Statistical learning analysis of thyroid cancer microarray data

IVÁN PETRINI, ROCÍO CECCHINI,
IGNACIO PONZONI AND JESSICA CARBALLIDO

I C I C

CONICET

U N S


UNS
ICIC
DCIC
ICIC
CONICET
DEPARTAMENTO DE CIENCIAS E INGENIERÍA DE LA COMPUTACIÓN

Motivation

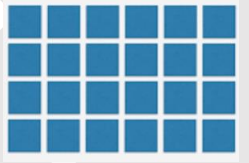
Many Data – Many ways

Motivation

Many Data – Many ways

Unstructured

Structured



Qualitative

Quantitative

Nominal

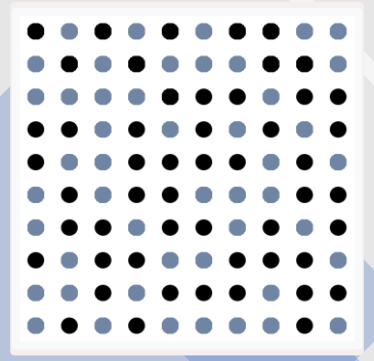
Ordinal

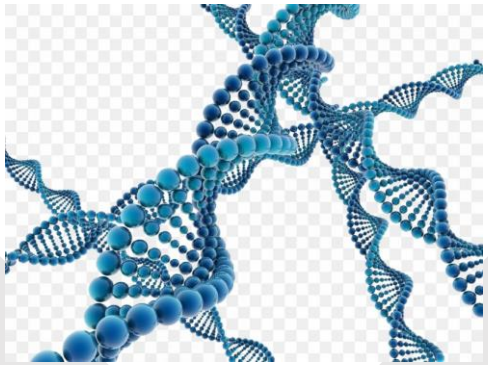
Continuous

Discrete

Motivation

From gene information to phenotype information





Motivation Scenario

- Microarray Data

Precise, productive, efficient.

Can be used for gene selection and classification

Provides massive amount of data

Motivation Scenario

- Microarray Data

Powerful across many disciplines

Is used for feature selection, classification, clustering

Gives a way to automate and works on massive data scenarios

Motivation Scenario

- Machine Learning and Gene Expression Data

Identify strong correlations

Grouping

Identify associations

Motivation Scenario

- Thyroid Cancer

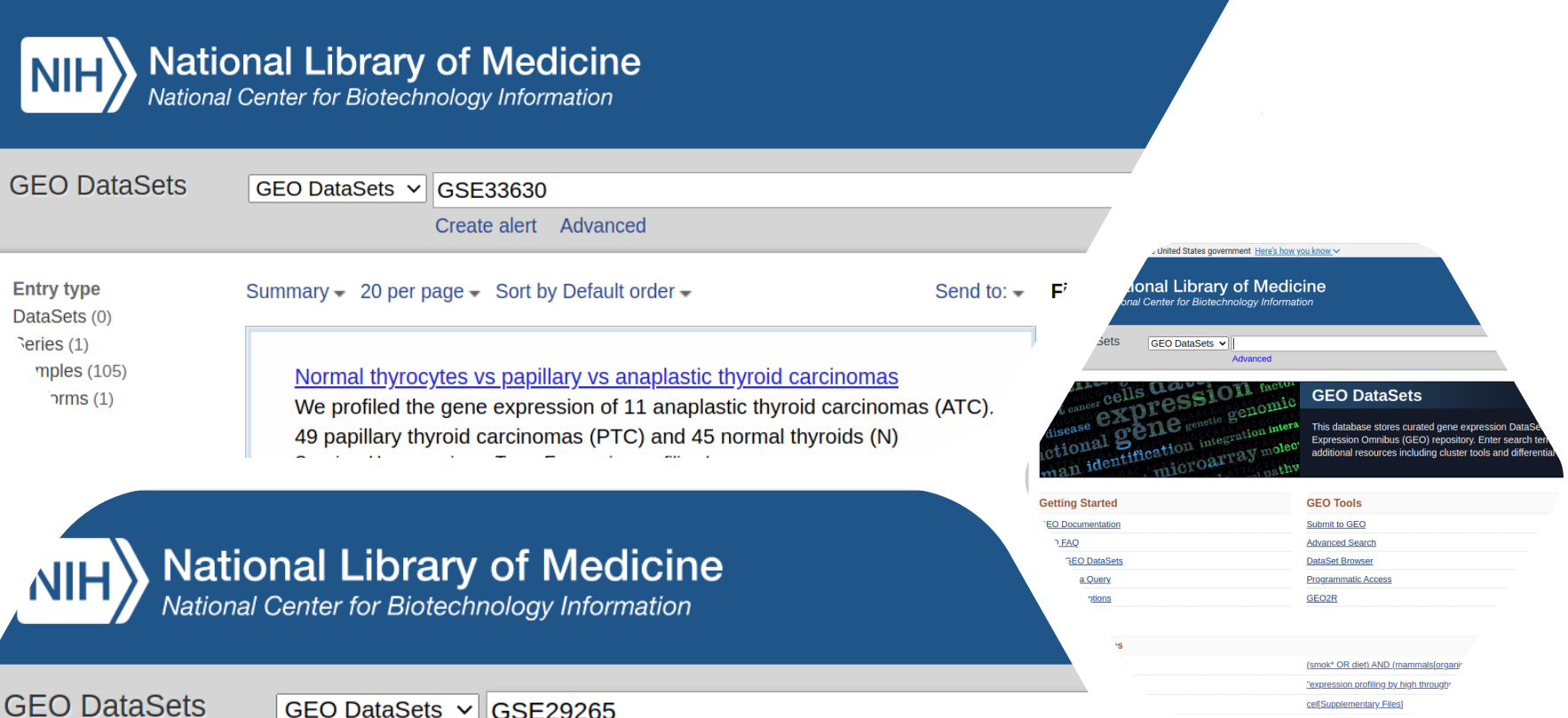
Most common malignant endocrine disease

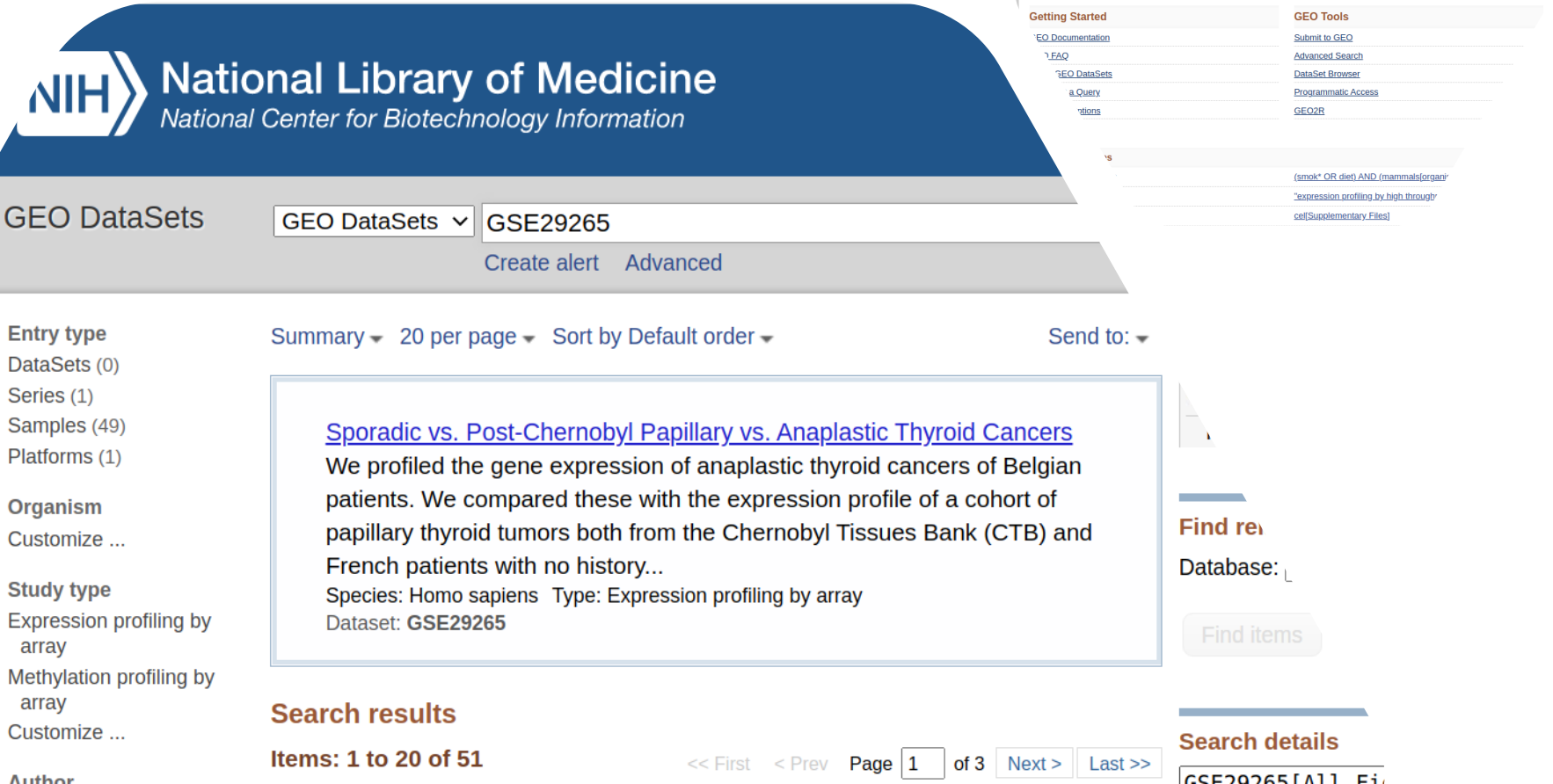
PTC constitutes between 70% and 80% of all cases

