Hackathon 2018

Background..

Data sets

We have two collections of data sets, one for melanoma tumors and one for lung cancer tumors. We will focus the melanoma tumors. Although we will produce melanoma-specific results, the underlying methods should be applicable to the lung cancer data set as the data schema's are identical.

For both the melanoma and the lung cancer data set we have:

- DNA, mutations: genomics
- DNA, Copy number variations: genomics
- DNA, Methylation: epigenomics
- RNA, genomic expression: epigenomics
- RNA, miRNA: transcriptomics
- Proteins: proteomics

DNA, Mutation

Literally, per genome and chromosome the change in the pair compared to a normal reference. Remember we (Adenine,Thymine) and (Guanine,Cytosine) as the base pairs.

The types of mutations include (taken from here):

Missense mutation: This type of mutation is a change in one DNA base pair that results in the substitution of c for another in the protein made by a gene.

Nonsense mutation: is also a change in one DNA base pair. Instead of substituting one amino acid for anothe altered DNA sequence prematurely signals the cell to stop building a protein. This type of mutation results in a protein that may function improperly or not at all.

Insertion: An insertion changes the number of DNA bases in a gene by adding a piece of DNA. As a result, the by the gene may not function properly.

Deletion: A deletion changes the number of DNA bases by removing a piece of DNA. Small deletions may rem few base pairs within a gene, while larger deletions can remove an entire gene or several neighboring genes. T DNA may alter the function of the resulting protein(s).

Duplication: A duplication consists of a piece of DNA that is abnormally copied one or more times. This type o alter the function of the resulting protein.

Frameshift mutation: This type of mutation occurs when the addition or loss of DNA bases changes a gene's A reading frame consists of groups of 3 bases that each code for one amino acid. A frameshift mutation shifts t

these bases and changes the code for amino acids. The resulting protein is usually nonfunctional. Insertions, d duplications can all be frameshift mutations.

Repeat expansion: Nucleotide repeats are short DNA sequences that are repeated a number of times in a row a trinucleotide repeat is made up of 3-base-pair sequences, and a tetranucleotide repeat is made up of 4-base- sequences. A repeat expansion is a mutation that increases the number of times that the short DNA sequence This type of mutation can cause the resulting protein to function improperly.

DATA FIELDS, shape (422553, 11)

```
ID | Location | Change | Gene | Mutation type| Var.Allele.Frequency | Amino acid

SampleID,| Chr, Start, Stop| Ref, Alt | Gene | Effect | DNA_VAF, RNA_VAF | Amino_Acid_Change

string |string, int, int | char, char | string | string | float, float | string
```

NOTE: this gives us direct insight in how genetic mutations lead to changes in amino-acids.

Copy Number Variations

A copy number variation (CNV) is when the number of copies of a particular gene varies from one individual to

DATA FIELDS, shape (24802, 372)

```
Gene | Chr, Start, Stop | Strand | SampleID 1..SampleID N

string |string, int, int | int | int..int
```

Methylation, gene expression regulation

Degree of methylation indicates addition of Methyl groups to the DNA. Increased methylation is associated with transcription of the DNA: Methylated means the gene is switched OFF, Unmethylated means the gene is switch

Alterations of DNA methylation have been recognized as an important component of cancer development.

DATA FIELDS, shape (485577, 483)

```
probeID | Chr, Start, Stop | Strand | Gene | Relation_CpG_island | SampleID 1..SampleID N

string |string, int, int | int | string | string | float..float
```

RNA, gene expression

Again four building blocks; Adenosine (A), Uracil (U), Guanine (G), Cytosine (C).

(DNA) --> (RNA)

A --> U

T --> A

C --> G

G --> C

Gene expression profiles, continuous values resulting from the normalisation of counts.

DATA FIELDS, shape (60531, 477)

```
Gene | Chr, Start, Stop | Strand | SampleID 1..SampleID N

string |string, int, int | int | float..float
```

miRNA, transcriptomics

The connection between the RNA production and protein creation. I.e. perhaps miRNA expression values can l with specific proteins.

DATA FIELDS, shape (2220, 458)

```
MIMATID | Name | Chr, Start, Stop | Strand | SampleID 1..SampleID N

string | string |string, int, int | int | float..float
```

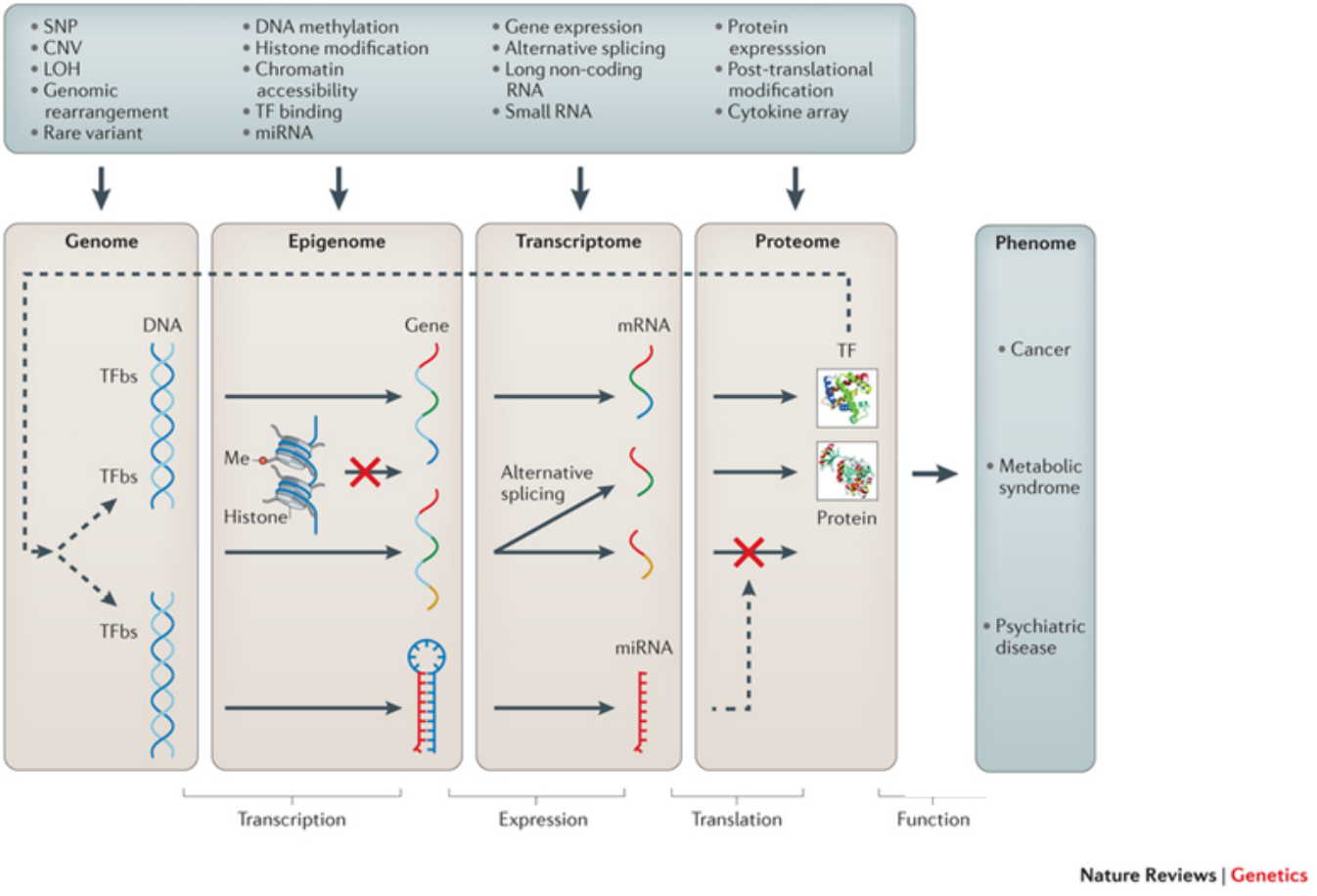
Proteomes

Proteine expression profiles, ditto, continuous values resulting from the normalisation of counts

