Internal Report CS Bioinformatics Track 15-02 August 2015

Leiden University

Computer Science

Bioinformatics Track

Characterizing mapk signaling in different cancers

Through large public datasets

Michael Liem

MASTER'S THESIS

Leiden Institute of Advanced Computer Science (LIACS)
Leiden University
Niels Bohrweg 1
2333 CA Leiden

Abstract

Melanoma is the most common form of cancer that originates in the skin; decades of intensive investigation generated large datasets that are currently provided to the open public. Large public datasets such as The Cancer Genome Atlas (referred to as TCGA), reveal that the majority of cases for both melanoma and papillary cancer (originating in the thyroid gland) harbor mutations affecting the *BRAF* gene. The most commonly occurring mutation is missense mutation p.Val600Glu (or p.V600E in one letter code), this particular mutation is associated to hyper activation of downstream protein MAPK1 that causes increased cell proliferation and inhibited apoptosis. Two drugs have been developed that specifically inhibit the mutated BRAF protein, thereby restoring the cell growth signaling to normal levels and effectively inhibiting tumor growth. It has however been observed that tumors can quickly become resistant to these treatment, particularly in colon *BRAF* mutated cancers.

We hypothesize that characterizing alternative signaling for specific subsets, for example those showing normal MAPK1 activity despite BRAF alteration p.V600E, would reveal alternative regulatory mechanisms which could lead to novel targets for drug development.

A popular choice to classify complex biological data, is the random forest technique, since it is nonparametric, interpretable and has high prediction accuracies in biologically relevant settings. Using TCGA data we generate four subsets from patients based on MAPK1 activity (high or low phosphorylation values) and presence or absence of BRAF mutation p.V600E.

We classify gene measurements with 10.000 trees, iterate 100 times and use the most important (top 10%) measurements to identify signaling routes that by-pass BRAF regulation. We perform classification on both melanoma and papillary cancer separately as well as on combined datasets.

In the comparison of patients with *BRAF* mutation p.V600E and decreased MAPK1 activity to other subsets we identified increased *BDKRB2, PCLB1, ITPR2* and *NOS3* gene expression, and decreased AKT phosphorylation values specifically for this subset. We hypothesize that this is a mechanism that activates cell-proliferation independent of BRAF by NOS3 activation through calcium second messenger systems.

Comparison of patients with unaltered BRAF genes and increased MAPK1 activity to other subsets was used to identify signaling that is not specifically BRAF-driven. Classification results show gene measurements to characterize this subset through gene expression and phosphorylation values of PTEN and AKT, MTOR, RPTOR, RBS6KB and EIF4EBP. This suggest circumvention of BRAF regulation through mtor signal transduction regardless of AKT inhibition mediated through PTEN activity.

Introduction

Biological background of mapk signaling in cancers

The molecular dynamics of cancers has been investigated thoroughly over the years, despite these efforts reasons explaining uncontrolled cell proliferation in many (types of) cancers remain unclear. It is known however that abnormal cell growth exhibited by many cancers requires deregulation of protein interaction and signaling networks such as the Mitogen Activated Protein Kinase pathway (referred to as mapk pathway). This is a large network that regulates cell proliferation, cell differentiation and apoptosis [5], which is frequently deregulated in cancers, amongst others papillary thyroid cancers and melanoma.

Melanoma is a cancerous disease that affects melanocytes; these cells produce pigment, grow relatively old and alterations of these cells are frequently observed in skin. Melanocytes are dendritic cells found in the lower region of the epidermis. Dendrites connect melanocytes to neighboring cells and transport pigment molecules know as melanin. Once arrived melanin umbrella's the nucleus of cells in the upper-layer of the epidermis, a process commonly observed as darkening of the skin after exposure to sunlight (see **Figure 1**). Surgical excision of small melanomas is the primary treatment and survival rates for early staged melanoma is 94% over five years [20]. However, patients with advanced metastasized melanoma have a very poor prognosis and show over 5 years only 10% survival rate for most severe forms of metastatic melanoma [21].

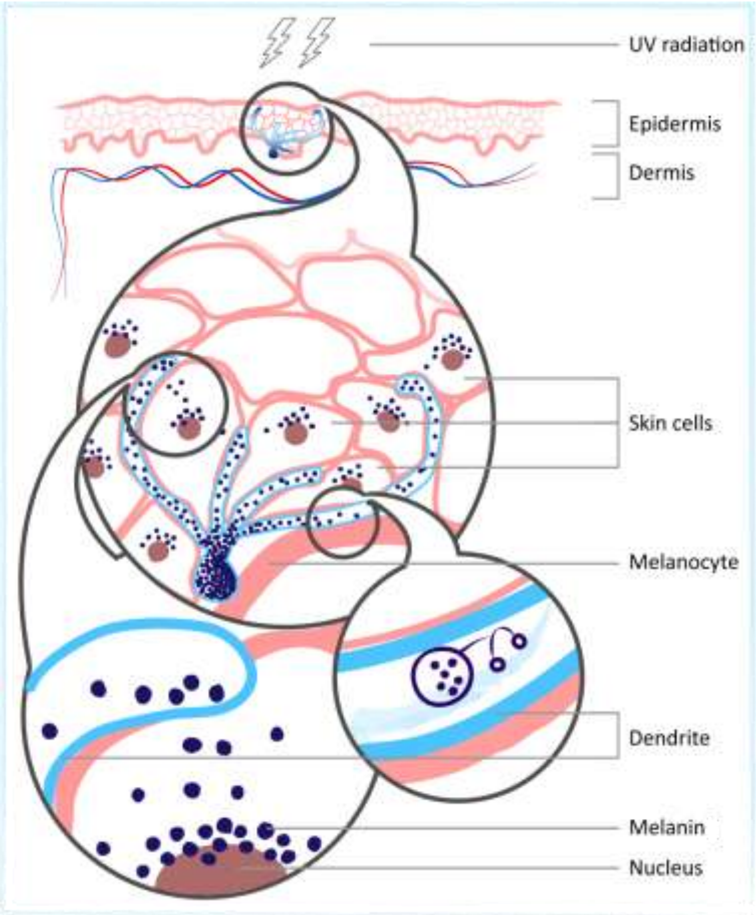


Figure 1: Representation of skin and locations of melanocytes.
From top to bottom; skin is exposed to UV-radiation from sunlight, and triggers melanocytes to produce pigment (melanin). These cells are located below epidermis cells, neighbouring cells are connected via long dendrites. Dendrites transport melanin from melanocytes to surrounding cells. Hereafter, melanin accumulates around the nucleus and protects DNA by absorbing UV radiation.

Deduce mapk functionality through various biological levels.

It is important to understand that regulation of mapk signaling involves several biological processes, among others, gene expression, DNA methylation, and protein activity. Furthermore, regulatory elements are influenced by copy number alterations (referred to as CNAs), and therefore considered important for mapk signaling characterization (i.e. genes deleted as result of CNAs cannot regulate or be regulated). Here we touch on the most important regulatory processes in the context of cancer. Firstly, an important factor in pathway regulation is the cell's ability to express genes and synthesize proteins that are required for functional cellular responses. The quantification of gene expression is attained through mRNA sequencing that can serve as approximation of protein quantities.

Secondly, CNAs of the gene structure is found to play a crucial role in many cancers, and causes deletions or amplifications that are particularly problematic for oncogenes. Exemplified; on the one hand, deleting tumor suppressors with inhibitory functions, or on the other hand, amplifying oncogenes that stimulates cell growth, results over activity of cell proliferation.

Thirdly, particular epigenetic changes, such as DNA methylation, have been documented through intensive studies and deregulation often occurs in cancers. Methyl groups bound to promoter regions of tumor suppressor genes are particularly problematic since this prohibits DNA polymerases to move along the genetic sequence. Which, in turn, prevents transcription of genes suppressing cell growth and emphasizes on the importance of epigenetic modifications in tumor genesis [8].

Finally, regulation of mapk signaling (or for signaling transduction cascades in general) involves phosphorylation and dephosphorylation of a large variety of proteins. This reversible post-translational modification adds or removes phosphate groups for phosphorylation and dephosphorylation, respectively. Transduction signals are consecutively transferred phosphate groups from one protein to another, proteins are considered active and inactive for bound and unbound phosphate groups, respectively.

Initial response of mapk signaling is triggered by binding of extra cellular compounds to receptors embedded in the cell's membrane. One such receptor is Epidermal Growth Factor Receptor (EGFR, see Figure 2 left-side), this tyrosine kinase receptor undergoes dimerization after binding of epidermal growth factors (EGFs), and is one 'entrance' to the mapk pathway [4]. Additionally, different intra- or extracellular stimuli can trigger different routes of pathways, and thereby allow cells to reach into different transduction cascades, at different parts of the cascade event (not shown in Figure 2).

Extensive whole-exome (sequencing all exonic regions) and RNA sequencing studies reveal a specific mapk activating mutation in the oncogene *BRAF* causing melanoma cells to develop hyper active mapk signaling and drug resistance. This mutation is found at the 600th amino acid of the protein chain (in the most common isoform) and causes the substitution of valine (one letter code V) to glutamic acid (one letter code E) and therefore named p.V600E (**Figure 2**). Previous work shows such mutations involve regulatory changes of cell proliferation, cell differentiation and can inhibit apoptosis through mapk signaling [6][7].

To date only two types of targeted treatment have been approved by the Administration of Drug and Food regarding melanoma [9] and both target BRAF [7]. These treatments show promising initial response, however, metastatic melanoma cells develops drug resistance after prolonged exposure to the treatment [15]. In cases of melanoma a number of pathways are known to play part in the regulation of cancerous development, among others, mapk, mtor, pi3k and the cell cycle (see **Figure**

2 right-side). Literature provides a core set of genes that are used for current targeted drug development that constructs our initial gene selection.

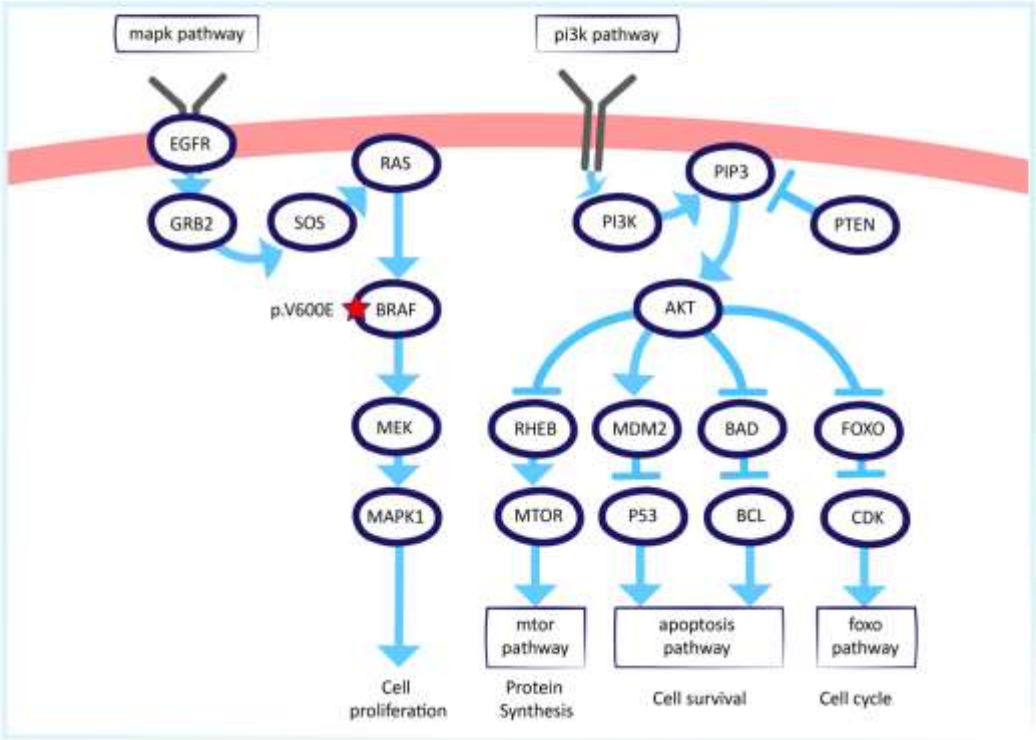


Figure 2: Simplified representation of mapk and pi3k pathway signaling. Membrane protein EGFR (entrance to mapk pathway at the left-hand side) activates GRB2 and SOS, which in turn activates membrane protein RAS, to finally signal through BRAF, MEK and MAPK1. At the right-side, PI3K stimulating PIP3, and activates AKT, that leads to four signaling routes.

