Research Project Proposal:

Machine Learning algorithms applied on RNA-seq data to classify subtypes of Colorectal Cancer

- Chiara Barbera
- chiara.barbera@mail.polimi.it
- Computer Science and Engineering

POLITECNICO MILANO 1863

HONOURS PROGRAMME HP-SR in Information Technology

Overview

- ✔ Research topic
- ✔ State of the art
  - CRIS classification
- ✔ Our contribute

ISTITUTO DI CANDIOLO

Research Topic

Background information

Key points

✔ **Improve classification of subtypes of ColoRectal Cancer**

✔ Use information on gene activity (**gene expression**)

✔ Cope with computational challenges

What is Colorectal Cancer (CRC)

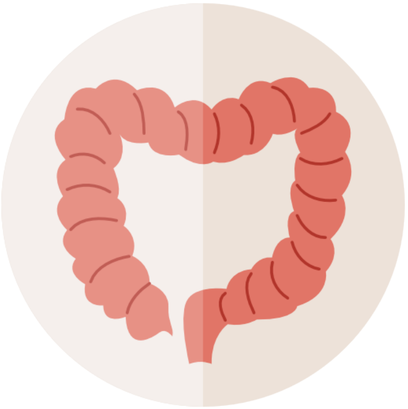
- ✔ Abnormal growth (polyps) in the colon rectum
- ✔ If cancerous, polyps can spread to lymph nodes and other organs[1]
- ✔ Heterogeneity of prognosis and therapy response

~90%

5-year survival rate (localized)[2]

~15%

5-year survival rate (distant)[2]



1) “What is colorectal cancer?.” https://www.cancer.org/cancer/colon-rectal-cancer/about/what-is-colorectal-cancer.html. Accessed: 2020-03-022.
2) “Key statistics for colorectal cancer.” https://www.cancer.org/cancer/colon-rectal-cancer/about/key-statistics.html. Accessed: 2020-03-022.

Images from pngwave.com

Genomics, epigenomics, genes

- ✔ **Gene[1]:** portion of the DNA that encodes for a product (RNA or protein). Placed on chromosomes, usually named with a short combination of letters (and possibly numbers)
- ✔ **Gene expression[1]:** measure of the activity of the genes
- ✔ **Epigenome[2]:** factors that modify how genes are expressed without modifying the DNA

23

Human pairs of chromosomes

20k – 25k

Genes in every person[3]

