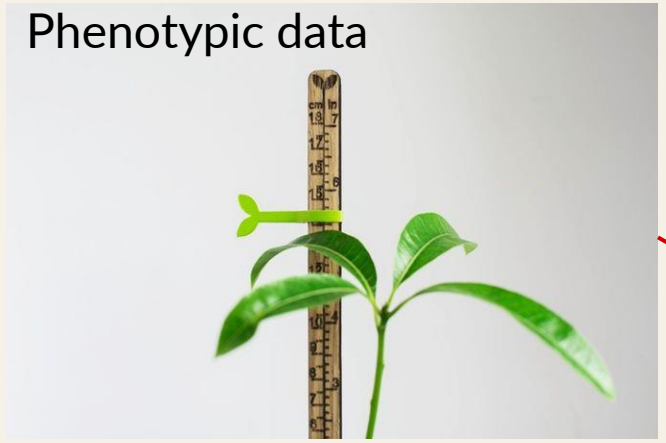
Evaluating Dimensionality Reduction for Genomic Prediction

Vamsi Manthena, Rajeev Varshney, Diego Jarquin, Reka Howard

Genomic Prediction (GP)



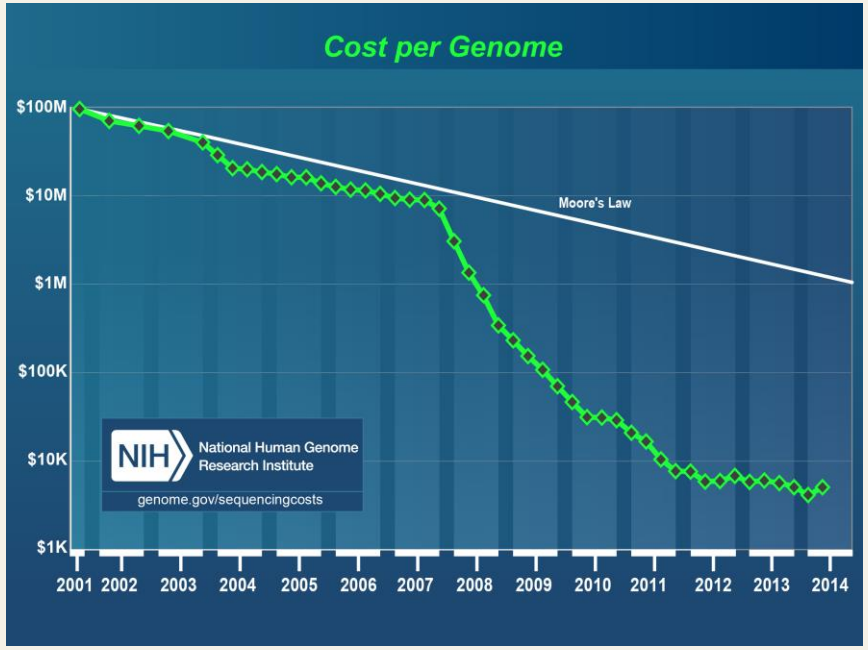
$$$

$

G-BLUP model

$y = g + e$

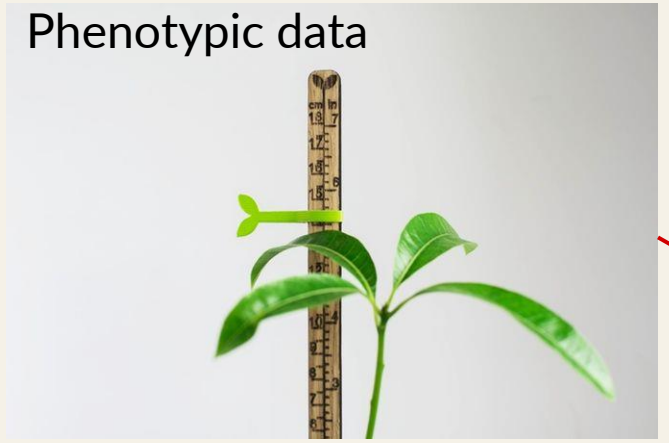
Prediction and Selection



High-dimensionality in GP

	Marker 1	Marker 2	Marker 3	…	…	…	Marker p
Line 1	1	1	0				1
Line 2	1	0	0				0
Line 3	1	0	0				1
…							
Line n	1	0	1				0

