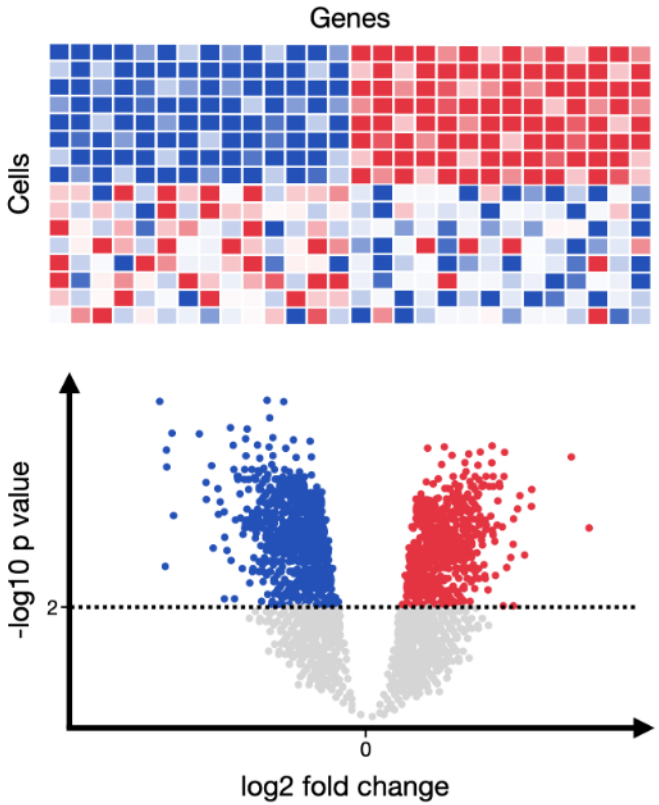
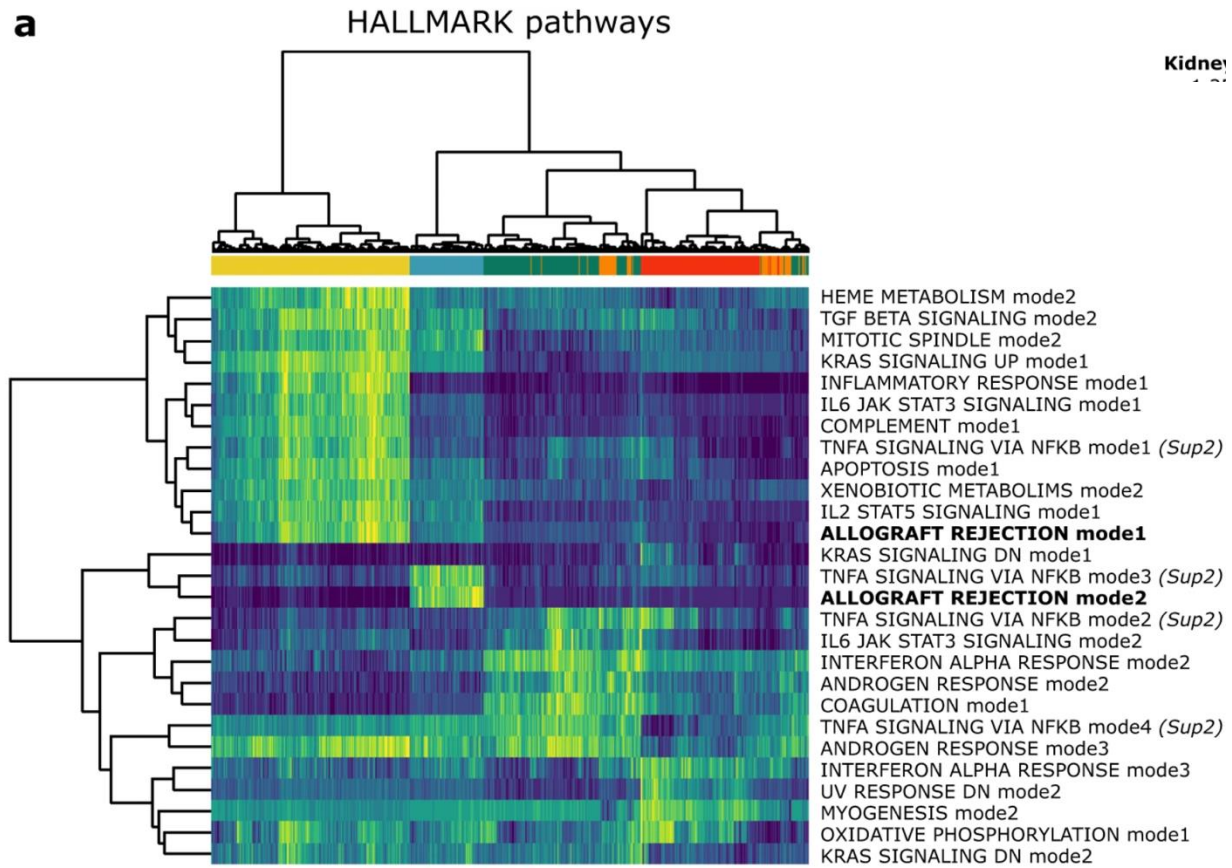
GSVA: gene set variation analysis for microarray and RNA-Seq data

