From bewilderment to enlightenment in cancer research… hopefully

Limsoon Wong

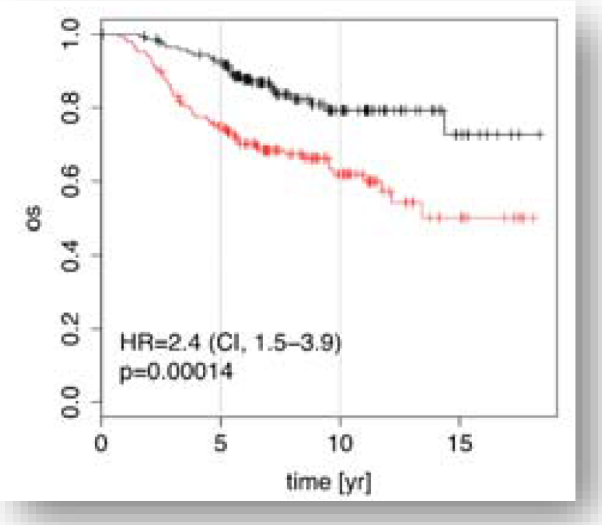
NUS
National University
of Singapore

A bewilderment

Breast cancer survival signatures are no better than random signatures

And maybe
some enlightenment at the end….

Venet et al., *PLOS Comput Biol*, 2011

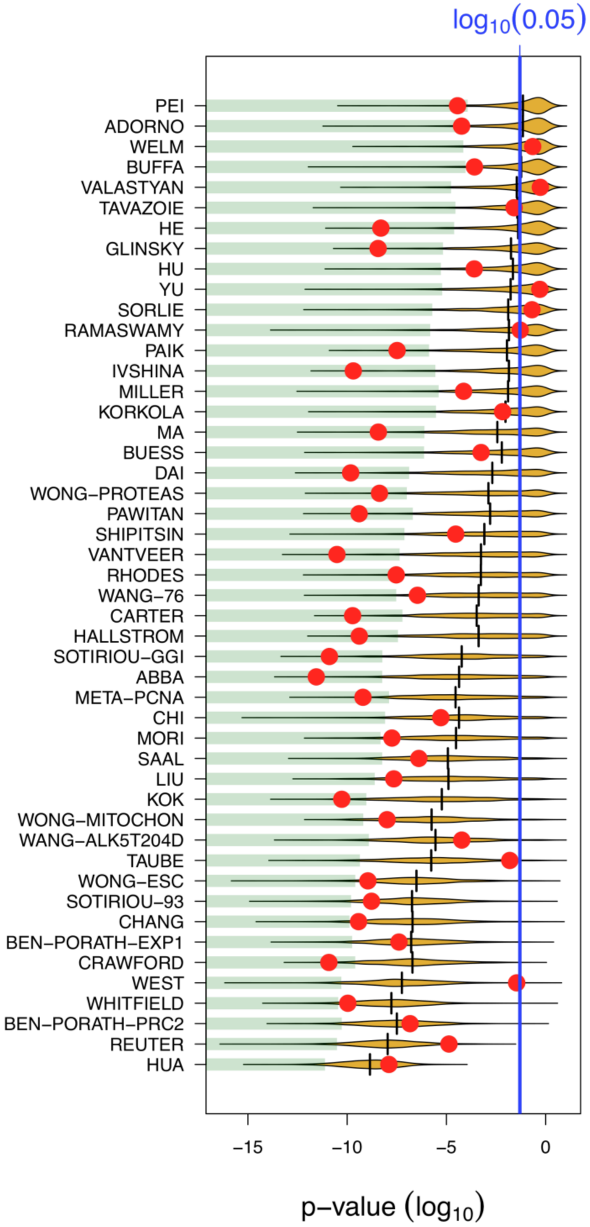


A seemingly obvious conclusion

A multi-gene signature (social defeat in mice) is claimed as a good biomarker for breast cancer survival

- Cox's survival model p-value << 0.05

A straightforward Cox's analysis. Anything wrong?