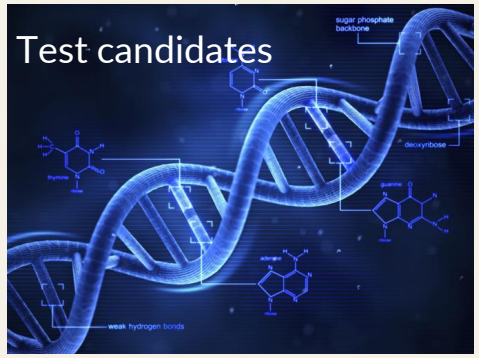
Genomic Prediction (GP)



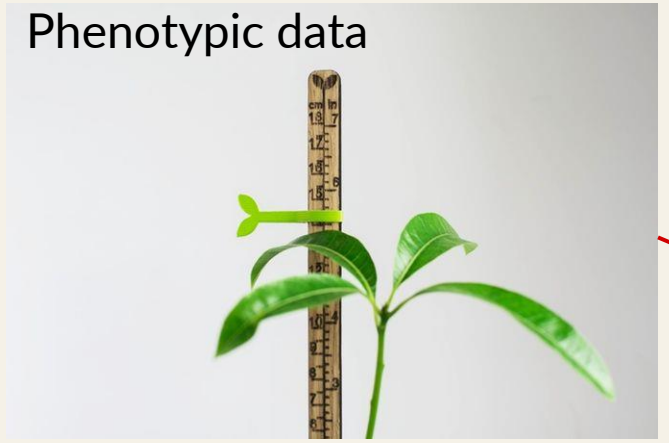
G-BLUP model

$y = g + e$

Prediction and Selection



Dimensionality Reduction in GP



Dimensionality Reduction

G-BLUP model
$y = g_r + e$

Prediction and Selection

Methods

- Random Projections
- Random Sampling
- Deterministic Sampling
- Ridge Regression based
- Clustering based

Data Description

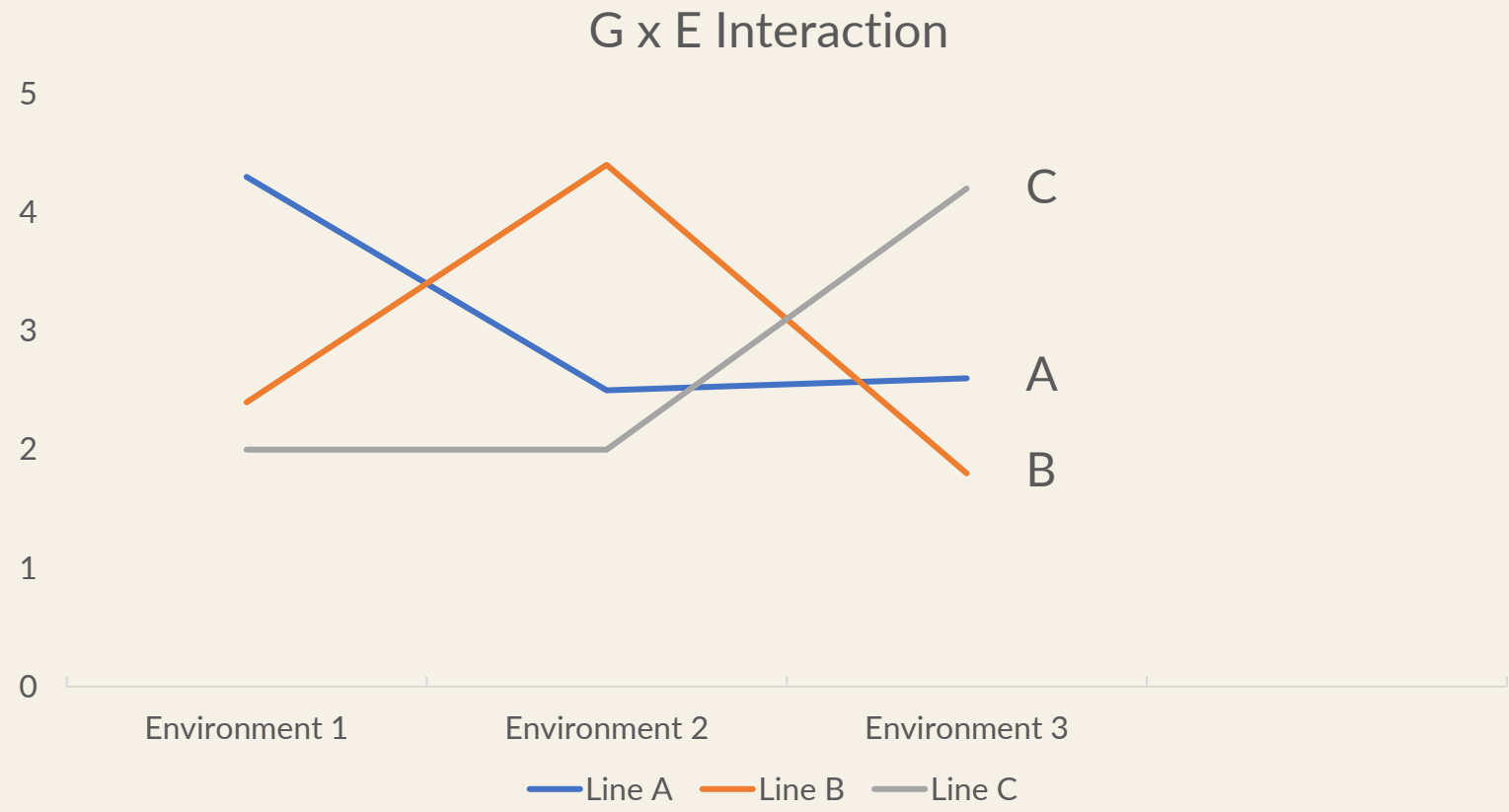
- Chickpea Data set collected by ICRISAT (Roorkiwal et al., 2016)

