Benchmarking of Mutational Signature Assignment Tools

Xi (Sam) Wang, Dr. Marcos Díaz-Gay, Dr. Raviteja Vangara,
Dr. Ludmil B. Alexandrov

UC San Diego

Somatic Mutations: changes in DNA

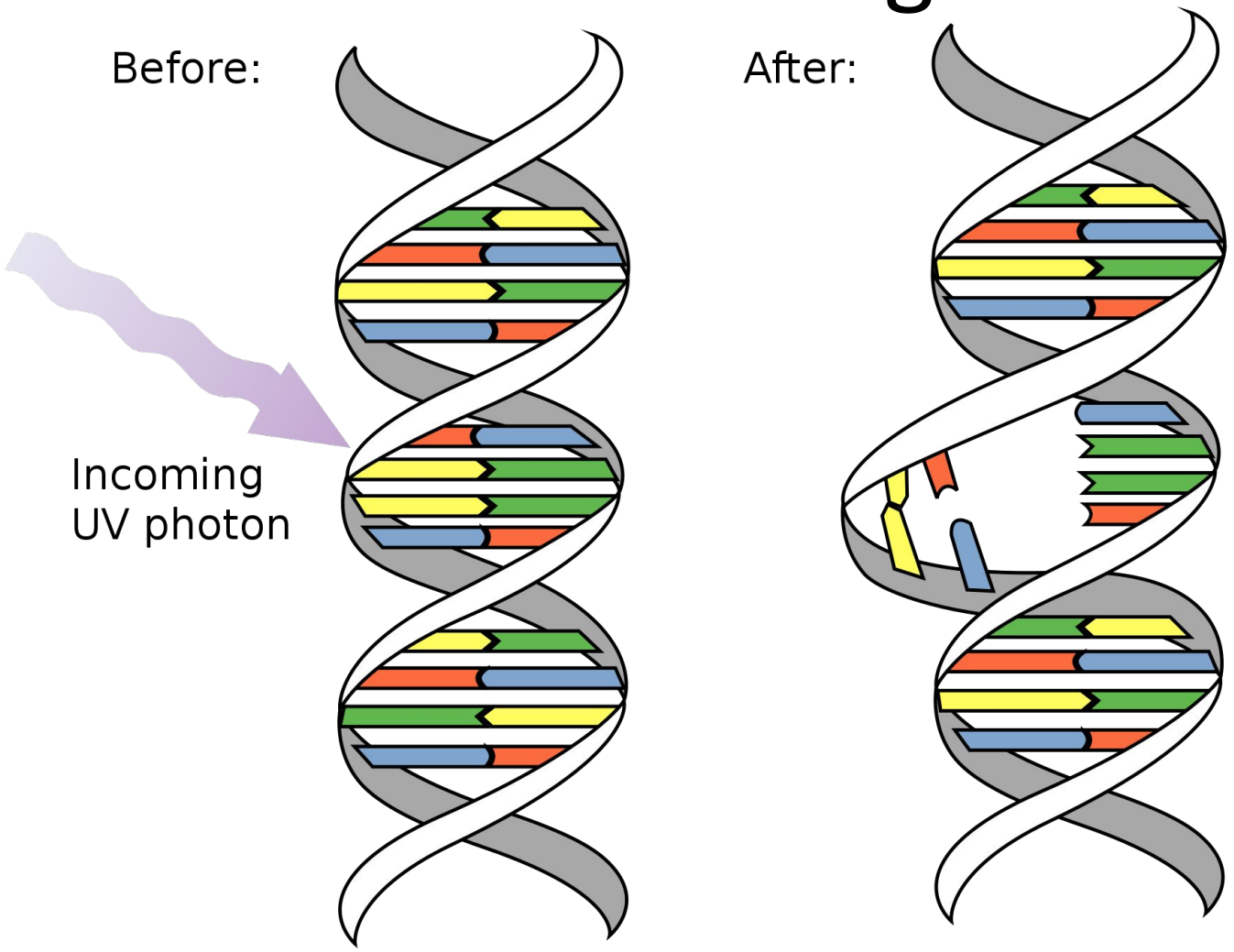
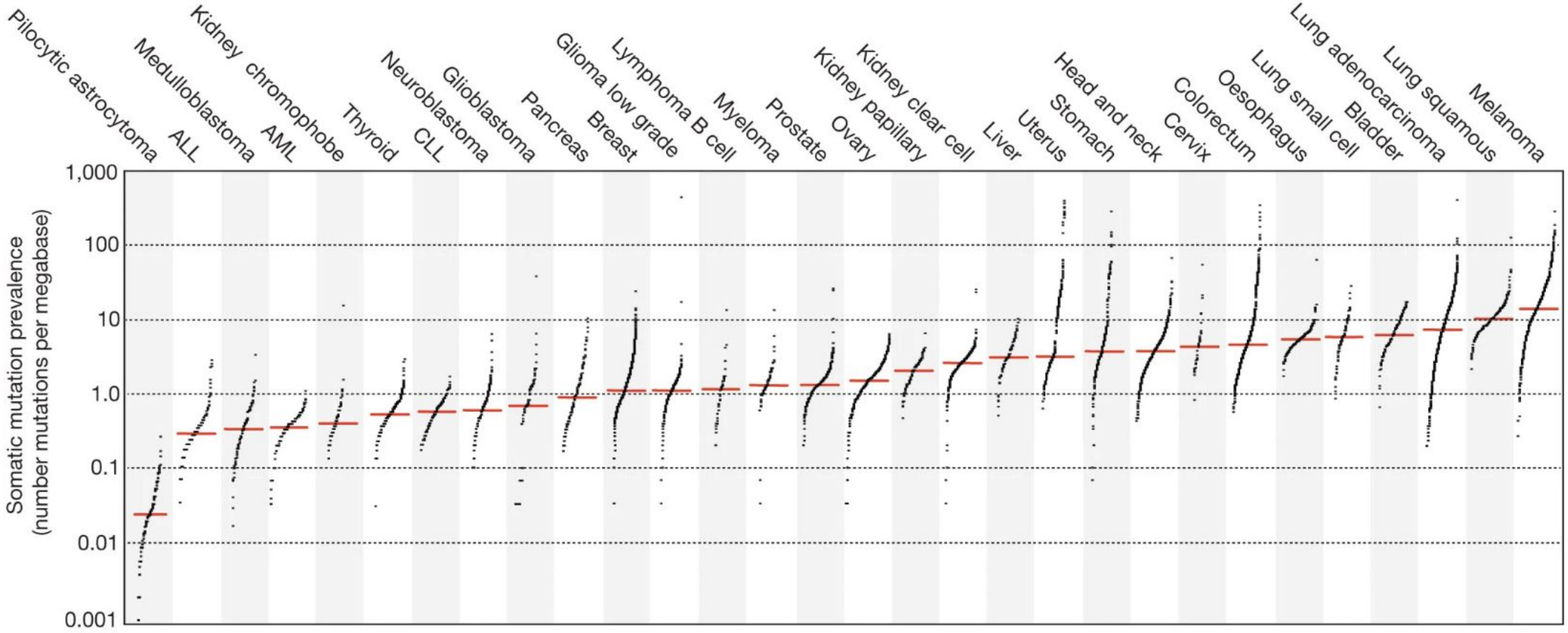


