OncodriveROLE classifies cancer driver genes in Loss of Function and Activating mode of action

Applying OncodriveROLE to the tumor suppressor genes from Zhao et al

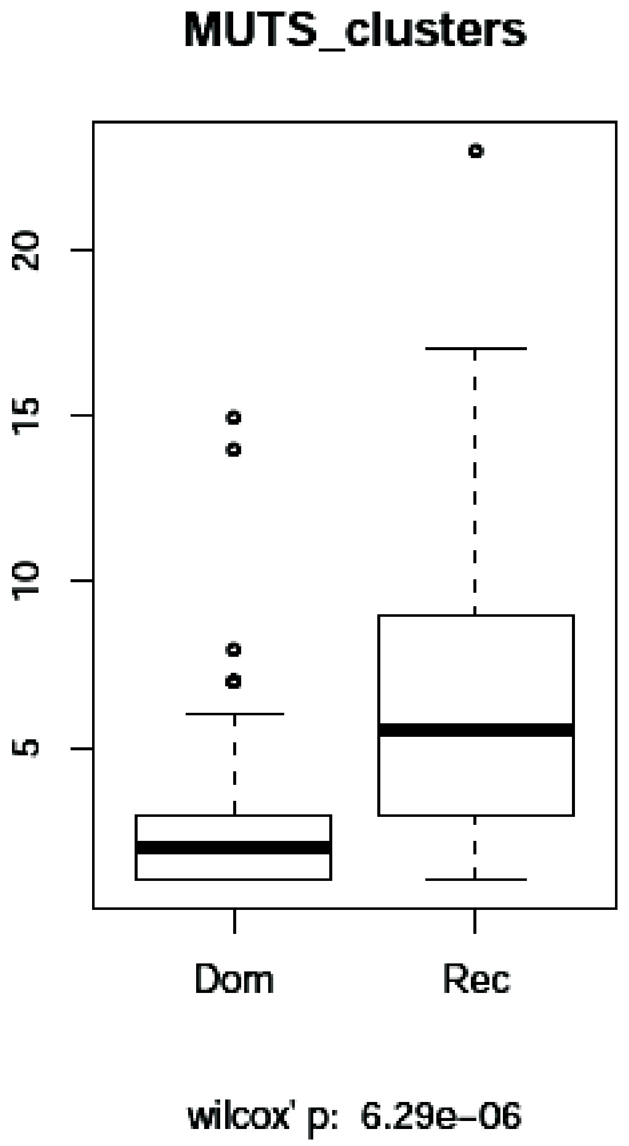
We tested the performance of OncodriveROLE on an independent set of genes annotated in the Tumor Suppressor Gene database (Zhao *et al.*, 2013). We downloaded the list of 716 human tumor suppressor gene entries, of which 610 are protein coding associated with an Ensembl gene id. Since OncodriveROLE is designed to classify previously predicted mutational drivers, we focused on the genes in the database that have previous evidence of being mutational drivers. In other words, we focused on genes of this list that are in the HCD or the Cancer5000 lists. Furthermore, we restricted our analysis to the genes not present in the CGC, which is our training set, to keep the validation employing this set completely independent from the training set. There are 27 genes in the TSGenes database with these properties, 22 of which are annotated as LoF drivers according to OncodriveROLE, 2 as Act drivers and 3 remain unclassified using the cut-offs 0.3 and 0.7. This represents an ACC of 91.7% and a COV of 89%. Note that one of the genes classified as Act is RHOA, which has been described early on as proto-oncogene and is annotated as such in the Uniprot database (Moscow *et al.*, 1994; Consortium, 2014), thus, this case is also probably correctly classified by OncodriveROLE.

