RESEARCH

Sensitivity analysis based ranking reveals unknown biological hypotheses for down regulated genes in time buffer during administration of PORCN-WNT inhibitor ETC-1922159 in CRC

shriprakash sinha*

Aspects of unpublished work were presented in a poster session at the recently concluded first ever Wnt Gordon Conference, from 6-11 August 2017, held in Stowe, VT 05672, USA.

Abstract

Background: The search and wet lab testing of unknown/unexplored/untested biological hypotheses in the form of combinations of various intra/extracellular factors that are involved in a signaling pathway, costs a lot in terms of time, investment and energy. Currently, a major problem in biology is to cherry pick the combinations based on expert advice, literature survey or guesses to investigate a particular combinatorial hypothesis.

Methods: In a recent development of the PORCN-WNT inhibitor ETC-1922159 for colorectal cancer, a list of down-regulated genes were recorded in a time buffer after the administration of the drug. The regulation of the genes were recorded individually but it is still not known which higher (≥ 2) order interactions might be playing a greater role after the administration of the drug. In order to reveal the priority of these higher order interactions among the down-regulated genes or the likely unknown biological hypotheses, a search engine was developed based on the sensitivity indices of the higher order interactions that were ranked using a support vector ranking algorithm and sorted.

Results: For example, LGR family (Wnt signal enhancer) is known to neutralize RNF43 (Wnt inhibitor). After the administration of ETC-1922159 it was found that using HSIC (and rbf, linear and laplace variants of kernel) the rankings of the interaction between LGR5-RNF43 were 61, 114 and 85 respectively. Rankings for LGR6-RNF43 were 1652, 939 and 805 respectively. The down-regulation of LGR family after the drug treatment is evident in these rankings as it takes bottom priorities for LGR5-RNF43 interaction. The LGR6-RNF43 takes higher ranking than LGR5-RNF43, indicating that it might not be playing a greater role as LGR5 during the Wnt enhancing signals. These rankings confirm the efficacy of the proposed search engine design.

Conclusion: Prioritized unknown biological hypothesis form the basis of further wet lab tests with the aim to reduce the cost of (1) wet lab experiments (2) combinatorial search and (3) lower the testing time for biologist who search for influential interactions in a vast combinatorial search forest. From in silico perspective, a framework for a search engine now exists which can generate rankings for n^{th} order interactions in Wnt signaling pathway, thus revealing unknown/untested/unexplored biological hypotheses and aiding in understanding the mechanism of the pathway. The generic nature of the design can be applied to any signaling pathway or phenomena under investigation where a prioritized order of interactions among the involved factors need to be investigated for deeper understanding. Future improvements of the design are bound to facilitate medical specialists/oncologists in their respective investigations.

Keywords: Combinatorial forest; Sensitivity analysis; Support vector ranking algorithm; Unknown biological hypotheses; PORCN-WNT inhibitors; ETC-1922159

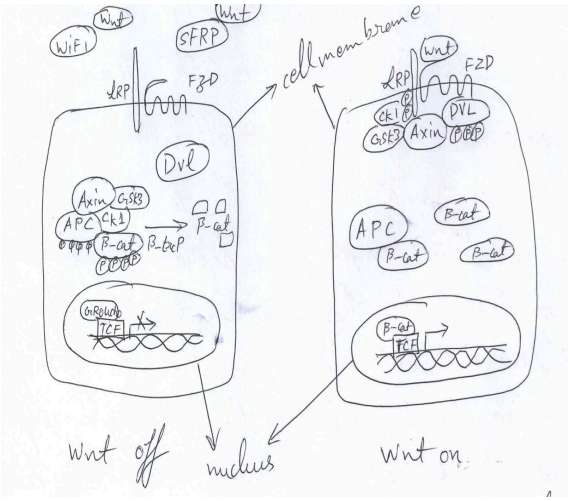

Figure 1 Cartoon of Wnt Signaling.

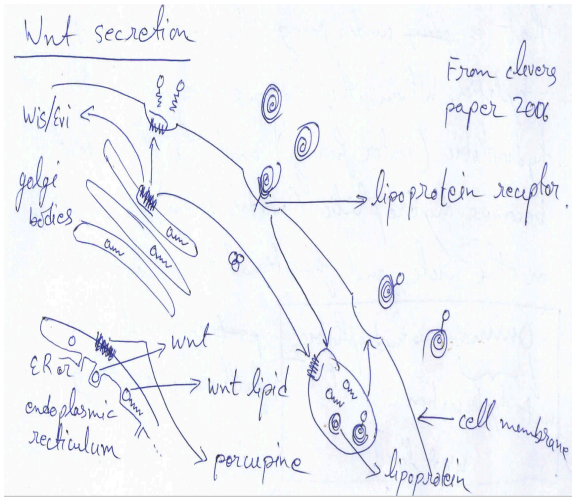

Figure 2 Cartoon of Wnt Secretion.

1 Background

Significance

Recent development of PORCN-WNT inhibitor ETC-1922159 cancer drug has lead to suppression of tumor in a specific type of colorectal cancer. After the administration of the drug, a list of genes were observed to know the affect of the drug. Down and up regulated list of genes have been provided but it is not known at the higher order ($\geq$ 2) level, which combination of these genes might be influential. A search engine has be developed to prioritise and reveal these unknown/untested/unexplored combinations to reduce the cost of wet lab tests in terms of time/investment/energy. These ranked biological hypotheses facilitate biologists to narrow down their investigation in a vast combinatorial search forest.

Wnt signaling and secretion

[1]'s accidental discovery of the Wingless played a pioneering role in the emergence of a widely expanding research field of the Wnt signaling pathway. A majority of the work has focused on issues related to • the discovery of genetic and epigenetic factors affecting the pathway [2] & [3], • implications of mutations in the pathway and its dominant role on cancer and other diseases [4], • investigation into the pathway's contribution towards embryo development [5], homeostasis [6] & [7] and apoptosis [8] and • safety and feasibility of drug design for the Wnt pathway [9], [10], [11], [12] & [13].

The Wnt phenomena can be roughly segregated into signaling and secretion part. The Wnt signaling pathway works when the WNT ligand gets attached to the Frizzled(FZD)/LRP coreceptor complex. FZD may interact with the Dishevelled (DVL) causing phosphorylation. It is also thought that Wnts cause phosphorylation of the LRP via casein kinase 1 (CK1) and kinase GSK3. These developments further lead to attraction of Axin which causes inhibition of the formation of the degradation complex. The degradation complex constitutes of AXIN, the β-catenin transportation complex APC, CK1 and GSK3. When the pathway is active the dissolution of the degradation complex leads to stabilization in the concentration of β-catenin in the cytoplasm. As β-catenin enters into the nucleus it displaces the GROUCHO and binds with transcription cell factor TCF thus instigating transcription of Wnt target genes. GROUCHO acts as lock on TCF and prevents the transcription of target genes which may induce cancer. In cases when the Wnt ligands are not captured by the coreceptor at the cell membrane, AXIN helps in formation of the degradation complex. The degradation complex phosphorylates β-catenin which is then recognised by F BOX/WD repeat protein β-TRCP. β-TRCP is a component of ubiquitin ligase complex that helps in ubiquitination of β-catenin thus marking it for degradation via the proteasome. A cartoon of the signaling transduction snapshot is shown in figure 1.

Contrary to the signaling phenomena, the secretion phenomena is about the release and transportation of

Correspondence: sinha.shriprakash@yandex.com
Independent Researcher
Full list of author information is available at the end of the article

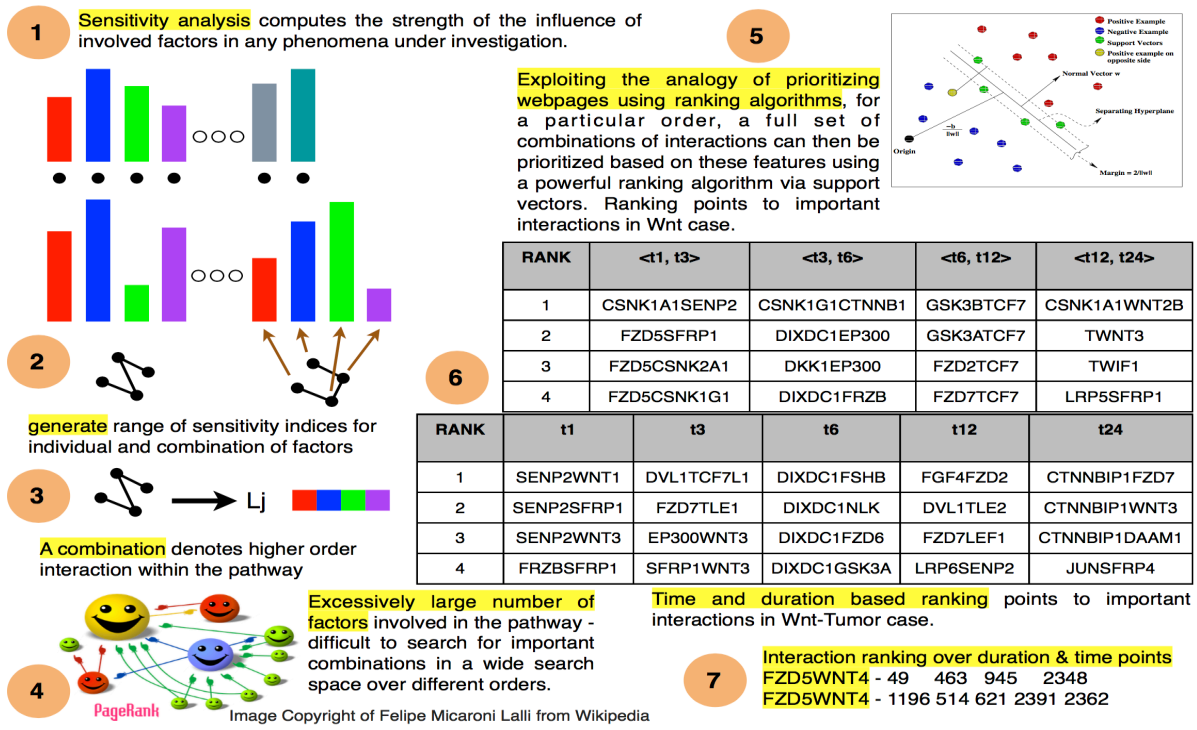

Figure 3 Graphical abstract of how the whole search engine works to find out crucial interactions within the components of the pathway under consideration as different time points and durations [14].

the WNT protein/ligand in and out of the cell, respectively. Briefly, the WNT proteins that are synthesized with the endoplasmic reticulum (ER), are known to be palmitoyleated via the Porcupine (PORCN) to form the WNT ligand, which is then ready for transportation [15]. It is believed that these ligands are then transported via the EVI/WNTLESS transmembrane complex out of the cell [16] & [17]. The EVI/WNTLESS themselves are known to reside in the Golgi bodies and interaction with the WNT ligands for the later's glycosylation [18] & [19]. Once outside the cell, the WNTs then interact with the cell receptors, as explained in the foregoing paragraph, to induce the Wnt signaling. Of importance is the fact that the EVI/WNTLESS also need a transporter in the from of a complex termed as Retromer. A cartoon of the signaling transduction snapshot is shown in figure 2.

PORCN-WNT inhibitors

The regulation of the Wnt pathway is dependent on the production and secretion of the WNT proteins. Thus, the inhibition of a causal factor like PORCN which contributes to the WNT secretion has been proposed to be a way to interfere with the Wnt cascade,

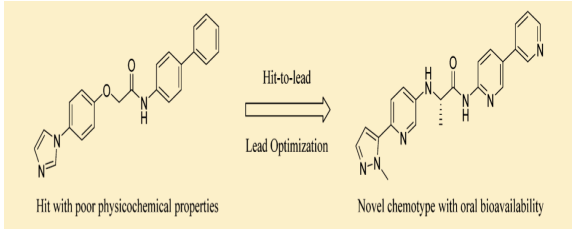

Figure 4 Hit to lead of ETC. Reprinted (adapted) with permission from (Duraiswamy, A.J., et al.; Discovery and optimization of a porcupine inhibitor. Journal of medicinal chemistry 58(15), 5889–5899 (2015)). Copyright (2015) American Chemical Society. [24]