Figure from Wikipedia: somatic mutation

Somatic mutations are known to be generated by exposures

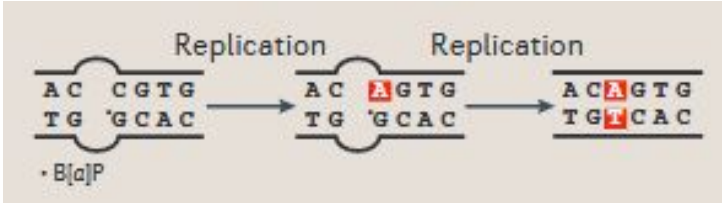


Alexandrov, L. B. *et al. Nature* **500**, 415–21 (2013).

Patterns of mutations are linked to sources of DNA damage

Environmental exposures

Tobacco smoking or chewing

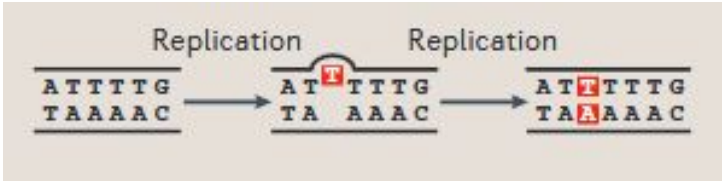


Normal cellular activities

Spontaneous deamination of methylated cytosines

Failure in DNA replication or repair

Aberrant mismatch repair pathway

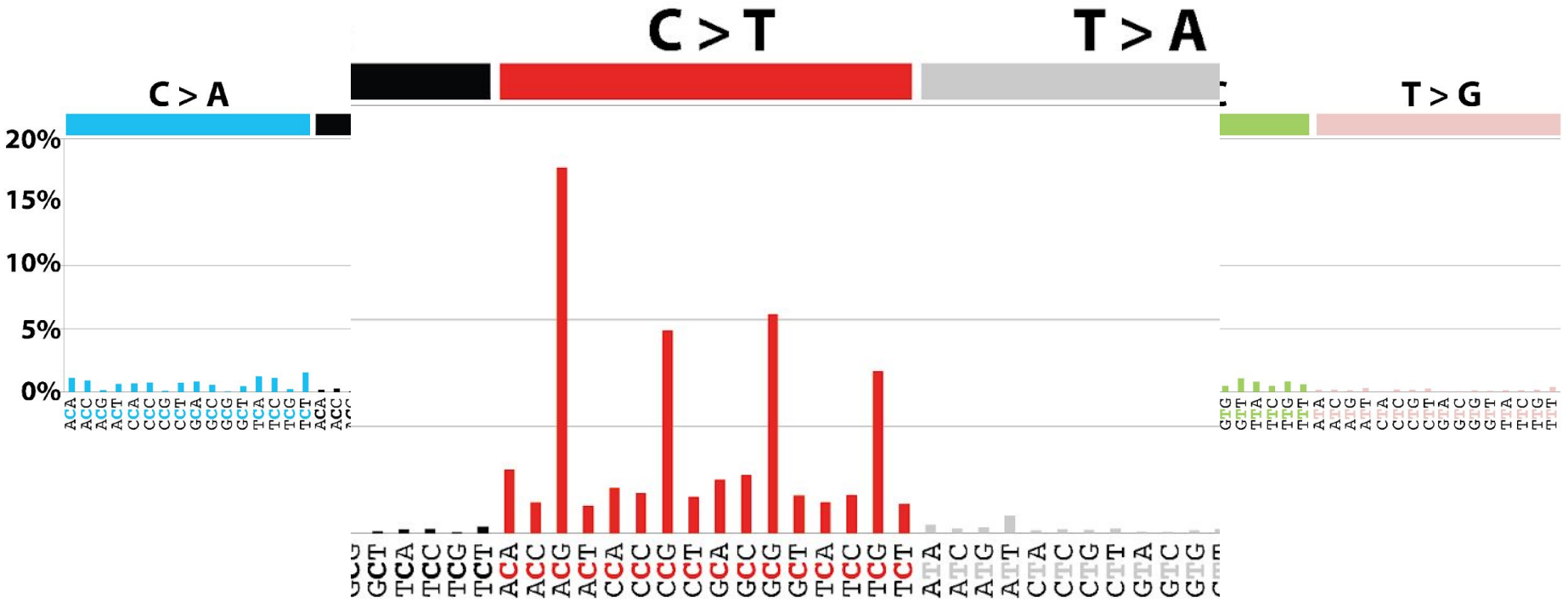


Helleday, T., Eshtad, S. & Nik-Zainal, S. *Nat. Rev. Genet.* **15**, 585–598 (2014).

Mutational Signature: defined by base substitutions and context

Six classes of single-base mutations
Reported by pyrimidine

Adding 5’ and 3’ adjacent bases
96 possibilities considering context



Tate, J. G. *et al. Nucleic Acids Res.* **47**, D941–D947 (2018). | https://cancer.sanger.ac.uk/cosmic/signatures_v2

Signature Assignment

- The MATRIX:

$$X = W \times H$$

Mutation Matrix (given)

Signatures Matrix (standard cosmic_v3)

Activities/Exposures Matrix (WHAT WE WANT!)

