



GenExSt: A Tool to Identify Correlation of Gene Expression after Normalization with Housekeeping Genes

Oliver BONHAM-CARTER and Yee Mon THU

Allegheny College

520 N. Main Street

Meadville, PA 16335

<https://allegheny.edu/>

Project Link: <https://github.com/developmentAC/genExSt>

14th Feb 2021

Cancer from Instability

Genomic instability characterizes many cancers

Instability

Research Interest

GenExSt

Method

Results and Conclusions

Thank You

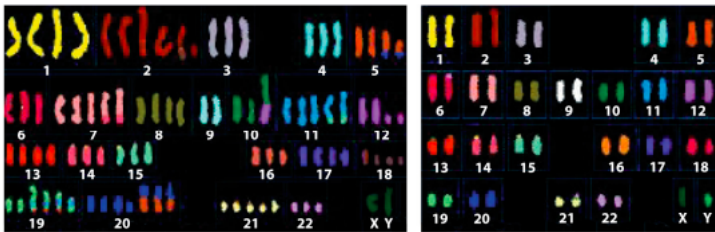


Figure: 1. Alberts, B. "Molecular biology of the cell Sixth edition Ch. 14, 755–756." Garland Science, Taylor and Francis Group (2015).

- Karyotypes of cancerous cells
 - left: very unstable
 - Right: unstable

Instability From Genes

Instability

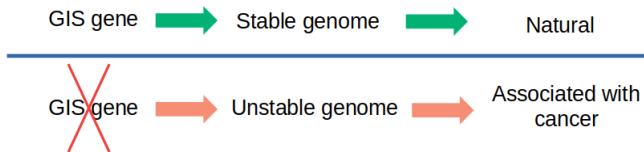
Research
Interest

GenExSt

Method

Results and
Conclusions

Thank You



- GIS genes: Genes that suppress genome instability
- When GIS gene(s) is /are disabled, then an elevated risk of cancer association may arise in genome
- Instability seems to enable some cancers to survive

Equilibrium and Instability

Instability

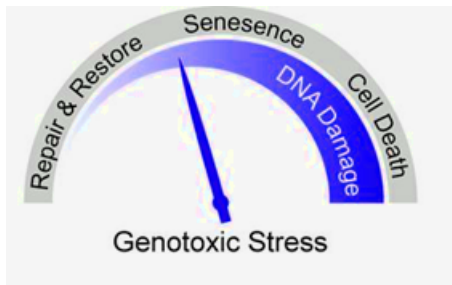
Research
Interest

GenExSt

Method

Results and
Conclusions

Thank You



- Relatively low levels of instability in the genome:
 - Repair and restore mechanisms function
 - Natural mutations to inspire genetic diversity
- High levels of instability:
 - Dysfunction, breakdown and *likely* cell death
 - Has cancer become dependent on some part of its unstable environment?

Correlations in Genome

A dependence (*correlation*) between genes

Instability

Research
Interest

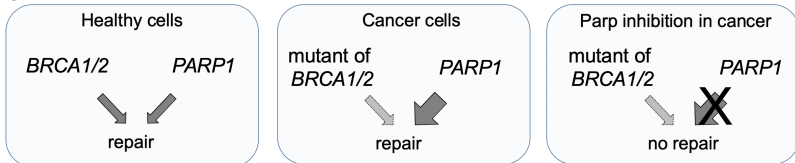
GenExSt

Method

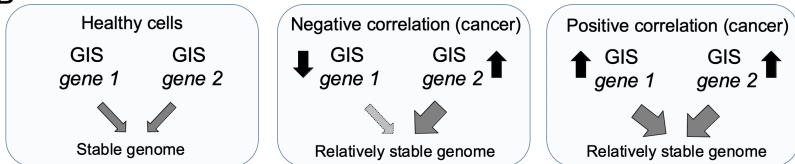
Results and
Conclusions

Thank You

A



B



- **A.** Gene interactions introduce stability (*instability*) in the genome
- **B.** GIS genes may work together to gain stability.

