Benchmarking Sparse Variable Selection for Genomic Data

Samiran Sinha

Department of Statistics, Texas A&M University

Kollipara, Maiti, Chakraborty & Sinha, Statistics in Medicine (2026)

Companion: Fletcher & Sinha, PLOS One (2026)

ACKNOWLEDGEMENTS

Collaborators and funding

Hema Kollipara

PhD graduate

Michigan State University

First author, Paper 1

Westley Fletcher

MS student

Texas A&M University

First author, Paper 2

Sanjukta Chakraborty

Faculty

Texas A&M University

Co-author, Paper 1

Tapabrata Maiti

Faculty

Michigan State University

Co-author, Paper 1

Paper 1: Kollipara, Maiti, Chakraborty & Sinha · Paper 2: Fletcher & Sinha

Funding: NIH R21HL181700 (PI: Sinha), 09/15/2025–08/31/2027

Development of statistical and computational tools to analyze rhythmic gene expressions for a deeper understanding of circadian rhythms

Selecting a few signals out of thousands

1. **$p \gg n$**
   Hundreds of patients, thousands of genes or miRNAs

2. **Sparsity**
   Only a few features truly matter

3. **Correlation**
   Co-expressed features confuse selection

4. **Censoring**
   Survival outcomes are only partly observed

Paper 1 · linear outcome

- 13 methods compared
- New method: RFSFS

Paper 2 · survival outcome

- 10 methods compared
- New screening threshold

TCGA kidney cancer data: 2,049 miRNAs to 247

606
tumour and normal samples with gene expression

311
also have miRNA data (2,049 miRNAs)

247
miRNAs with no missing values

240
primary tumours: 81 women, 159 men

Model, fitted separately by sex

- Outcome: $\log_2$ MET expression
- Predictors: 247 miRNAs
- More predictors than patients

Data

- TCGA clear-cell kidney cancer (KIRC)
- Exploratory: one cohort, no ground truth

MET's prognostic effect differs by sex

Term	Hazard ratio (95% CI)	p
MET, in women	0.72 (0.52–1.01)	0.055
MET, in men	1.60 (1.09–2.36)	0.016
MET × sex	**2.22 (1.33–3.70)**	**0.002**

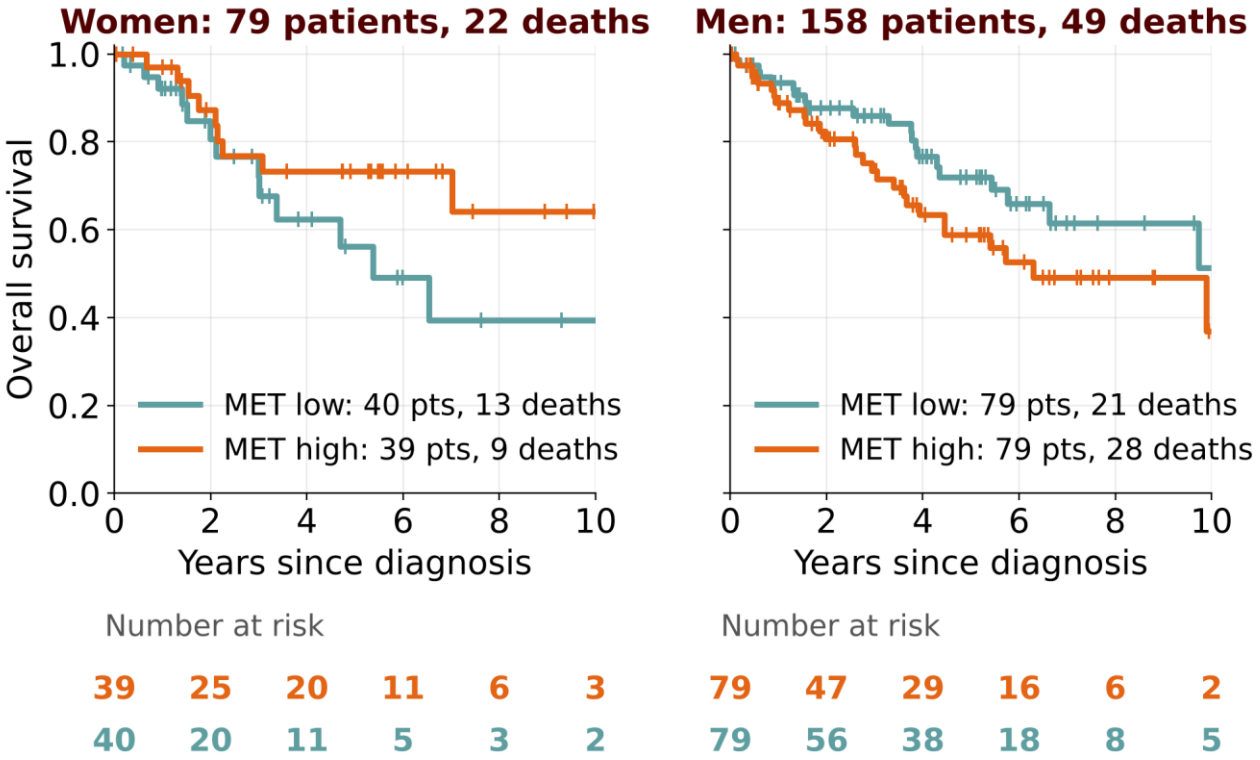
Group	Patients	Deaths
Women	79	22
Men	158	49