which might result in the growth of tumor. Several groups have been engaged in such studies and known PORCN-WNT inhibitors that have been made available till now are IWP-L6 [20] & [21], C59 [22], LGK974 [23] and ETC-1922159 [24]. In this study, the focus of the attention is on the implications of the ETC-1922159, after the drug has been administered. The drug is a enantiomer with a nanomolar activity and excellent bioavailability as claimed in [24]. Figure 4 shows the hit to lead stabilization of the ETC enantiomer.

```
> source("extractETCdata.R")
> library(sensitivity)
> source("manuscript-2-2.R")
Should i generate distribution of data [y/n] - y
Choose a file to process [1/2]

         1 - ../data/onc2015280x2-A.txt

             Genes down-regulated after ETC-159 treatment

         2 - ../data/onc2015280x2-B.txt

             Genes up-regulated after ETC-159 treatment

         File number - 1
(A) Genes down-regulated after ETC-159 treatment
Genesymbol+ENSEMBLgeneID+Genedescription+log2foldchange(VEH/ETC)+BH-adjustedP-value
RRM2+ENSG00000171848+ribonucleotide  reductase  M2  [Source:HGNC  Symbol;Acc:
10452]+4.93+1.2E-162
MKI67+ENSG00000148773+marker  of  proliferation  Ki-67  [Source:HGNC  Symbol;Acc:
7107]+4.35+1.0E-142
CLDN2+ENSG00000165376+claudin 2  [Source:HGNC  Symbol;Acc:2041]+5.03+2.3E-130
.
.
# followed by list of genes
.
ACO1 SF3B5 FAM117A FTSJ3 XX LMAN1 MED28 XX CEP76 ZNF629 ERAL1 KIAA1143 KIF13A
SMARCA5 PTRH2 POLR1D MRC2 XX ZNF691 HSPA4 LSM3 IFT27 PIFO COCH ZC2HC1C PPA2 FAM98A
C1QTNF9B-AS1 NKD1 TARBP2 SAYSD1

Enter name of gene to be evaluated - RAD51AP1
nCk - choosing k - 2
choosing  2  out of the 2744  genes!
---
Types of SA -  Fdiv.TV Fdiv.KL Fdiv.Chi2 Fdiv.Hellinger HSIC.rbf HSIC.linear
HSIC.laplace SB.2002 SB.2007 SB.jansen SB.martinez SBL
Enter a type of SA - rbf
generating sample combinations!
generating for sample -  1
computing estimate different indices ...
HSIC SA - rbf kernel
generating for sample -  2
computing estimate different indices …
.
.
# till computation for 50 samples are done.
>
```

order-2-SA-rbf-RAD51AP1-ranking-mean-DR.txt is generated after the completion of SVM-Ranking algorithm [8] on the right side.

```
> source("SVMRank-Results-S-mean.R")
Choose a file to process [1/2]

         1 - ../data/onc2015280x2-A.txt

             Genes down-regulated after ETC-159 treatment

         2 - ../data/onc2015280x2-B.txt

             Genes up-regulated after ETC-159 treatment

         File number - 1
pick a numeric for k in nCk - 2
Please enter the name of the gene to be processed - RAD51AP1
Please enter the name of the proposed SA from above list - rbf
Reading training examples...done
Training set properties: 2 features, 1 rankings, 2743 examples
NOTE: Adjusted stopping criterion relative to maximum loss: eps=3760.653000
Iter 1: .*(NumConst=1, SV=1, CEps=3760653.0000, QPEps=0.1620)
Iter 2: .*(NumConst=2, SV=2, CEps=3082595.0557, QPEps=0.2183)
.
.
Iter 17140: .*(NumConst=227, SV=182, CEps=17787.6300, QPEps=72767.8711)
Iter 17141: .(NumConst=227, SV=182, CEps=2802.6439, QPEps=72767.8711)
Final epsilon on KKT-Conditions: 72767.87109
Upper bound on duality gap: 291880.04091
Dual objective value: dval=55521783.89221
Primal objective value: pval=55813663.93311
Total number of constraints in final working set: 227 (of 17140)
Number of iterations: 17141
Number of calls to 'find_most_violated_constraint': 17141
Number of SV: 182
Norm of weight vector: |w|=4.41106
Value of slack variable (on working set): xi=2790302.46990
Value of slack variable (global): xi=2790682.71022
Norm of longest difference vector: ||Psi(x,y)-Psi(x,ybar)||=69552.50149
Runtime in cpu-seconds: 571.77
Compacting linear model...done
Writing learned model...done
Reading model...done.
Reading test examples...done.
Classifying test examples...done
Runtime (without IO) in cpu-seconds: 0.00
Average loss on test set: 0.3000
Zero/one-error on test set: 100.00% (0 correct, 1 incorrect, 1 total)
NOTE: The loss reported above is the fraction of swapped pairs averaged over
      all rankings. The zero/one-error is fraction of perfectly correct
      rankings!
Total Num Swappedpairs  : 1128107
Avg Swappedpairs Percent:  30.00
sorting in ascending order - top is lowest in ranking
>
```

Figure 5 Translation of a modification of the pipeline in figure 3 into code of execution in R.

Revealing higher order biological hypotheses via sensitivity analysis and in-silico ranking algorithm

In the trial experiments on ETC-1922159 [25], a list of genes (2500±) have been reported to be up and down regulated after the drug treatment and a time buffer of 3 days. Some of the transcript levels of these genes have been recorded and the experimental design is explained elaborately in the same manuscript. In the list are also available unknown or uncharacterised proteins that have been recorded after the drug was administered. These have been marked as "- -" in the list (Note - In this manuscript these uncharacterised proteins have been marked as "XXM", were $M = 1, 2, 3, ...$). The aim of this work is to reveal unknown/unexplored/untested biological hypotheses that form higher order combinations. For example, it is known that the combinations of WNT-FZD or RSPO-LGR-RNF play significant roles in the Wnt pathway. But the $n \geq 2, 3, ...$-order combinations out of $N(> n)$ genes forms a vast combinatorial search forest that is extremely tough to investigate due to the humongous amount of combinations. Currently, a major problem in biology is to cherry pick the combinations based on expert advice, literature survey or random choices to investigate a particular combinatorial hypothesis. The current work aims to reveal these unknown/unexplored/untested combinations by prioritising these combinations using a potent support vector ranking algorithm [26]. This cuts down the cost in time/energy/investment for any investigation concerning a biological hypothesis in a vast search space.

The pipleline works by computing sensitivity indicies for each of these combinations and then vectorising these indices to connote and form discriminative feature vector for each combination. The ranking algorithm is then applied to a set of combinations/sensitivity index vectors and a ranking score is generated. Sorting these scores leads to prioritization of the combinations. Note that these combinations are now ranked and give the biologists a chance to narrow down their focus on crucial biological hypotheses in the form of combinations which the biologists might want to test. Analogous to the webpage search engine, where the click of a button for a few key-words leads to a ranked list of web links, the pipeline uses sensitivity indices as an indicator of the strength of the influence of factors or their combinations, as a criteria to rank the combinations. The generic pipleline for the generation for ranking has been shown in figure 3 from one of the author's unpublished/submitted manuscript [14],

the time series data set for which was obtained from [27].

Translating the pipeline into code of execution in R

The pipeline depicted in figure 3 was modified and translated into code in R programming language [28]. Figure 5 shows one such computation and the main commands and some of the internal executions are shown in red. The execution begins by some preprocessing of the file containing the information regarding the down regulated genes via $source("extractETCdata.R")$. The library containing the sensitivity analysis methods is then loaded in the interface using the command $library(sensitivity)$ [29] & [30]. Next a gaussian distribution for a particular transcript level for a gene is generated to have a sample. This is needed in the computation of sensitivity index for that particular gene. A gene of particular choice is selected (in blue) and the choice of the combination is made (here k = 2). Later a kernel is used (here rbf for HSIC method). 50 different indices are generated which help in generating aggregate rank using these 50 indices. The Support vector ranking algorithm is employed to generate the scores for different combinations and these scores are then sorted out. This is done using the command $source("SVMRank-Results-S-mean.R")$. A file is generated that contains the combinations in increasing order of influence after the ETC-1922159 has been administered. Note that 1 means lowest rank and 2743 is the highest rank for 2^{nd} order combinations for gene RAD51AP1 using HSIC-rbf density based sensitivity index. The code has been depicted in figure 5.

Interpretation of 2^{nd} order ranking with RAD51AP1 using HSIC-rbf density based sensitivity index

For example, the ranking generated for second order combination for RAD51AP1 is enlisted below. What these rankings suggests is that after the treatment of ETC-1922159, genes along with their combination with RAD51AP1 were prioritised and this gives the biologists a clue to study the behaviour of RAD51AP1 along with other genes in colorectal cancer case.

1 HOXB9-RAD51AP1
2 NASP-RAD51AP1
3 MCM3-RAD51AP1
4 ASF1B-RAD51AP1
5 ZWINT-RAD51AP1
6 RAD51AP1-RETNLB
7 CDK1-RAD51AP1
8 AHCY-RAD51AP1
9 PTPLAD1-RAD51AP1
•
•
2741 RAD51AP1-CDX2
2742 RAD51AP1-LPCAT3
2743 RAD51AP1-EIF2B1

DNA repair is an important aspect in maintaining the proper and healthy functioning for the cells in the human body. Failure in DNA repair process can lead to aberrations as well as tumorous stages. There are various types of damages that a DNA can go through, one of which is the DNA double strand breaks (DSB) that can be repaired via homologous recombination (HR). RAD51 plays a central role in HR and is known to function in the three phases of HR namely : presynapsis, synapsis and post-synapsis [31]. Recently, RAD51 has been implicated as a negative/poor prognostic marker for colorectal adenocarcinoma and has been found to be highly expressed [32]. A negative/poor prognostic marker indicates that it is harder to control the malignancy. Since RAD51 helps in the repair of the DNA damage via HR and is implicated as a poor prognostic marker in colorectal adenocarcinoma, this suggests its functionality in maintaining genomic stability and therapeutic resistance to cancer drugs. Mechanistically, RAD51AP1 facilitates RAD51 during the repairing process by binding with RAD51 via two DNA binding sites, thus helping in the D-loop formation in the HR process [33] & [34].

In context of the ETC-1922159 treatment, a series of 2^{nd} order rankings have been generated for RAD51AP1. Using the above pipeline can help decipher biological implications. BRCA2 is known to be a main mediator in the HR process along with RAD51 and interact with RAD51 in various ways [35], [36], [37] & [38]. A relative low rank of 458 between RAD51AP1-BRCA2 states that after the ETC-1922159 treatment the combination gets low priority indicating that as there is suppression of CRC, the genomic stability of the cancer cells is affected by rendering the DNA repairing capacity of RAD51AP1-BRCA2 ineffective to a certain extent. PPA2 is a tumor suppressor that regulates many signaling pathways and plays crucial role in cell transformation [39]. PPA2 in known to be highly suppressed in colorectal cancer case [40]. The relatively high ranking of 2675 for RAD51AP1-PPA2 combination indicates two points (a) the ineffectiveness of RAD51AP1 to provide genomic stability to cancer cell and (b) the over expression of PPA2 after ETC-1922159 was administered. Also, SET is a known PPA2 inhibitor and is observed to be highly overexpressed in colorectal cancer cases. In relation to

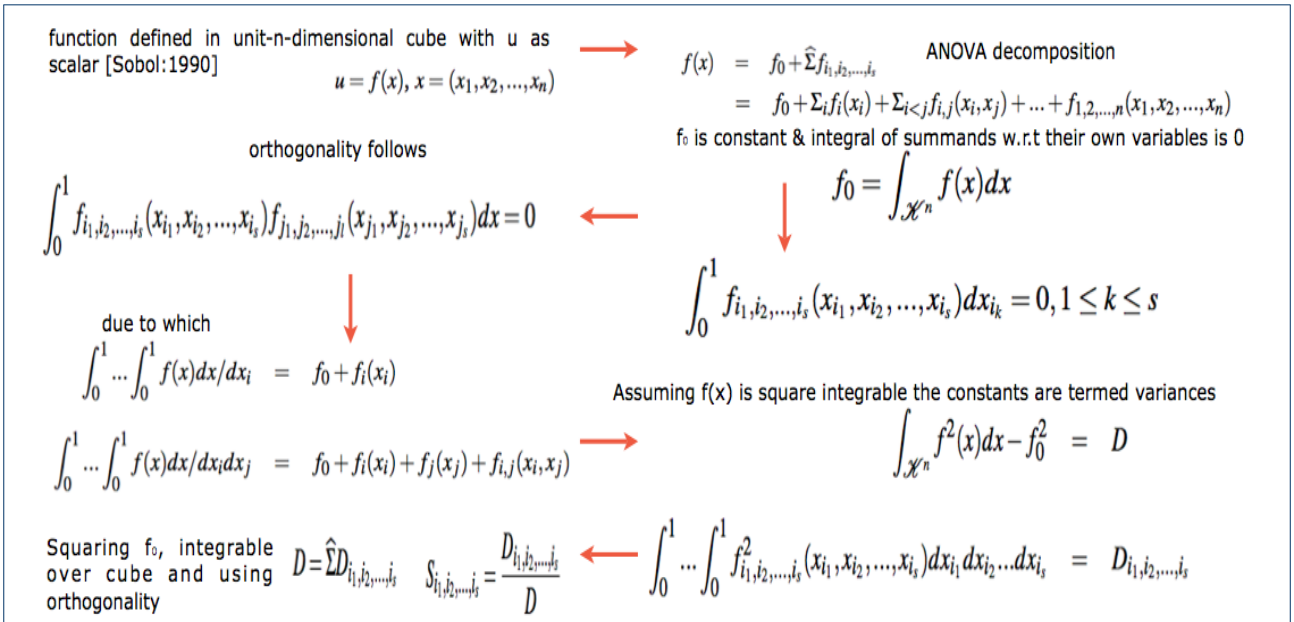

Figure 6 Computation of variance based sobol sensitivity indices.

the ranking of RAD51AP1-PPA2, after the administration of the drug, RAD51AP1-SET showed a lower rank of 2227.