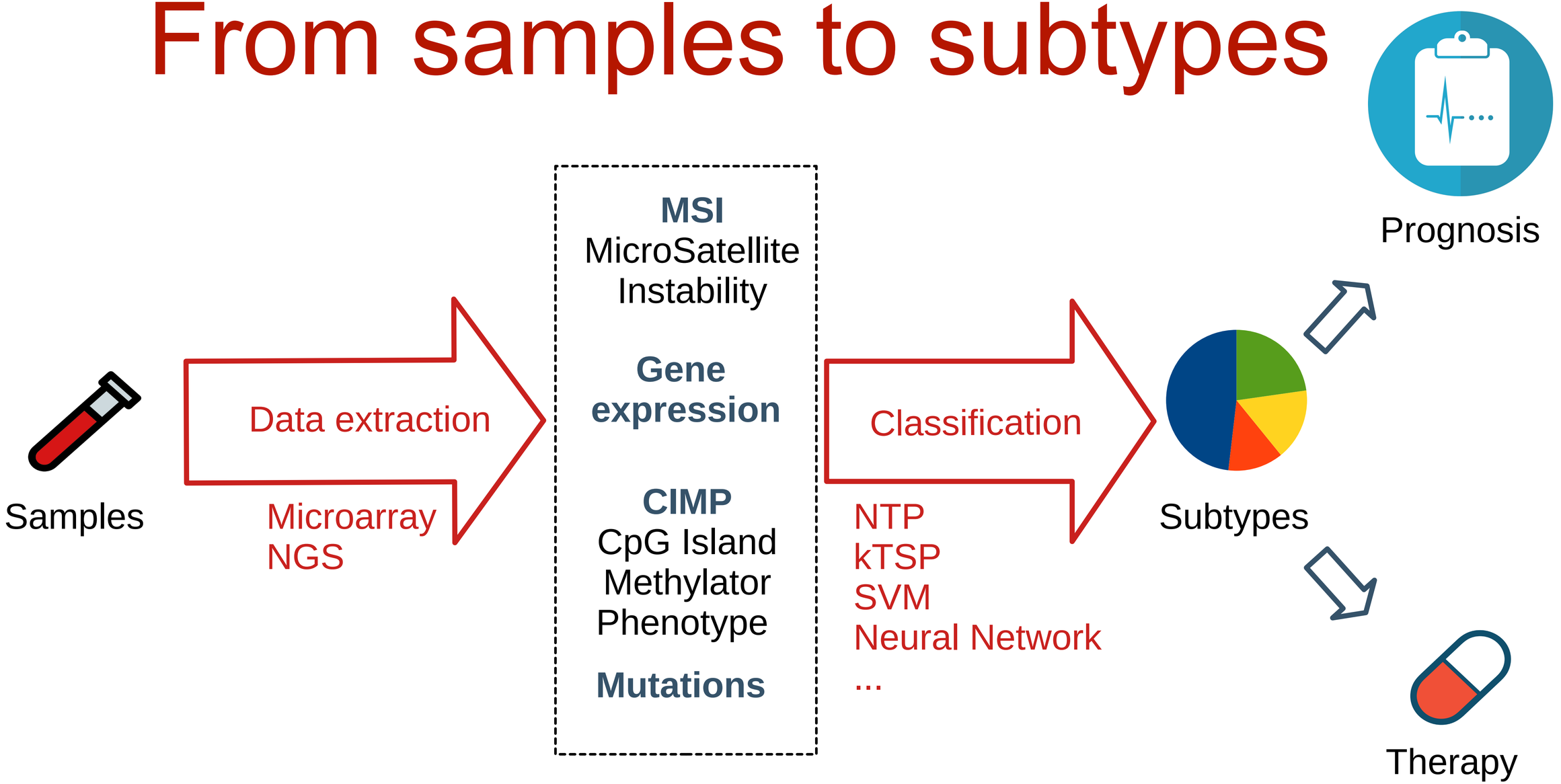
<1%

Different sequences of genes between individuals[3]

Images from pngwave.com;

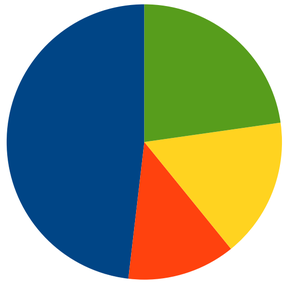
1) “Talking glossary of genetic terms.” https://www.genome.gov/genetics-glossary . Accessed: 2020-03-027.
2) “What is epigenetics?.” https://ghr.nlm.nih.gov/primer/howgeneswork/epigenome . Accessed: 2020-03-031.
3) Number of genes from https://ghr.nlm.nih.gov/primer/basics/gene

From samples to subtypes



Images from pngwave.com

Great, but...



