Journées Bioss Médecine Personnalisée 2019

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

Current occupation: CNRS & Institut Français de Bioinformatique (IFB) & LS2N
Previous occupation: Université de Rennes 1 & IRISA & IRSET

Co-authors: **MF, Vincent Legagneux, Arnaud Poret, Carito Guziolowski, Nathalie Théret**
Marie Lefebvre, MF, Jérémie Bourdon, Carito Guziolowski, Alban Gaignard

2019-07-01

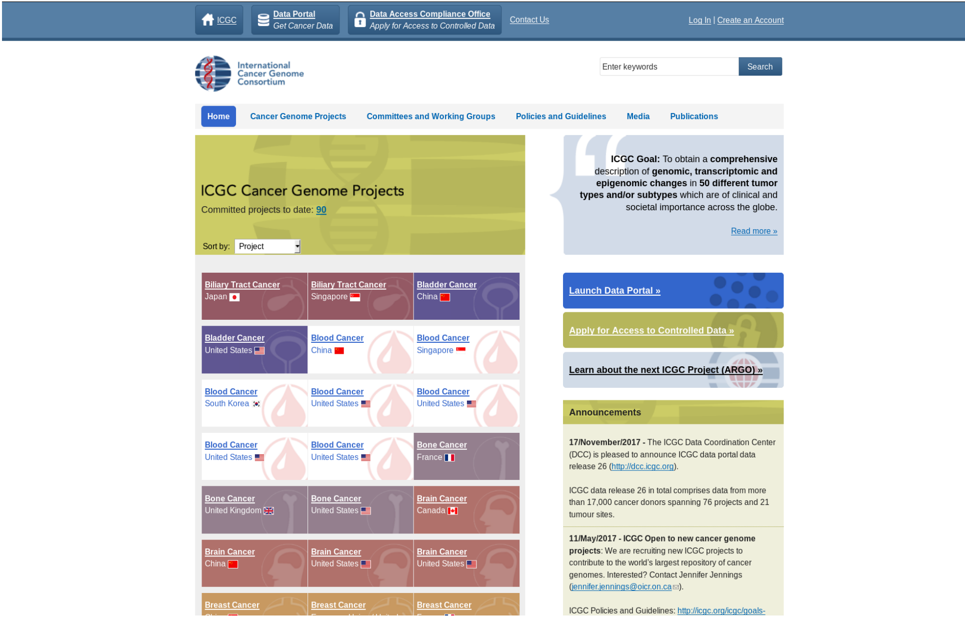
Context

Hepatocellular carcinoma (HCC)

- Most widespread liver cancer, 3rd most deadly cancer
- Mainly associated with chronic inflammation and fibrosis
- Late diagnosis and difficult to treat (resection, transplant, chemo-embolization)
- Very low survival rate

Objectives

- Build gene signaling networks associated with HCC aggressiveness
- Predict key molecules explaining changes in gene expression data between low and high aggressive HCC



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-

LIHC-US

LIHC-US in ICGC [Hudson et al., 2010]

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 for these samples

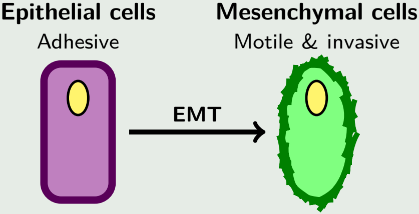
⇒ We need a **criterion** to distinguish aggressive and non-aggressive HCC

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

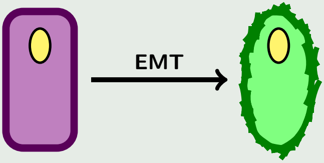
Epithelial-mesenchymal transition (EMT)

Epithelial cells

Adhesive

Mesenchymal cells

Motile & invasive



Low aggressiveness

High aggressiveness

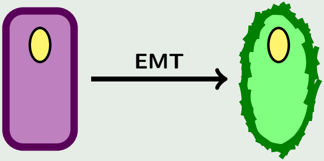
- De-differentiation of epithelial cells to mesenchymal cells
- Gain ability to remodel the extra-cellular matrix and migrate
- Invasive cancer cells ⇒ metastasis
- Indication of **tumor aggressiveness**

Epithelial-Mesenchymal Transition

Epithelial-mesenchymal transition (EMT)

Epithelial cells
Adhesive

Mesenchymal cells
Motile & invasive



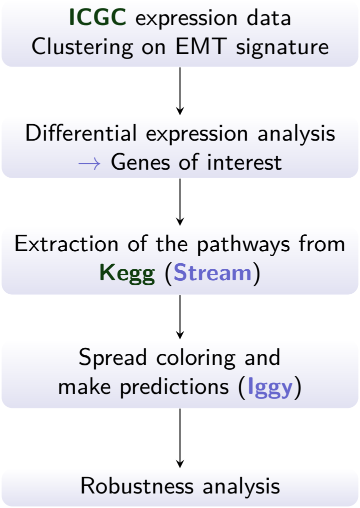
Low aggressiveness

High aggressiveness

- 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 aggressiveness**

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

Workflow of the Project



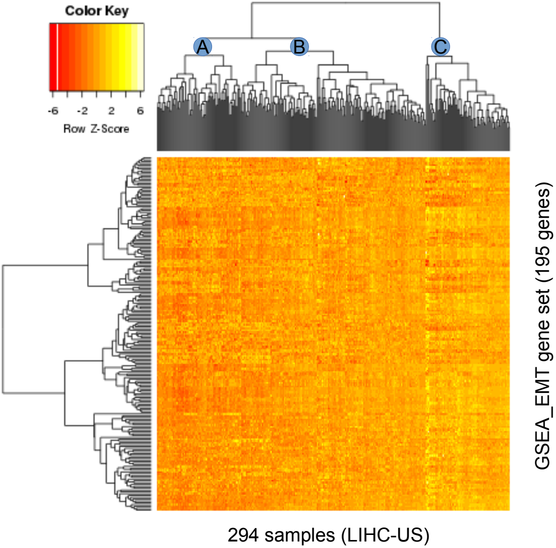
ICGC expression data
Clustering on EMT signature

