T.C.
İSTANBUL ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

M. Sc. THESIS

DETERMINING NOVEL TARGET GENES IN WNT/β-CATENIN SIGNALING PATHWAY USING MACHINE LEARNING

Cemre KEFELİ

Moleküler Biyoloji ve Genetik Anabilim Dalı

Moleküler Biyoloji ve Genetik Programı

Supervisor
Dr.Öğr.Üyesi Andrés Octavio ARAVENA DUARTE

Haziran, 2022

İSTANBUL

Bu çalışma 29.06.2022 tarihinde aşağıdaki jüri tarafından Moleküler Biyoloji ve Genetik Anabilim Dalı Moleküler Biyoloji ve Genetik Programında Yüksek Lisans Tezi olarak kabul edilmiştir.

Tez Jürisi

Dr.Öğr.Üyesi Andrés Octavio ARAVENA DUARTE (Danışman)
İstanbul Üniversitesi
Fen Fakültesi

Prof. Dr. Şule ARI
İstanbul Üniversitesi
Fen Fakültesi

Dr.Öğr.Üyesi Necla BİRGÜL
Boğaziçi Üniversitesi
Fen Fakültesi

- **İntihal Programı Beyanı**

20.04.2016 tarihli resmi gazetede yayımlanan Lisansüstü Eğitim ve Öğretim Yönetmeliğinin 9/2 ve 22/2 maddeleri gereğince; Bu Lisansüstü teze, İstanbul Üniversitesi'nin abonesi olduğu intihal yazılım programı kullanılarak Fen Bilimleri Enstitüsü'nün belirlemiş olduğu ölçütlere uygun rapor alınmıştır.

- **Tezden Üretilmiş Yayınların Künye Bilgileri**

Kefeli C., Aravena A. Novel targets of the WNT/β-catenin pathway discovered using gene expression data mining. In: Proceedings of the 5th International Symposium on Bioinformatics (InSyB2021). F1000Research 2022, 11:60 (https://doi.org/10.7490/f1000research.1118906.1)

PREFACE

I could not have undertaken this journey without my beloved friends Salva Asma ALGILANI, Selen YAPMIŞ, Sevim Beyza GÜRLER, Gözde Ervin KÖLE, Özgür Evrim TURSUN, Kübra KOÇAK and Merve ATAN.

I would thank my lovely little potato sac, my cousin/sister Cansu KEFELİ for just existing and bringing joy into my life.

Special thanks to my friends İrem EYİOKUR and Doğucan YAMAN for inspiring, encouraging and supporting me to study in the computer science field and believing that I can do it.

I would like to express my deepest gratitude to my advisor Assist. Prof. Andres Octavio ARAVENA DUARTE. He is an amazing scientist, a modern day philosopher and a wonderful human being, and I am extremely grateful to be his protégé.

This endeavour would not have been possible without my family. My parents and my brother. Thank you for believing and supporting me. I will be forever indebted to you.

Haziran, 2022 Cemre KEFELİ

CONTENTS

LIST OF FIGURES

Page

LIST OF TABLES

Page

LIST OF ABBREVIATIONS

Abbreviations	Description
CART	: Classification and regression tree
BMP	: Bone morphogenetic proteins
FGF	: Fibroblast growth factor
CRD	: Cystein rich domain
HMG	: High mobility group box
RSPO	: R-spondin proteins
APC	: Adenomatous polyposis coli
FAP	: Familial adenomatous polyposis

ÖZET

YÜKSEK LİSANS TEZİ

WNT/BETA-CATENIN SINYAL YOLAĞINDA MAKINE ÖĞRENMESI ILE HEDEF GENLER BELIRLENMESI

Cemre KEFELİ

İstanbul Üniversitesi

Fen Bilimleri Enstitüsü

Moleküler Biyoloji ve Genetik Anabilim Dalı

Danışman: Dr.Öğr.Üyesi Andrés Octavio ARAVENA DUARTE

Wnt/β-catenin, evrimsel açıdan korunmuş, hücre bölünmesinde ve farklılaşmasında önemli rol oynayan bir sinyal yolağıdır (Logan ve Nusse 2004). Wnt sinyal yolağı, hücre proliferasyonu, kök hücre farklılaşması, primer eksen oluşumu ve embriyonik gelişimden yetişkin doku homeostazına kadar çok sayıda süreci düzenler. Bu sinyal yolağında meydana gelen aksaklıklar, farklı kanser türlerine, Alzheimer hastalığına sebep olmakta, ayrıca embriyonik gelişimde sorunlar meydana getirmektedir (Clevers 2006; Ferrari ve Inestrosa 2000).

Literatürde birden fazla Wnt sinyal yolağının varlığı gösterilmiştir. Bunlar kanonik ve kanonik olmayan Wnt yolakları şeklinde ayrılmışlardır. Bütün bu yolaklar Wnt ligandları tarafından tetiklenmektedirler. Wnt/β-catenin yolağı β-catenin molekülünü içerir ve kanonik Wnt sinyal yolağıdır. Bu çalışmada Wnt/β-catenin sinyal yolağına odaklananılmıştır ve takip eden bölümlerde kısaca Wnt sinyal yolağı şeklinde adlandırılacaktır.

Wnt genleri tüm metazoalarda görülmektedir. Memelilerde her biri farklı işleve sahip 19 adet Wnt geni bulunmaktadır. Bu genlerin ürünü olan Wnt proteinleri ligand gibi hareket ederek hücre yüzeylerindeki reseptörlere bağlanırlar. Wnt ligandları frizzled reseptörleri tarafından algılanır (Niehrs 2012). Wnt sinyal iletimi otokrin ya da parakrin şekilde olabildiği gibi çoğunlukla bu sinyaller kısa mesafeli olarak hareket ederler. Wnt proteinlerinin, diğer bir deyişle sinyallerinin tetiklemesiyle başlayan bu sinyal yolağı hedef genlerin transkripsiyonel kontrolü ile sona erer.

Wnt sinyali ile tetiklenmemiş normal bir hücrede sitozolde bulunan serbest β-cateninler yıkım kompleksi tarafından yıkılır. Yıkım kompleksi Axin, adenomatöz polipozis koli

(APC), glikojen sentaz kinaz β (GSK3-β), protein fosfataz 2A (PP2A) ve kazein kinaz 1 (CK1) moleküllerinden oluşmaktadır. CK1 ve GSK3-β molekülleri β-catenini fosforiller. Devamında β-catenin β-Trcp adlı molekül tarafından tanınır ve ubikitinlenir (C. Liu ve diğerleri 2002). Bu süreç β-cateninlerin proteozomal yıkımını gerçekleştirir. Wnt tarafından tetiklenmemiş bir hücrede β-catenin seviyeleri oldukça düşük tutulur.

Hücre Wnt sinyali ile tetiklendiğinde yıkım kompleksi dağılır ve işlevsiz hale gelir. Bu durum sitozolde serbest β-cateninlerin birikmesine sebep olur (Cadigan ve Nusse 1997). Serbest β-cateninler nukleusa hareket eder ve nukleusta TCF/LEF transkripsiyon faktörleri ile etkileşime geçer (Polakis 2000). Tetiklenmemiş bir hücrede TCF/LEF transkripsiyon faktörleri korepresör groucho ile birlikte hedef genlerin anlatımını baskılar (Hoppler ve Kavanagh 2007). β-catenin TCF/LEF transkripsiyon faktörlerine bağlanarak birlikte Wnt hedef genlerinin transkripsiyonunu aktive ederler.

β-catenin ve TCF/LEF kompleksi Wnt hedef genlerinin yakınlarında bulunan cis-regülatör bölgelere bağlanırlar (Nusse 2012). Bu tip genlere doğrudan Wnt sinyali hedef geni olarak adlandırılır. Anlatımı Wnt sinyal yolağı tarafından dolaylı olarak etkilenen genlere ise dolaylı hedef gen adı verilir. Doğrudan hedef genlerin üstünde çok çalışılmış olmasına karşılık dolaylı hedef genler dair bulunan bilgiler azınlıktadır. Bu durum bu tip hedef genlerin tespitindeki zorluklardan kaynaklanmaktadır. Wnt hedef genleri dokuya veya gelişim evrelerine özeldir. Bir hücre tipinden diğer hücre tipine değişebilmektedirler (Boonekamp ve diğerleri 2021; Nakamura ve diğerleri 2016; Söderholm ve Cantù 2021).

Hedef genler çeşitli deneyler ile keşfedilir veya doğrulanırlar. Microarray, RNA-seq ve qPCR yöntemleri hedef genlerin anlatım düzeylerini ölçmede; ChIP, Raportör gen yöntemi ve EMSA TCF/LEF bağlanma bölgesinin varlığını saptamada kullanılır.

TCF/LEF bağlanma bölgesine sahip her gen hedef olmayabilir. Bu transkripsiyon faktörüne ait bağlanma bölgesi içeren bazı genler, β-catenin üretiminin tetiklendiği deneylerde yeterli değişim göstermeyebilirler. Bu durum henüz bilinmeyen interaksiyonlar sebebiyle gerçekleşebilir. Bu sebeple sadece tüm genom üzerinde bağlanma bölgesi aramak hedef gen tespitinde eksik kalmaktadır.

