UNIVERSIDADE FEDERAL DE SANTA MARIA
CENTRO DE CIÊNCIAS NATURAIS E EXATAS
PROGRAMA DE PÓS-GRADUAÇÃO EM FÍSICA

Carla Adriane Ramos Segatto Fontoura

UM MÉTODO MULTIESTATÍSTICO PARA IDENTIFICAÇÃO DE VIAS GENÉTICAS DIFERENCIALMENTE EXPRESSAS

Santa Maria, RS
2016

Carla Adriane Ramos Segatto Fontoura

UM MÉTODO MULTIESTATÍSTICO PARA IDENTIFICAÇÃO DE VIAS GENÉTICAS DIFERENCIALMENTE EXPRESSAS

Tese de Doutorado apresentada ao Programa de Pós-Graduação em Física, Área de Concentração em Sistemas Complexos, da Universidade Federal de Santa Maria (UFSM, RS), como requisito parcial para obtenção do grau de **Doutor em Física**.

ORIENTADOR: Prof. José Carlos Merino Mombach

Santa Maria, RS
2016

Ficha catalográfica elaborada através do Programa de Geração Automática
da Biblioteca Central da UFSM, com os dados fornecidos pelo(a) autor(a).

Fontoura, Carla Adriane Ramos Segatto
Um método multiestatístico para identificação de vias genéticas diferencialmente expressas / Carla Adriane Ramos Segatto Fontoura.- 2016.
87 p.; 30 cm

Orientador: José Carlos Merino Mombach
Tese (doutorado) - Universidade Federal de Santa Maria, Centro de Ciências Naturais e Exatas, Programa de Pós-Graduação em Física, RS, 2016

1. Vias genéticas 2. Microarranjo 3. Expressão Gênica 4. Software R I. Mombach, José Carlos Merino II. Título.

End. Eletr.: carladriani@yahoo.com.br

Carla Adriane Ramos Segatto Fontoura

UM MÉTODO MULTIESTATÍSTICO PARA IDENTIFICAÇÃO DE VIAS GENÉTICAS DIFERENCIALMENTE EXPRESSAS

Tese de Doutorado apresentada ao Programa de Pós-Graduação em Física, Área de Concentração em Sistemas Complexos, da Universidade Federal de Santa Maria (UFSM, RS), como requisito parcial para obtenção do grau de **Doutor em Física**.

Aprovado em 19 de agosto de 2016:

José Carlos Merino Mombach, Dr. (UFSM)
(Presidente/Orientador)

Gilberto Orengo, Dr. (UNIFRA)

Adriano Velasque Werhli, Dr. (FURG)

Lucio Strazzabosco Dorneles, Dr. (UFSM)

Giovani Rubert Librelotto, Dr. (UFSM)

Santa Maria, RS
2016

DEDICATÓRIA

À Deus e a todos que amo!

AGRADECIMENTOS

*Dos pensamentos que me motivam, o que mais me impacta é o que diz: "Seu trabalho vai preencher uma parte grande da sua vida, e a única maneira de ficar realmente satisfeito é fazer o que você acredita ser um ótimo trabalho. E a única maneira de fazer um excelente trabalho é amar o que você faz (*Steve Jobs*)".*

Faço aqui, não somente uma declaração de amor a este trabalho, que me inseriu novos desafios e experiências, amizades e horizontes, mas também a cada um que, de uma forma ou de outra, contribuíram para a concretização deste sonho.

Aos meus mestres, Dr. José C. M. Mombach e Dr. Gastone Castellani, meu orientadores no decorrer deste trabalho, pela oportunidade, pela paciência, pela partilha de conhecimento e pelos ensinamentos para a vida.

Aos meus irmãos, parentes e amigos, o meu sincero agradecimento pelos momentos de descontração, conselhos e incentivos. Obrigada por dividirem comigo as minhas angústias e alegrias. Foi bom poder contar com vocês!

Não poderia deixar de agradecer também à CAPES e ao CNPQ pelo importante apoio financeiro, tanto no Brasil quanto no exterior, e a Universidade Federal de Santa Maria, por abrir as portas do ensino para que eu pudesse evoluir pessoal e profissionalmente.

Ninguém vence sozinho... OBRIGADA A TODOS!

Agradecimentos Especiais

Aos meus pais, Valmir e Elisa, aos quais devo meu caráter e determinação, os meus mais profundos agradecimentos por suas sábias lições de esperança, sempre acreditando no meu potencial. Foi através da confiança por vocês depositada em mim que meus sonhos foram se tornando realidade. Espero estar retribuindo, não por obrigação, mas por puro prazer, tudo de bom que vocês me proporcionaram. À vocês, que cumpriram não somente o papel de pais, mas também o de grandes amigos, o meu muito obrigada!

Ao meu extraordinário marido, Sandro Roberto Fontoura, sempre paciente e generoso em meus momentos de desânimo e falta de estímulo. Tu foste o meu alicerce sobre o qual se assentaram a minha coragem de prosseguir e o meu desejo de vitória. São poucas as palavras que conheço e insuficiente o espaço que este documento me possibilita utilizar para descrever a tua importância nesta conquista. Muito obrigada por seres meu companheiro, não se resumindo em marido, mas um amigo, parceiro para os bons e maus momentos, meu melhor conselheiro e incentivador. Meu amor, a ti eu dedico esta tese e a minha vida. Muito obrigada!

Não lembro em que momento percebi que viver deveria ser uma permanente reinvenção de nós mesmos...

(Lya Luft)

RESUMO

UM MÉTODO MULTIESTATÍSTICO PARA IDENTIFICAÇÃO DE VIAS GENÉTICAS DIFERENCIALMENTE EXPRESSAS

AUTORA: Carla Adriane Ramos Segatto Fontoura
ORIENTADOR: José Carlos Merino Mombach

A determinação das causas e origens de uma determinada doença é uma tarefa complexa, considerando que existe um grande número de genes comprometidos que interagem entre si (WATSON, 2006). Especialistas em Bioinformática trabalham na busca de uma perfeita integração entre a biologia e a informação, com o intuito de compreender os prováveis fatores que desencadeiam determinadas doenças (PEVZNER, 2000). Para tal, a metodologia revolucionária de Microarranjos (LOCKHART et al., 1996), baseada na expressão gênica de pacientes, tem sido amplamente utilizada para medir simultaneamente as mudanças e regulação dos genes do genoma sob certas condições biológicas, resultando em uma lista de genes que podem ser considerados interessantes do ponto de vista biológico para uma determinada doença. Na presente tese, nós apresentamos um método multiestatístico destinado à detectar vias genéticas diferencialmente expressas em dados de microarranjos de DNA. Grande parte dos métodos de análise estatística são baseados no uso de apenas um teste estatístico. Acredita-se que associar métodos estatísticos baseados em testes diferentes diminui o número de falsos positivos. O método que nós desenvolvemos determina a atividade das vias avaliadas, e verifica se as alterações encontradas são estatisticamente significativas através dos testes de bootstrap, exato de Fisher e Wilcoxon. Este método pode ser aplicado à dados de transcriptoma para investigar quais vias apresentam mudanças na expressão de seus genes quando submetidos à algum tipo de perturbação. Implementado em linguagem R e disponibilizado para *download* no CRAN (do inglês, *Comprehensive R Archive Network*) como um pacote denominado PATHChange, nosso método demonstrou consistência entre os seus resultados com os previstos na literatura quando testado para dados públicos de microarranjos de câncer e pré-câncer de cólon. O método do PATHChange oferece um tipo alternativo de análise de vias de genes diferencialmente expressas para os pesquisadores que buscam apurar fenótipos de doenças, tais como o câncer.

Palavras-chave: Vias genéticas. Microarranjo. Expressão gênica. *software* R

ABSTRACT

A MULTI-STATISTIC METHOD FOR IDENTIFICATION OF DIFFERENTIALLY EXPRESSED GENETIC PATHWAYS

AUTHOR: Carla Adriane Ramos Segatto Fontoura
ADVISOR: José Carlos Merino Mombach

The determination of the causes and origins of a given disease is a complex undertaking, considering that there is a large number of genes engaged that interact with each other (Watson, 2006). Bioinformatics experts working in the search for a perfect integration between biology and information, in order to understand the likely factors that trigger certain diseases (Pevzner, 2000). To achieve this, the revolutionary methodology of Microarrays (LOCKHART et al., 1996) based on the gene expression of patients, it has been widely used to simultaneously measuring changes and regulation of the genes of the genome under certain biological conditions, resulting in a list of genes that may be considered interesting from a biological point of view for a particular disease. In this thesis, we present a multi-statistic method to detect differentially expressed genetic pathways in DNA microarray data. Many statistical methods of analysis are based on the use of a single statistical test. It is believed that the use of multiple tests decreases the number of false positive discoveries. Our method can be applied to transcriptome data to investigate which pathways have changes in expression when subjected to some type of disturbance. The method determines the activity of pathways evaluated, and verifies if the changes found are statistically significant through the bootstrap, Fisher exact and Wilcoxon tests. Implemented in R language and available for download from the Comprehensive R Archive Network (CRAN) as a package called PATHChange, our method showed consistency in its results with those predicted in the literature when tested for microarray of cancer and pre-cancer colon public data. The PATHChange method offers an alternative type of analysis of differentially expressed genes pathways for researchers seeking to determine phenotypes of diseases such as cancer.

Keywords: Gene pathways. Microarray. Gene expression. R software

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE SÍMBOLOS

DNA	Ácido desoxiribonucleico
$mRMA$	RMA mensageiro
$RNA-seq$	sequenciamento de RNA
$cDNA$	DNA complementar
$CRAN$	do inglês *Comprehensive R Archive Network*
$GPL-2$	do inglês *General Public Licence version 2*
GEO	do inglês *Gene Expression Omnibus*
$KEGG$	do inglês *Kyoto Encyclopedia of Genes and Genomes*
FDR	do inglês *False Discovery Rate*
$\in$	Pertence
$\notin$	Não pertence

SUMÁRIO

1 APRESENTAÇÃO

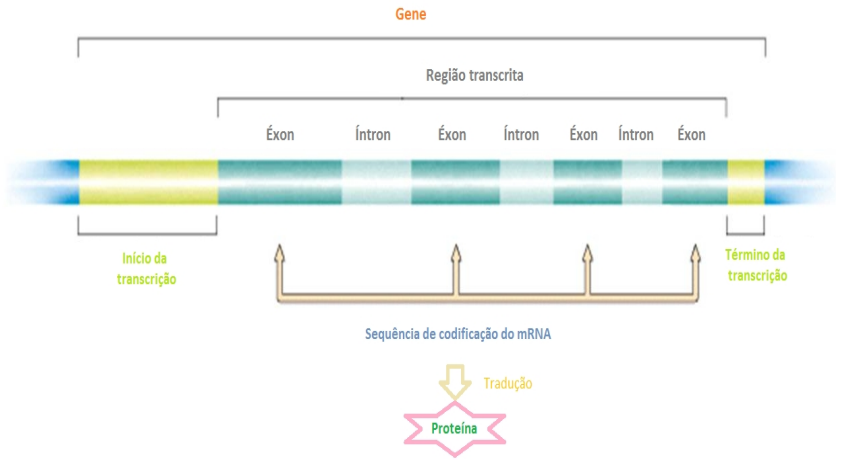
Descrever alterações fenotípicas em vias genéticas é uma tarefa importante para os pesquisadores que buscam interpretar dados biológicos. Para isso, utilizam-se métodos de análise estatística para determinar tais alterações. O fato de os métodos estatísticos disponíveis normalmente utilizarem apenas um teste estatístico no processamento dos dados nos motivou a desenvolver um método multiestatístico que combina os testes de bootstrap, exato de Fisher e Wilcoxon para avaliar as alterações em vias genéticas em dados de microarranjos de DNA. O método foi implementado em linguagem R através do pacote PATHChange, que está disponibilizado para download no CRAN (https://cran.r-project.org/web/packages/PATHChange/index.html).

1.1 INTRODUÇÃO

Com o advento do Projeto Genoma Humano, uma iniciativa dos Institutos Nacionais de Saúde dos Estados Unidos nos anos 90 que identificou e mapeou os genes que compõem o DNA das células do corpo humano, novas e promissoras tecnologias vem sendo desenvolvidas e aparecem como encorajadoras ferramentas metodológicas e científicas para os progressos na compreensão dos mecanismos que circundam várias doenças complexas, tais como as doenças degenerativas, doenças genéticas, doenças mentais e o câncer.

O genoma traz codificado no DNA as instruções que podem influenciar tanto os atributos físicos, como por exemplo a estrutura, o tamanho e a cor de um ser vivo, como também os aspectos comportamentais e a susceptibilidade a doenças. As informações necessárias para a criação de uma nova proteína está contida nos genes, que são as partes funcionais do DNA (éxons) conforme ilustração na Figura 1.1. Pra que uma nova proteína seja produzida, é necessário que as informações genéticas no DNA sejam transcritas em um produto intermediário, chamado RNA mensageiro. Nesse caso, a expressão gênica de um gene qualquer poderá ser medida através da quantidade de cópias de mRNA presente na célula. Assim, considera-se um gene altamente expresso quando há grandes quantidades de mRNA na célula, e pode-se investigar diferenças entre dois organismos (por exemplo, um doente e outro saudável) medindo as quantidades de mRNA transcritos nas duas condições e procurando aqueles genes que se expressam diferencialmente.

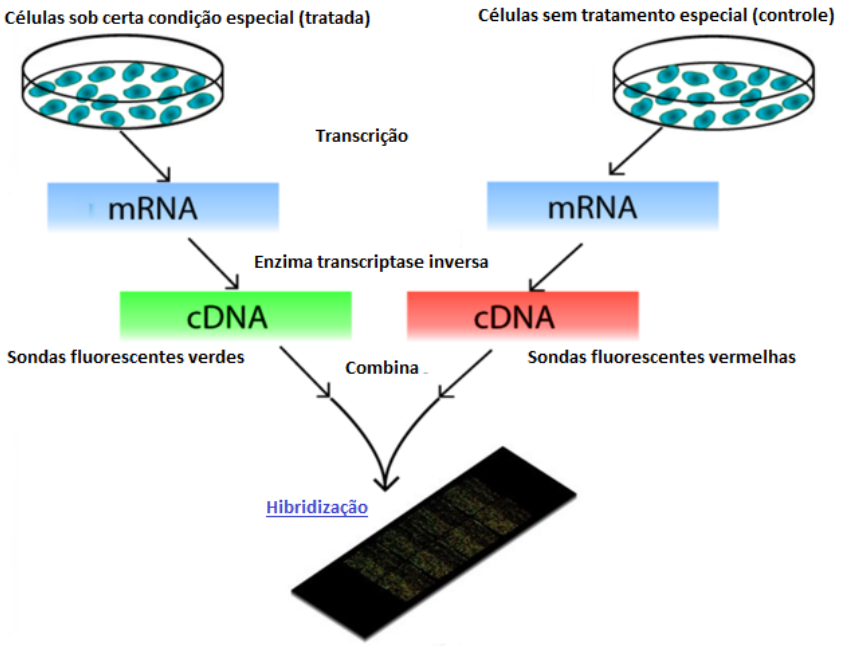
Figura 1.1 – Figura ilustrativa do processo de tradução de uma proteína à partir de um gene. A figura mostra o produto intermediário deste processo, mRNA, cujas quantidades de transcritos pode ser utilizada para medir a expressão de um gene.



Fonte: Próprio autor

Microarranjos de DNA têm a função de investigar os mecanismos genéticos de células através de análises dos níveis de expressão gênica. Esta abordagem é baseada no sequenciamento de transcritos (síntese do RNA a partir de um DNA molde), cujos conjuntos destes resulta em um transcriptoma. São chips contendo milhares de pequenos fragmentos de DNA (sondas), cada qual contendo uma informação sobre um fragmento de DNA desconhecido. Em linhas gerais, um experimento de microarranjos envolve duas fases principais: a deposição dos DNAs complementares nos *spots* (cavidades da lâmina do arranjo) e a preparação das amostras estudadas. A Figura 1.2 ilustra o processo de preparação de duas amostras de células sob duas condições (tratada e não tratada) das quais será retirado o material genético a ser hibridizado a um chip de microarranjo. Basicamente, o DNA inicialmente coletado das células é transcrito em RNA mensageiro. A enzima transcriptase inversa utiliza este RNA como molde para sintetizar o DNA complementar que será hibridizado ao chip. É possível perceber, ainda, que existe um passo adicional de identificação das amostras, onde é agregada coloração (fluorescente verde para a amostra tratada e fluorescente vermelha para a amostra controle). Esse passo é importante porque os níveis de expressão gênica serão obtidos a partir da leitura da intensidade luminosa de cada sonda (revertidos em valores numéricos) (DRĂGHICI, 2012; BARILLOT et al., 2012).

Figura 1.2 – Ilustração da metodologia de microarranjos de DNA. Partindo de células tratadas e não tratadas (controle), o DNA é transcrito em RMA mensageiro, que sofre a ação da enzima transcriptase inversa responsável pela síntese do DNA complementar. As sondas aparecem diferenciadas pelas cores verde (células tratadas) e vermelho (grupo controle) e serão hibridizadas no chip.



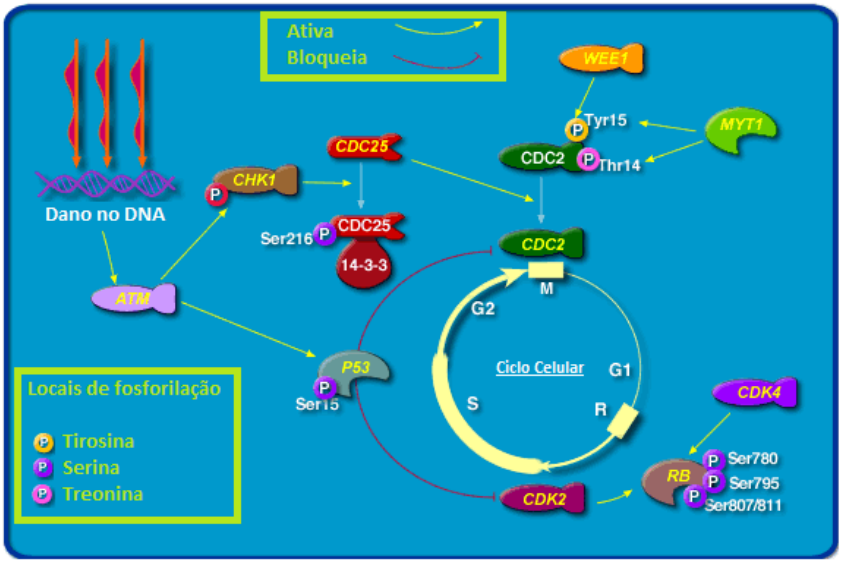
Fonte: Adaptado de Food Warden (2016)

De acordo com o interesse, o pesquisador pode utilizar um experimento de microarranjo para investigar genes, pacientes ou fenótipos (particularidades de um indivíduo determinadas pela expressão de seus genes, tais como cor da pele). Os mais comuns são os estudos focados nos genes, que visam basicamente aumentar a compreensão acerca de suas funções. Além disso, é comum estudos de microarranjos motivados pela condição clínica dos pacientes, que tem como objetivo favorecer o diagnóstico e prognóstico de uma doença. No entanto, estudos de microarranjos são considerados boas ferramentas para a investigação do fenótipo estudado, com a intenção de melhorar a compreensão da biologia da doença. Desse modo, informações sobre o fenótipo são obtidas a partir das funções dos genes, fornecendo uma ligação entre os genes e os processos biológicos associados a eles (DRĂGHICI, 2012; ANN; TALLOEN, 2010).

Genes com perfis de expressão semelhantes podem estar funcionalmente relacionados ou sob o mesmo mecanismo de controle. O conjunto de genes que se relacionam pelas suas funcionalidades é chamado de via (EISENBERG et al., 2000) e são responsáveis por conduzir a célula a um certo produto ou alteração, conforme o exemplo da Figura 1.3 em que a via do Retinoblastoma em resposta ao dano no DNA conduz a célula a parada do ciclo celular de crescimento através da inibição da quinase CDK2 resultante da ativação da via de ATM devido a um dano no DNA. A Figura 1.3 evidencia porque a análise de vias de genes é importante quando se pretende analisar fenótipos de doenças. Este tipo de análise possibilita obter informações à respeito das interações biológicas entre os genes envolvidos abrangendo os mecanismos mole-

culares causadores de doenças complexas (RITZ et al., 2016; YAP et al., 2016; KRIZKOVA et al., 2016). É, portanto, um método de análise de dados de transcriptoma mais específico e detalhado.

Figura 1.3 – Via da proteína do Retinoblastoma em resposta ao dano no DNA, cujo nome (do inglês) é *RB Tumor Suppressor/Checkpoint Signaling in response to DNA damage*. É a principal via supressora de tumor que controla as respostas celulares à estímulos potencialmente oncogênicos, tais como dano ao DNA. Na figura, as setas representam ativação e as barras, inibição. Através de um dano no DNA, a quinase CDK2 é inibida, resultando em uma parada do ciclo celular



Fonte: Adaptado de Cancer Genome Anatomy Project - informação da via fornecida por BioCarta (2016)

O fenótipo de organismos vivos é resultado de interações complexas envolvendo vias metabólicas e de sinalização. Além disso, a informação biológica é muito ruidosa, o que requer o uso de testes estatísticos para corrigir esses ruídos. Nesse sentido, ferramentas estatísticas de análise de amostras biológicas podem ser importantes aliados na identificação de vias alteradas em estudos comparativos (RITCHIE et al., 2015; THERNEAU; GRAMBSCH, 2000; VENABLES; RIPLEY, 2002). Em outras palavras, para uma análise de dados de expressão de vias genéticas realmente eficiente, faz-se necessário o uso de métodos estatísticos que irão garantir que uma dada via está diferencialmente regulada no estudo. O termo "estatístico" se refere a análise e interpretação dos dados biológicos visando avaliar com confiança a aleatoriedade e a incerteza das conclusões obtidas (ZAR, 1999).

À partir da criação da tecnologia de microarranjos, grande parte das pesquisas de biologia molecular passaram a adotá-la e as metodologias de análise estatística paralelamente progrediram, levando ao desenvolvimento constante de novos algoritmos para análise de mudanças de expressão gênica. No entanto, a maioria das ferramentas estatísticas desenvolvidas para analisar dados biológicos, tais como *limma* e *survival* (RITCHIE et al., 2015; THERNEAU; GRAMBSCH, 2000), ou ainda os pacotes específicos para análise de vias genéticas *GAGE* e *sigPathway* (LUO et al., 2009; LUO; BROUWER, 2013), são baseadas em apenas um

teste estatístico. Para aumentar a confiabilidade dos resultados, diminuindo os falsos positivos, acredita-se que métodos estatísticos podem combinar mais de um teste e obter uma maior precisão na significância resultante quando ela concorda com todos os testes usados (DRĂGHICI, 2012).

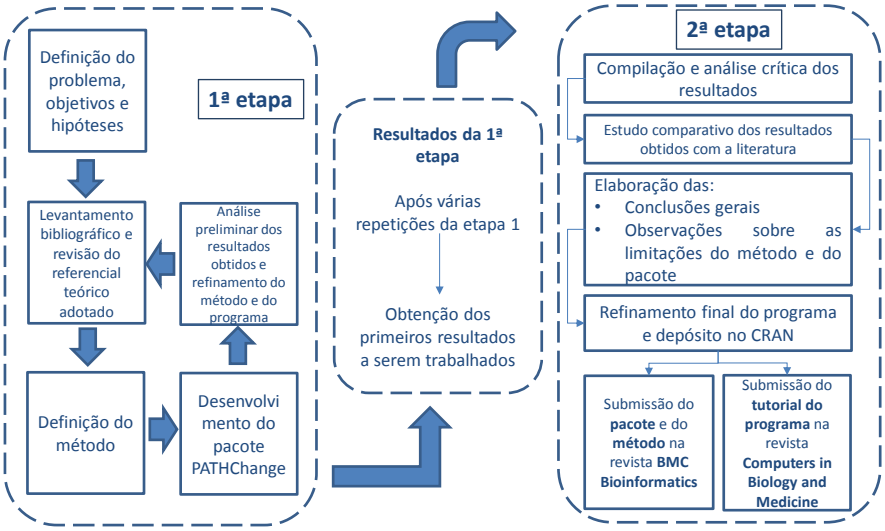
Neste contexto, a presente tese tem o propósito de apresentar e testar um método multiestatístico desenvolvido para detectar vias diferencialmente expressas em dados de microarranjos de DNA. Os objetivos do método aqui proposto são identificar as vias genéticas que manifestam consideráveis evidências de estarem diferencialmente expressas em dados de transcriptoma de câncer, identificar padrões nessas expressões diferenciais e correlacionar esses padrões com o câncer. O método propoe verificar essas alterações em dados de microarranjos através da combinação de três testes não paramétricos (assumindo que os nossos dados não se enquadram em nenhuma distribuição específica): Bootstrap (EFRON, 1979), que cria uma distribuição estatística através da reamostragem da amostra inicial; Wilcoxon pareado (WILCOXON, 1945), que busca semelhanças nas medidas das posições de duas amostras dependentes; e exato de Fisher (FISHER, 1934), que calcula um valor de p exato fazendo uso de uma tabela de contingência $2x2$. Fazendo uso dos benefícios da automação de modo a tornar mais produtivo e eficiente o uso do método, nós criamos PATHChange, um pacote em linguagem de programação R disponível para download no CRAN (<https://cran.r-project.org/web/packages/PATHChange/index.html>) sob a Licença Pública Geral versão 2 GNU (GPL-2).

Este trabalho foi parcialmente desenvolvido no Departamento de Física da Universidade Federal de Santa Maria, sob a orientação do prof. Dr. José Carlos Mombach, e entre os anos de 2014 a 2015 no Departamento de Física e Astronomia da Universidade de Bologna na Itália, por meio do programa de Doutorado Sanduíche no Exterior (SWE), sob a orientação do prof. Dr. Gastone Castellani. Durante este período, eu participei do projeto SYSAGEOMICS (Systems Biology of Aging and Age related diseases by integration of Multiscale Experimental and Computational Methods), que visa desenvolver métodos experimentais e computacionais para estudar doenças relacionadas ao envelhecimento, tais como o câncer.

O processo para desenvolvimento deste trabalho passou pelas etapas de desenvolvimento do método e do pacote PATHChange, e posterior submissão do trabalho para publicação, conforme organigrama da Figura 1.4. Primeiramente, foram definidos o problema, objetivos e hipóteses, e realizado um levantamento bibliográfico. A partir do estudo dos referenciais teóricos adotados, o método foi definido e o pacote PATHChange passou a ser desenvolvido. Através dos primeiros resultados obtidos, o método e o pacote foram refinados e novos resultados foram gerados. Esta etapa foi repetida diversas vezes até a obtenção dos primeiros resultados interessantes do ponto de vista biológico. Nesse ponto, iniciou-se a segunda etapa do trabalho que compreendeu a análise crítica dos resultados e validação dos mesmos através de comparações com as previsões da literatura. Após validados os resultados, foram elaboradas as conclusões finais à respeito do método proposto e as observações sobre as limitações encontradas no método

e no pacote. Nesse momento, foram consumidos vários meses de adaptação do pacote até que fossem atendidos os critérios de qualidade, instalabilidade e diversidade de sistemas operacionais requeridos pelo repositório CRAN do R. Finalmente, o trabalho encontrava-se pronto para a submissão nas revistas BMC Bioinformatics, onde foram apresentados o método e o pacote PATHChange, e Computers in Biology and Medicine, onde nós apresentamos um tutorial do uso do programa.

