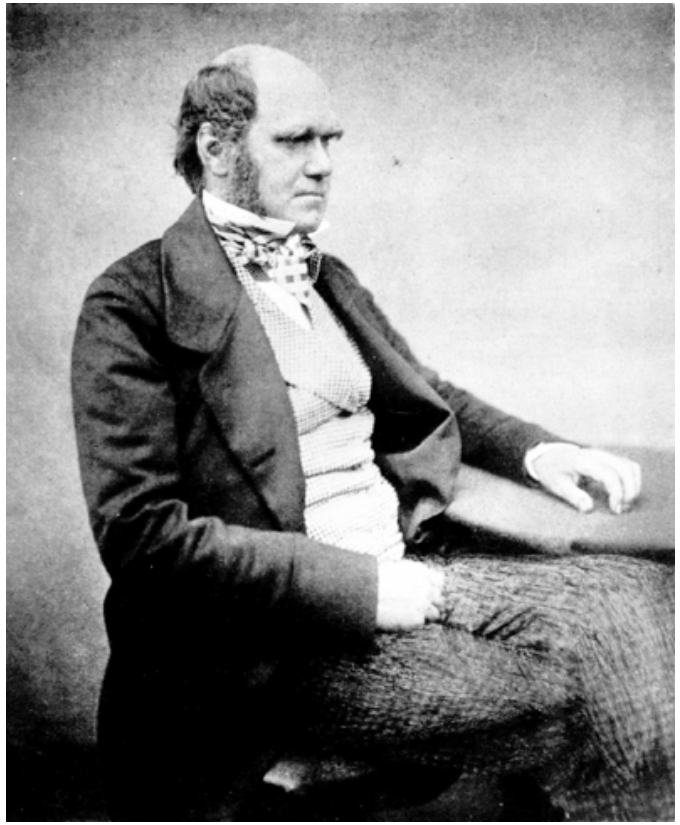
Probabilistic approaches to inference of mutation rate and selection in cancer

Donate Weghorn

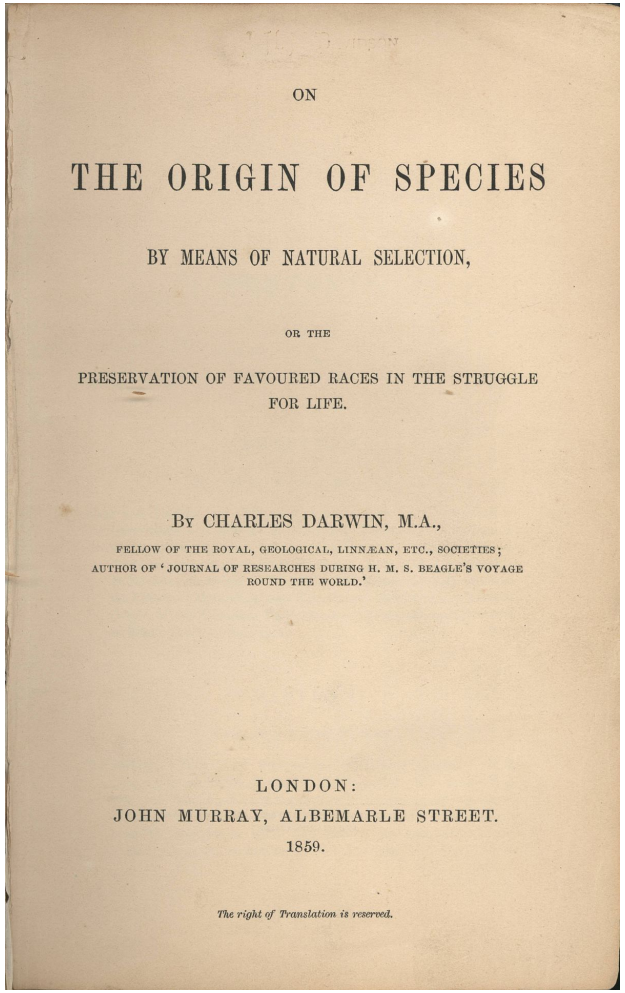
Centre for Genomic Regulation, Barcelona

25 October 2019

Darwinian evolution



Charles Darwin (1809-1882)



ON

THE ORIGIN OF SPECIES

BY MEANS OF NATURAL SELECTION,

OR THE

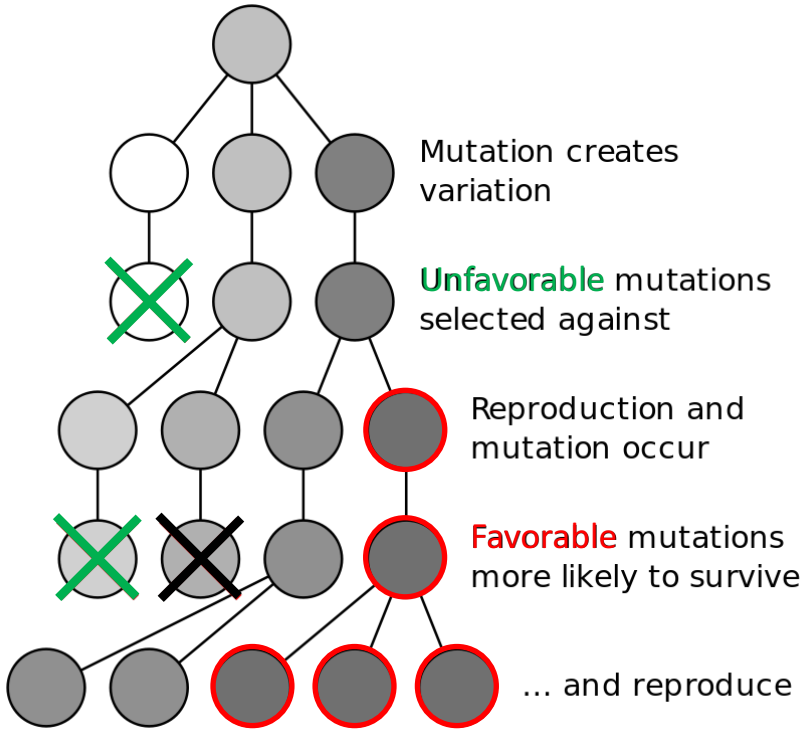
PRESERVATION OF FAVOURED RACES IN THE STRUGGLE FOR LIFE.

By CHARLES DARWIN, M.A.,

FELLOW OF THE ROYAL, GEOLOGICAL, LINNÆAN, ETC., SOCIETIES;
AUTHOR OF 'JOURNAL OF RESEARCHES DURING H. M. S. BEAGLE'S VOYAGE ROUND THE WORLD.'

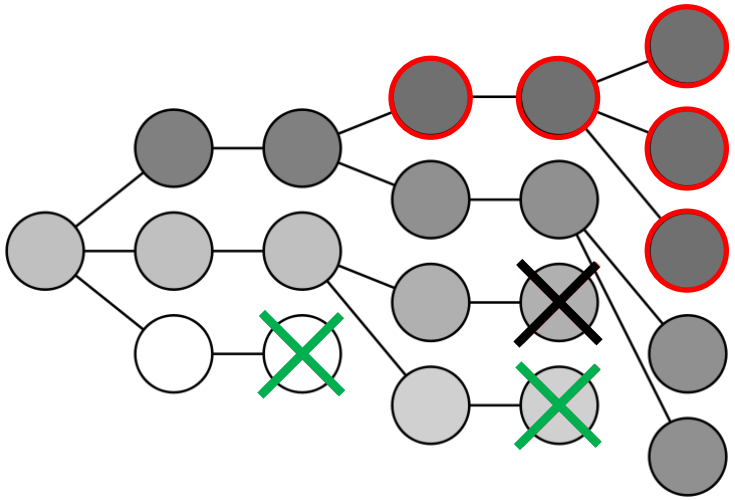
LONDON:
JOHN MURRAY, ALBEMARLE STREET.
1859.

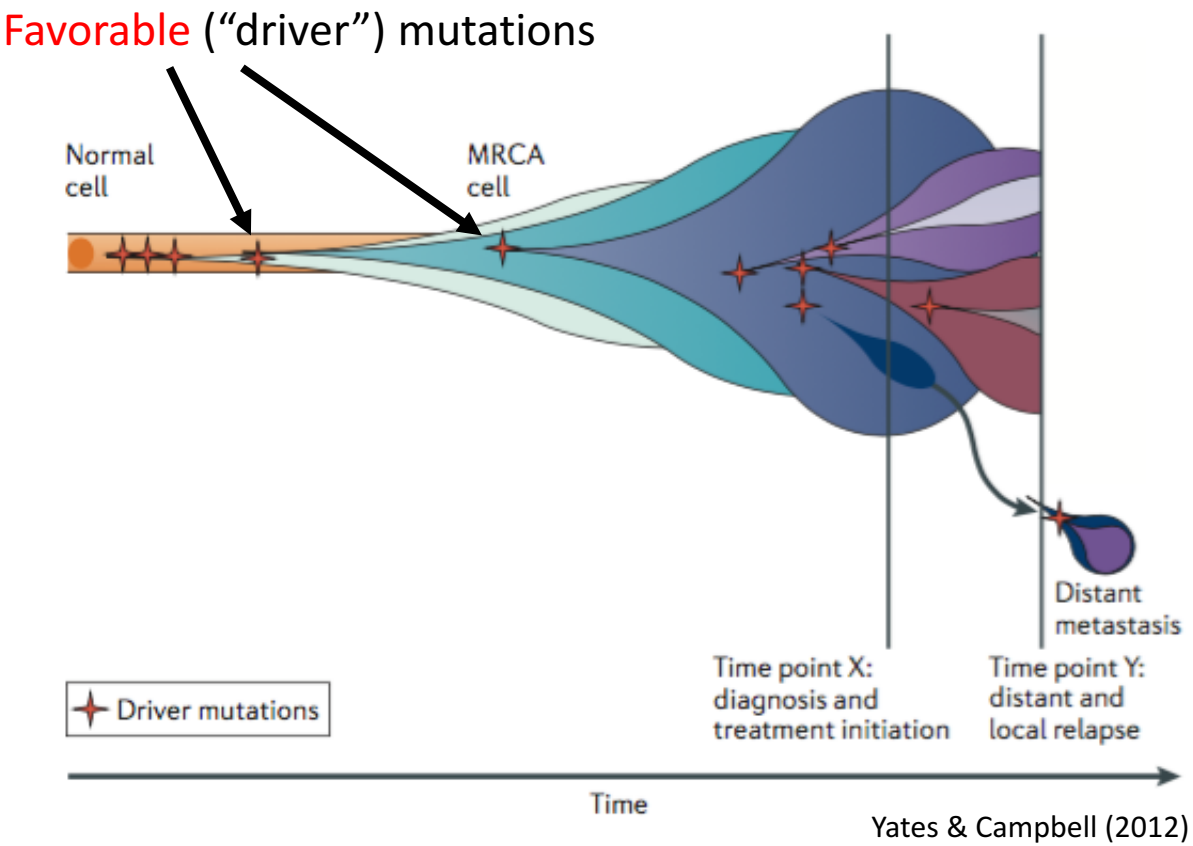
The right of Translation is reserved.