Venet et al., *PLOS Comput Biol*, 2011

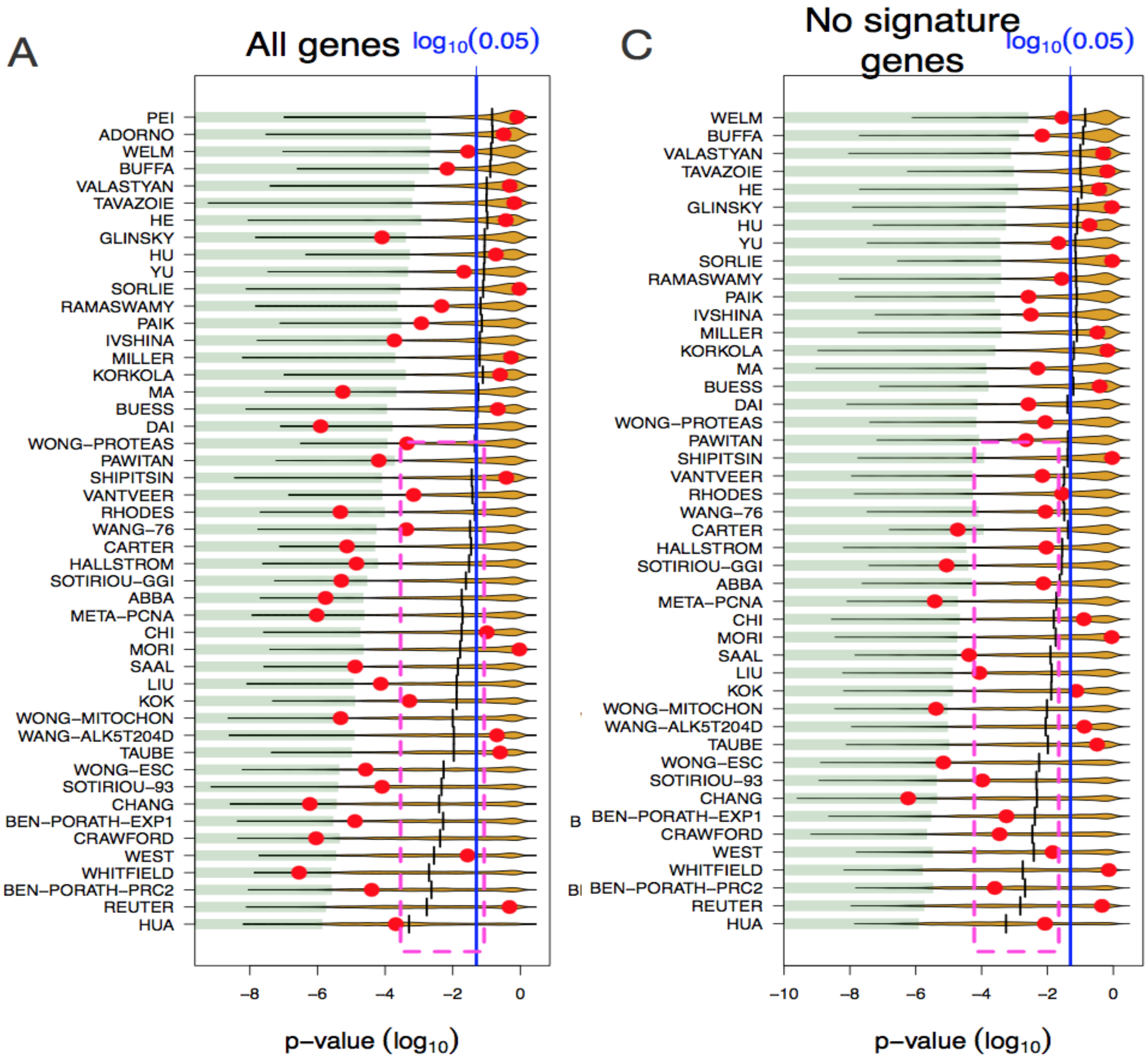
In fact, almost all random signatures also have p-value < 0.05;

And the larger a random signature is, the more likely this happens

Goh & Wong, Why breast cancer signatures are no better than random signatures explained. *Drug Discovery Today*, 2018

Maybe significant random signatures share genes with reported signatures?