Figura 1.4 – Organograma apresentando as duas etapas de desenvolvimento do método e do pacote PATHChange.



Fonte: Próprio autor

O conteúdo e organização desta tese seguem o formato de artigos científicos integrados. Os artigos em questão são *PATHChange: an R tool for identification of differentially expressed pathways usign multi-statistic comparison*, submetido para publicação na revista *BMC bioinformatics* em 1 de julho de 2016, e *The R implementation of the CRAN package PATHChange, a tool to study genetic pathway alterations in transcriptomic data*, submetido para publicação na revista *Computers in Biology and Medicine* em 14 de julho de 2016. Estes artigos estão inseridos nos capítulos 2 e 3 no formato BMC e *Elsevier*, respectivamente, e seguem a numeração sequencial da tese. Os textos aqui desenvolvidos servem de apoio a estes trabalhos, explicitando os métodos e técnicas utilizadas, além de inserir discussões e conclusões a respeito dos resultados obtidos.

A descrição a seguir trata da apresentação matemática do método desenvolvido durante o meu trabalho de doutorado para detectar as vias diferencialmente expressas. Os detalhes apresentados nesta sessão podem ser encontrados no Artigo do Capítulo 2 desta tese (*Artigo submetido para publicação*).

1.2 REFERENCIAL TEÓRICO

Estudos envolvendo microarranjos medem simultaneamente os níveis de expressão de milhares de genes, ignorando a existência de correlações entre eles. No entanto, genes não trabalham isoladamente, mas em vias, e os dados de expressão medidos contem genes correlacionados e outros anticorrelacionados. Para solucionar esta discordância, sugere-se que as análises de microarranjos devem conter, além da análise dos níveis de expressão dos genes, uma análise de vias de genes (ANN; TALLOEN, 2010). Conforme já mencionado, o pacote PATHChange visa detectar alterações em vias genéticas e, para isso, utiliza os diversos bancos disponíveis, tais como KEGG (KANEHISA; GOTO, 2000), PathwayCommons (CERAMI et al., 2011), Reactome (MATTHEWS et al., 2009) e Ontocancro (LIBRELOTTO, 2009). Em especial, a Ontocancro é um banco de dados públicos que agrupa as principais vias de genes envolvidas na manutenção do genoma, tais como reparo do DNA, ciclo celular, apoptose e senescência.

A análise de vias de genes pode ser feita através do cálculo da atividade relativa (CASTRO et al., 2007), onde os níveis de expressão de conjuntos de genes podem ser determinados. Essa quantidade é dita relativa porque é baseada em comparações entre duas amostras de interesse. Em seu trabalho, Castro et al. (2007) consideraram que a medida da expressão da atividade relativa dos genes, n_α, pertencentes a via α poderia ser obtida pela relação:

$$n_\alpha = \frac{N_\alpha^e}{N_\alpha^e + N_\alpha^y}, \tag{1.1}$$

onde N_α^e é a soma da expressão dos genes para a amostra experimento e N_α^y é a soma da expressão dos genes para a amostra controle. n_α varia entre 0 e 1, de modo que $n_\alpha < 0.5$ significa que a amostra alterada apresenta uma atividade relativa menor do que a amostra controle, e $n_\alpha > 0.5$ representa o caso inverso.

A seguir, iremos descrever o método utilizado por PATHChange para detectar vias diferencialmente expressas em dados de transcriptoma.

1.2.1 Inferência Estatística

Considerando o log(sinal+1) nos dados de expressão em estudo, o método estatístico parte da quantificação e caracterização da expressão da atividade das vias através da equação 1.1. Para sabermos qual a precisão desses resultados e com que probabilidade podemos confiar nas conclusões obtidas, é indicado realizar um teste de hipóteses apropriado e assumir duas observações a respeito dos resultados (hipótese nula e hipótese alternativa), de modo que ambas sejam antagônicas. Conforme já mencionado, nosso objetivo principal é detectar as vias diferencialmente expressas em um estudo de microarranjos. Nesse sentido, para dar suporte a nossa afirmação de que aquelas vias são realmente diferencialmente expressas, nós à esco-

lhemos como hipótese alternativa. A hipótese nula testada, portanto, é de que as vias não são diferencialmente expressas.

Existem muitos métodos estatísticos diferentes para calcular a estatística p neste caso. No caso dos testes paramétricos, a exemplo do teste t, é preciso assumir que a distribuição de probabilidades do arranjo estudado seja conhecida. Caso essa suposição não seja satisfeita, faz-se necessário o uso de métodos não paramétricos de dedução estatística (ZAR, 1999). Dados de microarranjos nem sempre possuem uma distribuição normal. Assim, nós propomos um método estatístico baseado na combinação de três diferentes testes estatísticos não paramétricos (Bootstrap (EFRON, 1979), Fisher (FISHER, 1934) e Wilcoxon (WILCOXON, 1945)) visando garantir uma maior precisão na escolha das vias alteradas. p é a probabilidade de rejeição da hipótese nula ser verdadeira. No entanto, no caso de múltiplas comparações p pode também estar associado aos genes falsamente declarados diferencialmente expressos, o que chamamos de falsos positivos. Nesse caso, para evitar este tipo de erro, o método estatístico avalia a ocorrência de falsos positivos através de um teste de correção FDR (BENJAMINI; HOCHBERG, 1995; DRĂGHICI, 2012).

1.2.1.1 Método de bootstrap

O teste de bootstrap (EFRON, 1979) visa construir uma distribuição estatística através de um grande número de reamostragens (repetidas amostragens a partir de uma pequena parte de uma população) dos elementos provenientes de uma amostra de uma população inicial. Nós procedemos o bootstrap conforme a proposta de Castro et al. (2007), onde a população inicial, que seriam os genes do microarranjo que pertencem à via, será reamostrada n_{Boot} vezes, assumindo que o número de elementos de cada reamostragem será igual ao número de elementos presentes na via. Portanto, a fim de estimar a significância estatística (valor de p) de uma alteração em uma via α, o nosso método executa o bootstrap através da reamostragem aleatória com repetição dos genes pertencentes a esta via. Seja A a amostra de genes presentes no microarranjo, B a amostra de genes pertencentes a via e X_{AB} a interseção dos genes do microarranjo com os genes da via, a amostra aleatória gerada no bootstrap será X^*,

$$A = \begin{pmatrix} a_1 \\ \vdots \\ a_j \end{pmatrix}; B = \begin{pmatrix} b_1 \\ \vdots \\ b_m \end{pmatrix}$$

$$X_{AB} = \begin{pmatrix} x_1 \\ \vdots \\ x_\beta \end{pmatrix} \longrightarrow X^*_{reamostragem} = \begin{pmatrix} x^*_1 \\ \vdots \\ x^*_m \end{pmatrix}_{n_{Boot}},$$

onde j é o número de genes no microarranjo, m é o número de genes na via α, β corresponde ao

numero de genes pertencentes a X_{AB} e n_{Boot} é o número de reamostragens do bootstrap. Perceba que X^* terá o mesmo número de genes de α.

Sendo n_γ um vetor de atividades relativas obtidas através da Equação 1.1 aplicada a cada reamostragem gerada no bootstrap, o valor de p será dado por:

$$p = \frac{\sum_{i=1}^{n_{Boot}} (n_{\gamma_i} > n_\alpha)}{n_{Boot}}. \tag{1.2}$$

1.2.1.2 Teste de Wilcoxon

Assumindo nossos dados pareados, visto que, para cada amostra tida como experimento, nós teremos dados de expressão de uma amostra controle para confrontar, é possível estimar a significância estatística da alteração na via realizando um teste pareado. Nós escolhemos realizar um teste não paramétrico equivalente ao teste t para amostras pareadas, chamado teste de Wilcoxon pareado (WILCOXON, 1945), não assumindo hipóteses sobre a distribuição de probabilidades da população inicial. No teste de Wilcoxon, para cada via é calculado o valor absoluto da diferença d_j de expressão entre o experimento e controle de cada gene.

$$d_j = x_{1j} - x_{2j}. \tag{1.3}$$

Os valores de d_j resultantes serão ranqueados em ordem crescente (ignorando os sinais). Na sequência, é determinado o número de diferenças positivas (T_+) e negativas (T_-). O valor da estatística p é determinado de acordo com o valor crítico, $T_{\alpha(2).n}$ (ZAR, 1999). Para vias com um grande número de genes, a estatística será calculada para uma distribuição normal:

$$\mu_T = \frac{n(n+1)}{4},$$

$$\sigma_T = \sqrt{\frac{n(n+1)(2n+1}{24}},$$

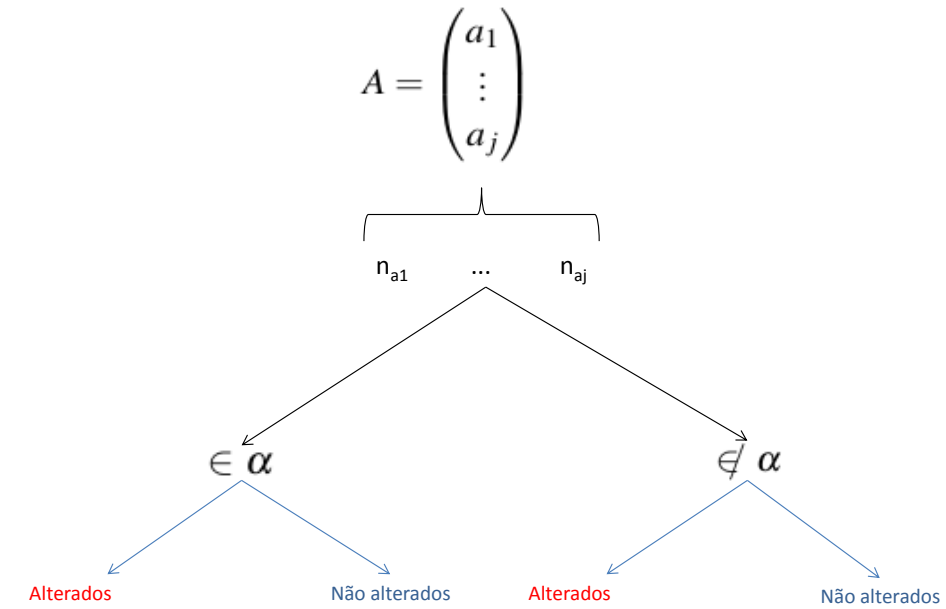
onde n é o número de genes presentes na via, μ_T é a média e σ_T é o desvio padrão. O valor de z será obtido através da seguinte relação:

$$z = \frac{T - \mu_t}{\sigma_t}. \tag{1.4}$$

1.2.1.3 Teste de Fisher

Outra importante ferramenta para estimar a alteração significativa de uma via é realizar um teste proposto por Ronald Fisher, conhecido como teste exato de Fisher (FISHER, 1934), que pode ser utilizado para classificar objetos em diferentes categorias (no nosso caso, alterados e não alterados). Para detectar a diferença específica nas proporções dos genes alterados na via α e no arranjo, diferentemente do procedimento realizado no bootstrap, onde eram descartados os genes que não faziam parte da via e o cálculo da atividade relativa dos genes da via considerava a média dos valores de expressão dos casos experimento e controle, para o teste de Fisher nós consideramos o conjunto total de genes pertencentes ao microarranjo (incluindo os genes não pertencentes a via) e aplicamos o cálculo da Equação 1.1 para cada gene individualmente, de modo que N_α^e e N_α^y são os logaritmos dos dados de expressão de cada gene do experimento e controle, respectivamente. Os resultados obtidos foram agrupados em genes pertencentes à via e genes não pertencentes a via, e em genes alterados e genes não alterados, conforme a Figura 1.5.

Figura 1.5 – Processo de agrupamento dos genes do microarranjo no teste exato de Fisher.



Fonte: Próprio autor

Em situações onde os dados foram coletados simultaneamente para duas variáveis e se deseja verificar as frequências de ocorrências das variáveis testadas, é indicado organizar esses dados em uma tabela de contingência *r*x*c*, onde *r* denota o número de linhas e *c* é o número de

colunas da tabela (ZAR, 1999). O menor arranjo de uma tabela experimental possível consiste em duas linhas e duas colunas (2x2). Existem diferentes formas de delineamento experimental que resultam em uma tabela de contingência 2x2. Uma delas ocorre quando ambas as margens na tabela de contingência são fixadas, cuja análise pode ser mais eficientemente feita através de um teste exato de Fisher (FISHER, 1934), que é baseado na construção de uma distribuição hipergeométrica.

Nesse caso, o teste exato de Fisher foi construído sobre uma tabela de contingência 2x2 (Tabela 1.1), onde o grupo total de genes do arranjo, N_{Tot}, foi rearranjado nos subconjuntos de genes que pertencem a via α e são diferencialmente expressos, e aqueles que não são, e o subconjunto remanescente de genes que não pertencem a via α.

Tabela 1.1 – Tabela de Contingência 2x2: o conjunto de genes no arranjo (N_{Tot}) podem ser agrupados em genes com um aumento na expressão $\in \alpha$, genes com um decréscimo na expressão $\in \alpha$, genes com um aumento na expressão $\notin \alpha$ e genes decréscimo na expressão $\notin \alpha$.

	Aumento na Expressão	Decréscimo na Expressão	
$\in \alpha$	n_{11}	n_{12}	$N_{l1} = \sum_{j=1}^{2} n_{1j}$
$\notin \alpha$	n_{21}	n_{22}	$N_{l2} = \sum_{j=1}^{2} n_{2j}$
	$N_{c1} = \sum_{i=1}^{2} n_{i1}$	$N_{c2} = \sum_{i=1}^{2} n_{i2}$	$N_{Tot} = \sum_{i,j} n_{ij}$

Fonte: Adaptado de Drăghici et al. (2003)

onde:
n_{11} = número de genes com aumento na expressão $\in \alpha$
n_{12} = número de genes com decréscimo na expressão $\in \alpha$
n_{21} = número de genes com aumento na expressão $\notin \alpha$
n_{22} = número de genes com decréscimo na expressão $\notin \alpha$
$N_{l1} = \sum_{j=1}^{2} n_{1j}$ = número de genes $\in \alpha$
$N_{l2} = \sum_{j=1}^{2} n_{2j}$ = número de genes $\notin \alpha$
$N_{c1} = \sum_{i=1}^{2} n_{i1}$ = número de aumentos na expressão dos genes no arranjo
$N_{c2} = \sum_{i=1}^{2} n_{i2}$ = número de decréscimos na expressão dos genes no arranjo
$N_{Tot} = \sum_{i,j} n_{ij}$ = número total de genes no arranjo

As alterações observadas em cada via são obtidas a partir da Tabela 1.1, sobre a qual uma distribuição hipergeométrica é construída baseada na equação hipergeométrica 1.7. A probabilidade de se obter uma amostra com N_{l1} genes pertencentes a via a partir de uma distribuição hipergeométrica, de modo que a amostra contenha n_{l1} genes com aumento de expressão e n_{l2} com diminuição na expressão, é

$$p = \frac{\binom{N_{l1}}{n_{11}}\binom{N_{l2}}{n_{21}}}{\binom{N_{Tot}}{N_{c1}}}, \tag{1.5}$$

que é idêntica a

$$p = \frac{\binom{N_{c1}}{n_{11}}\binom{N_{c2}}{n_{12}}}{\binom{N_{Tot}}{N_{l1}}}. \tag{1.6}$$

As equações 1.5 e 1.6 podem ser reescritas na forma reduzida

$$p = \frac{N_{l1}!N_{l2}!N_{c1}!(N_{Tot}-N_{c1})!}{n_{11}!(N_{l1}-n_{11})!(N_{c1}-n_{11})!(N_{l2}-N_{c1}+n_{11})!N_{Tot}!},$$

$$p = \frac{N_{l1}!N_{l2}!N_{c1}!N_{c2}!}{n_{11}!n_{12}!n_{21}!n_{22}!N_{Tot}!} \tag{1.7}$$

resultando em um valor de p exato.

1.2.1.4 Correção para falsos positivos

Conforme citado anteriormente, nós assumimos que a hipótese nula nos testes acima mencionados seria de que as vias não são realmente diferencialmente expressas. No entanto, realizar múltiplas comparações de vias sem um tratamento especial pode levar a rejeição da hipótese nula, quando na verdade ela era verdadeira. Esse resultado é um falso positivo. Quando isso ocorre, chamar essas vias genéticas de diferencialmente expressas será uma decisão errônea, o que compromete a conclusão final do método (DRĂGHICI, 2012).

O problema de múltiplas comparações surge do fato de que, quando todas as vias são comparadas com mesmo nível de significância de 5%, por exemplo, a probabilidade p de que pelo menos um resultado significativo seja devido ao acaso aumenta com n, conforme a Equação 1.8:

$$p = 1 - (1 - 0.05)^n. \tag{1.8}$$

Por exemplo, para n=20, a probabilidade de que tenha pelo menos um resultado falso positivo é aproximadamente 64.15%. Para evitar esse tipo de erro sugere-se o uso de algumas correções, tais como a correção de Bonferroni, Holm's, Permutação, FDR, entre outros (DRĂGHICI, 2012). Em se tratando de redes de genes, nós escolhemos a correção FDR (BENJAMINI; HOCHBERG, 1995) por ser um teste mais simples.

Na prática, o método de controle FDR (BENJAMINI; HOCHBERG, 1995) faz um ajuste dos valores de p através do ordenamento crescente das vias de genes de acordo com as probabilidades obtidas nos testes independentes (que, no caso do PATHChange, serão os teste de bootstrap, Wilcoxon e Fisher). Considerando que existem n vias significativas, para um dado nível de significância α_e escolhido, o método ajusta os valores de p_i de cada via com o limiar calculado na equação 1.9 (DRĂGHICI, 2012):

$$p_i < \frac{i}{n}\alpha_e, \tag{1.9}$$

onde i é a posição da via no ordenamento.

1.2.2 A ferramenta PATHChange

A análise estatística de experimentos envolvendo microarranjos de DNA é complexa em função da magnitude dos dados que envolve alguns milhares de genes. Uma ferramenta largamente utilizada em estudos envolvendo análises estatísticas e gráficas, R (R Core Team, 2014) é uma linguagem e um ambiente que facilitam a manipulação e geração de dados (DRĂGHICI, 2012; ANN; TALLOEN, 2010). É uma linguagem poderosa quando aplicada no desenvolvimento de rotinas de bioinformática, devido a confiabilidade e facilidade no tratamento de dados de experimentos biológicos. Dentre suas inúmeras vantagens, destacam-se a sua vasta biblioteca de funções de suporte estatístico normalmente utilizadas em análises de dados, tais como cálculo de médias e desvio padrão, além da capacidade de otimização de seu uso através de pacotes adicionais disponibilizados nos repositórios suporte do R, tais como CRAN, Bioconductor (GENTLEMAN et al., 2004) e GitHub (TEAM, 2014). Esses pacotes extras podem contemplar os mais variados propósitos, permitindo que o usuário possa desenvolver ferramentas incorporando à sua rotina funções previamente desenvolvidas. Por exemplo, o processo de hibridização dos microarranjos de DNA gera impurezas que devem ser eliminadas através da normalização dos dados, o que pode ser feito através de pacotes disponibilizados para download no Bioconductor (ANN; TALLOEN, 2010; BECKER; FEIJÓ[1],).

Visando transpor a complexidade de uma análise estatística conforme a proposta pelo nosso método, nós desenvolvemos o pacote PATHChange que agrega as vantagens da ferramenta R aliado aos benefícios da técnica de Microarranjos. PATHChange está disponível para download no CRAN (<https://cran.r-project.org/web/packages/PATHChange/index.html>) sob a Licença Pública Geral versão 2 GNU (GPL-2). O pacote foi criado para realizar a análise sobre dados experimentais, incluindo amostras "controle". Estes dados são disponibilizados pelo GEO (EDGAR; DOMRACHEV; LASH, 2002), um banco de dados criado para oferecer suporte público para dados de expressão de genes formados principalmente pela tecnologia de microarranjos (BARRETT et al., 2011).

Em termos breves, os dados do GEO são organizados nas seguintes seções: *Platform*, *samples* e *series* (BARRETT; EDGAR, 2006). Identificados por um código de acesso semelhante a "GPLxxx", *Platform* é uma lista dos elementos (e suas descrições) que podem estar contidos em um certo experimento. *Samples* se refere a condição especial sob a qual uma amostra foi submetida. Cada *sample* é identificado por um registro de acesso do tipo "GSMxxx". Por fim, *series* são a descrição resumida do estudo científico, referenciadas pelo código de acesso "GSExxx". No decorrer do presente trabalho, falaremos em *DataSets*, que nada mais são do

que o agrupamento das amostras de uma *serie* (EDGAR; DOMRACHEV; LASH, 2002).

A seguir, vamos apresentar alguns detalhes do pacote PATHChange, tais como a estrutura e suas funcionalidades, conforme o Artigo do Capítulo 3 (*Artigo submetido para publicação*).

1.2.2.1 Estrutura e funcionalidade

A análise de PATHChange é feita sobre dados de expressão que incluem amostras que sofreram algum tipo de perturbação (experimento) e amostras de tecido normal (controle). Compreendendo que o ponto de partida do usuário do nosso método seriam os dados normalizados do GEO (conforme exemplo da Tabela 1.2), nós percebemos a necessidade de desenvolver algumas funções que fossem capazes de dar suporte à análise estatística. Por isso nós incluímos no pacote duas funções prévias à análise estatística e uma função adicional que fornecerá uma melhor visualização dos resultados obtidos pelo método. PATHChange foi, portanto, estruturado em quatro funções (veja a Figura 1.6): `PATHChangeDat`, `PATHChangeList`, `PATHChange`, `PATHChangeVenn`.

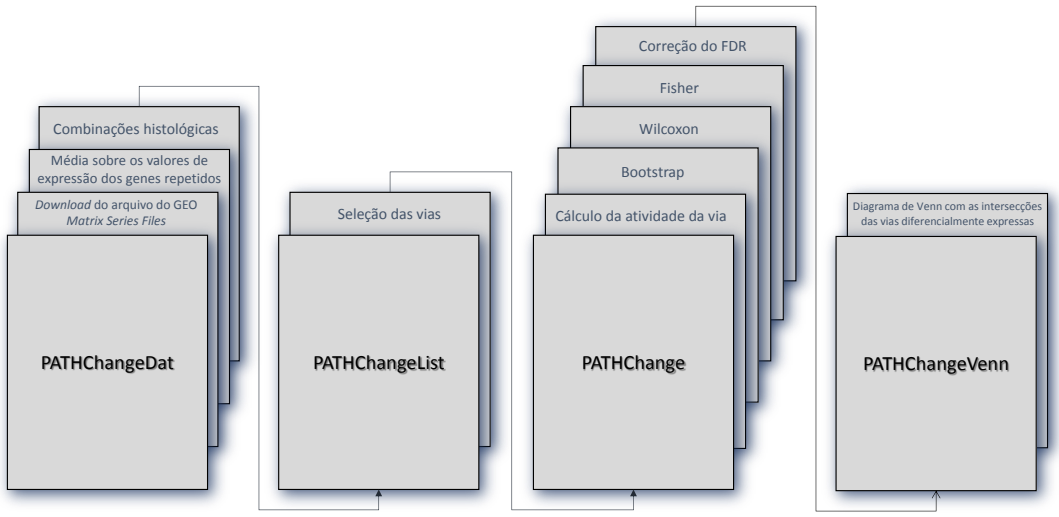
Tabela 1.2 – Exemplo de uma matriz de expressão normalizada com as linhas representando as sondas, uma coluna com os símbolos oficiais dos genes e as demais colunas (*samples* do GEO) representando as condições particulares a que foram submetidas as células em estudo. Cada célula da tabela exibe um valor fictício, que corresponde ao nível de expressão de um gene naquela condição.

Sondas	Símbolo	*sample* 1	*sample* 2	*sample* 3	*sample* 4
Sonda A	Gene A	5614	6446	5756	5498
Sonda B	Gene B	592	401	459	619
Sonda C	Gene C	246	238	261	207
Sonda D	Gene D	1233	813	647	663

Fonte: Próprio autor

A Figura 1.6 será utilizada para orientar o leitor sobre a ordem do processamento dos dados no pacote. Cada uma dessas funções realiza uma etapa importante do processamento dos dados do microarranjo, desde o pré-processamento desses dados e vias até a análise estatística e apresentação dos resultados. É importante ressaltar que, algumas funções servem de suporte para outras. No entanto, desde que o usuário utilize os argumentos das funções nos formatos adequados (conforme Artigo do Capítulo 3), ele não necessita utilizar todas as quatro funções.

Figura 1.6 – Quatro funções que estruturam o pacote PATHChange. Da esquerda para a direita: PATHChangeDat, o passo de pré-processamento dos dados; PATHChangeList, passo de seleção das vias estudadas; PATHChange, função principal do pacote que realiza o cálculo da atividade da via e aplica os testes estatísticos; PATHChangeVenn, que apresenta os resultados da análise da função PATHChange em forma de diagramas de Venn.



Fonte: Próprio autor

Agora, vamos descrever as finalidades de cada função de PATHChange. A primeira função do pacote PATHChange, `PATHChangeDat`, é responsável por proceder o pré-processamento dos dados, onde o arquivo suplementar *Matrix Series Files* disponibilizado para *download* no GEO informa a quais condições especiais o estudo foi submetido. Esta informação é importante porque, frequentemente, *DataSets* apresentam várias condições experimentais e controles. Durante o desenvolvimento do pacote, ficou claro que seria interessante possibilitar ao usuário o agrupamento de todas as possíveis combinações das condições experimentais em uma única análise. Convenientemente, `PATHChangeDat` informa ao usuário as condições experimentais do estudo e pergunta quais combinações ele gostaria de realizar. Maiores detalhes sobre a inserção dos argumentos no programa podem ser encontrados no Capítulo 3. A função `PATHChangeDat` ainda realiza uma média sobre os valores de expressão dos genes repetidos no microarranjo. A presença de genes repetidos afeta a probabilidade de "pesca" de cada gene no bootstrap, e este passo corrige o problema, igualando as probabilidades de cada gene.

Seguindo a Figura 1.6, a próxima função do pacote PATHChange é `PATHChangeList`. A inserção desta função se fez necessária para tornar mais eficiente a análise das vias no programa. Perceba que, quando estamos interessados em investigar as vias envolvidas em uma certa condição experimental, fica praticamente impossível pensar em apenas uma via, considerando que, certamente, várias delas foram recrutadas para que aquela condição experimental pudesse evoluir. Com o propósito de permitir a inserção de um número maior de vias para uma

mesma análise, foi criada a função `PATHChangeList`, onde o usuário pode informar quais vias ele pretende avaliar naquele momento. A função vai, basicamente, listar todas as vias inseridas, de modo que o programa poderá realizar os cálculos para cada uma das vias individualmente. Obviamente, quanto maior for o número de vias, bem como o número de comparações das condições experimentais, maior será o tempo de cálculo do programa.

A seguir, `PATHChange` realiza a análise estatística descrita na sessão 1.2. A intenção das funções `PATHChangeDat` e `PATHChangeList` é justamente fornecer os arquivos formatados de acordo com as necessidades da função `PATHChange`. Pensando em fornecer ao usuário um arquivo final manipulável, unicamente na função `PATHChange` o usuário terá uma opção adicional de salvar os resultados em formato ".csv". O arquivo resultante da função `PATHChange` será uma tabela com quatro colunas (atividades, e valores de *p* dos testes de bootstrap, Fisher e Wilcoxon corrigidos), cujas linhas serão a lista dos nomes das vias consideradas na análise.