A hypothesis

Finding correlations between genes

Instability

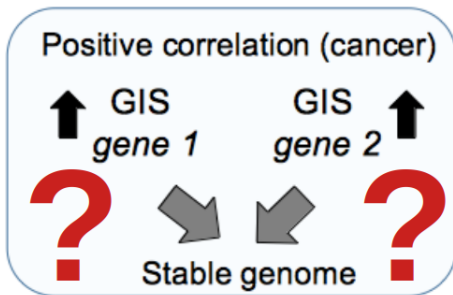
Research
Interest

GenExSt

Method

Results and
Conclusions

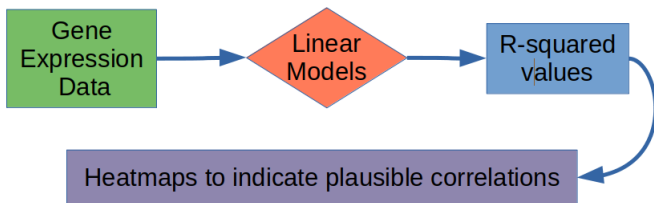
Thank You



- Change in expression of a GIS gene makes cancer cells more dependent on the function of other GIS gene(s) to prevent genome instability
- *Can there be a correlation between GIS genes?*
- *Goal: Discover evidence of (plausible) correlations*

Finding Correlations

We present GenExSt to automate the search



- A large-scale approach to discovering correlations between the genes of our study groups
- R^2 values from linear regression for correlation
- Heatmaps to visualize the correlations for tests

We present a tool (written in Python) for the analysis of gene expression data to determine plausible correlations.

24 July 2020 | version: 2_ii | Oliver Bonham-Carter | obonhamcarter@allegheny.edu

The GeneExPy program to perform linear regression over GDP datasets.

Library installation notes:

plotly:

pip3 install plotly, or try running python3 -m pip install scikit-learn

scikit-learn:

python3 -m pip install scikit-learn, maybe necessary: pip3 install scipy

+ 😊 USAGE: programName <any key to launch>

+ INPUT directory: (your data files are here) : data/

+ OUTPUT directory: (your output is placed here) : 0out/

+ Note:

Use parameter <<heatmap>> or <rsqu>> to ensure that

Plotly and the statistical libraries have been correctly installed.

Note: the data directory cannot handle subdirectories holding data. Please place the text files into this data directory without using subdirectories.

Instability

Research
Interest

GenExSt

Method

Results and
Conclusions

Thank You

GenExSt; Gene Expression Analysis

Date: 30 July 2020, Ver: 1_i



What are we doing with this data?

Gene Expression Analysis

Default data directory is :data/

Enter the path to data file.

data/

The input data directory : data/

☐ Show raw data



Gene Expression

Threshold of R-squared values



threshold : [0.93]

Instability Suppressing (GIS) genes

Choose a csv file of genes for correlation analysis.

pickThese.csv

Entered filename: pickThese.csv

HouseKeeping (HK) genes

Choose a csv file of genes for normalization. Note: this file must include all data files and those used for normalizing.

normNames_i.csv

Entered filename: normNames_i.csv

Compute

🐾 WOO WOO!! 🐾

ok :: computer

GitHub

Some Points

- <https://github.com/developmentAC/genExSt>
- Bash version: command-line execution
- Browser version: clicking and dragging utilization
- Same algorithms and output in each version
- Run in containers to setup libraries (Docker scripts are provided)



Genes for Normalization (for example)

Input parameter: `normNames.csv`

Instability

Research
Interest

GenExSt

Method

Normalization and
Models

Results and
Conclusions

Thank You

Table: The first set of HK genes for multiple expression normalization. In the group, the expression values of all ten were averaged in each set d to determine its normalizing factor. Here, we provide the ontology group, Uniprot ID, *human* gene ID and Ensembl IDs.