ATC accounts for 1-2% of all cases

> 50% of deaths are related to this type

mean survival rate is six months

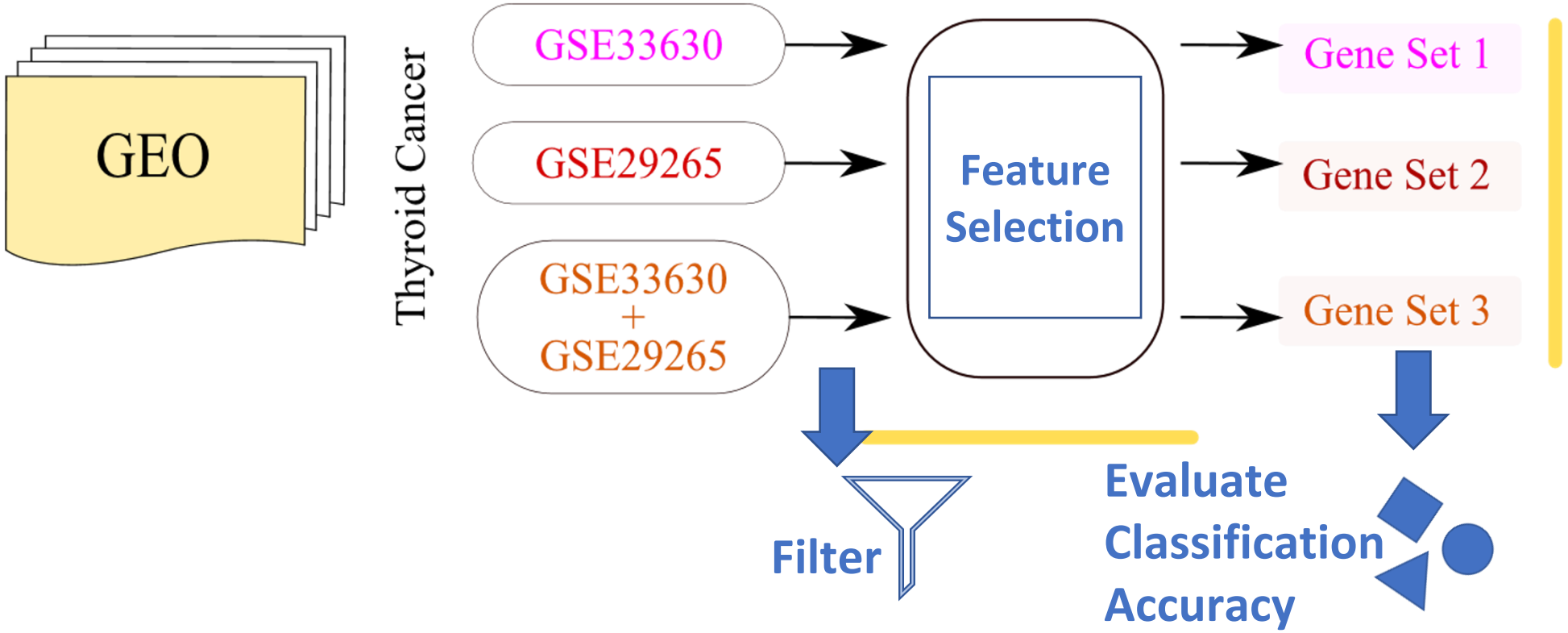

National Library of Medicine
National Center for Biotechnology Information
GEO DataSets
GEO DataSets
GSE33630
Create alert
Advanced
Entry type
DataSets (0)
eries (1)
mples (105)
rms (1)
Summary
20 per page
Sort by Default order
Send to:
Normal thyrocytes vs papillary vs anaplastic thyroid carcinomas
We profiled the gene expression of 11 anaplastic thyroid carcinomas (ATC).
49 papillary thyroid carcinomas (PTC) and 45 normal thyroids (N)


National Library of Medicine
National Center for Biotechnology Information
GEO DataSets
GEO DataSets
GSE29265
Create alert
Advanced
Entry type
DataSets (0)
Series (1)
Samples (49)
Platforms (1)
Organism
Customize ...
Study type
Expression profiling by array
Methylation profiling by array
Customize ...
Summary
20 per page
Sort by Default order
Send to:
Sporadic vs. Post-Chernobyl Papillary vs. Anaplastic Thyroid Cancers
We profiled the gene expression of anaplastic thyroid cancers of Belgian patients. We compared these with the expression profile of a cohort of papillary thyroid tumors both from the Chernobyl Tissues Bank (CTB) and French patients with no history...
Species: Homo sapiens Type: Expression profiling by array
Dataset: GSE29265
Search results
Items: 1 to 20 of 51
<< First
< Prev
Page 1 of 3
Next >
Last >>
Find re
Database:
Find items
Search details

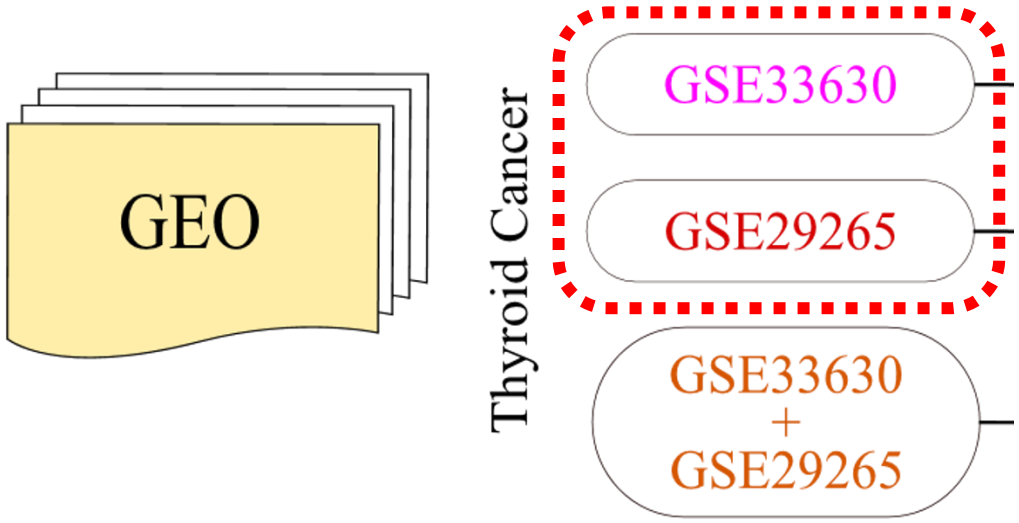
GEO DataSets
Getting Started
GEO Tools
Submit to GEO
Advanced Search
DataSet Browser
Programmatic Access
GEO2R

Datasets

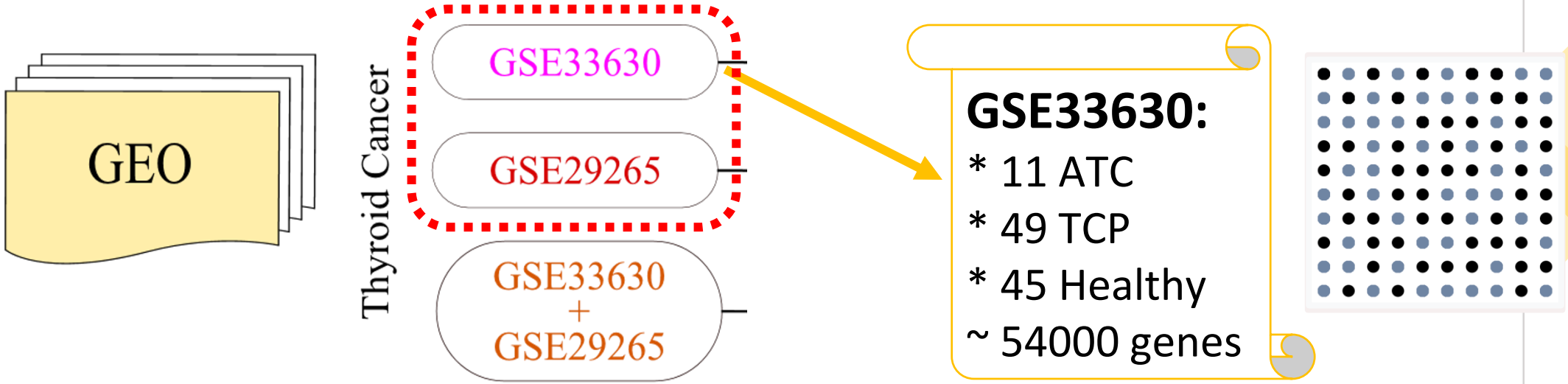
Workflow


GEO
Thyroid Cancer
GSE33630
GSE29265
GSE33630
+
GSE29265
Feature
Selection
Gene Set 1
Gene Set 2
Gene Set 3
Filter
Evaluate
Classification
Accuracy

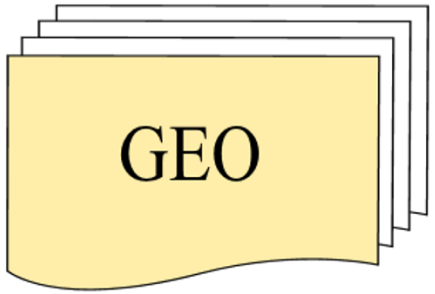
Workflow – Phase 1 – Individual Analysis



