2^{e} édition des Journées Bioss-IA

Search of Therapeutic Targets on the Hepatocellular Carcinoma with Database Extraction and Graph Coloring Methods

Recherche de cibles thérapeutiques pour le carcinome hépatocellulaire à l'aide d'extraction de bases de données et de méthodes de coloration de graphes

Maxime FOLSCHETTE

Current occupation:
CNRS & Institut Français de Bioinformatique (IFB)
Laboratoire des Sciences du Numérique de Nantes (LS2N)
groups: GDD & COMBI

Previous occupation:
Université de Rennes 1
IRISA & IRSET
groups: Dyliss & Dymec

maxime.folschette@ls2n.fr
http://maxime.folschette.name/

2018-12-19

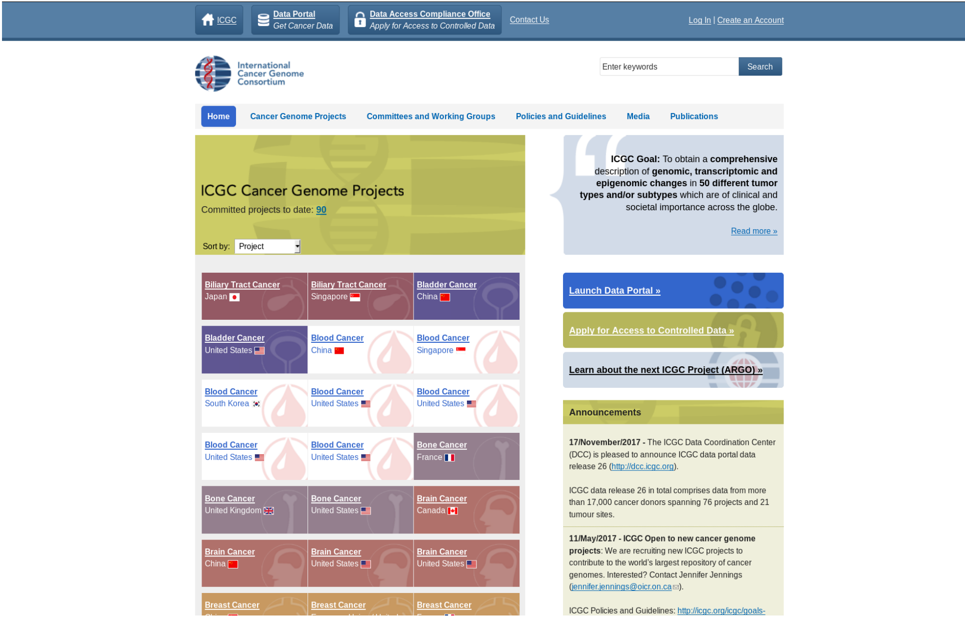
Context

Hepatocellular carcinoma (HCC)

- Most widespread liver cancer, 3rd most deadly cancer
- Mainly provoked by hepatitis and fibrosis
- Late diagnosis and difficult to treat (resection, transplant, ablation, chemo-embolization)
- Very low survival rate: from weeks to months

Objective

- Gather data from **ICGC** [Hudson and The International Cancer Genome Consortium, 2010]
- Distinguish two tumor stages: early stage vs. late (invasive) stage
- Watch expression change (differential expression)



ICGC
Data Portal
Get Cancer Data
Data Access Compliance Office
Apply for Access to Controlled Data
Contact Us
Log In | Create an Account
International Cancer Genome Consortium
Enter keywords
Search
Home
Cancer Genome Projects
Committees and Working Groups
Policies and Guidelines
Media
Publications

ICGC Cancer Genome Projects

Committed projects to date: 90

Sort by: Project

Biliary Tract Cancer Japan	Biliary Tract Cancer Singapore	Bladder Cancer China
Bladder Cancer United States	Blood Cancer China	Blood Cancer Singapore
Blood Cancer South Korea	Blood Cancer United States	Blood Cancer United States
Blood Cancer United States	Blood Cancer United States	Bone Cancer France
Bone Cancer United Kingdom	Bone Cancer United States	Brain Cancer Canada
Brain Cancer China	Brain Cancer United States	Brain Cancer United States
Breast Cancer	Breast Cancer	Breast Cancer

ICGC Goal: To obtain a **comprehensive** description of **genomic, transcriptomic and epigenomic changes** in **50 different tumor types and/or subtypes** which are of clinical and societal importance across the globe.

Read more »

Launch Data Portal »
Apply for Access to Controlled Data »
Learn about the next ICGC Project (ARGO) »

Announcements

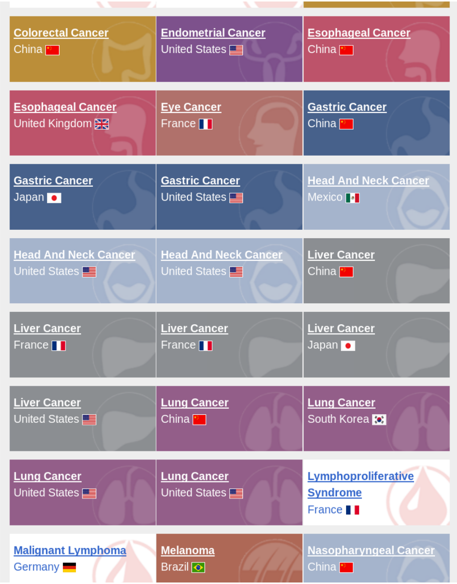
17/November/2017 - The ICGC Data Coordination Center (DCC) is pleased to announce ICGC data portal data release 26 (http://dcc.icgc.org).

ICGC data release 26 in total comprises data from more than 17,000 cancer donors spanning 76 projects and 21 tumour sites.

11/May/2017 - ICGC Open to new cancer genome projects: We are recruiting new ICGC projects to contribute to the world's largest repository of cancer genomes. Interested? Contact Jennifer Jennings (jennifer.jennings@oicr.on.ca).

ICGC Policies and Guidelines: http://icgc.org/icgc/goals-

https://dcc.icgc.org/projects/LIHC-US



Colorectal Cancer China	Endometrial Cancer United States	Esophageal Cancer China
Esophageal Cancer United Kingdom	Eye Cancer France	Gastric Cancer China
Gastric Cancer Japan	Gastric Cancer United States	Head And Neck Cancer Mexico
Head And Neck Cancer United States	Head And Neck Cancer United States	Liver Cancer China
Liver Cancer France	Liver Cancer France	Liver Cancer Japan
Liver Cancer United States	Lung Cancer China	Lung Cancer South Korea
Lung Cancer United States	Lung Cancer United States	Lymphoproliferative Syndrome France
Malignant Lymphoma Germany	Melanoma Brazil	Nasopharyngeal Cancer China

LIHC-US

LIHC-US in ICGC

Project for liver HCC (USA)

- 294 samples with gene expression data
- Primary tumor on solid tissue only
- 20502 genes
- 16282 genes when excluding low expression

But no tumor grade annotation!