3 of 240 excluded: no follow-up time

Hazard ratio per $\log_2$ unit of MET. Outcome: overall survival in years from diagnosis.

Higher MET goes with shorter survival in men, but not in women.
The sex difference is significant (interaction p = 0.002).

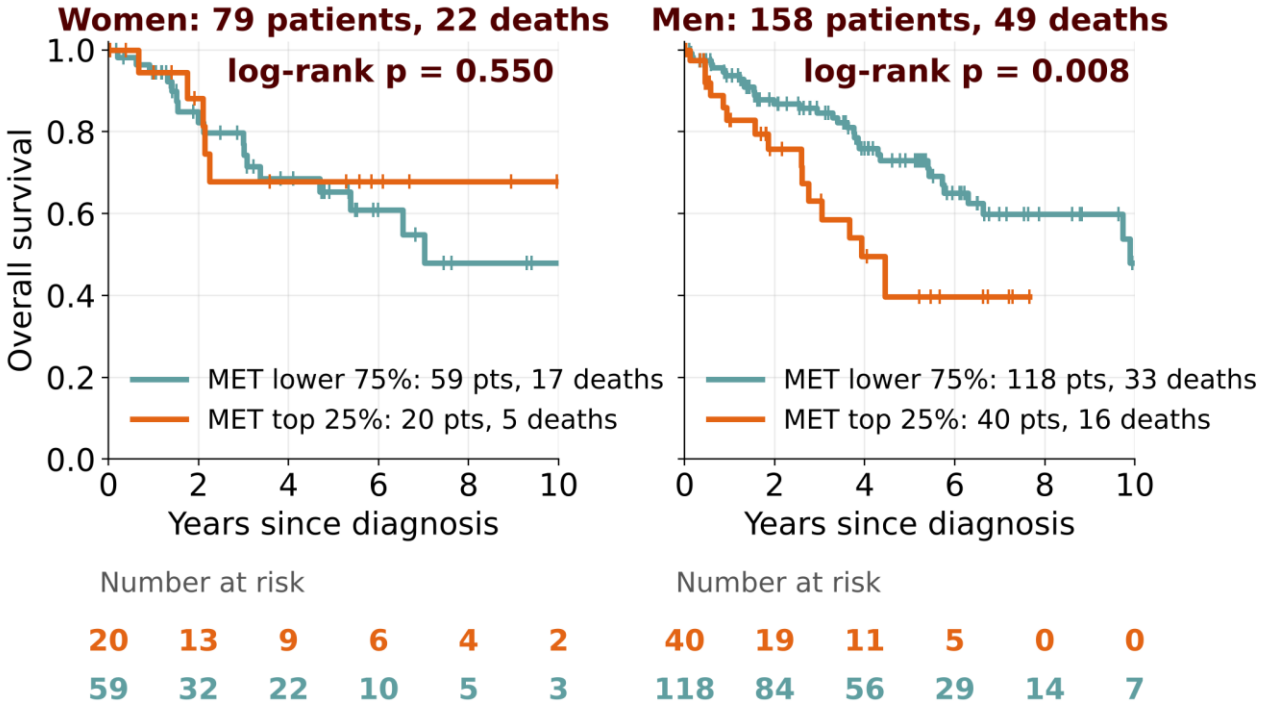
Survival by MET level in women and men



Test	Women	Men
Log-rank	0.14	0.11
FH early	0.18	0.10
FH late	0.11	0.26
FH middle	0.12	0.21

p-values, MET high vs low (median split). FH = Fleming–Harrington test.