Differential expression analysis
→ Genes of interest

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

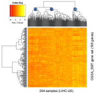
Spread coloring and
make predictions (**Iggy**)

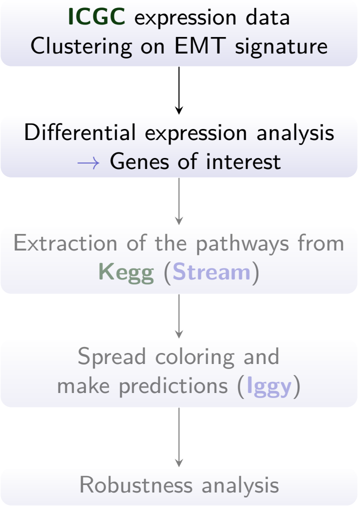
Robustness analysis



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
→ Genes of interest

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

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

Robustness analysis

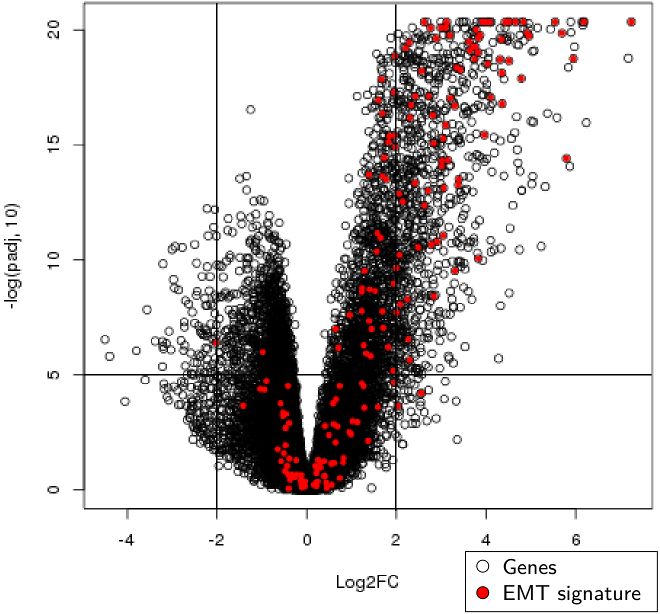
2 groups

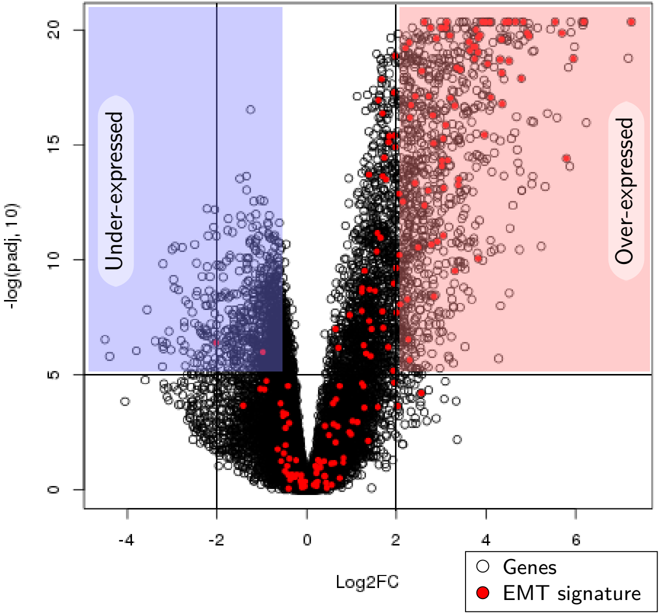
Differential Analysis

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

Genes of Interest

Genes of interest

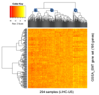
- 821 up-regulated genes
- 1092 down-regulated genes

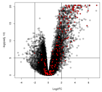
$=$ **1913 genes**

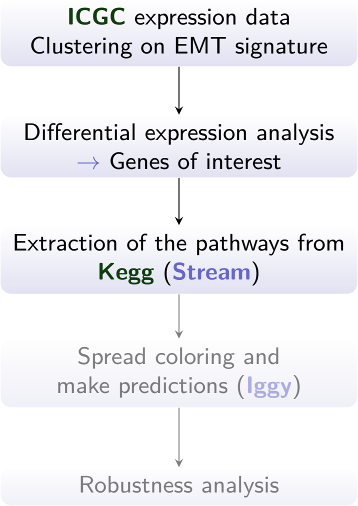
Objectives

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
→ Genes of interest

≃ 2’000 genes

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

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

Robustness analysis

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%)