Sonja Hänzelmann[1,2], Robert Castelo[1,2*] and Justin Guinney[3*]

Hui Huang
Data Science Journal Club
10.16.2025

Why pathway-level analysis?





noisy, small effects, hard to interpret.

stable and interpretable

Classic Gene Set Enrichment (GSE) workflow: rank genes → map to gene sets → compute pathway scores for downstream analyses

GSE: What to Know

- **Two GSE tests**: Competitive (set vs rest) and Self-contained (set alone).
- **Limits**: many methods are group-level/supervised; don't fit multi-phenotype/censored data.
- **Single-sample** (PLAGE, ssGSEA, z-score) helps, but relies on strong assumptions.
- **Need:** per-sample, assumption-light pathway scores → **GSVA**.

Where Gene Set Variation Analysis (GSVA) fits

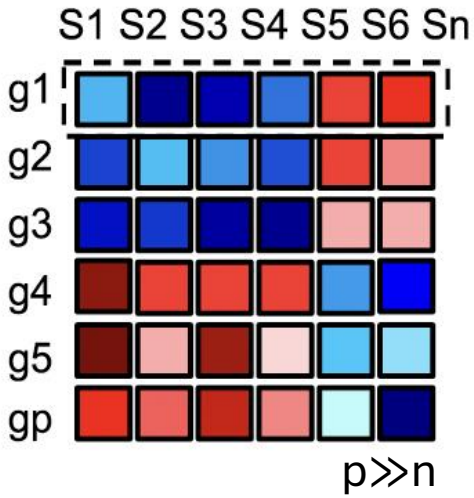
GSVA turns many gene expressions into one **pathway score per sample**, with no phenotype needed.

GSVA methods outline

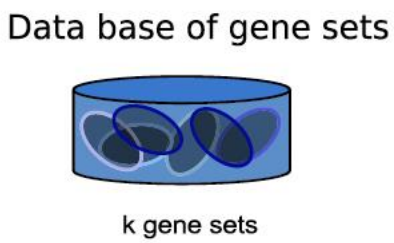
Input

- log2 microarray expression values
- RNA-seq counts

Gene expression matrix



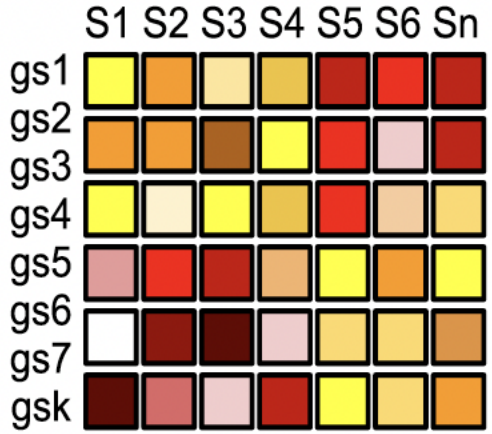
g: gene



GSVA Algorithm

Output

GSVA score matrix



gs: gene set

How to get the gene set score for each sample?

GSVA Algorithm

1. Gene expression level statistic 2. Rank order per sample

One gene in all samples (6 samples here)