DATA FIELDS, shape (282, 355)

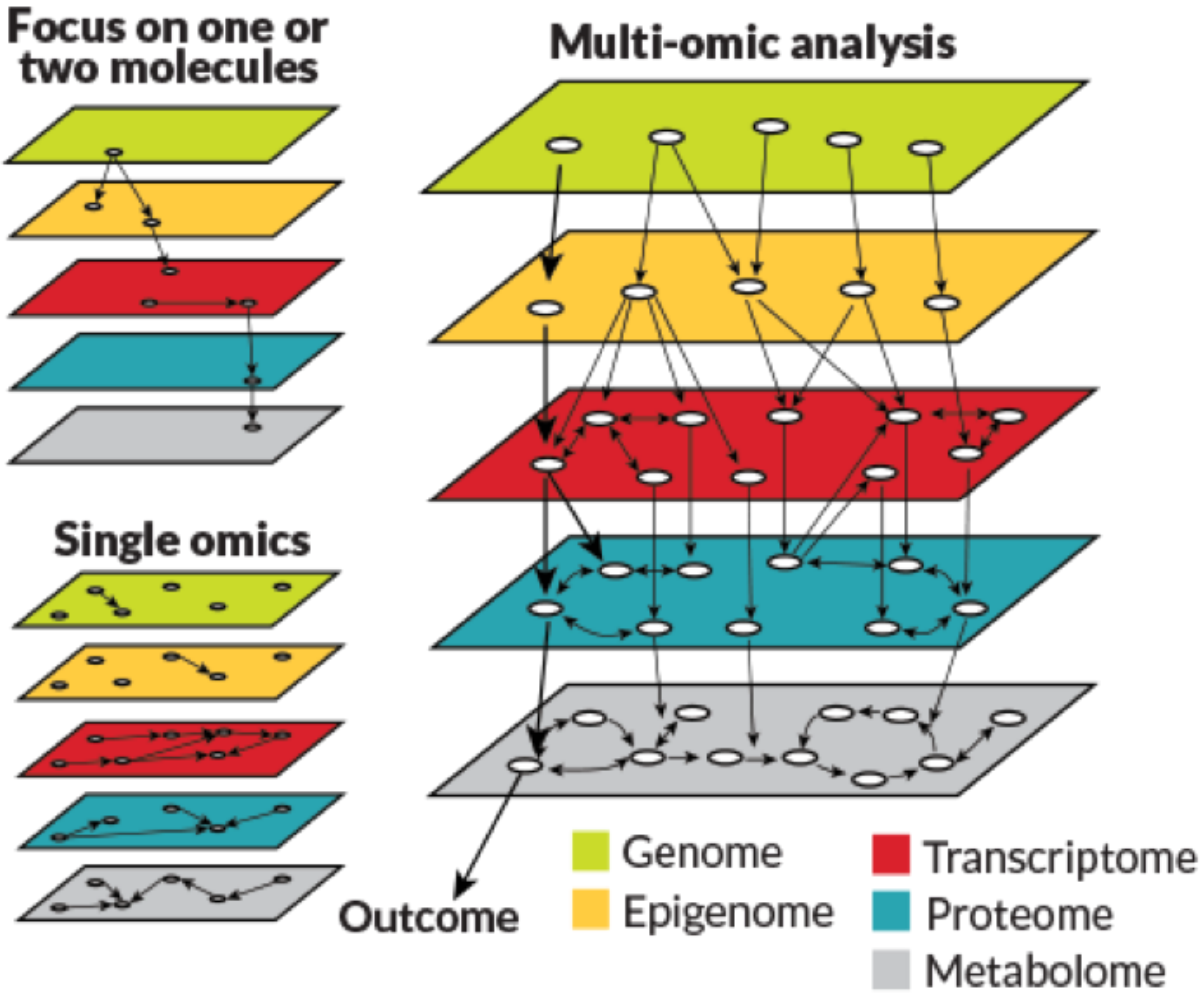
```
ProteinID | SampleID 1..SampleID N

string | float..float
```

QUIZ, identify our data sets in the following image!



GOAL

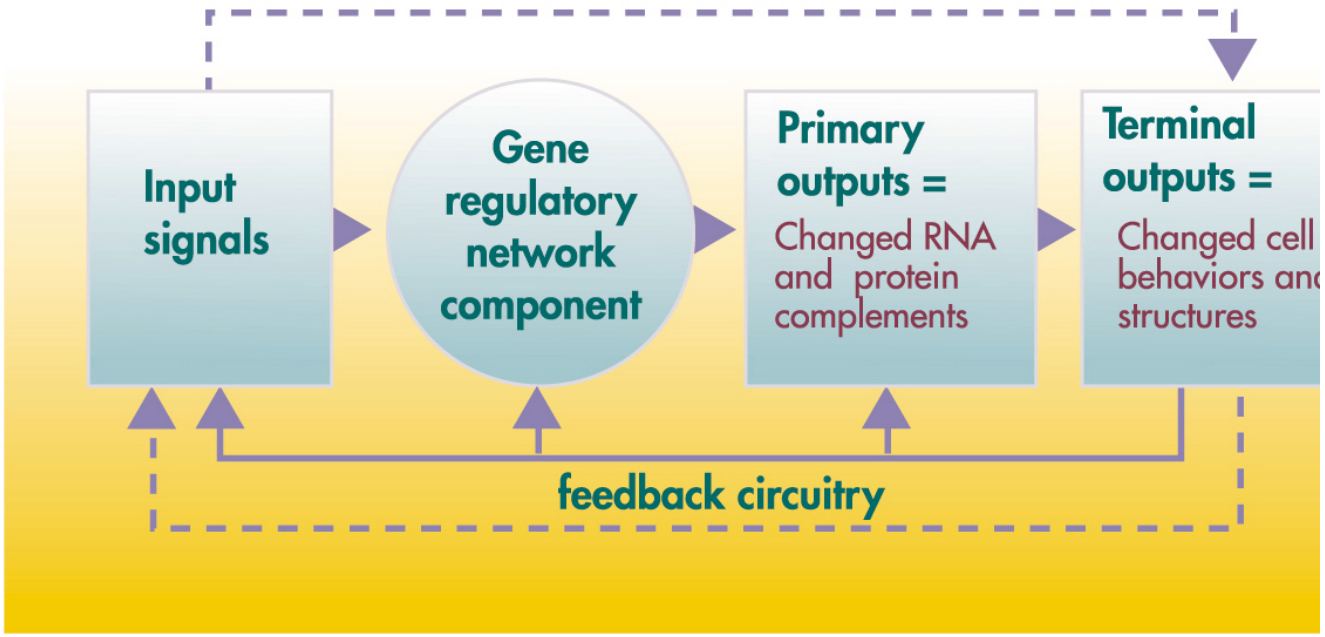
Some degree of multi-omic analysis and identification of pathways.



