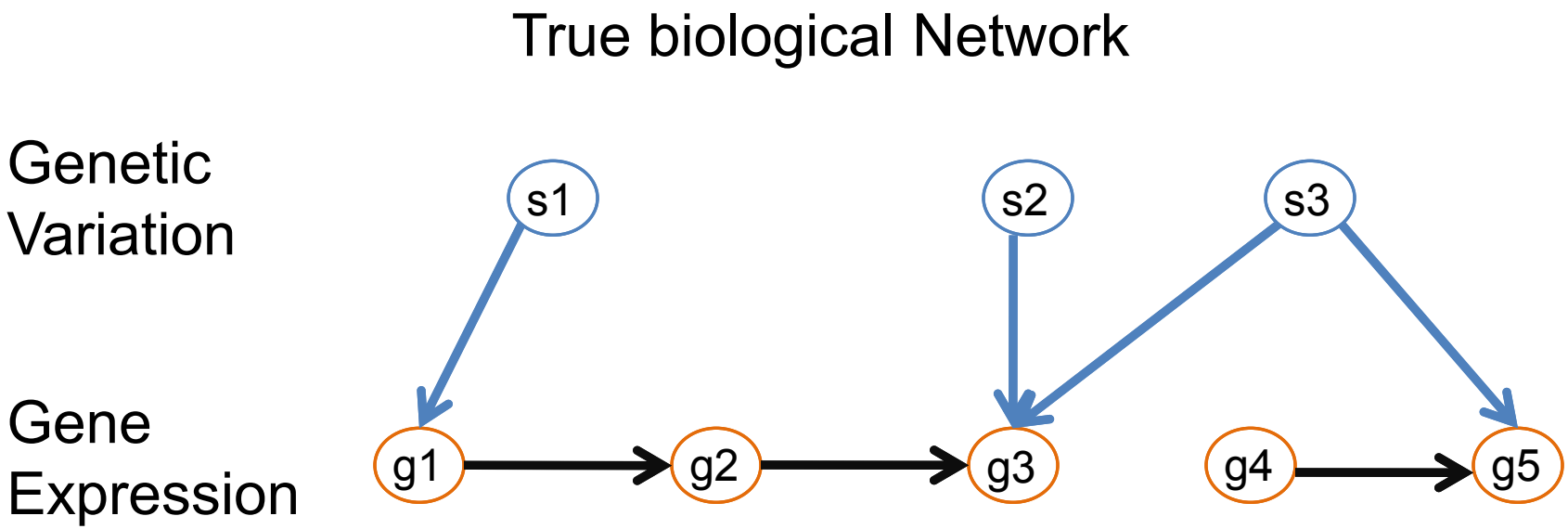
Detecting Presence and Absence of Causal Relationships Between Expression of Yeast Genes with Very Few Samples

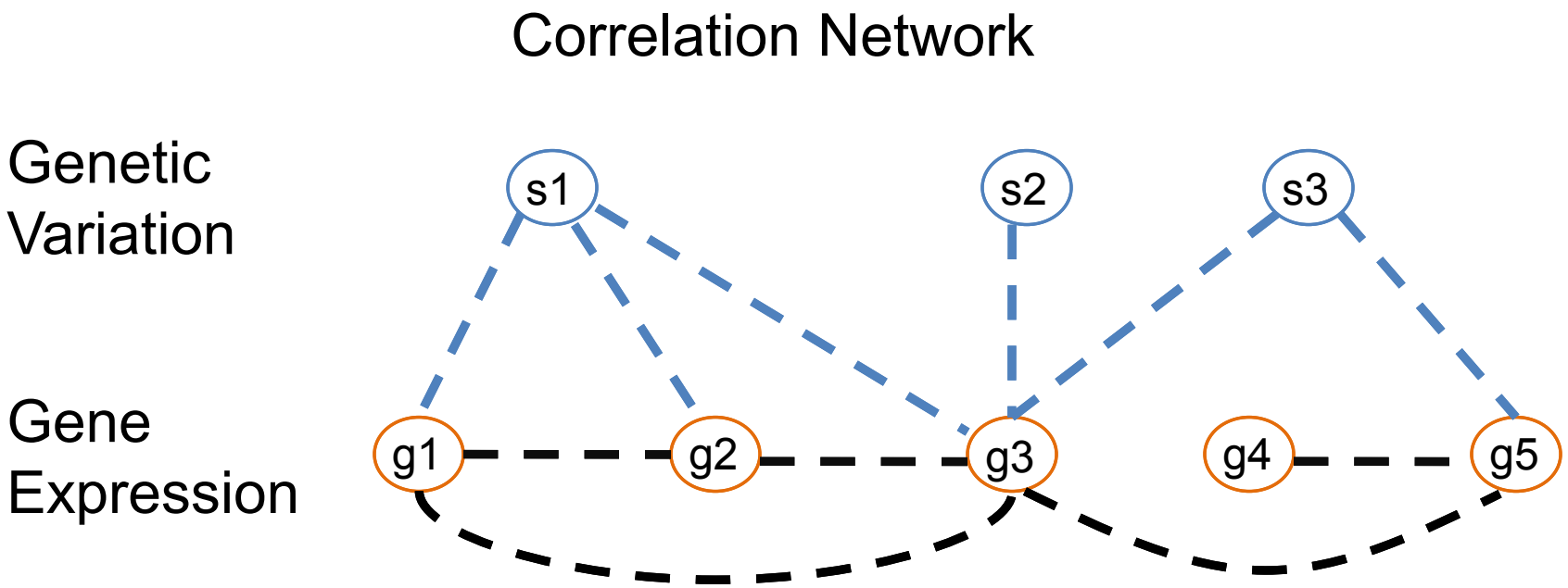
Eun Yong Kang, Ilya shpitser,
Hyun Min Kang, Chun Ye,
Eleazar Eskin