Data access and availability – The Cancer Genome Atlas

Previous work collected significant amounts of data in The Cancer Genome Atlas (TCGA), among others, aimed to investigate melanoma. This enormous collection contains a maximum of 11 data types [14] (from here on referred to as gene measurements) such as: DNA copy-number data, mRNA and microRNA expression data, non-synonymous mutations, protein-levels, DNA methylation data, limited clinical data related to survival, and at the point of writing >66 TB in volume [11]. It contains both raw and processed data, and is structured in 4 levels; raw, processed, interpreted and summary, for levels 1 to 4 respectively. Level 3 data contains a normalized data aggregation [12], that is used in our analysis. Data present in TCGA is available through cbioportal [22] (an online web-based interface) that allows access by specified querying per patient, gene and cancer type.

In silico analyses – Inactive MAPK1 functionality and acquired *BRAF* alteration p.V600E

We assess a number of popular *in silico* methods for analyzing TCGA data, namely; random forest, principle component analysis and hierarchical clustering. These are representatives for both supervised and unsupervised learning, allows for introducing prior knowledge to classifications and investigate numerical intrinsic trends, respectively. Random forest biggest advantage is the ability to guide classification instead of extracting general intrinsic trends and aims to find data trends that specifically separates biologically relevant subsets. This technique performs well even for small-sample sizes and unknown data distributions, that typically describes biological datasets.

On the one hand mapk hyper activation is commonly observed in cases of cancer and associated to altered *BRAF*, on the other hand various patients suffering melanoma show low or normal activity of the downstream effecter MAPK1 under *BRAF* altered conditions. From this we deduce mapk signaling that circumvents BRAF-controlled MAPK1 stimulation and gene measurements from this subset might indicate mechanisms regulating development of drug resistance (i.e. resistance to BRAF inhibitors). To investigate signaling routes that increases MAPK1 activity other than stimulated through BRAF, we focus on differentiated gene and protein measurements that particularly characterizes patients with active MAPK1 functionality despite absence of *BRAF* alteration. Alternative signaling that characterizes this subset might reveal routes that increases mapk signaling, that in turn, may provide potential drug targets.

The Cancer Genomics.nl (CGC) is an initiative of seven cancer research groups in The Netherlands, with united forces they aim to understand genetic alterations in individual tumors to increase development of personal medicine. The flagship project aims to measure and understand the rapid drug resistance to targeted BRAF treatment in colon tumors by studying a tumor derived cell line. They were willing to provide the available measurements for our genes of interest. This data includes mRNA, protein and phosphorylation measurements under control, BRAF inhibitor and EGFR inhibitor conditions.

Methods

For a complete visualization of our implementation please find a diagram representation on the page 10.

Selecting resources, gene set and corresponding nomenclature

We retrieve cancer study statistics through cbioportal and select data resources based on volume and presence of recorded *BRAF* alteration. We initially selected a set of 30 genes that are often altered in cases of melanoma and are targets for drug development in current studies. The gene selection covers; 3 genes from the mapk pathway, 5 genes from the cell cycle pathway, 12 genes from the mtor pathway, 6 genes from the pi3k pathway and finally 4 genes that are tyrosine kinase receptors and potential entrances of the mapk pathway. This selection contains the major elements of signal transduction under natural conditions, however, picking-up alternative signaling requires larger gene selections since one signaling route easily contains over a dozen different proteins. Therefore we expand our gene selection using parameter ‘optimized gene-interconnectivity’, through geneMania [23], a score that is based on physical interactions, co-expression, co-localization, protein domain similarity, genetic interactions, participate in identical pathways and predicted interactions (*for specific details on gene selection see* ***supplementary methods*** *– gene selection*).

Data acquisition, data types and normalization

To query TCGA data a basic collection of functions is provided through R package CGDS-R package, that is part of the R platform for statistical computing [24] and Cbioportal provides hands-on descriptions to implement functions. To retrieve and post-process data both platforms R and Ipython notebook [25] are used. Our setup requires various parameters to download data from Cbioportal; 1) online resource directory, 2) integer value that indicates cancer studies, 3) list of genes and 4) specification of desired gene measurements. For papillary thyroid cancer and melanoma studies five gene measurements are available, mRNA expression, methylation, copy number alteration, phosphorylation and mutation data.

Within-sample normalized gene expression values represent Z-scores that describe the number of standard deviations that the value is distanced from the average mRNA expression value across all patients per gene for a particular type of cancer.

Next, normalized methylation data is expressed through values between zero and one, zero indicating hypo-methylation of genes, meaning no methyl-groups bound to promoter regions, and one indicating hyper-methylation that resembles heavily silenced genes.

TCGA provides CNA values in a per-gene fashion, values 2, 1 ,0 ,-1 and -2 indicate nullizygosity, hemizygosity, unchanged, gain or heavily amplified, respectively. This indicates homozygous deletions, loss of single alleles, normal state (two copies), one extra copy or multiple copies of the allele, respectively. TCGA is able to putatively approximate these copy number alterations due to the purity and known ploidy of samples and additionally provides CNA values in log scale that are used in our analysis.

Phosphorylation measurements are collected using fluorescent labels using custom designed antibodies that target phosphorylation sites. Nearly a couple of hundred antibodies are carefully selected and used for testing. Phosphorylation data is both within-gene and within-sample normalized.

And finally, mutations are collected through second-generation whole exome sequencing, and TCGA provides mutations per gene, per patient and per cancer study. Moreover, TCGA provides effects for several kinds of mutations through a suffix, such as; ‘*’, ‘fs’ or ’_splice’ at the end of mutation names that we use for mutation discretization. (*For specific details on normalization of gene measurements see* ***supplementary methods*** *– gene measurement normalization*)

Data discretization, subsets and merging cancer studies

A large variety of mutations is observed for different genes from different patients, to direct our classification towards p.V600E we label every mutation based on their effect. Initially, mutations are provided through 5 events; 1) not available (NA), 2) nonsense, 3) frame-shift, 4) splice site and 5) missense mutations. Since frame shift, splice site and nonsense mutations all result in different transcripts they are identically labeled. Additionally we check for p.V600E mutations that are labeled ‘hotspot’ and finally, any other mutation is regarded missense.

Since activation of genes through phosphorylation is not standardized we take phosphorylation, in conceptual terms, to indicate on/off states, that is ‘on’ for increased and ‘off’ for decreased phosphorylation values. We recapitulate the on/off state by thresholding phosphorylation values with an upper and lower boundary, and take values to indicate active (on) and inactive (off) states for values above and below thresholds, respectively. Thresholds are defined by calculating maxima and minima of phosphorylation values per protein across all patients, subtracting minima from maxima results the total range in which phosphorylation values vary in a per-gene fashion. We set thresholds in percentage and use equal values for upper and lower boundaries, exemplified; for upper and lower thresholds = 0.2 we label proteins ‘active’ with phosphorylation values that range in the upper 20% of the total phosphorylation range per protein, and values in the lower 20% are labeled ‘inactive’.

We generate different gene selection subsets to provide complete signaling routes or add genes that potentially play part in alternative signaling, gene selection mainly relies on interconnectivity between our gene selection and available data. That follows the rational; genes involving regulation of multiple processes are most potent in offering alternative routes that provide escapes from BRAF-controlled MAPK1 regulation. Furthermore, we subset patients with similar properties to indentify particular signalings that might characterizes subsets. Exemplified; we cluster subsets that contain patients with elevated *BRAF* CNA (CNA >= 2) in order to find gene measurements involving hyper activation of MAPK1 (*for specific details on subsets see **supplementary methods***). Finally, to perform random forest classification on merged cancer studies we intersect gene measurements of complete case datasets from individual cancers. Gene measurements that are present exclusively in either one of the cancer studies are excluded from the merged dataset.

Implementation of principle component analysis and agglomerative hierarchical clustering – unsupervised clustering

To evaluate the spread of our data we designed a principle component analysis implementation using R statistics. We remove all gene measurements that contain 'NA' values, and parse the set of complete cases to the princom function with default settings. We evaluate eigenvector loadings that indicate data spread captured for every principle component separately and visualize their directions.

Additionally we parse the complete-case data set to the hierarchical clustering function, we define two parameters; 1) clustering method and 2) distance metric. We can generate a tree that connects data points together either starting from one big cluster separating one data point after another, or starting with all data points separately clustering them together one by one. The later is referred to as agglomerative clustering and is set through the first parameter. This method measures distances between data points in multidimensional space using Euclidian distance metric, that describes a space in which a straight line is the shortest distance between x and y, and is set with the second parameter.

Random forest classification – supervised clustering

Random forest is functionally related to its name and generates large numbers of random decision trees (i.e. a forest), the nodes of every tree are split based on optimization using a set of randomly chosen measurements. We set our random forest implementation to generate 10.000 trees and classify based on MAPK1 activity (high or low phosphorylation values) and present or absent BRAF mutation p.V600E

The main principle behind this algorithm is perform 'strong' classification taken together the sum of 'weak-classifiers' (single trees). However this requires balanced class distributions to avoid bias towards over represented classes, hence we inspect distributions for imbalances. Since datasets are relatively small phosphorylation threshold values are set to 0.5 and quantifies most balanced class distributions. (*For specific details on class distributions see **supplementary methods** – class distributions*).

The solution offered by random forest represents the highest ranking tree among all randomly generates trees. Our random forest provides two ways to assess the importance of measurements for node splitting. Firstly, the score for 'meanDecreaseAccuracy' indicates whether accuracies drop if measurements are left out during classification. Returning high and low values that indicate the importance of measurements regarding low and high error-rates, respectively. Secondly, the ‘Gini’ score indicates whether measurements generate pure nodes, resulting high and low values indicate measurements generate pure nodes or cause node impurity, respectively. Here, measurement importance is expressed through Gini scores.

Furthermore, for learning purposes random forest requires a training set that is used to generate all random decision trees and contains randomly chosen gene measurements from complete case datasets. Simultaneously resulting a randomly generated dataset (i.e. the measurements that were not selected for learning) that we use for testing. This Out Of Bag (OOB) method mimics commonly known train/ test-set procedures that are currently used in most classification algorithms

Define top-10% measurements in dynamic results due to randomness and data complexity

We focus on top-10% measurements as indicators for class separation, rather than to optimize correct classification of patients to their corresponding class. This method ranks measurements based on their importance in generating node purity, even for high complexity biological data sets. High complexity data results in rather low classification performance and causes varying results regarding measurement importance indicated by random forest. This means Gini score results vary for every classification run, however, the strongest separators will rank higher more frequently compared to others. Therefore we iterate our procedure 100 times and rank Gini score means to generate the top-10% separating measurements (*for specific details on error-rates en Gini score rankings see **supplementary methods***).