Workflow – Phase 1 – Individual Analysis



Workflow – Phase 1 – Individual Analysis – Filtering



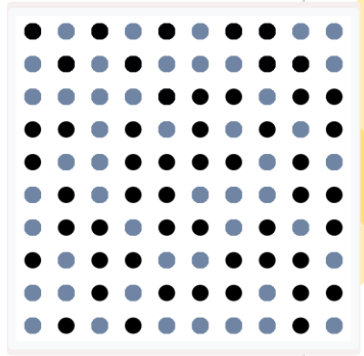
Thyroid Cancer

- GSE33630
- GSE29265
- GSE33630 + GSE29265

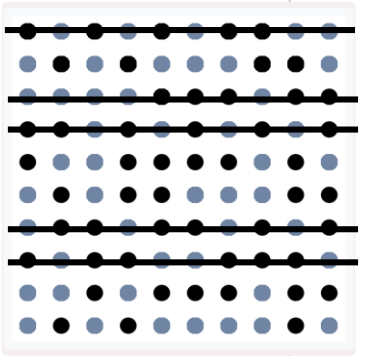
GSE33630:

* 11 ATC
* 49 TCP
* 45 Healthy

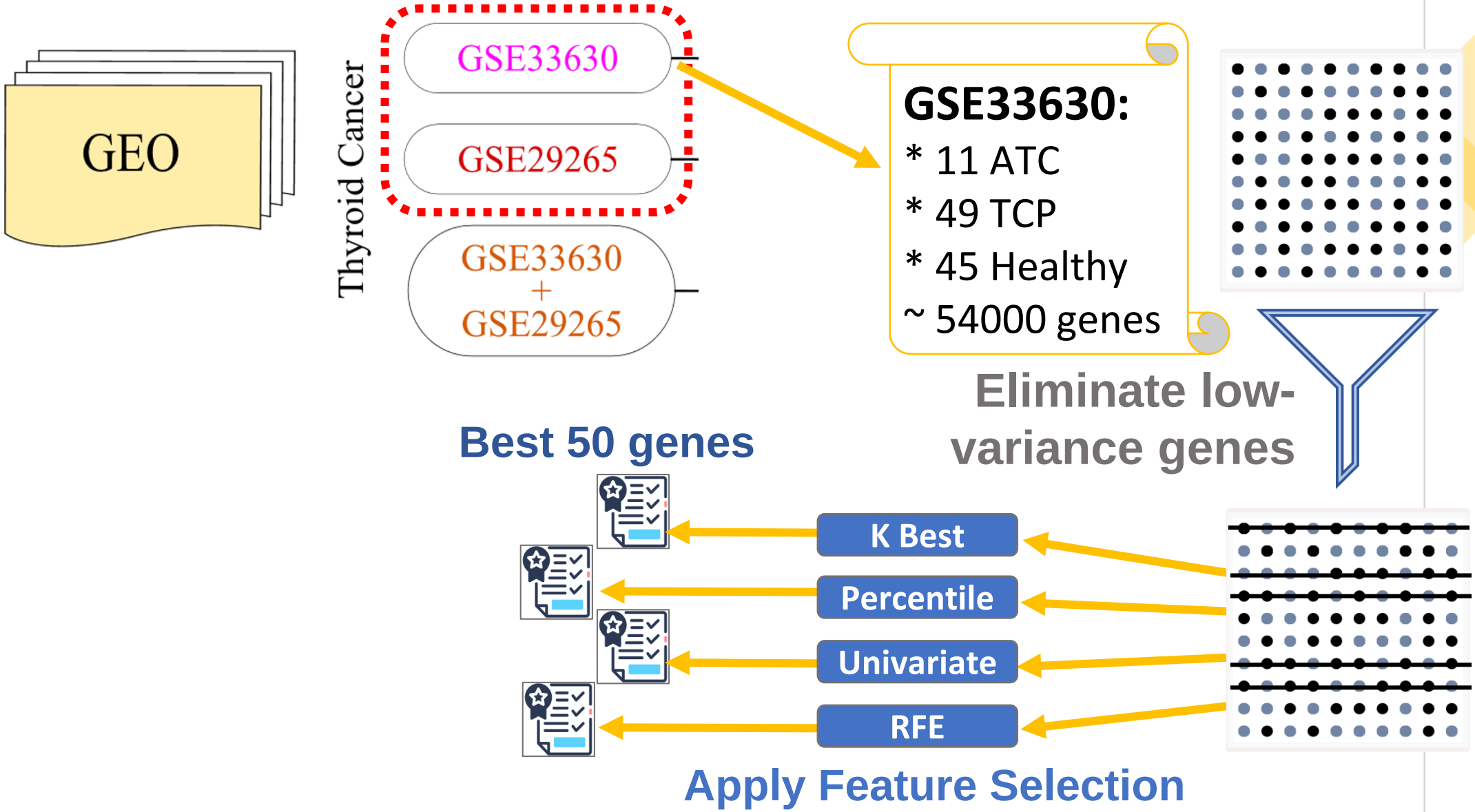
~ 54000 genes



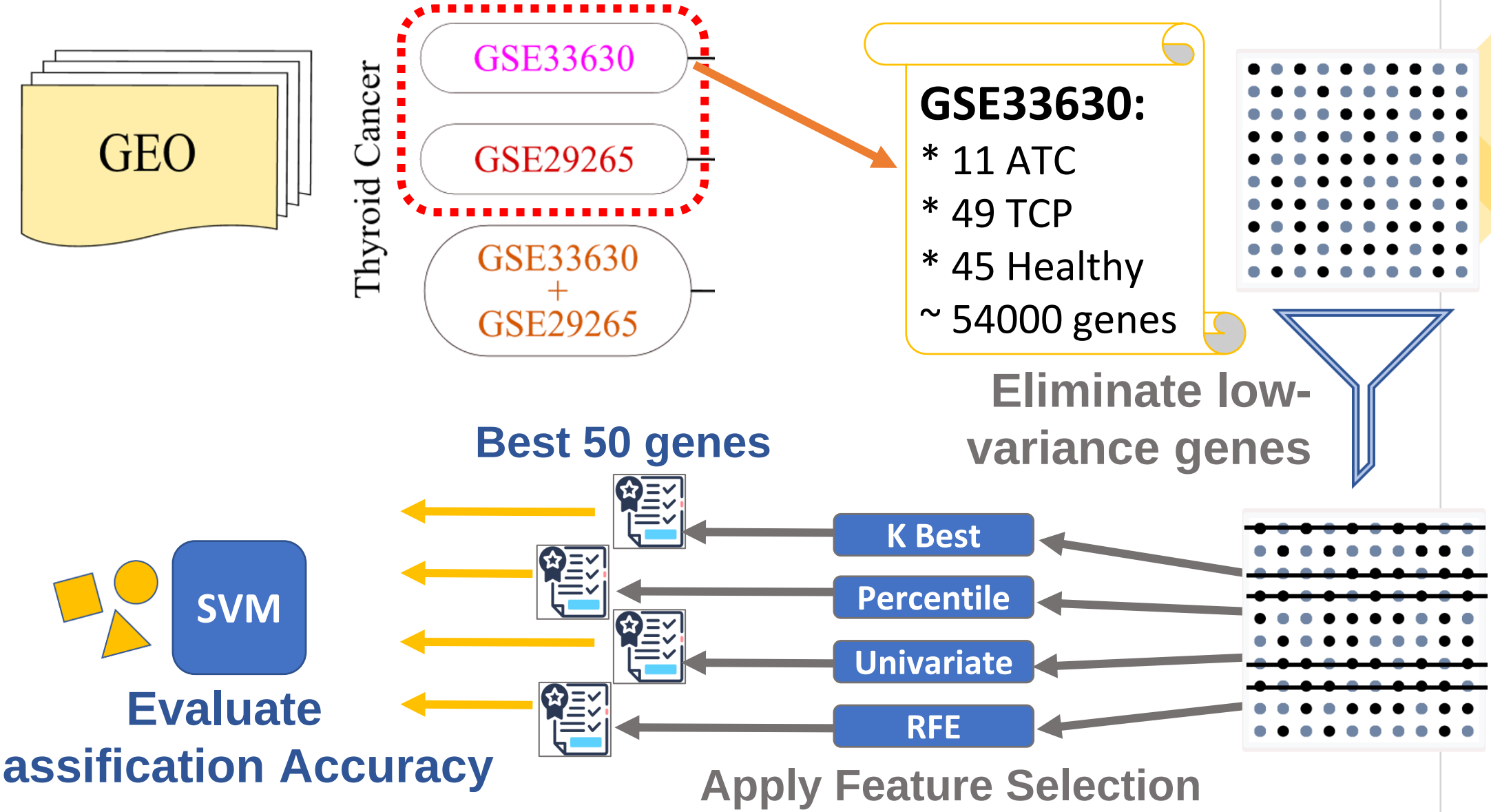
Eliminate low-variance genes



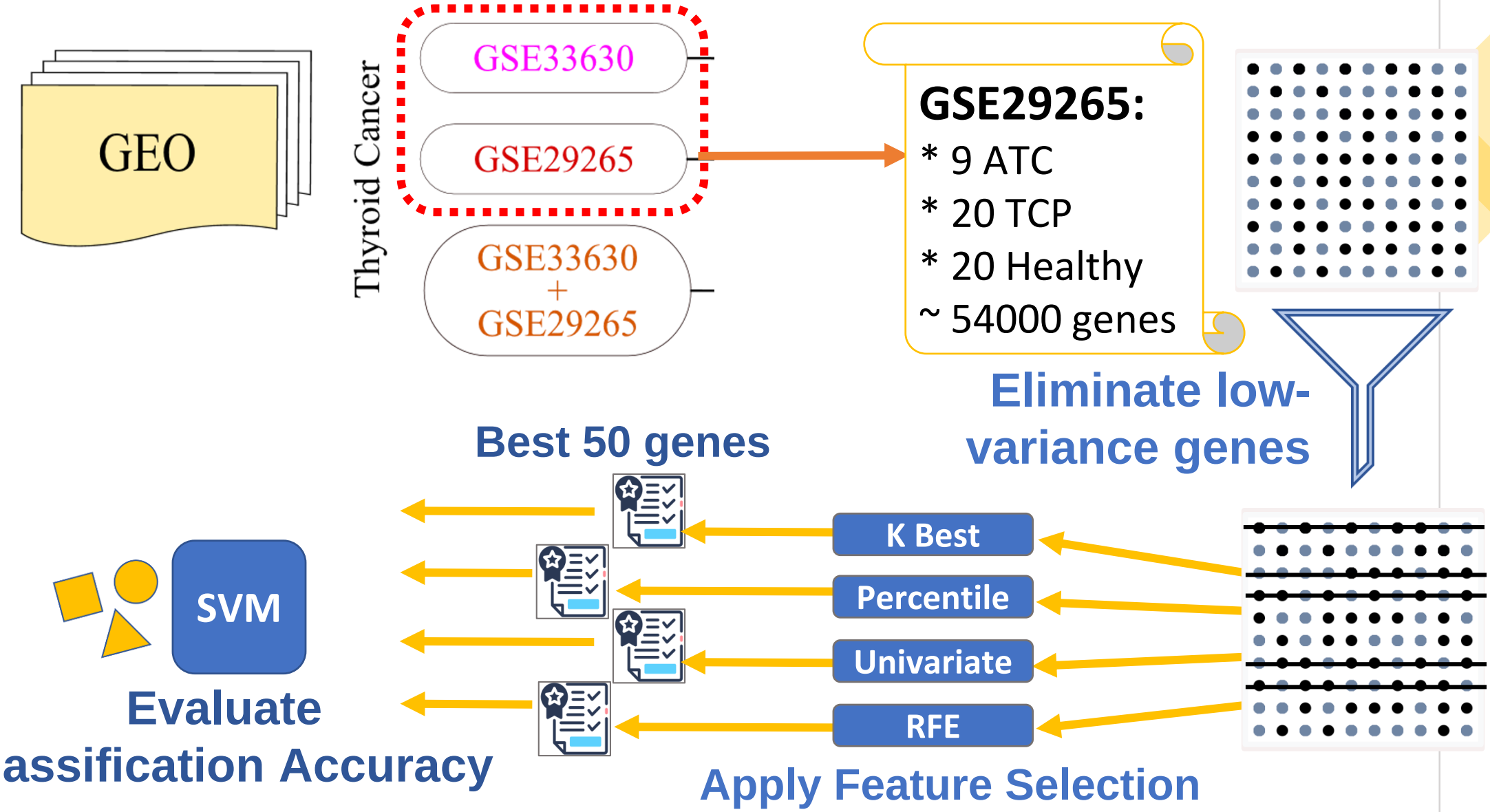
Workflow – Phase 1 – Individual Analysis – Feature Selection



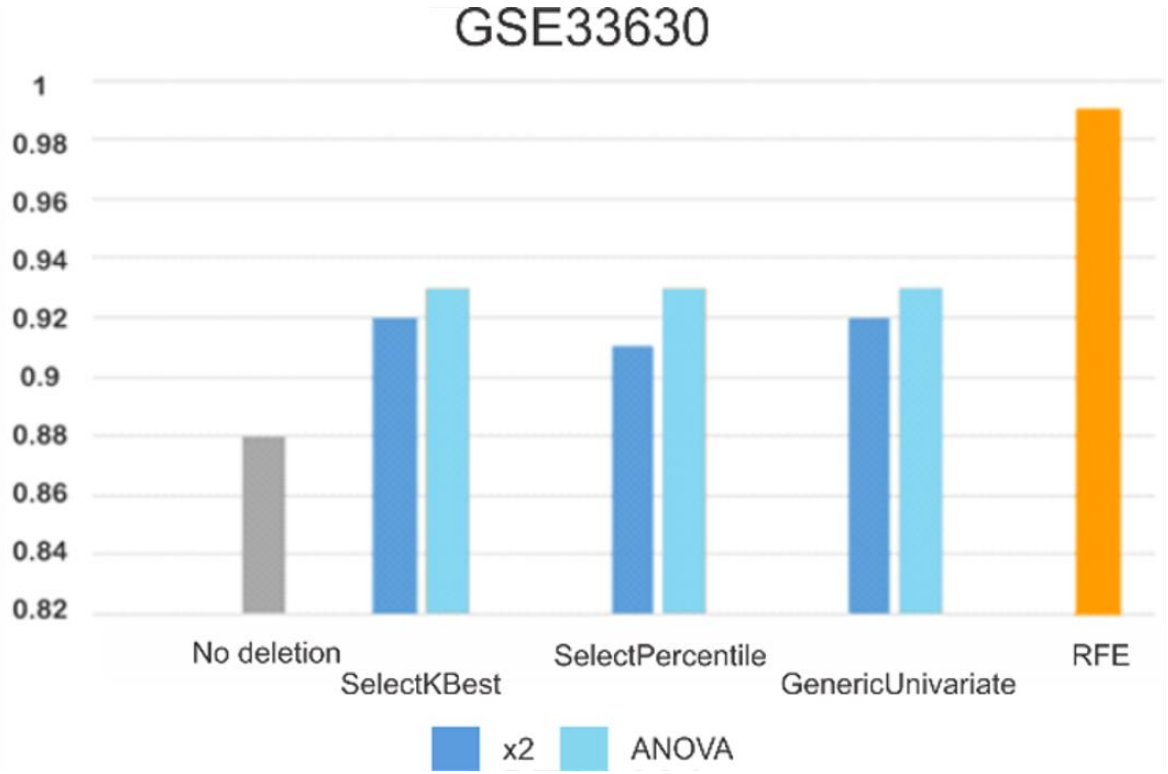
Workflow – Phase 1 – Individual Analysis – Feature Selection