Training the OncodriveROLE classifier with different combinations of features

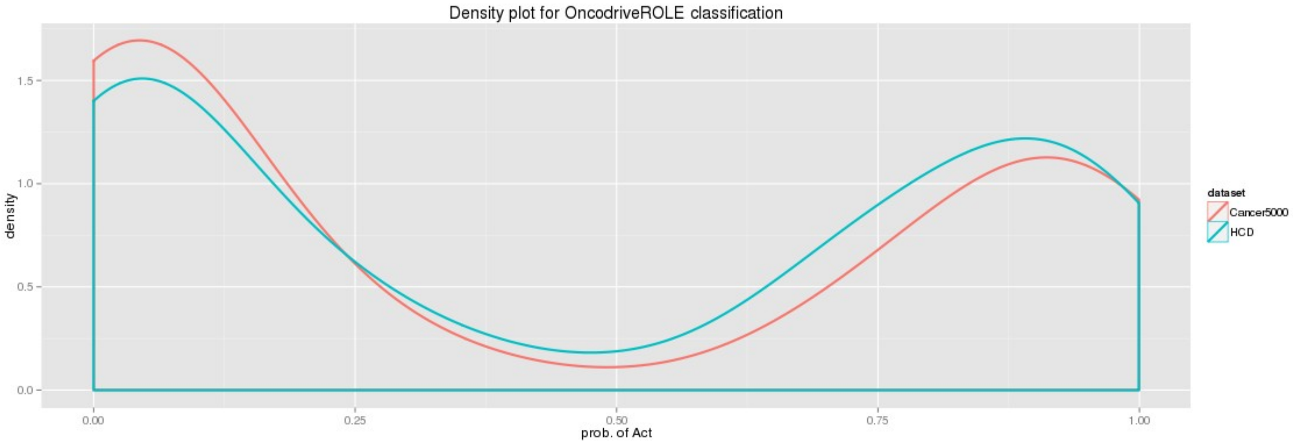
Given the amount of features developed and evaluated in our work, it is possible to choose alternative features combinations to train the OncodriveROLE classifier, depending on the set of driver genes that the researcher intends to classify.

We have evaluated the performance of different classifiers employing several combinations of the predictive features. These combinations of features and the results of the classifier trained on them are shown in Table S2. The performance of the classifiers trained on different features sets presents only very slight variations, with one or two genes in the cross-validation changing from one class to another. Although several classifiers with combination of only two features perform similarly as the classifiers with three features we have chosen to maintain all three in OncodriveROLE. We think that the incorporation of the information of CNAs may help classify novel driver genes with scarce mutational information. In other words, we have preferred to maintain a more versatile classifier. Nevertheless, the fact remains that probably most mutational drivers may be correctly classified only on the basis of their abundance of truncating mutations and one of the features measuring the enrichment for missense or clustered mutations.