Unfavorable mutations are under negative selection
Favorable mutations are under positive selection

Role of positive selection in tumorigenesis





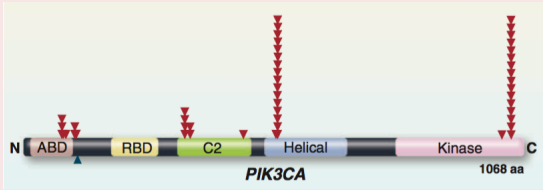
Yates & Campbell (2012)

→ From the perspective of the tumor, "drivers" are beneficial mutations, because they confer a growth advantage.

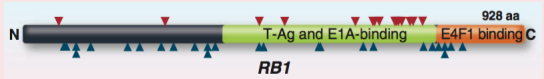
What are driver mutations?

Positive selection on driver mutation(s):

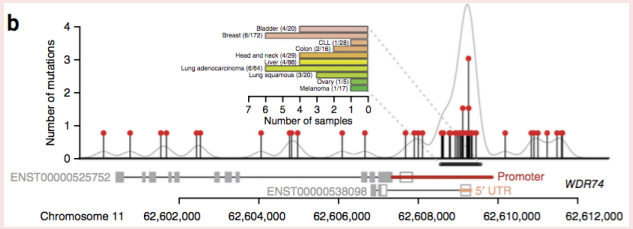
1. **Oncogenes** → Protein gain of function



2. **Tumor suppressors** → Protein loss of function

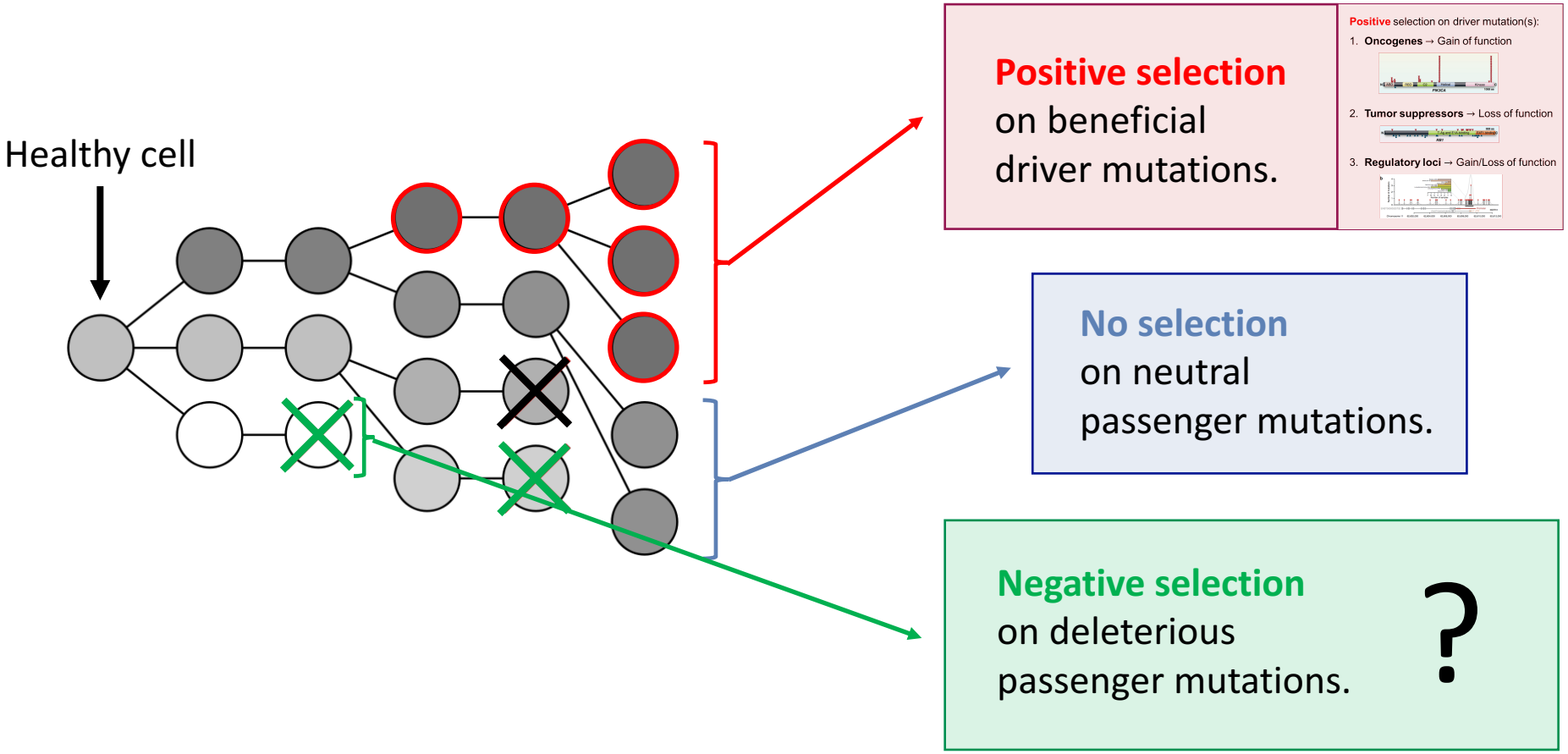


3. **Regulatory loci** → Gene expression changes



Vogelstein et al. (2013), Weinhold et al. (2014)

Selection during tumorigenesis



How much negative selection do we expect to see in cancer?

Bacteria: 20-90% of coding mutations deleterious

Mutation rate dynamics in a bacterial population reflect tension between adaptation and genetic load

Sébastien Wielgoss[a,b,1,2], Jeffrey E. Barrick[c,d,1], Olivier Tenaillon[e,f,1], Michael J. Wiser[d,g], W. James Dittmar[h], Stéphane Cruveiller[i,j], Béatrice Chane-Woon-Ming[i,j], Claudine Médigue[i,j], Richard E. Lenski[d,g,h,3], and Dominique Schneider[a,b,3]

(*dS*) substitutions. We observed a *dN*/*dS* ratio of ~0.80 for all mutators, implying that ~20% of all nonsynonymous mutations were deleterious. Confounding factors that can,

Humans: >70% of coding mutations deleterious

Initial sequence of the chimpanzee genome and comparison with the human genome

The Chimpanzee Sequencing and Analysis Consortium*

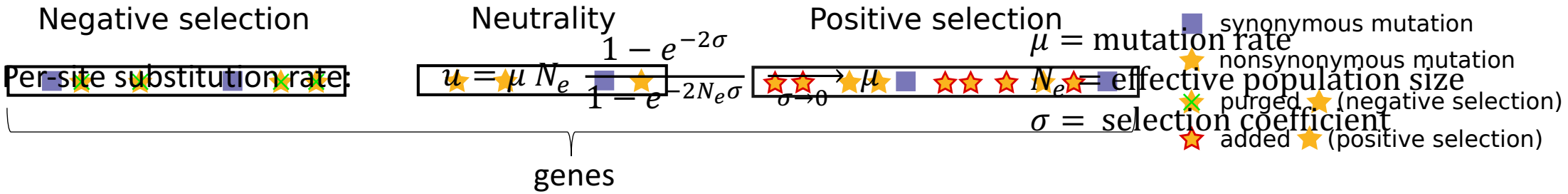
are selectively neutral, the results imply that 77% of amino acid alterations in hominid genes are sufficiently deleterious as to be eliminated by

Instead, negative selection in cancer is highly elusive, mainly because of adaptation.

Detecting selection on coding sequence in cancer

Simplest signature of genic selection

Increase (**positive**) or decrease (**negative**) in the number of observed nonsynonymous mutations relative to the **neutral expectation**.



Problem

What is the **neutral expectation** for the number of **nonsynonymous** mutations?