UCLA COMPUTER SCIENCE DEPARTMENT
Henry Samueli School of Engineering and Applied Science

Network Reconstruction of Biological System



Network Reconstruction Based on Correlation

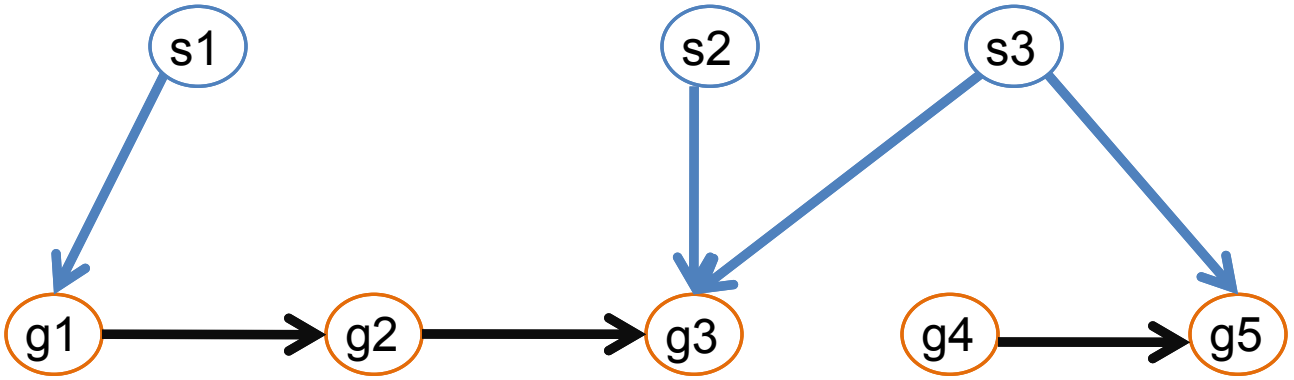


Network Reconstruction Based on Causal Inference

True biological Network

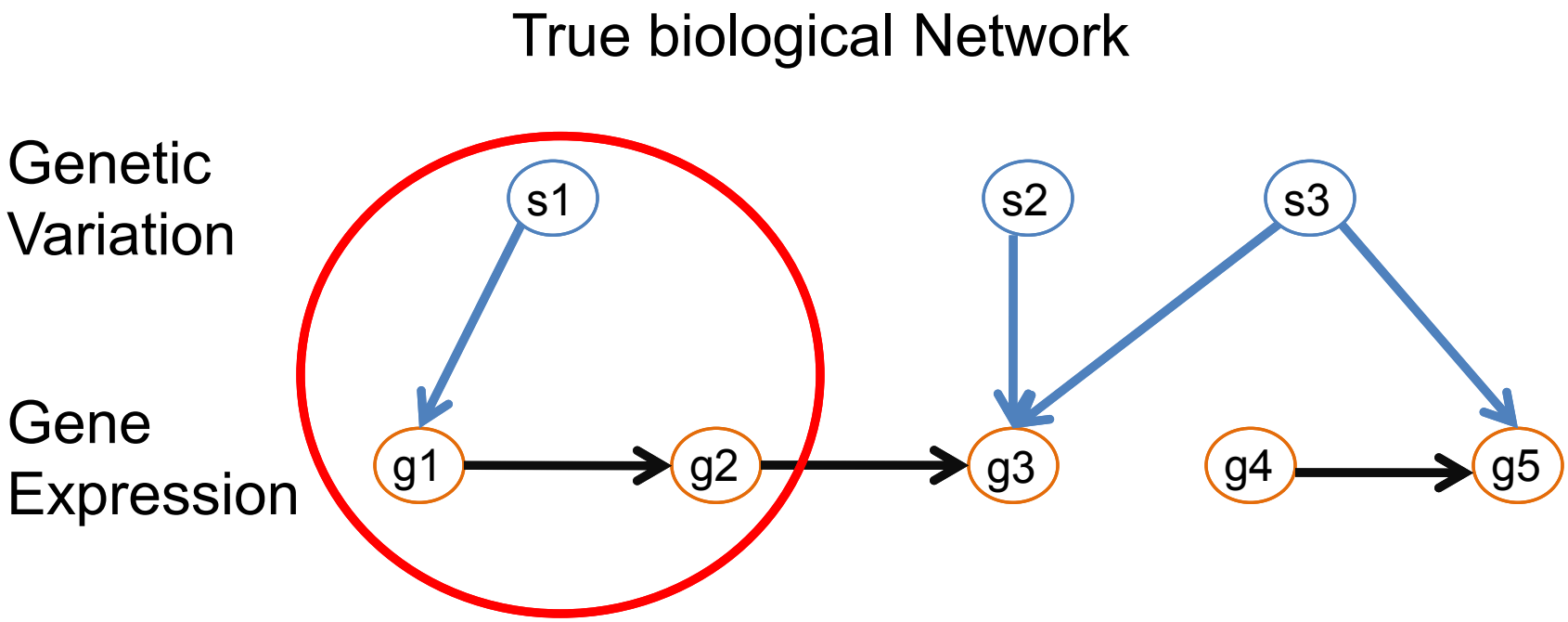
Genetic Variation

Gene Expression

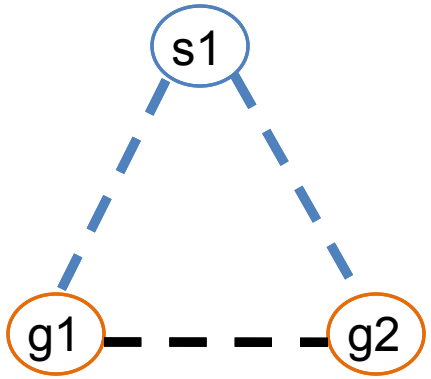


How can we re-construct true biological network?