The x-ray repair cross complementing XRCC family is known to work as a mediator or stabilizer for RAD51 during the HR process [31]. The relatively low rank of 366 for the RAD51AP1-XRCC2 indicates that the effectiveness for DNA repair capability is affected by ETC-1922159 drug as the tumor growth is suppressed. This is further confirmed by the fact that XRCC2 forms complex with paralogues of RAD51, i.e RAD51C–RAD51D–XRCC2 for DNA repair [41] & [42]. Other members of XRCC like XRCC1, XRCC6BP1 and XRCC6, showed relatively higher rankings of 1874, 2398 and 2558. Not much is known about these 3 factors in colorectal cancer case. XRCC1 may be weakly implicated in the colorectal cancer cases as have been found in the case studies in Taiwan [43]. Very little is known about XRCC6BP1 and in a risk score based analysis it was found that XRCC6BP1 acts as a tumor repressor with very low expression profile in colorectal cancer [44]. Hight rankings suggests that after ETC-1922159 treatment, the expression level of XRCC6BP1 is extremely high, indicating the establishment of genomic stability of health via suppression of cancer cells. Finally, the very high priority of XRCC6 only indicates its role as a tumor repressor and further wet lab analysis needs to be conducted to verify the effectiveness of the pipeline.

RNF43/ZNRF3 are known to negatively regulate the Wnt pathway by targeting the FZD family and leading to its degradation and thus acting as a hinderance to the Wnt pathway [45]. In presence of members from RSPO and LGR family, the RNF43/ZNRF3 is degraded in the cell and this leads to enhancement of the Wnt signaling [46]. In relation to the RAD51AP1, the ranking of RNF43 was found to be 1893 and the ranking of ZNRF3 was found to be 549, respectively, after the treatment of ETC-1922159. The lower ranking of ZNRF43 with RAD51AP1 might indicate some affinity between RAD51AP1-ZNRF43 in comparison to the high ranking of RAD51AP1-RNF43. This needs to be tested. LGR5 and LGR6 had associated low ranks for 327 and 161, respectively. After the administration of the drug, the LGR5, RNF43 and ZNRF43 were known to be downregulated [25]. These are evident from the low rankings in combination with RAD51AP1, except for RNF43. An implication of the above combination of rankings can be the fact that as the drug takes its affect, it has multiple consequences wherein the RNF43/ZNRF3 along with LGR5 is being degraded so that the signaling is inhibited and also the genomic instability of the cancer cells in instigated by the ineffectiveness of RAD51AP1 to help in DNA repair in cancer cells.

Such rankings hold promise for any biologists who is investigating a pathway for a particular phenomena and is faced with a vast combinatorial search forest. These rankings also provide confirmatory results for existing published wet lab affirmations.

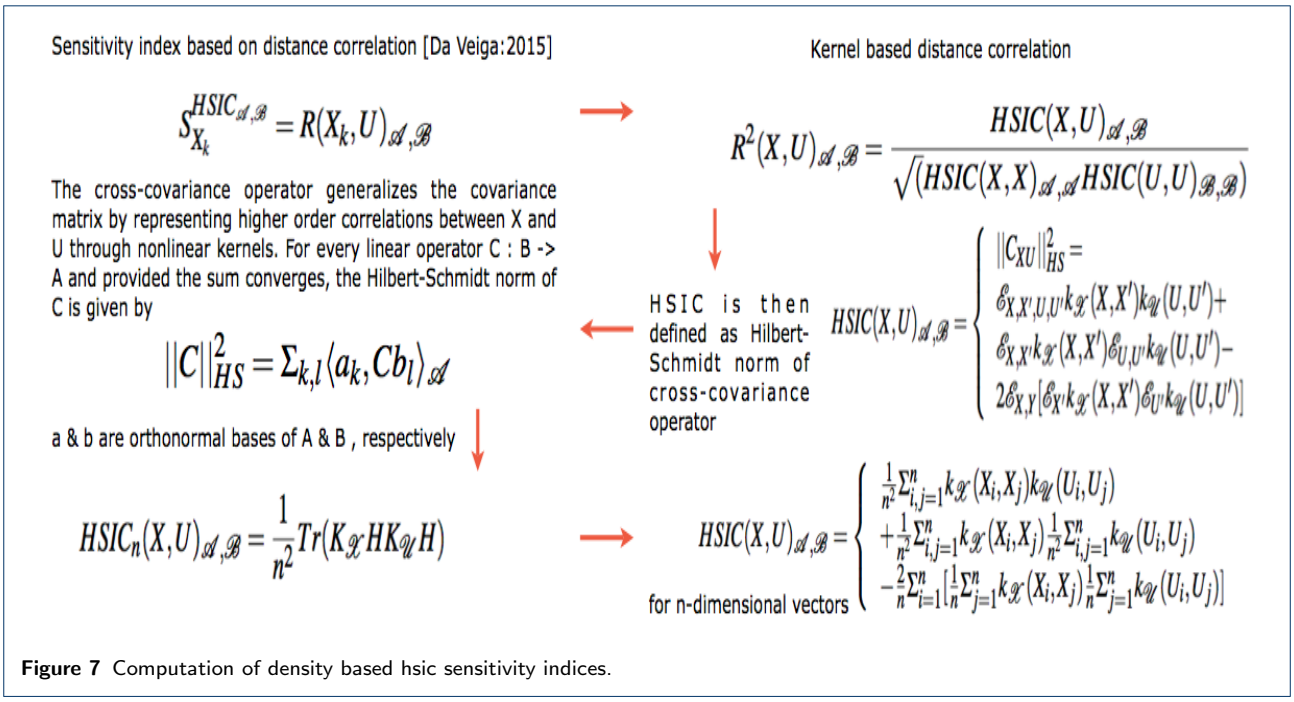

Figure 7 Computation of density based hsic sensitivity indices.

Tools of study

Sensitivity analysis

In order to address the above issues, sensitivity analysis (SA) is performed on either the datasets or results obtained from biologically inspired causal models. The reason for using these tools of sensitivity analysis is that they help in observing the behaviour of the output and the importance of the contributing input factors via a robust and an easy mathematical framework. In this manuscript both local and global SA methods are used. Where appropriate, a description of the biologically inspired causal models ensues before the analysis of results from these models.

Seminal work by Russian mathematician [47] lead to development as well as employment of SA methods to study various complex systems where it was tough to measure the contribution of various input parameters in the behaviour of the output. A recent unpublished review on the global SA methods by [30] categorically delineates these methods with the following functionality • screening for sorting influential measures ([48] method, Group screening in [49] & [50], Iterated factorial design in [51], Sequential bifurcation design in [52] and [53]), • quantitative indicies for measuring the importance of contributing input factors in linear models ([54], [55], [56] and [57]) and nonlinear models ([58], [59], [60], [61], [62], [63], [64], & [65], [66], [67], [68], [69], [70], [71], [72], [73] and [74]) and • exploring the model behaviour over a range on input values ([75] and [76], [77] and [78]). [30] also provide various criteria in a flowchart for adapting a method or a combination of the methods for sensitivity analysis. Figure 6 shows the general flow of the mathematical formulation for computing the indices in the variance based Sobol method. The general idea is as follows - A model could be represented as a mathematical function with a multidimensional input vector where each element of a vector is an input factor. This function needs to be defined in a unit dimensional cube. Based on ANOVA decomposition, the function can then be broken down into f_0 and summands of different dimensions, if f_0 is a constant and integral of summands with respect to their own variables is 0. This implies that orthogonality follows in between two functions of different dimensions, if at least one of the variables is not repeated. By applying these properties, it is possible to show that the function can be written into a unique expansion. Next, assuming that the function is square integrable variances can be computed. The ratio of variance of a group of input factors to the variance of the total set of input factors constitute the sensitivity index of a particular group.

Besides the above [47]'s variance based indicies, more recent developments regarding new indicies based on density, derivative and goal-oriented can be found in [79], [80] and [81], respectively. In a latest development, [82] propose new class of indicies based on density ratio estimation [79] that are special cases of dependence measures. This in turn helps in exploiting measures like distance correlation [83] and Hilbert-Schmidt independence criterion [84] as new sensitivity indicies.

The framework of these indicies is based on use of [85] f-divergence, concept of dissimilarity measure and kernel trick [86]. Finally, [82] propose feature selection as an alternative to screening methods in sensitivity analysis. The main issue with variance based indicies [47] is that even though they capture importance information regarding the contribution of the input factors, they • do not handle multivariate random variables easily and • are only invariant under linear transformations. In comparison to these variance methods, the newly proposed indicies based on density estimations [79] and dependence measures are more robust. Figure 7 shows the general flow of the mathematical formulation for computing the indices in the density based HSIC method. The general idea is as follows - The sensitivity index is actually a distance correlation which incorporates the kernel based Hilbert-Schmidt Information Criterion between two input vectors in higher dimension. The criterion is nothing but the Hilbert-Schmidt norm of cross-covariance operator which generalizes the covariance matrix by representing higher order correlations between the input vectors through nonlinear kernels. For every operator and provided the sum converges, the Hilbert-Schmidt norm is the dot product of the orthonormal bases. For a finite dimensional input vectors, the Hilbert-Schmidt Information Criterion estimator is a trace of product of two kernel matrices (or the Gram matrices) with a centering matrix such that HSIC evaluates to a summation of different kernel values.

It is this strength of the kernel methods that HSIC is able to capture the deep nonlinearities in the biological data and provide reasonable information regarding the degree of influence of the involved factors within the pathway. Improvements in variance based methods also provide ways to cope with these nonlinearities but do not exploit the available strength of kernel methods. Results in the later sections provide experimental evidence for the same.

Application in systems biology

Recent efforts in systems biology to understand the importance of various factors apropos output behaviour has gained prominence. [87] compares the use of [47] variance based indices versus [48] screening method which uses a One-at-a-time (OAT) approach to analyse the sensitivity of $GSK3$ dynamics to uncertainty in an insulin signaling model. Similar efforts, but on different pathways can be found in [88] and [89].

SA provides a way of analysing various factors taking part in a biological phenomena and deals with the effects of these factors on the output of the biological system under consideration. Usually, the model equations are differential in nature with a set of inputs and the associated set of parameters that guide the output. SA helps in observing how the variance in these parameters and inputs leads to changes in the output behaviour. The goal of this manuscript is not to analyse differential equations and the parameters associated with it. Rather, the aim is to observe which input genotypic factors have greater contribution to observed phenotypic behaviour like a sample after treatment of the drug.

There are two approaches to sensitivity analysis. The first is the local sensitivity analysis in which if there is a required solution, then the sensitivity of a function apropos a set of variables is estimated via a partial derivative for a fixed point in the input space. In global sensitivity, the input solution is not specified. This implies that the model function lies inside a cube and the sensitivity indices are regarded as tools for studying the model instead of the solution. The general form of g-function (as the model or output variable) is used to test the sensitivity of each of the input factor (i.e expression profile of each of the genes). This is mainly due to its non-linearity, non-monotonicity as well as the capacity to produce analytical sensitivity indices. The g-function takes the form -

$$f(x) = \Pi_{i=1}^{d} \frac{|4 * x_i - 2| + a_i}{1 + a_i} \tag{1}$$

were, d is the total number of dimensions and $a_i \geq 0$ are the indicators of importance of the input variable x_i. Note that lower values of a_i indicate higher importance of x_i. In our formulation, we randomly assign values of $x_i \in [0, 1]$. For the static data $d = 2500\pm$ (factors affecting the pathway). Thus the expression profiles of the various genetic factors in the pathway are considered as input factors and the global analysis conducted. Note that in the predefined dataset, the working of the signaling pathway is governed by a preselected set of genes that affect the pathway. For comparison purpose, the local sensitivity analysis method was also used to study how the individual factor is behaving with respect to the remaining factors while working of the pathway is observed in terms of expression profiles of the various factors. Thus, both global and local analysis methods were employed to observe the entire behaviour of the pathway as well as the local behaviour of the input factors with respect to the other factors, respectively, via analysis of fold changes. Given the range of estimators available for testing the sensitivity, it might be useful to list a few which are going to be employed in this research study. These have been described in the Appendix. For brevity, we report results from HSIC method only. This is not by random

choice but it has been shown that density based methods are known to superior to their counterparts the variance based methods.

Support vector ranking machines

Learning to rank is a machine learning approach with the idea that the model is trained to learn how to rank. A good introduction to this work can be found in [90]. Existing methods involve pointwise, pairwise and listwise approaches. In all these approaches, Support Vector Machines (SVM) can be employed to rank the required query. SVMs for pointwise approach build various hyperplanes to segregate the data and rank them. Pairwise approach uses ordered pair of objects to classify the objects and then utilize the classifier to rank the objects. In this approach, the group structure of the ranking is not taken into account. Finally, the listwise ranking approach uses ranking list as instances for learning and prediction. In this case the ranking is straightforward and the group structure of ranking is maintained. Various different designs of SVMs have been developed and the research in this field is still in preliminary stages. In context of the gene expression data set employed in this manuscript, the objects are the genes with their RECORDED EXPRESSION VALUES AFTER THE ETC-1922159 TREATMENT.

Note that rankings algorithms have been developed to be employed in the genomic datasets but to the author's awareness, these algorithms do not rank the range of combinations in a wide combinatorial search space in time. Also, they do not take into account the ranking of unexplored biological hypothesis which are assigned to a particular sensitivity value or vector that can be used for prioritization. For example, [91] presents a ranking algorithm that betters existing ranking model based on the assignment of P-value. As stated by [91] it detects genes that are ranked consistently better than expected under null hypothesis of uncorrelated inputs and assigns a significance score for each gene. The underlying probabilistic model makes the algorithm parameter free and robust to outliers, noise and errors. Significance scores also provide a rigorous way to keep only the statistically relevant genes in the final list. The proposed work here develops on sensitivity analysis and computes the influences of the factors for a system under investigation. These sensitivity indices give a much realistic view of the biological influence than the proposed P-value assignment and the probabilistic model. The manuscript at the current stage does not compare the algorithms as it is a pipeline to investigate and conduct a systems wide study. Instead of using SVM-Ranking it is possible to use other algorithms also, but the author has restricted to the development of the pipeline per se. Finally, the current work tests the effectiveness of the variance based (SOBOL) sensitivity indices apropos the density and kernel based (HSIC) sensitivity indices. Finally, [92] provides a range of comparison for 10 different regression methods and a score to measure the models. Compared to the frame provided in [92], the current pipeline takes into account biological information an converts into sensitivity scores and uses them as discriminative features to provide rankings. Thus the proposed method is algorithm independent.