A última função do pacote PATHChange é `PATHChangeVenn`, onde o usuário poderá visualizar em um diagrama de Venn os consensos (intersecções) das vias alteradas significativamente (no nível de significância escolhido) para os três testes estatísticos.

O pacote importa automaticamente alguns pacotes adicionais, tais como "stringr", "rlist", "VennDiagram", "grDevices", "stats", "utils" e "grid", que desempenharão funções de suporte específicas no código de PATHChange. Existem algumas características comuns a todas as funções do pacote. Com o intuito de controlar o depósito de arquivos no espaço interno dos computadores dos usuários do método, todos os arquivos resultantes das funções do pacote PATHChange serão salvos em uma pasta temporária, podendo ser acessados ao término do processamento da função. Caso o usuário pretenda salvar esses arquivos em uma pasta definitiva, ele deverá informar isso ao programa utilizando o argumento `writeRDS` (ou `writeCSV` unicamente para a função `PATHChange`), além de informar o endereço desta pasta através do argumento `destDIR`. Como todos os arquivos resultantes do programa são salvos em formato .rds, para acessá-los é indicado utilizar o pacote "rlist". Os códigos fonte de cada uma das funções encontram-se nos Anexos B, C, D, E.

O Artigo do Capítulo 3 possui demonstrações de como inserir os argumentos nas funções, bem como a preparação dos arquivos *input* e visualização dos arquivos finais resultantes do método.

2 ARTIGO 1: PATHChange: an R tool for identification of differentially expressed pathways using multi-statistic comparison

Artigo submetido à publicação que apresenta a descrição matemática do método multi-estatístico desenvolvido, bem como a aplicação do método em quatro *DataSets* do GEO que comparam o fenótipo de pré-cancer com câncer de cólon. O método claramento mostra uma distinção entre os fenótipos de câncer e pré-cancer. O material suplementar referenciado neste artigo encontra-se no anexo A da presente tese.

RESEARCH

PATHChange: an R tool for identification of differentially expressed pathways using multi-statistic comparison

Carla ARS Fontoura[1†], Enrico Giampieri[3], Daniel Remondini[3], Giovani Librelotto[2], Gastone Castellani[3*] and Jose CM Mombach[1†]

*Correspondence:
gastone.castellani@unibo.it
[3]Departament of Physics and Astronomy, Bologna University, Viale Berti Pichat, 40127 Bologna, Italy
Full list of author information is available at the end of the article
[†]Equal contributor

Abstract

Background: An important objective in post-genomics is the description of phenotype alterations in terms of gene pathways behavior. Statistical analysis provide methods to determine alterations in data obtained from transcriptomic studies. Usually these methods are based on a single test and can present false positive discoveries when compared with other methods based in different tests. To avoid that, methods that consider more than one test are more efficient to detect true alterations and are preferred. Multi-statistic tools that evaluate differential expression of gene pathways are still lacking.

Results: We developed a multi-statistic R tool suited to Affymetrix microarray data to evaluate pathway alterations through the combination of three different non-parametric tests: Bootstrap, Fisher's and Wilcoxon signed rank. Application of the tool to microarray data of pre-cancer and cancer are presented showing good agreement with experimental models proposed to explain their specific phenotype.

Conclusion: PATHChange is multi-statistic approach developed in R designed to improve the detection of differentially expressed pathways in transcriptomic studies providing an additional support to researchers in their phenotype investigations. The code is available under GNU General Public License 2 and can be downloaded from Comprehensive R Archive Network (CRAN).

Keywords: Gene pathways; Functional genomics; Microarray; Cancer; Gene Expression

Background

In cells genes work in orchestrated groups called pathways whose alterations provide an improved description of phenotype changes than the analysis of the behavior of single altered genes. Statistical tools that deal with functional information like gene pathways can provide better insights of the causal origins of a perturbed state, for example, a pathological state, and its normal (unperturbed) state. So, alternative methods and tools that can robustly identify gene pathway alterations in data obtained from high-throughput technology are essential to understand cellular responses. Most statistical methods employed in the analysis of biological data are based on a single statistical test and can present discoveries that are not confirmed by a different test when applied to the same data. In order to improve the detection of alterations some approaches combine more than one test and consider the prediction of an alteration robust when it is detected by all tests used [1] based on

the idea that a comprehensive statistical tool can outperform the results of a single one. To our knowledge, techniques and/or tools employing multi-statistics analysis of pathways alterations are still missing.

Methods like Gene Set Enrichment Analysis (GSEA) based on single statistical test are the most used to study pathway alterations. For a given phenotype, they rank genes based on the correlation between their expression and calculate a score that express the degree to which a given pathway appears at the top or the bottom of a ranking list [2].

Here we present a multi-statistic approach developed in R language [3] based on a combination of 3 non-parametric tests: Bootstrap [4], Fisher's exact [5] and Wilcoxon signed-rank [6]. These tests were selected because transcriptomic data can significantly deviate from a normal distribution. Other choice of non-parametric tests would also be valid but we think it is interesting to use those because the Bootstrap is used in some tools to analyze pathways [7], Fisher's test yields an exact p-value and Wilcoxon signed rank allows to compare related samples. This proposed approach is suited for comparative studies with dichotomic data: control (or normal) and altered (or perturbed).

Implementation

We developed an open-source R package, PATHChange, that detects alterations in genetic pathways provided by the user using three statistical tests. PATHChange can be downloaded from CRAN and the source code is available under the GNU General Public License 2.

The script was designed to perform the analysis over experimental data that include control samples. The tests used are non-parametric as we do not make any assumption about the parameters of the population from which the data were obtained. Consider a list of pathways of interest drawn, for example, from databases like KEGG [8], Reactome [9], PathwayCommons [10], Ontocancro [11] and others.

The Bootstrap method

This method is used to build a statistical distribution by resampling a large number of times the elements of a sample belonging to a given population. Consider a list of pathways where a given pathway α has n genes. This test considers the sum of the activities of the genes belonging to the pathway as proposed in [12]. The Bootstrap estimates the statistical significance (p-value) of an alteration in the α pathway by generating a large number of random samples (with replacement) of its n genes to estimate the statistical distribution of the population studied [1]. For each random sample we calculate the relative expression activity as follows. For the n_α genes in the α-pathway we calculate the ratio

$$n_\alpha = \frac{N_\alpha^e}{N_\alpha^e + N_\alpha^y}, \tag{1}$$

where N_α^e is the sum of gene expression activity of altered tissue and N_α^y is the sum of gene expression activity of control [12]. n_α varies between 0 and 1, where

$n_\alpha < 0.5$ means that the altered sample has lower gene activity relative to the control sample, and $n_\alpha > 0.5$, represents the contrary case.

Using the distribution for n_α for the α pathway obtained with the resampling above we obtain its p-value.

Fisher's exact test

Another approach to estimate the statistical significance of a change in a pathway is to perform Fisher's exact test which was proposed by Ronald Fisher [5] to classify objects in different categories, in our case, altered or unaltered pathway. The test detects differences in the proportion of altered genes in the α-pathway and in the whole dataset through a 2x2 contingency table. In this case we calculate N_α^e and N_α^y over the expression information of each gene for experiment and control, respectively. The group of N_{Tot} genes of the dataset is sorted in 4 groups and counted: n_{11} is number of genes in the pathway with increased relative expression and n_{12} with decreased relative expression. n_{21} is the number of genes not belonging to the pathway with increased expression and n_{22} with decreased expression. The changes observed in each pathway are obtained from the contingency table (Table 1).

Table 1: Contingency Table. The N_{Tot} genes in the dataset are sorted in 4 groups and counted: number of genes with (1) increased expression $\in$ pathway, (2) with decreased expression $\in$ pathway, (3) with increased expression $\notin$ pathway and (4) with decreased expression $\notin$ pathway. The sums are performed over the columns and rows of the table.

	Increased Expression	Decreased Expression	
$\in$ Pathway	n_{11}	n_{12}	$N_{l1} = \sum_{j=1}^{2} n_{1j}$
$\notin$ Pathway	n_{21}	n_{22}	$N_{l2} = \sum_{j=1}^{2} n_{2j}$
	$N_{c1} = \sum_{i=1}^{2} n_{i1}$	$N_{c2} = \sum_{i=1}^{2} n_{i2}$	$N_{Tot} = \sum_{i,j} n_{ij}$

Then the hypergeometric distribution is calculated over the elements of the contingency table yielding an exact p-value (Eq. 2) for a change of the pathway [5]:

$$p = \frac{N_{l1}!N_{l2}!N_{c1}!N_{c2}!}{N_{Tot}!n_{11}!n_{12}!n_{21}!n_{22}!} \tag{2}$$

Wilcoxon signed rank test

To conclude we apply another test to compare related samples, the Wilcoxon signed rank test, which is an alternative to the paired samples Student's t-test when the differences can severely depart from normally distributed. For each α pathway we calculate the absolute value of the difference of expression between experiment and control for each gene (T) and rank the genes in increasing order. We then determine the number of positive (T_+) and negative (T_-) differences and use the Wilcoxon rank sum test to obtain the p-value of the alteration of the α pathway.

Finally, since this approach involves multiple comparison (one for each pathway) the false discovery rate (FDR) must be controlled. In our case by means of Benjamini-Hochberg correction [13].

The final result of the approach is the consensus set of pathways whose adjusted p-value is significant and which we propose to be the best choice for a robust analysis. In the application below we used a level of significance of 5% in all tests described above.

Results and Discussion

Colon (or colorectal) adenoma (CRA) is a type of pre-cancer that easily evolves to cancer. Colon cancer development is well understood, it progresses gradually from inflammatory bowel disease to adenoma and then to colon carcinoma (CRC) making it a good system for study of phenotype alterations during tissue transformation [14, 15, 16]. A recent model based on several experimental studies of pre-cancerous lesions propose how these tissues evolve to cancer [17]. According to the model, pre-cancerous lesions have increased DNA repair response, apoptosis and senescence pathways playing the role of a barrier to tumor progression as they block the propagation of mutated cells that could give rise to a cancer. However, later the barrier is lowered by means of specific mutations, such as p53, that disrupt some these pathways [17]. This implies that we should observe more of these pathways activated in pre-cancer than in cancer tissues. In order to check if our tool can detect these phenotypic differences we analyzed public gene expression data obtained from GEO (Gene Expression Omnibus database, www.ncbi.nlm.nih.gov/geo) on colorectal adenoma, colon carcinoma and other cancer tissues such as breast, ovarian and prostate.

We selected data containing 5 adenoma and 4 carcinoma studies with respective GEO ID's: GSE8671, GSE15960, GSE20916 and GSE37364 [18, 19, 20, 21].

Detailed information of each dataset is presented below. The pathways used in the analysis were obtained from the Ontocancro project (http://ontocancro.inf.ufsm.br). This database provides a repository of pathways information obtained from curated databases on inflammatory and genome maintenance mechanisms that are important in pathologies associated with aging, like cancer [11, 22]. Currently, the database stores information about 51 pathways and 1105 genes. 49 pathways were tested and the results are presented in what follows.

Details of GEO datasets used:

- GSE8671 - Biopsy of colorectal adenomas and normal mucosa from 32 patients.
- GSE15960 - Gene expression profiles of normal, adenoma and carcinoma tissues from 18 human colonic epithelial cells;
- GSE20916 - Gene expression analysis data from RNA obtained from 105 macro- and 40 microdissected specimens, representing normal, adenomas and carcinomas colon samples. For microdissected specimens, we use 5 normal, 5 adenoma and 5 carcinoma mucosa samples.
- GSE37364 - Total RNA extracted from colonic biopsy samples of 38 healthy patients, 29 with adenoma and 27 with colorectal cancer;

Fig. 1 shows the Venn diagrams of altered pathways for adenomas and carcinomas dataset (GSE20916) with the selected consensus sets common to all 3 methods. The consensus pathways for the other datasets can be found in the supplementary material. Table 4 presents a summary of the biological function of each pathway. For adenoma (Fig. 1a) 12 pathways are detected by all tests and 10 for carcinoma (Fig. 1b). In Fig. 2 we present the relative activity eq. 1 for all pathways in Table 2 for adenoma and carcinoma showing an increased of activity of cancer in relation to adenoma, as expected. Table 2 shows a summary of the results: It lists all consensus pathways identified in all available datasets for adenoma (5 studies) and carcinoma (4 studies). Most of the altered pathways are related to cell cycle activity what is expected due to the proliferative activity of these pathologies. The main difference between pre-cancer and cancer in Table 2 is the presence of several DNA damage response pathways and a pathway of activation of apoptosis (Caspase cascade in apoptosis) for pre-cancer which characterizes the barrier proposed by the biological model mentioned above providing support to our method [17]. Table 3 shows all altered pathways found with PATHChange for prostate, breast and ovarian cancer. Here we observe a result similar to colon carcinoma, where different pathways related to cell cycle activity predominate and DNA damage response pathways are missing.
In summary, our results clearly show that pre-cancer and cancer phenotypes match the predictions of the biological model of cancer development.

Conclusion

We presented PATHChange a multi-statistic approach designed to improve the detection of differentially expressed pathways in transcriptomic studies. The application of the tool to cancer development data can distinguish, at transcriptional level, the phenotype of pre-cancer and cancer tissues matching the predictions of a general model of cancer evolution.

Availability and requirements

- **Project name:** PATHChange
- **Project home page:** https://cran.rstudio.com/web/packages/PATHChange/
- **Programming language:** R (version 3.3.0)
- **Other requirements:** stringr, rlist, VennDiagram, grDevices, stats, utils, grid
- **License:** GNU General Public License 2
- **Restrictions to use by non-academics:** According to GNU General Public License 2

Additional Files
Additional file 1 — Software manual
CRAN package description of PATHChange.

Additional file 2 — Supplementary results
Complete results of processed GEO datasets.

Ethics statement and consent
Does not apply.

Consent to publish
Does not apply.

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

Author's contributions
CARSF, GC, DR, EG, GRL, JCMM conceived the idea. CARSF developed the package and the analysis. CARSF, EG, DR and JCMM wrote the paper. All authors read and approved the manuscript.

List of abbreviations
CRAN, comprehensive R archive network; GSEA, gene set enrichment analysis; KEGG, Kyoto encyclopedia of genes and genomes; FDR, false discovery rate; CRA, colorectal adenoma; CRC, colorectal carcinoma; DNA, deoxyribonucleic acid; GEO, gene expression omnibus.

Acknowledgements
Work supported by Coordenação de Aperfeiçoamento de Pessoal de Ensino Superior (CAPES) and Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) (249374/2013-7, 402547/2012-8, 304805/2012-2).

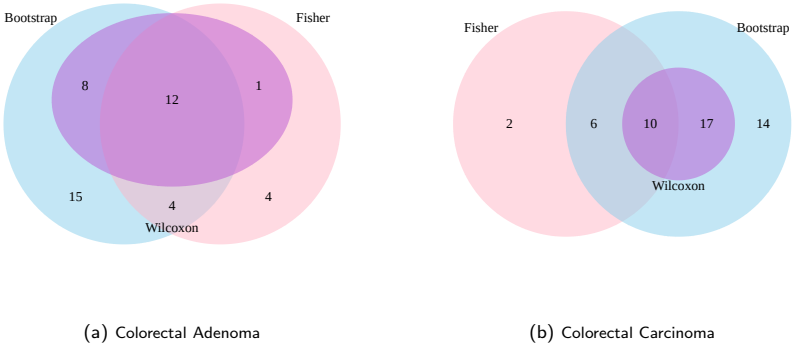
Author details
[1]Departament of Physics, Universidade Federal de Santa Maria, Avenida Roraima, 97105-900 Santa Maria, Brazil. [2]Departament of Electronics and Computing, Universidade Federal de Santa Maria, Avenida Roraima, 97105-900 Santa Maria, Brazil. [3]Departament of Physics and Astronomy, Bologna University, Viale Berti Pichat, 40127 Bologna, Italy.

References

1. Drăghici, S.: Statistics and Data Analysis for Microarrays Using R and Bioconductor. Chapman & Hall/CRC Mathematical and Computational Biology Series, London (2012)
2. Subramanian, A.e.a.: Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proceeding of the National Academy of Sciences of the USA **102(43)**, 15545–15550 (2005)
3. R Core Team: R: a Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria (2014). R Foundation for Statistical Computing. http://www.R-project.org/
4. Efron, B.: Bootstrap methods: another look at the jacknife. The Annals of Statistics **7** (1979)
5. Fisher, R.A.: Statistical methods for research workers. In: Crew, F.A.E., Cutler, D.W. (eds.) Biological Monographs and Manuals. Oliver And Boyd Tweeddale Court ; Edinburgh ; Paternoster Row ; London, ??? (1934)
6. Wilcoxon, F.: Individual comparisons by ranking methods. Biometrics bulletin **1**(6), 80–83 (1945)
7. Castro, M.A.A.e.a.: Viacomplex: software for landscape analysis of gene expression networks in genomic context. Bioinformatics **25(11)** (2009)
8. Kanehisa, M., Goto, S.: Kegg: kyoto encyclopedia of genes and genomes. Nucleic Acids Research **28**(1), 27–30 (2000)
9. Matthews, L., Gopinath, G., Gillespie, M., Caudy, M., Croft, D., de Bono, B., Garapati, P., Hemish, J., Hermjakob, H., Jassal, B., *et al.*: Reactome knowledgebase of human biological pathways and processes. Nucleic Acids Research **37**(suppl 1), 619–622 (2009)
10. Cerami, E.G., Gross, B.E., Demir, E., Rodchenkov, I., Babur, Ö., Anwar, N., Schultz, N., Bader, G.D., Sander, C.: Pathway commons, a web resource for biological pathway data. Nucleic Acids Research **39**(suppl 1), 685–690 (2011)
11. Librelotto, G.R.e.a.: An ontology to integrate transcriptomics and interatomics data involved in gene pathways of genome stability. In: Guimarães, K.S., Panchenko, A., Przytycka, T.M. (eds.) 4th Brazilian Symposium on Bioinformatics, BSB 2009, Porto Alegre, Brazil, July 29-31, 2009. Proceedings, pp. 164–167 (2009). Springer Berlin Heidelberg
12. Castro, M.A., Mombach, J.C., de Almeida, R.M., Moreira, J.C.: Impaired expression of ner gene network in sporadic solid tumors. Nucleic Acids Research **35**(6), 1859–1867 (2007)
13. Benjamini, Y., Hochberg, Y.: Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society, Series B (Methodological) **57(1)**, 289–300 (1995)
14. Farraye, F.A., Odze, R.D., Eaden, J., Itzkowitz, S.H., *et al.*: AGA medical position statement on the diagnosis and management of colorectal neoplasia in inflammatory bowel disease. Gastroenterology **138**(2), 738–745 (2010)
15. van Schaik, F.D., Mooiweer, E., van der Have, M., Belderbos, T.D., ten Kate, F.J., Offerhaus, G.J.A., Schipper, M.E., Dijkstra, G., Pierik, M., Stokkers, P.C., *et al.*: Adenomas in patients with inflammatory bowel disease are associated with an increased risk of advanced neoplasia. Inflammatory Bowel Diseases **19**(2), 342–349 (2013)
16. Triantafillidis, J.K., Nasioulas, G., Kosmidis, P.A.: Colorectal cancer and inflammatory bowel disease: epidemiology, risk factors, mechanisms of carcinogenesis and prevention strategies. Anticancer Research **29**(7), 2727–2737 (2009)
17. Halazonetis, T.D., Gorgoulis, V.G., Bartek, J.: An oncogene-induced DNA damage model for cancer development. Science **319**, 1352–1355 (2008)
18. Sabates-Bellver, J.e.a.: Transcriptome profile of human colorectal adenomas. Molecular Cancer Research **5(12)** (2007)
19. Galamb, O., Spisák, S., Sipos, F., Toth, K., Solymosi, N., Wichmann, B., Krenacs, T., Valcz, G., Tulassay, Z., Molnar, B.: Reversal of gene expression changes in the colorectal normal-adenoma pathway by ns398 selective cox2 inhibitor. British journal of cancer **102**(4), 765–773 (2010)

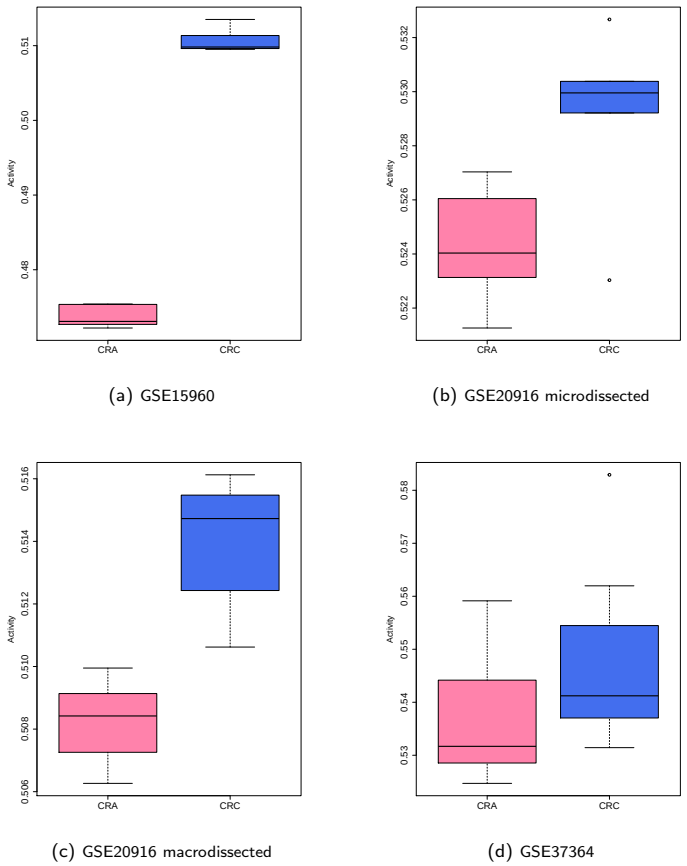
20. Skrzypczak, M., Goryca, K., Rubel, T., Paziewska, A., Mikula, M., Jarosz, D., Pachlewski, J., Oledzki, J., Ostrowsk, J.: Modeling oncogenic signaling in colon tumors by multidirectional analyses of microarray data directed for maximization of analytical reliability. PloS one **5**(10), 13091 (2010)
21. Valcz, G.e.a.: Myofibroblast-derived sfrp1 as potential inhibitor of colorectal carcinoma field effect. PLoS One **9(11)** (2014)
22. Campisi, J.: Aging, cellular senescence, and cancer. Annual Review of Physiology **75**, 685–705 (2013)

Figures



(a) Colorectal Adenoma (b) Colorectal Carcinoma

Figure 1: Diagrams showing the number of altered consensus pathways for (a) adenoma and (b) carcinoma for dataset GSE20916 macrodissected. Blue, pink and violet circles represent Bootstrap, Fisher and Wilcoxon tests, respectively.



(a) GSE15960

(b) GSE20916 microdissected

(c) GSE20916 macrodissected

(d) GSE37364

Figure 2: Diagrams of the relative activity for carcinoma (blue) and adenoma (pink) detected by PATHChange for (a) GSE15960, (b) GSE20916 (microdissected) (c) GSE20916 (macrodissected) and (d) GSE37364 datasets.

Tables

Table 2: PATHChange results of the analysis of the top significantly altered pathways of 5 pre-cancer and 4 cancer datasets from GEO database. The column 'Frequency' exhibits the number of datasets in which the pathway was detected as altered.

colorectal adenoma(CRA)		**colorectal carcinoma (CRC)**	
Pathway	Frequency	Pathway	Frequency
Cell Cycle Checkpoints	5	Cell Cycle Checkpoints	4
Cell Cycle, Mitotic	5	Cell Cycle, Mitotic	4
Mitotic Spindle Checkpoint	5	Mitotic Spindle Checkpoint	4
S Phase	5	S Phase	4
Regulation of Mitotic Cell Cycle	4	Regulation of Mitotic Cell Cycle	4
G1/S DNA Damage Checkpoints	4	G1/S DNA Damage Checkpoints	4
G2/M checkpoint	4	G2/M checkpoint	4
Mismatch Repair	4	Mismatch Repair	3
Caspase Cascade in Apoptosis	4		
Rb Tumor suppressor/Check. P. Sign. in Response to Damage	3	Rb Tumor suppressor/Check. P. Sign. in Response to Damage	3
Double-Strand Break Repair	3		
Non-Homologous end Joining	3		
Mitotic M-M/G1 Phases	2		
Fanconi Anemia Pathway	2		
		ATR signaling	3

Table 3: PATHChange results of the analysis of the significantly altered pathways of 2 datasets of prostate, breast and ovarian cancer.

GSE26910		**GSE22600**
Prostate	Breast	Ovarian
Caspase Cascade in Apoptosis	Caspase Cascade in Apoptosis	
Cell Cycle Checkpoints		Cell Cycle Checkpoints
Cell Cycle, Mitotic		Cell Cycle, Mitotic
G1/S DNA Damage Checkpoints	G1/S DNA Damage Checkpoints	G1/S DNA Damage Checkpoints
G2/M checkpoint	G2/M checkpoint	
		IL8- and CXCR2-mediated signaling events
		Mitotic Spindle Checkpoint
		Rb Tumor suppressor/Check. P. Sign. in Response to Damage
Regulation of Mitotic Cell Cycle	Regulation of Mitotic Cell Cycle	Regulation of Mitotic Cell Cycle
S Phase	S Phase	S Phase

Table 4: Short description of biological pathways found altered in our analysis.

Pathway	Description
Cell Cycle Pathways	
Cell Cycle Checkpoints	Regulates the arrests of the cell cycle at different checkpoints that occur during the transition from one phase of the cycle to another.
Cell Cycle, Mitotic	Genes involved in the cell cycle.
Mitotic Spindle Checkpoint	Regulation of the spindle checkpoint.
S Phase	S phase of the cell cycle.
Regulation of Mitotic Cell Cycle	Regulatory pathways of the cell cycle.
Mitotic M-M/G1 Phases	Regulation of mitosis and transition to G1 phase.
Apoptosis Pathways	
Caspase Cascade in Apoptosis	Caspase cascade of activation of apoptosis.
DNA Damage Response Pathways	
G1/S DNA Damage Checkpoints	Activation of the G1/S checkpoint due to DNA damage.
G2/M checkpoint	Activation of the G2/M checkpoint due to DNA damage.
Mismatch Repair	DNA repair of mismatched bases.
Rb Tumor suppressor/Check. P. Sign. in Response to Damage	Checkpoint activation of retinoblastoma pathway.
Double-Strand Break Repair	The two pathways that repair DNA double-strand breaks: homologous recombination (HR) and nonhomologous end-joining (NHEJ).
Non-Homologous end Joining	Double-strand breaks DNA repair.
Fanconi Anemia Pathway	Fanconi anemia DNA repair.
ATR signaling	ATR signaling due to DNA single-strand break.
Inflammatory Pathways	
IL8- and CXCR2-mediated signaling events	Interleukin 8 signaling pathway

3 ARTIGO 2: The R implementation of the CRAN package PATHChange, a tool to study genetic pathway alterations in transcriptomic data

Artigo submetido à publicação apresentando os detalhes da estrutura, implementação e um exemplo do uso do pacote PATHChange. O artigo apresenta a tabela de resultados do método proposto, um diagrama de Venn com as intersecções das vias diferente expressas nos testes estatísticos, além de uma tabela comparativa de outros pacotes de análise de vias genéticas.