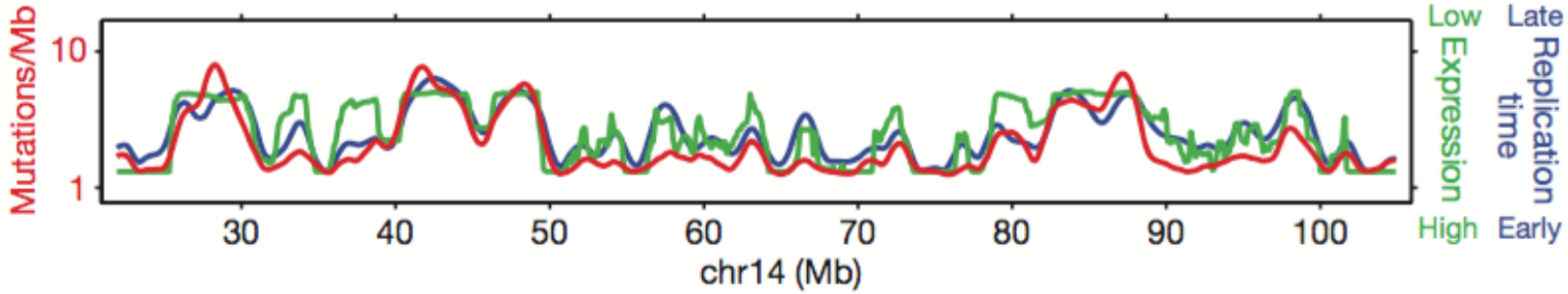
→ Selection inference in cancer is completely confounded by spatial *mutation rate heterogeneity*.

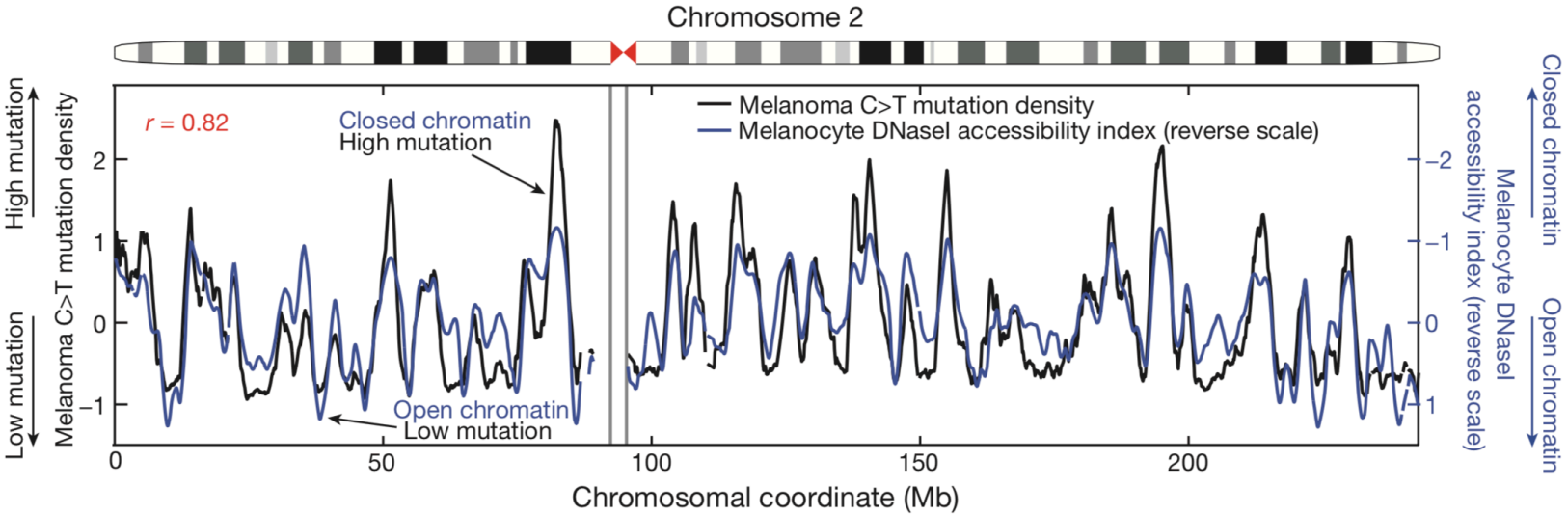
Dependence of mutation rate on external covariates

Local somatic mutation density is influenced by:

(1) expression level
(2) DNA replication time
(3) chromatin state
(4) sequence context.

⟹ The *genome-wide average* of the mutation density will lead to an incorrect estimate of the expected mutation density at the *local* gene level.





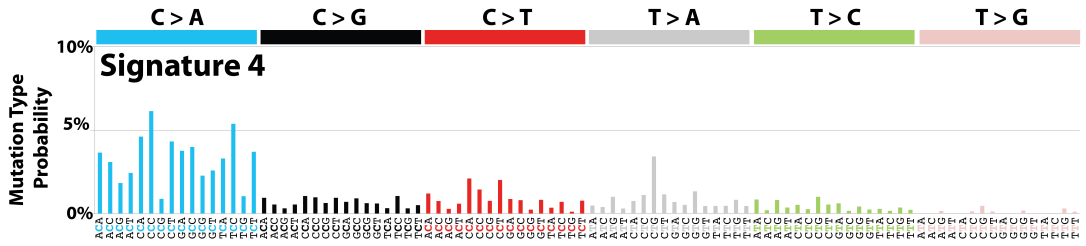
Lawrence et al. (2013); Polak et al. (2015)

Dependence of mutation rate on sequence context

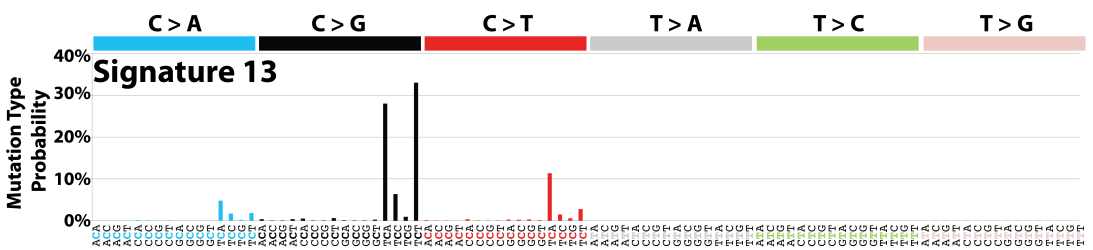
Local somatic mutation density is influenced by:

(1) expression level
(2) DNA replication time
(3) chromatin state
(4) sequence context, e.g. ATCGC**C**ATCGC > ATCGC**A**ATCGC

- Mutation probability depends on extended sequence context.
- Typically, one accounts for the trinucleotide context of a mutation.
- Some cancer types are subject to mutational processes that have a larger context dependence (e.g. pentameric or heptameric).

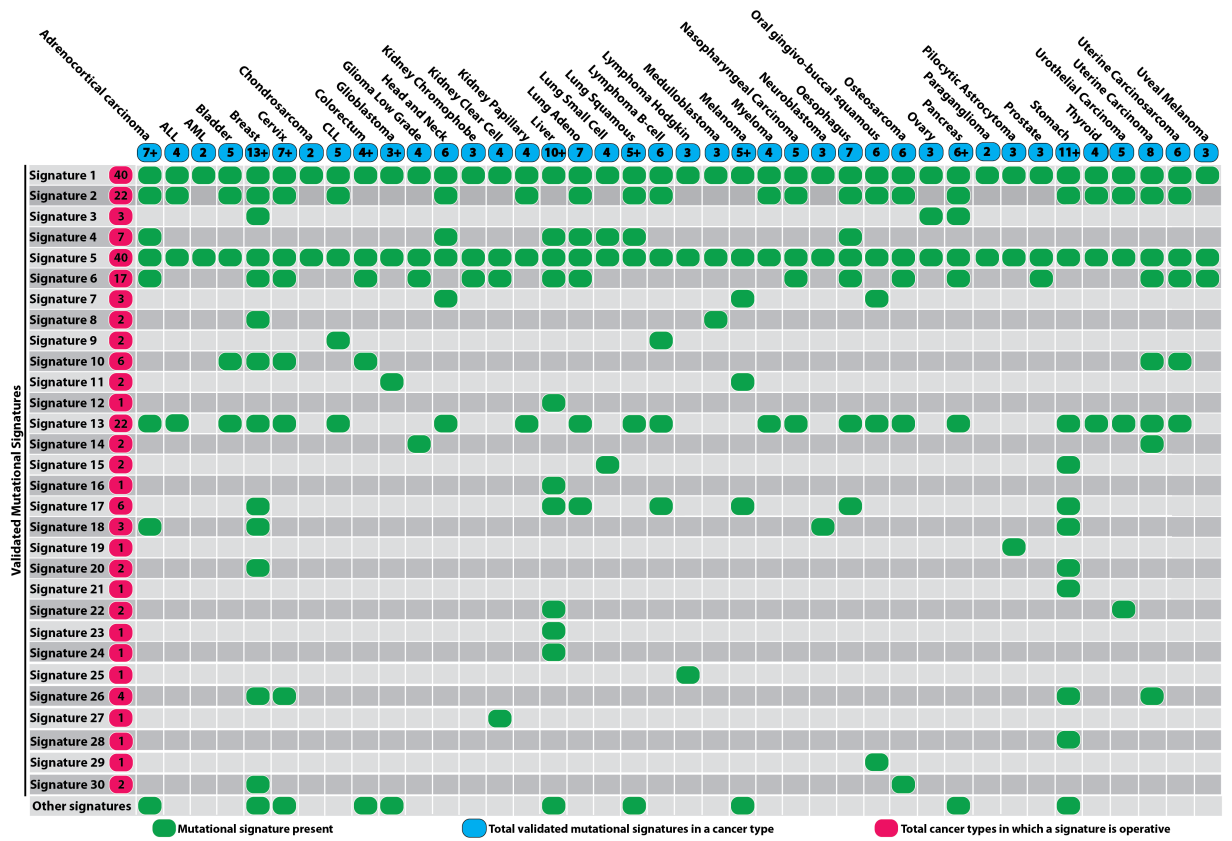


Signature 4 is associated with smoking and likely due to tobacco mutagens.