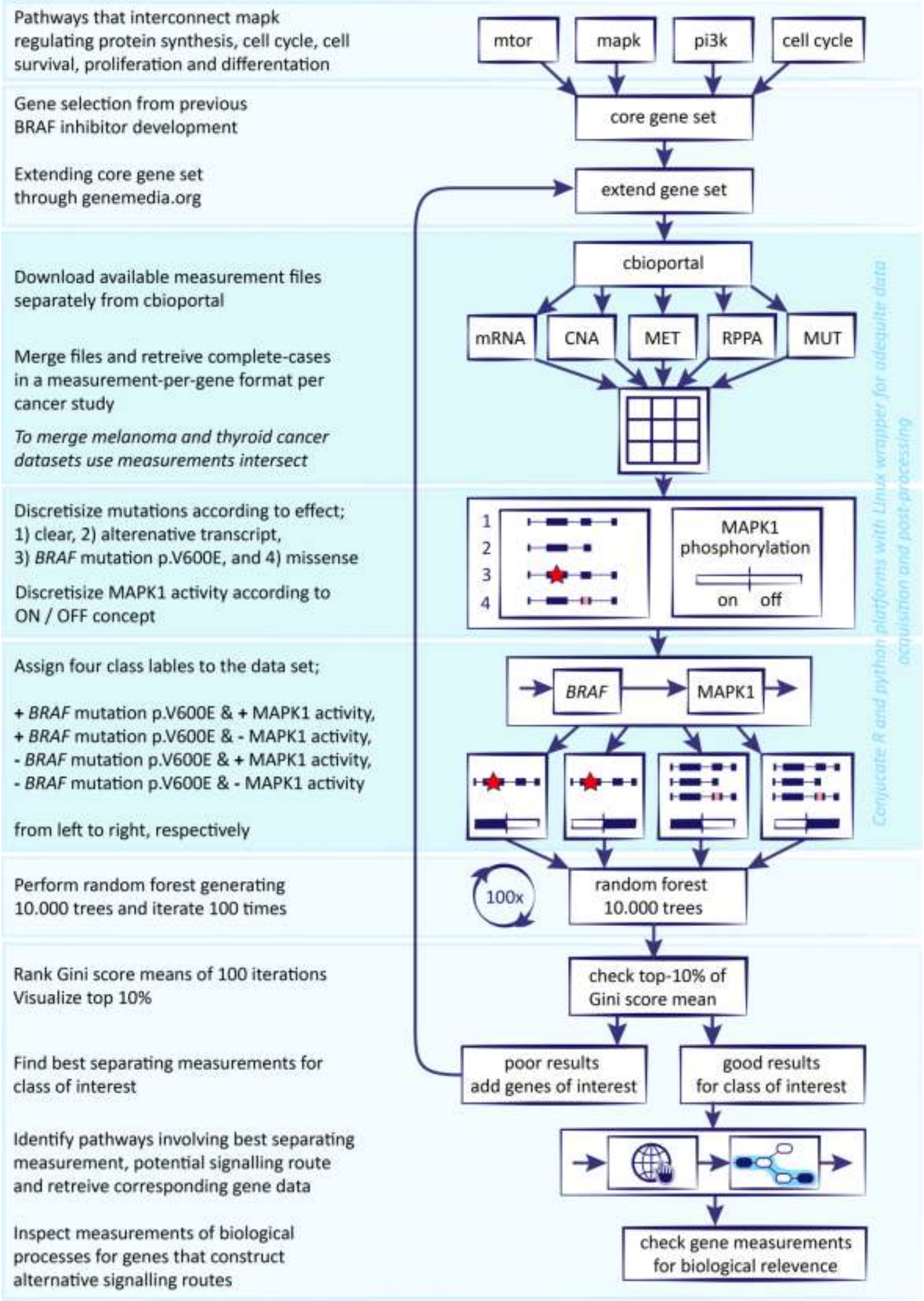
Adequate data acquisition and post-processing

To conjugate platforms such as R statistics and Python we design a shell script (i.e. ‘wrapper’) that embeds a set of command-line utilities, and allows post-processing data in an fast and adequate fashion. The wrapper directs data acquisition through R statistics and invokes custom made python code for basic post-processing and requires four parameters; 1) result folder directory, 2-3) invoking both R statistics and Python scripts, and 4) a text file that denotes all genes in a gene-per-line format.

Identify pathways involving best separating measurement and potential signaling routes

We inspect top-10% measurements and focus on characterizing alternative signaling for either *BRAF* mutation p.V600E under inactive mapk conditions or absent *BRAF* mutation p.V600E and MAPK1 activity. We identify pathways containing signaling routes that provide escapes for *BRAF*-controlled MAPK1 regulation, involving our best separating measurements for aforementioned classes. Pathways and signaling routes are collected from online databases, such as; wikipathways, KEGG and GeneCards [26][27][28]. Hereafter, we collect measurements for genes that construct potential alternative signaling routes and inspect biological relevance of our random forest prediction.

Gene selection, data acquisition, post-processing and random forest classification represented in a diagram. Separate steps are indicated through blue backgrounds (dark blue describing the wrapper)



Pathways that interconnect mapk regulating protein synthesis, cell cycle, cell survival, proliferation and differentation

Gene selection from previous BRAF inhibitor development

Extending core gene set through genemedia.org

Download available measurement files separately from cbioportal

Merge files and retreive complete-cases in a measurement-per-gene format per cancer study

To merge melanoma and thyroid cancer datasets use measurements intersect

Discretisize mutations according to effect; 1) clear, 2) alterenative transcript, 3) *BRAF* mutation p.V600E, and 4) missense

Discretisize MAPK1 activity according to ON / OFF concept

Assign four class lables to the data set;

+ *BRAF* mutation p.V600E & + MAPK1 activity,
+ *BRAF* mutation p.V600E & - MAPK1 activity,
- *BRAF* mutation p.V600E & + MAPK1 activity,
- *BRAF* mutation p.V600E & - MAPK1 activity

from left to right, respectively

Perform random forest generating 10.000 trees and iterate 100 times

Rank Gini score means of 100 iterations Visualize top 10%

Find best separating measurements for class of interest

Identify pathways involving best separating measurement, potential signalling route and retreive corresponding gene data

Inspect measurements of biological processes for genes that construct alternative signalling routes

mtor
mapk
pi3k
cell cycle
core gene set
extend gene set
cbioportal
mRNA
CNA
MET
RPPA
MUT
1
2
3
4
MAPK1 phosphorylation
on
off
BRAF
MAPK1
100x
random forest 10.000 trees
check top-10% of Gini score mean
poor results add genes of interest
good results for class of interest
check gene measurements for biological relevence
Conjucate R and python platforms with Linux wrapper for adequite data acquisition and post-processing

Results

Data resource and acquisition

Datasets from both Melanoma and Thyroid cancer studies are collected from TCGA, cbioportal reveals four and two available datasets, respectively. We retrieve TCGA provisional Melanoma data that shows >50% of 278 patients attained *BRAF* alterations during cancer development (**Table 1** and **Figure 3**) [10]. And for data homogeneity purposes we collect TCGA provisional Thyroid cancer data that indicates >60% of 399 patients either have mutations, amplifications, deletions or multiple alteration (**Figure 3** and **Table 1**). Both Thyroid cancer and Melanoma TCGA Provisional datasets provide five different measurements; mRNA expression, CNA, phosphorylation, methylation and mutations.

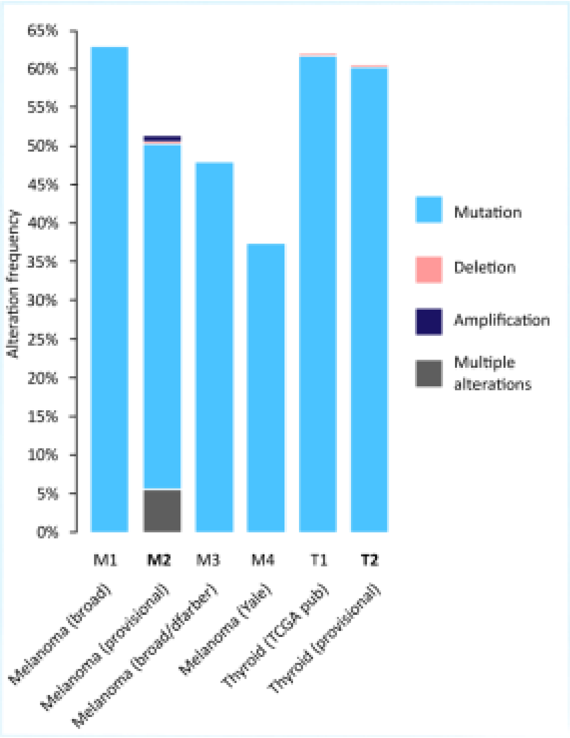


Figure 3: Data sets provided by TCGA for melanoma and papillary thyroid cancer. Datasets are represented in decreasing order of BRAF alteration frequency (in percentage), and primarily reveals alteration of gene structure (in light blue).

Table 1: Statistics of available TCGA resources

ID	cancer study	cases	*BRAF* mutation
M1	Melanoma	121	62.8%
M2	Melanoma	278	51.4%
M3	Melanoma	25	48.0%
M4	Melanoma	91	37.4%
T1	Thyroid	399	61.9%
T2	Thyroid	399	60.4%

We grow the core gene selection with 100 interconnected genes (*see **supplementary methods** – gene selection*), retrieve 601 gene measurements, and merge them to a measurement-per-gene format per cancer study. Hereafter, we inspect and remove gene measurements that contain missing values generating complete case datasets per cancer study. We request clinical data, for both melanoma and papillary thyroid cancer, and find tissues at stages from m0 to m1-a,b and c, and from n0 to n3. M stages indicate distant metastasis, the abbreviations m1 and m0 stand for observed or absent distant metastasis, respectively, with suffix denoting metastatic locations. N stage melanomas indicate spread to surrounding lymph nodes, and number of affected lymph nodes is indicated with an integer suffix. Cbioportal indicates control samples of melanoma and thyroid cancer are partially publicly available, however, package CGDS-R only allows data retrieval of cancerous tissues.

Data set statistics

Post-processed Melanoma and thyroid cancer datasets count 158 and 207 patients, 515 and 497 measurements, respectively (**Table 2**). Phosphorylation values are available for only a few genes compared to mRNA expression values (nearly all genes have mRNA expression data). Merging cancer studies slightly lowers patient count and gene measurements, however, intersecting both cancer studies results particularly decreased phosphorylation measurements.

Table 2: Summary statistics of melanoma papillary thyroid cancer and merged studies with core gene selection and 100 additional genes.

	Melanoma		Thyroid		Total
Description	TCGA	**Filtered**	TCGA	**Filtered**	**Filtered**
Patient count	278	158	397	207	365
Measurement count	616	515	601	497	491
genes with RPPA values	123	32	120	24	20

Class labels used for classification (**Table 3**), the second class indicates subsets that might reveal alternative signaling used to stimulate the cell cycle for low MAPK1 activity. The third class label indicates alternative signaling not specifically driven by *BRAF* alterations. First and last class labels indicate *BRAF* alteration and hyper active mapk signaling, and patients that developed melanoma despite intact mapk regulation (no *BRAF* mutation and normal MAPK1 activity).

Table 3: Class label description

	Prediction classes	Melanoma	Thyroid	Total
1	+ *BRAF* mutation p.V600E & + MAPK1 activity	22	5	27
2	+ *BRAF* mutation p.V600E & - MAPK1 activity	45	112	157
3	- *BRAF* mutation p.V600E & + MAPK1 activity	44	7	51
4	- *BRAF* mutation p.V600E & - MAPK1 activity	47	83	130

Post-processing melanoma datasets generates equally balanced data distributions for class labels 2 and 3 (**Table 3**), in contrast to class distributions of thyroid cancer data that shows bias towards inactive MAPK1 (**Table 3**, class labels 2 and 4). Merging melanoma and papillary thyroid cancer complete case data sets results the sum of class distribution of individual cancer studies.

Traditional classification methods and complex biological data (unsupervised classification)

Classical methods, such as principle component analysis and hierarchical clustering, are used to indicate spread and similarity of data sets, respectively. In order to describe data spread, for merged cancer datasets, PCA requires all principle components that indicates high complexity. Hierarchical clustering shows small separate clusters of patients with *BRAF* mutation p.V600E and inactive MAPK1 values, however noise levels cause interpretation of gene measurements importance to be problematic.

Difficulties of traditional classification arise due to a number of reasons; our extended gene selection is likely to describe signaling routes incompletely, genes from multiple signaling routes and other biological events are mixed in our dataset and clutters classification results. Additionally, signaling routes are described through a multitude of measurements increasing data complexity, and finally, these methods lack the ability to facilitate classifications with prior knowledge. Therefore results are not directed towards answering our biological question, but rather indicates most profound spread or similarities of our dataset in general. (*for specific details on traditional classification see **supplementary methods***)

Random forest classification for merged datasets with extended gene selection

Inspecting a number of the top-10% measurements reveals random forest finds measurements that separates distinct types of cancer. The five best results are mRNA expressions of ERBB3, *ELF3* and *MET* and methylation values of *EIF4G1* and *MAP3K1* (**see Figure 4**).