Ontology	UnitProtKB	Gene ID	EnsNum
Proteasome	Q9Y5K5	<i>UCHL5</i>	ENSG00000116750
Ribosome	Q96EL2	<i>MRPS24</i>	ENSG00000062582
ER	Q9P0I2	<i>EMC3</i>	ENSG00000125037
ER targeting	O76094	<i>SRP72</i>	ENSG00000174780
spilisosome	O15234	<i>CASC3</i>	ENSG00000108349
Prp19	Q9BZJ0	<i>CRNKL1</i>	ENSG00000101343
spilisosomal snRNP	P62310	<i>LSM3</i>	ENSG00000170860
Proteasome	P62195	<i>PSMC5</i>	ENSG00000087191
Methylsome	Q9BQA1	<i>WDR77</i>	ENSG00000116455
Translation	Q9UBQ5	<i>EIF3K</i>	ENSG00000178982

Genome Instability Suppressing genes (for example)

Input parameter: `pickThese.csv`

Instability

Research
Interest

GenExSt

Method

Normalization and
Models

Results and
Conclusions

Thank You

Table: The GIS genes. We provide the Gene ID and the Ensemble IDs. The gene expressions are taken from from Genome Data Commons Data Portal files.

Gene ID	EnsNum
XRCC4	ENSG00000152422.14
XRCC5	ENSG00000079246.14
XRCC6	ENSG00000196419.11
ZMIZ1	ENSG00000108175.15
ZSWIM7	ENSG00000214941.6
TUBB	ENSG00000196230.11
TUBA1A	ENSG00000167552.12
GAPDH	ENSG00000111640.13
MCM10	ENSG00000065328.15
...	...

Normalization

All values had to be normalized to enable comparisons

Instability

Research
Interest

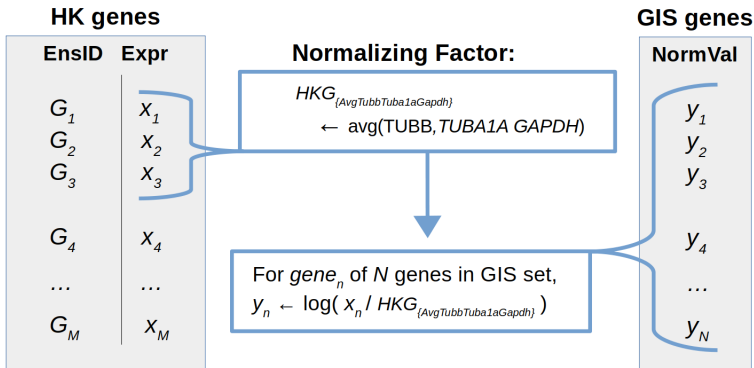
GenExSt

Method

Normalization and
Models

Results and
Conclusions

Thank You

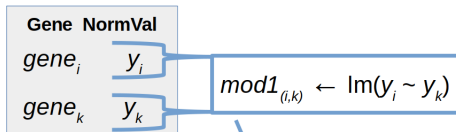


- The HK gene expressions were averaged to create the normalizing factors.
- The GIS gene expressions were divided by this value and log transformed for enhancement.

Linear Models and R^2 Values

R code and flow

GIS genes

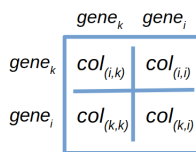


$V_{(i,k)} \leftarrow \text{summary}(mod1)\$r.squared$

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

Heatmap

	$gene_k$	$gene_i$
$gene_k$	$col_{(i,k)}$	$col_{(i,i)}$
$gene_i$	$col_{(k,k)}$	$col_{(k,i)}$



- Each normalized gene was regressed over all others in an *all-against-all* test
- The R^2 values were extracted and placed into heatmaps
- These R^2 values were used to inform of correlation

Linear Models and R^2 Values

R code and flow

Instability

Research
Interest

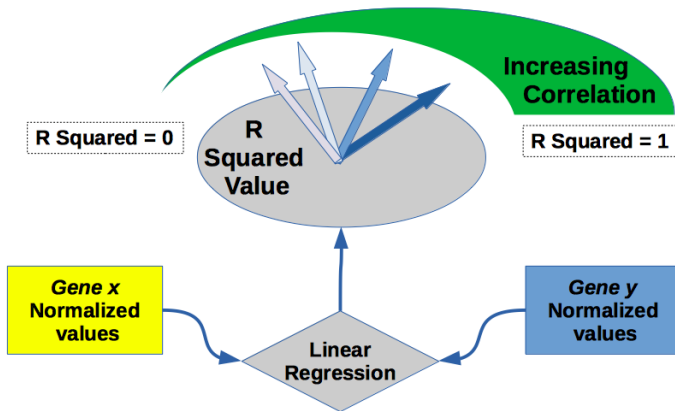
GenExSt

Method

Normalization and
Models

Results and
Conclusions

Thank You



- Scale $[0,1]$: High values of R^2 indicated strong correlations between values from each set of corresponding genes