Top quarter of MET: men fare worse



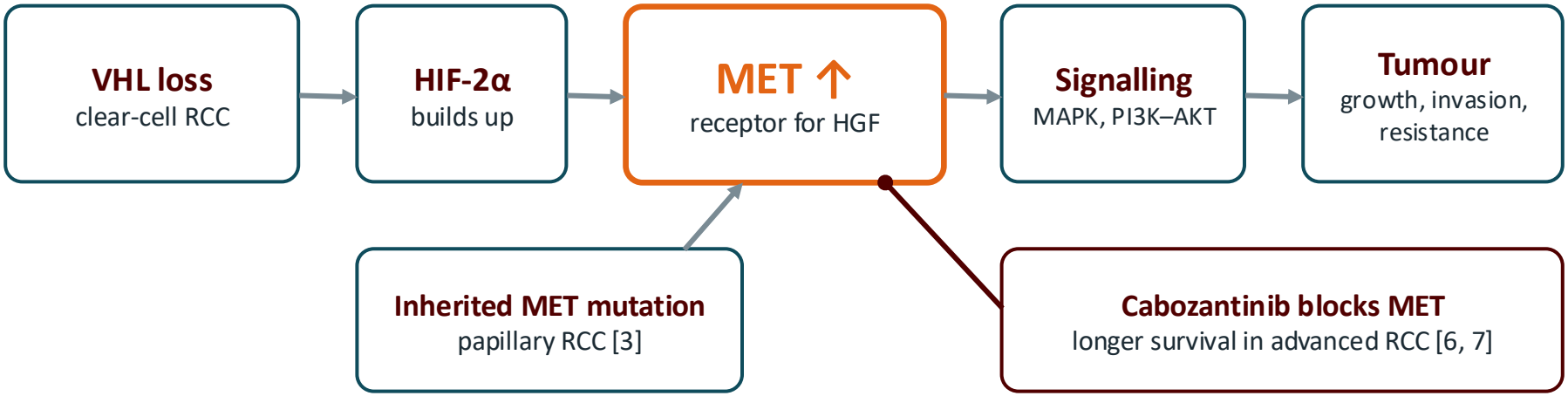
Cut	Women	Men
25th	0.13	0.89
Median	0.14	0.11
75th	0.55	**0.008**

Log-rank p-values by cut point.

Men, top 25%: HR 2.7

Cut chosen after looking: ×3 cuts gives p ≈ 0.02. Main evidence is the Cox model.

Why MET matters in kidney cancer



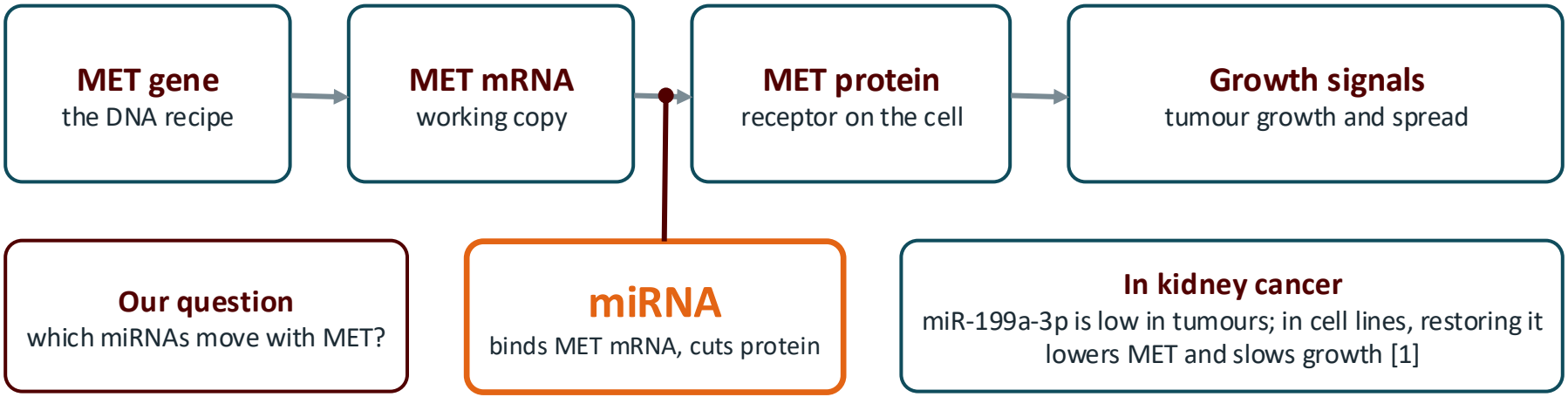
2.2×
higher in tumour than normal kidney (our TCGA data)

Worse
high MET: higher grade and stage, shorter survival [4, 5]

Prognostic
not an early-detection marker [5, 8]

[1] Mazumder, Cancers 2023 · [2] Huang, Tumor Biol 2014 · [3] Schmidt, Nat Genet 1997 · [4] Kim, Oncotarget 2017 · [5] Macher-Goeppinger, Oncotarget 2017 · [6] Choueiri, Lancet Oncol 2016 · [7] Pal, Lancet 2021 · [8] Kidney Cancer 2022 review

Why find the miRNAs that regulate MET?