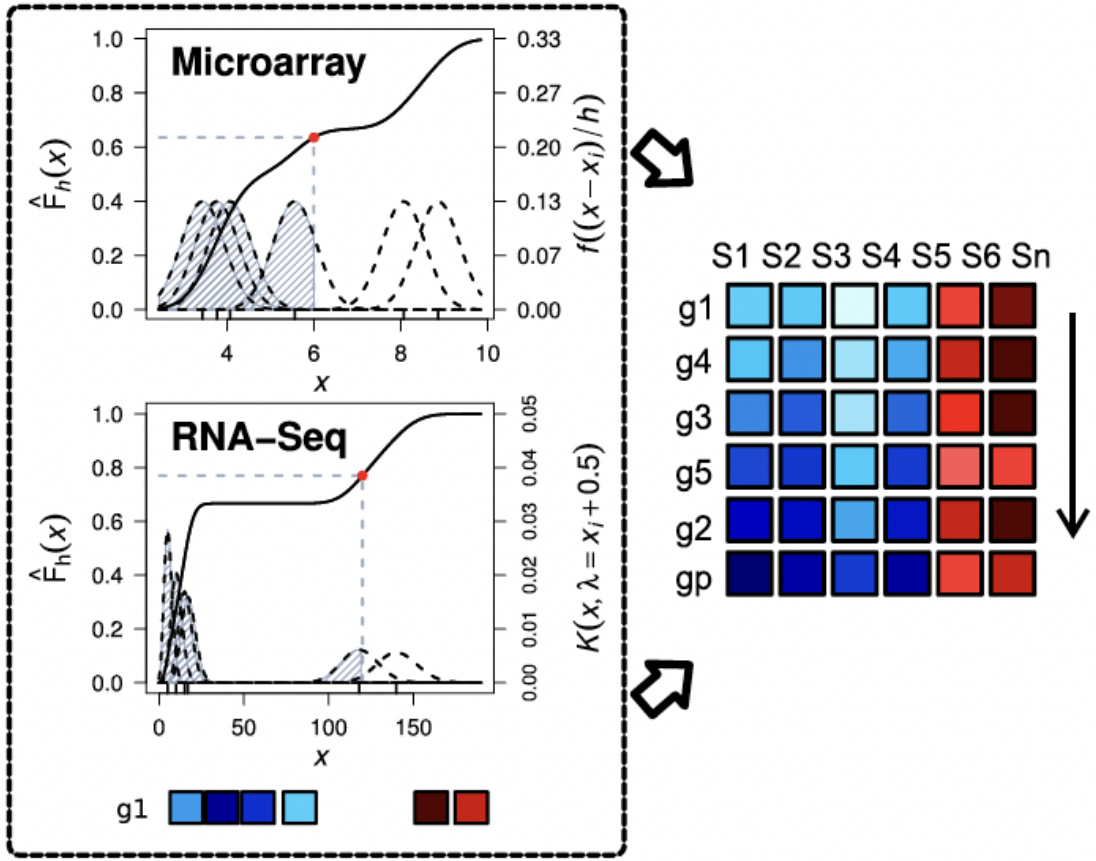
array→Gaussian kernel
RNA-seq→Poisson kernel

Dashed peaks: One kernel is placed at each sample.

sum and integrate

Solid black curve: smoothed cumulative distribution function (CDF)

Simulated expression profiles for six samples



Blue = low percentile;
Red = high percentile (high expression).

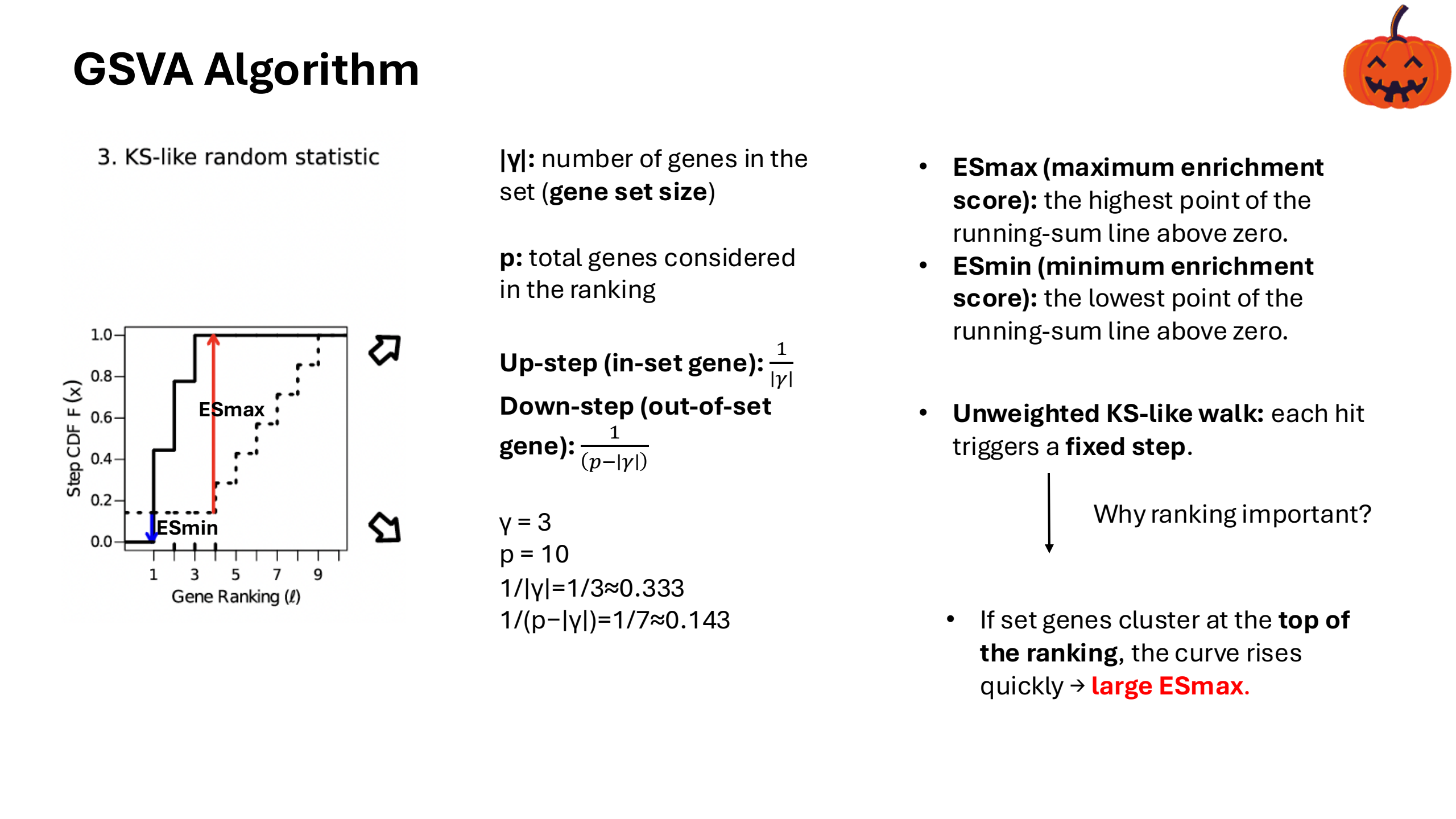
Genes × Samples = Percentiles (0–1)