The R implementation of the CRAN package PATHChange, a tool to study genetic pathway alterations in transcriptomic data

Carla A. R. S. Fontoura[a,*], Gastone Castellani[b], José C. M. Mombach[a]

[a] *Departament of Physics, Universidade Federal de Santa Maria, Avenida Roraima, 97105-900, Santa Maria, Brazil*
[b] *Department of Physics and Astronomy, Bologna University, Viale B. Pichat, 40123, Bologna, Italy*

Abstract

Tools that extract phenotype alterations from transcriptomic data are important to improve the interpretation of biological studies. PATHChange is a statistical CRAN package designed to work with data downloaded from the Gene Expression Omnibus database (GEO) to determine differential pathway expression in comparative studies including control samples. In this paper we present details of the structure, implementation and an example of use of the package.

Keywords: CRAN package, R, pathway expression, microarray, RNA-seq

1. Introduction

The large number of publications reporting results obtained with high-throughput data increases the confidence that microarray and RNA-seq tools are essential to measure the expression levels of large numbers of genes simultaneously [1, 2, 3]. These tools are used successfully in the investigation of the genetic mechanisms in living cells at transcription level.

Due to the fact that the phenotypes of living organisms are the result of thousands of complex interactions involving many metabolic and signaling pathways

*Corresponding author
Email addresses: carladriani@yahoo.com.br (Carla A. R. S. Fontoura), gastone.castellani@unibo.it (Gastone Castellani), jcmombach@ufsm.br (José C. M. Mombach)

[4], it is important to develop tools that allow us to identify altered pathways in comparative studies of biological samples.

In this paper, we present the implementation details of `PATHChange`, an R package designed to detect differentially expressed pathways in transcriptomic data based on three different statistical tests. It was developed using the statistical scripting language R [5], and available under the GNU General Public Licence 2. `PATHChange` is available for download from CRAN (`https://cran.rstudio.com/web/packages/PATHChange/`), providing a single flexible environment to evaluate genetic pathway alterations. The tool was conceived as a complementary source of information, then we recommend that any preprocessing of the data like normalization and/or batch-effect corrections to be performed before the use of PATHChange. The mathematical description of the statistical methods employed by the package will be published elsewhere [6].

Transcriptomic data is frequently used to describe comparative data like normal (or control) and altered (or perturbed), to search for different expression profiles in groups of genes. For that `PATHChange` package combines three different statistical tests (Bootstrap [7], Fisher exact [8] and Wilcoxon signed rank [9]) to evaluate significant alterations and reduce the number of false discoveries.

2. The `PATHChange` package structure

The core of the `PATHChange` package consists of four functions (see Figure 1): (i) `PATHChangeDat`, a data pre-processing step for download of Matrix Series Files the Gene Expression Omnibus (GEO, `http://www.ncbi.nlm.nih.gov/geo/`) [10], that calculates the average over the expression values of repeated genes and histology combinations; (ii) `PATHChangeList`, which selects the pathways to be evaluated; (iii) `PATHChange`, which performs the calculation of pathway activity, applies the bootstrap, Wilcoxon and Fisher tests, and computes the false discovery rate (FDR) correction [11]; and finally, (iv) `PATHChangeVenn`, which presents the results in the form of Venn diagrams of the differentially expressed pathways consensus.

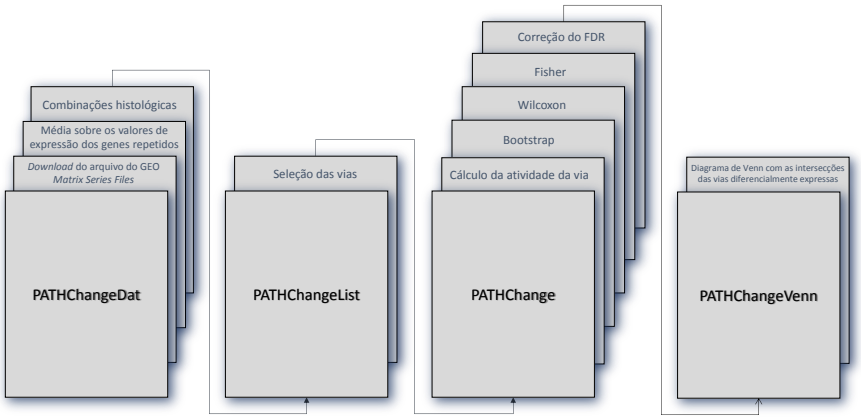


Figure 1: The four functions of the PATHChange package. From left to right: `PATHChangeDat`, a data pre-processing step; `PATHChangeList`, which selects the pathways to be evaluated; `PATHChange`, which performs the calculation of pathway activity, applies the statistical tests; `PATHChangeVenn`, which presents the results in the form of Venn diagrams.

2.1. Details of the package functions

In this section we detail other features of the package functions. Usually, datasets can have several experimental conditions and controls. Conveniently, `PATHChangeDat` detects these experimental conditions available in the Matrix Series Files downloaded from GEO and asks the user which comparisons he/she wants to make, offering the user the possibility to perform in the same analysis multiple comparative studies. Transcriptomic data can present repeated genes which affects the probability of choosing each gene in the Bootstrap. In this case, `PATHChangeDat` performs an average over the expression values of repeated genes.

As the focus of the analysis is pathways, it is interesting to investigate many pathways involved in that experimental condition at the same time. `PATHChangeList` reads the pathways from an input file (see below) and organizes them separately in lists, so that the statistical calculations will be pro-

cessed separately for each pathway considered.

The `PATHChange` function is designed to perform the statistical analysis. Firstly, `PATHChange` computes the relative activity [12] of each pathway (Eq. 1) to determine an increase or decrease in expression of the pathway in respect to the control samples:

$$n = \frac{N_e}{N_e + N_y}, \tag{1}$$

where, n is the relative activity of the pathway, N_e is the sum of the logarithm of the expression signal of each gene ($Log[signal+1]$) in the experimental condition data and N_y is the sum of the logarithm of the expression signal of each gene in the control data. Eq. 1 has monotonic behavior, for any combination of high or low values of N_e and N_y it remains between 0 and 1. The significance of the alterations can be assessed by different statistical methods, in `PATHChange` we use three different ones to improve statistical results: the Bootstrap method, Fisher exact test and Wilcoxon signed rank test. For all three statistical tests the null hypothesis is the same 'not differentially expressed pathway' with the alternative hypothesis 'differentially expressed pathway'. Eq. 1 is employed as it is in the bootstrap test, for Fisher's test the calculations consider single genes and the N's in Eq. 1 corresponds to the logarithm of the gene's signal. It is not used in Wilcoxon's test which uses the difference between the logarithm of the signal of each gene in control and altered samples.

It is important to note that this implies multiple comparisons, thus the function `PATHChange` corrects the false discovery rate (FDR) through the Benjamini-Hockberg algorithm.

The final results can be visualized with the `PATHChangeVenn` function that shows the results through Venn diagrams of the consensus of the differentially expressed pathways (at the significance level chosen). The user has the flexibility to choose which statistical tests he/she wants to combine.

In the next section, we present an example to demonstrate how easy is to perform a pathway data analysis with `PATHChange`.

2.2. Data input preparation functions

Here we use as sample test dataset GSE26910[13] from GEO to show how to define the inputs in the `PATHChange` package, it contains data from four different histologies: 6 breast primary tumor samples and 6 matched samples of normal stroma, 6 prostate primary tumor samples and 6 matched samples of normal stroma. The pathways used in the analysis correspond to genome maintenance mechanisms that are important in pathologies associated with cancer, they were extracted from the Ontocancro project (http://ontocancro.inf.ufsm.br) [14].

After the appropriate normalization procedure chosen by the user, the expression data matrix results is exhibited in a table similar to Table 1. Each row in the matrix corresponds to a particular probe and each column, called sample, corresponds to an experimental condition. Observe that additional annotation information (gene symbols of the probes) are included in the matrix.

Table 1: Schematic example of a expression matrix with rows representing probes, the genes (official symbols), and columns representing particular conditions. Each cell exhibit an fictitious value, corresponding to the expression level of a gene in that condition.

Probe	Symbol	Sample 1	Sample 2	Sample 3	Sample 4
Probe A	Gene A	5614	6446	5756	5498
Probe B	Gene B	592	401	459	619
Probe C	Gene C	246	238	261	207
Probe D	Gene D	1233	813	647	663

2.2.1. The `PATHChangeDat` function

Once the expression matrix has been created, it has to be saved in the tab delimited (*.csv) format to be used as the input argument `eDat` for the `PATHChangeDat` function.

To perform the calculations, `PATHChange` package selects only the genes that belong to the database where the pathways were obtained as the statistical

population of study. The `Genes` argument consists in a text formatted file organized in a single column called '*Symbol*', containing all genes members of the population (e.g., see the file `http://ontocancro.inf.ufsm.br/download/genes_ontocancro.zip` with the genes of Ontocancro database). Choosing `HistComp=TRUE` as input command during execution, `PATHChange` will exhibit the different sample types available in the dataset. However, if the user is interested in a specific comparison, then the argument `HistComp=FALSE` (a character string 'hc' is required as input) specifies which histological comparison he/she wishes to perform e.g., hc = c("breast normal", "breast tumor"). The results of `PATHChangeDat` (see Table 2) is an list of tables (named 'MeanData') saved in 'rds' format with the histological comparisons selected.

Table 2: The first six rows of the file MeanData generated by `PATHChangeDat` corresponding to a comparative study involving 'normal' and 'breast'.

Symbol	Control	Experiment
ABL1	9.748	9.058
ACIN1	6.125	6.222
ACOT13	6.109	7.095
ACR	3.413	3.476
ACTA1	3.397	3.438
ACTG1	13.054	13.140

Each file produced by `PATHChange` package are saved in the temporary folder, optionally you can select another destination folder to save these files.

2.2.2. The `PATHChangeList` function

In databases such as KEGG [15], PathwayCommons [16], Reactome [17] or Ontocancro [14], we can find a large number of metabolic and signalling pathways which we will be interested in checking alterations. For example, imagine that you are investigating a cancer microarray study. In this case multiple

pathways can be involved like Cell Cycle and DNA Damage Response. For that purpose the first argument of `PATHChangeList` is a text formatted file with two columns: '*Pathway*' and '*ApprovedSymbol*', so the user can inform to the package which pathways will be considered and which genes belong to each them. `PATHChangeList` produces a file of the class 'list' saved in 'rds' format. In this particular example, we use 49 pathways from the Ontocancro database (`http://ontocancro.inf.ufsm.br/download/BigPathways.zip`). The R command below is included in `PATHChange` as an example in its manual which is used here to exemplify the use of the function `PATHChangeList`. At this point, the `PATHChange` function can be runned. path = list.load(system.file("extdata", "path.rds", package = "PATHChange"))

2.2.3. The `PATHChange` function

As we have seen, the functions `PATHChangeDat` and `PATHChangeList` provide the formatted files to be used as arguments for the `PATHChange` function. To read them we use the R package 'rlist' required by the `PATHChange` function. With the `PATHChange` function we have the additional option of saving the results in '.csv' format. Table 3 shows a typical result where for each pathway it is presented its activity and the p-value of each statistical test. The results show activation of the following pathways: G1/S DNA Damage Checkpoints, G2/M Checkpoint, Regulation of Mitotic Cell Cycle, S Phase and Caspase Cascade in Apoptosis. The first four pathways correspond to the upregulation of cell cycle mechanisms which are induced by genetic instability, one of the hallmarks of cancer, see Table 4. DNA damage activates cell cycle checkpoints which are aberrant in cancer and well documented, see for example the review by Bartek and Lucas [18, 19]. However, the upregulation of the latter pathway, Caspase cascade in Apoptosis, is a feature specific of breast cancer since it is well documented that cancers cells are considered immortal because the caspase apoptotic pathways are supressed, with the exception of breast, where an upregulated caspase activity in relation to normal tissue is observed, see references [20, 21, 22]. So, the method is able

to detect that specific feature of breast cancer tissues.

The `PATHChange` function can be used with RNA-seq data provided that its input files are structured as described in Table 2. We are currently developing a new function of PATHChange that will make easier to process this type of data and it will be available in the next release of PATHChange.

Table 3: Table of results obtained from the `PATHChange` function for a breast cancer study, GSE26910 dataset including the name of the pathway, the relative activity and the tests. The consensus of the three tests for the significantly altered pathways are highlighted in bold fonts.

Pathway	Activity	Bootstrap	Fisher	Wilcoxon
Canonical NF-kappaB	0.5018	0.0312	1	0.3092
IL10 Anti-inflammatory Signaling Pathway	0.5003	0.1074	1	0.58
IL6-mediated signaling events	0.4991	0.6244	0.7628	1
IL8- and CXCR1-mediated signaling events	0.5008	0.0164	0.5389	0.2703
IL8- and CXCR2-mediated signaling events	0.5007	0.0042	0.2498	0.1024
mtor signaling	0.4996	0.6011	0.0331	0.267
Nemo	0.5002	0.0895	0.7628	0.3117
NF-kappaB pathway	0.5003	0.0092	0.7628	0.5685
nf-kb signaling	0.5003	0.05	1	0.6141
p53	0.4999	0.2516	0.4512	0.6312
Regulation of p38-alpha and p38-beta	0.5000	0.0973	0.7628	0.8355
Replicative Senescence	0.5001	0.1684	0.7628	0.9695
RIG-I	0.4998	0.5611	0.5389	0.6312
SASP	0.5001	0.061	0.7628	1
TGF	0.5000	0.0632	0.7628	0.6312
TNF	0.5000	0.0756	1	0.8094
ATM signaling	0.5000	0.2516	1	1
ATR signaling	0.5001	0.0087	1	0.267
Base Excision Repair	0.5001	0.0459	0.7651	0.491
Double-Strand Break Repair	0.5001	0.0737	1	0.6715
Fanconi Anemia Pathway	0.5002	0.0049	0.7164	0.0677
Homologous Recombination	0.5001	0.0309	1	0.4882
Hr Repair of Replication-Independent DSB	0.5001	0.0973	1	0.8613
Mismatch Repair	0.5002	0	1	0.0336
Non-Homologous end Joining	0.5001	0.0223	1	0.5002
Nucleotide Excision Repair	0.5001	0.0672	1	0.58
Processing of DNA DSB ends Recruitment of Repair and Sig. Proteins	0.5001	0.0843	1	0.58
Cell Cycle Checkpoints	0.5002	0	0.0514	1.02×10^{-5}
Cell Cycle, Mitotic	0.5001	0	0.2437	1.46×10^{-8}
Cyclins and Cell Cycle Regulation	0.5005	0.0004	0.7628	0.0677
G1/S DNA Damage Checkpoints	**0.5002**	**0**	**0.0331**	$\mathbf{5.46}\times10^{-5}$
G2/M checkpoint	**0.5002**	**0**	**0.0195**	$\mathbf{1.08}\times10^{-5}$
Mitotic M-M/G1 Phases	0.5000	0.0164	0.6274	0.58
Mitotic Spindle Checkpoint	0.5002	0	0.2498	1.46×10^{-7}
Rb Tumor suppressor/Check. P. Sign. in Response to Damage	0.5002	0.0028	0.6596	0.1056
Regulation of Mitotic Cell Cycle	**0.5002**	**0**	**0.0112**	$\mathbf{2.62}\times10^{-6}$
S Phase	**0.5001**	**0**	**0.0195**	$\mathbf{7.26}\times10^{-7}$
Apoptosis - Homo sapiens (human)	0.5001	0.0028	0.2533	0.0162
Apoptotic signaling in response to dna damage	0.5000	0.0049	0.7628	0.58
Caspase Cascade in Apoptosis	**0.5001**	**0**	**0.0195**	$\mathbf{2.37}\times10^{-5}$
Death Receptor Signalling	0.5000	0.4282	0.6617	0.6141
Extrinsic Pathway for Apoptosis	0.5001	0.0117	0.7628	0.4943
Granzyme a Mediated Apoptosis Pathway	0.5001	0.0517	0.7628	0.2703
Induction of apoptosis through dr3 and dr4/5 death receptors	0.5001	0.0033	1	0.5002
Intrinsic Pathway for Apoptosis	0.5000	0.0672	0.6596	0.8094
Regulation of Apoptosis	0.5000	0.1162	0.7628	0.58
TNF Receptor Signaling Pathway	0.5001	0.0226	0.3488	0.0791
tnfr1 Signaling Pathway	0.5001	0.0042	0.8697	0.3525
tnfr2 Signaling Pathway	0.5000	0.1162	0.7628	0.58

Table 4: Short description of biological pathways found altered in our analysis.

Pathway	Description
Cell Cycle Pathways	
S Phase	S phase of the cell cycle.
Regulation of Mitotic Cell Cycle	Regulatory pathways of the cell cycle.
Apoptosis Pathways	
Caspase Cascade in Apoptosis	Caspase cascade of activation of apoptosis.
Cell Cycle Checkpoints	
G1/S DNA Damage Checkpoints	Activation of the G1/S checkpoint due to DNA damage.
G2/M checkpoint	Activation of the G2/M checkpoint due to DNA damage.

2.2.4. The `PATHChangeVenn` function

Finally, the altered consensus pathways resulting from the three tests can be visualized using the `PATHChangeVenn` function (see Figure 2). The arguments of the function, `p.value` and `p`, are, respectively, the 'rds' file produced by the `PATHChange` function and the significance level used.

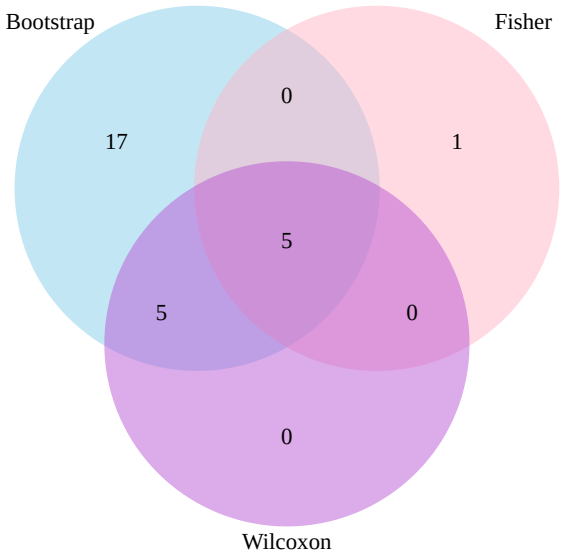


Figure 2: Venn Diagram showing the consensus of the three statistical tests for a breast cancer study, dataset GSE26910.

3. Discussion

In Table 5 we present a comparison of PATHChange with other similar packages. We considered several features of the tools, including language of development, repository for download, statistical tests employed and transcriptomic data applicable. Differently from the other packages, PATHChange employs three statistical tests and allow the user to build a custom pathway.

In summary, the `PATHChange` package offers an alternative type of detection of differentially expressed pathways in transcriptomic data. Application of the tool to an investigation breast cancer phenotype presents results consistent with the literature. Due to its high level of stringency, the method is more efficient when used with studies that present several altered pathways like cancer, other-

wise few alterations can be found. However, we expect to improve its efficiency in its next releases. The examples in this paper were generated with a PC running the Windows 8 operating system with 64 bit architecture but they can also be reproduced in PCs running Ubuntu 14.04 without significant performance differences.

Table 5: Comparison of tools for transcriptomic pathway analysis [23, 24, 25, 12].

Tool	Language	Repository	Pathways	Input data	Tests	User custom pathway
GAGE	R	Bioconductor	KEGG, GO	Microarray, RNA-seq	t-test, Mann Whitney U test (if gene set sizes smaller than 10)	no
sigPathway	R	Bioconductor	GO, BioCyc, BioCarta, KEGG	Microarray	t-test	no
ToPASeq	R	Bioconductor	topology-based pathway analysis method	Microarray, RNA-seq	t-test or z-test	no
ViaComplex	Fortran	webpage	any	Microarray	bootstrap	yes
PATHChange	R	CRAN	any	Microarray, RNA-seq	bootstrap, Fisher and Wilcoxon	yes

Conflicts of interest

None declared.

List of abbreviations

CRAN, Comprehensive R Archive Network; GO, gene ontology; KEGG, Kyoto encyclopedia of genes and genomes.

Acknowledgments

Work supported by Capes and CNPq (249374/2013-7, 402547/2012-8).

References

[1] D. J. Haustead, A. Stevenson, V. Saxena, F. Marriage, M. Firth, R. Silla, L. Martin, K. F. Adcroft, S. Rea, P. J. Day, et al., Transcriptome analysis of human ageing in male skin shows mid-life period of variability and central role of nf-κb, Scientific reports 6.

[2] Y.-J. Hu, A. N. Imbalzano, Global gene expression profiling of jmjd6-and jmjd4-depleted mouse nih3t3 fibroblasts, Scientific data 3.

[3] E. Korpelainen, J. Tuimala, P. Somervuo, M. Huss, G. Wong, RNA-seq Data Analysis: A Practical Approach, CRC Press, 2014.

[4] S. Drăghici, Statistics and data analysis for microarrays using R and bioconductor, CRC Press, 2011.

[5] R Core Team, R: a language and environment for statistical computing, R Foundation for Statistical Computing, Vienna, Austria (2014).
URL `http://www.R-project.org/`

[6] C. Fontoura, E. Giampieri, D. Remondini, G. Librelotto, G. Castellani, J. Mombach, PATHChange: an R tool for identification of differentially expressed pathways using multi-statistic comparison, manuscript submitted for publication (2016).

[7] B. Efron, Bootstrap methods: another look at the jacknife, The Annals of Statistics 7 (1979) 1–26.

[8] R. A. Fisher, Statistical methods for research workers, in: F. Crew, D. W. Cutler (Eds.), Biological monographs and manuals, no. 5, Oliver And Boyd Tweeddale Court ; Edinburgh ; Paternoster Row ; London, 1934.

[9] F. Wilcoxon, Individual comparisons by ranking methods, Biometrics bulletin 1 (6) (1945) 80–83.

[10] T. Barrett, S. E. Wilhite, P. Ledoux, C. Evangelista, I. F. Kim, M. Tomashevsky, K. A. Marshall, K. H. Phillippy, P. M. Sherman, M. Holko, et al.,

Ncbi geo: archive for functional genomics data setsupdate, Nucleic acids research 41 (D1) (2013) D991–D995.

[11] Y. Benjamini, Y. Hochberg, Controlling the false discovery rate: a practical and powerful approach to multiple testing, Journal of the Royal Statistical Society, Series B (Methodological) 57(1) (1995) 289–300.

[12] M. A. Castro, J. C. Mombach, R. M. de Almeida, J. C. Moreira, Impaired expression of ner gene network in sporadic solid tumors, Nucleic Acids Research 35 (6) (2007) 1859–1867.

[13] A. Planche, M. Bacac, P. Provero, C. Fusco, M. Delorenzi, J.-C. Stehle, I. Stamenkovic, Identification of prognostic molecular features in the reactive stroma of human breast and prostate cancer, PloS one 6 (5) (2011) e18640.

[14] G. R. e. a. Librelotto, An ontology to integrate transcriptomics and interatomics data involved in gene pathways of genome stability, in: K. S. Guimares, A. Panchenko, T. M. Przytycka (Eds.), 4th Brazilian Symposium on Bioinformatics, BSB 2009, Porto Alegre, Brazil, July 29-31, 2009. Proceedings, Springer Berlin Heidelberg, 2009, pp. 164–167.

[15] M. Kanehisa, S. Goto, Kegg: kyoto encyclopedia of genes and genomes, Nucleic Acids Research 28 (1) (2000) 27–30.

[16] E. G. Cerami, B. E. Gross, E. Demir, I. Rodchenkov, Ö. Babur, N. Anwar, N. Schultz, G. D. Bader, C. Sander, Pathway commons, a web resource for biological pathway data, Nucleic Acids Research 39 (suppl 1) (2011) D685–D690.

[17] L. Matthews, G. Gopinath, M. Gillespie, M. Caudy, D. Croft, B. de Bono, P. Garapati, J. Hemish, H. Hermjakob, B. Jassal, et al., Reactome knowledgebase of human biological pathways and processes, Nucleic Acids Research 37 (suppl 1) (2009) D619–D622.

[18] J. Bartek, J. Lukas, Mammalian g1-and s-phase checkpoints in response to dna damage, Current opinion in cell biology 13 (6) (2001) 738–747.

[19] T. D. Halazonetis, V. G. Gorgoulis, J. Bartek, An oncogene-induced DNA damage model for cancer development, Science 319 (2008) 1352–1355.

[20] M. Vakkala, P. Pääkkö, Y. Soini, Expression of caspases 3, 6 and 8 is increased in parallel with apoptosis and histological aggressiveness of the breast lesion, British journal of cancer 81 (4) (1999) 592.

[21] L. Nakopoulou, P. Alexandrou, K. Stefanaki, E. Panayotopoulou, A. C. Lazaris, P. S. Davaris, Immunohistochemical expression of caspase-3 as an adverse indicator of the clinical outcome in human breast cancer, Pathobiology 69 (5) (2002) 266–273.

[22] N. ODonovan, J. Crown, H. Stunell, A. D. Hill, E. McDermott, N. OHiggins, M. J. Duffy, Caspase 3 in breast cancer, Clinical Cancer Research 9 (2) (2003) 738–742.

[23] W. Luo, M. S. Friedman, K. Shedden, K. D. Hankenson, P. J. Woolf, Gage: generally applicable gene set enrichment for pathway analysis, BMC bioinformatics 10 (1) (2009) 1.

[24] W. Lai, L. Tian, P. Park, sigPathway: Pathway Analysis, http://www.pnas.org/cgi/doi/10.1073/pnas.0506577102, http://www.chip.org/ ppark/Supplements/PNAS05.html (2008).

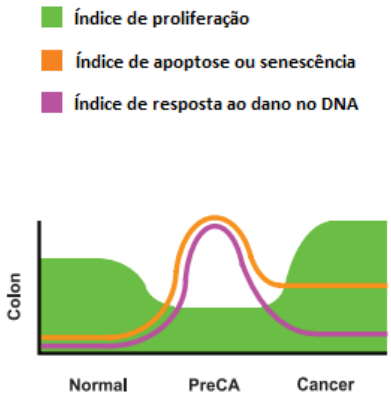
[25] I. Ihnatova, E. Budinska, Topaseq: an r package for topology-based pathway analysis of microarray and rna-seq data, BMC bioinformatics 16 (1) (2015) 1.

4 DISCUSSÃO

Visando melhorar a eficiência na análise de dados de transcriptoma, onde a maioria dos métodos tradicionais utiliza apenas um teste estatístico (RITCHIE et al., 2015; THERNEAU; GRAMBSCH, 2000), nós desenvolvemos um método multiestatístico para a detecção de vias genéticas diferencialmente expressas. Complementarmente, nós desenvolvemos PATHChange, um pacote do R voltado para pesquisadores que utilizam em seus estudos dados de transcriptoma. O pacote foi criado para implementar o método estatístico desenvolvido, que parte do cálculo da atividade das vias genéticas envolvidas em certas doenças e determina a significância estatística das alterações encontradas. A principal contribuição do pacote é fornecer um tipo alternativo de detecção de vias diferencialmente expressas em dados de expressão de microarranjos de DNA.