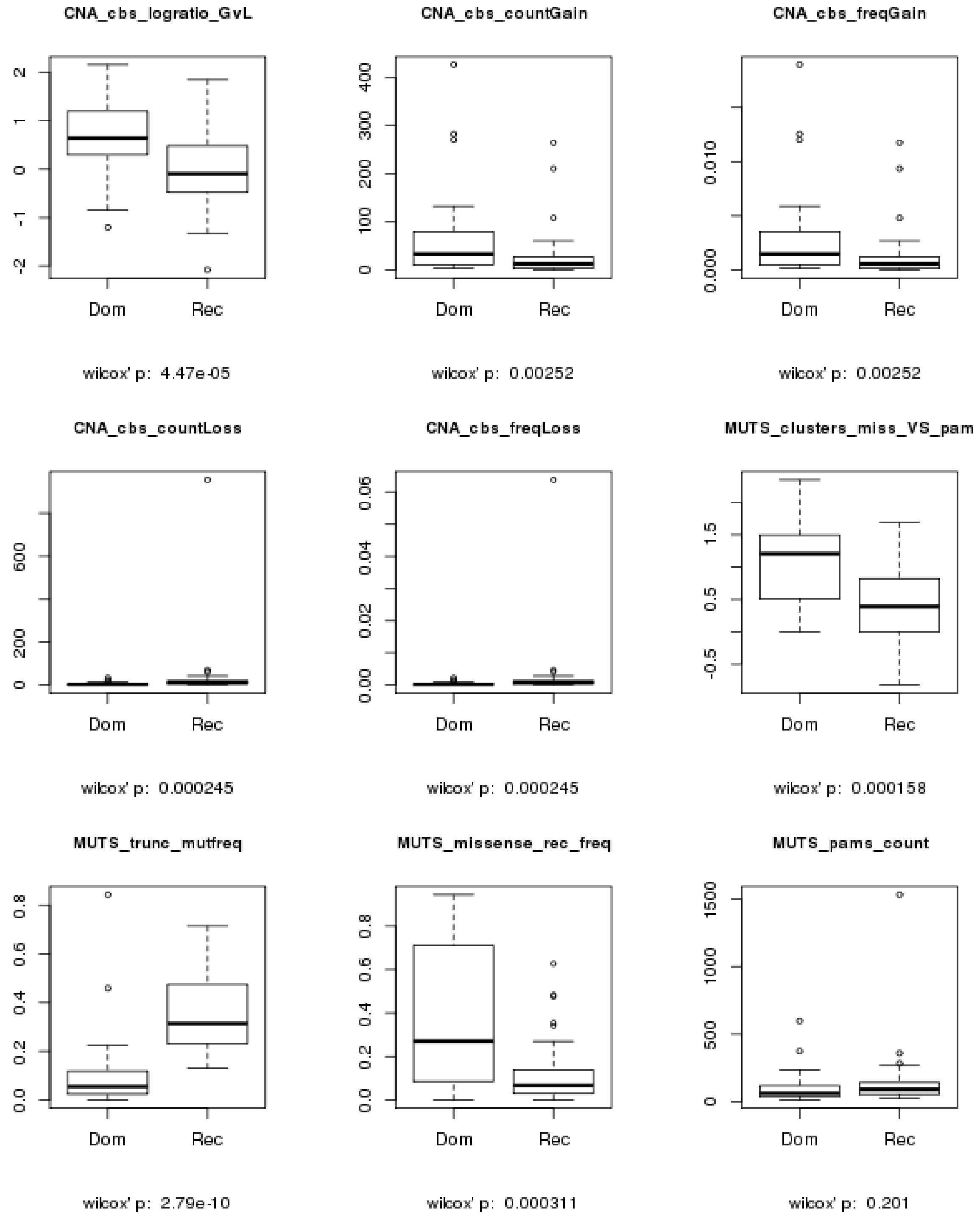
Suppl. Figure 1: CGC Dom genes within the OncodriveROLE training set tend to have fewer sites of clustered mutations (clusters) than CGC Rec genes. The clusters were detected via the computational method OncodriveCLUST (Tamborero *et al.*, 2013a).

