Transcriptomic Data Analysis of Racial Differences in Colorectal Cancer

R. Giuliani, Q. F. Lotito, A. Massacci, A. Sartori

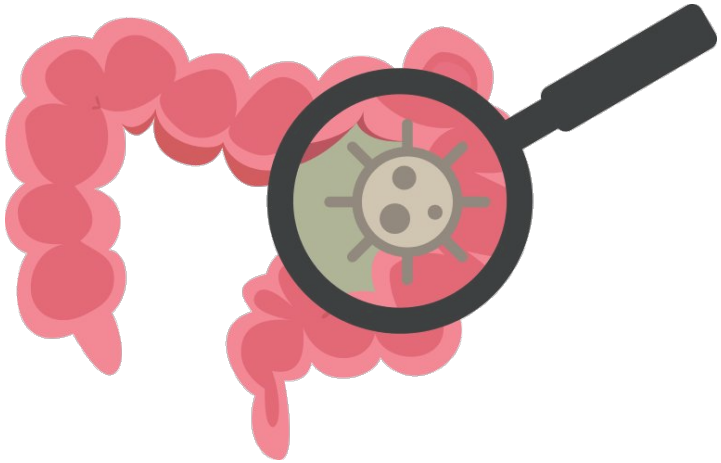


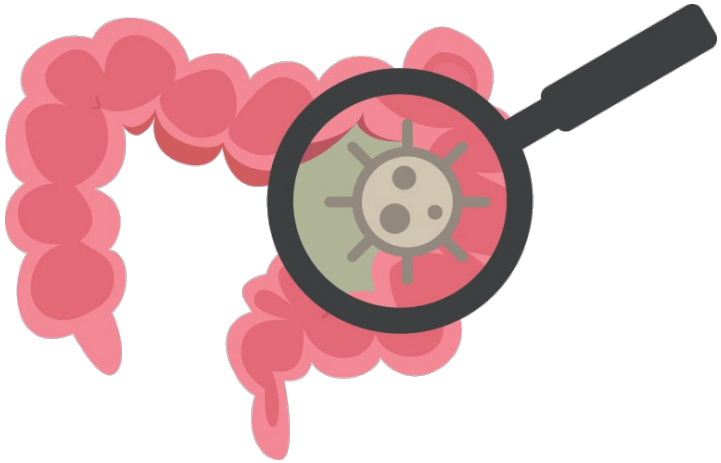
Table of Contents

01

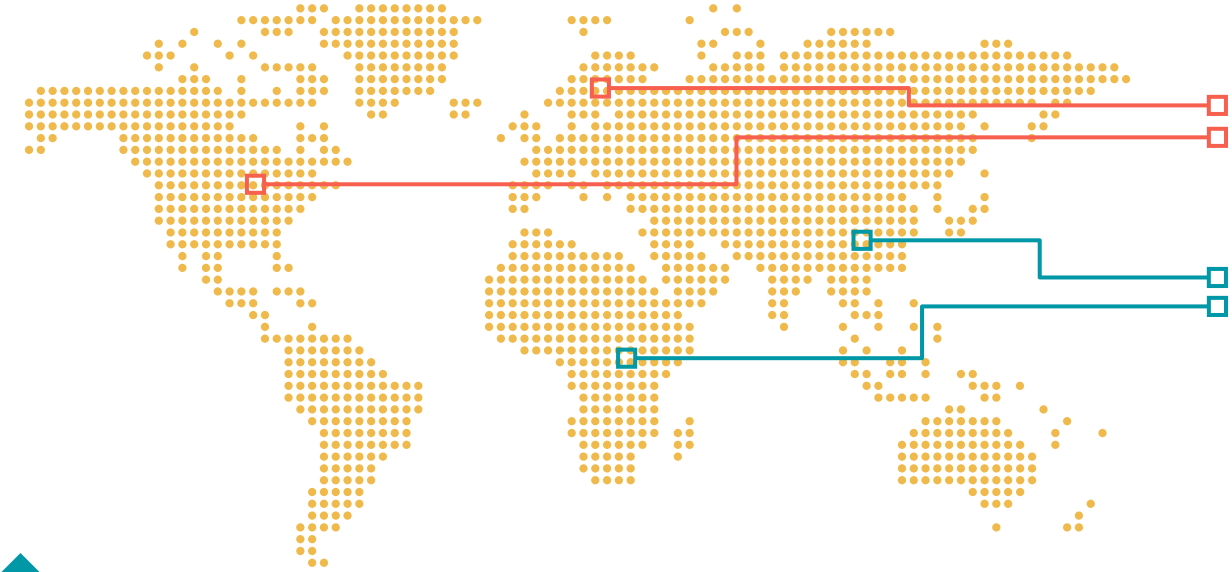
Introduction

Introduction

- Development of cancer from the colon or rectum
- Second most frequently diagnosed cancer in women (614,000 cases, 9.2% of the total female population)
- Third most frequently diagnosed cancer in men (746,000 cases, 10% of the total male population)
- Big incidence of geography!

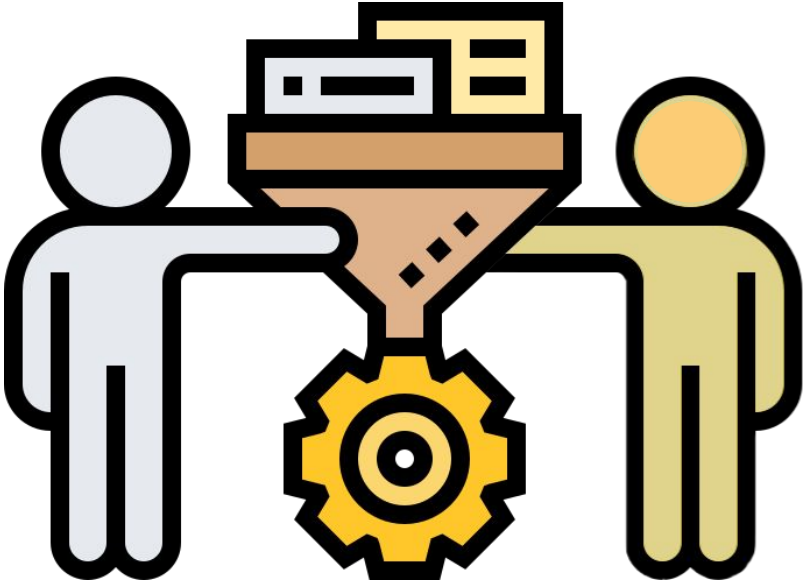


GEOGRAPHICAL INCIDENCE OF CRC



Higher incidence

Lower incidence



02

Data and preprocessing

Data

We used 3 datasets merged together:

- TCGA (RNA-seq), divided in 246 White and 25 Asian patients
- GEO (RNA-seq), (GSE154548) with 40 Korean patients
- GEO (Affymetrix), (GSE101896) with 90 Japanese patients

Features considered:

- Ethnicity
- Transcribed genes

Data

Different normalization units among RNA-Seq datasets

- TCGA: non-normalized integer counts
- GEO (Korean): log2(CPM+1), CPM stands for counts per million

Expression values in the GEO dataset did not appear to be CPM-normalized, as instead stated by the owner of the data

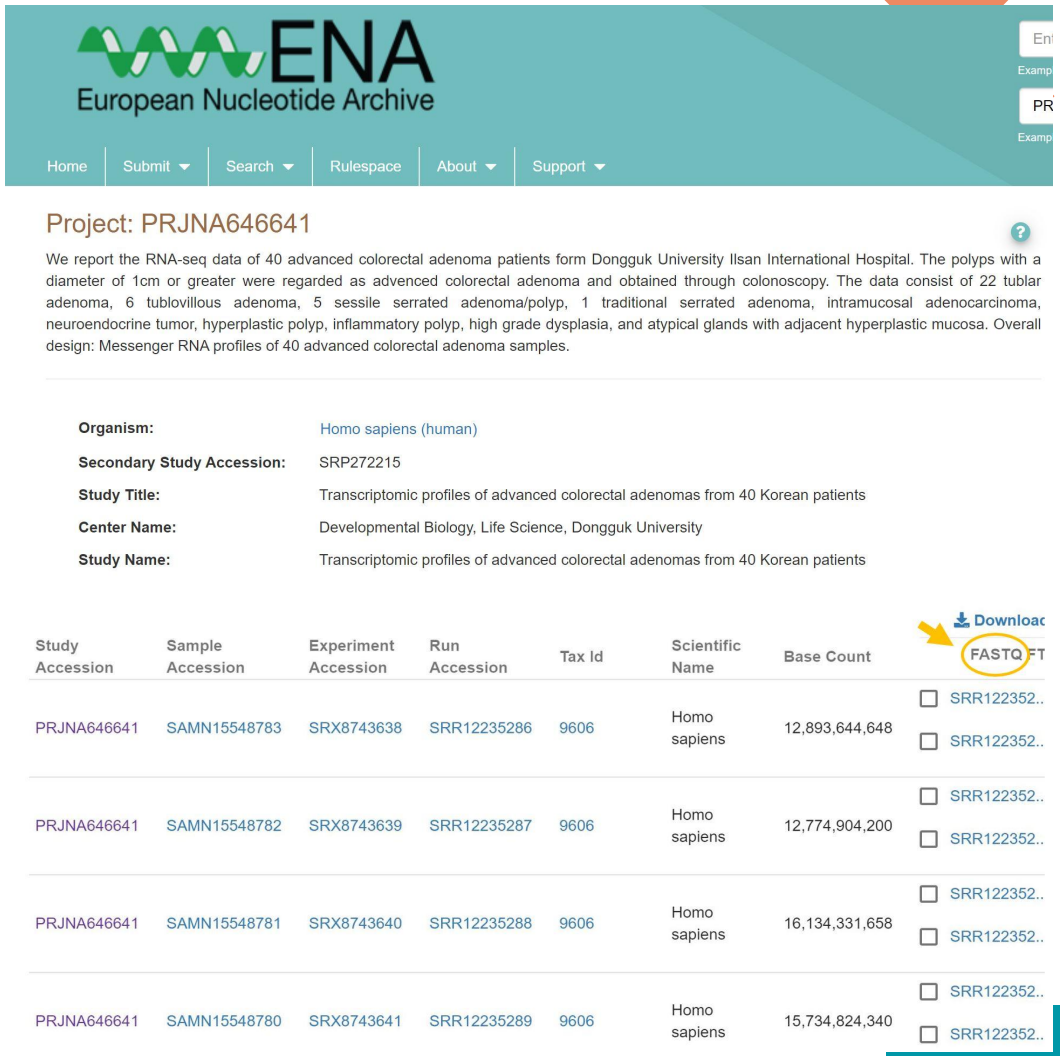
We decided to re-generate ourselves the expression matrix starting from the FASTQ files

Data

- The GSE154548 dataset has an associated PRJNA646641 BioProject code that we used to retrieve the FASTQ files from the ENA (European Nucleotide Archive) Browser

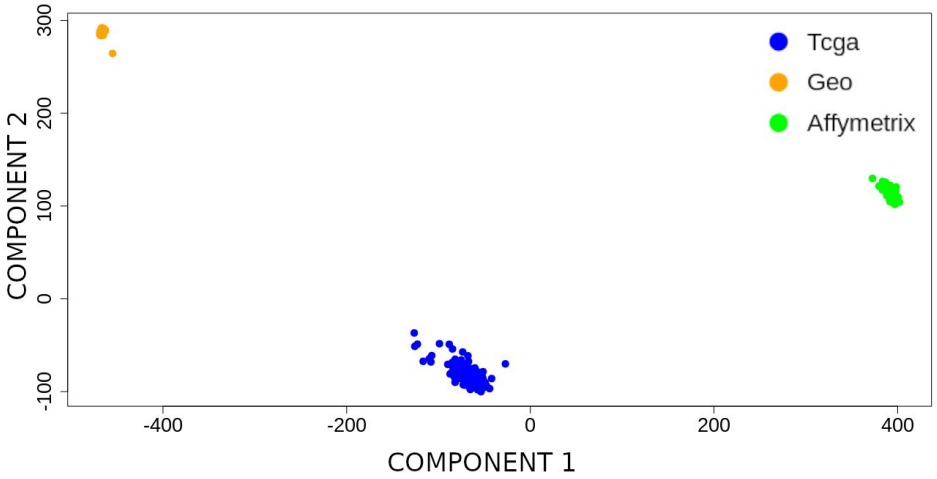
- The paired-end FASTQs associated to each patient were analysed using Kallisto, a bioinformatic tool that performs transcript quantification

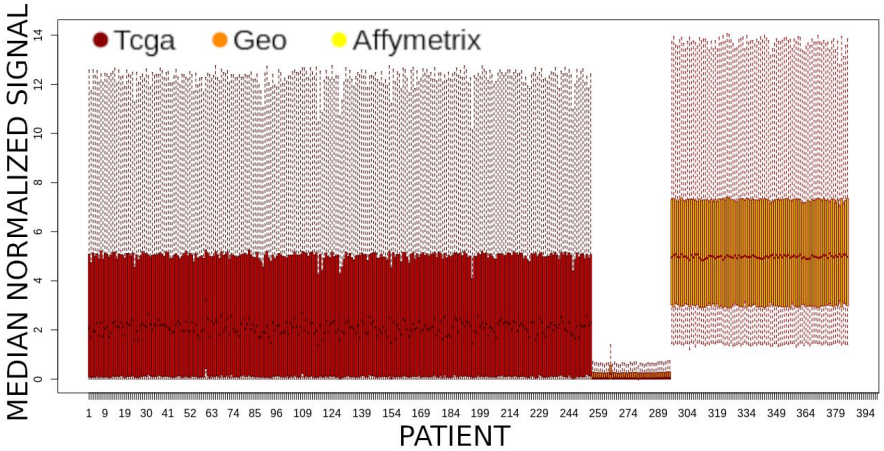
- In order to perform quantification a Homo sapiens reference transcriptome file was needed. We downloaded it from the Ensembl ftp site



Data Preprocessing

We had many problems dealing with our data!