Halazonetis, Gorgoulis e Bartek (2008) desenvolveram um modelo baseado em estudos experimentais da evolução de lesões pré-cancerosas para o câncer (Figura 4.1), constatando que as lesões de pré-câncer apresentam uma barreira contra o desenvolvimento do câncer, caracterizado por um aumento das vias de apoptose (morte celular) ou senescência (conjunto de fenômenos associados ao envelhecimento celular) e reparo de DNA (identificação e correção de danos no DNA). Halazonetis, Gorgoulis e Bartek (2008) adicionalmente perceberam que algumas mutações (tais como p53) podem perturbar essas vias, reduzindo essa barreira e induzindo a transição para o câncer.

Figura 4.1 – Ilustração de um dano no DNA induzindo uma barreira contra o desenvolvimento do câncer em lesões de pré-câncer de cólon, através do aumento nos índices de apoptose ou senescência e resposta ao dano no DNA, e diminuição nos índices de proliferação celular. Quando essa barreira é rompida, as lesões evoluem para câncer resultando em um aumento nos níveis de proliferação e diminuição das atividades de apoptose ou senescência e resposta ao dano no DNA.



Fonte: Adaptado de (HALAZONETIS; GORGOULIS; BARTEK, 2008)

Com o intuito de testar o método, foram selecionados 4 *DataSets* do GEO (SABATES-BELLVER, 2007; GALAMB et al., 2010; SKRZYPCZAK et al., 2010; VALCZ, 2014) contendo 4 estudos de adenoma (tipo de pré-câncer que pode facilmente progredir para o câncer) e 3 estudos de carcinoma (câncer) de cólon. e avaliadas as alterações de 49 vias da Ontocancro. As informações detalhadas a respeito dos *DataSets* podem ser encontrados no artigo do Capítulo 2, bem como um resumo das funções biológicas de cada via detectada pelo método.

O primeiro resultado importante são os diagramas de Venn da Figura 1 do artigo do Capítulo 2. Considerando os resultados do teste de bootstrap, que é um teste menos restritivo, é possível identificar um grande número de vias significativas em ambos os estudos. Por exemplo, para o caso de adenoma (Figura 1a do artigo), de acordo com o teste de bootstrap 39 vias poderiam ser consideradas alteradas significativamente, enquanto que os testes de Wilcoxon e Fisher identificaram 21 vias significativas. O nosso método propõe que a escolha das vias alteradas nesse caso seja feita considerando aquelas vias que aparecem significativamente alteradas nos três testes estatísticos. Nesse caso, considerando apenas a interseção entre os três métodos, o número de vias alteradas significativamente seria reduzido para 12, indicando que mais da metade das vias alteradas significativamente para o teste de bootstrap seriam falsos positivos na resposta. Ainda, considerando o diagrama de Venn do caso de câncer de cólon (Figura 1b do artigo), 10 vias apresentariam diferenças significativas na expressão para os três testes estatísticos simultaneamente, de modo que 37 vias apresentadas pelo teste de bootstrap seriam falsos positivos.

A análise da Figura 1 do artigo do Capítulo 2 nos instigou a verificar quais são as vias que apresentam alterações significativas nessas intersecções, a fim de analisar se elas concordam com o modelo previsto por Halazonetis, Gorgoulis e Bartek (2008). Esses resultados podem ser encontrados na Tabela 2 do artigo do Capítulo 2. Os resultados exibem alteração em 4 vias de reparo e uma via de apoptose nos estudos de pré-câncer de cólon, indicando que o método proposto no artigo do Capítulo 2 foi capaz de detectar a presença da barreira contra o câncer. Os resultados apresentados na Tabela 2 do artigo do Capítulo 2 evidenciam, ainda, que essa barreira perde atividade quando as células evoluem para câncer, apresentando uma brusca diminuição dessas vias nesses estudos. As demais vias alteradas nos resultados são as vias do ciclo celular, que ocorrem devido o caráter proliferativo dessas células.

Ainda de acordo com o modelo de Halazonetis, Gorgoulis e Bartek (2008), a atividade das vias do ciclo celular aumentam nas lesões de câncer. Para verificar isso, foi gerado o gráfico da Figura 2 do artigo do Capítulo 2 que apresenta as atividades globais das vias alteradas significativamente nos 3 *DataSets* que possuem estudos comparativos de câncer e pré-câncer. Considerando que as vias do ciclo celular predominam nos resultados, esse diagrama exibe um aumento global das vias de câncer em relação as vias de pré-câncer em todos os *DataSets* avaliados.

O método foi testado, adicionalmente, para outros estudos de câncer (tais como câncer de mama e próstata (PLANCHE et al., 2011), e ovário (SPILLMAN et al., 2010)), a fim de veri-

ficar se as vias diferencialmente expressas resultantes de PATHChange não seriam resultantes de um vício do programa. Os resultados da Tabela 3 do artigo do Capítulo 2 descartam essa opção. Muito embora a grande maioria das vias alteradas significativamente nos estudos sejam relacionadas ao Ciclo Celular (semelhantemente ao que ocorre nos estudos de câncer de cólon), as vias alteradas em cada um dos estudos são diferentes entre si e também diferentes das que ocorrem nos estudos de câncer de cólon, indicando que a hipótese de viciosas repetições nas respostas do programa não é verdadeira. O Artigo 2 (Capítulo 3) apresenta uma discussão à respeito do aparecimento da via de *Caspase cascade in Apoptosis* alterada nos resultados de câncer de mama e próstata, conforme pode ser visualizado na Tabela 3 do Artigo 1. Normalmente, apenas uma pequena parcela dos cânceres humanos apresentam a expressão das caspases. Estudos apontam para a ativação dessa via nos tipos de câncer de próstata e mama (VAKKALA; PÄÄKKÖ; SOINI, 1999; NAKOPOULOU et al., 2002; O'DONOVAN et al., 2003). A alteração dessa via específica demonstra que o método é capaz de detectar características específicas dos diferentes tecidos cancerosos.

Foram observadas mais vias alteradas em lesões de pré-câncer do que em câncer. Enquanto as alterações das vias de ciclo celular aparecem devido o caráter proliferativo das células cancerígenas, as demais vias aparecem alteradas devido os mecanismos de proteção do DNA. Quando comparadas as atividades de pré-câncer e câncer, as vias testadas apresentam um aumento global em todos os casos. Diferentes tipos de câncer apresentam vias genéticas diferentes, ainda mantendo a concordância com o modelo de Halazonetis, Gorgoulis e Bartek (2008). Em suma, os nossos resultados apresentam clara concordância com as previsões da literatura para os fenótipos de pré-câncer e câncer(HALAZONETIS; GORGOULIS; BARTEK, 2008).

Aparentemente, devido ao seu alto nível de restrição resultante do uso de três testes estatísticos ao invés de apenas um, o método apresenta maior eficiência quando utiliza estudos envolvendo um grande número de vias alteradas, como é o caso do câncer (BARILLOT et al., 2012). No entanto, PATHChange está em sua primeira versão e acredita-se que isto poderá ser aprimorado para as próximas versões, que deverá inserir um pacote destinado ao pré-processamento dos dados de RNA-seq (KORPELAINEN et al., 2014).

Para provar que nenhum dos testes isoladamente poderia ser suficiente para produzir o resultado, foram comparadas as interseções entre os testes das vias diferencialmente expressas em todos os diferentes *DataSets*. Aparentemente, o teste exato de Fisher define o resultado em grande parte das análises. No entanto, foram identificadas intersecções entre os testes de bootstrap e Fisher, por exemplo, onde o teste de Wilcoxon seria o determinante para a escolha da via, de modo que nem sempre o teste de Fisher caracteriza o resultado.

O artigo do Capítulo 3 apresenta uma tabela comparativa entre PATHChange e outros pacotes similares (Tabela 4), incluindo informações sobre a linguagem, repositório para download do *software*, testes estatísticos empregados e aplicabilidade dos pacotes a dados de transcriptoma (tais como RNA-seq e Microarranjos). Um dos benefícios do pacote PATHChange é o fato de permitir ao usuário a análise de vias construídas ou combinadas à partir de diversos

bancos de dados. Além disso, a grande vantagem do uso do método proposto é o fato de ele permitir que o usuário escolha as vias diferencialmente expressas em dados de transcriptoma com uma maior precisão, diminuindo as chances de ocorrências de falsos positivos devido ao fato de utilizar três testes estatísticos na decisão. No entanto, a alternativa de considerar os três testes fica à cargo da necessidade do usuário, de maneira que ele possa julgar as vias como diferencialmente expressas quando elas aparecem em um, dois ou nos três testes.

Em linhas gerais, o pacote parece ser de fácil uso para os usuários do programa R. Detalhes a respeito da facilidade no uso do programa aparecem descritos no artigo do Capítulo 3. As funções que compõem PATHChange fornecem o suporte necessário para a realização da análise dos dados de transcriptoma e visualização dos resultados. O pacote foi desenvolvido para ser funcional. Pensando nisso, foram incrementados detalhes nas funções, tais como a leitura do arquivo suplementar do GEO *Matrix Series Files*, que possibilita o uso das informações referentes ao *DataSet* utilizado. Através dessas informações, o usuário poderá realizar várias combinações das condições experimentais do estudo em uma mesma análise, conforme o seu interesse. Outra vantagem do pacote é a possibilidade de inserir um grande número de vias para análise em um único arquivo.

Sem dúvida, analisar diversas vias para diferentes tipos de estudos ao mesmo tempo é uma vantagem apreciável do programa. No entanto, um custo é pago em eficiência computacional. Para uma análise simples, PATHChange costuma utilizar apenas poucos minutos para processar os dados. Por exemplo, a análise estatística de dados de microarranjos de 5 vias onde se compara apenas uma condição experimental decorre cerca de 2 à 3 minutos. Se este número subir para 50 vias e duas comparações experimentais, o tempo de cálculo sobe para cerca de 15 minutos. De fato, o tempo de cálculo varia com o número de vias e de comparações das condições experimentais.

Como a análise de PATHChange parte dos dados normalizados do GEO, e com as sondas já identificadas pelos seus respectivos genes, o pacote pode ser utilizado para os diferentes chips e *platforms* desde que os arquivos *input* estejam nos formatos requisitados para a análise (conforme descrito no artigo do Capítulo 3). Para garantir que esta afirmação seja verdadeira, PATHChange foi testado para variados chips de microarranjos, incluindo *Affymetrix* (LOCKHART et al., 1996), e o resultado não parece sofrer interferências da tecnologia do chip utilizada. PATHChange funciona perfeitamente nos sistemas operacionais *Windows 8* e *Ubuntu 14.04*. Não foram realizados testes para sistemas operacionais diferentes destes.

Em síntese, o equilíbrio entre a eficiência do método, que apresentou resultados que mostram claramente as alterações fenotípicas entre pré-câncer e câncer em concordância com as observadas no modelo biológico de desenvolvimento do câncer, e a funcionalidade do pacote PATHChange, fazem desta ferramenta uma boa alternativa para os estudos envolvendo investigações fenotípicas em dados de transcriptoma, podendo ser um aliado no entendimento dos processos genéticos envolvidos na progressão do câncer.

No próximo capítulo, apresentaremos as conclusões e perspectivas futuras do trabalho.

5 CONCLUSÃO

Esta tese teve por objetivo apresentar e testar um método multiestatístico desenvolvido por nós para detectar vias diferencialmente expressas em estudos de microarranjos. O método parte de dados de transcriptoma para determinar a atividade das vias genéticas envolvidas em certas doenças e verificar a significância estatística das alterações encontradas através do uso dos testes de bootstrap, exato de Fisher e Wilcoxon.

Tendo em vista os aspectos apresentados, o método se mostrou capaz de detectar diferenças fenotípicas relevantes em estudos de câncer, oferecendo um tipo alternativo de análise de vias para pesquisadores que investigam fenótipos de doenças. O método proposto foi aplicado a alguns dados de transcriptoma de pré-câncer e câncer de cólon, a fim de verificar a sua eficiência como instrumento de detecção de vias diferencialmente expressas, capaz de obter diferenças fenotípicas relevantes entre esses dois tipos de tumor. A aplicação do método, portanto, corrobora o modelo de Halazonetis. Além disso, o método se mostrou capaz de detectar características específicas dos diferentes tipos de câncer. Utilizando o método nós percebemos que, devido ao alto nível de restrição, os melhores resultados são obtidos quando utiliza estudos envolvendo um grande número de vias alteradas (como é o caso do câncer), o que ocorre devido ao grande número de vias genéticas alteradas nesta condição.

Para tornar o método mais proveitoso, nós desenvolvemos o pacote PATHChange capaz de realizar a análise estatística em experimentos envolvendo a magnitude dos microarranjos de DNA. O pacote configura funcionalidade, permitindo ao usuário economizar etapas de uma investigação através da possibilidade de realizar várias comparações de condições experimentais para uma variedade de vias em uma única análise, além de ser disponibilizado para download totalmente gratuito no CRAN. O pacote pode ser utilizado para os diferentes *chips* e *platforms*, e funciona perfeitamente nos sistemas operacionais *Windows 8* e *Ubuntu 14.04*.

Em suma, o método proposto demonstrou eficiência na detecção de vias alteradas em dados de pré-câncer e câncer, indicando que PATHChange é uma boa alternativa para os estudos envolvendo investigações fenotípicas em dados de transcriptoma. Nós acreditamos que este trabalho pode vir a contribuir para os estudos envolvendo investigações fenotípicas de doenças, podendo colaborar para um melhor entendimento dos processos genéticos que compreendem o desenvolvimento do câncer. Como sugestão de pesquisa futura, nós acreditamos que seja possível adaptar o modelo aqui proposto para dados de sequenciamento de RNA, bem como para os estudos envolvendo fenótipos de outros tipos de doenças.

REFERÊNCIAS BIBLIOGRÁFICAS

ANN, H. github; TALLOEN, W. **Gene expression studies using Affymetrix microarrays**. [S.l.]: CRC Press, 2010.

BARILLOT, E. et al. Computational systems biology of cancer. CRC Press, 2012.

BARRETT, T.; EDGAR, R. Mining microarray data at ncbi's gene expression omnibus (geo)*. In: **Gene Mapping, Discovery, and Expression**. [S.l.]: Humana Press, 2006. p. 175–190.

BARRETT, T. et al. Ncbi geo: archive for functional genomics data sets—10 years on. **Nucleic acids research**, Oxford Univ Press, v. 39, n. suppl 1, p. D1005–D1010, 2011.

BECKER, J. D.; FEIJÓ[1], J. A. Profiling genomes with oligonucleotide arrays.

BENJAMINI, Y.; HOCHBERG, Y. the false discovery rate: a practical and powerful approach to multiple testing. **Journal of the Royal Statistical Society, Series B (Methodological)**, v. 57(1), p. 289–300, 1995.

Cancer Genome Anatomy Project - informação da via fornecida por BioCarta. **RB Tumor Suppressor/Checkpoint Signaling in response to DNA damage**. 2016. [Online; accessed Jully, 2016]. Disponível em: <http://cgap.nci.nih.gov/Pathways/BioCarta/h_rbPathway>.

CASTRO, M. A. A. et al. Impaired expression of ner gene network in sporadic solid tumors. **Nucleic Acids Research**, Nucleic Acids Research, v. 35, n. 6, p. 1859–1867, 2007.

CERAMI, E. G. et al. Pathway commons, a web resource for biological pathway data. **Nucleic Acids Research**, Oxford Univ Press, v. 39, n. suppl 1, p. D685–D690, 2011.

DRĂGHICI, S. **Statistics and Data Analysis for Microarrays Using R and Bioconductor**. London: Chapman & Hall/CRC Mathematical and Computational Biology Series, 2012. 1042 p.

DRĂGHICI, S. et al. Global functional profiling of gene expression. **Genomics**, v. 81, p. 98–104, 2003.

EDGAR, R.; DOMRACHEV, M.; LASH, A. E. Gene expression omnibus: Ncbi gene expression and hybridization array data repository. **Nucleic acids research**, Oxford Univ Press, v. 30, n. 1, p. 207–210, 2002.

EFRON, B. Bootstrap methods: another look at the jacknife. **The Annals of Statistics**, v. 7, p. 1–26, 1979.

EISENBERG, D. et al. Protein function in the post-genomic era. **Nature**, Nature Publishing Group, v. 405, n. 6788, p. 823–826, 2000.

FISHER, R. A. Statistical methods for research workers. In: CREW, F.; CUTLER, D. W. (Ed.). **Biological monographs and manuals**. [S.l.]: Oliver And Boyd Tweeddale Court ; Edinburgh ; Paternoster Row ; London, 1934.

Food Warden. **Introduction to the Microarray**. 2016. [Online; accessed Jully, 2016]. Disponível em: <http://2012.igem.org/Team:Groningen/Wetwork>.

GALAMB, O. et al. Reversal of gene expression changes in the colorectal normal-adenoma pathway by ns398 selective cox2 inhibitor. **British journal of cancer**, Nature Publishing Group, v. 102, n. 4, p. 765–773, 2010.

GENTLEMAN, R. C. et al. Bioconductor: open software development for computational biology and bioinformatics. **Genome biology**, BioMed Central, v. 5, n. 10, p. 1, 2004.

HALAZONETIS, T. D.; GORGOULIS, V. G.; BARTEK, J. An oncogene-induced DNA damage model for cancer development. **Science**, v. 319, p. 1352–1355, 2008.

KANEHISA, M.; GOTO, S. Kegg: kyoto encyclopedia of genes and genomes. **Nucleic Acids Research**, Oxford Univ Press, v. 28, n. 1, p. 27–30, 2000.

KORPELAINEN, E. et al. **RNA-seq Data Analysis: A Practical Approach**. [S.l.]: CRC Press, 2014.

KRIZKOVA, S. et al. Microarray analysis of metallothioneins in human diseases—a review. **Journal of pharmaceutical and biomedical analysis**, Elsevier, v. 117, p. 464–473, 2016.

LIBRELOTTO, G. R. e. a. An ontology to integrate transcriptomics and interatomics data involved in gene pathways of genome stability. In: aES, K. S. G.; PANCHENKO, A.; PRZYTYCKA, T. M. (Ed.). **4th Brazilian Symposium on Bioinformatics, BSB 2009, Porto Alegre, Brazil, July 29-31, 2009. Proceedings**. [S.l.], 2009. p. 164–167.

LOCKHART, D. J. et al. Expression monitoring by hybridization to high-density oligonucleotide arrays. **Nature biotechnology**, v. 14, n. 13, p. 1675–1680, 1996.

LUO, W.; BROUWER, C. Pathview: an r/bioconductor package for pathway-based data integration and visualization. **Bioinformatics**, Oxford Univ Press, v. 29, n. 14, p. 1830–1831, 2013.

LUO, W. et al. Gage: generally applicable gene set enrichment for pathway analysis. **BMC bioinformatics**, BioMed Central, v. 10, n. 1, p. 1, 2009.

MATTHEWS, L. et al. Reactome knowledgebase of human biological pathways and processes. **Nucleic Acids Research**, Oxford Univ Press, v. 37, n. suppl 1, p. D619–D622, 2009.

NAKOPOULOU, L. et al. Immunohistochemical expression of caspase-3 as an adverse indicator of the clinical outcome in human breast cancer. **Pathobiology**, Karger Publishers, v. 69, n. 5, p. 266–273, 2002.

O'DONOVAN, N. et al. Caspase 3 in breast cancer. **Clinical Cancer Research**, AACR, v. 9, n. 2, p. 738–742, 2003.

PEVZNER, P. A. **Computational Molecular Biology**: An algorithmic approach. [S.l.]: Massachusetts Institute of Technology Press, 2000. 325 p.

PLANCHE, A. et al. Identification of prognostic molecular features in the reactive stroma of human breast and prostate cancer. **PloS one**, Public Library of Science, v. 6, n. 5, p. e18640, 2011.

R Core Team. **R: a language and environment for statistical computing**. Vienna, Austria, 2014. Disponível em: <http://www.R-project.org/>.

RITCHIE, M. et al. limma powers differential expression analyses for rna-sequencing and microarray studies. **Nucleic Acids Research**, v. 43, n. 7, p. e47, 2015.

RITZ, M.-F. et al. Identification of inflammatory, metabolic, and cell survival pathways contributing to cerebral small vessel disease by postmortem gene expression microarray. **Current neurovascular research**, Bentham Science Publishers, v. 13, n. 1, p. 58–67, 2016.

SABATES-BELLVER, J. e. a. Transcriptome profile of human colorectal adenomas. **Molecular Cancer Research**, v. 5(12), p. 1263–1275, 2007.

SKRZYPCZAK, M. et al. Modeling oncogenic signaling in colon tumors by multidirectional analyses of microarray data directed for maximization of analytical reliability. **PloS one**, Public Library of Science, v. 5, n. 10, p. e13091, 2010.

SPILLMAN, M. A. et al. Tissue-specific pathways for estrogen regulation of ovarian cancer growth and metastasis. **Cancer research**, AACR, v. 70, n. 21, p. 8927–8936, 2010.

TEAM, R. C. **R: a language and environment for statistical computing**. Vienna, Austria, 2014. Disponível em: <http://www.R-project.org/>.

THERNEAU, T. M.; GRAMBSCH, P. M. **Modeling Survival Data: Extending the Cox Model**. New York: Springer, 2000. ISBN 0-387-98784-3.

VAKKALA, M.; PÄÄKKÖ, P.; SOINI, Y. Expression of caspases 3, 6 and 8 is increased in parallel with apoptosis and histological aggressiveness of the breast lesion. **British journal of cancer**, Nature Publishing Group, v. 81, n. 4, p. 592, 1999.

VALCZ, G. e. a. Myofibroblast-derived sfrp1 as potential inhibitor of colorectal carcinoma field effect. **PLoS One**, v. 9(11), p. e106143, 2014.

VENABLES, W. N.; RIPLEY, B. D. **Modern Applied Statistics with S**. Fourth. New York: Springer, 2002. ISBN 0-387-95457-0. Disponível em: <http://www.stats.ox.ac.uk/pub/MASS4>.

WATSON, J. D. **Biologia molecular do gene**. Porto Alegre: Artmed, 2006. 760 p.

WILCOXON, F. Individual comparisons by ranking methods. **Biometrics bulletin**, JSTOR, v. 1, n. 6, p. 80–83, 1945.

YAP, Y. W. et al. Comparative microarray analysis identifies commonalities in neuronal injury: Evidence for oxidative stress, dysfunction of calcium signalling, and inhibition of autophagy–lysosomal pathway. **Neurochemical research**, Springer, v. 41, n. 3, p. 554–567, 2016.

ZAR, J. H. **Biostatistical Analysis**. New Jersey: Prentice Hall, 1999. 929 p.

ANEXO A – MATERIAL SUPLEMENTAR DO ARTIGO 1

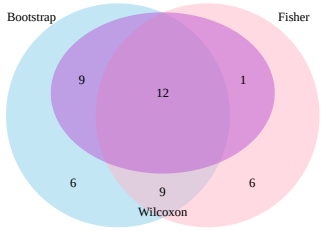
Apresentação completa dos resultados de todos os estudos utilizados no Artigo 1 (Capítulo 2).

RESEARCH

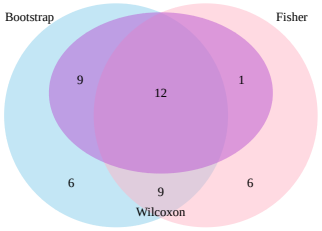
Supplementary material PATHChange: an R tool for identification of differentially expressed pathways using multi-statistic comparison

Full list of author information is available at the end of the article

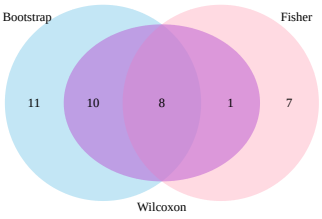
1 Venn Diagrams



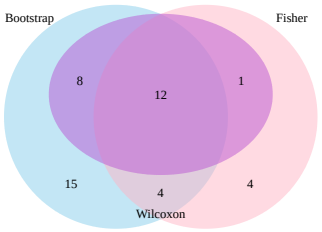
(a) GSE8671



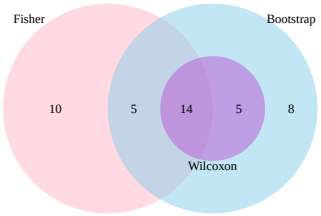
(b) GSE15960



(c) GSE20916 microdissected

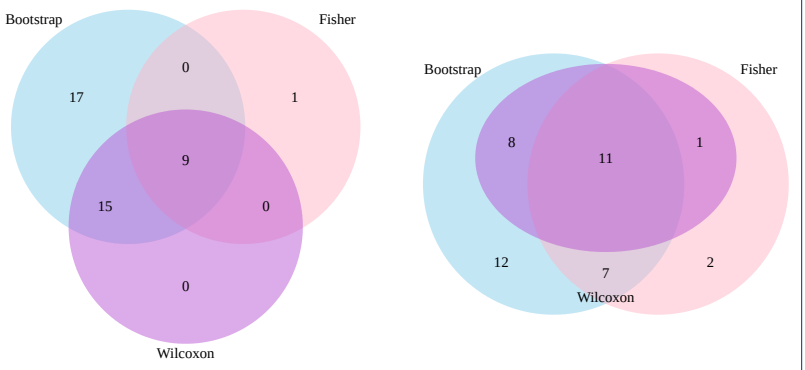


(d) GSE20916 macrodissected



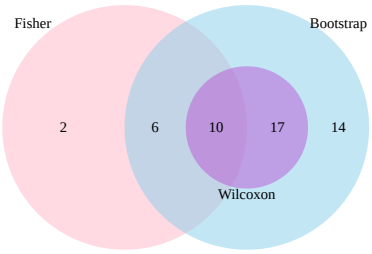
(e) GSE37364

Figure 1: Venn diagrams of colorectal adenoma series

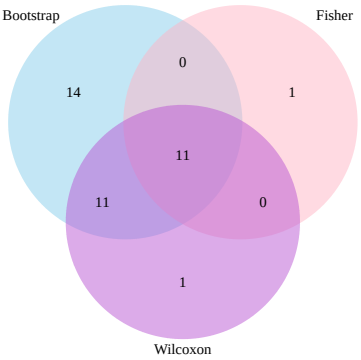


(a) GSE15960

(b) GSE20916 microdissected



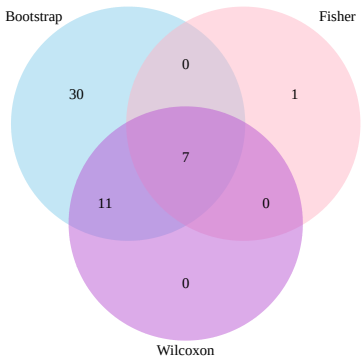
(c) GSE20916 macrodissected



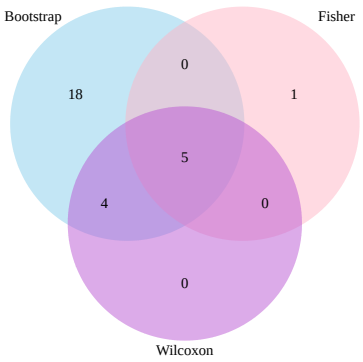
(d) GSE37364

Figure 2: Venn diagrams of colorectal carcinoma series

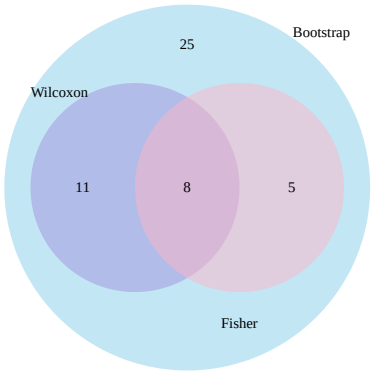
2 Tables



(a) GSE26910 Prostate



(b) GSE26910 Breast



(c) GSE22600 Ovarian

Figure 3: Venn diagrams of datasets of prostate, breast and ovarian.

