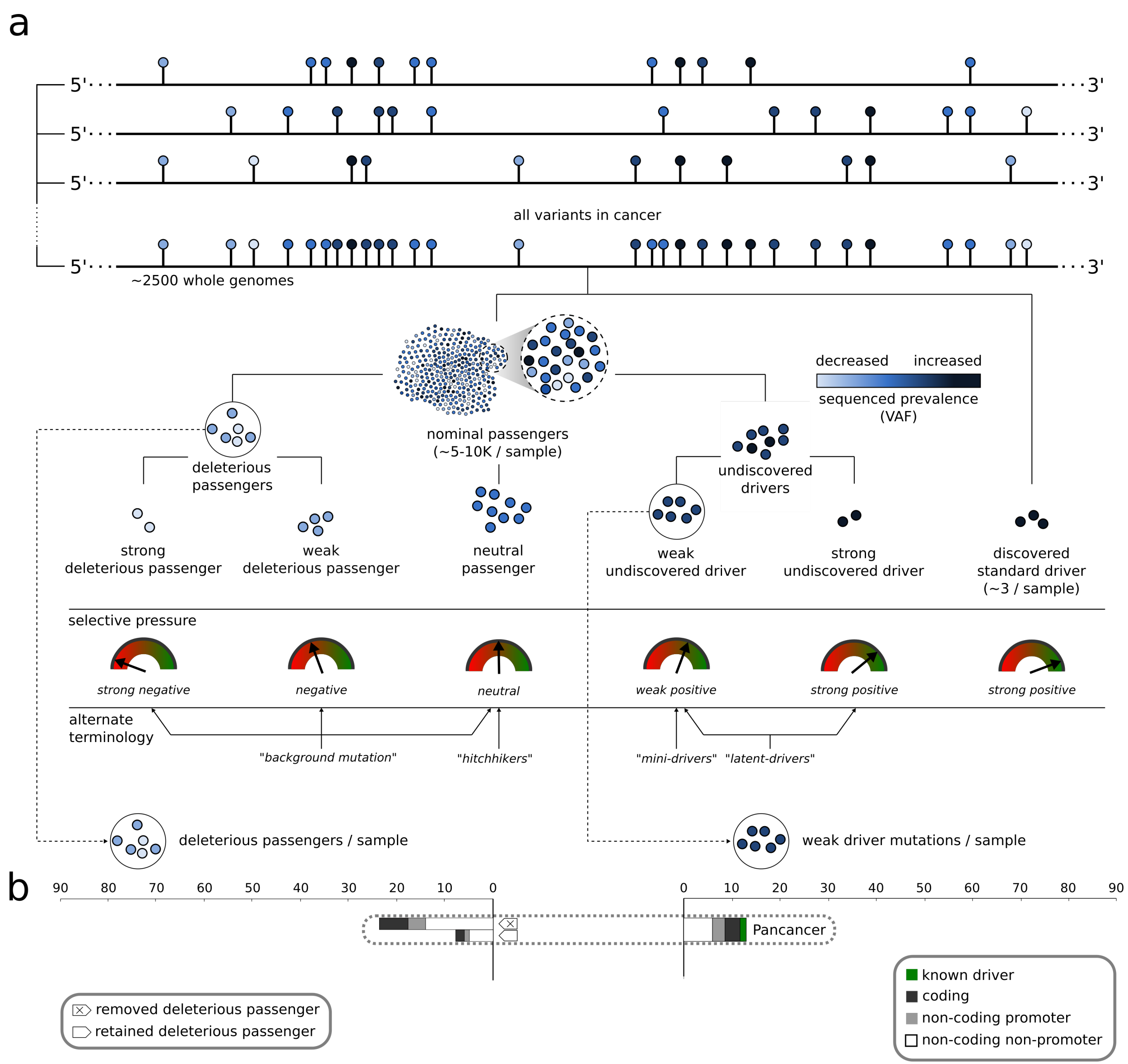
Signature landscape of nominal passengers



Subclonal architecture of nominal passengers



Categorizing nominal passengers and estimated frequencies



Conclusion and Discussion

Functional impact distribution has a multi-modal characteristic with significant number of nominal passengers with intermediate functional impact.