Explain

why MET is high, and why men and women differ

Treat

a mimic replaces a lost miRNA; an anti-miR blocks a harmful one [2, 3]

Test first

delivery and safety are hurdles [4]; our leads need lab checks

[1] Huang, Tumor Biol 2014 · [2] Janssen, N Engl J Med 2013 · [3] van Zandwijk, Lancet Oncol 2017 · [4] Hong, Br J Cancer 2020

miRNAs linked to MET in men and women

Men (n = 159)

- miR-145-3p
- miR-29b-1-5p
- miR-183-5p
- miR-30b-5p
- miR-29c-5p

chosen by both SuSiE and RFSFS

In both sexes

miR-145-3p

Women (n = 81)

- No miRNA chosen by both SuSiE and RFSFS
- miR-145-3p: SuSiE + 8 others
- let-7e-3p: RFSFS + 7 others

What is already known about these miRNAs

miRNA	Selected in	Earlier finding in kidney cancer	Source
miR-145-3p	Men, women	Low in tumours that relapse	Slaby 2012
miR-29b-1-5p	Men	Higher in tumours; linked to survival	Xu 2015
miR-29c-5p	Men	Tumour suppressor; regulates enzyme-coding gene LOXL2	Qu 2018
miR-183-5p	Men	High level: shorter survival	Li 2019
miR-30b-5p	Men	Lower in tumours; suppresses invasion (GNA13)	Liu 2017
let-7e-3p	Women	Raised in patients' urine	Fedorko 2017

Most earlier work studied the miRNA family or its other arm, and none studied MET.

Earlier studies of these data did not examine MET

Earlier study	Question they studied	MET?	Our miRNAs reported?
TCGA Network 2013	Molecular profile; survival	No	No
Mlcochova 2016	Metastasis	No	miR-30b
Liang 2017	Diagnosis; survival	No	No
Li 2019	Survival	No	miR-183-5p
Zhan 2021	Survival model	No	No

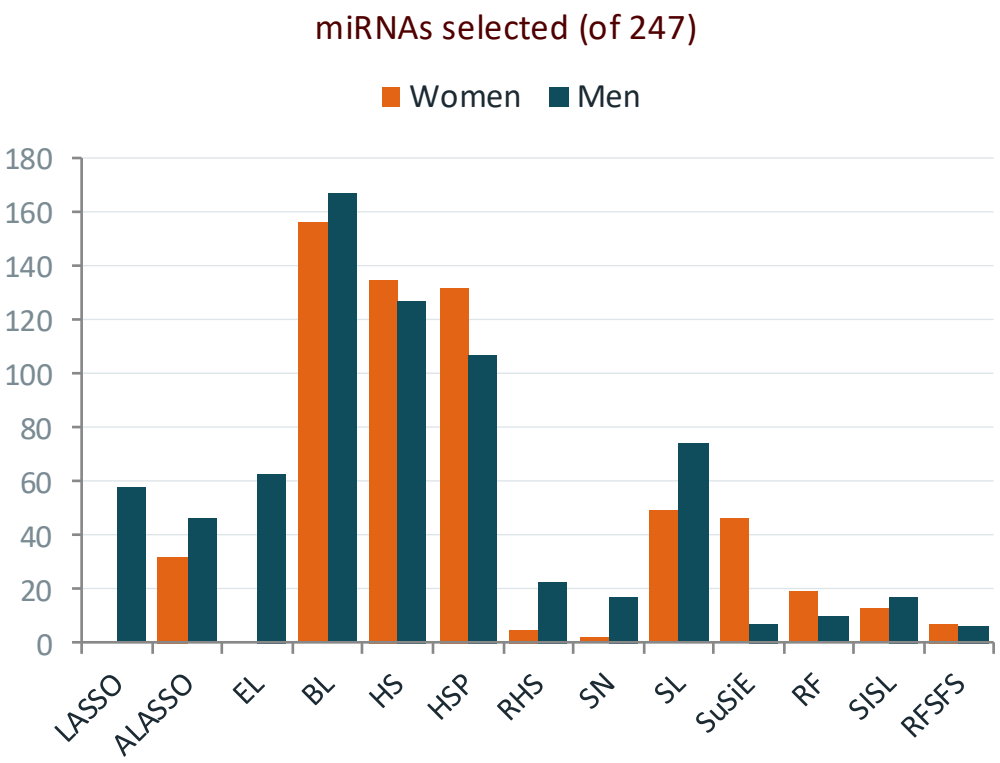
What is new

First to link miRNAs to MET in these data, and to analyse men and women separately.