Table 1: Details of data analysis of GSE8671 series using PATHChange. Consensus pathways are marked in light gray color.

Colorectal Adenoma (CRA) - GSE8671				
Pathway	Activity	Bootstrap	Fisher	Wilcoxon
Canonical NF-kappaB	0.5358461724538660 1	1.0888888888888901E-3	1.46404597835215E-4	0.18792282044042699
IL10 Anti-inflammatory Signaling Pathway	0.5617262370510970 3	0	0.60325439143623105	9.2632085084915095E-3
IL6-mediated signaling events	0.5620134427657990 1	0	0.70132792479988104	0.17859702640109601
IL8- and CXCR1-mediated signaling events	0.4873087311738840 1	6.1378947368421102E-2	4.67081794718524E-4	0.158679550656906
IL8- and CXCR2-mediated signaling events	0.4971750879801110 1	1.0952941176470599E-2	5.9541924934691803E-4	0.54903371710526305
mtor signaling	0.5281108171729850 3	0	1.6660318956685599E-2	0.967085702017357
Nemo	0.5140199360414370 3	1.0952941176470599E-2	3.7342291120331203E-2	0.92296504974365201
NF-kappaB pathway	0.5607981721208099 6	0	0.63118314839489398	4.4677236643763703E-4
nf-kb signaling	0.4966999387289269 8	5.8932432432432402E-2	0.35010901960109703	0.79899708817644799
p53	0.4912001092538170 2	3.5700000000000003E-2	2.3811631461693499E-2	0.86848397674255695
Regulation of p38-alpha and p38-beta	0.486458429703436	4.2194444444444403E-2	2.72884325190959E-4	0.18792282044042699
Replicative Senescence	0.4684769476432160 1	0.40603181818181799	5.1695029637032902E-2	0.54903371710526305
RIG-I	0.5523266441934939 8	9.4230769230769205E-4	0.69169044539114499	0.54903371710526305
SASP	0.4532965360110279 8	0.70434893617021299	6.9412901627576994E-5	0.10947230696678199
TGF	0.5137008392651679 6	0	3.5852990961307402E-3	0.60729735916582495
TNF	0.480001016435537	9.4570000000000001E-2	1.6405458403655499E-2	0.79953344317618802
ATM signaling	0.5536020556481330 3	1.1827586206896599E-3	1	2.4281819661458301E-2
ATR signaling	0.554985872120206	0	2.37711737854546E-2	7.29362826942552E-7
Base Excision Repair	0.5162956570594330 3	5.8187500000000001E-3	0.23773423264081101	8.1058175940262704E-3
Double-Strand Break Repair	0.6046433734439190 5	0	0.23773423264081101	4.4677236643763703E-4
Fanconi Anemia Pathway	0.5960517137815509 7	0	6.1760173214593497E-2	2.4374690838158198E-5
Homologous Recombination	0.6141323311931740 1	0	0.23773423264081101	3.5164356231689602E-4
Hr Repair of Replication-Independent DSB	0.5982429051755410 2	6.1249999999999998E-4	0.42991839661057102	6.6460503472222203E-3
1 Mismatch Repair	0.5925605299108720 5	0	4.7823714260875104E-3	2.0483261842204299E-7
Non-Homologous end Joining	0.5938066743348949 8	0	0.15050011268742 5	7.1763992309570394E-5
Nucleotide Excision Repair	0.5736253208274250 1	9.4230769230769205E-4	1	0.10532545006793501
Processing of DNA DSB ends Recruitment of Repair and Sig. Proteins	0.4539448339259959 9	0.61793260869565203	6.7617377779459595E-4	3.0450994318181799E-2
Cell Cycle Checkpoints	0.5645116578369400 2	0	2.86618675005789E-9	6.0721377707522598E-16
Cell Cycle, Mitotic	0.5651208276254570 1	0	1.60275600645775E-15	6.8096449579455799E-34
Cyclins and Cell Cycle Regulation	0.4860882720859209 8	0.13552682926829299	1	0.234397888183594
G1/S DNA Damage Checkpoints	0.5528056288499240 1	0	1.6944437168149899E-4	1.168791775571 7E-8
G2/M checkpoint	0.5561679112103059 5	0	6.8715209159306704E-5	6.9060147253943301E-10
Mitotic M-M/G1 Phases	0.5606875128540019 8	0	1.1684433366965299E-2	4.6399979699348599E-8
Mitotic Spindle Checkpoint	0.5585524013651940 2	0	4.0036243659853001E-9	2.9776693384053298E-21
Rb Tumor suppressor/Check. P. Sign. in Response to Damage	0.6073960398167860 3	0	4.7823714260875104E-3	1.5576680501302099E-5
Regulation of Mitotic Cell Cycle	0.5611323629809510 4	0	2.8094224036435E-7	7.1984591452393097E-12
S Phase	0.5662566832629130 1	0	4.0036243659853001E-9	1.7156405470305 3E-14
Apoptosis - Homo sapiens (human)	0.4839874209258990 3	7.7520512820512796E-2	0.10323138560030801	0.98349067629169495
Apoptotic signaling in response to DNA Damage	0.4609783059114639 9	0.57972444444444404	6.8769998829947104E-6	0.18792282044042699
Caspase Cascade in Apoptosis	0.5513166503094589 8	0	1.2152743356312499E-2	7.0226890158846294E-8
Death Receptor Signalling	0.476731965250967	0.21457441860465101	0.13326160894741201	0.54903371710526305
Extrinsic Pathway for Apoptosis	0.5408575934770779 6	4.7419354838709703E-3	0.35010901960109703	0.18792282044042699
Granzyme a Mediated Apoptosis Pathway	0.6112885082800749 7	4.2608695652173898E-4	1	0.10947230696678199
Induction of apoptosis through dr3 and dr4/5 death receptors	0.5144767213684650 1	0	5.0620874498783696E-3	0.98349067629169495
Intrinsic Pathway for Apoptosis	0.498665713859479	2.7766666666666699E-3	0.168996996059633	0.60729735916582495
Regulation of Apoptosis	0.4366322478960080 2	0.82020000000000004	1.1741796231878999E-2	0.48942296645220601
TNF Receptor Signaling Pathway	0.4878883249479660 2	0.14093333333333299	0.26894659779959401	0.968994140625
tnfr1 Signaling Pathway	0.5151070692915760 1	1.1827586206896599E-3	1.06807454150546E-2	0.98349067629169495
tnfr2 Signaling Pathway	0.4366322478960080 2	0.82020000000000004	1.1741796231878999E-2	0.48942296645220601

Table 2: Details of data analysis of GSE15960 series using PATHChange. Consensus pathways are marked in light gray color.

Colorectal Adenoma (CRA) - GSE15960				
Pathway	Activity	Bootstrap	Fisher	Wilcoxon
Canonical NF-kappaB	0.49036598864305098	1	0.31865192209800902	0.25869342088699399
IL10 Anti-inflammatory Signaling Pathway	0.49997651867395698	1	3.0719759276323099E-2	0.74844732880592402
IL6-mediated signaling events	0.498710964857838	1	0.13470745269871001	0.82504173983698303
IL8- and CXCR1-mediated signaling events	0.49472289350432702	1	5.8270550085766197E-2	0.77744236224227403
IL8- and CXCR2-mediated signaling events	0.49139232116477299	1	4.4639627668208601E-2	0.45265101111726902
mtor signaling	0.48972876021004902	1	0.59811886997645602	1.0772383858308201E-2
Nemo	0.48929612298006497	1	1	0.12567443847656201
NF-kappaB pathway	0.48692169655890399	1	1	2.4101041620708599E-4
nf-kb signaling	0.49820360044273598	1	7.8635826074515899E-2	0.91868146260579397
p53	0.49430914953145	1	0.29894890641157701	0.25869342088699399
Regulation of p38-alpha and p38-beta	0.49239786970857802	1	7.1289466067099794E-2	0.21711295650579701
Replicative Senescence	0.47942996803890098	1	0.59811886997645602	0.53415697674418605
RIG-I	0.49456320330488501	1	4.0231968321483201E-2	0.89615095422622904
SASP	0.49871647444732697	1	1.6059268602808E-2	0.94410717487335205
TGF	0.49275027799687099	1	0.122482432031222	0.12265366949394101
TNF	0.48806265654087699	1	1	9.2022235576923097E-3
ATM signaling	0.48768384809419901	1	0.87527992735256399	9.2911783854166693E-2
ATR signaling	0.47270133587871099	1	0.15917531706588001	3.6857884631238202E-7
Base Excision Repair	0.48231782644943	1	0.78158246371915796	5.1140189170837498E-3
Double-Strand Break Repair	0.48129753345573101	1	0.12867209638354299	1.10184518914474E-3
Fanconi Anemia Pathway	0.47876330974223102	1	1.4952779412906E-2	3.1259842216968598E-7
Homologous Recombination	0.47844942248521899	1	0.15917531706588001	4.7768486870659698E-4
Hr Repair of Replication-Independent DSB	0.48508634324473798	1	0.33500606835533098	9.2022235576923097E-3
Mismatch Repair	0.47909589318656798	1	0.108472095270243	1.56058019118106E-6
Non-Homologous end Joining	0.48430802686705599	1	0.52761429728643405	2.1963119506835898E-3
Nucleotide Excision Repair	0.485532618692953	1	1	0.11926639441287901
Processing of DNA DSB ends Recruitment of Repair and Sig. Proteins	0.49444304840529202	1	0.67051659234040395	0.44817073170731703
Cell Cycle Checkpoints	0.47307296227773099	1	1.0555329390825299E-5	1.4047824423466299E-15
Cell Cycle, Mitotic	0.47539297994839202	1	3.2712860481144602E-9	1.224318450897E-36
Cyclins and Cell Cycle Regulation	0.479260850905039	1	0.87527992735256399	0.103553771972656
G1/S DNA Damage Checkpoints	0.47541048432016098	1	6.0920930949988403E-5	2.5473916589656501E-9
G2/M checkpoint	0.47308224703811402	1	2.15318950152178E-5	2.01157434326543E-10
Mitotic M-M/G1 Phases	0.482116611731565	1	0.31584688294498497	1.7241931842605401E-8
Mitotic Spindle Checkpoint	0.47229505557428503	1	3.2712860481144602E-9	1.6073636285458101E-24
Rb Tumor suppressor/Check. P. Sign. in Response to Damage	0.477199619167883	1	0.47704568036568801	1.6448974609375E-3
Regulation of Mitotic Cell Cycle	0.47533081566507202	1	4.4897476234950504E-3	2.9669925394648498E-11
S Phase	0.472200202175307	1	1.19297298732753E-5	8.0386957824928504E-15
Apoptosis - Homo sapiens (human)	0.47580298506668001	1	0.343034592814211	4.9455960591634102E-5
Apoptotic signaling in response to DNA Damage	0.48322306145806199	1	0.31865192209800902	4.9273605189641701E-5
Caspase Cascade in Apoptosis	0.47722997154508401	1	3.10480638168152E-2	6.0574875583281295E-8
Death Receptor Signalling	0.48319426818476702	1	0.87527992735256399	1.250362396240 23E-2
Extrinsic Pathway for Apoptosis	0.48330402582739201	1	0.59811886997645602	4.2179502289870698E-2
Granzyme a Mediated Apoptosis Pathway	0.47928163035987897	1	0.87527992735256399	0.103553771972656
Induction of apoptosis through dr3 and dr4/5 death receptors	0.48461136986729703	1	0.87527992735256399	3.6559431809176702E-3
Intrinsic Pathway for Apoptosis	0.485868248825574	1	0.68094143628124504	3.0280135040720499E-4
Regulation of Apoptosis	0.49154722893737302	1	0.68094143628124504	0.193992820945946
TNF Receptor Signaling Pathway	0.46893981137378299	1	0.33500606835533098	3.2626065340909099E-3
tnfr1 Signaling Pathway	0.48422676838561401	1	0.33500606835533098	3.3788405394611498E-5
tnfr2 Signaling Pathway	0.49154722893737302	1	0.68094143628124504	0.193992820945946

Table 3: Details of data analysis of GSE20916 microdissected series using PATHChange. Consensus pathways are marked in light gray color.

Colorectal Adenoma (CRA) - GSE20916 Micro				
Pathway	Activity	Bootstrap	Fisher	Wilcoxon
Canonical NF-kappaB	0.497497365551195	0.200009090909091	0.35883601490081901	0.75950634733159506
IL10 Anti-inflammatory Signaling Pathway	0.50780848376944898	1.12291666666667E-2	0.468524827146578	0.75950634733159506
IL6-mediated signaling events	0.50643292910827797	2.5407407407407399E-2	0.57108133894481095	0.69092987804878003
IL8- and CXCR1-mediated signaling events	0.49161637054368701	0.45263750000000003	3.0731950470975801E-2	0.313228130340577
IL8- and CXCR2-mediated signaling events	0.49564309776904802	0.22583235294117601	7.8591148084964906E-3	0.28313577690949898
mtor signaling	0.489971488600001	0.519166666666667	1.37104313877146E-3	8.2405205259274902E-2
Nemo	0.50490339997529099	4.3931034482758598E-2	0.69469416352616298	0.75950634733159506
NF-kappaB pathway	0.52029328976245703	0	0.15881670079764201	4.3968720869584502E-4
nf-kb signaling	0.486891416580984	0.66729090909090905	0.249134151003297	0.65151742788461497
p53	0.50900182208379696	4.4545454545454497E-3	0.22636873500358001	0.71755021810531605
Regulation of p38-alpha and p38-beta	0.498499656179963	9.4206451612903197E-2	0.35883601490081901	0.95012537851774403
Replicative Senescence	0.48215691343753903	0.72509999999999997	0.105331516515031	0.3828125
RIG-I	0.50694376152509302	3.1850000000000003E-2	0.82955523655517605	0.61833190917968806
SASP	0.48691164515808799	0.670891666666667	8.4642785888934501E-3	3.36456298828125E-2
TGF	0.50383230459566397	2.0999999999999999E-3	8.3734828642258799E-2	0.75950634733159506
TNF	0.48720331034943398	0.670891666666667	5.5228245432612497E-2	0.43839615024626299
ATM signaling	0.51760569298954395	2.0999999999999999E-3	0.82955523655517605	0.228670654296875
ATR signaling	0.53826591832715198	0	8.3734828642258799E-2	3.5672972444445003E-5
Base Excision Repair	0.537936558634593	0	8.3734828642258799E-2	1.9276142120361301E-4
Double-Strand Break Repair	0.54203901215325101	0	2.5996133604395202E-2	6.2306722005208304E-4
Fanconi Anemia Pathway	0.53069859465399505	0	0.186903015656325	1.4873817563057E-3
Homologous Recombination	0.53732124094037803	0	8.3734828642258799E-2	9.0943850003755904E-4
Hr Repair of Replication-Independent DSB	0.53332478126536098	2.7222222222222198E-4	8.3734828642258799E-2	7.0369944852941204E-3
Mismatch Repair	0.52363258541246804	0	2.0857918398400802E-2	1.96703891693536E-5
Non-Homologous end Joining	0.53388313110279495	0	2.0857918398400802E-2	4.3968720869584502E-4
Nucleotide Excision Repair	0.52772181978535804	2.7222222222222198E-4	0.33582333833172601	0.11442764945652199
Processing of DNA DSB ends Recruitment of Repair and Sig. Proteins	0.49472466519331398	0.39790512820512802	0.519130299581681	0.69092987804878003
Cell Cycle Checkpoints	0.52126296192674804	0	1.5685298182428599E-3	3.9555013215225798E-8
Cell Cycle, Mitotic	0.52604762165251695	0	1.68893143003619E-9	3.97325343574543E-23
Cyclins and Cell Cycle Regulation	0.50407716811699299	9.2119999999999994E-2	0.62743337145826505	0.47509765625
G1/S DNA Damage Checkpoints	0.50610796583681905	5.15789473684211E-4	0.43387459095351899	4.5920087888107301E-2
G2/M checkpoint	0.512729124166314	0	0.14670544221048701	1.5531265315500799E-3
Mitotic M-M/G1 Phases	0.52443543853066299	0	2.0857918398400802E-2	2.0830142800796401E-6
Mitotic Spindle Checkpoint	0.52703436753805799	0	8.9947658165228594E-6	1.5147907820120499E-14
Rb Tumor suppressor/Check. P. Sign. in Response to Damage	0.52832526283181902	0	8.3734828642258799E-2	5.1403045654296901E-3
Regulation of Mitotic Cell Cycle	0.51214624757219196	0	0.14580627042091601	4.3968720869584502E-4
S Phase	0.52313223878200998	0	9.0158414179054101E-3	4.93895403664354E-8
Apoptosis - Homo sapiens (human)	0.49472429306006699	0.28517999999999999	9.0158414179054101E-3	0.3797253024193549 8
Apoptotic signaling in response to DNA Damage	0.49140533574417899	0.38271578947368401	3.5648706967610902E-4	6.4455255802974398E-2
Caspase Cascade in Apoptosis	0.49953066324795897	2.5407407407407399E-2	1	0.65151742788461497
Death Receptor Signalling	0.49389812848331899	0.365910810810811	0.20279342826650901	0.75950634733159506
Extrinsic Pathway for Apoptosis	0.50942403752475396	2.3519999999999999E-2	1	0.39731297348484901
Granzyme a Mediated Apoptosis Pathway	0.51922914068235304	8.30869565217391E-3	0.82232843361219399	5.9814453125E-2
Induction of apoptosis through dr3 and dr4/5 death receptors	0.49259325170273	0.34871666666666701	9.0158414179054101E-3	0.31238190589251602
Intrinsic Pathway for Apoptosis	0.48936651697331701	0.55506744186046497	7.8591148084964906E-3	0.17235866912293801
Regulation of Apoptosis	0.48532121063374001	0.670891666666667	0.44684336035083799	0.37972530241935498
TNF Receptor Signaling Pathway	0.49216088894543197	0.47063902439024402	0.12942995892211601	0.65151742788461497
tnfr1 Signaling Pathway	0.49666195385488099	0.15205312500000001	0.45333469397870102	0.985936868935824
tnfr2 Signaling Pathway	0.48532121063374001	0.670891666666667	0.44684336035083799	0.37972530241935498

Table 4: Details of data analysis of GSE20916 macrodissected series using PATHChange. Consensus pathways are marked in light gray color.

Colorectal Adenoma (CRA) - GSE20916 Macro				
Pathway	Activity	Bootstrap	Fisher	Wilcoxon
Canonical NF-kappaB	0.49836300384499399	0.13696666666666699	9.6175419942705707E-3	0.5522236451212595
IL10 Anti-inflammatory Signaling Pathway	0.50644872867236501	0	0.75127556130371997	2.1614688634872401E-2
IL6-mediated signaling events	0.50475741262146001	0	0.83310182287536705	0.13388935724894199
IL8- and CXCR1-mediated signaling events	0.49827657102608103	0.136721951219512	0.106495462625424	0.67154450112200803
IL8- and CXCR2-mediated signaling events	0.49871874743154399	7.2887499999999994E-2	5.8173371240194699E-2	0.662385879001684
mtor signaling	0.49566130813989101	0.49224583333333299	5.6081699204125898E-6	3.0736694939908799E-2
Nemo	0.50148315941796595	2.0176470588235299E-2	0.109578427938799	0.502545250786676
NF-kappaB pathway	0.50742687289881006	0	0.46857034217173998	2.9833931247817199E-5
nf-kb signaling	0.49653818825649798	0.37135744680851102	0.49234596015834398	0.52286585365853699
p53	0.50275137567941397	1.8148148148148099E-3	0.27440718432126099	0.50349993641312096
Regulation of p38-alpha and p38-beta	0.49969799975330997	2.5760000000000002E-2	4.7482100243431503E-2	0.90179570321925095
Replicative Senescence	0.50523336371918504	1.7818181818181799E-2	1	0.52286585365853699
RIG-I	0.502760112275497	7.1866666666666702E-3	0.82939173306477199	0.43205973307291701
SASP	0.49748327689410299	0.22528604651162801	5.0637764703790999E-2	0.77191992600758896
TGF	0.50377584845960199	0	0.27440718432126099	6.6576086956521702E-2
TNF	0.497188535840068	0.22963181818181799	8.0244103419866306E-2	0.52158501768778898
ATM signaling	0.50664764504166504	2.2272727272727301E-4	0.82939173306477199	6.8475087483723995E-2
ATR signaling	0.51077462742823299	0	9.6708923041332995E-2	1.2749223969876799E-7
Base Excision Repair	0.51158633036814305	0	6.2558744417278198E-2	9.34600830078125E-5
Double-Strand Break Repair	0.50973604913291404	0	2.0447680760448301E-2	1.7592486213235299E-4
Fanconi Anemia Pathway	0.50829888149560298	0	6.5205640931745801E-2	4.8808180368863601E-5
Homologous Recombination	0.51085453341694997	0	8.8788265602089494E-2	7.7883402506510204E-5
Hr Repair of Replication-Independent DSB	0.50813423504194299	1.225E-3	7.5436777064572705E-2	2.5185032894736799E-3
Mismatch Repair	0.50964692721000404	0	3.51151114310034E-4	5.78402274224268E-9
Non-Homologous end Joining	0.50782087938250597	0	2.0447680760448301E-2	9.6315807766384604E-4
Nucleotide Excision Repair	0.50756054753553903	4.2608695652173898E-4	0.62848846296819105	8.3261718750000005E-2
Processing of DNA DSB ends Recruitment of Repair and Sig. Proteins	0.49383140081933302	0.68469999999999998	1.11229180947344E-2	6.6576086956521702E-2
Cell Cycle Checkpoints	0.50841997237791803	0	1.9776357880515101E-7	6.1734916571483905E-16
Cell Cycle, Mitotic	0.50841668624659497	0	1.4584576685843501E-8	5.3228747180587397E-30
Cyclins and Cell Cycle Regulation	0.50161153328990904	3.05605263157895E-2	0.82482598489280501	0.29783472521551702
G1/S DNA Damage Checkpoints	0.50626547558441704	0	9.0397412699825698E-5	3.5263570782624501E-9
G2/M checkpoint	0.50662736745404502	0	4.9207746824733503E-5	2.0930271623786599E-9
Mitotic M-M/G1 Phases	0.50702410738483095	0	0.47974434204049499	1.0871659708864301E-6
Mitotic Spindle Checkpoint	0.50913593995003104	0	5.8425689051225098E-8	2.6803067879492001E-20
Rb Tumor suppressor/Check. P. Sign. in Response to Damage	0.50995091915301405	0	4.7482100243431503E-2	5.340576171875E-5
Regulation of Mitotic Cell Cycle	0.50725811120421405	0	1.02025777069087E-5	4.1937363217683198E-11
S Phase	0.50865498404517995	0	5.8425689051225098E-8	6.1795454588972403E-15
Apoptosis - Homo sapiens (human)	0.50207912699528001	3.8500000000000001E-3	0.50071464991431902	0.45851556764495
Apoptotic signaling in response to DNA Damage	0.49880048016738499	1.05903225806452E-2	2.38043795360069E-6	0.52286585365853699
Caspase Cascade in Apoptosis	0.50503364897261305	0	4.8732155128232398E-3	5.9043140746932099E-11
Death Receptor Signalling	0.50099856190509395	2.7494444444444401E-2	0.27100588093251299	0.67154450112200803
Extrinsic Pathway for Apoptosis	0.50399577660539796	7.0965517241379297E-3	0.66653275419108704	0.502545250786676
Granzyme a Mediated Apoptosis Pathway	0.50291515175119605	3.05605263157895E-2	0.62848846296819105	0.21533203125
Induction of apoptosis through dr3 and dr4/5 death receptors	0.50059121182787303	1.8148148148148099E-3	1.06312644121264E-2	0.645212566875902
Intrinsic Pathway for Apoptosis	0.50095211996398503	1.5679999999999999E-3	0.49234596015834398	9.0693491679128602E-2
Regulation of Apoptosis	0.49711724824889603	0.32563695652173902	1.1122918094734E-2	0.502545250786676
TNF Receptor Signaling Pathway	0.50380016719036802	1.4853125E-2	0.82482598489280501	0.461318969726562
tnfr1 Signaling Pathway	0.49890899812772899	4.7115384615384601E-2	3.3746271776840301E-3	0.58513337490148898
tnfr2 Signaling Pathway	0.49711724824889603	0.31316444444444402	1.1122918094734E-2	0.502545250786676

Table 5: Details of data analysis of GSE37364 series using PATHChange. Consensus pathways are marked in light gray color.