Wnt sinyal yolağı bağırsaklardaki homeostasinin korunması adına hücre bölünmesini ve farklılaşmasını yöneten önemli faktörlerden biridir. Wnt sinyal iletimindeki bozukluklar bu hızla değişen sistemde tümörlerin oluşmasına yol açabilir. İnce bağırsak villus ve kript adı verilen yapılardan oluşur. Villuslar besinlerin emiliminden sorumlu iken kriptlerde ise yeni hücreler üretilir. Wnt sinyal yolağı bu yeni hücre üretiminde rol alan sinyal yolaklarından biridir.

Bu çalışma, insanda kolorektal kanser bağlamında Wnt sinyal yolağında yeni hedef genleri tanımlamayı amaçlamaktadır. Daha önceki çalışmalar, Wnt sinyal yolağındaki hedef genlerin anlatımının dokuya ve hastalığa bağlı olarak değişebildiğini öne sürmektedir. Bu nedenle, hedef genler dokuya özgü olabilmektedir. Bu fikri takiben bu çalışmada, Wnt sinyali yolağı kolorektal kanser bağlamında çokça çalışıldığı için odak noktası olarak kolorektal kanseri seçilmiştir. Bu alanda gen ekspresyon verileri diğer doku ve kanser türlerine göre daha fazladır. Ayrıca, bu bağlamda çok sayıda deneysel olarak doğrulanmış hedef gen bulunmaktadır.

Bu tezde, genomik ve transkriptomik verileri kullanarak bir makine öğrenmesi modeli eğitilmiştir ve bu model yeni hedef genler tahmin etmek için kullanılmıştır. Klasik makine öğrenmesi yöntemleri pozitif ve negatif örnekler kullanır, ancak bu bağlamda yalnızca pozitif

örnekler mevcut bulunmaktadır. Bu çalışmada bu duruma özel bir yöntem kullanılmıştır. Bu çalışmada Hödar ve diğerleri (2010) 'ın bu durum için kullandığı yöntem genişletilerek yeni Wnt sinyal yolağı hedef genlerinin tespitinde kullanılmıştır. Eğitim için, hakemli dergilerde yayınlanmış, bazı kolorektal kanser hücre hatlarından elde edilen transkriptomik diferansiyel anlatım verilerini kullanılmıştır. Ön araştırmada kolorektal kanser hücre hatlarında Wnt/β-katenin yolağının uyarıldığı gen anlatımı deneyleri (Microarray ve RNA-seq) derlenmiştir (Tablo 2.2). Bu veri kümeleri makine öğrenmesi modelini eğitmek için kullanılmıştır. Literatür araştırmasında kolorektal kanser ile ilişkili 93 adet doğrulanmış hedef gen tespit edildi ve bu genler pozitif örnek olarak kullanılmıştır. Bu bilgiler, daha önce aday olarak olarak gösterilmemiş yeni hedef genlerin tespiti için kullanılmıştır.

Makine öğrenmesi, modellere dayanarak verileri sınıflandırılabilen ve üzerinde tahminlerde bulunabilen algoritmalar oluşturulması ve test edilmesi gibi konulara odaklanan bir Yapay Zeka dalıdır. Modelleri "eğitmek" için mevcut verileri kullanılır, ardından bu modeller tahminler yapmak ve yeni vakaları sınıflandırmak için kullanılır. Makine öğrenmesi modelleri, başka yollarla keşfedilmesi zor olabilecek örüntüleri "görebilir" ve ilgilendiğimiz konular hakkında bazı içgörüler sağlayabilir.

Genel anlamda veriler, her biri özellikler veya değişkenler olarak da adlandırılan çeşitli nitelikler içeren bir dizi bireysel gözlemden oluşur. Makine öğreniminde, gözlemlere dayalı tümevarımsal bir akıl yürütme yapmak için değişkenler veya nitelikler kullanılır (H. Liu ve Motoda 2013). Veri, sayısal ya da kategorik değerler olabilen özelliklerin bir koleksiyonudur. Bu küme ayrıca gözlemin sınıfını belirten bir etiket özelliğini de içerir. Eğitilen sınıflandırıcı, eğitim sırasında kullanılan özelliklere dayalı olarak yeni gözlemlerin etiketlerini tahmin eder.

Bu çalışmadaki sınıflandırma problemi klasik sınıflandırma problemlerine kıyasla bir takım farklılıklar göstermektedir. Bir dizi pozitif ve negatif örnek yerine, gerçekte daha fazlasının olduğu düşünülen yalnızca pozitif örneklerin bir alt kümesi mevcut bulunmaktadır. Pozitif örnekler, literatür araştırmasında derlenen, Wnt sinyal yolağının deneysel olarak doğrulanmış hedef genleridir. Gerçek anlamda negatif örnekler bulunmamaktadır. Sadece pozitif örnekler bulunmakta ve genlerin geri kalanının Wnt sinyal yolağı hedef geni olup olmadığı ise belirsizdir.

Bu sorunun üstesinden gelebilmek için doğrulanmış hedef genleri pozitif örnek olarak, negatif örnekleri ise doğrulanmış hedef genlerin dışında kalan tüm bilinen genler içerisinden çeşitli büyüklükte alt kümeler olarak seçilmiştir. Buradaki pozitif ve negatif adlandırması eğitime dair verinin sınıfını belirtmeye yönelik bir terim olarak kullanılmıştır. Herhangi bir biyolojik anlam içermemektedir. Ortaya koyduğumuz hipoteze göre bilinmeyen hedef genlerin sayısı çoğunlukta değildir. Bu hipotez, rastgele bir örneğin muhtemelen yalnızca hedef olmayan genleri içereceğini, ancak bazen bunlardan bazılarının gerçek hedefler olabileceğini fikrini desteklemektedir.

Bu çalışmada rastgele örnekleme kullanıldığı için varyasyonu kontrol altına almak adına her biri rastgele seçilmiş negatif örnekler içeren (her bağımsız sınıflandırıcı için farklı negatif örnekler) 500 ya da 1000 adet birbirinden bağımsız sınıflandırıcı havuzu eğitilmiştir. Havuzdaki her sınıflandırıcı, hangi genlerin "pozitif" örneklerin özelliklerini paylaştığına bağımsız olarak karar verir. Deneyin sonunda, her gen, onu hedef gen olarak ilan eden sınıflandırıcıların sayısına karşılık gelen bir puan alır. Bu puan 0 ile 500 (veya havuzdaki sınıflandırıcı sayısına bağlı olarak 1000) arasında bir sayıdır. Daha yüksek puanlar, daha yüksek bir hedef olma olasılığına karşılık gelir.

İzlenilen yöntemin doğruluğunu test etmek adına negatif kontrol kullanılmıştır. Bu sefer tüm genlerin içerisinden rastgele "sahte" pozitif örnekler seçilmiştir. Önceki eğitim aşamaları, pozitif örneklerin rastgele seçilmiş olması haricinde aynı şekilde tekrar edilmiştir. Bu kontrolü, yöntemin ortaya koyduğu sonuçların rastgele olmadığını, sonuçların sadece doğrulanmış hedef genler örnek olarak kullanıldığında ortaya çıkacağını göstermek için gerçekleştirilmiştir.

Makine öğrenme sürecini doğrudan etkileyen bazı parametreler bulunmaktadır. Bu tip parametrelere hiper parametre adlandırılmaktadır. Optimal hiper parametreleri bulmak için sınıflandırıcıları farklı hiper parametre değerleriyle eğitilmiştir. Örnek boyutu hiper parametresi seçilen negatif örneklerin sayısını belirtmektedir. Sayıları 800, 1000, 1200, 1500, 2000, 3000, 5000 ve 8000 olan örnekler farklı eğitim turlarında test edilmiştir. Elimizdeki veri küçük sayıdaki pozitif örneklere (N=93) karşılık büyük sayıda negatif örnekler (geriye kalan tüm genler N≈20.000) içerdiğinden çok sayıda negatif örnek sonuçları saptırabilir. Bu durumu telafi etmek adına yanlış sınıflandırma için farklı penalizasyonlara (bazen ağırlıklar da denir) sahip bir kayıp fonksiyonu kullanılmıştır. Farklı eğitim turlarında 1, 5, 10, 20, 50, 100 ve 200 değerlerini penalizasyon olarak test edilmiştir. Belirtilen yöntemler kullanılarak eğitimler 4 turda gerçekleştirilmiştir.

Birinci eğitimde Hödar ve diğerlerinin (2010) çalışmasında kullandığı veri kullanılmıştır. İkinci eğitimde yine aynı çalışmada kullanılan transkripsiyon faktörü bağlanma bölgesi parmak izi matrisi ve yapılan ön araştırmada derlenen doğrulanmış Wnt sinyali hedef genleri kullanılmıştır. Üçüncü ve dördüncü eğitimlerde derlenen doğrulanmış hedef genleri ve sırasıyla analiz edilen RNAseq ve RNAseq-Microarray verileri kullanılmıştır. Her bir oturumda, hiper parametre kombinasyonuna bağlı olarak en iyi performans gösteren sınıflandırıcılar seçilmiştir. Toplamda yeni hedef genler belirten 4 adet liste elde edilmiştir.

