ncdDetect-2: Improved models of the site-specific mutation rate in cancer and driver detection with robust significance evaluation

Supplementary document

Malene Juul[1,2] (malene.juul.rasmussen@clin.au.dk)
Tobias Madsen[1,2] (tobias.madsen@clin.au.dk)
Qianyun Guo[2] (guo@cs.au.dk)
Johanna Bertl[1] (johanna.bertl@clin.au.dk)
Asger Hobolth[2] (asger@birc.au.dk)
Manolis Kellis[3] (manoli@mit.edu)
Jakob Skou Pedersen[1,2] (jakob.skou@clin.au.dk)

[1] *Department of Molecular Medicine, Aarhus University, Palle Juul-Jensens Boulevard 99, DK-8200 Aarhus N, Denmark*
[2] *Bioinformatics Research Centre, Aarhus University, C.F. Mollers Alle 8, DK-8000 Aarhus C, Denmark*
[3] *Computer Science and Artificial Intelligence Laboratory, MIT, 32 Vassar St, Cambridge, MA 02139, USA*

Contents

S1 Supplementary figures

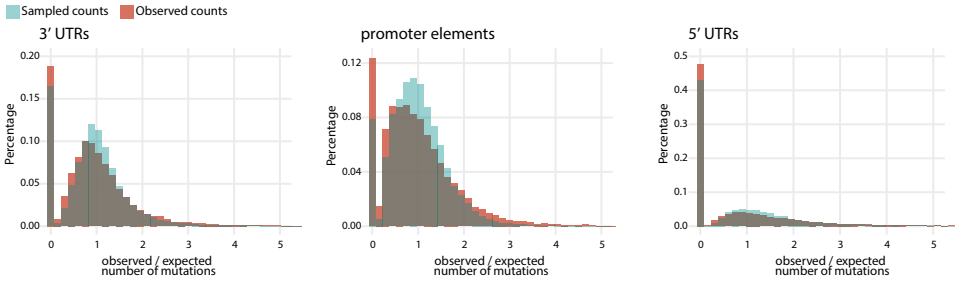

Figure S1: Observed-to-expected number of mutations shown for 3' UTRs, promoter elements, and 5' UTRs. For details, see Figure 2B legend description. Note that splice sites are omitted from these plots as they are short and heavily dominated by observed counts of zero.

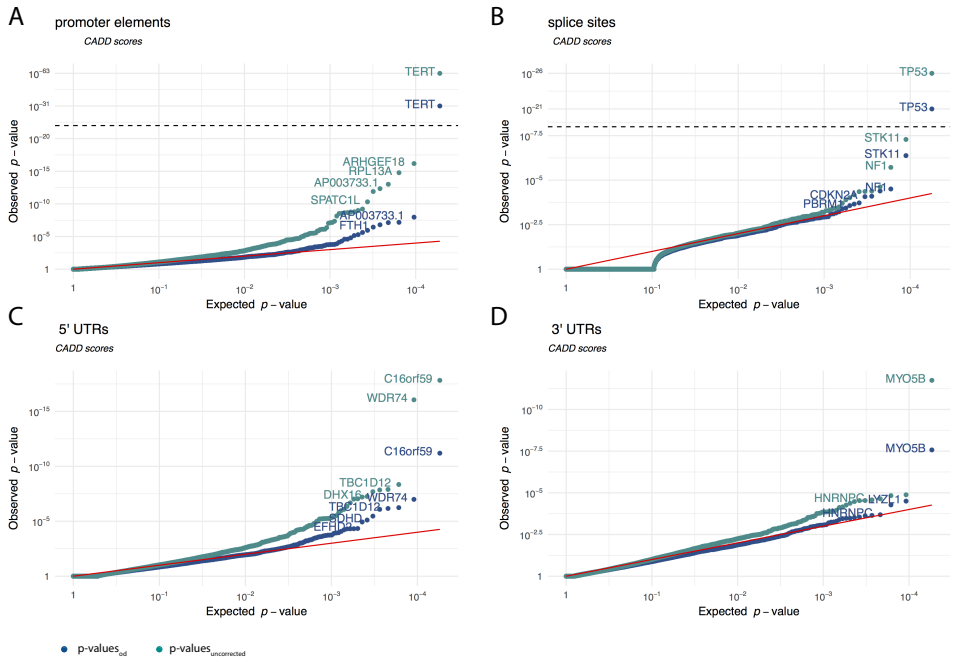

Figure S2: CADD-score based QQplots of p-values for the considered non-coding regulatory regions obtained with and without overdispersion. QQplot for protein-coding genes is shown in Figure 4G.