Colorectal Adenoma (CRA) - GSE37364				
Pathway	Activity	Bootstrap	Fisher	Wilcoxon
Canonical NF-kappaB	0.50751576123332398	2.87E-2	1.3283274615991001E-2	0.39898866144093598
IL10 Anti-inflammatory Signaling Pathway	0.54124808156832904	2.5789473684210501E-4	0.88308082557973999	9.6130933450615894E-2
IL6-mediated signaling events	0.53698409828364602	4.8999999999999998E-4	1	0.34688293933868403
IL8- and CXCR1-mediated signaling events	0.48939850096160098	0.223094117647059	6.8217617389462002E-3	0.39898866144093598
IL8- and CXCR2-mediated signaling events	0.486914280717619	0.24737999999999999	7.3618833363560095E-4	0.361714832396408
mtor signaling	0.537667178190117	0	1.9571709938907798E-3	0.86676611771724199
Nemo	0.506645953365782	4.2203225806451597E-2	0.10639011768933	0.91563249670940805
NF-kappaB pathway	0.53377523355213896	0	0.42472446928066698	2.0210082955611701E-3
nf-kb signaling	0.48611441214940498	0.335895	7.8203278653440705E-2	0.979949951171875
p53	0.48736979221440901	0.2646	0.131844773391254	0.95701234958445103
Regulation of p38-alpha and p38-beta	0.48767269393981399	0.223094117647059	3.2425380764907602E-4	0.28459194565463503
Replicative Senescence	0.490757079911609	0.30782051282051298	1.4104884429256E-2	0.39898866144093598
RIG-I	0.52899403440437498	3.6217391304347799E-3	0.10639011768933	0.91563249670940805
SASP	0.48588341199826302	0.306894736842105	1.9711929483140101E-6	6.2109336256981E-2
TGF	0.500498833873537	5.3083333333333298E-3	2.3887170724148699E-3	0.86676611771724199
TNF	0.48108154149101501	0.42527441860465098	1.21176212941056E-2	0.70702366833575103
ATM signaling	0.51427761117556503	3.1033333333333302E-2	0.50529438484672495	0.122370402018229
ATR signaling	0.52272231192241303	2.5789473684210501E-4	0.19652969090447001	2.5782335651456398E-4
Base Excision Repair	0.50664885224513601	2.87E-2	6.6971047226270605E-2	6.4896345138550004E-3
Double-Strand Break Repair	0.56515311881535002	2.5789473684210501E-4	4.8186508329563699E-2	2.3005558894230801E-4
Fanconi Anemia Pathway	0.554370902760006	0	1.8322740015368499E-2	2.8806452808732399E-5
Homologous Recombination	0.56179388811005204	0	2.5329655999038601E-2	2.5782335651456398E-4
Hr Repair of Replication-Independent DSB	0.56360119850860202	6.6818181818181798E-4	0.14700560685344899	2.6584201388888899E-3
Mismatch Repair	0.55505055807656001	0	7.0111333980833705E-5	1.56079974915452E-7
Non-Homologous end Joining	0.55208264704503096	0	3.4624713490358103E-2	1.07089678446452E-4
Nucleotide Excision Repair	0.53299260678349203	8.2319999999999997E-3	0.28834038268759	6.8929036458333301E-2
Processing of DNA DSB ends Recruitment of Repair and Sig. Proteins	0.47963060201482799	0.47867555555555602	2.6629504641715599E-2	0.24661959134615399
Cell Cycle Checkpoints	0.53365038617661198	0	4.1118222727917701E-9	5.0096798345908399E-14
Cell Cycle, Mitotic	0.53168206131816298	0	3.0264265749542798E-13	2.6443145401871097E-29
Cyclins and Cell Cycle Regulation	0.482302705529364	0.42527441860465098	1	0.39898866144093598
G1/S DNA Damage Checkpoints	0.52568947535312005	0	2.6467152985503601E-5	1.4630783036988699E-7
G2/M checkpoint	0.52764448060946501	0	3.1711663340192998E-5	1.1660993316406E-8
Mitotic M-M/G1 Phases	0.52446665524869196	0	5.7599382045663498E-2	6.4449309252059404E-7
Mitotic Spindle Checkpoint	0.52943047100955798	0	2.2390901234620698E-9	5.9675228229361302E-19
Rb Tumor suppressor/Check. P. Sign. in Response to Damage	0.55913971010432095	0	2.24503190561916E-2	6.4253807067871105E-4
Regulation of Mitotic Cell Cycle	0.53102106639364399	0	1.1393947363866299E-6	1.3049593722863599E-10
S Phase	0.536261980450501	0	2.7626948569481302E-9	2.6265554045322302E-13
Apoptosis - Homo sapiens (human)	0.48096287834036699	0.42540909090909101	1.8322740015368499E-2	0.67362559586763404
Apoptotic signaling in response to DNA Damage	0.47477743287337498	0.71800638297872299	3.0069594198341201E-7	0.136633643463694
Caspase Cascade in Apoptosis	0.52469434350427002	0	2.5946542808950999E-2	1.3722683125645401E-6
Death Receptor Signalling	0.482247327075084	0.42527441860465098	0.120832063922181	0.87323154102672196
Extrinsic Pathway for Apoptosis	0.51048910215555399	4.6703124999999998E-2	1	0.50208955652573495
Granzyme a Mediated Apoptosis Pathway	0.56753696139135401	6.6818181818181798E-4	0.56393081460875305	9.3528053977272693E-2
Induction of apoptosis through dr3 and dr4/5 death receptors	0.49584746455098699	3.0244827586206901E-2	1.3890280592058201E-2	0.86676611771724199
Intrinsic Pathway for Apoptosis	0.48413016091167499	0.29135135135135098	8.5889608216474407E-3	0.53894291398726502
Regulation of Apoptosis	0.46802032913059899	0.72030000000000005	5.0901924781306701E-2	0.60234770903716195
TNF Receptor Signaling Pathway	0.47527054270142999	0.58107608695652202	0.2549554385151320 3	0.94837724401595702
tnfr1 Signaling Pathway	0.50267049430281097	1.4511538461538499E-2	1.21499172715388E-2	0.64825140462792397
tnfr2 Signaling Pathway	0.46802032913059899	0.72030000000000005	5.0901924781306701E-2	0.60234770903716195

Table 6: Details of data analysis of GSE15960 series using PATHChange. Consensus pathways are marked in light gray color.

Colorectal Carcinoma (CRC) - GSE15960				
Pathway	Activity	Bootstrap	Fisher	Wilcoxon
Canonical NF-kappaB	0.50667269117517399	1.781818181818818E-3	0.79719969286123205	0.3789465362961202
IL10 Anti-inflammatory Signaling Pathway	0.51031769264547999	0	0.70166426234838797	4.1749574921347801E-2
IL6-mediated signaling events	0.51486396241605603	1.9599999999999999E-4	0.38890102081944	5.1227807998657096E-3
IL8- and CXCR1-mediated signaling events	0.50098382585738499	5.5609302325581401E-2	0.48944924055012001	0.92148147183250695
IL8- and CXCR2-mediated signaling events	0.50213545377516	1.47E-2	0.653727457106559	0.61673656729764703
mtor signaling	0.50615577431022996	0	0.216917457199823	0.78489334703042901
Nemo	0.51059807492571796	0	0.216917457199823	3.9397019606370199E-3
NF-kappaB pathway	0.50838539598698496	0	0.216917457199823	1.85961991584155E-3
nf-kb signaling	0.50144501765047	8.6022222222222194E-2	0.39593737647383298	0.89686681869182205
p53	0.504332005835196	4.7600000000000003E-3	0.70166426234838797	0.124251617696779
Regulation of p38-alpha and p38-beta	0.50613743625559005	1.9599999999999999E-4	1	6.1650042885538402E-2
Replicative Senescence	0.50866113265392898	1.5076923076923101E-2	1	0.56021341463414598
RIG-I	0.51449763831089601	0	0.51225710495398502	4.3487382971722201E-2
SASP	0.50303817884096302	1.4957894736842101E-2	0.97129395331776403	0.537689876556397
TGF	0.50672862156666798	0	1	1.6774415969848602E-2
TNF	0.50465730417393195	1.26451612903226E-3	1	0.49794439783344402
ATM signaling	0.50848667454997998	1.5312500000000001E-3	1	0.130146280924479
ATR signaling	0.51356752077634904	0	5.9088735021829604E-3	1.458293506099953E-6
Base Excision Repair	0.50695842101674604	1.1827586206896599E-3	0.216917457199823	1.2753407160441099E-2
Double-Strand Break Repair	0.50928033955033503	1.26451612903226E-3	0.48944924055012001	1.6212864925986802E-2
Fanconi Anemia Pathway	0.50520036176602801	1.1827586206896599E-3	0.97342031586237798	0.21110983455882401
Homologous Recombination	0.51133034280005896	0	0.40083812297578703	1.6774415969848602E-2
Hr Repair of Replication-Independent DSB	0.50929074907429805	4.8999999999999998E-3	0.54490025156753796	5.2636718749999999E-2
Mismatch Repair	0.51087146152340801	0	7.2972539554626098E-2	6.8722476221141503E-4
Non-Homologous end Joining	0.50970012589891001	0	0.38890102081944	4.4159889221191398E-3
Nucleotide Excision Repair	0.51249806735868697	1.9599999999999999E-4	0.345749815913171	4.3863932291666699E-2
Processing of DNA DSB ends Recruitment of Repair and Sig. Proteins	0.51909634940244997	0	4.2743039742736202E-2	5.6295955882352897E-3
Cell Cycle Checkpoints	0.51172097662277305	0	1.82902733282307E-4	2.5311909084275899E-10
Cell Cycle, Mitotic	0.509512981212076	0	1.3357153172388499E-6	2.139545074031131E-16
Cyclins and Cell Cycle Regulation	0.50490373149831302	2.2662499999999999E-2	0.44108309214406299	0.21430664062499999
G1/S DNA Damage Checkpoints	0.50981984961079096	0	2.6883135826493799E-2	2.0534396757700099E-5
G2/M checkpoint	0.50956763118246595	0	2.474735011271 96E-3	9.5866377848983 97E-6
Mitotic M-M/G1 Phases	0.508328849311667	0	0.28163975623049903	1.3981525412203499E-4
Mitotic Spindle Checkpoint	0.50966342911610896	0	1.3369952496000001E-3	2.5311909084275899E-10
Rb Tumor suppressor/Check. P. Sign. in Response to Damage	0.51537629633453097	0	0.216917457199823	5.1227807998657096E-3
Regulation of Mitotic Cell Cycle	0.51099420863859901	0	2.474735011271 96E-3	1.0611039095071001E-7
S Phase	0.51351140792124605	0	6.0179057636512797E-7	3.6104773059583202E-11
Apoptosis - Homo sapiens (human)	0.50710383711198204	5.4444444444444397E-4	0.70166426234838797	0.23262214660644501
Apoptotic signaling in response to DNA Damage	0.49472298321556302	0.56520434782608697	2.2576897568880299E-3	0.19829126709033401
Caspase Cascade in Apoptosis	0.508172682988141	0	0.216917457199823	6.8722476221141503E-4
Death Receptor Signalling	0.50123708092836605	7.2275000000000006E-2	1	0.61673656729764703
Extrinsic Pathway for Apoptosis	0.508374273842646	2.7382352941176498E-3	0.75341377703821	8.8653564453125E-2
Granzyme a Mediated Apoptosis Pathway	0.51625002971471901	1.9599999999999999E-4	0.40083812297578703	7.6208043981481496E-2
Induction of apoptosis through dr3 and dr4/5 death receptors	0.49965507572133999	4.3143902439024398E-2	0.40083812297578703	0.89686681869182205
Intrinsic Pathway for Apoptosis	0.49947535480673899	5.0049999999999997E-2	0.53066526699372196	0.92148147183250695
Regulation of Apoptosis	0.49070141069671203	0.78420000000000001	0.150136899404855	0.21110983455882401
TNF Receptor Signaling Pathway	0.49252411784161099	0.70101276595744699	0.40083812297578703	0.53250200320512797
tnfr1 Signaling Pathway	0.50510674666153499	5.4444444444444397E-4	0.40083812297578703	0.85850910083002696
tnfr2 Signaling Pathway	0.49070141069671203	0.78420000000000001	0.150136899404855	0.21110983455882401

Table 7: Details of data analysis of GSE20916 microdissected series using PATHChange. Consensus pathways are marked in light gray color.

Colorectal Carcinoma (CRC) - GSE20916 Micro				
Pathway	Activity	Bootstrap	Fisher	Wilcoxon
Canonical NF-kappaB	0.5000628836200250 4	3.9620000000000002E-2	3.8061473264116401E-2	0.83344305947769504
IL10 Anti-inflammatory Signaling Pathway	0.4992439781385860 1	4.1513888888888899E-2	2.34735261884172E-2	0.83344305947769504
IL6-mediated signaling events	0.5136532151021310 2	4.45454545454545E-4	0.44879404477863799	0.53897183736165299
IL8- and CXCR1-mediated signaling events	0.4891700347374959 8	0.34136666666666698	3.03454190367937E-3	0.16187083202859601
IL8- and CXCR2-mediated signaling events	0.4905594185986070 2	0.255888888888889	1.9279709531740601E-3	7.3987841606140206E-2
mtor signaling	0.4951539976034249 9	4.82263157894737E-2	1.10081076704264E-4	0.219004419205134
Nemo	0.5034919826204290 5	2.18866666666667E-2	0.296448812372962	0.83344305947769504
NF-kappaB pathway	0.5183856068945970 1	0	0.65494639642430397	1.73371251208199E-3
nf-kb signaling	0.4901656128154490 1	0.33893404255319098	0.68098523114519505	1
p53	0.5041999983453450 97	4.3346153846153803E-3	6.0396508725436199E-2	0.88950590503976701
Regulation of p38-alpha and p38-beta	0.4978594381549569 8	3.4146875E-2	0.1531656779683601	0.82068926682695698
Replicative Senescence	0.4914178619034940 1	0.33893404255319098	0.35917022067098697	0.83522727272727304
RIG-I	0.5123212840095949 8	1.225E-3	0.68101734479414699	0.36358642578125
SASP	0.4848698662875570 2	0.56379999999999997	3.00419745206153E-2	4.4300079345703097E-2
TGF	0.5048784487308469 7	4.45454545454545E-4	6.0396508725436199E-2	0.78681433741795503
TNF	0.4913613744404660 1	0.22015000000000001	0.186831137282084	0.82068926682695698
ATM signaling	0.5160183753884789 4	6.3913043478260902E-4	0.862108354248318	0.34991455078125
ATR signaling	0.5423376259309680 1	0	3.8623894803703801E-3	3.8112193578854299E-7
Base Excision Repair	0.5385632270064220 1	0	0.26791395538576601	5.6409835815429601E-4
Double-Strand Break Repair	0.5361667400445160 3	0	0.68101734479414699	1.1381361219618099E-2
Fanconi Anemia Pathway	0.528873203308039	0	0.89518088200469603	1.3349669936456201E-2
Homologous Recombination	0.5413365128888659 7	0	0.118960165560721	1.674493153889 97E-4
Hr Repair of Replication-Independent DSB	0.5233325392283000 3	4.45454545454545E-4	1	0.14246715198863599
Mismatch Repair	0.5298482890614180 1	0	3.00419745206153E-2	3.3226719153845003E-8
Non-Homologous end Joining	0.5336061262001680 1	0	0.26791395538576601	5.7789484659830602E-4
Nucleotide Excision Repair	0.5341611462456430 4	0	0.186831137282084	8.4443933823529407E-3
Processing of DNA DSB ends Recruitment of Repair and Sig. Proteins	0.4950416055009870 3	0.23241590909090901	0.32727022446558002	0.80937499999999996
Cell Cycle Checkpoints	0.5292139751178279 7	0	1.2018827912118E-6	4.9436278927507201E-14
Cell Cycle, Mitotic	0.5303822435467869 8	0	1.4990948045020601E-13	6.2183881006815E-30
Cyclins and Cell Cycle Regulation	0.5127155324248650 1	4.3119999999999999E-3	0.84644625116640204	0.44963968211206901
G1/S DNA Damage Checkpoints	0.5161081948314439 9	0	2.5689399373439899E-2	6.6642083234585103E-7
G2/M checkpoint	0.5209334479590229 7	0	6.2943656469436797E-3	3.3170639443156098E-8
Mitotic M-M/G1 Phases	0.5230325066042810 2	0	2.5689399373439899E-2	9.5137350114871603E-7
Mitotic Spindle Checkpoint	0.5326706740817279 7	0	2.3535281540403601E-9	2.6144548211819301E-20
Rb Tumor suppressor/Check. P. Sign. in Response to Damage	0.5371364109860999 9	0	3.8061473264116401E-2	2.3005558894230801E-4
Regulation of Mitotic Cell Cycle	0.5231726076421739 9	0	6.1084139753431199E-4	2.2626269867984299E-10
S Phase	0.5300621905239140 1	0	1.8888740961333501E-5	5.32798221066343E-14
Apoptosis - Homo sapiens (human)	0.4981780257221710 1	4.35702702702703E-2	2.2689699066253399E-2	0.88396445757763897
Apoptotic signaling in response to DNA Damage	0.4938365955452109 8	3.5042424242424201E-2	3.1678724772039399E-5	0.25847968653909897
Caspase Cascade in Apoptosis	0.5104950883506490 5	0	0.27229967998610399	7.2297360148992798E-5
Death Receptor Signalling	0.499101696644078	6.5835897435897406E-2	0.27574841372235098	1
Extrinsic Pathway for Apoptosis	0.5096664517667880 4	7.0000000000000001E-3	1	0.56515010710685498
Granzyme a Mediated Apoptosis Pathway	0.5152668046606609 9	7.0000000000000001E-3	1	0.32943960336538503
Induction of apoptosis through dr3 and dr4/5 death receptors	0.4989239664945990 2	7.7724137931034498E-3	2.2689699066253399E-2	0.88396445757763897
Intrinsic Pathway for Apoptosis	0.495624740647992	3.5597058823529398E-2	9.72654363720994E-3	0.58273143293763496
Regulation of Apoptosis	0.4968549537250430 1	0.16946829268292701	0.65494639642430397	0.82068926682695698
TNF Receptor Signaling Pathway	0.4941876896814719 8	0.23241590909090901	5.8743258818634701E-2	0.76997514204545503
tnfr1 Signaling Pathway	0.4973242575526379 8	3.1929032258064503E-2	7.3789870964732301E-2	0.82068926682695698
tnfr2 Signaling Pathway	0.4968549537250430 1	0.1637825	0.65494639642430397	0.82068926682695698

Table 8: Details of data analysis of GSE20916 macrodissected series using PATHChange. Consensus pathways are marked in light gray color.

Colorectal Carcinoma (CRC) - GSE20916 Macro				
Pathway	Activity	Bootstrap	Fisher	Wilcoxon
Canonical NF-kappaB	0.50575862529975801	0	0.56937990234137303	0.16490465221983
IL10 Anti-inflammatory Signaling Pathway	0.51313497260298901	0	0.50036033074418995	6.44228960338391E-4
IL6-mediated signaling events	0.512452970849955	0	0.72306696583977004	7.1954917907714796E-3
IL8- and CXCR1-mediated signaling events	0.504152820429142	0	4.1045900355149698E-2	0.731675624847412
IL8- and CXCR2-mediated signaling events	0.50423963973159502	0	4.1262263991218703E-3	0.57646423193315699
mtor signaling	0.50696055531587103	0	0.32900632006631197	5.6784808049902998E-2
Nemo	0.50544386774469696	0	0.85379307066993004	6.2867482503255301E-2
NF-kappaB pathway	0.51195299340749501	0	0.451544225053158	4.1572704224298197E-6
nf-kb signaling	0.50598394279811998	0	0.22586574060937101	0.4561564127604170 2
p53	0.50847797585875998	0	0.397513535243672	5.1014080643653897E-2
Regulation of p38-alpha and p38-beta	0.50500043536323902	0	0.78218117991243696	2.4368851273148199E-2
Replicative Senescence	0.50863768727418501	6.5333333333333302E-4	0.72018640574877502	0.49854651162790697
RIG-I	0.51400611225884796	0	1	6.8693161010742196E-3
SASP	0.50181058977815496	1.1136363636363601E-4	4.1129269865243703E-3	0.98543548583984397
TGF	0.50905709023421497	0	0.56937990234137303	1.09178098677127E-3
TNF	0.50196036882296102	0	0.88449305528794397	0.25755673088133402
ATM signaling	0.51065291577153404	0	0.52030836465924002	7.476806640625E-4
ATR signaling	0.520024673572267	0	1.6660285364683801E-2	2.49565346166492E-9
Base Excision Repair	0.51880215042925004	0	0.40889302588084397	4.1537814670138899E-4
Double-Strand Break Repair	0.51651588442430696	0	7.12560894340281E-2	1.86920166015625E-4
Fanconi Anemia Pathway	0.51444785681048499	0	6.63835897295042E-2	1.98706984519959E-5
Homologous Recombination	0.517085463950402	0	8.2960384771564102E-2	3.0237085678997999E-4
Hr Repair of Replication-Independent DSB	0.51243091747994896	0	0.14099102890245599	2.0805027173913101E-3
Mismatch Repair	0.51401552287716901	0	4.1262263991218703E-3	4.8908610494891098E-10
Non-Homologous end Joining	0.51445222716579997	0	8.2960384771564102E-2	7.1892371544471203E-6
Nucleotide Excision Repair	0.51598844244434705	0	0.451544225053158	2.0805027173913101E-3
Processing of DNA DSB ends Recruitment of Repair and Sig. Proteins	0.502644235572132	5.3260869565217401E-3	0.13333932871561399	0.82942708333333304
Cell Cycle Checkpoints	0.51472820015673504	0	8.0035616869568404E-8	4.8448080458787401E-16
Cell Cycle, Mitotic	0.51490793003633295	0	4.3248892231783602E-7	5.9304203922116904E-35
Cyclins and Cell Cycle Regulation	0.51229004655777599	0	1	0.10110572076612 9
G1/S DNA Damage Checkpoints	0.51062166199259995	0	3.6622927321917198E-4	5.1829895045421099E-9
G2/M checkpoint	0.51235323254169096	0	3.581981456422223E-4	1.1410942119961601E-9
Mitotic M-M/G1 Phases	0.51225516039042596	0	0.92544121349455399	5.6600049189822502E-8
Mitotic Spindle Checkpoint	0.51567645441231302	0	1.427596751834 19E-7	3.7845863342520901E-23
Rb Tumor suppressor/Check. P. Sign. in Response to Damage	0.51547765579756699	0	3.3565854246056299E-2	2.4922688802083299E-5
Regulation of Mitotic Cell Cycle	0.51242855147422295	0	1.3683886919326E-4	5.53128827716767E-11
S Phase	0.51612673049313995	0	1.981537028701 01E-6	5.3282014467005299E-15
Apoptosis - Homo sapiens (human)	0.50453512027118896	0	0.13333932871561399	0.320225352911573
Apoptotic signaling in response to DNA Damage	0.502443585239426	0	3.5806633976951698E-2	0.22740091379263599
Caspase Cascade in Apoptosis	0.50749474172452003	0	8.5843269697702507E-2	4.7373912437365201E-8
Death Receptor Signalling	0.50308248695289604	0	0.25648352534151703	0.57585005326704497
Extrinsic Pathway for Apoptosis	0.50959701507189104	0	0.65696761933532499	2.4368851273148199E-2
Granzyme a Mediated Apoptosis Pathway	0.51026769827529705	1.1136363636363601E-4	0.82301135926768398	0.127105712890625
Induction of apoptosis through dr3 and dr4/5 death receptors	0.50260101855468098	0	2.57892540643293E-3	0.41106384309560201
Intrinsic Pathway for Apoptosis	0.50061632779325804	0	4.28243963404904E-2	0.42379392223110901
Regulation of Apoptosis	0.49472414740305398	0.1608	4.1262263991218703E-3	0.27288323479729698
TNF Receptor Signaling Pathway	0.50099377449400495	9.59148936170213E-3	0.63445411725331002	0.76460688164893598
tnfr1 Signaling Pathway	0.50160789903984504	0	0.25648352534151703	0.40390279533293799
tnfr2 Signaling Pathway	0.49472414740305398	0.1608	4.1262263991218703E-3	0.27288323479729698

Table 9: Details of data analysis of GSE37364 series using PATHChange. Consensus pathways are marked in light gray color.

Colorectal Carcinoma (CRC) - GSE37364				
Pathway	Activity	Bootstrap	Fisher	Wilcoxon
Canonical NF-kappaB	0.54163294447814403	0	0.172676565510267	0.75552219297827705
IL10 Anti-inflammatory Signaling Pathway	0.58318376571814001	0	0.35977546815193701	1.8058824539184601E-2
IL6-mediated signaling events	0.56319079901579305	0	0.85659423770183396	8.7627172470092704E-2
IL8- and CXCR1-mediated signaling events	0.51910507417989304	6.7735294117647104E-3	0.18959782015495899	0.45869745612144402
IL8- and CXCR2-mediated signaling events	0.51196388113985203	1.7919999999999998E-2	5.0729798721153199E-2	0.71888850936188597
mtor signaling	0.58846480359595799	0	0.75984117026505504	5.7008479929624697E-2
Nemo	0.51283260556094601	4.5733333333333299E-2	0.38102670420205298	0.78834375422051595
NF-kappaB pathway	0.54530359679873996	0	0.15669214553838501	3.2643282656002603E-4
nf-kb signaling	0.51463840104366998	5.4694594594594602E-2	0.844041764538772	0.75552219297827705
p53	0.53498491335667098	1.8148148148148101E-4	0.80127311446791305	6.7382178340966895E-2
Regulation of p38-alpha and p38-beta	0.53884779282754802	0	0.74269689460660104	9.3940763715213094E-2
Replicative Senescence	0.51856737750839299	7.9907692307692305E-2	0.170253647096032	0.76106770833333304
RIG-I	0.60635762221567602	0	1	8.0867202193648693E-2
SASP	0.52033574045063402	4.2874999999999996E-3	0.36433250920627802	0.59922034837104199
TGF	0.53552080840364003	0	0.90956313430926705	2.5431242233159498E-2
TNF	0.516326264300497	4.1096774193548399E-3	0.52890220655923803	0.50296473511520501
ATM signaling	0.537331165738126	2.0275862068965499E-3	0.26174830245473102	1.223835192228 62E-2
ATR signaling	0.54608760490435404	0	5.0729798721153199E-2	2.056667093783 11E-7
Base Excision Repair	0.49463727899932802	0.33308064651162799	0.17576099697620701	4.8855242521866502E-2
Double-Strand Break Repair	0.58420215454784397	0	5.0729798721153199E-2	3.1153361002604201E-4
Fanconi Anemia Pathway	0.571089863492941	0	0.59547493933067297	1.596485425789 77E-4
Homologous Recombination	0.57711104506149002	0	0.17576099697620701	6.8961171542896999E-4
Hr Repair of Replication-Independent DSB	0.58176084474967804	0	0.170253647096032	3.9876302083333296E-3
Mismatch Repair	0.561929748119395	0	6.6695071123042203E-4	2.056667093783 11E-7
Non-Homologous end Joining	0.56196909236190096	0	4.7003553648660198E-2	2.1240927956321E-5
Nucleotide Excision Repair	0.53598046142745104	5.0484848484848499E-3	0.844041764538772	6.7382178340966895E-2
Processing of DNA DSB ends Recruitment of Repair and Sig. Proteins	0.56228321563113504	1.575E-3	0.844041764538772	0.287109375
Cell Cycle Checkpoints	0.54392695979911898	0	2.7832752986470302E-7	1.427759793098 58E-13
Cell Cycle, Mitotic	0.541216807940611	0	9.554248337225421E-10	3.5137828008833297E-27
Cyclins and Cell Cycle Regulation	0.49816432229698798	0.29551666666666698	0.844041764538772	0.241959110383065
G1/S DNA Damage Checkpoints	0.53157457651762796	0	6.6695071123042203E-4	5.4933983388115096E-7
G2/M checkpoint	0.53568372522697905	0	6.6695071123042203E-4	5.7633042426618398E-8
Mitotic M-M/G1 Phases	0.53213204003955605	0	0.75984117026505504	1.208352035044 62E-4
Mitotic Spindle Checkpoint	0.53838239293401702	0	4.3583381310947303E-9	9.0176304115553899E-18
Rb Tumor suppressor/Check. P. Sign. in Response to Damage	0.582917659683025	0	3.2503269357187899E-2	3.1153361002604201E-4
Regulation of Mitotic Cell Cycle	0.54010931943587903	0	5.0635239792078303E-6	5.6639813863590898E-10
S Phase	0.54700427674143004	0	5.2727838948433302E-8	7.8780202851259197E-13
Apoptosis - Homo sapiens (human)	0.48263116675849299	0.64252553191489403	0.56562367665683599	0.67327411659061898
Apoptotic signaling in response to DNA Damage	0.49285759766836501	0.28898048780487801	4.7003553648660198E-2	0.78834375422051595
Caspase Cascade in Apoptosis	0.53143577104616302	0	4.7003553648660198E-2	2.4903464862883499E-6
Death Receptor Signalling	0.46613587043963001	0.83994166666666703	5.0729798721153199E-2	0.78834375422051595
Extrinsic Pathway for Apoptosis	0.51158713378435605	7.9560526315789498E-2	0.455307632603908	0.13597819010416701
Granzyme a Mediated Apoptosis Pathway	0.577842801609454	1.8148148148148101E-4	0.21683925969787801	2.2786458333333301E-2
Induction of apoptosis through dr3 and dr4/5 death receptors	0.49742998756094398	0.1836275	0.45551900402106799	0.45869745612144402
Intrinsic Pathway for Apoptosis	0.48509163005595801	0.62697727272727299	5.9379616688272401E-2	0.78834375422051595
Regulation of Apoptosis	0.47756360880181498	0.64252553191489403	0.52890220655923803	1
TNF Receptor Signaling Pathway	0.44970852982966802	0.86009999999999998	0.38466958959394199	0.78834375422051595
tnfr1 Signaling Pathway	0.51638543483657096	2.6133333333333299E-3	0.52890220655923803	0.27580723015955799
tnfr2 Signaling Pathway	0.47756360880181498	0.64252553191489403	0.52890220655923803	1

Table 10: Details of data analysis of GSE26910 series using PATHChange. Consensus pathways are marked in light gray color.