Bu çalışmanın sonucunda pek çok ilginç aday hedef genler belirlenmiştir. Bunlardan en önemli iki tanesi PTCH1 ve GLI3 genleridir. Bu genler Hedgehog sinyal yolağının 2 önemli bileşenini oluşturmaktadır. Özellikle yakın zamanda yapılan bir çalışma PTCH1 geninin kolon spesifik Wnt sinyali hedef geni olduğunu öne sürmektedir. Bu durum Wnt sinyal yolağının bu genlerin transkripsiyonunu kontrol ederek iki sinyal yolağı arasında kolorektal kanser bağlamında bir iletişimin gerçekleşebileceğini göstermektedir.

CDH1, FN1, VIM, TBX3, TNFRSF19, GLUT3 tespit ettiğimiz aday Wnt sinyali hedef genleri içerisinde yer almaktadır. Bu genler kolorektal kanser hücrelerinde doğrulanmadıkları için doğrudan Wnt sinyali hedef genleri listemizde yer almamakla birlikte başka dokularda doğrudan Wnt sinyali hedef geni olarak tespit edilmişlerdir. Yapılan bir kaç farklı kolorektal kanser araştırmasında tespit ettiğimiz hedef genler ile bu çalışmalarda sözü edilen genler arasında örtüşme tespit edilmiştir. Belirlediğimiz Wnt sinyal yolağı hedef genlerinin çeşitli biyolojik deneyler ile doğrulanması gerekmektedir. Bulgularımız kullandığımız metod ve verinin doğruluğunu destekler niteliktedir.

Haziran 2022, 144 sayfa.

Anahtar kelimeler: WNT sinyal yolağı, hedef genler, kolorektal kanseri, makine öğrenimi, CART

SUMMARY

M.Sc. THESIS

DETERMINING NOVEL TARGET GENES IN WNT/β-CATENIN SIGNALING PATHWAY USING MACHINE LEARNING

Cemre KEFELİ

Istanbul University

Institute of Graduate Studies in Sciences

Department of Molecular Biology and Genetics

Supervisor: Assist.Prof. Andrés Octavio ARAVENA DUARTE

The Wnt signalling pathway is a driving force of proliferation and differentiation. Aberrant behaviour in this pathway may lead to several types of cancers and Alzheimer's disease. This pathway controls the transcription of target genes via modulating the presence of Beta-catenin in cytosol, which triggers the TCF/LEF transcription factor. We found 93 target genes identified experimentally in the colorectal cancer context, and new target genes are constantly being discovered. This study aims to identify novel target genes of the Wnt/Beta-catenin signalling pathway using a machine learning approach.

We analysed several publicly available Microarray and RNA-seq experiments and used the differential gene expression data to represent the genes. We used the experimentally validated target genes as "positive" examples in training. We chose the "negative" examples randomly from the rest of the genes. We trained pools of 1000 independent classifiers using the Classification and Regression Tree (CART) method. Then each trained classifier was used to assign a "positive" or "negative" label for each gene. The number of times each gene is classified as "positive" is a score that can be tested using the Fisher method. Thus, we found a set of putative target genes having an expression pattern very similar to known target genes. The pool of trained classifiers predicted 144 putative novel target genes. Some of the highest scoring genes are PTCH1, GLI3 and SOX4. The first two predictions, PTCH1 and GLI3, are important components of the Hedgehog Signalling. This suggests a possible interplay between Wnt and Hedgehog signalling in colorectal cancer. In parallel to our study, experimental researchers have reported that PTCH1 is a colon specific Wnt target.

We present a bioinformatic method that can be used to predict target genes of the canonical

Wnt signalling pathway, and eventually other pathways, based only on gene expression data and a sample of experimentally validated targets. This method narrows the set of genes that should be experimentally validated. Moreover, some of our predictions have already been validated by other studies.

June 2022, 144 pages.

Keywords: Wnt signalling, target genes, colorectal cancer, one-class machine learning, CART

1. INTRODUCTION

This thesis deals with discovering novel target genes in the Wnt signaling pathway.

Cells sense the signals in their environment. After processing the external stimuli, cells produce an appropriate response to these signals. The external signals are received by proteins, called receptors, that lie at the cell's surface, and are transmitted to the target through a series of molecular interactions and changes. This information transmission and processing is called a *"signaling pathway"*. Signaling pathways consist of specific molecules, including proteins (receptors, transcription factors, hormones), and metabolites (molecule complexes, ligands, hormones). These molecules interact with each other following a path and transmit signals to one or more targets. Transmission and processing of these signals are important for the appropriate functioning of the cells because these signals determine the cell fate. Aberrations in signaling pathways can lead to several diseases or disorders.

1.1. WNT SIGNALING PATHWAY

Wnt signaling is an evolutionarily conserved pathway. There are seven known signaling pathways that are evolutionarily conserved and use similar mechanisms to control the organism's development in metazoa. These pathways are responsible for cell differentiation through the control of target gene expressions (Babonis and Martindale, 2017).

The cascades of reactions in the Wnt signaling pathway regulate the numerous processes ranging from cell proliferation, stem cell differentiation, primary axis formation, embryonic development, to adult tissue homeostasis (Logan and Nusse, 2004). Misregulations in this signaling pathway can cause aberrations in development, and diseases such as different types of cancers and Alzheimer's disease (Ferrari and Inestrosa, 2000; Clevers, 2006).

The scientific literature describes several Wnt signaling pathways, traditionally called canonical and non-canonical. The naming indicates the chronological order of their discovery. All of these pathways are triggered by the Wnt ligands. The canonical Wnt signaling involves the molecule β-catenin, which is at the center of this signaling cascade. The non-canonical pathways (Wnt/Ca^{2+} and Wnt/planar cell polarity) are β-catenin independent. In this study, we will focus on the canonical Wnt/β-catenin signaling pathway. We will refer to the canonical Wnt/β-catenin signaling pathway as Wnt signaling from now on.

At the beginning of the Wnt signaling, there are Wnt proteins that act as Wnt signals. In the canonical pathway, there are target genes downstream of this signaling. Their activity is regulated by β-catenin and TCF/LEF family of transcription factors. These molecules are vital components of Wnt signaling.

1.2. IMPORTANCE OF WNT SIGNALING PATHWAY

Wnt signaling controls many important events in cells. Malfunctions in this pathway can lead to diseases such as cancer. Mutations in the Wnt signaling cascade elements, such as APC, Axin1, and β-catenin, are associated with several cancers (Logan and Nusse, 2004). β-catenin is an essential part of this pathway and an important target for cancer drug studies.

Development and tumorigenesis resemble one another in some aspects. In both phenomena, we see differentiation and proliferation. In a sense, they are two sides of the same coin. Therefore it is not surprising that viral oncogenes control developmental switches that cause such events. We can see this duality in healthy and misregulated Wnt signaling.

During embryogenesis, a single cell develops into an organism that has a differentiated body structure. This differentiation is carried out by molecules, called morphogenes, that form a concentration gradient specific to the tissue and results in differential gene expression, and determination of the particular features of the tissue. Wnt signals are also morphogenes.

Wnt signaling plays many important roles in several processes of embryogenesis. Logan and Nusse (2004) showed that knocking out *Wnt* genes in mice resulted in serious phenotypic defects. Many studies indicate that Wnt signaling is responsible for primary body axis formation in bilateral organisms, an embryonic process that defines the anteroposterior and dorsoventral body axes in many organisms (Molenaar *et al.*, 1996; Sokol, 1999; Kiecker and Niehrs, 2001; Petersen and Reddien, 2009; Alexander *et al.*, 2014).

In embryogenesis, Wnt signals trigger the differentiation of embryonic layers. When Wnt signals are present, they block ectodermal progenitor cells' ability to respond to fibroblast growth factors (FGF). Without the FGF response, these cells respond to bone morphogenetic proteins (BMP), and cells differentiate into the embryonic epidermis. When the Wnt signal is absent, ectodermal progenitor cells respond to FGF and block the BMP resulting in neurogenesis (Wilson and Hemmati-Brivanlou, 1995; Wilson *et al.*, 2001) (Fig. 1.1). Wnt signals also determine the differentiation of the embryonic epidermis. Cells responding

to Wnt signals differentiate into placode, which is the primary progenitor of hair follicles (Fig. 1.1). Additionally, several studies have reported that increased expression of β-catenin results in the formation of new hair follicles (Gat *et al.*, 1998; Noramly, Freeman and Morgan, 1999; Widelitz *et al.*, 2000; Millar, 2002). Furthermore, loss of function in β-catenin causes failure in the hair follicle formation (Huelsken *et al.*, 2001). It has also been reported that the TCF/LEF family of transcription factors have important roles in follicle development (Genderen *et al.*, 1994; Zhou *et al.*, 1995; Kratochwil *et al.*, 1996; DasGupta and Fuchs, 1999).