$$P_{ij} = F_i(x_{ij})$$

$$P = \begin{bmatrix} 0.10 & 0.85 & 0.40 & 0.60 \\ 0.70 & 0.20 & 0.95 & 0.55 \\ 0.30 & 0.45 & 0.10 & 0.80 \end{bmatrix}$$

eg.: $P_{1,2} = 0.85$means gene 1 is high in Sample 2 (higher than 85% of samples).

GSVA Algorithm



3. KS-like random statistic

|γ|: number of genes in the set (**gene set size**)

p: total genes considered in the ranking

Up-step (in-set gene): $\frac{1}{|\gamma|}$

Down-step (out-of-set gene): $\frac{1}{(p-|\gamma|)}$

γ = 3
p = 10
1/|γ|=1/3≈0.333
1/(p−|γ|)=1/7≈0.143

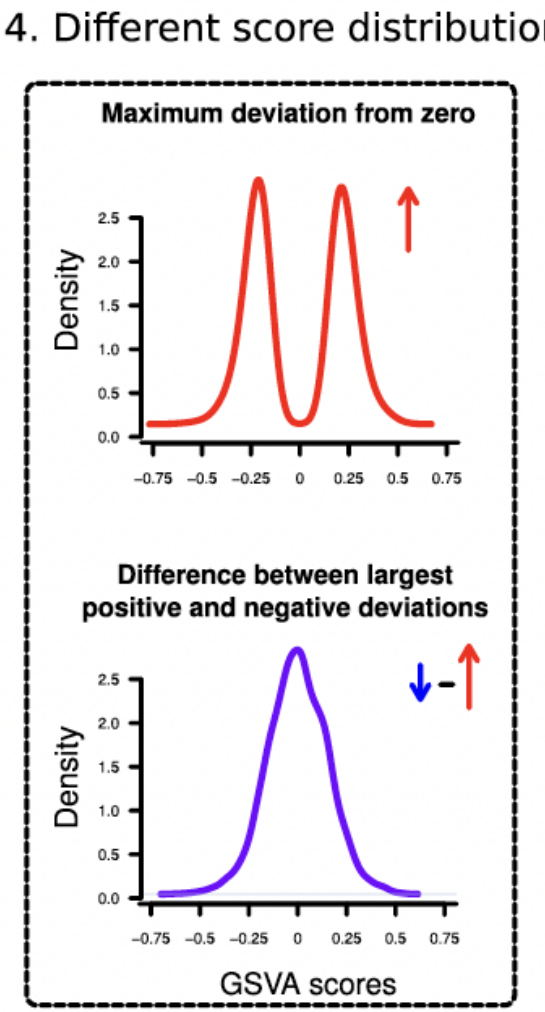
- **ESmax (maximum enrichment score):** the highest point of the running-sum line above zero.
- **ESmin (minimum enrichment score):** the lowest point of the running-sum line above zero.

- **Unweighted KS-like walk:** each hit triggers a **fixed step**.

↓ Why ranking important?