Signature 13 has been attributed to activity of the AID/APOBEC family of cytidine deaminases.

Dependence of mutation rate on sequence context



COSMIC database

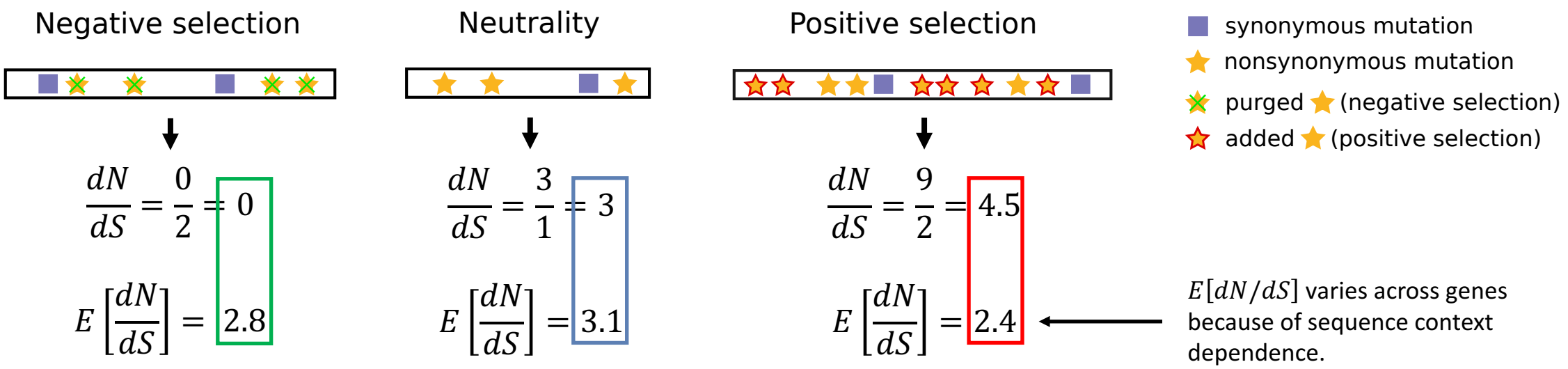
What to do about mutation rate heterogeneity?

Estimating local mutation rate in cancer

Classical approach I

Gauge mutation density by observed putatively neutrally evolving mutations, e.g. synonymous or nearby intronic mutations

- Very noisy.
- Not applicable for genes with no synonymous or intronic mutations.

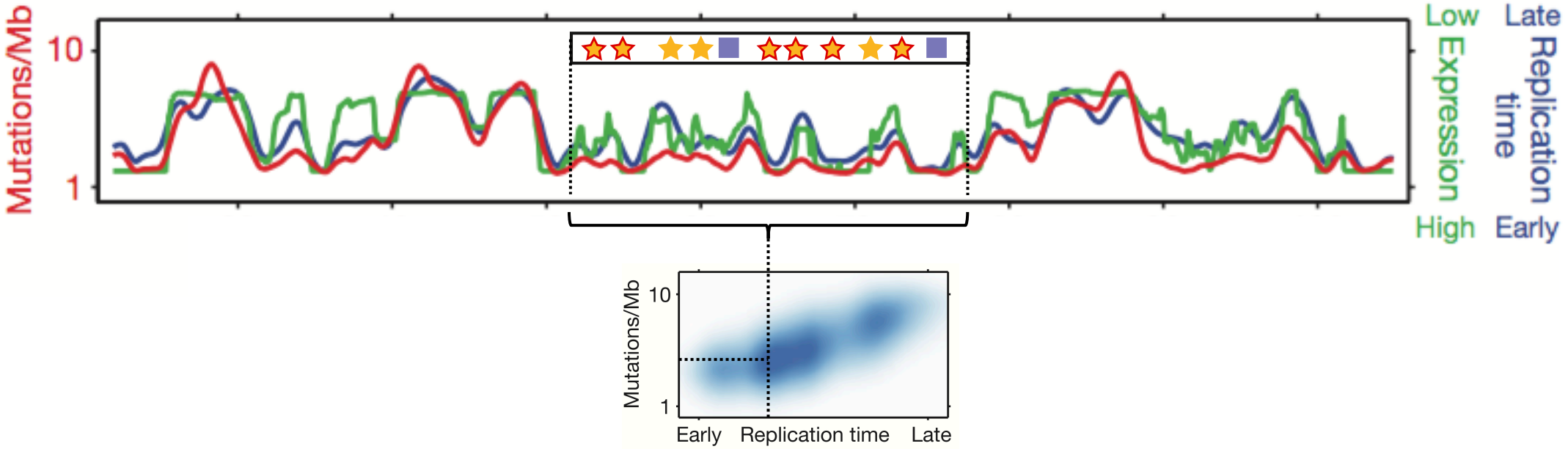


Estimating local mutation rate in cancer

Classical approach II

Cluster loci by covariates of mutation rate (replication time, expression, chromatin, ...), e.g. state-of-the-art MutSigCV (Lawrence et al., 2013)

- Need meta-data about mutation rate covariates as input.
- Cannot model unknown factors affecting mutation rate.
- Delivers error-prone point estimates of mutation rate.



Estimating local mutation rate in cancer

Probabilistic approach

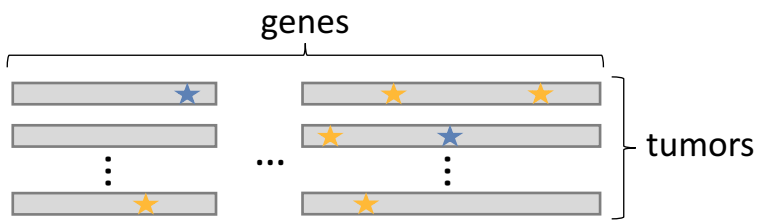
Model the distribution of mutation density across all genes by fitting the observed distribution of per-gene synonymous mutation count s:

$$P(s;\hat{\boldsymbol{\theta}}) = \int \mathrm{d}\lambda_s \, P(s|\lambda_s) \, P(\lambda_s;\hat{\boldsymbol{\theta}})$$

Parameters $\boldsymbol{\theta}$ estimated from ML $\quad \mathrm{Pois}(\lambda_s) \quad \lambda_s \propto$ local mutation rate

Data set and data structure

- TumorPortal (Lawrence et al., *Nature,* 2014).
- 17 cancer types (4478 patients).
- Curated nonsynonymous and synonymous mutation calls.



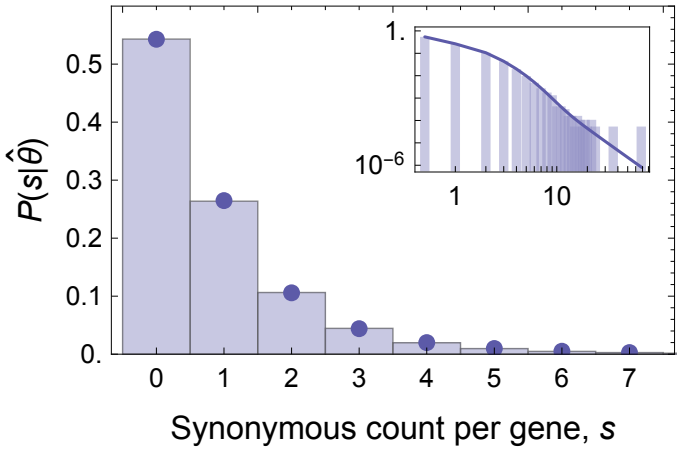
$$s = \sum_{t=1}^{|\{\text{tumors}\}|} s_t$$

$$x = \sum_{t=1}^{|\{\text{tumors}\}|} x_t$$

Expected and observed genome-wide mutation count distributions

Results: Head-neck squamous cell carcinoma

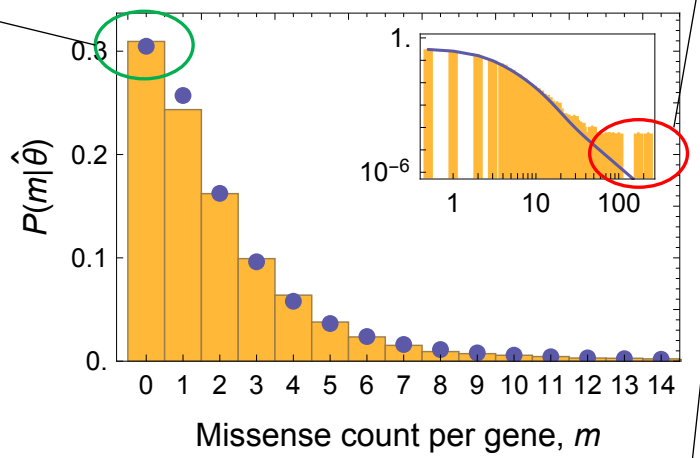
$$P(s;\hat{\boldsymbol{\theta}}) = \int \mathrm{d}\lambda_s \, P(s|\lambda_s) \, P(\lambda_s;\hat{\boldsymbol{\theta}})$$

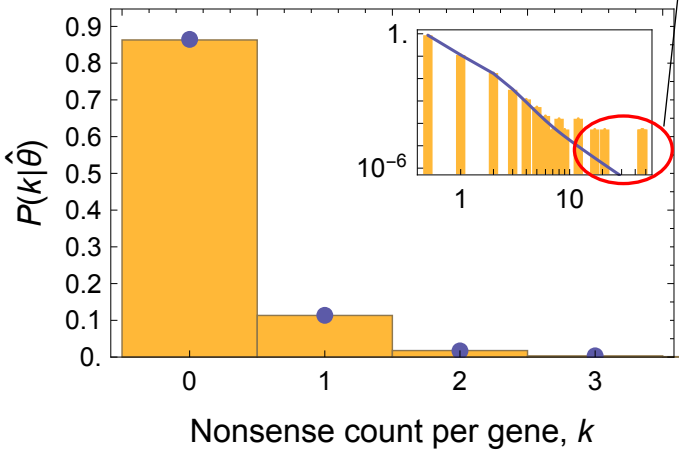


For missense and nonsense mutations: rescale λ_s with target size ratio, accounting for cancer type-specific mutational signature.