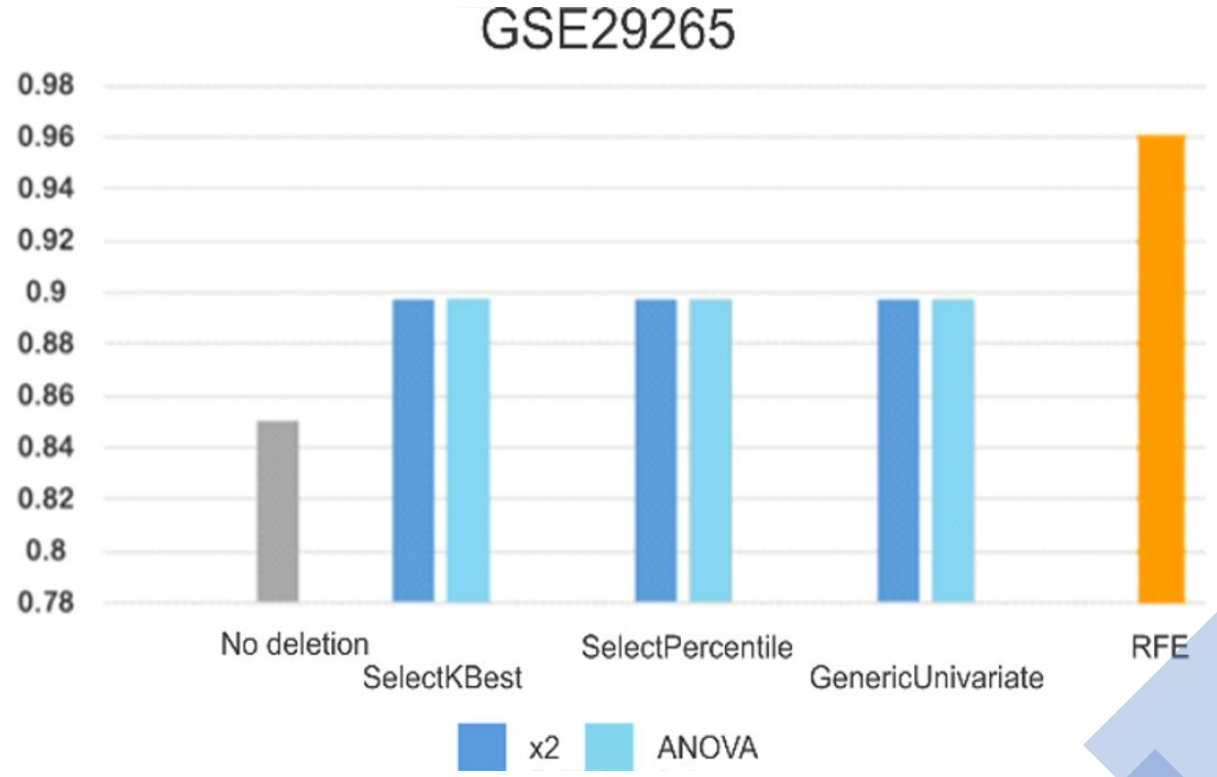


Workflow – Phase 1 – Individual Analysis – Feature Selection

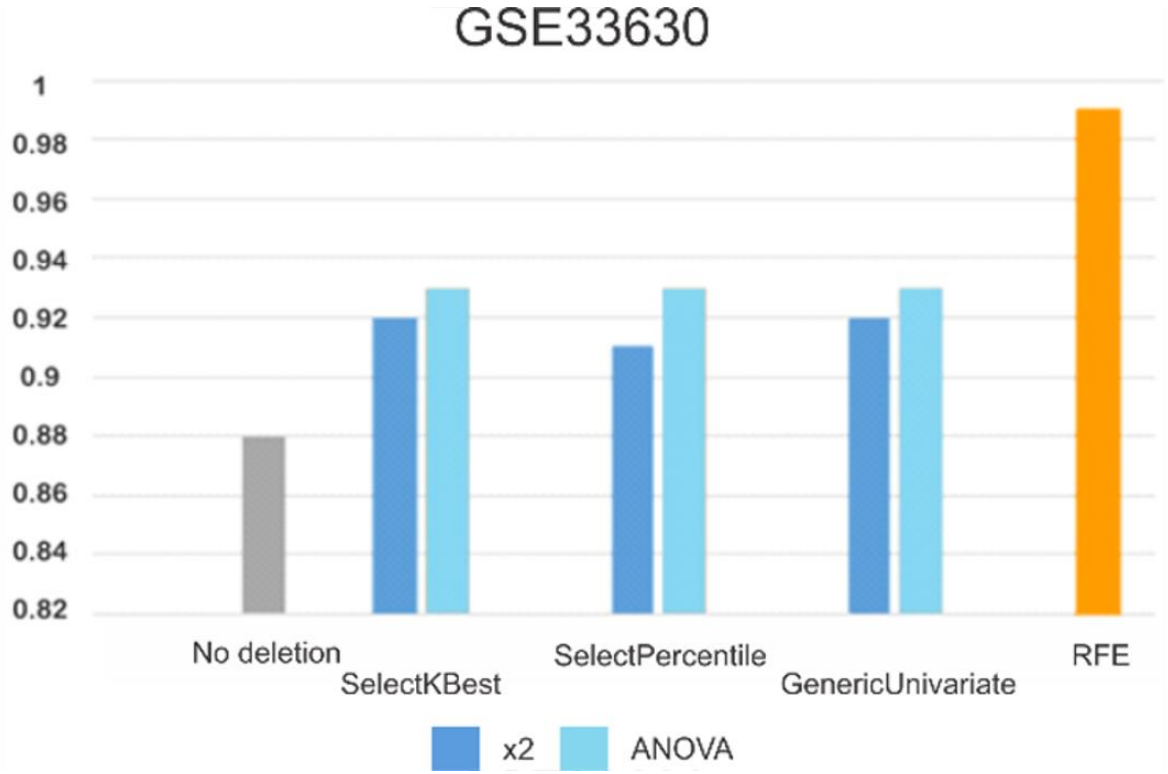


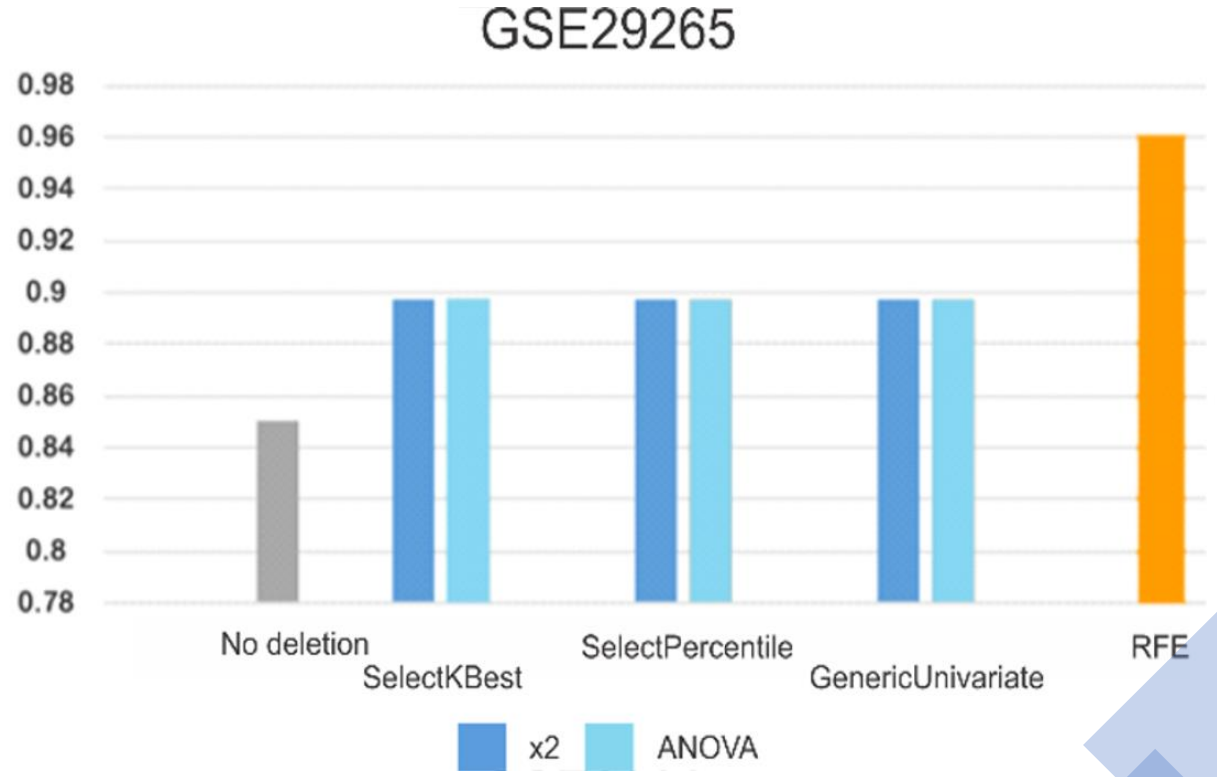
Results – Phase 1



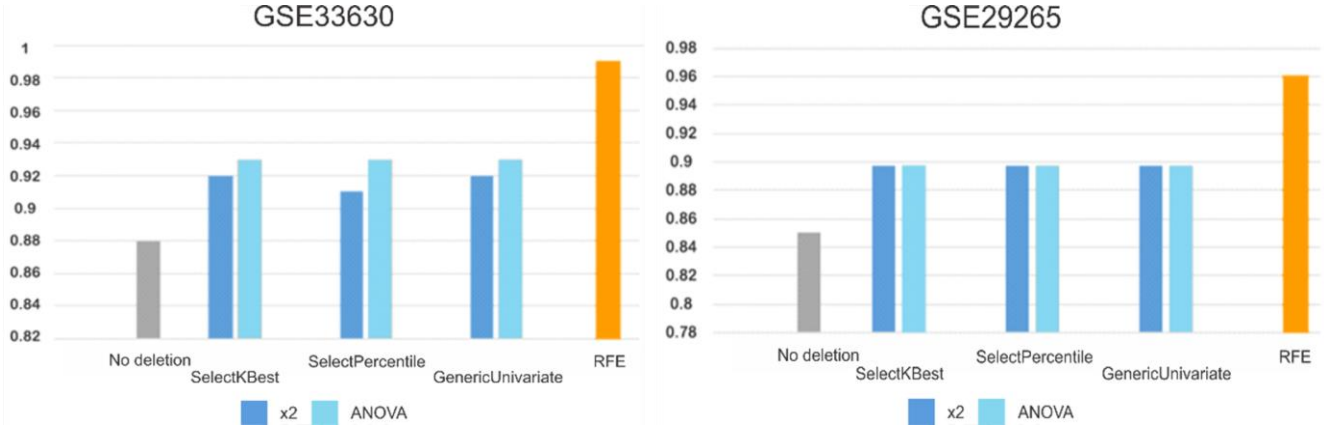


Results – Phase 1



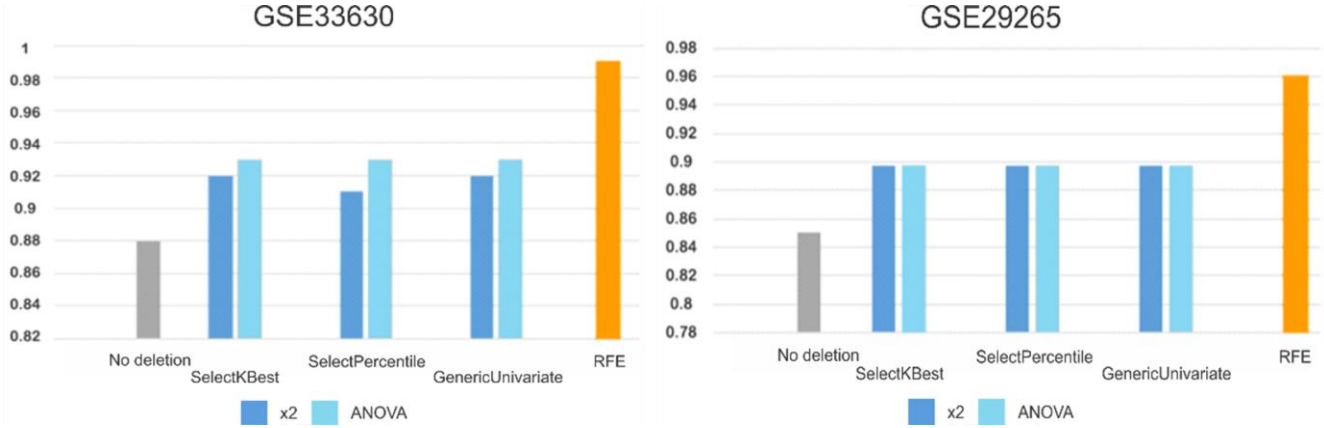


Results – Phase 1



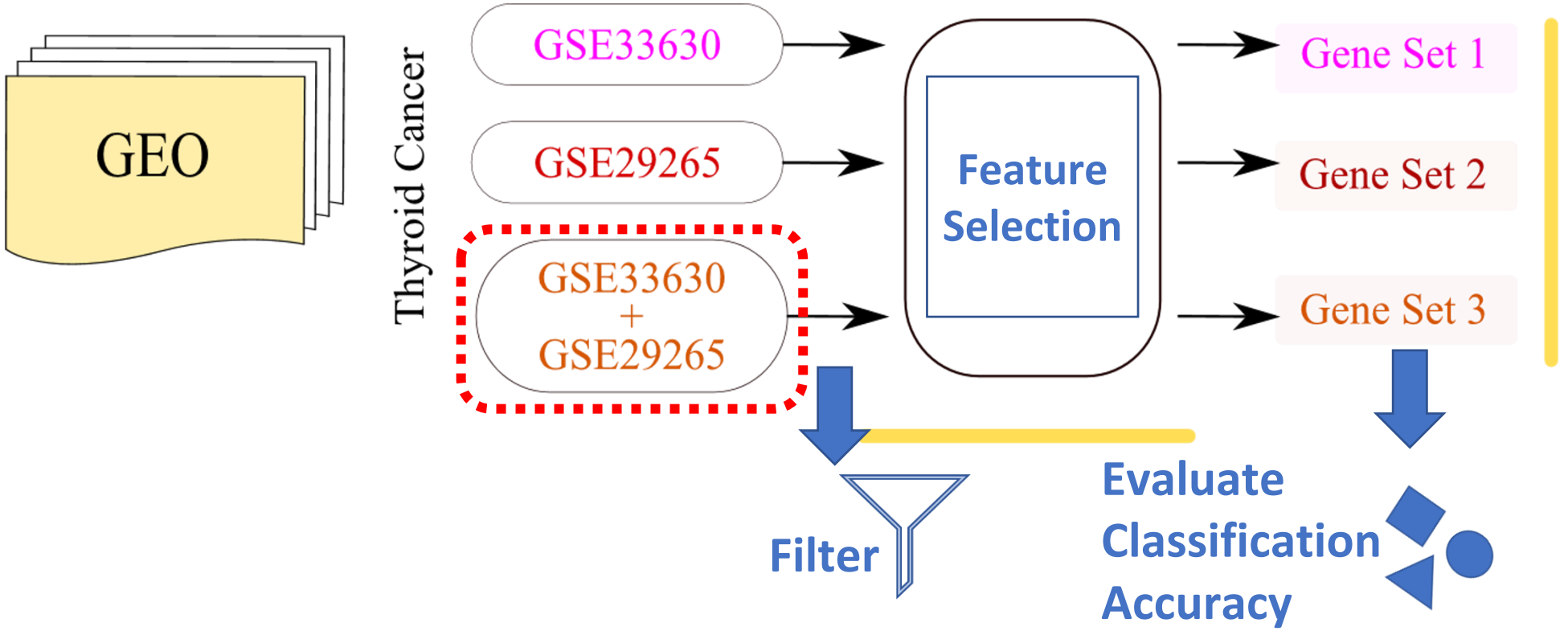
- **GSE33630**: GABRB2, Hs.544373, CDH2, CCL21, TFF3, BCHE, MMRN1, COMP, DPT, NFAT5, TFPI2, CHI3L1, GRB10, NR4A3, SCN3A, TFPI, POSTN, APLNR, MFAP5, HRC13275, CNTNAP2, BMS1P20, IGLV1-44, COL10A1, SFTPA2, LYVE1, FLRT3, CLIC3, TRIM36, SLC27A6, COLEC12, C2orf40, SLC1A2, ENTPD1, HINT3, GJC1, CPNE4, WISP1, F2RL2, COL3A1, SNORD3D, HECW2, PLEKHA2, ARHGAP36, LPP, and COL11A1.
- **GSE29265**: WASIR2, LTF, TACSTD2, ATF3, ADM, NELL2, DPP4, BUB1B, IGF2BP3, DUSP4, MMP7, PROM1, EIF1AY, GAP43, PLN, DEFA1B, PRSS2, RASGRP1, TENM1, EGR3, RYR3, GRP, HMGA2, PLAUR, TMEM158, LRRC15, CXCL5, YME1L1, GREM1, RERGL, MECOM, ANLN, HHATL, INMT, FOXQ1, ZNF595, AGR3, TDRD9, HINT3, RIMS2, TCERG1L, Hs.720692, LOC101930164, LIPH, Hs.443967, SCARNA2, LCN10, Hs.553068, and COL11A1.

Results – Phase 1

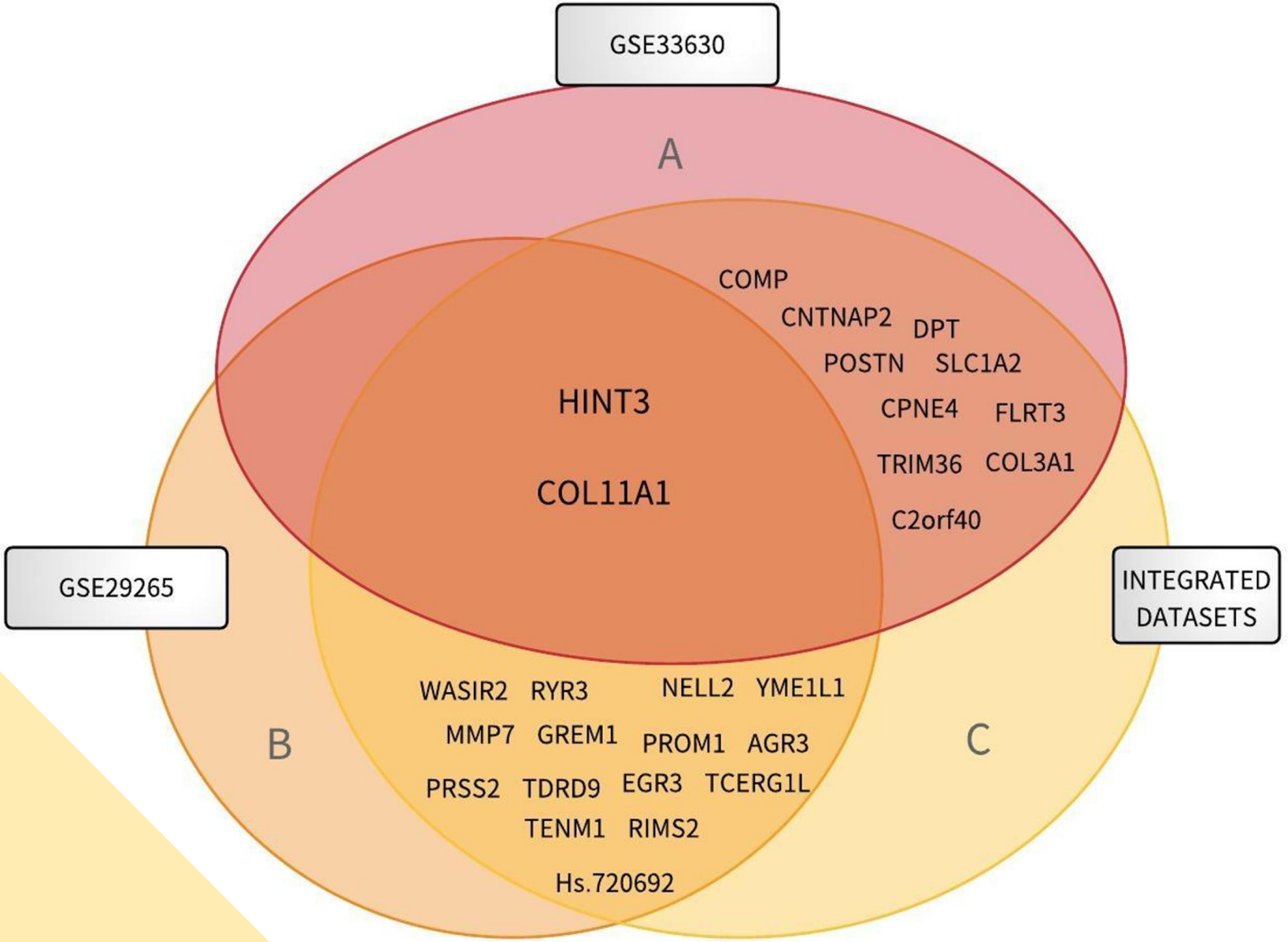


- **GSE33630**: GABRB2, Hs.544373, CDH2, CCL21, TFF3, BCHE, MMRN1, COMP, DPT, NFAT5, TFPI2, CHI3L1, GRB10, NR4A3, SCN3A, TFPI, POSTN, APLNR, MFAP5, HRC13275, CNTNAP2, BMS1P20, IGLV1-44, COL10A1, SFTPA2, LYVE1, FLRT3, CLIC3, TRIM36, SLC27A6, COLEC12, C2orf40, SLC1A2, ENTPD1, HINT3, GJC1, CPNE4, WISP1, F2RL2, COL3A1, SNORD3D, HECW2, PLEKHA2, ARHGAP36, LPP, and COL11A1.
- **GSE29265**: WASIR2, LTF, TACSTD2, ATF3, ADM, NELL2, DPP4, BUB1B, IGF2BP3, DUSP4, MMP7, PROM1, EIF1AY, GAP43, PLN, DEFA1B, PRSS2, RASGRP1, TENM1, EGR3, RYR3, GRP, HMGA2, PLAUR, TMEM158, LRRC15, CXCL5, YME1L1, GREM1, RERGL, MECOM, ANLN, HHATL, INMT, FOXQ1, ZNF595, AGR3, TDRD9, HINT3, RIMS2, TCERG1L, Hs.720692, LOC101930164, LIPH, Hs.443967, SCARNA2, LCN10, Hs.553068, and COL11A1.

