Systematic Normalization with Multiple Housekeeping Genes for the Discovery of Genetic Dependencies in Cancer

Oliver Bonham-Carter[a] and Yee Mon Thu[b]
Depts of Computer Science[a] and Biology[b], Allegheny College
Meadville, PA

ALLEGHENY COLLEGE

https://www.cs.allegheny.edu
obonhamcarter@allegheny.edu
ythu@allegheny.edu

Project Objectives

We analyze gene expression data to discover pairs of genes whose expressions demonstrate patterns of correlation using a computational approach.
This project presents:

- A focus on genes suppressing genome instability (GIS genes) since function or expression may often be altered in cancer.
- A computational method to determine normalizing factors that make it possible to discover pairs of GIS genes which show consistent correlation.
- Normalizing factors, created by a selection of cancer-specific housekeeping genes providing ability to compare gene expressions data and to treat these values by linear regression.

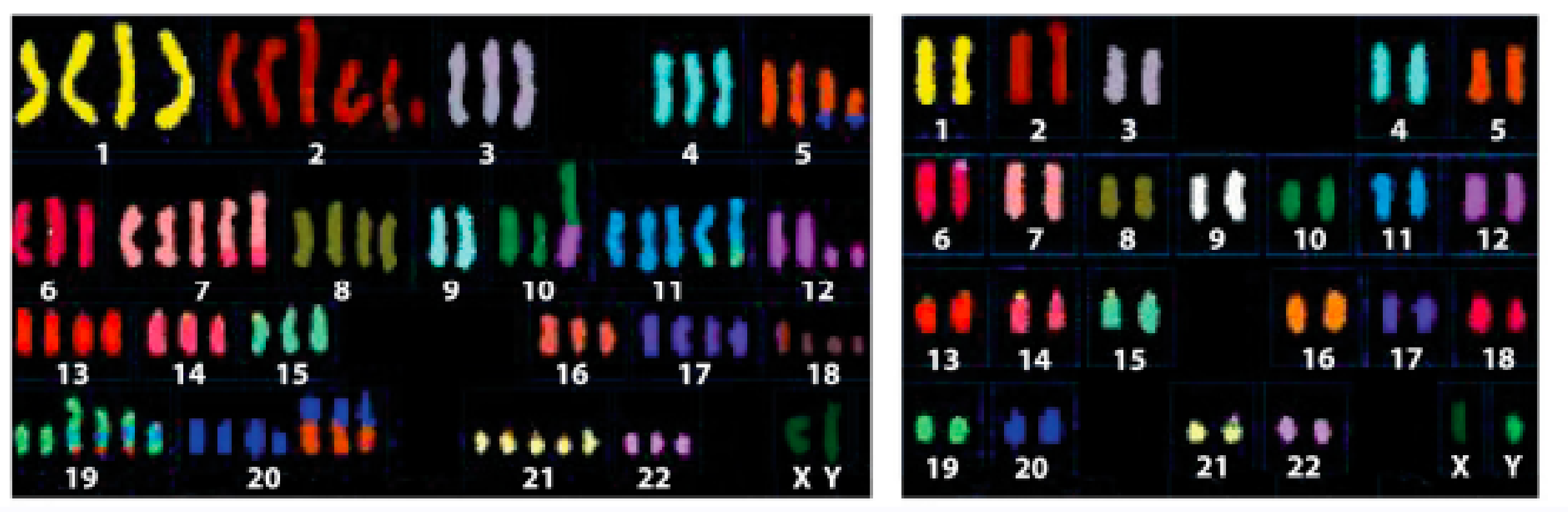


Figure: 1. A karyotype of reduced stability of typical cancer cells, left. Increased stability in cells, right.