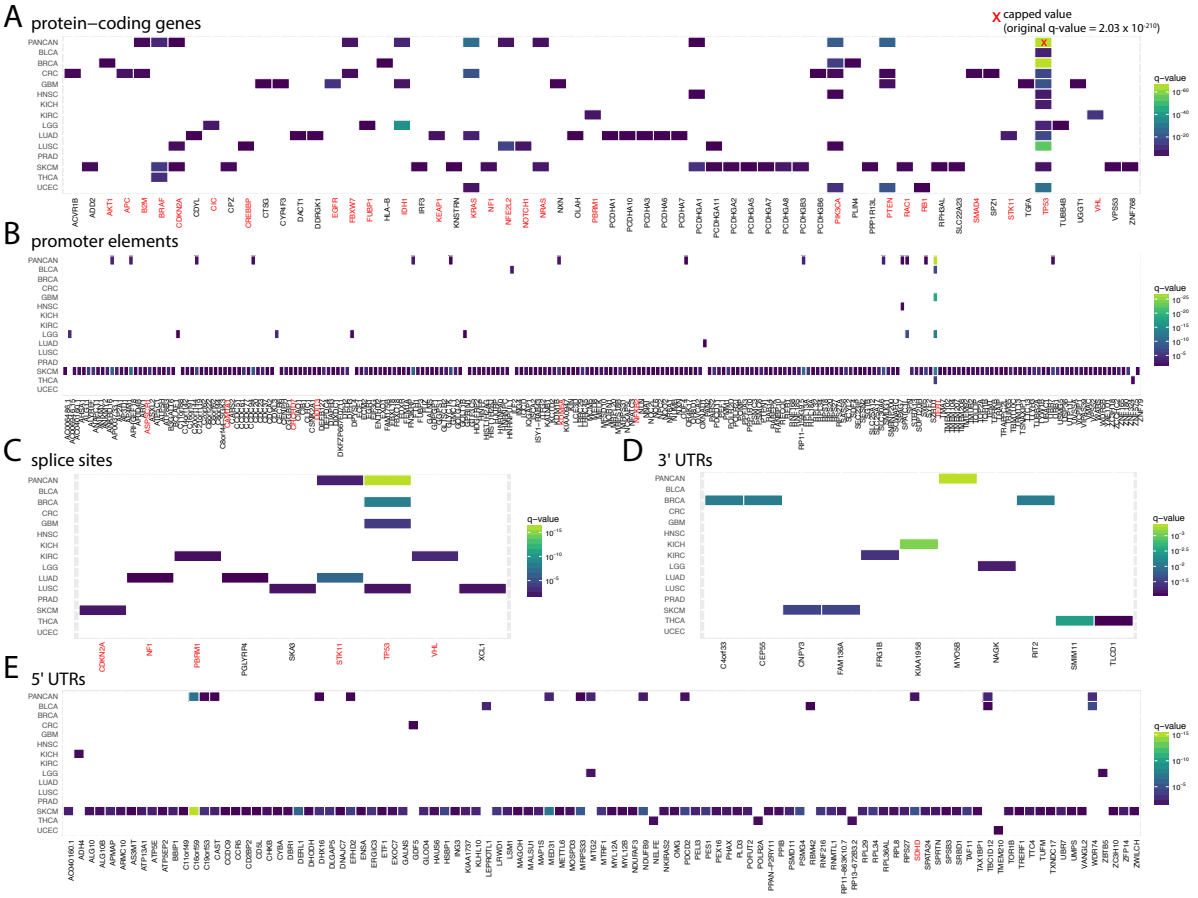

Figure S3: Potential cancer driver candidates identified with ncdDetect using overdispersion and CADD scores, including melanoma results. All elements with a colored tile has a q-value less than 10%.