Not quite…

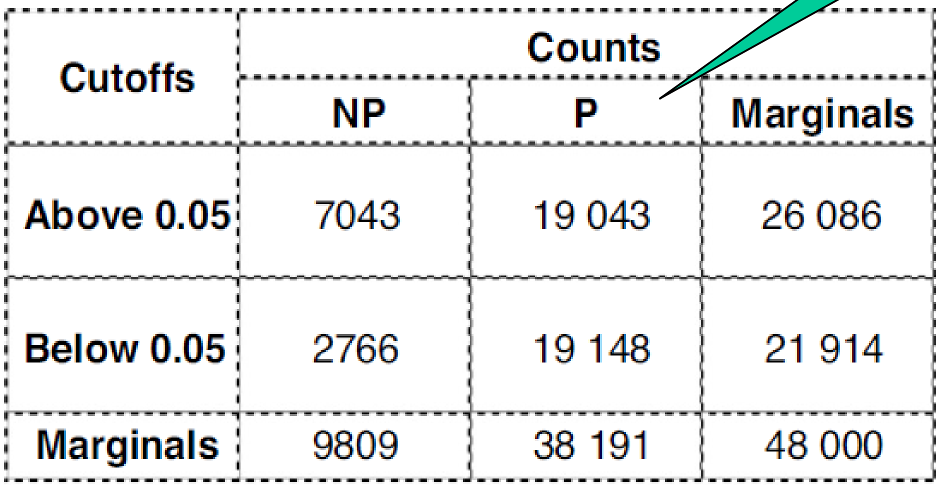


Perhaps instead of asking whether a signature is significant, ask what makes a signature significant

Proliferation is a hallmark of cancer

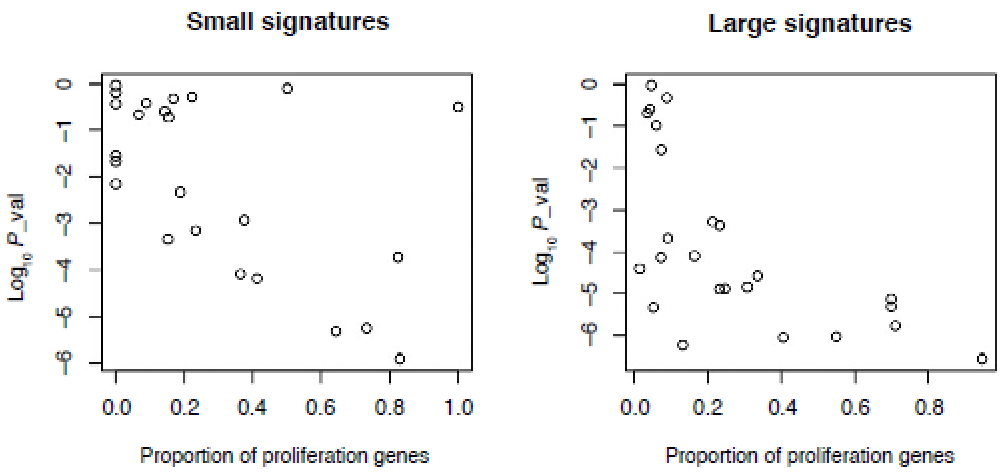
Hypothesis a la Venet et al.: **Proliferation-associated genes make a signature significant**