Identifying Causal Relationships between Genes



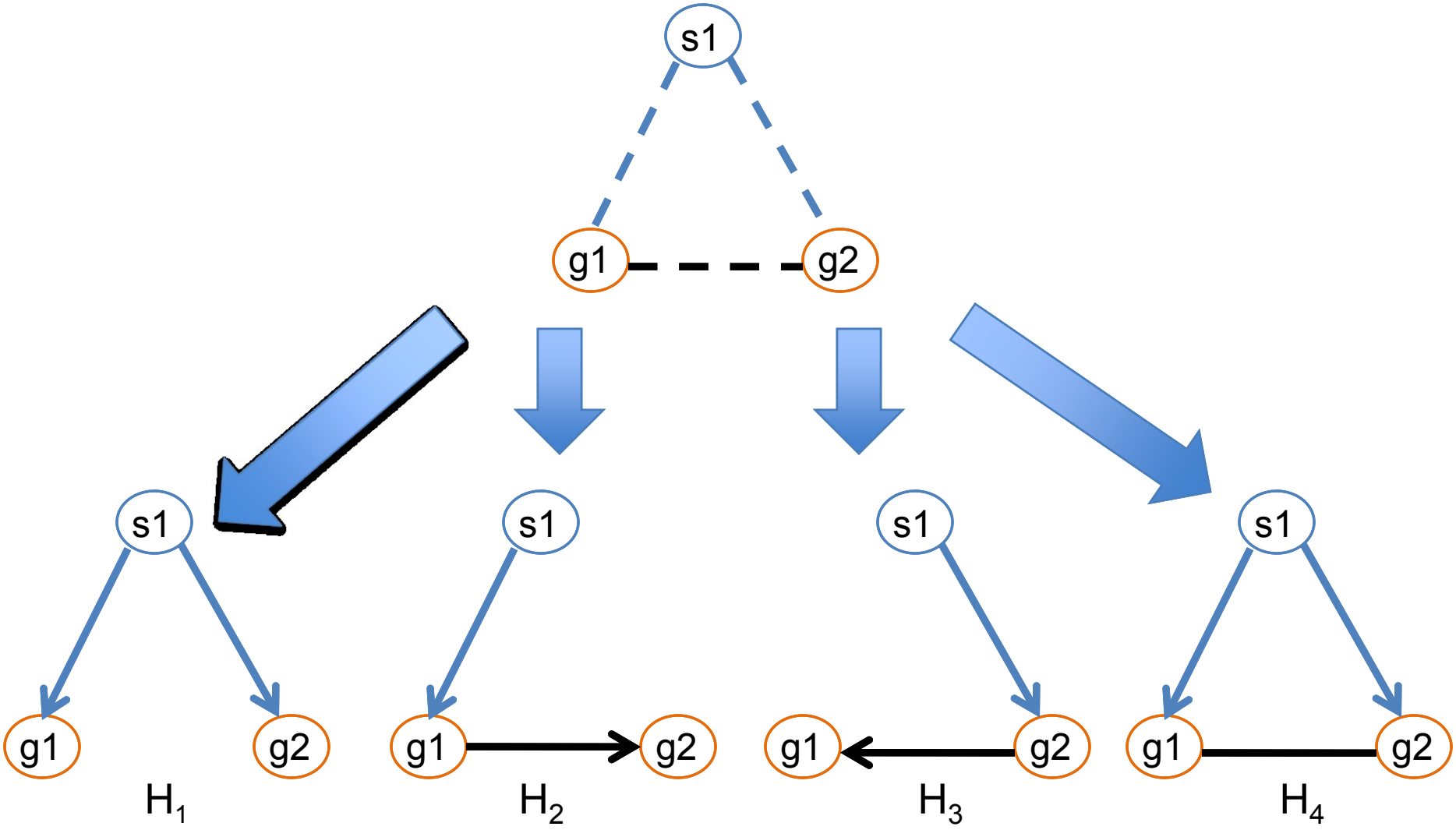
Correlation Network for this triplet



We have a hint for certain edges!