Methods

1.1 Variance based sensitivity indices

The variance based indices as proposed by [47] prove a theorem that an integrable function can be decomposed into summands of different dimensions. Also, a Monte Carlo algorithm is used to estimate the sensitivity of a function apropos arbitrary group of variables. It is assumed that a model denoted by function $u = f(x)$, $x = (x_1, x_2, ..., x_n)$, is defined in a unit n-dimensional cube $\mathcal{K}^n$ with u as the scalar output. The requirement of the problem is to find the sensitivity of function $f(x)$ with respect to different variables. If $u^* = f(x^*)$ is the required solution, then the sensitivity of u^* apropos x_k is estimated via the partial derivative $(\partial u/\partial x_k)_{x=x^*}$. This approach is the local sensitivity. In global sensitivity, the input $x = x^*$ is not specified. This implies that the model $f(x)$ lies inside the cube and the sensitivity indices are regarded as tools for studying the model instead of the solution. Detailed technical aspects with examples can be found in [58] and [93].

Let a group of indices $i_1, i_2, ..., i_s$ exist, where $1 \leq i_1 < ... < i_s \leq n$ and $1 \leq s \leq n$. Then the notation for sum over all different groups of indices is -

$$\widehat{\Sigma} T_{i_1,i_2,...,i_s} = \Sigma_{i=1}^{n} T_i + \Sigma_{s=1}^{n} \Sigma_{1 \leq i < j \leq n} T_{i,j} + ... + T_{1,2,...,n} \quad (2)$$

Then the representation of $f(x)$ using equation 2 in the form -

$$\begin{aligned} f(x) &= f_0 + \widehat{\Sigma} f_{i_1,i_2,...,i_s} && (3) \\ &= f_0 + \Sigma_i f_i(x_i) + \Sigma_{i<j} f_{i,j}(x_i, x_j) + ... \\ &\quad + f_{1,2,...,n}(x_1, x_2, ..., x_n) && (4) \end{aligned}$$

is called ANOVA-decomposition from [71] or expansion into summands of different dimensions, if f_0 is a

constant and integrals of the summands $f_{i_1,i_2,...,i_s}$ with respect to their own variables are zero, i.e,

$$f_0 = \int_{\mathcal{K}^n} f(x)dx \tag{5}$$

$$\int_0^1 f_{i_1,i_2,...,i_s}(x_{i_1}, x_{i_2}, ..., x_{i_s})dx_{i_k} = 0, 1 \leq k \leq s \tag{6}$$

It follows from equation 4 that all summands on the right hand side are orthogonal, i.e if at least one of the indices in $i_1, i_2, ..., i_s$ and $j_1, j_2, ..., j_l$ is not repeated i.e

$$\int_0^1 f_{i_1,i_2,...,i_s}(x_{i_1}, ..., x_{i_s}) f_{j_1,...,j_l}(x_{j_1}, x_{j_2}, ..., x_{j_s})dx = 0 \tag{7}$$

[47] proves a theorem stating that there is an existence of a unique expansion of equation 4 for any $f(x)$ integrable in $\mathcal{K}^n$. In brief, this implies that for each of the indices as well as a group of indices, integrating equation 4 yields the following -

$$\int_0^1 .. \int_0^1 f(x)dx/dx_i = f_0 + f_i(x_i) \tag{8}$$

$$\int_0^1 .. \int_0^1 f(x)dx/dx_i dx_j = f_0 + f_i(x_i) + f_j(x_j) + f_{i,j}(x_i, x_j) \tag{9}$$

were, dx/dx_i is $\prod_{\forall k \in \{1,..,n\}; i \notin k} dx_k$ and $dx/dx_i dx_j$ is $\prod_{\forall k \in \{1,..,n\}; i,j \notin k} dx_k$. For higher orders of grouped indices, similar computations follow. The computation of any summand $f_{i_1,i_2,...,i_s}(x_{i_1}, x_{i_2}, ..., x_{i_s})$ is reduced to an integral in the cube $\mathcal{K}^n$. The last summand $f_{1,2,...,n}(x_1, x_2, ..., x_n)$ is $f(x) - f_0$ from equation 4. [58] stresses that use of Sobol sensitivity indices does not require evaluation of any $f_{i_1,i_2,...,i_s}(x_{i_1}, x_{i_2}, ..., x_{i_s})$ nor the knowledge of the form of $f(x)$ which might well be represented by a computational model i.e a function whose value is only obtained as the output of a computer program.

Finally, assuming that $f(x)$ is square integrable, i.e $f(x) \in \mathcal{L}_2$, then all of $f_{i_1,i_2,...,i_s}(x_{i_1}, x_{i_2}, ..., x_{i_s}) \in \mathcal{L}_2$. Then the following constants

$$\int_{\mathcal{K}^n} f^2(x)dx - f_0^2 = D \tag{10}$$

$$\int_0^1 .. \int_0^1 f^2_{i_1,..,i_s}(x_{i_1}, .., x_{i_s})dx_{i_1}..dx_{i_s} = D_{i_1,...,i_s} \tag{11}$$

are termed as variances. Squaring equation 4, integrating over $\mathcal{K}^n$ and using the orthogonality property in equation 7, D evaluates to -

$$D = \widehat{\Sigma} D_{i_1,i_2,...,i_s} \tag{12}$$

Then the global sensitivity estimates is defined as -

$$S_{i_1,i_2,...,i_s} = \frac{D_{i_1,i_2,...,i_s}}{D} \tag{13}$$

It follows from equations 12 and 13 that

$$\widehat{\Sigma} S_{i_1,i_2,...,i_s} = 1 \tag{14}$$

Clearly, all sensitivity indices are non-negative, i.e an index $S_{i_1,i_2,...,i_s} = 0$ if and only if $f_{i_1,i_2,...,i_s} \equiv 0$. The true potential of Sobol indices is observed when variables $x_1, x_2, ..., x_n$ are divided into m different groups with $y_1, y_2, ..., y_m$ such that $m < n$. Then $f(x) \equiv f(y_1, y_2, ..., y_m)$. All properties remain the same for the computation of sensitivity indices with the fact that integration with respect to y_k means integration with respect to all the x_i's in y_k. Details of these computations with examples can be found in [47]. Variations and improvements over Sobol indices have already been stated in section 1.

1.2 Density based sensitivity indices

As discussed before, the issue with variance based methods is the high computational cost incurred due to the number of interactions among the variables. This further requires the use of screening methods to filter out redundant or unwanted factors that might not have significant impact on the output. Recent work by [82] proposes a new class of sensitivity indicies which are a special case of density based indicies [79]. These indicies can handle multivariate variables easily and relies on density ratio estimation. Key points from [82] are mentioned below.

Considering the similar notation in previous section, $f : \mathcal{R}^n \to \mathcal{R}$ $(u = f(x))$ is assumed to be continuous. It is also assumed that X_k has a known distribution and are independent. [94] state that a function which measures the similarity between the distribution of U and that of $U|X_k$ can define the impact of X_k on U. Thus the impact is defined as -

$$S_{X_k} = \mathcal{E}(d(U, U|X_k)) \tag{15}$$

were $d(\cdot, \cdot)$ is a dissimilarity measure between two random variables. Here d can take various forms as long as it satisfies the criteria of a dissimilarity measure. [85]'s f-divergence between U and $U|X_k$ when all input random variables are considered to be absolutely

continuous with respect to Lebesgue measure on $\mathcal{R}$ is formulated as -

$$d_F(U||U|X_k) = \int_{\mathcal{R}} F(\frac{p_U(u)}{p_{U|X_k}(u)})p_{U|X_k}(u)du \quad (16)$$

were F is a convex function such that $F(1) = 0$ and p_U and $p_{U|X_k}$ are the probability distribution functions of U and $U|X_k$. Standard choices of F include Kullback-Leibler divergence $F(t) = -\log_e(t)$, Hellinger distance $(\sqrt{t}-1)^2$, Total variation distance $F(t) = |t-1|$, Pearson χ^2 divergence $F(t) = t^2 - 1$ and Neyman χ^2 divergence $F(t) = (1-t^2)/t$. Substituting equation 16 in equation 15, gives the following sensitivity index -

$$\begin{aligned} S^F_{X_k} &= \int_{\mathcal{R}} d_F(U||U|X_k)p_{X_k}(x)dx \\ &= \int_{\mathcal{R}}\int_{\mathcal{R}} F(\frac{p_U(u)}{p_{U|X_k}(u)})p_{U|X_k}(u)p_{X_k}(x)dxdu \\ &= \int_{\mathcal{R}^2} F(\frac{p_U(u)p_{X_k}(x)}{p_{U|X_k}(u)p_{X_k}(x)})p_{U|X_k}(u)p_{X_k}(x)dxdu \\ &= \int_{\mathcal{R}^2} F(\frac{p_U(u)p_{X_k}(x)}{p_{X_k,U}(x,u)})p_{X_k,U}(x,u)dxdu \quad (17) \end{aligned}$$

were p_{X_k} and $p_{X_k,Y}$ are the probability distribution functions of X_k and (X_k, U), respectively. [85] f-divergences imply that these indices are positive and equate to 0 when U and X_k are independent. Also, given the formulation of $S^F_{X_k}$, it is invariant under any smooth and uniquely invertible transformation of the variables X_k and U [95]. This has an advantage over Sobol sensitivity indices which are invariant under linear transformations.

By substituting the different formulations of F in equation 17, [82]'s work claims to be the first in establishing the link that previously proposed sensitivity indices are actually special cases of more general indices defined through [85]'s f-divergence. Then equation 17 changes to estimation of ratio between the joint density of (X_k, U) and the marginals, i.e -

$$\begin{aligned} S^F_{X_k} &= \int_{\mathcal{R}^2} F(\frac{1}{r(x,u)})p_{X_k,U}(x,u)dxdu \\ &= \mathcal{E}_{(X_k,U)}F(\frac{1}{r(X_k,U)}) \quad (18) \end{aligned}$$

were, $r(x,y) = (p_{X_k,U}(x,u))/(p_U(u)p_{X_k}(x))$. Multivariate extensions of the same are also possible under the same formulation.

Finally, given two random vectors $X \in \mathcal{R}^p$ and $Y \in \mathcal{R}^q$, the dependence measure quantifies the dependence between X and Y with the property that the measure equates to 0 if and only if X and Y are independent. These measures carry deep links [96] with distances between embeddings of distributions to reproducing kernel Hilbert spaces (RHKS) and here the related Hilbert-Schmidt independence criterion (HSIC by [84]) is explained.

In a very brief manner from an extremely simple introduction by [97] - "We first defined a field, which is a space that supports the usual operations of addition, subtraction, multiplication and division. We imposed an ordering on the field and described what it means for a field to be complete. We then defined vector spaces over fields, which are spaces that interact in a friendly way with their associated fields. We defined complete vector spaces and extended them to Banach spaces by adding a norm. Banach spaces were then extended to Hilbert spaces with the addition of a dot product." Mathematically, a Hilbert space $\mathcal{H}$ with elements $r, s \in \mathcal{H}$ has dot product $\langle r, s\rangle_{\mathcal{H}}$ and $r \cdot s$. When $\mathcal{H}$ is a vector space over a field $\mathcal{F}$, then the dot product is an element in $\mathcal{F}$. The product $\langle r, s\rangle_{\mathcal{H}}$ follows the below mentioned properties when $r, s, t \in \mathcal{H}$ and for all $a \in \mathcal{F}$ -

- Associative : $(ar) \cdot s = a(r \cdot s)$
- Commutative : $r \cdot s = s \cdot r$
- Distributive : $r \cdot (s + t) = r \cdot s + r \cdot t$

Given a complete vector space $\mathcal{V}$ with a dot product $\langle \cdot, \cdot \rangle$, the norm on $\mathcal{V}$ defined by $||r||_{\mathcal{V}} = \sqrt{(\langle r, r\rangle)}$ makes this space into a Banach space and therefore into a full Hilbert space.

A reproducing kernel Hilbert space (RKHS) builds on a Hilbert space $\mathcal{H}$ and requires all Dirac evaluation functionals in $\mathcal{H}$ are bounded and continuous (on implies the other). Assuming $\mathcal{H}$ is the $\mathcal{L}_2$ space of functions from X to $\mathcal{R}$ for some measurable X. For an element $x \in X$, a Dirac evaluation functional at x is a functional $\delta_x \in \mathcal{H}$ such that $\delta_x(g) = g(x)$. For the case of real numbers, x is a vector and g a function which maps from this vector space to $\mathcal{R}$. Then δ_x is simply a function which maps g to the value g has at x. Thus, δ_x is a function from $(\mathcal{R}^n \mapsto \mathcal{R})$ into $\mathcal{R}$.