- If set genes cluster at the **top of the ranking**, the curve rises quickly → **large ESmax.**

GSVA Algorithm



ESmax

top-ranked in-set genes ⇒ positive

bottom-skewed ⇒ negative

- **Use ESmax** when you care about **direction (up vs down)**

Exdiff (Esmax – Esmin)

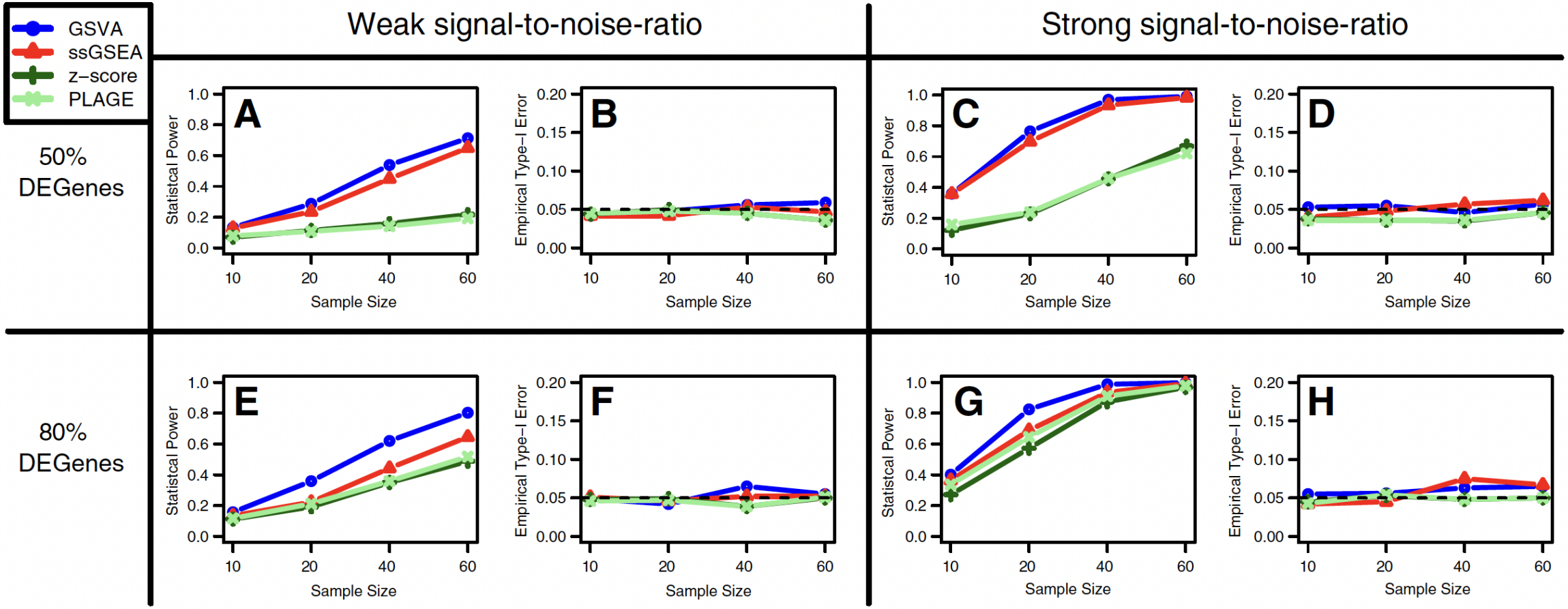
single-peaked/near-normal

- **Use ESdiff** when you need **model-friendly, stable scores** (add ESmax sign for direction)

GSVA Shows Higher Power in Simulations

easier to detect, higher power, smaller sample size needed.