SNPs can cause changes in gene expression and not vice versa

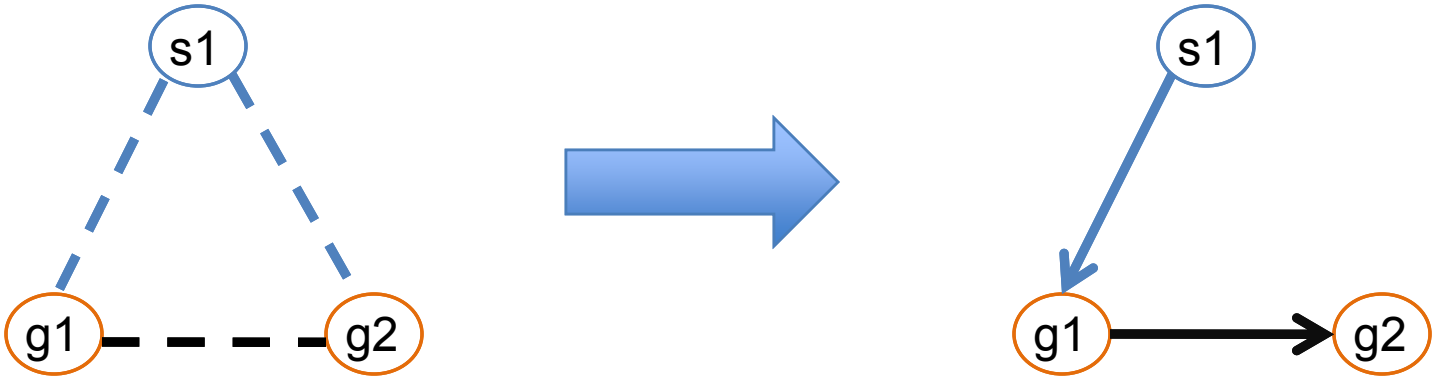
4 Possible Models



Theorem 1 (Infinite Sample Case)