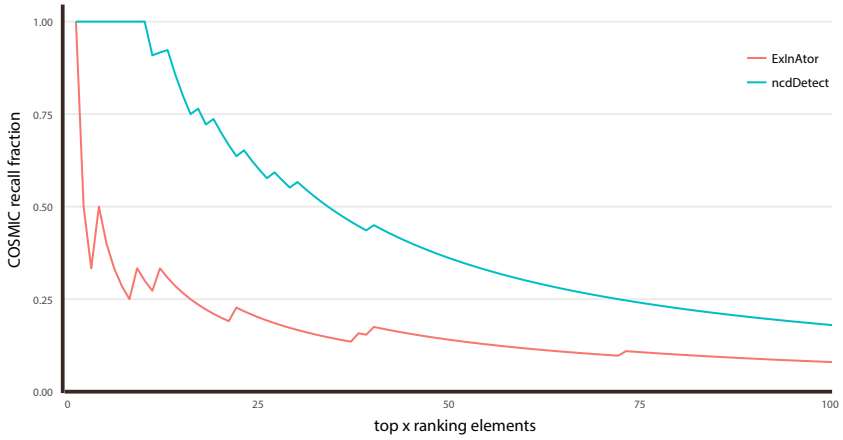

Figure S4: COSMIC CGC recall plot. Fraction of COSMIC CGC genes recalled in top candidates by ncdDetect and ExInAtor.

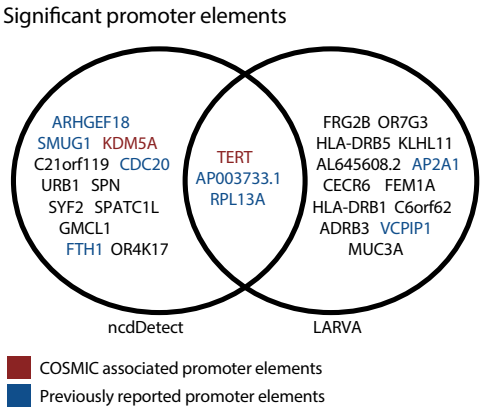

Figure S5: Promoter elements called significant by LARVA [3] and ncdDetect. Promoter elements highlighted in green are associated to COSMIC CGC genes [2]. Promoter elements highlighted in blue have previously been reported by other non-coding cancer driver screens [6, 4]. For the specific ranking of promoter elements by ncdDetect and LARVA, please refer to supplementary table 6.

S2 Autocorrelation distance between features

We want to explore the similarity between the pattern of values of a genomic feature and that of the observed-to-expected mutation rate. Features similar to the observed-to-expected mutation rate, are good candidate features able to explain part of the additional variance in the mutation rate model.

The autocorrelation distance is a distance measure between time series, quantifying the dissimilarity between the autocorrelation functions of two time-series(Montero and Vilar 2014). Given two time series with autocorrelation functions $\rho_1 = (\rho_{11}, \rho_{12}, \ldots)$ and $\rho_2 = (\rho_{21}, \rho_{22}, \ldots)$ respectively, the distance is given by

$$d(x, y) = (1 - p) \sum_{i=1}^{K} p^i (\rho_{1i} - \rho_{2i})^2$$

The parameter p is a weight decay, such that more distant autocorrelation have smaller impact on the distance measure. We used $p = 0.99$, but our results are consistent over a wide range of values for p.

We compute the pairwise autocorrelation distance between the features (Dnase hypersensitivity, h3k9me3, replication timing and xr-seq), the expected-to-observed mutation rate and expected-to-observed mutation rate where the values of the time series were shuffled. The feature closest to expected-to-observed mutation rate was XR-seq (Figure S6).