# of random signatures w/ ≥1 prolif gene

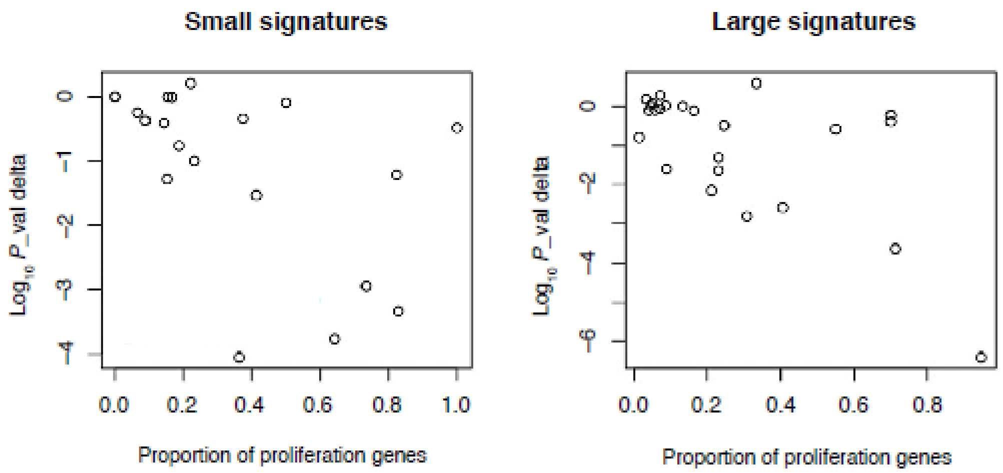


Cutoffs	Counts		
	NP	P	Marginals
Above 0.05	7043	19 043	26 086
Below 0.05	2766	19 148	21 914
Marginals	9809	38 191	48 000

Impact of proliferation genes on reported signatures



P-value of reported signatures, before removing proliferation genes



P-value of reported signatures, after removing proliferation genes

Many random signatures with proliferation genes are not significant;

Which proliferation genes make many random signatures significant?

Leverage background knowledge

Proliferation is a cancer hallmark

Good signatures with high diff in p-values before vs after removing proliferation genes:

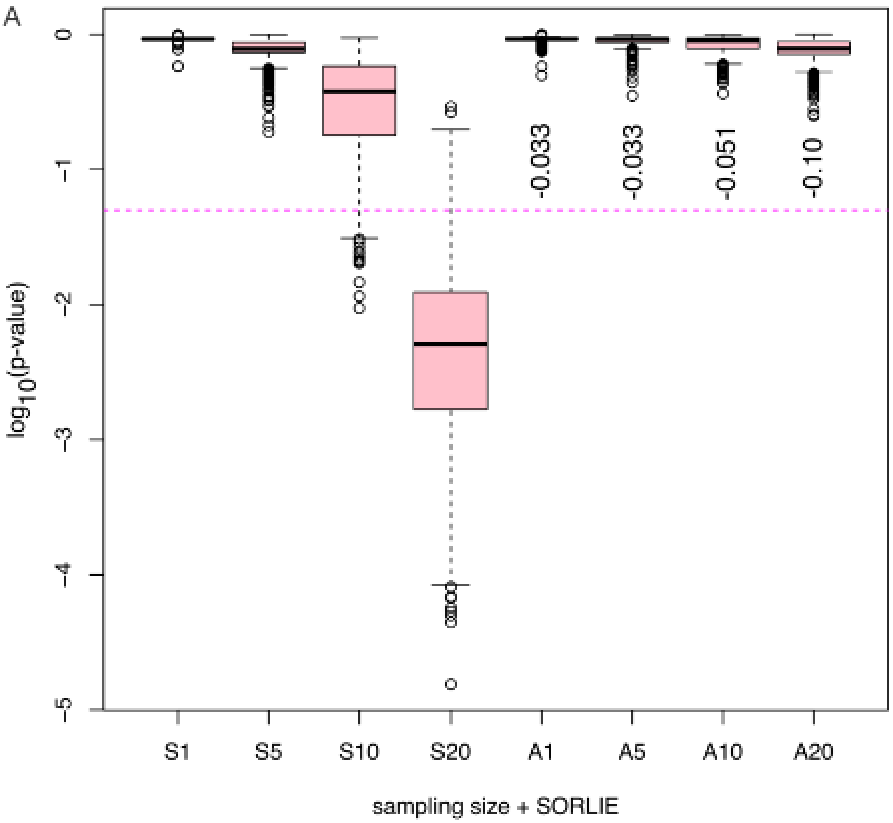
GLINSKY, DAI, RHODES, ABBA, WHITFIELD

SPS = { genes appearing in at least two of these good signatures }:

83 genes in total

81 of these are proliferation associated

Systematic evaluation



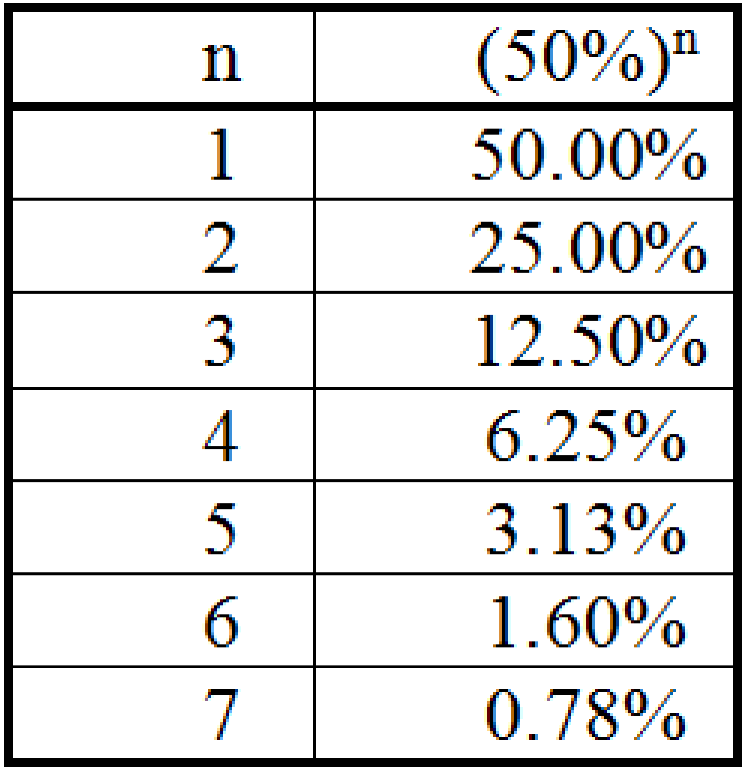
SPS genes show additive effect,

other proliferation genes don't

Test on many datasets

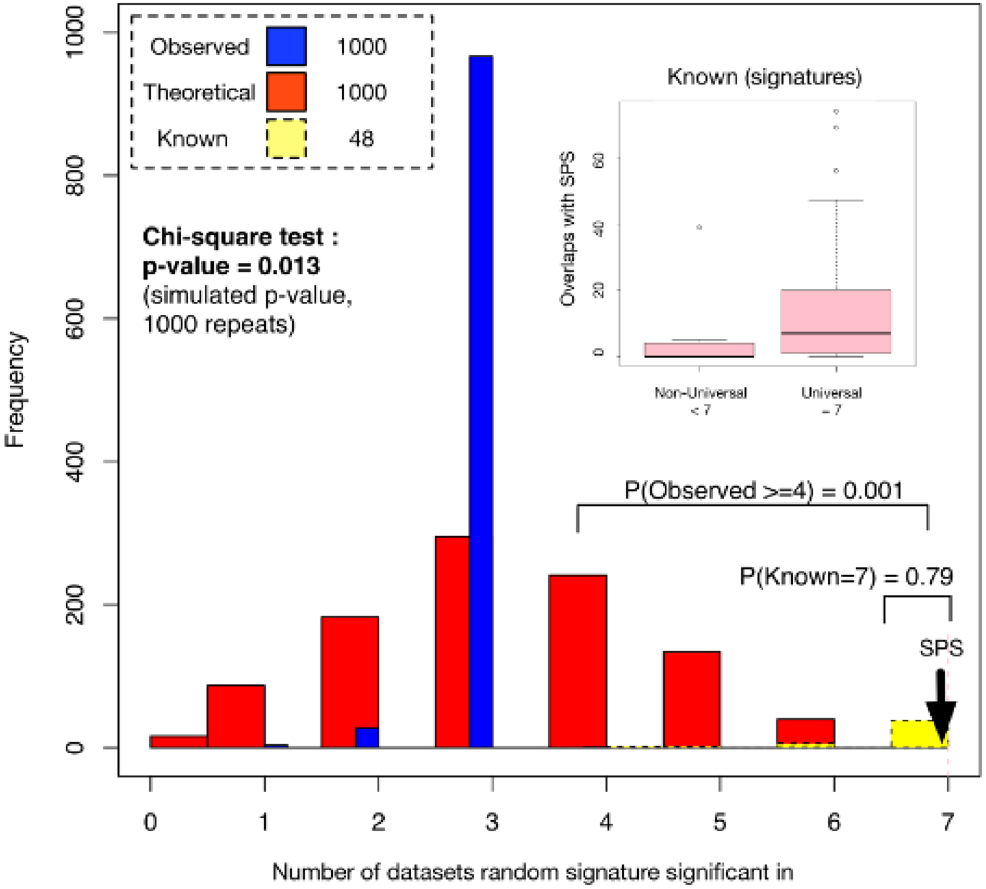
For any independent dataset, a random signature has ~50% chance to be significant in it