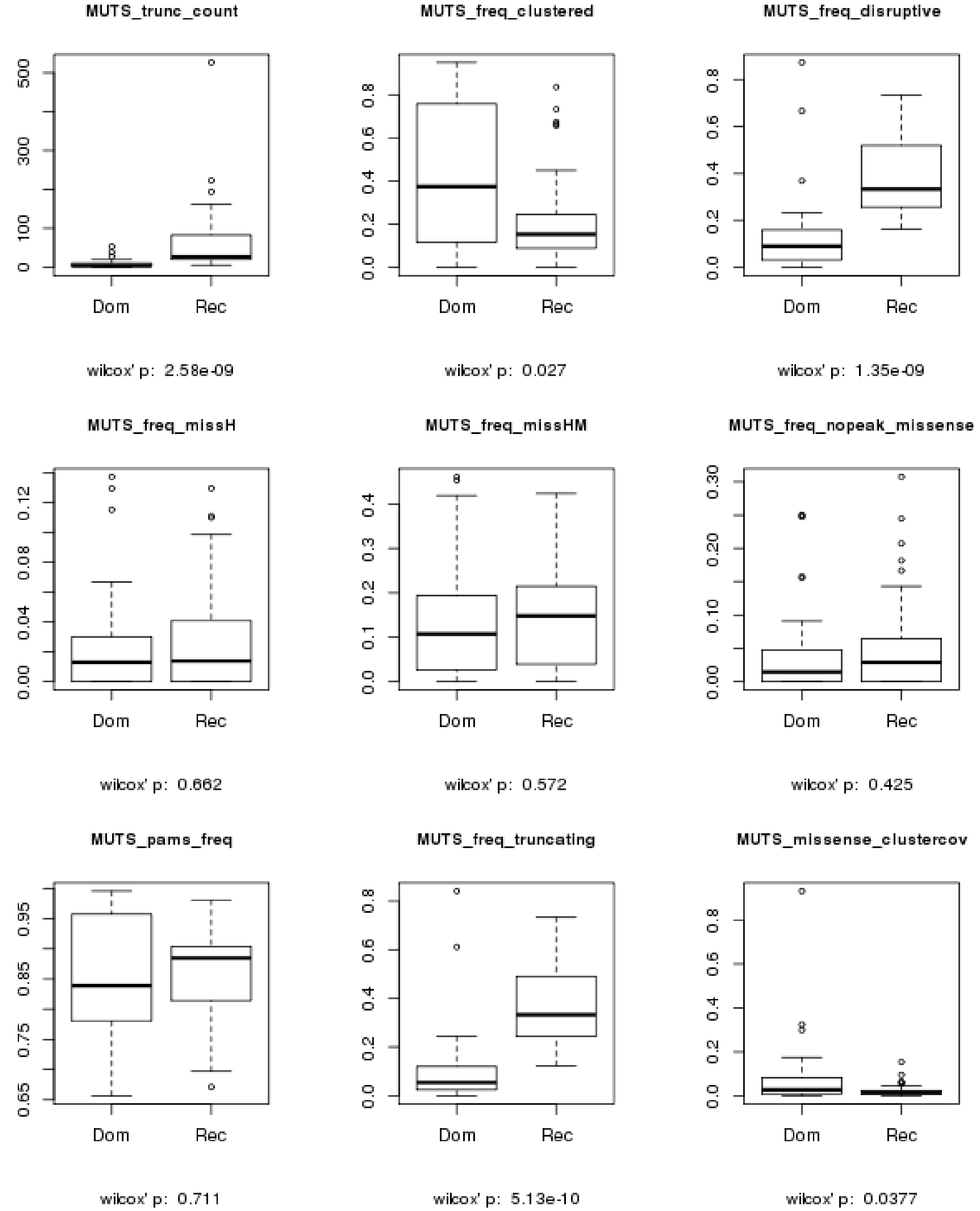


Suppl. Figure 2: A density plot of the results from the classification performed with OncodriveROLE on the two selected driver lists HCD (Tamborero *et al.*, 2013b) and Cancer5000 (Lawrence *et al.*, 2014). The plot shows a bimodal distribution in both result data sets and therefore allows selecting symmetric cut-offs.



Suppl. Figure 3: Boxplots for 45 CGC Rec and 31 CGC Dom genes for all 30 features tested when building the OncodriveROLE classifier.