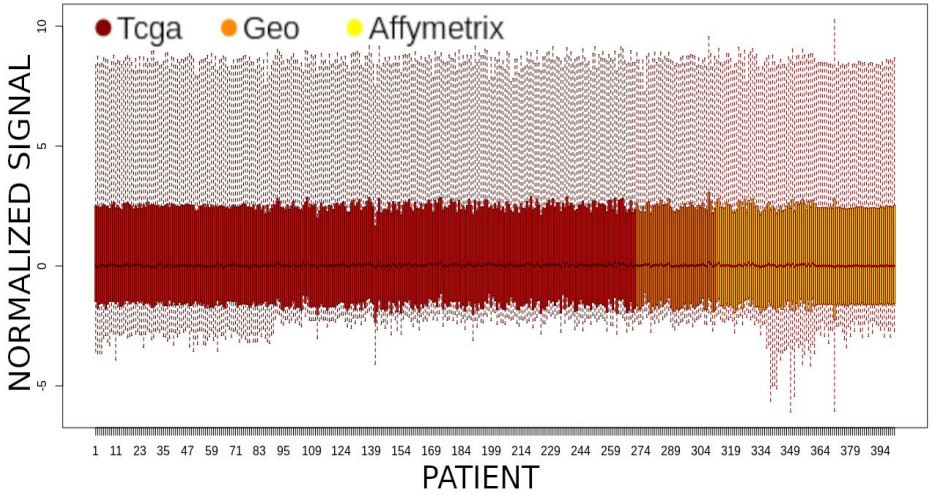


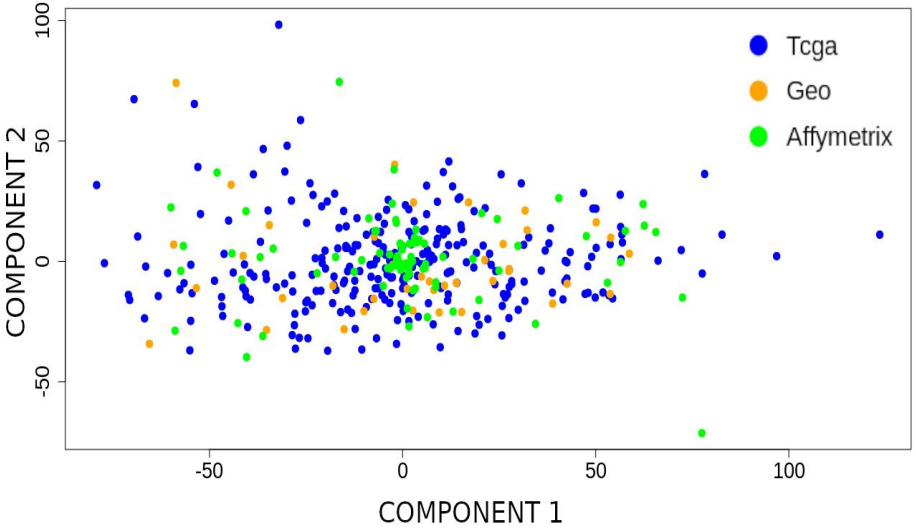


- Different sources
- Different technologies

Data Preprocessing

Thanks to filtering methods and «Combat» normalization





03

Methods and Results

Feature Selection

Mutual Information employed to reduce our features

- Initial n° of features: 19811
- Final n°: 300 features

CV-RF Classifier to validate the selection method

→ 300 genes: 87% accuracy

→ No overfitting/learning datasets

Cancer genes

OncoSearch db to retrieve cancer genes in our lists

300 genes → 12 cancer genes

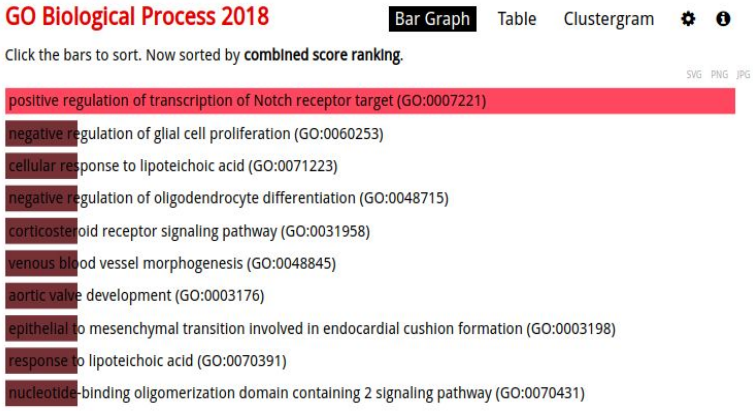
Gene	Cancer	Gene Expression	Expected Class	Cancer Change
FASN	Colorectal Carcinoma	up-regulated	Biomarker	progression
RELA	Colorectal Carcinoma	up-regulated	Biomarker	progression
ARHGDIA	Colorectal Carcinoma	up-regulated	Biomarker	progression
NOTCH1	Colorectal Carcinoma	up-regulated	Biomarker	progression
FASN	Colon Carcinoma	up-regulated	Biomarker	progression
NCOA6	Colon Carcinoma	up-regulated	Biomarker	progression
HSPG2	Colorectal Carcinoma	up-regulated	Biomarker	progression