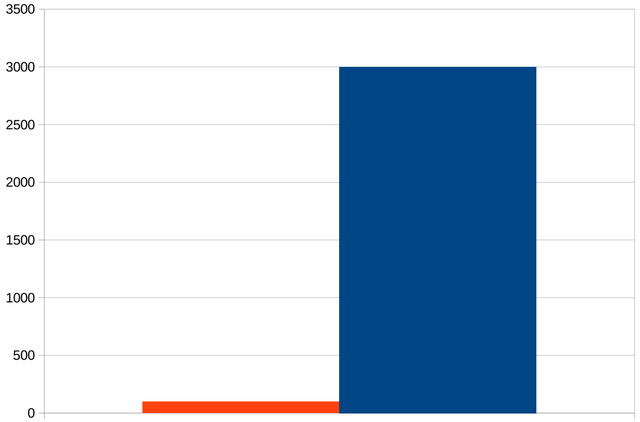
Unbalanced dataset

Ambiguity of data

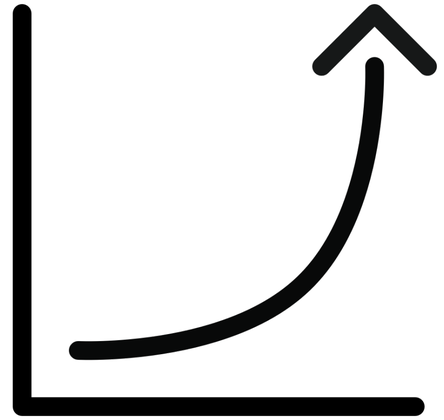
Variety of platforms

FPKM
RPKM **TPM**

Variety of normalizations



Few data, too many features



Curse of dimensionality

Images from pngwave.com

Platforms

MICROARRAY[1]

- ✔ Older
- ✔ Uses fluorochrome to mark binded sequences on a chip
- ✔ Can use only a limited set of genes on the chip

Next-Generation Sequencing

NGS[2]

- ✔ More recent
- ✔ Parallel sequencing
- ✔ Can sequence whole genome
- ✔ RNA-seq, DNA-seq and other

Images from pngwave.com

1) “Talking glossary of genetic terms.” https://www.genome.gov/genetics-glossary . Accessed: 2020-03-027.
2) “Definition of next-generation sequencing.” https://www.cancer.gov/publications/dictionaries/genetics-dictionary/def/next-generation-sequencing . Accessed: 2020-03-024.

State of the art

How did researchers face these issues?

Consensus Molecular Subtype Classification

✔ 4 subtypes
✔ Aggregated 6 different classifications and 18 datasets (Network approach + Markov Clustering Algorithm) and Random Forest Classifier
✗ 22 % = non-consensus samples, classified through a probability threshold

SUBTYPE	SOME CHARACTERISTICS	DISTRIBUTION OF CMS
CMS1	MSI	14 %
CMS2	Chromosome Instability	37 %
CMS3	Methabolic disregulation	13 %
CMS4	**Stromal influence**	23 %
		87% of samples classified

Ambiguity of data

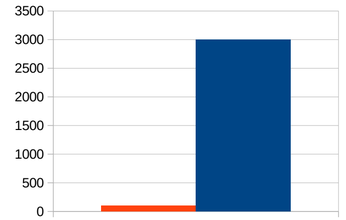
J. Guinney, R. Dienstmann, X. Wang, A. Reyniès, A. Schlicker, C. Soneson, L. Marisa, P. Roepman, G. Nyamundanda, P. Angelino, B. Bot, J. Morris, I. Simon, S. Gerster, E. Fessler, F. De Sousa E Melo, E. Missiaglia, H. Ramay, D. Barras, and S. Tejpar, “The consensus molecular subtypes of colorectal cancer,” Nature Medicine, vol. 21, pp. 1350–1356, 2015.

Study for chemotherapy response

- ✔ Differences in prognosis and chemotherapy response according to subtype
- ✗ Only 71% of classified samples
- ✗ Few samples respecting therapy requirements for the analysis

	MSI STATUS	CIMP	BRAF MUTATION	KRAS MUTATION
SUBTYPE 1	INSTABLE	+	+	-
SUBTYPE 2	STABLE	+	+	-
SUBTYPE 3	STABLE	-	-	+
SUBTYPE 4	STABLE	-	-	-
SUBTYPE 5	INSTABLE	-	-	-

Ambiguity of data



Few samples

O. Murcia, M. Juárez, M. Rodríguez-Soler, E. Hernández-Illán, M. Giner-Calabuig, M. Alustiza, C. Egoavil, A. Castillejo, C. Alenda, V. Barberá, C. Mangas-Sanjuan, A. Yuste, L. Bujanda, J. Clofent, M. Andreu, A. Castells, X. Llor, P. Zapater, and R. Jover, “Colorectal cancer molecular classification using braf, kras, microsatellite instability and cimp status: Prognostic implications and response to chemotherapy,” Plos One, vol. 13, p. e0203051, 2018.