Observations – 306 lines of chickpea

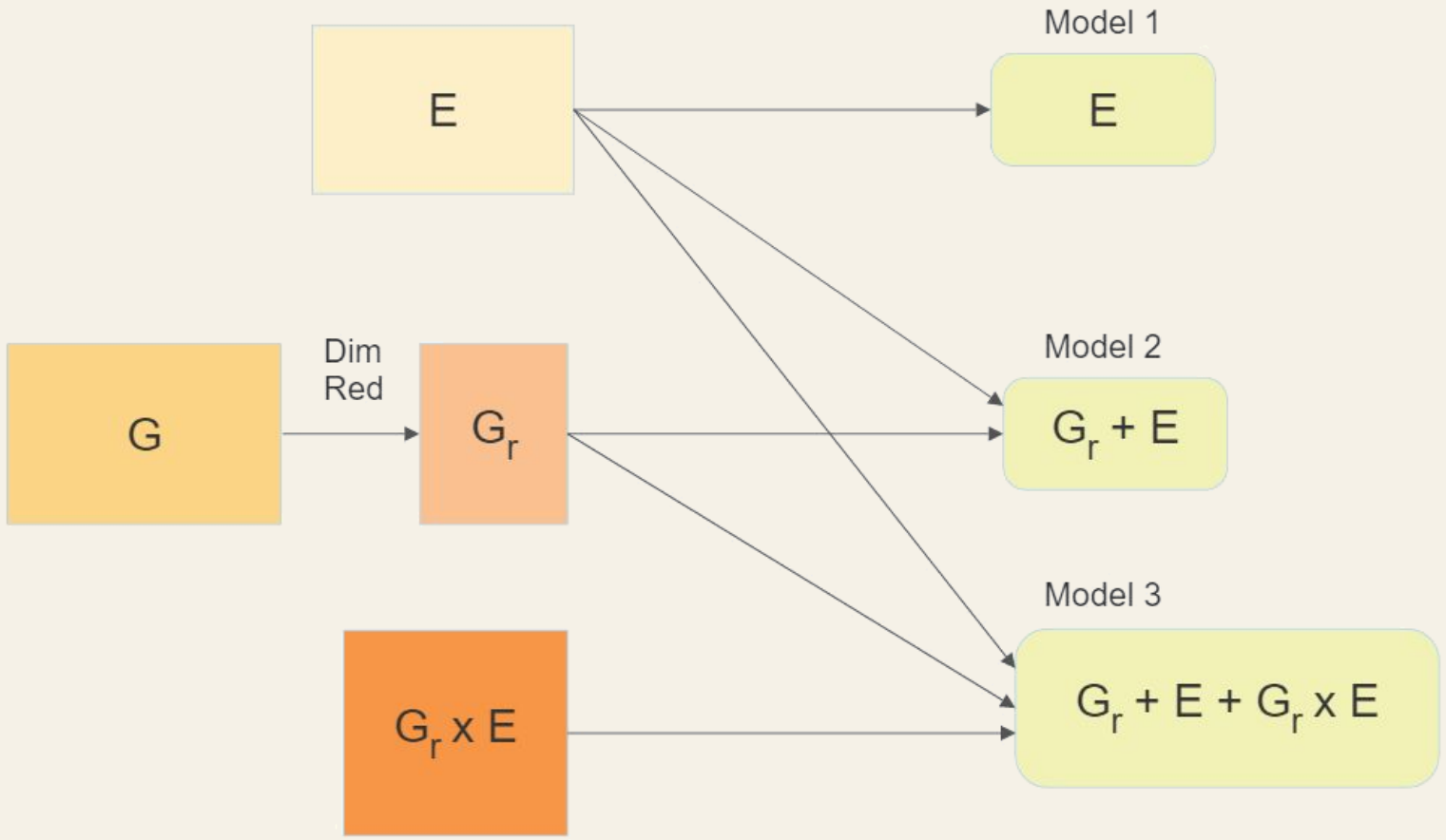
Environment – 9 environments

Marker Data – 14928 markers

Genotype-by-Environment Interaction (G × E)



Prediction Models



Cross-Validation

- Performance of prediction models assessed by correlations between observed and predicted yield values.

- We assessed the performance of all three prediction models in three cross-validation schemes.
  - CV0
  - CV1
  - CV2

Cross-Validation – CV0

Performance of all lines in a new untested environment

	CV0				
	E1	E2	E3	E4	E5
Line 1	Y_{11}	NA	Y_{13}	Y_{14}	Y_{15}
Line 2	Y_{21}	NA	Y_{23}	Y_{24}	Y_{25}
Line 3	Y_{31}	NA	Y_{33}	Y_{34}	Y_{35}
Line 4	Y_{41}	NA	Y_{43}	Y_{44}	Y_{45}
Line 5	Y_{51}	NA	Y_{53}	Y_{54}	Y_{55}

Cross-Validation – CV1

Performance of a new line in all the environments

	CV1				
	E1	E2	E3	E4	E5
Line 1	Y_{11}	Y_{12}	Y_{13}	Y_{14}	Y_{15}
Line 2	Y_{21}	Y_{22}	Y_{23}	Y_{24}	Y_{25}
Line 3	NA	NA	NA	NA	NA
Line 4	Y_{41}	Y_{42}	Y_{43}	Y_{44}	Y_{45}
Line 5	Y_{51}	Y_{52}	Y_{53}	Y_{54}	Y_{55}