Translational impact

- Panel to flag MET-high tumours
- Confirm in the laboratory

How many miRNAs each method selects



LASSO	Least absolute shrinkage and selection operator
ALASSO	Adaptive LASSO
EL	Elastic net
BL	Bayesian LASSO
HS	Horseshoe prior
HSP	Horseshoe+ prior
RHS	Regularized horseshoe
SN	Spike-and-slab, normal slab
SL	Spike-and-slab LASSO
SuSiE	Sum of single effects
RF	Random forest
SISL	Sure independence screening + LASSO
RFSFS	Forest screening + forward selection

How we compared the methods

Factor	Levels
Size	n = 200, p = 1000; 100 replicates
Sparsity	1%, 3%, 10%, 30% non-zero
Signal-to-noise	1 or 5
Errors	Normal or t_2 (heavy tails)
Features	Independent, binary, or correlated
Truth	Linear, or nonlinear

Metrics

- False discovery rate
- False negative rate
- F-score
- Prediction error
- Runtime
- Selection stability

Thirteen methods from five families

Penalized

LASSO, elastic net, adaptive LASSO

Global–local shrinkage

Bayesian LASSO [1]; horseshoe [2], horseshoe+, regularized horseshoe

Spike-and-slab

Normal slab (SN); LASSO slab (SL) [3]

SuSiE

Sum of single effects [4]

Screening-based

Random forest, SIS + LASSO [5], RFSFS

Prior density at $\beta = 0$

- Bayesian LASSO: sharp, but bounded
- Horseshoe: $\to +\infty$ as $\beta \to 0$

No exact zeros: select by clustering $|$posterior means$|$. SN and SuSiE: inclusion probability > 0.5.

[1] Park & Casella, JASA 2008 · [2] Carvalho, Polson & Scott, Biometrika 2010 · [3] Ročková & George, JASA 2018 · [4] Wang, Sarkar, Carbonetto & Stephens, JRSS-B 2020 · [5] Fan & Lv, JRSS-B 2008

RFSFS: screen with a forest, then select

- Fan and Lv (2008): screen by marginal correlation, then select with LASSO.
- Marginal correlation misses joint and nonlinear effects, so RFSFS screens with a random forest.

1 Random forest
Rank features by minimum depth

2 Screen
Keep the top n / log n

3 Select
Forward selection with BIC

Caveat

No consistency proof yet

Evidence

Best false-discovery and prediction ranks with correlated features

Fan and Lv (2008): sure independence screening

- Problem: when p grows exponentially with n, penalized methods become slow and unstable.
- Idea: keep the $d < n$ features most correlated with Y, then run LASSO or a similar method.
- Limit: misses features related to Y only jointly; iterated screening helps.

Theory (sure screening): P(all true features kept) → 1

Rate $1 - O(\exp\{-C\, n^{(1-2\kappa)} / \log n\})$; needs Gaussian errors, $\min |\beta_j| \geq c\, n^{(-\kappa)}$, $\lambda_{\max}(\Sigma) \leq c\, n^{\tau}$ with $2\kappa + \tau < 1$.

So the true number of features must be $s = O(n^{(2\kappa+\tau)})$, far fewer than the n / log n kept.

Fan J, Lv J. Sure independence screening for ultrahigh dimensional feature space. J R Stat Soc B 2008;70:849–911.