Feature Specific Quantile Normalization (FSQN)

- ✔ Data on Breast and Colorectal cancer (Consensus Molecular Subtype)
- ✔ Train on microarray, validation on RNA-seq data
- ✔ Several normalization procedure. The FSQN showed the highest performances

FPKM
RPKM **TPM**

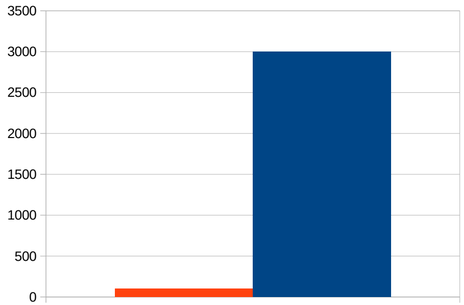
Variety of normalizations

Variety of platforms

J. Franks, G. Cai, and M. Whitfield, “Feature specific quantile normalization enables cross-platform classification of molecular subtypes using gene expression data,” Bioinformatics (Oxford, England), vol. 34, p. 1868–1874, 2018.

Binary and Pan classifier

- ✔ First Classifier: Normal VS cancerous cells
- ✔ Second Classifier: for 21 types of tumors
- ✔ Selected different set of genes and applied different algorithms
  - ✔ **Neural Network**
  - ✔ Linear Support Vector Machine
  - ✔ Radial Basis Function Support Vector Machine
  - ✔ K-Nearest Neighbours
  - ✔ Random Forest
- ✗ Some samples were misclassified by pan classifier (near regions)



Few samples,
too many features

Ambiguity of data

B.-H. Kim, K. Yu, and P. Lee, “Cancer classification of single cell gene expression data by neural network,”Bioinformatics (Oxford, England), vol. 36, p. 1–7, 2019.

Our objective

Single-sample classifier for CRIS[1] subtypes of Colorectal Cancer (CRC)

Why ?

- ✔ Improve the current approaches and results
- ✔ Only RNA-seq data (NGS)
- ✔ Necessary step towards a clinical application

1) C. Isella, F. Brundu, S. E. Bellomo, F. Galimi, E. Zanella, R. Porporato, C. Petti, A. Fiori, F. Orzan, R. Senetta, C. Boccaccio, E. Ficarra, L. Marchionni, L. Trusolino, E. Medico, and A. Bertotti, “Selective analysis of cancer-cell intrinsic transcriptional traits defines novel clinically relevant subtypes of colorectal cancer,” Nature Communications, vol. 8, p. 15107, 2017.