Cross-Validation – CV2

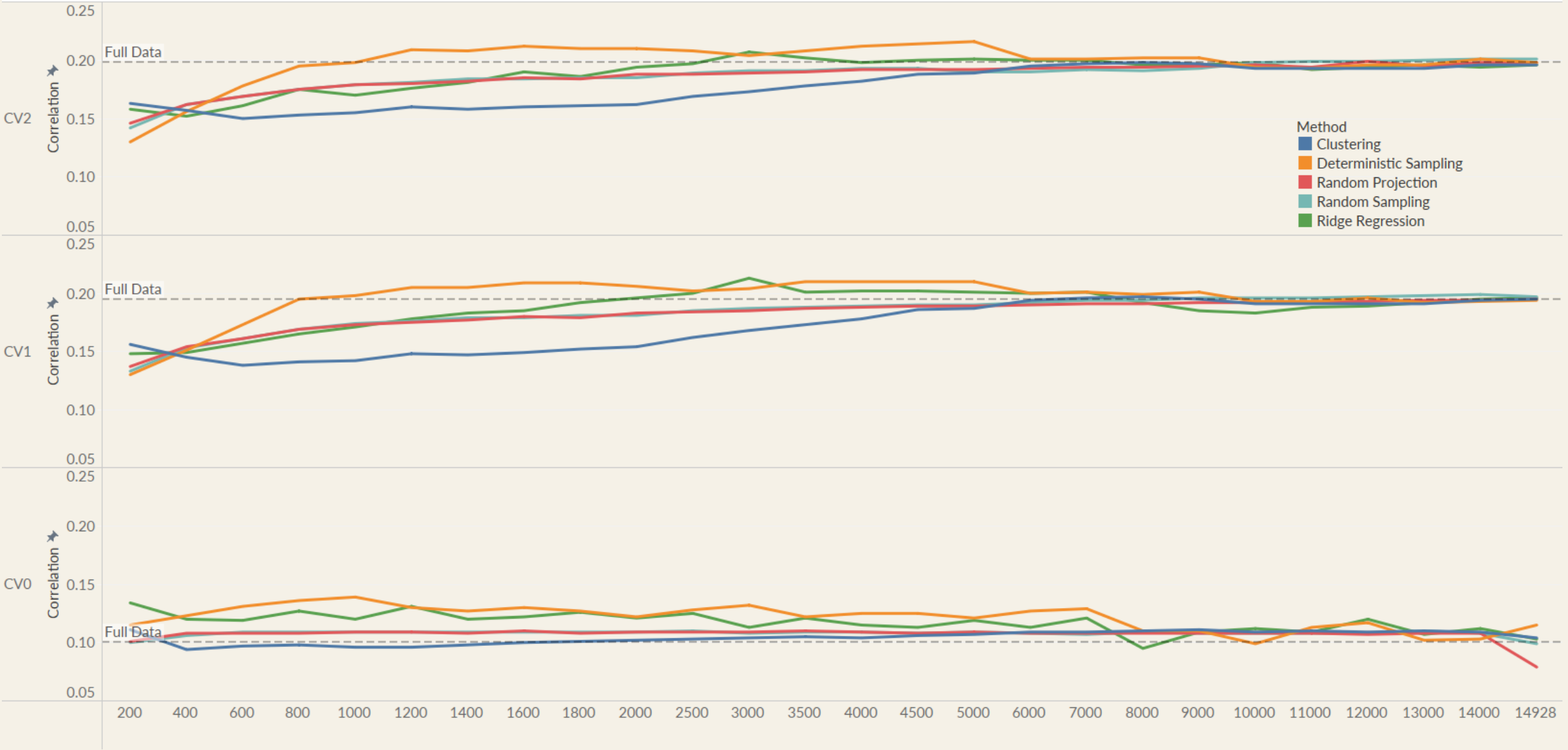
Performance of some lines that have been observed in some environments but not others

	CV2				
	E1	E2	E3	E4	E5
Line 1	Y_{11}	NA	Y_{13}	NA	Y_{15}
Line 2	Y_{21}	Y_{22}	Y_{23}	Y_{24}	Y_{25}
Line 3	Y_{31}	Y_{32}	NA	Y_{34}	Y_{35}
Line 4	NA	Y_{42}	Y_{43}	Y_{44}	Y_{45}
Line 5	Y_{51}	Y_{52}	Y_{53}	Y_{54}	NA