The requirement of Dirac evaluation functions basically means (via the [98] representation theorem) if ϕ is a bounded linear functional (conditions satisfied by the Dirac evaluation functionals) on a Hilbert space $\mathcal{H}$, then there is a unique vector ℓ in $\mathcal{H}$ such that $\phi g = \langle g, \ell\rangle_{\mathcal{H}}$ for all $\ell \in \mathcal{H}$. Translating this theorem back into Dirac evaluation functionals, for each δ_x there is a unique vector k_x in $\mathcal{H}$ such that $\delta_x g = g(x) = \langle g, k_x\rangle_{\mathcal{H}}$. The reproducing kernel K for $\mathcal{H}$ is then defined as : $K(x, x') = \langle k_x, k_{x'}\rangle$, were k_x and $k_{x'}$ are unique representatives of δ_x and $\delta_{x'}$. The main property of interest is $\langle g, K(x, x')\rangle_{\mathcal{H}} = g(x')$. Furthermore, k_x is defined to be a function $y \mapsto K(x, y)$ and thus the reproducibility is given by $\langle K(x, \cdot), K(y, \cdot)\rangle_{\mathcal{H}} = K(x, y)$.

Basically, the distance measures between two vectors represent the degree of closeness among them. This

degree of closeness is computed on the basis of the discriminative patterns inherent in the vectors. Since these patterns are used implicitly in the distance metric, a question that arises is, how to use these distance metric for decoding purposes?

The kernel formulation as proposed by [86], is a solution to our problem mentioned above. For simplicity, we consider the labels of examples as binary in nature. Let $\mathbf{x}_i \in \mathcal{R}^n$, be the set of n feature values with corresponding category of the example label (y_i) in data set $\mathcal{D}$. Then the data points can be mapped to a higher dimensional space $\mathcal{H}$ by the transformation ϕ:

$$\phi : \mathbf{x}_i \in \mathcal{R}^n \mapsto \phi(\mathbf{x}_i) \in \mathcal{H} \tag{19}$$

This $\mathcal{H}$ is the *Hilbert Space* which is a strict inner product space, along with the property of completeness as well as separability. The inner product formulation of a space helps in discriminating the location of a data point w.r.t a separating hyperplane in $\mathcal{H}$. This is achieved by the evaluation of the inner product between the normal vector representing the hyperplane along with the vectorial representation of a data point in $\mathcal{H}$ (Figure 8 represents the geometrical interpretation). Thus, the idea behind equation(19) is that even if the data points are nonlinearly clustered in space $\mathcal{R}^n$, the transformation spreads the data points into $\mathcal{H}$, such that they can be linearly separated in its range in $\mathcal{H}$.

Often, the evaluation of dot product in higher dimensional spaces is computationally expensive. To avoid incurring this cost, the concept of kernels in employed. The trick is to formulate kernel functions that depend on a pair of data points in the space $\mathcal{R}^n$, under the assumption that its evaluation is equivalent to a dot product in the higher dimensional space. This is given as:

$$\kappa(\mathbf{x}_i, \mathbf{x}_j) =< \phi(\mathbf{x}_i), \phi(\mathbf{x}_j) > \tag{20}$$

Two advantages become immediately apparent. First, the evaluation of such kernel functions in lower dimensional space is computationally less expensive than evaluating the dot product in higher dimensional space. Secondly, it relieves the burden of searching an appropriate transformation that may map the data points in $\mathcal{R}^n$ to $\mathcal{H}$. Instead, all computations regarding discrimination of location of data points in higher dimensional space involves evaluation of the kernel functions in lower dimension. The matrix containing these kernel evaluations is referred to as the *kernel* matrix. With a cell in the kernel matrix containing a kernel evaluation between a pair of data points, the kernel matrix is square in nature.

As an example in practical applications, once the kernel has been computed, a pattern analysis algorithm uses the kernel function to evaluate and predict the nature of the new example using the general formula:

$$\begin{aligned} f(\mathbf{z}) &= < \mathbf{w}, \phi(\mathbf{z}) > +b \\ &= < \sum_{i=1}^{N} \alpha_i \times y_i \times \phi(\mathbf{x}_i), \phi(\mathbf{z}) > +b \\ &= \sum_{i=1}^{N} \alpha_i \times y_i \times < \phi(\mathbf{x}_i), \phi(\mathbf{z}) > +b \\ &= \sum_{i=1}^{N} \alpha_i \times y_i \times \kappa(\mathbf{x}_i, \mathbf{z}) + b \end{aligned} \tag{21}$$

where $\mathbf{w}$ defines the hyperplane as some linear combination of training basis vectors, $\mathbf{z}$ is the test data point, y_i the class label for training point $\mathbf{x}_i$, α_i and b are the constants. Various transformations to the kernel function can be employed, based on the properties a kernel must satisfy. Interested readers are referred to [99] for description of these properties in detail.

The Hilbert-Schmidt independence criterion (HSIC) proposed by [84] is based on kernel approach for finding dependences and on cross-covariance operators in RKHS. Let $X \in \mathcal{X}$ have a distribution P_X and consider a RKHS $\mathcal{A}$ of functions $\mathcal{X} \to \mathcal{R}$ with kernel $k_{\mathcal{X}}$ and dot product $\langle \cdot, \cdot \rangle_{\mathcal{A}}$. Similarly, Let $U \in \mathcal{Y}$ have a distribution P_Y and consider a RKHS $\mathcal{B}$ of functions $\mathcal{U} \to \mathcal{R}$ with kernel $k_{\mathcal{B}}$ and dot product $\langle \cdot, \cdot \rangle_{\mathcal{B}}$. Then the cross-covariance operator $C_{X,U}$ associated with the joint distribution P_{XU} of (X, U) is the linear operator $\mathcal{B} \to \mathcal{A}$ defined for every $a \in \mathcal{A}$ and $b \in \mathcal{B}$ as -

$$\langle a, C_{XU} b \rangle_{\mathcal{A}} = \mathcal{E}_{XU}[a(X), b(U)] - \mathcal{E}_X a(X) \mathcal{E}_U b(U) \tag{22}$$

The cross-covariance operator generalizes the covariance matrix by representing higher order correlations between X and U through nonlinear kernels. For every linear operator $C : \mathcal{B} \to \mathcal{A}$ and provided the sum converges, the Hilbert-Schmidt norm of C is given by -

$$||C||^2_{HS} = \Sigma_{k,l} \langle a_k, C b_l \rangle_{\mathcal{A}} \tag{23}$$

were a_k and b_l are orthonormal bases of $\mathcal{A}$ and $\mathcal{B}$, respectively. The HSIC criterion is then defined as the Hilbert-Schmidt norm of cross-covariance operator -

$$HSIC(X, U)_{\mathcal{A},\mathcal{B}} = \begin{cases} ||C_{XU}||^2_{HS} = \\ \mathcal{E}_{X,X',U,U'} k_{\mathcal{X}}(X, X') k_{\mathcal{U}}(U, U') + \\ \mathcal{E}_{X,X'} k_{\mathcal{X}}(X, X') \mathcal{E}_{U,U'} k_{\mathcal{U}}(U, U') - \\ 2\mathcal{E}_{X,Y}[\mathcal{E}_{X'} k_{\mathcal{X}}(X, X') \mathcal{E}_{U'} k_{\mathcal{U}}(U, U')] \end{cases} \tag{24}$$

were the equality in terms of kernels is proved in [84]. Finally, assuming (X_i, U_i) $(i = 1, 2, ..., n)$ is a sample of the random vector (X, U) and $K_{\mathcal{X}}$ and $K_{\mathcal{U}}$ denote the Gram matrices with entries $K_{\mathcal{X}}(i,j) = k_{\mathcal{X}}(X_i, X_j)$ and $K_{\mathcal{U}}(i,j) = k_{\mathcal{U}}(U_i, U_j)$. [84] proposes the following estimator for $HSIC_n(X,U)_{\mathcal{A},\mathcal{B}}$ -

$$HSIC_n(X,U)_{\mathcal{A},\mathcal{B}} = \frac{1}{n^2} Tr(K_{\mathcal{X}} H K_{\mathcal{U}} H) \tag{25}$$

were H is the centering matrix such that $H(i,j) = \delta_{i,j} - \frac{1}{n}$. Then $HSIC_n(X,U)_{\mathcal{A},\mathcal{B}}$ can be expressed as -

$$HSIC(X,U)_{\mathcal{A},\mathcal{B}} = \begin{cases} \frac{1}{n^2}\Sigma_{i,j=1}^{n} k_{\mathcal{X}}(X_i, X_j) k_{\mathcal{U}}(U_i, U_j) \\ + \frac{1}{n^2}\Sigma_{i,j=1}^{n} k_{\mathcal{X}}(X_i, X_j) \times \\ \frac{1}{n^2}\Sigma_{i,j=1}^{n} k_{\mathcal{U}}(U_i, U_j) \\ - \frac{2}{n}\Sigma_{i=1}^{n} [\frac{1}{n}\Sigma_{j=1}^{n} k_{\mathcal{X}}(X_i, X_j) \times \\ \frac{1}{n}\Sigma_{j=1}^{n} k_{\mathcal{U}}(U_i, U_j)] \end{cases} \tag{2}$$

Finally, [82] proposes the sensitivity index based on distance correlation as -

$$S_{X_k}^{HSIC_{\mathcal{A},\mathcal{B}}} = R(X_k, U)_{\mathcal{A},\mathcal{B}} \tag{27}$$

were the kernel based distance correlation is given by -

$$R^2(X,U)_{\mathcal{A},\mathcal{B}} = \frac{HSIC(X,U)_{\mathcal{A},\mathcal{B}}}{\sqrt{(HSIC(X,X)_{\mathcal{A},\mathcal{A}} HSIC(U,U)_{\mathcal{B},\mathcal{B}})}} \tag{28}$$

were kernels inducing $\mathcal{A}$ and $\mathcal{B}$ are to be chosen within a universal class of kernels. Similar multivariate formulation for equation 25 are possible.

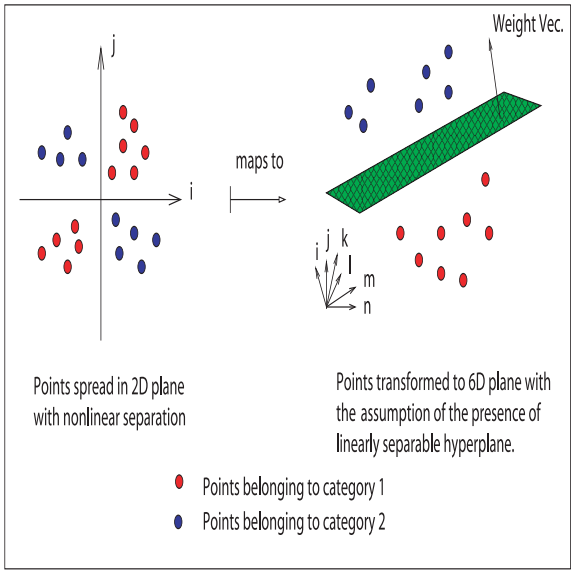

Figure 8 A geometrical interpretation of mapping nonlinearly separable data into higher dimensional space where it is assumed to be linearly separable, subject to the holding of dot product.

Results & discussion

Identified/identified 2^{nd}-order combinations

Table 1 shows the relative rankings apropos RNF43 for three different kernels. The genes that have been shown along with the RNF43 are known to be highly expressed in colorectal cancer cases and after the ETC-1922159 was administered. Here we show only a few of the rankings as a confirmatory result that supports the in silico findings of the pipeline. In the table, except for BMP7 and NKD1, while considering the majority rankings over the three different columns representing the kinds of kernels used with the HSIC density based method, we find that each one of them has been assigned a relatively low priority in the order of 2743 combinations. The implication of these low rankings indicate or confirm the effectiveness of the pipeline in assigning appropriate biologically induced priority to the ETC-1922159 influenced down regulation of these genes. Surprisingly, BMP7 [100] and NKD1 [101] are known to be highly expressed in colorectal cancer case and were down regulated after the drug treatment. But these were assigned high rank with RNF43. Mutations in RNF43 could be subdued after ETC-1922159 has suppressed the Wnt pathway as has been shown in [25]. NKD1 is enigmatic in nature and known to be a negative regulator of the Wnt pathway [102]. Mutations in NKD1 have been found to be prevalent in colorectal cancer cases [103]. High rank might suggest that after the suppression of the cancer cells, the NKD1 is highly activated or the mutated version of NKD1, if present, are highly ineffective. Thus the RNF43-NKD1 acquires a high ranking by the engine and it is more likely that NKD1 is highly activated.

