LARVA-SAM: The relationship between “number of random variant datasets” (nrand) and *p*-value stability

~~Jor-El, Kal-El,~~ El-El

The subgroup formerly known as Annotation

One day late to the April Fools’ party ☹

Overview

- LARVA-SAM identifies significantly higher and lower patterns of recurrent variation relative to the pattern seen in random variant datasets
- Creation of random variant datasets is computationally expensive
- What is the lowest number of random variant datasets (*nrand*) to generate that is sufficient to achieve stable *p*-values in the LARVA-SAM significance test?

Methods

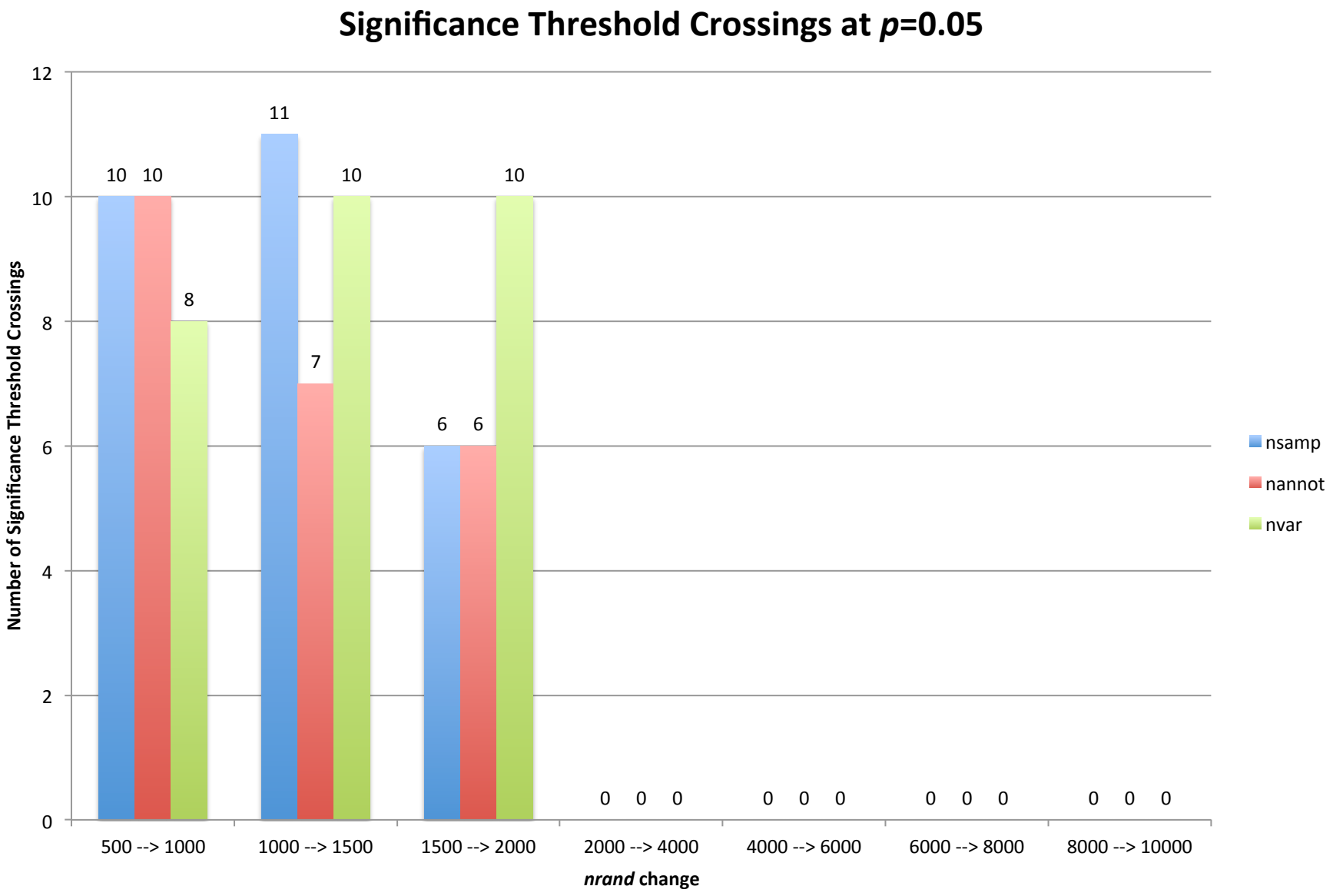
- Run the same query through LARVA-SAM at different *nrand* settings
  - LARVA-SAM(all prostate, KEGG)
- *nrand* range: 500 to 10,000
- Between consecutive runs r1 and r2, where r1.nrand < r2.nrand, look at how many "significance threshold crossings" there are