Who wins depends on the setting

Very sparse, strong

HS, RHS, SuSiE perfect (F-score 1.0)

Denser signals

Adaptive LASSO best (F-score)

Heavy-tailed errors

RFSFS 1st when correlated (F-score)

Prediction

RFSFS 1st–2nd in 9 of 11 (prediction error)

Failure modes

SuSiE selects nothing in hard cases (FNR = 1)

Nonlinear truth

All methods poor (F-score < 0.30)

Nonlinear truth: every method struggles

$$Y = \sin(\pi X_1 X_2) + X_3^2 + \log(|X_4| + 1) - X_5 X_6 + \varepsilon$$ n = 200, p = 1,000

Why it is hard

- All 13 methods assume linear effects
- Two pure interactions: X_1X_2, X_5X_6
- Two symmetric terms: X_3^2, $\log(|X_4| + 1)$
- No term is correlated with any single X_j: the best linear fit is flat

Method	F-score	FDR rank
SN (spike-and-slab)	**0.29**	**1**
LASSO	0.28	4
RFSFS	0.26–0.27	3
HSP (horseshoe+)	0.26–0.27	2
SuSiE	selects none	13

Best: SN. Yet all F-scores < 0.30, FNR > 0.66.

Kollipara et al. 2026, Section 4.4.6 and Table 4. SN was given the true model size.

Selection accuracy across replications

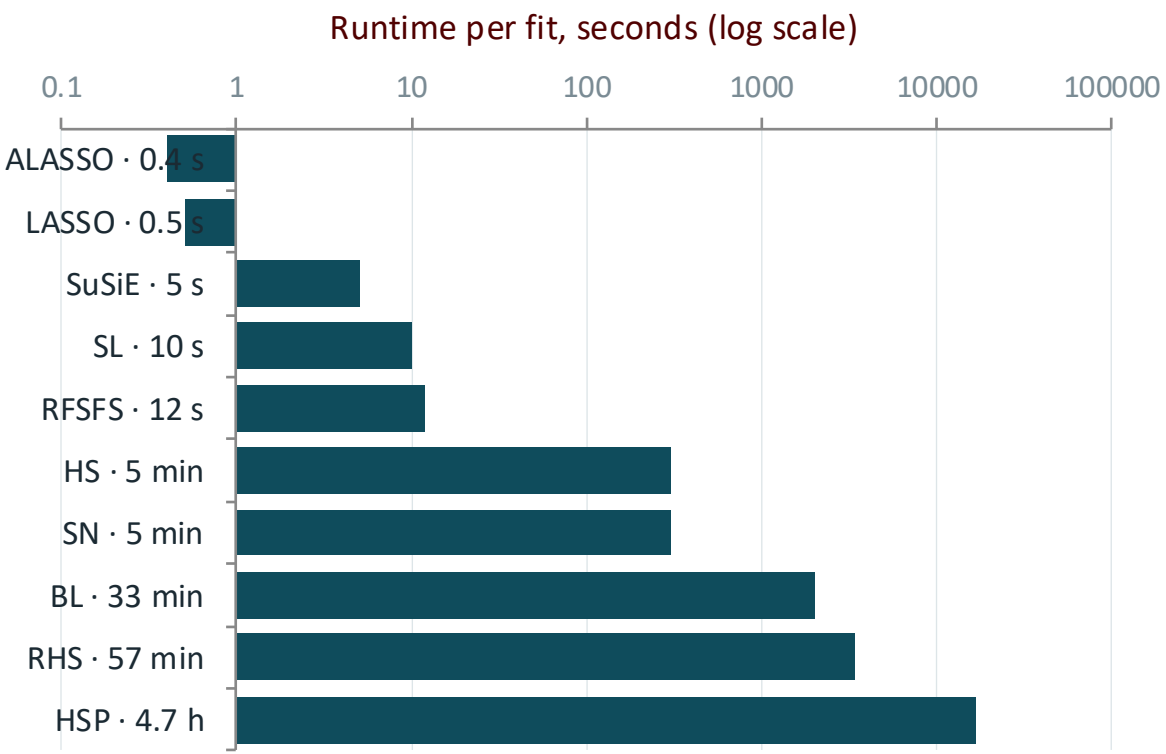
Model: $Y = X\beta + \varepsilon$; $n = 200$, $p = 1{,}000$; $\beta = \pm 1$ on $q = 10$–300 true features, else 0; $\varepsilon \sim t_2$; $X \sim N(0, \Sigma)$

Method	1%	3%	10%	30%	Noise
RFSFS	0.90	0.99	1.00	1.00	0.00
SISL	0.83	0.81	1.00	1.00	0.00
RF	0.52	0.76	1.00	1.00	0.00
SuSiE	0.76	0.78	0.75	0.75	0.10
LASSO	0.89	0.84	0.70	1.00	0.02
EL	0.89	0.82	0.66	0.98	0.03
ALASSO	0.52	0.62	0.62	1.00	0.03
SL	0.38	0.52	0.80	1.00	0.02
HS	0.71	0.57	0.34	0.39	0.26
RHS	0.80	0.70	0.41	0.43	0.25
SN	0.64	0.36	0.22	0.21	0.33