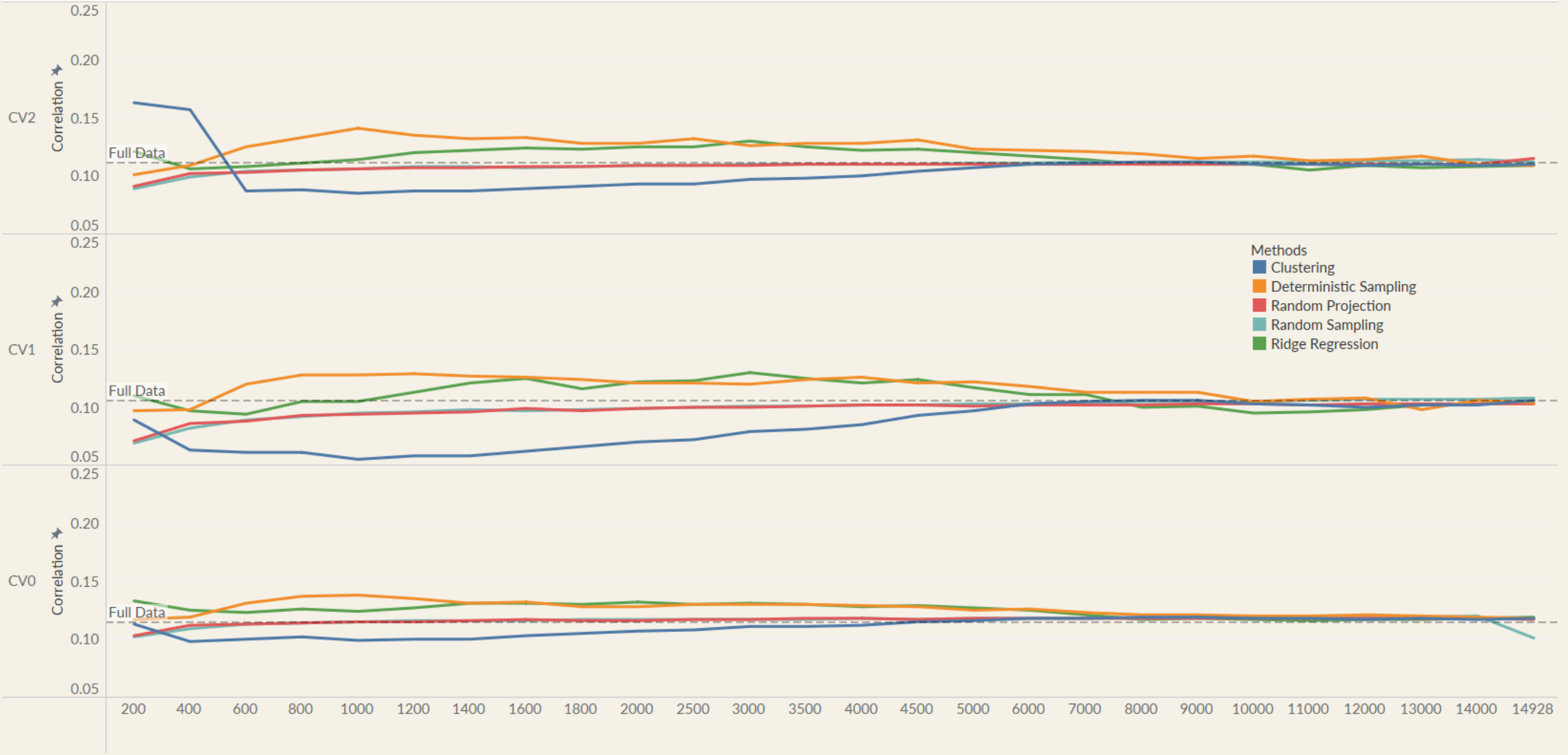
Dimensionality Reduction Methods

- Random projections [Ailon and Chazelle., 2009]
- Random sampling [Boutsidis et al., 2008]
- Deterministic sampling [Papiliopoulos et al., 2014]
- Clustering based [Sneath and Sokal, 1973]
- Ridge regression based [Hoerl and Kennard, 1970]

Cross-Validated Correlation Comparisons for the G + E + G x E model across DR Methods



Cross-Validated Correlation Comparisons for the G + E model across DR Methods



Conclusions

- Irrespective of DR method, full data correlation achieved with just a fraction of markers.
- Holds across models and cross-validation scheme combinations.
- Deterministic sampling gave greatest reduction for this chickpea data set.

Evaluating Dimensionality Reduction for Genomic Prediction

Contact:

vamsi.manthena@gmail.com

www.vamsimanthena.com

Select References

- Jarquin, D., Crossa, J., Lacaze, X., Du Cheyron, P., Daucourt, J., Lorgeou, J., Piraux, F., Guerreiro, L., Perez, P., Calus, M., Burgueno, J., and de los Campos, G. (2014), “A reaction norm model for genomic selection using high-dimensional genomic and environmental data”, *Theoretical and Applied Genetics*, *127(3)*, 595 - 607.