Various functional elements in a cancer genome are differentially burdened with distinct functional impact.

Differential functional burdening between early and late subclones in a cancer. An overall enrichment and depletion of nominal passenger variants among TSGs and oncogenes, respectively.

Additive effects model shows that aggregating nominal passengers in a cancer genome can provide significant predictive ability to distinguish cancer phenotype from non-cancerous ones.

Extra Slides

Linear model with random effects

- Model for the effect of an individual SNP on a phenotype

$$y_j = \mu + x_{ij}a_i + e_j$$

where: y=phenotype; x_ij is the ’genetic dosage’ of the i’th SNP in individual j, taking values {0,1}; a_i is the fixed effect size of SNP i, and e_j is the environmental effect

$$e_j \sim N(0, \sigma_e^2)$$

- Extension to model the combined effects of multiple SNPs

$$y_j = \mu + g_j + e_j \text{ and } g_j = \sum_{i=1}^{m} z_{ij}u_i$$

where: z_ij is a ‘normalized genetic dosage’, i.e. the z-score of x_ij; u_i is the effect size of SNP i treated as a random variable; g_j is the combined effect of all SNPs for individual j

$$\mathbf{u} \sim N(\mathbf{0}, \mathbf{I}\sigma_u^2)$$

$$g_j \sim N(0, \sigma_g^2 = m\sigma_u^2)$$

Linear model with random effects

- The variance-covariance matrix of y can be expressed using matrix notation

$$y_j = \mu + g_j + e_j \text{ and } g_j = \sum_{i=1}^{m} z_{ij} u_i$$ [from previous slide]

$$\mathrm{var}(\mathbf{y}) = \mathbf{ZZ}'\sigma_u^2 + \mathbf{I}\sigma_e^2 = \frac{\mathbf{ZZ}'\sigma_g^2}{m} + \mathbf{I}\sigma_e^2 = \mathbf{G}\underbrace{\sigma_g^2}_{\text{additive variance}} + \mathbf{I}\sigma_e^2$$

$$z_{ij} = (x_{ij} - p_i)/(p_i(1 - p_i))$$

Model with gene-level priors

$$y_j = \mu + g_j + e_j \text{ and } g_j = \sum_{i=1}^{m} z_{ij} u_i$$

For simple gene-level prior:

$$z_{ij} = \frac{x_{ij} - (1/n_\gamma) M_{\gamma_i}}{V_{\gamma_i}}$$

For gene-level prior with normalized effects:

$$z_{ij} = \frac{(x_{ij}/T_j) - (1/n_\gamma) \tilde{M}_{\gamma_i}}{\tilde{V}_{\gamma_i}} \qquad T_j = \sum_{i=1}^{m} x_{ij}$$

where for both:

$$i, j \in \gamma \Rightarrow u_i = u_j = u_\gamma, u_\gamma \sim N(0, \sigma_u^2)$$

$$y_j = \mu + \sum_{i=1}^{m} x_{ij} \beta_i + e_j$$

$$y_j = \mu + \frac{1}{T_j} \sum_{i=1}^{m} x_{ij} \beta_i + e_j$$

Model with gene-level priors

Normalizing constants:

$$M_{\gamma_i} = \frac{1}{N}\sum\nolimits_{j,k\in\gamma_i} x_{kj}$$ (mean burden)

$$V_{\gamma_i} = \frac{1}{N}\sum\nolimits_{j}\left(\left(\sum\nolimits_{k\in\gamma_i} x_{kj}\right) - M_{\gamma_i}\right)$$ (variance of burden)

$$\tilde{M}_{\gamma_i} = \frac{1}{N}\sum\nolimits_{j,k\in\gamma_i}(x_{kj}/T_j)$$

$$\tilde{V}_{\gamma_i} = \frac{1}{N}\sum\nolimits_{j}\left(\left(\sum\nolimits_{k\in\gamma_i}(x_{kj}/T_j)\right) - \tilde{M}_{\gamma_i}\right)$$