Accuracy

- Proportion of 100 replications in which a feature is selected
- Columns 1%–30%: median over true features (higher is better)
- Noise: median over noise features at 30% (lower is better)
- Σ: correlation 0.3 true–true, 0.5 true–noise, 0.8 noise–noise

Optimization runs far faster than MCMC



Fast and good

- SuSiE: seconds
- RFSFS: seconds

Slow

- MCMC priors: minutes to hours

What we recommend

Your setting	Use	Evidence
Correlated features typical genomic data	**RFSFS; SuSiE**	RFSFS: best FDR rank (2.0); both most stable
Independent features	**SN to find markers; RHS to avoid false ones**	SN: best F-score rank (2.2); RHS: best FDR rank (2.2)
Very sparse, strong signals 1% active, high SNR	**Horseshoe priors; SuSiE**	F-score ≈ 1.0, but slow
Denser signals 10%–30% active	**Adaptive LASSO**	Ranked 1st–3rd
Goal: prediction	**RFSFS; SIS + LASSO**	Lowest prediction error

No single winner. Trust markers that several methods agree on. Nonlinear truth: no method is adequate.

Selecting biomarkers for survival outcomes

Question and methods

- Which methods find true biomarkers for survival?
- Penalized Cox regression (LASSO, elastic net, adaptive LASSO)
- Boosting; random survival forests
- Filters: false discovery rate, q-value, correlation-adjusted scores

Data and metrics

- Simulations: 18 settings × 200 replicates
- Realistic simulation from bladder cancer genes
- Real data: 423 bladder cancer patients
- False discovery rate, F1, concordance, Brier score

CoxBoost: grow a Cox model one gene at a time

1 Start empty
All gene coefficients $\beta = 0$. Clinical covariates, if any, are fitted without penalty.

2 Try each gene
One penalized Newton step from the current fit: $\gamma_j = U_j / (I_j + \lambda)$, with score U and information I.

3 Update the best
Pick the gene with the largest $U_j^2 / (I_j + \lambda)$; update only it: $\beta_j \leftarrow \beta_j + \gamma_j$.

4 Stop early
Repeat for M steps, with M chosen by cross-validation. Genes never picked stay at 0, so the model is sparse.

Role of the penalty λ

- Large λ: small, cautious steps
- Set so that M > 50 steps
- Package default: λ = 9 × number of deaths

In our survival study

- Lowest false discovery rate (0.26) with realistic correlation

Binder H, Schumacher M. BMC Bioinformatics 2008;9:14 (CoxBoost) · Tutz G, Binder H. Biometrics 2006;62:961–971 (likelihood-based boosting)

What the survival benchmark shows

Simple simulation
n = 300, p = 1,000; 18 settings × 200 data sets

Metrics
False discovery rate, F1-score; concordance, Brier score, error of log survival time

Adaptive LASSO
ranked 1st–4th on every metric

Realistic simulation
bladder cancer genes: n = 423, p = 3,000, 53 true genes

Metrics
Same five metrics

CoxBoost: fewest false discoveries
false discovery rate 0.26; p-value filters fail (0.97)

Real bladder cancer data
423 patients; 30 genes known from literature

Metrics
Concordance, Brier score

All methods predict modestly
concordance 0.49–0.59 (0.5 = chance)

Adaptive LASSO is the best all-rounder. Same lesson as our study: no universal winner.

THANK YOU

Questions and discussion

Kollipara HSS, Maiti T, Chakraborty S, Sinha S. Benchmarking sparse variable selection methods for genomic data analyses. Stat Med 2026;45:e70428. doi:10.1002/sim.70428

Fletcher W, Sinha S. Benchmark of biomarker identification and prognostic modeling methods on diverse censored data. PLoS One 2026;21(6):e0351429. doi:10.1371/journal.pone.0351429

Code: github.com/Kollipara-Hema/BayesianVarSelGenomics · github.com/WesLFletch/censored-data-regression-benchmark

References

Binder H, Schumacher M. Allowing for mandatory covariates in boosting estimation of sparse high-dimensional survival models. BMC Bioinformatics 2008;9:14.

Carvalho CM, Polson NG, Scott JG. The horseshoe estimator for sparse signals. Biometrika 2010;97:465–480.

Choueiri TK, et al. Cabozantinib versus everolimus in advanced renal cell carcinoma (METEOR): final results. Lancet Oncol 2016;17:917–927.