MUTS_trunc_count

wilcox' p: 2.58e-09

MUTS_freq_clustered

wilcox' p: 0.027

MUTS_freq_disruptive

wilcox' p: 1.35e-09

MUTS_freq_missH

wilcox' p: 0.662

MUTS_freq_missHM

wilcox' p: 0.572

MUTS_freq_nopeak_missense

wilcox' p: 0.425

MUTS_pams_freq

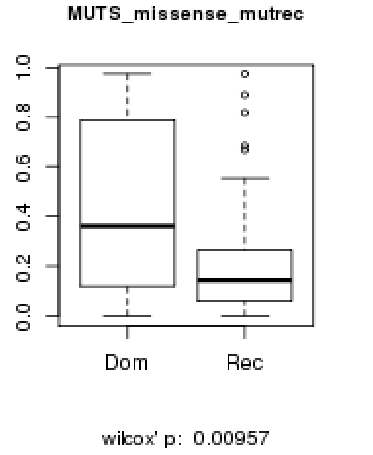
wilcox' p: 0.711

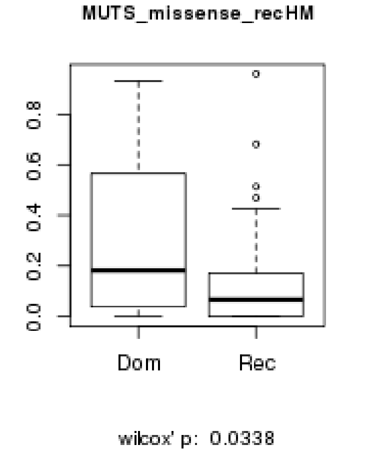
MUTS_freq_truncating

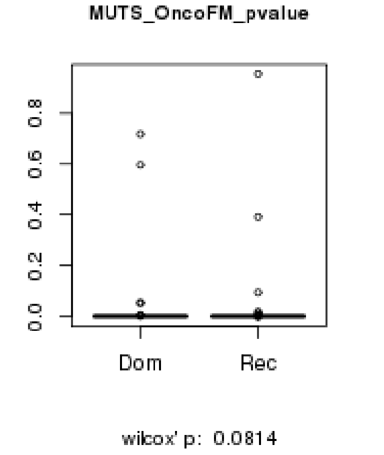
wilcox' p: 5.13e-10

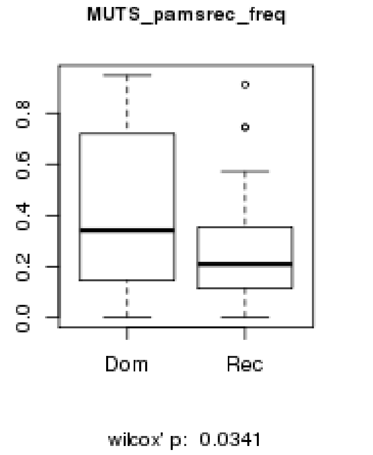
MUTS_missense_clustercov

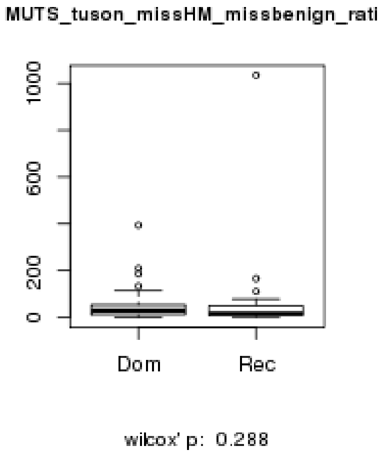
wilcox' p: 0.0377

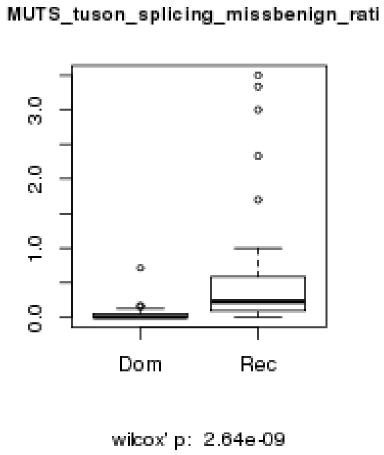

MUTS_missense_mutrec
0.0
0.2
0.4
0.6
0.8
1.0
Dom
Rec
wilcox' p: 0.00957

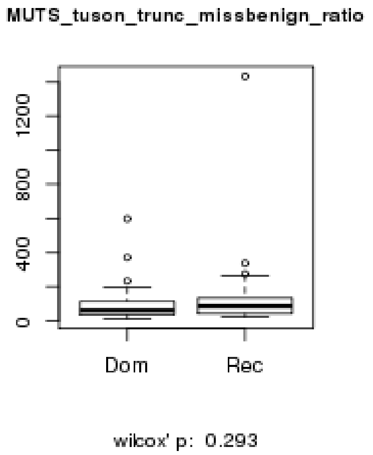

MUTS_missense_recHM
0.0
0.2
0.4
0.6
0.8
Dom
Rec
wilcox' p: 0.0338

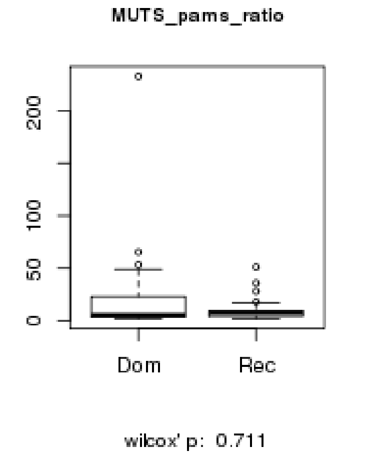

MUTS_OncoFM_pvalue
0.0
0.2
0.4
0.6
0.8
Dom
Rec
wilcox' p: 0.0814

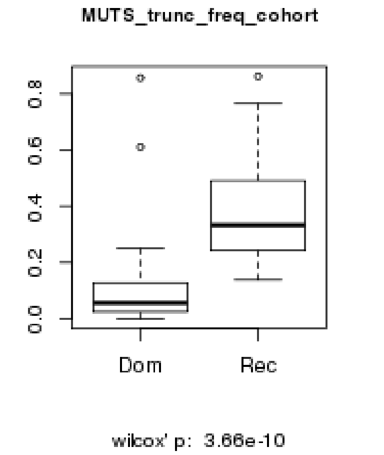

MUTS_pamsrec_freq
0.0
0.2
0.4
0.6
0.8
Dom
Rec
wilcox' p: 0.0341

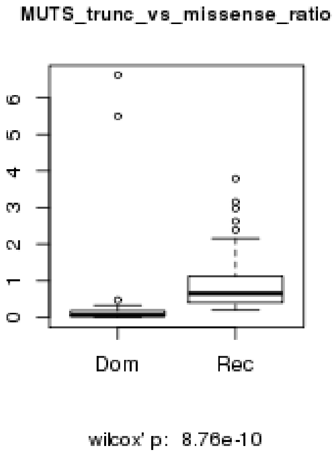

MUTS_tuson_missHM_missbenign_rati
0
200
600
1000
Dom
Rec
wilcox' p: 0.288

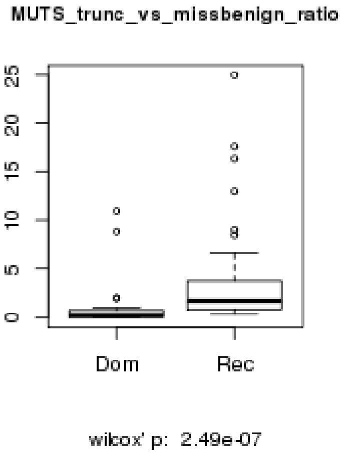

MUTS_tuson_splicing_missbenign_rati
0.0
1.0
2.0
3.0
Dom
Rec
wilcox' p: 2.64e-09

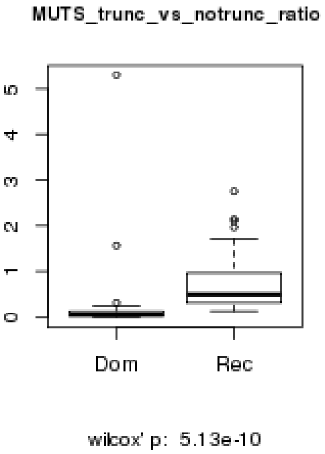

MUTS_tuson_trunc_missbenign_ratio