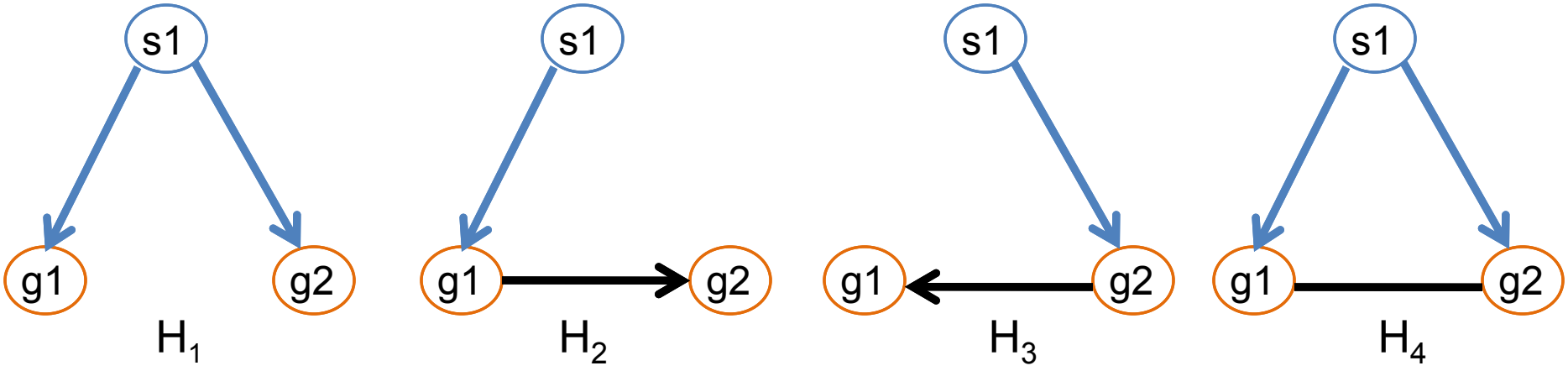
- Given a causal graph G where:
  1. s_1 is correlated to g_1
  2. s_1 is correlated to g_2
  3. s_1 and g_2 are conditionally independent given g_1, Then g_1 causes g_2.

(For more formal definition, look at our paper)

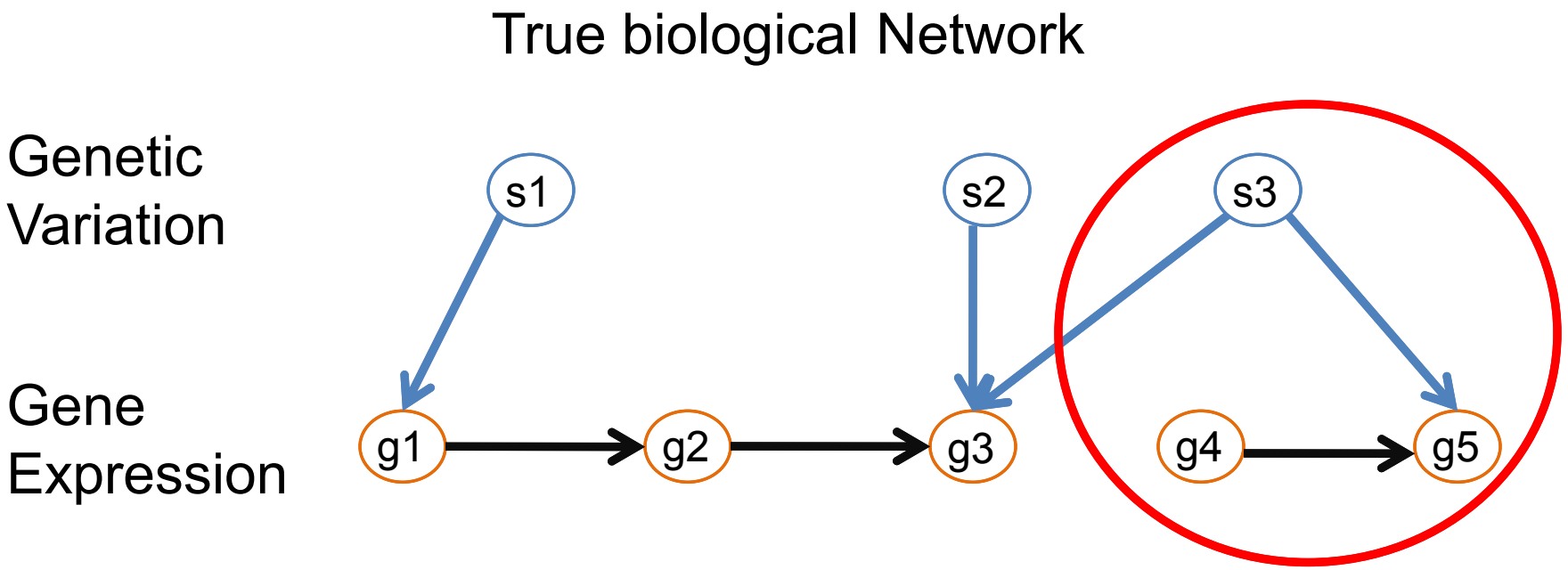


Model Selection among Possible Models (Finite Sample Case)

- likelihood ratio test : For each H_1, H_2, H_3 model against H_4(Full Model)
- Conclude that hypothesis is likely true, if corresponding ratio is close to 1.