Negative selection signal

Positive selection signal





Observed distribution

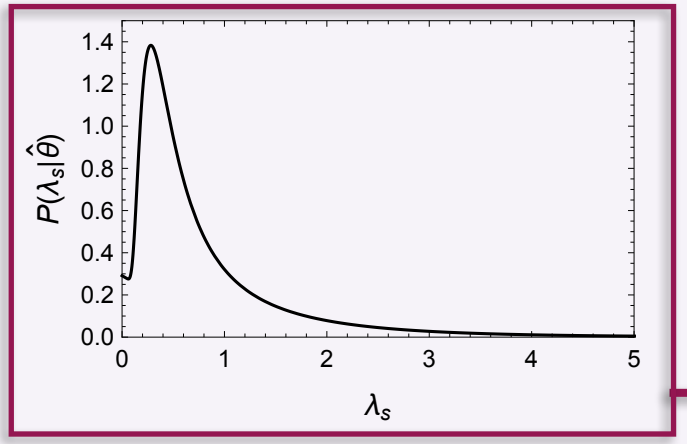
Expected neutral distribution

Weghorn & Sunyaev, *Nat. Genet.* (2017)

Per-gene inference of selective growth (dis)advantage

Inferred distribution of per-gene expected values for s:

$$P(s;\widehat{\boldsymbol{\theta}}) = \int \mathrm{d}\lambda_s \, P(s|\lambda_s) \, P(\lambda_s;\widehat{\boldsymbol{\theta}})$$



Conditional probability of x nonsynonymous mutations on gene, given s synonymous mutations and target size ratio r_x:

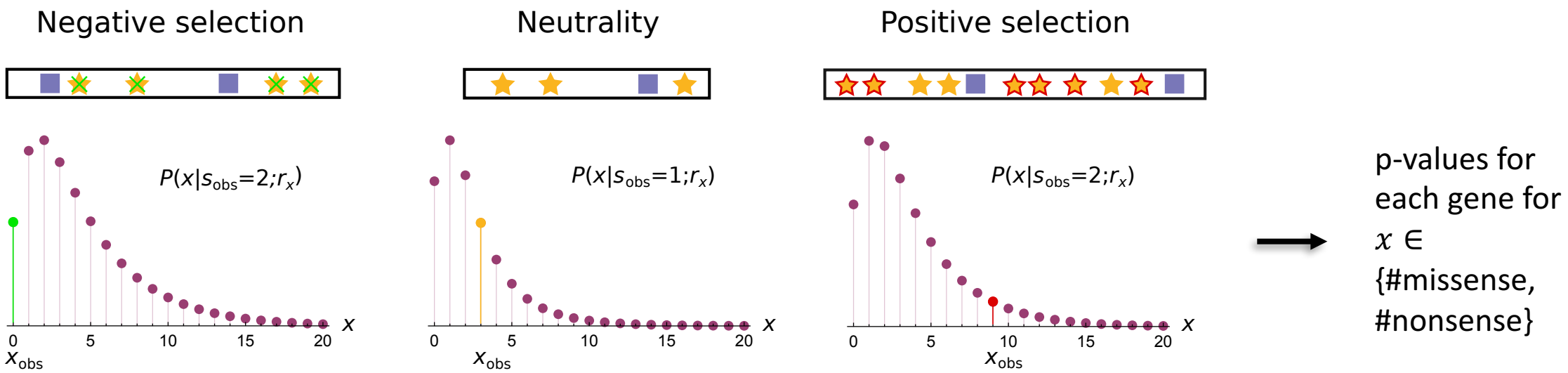
$$P(x|s;r_x,\widehat{\boldsymbol{\theta}}) = \int \mathrm{d}\lambda_s \boxed{P(x|\lambda_s r_x)}\boxed{P(\lambda_s|s;\widehat{\boldsymbol{\theta}})}$$

$$\mathrm{Pois}(\lambda_s r_x) \quad \frac{P(s|\lambda_s)P(\lambda_s;\widehat{\boldsymbol{\theta}})}{P(s)}$$

informs

Weghorn & Sunyaev, *Nat. Genet.* (2017)

Per-gene inference of selective growth (dis)advantage: $P(x|s; r_x, \hat{\boldsymbol{\theta}})$



p-values for each gene for $x \in$ {#missense, #nonsense}

synonymous mutation
nonsynonymous mutation
purged ★ (negative selection)
added ★ (positive selection)

r_x = nonsyn/syn target size ratio on gene
x = # nonsynonymous mutations
s = # synonymous mutations
$\hat{\boldsymbol{\theta}}$ = estimated parameter vector

Weghorn & Sunyaev, *Nat. Genet.* (2017)

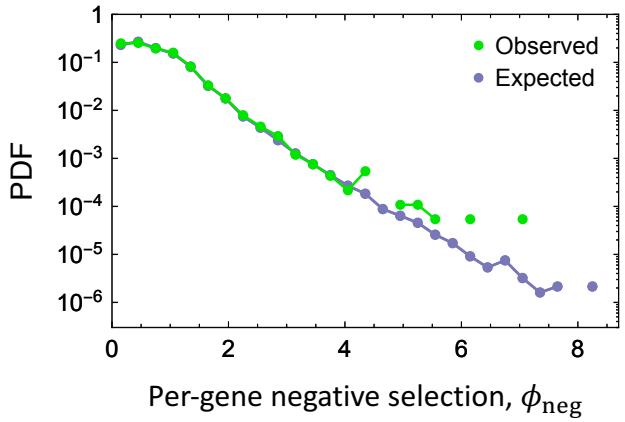
Per-gene inference of selection

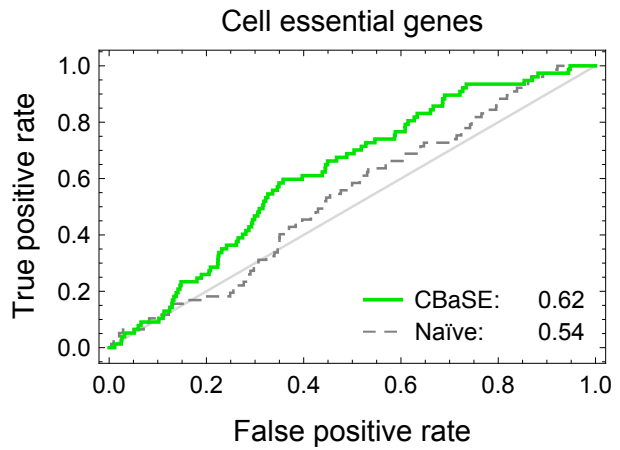
Results: Head-neck squamous cell carcinoma

Negative selection

Hypomutation ∝

$\phi_{\text{neg}} = -\log[p_{m,\text{neg}}] - \log[p_{k,\text{neg}}]$

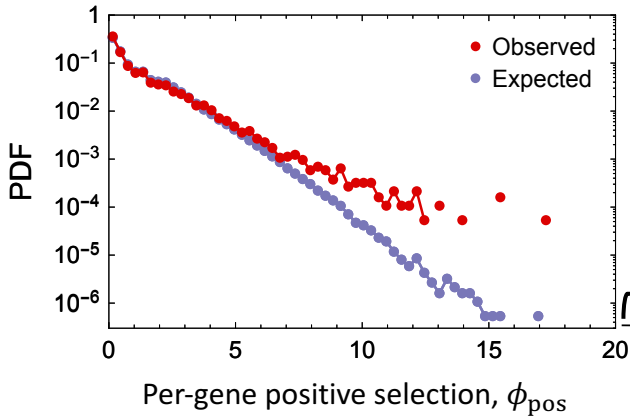


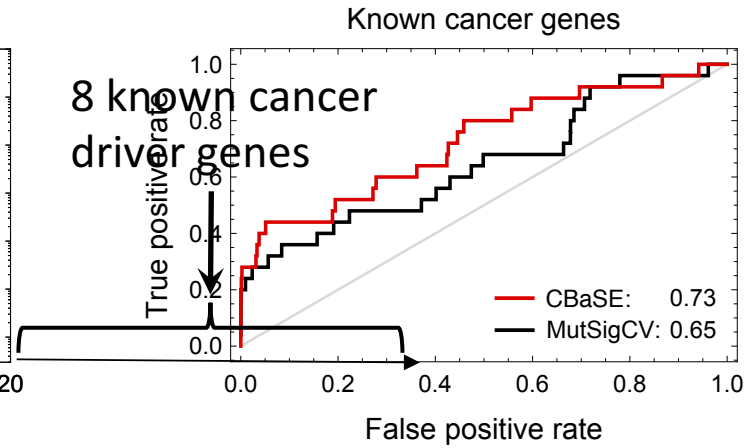


Positive selection

Hypermutation ∝

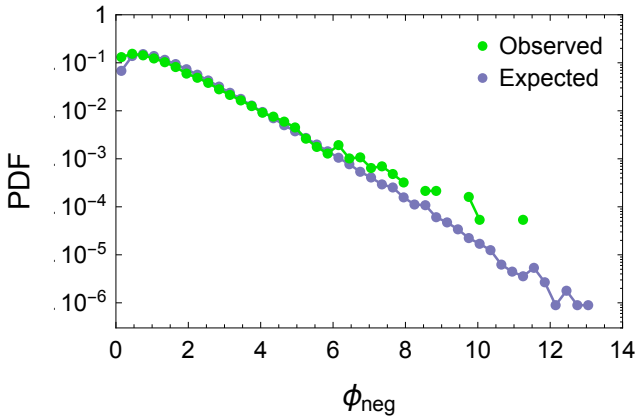
$\phi_{\text{pos}} = -\log[p_{m,\text{pos}}] - \log[p_{k,\text{pos}}]$