Feature Selection

Enrichment pathway analysis to understand biological implications

Most of these genes belong to Notch signalling pathway or closely related

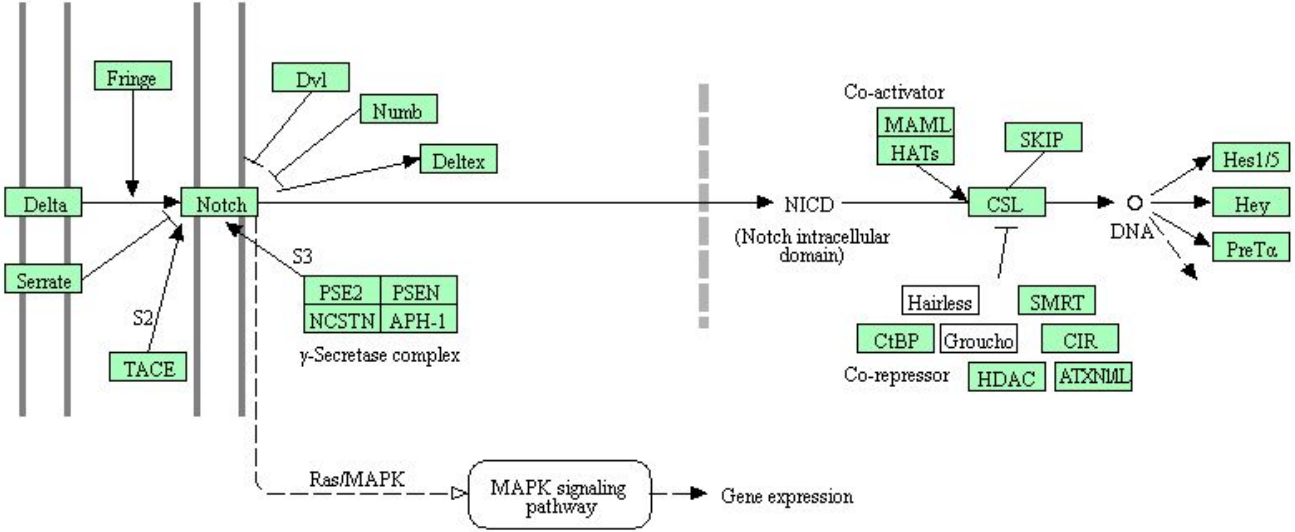
NOTCH1, NOTCH3, RELA, CREBBP, MAP2K2, CASP2, HSPG2, ARHGDIA



clr all		Gene ID	Mapped IDs	Gene Name Gene Symbol Ortholog	PANTHER Family/Subfamily	PANTHER Protein Class	Species
	1.	HUMAN|HGNC=7881|UniProtKB=P46531	NOTCH1	Neurogenic locus notch homolog protein 1 NOTCH1 ortholog	NEUROGENIC LOCUS NOTCH HOMOLOG PROTEIN 1 (PTHR45836:SF12)	-	Homo sapiens
	2.	HUMAN|HGNC=678|UniProtKB=P52565	ARHGDIA	Rho GDP-dissociation inhibitor 1 ARHGDIA ortholog	RHO GDP-DISSOCIATION INHIBITOR 1 (PTHR10980:SF9)	G-protein modulator	Homo sapiens
	3.	HUMAN|HGNC=1503|UniProtKB=P42575	CASP2	Caspase-2 CASP2 ortholog	CASPASE-2 (PTHR10454:SF151)	protease	Homo sapiens
	4.	HUMAN|HGNC=5273|UniProtKB=P98160	HSPG2	Basement membrane-specific heparan sulfate proteoglycan core protein HSPG2 ortholog	BASEMENT MEMBRANE-SPECIFIC HEPARAN SULFATE PROTEOGLYCAN CORE PROTEIN (PTHR10574:SF273)	extracellular matrix protein	Homo sapiens
	5.	HUMAN|HGNC=9955|UniProtKB=Q04206	RELA	Transcription factor p65 RELA ortholog	TRANSCRIPTION FACTOR P65 (PTHR24169:SF1)	Rel homology transcription factor	Homo sapiens
	6.	HUMAN|HGNC=2348|UniProtKB=Q92793	CREBBP	CREB-binding protein CREBBP ortholog	CREB-BINDING PROTEIN (PTHR13808:SF34)	-	Homo sapiens