Gene Correlation

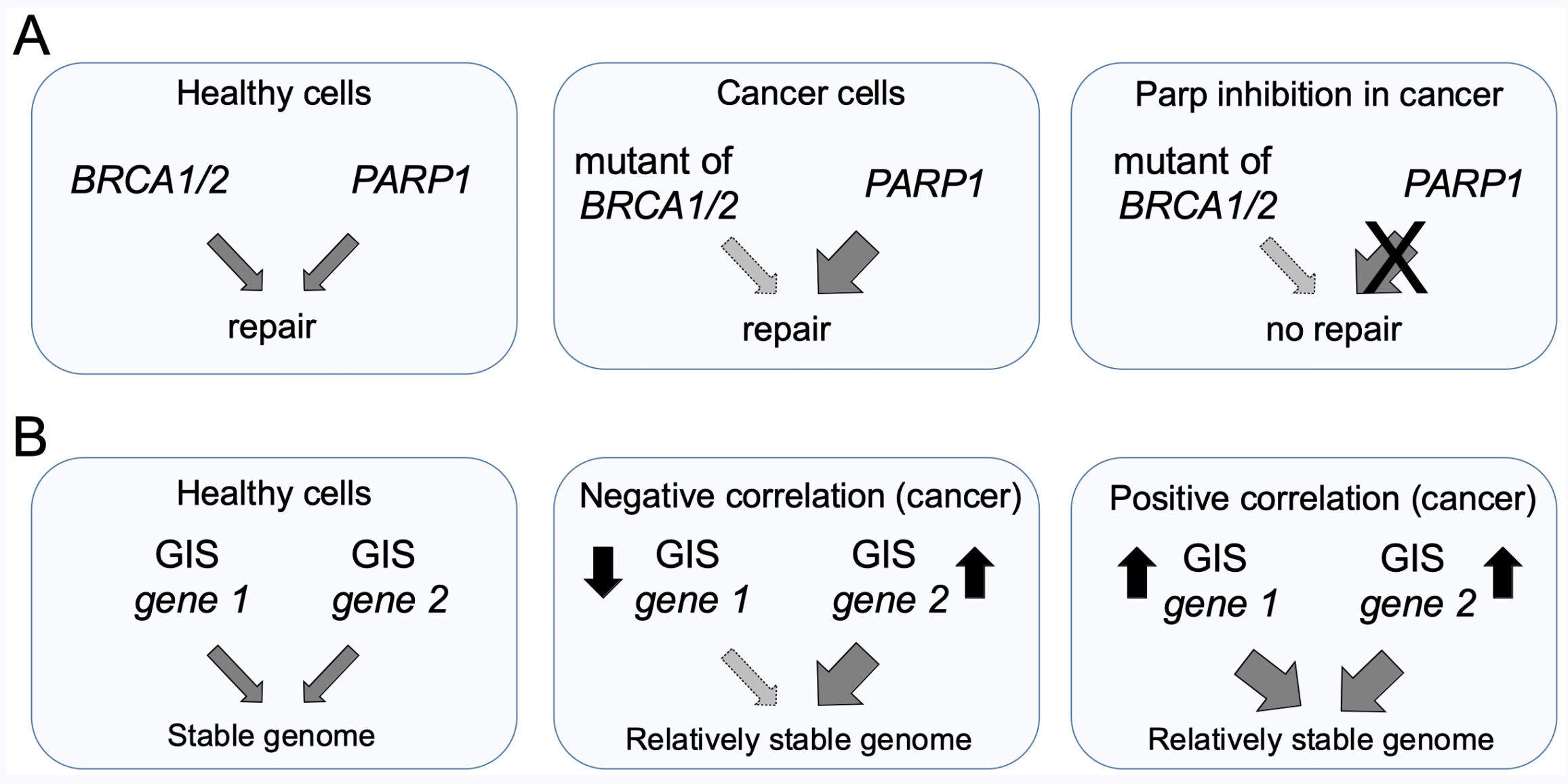


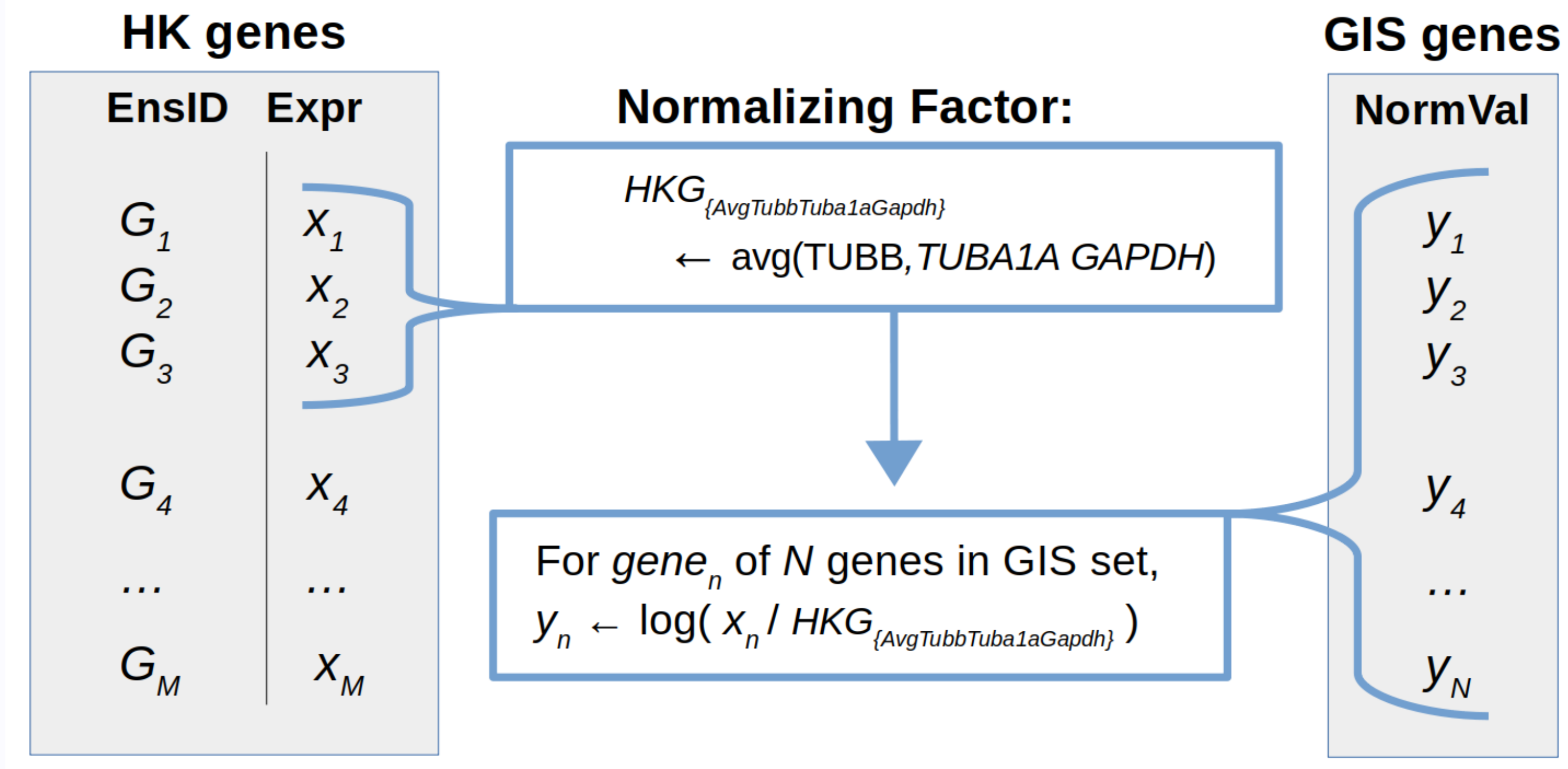
Figure: 2. Co-Dependance by genes

Genetic interactions between genes are often responsible for DNA repair or genome stability.

- FPKM datasets are used to determine the existence of a positive or negative correlation between the expression of two given GIS (suppress genome instability) genes in cancer, Shown in Figure 2.
- Correlations can reveal if two GIS genes coordinate or if an alteration in the expression of one GIS gene increases dependency of cancer cells on another GIS gene.

Method: Proof of Concept

Random selection of ten data sets of breast cancer gene expression.



HK genes

EnsID Expr

G_1 x_1
G_2 x_2
G_3 x_3
G_4 x_4
… …
G_M x_M

Normalizing Factor:

$HKG_{\{AvgTubbTuba1aGapdh\}} \leftarrow$ avg(TUBB, *TUBA1A GAPDH*)