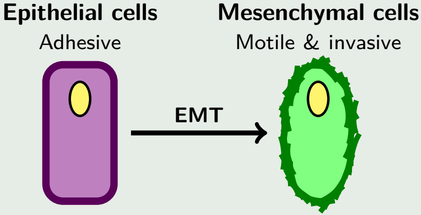
⇒ We need a **criterion** to distinguish tumor stages

Objectives

1) Clustering on the **criterion** ⇒ Two groups
2) Differential analysis on the two groups

Epithelial-Mesenchymal Transition

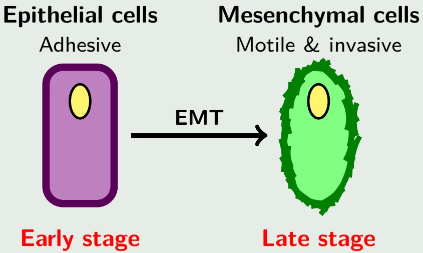
Epithelial-mesenchymal transition (EMT)



- De-differentiation of epithelial cells to mesenchymal cells
- Gain ability to **remodel** the extra-cellular matrix and **migrate**
- Invasive cancer cells ⇒ **metastasis**

Epithelial-Mesenchymal Transition

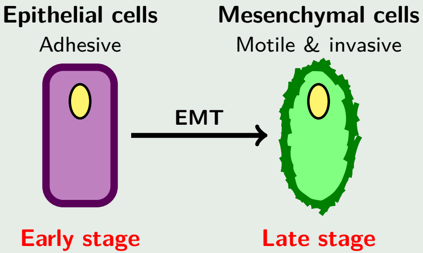
Epithelial-mesenchymal transition (EMT)



- De-differentiation of epithelial cells to mesenchymal cells
- Gain ability to **remodel** the extra-cellular matrix and **migrate**
- Invasive cancer cells ⇒ **metastasis**
- Indication of **tumor stage**
  ⇒ **Criterion**

Epithelial-Mesenchymal Transition

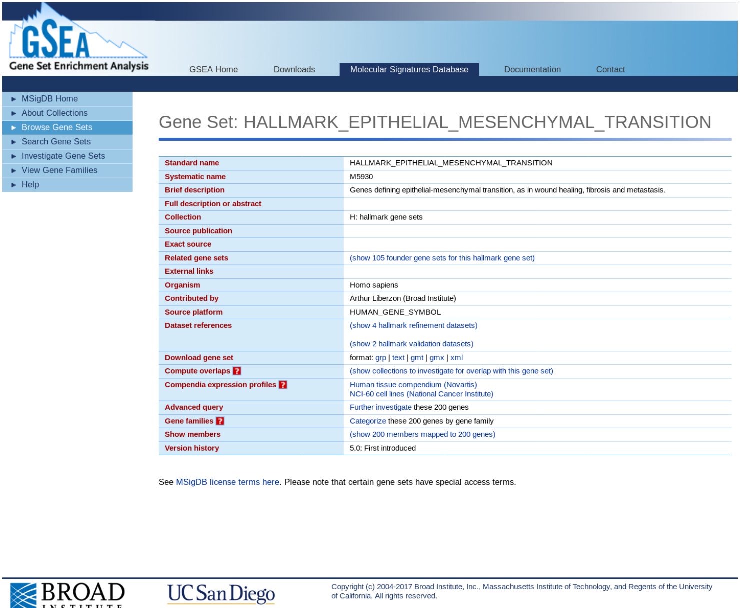
Epithelial-mesenchymal transition (EMT)



- De-differentiation of epithelial cells to mesenchymal cells
- Gain ability to **remodel** the extra-cellular matrix and **migrate**
- Invasive cancer cells ⇒ **metastasis**
- Indication of **tumor stage** ⇒ **Criterion**

- **EMT signature** = Set of genes that are over-expressed during EMT (includes TGF-β)
- Downloaded on **GSEA** [Subramanian et al., 2005]

http://software.broadinstitute.org/gsea/msigdb/cards/HALLMARK_EPITHELIAL_MESENCHYMAL_TRANSITION.html

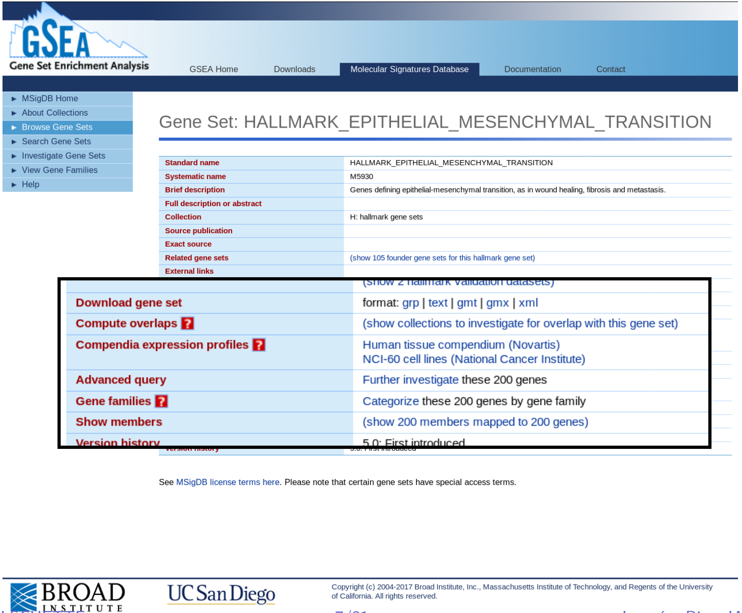


Gene Set: HALLMARK_EPITHELIAL_MESENCHYMAL_TRANSITION

Standard name	HALLMARK_EPITHELIAL_MESENCHYMAL_TRANSITION
Systematic name	M5930
Brief description	Genes defining epithelial-mesenchymal transition, as in wound healing, fibrosis and metastasis.
Full description or abstract	
Collection	H: hallmark gene sets
Source publication	
Exact source	
Related gene sets	(show 105 founder gene sets for this hallmark gene set)
External links	
Organism	Homo sapiens
Contributed by	Arthur Liberzon (Broad Institute)
Source platform	HUMAN_GENE_SYMBOL
Dataset references	(show 4 hallmark refinement datasets) (show 2 hallmark validation datasets)
Download gene set	format: grp | text | gmt | gmx | xml
Compute overlaps ?	(show collections to investigate for overlap with this gene set)
Compendia expression profiles ?	Human tissue compendium (Novartis) NCI-60 cell lines (National Cancer Institute)
Advanced query	Further investigate these 200 genes
Gene families ?	Categorize these 200 genes by gene family
Show members	(show 200 members mapped to 200 genes)
Version history	5.0: First introduced