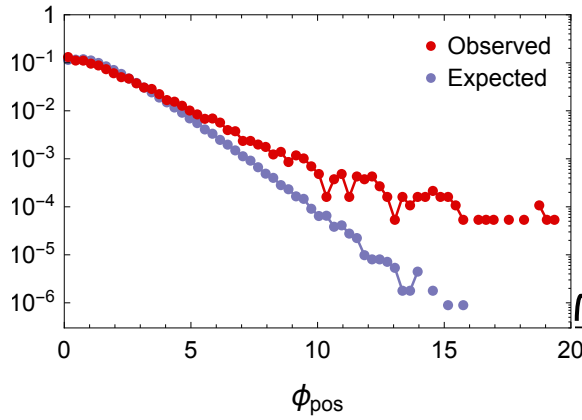




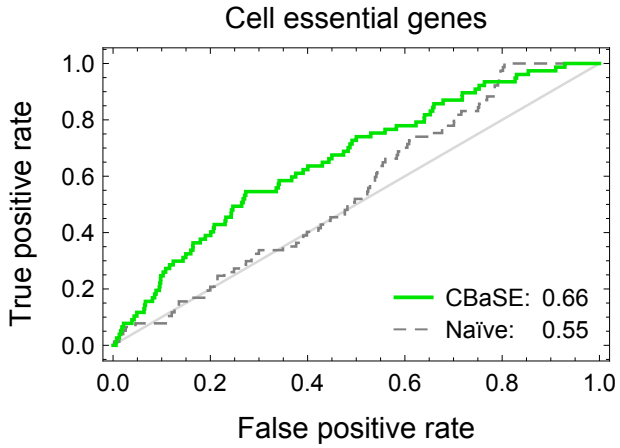
Weghorn & Sunyaev, *Nat. Genet.* (2017)

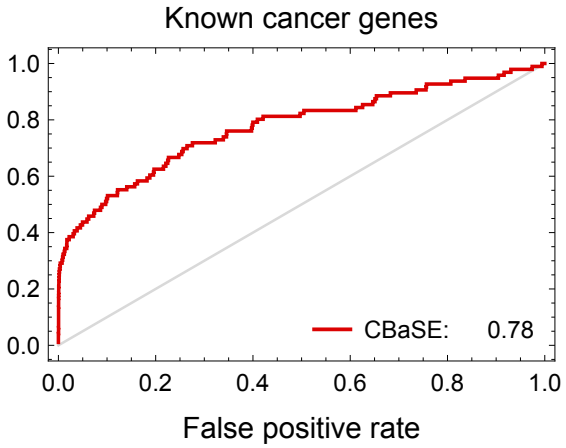
Pan-cancer





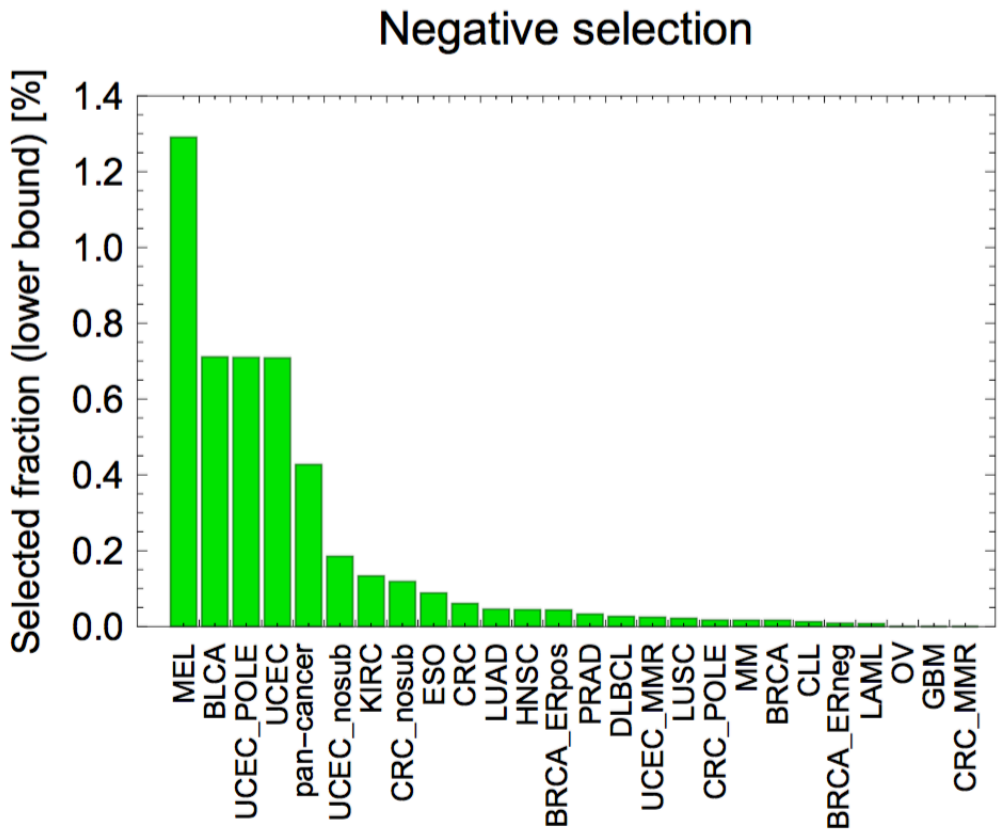
30 known cancer driver genes

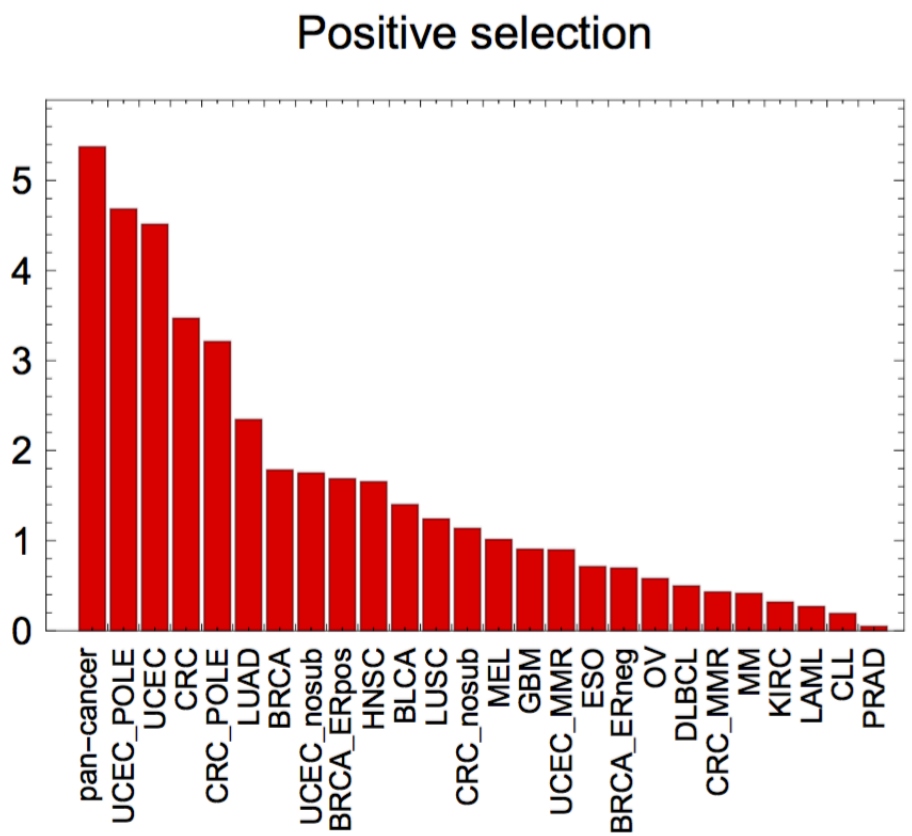




Weghorn & Sunyaev, *Nat. Genet.* (2017)

Lower bound on genome-wide selected fraction of genes





Weghorn & Sunyaev, *Nat. Genet.* (2017)

Gene findings

Top ten negatively selected genes contain:

BOD1L1	→ Repairs stalled replication fork
BCL2 *BCL11B* *PREX2*	→ Overexpressed in their respective cancer types ("oncogenes")

Among top 5 negatively selected genes in melanoma:

MKL1, *NPY5R*, *RMDN2*, and *DIAPH1*.

Among top 13 pan-cancer negatively selected genes:

ATAT1, *BCL2*, *CLIP1*, *GALNT6*, *CKAP5*, and *REV1*.

Across cancer types:

74% of genes with q_{pos}<0.002 are already members of the COSMIC cancer gene census (CGC; 43 genes).