Methods with Example

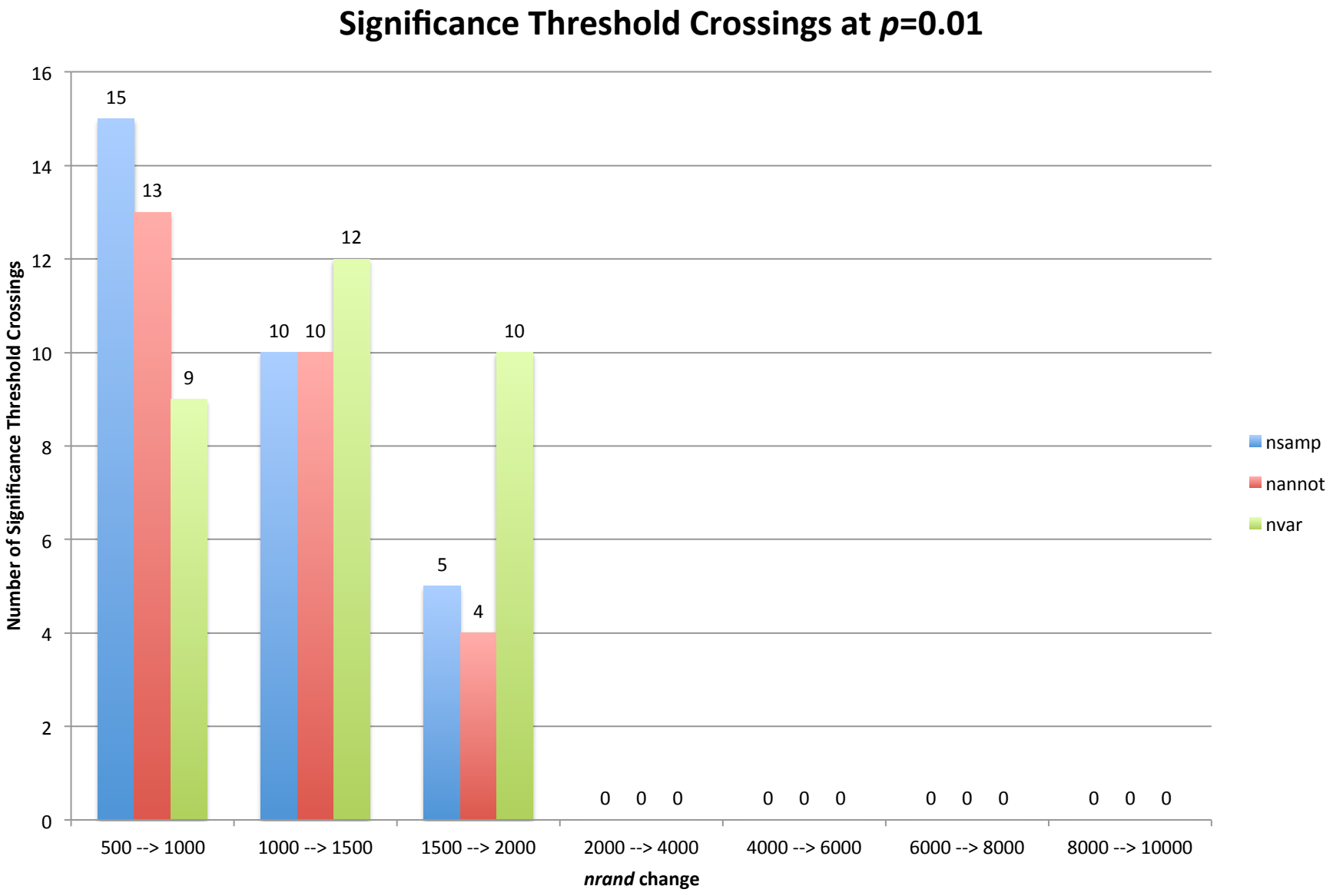
- Consider the following example data:

nrand	Pathway	Observed nsamp	Observed nannot	Observed nvar	nsamp expected avg	nsamp *p*-value	nannot expected avg	nannot *p*-value	nvar expected avg	nvar *p*-value
500	kegg_pathways_in_cancer.txt	154	30	8	1.15E+02	2.20E-08	1.70E+01	1.78E-04	1.25E-01	1.26E-125
	kegg_focal_adhesion.txt	131	20	1	9.60E+01	6.02E-31	6.38E+00	1.30E-09	0.25	0.04163226
	kegg_neuroactive_ligand_receptor_interaction.txt	122	31	3	1.40E+02	1.88E-03	40.75	0.0573091	1.25E-01	1.76E-18
1000	kegg_pathways_in_cancer.txt	154	30	8	1.16E+02	6.60E-10	1.69E+01	6.67E-04	2.50E-01	5.26E-44
	kegg_focal_adhesion.txt	131	20	1	9.27E+01	3.09E-08	5.94E+00	2.25E-11	1.88E-01	0.01868649
	kegg_neuroactive_ligand_receptor_interaction.txt	122	31	3	1.42E+02	1.21E-03	41.8125	0.01921396	6.25E-02	3.43E-34

- At *p*=0.05, the "kegg_neuroactive_ligand_receptor_interaction" pathway's "nannot *p*-value" changes from 0.0573091 to 0.01921396 when nrand is increased from 500 to 1000
  - In this data, there is one significance threshold crossing at *p*=0.05
- Do this analysis for all 185 KEGG pathways
  - 3 *p*-values produced for each pathway: nsamp, nannot, nvar (555 *p*-values total)
  - *nrand* values covered: 500, 1000, 1500, 2000, 4000, 6000, 8000, 10,000



nrand = 2000 is the sweet spot



nrand = 2000 is the sweet spot