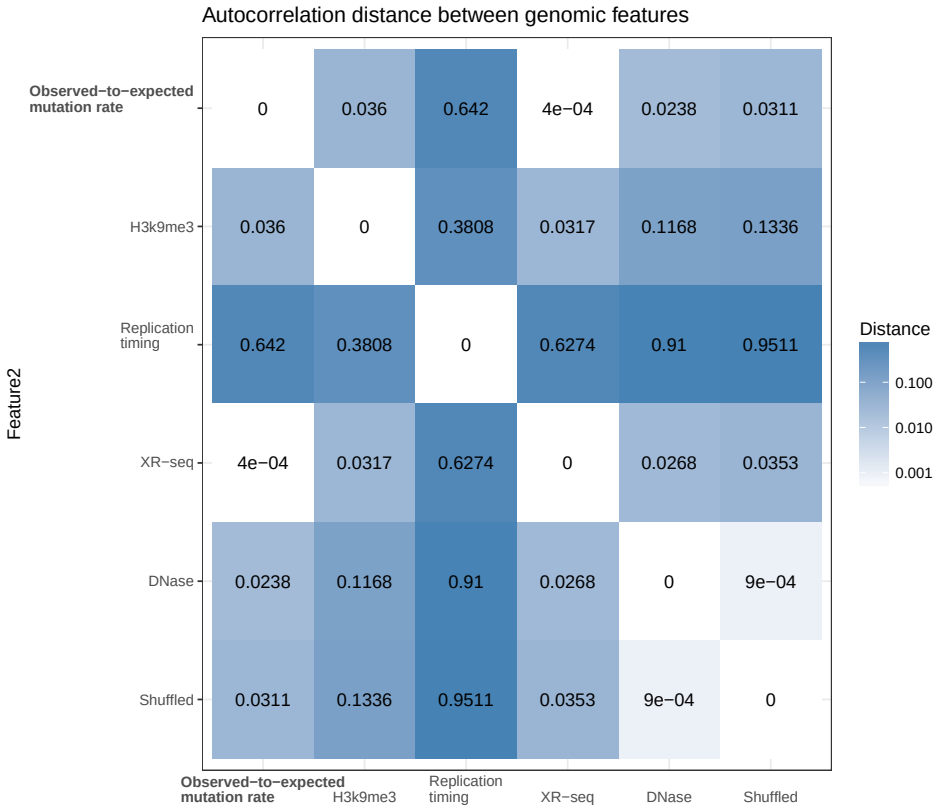

Figure S6: Auto-correlation distance between genomic features.

S3 Benjamini-Hochberg step-up procedure on subset of p-values is conservative

Given m hypothesis $H_1, \ldots, H_m$, with corresponding p-values $p_1, \ldots, p_m$, let $p_{(1)}, \ldots, p_{(m)}$ be the p-values sorted from smallest to largest. The Benjamini-Hochberg step-up procedure [1] requires us to find the maximal k such that

$$p_{(k)} \leq \frac{k}{m}\alpha,$$

where α is the level at which FDR rate is controlled.

Suppose now that we have only computed the p-value for a subset of the hypotheses. Let $n \leq m$ be the size of this subset. Denote by $r_{(1)}, \ldots, r_{(n)}$, the sorted p-values for this subset. Clearly $r_{(j)} >= p_{(j)}$, hence the maximal k' such that

$$r_{(k')} \leq \frac{k'}{m}\alpha,$$

is smaller than k. Thus fewer hypothesis are rejected. It can also be seen that all hypothesis rejected while knowing only a subset of the p-values would also have been rejected had we known all p-values.

S4 OD estimates in sub-samples and sub-datasets

Overdispersion measures the regional deviation between the observed and expected mutation count accumulated across a large number of samples. Sample-specific deviations will cancel out as a larger number of samples is accumulated. Cancer-type specific deviations will cancel out as we accumulate samples from different cancer types. Pan-cancer deviations do not cancel out and will remain as more samples are included.

First, we compute the overdispersion in sub-datasets sampled randomly without replacement from the 505 samples. We note that the od-estimates decreases as we include more samples, allowing sample-specific deviations in the mutation rate to cancel out. In concordance with the idea of pan-cancer deviation it appears that the over-dispersion asymptote to a value larger than zero. Note also that as more samples are included, the estimate of overdispersion stabilizes with a small variance.

Next, we compute the overdispersion in each of the 14 cancer-types and compare them to the overdispersion estimates obtained for a similar sized sub-sampled dataset. The larger overdispersion in the single cancer-type datasets compared to the overdispersion in similar sized sub-sampled datasets indicate that there are cancer-type specific deviations in the mutation rate.