Mutation Type	Sample 1	Sample 2	…
A[C>A]A			
A[C>A]C			
…			
T[C>G]T			

=

Mutation Type	SBS1	SBS2	…
A[C>A]A			
A[C>A]C			
…			
T[C>G]T			

X

Signatures	Sample 1	Sample 2	…
SBS1			
SBS2			
…			
SBS60			

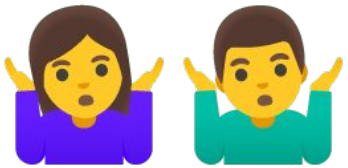
In other words:

Mutational signature assignment is the process of finding the most contributing factors associated with an individual's potential or existing cancer causes.

An accurate assignment means accurate discovery of causes of cancer!

Problem

There are no systematic benchmarks for all published mutational signature assignment tools.



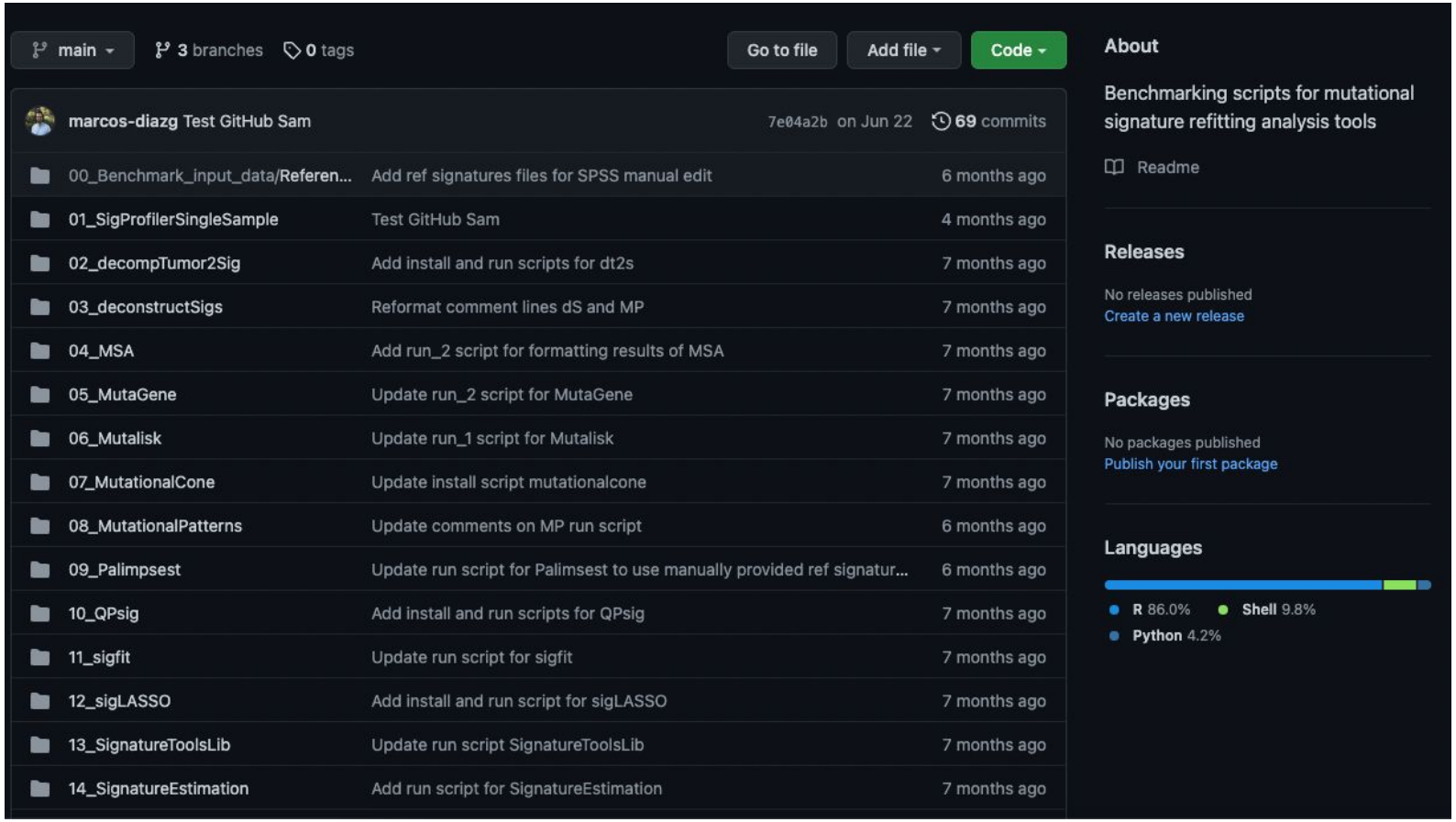
So, what do we do in this project: Benchmark them all!