ZNRF3/RNF43 is known to be implicated in inhibiting the Wnt signaling via degradation of FZDs at the cell surface level [46] & [104]. Mutations in these can lead to abberent signaling and mutated RNF43 were found to be suppressed after the ETC-1922159 treatment and the relatively low rank assigned to the combination points to effectiveness of the pipeline. Similarly, RRM2 which is involved in DNA repair [105], is implicated in the metastasis of colon cancer [106] and was observed to be highly expressed in colorectal cancer cases [25]. With mutated RNF43, as the Wnt signaling is enhanced, there is possibility of increased

Ranking using HSIC w.r.t RNF43

2^{nd} odr comb.	rbf	laplace	linear
LGR5-RNF43	61	85	114
LGR6-RNF43	1652	805	939
RAD51AP1-RNF43	175	469	11
ZNRF3-RNF43	570	1603	404
RRM2-RNF43	556	667	442
MKI67-RNF43	1575	278	92
AXIN2-RNF43	531	51	482
ASCL2-RNF43	116	441	197
MCM4-RNF43	131	1142	204
RNF43-BMP7	2723	2701	2530
RNF43-NKD1	2635	2169	2226

Table 1 2^{nd} order interaction ranking using HSIC for different kernels. Total number of interactions per gene were 2743.

tumorigenesis and RRM2 could synergistically work in tandem with RAD51AP1 (another contributor of genomic stability) to enhance the metastasis of CRC. Coincidently, the pipeline gives a preferred low rank of 139 to RRM2-RAD51AP1 combination after the drug treatment, thus indicating that its inhibition in the suppressed cancer cells. The lower ranking of RRM2 along with the RNF43 points to the correct prioritization by the pipeline. Similar results were found for KI-67 (MKI67) which as an independent prognostic marker for CRC [107] and MCM4, which play essential role in DNA replication [108], are both highly expressed in colorectal cancer cases. AXIN2, like AXIN1, helps in the assembly of the destruction complex that facilitates in the degradation of β-catenin in the cytoplasm, thus negatively regulating the Wnt pathway [109]. Mutations in AXIN family can lead to different subtypes of CRC. AXIN2 is also a transcriptional target for β-catenin and changes in protein levels of AXIN2 due to excessive β-catenin can negatively regulate the Wnt pathway in cancer cases. Mutations in RNF43 can lead to enhanced Wnt signaling which can then target the over expression of AXIN2 via β-catenin that might lead to negative feedback to the pathway at a later stage. After the ETC-1922159 treatment, AXIN2 was found to be down regulated. This implies that the drug is working at multiple levels to inhibit the Wnt pathway and the low ranking of the AXIN2-RNF43 combination assigned by the pipeline accounts for this fact.

ASCL2 has been found to play a major role in stemness in colon crypts and is implicated in colon cancer [110]. Switching of the ASCL2 leads to a literal blockage of the stemness process and vice versa. At the downstream level, ASCL2 is regulated by TCF4/β-catenin via non-coding RNA target named WiNTRLINC1 [111]. Activation of ASCL2 leads to feedforward transcription of the non-coding RNA and thus a loop is formed which helps in the stemness and is highly effective in colon cancer. At the upstream level, ASCL2 is known act as a WNT/RSPONDIN switch that controls the stemness [112]. It has been shown that removal of RSPO1 lead to decrease in the Wnt signaling due to removal of the FZD receptors that led to reduced expression of ASCL2. Also, low levels of LGR5 were observed due to this phenomena. The opposite happened by increasing the RSPO1 levels. After the drug treatment, it was found that ASCL2 was highly suppressed pointing to the inhibition of stemness in the colorectal cancer cells. Also, [112] show that by genetically disrupting PORCN or inducing a PORCN inhibitor (like IWP-2), there is loss of stem cell markers like LGR5 and RNF43, which lead to disappearance of stem cells and moribund state of mice. A similar affect can be found with ETC-1922159, where there is suppression of RNF43 and LGR5 that lead to inhibition of the Wnt pathway and thus the ASCL2 regulation. These wet lab evidences are confirmed in the relatively low ranking of the combination ASCL2-RNF43 via the inhibition of PORCN-WNT that leads to blocking of the stemness that is induced by ASCL2. Since ASCL2 is directly mediated by the WNT proteins, the recorded ASCL2-WNT10B combination showed low priority ranking of 488, 497 and 321 for rbf, laplace and linear kernels, respectively, thus indicating a possible connection between WNT10B and ASCL2 activation. WNT10B might be playing a crucial role in stemness. This is further confirmed by wet lab experiments in [113], which show BVES deletion results in amplified stem cell activity and Wnt signaling after radiation. WNT10B has been implicated in colorectal cancer [114].

The foregoing descriptions confirm the effectiveness of the pipeline for a few cases and it is not possible to elucidate each and every combination in a single manuscript. The rankings have been made available for further tests and will help the investigator in narrowing down their focus on particular aspects of the signaling pathway in cancer cases.

Unidentified/(Un)identified 2^{nd}-order combinations - one blade double edged sword

Hitherto, the 2^{nd} order combinations pertaining to known or recorded factors that have been affected by the ETC-1922159 have been prioritized and discussed. But there were some uncharacterized proteins that were also affected after the drug treatment and their behaviour might not be known in terms of higher order interactions. We now traverse the unchartered territory of unknown/unexplored/untested proteins whose

transcript levels were recorded after the drug treatment. Note that we will be dealing with two different cases over here - (1) Unidentified/identified 2^{nd} order combinations and (2) Unidentified/unidentified 2^{nd} order combinations. In total 234 such unidentified proteins were recorded of which we show the interpretations of only a few. Note that the numbers like 1,13, 121 etc associated with XX DO NOT imply any sort of association and we do not assume anything about them. We just interpret the rankings of these unidentified factors with indentified and unidentified factors. Table 2 represents these combinations, with the first 20 describing XX1-identified factor combinations and the next 16 describing XX1-unidentified factor combinations.

UNIDENTIFIED-IDENTIFIED COMBINATIONS - We first begin with the analysis of the unknown factor XX1 with some of the identified and comfirmed factors that are implicated in the pathway either positively or negatively. As has been seen earlier many of the factors like RAD51AP1, RRM2, ASCL2, AXIN2, MKI67, LGR5 and NDK1 that are highly implicated in colorectal cancer case were found to be suppressed after the drug treatment. XX1 pairs with these factors and the pipeline assigns low rank to these combinations. One of the interpretations can be that XX1, like the above factors, is highly implicated in colorectal cancer case and its association with some of these factors might be possible in reality and entails verification of the same. For example, the minichromosome maintenance (MCM) proteins are essential replication factors and have been found to be overexpressed in colorectal cancer [115]. After the administration of the drug, MCM3 was found to be highly suppressed and the pipeline points to its low rank along with XX1. Ubiquitin-conjugating enzyme E2C gene (UBE2C) [116] is known to be overexpressed in colorectal cancer [117] and its apparent downregulation after ETC-1922159 has been assigned a low rank with XX1. Genetic alterations in tyrosine phosphatases have been found in colorectal cancer and serve as tumor suppressors in colorectal cancer [118]. PTPRO is one such family member and was found to be suppressed after drug treatment and assigned low ranks with XX1. There might be a possibility that mutations in PTPRO were present and there was overexpression of these mutated versions in the tumor cells before the drug administration. After the treatment, the in silico low ranks point to the observed suppressions in PTPRO. The HOX family, is known to play multiple roles in various tumor cases and varied affects have been found in colorectal tumor and normal cases [119]. HOXB8 and HOXB9 are highly expressed in colorectal cancer cases and were found to be downregulated after the drug treatment.

RANKING USING HSIC W.R.T XX1

2^{nd} odr comb.	rbf	laplace	linear
UNIDENTIFIED-IDENTIFIED COMBINATIONS			
XX1-RAD51AP1	199	434	36
RRM2-XX1	131	185	138
ASCL2-XX1	476	328	217
XX1-AXIN2	552	331	419
MKI67-XX1	67	299	48
MCM3-XX1	188	3	26
UBE2C-XX1	10	204	21
XX1-LGR5	40	208	39
XX1-PTPRO	130	21	12
XX1-NKD1	763	801	539
XX1-HOXB8	740	178	411
XX1-HOXB9	436	464	287
XX1-PPA2	2399	2320	2659
XX1-RNF43	2127	1488	2140
XX1-XRCC6BP1	1475	1916	1732
XX1-XRCC6	2569	2278	2734
XX1-HOXB7	2533	2554	2708
XX1-HOXB13	2631	2463	1579
XX1-HOXA9	2006	1091	1823
XX1-HOXA11	1860	1984	2326
UNIDENTIFIED-UNIDENTIFIED COMBINATIONS			
XX1-XX15	35	107	245
XX1-XX110	112	763	101
XX1-XX182	2734	2582	2669
XX1-XX196	2689	2615	1964
XX1-XX2	86	190	231
XX1-XX20	105	386	15
XX1-XX205	2669	2499	2610
XX1-XX207	2671	2739	2733
XX1-XX33	194	194	360
XX1-XX35	430	600	654
XX1-XX34	1969	1397	2173
XX1-XX38	1968	1445	1509
XX1-XX49	21	340	303
XX1-XX47	115	814	119
XX1-XX44	1690	465	1070
XX1-XX45	1417	1830	2106

Table 2 2^{nd} order interaction ranking using HSIC for different kernels. Total number of interactions per gene were 2743.

Along with XX1, they were assigned low ranks as desired.

The opposite interpretations are that there are very high ranks assigned to many of the factors that are known to play tumor suppressor roles like HOXB7, HOXB13, HOXA9, HOXA11 [119], PPA2 [40] and XRCC6BP1 and mutations in these could lead to enhancement in colorectal cancer. These were found to be down regulated after the drug treatment in can-

cer cases and were assigned high ranks along with the XX1. XX1 might be a tumor suppressor gene also that might be mutated in colorectal cancer case. But these high ranks provide an alternative insight to the functioning of XX1. Thus the interpretations are like two edges of one blade and biologists can use these rankings to see the multifaceted aspects of the hitherto unidentified XX1.

UNIDENTIFIED-UNIDENTIFIED COMBINATIONS - Now we analyse the unknown factor XX1 with some of the unknown/unexplored factors that are implicated in the pathway either positively or negatively. Observing table 2, the rankings of unidentified-unidentified factors after the ETC-1922159 were also generated.

In each of the series starting with XXM, were $M \in 1, 2, 3, 4$, two interactions are shown with very low ranks assigned to them and two interactions are shown with very high ranks assigned to them. For example, XX1-XX20 and XX1-XX2 both show low priority ranks while XX1-XX205 and XX1-XX207 high ranks. The low priority ranks indicate that their down regulation was important after the ETC-1922159 treatment and these might have been highly expressed in colorectal cancer case before the treatment of the drug. The high priority ranks indicate that these might be tumor suppressor genes (might be mutated) in colorectal cancer and would be highly expressed after the administration of the drug, had the mutations not happened in them. Their down regulation and yet high rank points to their mutated versions in the cancerous cells, a possibility that needs to be verified. Similar interpretations can be made for the different ranked unidentified-unidentified combinations that the pipeline generated on the list of downregulated genes after the ETC-1922159 treatment.

Conclusion

A theoretically sound and a practical framework has been developed to prioritize higher order combinations of downregulated genes after the administration of ETC-1922159 PORCN-WNT inhibitor in cancer cells. The prioritization uses advanced density based sensitivity indices that exploit nonlinear relations in reproducing kernel hilbert spaces via kernel trick and support vector ranking method to rank and reveal various combinations of identified and unidentified factors that are affected after the drug treatment. This gives medical specialists/oncologists as well as biologists a way to navigate in a guided manner in a vast combinatorial search forest thus cutting down cost in time/investment/energy as well as avoid cherry picking unknown biological hypotheses.

Competing interests
The author declare that they have no competing interests.

Dedication
The author dedicates this small work to his long time friend Joana Carolina Martins who was diagnosed with tumor at the age of 35 and to other cancer patients. Joana's diagnosis as well as communications propelled the author to delve into the research work related to cancer biology and therapeutics.

Author's contributions
SS designed, developed and implemented the insilico experimental setup, wrote the code, generated and analysed the results and wrote the manuscript.

Acknowledgements
Special thanks to Mrs. Rita Sinha and Mr. Prabhat Sinha for supporting the author financially, without which this work could not have been made possible. Thomas F. Rosenbaum and Stephan L. Mayo for providing hope and encouragement regarding the work on biosearch engine design as well as the related rankings that were generated from the various data sets. The author also acknowledges fruitful informal discussions with Kibeom Kim, Mengsha Gong, Christopher Frick, Michael Abrams, Ty Basinger, Lea Goentoro, Frances Arnold, Viviana Gardinaru and Richard Murray. Marco Wiering, Marc Thioux and Silja Renooij for continued support during the years of independent research work.

Code Availability
Code on google drive at `https://drive.google.com/drive/folders/0B7Kkv8wlhPU-WDgzdUVfTzA2cW8`
Some of the 2^{nd} order rankings in `https://drive.google.com/file/d/0B7Kkv8wlhPU-UkVNVkx3TzNQcU0/view`

Source of Data
Data used in this research work was released in a publication in [25]. The ETC-1922159 was released in Singapore in July 2015 under the flagship of the Agency for Science, Technology and Research (A*STAR) and Duke-National University of Singapore Graduate Medical School (Duke-NUS).