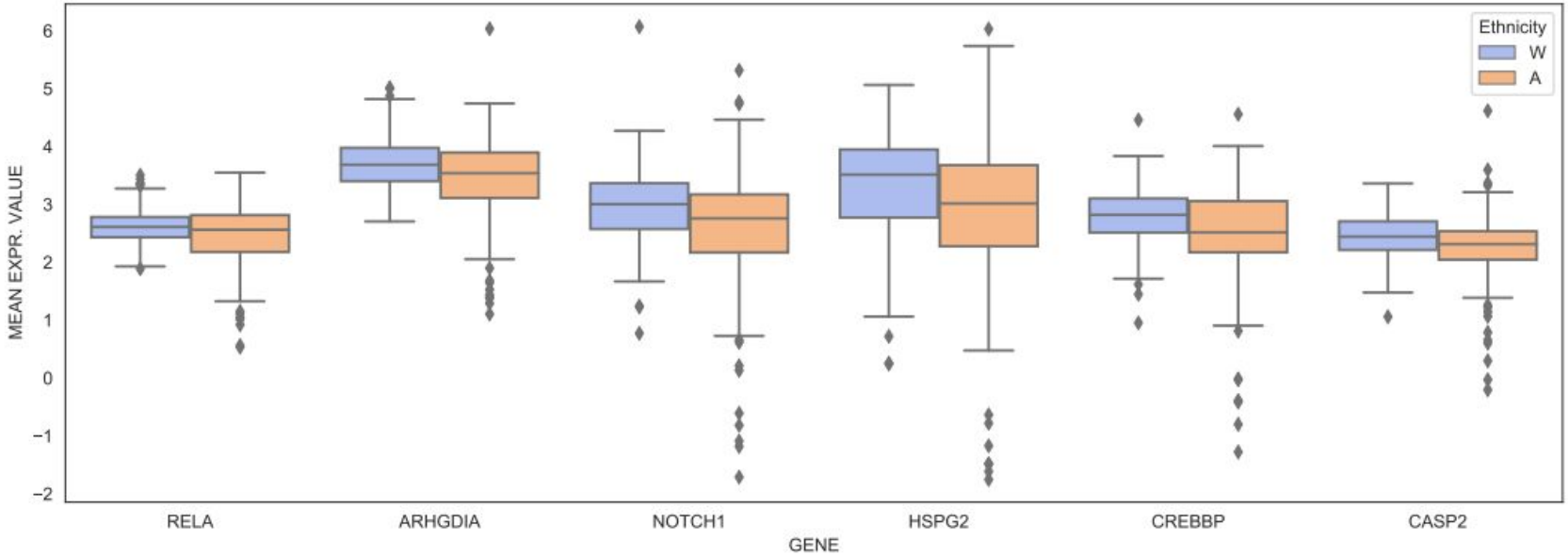
Notch signaling pathway

- Cell death regulation
- Highly conserved pathway
- Cancer association



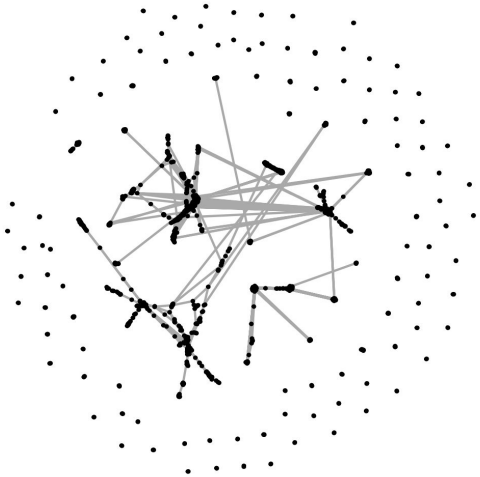
Notch signaling pathway

Upregulation in White patients

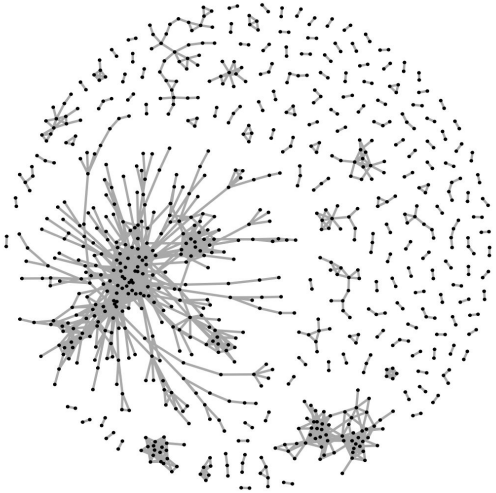


Gene co-expression network analysis

- Complete set of genes to describe each network
- Pearson’s to determine correlation (threshold of 0.75)
- Betweenness centrality ranking criterion



White GCN

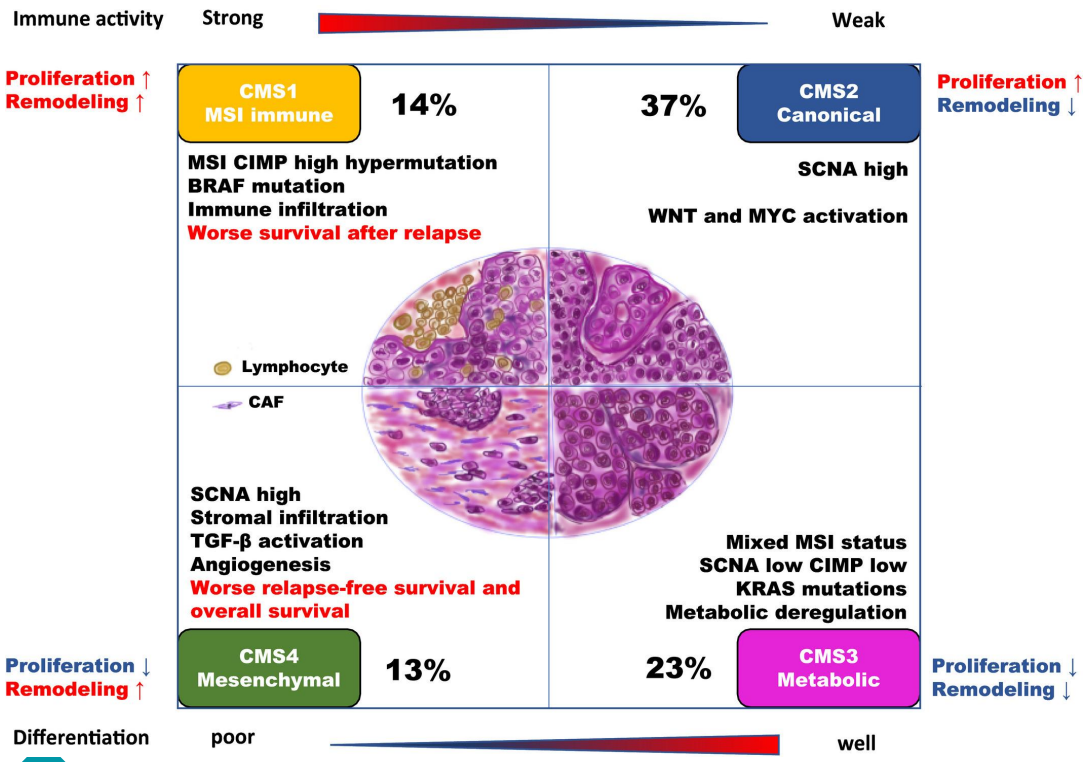


Asian GCN

Gene co-expression network analysis

- **White patients' network results:**
  - Found again NOTCH3 among the top 5 genes ranked according to their betweenness centrality
  - Found the gene MAP2K2 in top genes, a kinase closely related with Notch pathway
- **Asian patients' network results:**
  - Found the gene CDKN3 among the highest ranked genes
  - From literature, "*CDKN3 had effects in suppressing colorectal cancer cell proliferation and migration, inducing cell cycle arrest and apoptosis in a colorectal cancer cell line, SW480 cells*".
  - From literature, CDKN3 is considered a possible target of novel treatments for colorectal cancer.

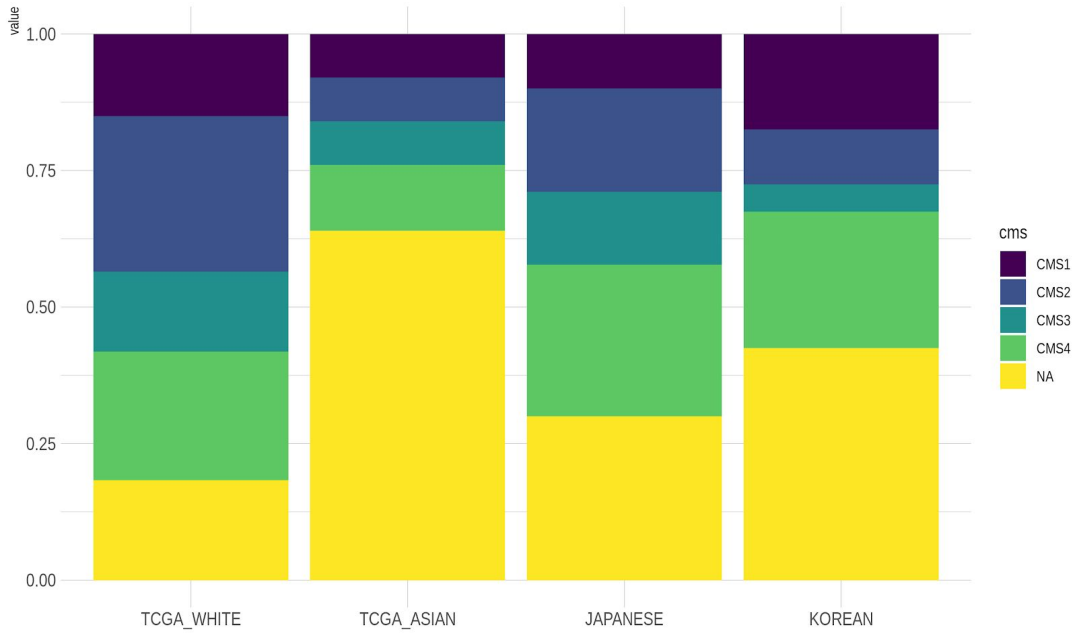
CMS analysis



Colorectal cancers can be classified into four biologically distinct consensus molecular subtypes (CMS) based on their gene expression patterns.