Model with partitioned variance (for nested hypothesis testing)

$$\mathrm{var}(\mathbf{y}) = \mathbf{G}_1\sigma_{A_1}^2 + \mathbf{G}_2\sigma_{A_2}^2 + \mathbf{G}_3\sigma_{A_3}^2 + \mathbf{I}\sigma_e^2$$

$\sigma_{A_1}^1, \sigma_{A_2}^2, \sigma_{A_3}^3$: Additive variances for known drivers, non-coding nominal passengers and coding nominal passengers respectively

$\mathbf{G}_1, \mathbf{G}_2, \mathbf{G}_3$: Genetic relationship matrix for known drivers, non-coding nominal passengers and coding nominal passengers respectively

Model for known drivers

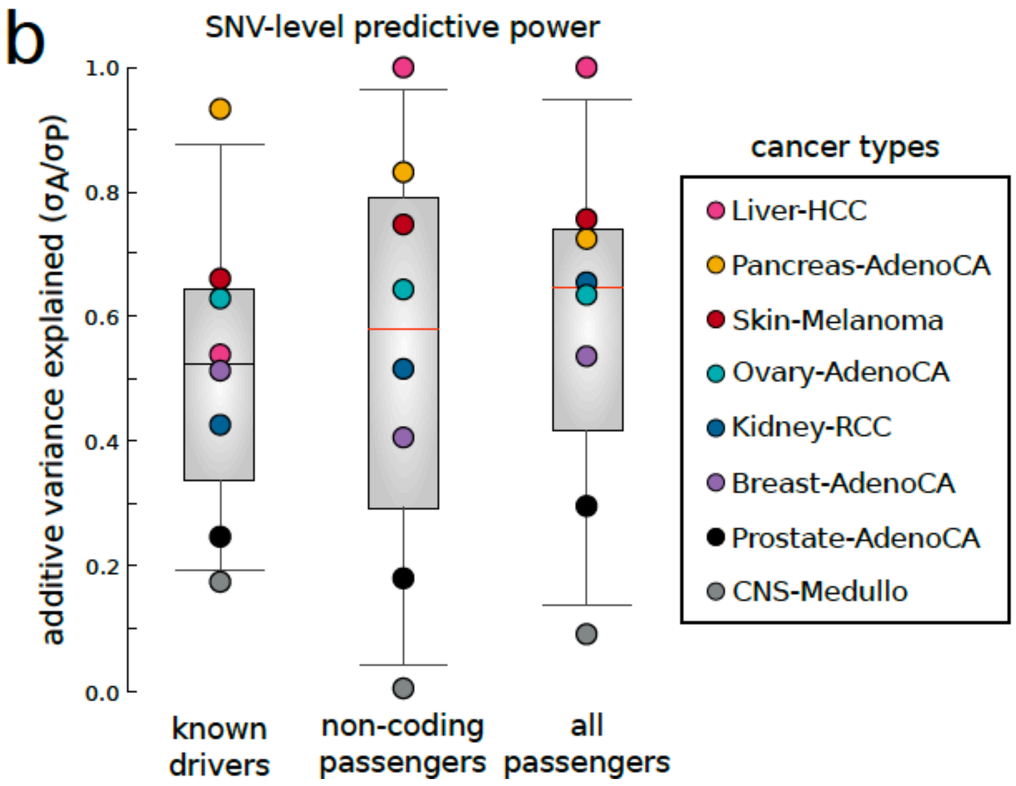
Bayes optimal predictor:

$$y_j = \hat{y}_j + e_j$$

$$\hat{y} = \begin{cases} 1 \text{ if } d_j > 0 \\ p \text{ otherwise} \end{cases}$$

$$p = P(y = 1 | d = 0) = \frac{\overline{D}_{obs}}{\overline{D}_{obs} + \overline{D}_{null}}$$

where $\overline{D}_{obs}$ is the number of observed samples not containing a driver, and $\overline{D}_{null} = N_{null}$ is the number of null samples.