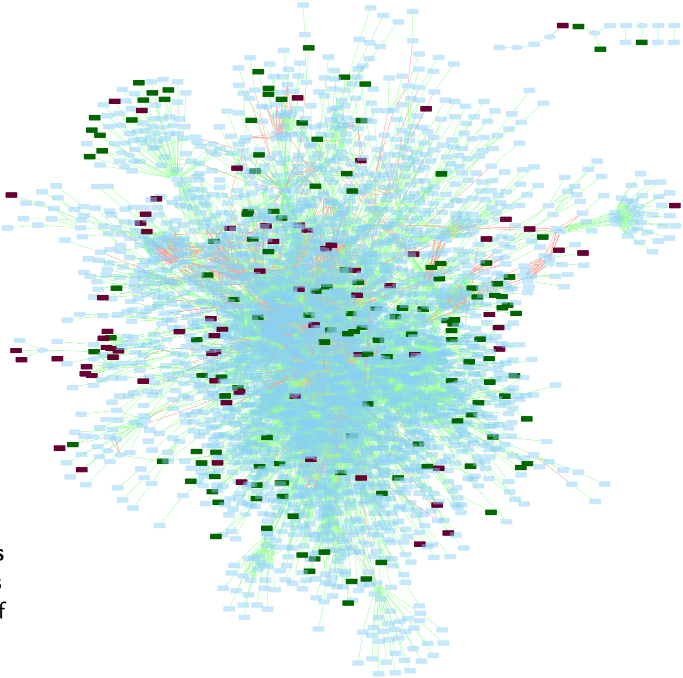
Graph content:

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

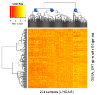
1913 genes from the differential expression **Only 209 are found in Kegg:**

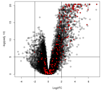
- 🟩 138 up-regulated
- 🟥 71 down-regulated
- 🟦 3174 new nodes

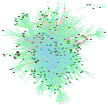
Nodes with up to:
92 incoming influences
79 outgoing influences
→ Nodes with a lot of impact on the network

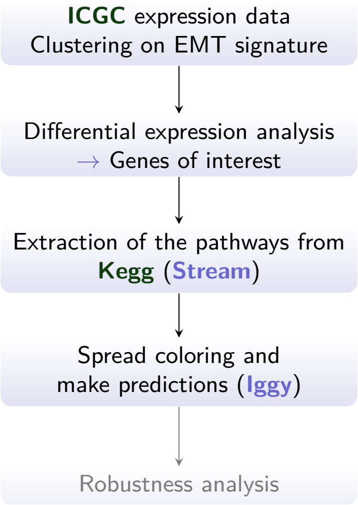


Workflow of the Project









ICGC expression data
Clustering on EMT signature

Differential expression analysis
→ Genes of interest

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

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

Robustness analysis

2 groups

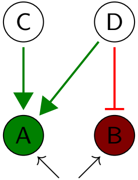
≃ 2’000 genes

≃ 3’400 nodes, 14’000 edges

Graph Coloring

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

X = over-expressed   Y = under-expressed

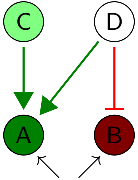


Given by the experimental data

Graph Coloring

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

X = over-expressed   Y = under-expressed

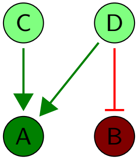


Given by the experimental data

Graph Coloring

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

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

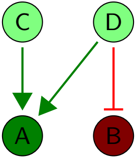


Consistent

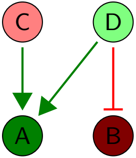
Graph Coloring

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

X = over-expressed   Y = under-expressed



Consistent

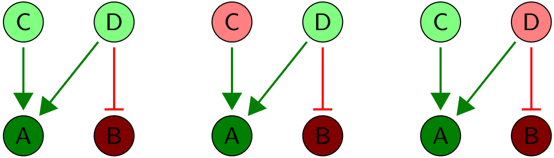


Consistent

Graph Coloring

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

X = over-expressed Y = under-expressed

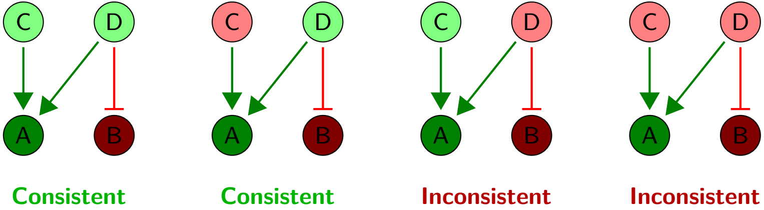


Consistent **Consistent** **Inconsistent**

Graph Coloring

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

X = over-expressed Y = under-expressed

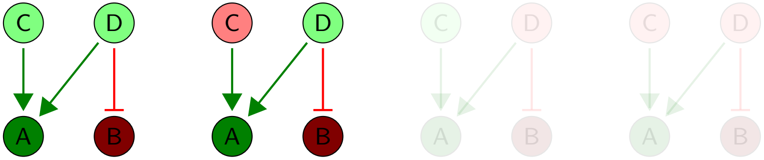


Consistent **Consistent** **Inconsistent** **Inconsistent**

Graph Coloring

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

X = over-expressed Y = under-expressed



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)

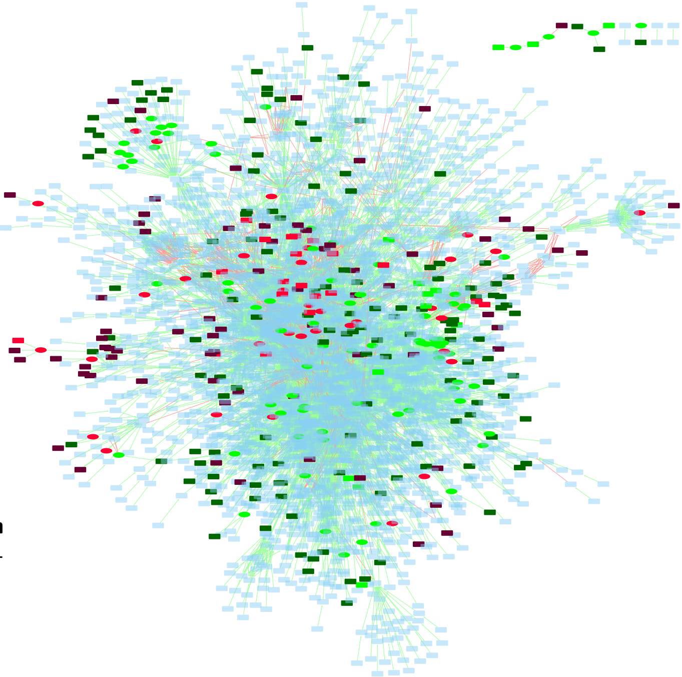
Knowledge from experiments:

- 🟩 138 up-regulated
- 🟥 71 down-regulated

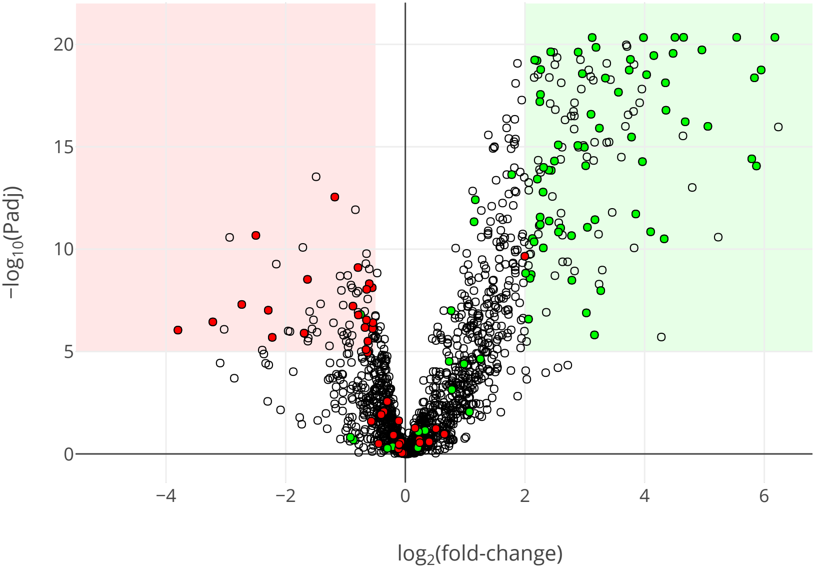
Computational predictions:

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

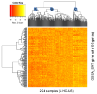
70% more information compared to only knowledge from experiments



Computational predictions (results of Iggy)

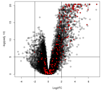


Workflow of the Project



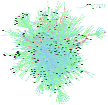
ICGC expression data
Clustering on EMT signature

2 groups



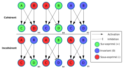
Differential expression analysis
→ Genes of interest

≃ 2’000 genes



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

≃ 3’400 nodes, 14’000 edges



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

Predictions:
92 ⊕ and 54 ⊖

Robustness analysis

- ⬛ 209 inputs

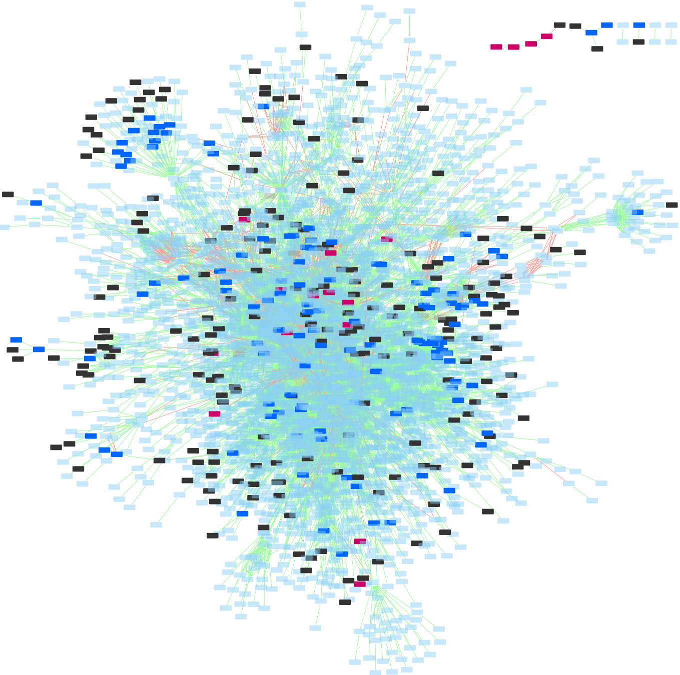
Matching between comp$^{\text{al}}$ predictions and ICGC expression data:

- 🟦 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

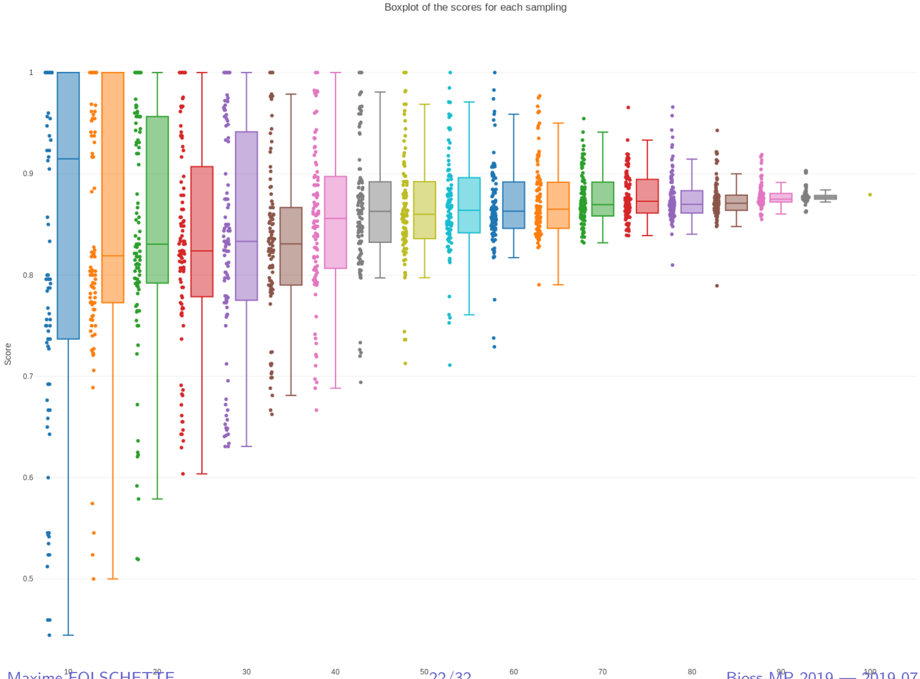
- Compare the predictions to the original ICGC data

  → Scores converge to the final score at 100%

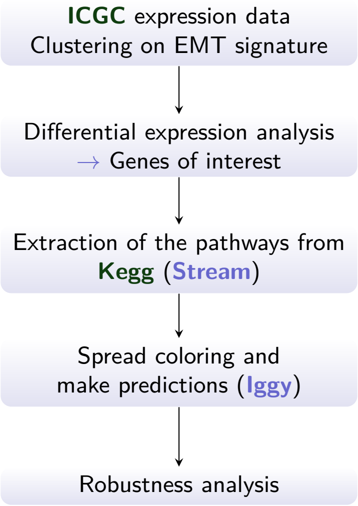
Robustness of the prediction of each node

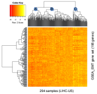
- Compare the predictions to the final sampling of 100%

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


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

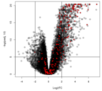
Workflow of the Project





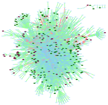
ICGC expression data
Clustering on EMT signature

2 groups



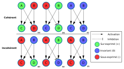
Differential expression analysis
→ Genes of interest

≃ 2’000 genes



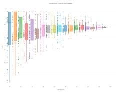
Extraction of the pathways from **Kegg** (**Stream**)

≃ 3’400 nodes, 14’000 edges



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

Predictions:
92 ⊕ and 54 ⊖



Robustness analysis

Robust predictions

Prediction Results

New results compared to ICGC : complexes

Complexes predicted:

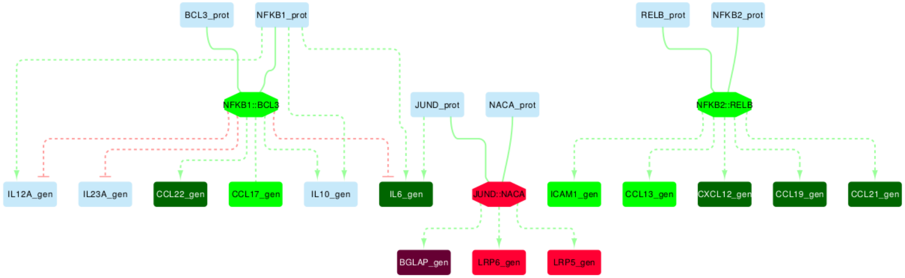
- NFKB1::BCL3 (⊕)
- NFKB2::RELB (⊕)
- JUND::NACA (⊖)

Results conflicting with ICGC data

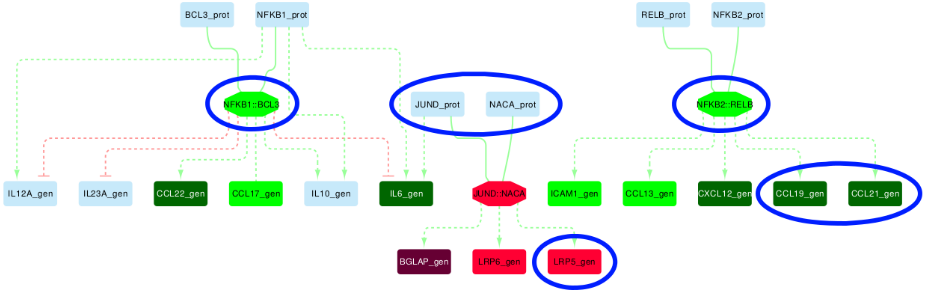
Computational predictions which are **different from differential analysis**:

- 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

Biological Validation

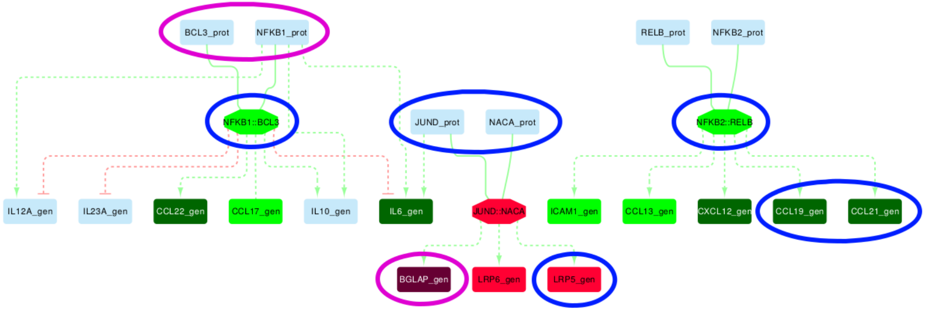


Biological Validation



○ Validated by experimental data

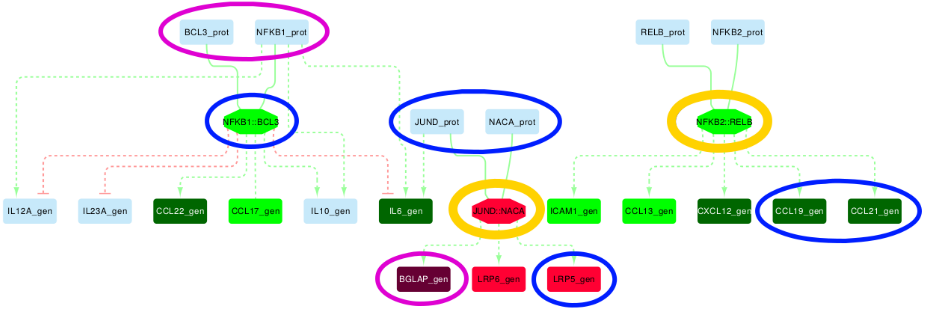
Biological Validation



○ Validated by experimental data

○ Link with cancer validated by literature

Biological Validation



○ Validated by experimental data

○ Link with cancer validated by literature

○ New knowledge in aggressive HCC:

- up-regulation of NFKB2::RELB
- down-regulation of JUND::PACA

Using Pathway Commons

Pathway Commons

Pathway Commons [Cerami et al., 2010]

- A gathering of **25 pathway databases**
- Contains: PID, Kegg, Reactome, CTD, Panther, ...

Benefits

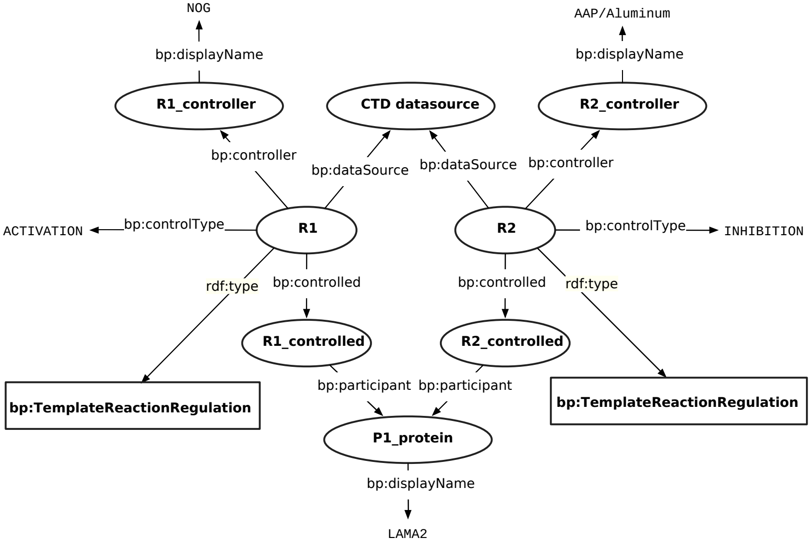
- Common ontology (BioPAX)
- Freely available *via* a SPARQL endpoint (programmatic queries)
- Much more data than Kegg alone ⇒ better coverage

Drawbacks

- **Very heterogeneous data**, duplicated entities
- The BioPAX ontology is big and difficult to use
- Made for biologists rather that modelers or computer scientists

→ Could Pathway Commons be used instead of Kegg?

The BioPAX Ontology



BRAvo

BRAvo [Lefebvre et al., 2017]

- **Interrogates** Pathway Commons with SPARQL queries
- Upstream graph reconstruction
- Written in Python; available soon (open source)

Benefits of BRAvo

- Fast reconstruction: 10 mins for a graph with 1402 nodes
- Unification of duplicated nodes based on synonyms
- Regulation or signaling
- Source selection
- SIF output

910 genes of interest
Regulation graph

Graph content:

- 1’402 nodes
- 2’804 edges

641/910 genes found

Computational predictions:

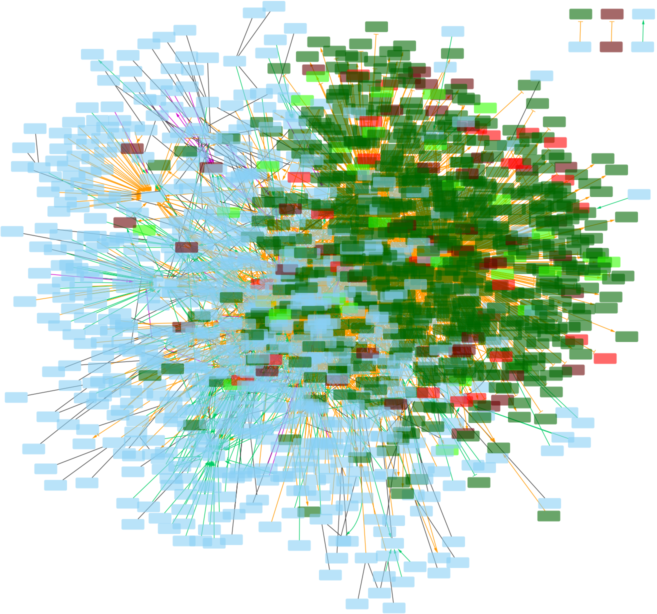
- 🟩 40 predicted ⊕
- 🟥 32 predicted ⊖

12% more information

Using additional synonyms unification:

- 🟦 26 match
- 🟪 28 do **not** match

47% matching



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)
- Computational & biological validations

Ongoing work

- General pipeline of the whole method
- Try other sources (Pathway Commons with BRAvo)
- Finalizing manuscripts; submission soon

Other objectives (to do)

- Explore survival curves compared to most robust genes
- Search for proliferation signatures
- Try the same workflow on a different type of cancer (breast?)