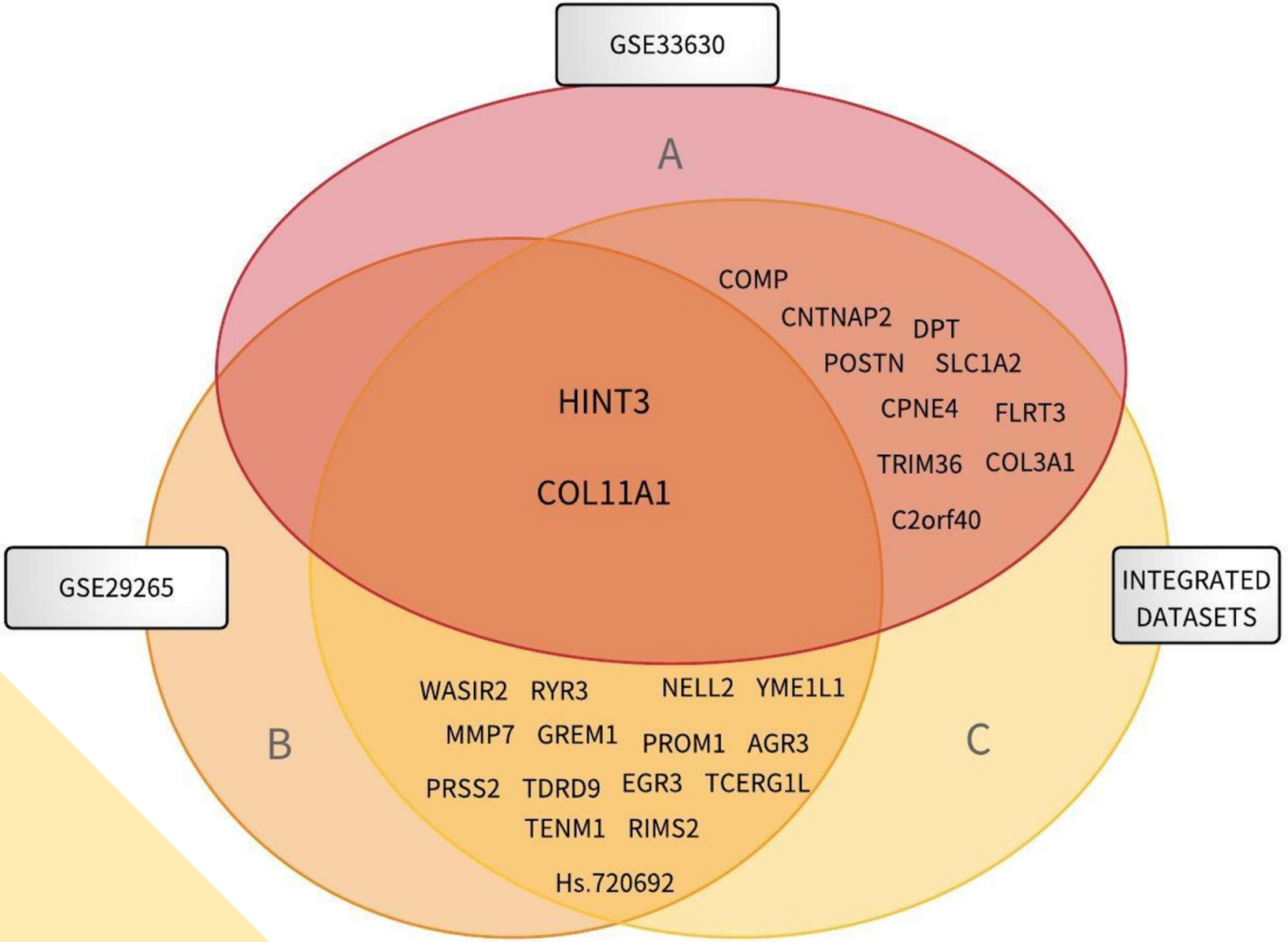
Workflow – Phase 2 – Integrated Analysis