Current published signature assignment tools:

- Our own tools at the Alexandrov Lab: SigProfilerSingleSample, SigProfilerExtractor (decomposition module)
- 15 other tools using following algorithms:
  - NNLS (non-negative least squares)
  - Multiple linear regression
  - Quadratic programming
  - Cone projection
  - Bootstrapping

How do we benchmark?

- Git! Automation scripts! Benchmarking performance scripts!
- Qualitative & Quantitative Performance Analysis



Current Progress?

Qualitative & Quantitative

Qualitative Analysis Method

- For the assignment results of each tool, we want:
  - True positive, true negative, false positive, false negative to get:
  - Average precision, sensitivity, and specificity across all scenarios per noise level
- A yes or no questions:

	SP.Syn.Panc-	SP.Syn.Panc-	SP.Syn.Panc-	SP.Syn.Panc-	SP.Syn.Panc-
SBS1	554.298921	1378.31487	827.105158	1116.7868	1576.23622
SBS2	204.743328	0	256.835317	61.8438701	0
SBS3	0	0	0	0	0
SBS5	4410.98461	2156.46037	7293.59929	2590.98303	1340.81445
SBS13	329.166888	133.140279	70.0235047	225.297873	108.256413
SBS17a	0	0	0	0	0
SBS17b	0	0	0	464.135025	0
SBS18	1894.12041	0	0	0	706.566553
SBS28	0	0	0	0	0
SBS30	0	0	475.873322	0	0
SBS40	0	2160.76295	2475.98966	0	0

Ground Truth

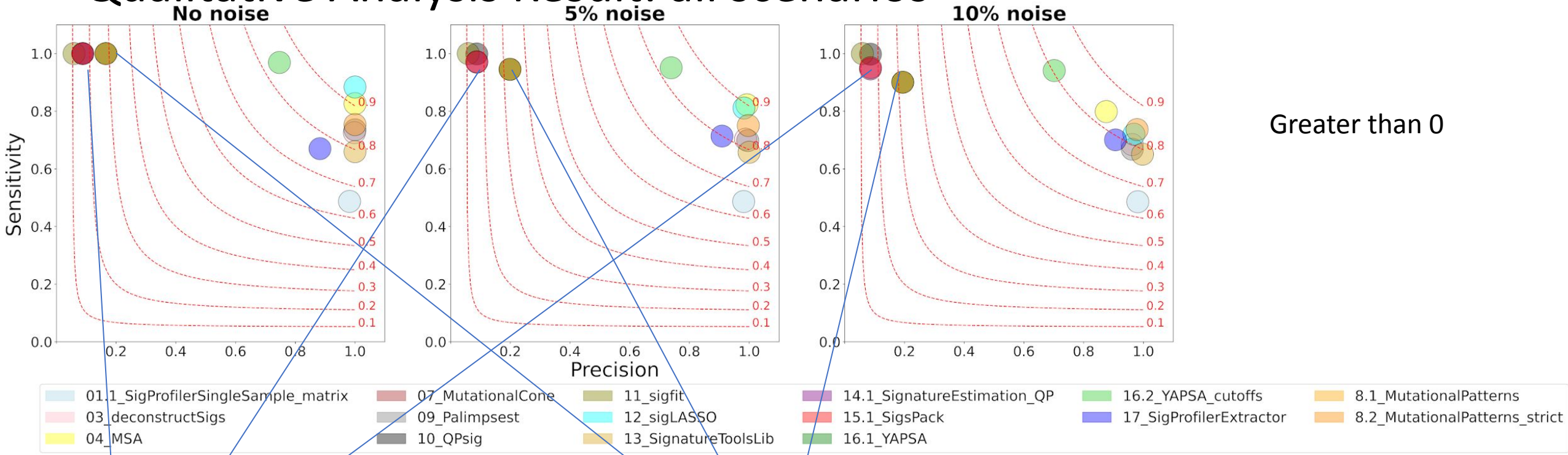
vs.