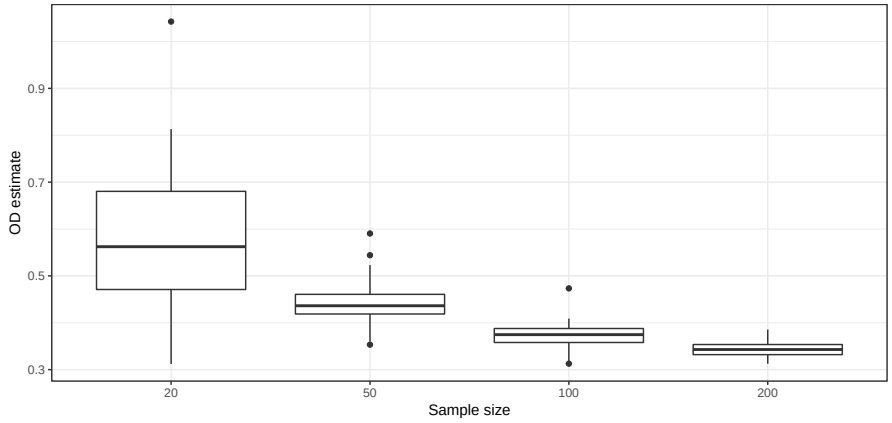

Figure S7: OD estimates as the number of samples increase. The samples are sampled randomly from all cancer-types.

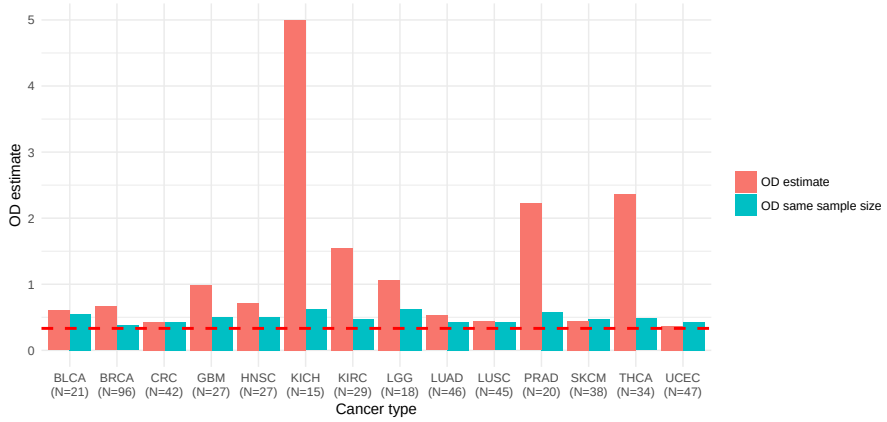

Figure S8: OD estimates for each cancer-type. The OD estimate is compared to the average OD estimate obtained by sampling the same number of samples across all cancer-types and computing the overdispersion.

S5 Logit-normal distribution

We look at a fixed predicted mutation probability of $\hat{p} = 1 \cdot 10^{-6}$. Let $r = \text{logit}(\hat{p}) = \log\left(\frac{\hat{p}}{1-\hat{p}}\right)$. The underlying true mutation probability, p, is modelled as

$$p = \text{logistic}(r + \gamma) = \frac{1}{1 + \exp(-r - \gamma)}$$

where $\gamma \sim \mathcal{N}(0, \sigma^2)$. Using transformation of random variables, the probability density function of p can be found as

$$\begin{aligned} f_p(p) &= \varphi\left(\text{logit}(p) - r; \sigma^2\right) \left| \frac{d}{dp}\text{logit}(p) - r \right| \\ &= \varphi\left(\text{logit}(p) - r; \sigma^2\right) \frac{1}{p(1-p)}. \end{aligned}$$

where $\varphi(\cdot; \sigma^2)$ is the density function for a normal distribution with variance σ^2.

S6 Simulation study of the effects of overdispersion

To investigate how ncdDetect performs on simulated data, we randomly sampled 500 non-COSMIC CGC genes, and added overdispersion to the mutation probabilities predicted by the background model for these genes. Concretely, we added a random effect $\gamma \sim \mathrm{N}(0, \rho^2)$ to the mutation probabilities, where $\rho \in \{0, 0.2, 0.4, 0.6\}$. We randomly selected 50 of the sampled genes to be true positives, and increased the number of mutations observed for these genes, to make their resulting scores similar to the mean observed score for ncdDetect-recalled COSMIC CGC genes. We ran ncdDetect to learn the degree of overdispersion, and accounted for this in the significance evaluation of the sampled genes. We also performed significance evaluation without accounting for overdispersion.