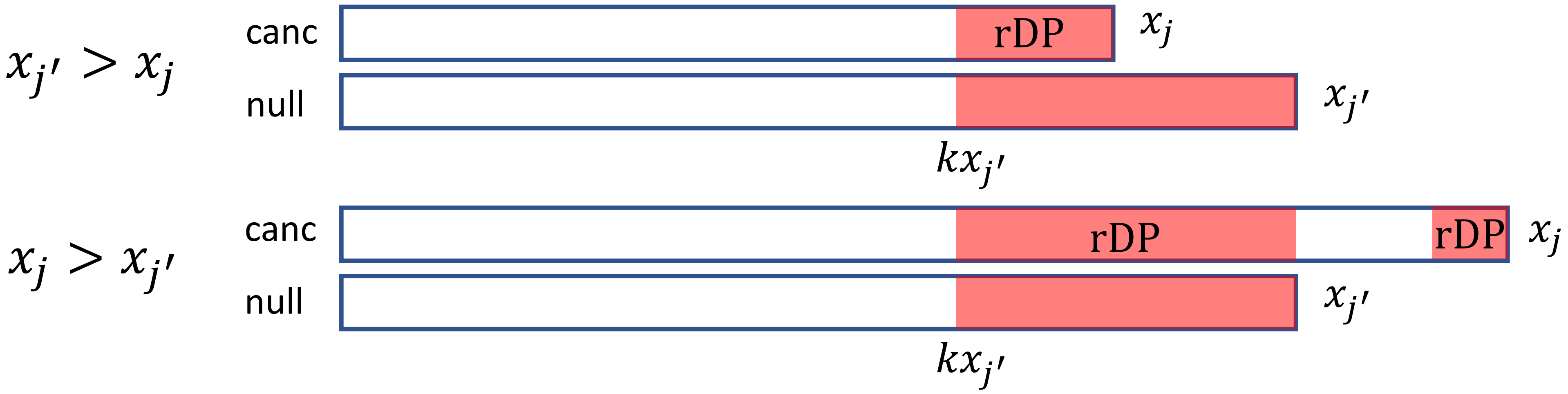
Medians:
0.525, 0.579, 0.6453

Estimating the number of retained DPs

For a putative DP gene i, $k = \frac{\mathrm{E}_{canc}[x_i]}{\mathrm{E}_{null}[x_i]} < 1$