Identifying Absence of Causal Relationship between Gene Expression Levels

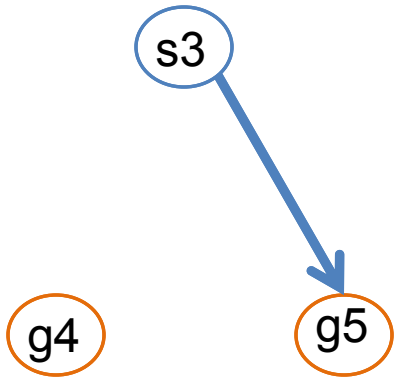


Theorem 2

- Given a causal graph G where:
  1. s_3 is correlated to g_5
  2. s_3 is not correlated to g_4.

  Then g_5 cannot cause g4.

  (For more formal definition, look at paper)

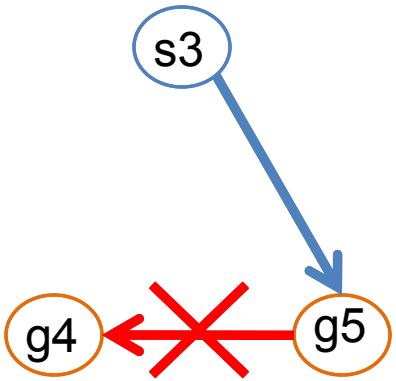


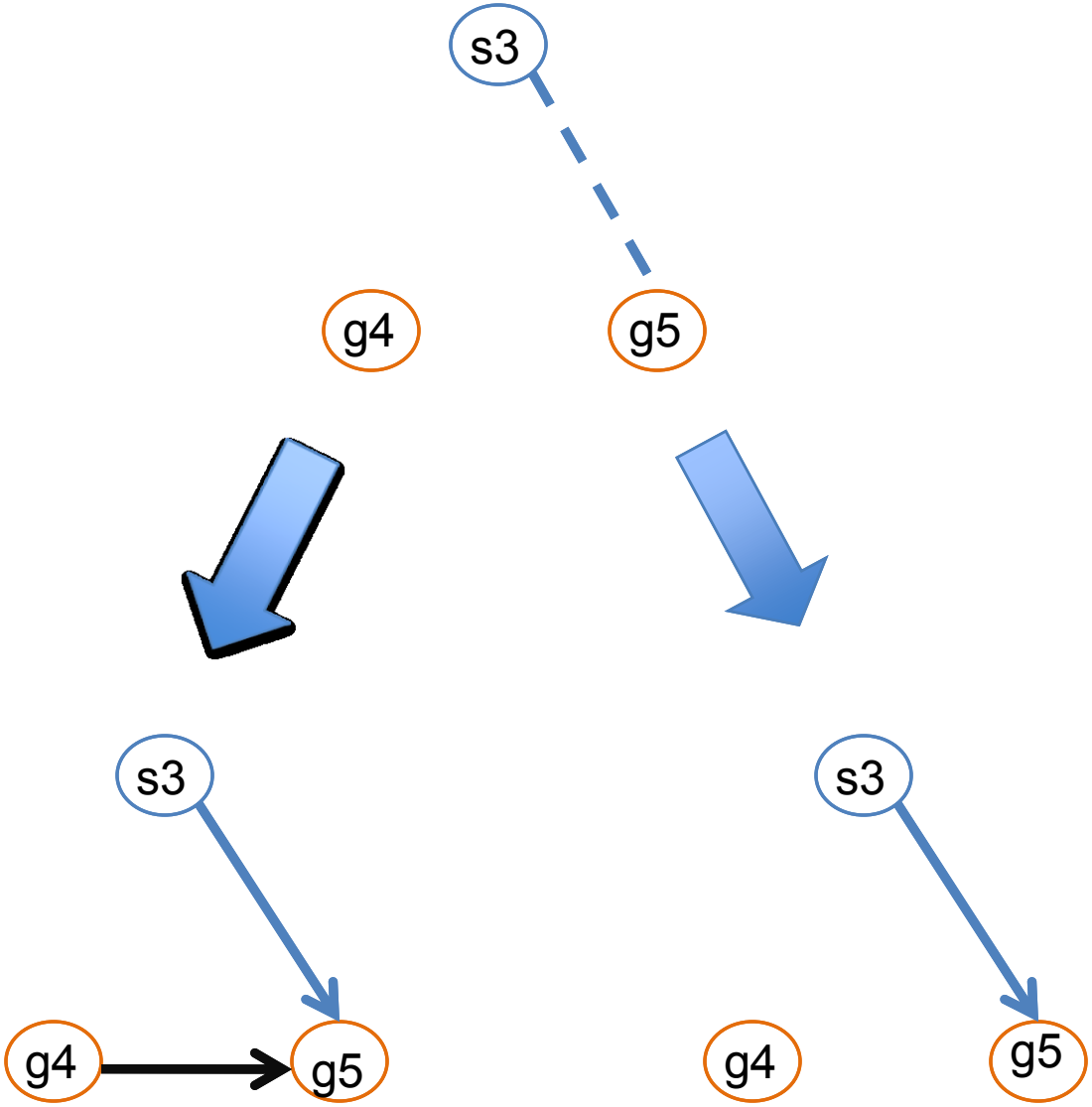
Theorem 2

- Given a causal graph G where:
  1. s_3 is correlated to g_5
  2. s_3 is not correlated to g_4.

  Then g_5 cannot cause g4.

  (For more formal definition, look at paper)




s3
g4
g5
s3
g4
g5
s3
g4
g5

Experiment for the validation

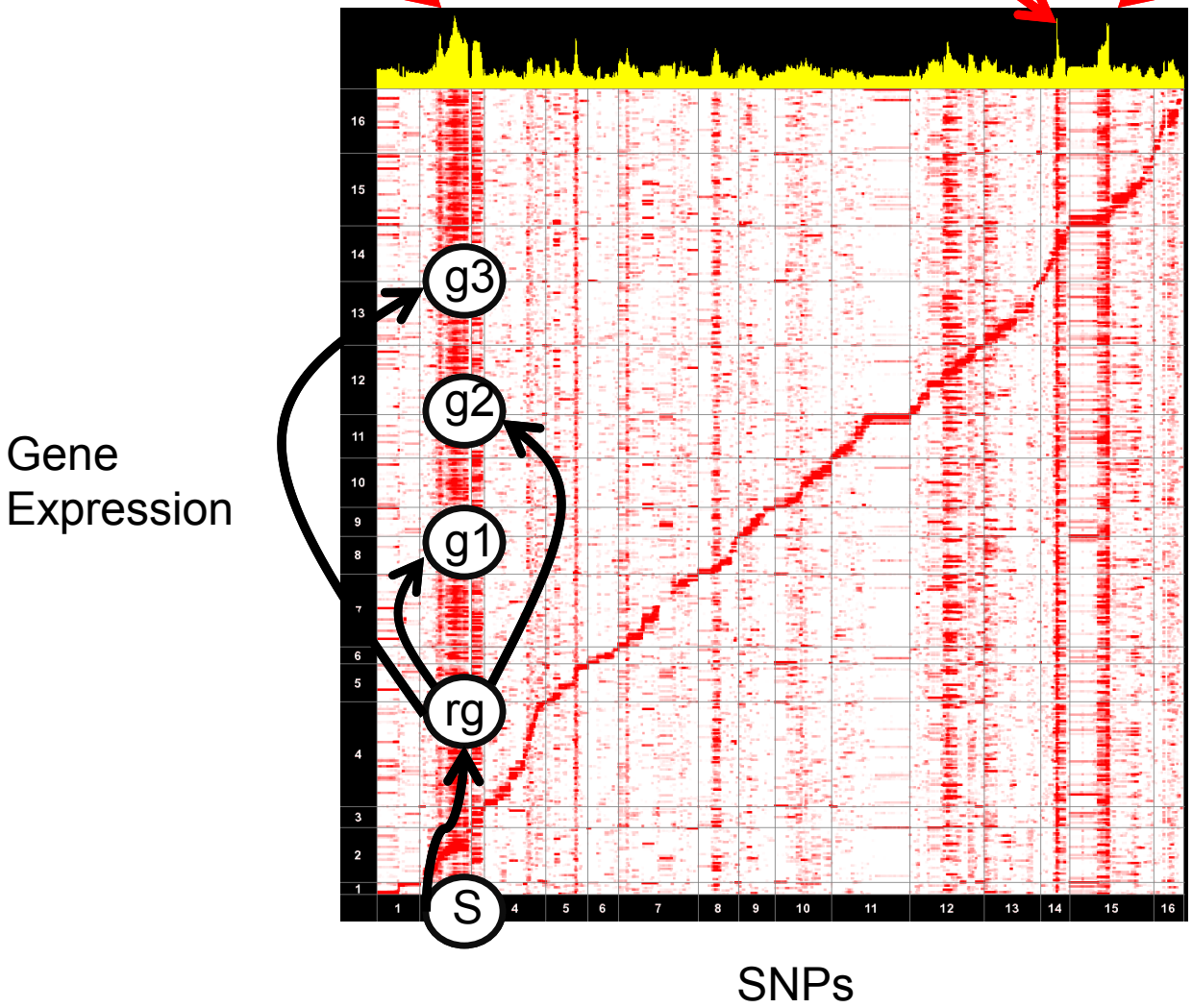
- 112 genetically distinct segregants of yeast
- Gene expression dataset contains 5534 genes
- Genotyping dataset of 2956 SNPs from each 112 segregants of yeast