Our starting point

The CRIS Classification

CRIS Classification

Next-Generation Sequencing

Variety of platforms

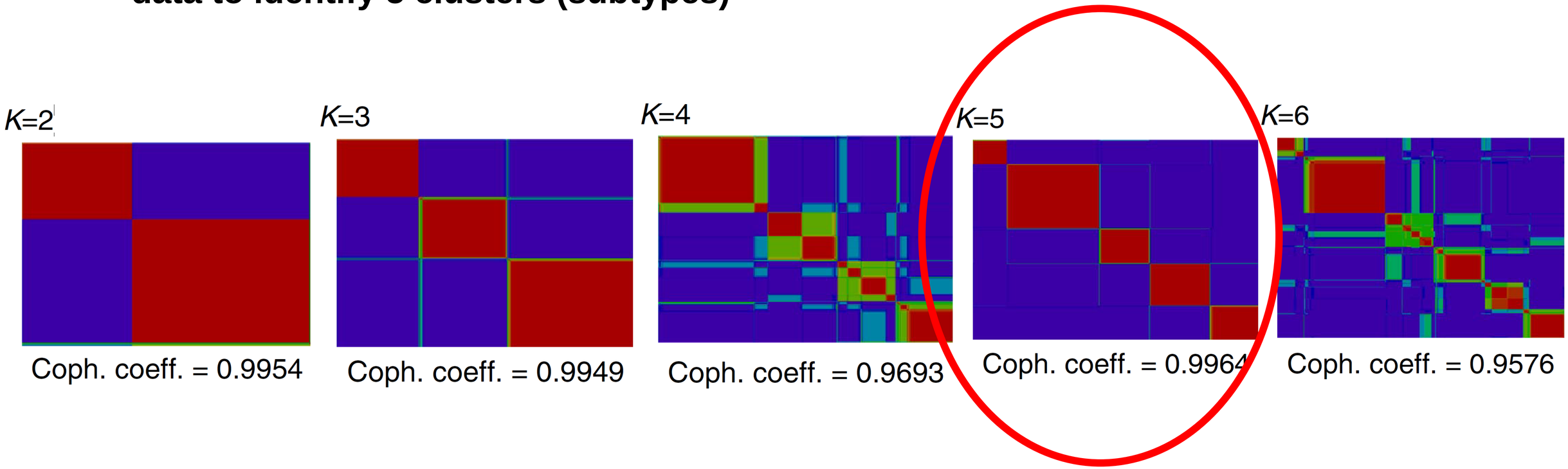
- ✔ 5 CRIS subtypes (ColoRectal Intrinsic Subtypes)
- ✔ Training on microarray, testing on both microarray and RNA-seq

SUBTYPE	SOME CHARACTERISTICS
CRIS-A	Enriched for MSI or KRAS gene mutation
CRIS-B	Poor prognosis
CRIS-C	Sensitivity to EFGR inibitors (responsive to cetuximab)
CRIS-D	IGF2 gene overexpression
CRIS-E	TP53 gene mutation

1) C. Isella, F. Brundu, S. E. Bellomo, F. Galimi, E. Zanella, R. Porporato, C. Petti, A. Fiori, F. Orzan, R. Senetta, C. Boccaccio, E. Ficarra, L. Marchionni, L. Trusolino, E. Medico, and A. Bertotti, “Selective analysis of cancer-cell intrinsic transcriptional traits defines novel clinically relevant subtypes of colorectal cancer,” Nature Communications, vol. 8, p. 15107, 2017.

Machine learning analysis

1) Non-negative Matrix Factorization (NMF) on PDX (Patient Derived Xenografts) data to identify 5 clusters (subtypes)



1) C. Isella, F. Brundu, S. E. Bellomo, F. Galimi, E. Zanella, R. Porporato, C. Petti, A. Fiori, F. Orzan, R. Senetta, C. Boccaccio, E. Ficarra, L. Marchionni, L. Trusolino, E. Medico, and A. Bertotti, "Selective analysis of cancer-cell intrinsic transcriptional traits defines novel clinically relevant subtypes of colorectal cancer," Nature Communications, vol. 8, p. 15107, 2017.

Machine learning analysis

1) Non-negative Matrix Factorization (NMF) on PDX (Patient Derived Xenografts) data to identify 5 clusters (subtypes)

2) Definition of templates of each subtype, based on NMF

1) C. Isella, F. Brundu, S. E. Bellomo, F. Galimi, E. Zanella, R. Porporato, C. Petti, A. Fiori, F. Orzan, R. Senetta, C. Boccaccio, E. Ficarra, L. Marchionni, L. Trusolino, E. Medico, and A. Bertotti, "Selective analysis of cancer-cell intrinsic transcriptional traits defines novel clinically relevant subtypes of colorectal cancer," Nature Communications, vol. 8, p. 15107, 2017.

Machine learning analysis

1) Non-negative Matrix Factorization (NMF) on PDX (Patient Derived Xenografts) data to identify 5 clusters (subtypes)

2) Definition of templates of each subtype, based on NMF

3) NTP (Nearest Template Prediction) classification

- ✔ **On dataset (microarray + RNA-seq data)**
- ✔ **Good accuracy**

1) C. Isella, F. Brundu, S. E. Bellomo, F. Galimi, E. Zanella, R. Porporato, C. Petti, A. Fiori, F. Orzan, R. Senetta, C. Boccaccio, E. Ficarra, L. Marchionni, L. Trusolino, E. Medico, and A. Bertotti, “Selective analysis c
cancer-cell intrinsic transcriptional traits defines novel clinically relevant subtypes of colorectal cancer,” Nature Communications, vol. 8, p. 15107, 2017.