Significance evaluation in this setup illustrates that the number of false positives increases with the amount of overdispersion (Figure S9). We further observe that as the amount of overdispersion decreases, the recall rate of ncdDetect improves (Figure S10).

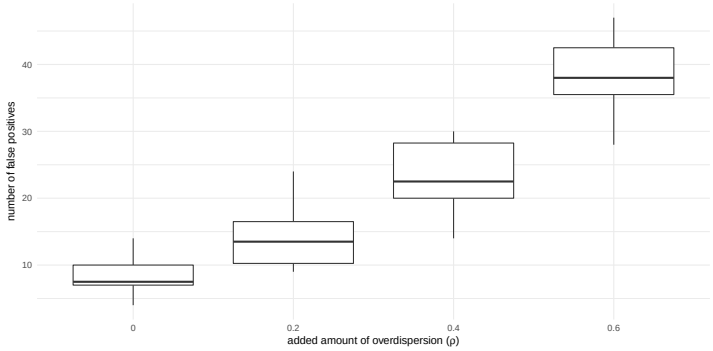

Figure S9: Number of false positives detected with ncdDetect as a function of the amount of overdispersion. P-values are obtained without accounting for overdispersion.

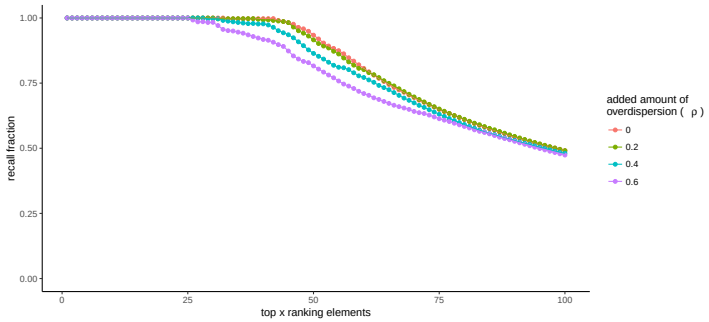

Figure S10: The ability of ncdDetect v.2 to recall the simulated true positive elements as the amount of added overdispersion increases. Each point of the plot is the average result of ten simulations. P-values are obtained by accounting for overdispersion.

References

[1] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the royal statistical society. Series B (Methodological)*, pages 289–300, 1995.

[2] Simon A. Forbes, David Beare, Harry Boutselakis, Sally Bamford, Nidhi Bindal, John Tate, Charlotte G. Cole, Sari Ward, Elisabeth Dawson, Laura Ponting, et al. COSMIC: somatic cancer genetics at high-resolution. *Nucleic Acids Research*, 45(D1):D777–D783, 01 2017.

[3] Lucas Lochovsky, Jing Zhang, Yao Fu, Ekta Khurana, and Mark Gerstein. LARVA: an integrative framework for large-scale analysis of recurrent variants in noncoding annotations. *Nucleic Acids Res*, 43(17):8123–8134, Sep 2015.

[4] Collin Melton, Jason A Reuter, Damek V Spacek, and Michael Snyder. Recurrent somatic mutations in regulatory regions of human cancer genomes. *Nat Genet*, 47(7):710–716, 07 2015.

[5] Serena Nik-Zainal, Helen Davies, Johan Staaf, Manasa Ramakrishna, Dominik Glodzik, Xueqing Zou, Inigo Martincorena, Ludmil B. Alexandrov, Sancha Martin, David C. Wedge, Peter Van Loo, et al. Landscape of somatic mutations in 560 breast cancer whole-genome sequences. *Nature*, 534(7605):47–54, 06 2016.

[6] Nils Weinhold, Anders Jacobsen, Nikolaus Schultz, Chris Sander, and William Lee. Genome-wide analysis of noncoding regulatory mutations in cancer. *Nat Genet*, 46(11):1160–1165, 11 2014.