References

1. Sharma, R.: Wingless a new mutant in drosophila melanogaster. Drosophila information service **50**, 134–134 (1973)
2. Thorstensen, L., Lind, G.E., Løvig, T., Diep, C.B., Meling, G.I., Rognum, T.O., Lothe, R.A.: Genetic and epigenetic changes of components affecting the wnt pathway in colorectal carcinomas stratified by microsatellite instability. Neoplasia **7**(2), 99–108 (2005)
3. Baron, R., Kneissel, M.: Wnt signaling in bone homeostasis and disease: from human mutations to treatments. Nature medicine **19**(2), 179–192 (2013)
4. Clevers, H.: Wnt/[β]-catenin signaling in development and disease. Cell **127**(3), 469–480 (2006)
5. Sokol, S.: Wnt Signaling in Embryonic Development vol. 17. Elsevier, ??? (2011)
6. Pinto, D., Gregorieff, A., Begthel, H., Clevers, H.: Canonical wnt signals are essential for homeostasis of the intestinal epithelium. Genes & development **17**(14), 1709–1713 (2003)
7. Zhong, Z., Ethen, N.J., Williams, B.O.: Wnt signaling in bone development and homeostasis. Wiley Interdisciplinary Reviews: Developmental Biology **3**(6), 489–500 (2014)
8. Pećina-Šlaus, N.: Wnt signal transduction pathway and apoptosis: a review. Cancer Cell International **10**(1), 1–5 (2010)
9. Kahn, M.: Can we safely target the wnt pathway? Nature Reviews Drug Discovery **13**(7), 513–532 (2014)
10. Garber, K.: Drugging the wnt pathway: problems and progress. Journal of the National Cancer Institute **101**(8), 548–550 (2009)
11. Voronkov, A., Krauss, S.: Wnt/beta-catenin signaling and small molecule inhibitors. Current pharmaceutical design **19**(4), 634 (2012)
12. Blagodatski, A., Poteryaev, D., Katanaev, V.: Targeting the wnt pathways for therapies. Mol Cell Ther **2**, 28 (2014)
13. Curtin, J.C., Lorenzi, M.V.: Drug discovery approaches to target wnt signaling in cancer stem cells. Oncotarget **1**(7), 552 (2010)
14. Sinha, S.: Prioritizing 2nd order interactions via support vector ranking using sensitivity indices on time series wnt measurements. bioRxiv, 060228 (2017)

15. Tanaka, K., Okabayashi, K., Asashima, M., Perrimon, N., Kadowaki, T.: The evolutionarily conserved porcupine gene family is involved in the processing of the wnt family. The FEBS Journal **267**(13), 4300–4311 (2000)
16. Bänziger, C., Soldini, D., Schütt, C., Zipperlen, P., Hausmann, G., Basler, K.: Wntless, a conserved membrane protein dedicated to the secretion of wnt proteins from signaling cells. Cell **125**(3), 509–522 (2006)
17. Bartscherer, K., Pelte, N., Ingelfinger, D., Boutros, M.: Secretion of wnt ligands requires evi, a conserved transmembrane protein. Cell **125**(3), 523–533 (2006)
18. Kurayoshi, M., Yamamoto, H., Izumi, S., Kikuchi, A.: Post-translational palmitoylation and glycosylation of wnt-5a are necessary for its signalling. Biochemical Journal **402**(3), 515–523 (2007)
19. Gao, X., Hannoush, R.N.: Single-cell imaging of wnt palmitoylation by the acyltransferase porcupine. Nature chemical biology **10**(1), 61–68 (2014)
20. Chen, B., Dodge, M.E., Tang, W., Lu, J., Ma, Z., Fan, C.-W., Wei, S., Hao, W., Kilgore, J., Williams, N.S., *et al.*: Small molecule–mediated disruption of wnt-dependent signaling in tissue regeneration and cancer. Nature chemical biology **5**(2), 100–107 (2009)
21. Wang, X., Moon, J., Dodge, M.E., Pan, X., Zhang, L., Hanson, J.M., Tuladhar, R., Ma, Z., Shi, H., Williams, N.S., *et al.*: The development of highly potent inhibitors for porcupine. Journal of medicinal chemistry **56**(6), 2700–2704 (2013)
22. Proffitt, K.D., Madan, B., Ke, Z., Pendharkar, V., Ding, L., Lee, M.A., Hannoush, R.N., Virshup, D.M.: Pharmacological inhibition of the wnt acyltransferase porcn prevents growth of wnt-driven mammary cancer. Cancer research **73**(2), 502–507 (2013)
23. Liu, J., Pan, S., Hsieh, M.H., Ng, N., Sun, F., Wang, T., Kasibhatla, S., Schuller, A.G., Li, A.G., Cheng, D., *et al.*: Targeting wnt-driven cancer through the inhibition of porcupine by lgk974. Proceedings of the National Academy of Sciences **110**(50), 20224–20229 (2013)
24. Duraiswamy, A.J., Lee, M.A., Madan, B., Ang, S.H., Tan, E.S.W., Cheong, W.W.V., Ke, Z., Pendharkar, V., Ding, L.J., Chew, Y.S., *et al.*: Discovery and optimization of a porcupine inhibitor. Journal of medicinal chemistry **58**(15), 5889–5899 (2015)
25. Madan, B., Ke, Z., Harmston, N., Ho, S.Y., Frois, A., Alam, J., Jeyaraj, D.A., Pendharkar, V., Ghosh, K., Virshup, I.H., *et al.*: Wnt addiction of genetically defined cancers reversed by porcn inhibition. Oncogene **35**(17), 2197 (2016)
26. Joachims, T.: Training linear svms in linear time. In: Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 217–226 (2006). ACM
27. Gujral, T.S., MacBeath, G.: A system-wide investigation of the dynamics of wnt signaling reveals novel phases of transcriptional regulation. PloS one **5**(4), 10024 (2010)
28. R Development Core Team: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria (2008). R Foundation for Statistical Computing. ISBN 3-900051-07-0. http://www.R-project.org
29. Faivre, R., Iooss, B., Mahévas, S., Makowski, D., Monod, H.: Analyse de Sensibilité et Exploration de Modèles: Application aux Sciences de la Nature et de L'environnement. Editions Quae, ??? (2013)
30. Iooss, B., Lemaître, P.: A review on global sensitivity analysis methods. arXiv preprint arXiv:1404.2405 (2014)
31. Krejci, L., Altmannova, V., Spirek, M., Zhao, X.: Homologous recombination and its regulation. Nucleic acids research **40**(13), 5795–5818 (2012)
32. Tennstedt, P., Fresow, R., Simon, R., Marx, A., Terracciano, L., Petersen, C., Sauter, G., Dikomey, E., Borgmann, K.: Rad51 overexpression is a negative prognostic marker for colorectal adenocarcinoma. International journal of cancer **132**(9), 2118–2126 (2013)
33. Dunlop, M.H., Dray, E., Zhao, W., San Filippo, J., Tsai, M.-S., Leung, S.G., Schild, D., Wiese, C., Sung, P.: Mechanistic insights into rad51-associated protein 1 (rad51ap1) action in homologous dna repair. Journal of Biological Chemistry **287**(15), 12343–12347 (2012)
34. Modesti, M., Budzowska, M., Baldeyron, C., Demmers, J.A., Ghirlando, R., Kanaar, R.: Rad51ap1 is a structure-specific dna binding protein that stimulates joint molecule formation during rad51-mediated homologous recombination. Molecular cell **28**(3), 468–481 (2007)
35. Davies, O.R., Pellegrini, L.: Interaction with the brca2 c-terminus protects rad51–dna filaments from disassembly by brc repeats. Nature structural & molecular biology **14**(6), 475 (2007)
36. Esashi, F., Galkin, V.E., Xiong, Y., Egelman, E.H., West, S.C.: Stabilization of rad51 nucleoprotein filaments by the c-terminal region of brca2. Nature structural & molecular biology **14**(6), 468 (2007)
37. Ayoub, N., Rajendra, E., Su, X., Jeyasekharan, A.D., Mahen, R., Venkitaraman, A.R.: The carboxyl terminus of brca2 links the disassembly of rad51 complexes to mitotic entry. Current Biology **19**(13), 1075–1085 (2009)
38. Schlacher, K., Christ, N., Siaud, N., Egashira, A., Wu, H., Jasin, M.: Double-strand break repair-independent role for brca2 in blocking stalled replication fork degradation by mre11. Cell **145**(4), 529–542 (2011)
39. Seshacharyulu, P., Pandey, P., Datta, K., Batra, S.K.: Phosphatase: Pp2a structural importance, regulation and its aberrant expression in cancer. Cancer letters **335**(1), 9–18 (2013)
40. Cristóbal, I., Manso, R., Rincón, R., Caramés, C., Senin, C., Borrero, A., Martínez-Useros, J., Rodriguez, M., Zazo, S., Aguilera, O., *et al.*: Pp2a inhibition is a common event in colorectal cancer and its restoration using fty720 shows promising therapeutic potential. Molecular cancer therapeutics **13**(4), 938–947 (2014)
41. Masson, J.-Y., Tarsounas, M.C., Stasiak, A.Z., Stasiak, A., Shah, R., McIlwraith, M.J., Benson, F.E., West, S.C.: Identification and purification of two distinct complexes containing the five rad51 paralogs. Genes & development **15**(24), 3296–3307 (2001)
42. Yokoyama, H., Sarai, N., Kagawa, W., Enomoto, R., Shibata, T., Kurumizaka, H., Yokoyama, S.: Preferential binding to branched dna strands and strand-annealing activity of the human rad51b, rad51c, rad51d and xrcc2 protein complex. Nucleic acids research **32**(8), 2556–2565 (2004)
43. Yeh, C.-C., Sung, F.-C., Tang, R., Chang-Chieh, C.R., Hsieh, L.-L.: Polymorphisms of the xrcc1, xrcc3, & xpd genes, and colorectal cancer risk: a case-control study in taiwan. BMC cancer **5**(1), 12 (2005)
44. Diao, R., Mu, X., Wang, T., Li, S.: Risk score based on ten lncrna-mrna expression predicts the survival of stage ii-iii colorectal carcinoma. PloS one **12**(8), 0182908 (2017)
45. Feng, Q., Gao, N.: Keeping wnt signalosome in check by vesicular traffic. Journal of cellular physiology **230**(6), 1170–1180 (2015)
46. de Lau, W., Peng, W.C., Gros, P., Clevers, H.: The r-spondin/lgr5/rnf43 module: regulator of wnt signal strength. Genes & development **28**(4), 305–316 (2014)
47. Sobol', I.M.: On sensitivity estimation for nonlinear mathematical models. Matematicheskoe Modelirovanie **2**(1), 112–118 (1990)
48. Morris, M.D.: Factorial sampling plans for preliminary computational experiments. Technometrics **33**(2), 161–174 (1991)
49. Moon, H., Dean, A.M., Santner, T.J.: Two-stage sensitivity-based group screening in computer experiments. Technometrics **54**(4), 376–387 (2012)
50. Dean, A., Lewis, S.: Screening: Methods for Experimentation in Industry, Drug Discovery, and Genetics. Springer, ??? (2006)
51. Andres, T.H., Hajas, W.C.: Using iterated fractional factorial design to screen parameters in sensitivity analysis of a probabilistic risk assessment model (1993)
52. Bettonvil, B., Kleijnen, J.P.: Searching for important factors in simulation models with many factors: Sequential bifurcation. European Journal of Operational Research **96**(1), 180–194 (1997)
53. Cotter, S.C.: A screening design for factorial experiments with interactions. Biometrika **66**(2), 317–320 (1979)
54. Christensen, R.: Linear Models for Multivariate, Time Series, and Spatial Data. Springer, ??? (1991)
55. Saltelli, A., Chan, K., Scott, E.: Sensitivity analysis wiley series in probability and statistics. Willey, New York (2000)
56. Helton, J.C., Davis, F.J.: Latin hypercube sampling and the propagation of uncertainty in analyses of complex systems. Reliability Engineering & System Safety **81**(1), 23–69 (2003)