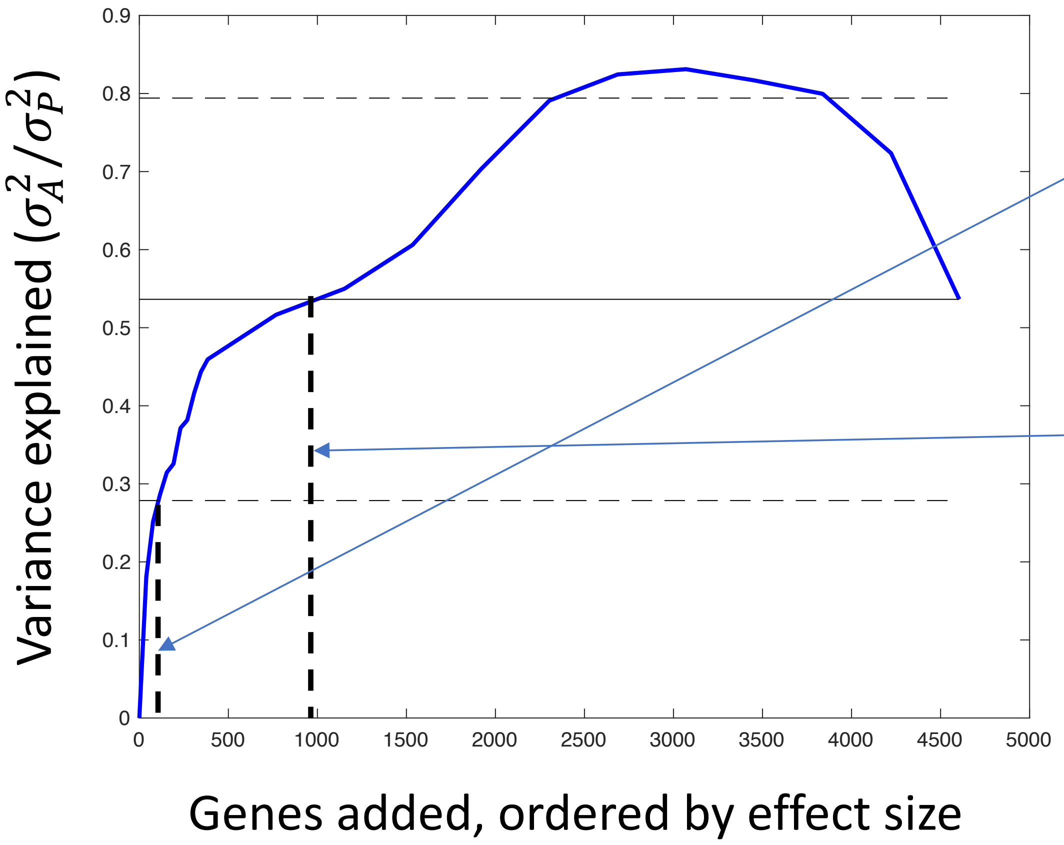
For cancer sample j and matching null sample j', estimate # retained DPs as:

$$\mathrm{rDP}_j = \begin{cases} \max(x_j - k \cdot x_{j'}, 0) \text{ if } x_{j'} > x_j \\ k \cdot x_j \qquad \text{otherwise} \end{cases}$$



Lower-bound for # WDs and DPs

Breast Adeno-carcinoma



Lower estimates:
116 genes total
47 WD genes (1.9 SNVs per tumor)
69 DP genes (4.0 SNVs removed per tumor)

Higher estimates:
968 genes total
423 WD genes (8.4 SNVs per tumor)
545 DPs genes (12.6 SNVs removed per tumor)

Estimate for WD SNVs per tumor: $\sum_{g \in \mathrm{WD}} \mathrm{mean}_{canc}(x_g) - \mathrm{mean}_{null}(x_g)$

Estimate for DP SNVs removed per tumor: $\sum_{g \in \mathrm{DP}} \mathrm{mean}_{null}(x_g) - \mathrm{mean}_{canc}(x_g)$

SNV-level additive variances & q-values

	Non-coding	Coding
Breast	46.5% q = 3.6e-9	46.4% q = 4e-9
CNS	18.4% q = 2e-4	19.2% q = 1e-4
Kidney	56.6% q = 1.5e-11	56.7% q = 1.6e-11
Liver	74.6% q = ~0	74.8% q = ~0
Ovary	67.8% q = 4.4e-12	67.7% q = 4.8e-12
Pancreas	90.1% q = ~0	90% q = ~0
Prostate	66.7% q = 1.8e-10	66.7% q = 1.8e-10
Skin	21.6% q = 9e-4	21.6% q = 9e-4
	FDR < 0.001	FDR < 0.001

Gene-level additive variances & q-values

	Non-coding	Coding
Breast	40.6% q = 0.094	53.6% q= 0.0013
CNS	0.3% q = 0.47	9.2% q = 0.097
Kidney	51.7% q = 0.026	65.6% q = 6.4e-9
Liver	99.9% q = 2e-10	100% q = 1.9e-9
Ovary	64.2% q = 0.12	63.5% q = 5.4e-6
Pancreas	83.2% q = 4.1e-5	72.5% q = 0.0012
Prostate	18% q = 0.27	29.6% q = 1.4e-5
Skin	74.8% q = 1e-4	75.5% q = 7.2e-5
	FDR < 0.1 (Except CNS,Ov,Prost)	FDR < 0.1