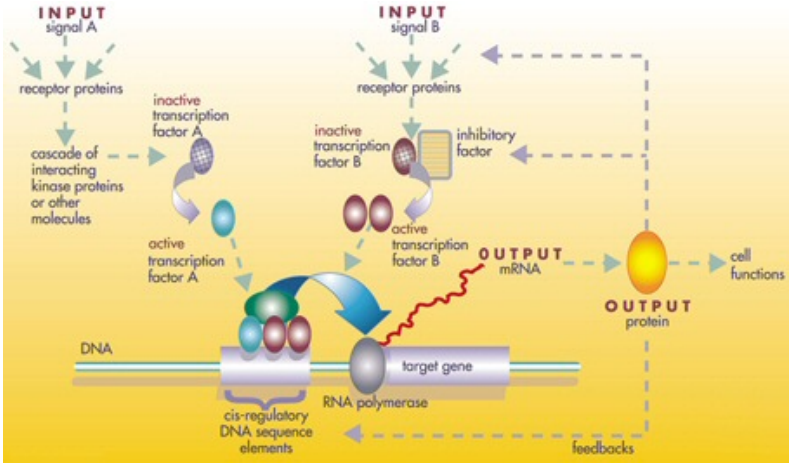
Targets

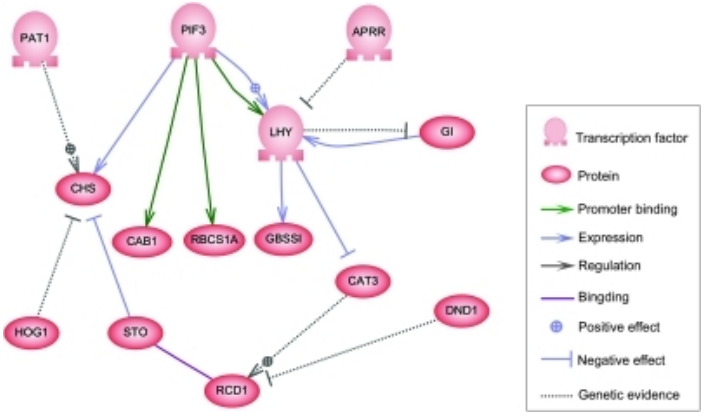
First we need get a picture of what a "signature" actually means in this context. We basically have hierarchicall data with "pathways" going through those layers, those pathways are connected by mutations on the one end (proteins on the other end. How to find those pathways is the main question, because once we can do that, we identify either which pathways are typical for people that do or do not respond well to immunotherapy, or what p pathway is typically different for those patients.

So what is a pathway? A pathway is a chain of molecular changes that leads to, in our case, the production of proteines, or (since we don't have many proteomic measurements) certain RNA codes. In the simplest form it is but it is more likely similar to a bi-directed graph, in it's simplest form; DNA mutation <--> RNA <--> mRNA <--> The collection of molecular regulators that govern the genomic expression levels of mRNA and proteins is calle regulatory network (GRN). So, instead of chain, it is better to say a network of molecular changes.