Machine learning analysis

1) Non-negative Matrix Factorization (NMF) on PDX (Patient Derived Xenografts) data to identify 5 clusters (subtypes)

2) Definition of templates of each subtype, based on NMF

3) NTP (Nearest Template Prediction) classification

✔ On dataset (microarray + RNA-seq data)

✔ Good accuracy

4) Attempt of single-sample classifier through kTSP (k Top Scoring Pairs)

✗ **Suboptimal performances**

1) C. Isella, F. Brundu, S. E. Bellomo, F. Galimi, E. Zanella, R. Porporato, C. Petti, A. Fiori, F. Orzan, R. Senetta, C. Boccaccio, E. Ficarra, L. Marchionni, L. Trusolino, E. Medico, and A. Bertotti, “Selective analysis of cancer-cell intrinsic transcriptional traits defines novel clinically relevant subtypes of colorectal cancer,” Nature Communications, vol. 8, p. 15107, 2017.

Machine learning analysis

1) Non-negative Matrix Factorization (NMF) on PDX (Patient Derived Xenografts) data to identify 5 clusters (subtypes)

2) Definition of templates of each subtype, based on NMF

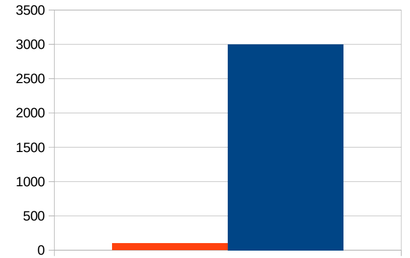
3) NTP classification

✔ On dataset (microarray + RNA-seq data)

✔ Good accuracy

4) Attempt of single-sample classifier through kTSP (k Top Scoring Pairs)