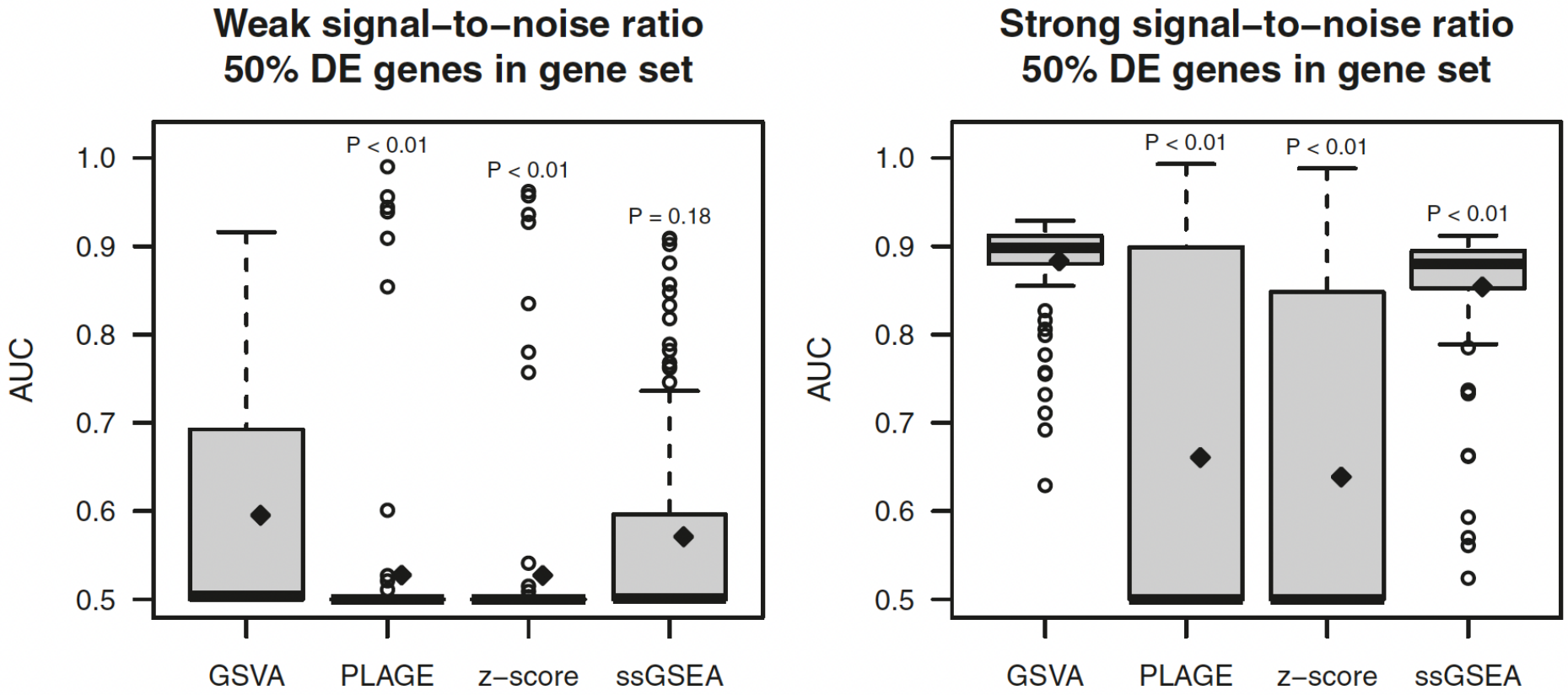
harder to detect, lower power, larger sample size needed.



- **Power:** chance to **detect a real effect** $(1-\beta)$;higher power = fewer missed true differences.
- **Type-I error:** chance of a **false positive** (α) ;often controlled at **0.05**.

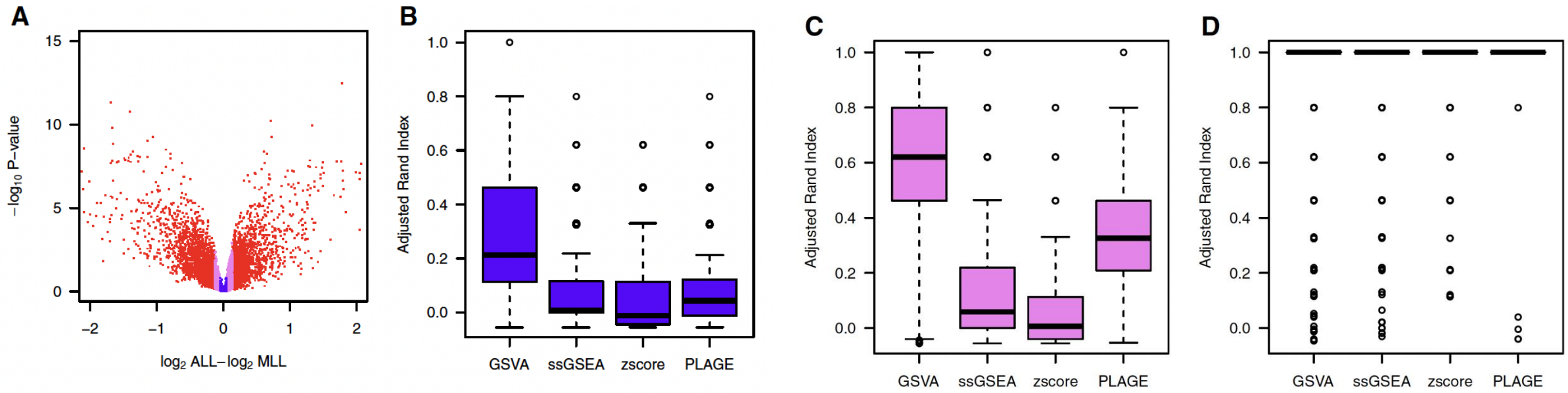
GSVA Delivers the Highest ROC curve AUC in Simulations

For each gene set, compare GSVA scores between the two groups using a two-sample t-test; use the **absolute t-statistic** | t |(or $1 - p$) as the ranking score to build the ROC and compute AUC.