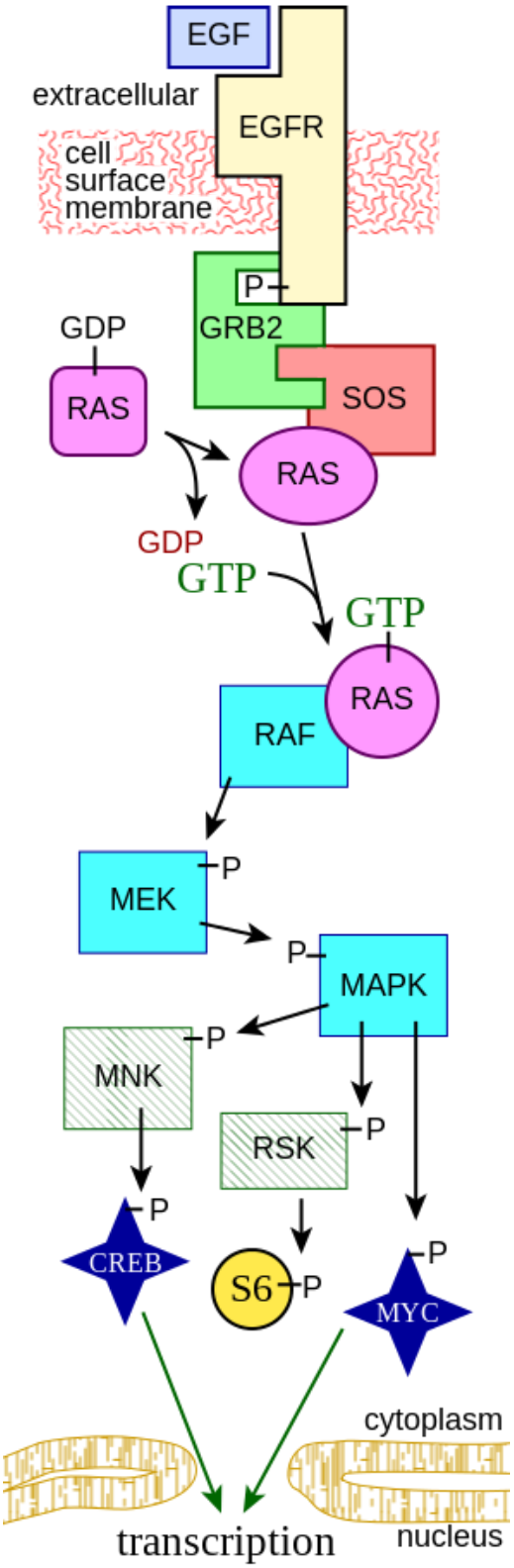
YGG 01-0086



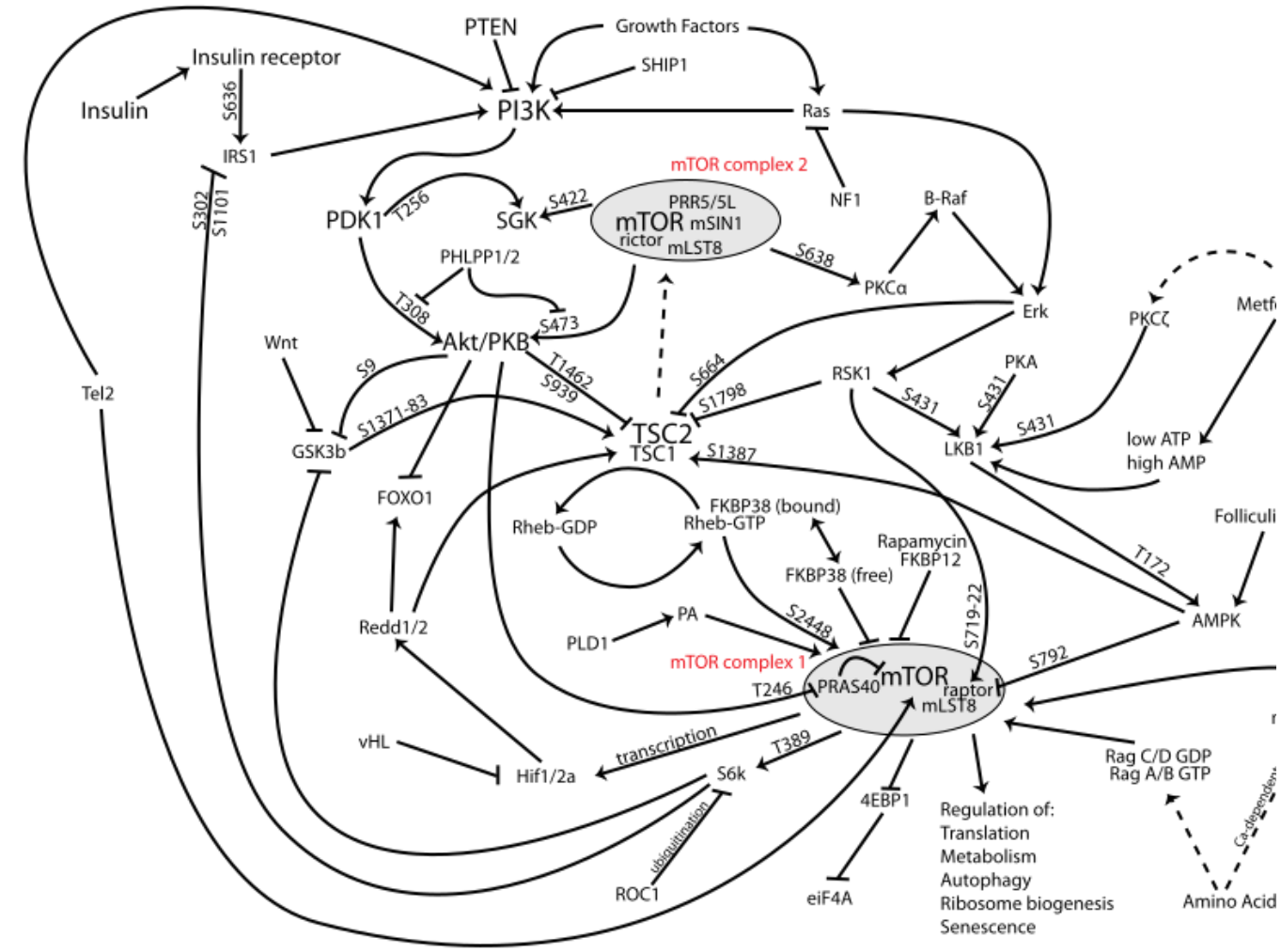


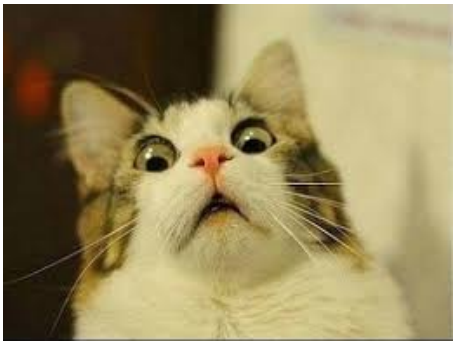
Specifically, given that we find a pathway, the genomes that, given a mutation, will lead to a proto-oncogenic or effect on tumor development are denoted as proto-oncogenics and inhibitors, the former promotes tumor growt slows it down.

One such pathway is the Mitogen Activated Protein Kinase (MAPK) pathway. This pathway connects certain m BRAF) oncogene (i.e. DNA) to the generation of certain proteins that lead to the promotion of cell growth. This pathway looks as follows:


EGF
extracellular
EGFR
cell
surface
membrane
P
GRB2
GDP
RAS
SOS
RAS
GDP
GTP
GTP
RAS
RAF
MEK
P
P
MAPK
P
MNK
P
RSK
P
CREB
S6
P
P
MYC
cytoplasm
transcription
nucleus

Just to show the complexity, the PI3K/AKT/mTOR pathway looks as follows:





Don't worry, this is not expected from us..although I can produce this stuff in paint, hands down (I can actually u draw..). Anyways, back to reality: we only have a few thousand protein measurements, and we do not have any data so it is practically impossible to extract any feedback effects from downstream changes. Ooofff, that leaves down approach, from instruction/mutation to RNA and in some cases, proteins. Finding any feedback effect is s perhaps for continued work after the hackathon.

Now, given such a pathway we can frustrate the signal anywhere on the chain, as long as it prevents the cell gr stimulation.

We should keep in mind that the inhibitors/proto-oncogenes are likely specific to the type of melanoma's, we dis least the following by their genomic mutations:

- (proto-oncogenes) BRAF wild-type
- Triple Wild-type
- NF1

- KIT
- MITF
- RAS
- (inhibitor) PD-1/PD-L1

The current inhibitor, the one they likely use in the immunotherapy is PD-L1. It is called a checkpoint inhibitor, r from our models would be a good validator. In fact, PD-L1 itself is a proto-oncogene, but apparently it can be in instead of searching for inhibitors, we should be looking for proto-oncogenes.

The main questions:

- Why do some patients respond to immunotherapy and others not? --> can we predict who will not respond
- What are the pathways related to melanoma (firstly), to immunotherapy response (secondly)

"Official" supporting questions:

- Can you show and visualize the correlations and concepts between the different datasets?
- As melanoma is a set of diverse diseases, can you stratify the patients based on all the data in to subgrou
- Can you integrate all the data to make more accurate predictions for each patient than you would by only l data source?
- Can you select a list of most informational variables that drive the predictions?
- Can you select a list of most informational variables distinctive for each patient subgroup?
- Can you identify a signature based on an integrative approach that can predict response to immunotherap
- Can you identify a signature that correlates with the prognosis of immunotherapy?

Basic hypotheses that would be nice to confirm (i.e. nice to haves, feel free to ignore)

- T(tumor), increased Bresow-thickness correlates with more malignancy (Tis, T1a/b, T2a/b, T3a/b, T4a/b), i survival rate
- N(nodal stage), Local spread correlates with more malignancy (N0, N1a/b, N2a/b/c, N3)
- M(metastasis location), distant metastasis (beyond regional lymph nodes) corresponds with mmore malign
- BRAF proto-oncogenic mutations should occcur in about 50% of all cutaneous melanomas.
- We should be able to identify 4 subtypes of cutaneous melanomas: BRAS/RAS(N/H/K)/NF1/Triple-WT
- order 3 clusters in the mRNA profiles of the most variant genes (keratin, immune, MITF-low)
- inhibitor: PD-L1/PD-1, our method should be able to retrieve this specific mutation as an inhibitor
- inhibitor: MEK, our method should be able to retrieve this specific mutation as an inhibitor for BRAF wild-ty mutant melanoma's
- inhibitor: PTEN/TP53/APC, our method should be able to retrieve this specific mutation as an inhibitor
- proto-oncogenic: BRAF, our method should be able to retrieve this specific mutation as a proto-oncogene
- LCK protein expression: correlates positively with patient survival
- genetic markers for melatonine may be proxy for higher risk of melanoma

Basic information:

- Moles spatially near eachother, in combination with discolouring is indicative for a higher likelihood of meta
- (Breslow depth) under skin correlates with more malignancy
- RNA, each 3 letter combo is associated with a specific amino-acid --> an amino-acid is associated with mu combinations --> i.e. not every change in RNA coding leads to a delta amino acid.
- mutations may lead to a change in protein function, and a change in genetic expression of other genes