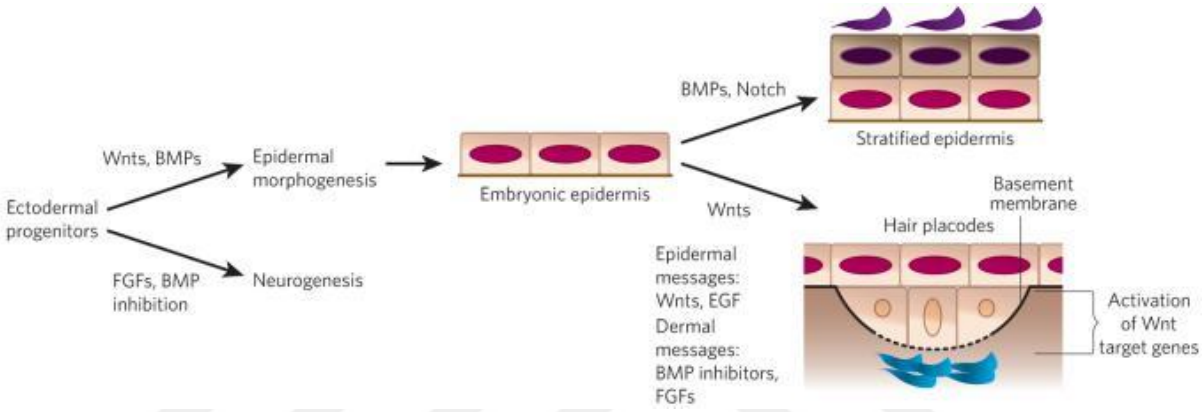

Figure 1.1: Wnt signaling controls several switches in embryonic development. Ectodermal progenitors respond to different switches/signals depending on the presence of a Wnt signal. When the Wnt signal is active, FGFs are blocked and cells respond to BMPs and differentiate into embryonic epidermis. Without the Wnt signals, cells respond to FGF, BMPs are blocked, and cells go through neurogenesis. Embryonic epidermis also responds to Wnt. Presence of Wnt promotes hair progenitors in embryonic epidermis (Fuchs, 2007).

Wnt signaling also regulates bone mass. Bone is a dynamic tissue that is constantly broken down and rebuilt by osteoclast and osteoblast cells. Osteoclast cells dissolve the surface of the bone and osteoblast rebuilds it. The bone mass homeostasis is the result of the action of these two types of cells. A receptor protein of the Wnt pathway, LRP5 is a key regulator of bone mass density. Depending on the type of mutation, LRP5 causes low bone density or increase in bone mass (Gong *et al.*, 2001; Boyden *et al.*, 2002; Wesenbeeck *et al.*, 2003; Ai *et al.*, 2005).

In adult tissues, Wnt signaling controls homeostasis and tissue renewal. These aspects of Wnt signaling are extensively studied in the intestinal epithelium. The intestines are covered with epithelial cells that absorb nutrients, and water and eliminate the waste that might harm the body. As a defense mechanism to this harsh environment, gut epithelium constantly renews itself via intestinal stem cells. Thus, mice lacking TCF4 —a member of the TCF/LEF

transcription factor family that controls the target gene expression in Wnt signaling— result in total absence of gut epithelial stem cells, demonstrating that Wnt signaling is essential for gut renewal (Korinek *et al.*, 1998). Adenovirus Dkk1 (Dickkopf-1) —a Wnt antagonist— causes a decrease in stem cell proliferation and loss of crypts in adult mice' intestines (Kuhnert *et al.*, 2003).

Wnt signaling is also important for the positioning of the differentiated cells in the intestines. Wnt signaling controls the positioning of Paneth cells through the Wnt target gene EphrinB (Batlle *et al.*, 2002).

1.3. HOW WNT SIGNALING WORKS

Wnt genes are present in all metazoans. There are 19 known *Wnt* genes in mammals, all with different functions. The products of these genes, the Wnt proteins, are exported and act as ligands to stimulate target cells by binding to receptors on the membrane. Although Wnt signaling can be autocrine or paracrine, most of the time they are short-range signals. Some works suggest that this is due to their lipid-modified nature, making it hard for them to move across the hydrophilic environment (Farin *et al.*, 2016). Several signaling reactions are carried inside the cells upon stimulation with Wnt proteins, leading to transcriptional control.

Wnt ligands are sensed by frizzled receptors,which are G protein-coupled, seven-transmembrane receptors (Niehrs, 2012). Wnt ligands bind to the cysteine-rich domain (CRD) of the frizzled receptor. This receptor is accompanied by co-receptors LRP5 and LRP6. These co-receptors belong to the LRP (transmembrane low-density lipoprotein receptor molecules) class of proteins.

In a normal cell that is not stimulated by a Wnt ligand, β-catenin is degraded by the destruction complex. This is composed by the tumor suppressor molecule Axin, adenomatous polyposis coli (APC), glycogen synthase kinase β (GSK3β), protein phosphatase 2A (PP2A), disheveled (DVL) and casein kinase (CK1). The molecules CK1 and GSK3β phosphorylate β-catenins. Subsequently, the β-catenins are recognized by β-Trcp, —an E3 ubiquitin ligase subunit— and ubiquitinated (Liu *et al.*, 2002). This process leads to proteasomal degradation of β-catenins. In a non-stimulated cell, β-catenin levels are kept relatively low (Logan and Nusse, 2004).

When a Wnt signal is present, it binds to the frizzled receptor. Frizzled recruits Disheveled

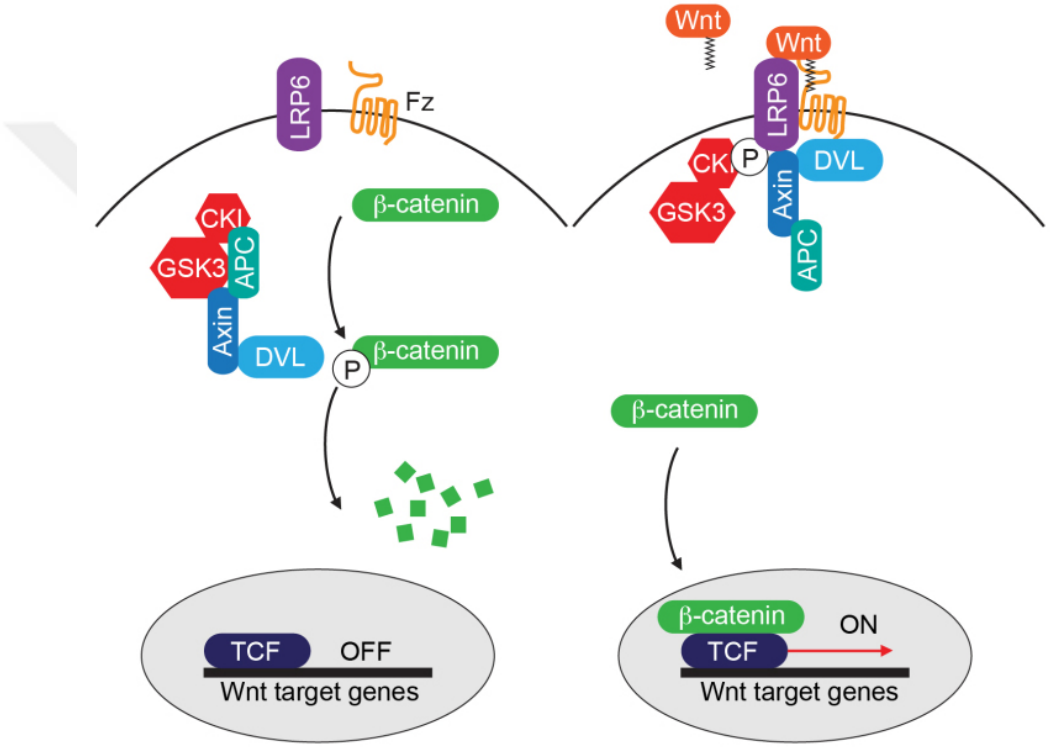

Figure 1.2: Wnt Signaling in General. Without the Wnt signal (on the left) the pathway is in an off state, and β-catenins tagged and degraded by the destruction complex. When a Wnt signal is present it is recognized by the frizzled and LRP5/6 receptors. The destruction complex is held and dismantled upon the stimulation. Free β-catenins travel to the nucleus, and interact with TCF. Then this complex regulates the transcription of the target genes.

(Dsh) to the membrane. This is followed by translocation of the destruction complex to the cell membrane, and the function of the complex is disrupted. This leads to the accumulation of β-catenin in the cytosol (Cadigan and Nusse, 1997). Free β-catenins are shuttled to the nucleus where they interact with TCF/LEF transcription factors (Polakis, 2000). In non-stimulated cells, these transcription factors, with corepressors groucho/TLE, repress the target genes (Hoppler and Kavanagh, 2007). When β-catenin binds to TCF/LEF transcription factors, they activate the transcription of the Wnt target genes (Fig. 1.2).

The expression of Wnt target genes is suppressed by TCF/LEF in combination with corepressors in the absence of β-catenin. When β-catenin is present, it binds to TCF/LEF, and this complex upregulates the target gene's expression. TCF/LEF contains a high mobility group box (HMG) domain that interacts with DNA (Oosterwegel *et al.*, 1991; Giese, Cox and Grosschedl, 1992). It binds to cis-regulatory elements: transcription factor binding sites (TFBS) in the promoter region of the target genes.