Novel driver candidates:

ARHGAP35, *TRAF3*, *EPHA2*, *AJUBA*, *RBL2*, and *MED23.*

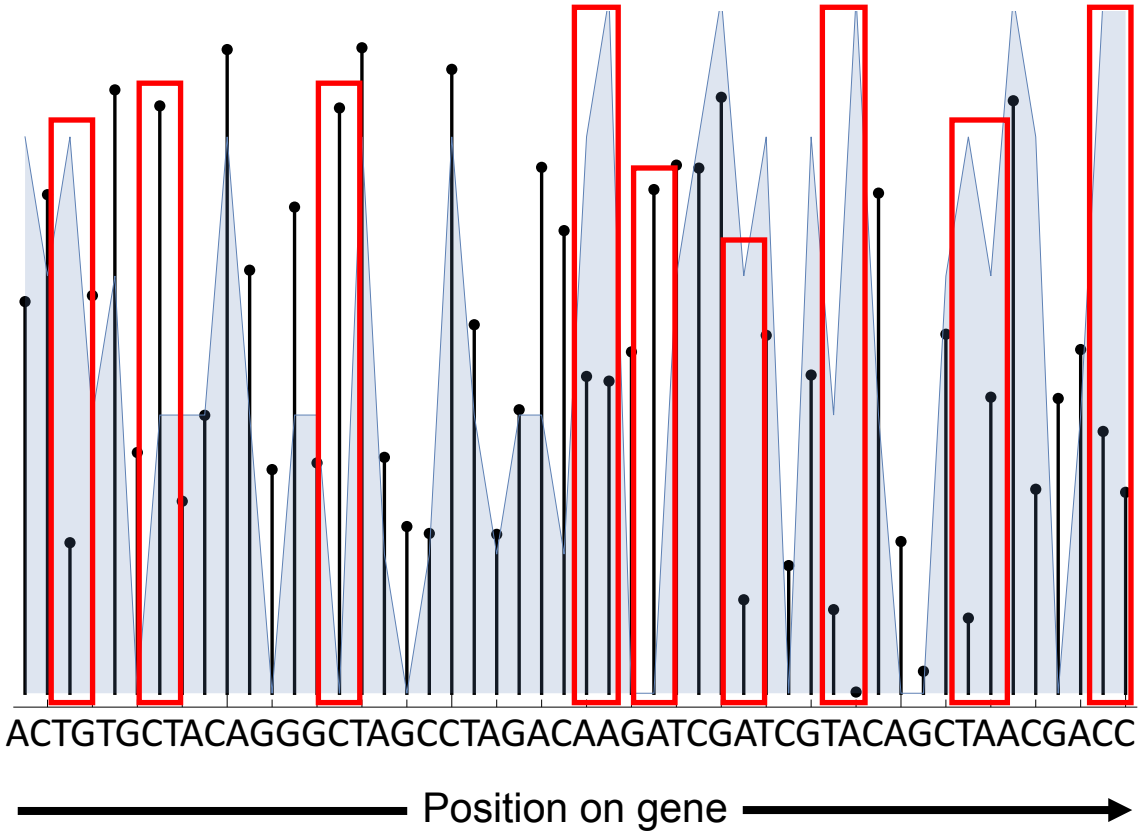
Pan-cancer:

97% of genes with $q_{\mathrm{pos}} = 0$ are members of the CGC (35 genes). Non-member: *ARHGAP35.*

Conclusions I

- Negative selection is disproportionately **hard to detect** in cancer (power, effect sizes, haplosufficiency, interference selection, ...).
- Overall signal of negative selection in cancer is small (≈**1% of genes**).
- Novel probabilistic approach enables for the first time:
  - estimation of negative selection at the individual **cancer type and gene level** and
  - **increased sensitivity** in detection of cancer driver genes compared to MutSigCV without input of external meta-data.
- Full method, **C**ancer **Ba**yesian **S**election **E**stimation (**CBaSE**), is available as a browser-based and standalone tool: http://genetics.bwh.harvard.edu/cbase

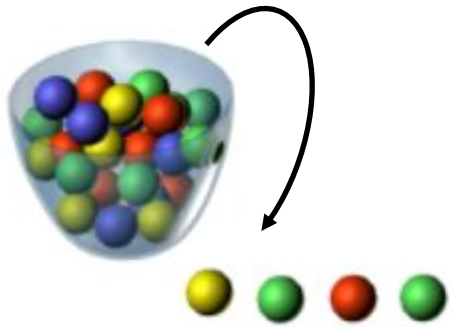
Pattern deviation test



- Expected mutation pattern, $\vec{p}$
- Observed mutation pattern, $\vec{v}$

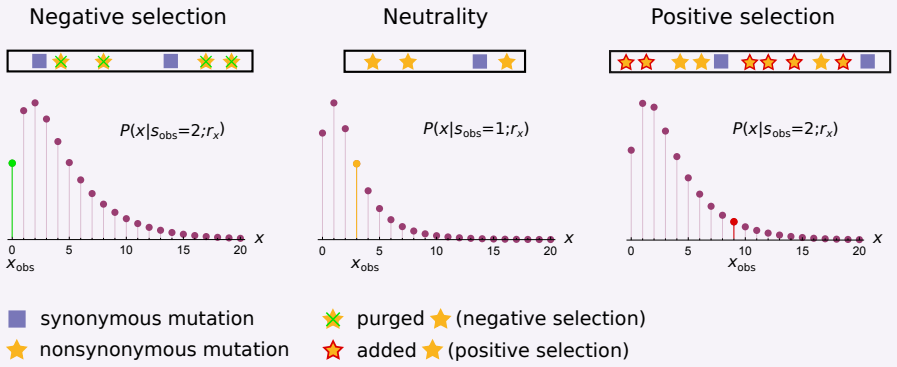
$$P(\boldsymbol{v}|x;\boldsymbol{p}) = \text{Multinom}(\boldsymbol{v};x,\boldsymbol{p}) \qquad x = \|\boldsymbol{v}\|_1$$

where $$p_i = \frac{\ell\left(b_i \to c\middle|\vec{b}\right)}{\sum_{i=1}^{L}\ell(b_i \to c|\vec{b})}$$



Two tests for selection

Mutational recurrence test



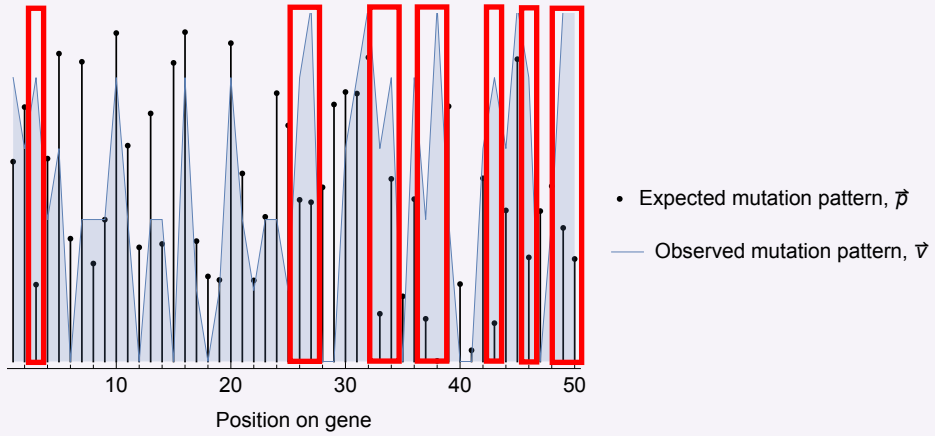
$$P\left(x\middle|s;r_x,\widehat{\boldsymbol{\theta}}\right)=\int \mathrm{d}\lambda_s\, P(x|\lambda_s r_x)\, P(\lambda_s|s;\widehat{\boldsymbol{\theta}})$$

$P(x|\lambda_s r_x)$: $\mathrm{Pois}(\lambda_s r_x)$

$P(\lambda_s|s;\widehat{\boldsymbol{\theta}})$: $\dfrac{P(s|\lambda_s)P(\lambda_s;\widehat{\boldsymbol{\theta}})}{P(s)}$

+

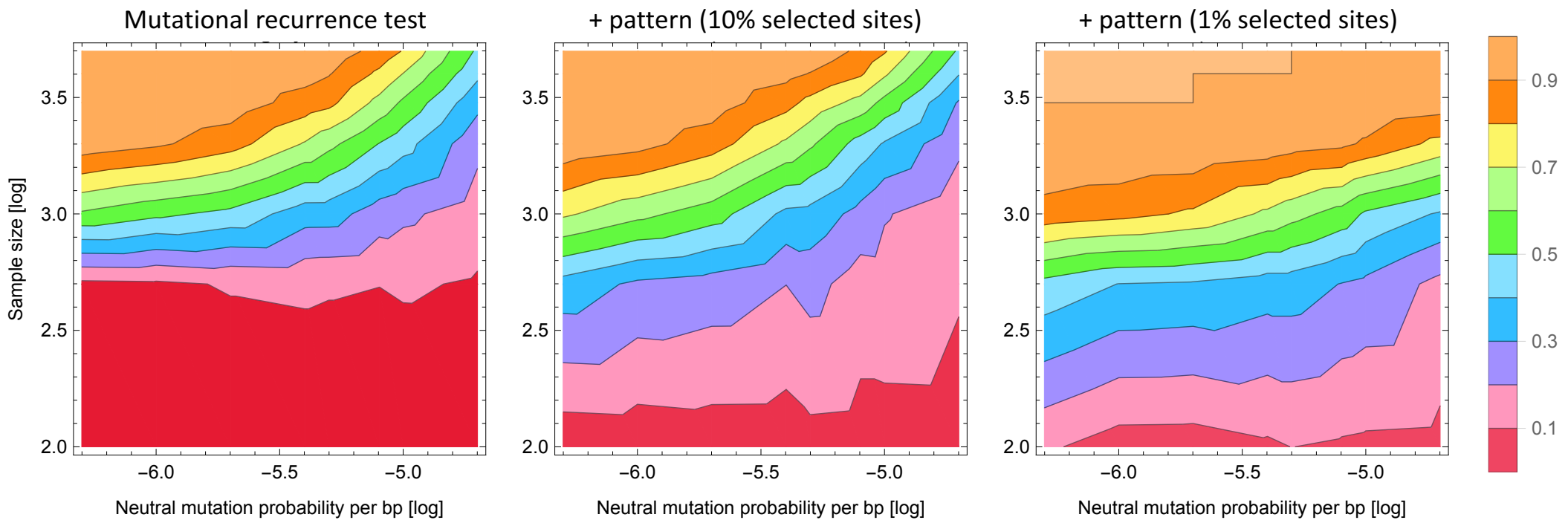
Pattern deviation test



$$P(\boldsymbol{v}|x;\boldsymbol{p})=\mathrm{Multinom}(\boldsymbol{v};x,\boldsymbol{p}) \qquad x=\|\boldsymbol{v}\|_1$$