Acknowledgments

Hepatocellular carcinoma computational models identify key protein-complexes associated to tumor progression

Vincent Legagneux: Inserm, Irset (Rennes)
Arnaud Poret: LS2N (Nantes)
Carito Guziolowski: École centrale de Nantes, LS2N (Nantes)
Nathalie Théret: Inserm, Irset, IRISA (Rennes)

Special thanks to **Anne Siegel**

BRAvo: Regulatory and signaling network assembly from Pathway Commons

Marie Lefebvre: INRA (Bordeaux)
Jérémie Bourdon: Université de Nantes, LS2N (Nantes)
Carito Guziolowski: École centrale de Nantes, LS2N (Nantes)
Alban Gaignard: CHU de Nantes, Institut du Thorax, LS2N (Nantes)

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 et al. (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/.

Biological Validation

NFKB1::BCL3 (⊕)

Part of the NFκB pathway

- Bcl3 alters NFκB signaling pathways, **associated with cancer**
- NFKB1 (p105) is activated to p50 ⇒ **early recurrence of HCC**
- Increased p50 and Bcl3 **reported in tumors**

NFKB2::RELB (⊕)

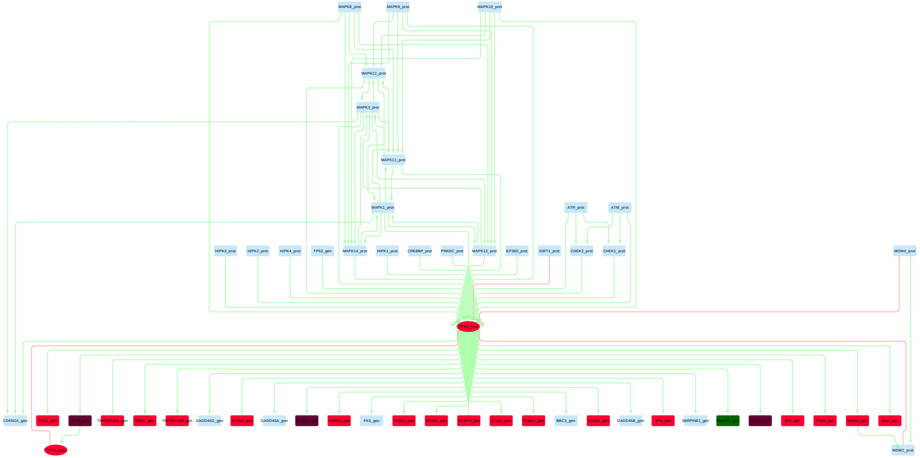
Activates CCL19 and CCL21

- Increase of CCL19 and CCL21 validated in experimental data

JUND::NACA (⊖)

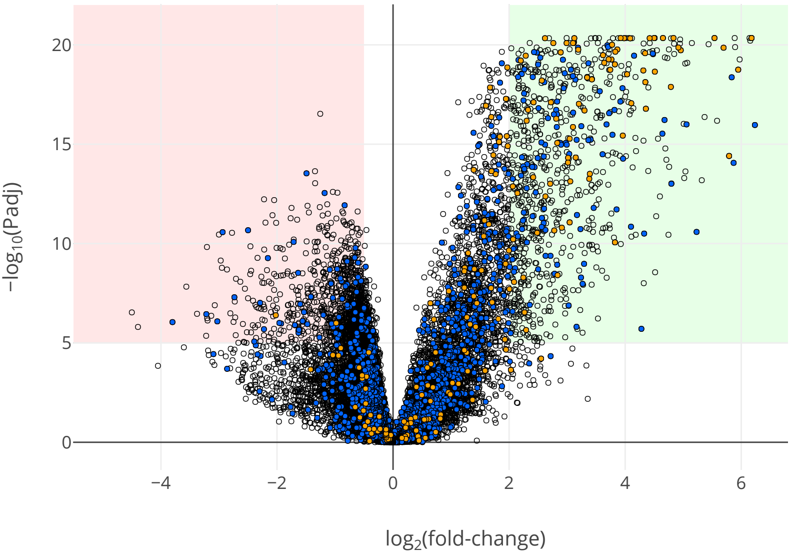
- Regulates BGLAP (osteocalcin), **down-regulated in the serum of HCC patients**
- Regulates LRP5, validated in experimental data (decrease)
- Decrease of JUND and NACA validated in experimental data