How many independent datasets are needed to avoid reporting random signatures as significant?



n	$(50\%)^n$
1	50.00%
2	25.00%
3	12.50%
4	6.25%
5	3.13%
6	1.60%
7	0.78%

Test on many datasets

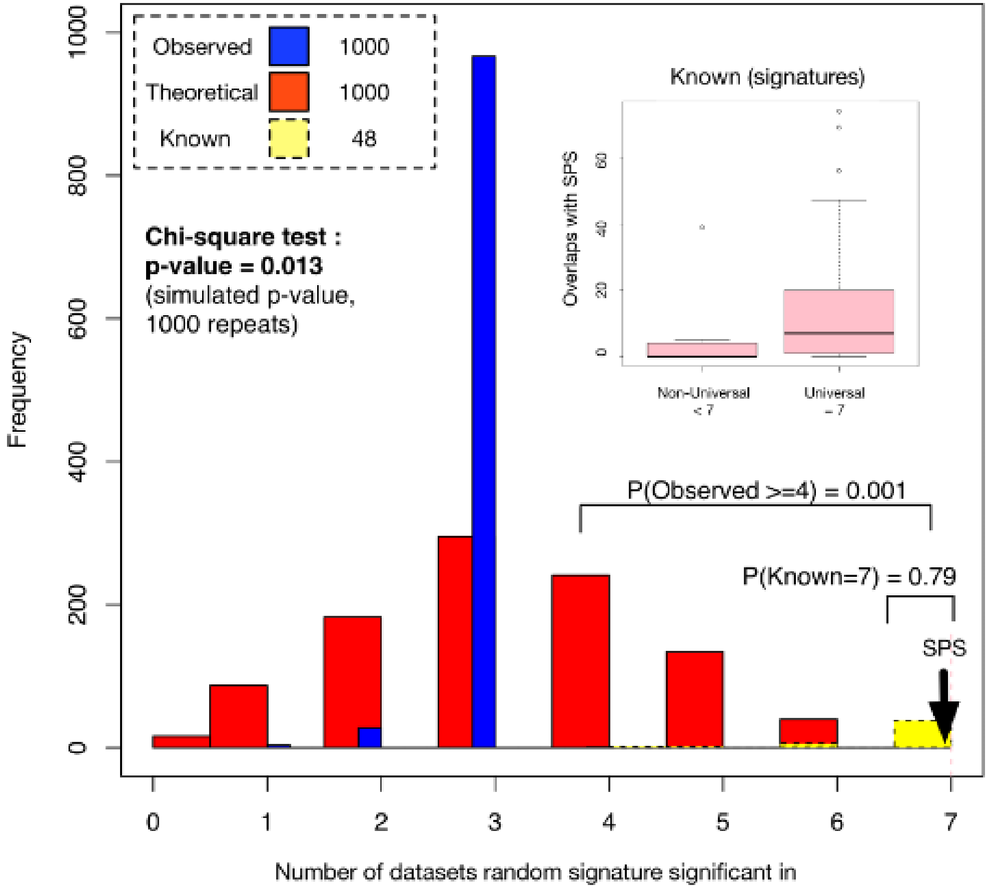


Chi-square test :
p-value = 0.013
(simulated p-value,
1000 repeats)

SPS is universally significant on 7 breast cancer datasets

Random signatures (same size as SPS) are hardly universal, even though they get better p-values than known signatures on some datasets

A theory-practice gap



~50% of random signatures are significant in 1 dataset

Red histogram is expected # of random signatures significant in n independent dataset (according to bionomial distribution)

Blue histogram is observed distribution

Closing remarks

Bewilderment: **Breast cancer survival signatures are no better than random signatures**

Enlightenment: **SPS genes**

Cautionary note 1: **Need to validate on many independent data sets**

Cautionary note 2: **Some independent data sets are not as independent as you think**

Goh & Wong. **Why breast cancer signatures are no better than random signatures explained**. *Drug Discovery Today*, 23(11):1818-1823, 2018
Goh & Wong. **Turning straw into gold: Building robustness into gene signature inference**. *Drug Discovery Today*, 24(1):31-36, 2019