- Ailon, N., and Chazelle, B. (2009), “The Fast Johnson-Lindenstrauss Transform and Approximate Nearest Neighbors”, *SIAM Journal on Computing*, *39(1)*, 302-322.
- Boutsidis, C., Mahoney, M. W., and Drineas, P. (2009), “An Improved Approximation Algorithm for the Column Subset Selection Problem”, *Proceedings of the Twentieth Annual ACM-SIAM Symposium on Discrete Algorithms*, *Society for Industrial and Applied Mathematics*, pp. 968-977.
- Papailiopoulos, D., Kyrillidis, A., and Boutsidis, C. (2014), “Provable Deterministic Leverage Score Sampling," *arXiv:1404.1530 [cs, math, stat]*, . arXiv: 1404.1530.
- Roorkiwal, M., Jarquin, D., Singh, M. K., Gaur, P. M., Bharadwaj, C., Rathore, A., Howard, R., Srinivasan, S., Jain, A., Garg, V., Kale, S., Chitikineni, A., Tripathi, S., Jones, E., Robbins, K. R., Crossa, J., and Varshney, R. K. (2018), “Genomic-enabled prediction models using multi-environment trials to estimate the effect of genotype × environment interaction on prediction accuracy in chickpea”, *Scientific Reports*, *8(1)*, 11701.