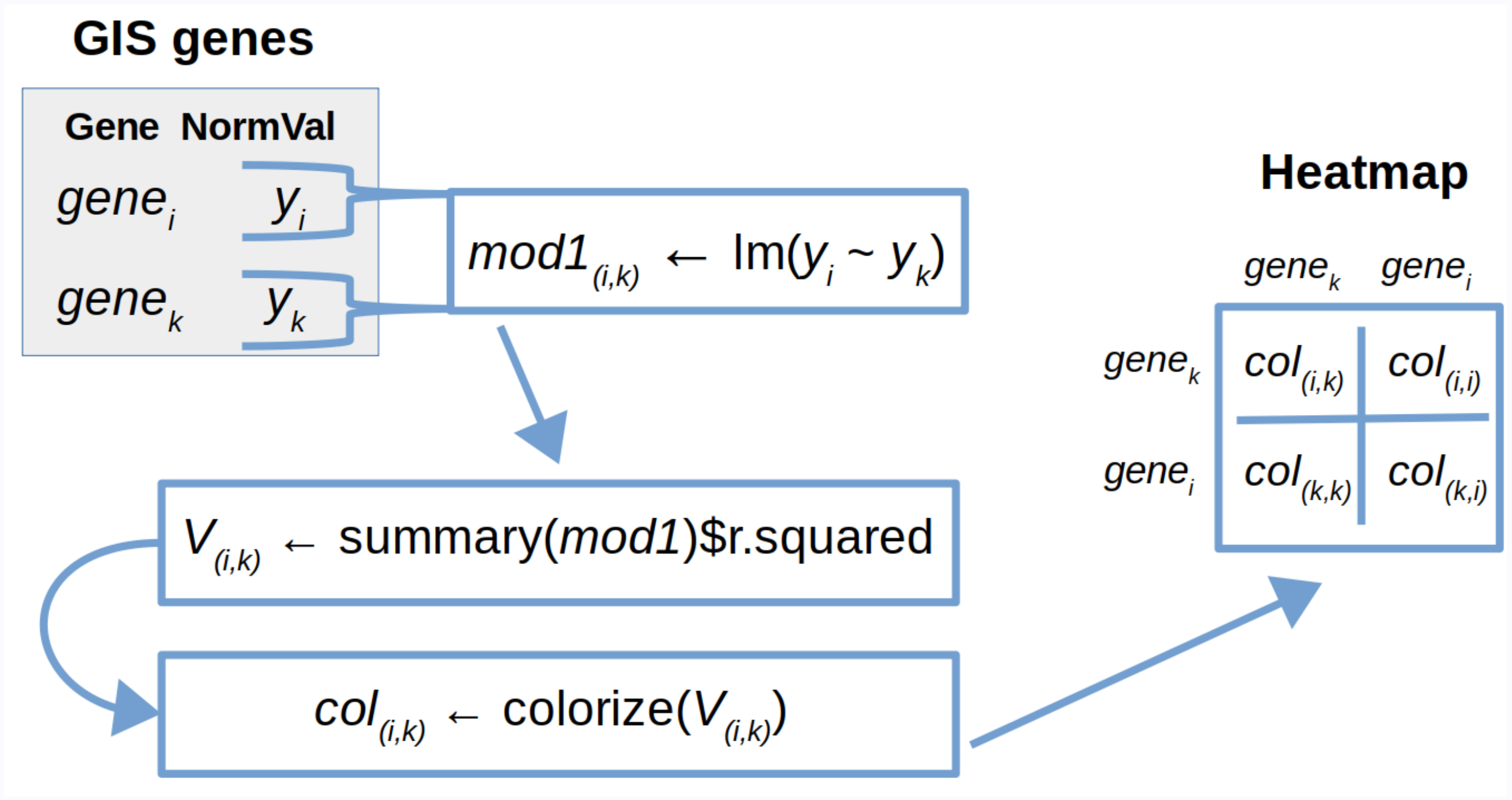
For *gene*$_n$ of *N* genes in GIS set,
$y_n \leftarrow \log(x_n / HKG_{\{AvgTubbTuba1aGapdh\}})$

GIS genes

NormVal

y_1
y_2
y_3
y_4
…
y_N

Figure: 3. Determining the normalizing factors



GIS genes

Gene NormVal
gene$_i$ y_i
gene$_k$ y_k

$mod1_{(i,k)} \leftarrow \mathrm{lm}(y_i \sim y_k)$

$V_{(i,k)} \leftarrow$ summary(*mod1*)$r.squared

$col_{(i,k)} \leftarrow \mathrm{colorize}(V_{(i,k)})$

Heatmap

	gene$_k$	*gene*$_i$
gene$_k$	$col_{(i,k)}$	$col_{(i,i)}$
gene$_i$	$col_{(k,k)}$	$col_{(k,i)}$

Figure: 4. HouseKeeping genes to normalize data sets

- Genomic Data Commons Data Portal (National Cancer Institute).
- Selected subset of genes limited search space, were specific to breast cancer research and reduced noise in results.