1. CMS1: MSI-immune
2. CMS2: epithelial and canonical
3. CMS3: epithelial and metabolic
4. CMS4: mesenchymal

CMS analysis



WHITE

CMS1: 15%
CMS2: 28%
CMS3: 15%
CMS4: 24%
NA: 18%

ASIAN TCGA

CMS1: 8%
CMS2: 8%
CMS3: 8%
CMS4: 10%
NA: 64%

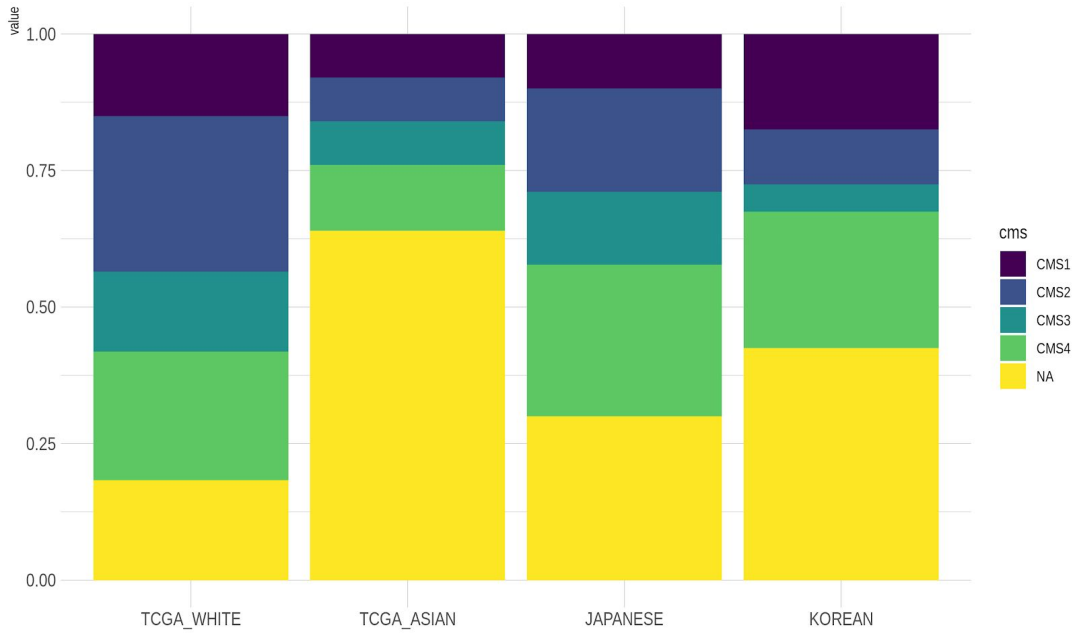
JAPANESE

CMS1: 10%
CMS2: 20%
CMS3: 12%
CMS4: 28%
NA: 30%

KOREAN

CMS1: 18%
CMS2: 10%
CMS3: 5%
CMS4: 25%
NA: 42%

CMS analysis

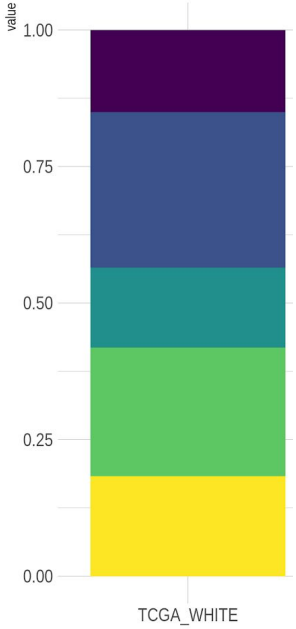


Tumors that could not be assigned to a consensus subtype had mixed gene expression signatures

The “not assigned” percentage was the highest (64%) among the 25 Asian TCGA samples, and the lowest (18%) among White

High intratumor heterogeneity and mixed gene expression signatures could be a characteristic of CRC in Asian patients

CMS analysis



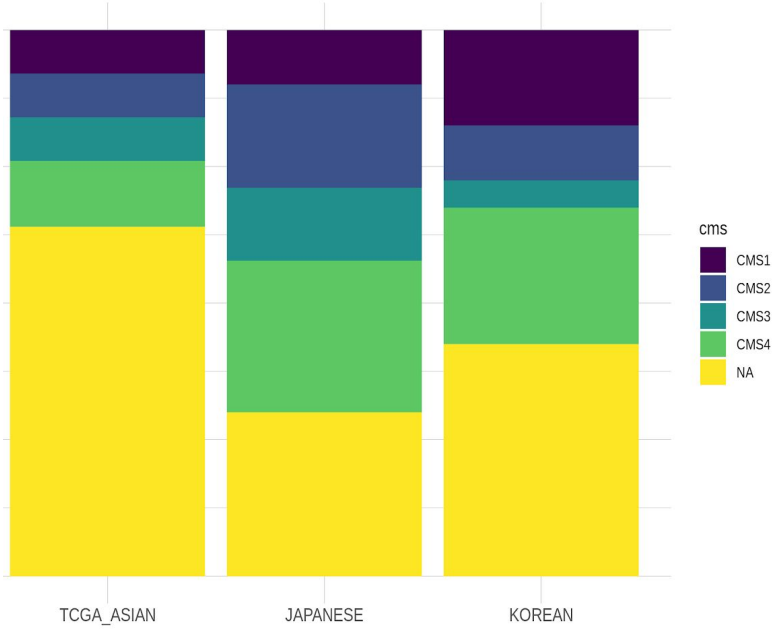
WHITE

CMS1: 15%
CMS2: 28%
CMS3: 15%
CMS4: 24%
NA: 18%

DISTRIBUTIONS BY CRCSC

CMS1: 14%
CMS2: 37%
CMS3: 13%
CMS4: 23%

White population matched reported distributions by the CRC Subtyping Consortium (CRCSC)



CMS analysis

CMS were mostly derived from a US/European population and could be not representative of other ethnic groups

"CMS subtype prevalence differs substantially by geographic region in CRC. These variations suggest that transcriptomic-defined disease biology in international populations may be more heterogeneous than previously appreciated

From Korphaisarn, Krittiya, et al. "Consensus molecular subtypes in colorectal cancer differ by geographic region." (2020): 4061-4061.