See MSigDB license terms here. Please note that certain gene sets have special access terms.

http://software.broadinstitute.org/gsea/msigdb/cards/HALLMARK_EPITHELIAL_MESENCHYMAL_TRANSITION.html

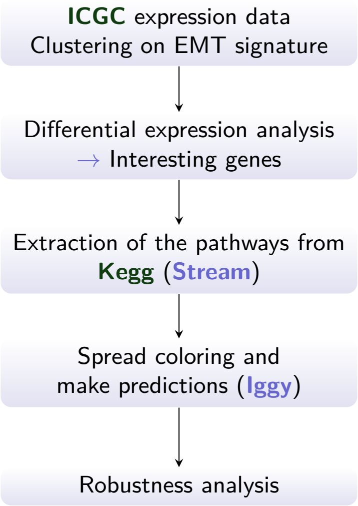


Gene Set: HALLMARK_EPITHELIAL_MESENCHYMAL_TRANSITION

Standard name	HALLMARK_EPITHELIAL_MESENCHYMAL_TRANSITION
Systematic name	M5930
Brief description	Genes defining epithelial-mesenchymal transition, as in wound healing, fibrosis and metastasis.
Full description or abstract	
Collection	H: hallmark gene sets
Source publication	
Exact source	
Related gene sets	(show 105 founder gene sets for this hallmark gene set)
External links	(show 2 hallmark validation datasets)
Download gene set	format: grp | text | gmt | gmx | xml
Compute overlaps ?	(show collections to investigate for overlap with this gene set)
Compendia expression profiles ?	Human tissue compendium (Novartis) NCI-60 cell lines (National Cancer Institute)
Advanced query	Further investigate these 200 genes
Gene families ?	Categorize these 200 genes by gene family
Show members	(show 200 members mapped to 200 genes)
Version history	5.0: First introduced

See MSigDB license terms here. Please note that certain gene sets have special access terms.

Workflow of the Project



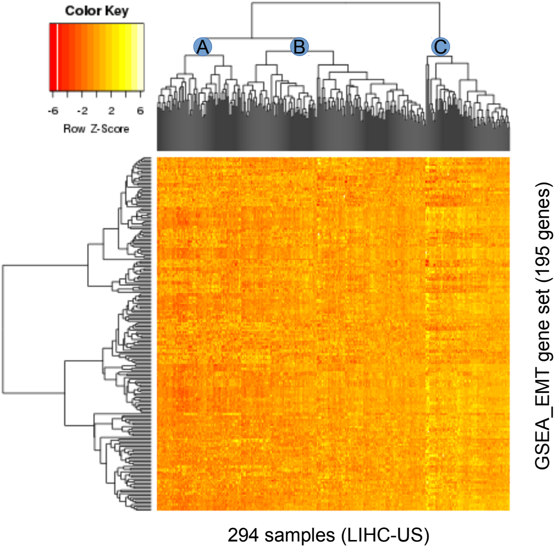
ICGC expression data
Clustering on EMT signature

Differential expression analysis
→ Interesting genes

Extraction of the pathways from
Kegg (**Stream**)

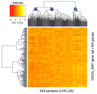
Spread coloring and
make predictions (**Iggy**)

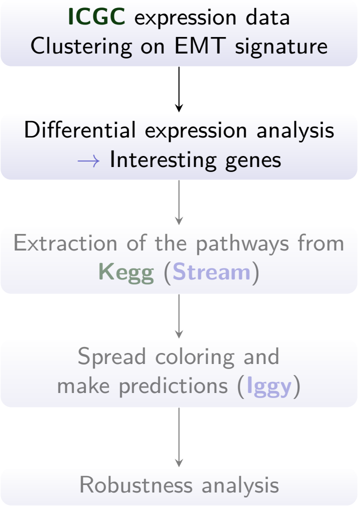
Robustness analysis



Group A = Low expression of the EMT signature
Group C = High expression of the EMT signature

Workflow of the Project





ICGC expression data
Clustering on EMT signature

Differential expression analysis
→ Interesting genes

Extraction of the pathways from
Kegg (**Stream**)

Spread coloring and
make predictions (**Iggy**)

Robustness analysis

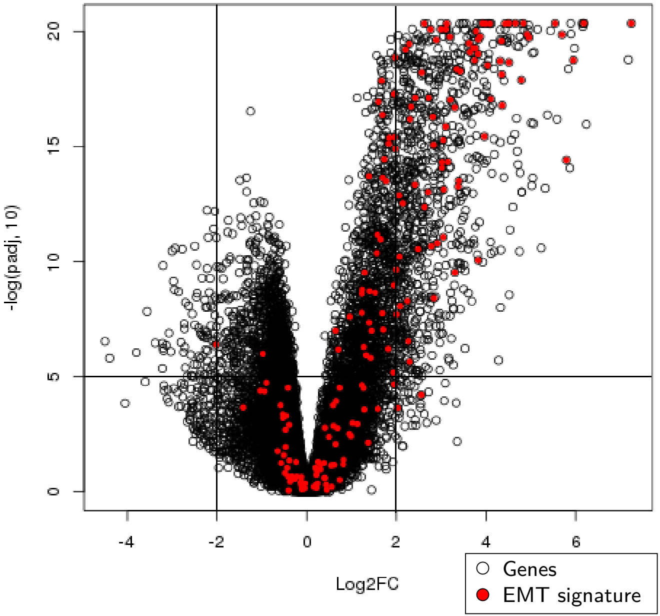
2 groups

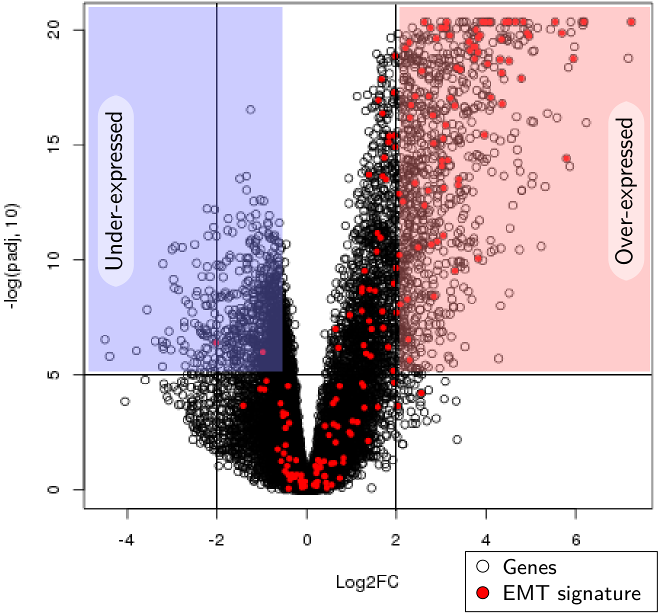
Differential Analysis

Fold-change definition

- Consider groups A (lowest expression of EMT) and C (resp. highest)
- For each gene g, compute mean value for group A (resp. C)
- Differential analysis:

$$\text{fold-change}(g) = \text{mean}_g(C) \ / \ \text{mean}_g(A)$$


-log(padj, 10)
0
5
10
15
20
-4
-2
0
2
4
6
Log2FC
○ Genes
● EMT signature


Under-expressed
Over-expressed
-log(padj, 10)
0
5
10
15
20
-4
-2
0
2
4
6
Log2FC
○ Genes
● EMT signature

Selected genes

Criteria

- Adjusted P-value $< 10^{-5}$
- $\log_2(\text{fold-change}) > 2$ (up-regulated genes)
- $\log_2(\text{fold-change}) < -0,5$ (down-regulated genes)

Selected genes

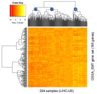
- 821 up-regulated genes
- 1092 down-regulated genes

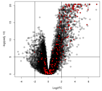
$= 1913$ genes

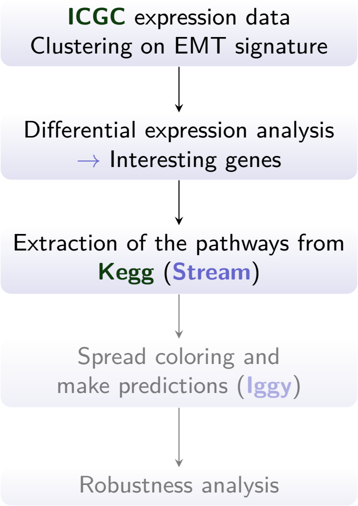
Objectives

1) Extract a graph from **Kegg** [Kanehisa et al., 2017] using these genes, with the tool **Stream**
2) Coloring + predictions with **Iggy** [Thiele et al., 2015]

Workflow of the Project







ICGC expression data
Clustering on EMT signature

2 groups

Differential expression analysis
→ Interesting genes

≃ 2’000 genes

Extraction of the pathways from
Kegg (**Stream**)

Spread coloring and
make predictions (**Iggy**)

Robustness analysis

Pathway Commons + Bravo

Pathway Commons [Cerami et al., 2010]

- A gathering of **25 pathway databases**
- Contains: PID, Kegg, Reactome, CTD, Panther, ...
- Common ontology (BioPAX)
- Freely available
- SPARQL endpoint

BRAvo [Lefebvre et al., 2017]

- **Interrogates** Pathway Commons with SPARQL queries
- Search and fusion of synonyms, optimizations
- (Incomplete) visualization tool

Problems with Pathway Commons

Problems with Pathway Commons

- **Very heterogeneous data**, curation depends on the data sources
- The BioPAX ontology is big and difficult to use
- Unification must be done by the user, based on gene unames (fast) or identifies (slow)
- Updated without history (bad for reproducibility)

Problems with BRAvo

- Still in development, only regulation at the time (no signaling)
- Struggles with the heterogeneous content of Pathway Commons
- Unification was still incomplete

Kegg + Stream

Kegg [Kanehisa et al., 2017]

- Homogeneous data
- Categories: 2. Genetic Information Processing
  3. Environmental Information Processing
  4. Cellular Processes
  5. Organismal Systems
- Already formatted and curated by Arnaud Poret

SIF format: $A \xrightarrow{+/-} B$ "*A* positively/negatively influences *B*"

- Genes (`XXX_gen`)
- Proteins (`XXX_prot`)
- Complexes (`XXX::YYY::ZZZ`)

Stream (Arnaud Poret)

- Ad-hoc program for upstream graph extraction
- Extract the part of the graph for which we have expression data (25%)

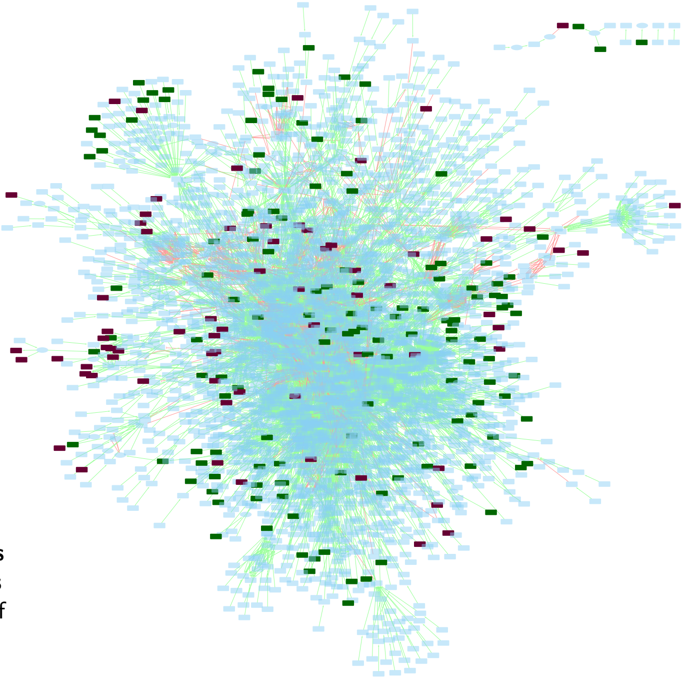
Graph content:

- 3’383 nodes
- 13’771 edges
  - — 11’661 activations
  - — 2’110 inhibitions

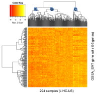
1913 genes from the differential expression
Only 209 are found in Kegg:

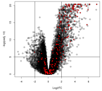
- 🟩 138 up-regulated
- 🟥 71 down-regulated
- 🟦 3174 new nodes

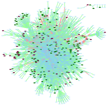
Nodes with up to:
92 incoming influences
79 outgoing influences
→ Nodes with a lot of impact on the network