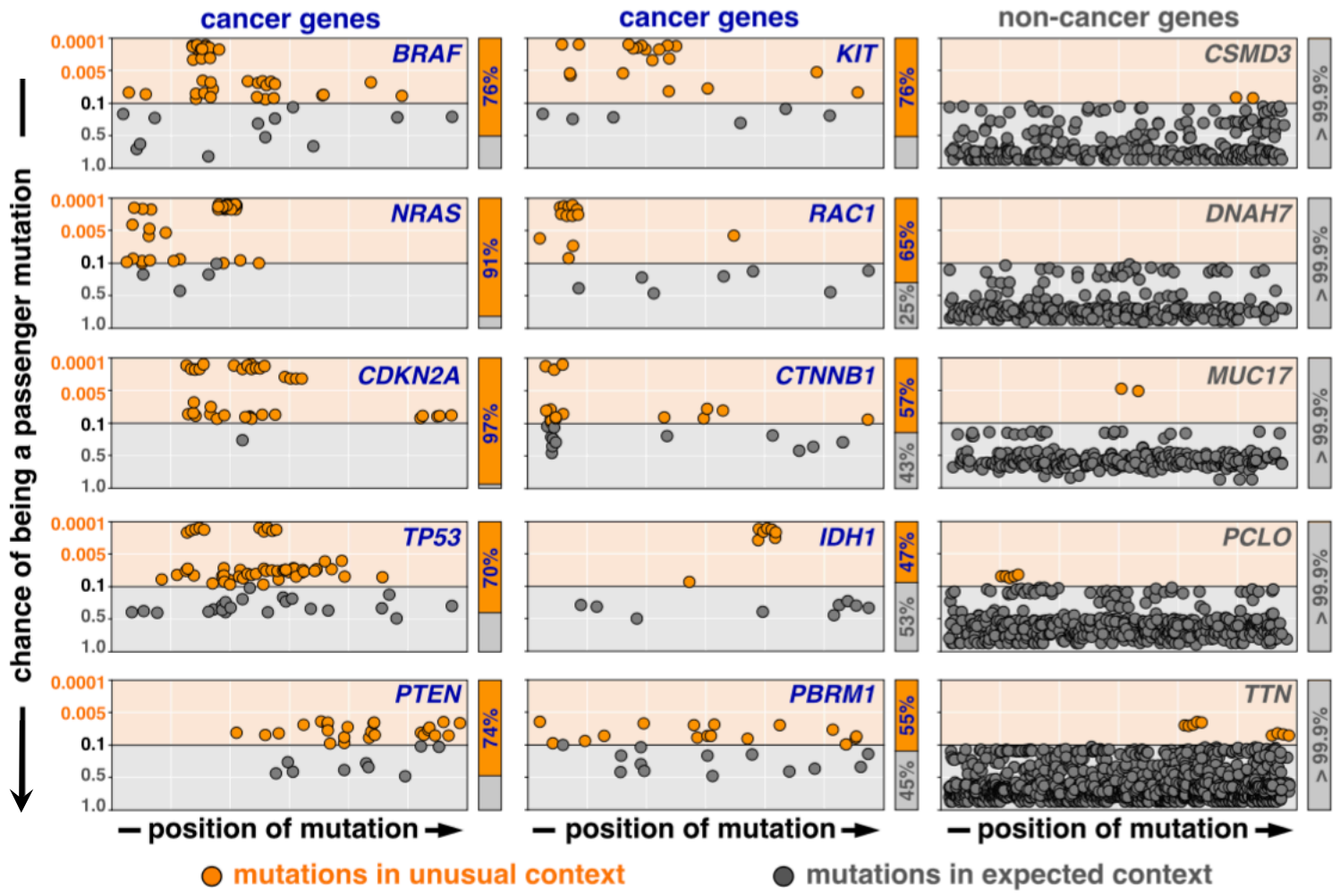
where $$p_i=\frac{\ell\left(b_i\to c\middle|\vec{b}\right)}{\sum_{i=1}^{L}\ell(b_i\to c|\vec{b})}$$

Power gain through additional test for pattern deviation



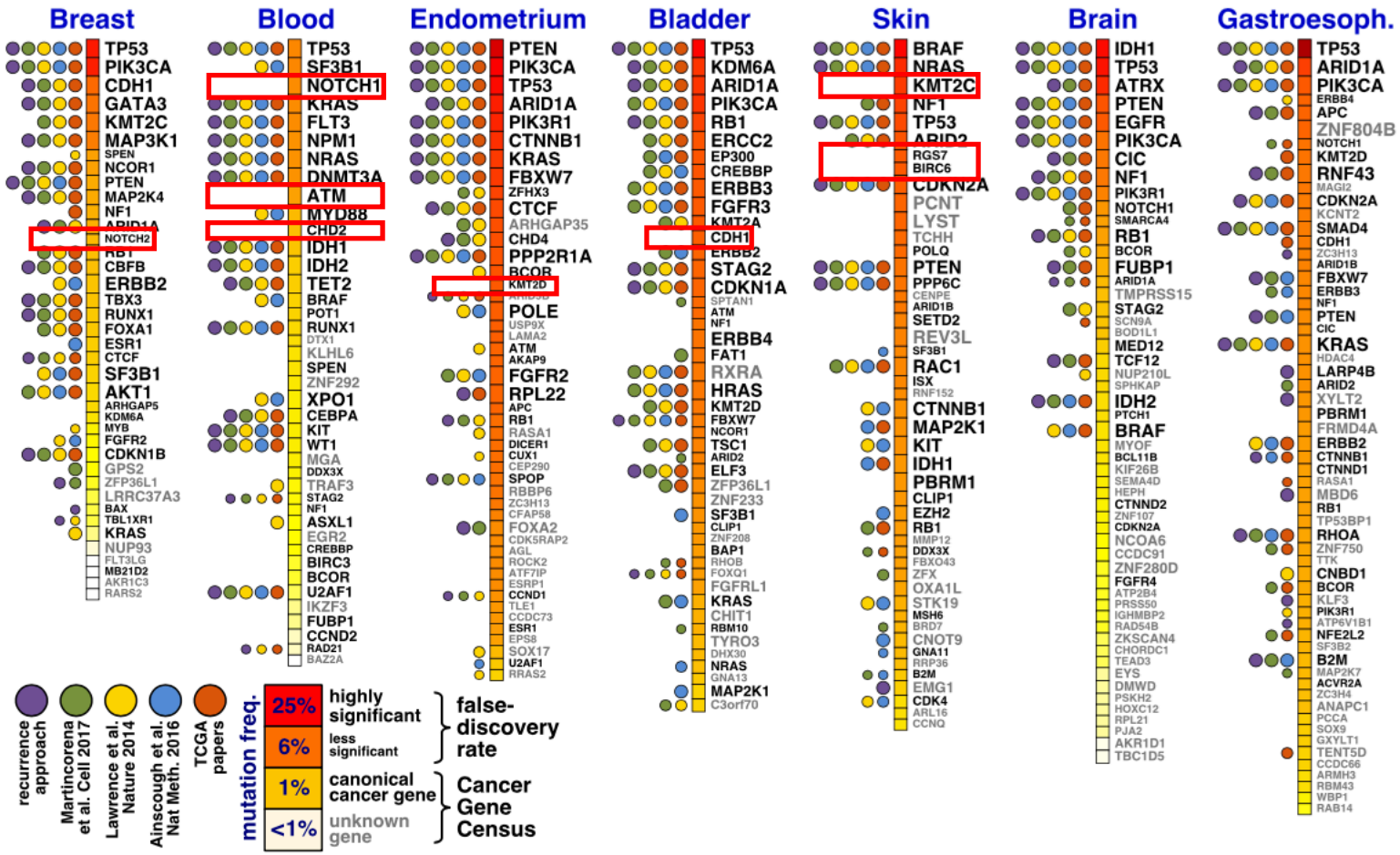
Simulation parameters: $2N_e s = 0.01/(l_{sel}\mu)$, $l_{sel} = f_{sel} l_{non}$, $l_{non} = 1250$, $l_{syn} = 500$

Putative cancer driver mutations stick out



Dietlein*, Weghorn* et al., bioRxiv (2018)

Identification of novel cancer driver gene candidates



Dietlein*, Weghorn* et al., bioRxiv (2018)

Conclusions II

- Adding a test for the **“selection mutational signature”** can boost power to detect selection on some genes by up to several tens of percent, compared to a mutation recurrence test alone.
- Bulk of signal comes from **recurrence test**.
- Joint probabilistic framework (MutPanning) to **incorporate both tests** predicts novel cancer driver gene candidates.
- **Caveat**: Signature test is sensitive to DNA sequencing artifacts.