Several negative feedback control mechanisms target Wnt pathway components. For instance, Axin acts as a negative feedback regulator and controls β-catenin levels in the cytosol through the destruction complex (Lustig *et al.*, 2002). Another negative feedback regulation is carried out by RNF43 (ring finger protein 43) and ZNRF3 (zinc and ring finger protein 3) (Koo *et al.*, 2012). These proteins are ubiquitin ligases and they ubiquitinate the Frizzled and LRP receptors. As a result, receptors are degraded. RSPO's (R-spondin proteins) act as Wnt agonists. RSPO binds to Frizzled and LRP receptors and mediates the clearance of RNF43 and ZNRF3 thus increasing the level of the Wnt signal.

1.4. WNT SIGNALING COMPONENTS

The main components of the Wnt signaling pathway are the following proteins.

1.4.1. β-catenin

β-catenin is a multifunctional protein, coded by the CTNNB1 gene, that takes part in both cell-to-cell communication and gene regulatory processes. It is a member of the catenin protein family that interacts with cadherins to perform cell adhesion. β-catenin is homologous *armadillo* which is a *Drosophila* protein (McCrea, Turck and Gumbiner, 1991). Moreover, β-catenin contains the armadillo repeat, a structure composed of 12 tandem repeats of motifs with approximately 42 amino acids (Riggleman, Wieschaus and Schedl, 1989). Through this

structure, β-catenin interacts with TCF/LEF, APC, and E-cadherin (Huber, Nelson and Weis, 1997).

1.4.2. APC

APC (adenomatous polyposis coli) is a tumor suppressor gene at the center of the Wnt/β-catenin signaling, especially in colorectal cancer. It has many functional domains that interact with several different molecules such as Axin, β-catenin, PP2A, EB1, and microtubules. APC mutations are often found in colorectal cancers. These mutations are mostly truncating mutations that create chromosomal instability that cause carcinogenesis in intestinal cells.

1.4.3. GSK3

Glycogen synthase kinase (GSK3) is a serine-threonine kinase that phosphorylates -OH groups in serine-threonine residues in molecules. It has two isoforms: GSK3-alpha and GSK3-β. GSK3 is constitutively active in most of the cells and regulated negatively by Wnt ligands. It was first shown to regulate glycogen metabolism by phosphorylation of glycogen synthase (Embi, Rylatt and Cohen, 2005). Later, its role in development in *Xenopus* embryos was discovered (He *et al.*, 1995;, Dominguez, Itoh and Sokol, 1995). This is followed by the discovery of β-catenin as a substrate of GSK3. Thus the role of GSK3 in Wnt signaling began to emerge.

1.4.4. TCF/LEF transcription factor

TCF is a family of transcription factors that has four members, TCF7, LEF1 (formerly known as TCF1), TCF7L2 (as activator) and TCF7L1 (as repressor) (Hoppler and Kavanagh, 2007). All members of this transcription factor family have a high mobility group (HMG) box that enables these proteins to bind the DNA. These transcription factors are the downstream gatekeepers of the Wnt signaling.

1.5. WNT SIGNALING PATHWAY TARGET GENES

Wnt signaling is composed of two parts. The upstream part includes Wnt signals, expressed from *Wnt* genes, frizzled receptors on the target cell surface for Wnt signal binding; intracellular components APC, Axin1, and GSK3β complex that regulates the β-catenin level in the cytosol. β-catenin is translocated to the nucleus depending on its availability in the cytosol.

The downstream part of this pathway is composed of the target genes whose activity is regulated by β-catenin and TCF/LEF interactions. β-catenin interacts with TCF/LEF, and

then this complex binds to cis-regulatory elements in the vicinity of the Wnt target genes (Nusse, 2012). These genes are direct Wnt targets. While direct target genes are extensively studied, there is less information about the indirect target genes. This might be due to the difficulties in distinguishing them. Expression of the target genes changes upon stimulation or repression of the Wnt signaling.

In the basic model of a Wnt target gene there is a TCF/LEF or β-catenin binding site on the promoter region. But most of the predicted binding sites for any transcription factor are not functional or are false positives (Wasserman and Sandelin, 2004). Moreover, there might be other influences defining the Wnt target gene.

Wnt target genes are tissue and stage-specific (Nakamura *et al.*, 2016; Boonekamp *et al.*, 2021; Söderholm and Cantù, 2021). Target genes can vary from one cell type to another. Only less than 5% of all candidate target genes were expressed in all microarray studies (Vlad *et al.*, 2008). Therefore, it seems as if there are no universal Wnt targets.

The products of Wnt target genes have many vital roles in biological processes. They function as receptors, cell adhesion proteins, growth factors, cell cycle regulators, and transcriptional regulators. Malfunctions in the cell cycle regulators and growth factors can lead to abnormal proliferation or abnormal cell adhesion proteins, leading to tumor growth. Many Wnt signaling targets play a role in cancer stem cell formation (Kim *et al.*, 2017). Both epithelial and mesenchymal markers are to be a target of the Wnt signaling. Some works demonstrate that Wnt signaling takes part in epithelial-mesenchymal transition (EMT) (Basu *et al.*, 2018). Finding new targets can illuminate many cellular phenomena and possibly provide novel targets for cancer treatment.

1.6. HISTORY OF WNT STUDIES

The beginning of the Wnt research can be traced back to the study of oncogenic retroviruses by Nusse and Varmus (1982) and Nusse *et al.* (1984). Working with mouse mammary tissue, they investigated the sites where these viruses integrate into the host genome and their associated genes. In this search, they found the gene *int1*. This was followed by the discovery of a homolog of int1 in *Drosophila*, the *wingless*, a segment polarity gene that plays a role in development (Rijsewijk *et al.*, 1987). The effect of a mutation in wingless in *Drosophila* was first reported by Sharma and Chopra (Sharma and Chopra, 1976). Mutation in this gene

produces wingless flies. Sharma and Chopra also suggested that mutation in this gene caused several other failures during early development. In the following years, (Nüsslein-Volhard and Wieschaus, 1980) found that *wingless* is one of the genes that control segment number and polarity in *Drosophila* embryos. The int1/Wingless family was renamed into a combination of the two, the Wnt (Nusse *et al.*, 1991). The receptor for this protein family, a member of the frizzled family receptors, was also discovered in *Drosophila* (Bhanot *et al.*, 1996)

In 1991, adenomatous polyposis coli (APC), a tumor suppressor gene was identified (Groden *et al.*, 1991; Kinzler *et al.*, 1991; Nishisho *et al.*, 1991). Mutations in this gene cause familial adenomatous polyposis (FAP), an inherited condition characterized by a large number of benign tumors in gut epithelia that can turn malignant if left untreated. Molecular interactions of APC were unknown until this point. In 1993 the interaction between APC and β-catenin was discovered (Rubinfeld *et al.*, 1993). Also in 1991, McCrea, Turck and Gumbiner (1991) discovered that β-catenin is a homolog of Armadillo, a gene involved in segment polarity in *Drosophila*.

Around this time an important component of the destruction complex also emerged around this time in the Wnt scene. The first report on the GSK3-Wnt signaling relationship came from Siegfried, Chou and Perrimon (1992). They demonstrated that zest-white-3, the *Drosophila* homolog of GSK3, acts as a negative regulator in the pathway. This was followed by a study from Peifer *et al.* (1994) that showed the phosphorylation of Armadillo —a homolog of β-catenin— by zest-white-3 in *Drosophila*. A subsequent study from Rubinfeld *et al.* (1996) reported that, when β-catenin levels are high, GSK3 and APC form a complex that reduces the β-catenin levels. This study reports the early discoveries about the destruction complex. Through the mentioned studies, the Wnt signaling slowly began to emerge.

In the following years, researchers discovered the interaction between TCF/LEF transcription factors and β-catenin (Molenaar *et al.*, 1996; Behrens *et al.*, 1996). This paved the way for the discovery of the downstream part of the Wnt signaling cascade, the target genes. Soon after this, the first target gene, c-myc, was reported (He *et al.*, 1998).

Many Wnt target genes have been found in different species. The number of target genes grows over time, but most of these genes are tissue-specific or act on different stages of development (Söderholm and Cantù, 2021). Therefore the underlying mechanisms are still unknown.

1.7. WNT AND COLORECTAL CANCER

Wnt/β-catenin signaling is one of the major driving forces of stem cell proliferation and differentiation in the intestines. The proper functioning of Wnt signaling is essential for intestinal homeostasis (Wong, Rubinfeld and Gordon, 1998). Aberrant Wnt signaling can introduce tumors to this rapidly changing system.

The small intestine lining is composed of two compartments: the protruding finger-like ruffles, called villi; and the inward cavities between villi, called crypt. Villi cover the surface of the small intestine. They are composed of several cell types that absorb nutrients and also secrete mucus, hormones, and enzymes. The crypts are invaginations, positioning the opposite direction of the intestinal lumen. Inside the crypts, stem cells divide and differentiate into other cells of the intestine epithelium. The other type of cells at the crypt, called Paneth cells, secrete antimicrobial agents.