- **GSVA leads under weak signals** (AUC significantly higher, P < 0.01).
- **Under strong signals**, methods converge; **GSVA/ssGSEA** remain top.
- **Higher AUC** means the method **puts true positives ahead of nulls** more reliably.

GSVA is more sensitive to weak/moderate signals in real data



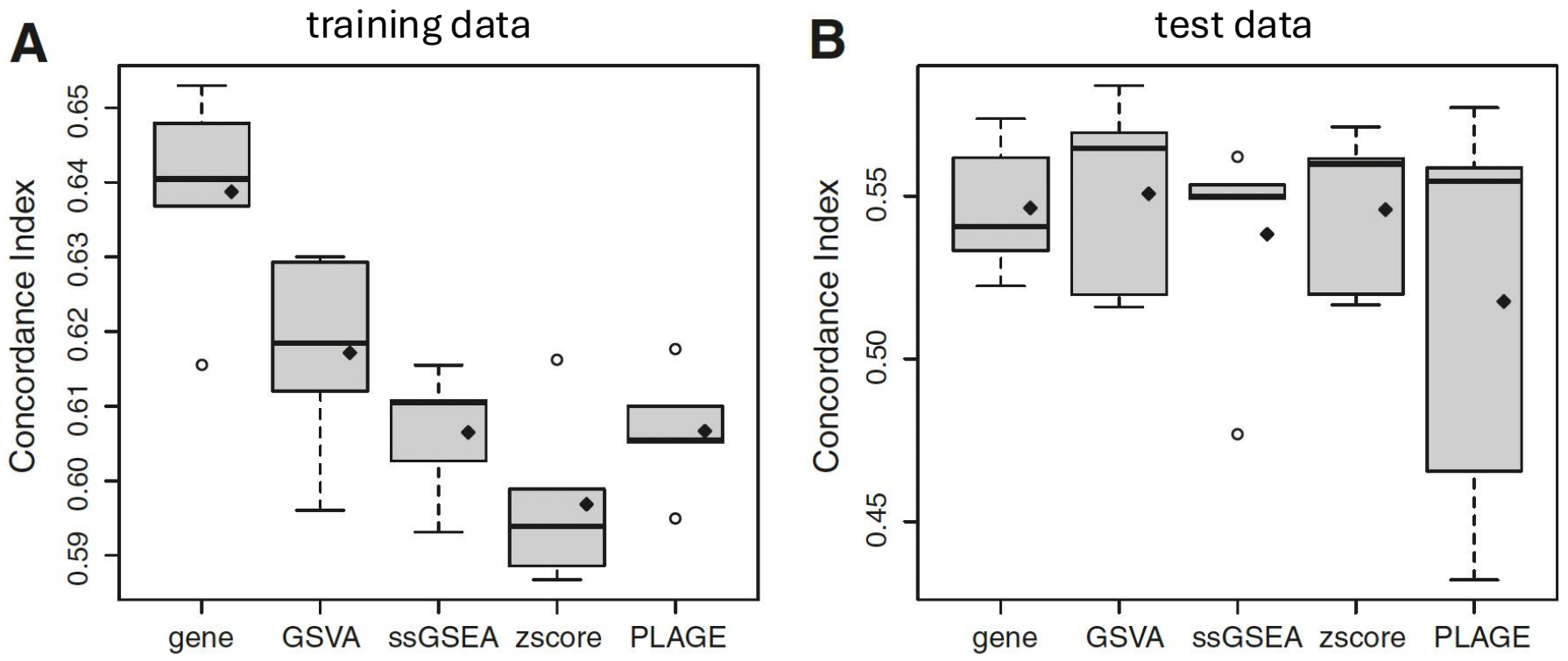
Rank all genes by fold change into three equal terciles: small, medium, and large effect.

Workflow: fold-change stratification → bootstrap sampling (10 samples, repeat 1000 times → compute GSE scores → **limma** select top-5 gene sets → hierarchical clustering → **ARI** evaluation.

Adjusted Rand Index (ARI): How well a clustering matches the true groups (0–1; **1 = perfect**, **0 = like random**)

- **Weak/medium signals:** GSVA attains the highest ARI (**$p \ll 0.01$**).
- **Strong signals:** all methods are near ceiling; differences are smaller.