Prostate Cancer - GSE26910				
Pathway	Activity	Bootstrap	Fisher	Wilcoxon
Canonical NF-kappaB	0.49413146434128202	1	0.318830503396827	1.0747909545898399E-2
IL10 Anti-inflammatory Signaling Pathway	0.49260595627209203	1	0.33078012945426	3.7235567967097002E-3
IL6-mediated signaling events	0.48783236435272198	1	0.35534432526813903	1.43475830554962E-2
IL8- and CXCR1-mediated signaling events	0.49416227319315797	1	0.33078012945426	1.0028873498623201E-2
IL8- and CXCR2-mediated signaling events	0.49535061715238698	1	0.20944634838429599	1.3996234598259201E-2
mtor signaling	0.49824146669094599	1	1	0.6388341000205896
Nemo	0.49392369335681902	1	0.85377576203868999	2.5732676188151001E-2
NF-kappaB pathway	0.49969603750786701	1	0.70326821286543595	0.66213787460378903
nf-kb signaling	0.498769060646016	1	0.9808596510479604	0.6388341000205896
p53	0.49491404441126602	1	1	7.6938227211174298E-2
Regulation of p38-alpha and p38-beta	0.49535107171269799	1	0.35534432526813903	1.9189166625865801E-2
Replicative Senescence	0.49437519242125399	1	0.69661368834642501	0.21750710227272699
RIG-I	0.49181661539469601	1	8.5012108768079603E-2	2.3280057040127801E-3
SASP	0.49551202794767102	1	0.73616294170851004	0.147069315115611
TGF	0.493064028671584	1	0.58293945575851303	1.3409039751222499E-3
TNF	0.49866092659007699	1	1	0.6388341000205896
ATM signaling	0.49738560993189601	1	0.85377576203868999	0.6388341000205896
ATR signaling	0.49903032161465599	1	0.73616294170851004	0.89278680107802
Base Excision Repair	0.49869376804964199	1	1	0.6388341000205896
Double-Strand Break Repair	0.49705597625254699	1	0.85377576203868999	0.45774671766493102
Fanconi Anemia Pathway	0.50020938809763205	1	0.77217069701053498	0.87316371448105201
Homologous Recombination	0.498596783571086	1	0.81672614598758697	0.6388341000205896
Hr Repair of Replication-Independent DSB	0.49583053186249298	1	0.58293945575851303	0.22430419921875
Mismatch Repair	0.49841813862658302	1	0.85377576203868999	0.6388341000205896
Non-Homologous end Joining	0.49777303993035499	1	0.99288411382446196	0.49037699863828399
Nucleotide Excision Repair	0.496463694089466	1	1	0.46826171875
Processing of DNA DSB ends Recruitment of Repair and Sig. Proteins	0.49429785491429401	1	1	0.22430419921875
Cell Cycle Checkpoints	0.49683007031515403	1	3.4851245426199898E-2	8.9657259258669206E-5
Cell Cycle, Mitotic	0.49693530775172101	1	1.3964770230948199E-2	1.1492329735863599E-5
Cyclins and Cell Cycle Regulation	0.50183301418229198	1	0.7361629417085 1004	0.63861897786458299
G1/S DNA Damage Checkpoints	0.49595026743667397	1	4.7552015902397196E-3	8.4336519348718396E-5
G2/M checkpoint	0.49578013358734302	1	6.7090401619447702E-4	5.03450035053326E-5
Mitotic M-M/G1 Phases	0.49516234413823701	1	6.41192072278471E-2	1.7488631769382601E-3
Mitotic Spindle Checkpoint	0.49706497465556798	1	0.15688117708844401	3.0291426867568797E-4
Rb Tumor suppressor/Check. P. Sign. in Response to Damage	0.49893347471588301	1	0.72030489287258403	0.39590371704102 01
Regulation of Mitotic Cell Cycle	0.49647900665574102	1	6.7090401619447702E-4	7.2223130592887107E-5
S Phase	0.49555303435849402	1	6.2246898017595507E-5	2.177504386623 55E-6
Apoptosis - Homo sapiens (human)	0.499130592229235	1	0.7361629417085 1004	0.83817350073921904
Apoptotic signaling in response to DNA Damage	0.49765863726616499	1	0.58293945575851303	0.778099543946561
Caspase Cascade in Apoptosis	0.49654246298778199	1	1.0775086728677099E-2	8.5824356184182801E-4
Death Receptor Signalling	0.49822013974187102	1	0.85377576203868999	0.93568605333027299
Extrinsic Pathway for Apoptosis	0.496806165422365	1	0.7361629417085 1004	0.6388341000205896
Granzyme a Mediated Apoptosis Pathway	0.50137702816252705	1	0.81672614598758697	0.9658203125
Induction of apoptosis through dr3 and dr4/5 death receptors	0.500668916113443	1	0.318830503396827	0.93568605333027299
Intrinsic Pathway for Apoptosis	0.50573992432233905	0.38219999999999998	3.4851245426199898E-2	0.140166759735949
Regulation of Apoptosis	0.49763038635123602	1	1	0.87628173828125
TNF Receptor Signaling Pathway	0.50031231393697695	1	0.7361629417085 1004	0.93568605333027299
tnfr1 Signaling Pathway	0.49763524342113602	1	0.93551797843415 95	0.36208640936573
tnfr2 Signaling Pathway	0.49763038635123602	1	1	0.87628173828125

Table 11: Details of data analysis of GSE26910 series using PATHChange. Consensus pathways are marked in light gray color.

Breast Cancer - GSE26910				
Pathway	Activity	Bootstrap	Fisher	Wilcoxon
Canonical NF-kappaB	0.50339100862255004	5.1624999999999997E-2	1	0.30146484374999999
IL10 Anti-inflammatory Signaling Pathway	0.50000868091806105	0.193083333333333	1	0.606475830078125
IL6-mediated signaling events	0.49737507475694698	0.47758666666666699	0.76287142152699705	1
IL8- and CXCR1-mediated signaling events	0.50573236679440803	1.20866666666667E-2	0.53890221911195102	0.267751455307007
IL8- and CXCR2-mediated signaling events	0.50612672893745203	3.0153846153846199E-3	0.2498295480668749 9	0.14188670762814601
mtor signaling	0.49479740184635801	0.84905531914893595	3.3139402134647902E-2	0.30146484374999999
Nemo	0.50339404139648602	5.6503125000000001E-2	0.76287142152699705	0.267751455307007
NF-kappaB pathway	0.50253280251221699	1.41235294117647E-2	0.76287142152699705	0.606475830078125
nf-kb signaling	0.50430043903984501	5.4237931034482802E-2	1	0.606475830078125
p53	0.49502884983119999	0.78485217391304396	0.45125842169212399	0.58965802192687899
Regulation of p38-alpha and p38-beta	0.49978634934050098	0.17309250000000001	0.76287142152699705	0.89656460803488003
Replicative Senescence	0.50014957563448503	0.29274651162790699	0.76287142152699705	0.94888630319148903
RIG-I	0.49216943589436102	0.89770000000000005	0.53890221911195102	0.606475830078125
SASP	0.50117312971139205	0.12418	0.76287142152699705	0.89656460803488003
TGF	0.50011921747171295	8.1858823529411803E-2	0.76287142152699705	0.70095402346122104
TNF	0.49977812470055899	0.18141951219512201	1	0.97962278003978998
ATM signaling	0.49253493741514798	0.89292291666666701	1	0.711035410563151
ATR signaling	0.50242171939769098	3.1473076923076902E-2	1	0.52791022669801602
Base Excision Repair	0.50277545300047599	5.6503125000000001E-2	0.76516088232200996	0.48603305491534299
Double-Strand Break Repair	0.504616859706286	4.990740740740 74E-2	1	0.711035410563151
Fanconi Anemia Pathway	0.50664220024087903	1.3363636363636401E-3	0.71643824113149002	8.8194762366280696E-2
Homologous Recombination	0.50554809745293205	1.20866666666667E-2	1	0.484438657760622
Hr Repair of Replication-Independent DSB	0.50531366475809203	5.6503125000000001E-2	1	0.77174554869186096
Mismatch Repair	0.50598378501380303	0	1	8.8194762366280696E-2
Non-Homologous end Joining	0.504894879653494	2.77666666666667E-2	1	0.58965802192687899
Nucleotide Excision Repair	0.501584339859246	0.169992307692308	1	0.61334063555743201
Processing of DNA DSB ends Recruitment of Repair and Sig. Proteins	0.50333916390641398	0.13075263157894701	1	0.606475830078125
Cell Cycle Checkpoints	0.50759678083974302	0	5.1479472984140802E-2	7.1534628503255401E-6
Cell Cycle, Mitotic	0.50601958008635695	0	0.24376116759906499	2.3879172921340299E-8
Cyclins and Cell Cycle Regulation	0.51627741079296197	0	0.76287142152699705	0.111346905048077
G1/S DNA Damage Checkpoints	0.50940057807121497	0	3.3139402134647902E-2	1.4898792823542999E-5
G2/M checkpoint	0.50944998962144705	0	1.9551633498738698E-2	4.1888088604642003E-6
Mitotic M-M/G1 Phases	0.50099677045647095	2.1827272727272699E-2	0.62741474615781001	0.606475830078125
Mitotic Spindle Checkpoint	0.50778961700134395	0	0.24982954806687499	8.9478832950449902E-8
Rb Tumor suppressor/Check. P. Sign. in Response to Damage	0.50652520264627099	1.8865E-2	0.65961656328433804	0.30146484374999999
Regulation of Mitotic Cell Cycle	0.50947655382896395	0	1.1201759629347699E-2	1.29452060519265E-6
S Phase	0.50860520971076195	0	1.9551633498738698E-2	3.6616506530362202E-7
Apoptosis - Homo sapiens (human)	0.50834596538779897	1.63333333333333E-3	0.25335083300795203	1.1718835681676899E-2
Apoptotic signaling in response to DNA Damage	0.50118677650575805	2.1827272727272699E-2	0.76287142152699705	0.60837883847817598
Caspase Cascade in Apoptosis	0.510717191922435 03	0	1.9551633498738698E-2	7.1534628503255401E-6
Death Receptor Signalling	0.49855791261593901	0.35402499999999998	0.66170893600854397	0.606475830078125
Extrinsic Pathway for Apoptosis	0.50749916035760101	1.8865E-2	0.76287142152699705	0.53791137695312496
Granzyme a Mediated Apoptosis Pathway	0.50760701602716196	3.1473076923076902E-2	0.76287142152699705	0.30146484374999999
Induction of apoptosis through dr3 and dr4/5 death receptors	0.50287689690855697	1.44277777777778E-2	1	0.584051891411837
Intrinsic Pathway for Apoptosis	0.50069006078548395	6.9639393939393904E-2	0.65961656328433804	0.77887676412990603
Regulation of Apoptosis	0.50264917205023496	0.12687027027027001	0.76287142152699705	0.68065780248397401
TNF Receptor Signaling Pathway	0.509926673666934	1.316875E-2	0.34886520245989699	7.8955078124999994E-2
tnfr1 Signaling Pathway	0.50280200048349399	2.4286956521739098E-2	0.86977642579793102	0.502129447118141
tnfr2 Signaling Pathway	0.50264917205023496	0.12687027027027001	0.76287142152699705	0.68065780248397401

Table 12: Details of data analysis of GSE22600 series using PATHChange. Consensus pathways are marked in light gray color.

Ovarian Cancer - GSE22600				
Pathway	Activity	Bootstrap	Fisher	Wilcoxon
Canonical NF-kappaB	0.49640749414289098	1	0.67574340887449302	0.56387246621621601
IL10 Anti-inflammatory Signaling Pathway	0.49518655454967703	1	1	0.49825642071664 3
IL6-mediated signaling events	0.49327287943884401	1	0.89713490555270303	0.29770339768508403
IL8- and CXCR1-mediated signaling events	0.51213430588500097	1	2.5698065772656502E-4	6.5626502037048401E-3
IL8- and CXCR2-mediated signaling events	0.50422213295843699	1	1.8961358219181199E-3	0.244546548920442
mtor signaling	0.49792028919056203	1	0.109584744684753	0.905262388987779
Nemo	0.49373711069732201	1	1	0.244546548920442
NF-kappaB pathway	0.49114747093050198	1	0.72691181615582001	5.9955896948451097E-2
nf-kb signaling	0.49960715014669899	1	0.21154665614961701	0.69198676215277799
p53	0.49597551720577199	1	0.58543413815863299	0.44924045522366801
Regulation of p38-alpha and p38-beta	0.50525161918775197	1	3.2036533180563901E-3	0.244546548920442
Replicative Senescence	0.50799776266408103	1	0.16353608805585801	0.56387246621621601
RIG-I	0.50348963346686704	1	0.12967823283633201	0.56908828333804495
SASP	0.51250537352401004	1	8.7461096988564603E-4	0.198291895064441
TGF	0.49622009032124798	1	0.49886083927865699	0.29355153564263398
TNF	0.50242706426122097	1	0.19690834149018599	0.60918890717496099
ATM signaling	0.48697284182806599	1	0.762321516469047	0.244546548920442
ATR signaling	0.47819421558479802	1	6.0306216034315997E-2	5.6802855397109101E-7
Base Excision Repair	0.48841539408968099	1	1	2.4695289762396599E-2
Double-Strand Break Repair	0.47915215428535302	1	0.14612120497586101	6.7970969460227296E-4
Fanconi Anemia Pathway	0.48323889186300301	1	0.84052753661545898	4.8387789592037296E-3
Homologous Recombination	0.47872124031843799	1	0.54445819058728295	2.29117719497543E-3
Hr Repair of Replication-Independent DSB	0.48036338217346802	1	0.14612120497586101	2.99072265625E-3
Mismatch Repair	0.48889186417198999	1	0.77862854396805303	2.5522009948263499E-3
Non-Homologous end Joining	0.47707493240892501	1	0.109584744684753	6.7970969460227296E-4
Nucleotide Excision Repair	0.479931058573978	1	0.86913309577687803	6.8929036458333301E-2
Processing of DNA DSB ends Recruitment of Repair and Sig. Proteins	0.50689314019555298	1	0.49886083927865699	0.69198676215277799
Cell Cycle Checkpoints	0.48276710306168202	1	1.1969013334781701E-5	8.1203052671710306E-12
Cell Cycle, Mitotic	0.48287451626883798	1	7.5719354485952595E-7	8.6575955370625905E-26
Cyclins and Cell Cycle Regulation	0.50718968685527199	1	0.19690834149018599	0.76146399456521696
G1/S DNA Damage Checkpoints	0.49027010401578403	1	4.3262553306693E-2	5.2710324592948997E-5
G2/M checkpoint	0.49083917560811302	1	0.21154665614961701	1.19735621851904E-4
Mitotic M-M/G1 Phases	0.48957924272352799	1	0.32408877092090899	3.1494209844638303E-5
Mitotic Spindle Checkpoint	0.47854875613095899	1	7.1796067974101495E-7	5.7848209217229198E-20
Rb Tumor suppressor/Check. P. Sign. in Response to Damage	0.477662898584921	1	1.42709015789144E-2	2.13400522867839E-3
Regulation of Mitotic Cell Cycle	0.48718270243195799	1	1.76918243560869E-3	1.59423017401782E-7
S Phase	0.48702924012981902	1	2.0103517719536002E-2	1.6424976062272002E-8
Apoptosis - Homo sapiens (human)	0.50066432732138899	1	9.7851256840465001E-2	0.865678125992417
Apoptotic signaling in response to DNA Damage	0.50284568546224095	1	9.2237543720406594E-3	0.56387246621621601
Caspase Cascade in Apoptosis	0.493096690563051	1	0.40992298413656297	2.29117719497543E-3
Death Receptor Signalling	0.50410378794969402	1	9.7851256840465001E-2	0.865678125992417
Extrinsic Pathway for Apoptosis	0.49492757829893202	1	0.75439895782048805	0.58265412159455099
Granzyme a Mediated Apoptosis Pathway	0.490671556476102	1	1	0.33655598958333299
Induction of apoptosis through dr3 and dr4/5 death receptors	0.50484935322941904	1	1.2336881000168801E-4	0.244546548920442
Intrinsic Pathway for Apoptosis	0.49853120520192401	1	0.23071837260686501	0.58922939121862195
Regulation of Apoptosis	0.50543407342751601	1	0.38324959182266799	0.69198676215277799
TNF Receptor Signaling Pathway	0.49947149961323301	1	0.86913309577687803	0.55806809179097505
tnfr1 Signaling Pathway	0.49468802703118803	1	0.109584744684753	0.55806809179097505
tnfr2 Signaling Pathway	0.50543407342751601	1	0.38324959182266799	0.69198676215277799

ANEXO B – PATHChangeDat FUNCTION

```
PATHChangeDat <- function(eDat, DataSet, NumbSample, Genes, HistComp, hc,
    writeRDS, destDIR){
  DataSet <- toupper(DataSet)
  tf <- tempfile() ; td <- tempdir()

  GSENumb <- as.numeric(str_extract(DataSet, "[0-9]+"))
  a<- if(str_length(GSENumb)>=4){paste0("GSE", substr(GSENumb, 1, str_length(
      GSENumb)-3), "nnn")}else{paste("GSE", "nnn", sep="")}

  url <- paste("ftp.ncbi.nlm.nih.gov/geo/series",a, DataSet, "matrix", paste(
      DataSet, "series_matrix.txt.gz", sep="_"), sep="/")
  download.file(url, destfile=paste0("./", paste(DataSet, "series_matrix.txt.gz
      ", sep="_")), method="libcurl",mode = "wb")
  MatrixGeo<-read.table(paste(DataSet, "series_matrix.txt.gz", sep="_"), header
      = FALSE, sep = "\t", col.names = paste0("V",1:(NumbSample+1)), fill = TRUE,
       strip.white = FALSE)
  file.remove(paste(DataSet, "series_matrix.txt.gz", sep="_"))

  geo_accession<-as.matrix(MatrixGeo[str_detect(MatrixGeo[,1], "ID_REF"),])
  title<-as.vector(as.matrix(MatrixGeo[str_detect(MatrixGeo[,1], "!Sample_title
      "),]))

  description<- cbind(title=title, geo_accession=as.vector(geo_accession))[-1,]
  detectLevels <- str_extract_all(title[-1], regex("[a-z]+|[[:alnum:]]+\\.*[[:
      digit:]]*", TRUE))
  level <- vector()
  for (i in 1:length(detectLevels)){level[i] <- str_c(detectLevels[[i]],
      collapse = " ")}
  description[,"title"]<-level; level <- unique(level)

  if(HistComp==TRUE){
    combinations <- NamesComb <- list()
    count = 0
    repeat{
      HistologyComp <- readline(c("Choose tissues to compare [ENTER]"))
      count <- count +1
      print(level)
      Ctrl <- readline(c("Please, choose the control groups: "))
```

```
    Control<-print(as.matrix(description[str_detect(description[,"title"], Ctrl
        ),]))
    print(level)
    Exp <- readline(c("Please, choose the Experimental groups: "))
    Experiment <- print(as.matrix(description[str_detect(description[,"title"],
         Exp),]))
    combinations[[count]] <- rbind(Control, Experiment)
    NamesComb[[count]] <- paste(c(Ctrl, Exp), sep="", collapse=" ")
    if(response <- readline(c("Would you like to compare another tissues? (yes/
        no) ")) == "no")break;
  }}else{Control <- print(as.matrix(description[str_detect(description[,"title
      "], hc[1]),]));
  Experiment <- print(as.matrix(description[str_detect(description[,"title"],
      hc[2]),]));
  combinations <- list(rbind(Control, Experiment));
  NamesComb <- paste(hc, collapse=" ")}


GenesSet<- read.table(Genes, header=TRUE)
eDat<-read.table(eDat, header = TRUE, sep = "/")
data <- merge(eDat, GenesSet, by.x ="Symbol", by.y ="ApprovedSymbol")
data <- aggregate(data[,3:dim(data)[2]], by = list(Symbol=data$Symbol), FUN =
    function(x) mean(as.numeric(as.character(x))))

MeanData <- list()
for(k in 1:length(combinations)){
  colnames(data) <- c("Symbol", as.vector(geo_accession)[-1])
  comb <- as.character(combinations[[k]][,"geo_accession"])
  dat <- matrix(0, dim(data)[1], length(comb))
  for(i in 1:length(comb)){
    for(j in 2:length(data)){
      if(colnames(data)[j]==comb[i]){dat[,i] <- data[,j]}
    }
  }
  MeanData[[k]]<-cbind.data.frame(Symbol=data[,1], cbind(Control=rowMeans(dat
      [,1:dim(Control)[1]]), Experiment=rowMeans(dat[,-c(1:dim(Control)[1])])))
}
names(MeanData)<-NamesComb
list.save(MeanData, file.path(td,"MeanData.rds"))
if(writeRDS==TRUE){list.save(MeanData, paste(destDIR, "MeanData.rds", sep="/")
    )}
```

```
  return(MeanData)
}
```

ANEXO C – PATHChangeList FUNCTION

```
PATHChangeList <- function(filePathway, writeRDS, destDIR){
  Pathway <- read.table(filePathway, header=T)
  colnames(Pathway)<-c("Pathway", "ApprovedSymbol")
  unique.path <- as.character(unique(Pathway$Pathway))

  path <- list()
  for (i in 1:length(unique.path)){
    path[i] <- list(Pathway[Pathway$Pathway == unique.path[i], ])
  }
  list.save(path, file.path(tempdir(),"path.rds"))
  if(writeRDS==TRUE){list.save(path, paste(destDIR, "path.rds", sep="/"))}
  return(path)
}
```

ANEXO D – PATHChange FUNCTION

```
PATHChange <- function(path, MeanData, writeCSV, writeRDS, destDIR){
  p.value<-list()
  for(j in 1:length(MeanData)){
    k=0
    result<-matrix(0,nrow=length(path), ncol=5,
                   dimnames = list(NULL, c("Pathway","Activity", "Bootstrap", "
                       Fisher", "Wilcoxon" )))
    repeat{
      k=k+1
      MeanData[[j]] <- cbind("Symbol"=MeanData[[j]][,1],log(MeanData[[j
          ]][,2:3]+1))
      Genes.path<-merge(MeanData[[j]],path[[k]],by.x=c("Symbol"), by.y=c("
          ApprovedSymbol"))[,-4]

      ################################################################
      ###                      BOOTSTRAP
      ################################################################
      nBoot <- 10000
      sample.e <- matrix(0, dim(path[[k]])[1], nBoot)
      sample.c <- matrix(0, dim(path[[k]])[1], nBoot)
      sample <- array(0,dim=c(dim(path[[k]])[1], 3, nBoot))
      for (i in 1:nBoot){
        sample[,,i] <- as.matrix(MeanData[[j]][sample(nrow(MeanData[[j]]), dim(
            path[[k]])[1], replace=TRUE), ])
        sample.e [,i] <- as.numeric(sample[1:dim(path[[k]])[1],2,i])
        sample.c [,i] <- as.numeric(sample[1:dim(path[[k]])[1],3,i])
      }

      exp <- Genes.path$Experiment
      ctrl <- Genes.path$Control
      activ <- function(exp,ctrl){                    # Activity expression
        nboot <- sum(exp)/(sum(exp)+sum(ctrl))
        nfisher <- exp/(exp+ctrl)
        return(c(nboot, nfisher))
      }
      n.1 <- as.matrix(activ(exp,ctrl))

      a <- matrix(0,dim(sample.e)[1]+1,nBoot)
```

```
for(i in 1:nBoot) {
  a[,i] <- activ(sample.e[,i],sample.c[,i])
}

activity <- sum(a[1,]>n.1[1,])/nBoot

################################################################
###                  Wilcoxon test
################################################################

Wilcoxon <- wilcox.test(exp, ctrl, paired=TRUE)$p.value

################################################################
###                   Fisher test
################################################################

Genes.out.path <- data.frame(Symbol=setdiff(MeanData[[j]][,1], Genes.path
    [,1]))
Genes.out.path <- merge(MeanData[[j]],Genes.out.path,by.x=c("Symbol"), by.y
    =c("Symbol"))

exp <- Genes.out.path$Experiment
ctrl <- Genes.out.path$Control

Out.Pathway <- activ(exp,ctrl)

Path.decrease.n <- sum(n.1[-1]<0.5); Path.increase.n <- sum(n.1[-1]>0.5)
OutPath.decrease.n <- sum(Out.Pathway[-1]<0.5); OutPath.increase.n <- sum(
    Out.Pathway[-1]>0.5)

cont.table.Act <- matrix(c(Path.increase.n, OutPath.increase.n, Path.
    decrease.n, OutPath.decrease.n),
                      2,2,dimnames = list(c("In Pathway", "Out Pathway"),
                                          c("Increase Expression", "Decrease
                                              Expression")))

Fisher <- fisher.test(cont.table.Act)$p.value


result[k,]<-c(as.character(unique(path[[k]]$Pathway)),n.1[1], activity,
    Fisher, Wilcoxon)
```

```
      if(k==length(path)) break;
    }
    p.valFDR <- cbind(result[,c(1,2)],apply(result[,-c(1,2)], 2, function(x) p.
        adjust(x, method="BH")))
    p.value[j]<-list(p.valFDR)
    if(writeCSV==TRUE){write.csv(p.value[j], file = paste(destDIR, paste(names(
        MeanData)[j], ".csv"), sep="/"), row.names=FALSE)}
  }
  list.save(p.value, file.path(tempdir(),"pValue.rds"))
  if(writeRDS==TRUE){list.save(p.value, paste(destDIR,"pValue.rds", sep="/"))}
  return(p.value)
}
```

ANEXO E – PATHChangeVenn FUNCTION

```
PATHChangeVenn <- function(p.value, p, writePDF, destDIR){
  Boot <- Fisher <- Wilc <- list()
  for (j in 1:length(p.value)){
    Boot[[j]] <- which(as.numeric(p.value[[j]][,"Bootstrap"])<=p|as.numeric(p.
        value[[j]][,"Bootstrap"])>=1-p)
    Fisher[[j]] <- which(as.numeric(p.value[[j]][,"Fisher"])<=p)
    Wilc[[j]] <- which(as.numeric(p.value[[j]][,"Wilcoxon"])<=p)
    grid.newpage()
    venn.plot <- paste("venn.plot", j, sep=".")
    venn.plot.j<- draw.triple.venn(area1=length(Boot[[j]]),area2=length(Fisher[[
        j]]),area3=length(Wilc[[j]]),
                                   n12 = length(grep(TRUE, (Boot[[j]]%in%Fisher[[j
                                       ]]))),
                                   n23 = length(grep(TRUE, (Fisher[[j]]%in%Wilc[[j
                                       ]]))),
                                   n13 = length(grep(TRUE, (Boot[[j]]%in%Wilc[[j]])
                                       )),
                                   n123 = length(intersect(Boot[[j]],intersect(
                                       Fisher[[j]],Wilc[[j]]))),
                                   category = c("Bootstrap", "Fisher", "Wilcoxon"),
                                   lty = "blank",
                                   margin = 0.05,
                                   cex = 1.5,
                                   cat.cex = 1.5,
                                   fill = c("skyblue","pink1","mediumorchid"))
    if(writePDF==TRUE){pdf(paste0(destDIR,"/",(paste(paste("VennDiagram", j, sep
        ="_") ,".pdf", sep=""))))}
    grid.draw(venn.plot.j);
    dev.off()
  }
  return(grid.draw(venn.plot.j))
}
```