0
400
800
1200
Dom
Rec
wilcox' p: 0.293


MUTS_pams_ratio
0
50
100
200
Dom
Rec
wilcox' p: 0.711


MUTS_trunc_freq_cohort
0.0
0.2
0.4
0.6
0.8
Dom
Rec
wilcox' p: 3.66e-10


MUTS_trunc_vs_missense_ratio
0
1
2
3
4
5
6
Dom
Rec
wilcox' p: 8.76e-10


MUTS_trunc_vs_missbenign_ratio
0
5
10
15
20
25
Dom
Rec
wilcox' p: 2.49e-07


MUTS_trunc_vs_notrunc_ratio
0
1
2
3
4
5
Dom
Rec
wilcox' p: 5.13e-10

Consortium,T.U. (2014) Activities at the Universal Protein Resource (UniProt). *Nucleic Acids Res.*, **42**, D191–D198.

Frezza,C. *et al.* (2010) IDH1 Mutations in Gliomas: When an Enzyme Loses Its Grip. *Cancer Cell*, **17**, 7–9.

Lawrence,M.S. *et al.* (2014) Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature*, **505**, 495–501.

Moscow,J.A. *et al.* (1994) Utilization of multiple polyadenylation signals in the human RHOA protooncogene. *Gene*, **144**, 229–236.

Tamborero,D. *et al.* (2013b) Comprehensive identification of mutational cancer driver genes across 12 tumor types. *Sci. Rep.*, **3**.

Tamborero,D. *et al.* (2013a) OncodriveCLUST: exploiting the positional clustering of somatic mutations to identify cancer genes. *Bioinforma. Oxf. Engl.*, **29**, 2238–2244.

Zhao,M. *et al.* (2013) TSGene: a web resource for tumor suppressor genes. *Nucleic Acids Res.*, **41**, D970–D976.