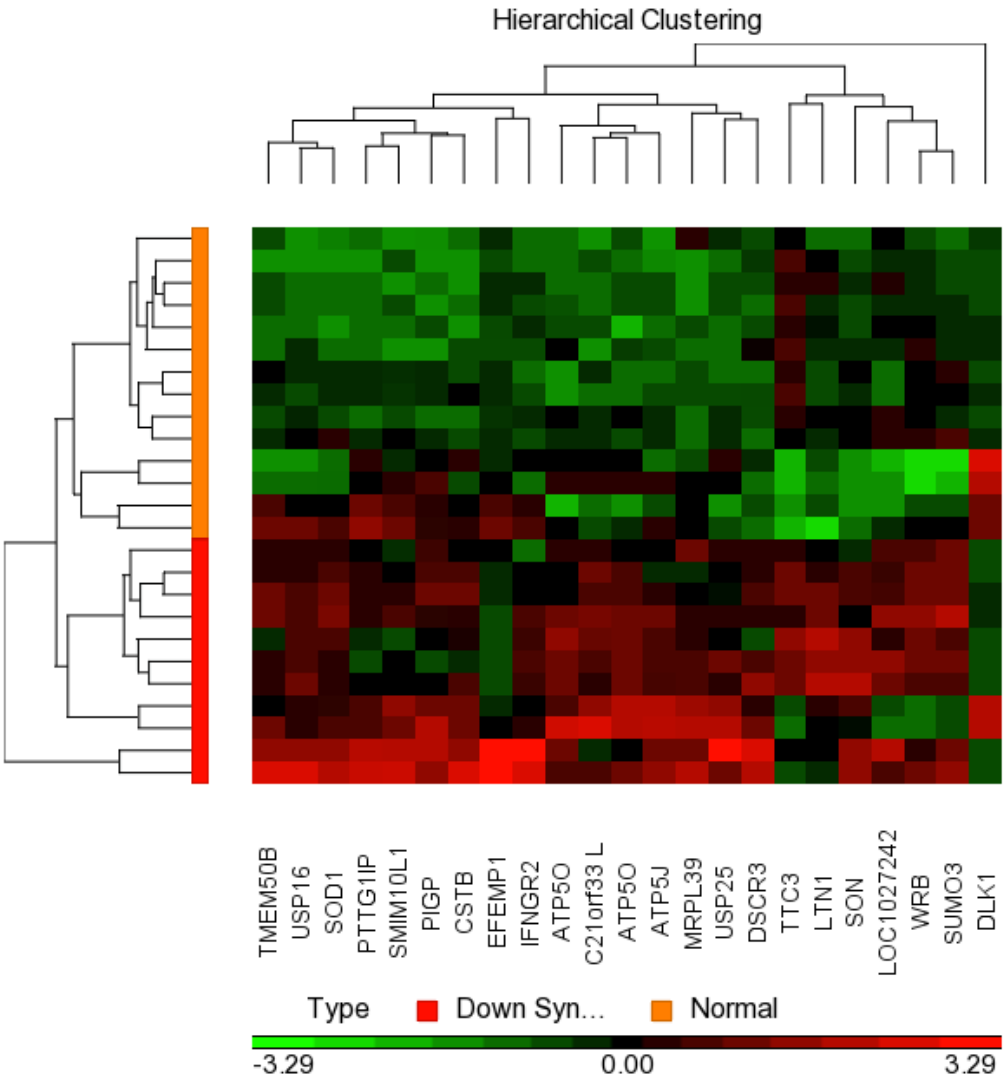
- The multi-omic information we have in our hands says nothing about the environmental factors influencing clinical information may give us a hint; age is likely important, also, someone who has stage 4 cancer is lik significantly different level of immune function.

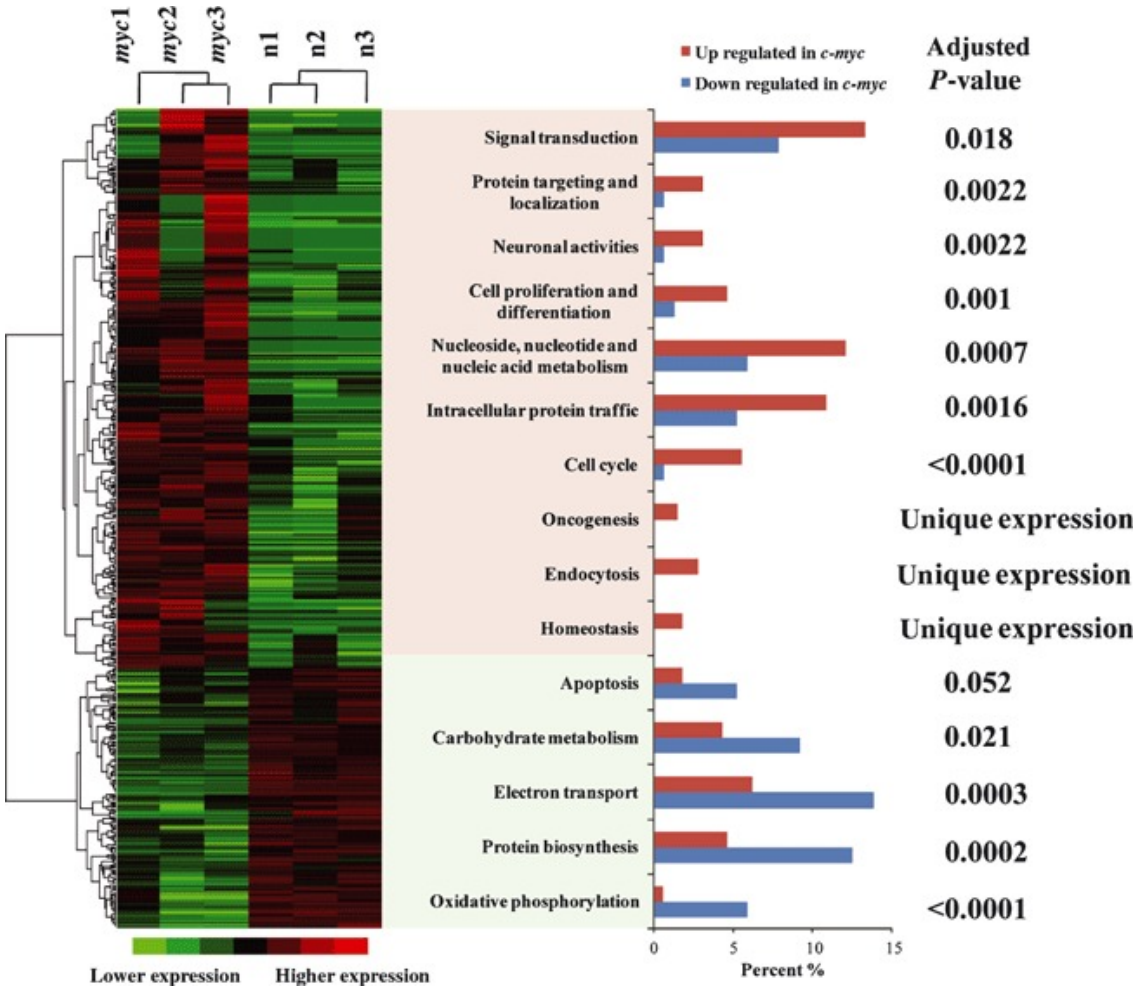
Questions to people from Erasmus:

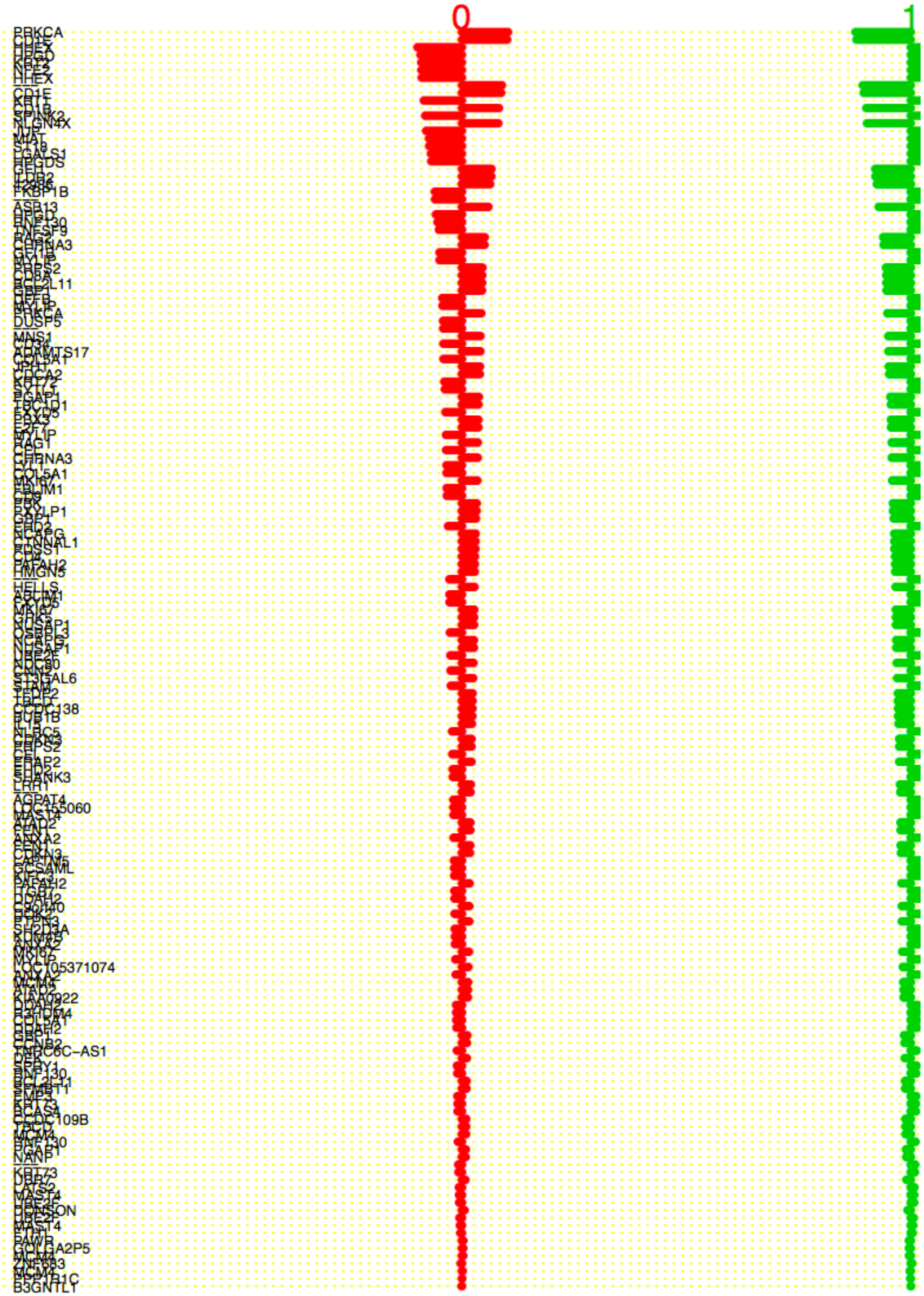
- can we get mappings from genomes to genome groups ? I added a folder to the google drive: mapping_da a mapping information from RNA probesets to genomes.
- how can we couple miRNA to proteomics proteine/miRNA-wise (so not sample wise..)?