Result on Yeast Dataset

- We found 24620 statistically significant association from SNP to gene expression levels

- By applying our method, we found 4684 causal relationships between genes and 292 regulator genes and 2217 affected target genes.

Regulatory Hotspot Analysis



SNP Chr	SNP Loc	Regulators	# Affected Genes
2	390000	TAT1(128)	128
2	560000	**AMN1*+(187)**, SDS24(146), *YSW1+(128)*, *TBS1*(122)*, CNS1*+(117), *ARA1*(110)*, *SUP45*(64)*, LYS2(43), RPS9B(42), *TOS1*(31)*, YBR187W(29)	339
3	100000	*NFS1*(121)*, *CIT2*(118)*, **LEU2*+(105)**, HIS4(83), **ILV6*+(29)**	231
3	230000	**MATALPHA1*(57)**, MATALPHA2(37)	63
4	150000	YRF1-4(32), YRF1-1(25)	38
5	130000	**URA3*(25)**	25
8	130000	*SPO11*(26)*, **GPA1*+(12)**	36
9	130000	HIS4(52),	52
12	110000	ASP3-1(22), ASP3-3(18), ASP3-2(15),	27
12	680000	*MAP1*(34)*, **HAP1*(33)**	62
12	800000	STP3(35), GAS2(33)	68
12	1070000	YML133C(36), *YRF1-4*(32)*, YLR464W*(31), *YRF1-5*(24)*,	39
13	70000	*SMA2*(25)*,	25
14	503000	**SAL1*+(154)**, LAT1(113), *TOP2*(94)*, COG6(93), YNL035C(82), MSK1(48), PHO91(46), NMA111(44), NAM9+(36), MKT1(28), MRP7(26)	433
15	180000	**PHM7*+(116)**, *HAL9*(92)*, *NDJ1*(91), RFC4(81), ZEO1(78), WRS1(67), SKM1(55)	296

Table 2: Regulatory Hotspots and Corresponding Regulators

Regulators with an asterisk (*) were found by Zhu et al. (2008). Regulators marked with a plus (+) were found in the Chen et al. (2007) study and unlabeled regulators are novel predictions.

Regulatory hotspot analysis

- **There are a total of 12 causal regulators with some experimental evidence (Gold Standard). -AMN1, MAK5, LEU2, MATALPHA1, URA3, GPA1, HAP1, SIR3 and CAT5 (Yvert et al., 2003)**
- **ILV6, SAL1 and PHM7(Zhu et al., 2008)**

- **All three methods : AMN1, LEU2, ILV6, GPA1, SAL1,PHM7**
- **Zhu & Our method : METALPHA1, URA3, HAP1**

- **The best validation of our method is that we were able to find ILV6 which was experimentally validated in (Zhu et al., 2008). However, Zhu et al. (2008) used additional types of data (incorporating TFBS data from ChIP-chip experiments, phylogenetic conservation, and protein protein interaction data (PPI)) in order to discover ILV6 and they claim that they would not have been able to discover ILV6 if they used only the data that we used.**

Conclusion

- We infer the presence or absence of causal relationship between genes.
- The fact that SNP only cause the gene help us to reduce the number of possible search space.
- Reduced search space allow us to overcome the disadvantage of limited samples.
- Our method is applicable to any causal inference on sparse graph where there is causal anchors (SNP).

Acknowledgement

Ilya Shpitser - Harvard

Hyun Min Kang - UC San Diego

Chun Ye - UC San Diego

Eleazar Eskin - UCLA