R-Squared Values from Linear Models

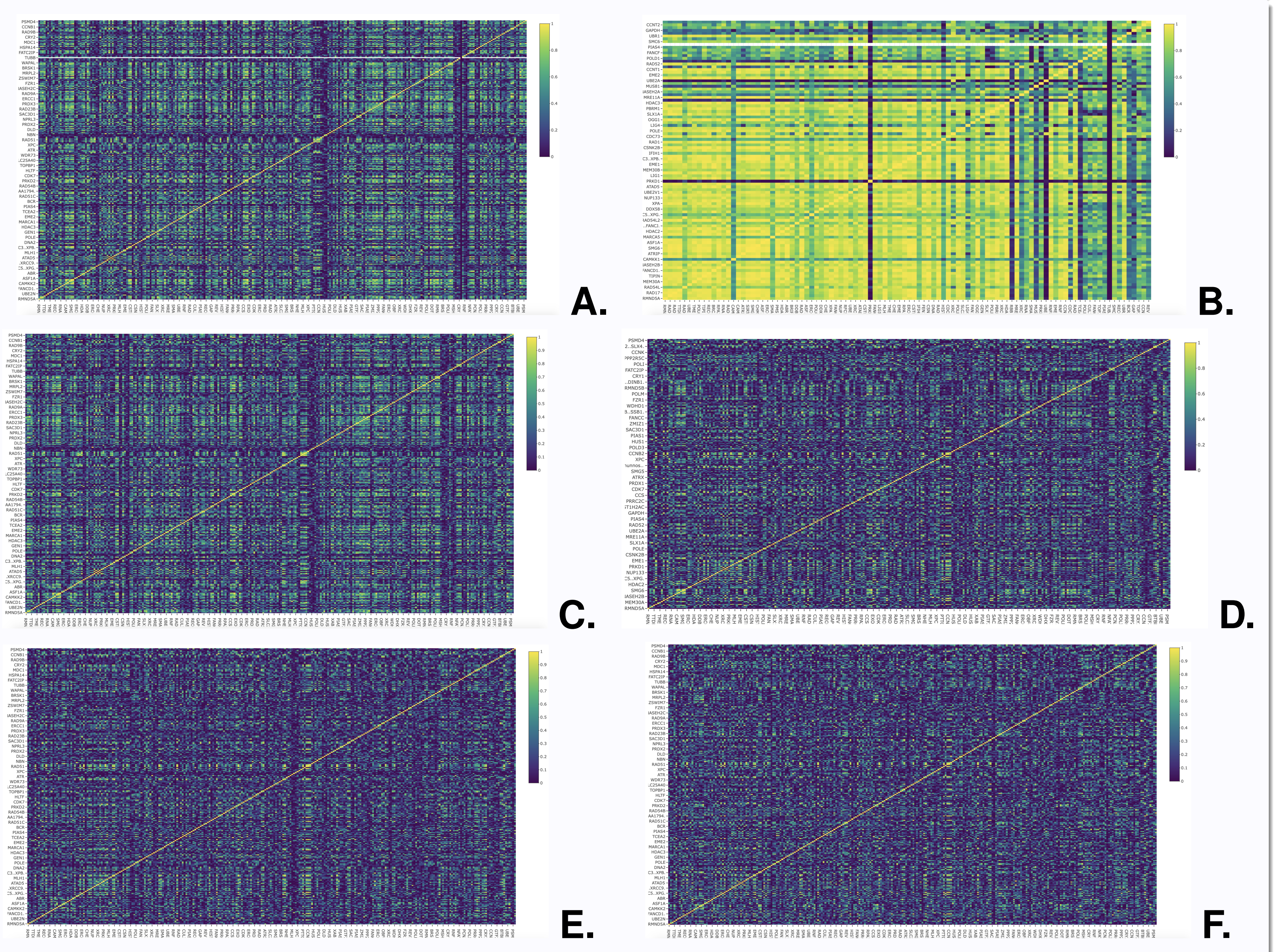


Figure: 5. Heatmaps of R^2 values, derived from normalizing factors

- R^2 values in heatmaps from linear regression models, *all-against-all* regressions of GIS genes.
- Left to right, R^2 values, derived from single housekeeping genes to create normalizing factors; HKG_{Tubb}, HKG_{Tuba1a}, HKG_{Tubb}, $HKG_{AvgTubbTuba1aGapdh}$, see Figure 5,{**A** - **C**}, resp.
- Ten housekeeping genes: HKG_{AvgG1}, HKG_{AvgG3}, HKG_{AvgG3}, See Figure 5,{**D** - **F**}, resp.

Results

- HKG_{Tuba1a} (**B**): high diversity of R^2 values indicated poor normalizing, biologically improbable.
- According to regression model results, normalizing factors that were created from larger groups of housekeeping gene produced correlations implying more biological relevance.

Conclusions

- Our results (examples shown in heatmaps of Figure 5) indicate normalized data and contain biologically probable findings.
- **Single Expression Normalization**: Normalizing factors derived from single housekeeping genes did not provide generally consistent correlations that were biologically probable.
- **Multiple Expression Normalization**: Using the averaged expression values of multiple housekeeping genes was an effective approach to finding biologically relevant consistency across our data sets.
- Our method enabled us to identify the co-expression of gene pairs in breast cancer tissues and compared diverse normalization factors.
- Our study also allows reproducibility across data sets and allows for scalability in gene correlation research.