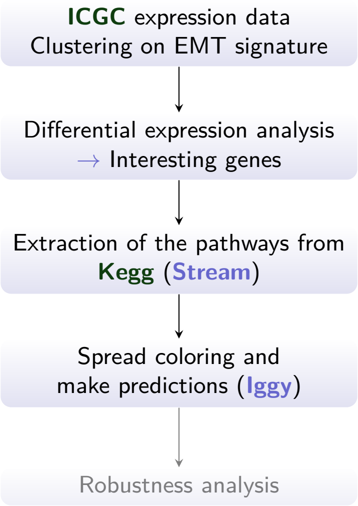


Workflow of the Project









ICGC expression data
Clustering on EMT signature

Differential expression analysis
→ Interesting genes

Extraction of the pathways from
Kegg (**Stream**)

Spread coloring and
make predictions (**Iggy**)

Robustness analysis

2 groups

≃ 2'000 genes

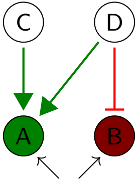
≃ 3'400 nodes, 14'000 edges

Graph Coloring

- Coloring = information attached to nodes about over- or under-expression

  X = over-expressed   Y = under-expressed

- Provenance = experimental (expression data) & computational (inference)



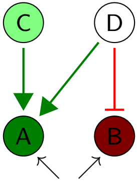
Given by the experimental data

Graph Coloring

- Coloring = information attached to nodes about over- or under-expression

  X = over-expressed   Y = under-expressed

- Provenance = experimental (expression data) & computational (inference)



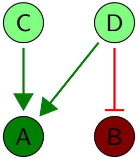
Given by the experimental data

Graph Coloring

- Coloring = information attached to nodes about over- or under-expression

  🟩 X = over-expressed   🟥 Y = under-expressed

- Provenance = experimental (expression data) & computational (inference)



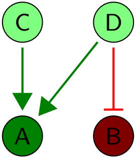
Consistent

Graph Coloring

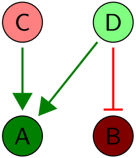
- Coloring = information attached to nodes about over- or under-expression

  X = over-expressed   Y = under-expressed

- Provenance = experimental (expression data) & computational (inference)



Consistent



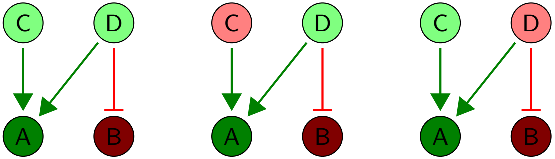
Consistent

Graph Coloring

- Coloring = information attached to nodes about over- or under-expression

  X = over-expressed   Y = under-expressed

- Provenance = experimental (expression data) & computational (inference)



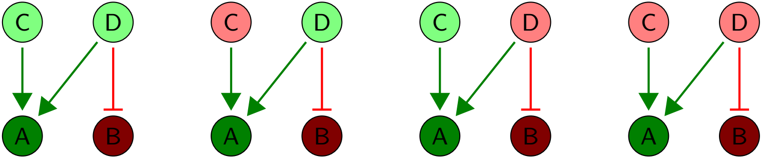
Consistent   **Consistent**   **Inconsistent**

Graph Coloring

- Coloring = information attached to nodes about over- or under-expression

  X = over-expressed   Y = under-expressed

- Provenance = experimental (expression data) & computational (inference)



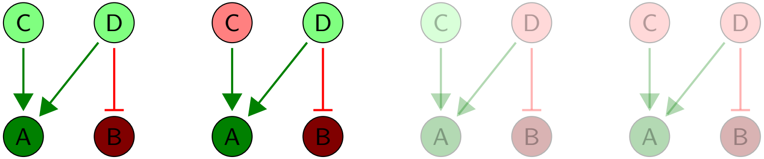
Consistent **Consistent** **Inconsistent** **Inconsistent**

Graph Coloring

- Coloring = information attached to nodes about over- or under-expression

  X = over-expressed   Y = under-expressed

- Provenance = experimental (expression data) & computational (inference)



Consistent   **Consistent**   **Inconsistent**   **Inconsistent**

- Compute all colorings without inconsistencies
- **Prediction** = a node that is always colored the same

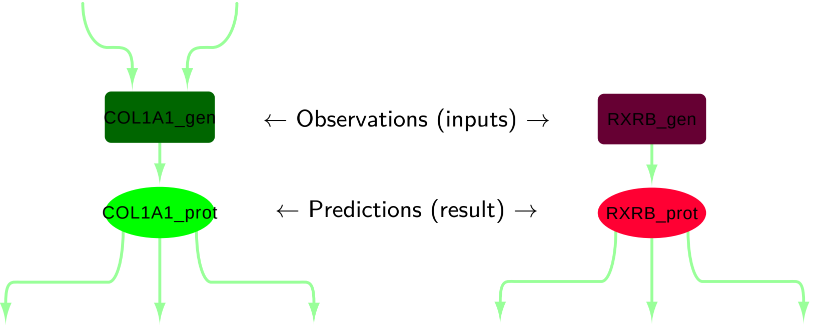
  Here, only 1 prediction: D

- All computed by **Iggy** [Thiele et al., 2015] (Answer Set Programming)

Trivial Predictions

“Trivial” prediction

- Protein predicted the same as its observed gene
- Rarely brings new information
- Useful for validation



← Observations (inputs) →

← Predictions (result) →

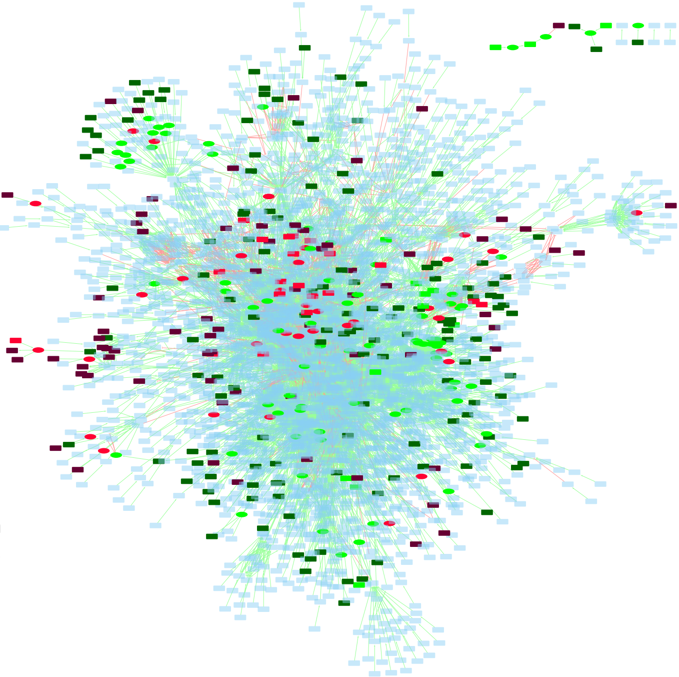
Knowledge from experiments:

- 🟩 138 up-regulated
- 🟫 71 down-regulated

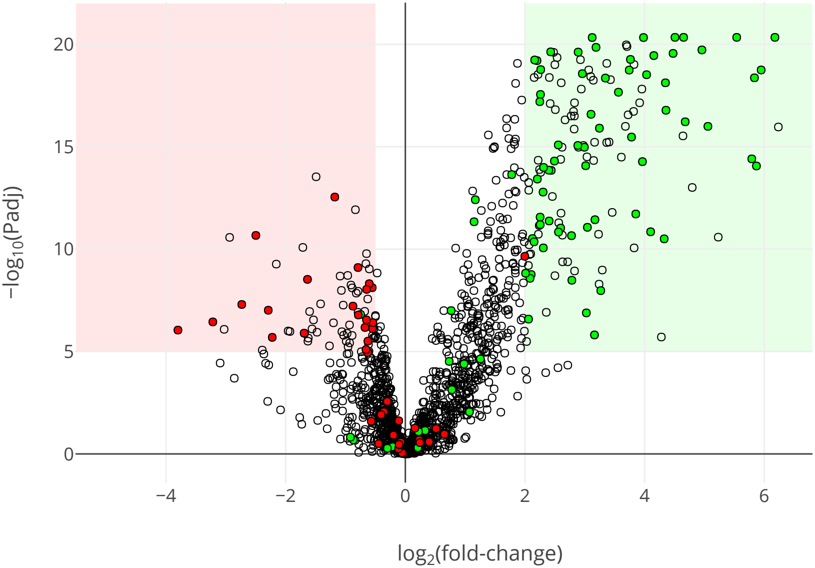
Computational predictions:

- 🟩 92 predicted ⊕
  - 🟩 24 non-trivial
- 🟥 54 predicted ⊖
  - 🟥 33 non-trivial

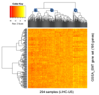
70% more information compared to only knowledge from experiments

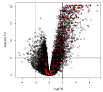


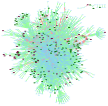
Computational predictions (results of Iggy)

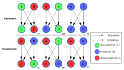


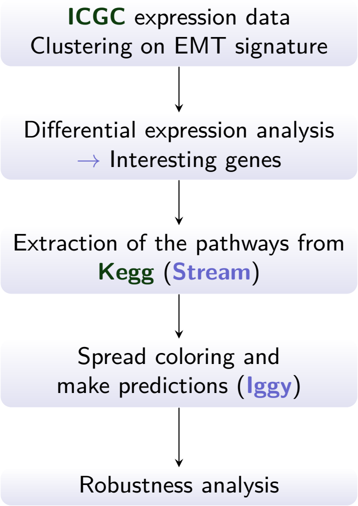
Workflow of the Project











ICGC expression data
Clustering on EMT signature

Differential expression analysis
→ Interesting genes

Extraction of the pathways from
Kegg (**Stream**)

Spread coloring and
make predictions (**Iggy**)

Robustness analysis

2 groups

≃ 2’000 genes

≃ 3’400 nodes, 14’000 edges

Predictions:
92 ⊕ and 54 ⊖

- ⬛ 209 inputs

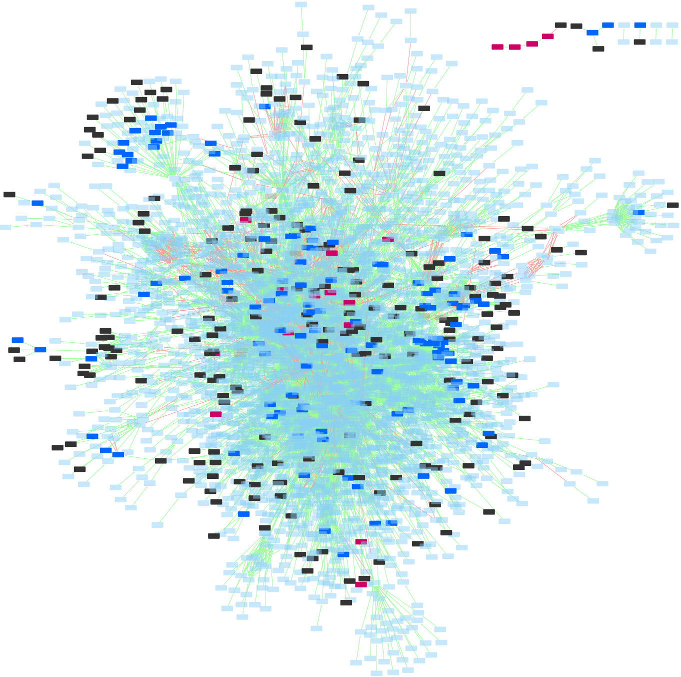
Matching between comp^{al} predictions and ICGC expression data:

- 🟦 124 match
  - 🟦 36 non-trivial
- 🟥 17 do **not** match
  - 🟥 16 non-trivial
- ⬜ 5 not found in ICGC data

88% matching

69% non-trivial

→ Good overlap



Cross-Validation

Sampling

- Consider a range of samplings (10%, 15%, 20%, ... 95%)
- Randomly pick x% of under- and over-expressed genes (observations)
- Compute the predictions on this sample ; repeat 100 times

Score compared to the original data

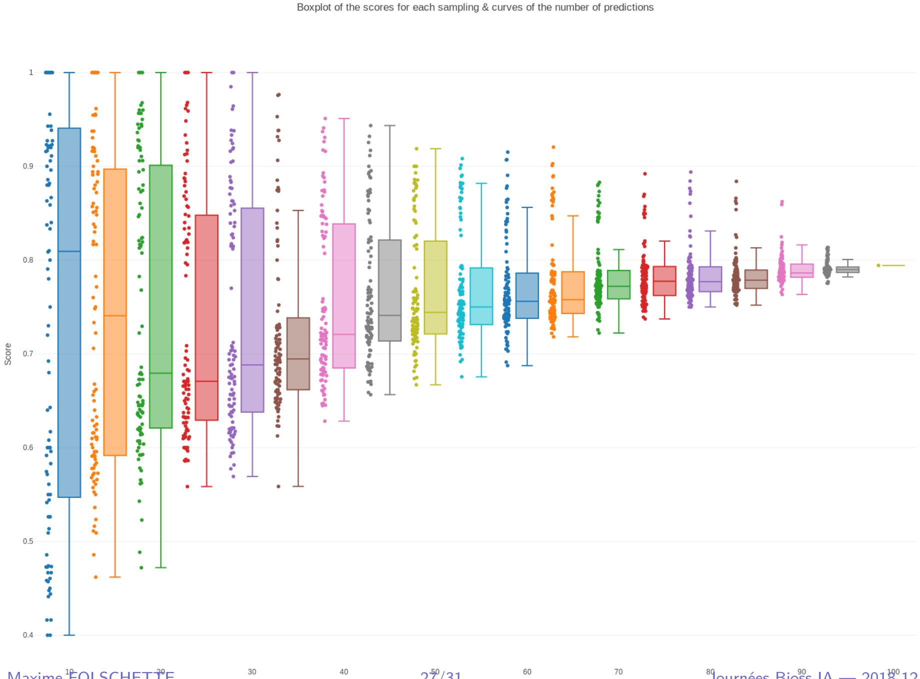
- Compare the predictions to the original ICGC data
- Give a score to each set of predictions