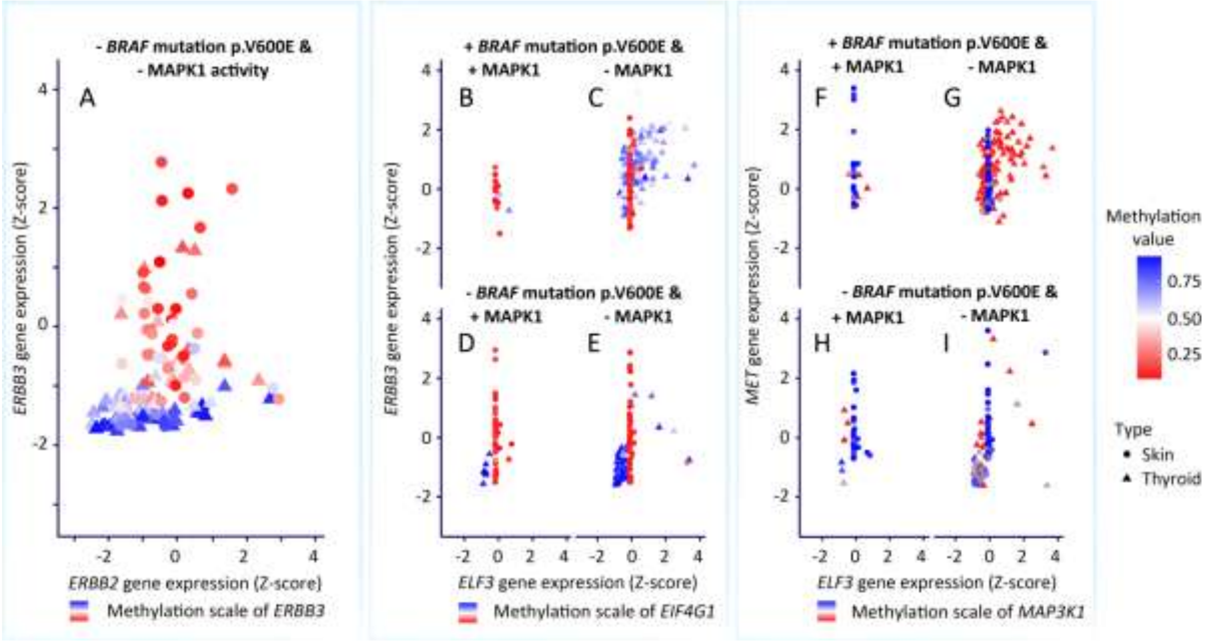


Figure 4: A, indicates gene expression of *ERBB2* and *ERBB3*, dot shapes indicate cancer studies, circles and triangles indicate melanoma and thyroid cancer cases, respectively. Here we observe decreased *ERBB3* gene expression in addition to hyper methylation. **B**, separable power of *ELF3* gene expression results from tight regulation compared to increased values for melanoma and papillary thyroid cancer, respectively. **C**, hypo methylation of *MAP3K1* shows separable power towards cancer studies.

Visualization of *ERBB2* and *ERBB3* mRNA expression reveals values of both melanoma and papillary thyroid cancer patients, that emphasizes on tight epigenetic regulation of *ERBB3* gene expression for patients suffering papillary thyroid cancer (***Figure 4A***). These results explain separable power gained from gene expression of *ERBB3* facilitates separating cancer types. ERBB3 is a unit of heterodimer ERBB2-ERBB3 (with downstream protein ELF3) that is embedded in the cell membrane and phosphorylates protein from both pi3k and mapk pathways. Homodimer formation of ERBB2 is previously observed under decreases *ERBB3* gene expression conditions and is described as cancer driving mechanism in diseases such as, papillary thyroid cancer and breast cancer [1][2][3].

Additionally, we observe similar results for gene expression of *ELF3* and methylation values of *EIF4G1* (***Figure 4B,C,D and E***). Gene expression of *ELF3*, in cases of melanoma, rarely deviates from the mean indicating tight regulation, in addition to low methylation values of *EIF4G1* across all patients. In contrast to melanoma, patients suffering from papillary thyroid cancer show *ELF3* mRNA expression ranging from -1 to 4 standard deviations, rarely accumulate around the mean and are characterized by high methylation values of *EIF4G1* across all patients. And finally, separable power of *MET* gene expression and *MAP3K1* methylation values is most profound between presence and absence of *BRAF* mutation p.V600E and inactive MAPK1 (***Figure 4G and I***). Gene expression values of *MET* range from -1 to 3 standard deviations from the mean for patients without *BRAF* mutation p.V600E and inactive MAPK1 (***Figure 4G***), this contrasts to patients without *BRAF* mutation p.V600E and activated MAPK1 that typically shows expression values below zero (***Figure 4I***). From this we take profound differences (e.g. ERBB2 homodimer formation) to overshadow separation towards minority subsets. Hence, we continue in a per-cancer-study fashion by focusing on melanoma cases to lower data complexity.

Alternative signaling possibly involving drug resistance – *BRAF* mutation p.V600E and inactive MAPK1

The top-10% of our random forest results for patients with altered BRAF and low MAPK1 activities identified increased gene expression and CNA values for genes *BDKRB2* and *NOS3*, respectively.

BDKRB2 is a membrane protein that associates with G proteins stimulating calcium second messenger systems (**Figure 5**) and increases cytosolic calcium concentrations [19]. BDKRB2 gene expression appears particularly increased compared to other subsets (***Figure 6A***). The mean of BDKRB2 gene expression is zero standard deviations that indicates expression levels are similar compared to *BDKRB2* expression levels across all patients in the unfiltered the TCGA dataset. These differences reveal separation originates from decreased *BDKRB2* gene expression values of other subsets.

NOS3 is a member of the pi3k pathway and is known to activate under elevated calcium concentrations [18]. This effect is previously observed in human melanomas and for human *NOS3* gene transfer in mice, that showed mapk inhibiting effects that corresponds to the characteristics of this particular subset [16][17]. The separable characteristics picked-up by our classification indicate >77% of BRAF mutated patients have at least one extra copy of the NOS3 gene, in contrast to cases of unaltered BRAF that count <48% patients with elevated NOS3 CNAs.

We investigate available measurements of both genes in addition to genes that construct the alternative signaling route (i.e. genes from the pi3k and calcium signaling pathways). Visualizing gene expressions of *BDKRB2* and *NOS3* reveals the widest spread of these two genes specific for patients with BRAF mutation p.V600E and inactive MAPK1 (***Figure 6C***)

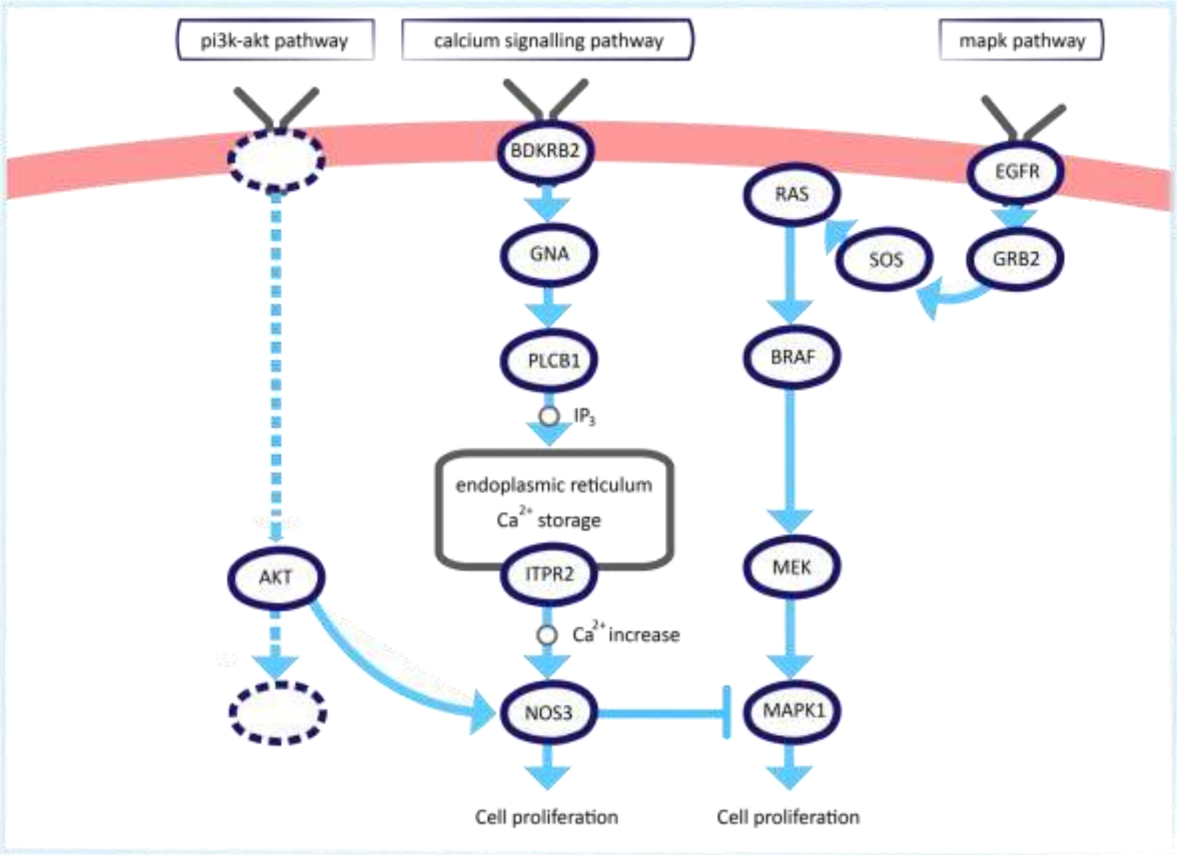


Figure 5: simplified representation of alternative signaling route through second messenger system and nos3. At the right-hand side mapk signaling under normal conditions. On the left-hand side signaling routes regulated through AKT. In the middle section, activation of NOS3 through calcium increase activated by BDKRB2, that stimulates, GNA and PLCB1 and cause ITPR2 to release calcium from the ER to the cytosol activating NOS3.

From the pi3k pathway we find decreased phosphorylation values of NOS3 upstream activators AKT1 and AKT2 ***(Figure 6A)***, that indicates NOS3 is not stimulated via AKT. However elevated gene expression levels for *PLCB1* and *ITPR2* suggests activation of second messenger systems that releases calcium from the endoplasmic reticulum into the cytosol activating NOS3 (**Figure 5** middle section and ***Figure 6A***). This would allow cells to escape BRAF-controlled mapk regulation and stimulates proliferation under BRAF inhibited conditions.

The comparison of read count medians across all patients from the filtered dataset reveals median = 51, 120 and 222 for *BDKRB2*, *PLCB1* and *NOS3* gene expression, respectively, to read count medians of this subset (median = 103, 231 and 334 for BDKRB2, PLCB1 and NOS3, respectively) reveals a nearly two-fold increase. Albeit read count medians of these genes are specifically increased for this subset they indicate relatively low expression compared to the read count median across all genes (median = 1473 reads across all genes in our dataset).

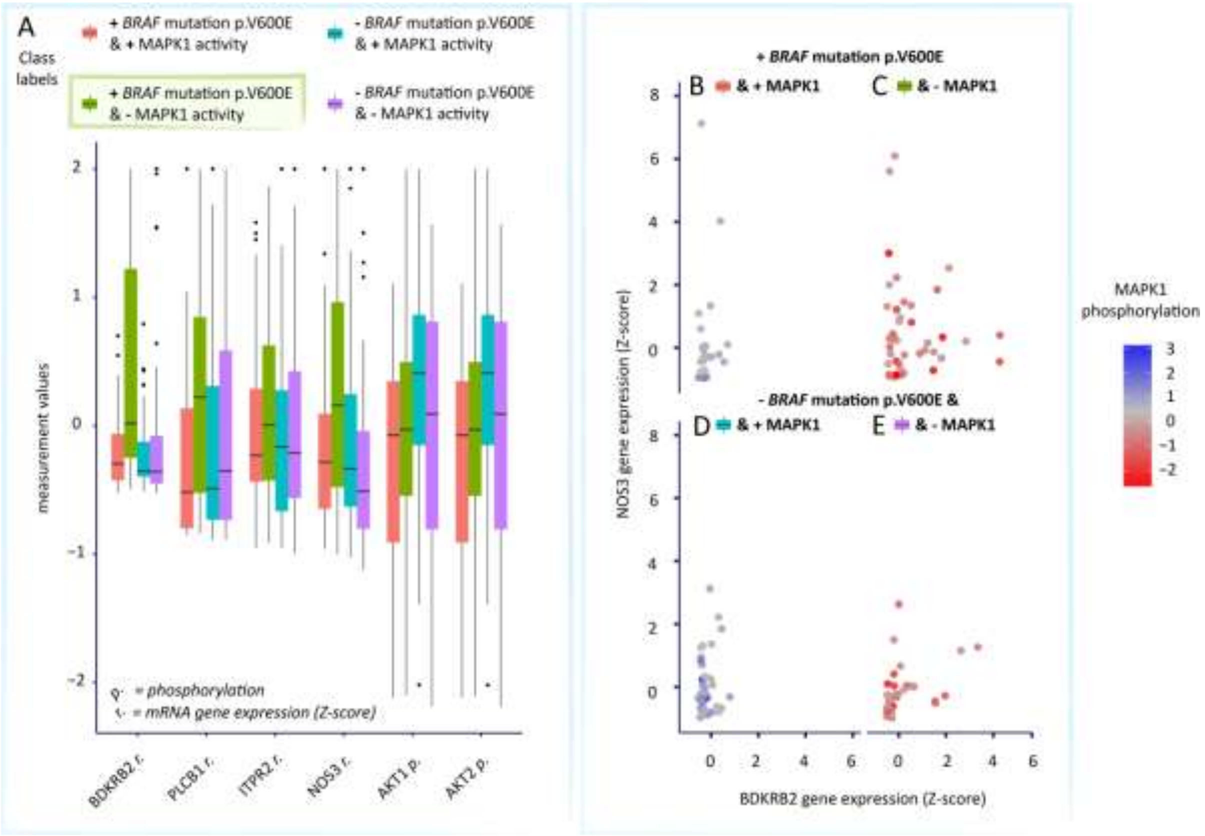


Figure 6: gene expression and phosphorylation values of alternative signaling route. A, From left to right; gene expression values of *BDKRB2*, *PLCB1*, *ITPR2* and *NOS3*, indicating increased levels specific for the class indicated in green. Furthermore, phosphorylation values of AKT1 and AKT2 appear low, indicating NOS3 is not stimulated via AKT. **B**, gene expression of NOS3 and BDKRB2, including MAPK1 activity that acts as intermediate quality controle visualizing separation of active and inactive MAPK1.