CIBERSORTx deconvolution analysis

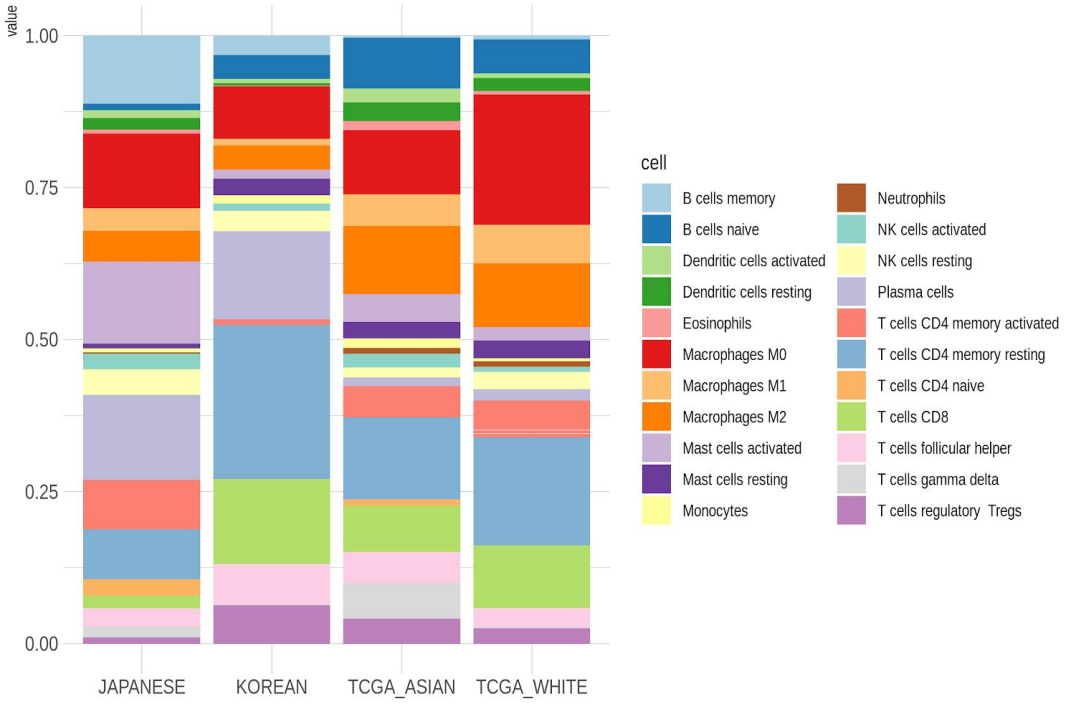
To investigate the differences in immune cell infiltration within the tumor microenvironment between White and Asian samples, we employed the CIBERSORTx algorithm

CIBERSORTx uses a signature matrix of 547 genes called LM22 for the deconvolution of 22 types of infiltrating immune cells

LM22 uses HUGO gene symbols, thus we used biomaRt for the conversion

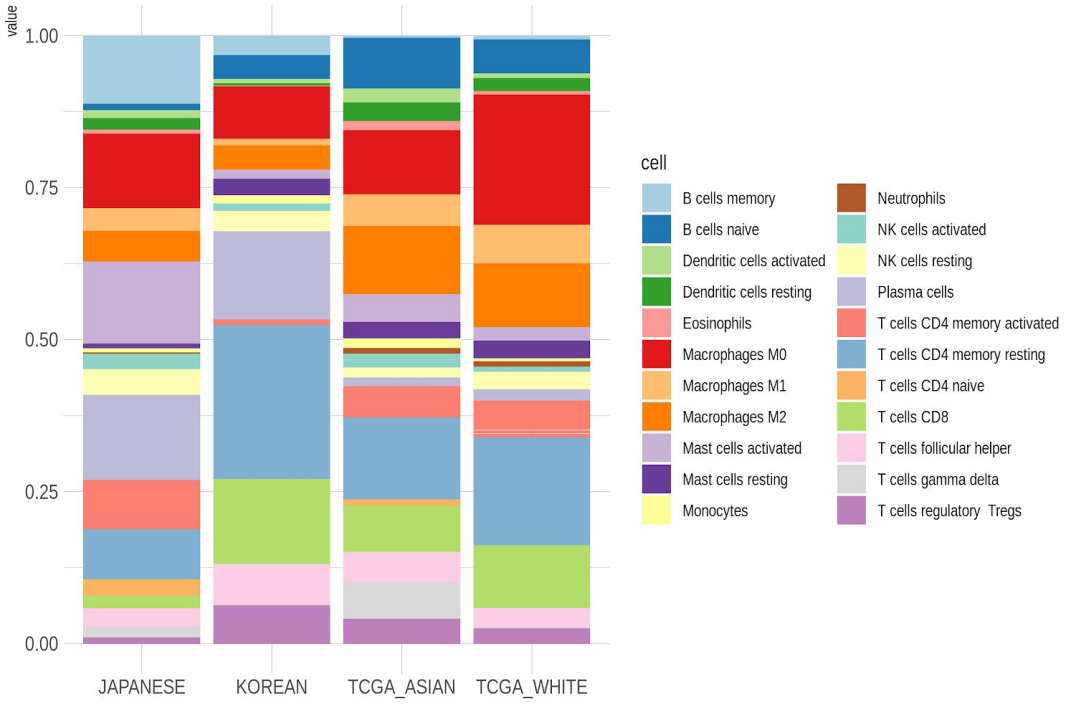
We transformed the Affymetrix CEL files belonging to the Japanese dataset into a tabular format suitable for analysis with CIBERSORTx. For this purpose we run an R script provided by the CIBERSORTx website and we downloaded a CDF (Chip Description File) compatible with the HGU133 Plus 2.0 microarray platform from BrainArray

CIBERSORTx deconvolution analysis



- Memory B cells and plasma cells were more infiltrated in Japanese and Korean GEO cohorts compared with both White and Asian TCGA
- M1 and M2 macrophages and naïve B cells were less infiltrated
- M0 macrophages were the highest in White TCGA tumors
- T cells gamma delta and naïve CD4 T cells were only expressed among Japanese and Asian TCGA

CIBERSORTx deconvolution analysis



- Notch signaling is involved in Macrophage activation and its effector functions
- Macrophages are involved in the creation of a tumor microenvironment that supports tumor growth
- M0,M1,M2 Macrophages were more infiltrated in White samples compared to Asian samples
- This is consistent with Notch1 upregulation among White patients