✗ Suboptimal performances

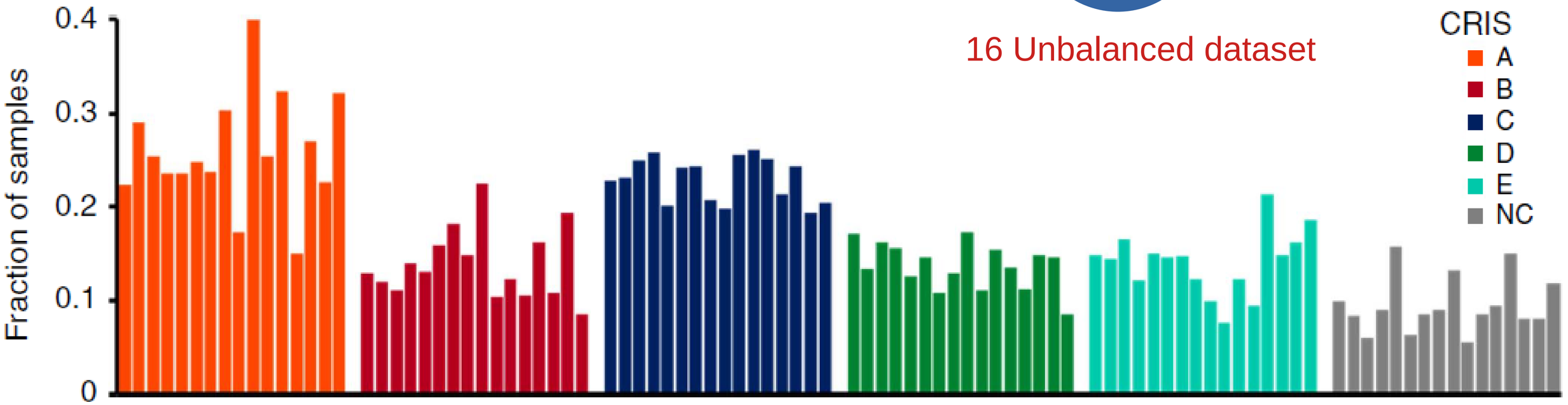


Few samples,
too many features

1) C. Isella, F. Brundu, S. E. Bellomo, F. Galimi, E. Zanella, R. Porporato, C. Petti, A. Fiori, F. Orzan, R. Senetta, C. Boccaccio, E. Ficarra, L. Marchionni, L. Trusolino, E. Medico, and A. Bertotti, “Selective analysis of cancer-cell intrinsic transcriptional traits defines novel clinically relevant subtypes of colorectal cancer,” Nature Communications, vol. 8, p. 15107, 2017.

Results of NTP analysis

16 Unbalanced dataset



1) C. Isella, F. Brundu, S. E. Bellomo, F. Galimi, E. Zanella, R. Porporato, C. Petti, A. Fiori, F. Orzan, R. Senetta, C. Boccaccio, E. Ficarra, L. Marchionni, L. Trusolino, E. Medico, and A. Bertotti, "Selective analysis of cancer-cell intrinsic transcriptional traits defines novel clinically relevant subtypes of colorectal cancer," Nature Communications, vol. 8, p. 15107, 2017.

Pros and Cons of CRIS

- ✔ Removed stromal influence
- ✔ Good performances for the dataset classifier
- ✔ Cross – platform (Microarray and RNA-seq data)

- ✗ **Single–sample classifier is suboptimal**
- ✗ Microarray do not provide as many features as RNA-seq data
- ✗ **Instability of classification** for some samples due to dataset composition dependency (z-score)
- ✗ Some genes may still be stromal

Ambiguity of data

1) C. Isella, F. Brundu, S. E. Bellomo, F. Galimi, E. Zanella, R. Porporato, C. Petti, A. Fiori, F. Orzan, R. Senetta, C. Boccaccio, E. Ficarra, L. Marchionni, L. Trusolino, E. Medico, and A. Bertotti, “Selective analysis of cancer-cell intrinsic transcriptional traits defines novel clinically relevant subtypes of colorectal cancer,” Nature Communications, vol. 8, p. 15107, 2017.

Our contribution

What we want to do

What do we want to investigate?

✔ **Is CRIS classification stable on RNA-seq only?**

What do we want to investigate?

✔ Is CRIS classification stable on RNA-seq only?

✔ **Are there any new features (more meaningful) on RNA-seq data?**

What do we want to investigate?

✔ Is CRIS classification stable on RNA-seq only?

✔ Are there any new features (more meaningful) on RNA-seq data?

✔ **Can we obtain a single-sample classifier with high accuracy?**

What do we want to investigate?

✔ Is CRIS classification stable on RNA-seq only?

✔ Are there any new features (more meaningful) on RNA-seq data?

✔ Can we obtain a single-sample classifier with high accuracy?

✔ **If new features are selected, what is their biological meaning?**

What do we want to investigate?

✔ Is CRIS classification stable on RNA-seq only?

✔ Are there any new features (more meaningful) on RNA-seq data?

✔ Can we obtain a single-sample classifier with high accuracy?

✔ If new features are selected, what is their biological meaning?

✔ **Are there other subtypes for CRIS? If so, does the accuracy improve?**

How will we face the issues?

✔ **Curse of dimensionality and small dataset?**
Apply feature selection to select only the important features

✔ **Heterogeneity of platform and data?**
Only RNA-seq data and uniform type of normalization

✔ **Unbalanced dataset?**
Sample selection

What is new in the research?