Alternative routes regulating proliferation - other than BRAF driven mapk signaling

After inspecting the top-10% measurements of the extended gene selection no biologically relevant signaling routes were identified. Therefore we extended the gene set with another 100 genes and re-execute the random forest classification. After data acquisition and post-processing the dataset counts 147 patients and 1096 measurements among which protein activity data for 52 genes.

These top-10% measurements reveals PTEN functionality is particularly active for those showing MAPK1 activity despite absence of *BRAF* mutation p.V600E (***Figure 8A***). *PTEN* is a tumor suppressor gene from the pi3k pathway and its protein product inhibits PIP3, that in turn stimulates AKT (**Figure 7**). However, despite high inhibiting activity of PTEN phosphorylation means of AKT1, AKT2 and ATK3 are increased for this subset (***Figure 8B***). Since AKT stimulates various other protein we check activity of various AKT regulated signaling routes by inspecting FOXO, BAD, BCL2, P53 and MTOR phosphorylation values and gene expression (**Figure 7**).

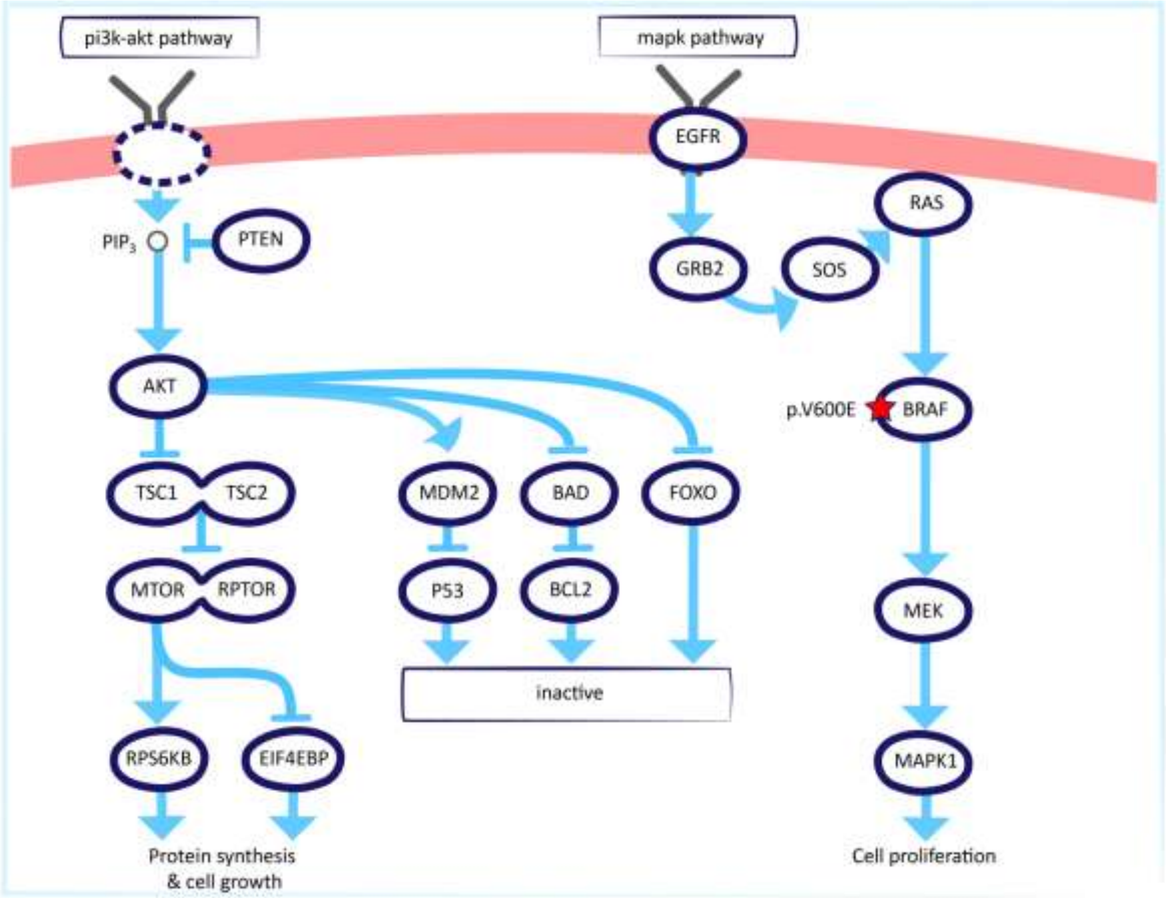


Figure 7: Alternative signaling route through MTOR signaling. Four signaling routes are inspected downstream of AKT, FOXO, BAD and MDM2 appear inactive on both gene expression and phosphorylation levels. Furthermore, increased AKT phosphorylation indicates inhibition of downstream activator TSC2, that is in complex with TSC1 and inhibits RHEB. RHEB activates MTOR and RPTOR that signals further down the mtor pathway.

Both gene expression and phosphorylation values of FOXO reflect low activity (with means around -0.5 standard deviations and activity 0, respectively) that indicates this particular signaling route is inhibited by AKT and appears inactive.

The means of *BAD* and *BCL2* gene expression are -0.5 and 0 standard deviations, respectively. Since AKT inhibits BAD, which in turn inhibits BCL2, increased AKT activity suggest BCL2 activation. It has been observed however that the majority of patients in this subset have inactive BCL2 suggesting no particular signaling further downstream of BCL2. Gene expression values of *P53* appear stable amongst all four subsets (with means slightly below 0 standard deviations) and a phosphorylation mean around -0.5, that reveals inhibition of MDM2, that in turn, is stimulated by AKT and inactivates this signaling route.

Low phosphorylation values and gene expression of TSC1 (with means -1 and -1 standard deviations, respectively) agree on the inhibitory function of AKT. TSC2 is specifically increased for this subset (gene expression mean = 0 standard deviations compared to mean = -0.5 standard deviations for other subsets) and phosphorylation values reveal activated TSC2 (mean = 0.5 compared to mean = -0.5 for inactive MAPK1). The comparison of present or absent BRAF mutation p.V600E (in cases of active MAPK1) shows increased MTOR gene expression and phosphorylation (***Figure 8C***) that is characterizing for MAPK1 activity without BRAF mutation p.V600E and found even more explicit for

RPTOR gene expression and phosphorylation data (***Figure 8D***). MTOR and RPTOR form a complex that goes on to activate RPS6KB and inhibit EIF4EBP indicated by increased and decreased phosphorylation (***Figure 8E and F***).

Recapitulating these findings; typically increased PTEN activity with consecutive downstream signaling through MTOR, RPTOR and RPS6BK is characterizing for those with active MAPK1 signaling without BRAF mutation p.V600E in TCGA melanoma datasets.

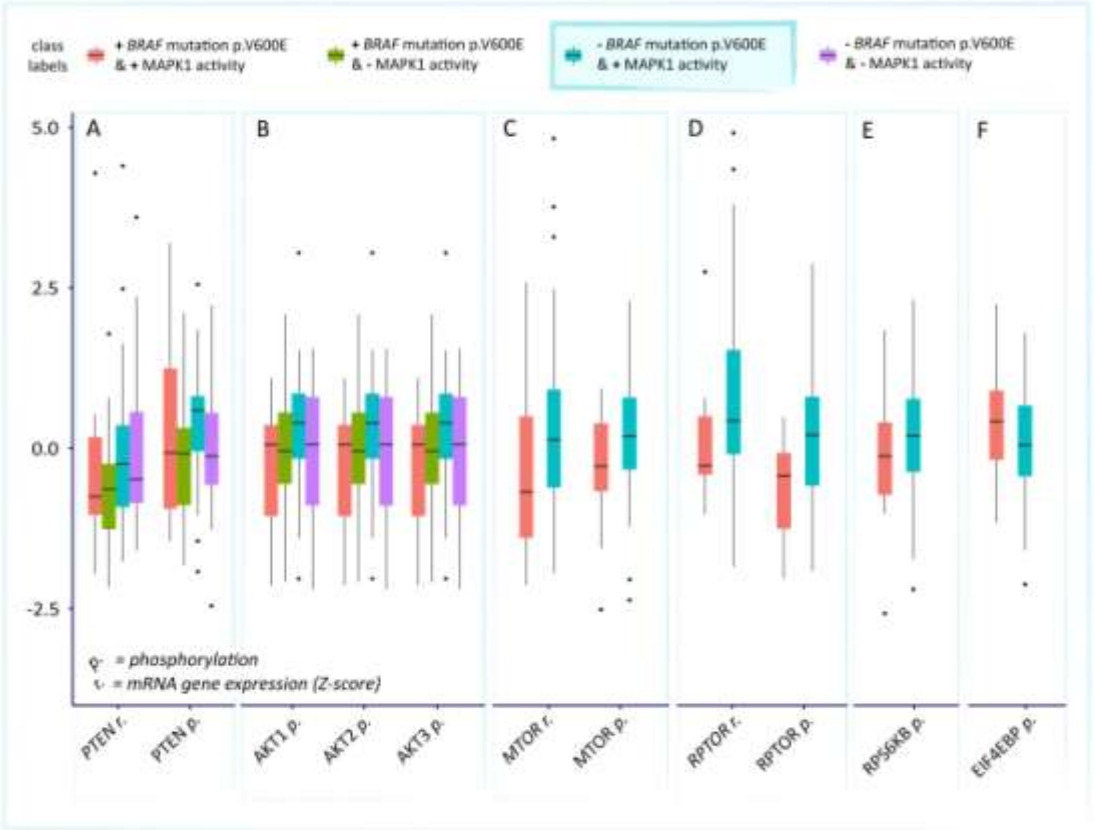


Figure 8: Gene measurement characterizing alternative signaling under PTEN inhibiting and AKT active conditions. From left to right; **A**, increased PTEN phosphorylation and gene expression values, that indicates increased PTEN activity. **B**, three units of AKT are in complex and reveal increased phosphorylation values, indicating activity despite PTEN inhibition. **C**, *MTOR* gene expression and activity values, show slightly increased. **D**, particularly gene expression increase is observed for RPTOR. **E and F,** highest and lowest phosphorylation values for downstream proteins of MTOR, activation of RPS6KB is associated to transcription and proliferation.

Comparing Random forest classification results to preliminary results CGC flagship project

Figure 9 panel A, C and D represents control samples (in grey) of *PLCB1*, *ITPR2* and *NOS3*, respectively, that are stimulated with EGFs and measured over 48 hours with t0 measured before stimulation. Red and blue indicate BRAF inhibitor Vemurafenib and EGFR inhibitor Gefitinib, respectively.