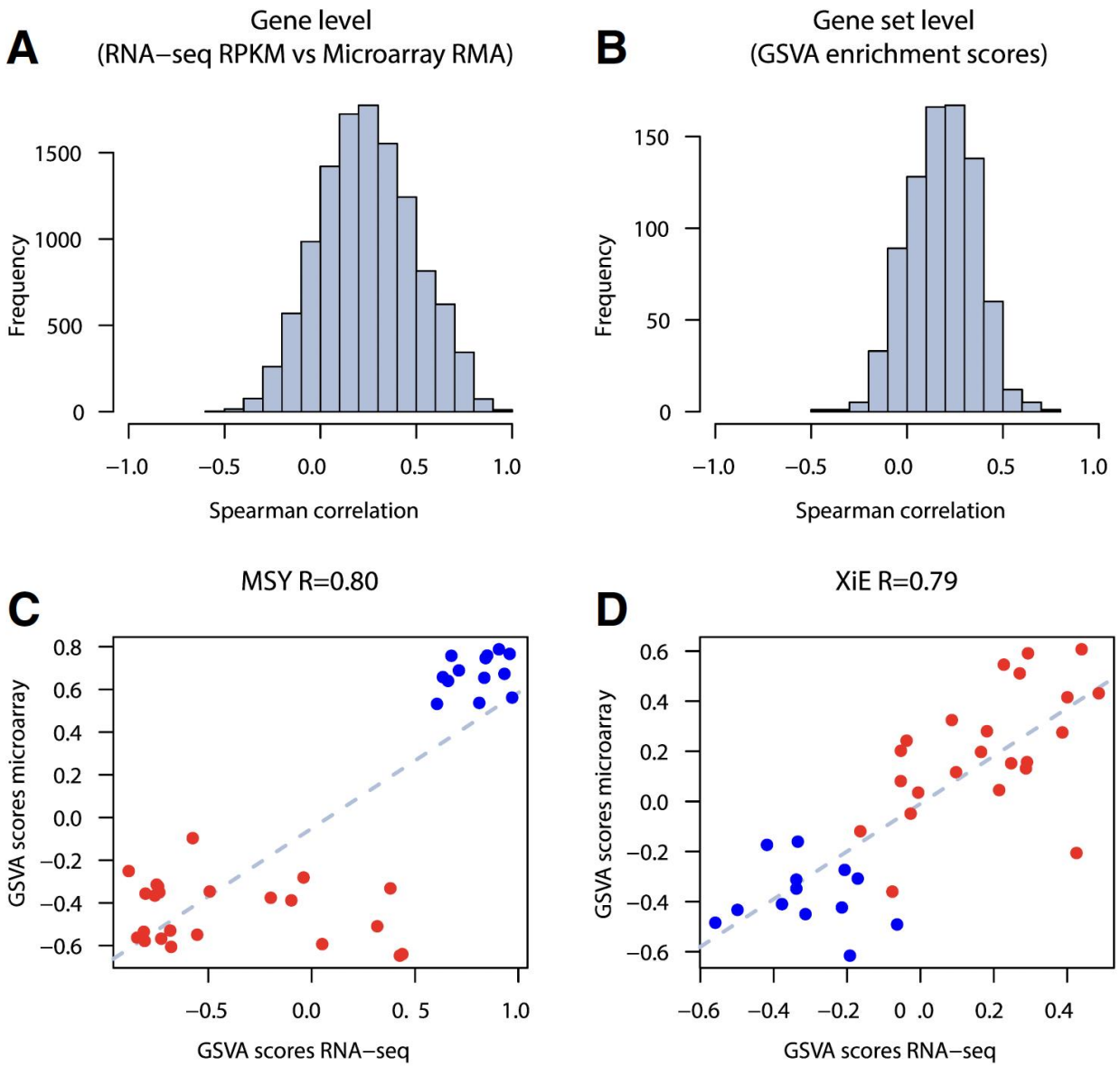
GSVA improves survival prediction in ovarian cancer



C-index (0.5 = random; higher is better).

- **Train (A):** Gene looks best → likely **overfitting**.
- **Test (B): GSVA/ssGSEA** show **higher C-index** with **smaller spread**.

GSVA Cross-platform Consistency



- **Pathway > gene correlation** — GSVA aggregation is more stable.
- **Sex signatures replicate** — MSY ↑ in males; XiE ↑ in females
- **High agreement (R ≈ 0.8)** — GSVA scores are portable across technologies.

Conclusion

- **What GSVA is:** Per-sample pathway scores via kernel-ECDF and KS-like running sum.
- **Sensitivity:** Better at detecting **weak/coordinated** signals than ssGSEA/PLAGE/z-score (with controlled Type-I error).
- **Real data wins:** Higher **ARI** (ALL vs MLL) and **C-index** (TCGA ovarian) vs alternatives.
- **Practical tips:** Use **ESdiff** for modeling, **ESmax** for direction (ensure **n>10**, adjust **batch/covariates**, pre-filter coherent gene sets (size ~10–500)).