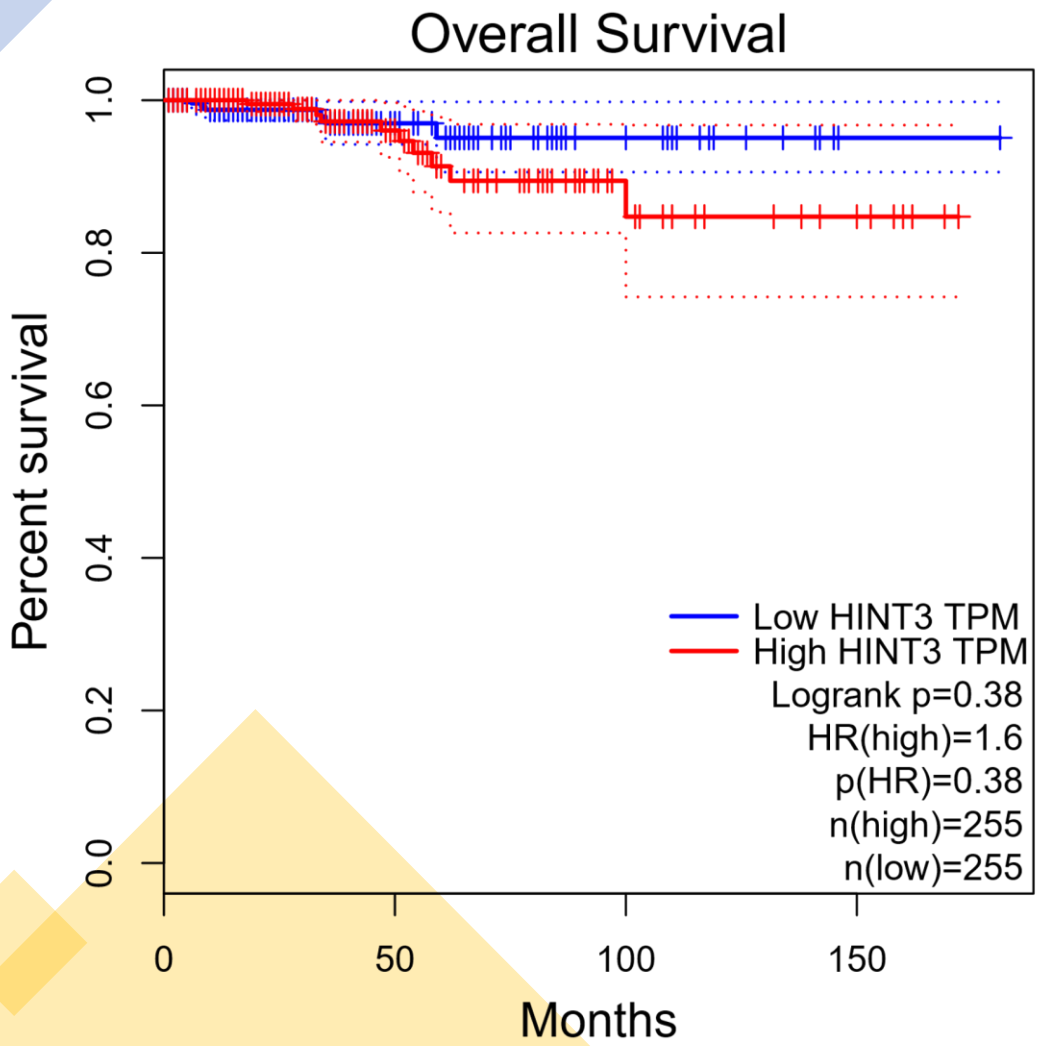
Results – Phase 2 – Integrated Analysis

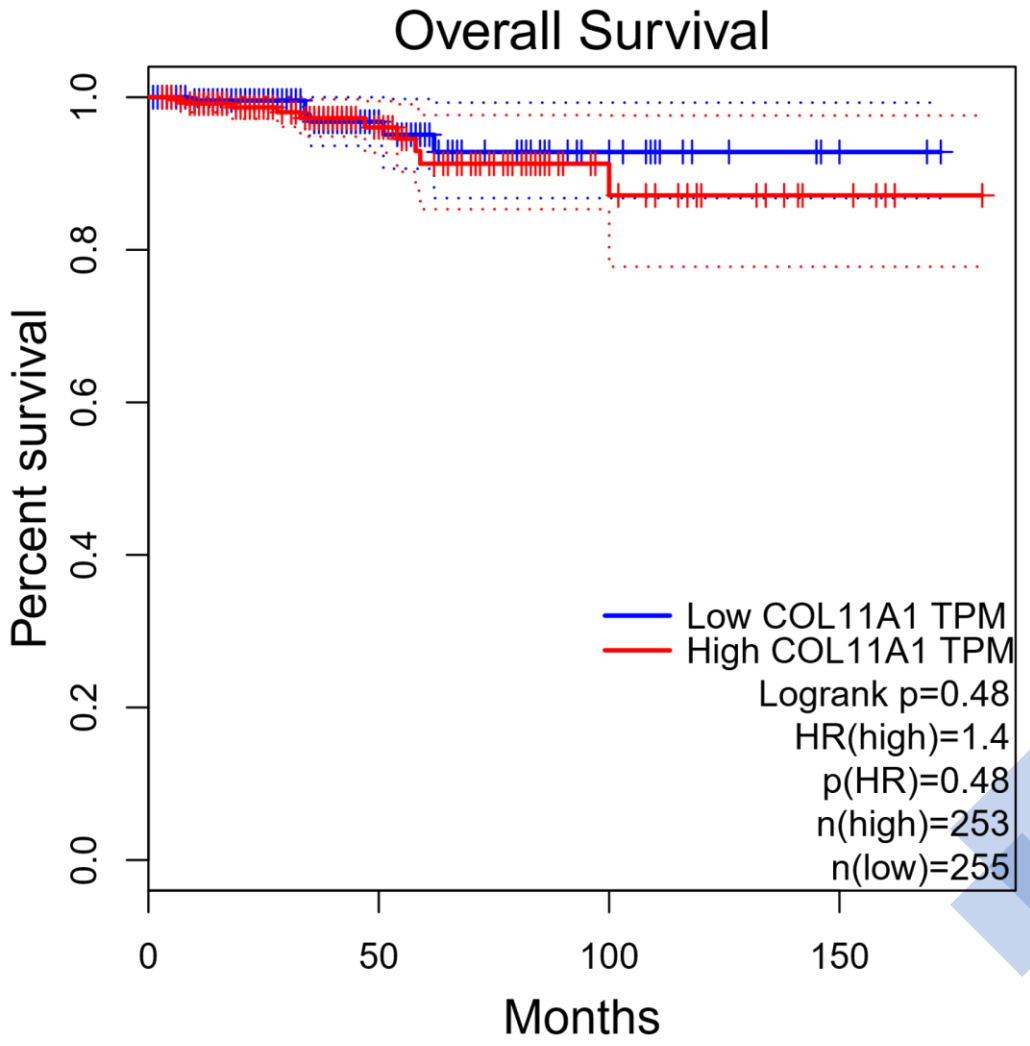


Results – Phase 2 – Integrated Analysis



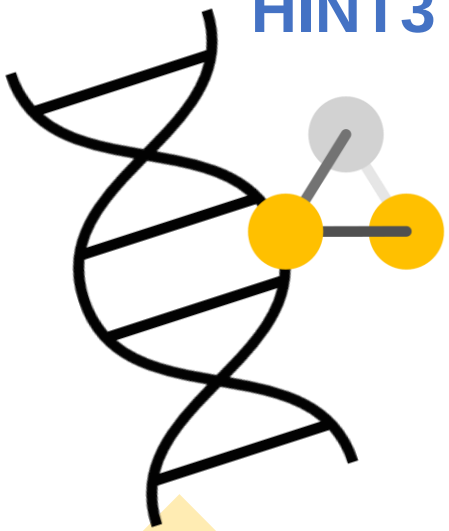
Results – Phase 2 – Integrated Analysis





Results – Statistical and Biological HINT3 Study

Up-regulated HINT3



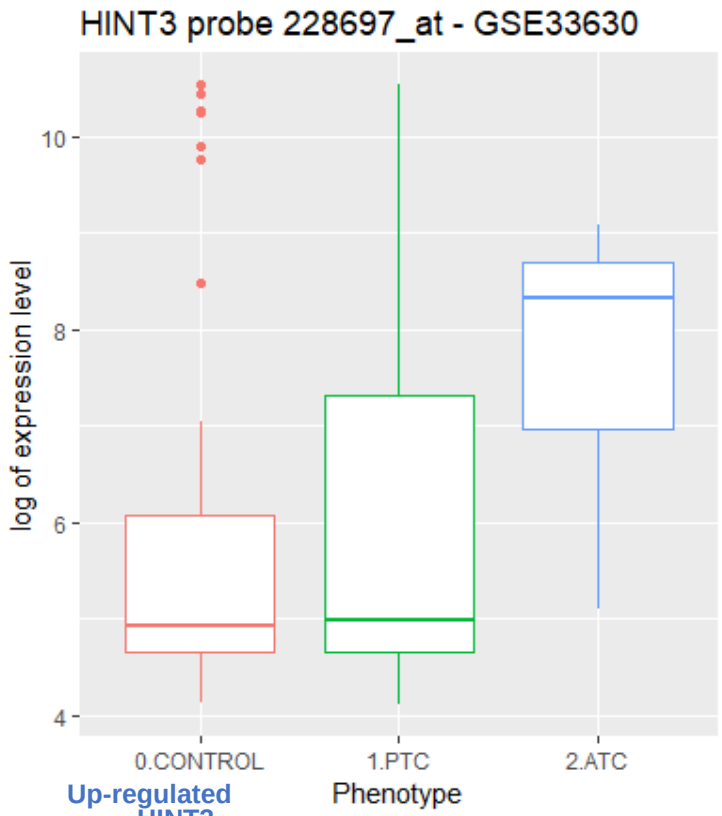
Reported in hepatocellular carcinoma and neurodegenerative disorders

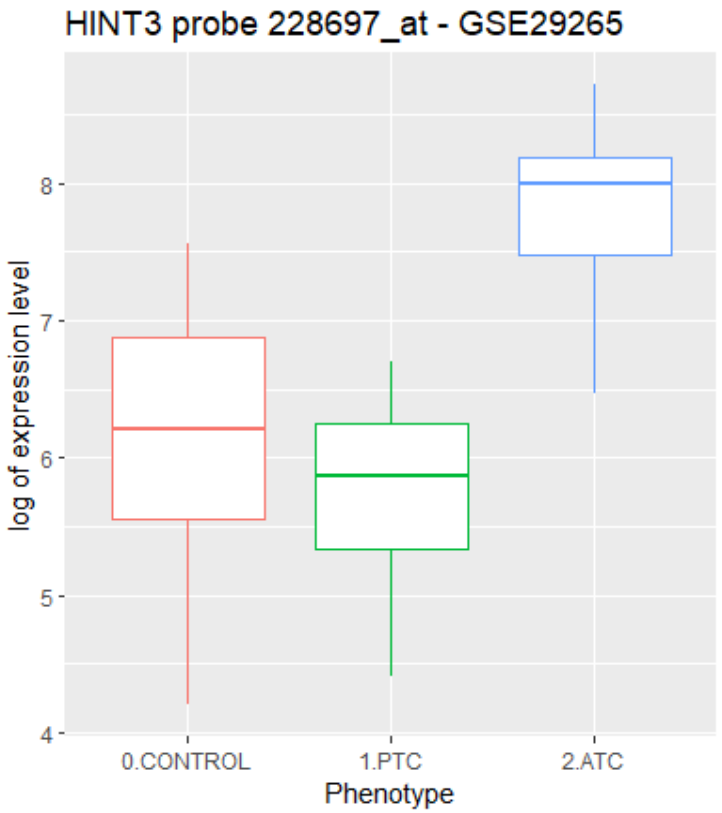
All-trans retinoic acid, which prevents cell death but induces cell migration and invasion

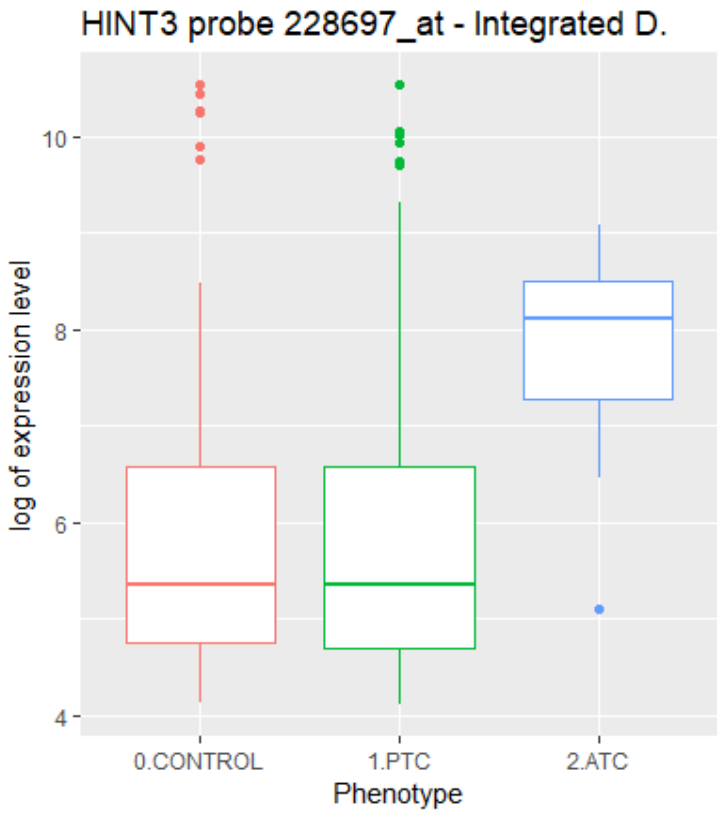
Related to induction of apoptosis of neurons in neurodegenrative disorders

We found it **upregulated** in human ATC tissue samples compared to healthy

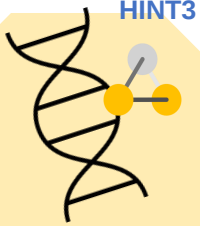
Results – Statistical and Biological HINT3 Study







Up-regulated HINT3



We found it upregulated in human ATC tissue samples compared to healthy

Results – Statistical and Biological HINT3 Study

	ANOVA	Tuckey test		
	p value	Control-PTC	Control-ATC	ATC-PTC