Stem cells in crypts divide and differentiate into mature epithelial cells as they move upwards (Fig. 1.3). These mature cells include enterocytes, enteroendocrine cells, and goblet cells. The journey ends at the tip of the villi. The cells detach from epithelium and the cells go through apoptosis. This cycle is completed in about 3 to 4 days.

Wnt has a key role in homeostasis and stem cell renewal at intestinal crypts (Fevr *et al.*, 2007). Korinek *et al.* (1998) reported that mice lacking TCF-4 resulted in loss of proliferative characteristics of intestinal crypt. Ireland *et al.* (2004) showed that β-catenin is needed for stem cell proliferation in intestines. Abnormal expression of Dkk1 (a Wnt inhibitor) in transgenic mice resulted in a loss of crypts and secretory cell lineage in the intestinal epithelium (Pinto *et al.*, 2003; Kuhnert *et al.*, 2003). For instance, the Wnt pathway target gene c-myc causes a rapid loss of crypts upon deletion (Muncan *et al.*, 2006).

According to global cancer statistics of the year 2020, colorectal cancer is the third most frequent type of cancer diagnosed worldwide, with 1.9 million new cases estimated (Sung *et al.*, 2021). About 15% of colorectal cancers arise from a germline mutation in the tumor suppressor APC gene, leading to familial adenomatous polyposis (FAP). People with this mutation develop several thousands of benign tumors in their lifetime. The remaining 85% of colorectal cancer cases are sporadic. Nearly 80% of colorectal cancers come from mutations of the APC gene (Markowitz and Bertagnolli, 2009). Colorectal tumor formation usually begins with APC mutation, but it is not sufficient alone for full-blown tumorigenesis. Additional

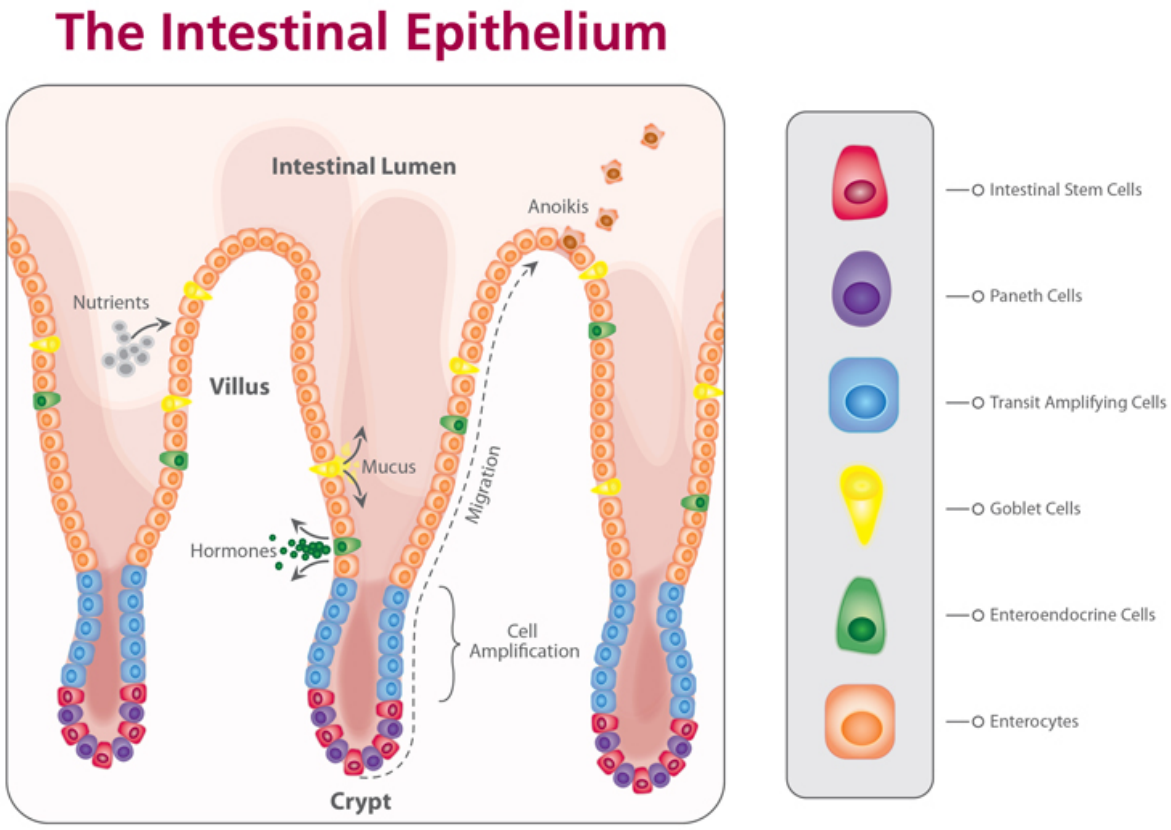

Figure 1.3: Illustration of the intestinal epithelium. It is composed of two sections: villus and crypt. Each of them has characteristic cell types. Intestinal stem cells divide, and differentiate into specific cell types in the intestinal epithelium. While they divide and differentiate, they migrate along the villus. When they reach the villus tip, they separate from the epithelium and go through apoptosis.

mutations in β-catenin, *K-ras,* and p53 genes are needed for invasive tumor formation (Pinto and Clevers, 2005). Initially, APC mutations lead to benign adenomas, then the additional K-ras mutations and p53 mutations, tumors develop into carcinoma.

Tumor introducing mutations in other Wnt components related to colorectal cancer have also been reported. One study suggests that 18% of colorectal cancers have RNF43 truncating mutations that do not coincide with APC mutations (Giannakis *et al.*, 2014). Another suggests that RSPO2 and RSPO3 (R-spondin family members) fusion occurs in 10% of colorectal cancers, and this fusion can trigger Wnt signaling and tumor growth (Seshagiri *et al.*, 2012).

1.8. TARGET GENE IDENTIFICATION

Traditionally, computational and transcriptomic/genomic approaches have been used to find novel target genes (Vlad *et al.*, 2008). In both approaches, researchers look for TCF/LEF binding sites in the vicinity of the target genes.

Transcription factor binding sites are characterized by short, degenerate sequences called motifs. Motifs can be described from either a consensus model or a PWM (position weight matrix, also known as position-specific weight matrix or position-specific scoring matrix) model (Singh and Hannenhalli, 2008). These methods have a high false-positive rate (Medina-Rivera *et al.*, 2011). Given the vastness of the genome and the degeneracy of binding sites motif, searching for functional binding sites of TFs is a challenging task (Blanchette *et al.*, 2006).

Computational search for TCF/LEF binding sites includes genome-wide scanning TCF/LEF binding motifs on the upstream region of genes (Hödar *et al.*, 2010). With computational search, it is possible to rapidly identify potential targets with matching binding sites on their upstream. However, these predicted binding sites may not be functional, and TCF/LEF may have other degenerate binding sites that can go unnoticed (Ramakrishnan and Cadigan, 2017). Therefore, computational findings must be backed with experimental validations.

Novel target genes for Wnt signaling are mostly discovered or validated with wet-lab experiments. In these experiments two questions are answered:

- Is there a TCF/LEF binding site in the promoter of the gene of interest?
- Comparing the Wnt signaling activated and repressed cells, Is the gene of interest

differentially expressed?

These questions are answered with several techniques in different studies. The presence of a TCF/LEF binding site can be determined with Electrophoretic Mobility Shift Assay (EMSA), ChIP (Chromatin Immunoprecipitation), and Reporter Assays. For determining the differential expression, gene expression is measured using microarrays, qPCR, or RNAseq.

1.8.1. Wnt signaling manipulation

There are several ways of manipulating the Wnt signaling pathway. Wnt proteins are used as activating agents, since these are the molecules that start the signaling cascade. Recombinant Wnt proteins are commercially available for experimental use. There are also Wnt surrogate molecules that mimic the action of Wnt and act as an agonist (Janda *et al.*, 2017; Tao *et al.*, 2019; Miao *et al.*, 2020). Since Wnt signaling plays a role in development, Wnt signals, agonists, and surrogates are also important in organoid studies.

Another strategy for the activation of Wnt signaling is targeting GSK3β. This molecule is part of the destruction complex and it is responsible for the phosphorylation of β-catenin which leads to its degradation. Inhibition of GSK3β means accumulation of β-catenin which regulates the target genes with TCF/LEF. LiCl is frequently used for this (Klein and Melton, 1996). There are several other GSK3β inhibitors (Cohen and Goedert, 2004).

GSK3β inhibitors are also heavily studied for therapeutic targets. This molecule is associated with several diseases and disorders, such as Alzheimer's disease, diabetes mellitus, bipolar disorder, and some cancers (Cohen and Goedert, 2004).

Components of the Wnt signaling pathway, such as β-catenin, APC, GSK3, and TCF/LEF, are silenced using RNA interference technology to contrast them with their normal state. Double negative constructs of TCF/LEF can be used to repress the Wnt signaling (Molenaar *et al.*, 1996). Mutations induced in β-catenin are also used to alter Wnt signaling.