Hub example: TP53_prot

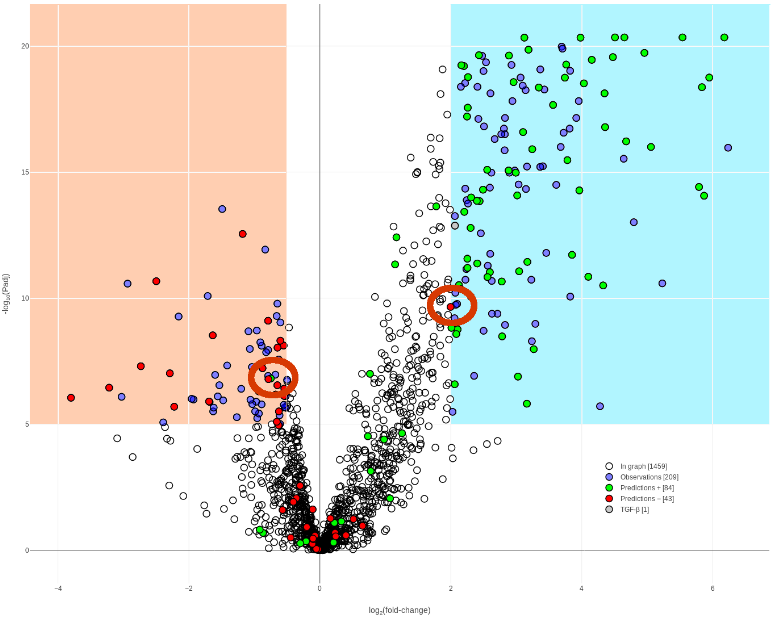


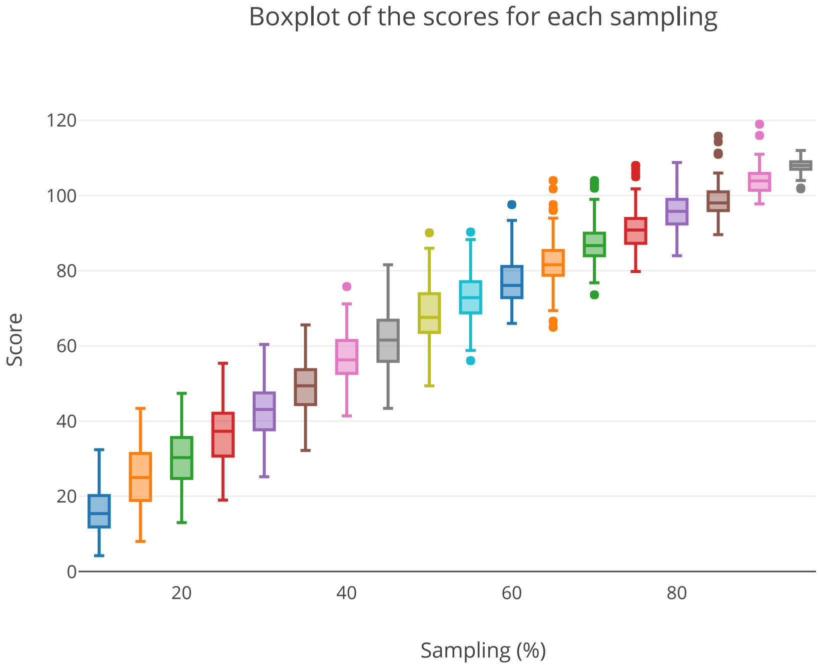
18 predictions directly depend of TP53_prot

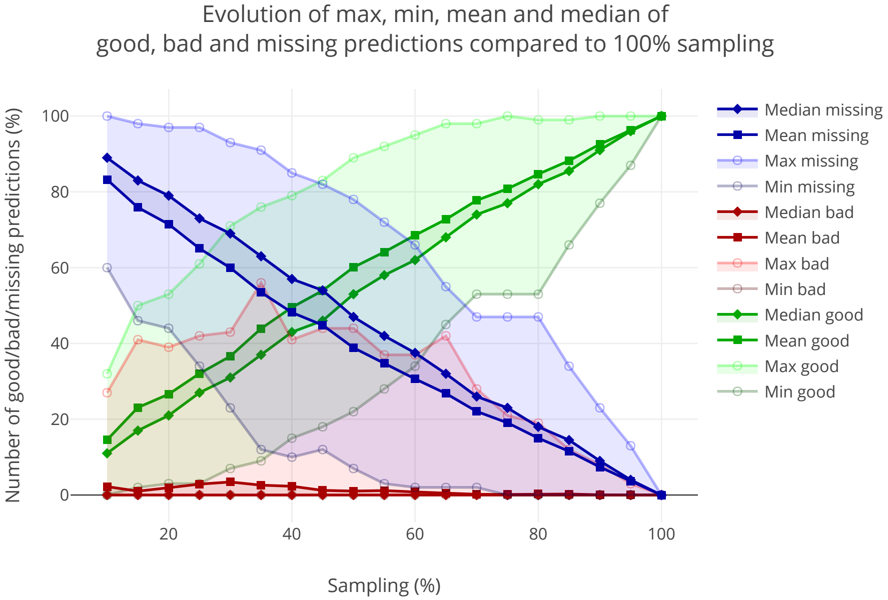
Initial ICGC data, EMT signature & genes found in Kegg



Results of Iggy (predictions)



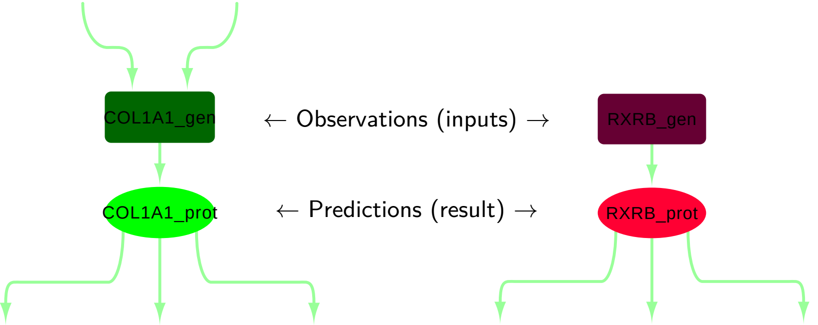

Boxplot of the scores for each sampling
Score
0
20
40
60
80
100
120
Sampling (%)
20
40
60
80


Evolution of max, min, mean and median of
good, bad and missing predictions compared to 100% sampling

Trivial Predictions

"Trivial" prediction

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



← Observations (inputs) →

← Predictions (result) →