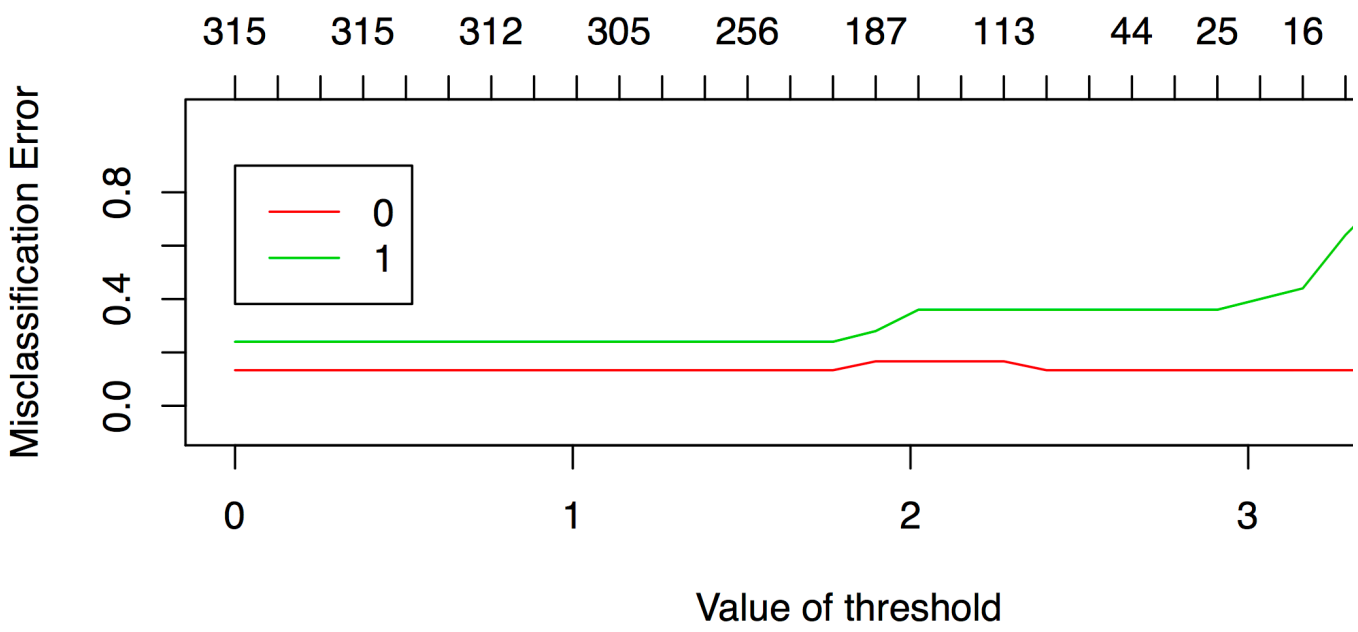
Things biologists like

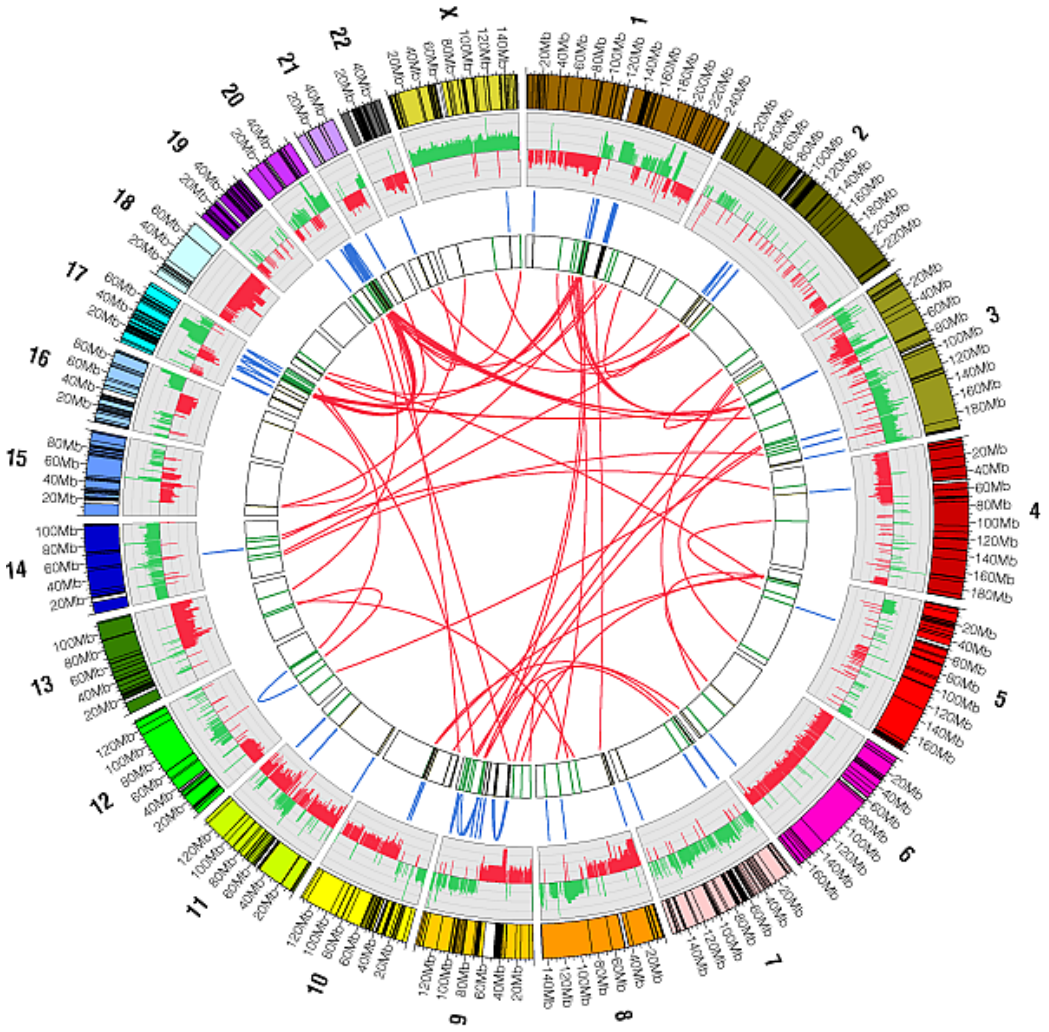
Hierarchical cluster diagrams, linked graphs, flow diagrams and simple tables/heatmaps with the most importar

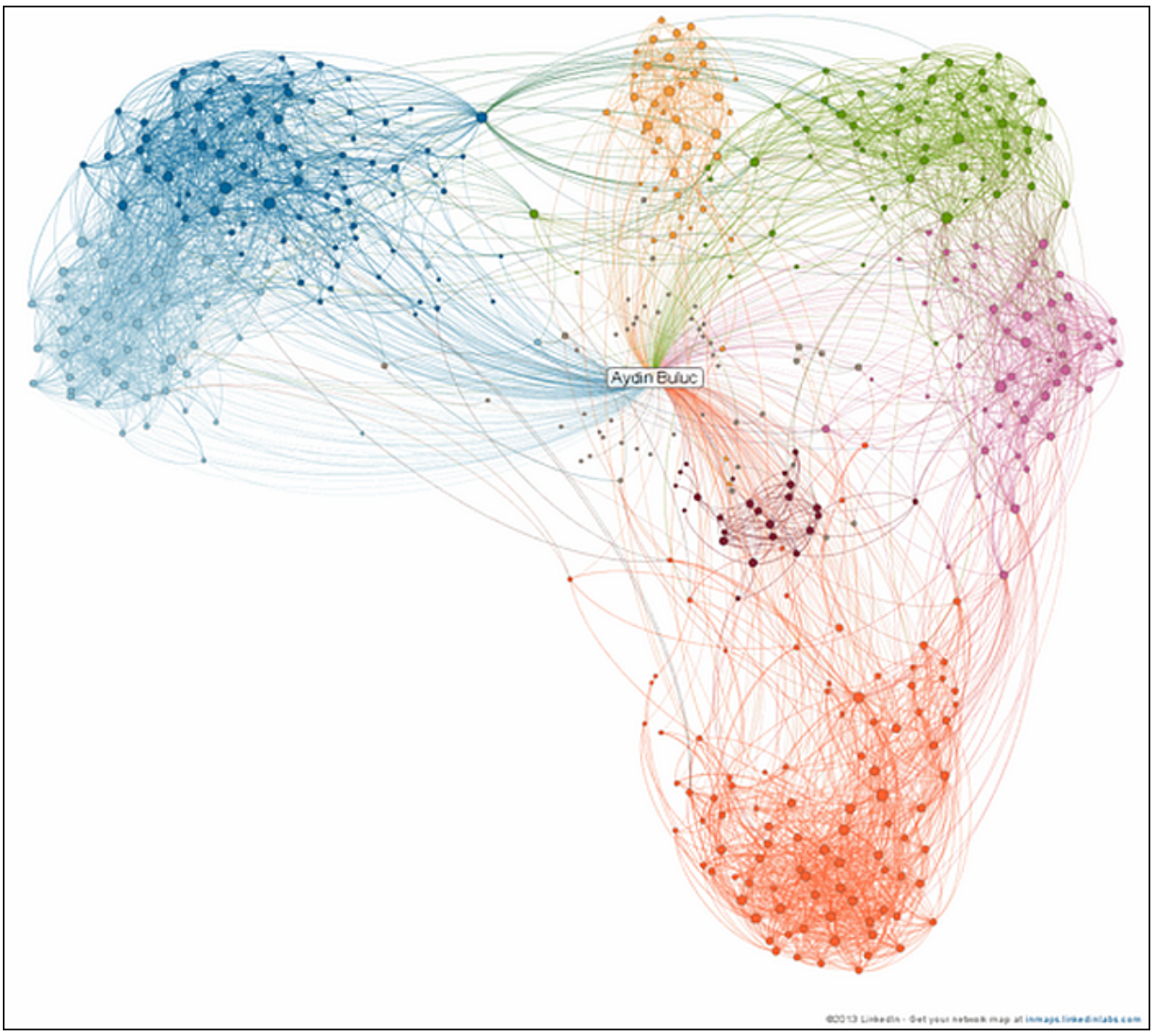



myc1
myc2
myc3
n1
n2
n3
Up regulated in c-myc
Down regulated in c-myc
Adjusted P-value
Signal transduction
0.018
Protein targeting and localization
0.0022
Neuronal activities
0.0022
Cell proliferation and differentiation
0.001
Nucleoside, nucleotide and nucleic acid metabolism
0.0007
Intracellular protein traffic
0.0016
Cell cycle
<0.0001
Oncogenesis
Unique expression
Endocytosis
Unique expression
Homeostasis
Unique expression
Apoptosis
0.052
Carbohydrate metabolism
0.021
Electron transport
0.0003
Protein biosynthesis
0.0002
Oxidative phosphorylation
<0.0001
0
5
10
15
Percent %
Lower expression
Higher expression


0
1
PRKCA
FKBP1B
DUSP5
KRT73
MAST4
MAST4
B3GNTL1


315
315
312
305
256
187
113
44
25
16
Misclassification Error
0.0
0.4
0.8
0
1
0
1
2
3
Value of threshold





Biologists like to understand the results, not very strange since they will base their laboratory work on it, medica derived from it and it will be applied to real patients. This is very important to keep in mind since it **excludes a r network-only approach**.