1.9. MACHINE LEARNING

Machine learning is a branch of Artificial Intelligence that focuses on creating *models* to classify data and make predictions about it. Machine learning uses example data to *train* models, then these trained models can be used to make predictions and classify new cases. Machine learning models can "see" patterns that may be hard to discover by other means, and

therefore yield some insights into the topics we investigate.

In general terms, data consists of a set of individual *observations*, each one containing several attributes, also called *features* or *variables*. In machine learning, variables or attributes are used to make an inductive reasoning based on observations (Liu and Motoda, 2013). The data is a collection of the features that can be numerical or categorical values. This set also contains a *label* feature that indicates the class of the observation. The job of a trained classifier is to predict labels of the new observations based on the features used while training.

We train a model by presenting examples of inputs and outputs to the machine learning algorithm, and the machine tries to find an approximate formula that encodes the relationship between the inputs and outputs (Samuel, 1959). In contrast to traditional methods, machine learning does this discovery of relationships without explicit programming. Once a machine learns the relationship, it can predict what would be the output for any novel input.

In broad terms, there are two types of machine learning methods, depending on the attributes present on the example data. If each example has a *label, class*, or *value* that may depend on the other features, then we can use supervised machine learning. Here "supervised" means that we know the labels of each observation, and we "supervise" the machine to learn them. In contrast, when there are no labels in the data, we can use an *unsupervised* method to classify (or cluster) the observations in a coherent way. Clustering is dividing data into groups regarding the distances between data points.

In this thesis we have pre-existing labels for some of the observations. Therefore, we will use a supervised machine learning method.

There are two main types of supervised machine learning: classification and regression. In the first case the training data is labeled with a categorical variable (let's say, color) and the machine learning model should be able to predict the label for new data. In the second case, the model aims to predict a numerical value (let's say . . .). In this work we focus on *classification*.

There are several standard machine learning tasks such as classification, regression, clustering, ranking, and dimensionality reduction. Since we would like to know the Wnt target labels (whether a gene is a Wnt target or not) of the genes, our desired output is categorical, hence ours is a binary classification task.

Our classification problem is different from the classical ones. Instead of a set of labeled examples, we only have a small subset of labeled examples and the rest of the observations do not have a label assigned to them. The labeled examples are experimentally validated target genes of the Wnt signaling pathway. Rest of the genes do not possess any label since we do not know whether they are Wnt target or not. We expect that there are more target genes and we are looking for them in the non-labeled observations. Our problem is analogous to spotting a genuine art piece among its forgeries.

So let's say we have class 1, the bonafide target genes, and class 0 which contains the rest of the genes. We hypothesize that a small number of genes that are in class 0 are actually belong to class 1.

1.9.1. Classification error and Generalization error

In supervised learning, the example data is split into two groups: training and test sets. Both of these sets contain the *labeled data*. In a set of labeled data, each entry has a label indicating the group of data. The machine is trained with the test set. After the training then the trained model is tested with the test set without its labels. Then we compare labels that are guessed by the model and the actual labels of the test set to assess the accuracy/success of the trained model.

A good machine learning model should be able to generalize (beyond the training set) the predictions when new data is presented. We want our model to perform well not just on training data, but also correctly classify the data that it has not seen before.

Generalization error is a measurement of how accurate the model is when making predictions about the unseen data. Since we are training only with small portions of data, rather than all the data there is, our model can be biased to our sample. We may introduce sampling error into our model.

Bias is the difference between the predicted value and the real value. If bias is high in the model then the model is too simple to convey the underlying pattern. Models with high bias *underfit* the data. they are too general, therefore models do not reflect the inner complexity of the data (Fig. 1.4 and Fig. 1.5).

Variance is a measure of how the model can change depending on variation of the training data. Models with high variance tend to *memorize* the training data rather than learn its inner

pattern. They fit the data perfectly so that they fail to generalize and correctly classify the new instances. Models that fail to generalize are said to be *overfitting*. Sometimes the data itself has some degrees of noise in it. A model that is too perfect fitting to the training data can also be conforming to this noise, and this can result in a highly incorrect model (Fig. 1.4 and Fig. 1.5).

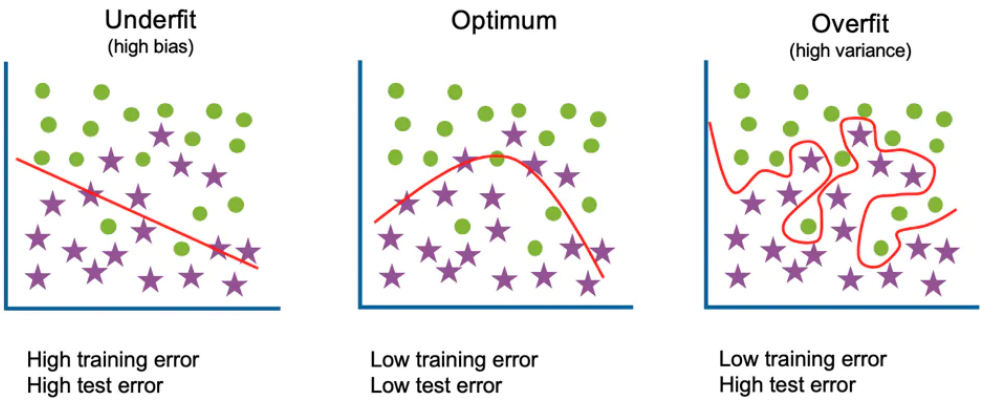

Figure 1.4: Performance of the models depending on the bias-variance tradeoff. The figure shows the same data explained with three different models. The one on the left is too simplistic to capture the complexity of the data so it is an underfit. This model has high bias, thus the differences between predicted and actual values is large. The plot on the right separates the two classes perfectly, but with a highly-complex model that suggests overfitting. The model would probably fail to correctly classify new data. The plot in the middle is an optimum model which is not too simple, not too complex, with a balance between bias and variance (IBM Cloud Education (2021)).

1.9.2. Decision Trees

A decision tree is a type of predictor that uses the method called recursive partitioning. In this type of predictor, data is split into subgroups based on predefined splitting rules. Starting with the **root node** the data is split into **decision nodes** based on different rules. Decision nodes can be split into other nodes. The ones that do not split further are called **leaf nodes**. The data is iteratively split with the decisions made and gradually a decision tree is formed along the way (Fig. 1.6).

When a decision tree predicts the category of data it is a **classification tree**. If it makes numerical predictions then it is a regression tree.

There are several different machine learning algorithms to build a decision tree, such as ID3, C4.5, CHAID, MARS, and CART. In this work we used CART, so we will only focus on this algorithm. A discussion of the other algorithms is beyond the scope of this thesis.

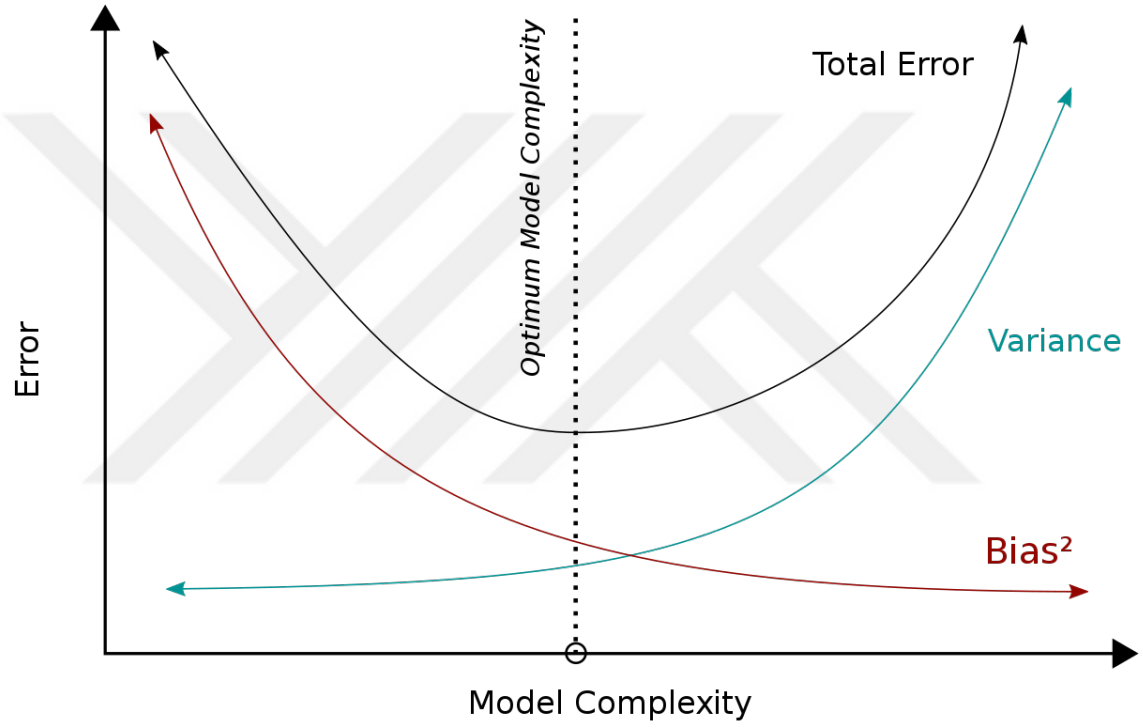

Figure 1.5: Bias-variance tradeoff in machine learning. When the model complexity increases, the variance also increases and the bias decreases. The generalization error can be decomposed as the sum of variance plus the square of bias. The optimal model complexity is achieved by minimizing the combination of bias squared and variance, so both are expressed in the same units. ('Bias–variance tradeoff', 2022).