Normalizing Techniques

Few HK genes to normalize GIS genes

Instability

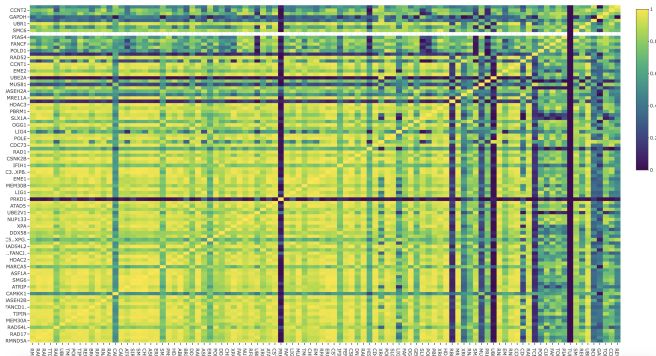
Research
Interest

GenExSt

Method

Results and
Conclusions

Thank You



- The R^2 values from normalization where too few Housekeeping genes were used to normalize provided poor results.
- Biologically unlikely: Too many R^2 values to suggest high correlations



Normalizing Techniques

Ten HK genes to normalize GIS genes: Trial 1

Instability

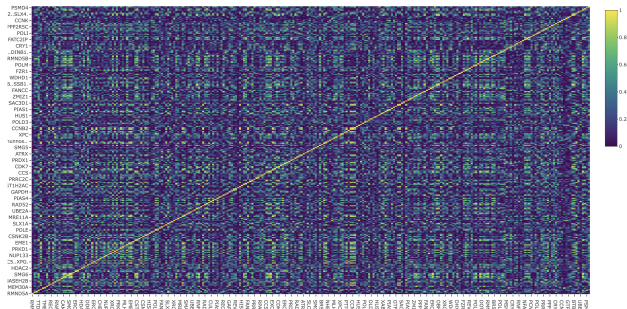
Research
Interest

GenExSt

Method

Results and
Conclusions

Thank You



- The R^2 values from normalization where ten housekeeping genes were used to normalize provided poor results.
- Biologically plausible; comparable with other sets of gene expressions, normalized by same ten housekeeping genes taken from their set.

Normalizing Techniques

Ten HK genes to normalize GIS genes: Trial 2

Instability

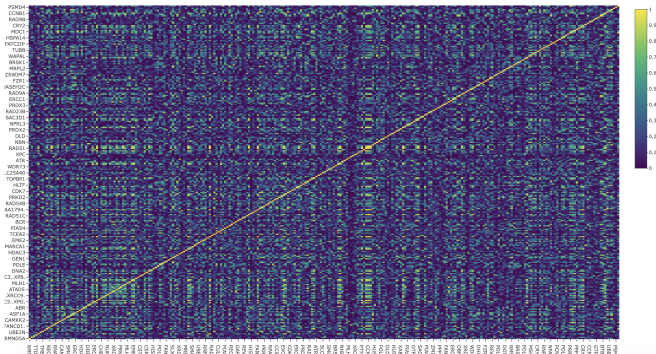
Research
Interest

GenExSt

Method

Results and
Conclusions

Thank You





Ten HK genes to normalize GIS genes: Trial 3

Thank You



Normalizing Techniques

Ten HK genes to normalize GIS genes: Trial 2

Instability

Research
Interest

GenExSt

Method

Results and
Conclusions

Thank You

- **Single Expression Normalization:** Too few housekeeping genes used to normalize was not effective: correlations were biologically improbable.
- **Multiple Expression Normalization:** Multiple housekeeping genes used to normalize was successful in providing means to compare genes from multiple sets and to find biologically relevant correlations
- Our method was a pilot study, provided a good means to normalize data, and identified the co-expression of gene pairs in breast cancer tissues

Thank You! Questions?

obonhamcarter@allegheny.edu
ythu@allegheny.edu

<https://www.cs.allegheny.edu/sites/obonhamcarter/>
<https://sites.allegheny.edu/bio/faculty/yee-mon-thu/>