Suggested approaches

Please add ideas with your name in the section header (change the readme.md file in the `_doc` folder, then us `readme.md` , to install just do `pip3 install grip`)

Overall, I see three paths:

1. unsupervised learning and general exploratory data analysis to identify promising target variables, plus wo hypotheses
2. feature engineering --> transposition of tables --> dimension reduction --> normalisation --> classification --
3. graph generation and identification of common paths and graph clusters per classification --> viz

Clusters per layer

For the non-graphs, use some density-based clustering algorithm like HDBSCAN and lower dimension embedd

For the graphs, assuming we can construct them we can try to find

- communities

- exemplars
- cliques

Suggested algorithms/tools are :

- Sparse Affinity Propagation, for exemplars and communities
- Markov Clustering for cliques
- t-SNE (in sklearn but not hierarchical): multicore, multicore2
- HDBSCAN (you can find that here)

Dimension reduction

I would suggest the golden oldies, because they work :D

- PCA
- LDA
- FDR with ANOVA

If anyone can whip up an autoencoder that we would be cool but likely the above methods will do fine..

Classifications

This should be easy to do. First we should define the targets that are relevant to our end goal, which is to recog pathways, and inhibitors on those pathways. I.e. we need to be able to predict the **level of malignancy**, the **su** and the **response** to immunotherapy.

This generates weights/importances per feature and gives the predictive power of each layer.

A novel thing to do is to apply a tool like Quiver to visualise which genomes are important per classification. Fo need to transform the 1-dimensional tensors into 2-dimensional tensors. This would only work per patient, but a transparency is one of the key-ingredients of personalised medicine, and in my view of ML applied to health ca would be a nice touch.

Suggested algorithms:

- lightGBM, boosting type = GBDT, or DART, or GOSS in case of scaling problems
- CNN in Keras, we already have something laying around from last year

In general I see two approaches here:

- classification per layer plus stacking of the classifiers
- classification of the merged layers [CNV, mutation, methylation and RNA expression]+[miRNA, proteins]

These classifications can provide use with input for finding the pathways! Obviously by looking at the fea importances.

In particular this will say per layer which features are important, which for the miRNA and the protein data may identifying the final pieces of the pathway puzzle.

Correlations between different layers,

Extract layer pairs: DNA-RNA, RNA-mRNa, mRNa-proteomics, using

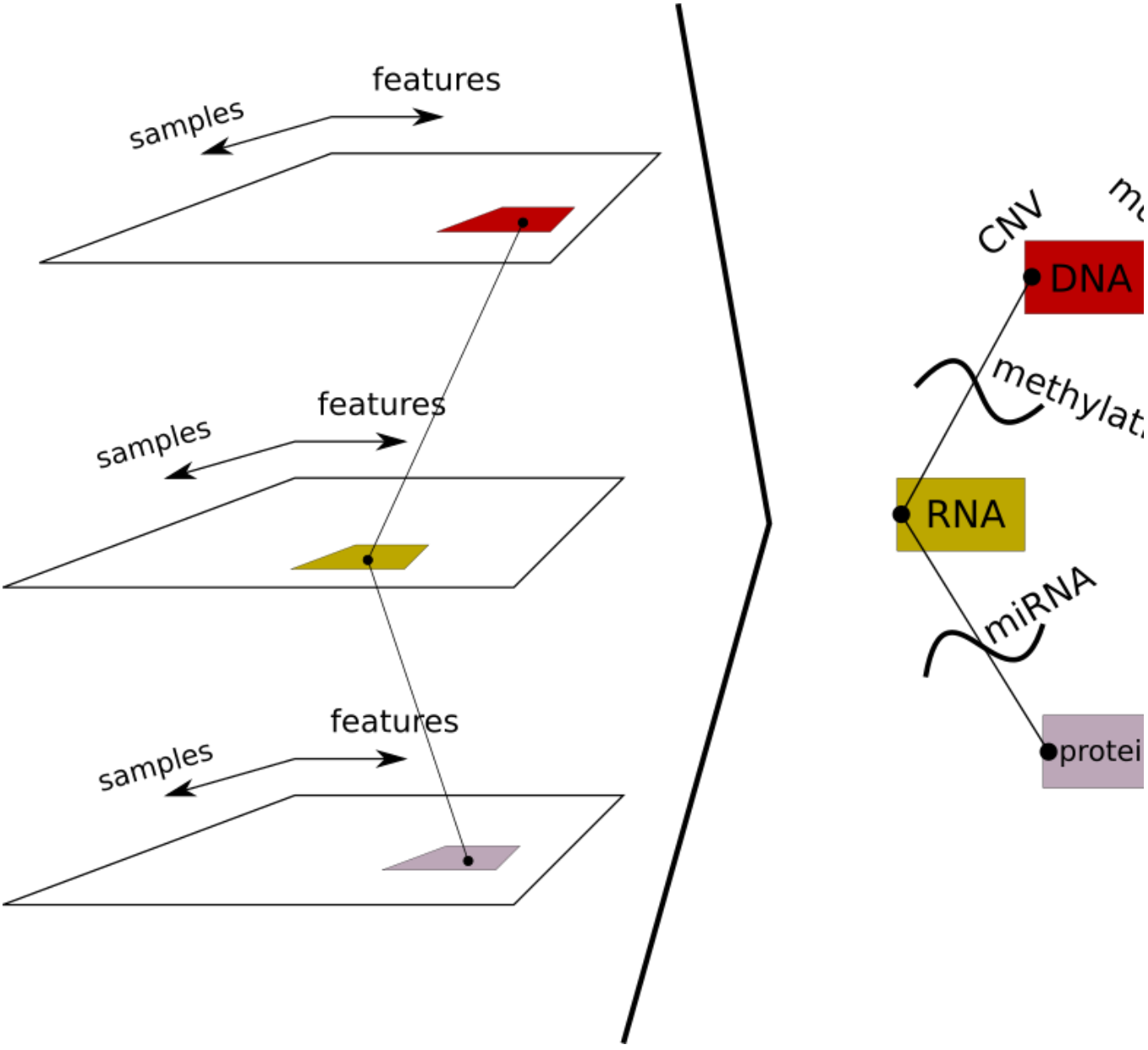
- raw features
- raw, filtered features (using variance over the different classes)
- reduced features per layer (PCA/LDA whatever)

-> generate new cross-layers that combine all possible layer pairs into single layer, train classifier and characte
multilayer pairs.

I.e. Correlated features link the layers pairwise, after which the layers can be connected into a single layer.

Bayesian Networks

Given potential pathways we can infer Bayesian Networks as approximations for the GRN and visualize them v
graph viz. tool. Watch, Read the wiki :D



Graphs, from the ground up

Per patient we have a graph connecting the CNV, mutation, methylation and RNA expression data using (Gene
stop). When looking at the gene-connectivity (i.e. counting occurrence of chr, start, stop, amino-acid change, ty

mutation), this graph will mostly be similar per patient in terms of the adjacency matrix but dissimilar in terms of matrix. This opens up some possibilities: We can

- determine clusters per patient graph: exemplars, communities, cliques. Then determine cluster overlap per
- create multi-layer graph per target label, count edges (or sum edge weights), normalise edge sums. Flatten graph and interpret normalised sums as edge weights. Determine characteristics clusters per target label. -> (N, *N*, 1)

The resulting clusters, and their characteristics can be used to feed a predictor. This has the benefit of

- transparency: it is clear why a target value is predicted
- compatibility: compared to simply merging the data into one matrix we have more guarantee to obtain biolo estimations
- it looks f*cking nice ;)

Suggested tools/algo's are:

- Markov Clustering,
- Neo4j-Bloom, networkx

Sources

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4731297/ https://www.ncbi.nlm.nih.gov/pmc/articles/PMC54963

https://www.nature.com/articles/nature13385

https://www.ncbi.nlm.nih.gov/pubmed/26091043 https://www.ncbi.nlm.nih.gov/pubmed/22960745