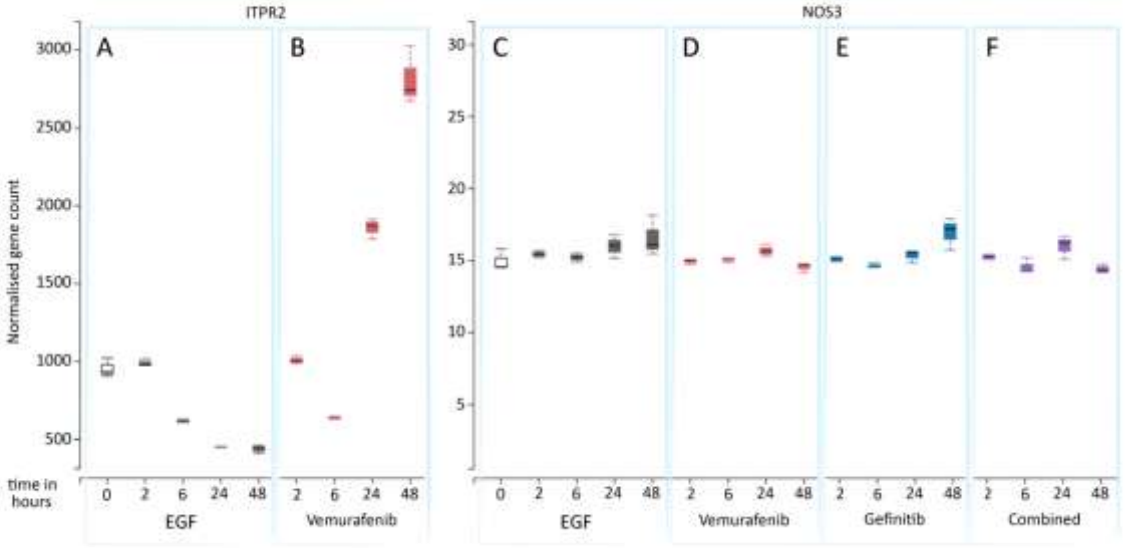


Figure 9: normalized gene expression count of genes of interest under different conditions over time. Control samples in grey are stimulated with EGF, in red cells stimulated with EGF and BRAF inhibitor Vemurafenib, in blue EGFR inhibition through Gefitinib and in purple an inhibition combination of these **A**, measurements for *ITPR2* under stimulation of EGF, **B** explicit increase of *ITPR2* gene expression through mapk activation. **C**, Control shows remarkable low gene expression of *NOS3* ***D, E and F*** that is unresponsive to any kind of inhibition.

Regardless of inhibition normalized gene counts of both *NOS3* and *BDKRB2* are remarkably low (***Figure 9C, D, E and F***), specifically BDKRB2 expression is decreased beyond meaningful visualization and therefore not shown. Hence both TCGA data (for cases of melanoma) and cell line samples derived from colon tumors show no evidence of either BDKRB2 or NOS3 gene expression.

Interestingly, decreased expression of *ITPR2* is observed under EGF stimulation and suggests low expression for MAPK1 active conditions (***Figure 9A***). Decreasing mapk signaling through BRAF inhibition causes *ITPR2* expression to increase nearly three-fold compared to original levels in colon cancers (***Figure 9B***). This is similarly observed in cases of melanoma and might indicate why patients with inactive MAPK1 show increased expression of this calcium channel subunit in TCGA data.

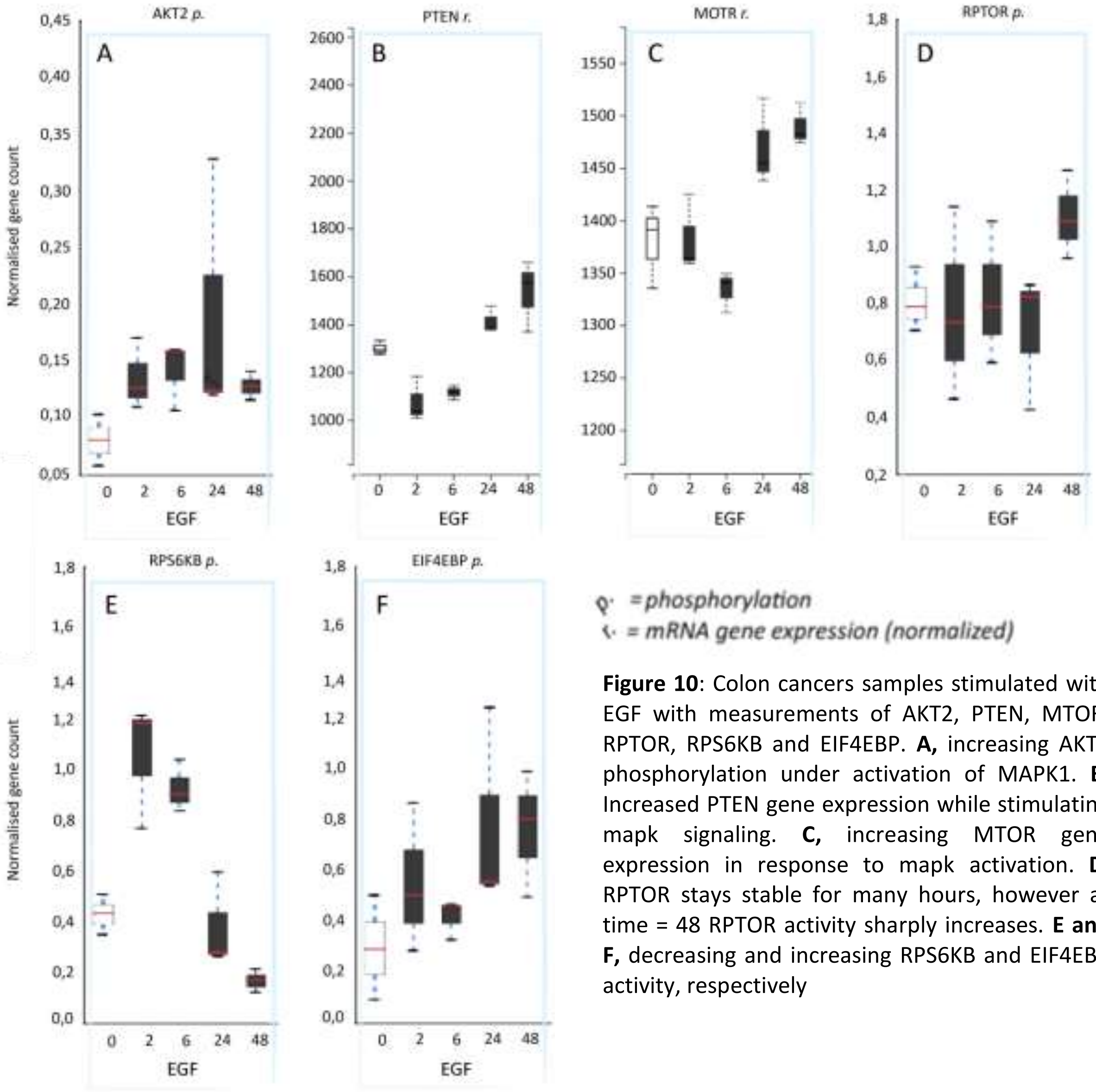


Figure 10: Colon cancers samples stimulated with EGF with measurements of AKT2, PTEN, MTOR, RPTOR, RPS6KB and EIF4EBP. **A,** increasing AKT2 phosphorylation under activation of MAPK1. **B,** Increased PTEN gene expression while stimulating mapk signaling. **C,** increasing MTOR gene expression in response to mapk activation. **D**, RPTOR stays stable for many hours, however at time = 48 RPTOR activity sharply increases. **E and F,** decreasing and increasing RPS6KB and EIF4EBP activity, respectively

Activating MAPK1 in colon tumor cells, by exposure to EGF, indicates to increase phosphorylation of AKT2 and RPTOR in addition to increased gene expression of PTEN and MTOR (***Figure 10A, B, C and D***). Thereby showing initial response of mtor signaling that is caused by mapk stimulation. However, inspecting downstream signaling of MTOR and RPTOR in colon cancers indicates decreased and increased RPS6KB and EIF4EBP phosphorylation, respectively (***Figure 10E and F***). RPS6KB gene expression reveals a strong initial response to EGF stimulation that gradually decreases over time resulting expression below levels observed before stimulation. Inhibition of RPS6KB while MTOR and RPTOR are activated indicates no further downstream signaling and suggests an inhibition mechanism that is specifically responsive to EGF stimulation in colon cancers. Furthermore the differently increased and decreased RPS6KB phosphorylation between cancer types indicate activation of protein synthesis and cell growth through mtor signaling may be specific for MAPK1 active cases of melanoma.

Conclusion and discussion

Our methods is a quick and intuitive approach to identify most important processes separating cancer types from large datasets such as TCGA. Classification of melanoma and papillary thyroid cancers shows to pick-up differences directly from patient data (e.g. differences between epigenetic regulation of MAP3K1 and EIF4G1 in addition to expression differences of ERBB2, ERBB3 and ELF3).

Here we consider patients without BRAF mutation p.V600E and active MAPK1 phosphorylation most prune to reveal alternative routes that increase MAPK1 activity independent of BRAF stimulation. Hence, during classification of melanoma patients we focus on those showing aforementioned requirements and find increased PTEN activity and evidence of downstream signaling thourgh MTOR, RPTOR and RPS6KB. Comparing melanoma gene expression and phosphorylation to CGC flagship provided of colon cancer samples shows comparable trends for AKT and RPTOR phosphorylation, however downstream signaling through EIF4EBP seems particular for cases of melanoma.

From previous work it is known that *BRAF* mutation p.V600E result in increased MAPK1 phosphorylation values, that is referred to as hyper-activity and characterizes a state that is separate from MAPK1 activity under natural conditions. Hence we directed our on/off state labeling method to uses upper and lower boundaries of 20, 30, and 40%, however, this quickly leads to class imbalances. Furthermore, this methodology results one extra MAPK1 state (in between 'on' and 'off' states) compared to our current on/off concept, the addition of an extra class introduces yet another layer of complexity that influences classification negatively (*for specific details on class distribution see **supplementary methods***). Hence, standardized phosphorylation values (either disease specific or under natural conditions) would circumvents this issue.

Alternative routes that constructs signaling cascades are specifically defined through phosphorylation events. However, our dataset contains phosphorylation values for only 10% to 20% of selected genes, therefore we expect our dataset to represent partial signaling routes (at least at the level of phosphorylation) that complicates classification. To indicate the impact of absent phosphorylation values we generate two datasets constructed from our core gene selection and expanded with 20 genes (that we use for explanatory purpose). Phosphorylation values are removed from one of the datasets and both datasets are clustered with hierarchical clustering, hereafter we compare dendrograms by randomly shuffling and repetitively verifying if their entanglement was lowered. This reveals increased clustering results for datasets including phosphorylation values and emphasizes on the importance of available protein activity data (*for specific details on comparing dendrograms and entanglement see **supplementary methods***).

Datasets generated from the core gene selection expanded by 100 genes contain 16 patients that acquired either *BRAF* mutation p.V600K or p.V600R. Hyper activation of mapk signaling is associated to *BRAF* mutation p.V600E, however, lysine (one letter code K) and arginine (one letter code R) are labeled 'hotspot' since they have similar molecule structures compared to glutamic acid.

Although random forest is able to highlight genes and gene measurements that characterize our two main classes of interest, reasons explaining elevated activity of the second messenger system or mtor signaling under AKT inhibiting and PTEN active conditions remain unclear. Here we use a semi-automated gene selection using geneMania, however our results indicate specific signaling routes to characterize subsets. Since PTEN phosphorylation levels is the strongest characterizer, we propose a more targeted approach to generate gene selection. Investigating proteins that stimulate AKT and possibly overrule PTEN inhibition, in addition to proteins activating mtor signaling under AKT activated conditions is particularly interesting. This follow-up study leads to better understanding of specific pi3k-akt regulation that can explain our contradicting results regarding mtor activation despite PTEN activity. Furthermore, targeted gene selection might reveal the feedback loop between PTEN and MAPK1 activity, that is currently not picked-up.

TCGA contains control sample data, which are partially protected from open public and therefore not available through CGDS-R, hence control data are not part of our analyses. Control data could strengthen findings of class specific deregulation (by comparing gene measurements under normal and cancerous circumstances) and might offer a solution to standardizing phosphorylation values that facilitates labeling of our on/off concept.