✔ **Application of CRIS on RNA-seq data only** because they have more features

✔ **No single-sample classifier** is currently available

✔ **Single-sample classifier** development is necessary for clinical application

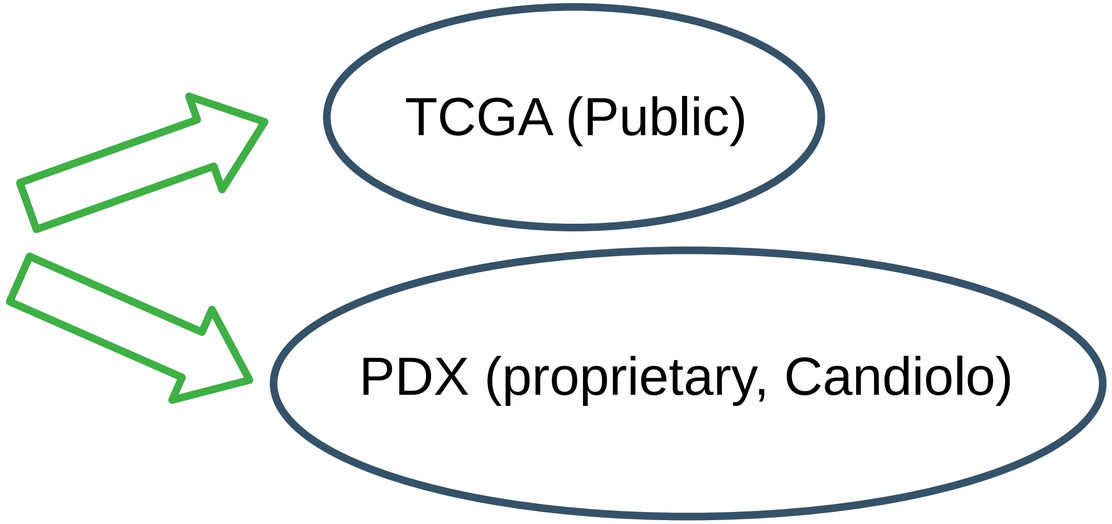
✔ Possibly develop a **new algorithm for feature selection**

✔ Possibly **hybrid approaches** (starting with unsupervised) **for classification**

Data and tools

RNA-seq data from

- ✔ More features
- ✔ More precise
- ✔ Current trend technology



Steps of our work

1) Collection of data (TCGA RNA-Seq data + PDX RNA-seq data)

Steps of our work

1) Collection of data (TCGA RNA-Seq data + PDX RNA-seq data)

2) Replication of CRIS NTP and kTSP classifier on these data

Steps of our work

1) Collection of data (TCGA RNA-Seq data + PDX RNA-seq data)

2) Replication of CRIS NTP and kTSP classifiers on these data

3) Study of possible alternative classifiers

- **1) Feature selection and sample selection**
- **2) Execution of classifiers**

Steps of our work

1) Collection of data (TCGA RNA-Seq data + PDX RNA-seq data)
2) Replication of CRIS NTP and kTSP classifiers on these data
3) Study of possible alternative classifiers
    1) Feature selection and sample selection
    2) Execution of classifiers
4) Performance comparison
    1) CRIS classification of Isella *et al.*
    2) CRIS classification on RNA-seq data only
    3) Alternative classifiers

Steps of our work

1) Collection of data (TCGA RNA-Seq data + PDX RNA-seq data)
2) Replication of CRIS NTP and kTSP classifiers on these data
3) Study of possible alternative classifiers
    1) Feature selection and sample selection
    2) Execution of classifiers
4) Performance comparison
    1) CRIS classification of Isella *et al.*
    2) CRIS classification on RNA-seq data only
    3) Alternative classifiers
5) Clinical and biological validation

Conclusions

1) Several issues to take into account

Conclusions

1) Several issues to take into account

2) Other studies coped with these issues and reached relevant results on classification and response to therapy

Conclusions

1) Several issues to take into account

2) Other studies coped with these issues and reached relevant results on classification and response to therapy

3) We want to improve the results of the CRIS classification by applying alternative classifiers on RNA-seq data only

Conclusions

1) Several issues to take into account

2) Other studies coped with these issues and reached relevant results on classification and response to therapy

3) We want to improve the results of the CRIS classification by applying alternative classifiers on RNA-seq data only (new feature selection and/or classification algorithm)

4) We would like to have a single-sample classifier with acceptable accuracy

Questions?

Thank you for your attention!