  → Scores converge to the final score at 100%

Robustness of the prediction of each node

- Compare the predictions to the final sampling of 100%

  → Not a lot of variability in the prediction types → Robust


Boxplot of the scores for each sampling & curves of the number of predictions
Score
1
0.9
0.8
0.7
0.6
0.5
0.4
10
20
30
40
50
60
70
80
90
100

Prediction Results

New results compared to ICGC : complexes

Complexes predicted:

- NFKB1::BCL3 (⊕)
- NFKB2::RELB (⊕)
- JUND::NACA (⊖)

Results conflicting with ICGC data

Computational predictions which are different from differential analysis:

- BAK1_gen, BMP4_gen, CREB1_prot, EIF4EBP2_prot, IGFBP3_gen, IGFBP3_prot, NR0B2_gen, NR0B2_prot, NR1H4_gen, NR1H4_prot, NR3C2_gen, NR3C2_prot, SESN3_gen, SESN3_prot, THBS1_gen, TNFRSF10A_gen, TP53_prot

Prediction Results

New results compared to ICGC : complexes

Complexes predicted:

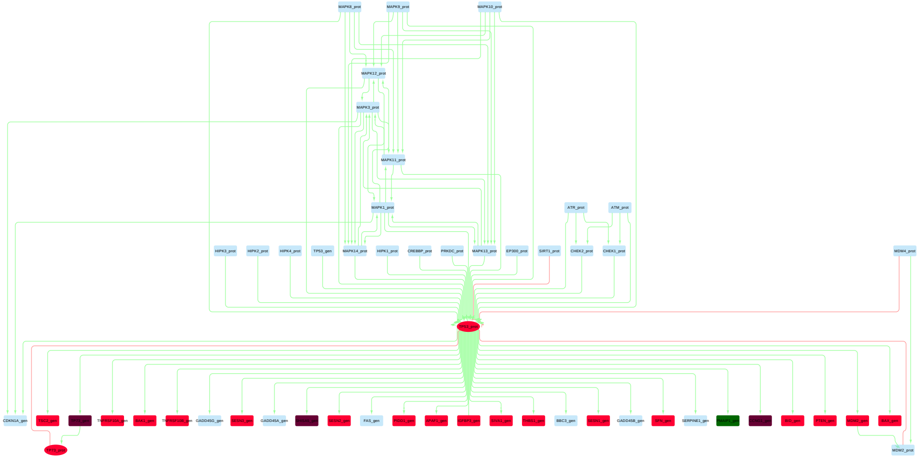
- NFKB1::BCL3 (⊕)
- NFKB2::RELB (⊕)
- JUND::NACA (⊖)

Results conflicting with ICGC data

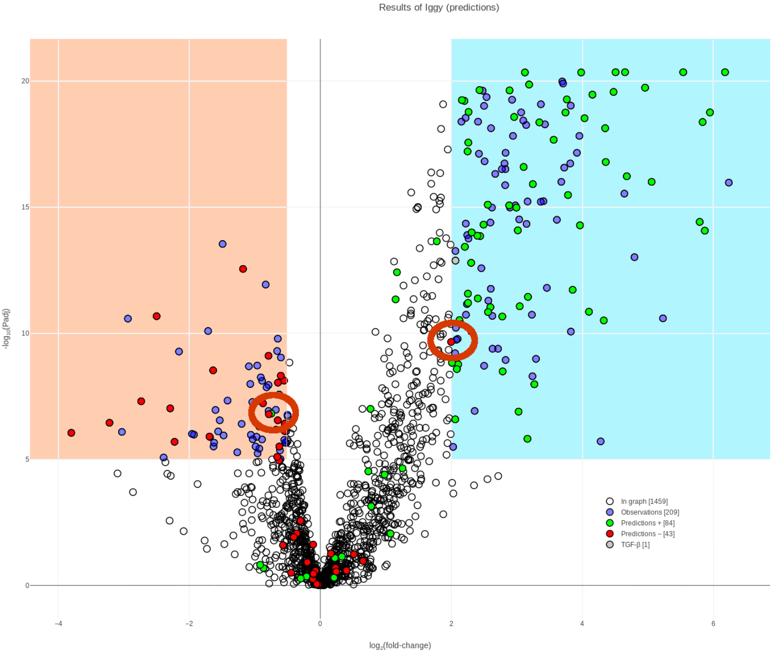
Computational predictions which are different from differential analysis:

- BAK1_gen, BMP4_gen, CREB1_prot, EIF4EBP2_prot, IGFBP3_gen, IGFBP3_prot, NR0B2_gen, NR0B2_prot, NR1H4_gen, NR1H4_prot, NR3C2_gen, NR3C2_prot, SESN3_gen, SESN3_prot, THBS1_gen, TNFRSF10A_gen, **TP53_prot**

Hub example: TP53_prot



18 predictions directly depend of TP53_prot


Results of Iggy (predictions)
-log10(Padj)
log2(fold-change)
20
15
10
5
0
−4
−2
0
2
4
6
○ In graph [1459]
● Observations [209]
● Predictions + [84]
● Predictions − [43]
● TGF-β [1]

Summary & Conclusion

Summary

- Clustering + diff analysis: 2 lists of over- and under-expressed genes
- Graph extracted from Kegg: regulation + signaling
- 146 computational predictions (57 non-trivial)
- Predictions seem robust

Objectives (to do)

- Explore survival curves compared to most robust genes
- Explore the literature regarding predicted complexes
  ⇒ New proliferation signature?
- Try the same workflow on a different type of cancer (breast?)
- PUBLISH

Bibliography I

Cerami, E. G., Gross, B. E., Demir, E., Rodchenkov, I., Babur, Ö., Anwar, N., Schultz, N., Bader, G. D., and Sander, C. (2010).
Pathway Commons, a web resource for biological pathway data.
Nucleic acids research, 39.
http://www.pathwaycommons.org/.

Hudson, T. J. and The International Cancer Genome Consortium (2010).
International network of cancer genome projects.
Nature, 464.
http://icgc.org/.

Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y., and Morishima, K. (2017).
KEGG: new perspectives on genomes, pathways, diseases and drugs.
Nucleic Acids Research, 45(D1):D353–D361.
http://https://www.kegg.jp/.

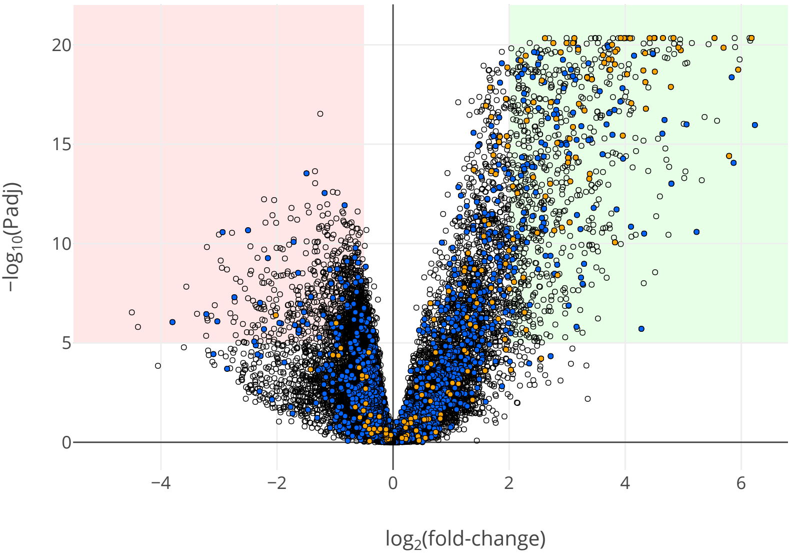
Lefebvre, M., Bourdon, J., Guziolowski, C., and Gaignard, A. (2017).
Regulatory and signaling network assembly through linked open data.
In Journées Ouvertes en Biologie, Informatique et Mathématiques.
Demo paper. https://github.com/symetric-group/bionets-demo.

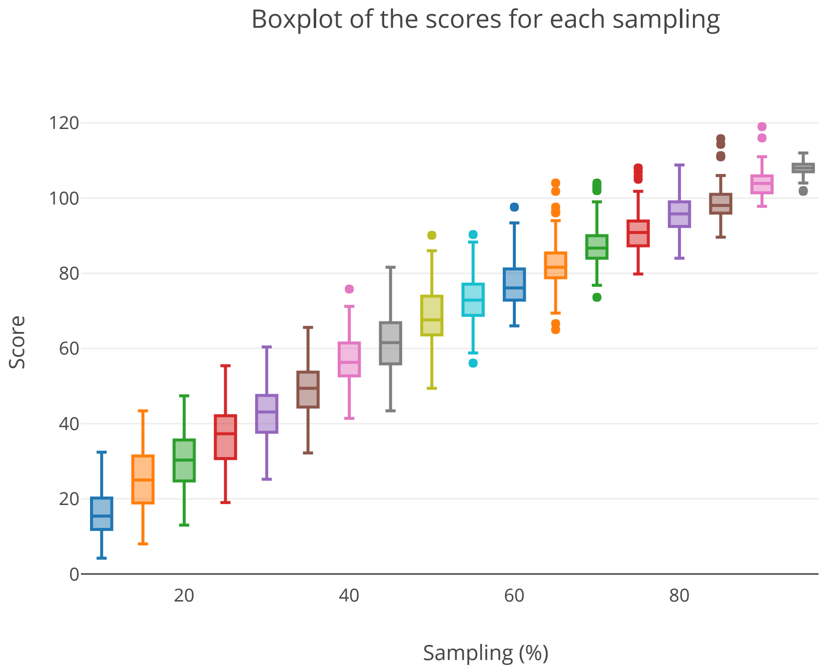
Bibliography II

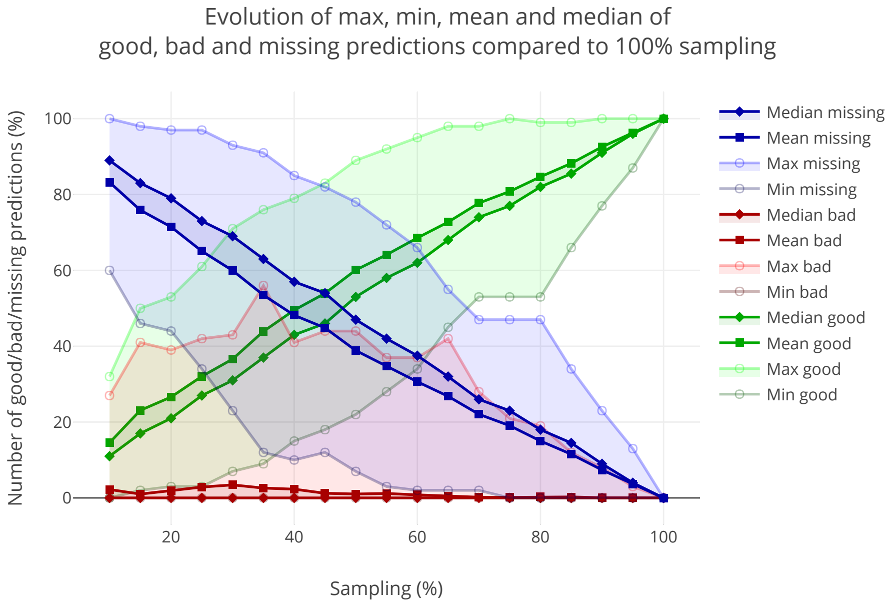
Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., and Mesirov, J. P. (2005).
Gene Set Enrichment Analysis: A knowledge-based approach for interpreting genome-wide expression profiles.
Proc. of the Nat. Ac. of Sci., 102(43).
http://software.broadinstitute.org/gsea/.

Thiele, S., Cerone, L., Saez-Rodriguez, J., Siegel, A., Guziołowski, C., and Klamt, S. (2015).
Extended notions of sign consistency to relate experimental data to signaling and regulatory network topologies.
BMC Bioinformatics, 16(1).
http://bioasp.github.io/iggy/.

Initial ICGC data, EMT signature & genes found in Kegg




Boxplot of the scores for each sampling
120
100
80
60
40
20
0
Score
20
40
60
80
Sampling (%)


Evolution of max, min, mean and median of
good, bad and missing predictions compared to 100% sampling
Number of good/bad/missing predictions (%)
0
20
40
60
80
100
Sampling (%)
20
40
60
80
100
Median missing
Mean missing
Max missing
Min missing
Median bad
Mean bad
Max bad
Min bad
Median good
Mean good
Max good
Min good