SP.Syn.Pan..	SP.Syn.Panc..	SP.Syn.Panc..	SP.Syn.Panc..	SP.Syn.Panc..	SP.Syn.Panc.Ac
SBS1	554.504682	1377.42657	824.560442	1116.68054	1569.43852
SBS2	0	0	0	0	0
SBS3	0	0	0	0	0
SBS5	4399.98309	2156.51546	7144.99565	2580.38002	1189.15106
SBS13	0	0	0	0	0
SBS17a	0	0	0	0	0
SBS17b	0	0	0	463.263627	0
SBS18	1893.90681	0	0	0	709.020081
SBS28	0	0	0	0	0
SBS30	0	0	0	0	0
SBS40	0	2124.91202	2455.60728	0	0
Unknown	545.605418	170.145944	976.83663	297.675817	259.390335

Assignment from 03_deconstructSigs

* Sample shown the scenario 2 without noise ground truth activities vs. activity results from tool 03_deconstructSigs

Qualitative Analysis Result: all scenarios

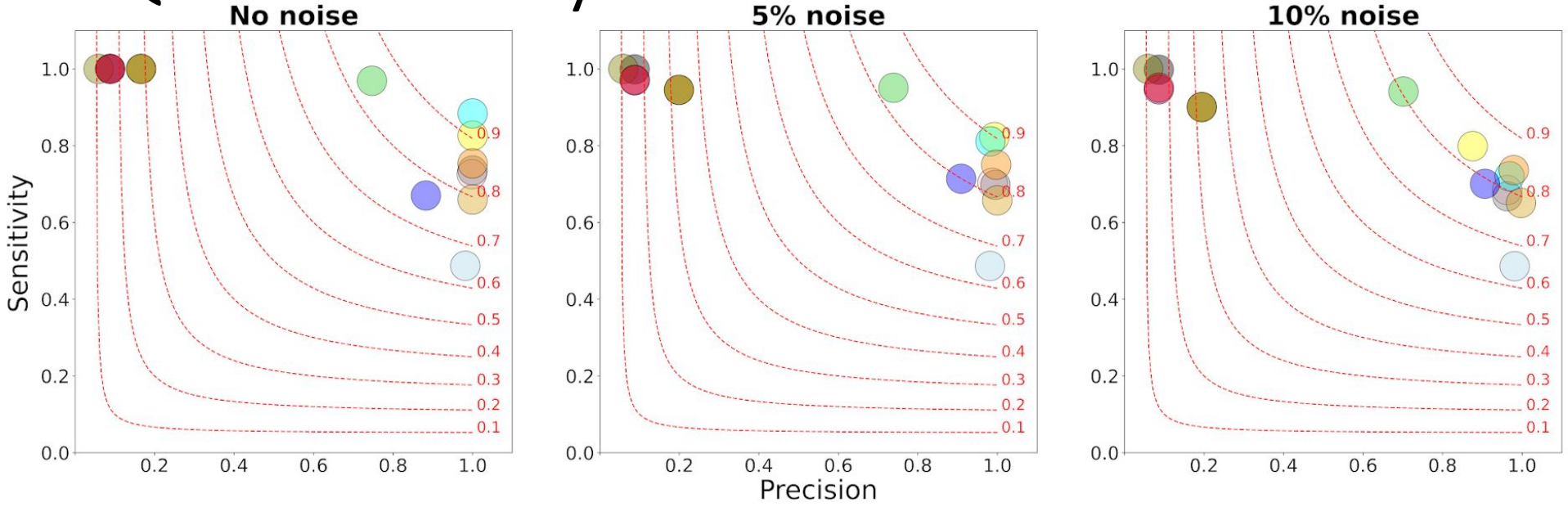


Greater than 0

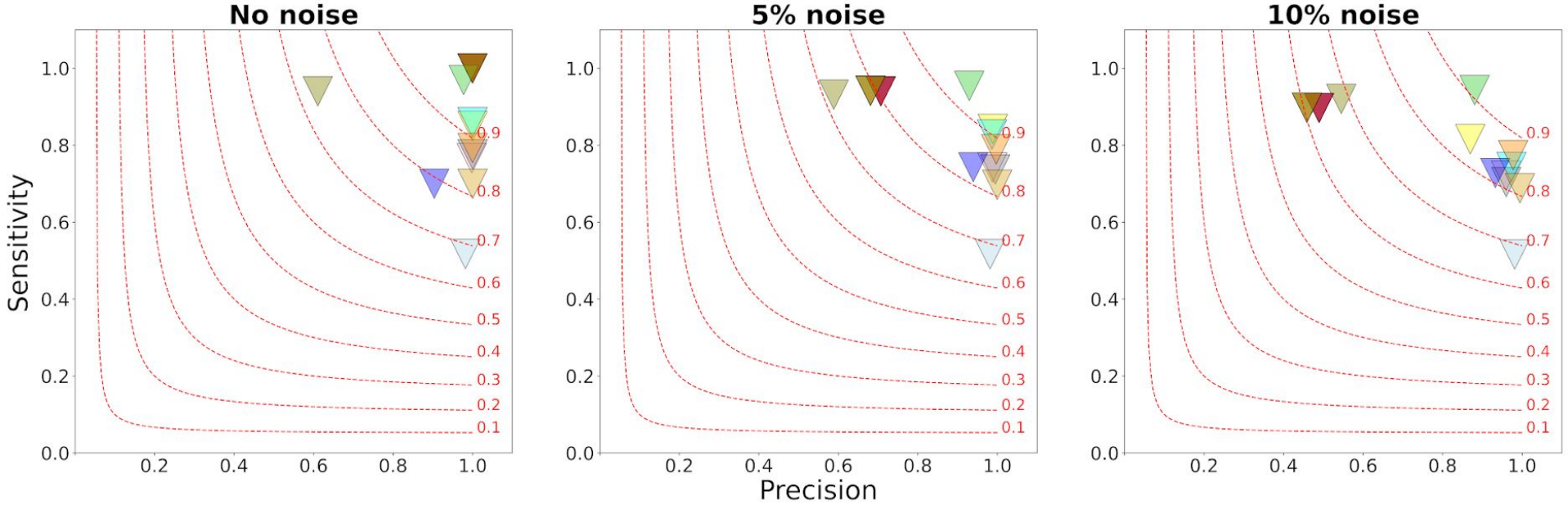
Cluster 1: QPSig, SigsPack, SignatureEstimation_QP

Cluster 2: MutationalPatterns, YAPSA (normal), MutationalCone

Qualitative Analysis Result: all scenarios



Greater than 0



Greater than 1% TMB

01.1_SigProfilerSingleSample_matrix
03_deconstructSigs
04_MSA
07_MutationalCone
09_Palimpsest
10_QPsig
11_sigfit
12_sigLASSO
13_SignatureToolsLib
14.1_SignatureEstimation_QP
15.1_SigsPack
16.1_YAPSA
16.2_YAPSA_cutoffs
17_SigProfilerExtractor
8.1_MutationalPatterns
8.2_MutationalPatterns_strict

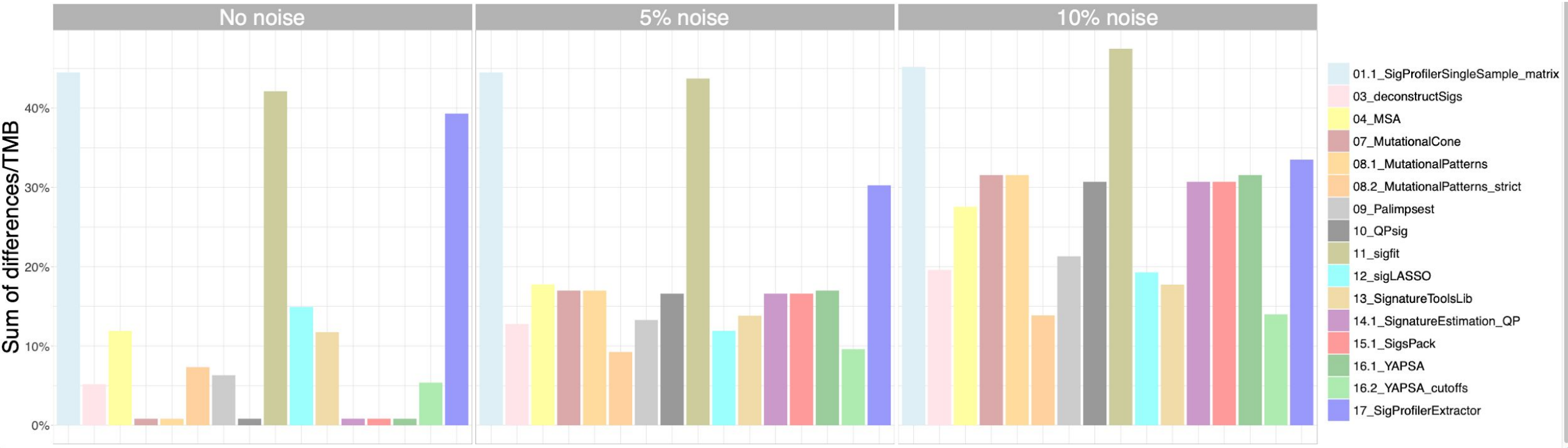
A step forward: Qualitative Analysis Method