GSE33630	0.00778	0.79	0.005	0.017
GSE29265	6.52e-07	0.301	0.000024	4.0e-7

We found it upregulated in human ATC tissue samples compared to healthy

Conclusions and Future Work

Statistics and Machine Learning to analyze microarray data.

Real-world data sets of thyroid cancer.

Different matrix reduction approaches were used to reduce the matrix from ~50000 to 50 genes.

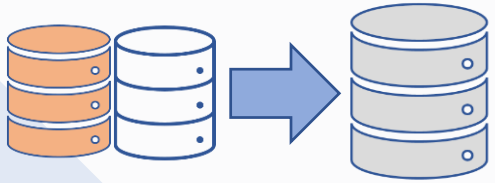
- Select K Best
- Generic
- Univariate
- Recursive Feature Elimination

Conclusions and Future Work

Biological analysis of selected genes

The importance of the used dataset.

- GSE33630
- GSE29265
- GSE33630 + GSE29265



Integrating? Or not?

Conclusions and Future Work

Classification results

- SVM
- KNN

Most representative genes:

- COL11A1
- HINT3

Further studies and discussion remain, but this work constitutes a promising achievement

INSTITUTE OF COMPUTER SCIENCE AND ENGINEERING

DEPARTAMENTO DE CIENCIAS E INGENIERÍA DE LA COMPUTACIÓN

CONICET (Grant number 12-2017-010082)
(UNS) (Grant number 24/N052)