Interestingly, retrieving data from cbioportal requires an integer specification of requested cancer study (78 for melanoma). However, during our analysis the parameter indicating melanoma changes from 76 to 78, indicating two new studies are added to the TCGA. Warranting manual inspection of data acquisition parameters to accurately specify cancer studies.

References

1. Thomas Holbro, **The ErbB2/ErbB3 heterodimer functions as an oncogenic unit: ErbB2 requires ErbB3 to drive breast tumor cell proliferation**, PNAS 2002 vol. 100 no. 15, 8933–8938
2. Xiaoping Huang, **Heterotrimerization of the Growth Factor Receptors erbB2, erbB3, and Insulin-like Growth Factor-I Receptor in Breast Cancer Cells Resistant to Herceptin,** doi: 10.1158/0008-5472.CAN-09-3321 *Cancer Research* 2010 *70; 1204*
3. Ivan Bièche, **Analyses of *MYC*, *ERBB2*, and *CCND1* Genes in Benign and Malignant Thyroid Follicular Cell Tumors by Real-Time Polymerase Chain Reaction**, Thyroid 2001, 11(2): 147-152. doi:10.1089/105072501300042802
4. H-W Lo and M-C Hung, **Nuclear EGFR signalling network in cancers: linking EGFR pathway to cell cycle progression, nitric oxide pathway and patient survival**, British Journal of Cancer (2006) 94, 184–188. doi:10.1038/sj.bjc.6602941
5. http://www.kegg.jp/kegg-bin/show_pathway?scale=1.0&query=STAT&map=ko04010&scale=1.0&image=%2Fshare%2Fwww%2Fmark_pathway142134763128425%2Fko04010.png&auto_image=&show_description=hide&multi_query=
6. Li Liu et al., **Sorafenib Blocks the RAF/MEK/ERK Pathway, Inhibits Tumor Angiogenesis, and Induces Tumor Cell Apoptosis in Hepatocellular Carcinoma Model PLC/PRF/5,** Cancer research 2006 doi: 10.1158/0008-5472.CAN-06-1377
7. Chong Sun et al., **Reversible and adaptive resistance to BRAF(V600E) inhibition in melanoma**, NATURE 2014 VOL 508, doi:10.1038/nature13121
8. Jinhua Wang, Sharon K. Huang, Diego M. Marzese, Sandy C. Hsu, Neal P. Kawas, Kelly K. Chong, Georgina V. Long, Alexander M. Menzies, Richard A. Scolyer, Sivan Izraely, Orit Sagi-Assif, Isaac P. Witz and Dave S.B. Hoon, **Epigenetic Changes of EGFR Have an Important Role in BRAF Inhibitor–Resistant Cutaneous Melanomas,** Journal of Investigative Dermatology advance online publication, 2014; doi:10.1038/jid.2014.418
9. Keith T. Flaherty et al., **Inhibition of Mutated, Activated BRAF in Metastatic Melanoma,** *The* new england, journal *of* medicine 2010 vol. 363 no. 9
10. http://www.cbioportal.org/cross_cancer.do?cancer_study_list=&cancer_study_id=all&data_priority=0&case_ids=&gene_set_choice=user-defined-list&gene_list=BRAF&clinical_param_selection=null&tab_index=tab_visualize & Action=Submit#crosscancer/overview/0/BRAF/thca_tcga%2Cthca_tcga_pub%2Cskcm_broad%2Cskcm_broad_dfarber%2Cskcm_tcga%2Cskcm_yale
11. https://tcga-data.nci.nih.gov/datareports/statsDashboard.htm]
12. https://wiki.nci.nih.gov/display/TCGA/Data+level]
13. http://cancergenome.nih.gov/abouttcga/overview
14. https://tcga-data.nci.nih.gov/tcga/tcgaDataType.jsp]
15. Paul B. Chapman et al., **Improved Survival with Vemurafenib in Melanoma with BRAF V600E Mutation,** England journal med 2011 364;26
16. Antonio Barbieri et al., **Role of endothelial nitric oxide synthase (eNOS) in chronic stress-promoted tumour growth**, Journal of Cellular and Molecular Medicine 2012 Volume 16, Issue 4, pages 920–926
17. Lei-Lei Chen et al., **Inhibition of MAPK signaling by eNOS gene transfer improves ventricular remodeling after myocardial infarction through reduction of inflammation**, Molecular Biology Reports 2010 Volume 37, Issue 7, pp 3067-3072
18. William C. Sessa**, eNOS at a glance,** Journal Cell Science 2004 doi: 10.1242/jcs.01165 *117, 2427-2429*

19. Bing Shen et al., The Bradykinin B2 Receptor Gene Is a Target of Angiotensin II Type 1 Receptor Signaling, JASN 2007 doi: 10.1681/ASN.2006101127 *vol. 18 no. 4 1140-1149*
20. Dr. Darrell S. Rigel MD et al., **Malignant melanoma: Prevention, early detection, and treatment in the 21st century**, CA: A Cancer Journal for Clinicians 2008 DOI: 10.3322/canjclin.50.4.215
21. Yasuo Koga et al., **Genome-wide screen of promoter methylation identifies novel markers in melanoma**, Genome Research 2009 doi: 10.1101/gr.091447.109
22. Gao et al. **Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal**. *Sci. Signal.* 2013 6, pl1
23. Warde-Farley D et al., **The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function,** Nucleic Acids Res. 2010 38 Suppl:W214-20
24. R Development Core Team (2008). **R: A language and environment for statistical computing. R Foundation for Statistical Computing**, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.
25. Fernando Pérez, Brian E. Granger***, IPython: A System for Interactive Scientific Computing*, Computing in Science and Engineering**, vol. 9, no. 3, pp. 21-29, May/June 2007, doi:10.1109/MCSE.2007.53. URL: http://ipython.org
26. Kelder T et al., (2011) **WikiPathways: building research communities on biological pathways**. NAR doi: 10.1093/nar/gkr1074
27. Kanehisa, M. et al., **Data, information, knowledge and principle: back to metabolism in KEGG**. Nucleic Acids Res. 42, D199–D205 (2014)
28. www.genecrads.org

supplementary method

Gene selection

We retrieve the expanded core gene selection from geneMania by downloading a list in csv format, removing gene descriptions and check data availability in TCGA. Both Cbioporatal and geneMania require HGNC nomenclature, and both interfaces offer innate converters that provide alternative suggestions for unrecognized gene names. The majority of our selection is recognized or converted correctly, however in some rare cases both resources are unable to properly convert the gene names. For these cases we alter our trajectory to an online approachable 'Multi-symbol Checker' that is part of the HGNC community and provides summary statistics on the provided gene selection. It indicates accepted or converted gene names, in addition to genes that are withdraw from the results, gene withdrawn from the results are removed from the analysis.

Gene measurement normalization

TCGA provides mRNA expression values (coming from second-generation sequencing) either through molecule count per patient per gene or using within-sample normalized values. Both values are calculated using RNA-seq by Expectation-Maximization (better known as RSEM). Read counts are calculated by aligning reads to the reference genome, reads that are mapped to multiple locations are ought to divide their score (every reads gets score = 1) over these locations resulting fractional values. Within-sample normalization values represent Z-scores that describe the number of standard deviations that the value is distanced from the average mRNA expression value across all patients per gene for a particular type of cancer.

Normalized methylation data is calculated through a Beta-Mixture Quantile dilation algorithm and expresses values between zero and one. Zero indicating hypo-methylation of genes meaning no methyl-groups bound to promoter regions, and one indicating hyper-methylation resembling heavily silenced genes. Methylation is measured using ChipSeq technology (HM450) that contains over 450.000 CpG islands, that span over 95% of all CpG islands and covers >99% of all the genes collected in the RefSeq database.

TCGA provides CNA values in a discrete fashion and are estimated using the GISTIC algorithm, a gene can either have values -2, -1, 0, 1 or 2, indicating nullizygosity, hemizygosity, unchanged, gain or heavily amplified, respectively. And indicate homozygous deletions, loss of single alleles, normal state (two copies), one extra copy or multiple copies of the allele, respectively. TCGA is able to putatively approximate these copy number alterations due to the purity and known ploidy of samples and additionally provides CNA values in log scale that are used in our analysis.

Phosphorylation measurements are collected using fluorescent labels, known as reverse phase protein array (RPPA). This test requires prior knowledge on the protein and custom designed antibodies that target phosphorylation sites, and nearly a couple of hundred antibodies are carefully selected and used for testing. Phosphorylation data is both within-gene and within-sample normalized using four steps; 1) calculate the median of each gene across all samples, 2) subtract the median from all samples per gene, 3) calculate median of each sample across all genes, and 4) subtract the median from all genes per sample.

Mutations are collected through second-generation whole exome sequencing, and the TCGA provides these mutations per gene per patient per cancer study. Mutations are coded similar to V600E (*see introduction - Deduce MAPK functionality through various biological levels*). Moreover, TCGA provides effects for several kinds of mutations through a suffix, such as; '*', 'fs' or '_splice' at the end of mutation names. Asterisk indicates nonsense mutations, 'fs' stands for frameshift and '_splice' for splice site, that we use for data discretization.

Class distributions

Varying phosphorylation threshold values between 0.5 and 0.3 results different class distributions that, for values 0.3 and 0.4 results in a rather large group with 'intermediate' phosphorylation state. It is important to understand decreased error rates for smaller phosphorylation threshold values are due to misclassification of almost all patients. Since the groups with label 'intermediate' is large, our classification simply suggests to classify all patients as the largest group, and since patients that are assigned class ID's 1 until 4 are low in numbers (for threshold value = 0.3), the error rate appears low. However this classification is not targeted towards separating the five groups, hence not biologically relevant. Using phosphorylation threshold value 0.5 results exactly four classes and no patients are considered intermediate, although this results the highest classification error rates, resulting gene measurements indicate trends that separate either of the four classes in a biologically relevant setting.

Table S1: Class distributions and corresponding error rates of phosphorylation threshold values

Phosphorylation threshold values	**0.3**	**0.4**	**0.5**

Class ID	**Patient count**		
1	1	5	22
2	11	23	45
3	8	20	44
4	9	25	47
Intermediate class	129	85	0
Error rates (%)	18,35	46,2	62,66

Unsupervised learning – PCA and hierarchical clustering

We visualize data spread using principle component analysis and reveal loadings of every component separately and observe loadings continuous decrease indicating high complexity or noisy data (see Figure S1 panel A). Top ranking gene measurements are angled towards multiple directions, indicating data is widely spread (see Figure S1 panel B), and most importantly, no prior knowledge is provided that facilitates principle component analysis. Therefore, directing classification towards biologically relevant groups, such as, patients simulating drug resistance, requires specific subset procedures for classes separately.

Additionally, since dataset volumes are relatively small classification difficulties arise due to increased gene selections. Principle component analysis classifies every patient based on the measurements that are provided, however increasing the number of measurements for small sets of patients results empty multidimensional space that limits separable power. Hence, for ideal principle component analysis data sets contain larger numbers of patients compared to gene measurements. Although similar results are obtained for larger subsets of our data, for explanatory purposes we use our core gene selection to visualize principle component analysis results.