- Simple!
- Compare the assignment results of each tools vs. ground truth
- By calculating:
  - Sum of absolute differences by TMB
    - percentage mutation mistakes detected by the tools
  - Average cos_sim scores
    - We are applying this on Activities matrix this time!

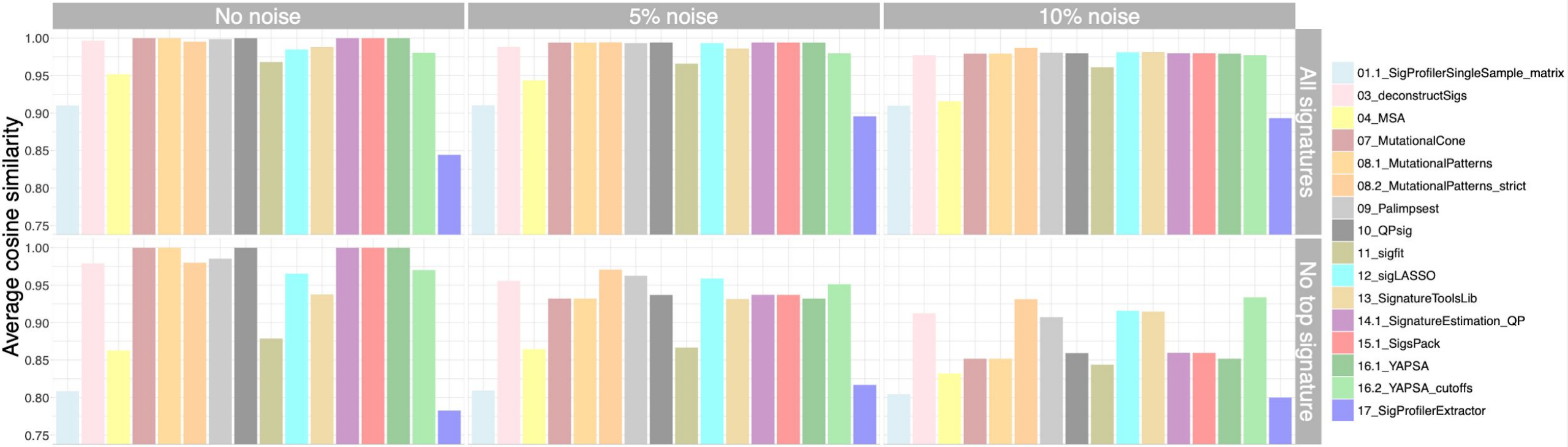
Sum of Absolute differences by TMB

	Ground Truth	Tool A	Abs. diff.	Abs. diff / TMB
SBS1	554.2989207	2000	1445.70108	19.6%
SBS2	204.7433281	200	4.74332809	0.1%
SBS3	0	0	0	0.0%
SBS5	4410.984607	4000	410.984607	5.6%
SBS13	329.1668878	300	29.1668878	0.4%
SBS17a	0	0	0	0.0%
SBS17b	0	0	0	0.0%
SBS18	1894.120409	1500	394.120409	5.3%
SBS28	0	0	0	0.0%
SBS30	0	0	0	0.0%
SBS40	0	1000	1000	13.5%
TMB	7393.314152			44.4%

Quantitative Analysis Result: sum of absolute differences by TMB



Qualitative Analysis Result: average cos_sim



Conclusion

- Consistent patterns are shown across all analysis
- All tools' accuracies decrease with noise in the samples
- Next step: need to do all the VCF tools and tools that require lots of time, finish run_time analysis, then possible finish benchmarking!

Future Impacts!

- Know where to improve
- Improve our own tool!
- Improve accuracy -> clinically applicable -> happy patients!

Thank you! Questions?