57. McKay, M.D., Beckman, R.J., Conover, W.J.: Comparison of three methods for selecting values of input variables in the analysis of output from a computer code. Technometrics **21**(2), 239–245 (1979)
58. Homma, T., Saltelli, A.: Importance measures in global sensitivity analysis of nonlinear models. Reliability Engineering & System Safety **52**(1), 1–17 (1996)
59. Sobol, I.M.: Global sensitivity indices for nonlinear mathematical models and their monte carlo estimates. Mathematics and computers in simulation **55**(1), 271–280 (2001)
60. Saltelli, A.: Making best use of model evaluations to compute sensitivity indices. Computer Physics Communications **145**(2), 280–297 (2002)
61. Saltelli, A., Ratto, M., Tarantola, S., Campolongo, F.: Sensitivity analysis for chemical models. Chemical reviews **105**(7), 2811–2828 (2005)
62. Saltelli, A., Ratto, M., Andres, T., Campolongo, F., Cariboni, J., Gatelli, D., Saisana, M., Tarantola, S.: Global Sensitivity Analysis: the Primer. John Wiley & Sons, ??? (2008)
63. Cukier, R., Fortuin, C., Shuler, K.E., Petschek, A., Schaibly, J.: Study of the sensitivity of coupled reaction systems to uncertainties in rate coefficients. i theory. The Journal of Chemical Physics **59**(8), 3873–3878 (1973)
64. Saltelli, A., Tarantola, S., Chan, K.-S.: A quantitative model-independent method for global sensitivity analysis of model output. Technometrics **41**(1), 39–56 (1999)
65. Tarantola, S., Gatelli, D., Mara, T.A.: Random balance designs for the estimation of first order global sensitivity indices. Reliability Engineering & System Safety **91**(6), 717–727 (2006)
66. Saltelli, A., Annoni, P., Azzini, I., Campolongo, F., Ratto, M., Tarantola, S.: Variance based sensitivity analysis of model output. design and estimator for the total sensitivity index. Computer Physics Communications **181**(2), 259–270 (2010)
67. Janon, A., Klein, T., Lagnoux, A., Nodet, M., Prieur, C.: Asymptotic normality and efficiency of two sobol index estimators. ESAIM: Probability and Statistics **18**, 342–364 (2014)
68. Owen, A.B.: Better estimation of small sobol'sensitivity indices. ACM Transactions on Modeling and Computer Simulation (TOMACS) **23**(2), 11 (2013)
69. Tissot, J.-Y., Prieur, C.: Bias correction for the estimation of sensitivity indices based on random balance designs. Reliability Engineering & System Safety **107**, 205–213 (2012)
70. Da Veiga, S., Gamboa, F.: Efficient estimation of sensitivity indices. Journal of Nonparametric Statistics **25**(3), 573–595 (2013)
71. Archer, G., Saltelli, A., Sobol, I.: Sensitivity measures, anova-like techniques and the use of bootstrap. Journal of Statistical Computation and Simulation **58**(2), 99–120 (1997)
72. Tarantola, S., Gatelli, D., Kucherenko, S., Mauntz, W., *et al.*: Estimating the approximation error when fixing unessential factors in global sensitivity analysis. Reliability Engineering & System Safety **92**(7), 957–960 (2007)
73. Saltelli, A., Annoni, P.: How to avoid a perfunctory sensitivity analysis. Environmental Modelling & Software **25**(12), 1508–1517 (2010)
74. Jansen, M.J.: Analysis of variance designs for model output. Computer Physics Communications **117**(1), 35–43 (1999)
75. Storlie, C.B., Helton, J.C.: Multiple predictor smoothing methods for sensitivity analysis: Description of techniques. Reliability Engineering & System Safety **93**(1), 28–54 (2008)
76. Da Veiga, S., Wahl, F., Gamboa, F.: Local polynomial estimation for sensitivity analysis on models with correlated inputs. Technometrics **51**(4), 452–463 (2009)
77. Li, G., Rosenthal, C., Rabitz, H.: High dimensional model representations. The Journal of Physical Chemistry A **105**(33), 7765–7777 (2001)
78. Hajikolaei, K.H., Wang, G.G.: High dimensional model representation with principal component analysis. Journal of Mechanical Design **136**(1), 011003 (2014)
79. Borgonovo, E.: A new uncertainty importance measure. Reliability Engineering & System Safety **92**(6), 771–784 (2007)
80. Sobol, I.M., Kucherenko, S.: Derivative based global sensitivity measures and their link with global sensitivity indices. Mathematics and Computers in Simulation **79**(10), 3009–3017 (2009)
81. Fort, J.-C., Klein, T., Rachdi, N.: New sensitivity analysis subordinated to a contrast. arXiv preprint arXiv:1305.2329 (2013)
82. Da Veiga, S.: Global sensitivity analysis with dependence measures. Journal of Statistical Computation and Simulation **85**(7), 1283–1305 (2015)
83. Székely, G.J., Rizzo, M.L., Bakirov, N.K., *et al.*: Measuring and testing dependence by correlation of distances. The Annals of Statistics **35**(6), 2769–2794 (2007)
84. Gretton, A., Bousquet, O., Smola, A., Schölkopf, B.: Measuring statistical dependence with hilbert-schmidt norms. In: Algorithmic Learning Theory, pp. 63–77 (2005). Springer
85. Csiszár, I., *et al.*: Information-type measures of difference of probability distributions and indirect observations. Studia Sci. Math. Hungar. **2**, 299–318 (1967)
86. Aizerman, M., Braverman, E., Rozonoer, L.: Theoretical foundations of the potential function method in pattern recognition learning. Automation and Remote Control **25**, 821–837 (1964)
87. Sumner, T., Shephard, E., Bogle, I.: A methodology for global-sensitivity analysis of time-dependent outputs in systems biology modelling. Journal of The Royal Society Interface **9**(74), 2156–2166 (2012)
88. Zheng, Y., Rundell, A.: Comparative study of parameter sensitivity analyses of the tcr-activated erk-mapk signalling pathway. IEE Proceedings-Systems Biology **153**(4), 201–211 (2006)
89. Marino, S., Hogue, I.B., Ray, C.J., Kirschner, D.E.: A methodology for performing global uncertainty and sensitivity analysis in systems biology. Journal of theoretical biology **254**(1), 178–196 (2008)
90. Li, H.: A short introduction to learning to rank. IEICE TRANS. INF. & SYST. **E94-D**(10) (2011)
91. Kolde, R., Laur, S., Adler, P., Vilo, J.: Robust rank aggregation for gene list integration and meta-analysis. Bioinformatics **28**(4), 573–580 (2012)
92. Blondel, M., Onogi, A., Iwata, H., Ueda, N.: A ranking approach to genomic selection. PloS one **10**(6), 0128570 (2015)
93. Sobol, S. IM andKucherenko: Global sensitivity indices for nonlinear mathematical models. review. Wilmott Magazine, 2–7 (2005)
94. Baucells, M., Borgonovo, E.: Invariant probabilistic sensitivity analysis. Management Science **59**(11), 2536–2549 (2013)
95. Kraskov, A., Stögbauer, H., Grassberger, P.: Estimating mutual information. Physical review E **69**(6), 066138 (2004)
96. Sejdinovic, D., Sriperumbudur, B., Gretton, A., Fukumizu, K., *et al.*: Equivalence of distance-based and rkhs-based statistics in hypothesis testing. The Annals of Statistics **41**(5), 2263–2291 (2013)
97. Daumé III, H.: From zero to reproducing kernel hilbert spaces in twelve pages or less. Citeseer (2004)
98. Riesz, F.: Sur une espèce de géométrie analytique des systèmes de fonctions sommables. CR Acad. Sci. Paris **144**, 1409–1411 (1907)
99. Taylor, J.S., Cristianini, N.: Properties of Kernels. Cambridge University Press, ??? (2004). Chap. 3
100. Motoyama, K., Tanaka, F., Kosaka, Y., Mimori, K., Uetake, H., Inoue, H., Sugihara, K., Mori, M.: Clinical significance of bmp7 in human colorectal cancer. Annals of surgical oncology **15**(5), 1530–1537 (2008)
101. Stancikova, J., Krausova, M., Kolar, M., Fafilek, B., Svec, J., Sedlacek, R., Neroldova, M., Dobes, J., Horazna, M., Janeckova, L., *et al.*: Nkd1 marks intestinal and liver tumors linked to aberrant wnt signaling. Cellular signalling **27**(2), 245–256 (2015)
102. Larraguibel, J., Weiss, A.R., Pasula, D.J., Dhaliwal, R.S., Kondra, R., Van Raay, T.J.: Wnt ligand–dependent activation of the negative feedback regulator nkd1. Molecular biology of the cell **26**(12), 2375–2384 (2015)
103. Guo, J., Cagatay, T., Zhou, G., Chan, C.-C., Blythe, S., Suyama, K., Zheng, L., Pan, K., Qian, C., Hamelin, R., *et al.*: Mutations in the human naked cuticle homolog nkd1 found in colorectal cancer alter wnt/dvl/β-catenin signaling. PloS one **4**(11), 7982 (2009)
104. Hao, H.-X., Jiang, X., Cong, F.: Control of wnt receptor turnover by r-spondin-znrf3/rnf43 signaling module and its dysregulation in cancer. Cancers **8**(6), 54 (2016)
105. Zhang, Y.-W., Jones, T.L., Martin, S.E., Caplen, N.J., Pommier, Y.: Implication of checkpoint kinase-dependent up-regulation of

ribonucleotide reductase r2 in dna damage response. Journal of Biological Chemistry **284**(27), 18085–18095 (2009)
106. Liu, X., Zhou, B., Xue, L., Yen, F., Chu, P., Un, F., Yen, Y.: Ribonucleotide reductase subunits m2 and p53r2 are potential biomarkers for metastasis of colon cancer. Clinical colorectal cancer **6**(5), 374–381 (2007)
107. Melling, N., Kowitz, C.M., Simon, R., Bokemeyer, C., Terracciano, L., Sauter, G., Izbicki, J.R., Marx, A.H.: High ki67 expression is an independent good prognostic marker in colorectal cancer. Journal of clinical pathology **69**(3), 209–214 (2016)
108. Nishihara, K., Shomori, K., Fujioka, S., Tokuyasu, N., Inaba, A., Osaki, M., Ogawa, T., Ito, H.: Minichromosome maintenance protein 7 in colorectal cancer: implication of prognostic significance. International journal of oncology **33**(2), 245–251 (2008)
109. Mazzoni, S.M., Fearon, E.R.: Axin1 and axin2 variants in gastrointestinal cancers. Cancer letters **355**(1), 1–8 (2014)
110. Zhu, R., Yang, Y., Tian, Y., Bai, J., Zhang, X., Li, X., Peng, Z., He, Y., Chen, L., Pan, Q., *et al.*: Ascl2 knockdown results in tumor growth arrest by mirna-302b-related inhibition of colon cancer progenitor cells. PloS one **7**(2), 32170 (2012)
111. Giakountis, A., Moulos, P., Zarkou, V., Oikonomou, C., Harokopos, V., Hatzigeorgiou, A.G., Reczko, M., Hatzis, P.: A positive regulatory loop between a wnt-regulated non-coding rna and ascl2 controls intestinal stem cell fate. Cell reports **15**(12), 2588–2596 (2016)
112. Schuijers, J., Junker, J.P., Mokry, M., Hatzis, P., Koo, B.-K., Sasselli, V., Van Der Flier, L.G., Cuppen, E., van Oudenaarden, A., Clevers, H.: Ascl2 acts as an r-spondin/wnt-responsive switch to control stemness in intestinal crypts. Cell stem cell **16**(2), 158–170 (2015)
113. Reddy, V.K., Short, S.P., Barrett, C.W., Mittal, M.K., Keating, C.E., Thompson, J.J., Harris, E.I., Revetta, F., Bader, D.M., Brand, T., *et al.*: Bves regulates intestinal stem cell programs and intestinal crypt viability after radiation. Stem Cells **34**(6), 1626–1636 (2016)
114. Yoshikawa, H., Matsubara, K., Zhou, X., Okamura, S., Kubo, T., Murase, Y., Shikauchi, Y., Esteller, M., Herman, J.G., Wang, X.W., *et al.*: Wnt10b functional dualism: β-catenin/tcf-dependent growth promotion or independent suppression with deregulated expression in cancer. Molecular biology of the cell **18**(11), 4292–4303 (2007)
115. Ha, S.-A., Shin, S.M., Namkoong, H., Lee, H., Cho, G.W., Hur, S.Y., Kim, T.E., Kim, J.W.: Cancer-associated expression of minichromosome maintenance 3 gene in several human cancers and its involvement in tumorigenesis. Clinical cancer research **10**(24), 8386–8395 (2004)
116. Hao, Z., Zhang, H., Cowell, J.: Ubiquitin-conjugating enzyme ube2c: molecular biology, role in tumorigenesis, and potential as a biomarker. Tumor Biology **33**(3), 723–730 (2012)
117. Takahashi, Y., Ishii, Y., Nishida, Y., Ikarashi, M., Nagata, T., Nakamura, T., Yamamori, S., Asai, S.: Detection of aberrations of ubiquitin-conjugating enzyme e2c gene (ube2c) in advanced colon cancer with liver metastases by dna microarray and two-color fish. Cancer genetics and cytogenetics **168**(1), 30–35 (2006)
118. Laczmanska, I., Sasiadek, M.M.: Tyrosine phosphatases as a superfamily of tumor suppressors in colorectal cancer. Acta Biochim Pol **58**(4), 467–470 (2011)
119. Kanai, M., Hamada, J.-I., Takada, M., Asano, T., Murakawa, K., Takahashi, Y., Murai, T., Tada, M., Miyamoto, M., Kondo, S., *et al.*: Aberrant expressions of hox genes in colorectal and hepatocellular carcinomas. Oncology reports **23**(3), 843–851 (2010)

Appendix

Choice of sensitivity indices

The SENSITIVITY PACKAGE ([29] and [30]) in R langauge provides a range of functions to compute the indices and the following indices will be taken into account for addressing the posed questions in this manuscript.

1 sensiFdiv - conducts a density-based sensitivity analysis where the impact of an input variable is defined in terms of dissimilarity between the original output density function and the output density function when the input variable is fixed. The dissimilarity between density functions is measured with Csiszar f-divergences. Estimation is performed through kernel density estimation and the function kde of the package ks. [79] and [82]

2 sensiHSIC - conducts a sensitivity analysis where the impact of an input variable is defined in terms of the distance between the input/output joint probability distribution and the product of their marginals when they are embedded in a Reproducing Kernel Hilbert Space (RKHS). This distance corresponds to HSIC proposed by [84] and serves as a dependence measure between random variables.

3 soboljansen - implements the Monte Carlo estimation of the Sobol indices for both first-order and total indices at the same time (all together 2p indices), at a total cost of (p+2) $\times$ n model evaluations. These are called the Jansen estimators. [74] and [66]

4 sobol2002 - implements the Monte Carlo estimation of the Sobol indices for both first-order and total indices at the same time (all together 2p indices), at a total cost of (p+2) $\times$n model evaluations. These are called the Saltelli estimators. This estimator suffers from a conditioning problem when estimating the variances behind the indices computations. This can seriously affect the Sobol indices estimates in case of largely non-centered output. To avoid this effect, you have to center the model output before applying "sobol2002". Functions "soboljansen" and "sobolmartinez" do not suffer from this problem. [60]

5 sobol2007 - implements the Monte Carlo estimation of the Sobol indices for both first-order and total indices at the same time (all together 2p indices), at a total cost of (p+2) $\times$ n model evaluations. These are called the Mauntz estimators. [73]

6 sobolmartinez - implements the Monte Carlo estimation of the Sobol indices for both first-order and total indices using correlation coefficients-based formulas, at a total cost of (p + 2) $\times$ n model evaluations. These are called the Martinez estimators.

7 sobol - implements the Monte Carlo estimation of the Sobol sensitivity indices. Allows the estimation of the indices of the variance decomposition up to a given order, at a total cost of (N + 1) $\times$ n where N is the number of indices to estimate. [47]