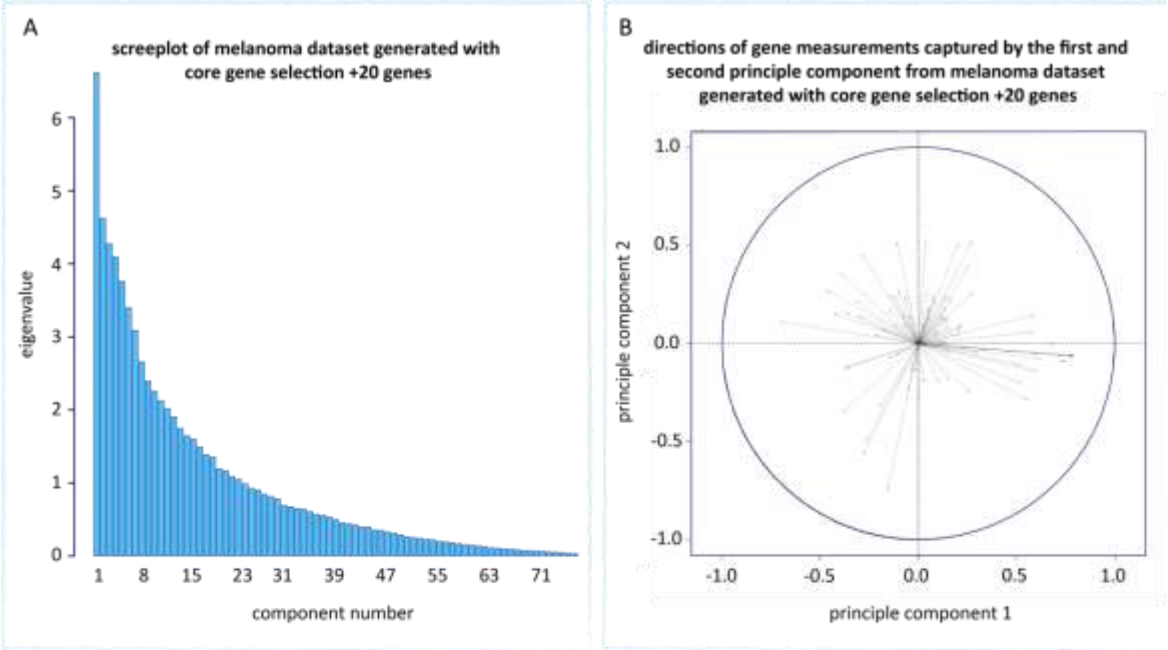


Figure S1, panel A: data spread captured with principle component analysis, every eigenvalue is plotted and indicates all eigenvalues are used to describe the spread of the data. **Panel B:** Directions of gene measurements as a result of PCA, here we observe gene measurement data is spread in many directions, making it difficult to find characterizing trends of gene measurements for particular subsets.

Clustering our core gene selection increased with 100 genes results difficult to interpret clusters and hierarchical trees. We cluster data sets with hierarchical clustering, using Euclidean space distance metric, and define clusters using dynamic tree cutting. This algorithm cuts dendrograms and stabilizes through iterating adaptive protocols of cluster decomposition. Without subsets patients that with *BRAF* mutation p.V600E appear scattered across the leafs (see Figure S2 panel A, p.V600E in red, other mutation types in blue, no BRAF alteration in grey), and gene measurements clustering appears noisy.

Furthermore, clustering patients without BRAF alteration or mutations other than p.V00E reveals noisy gene measurement clustering and patients appear clustered in multiple small groups. This indicates data similarities does not characterizes this subset profoundly (see Figure S2 panel B). Additionally, we subset patients with active MAPK1 despite absence of *BRAF* mutation p.V600E, however, we observe a similar patient clustering of smaller groups and noisy gene measurements (see Figure S2 panel C patients are labeled red, and the dendrogram reveals many small groups). This indicates patient-specific characterization through hierarchical clustering requires specific subsets to narrow down noise and mixed-in signals that are present in our dataset. Finally, we have performed clustering using a variety of discretization methods, that included CNA values in discrete fashion, mRNA expression values in molecule counts and several threshold values for our on/off state concept, however, none of these discretization methods seemed to improve patient or gene measurement clustering, or any other of our classifications methods.

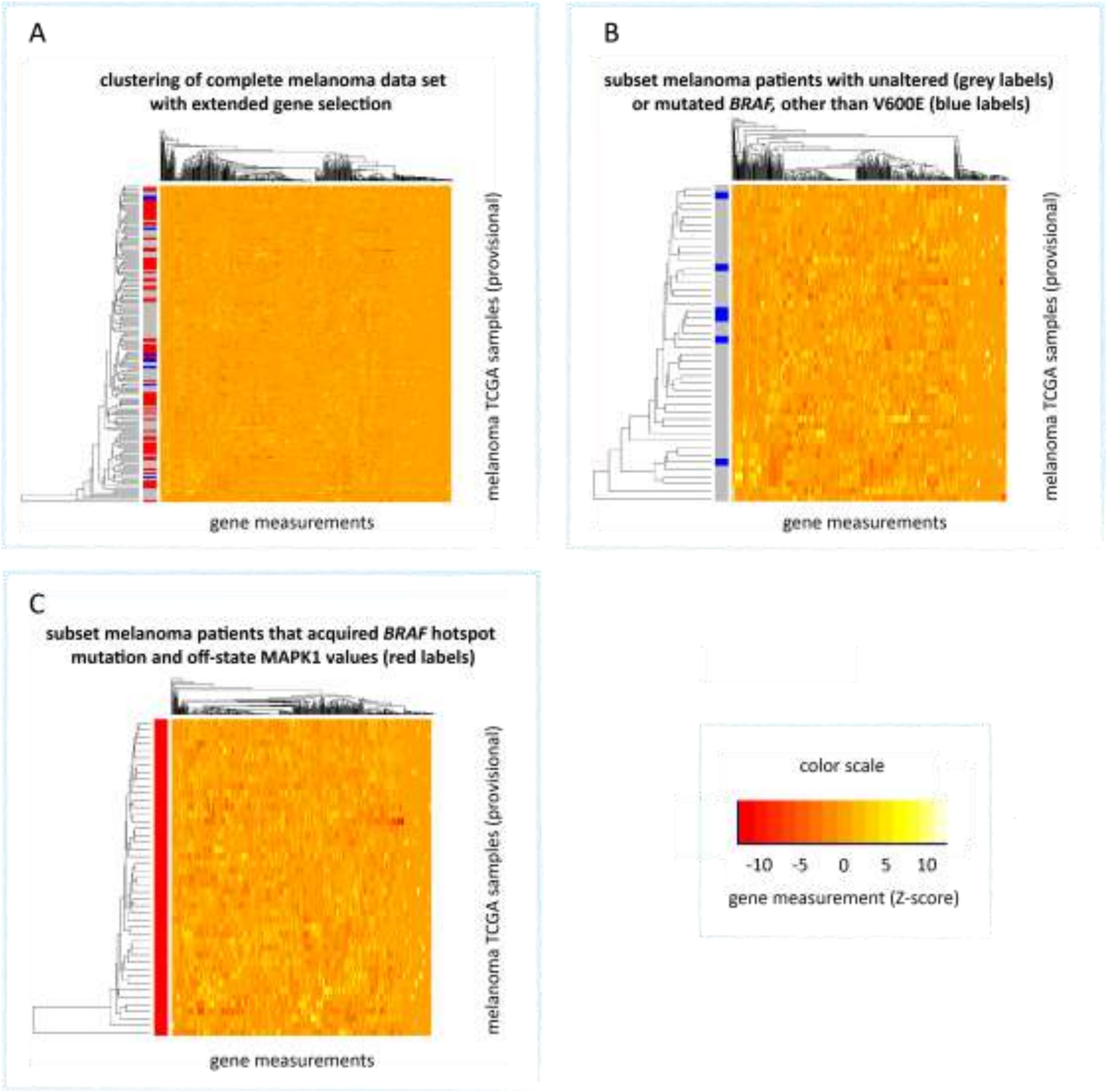


Figure S2, panel A: clustering of complete melanoma data set, colors at left hand side indicate patients without *BRAF* mutation (grey), *BRAF* mutation p.V600E (red) and missense mutations (blue). **Panel B:** clustering of subset from patients with mutations other than p.V600E or no mutation. **Panel C:** clustering of subset from patients with *BRAF* mutation p.V600E exclusively (all patients are labeled red)

Importance of phosphorylation in signal transduction classification

Phosphorylation is a key player in signaling transduction, nonetheless, values represent a minority measurement in our dataset, we are therefore interested to inspect the impact of absent phosphorylation values. Although similar results are obtained for larger gene selections, for the sake of simplicity, we use our core gene selection plus 20 additional genes for melanoma cases. We post-process data as described in the method section and generate data sets that include and exclude phosphorylation values (we remove phosphorylation values of 19 genes). Hereafter we perform hierarchical clustering on both data sets separately and cut clusters dynamically. To compare clustering results we use R package ‘dendextendRcpp’, that provides a set of functions to compare dendrograms by shuffling parts of the dendrogram randomly until researching an optimal ‘untanglement’ state, that maximally resolves dendrogram entanglement.

Clustering and dynamic cutting of data that includes phosphorylation values results three clusters and one outlier (see Figure S3 left-side in different shades of green, and one patient at the bottom representing a single cluster). Additionally, labeling patients with *BRAF* mutation p.V600E reveals three distinct clusters for datasets including phosphorylation values (see Figure S3 left-hand side vertical bars in red represents clusters of patients with *BRAF* mutation p.V600E), that contrast to clustering defined by dynamic cutting. The dataset excluding phosphorylation values results a different clustering, although dynamic cutting results the same number of clusters (three clusters and one outlier at the bottom), the dendrogram is structured differently. Untanglement score = 0.23, that is expressed on a scale from zero to one (zero indicating completely untangled and one indicates high entanglement) indicates rather resolved dendrograms. Furthermore, BRAF mutation clusters observed previously appear scattered for absent phosphorylation values, and indicates a negative impact on clustering. Taken together decreased clustering capacity through absence of activity measurements and difficulties around tree cutting indicate the problematic setup for biological data, that we circumvent with random forest.

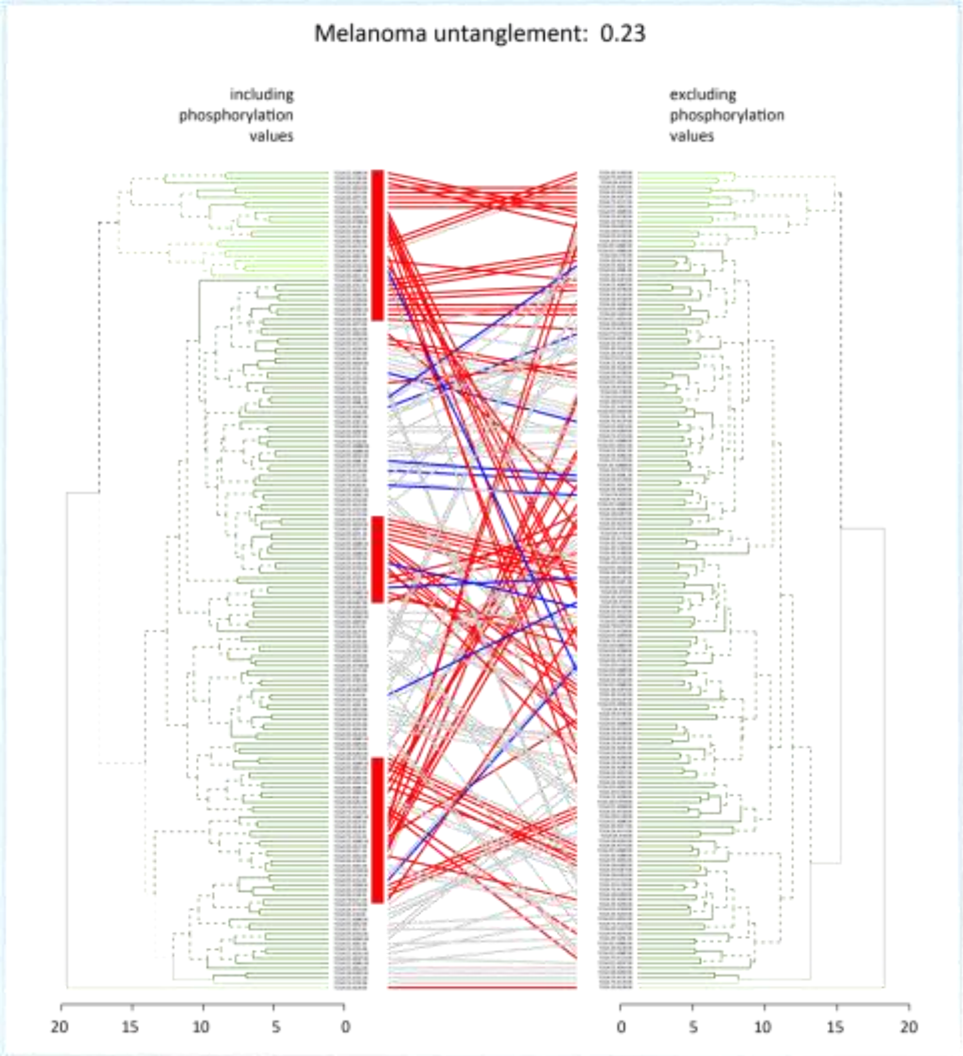


Figure S3: Untanglement of datasets with and without phosphorylation values, at the left hand side three distinct clusters are observed, that are lost when removing phosphorylation values. Untanglement = 0.23 reveals a nearly resolved comparison of both clustering trees.