04

Conclusions

Preprocessing recap

Three dataset merged

From different sources, only White and Asian patients, only cancer patients

One dataset regenerated

Errors in the RNA-seq data of one dataset, regenerated from FASTQ files

COMBATR

CombatR for normalizing data, validated by PCA and boxplots

Methods recap

Feature selection + OncoSearch

Focus only on the most important genes by mutual information (for the classification task) + oncosearch for finding cancer correlations

Network analysis

Gene co-expression networks (Pearson's correlation) and betweenness centrality to rank nodes

CMS and immune deconvolution

CMScaller for subtype classification and CIBERSORTx for quantification of immune populations

Results recap

Notch signaling pathway

Most important tumor related pathway results from our analysis. (White upregulation)

Network validation

Network confirmed high level regulation from notch pathway and CDKM3 tumor suppressor gene

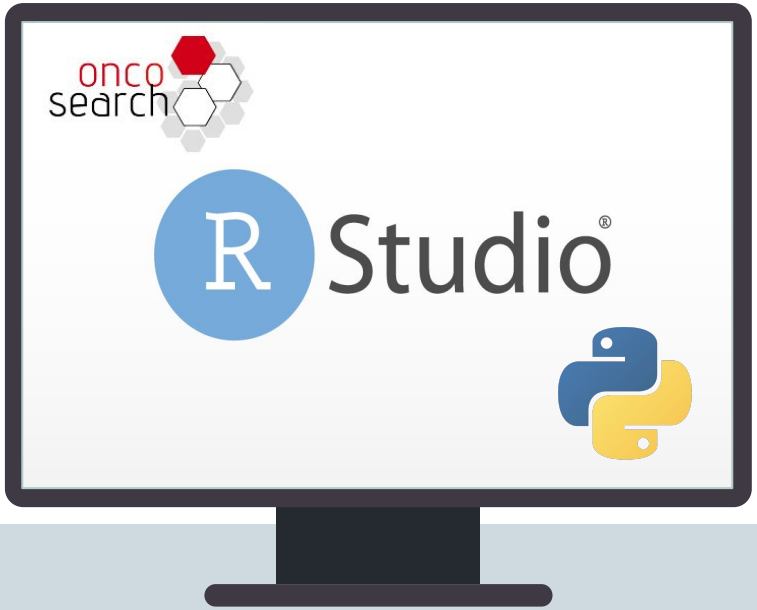
CMS and CIBERSORTx

CMS4 was more prominent as a function of race and CIBERSORTx confirmed role of Notch in modulating macrophage activation

Output of our project

- Our findings suggest **differences in Notch signaling among racially-distinct CRC patients** that may contribute to the **more aggressive clinical behavior of White patients**
- Studies related to different types of cancer (e.g., breast cancer) have supposed a **relation between ethnicity, mortality and the notch signaling pathway**
- This can motivate further study on the topic

Lessons Learned



Python / Scikit-learn

Feature selection
Machine Learning

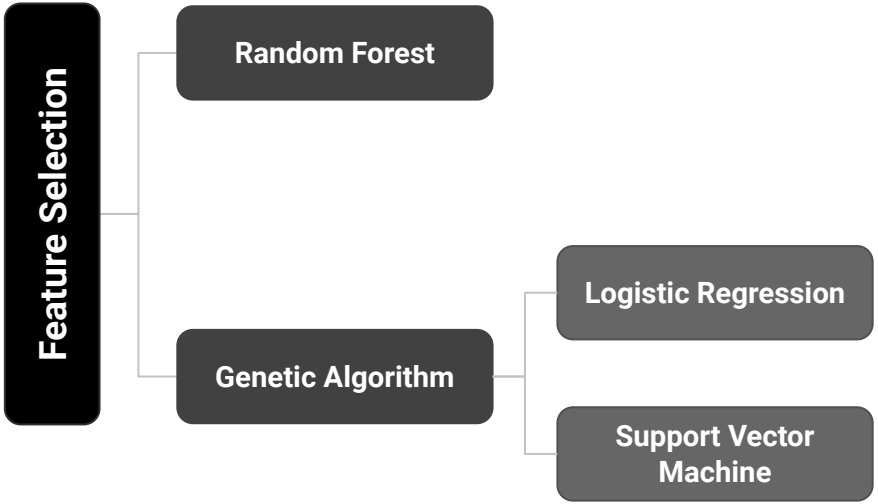
R

Preprocessing
Network Analysis

OncoSearch

Cancer gene search

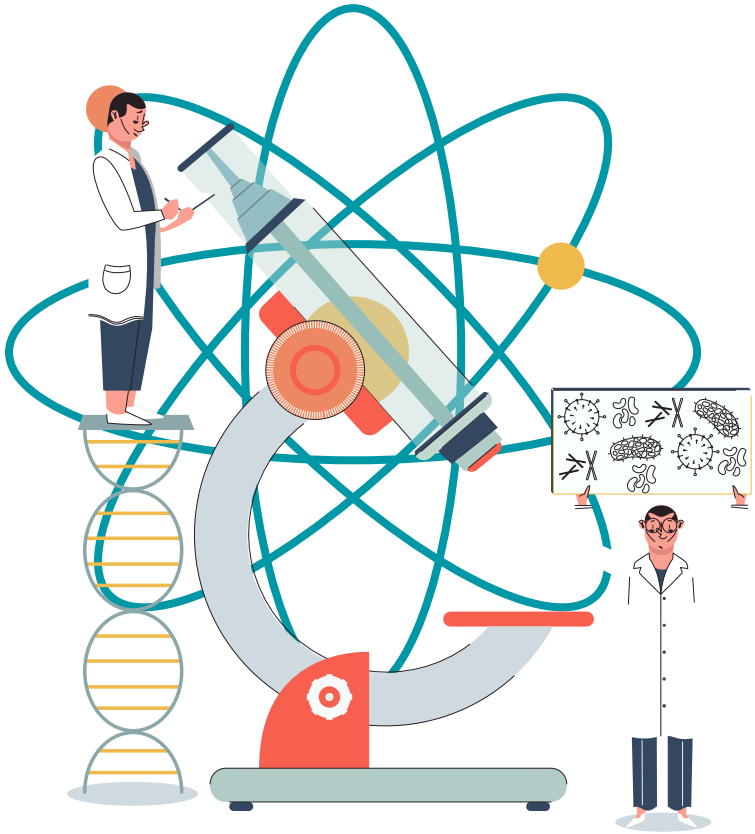
What did not work



Asians from TCGA were classified as White

Overfitting

Noisy genes were considered as important



Thanks!