Choueiri TK, et al. Efficacy of savolitinib vs sunitinib in MET-driven papillary renal cell carcinoma (SAVOIR). JAMA Oncol 2020.

Fan J, Lv J. Sure independence screening for ultrahigh dimensional feature space. J R Stat Soc B 2008;70:849–911.

Fedorko M, et al. Detection of let-7 miRNAs in urine supernatant… non-metastatic ccRCC. Biochem Med 2017;27:411–417.

Hong DS, et al. Phase 1 study of MRX34, a liposomal miR-34a mimic, in advanced solid tumours. Br J Cancer 2020;122:1630–1637.

Huang J, et al. miR-199a-3p inhibits hepatocyte growth factor/c-Met signaling in renal cancer carcinoma. Tumor Biol 2014;35:5833–5843.

Janssen HLA, et al. Treatment of HCV infection by targeting microRNA (miravirsen). N Engl J Med 2013;368:1685–1694.

Kim JH, Kim BJ, Kim HS. Clinicopathological impacts of high c-Met expression in renal cell carcinoma: a meta-analysis. Oncotarget 2017;8:75478–75487.

Li H, et al. Upregulation of miR-183-5p predicts worse survival in patients with renal cell cancer after surgery. Cancer Biomark 2019.

Liang B, Zhao J, Wang X. A three-microRNA signature as a diagnostic and prognostic marker in clear cell renal cancer. PLoS One 2017;12:e0180660.

References (continued)

Liu W, Li H, Wang Y, Zhao X, Guo Y, Jin J, Chi R. miR-30b-5p suppresses renal cell carcinoma by targeting GNA13. Gene 2017. PMID: 28536082.

Macher-Goeppinger S, et al. MET expression and copy number status in clear-cell renal cell carcinoma. Oncotarget 2017;8:1046–1057.

Mazumder S, Higgins PJ, Samarakoon R. Downstream targets of VHL/HIF-α signaling in renal clear cell carcinoma progression. Cancers 2023;15:1316.

miR-500a-3p is a potential prognostic biomarker in hepatocellular carcinoma. Int J Gen Med 2022. doi:10.2147/IJGM.S340629.

Mlcochova H, et al. EMT-associated microRNA/mRNA signature is linked to metastasis and prognosis in ccRCC. Sci Rep 2016;6:31852.

Müller S, Nowak K. Exploring the miRNA-mRNA regulatory network in ccRCC by NGS expression profiles. BioMed Res Int 2014;948408.

Pal SK, et al. A comparison of sunitinib with cabozantinib, crizotinib, and savolitinib for metastatic papillary RCC (PAPMET). Lancet 2021;397:695–703.

Park T, Casella G. The Bayesian Lasso. J Am Stat Assoc 2008;103:681–686.

Peired AJ, et al. Sex and gender differences in kidney cancer: clinical and experimental evidence. Cancers 2021;13:4588.

Qu Y, et al. Upregulation of MIAT regulates LOXL2 expression by competitively binding miR-29c in ccRCC. Cell Physiol Biochem 2018;48:1075–1087.

References (continued)

Ročková V, George EI. The spike-and-slab LASSO. J Am Stat Assoc 2018;113:431–444.

Schmidt L, et al. Germline and somatic mutations in the tyrosine kinase domain of the MET proto-oncogene in papillary renal carcinomas. Nat Genet 1997;16:68–73.

Slaby O, et al. Identification of microRNAs associated with early relapse after nephrectomy in RCC patients. Genes Chromosomes Cancer 2012;51:707–716.

TCGA Research Network. Comprehensive molecular characterization of clear cell renal cell carcinoma. Nature 2013;499:43–49.

The current status of kidney cancer urine markers: a systematic review. Kidney Cancer 2022.

Tutz G, Binder H. Generalized additive modeling with implicit variable selection by likelihood-based boosting. Biometrics 2006;62:961–971.

van Zandwijk N, et al. Safety and activity of microRNA-loaded minicells in malignant pleural mesothelioma. Lancet Oncol 2017;18:1386–1396.

Wang G, Sarkar A, Carbonetto P, Stephens M. A simple new approach to variable selection in regression, with application to genetic fine mapping. J R Stat Soc B 2020;82:1273–1300.

Xu Y, et al. Expression and functional role of miR-29b in renal cell carcinoma. Int J Clin Exp Pathol 2015;8:14161–14170.

Zhan Y, et al. A microRNA-clinical prognosis model to predict overall survival for kidney renal clear cell carcinoma. Cancer Med 2021;10:6128–6139.