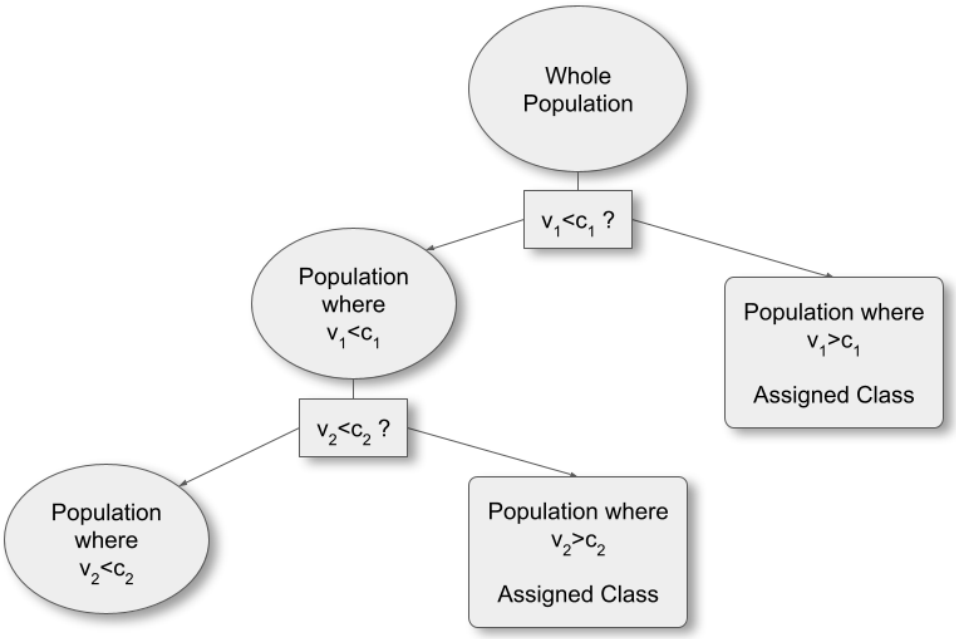

Figure 1.6: Basic structure of a decision tree. The node on the top of the is the root node. Then this first node of the tree subdivided into two nodes depending on whether the variable v1 is less than the threshold c1. The nodes that can be split further with a new rule are decision nodes. The nodes that are not divided further are leaf nodes. Adapted from Aravena Duarte (2013).

CART, short for Classification and Regression Trees, is a supervised machine learning algorithm proposed by Breiman *et al.* (1984) in which each node in the tree is defined by comparing a single variable against a threshold. Thus, this splits the data into two smaller sets. The CART algorithm specifies how to choose those variables and thresholds.

The splitting of the tree is carried out in a way that minimizes a measure called Gini impurity. Gini impurity is the probability of misclassification of new data (Breiman *et al.*, 1984). Gini impurity is measured with this:

$$\text{Gini Impurity} = \sum_{i=1}^{j} p_i(1 - p_i).$$

The j index indicates the number of classes in the node and p_i is the probability of class i in that node. Gini impurity is calculated for each possible split considering every value on each column of the data. Splitting from the root node based on the column that has the smallest Gini impurity.

The combination of column and value resulting in the smallest Gini impurity defines the

variable and threshold of a new inner node in the tree, that is, a split.

1.10. AIM OF THIS STUDY

This study aims to identify novel target genes in the Wnt signaling pathway in the context of colorectal cancer in humans. Earlier works suggest that Wnt signaling patterns can change depending on the tissue and disease. Therefore, target genes may be tissue-specific. Following this idea, we chose colorectal cancer as our tissue and disease of interest, mainly because Wnt signaling is extensively studied in this context. Therefore, gene expression data is more abundant in this field compared to other tissues and cancer types. Moreover, there is a good number of experimentally validated target genes in this context.

Regardless of tissue-specificity or developmental stage, novel target genes are usually discovered one gene at a time with wet-lab experiments. In these experiments most of the time two things are demonstrated: a presence of the transcription factor binding site for TCF/LEF on the gene's promoter, and whether the gene is differentially expressed upon activation or repression of the Wnt signaling.

We will take a different approach and use the existing data to discover novel target genes computationally. Following the work of Hödar *et al.* (2010) we will train a pool of classifiers. Hödar used genomic data to represent the genes, which is the information of transcription factor binding sites on the known genes' promoters. We will differ from Hödar *et al.* (2010) by representing genes with transcriptomic gene expression data. Since we focused on colorectal cancer we will use gene expression data from colon cells or colorectal cancer cells.

2. MATERIALS AND METHODS

Our goal is to discover novel Wnt *target genes*. Our general strategy for this goal is to collect relevant data, train machine-learning *classifiers* with this data, and use them to predict putative novel target genes. In this chapter, we will describe the protocols used to find a set of experimentally validated target genes, to collect the data used to characterize genes, and to train a pool of classifiers. The set of experimentally validated genes will be called *bonafide targets* in the rest of the text. In some cases, genes will be represented by their differential expression under several conditions. This requires a specific statistical analysis that we will also describe.

The genes that we want to classify will be represented by a list of *features*, such as their differential expression values under several conditions, or the presence of specific binding sites in their upstream region. This last characterization is also known as *transcription factor binding site fingerprint*.

We initially follow and then extend the protocol described in Hödar *et al.* (2010), using updated data and upgraded methods. First, we will describe how we compiled the set of bonafide target genes to be used as *positive* examples to train the *classifiers*. Then we will explain how we trained our classifiers with these bonafide target genes and transcription factor binding site (TFBS) fingerprints, and how we evaluated the results of this training. Following this, we will describe our search for a new appropriate dataset.

In the second part of our analysis, in contrast to Hödar *et al.* (2010), we represented each gene with its differential expression data instead of TFBS fingerprints. We will explain how we built the differential expression matrices and used them as the input data for training a set of classifiers.

2.1. GENERAL IDEA

In general terms, the input data used to train and test a machine learning classifier is usually a matrix containing categorical or numerical values. Each row is an observation, each column is a feature and each cell contains the value of the given feature for the given observation. The output of these classifiers are categorical values indicating the class assigned to each test case.

For our purposes, the set of genes will be represented by numeric matrices containing experimental values for each gene in several conditions. The classifiers will be trained to classify genes as either *targets* or *non-targets*.

The set of bonafide genes that we compiled will be used as positive examples in the training of machine-learning classifiers, and the set of negative examples will be chosen randomly from the rest of the genes.

2.2. COMPILING A REFERENCE LIST OF BONAFIDE TARGET GENES

For the selection of the reference database of bonafide target genes, we used several criteria. Since we focused on colorectal cancer, we chose genes that have been described as Wnt targets in human colorectal cancer cell lines.

2.2.1. References for bonafide target genes

We used the Wnt Stanford webpage ('Wnt target genes', no date) and the work of (Herbst *et al.*, 2014) to compile a list of genes discovered in the colorectal cancer context. That gave us 66 bonafide target genes. We found 29 bonafide target genes in the colorectal cancer context which was published after Herbst's (Herbst *et al.*, 2014) work.

2.2.2. Protocol for bonafide target gene selection

We added other target genes that we found searching into the NCBI PubMed Central database (Biotechnology Information, no date). We used keywords such as "WNT", "WNT target", "beta-catenin target", "tcf/lef target", "colon cancer", "colorectal cancer" in our search of publications for target gene discovery.

We checked these publications for discussions about differential expression of the gene of interest, and the presence of a TCF/LEF binding motif in its promoter. If the gene is differentially expressed upon direct activation or blocking the Wnt signaling and has one or more TFBS for TCF/LEF on its promoter in the same paper we consider it is a bonafide Wnt target in colorectal cancer.

We identified 93 target genes matching our interests. They can be seen in Appendix 1 with their corresponding bibliographical references.

2.3. USING TFBS FINGERPRINTS TO REPRESENT GENES

2.3.1. What are TFBS fingerprints?

The theory indicates that if a gene is directly regulated by a transcription factor, a *binding site* for that transcription factor must be present in the upstream region of the gene. The presence of a binding site can be inferred by matching the genomic sequence to a *motif* that characterizes the multiple instances of binding sites for a given transcription factor.

A TFBS fingerprint is the number of times that each transcription factor binding site motif is found in the upstream region of each gene. The fingerprints of all genes can be written as a matrix, one row for every gene and one column for each TFBS motif.

These fingerprint matrices depend on: the set of genes selected, the length of the upstream region considered, the set of motifs to search inside the upstream regions, and the software used to identify the presence of every instance of each motif.

Hödar *et al.* (2010) used TFBS fingerprints described in 2004 by Ron Shamir (Elkon *et al.*, 2003) to characterize each human gene.

2.3.2. How did Shamir lab build this matrix

Ron Shamir *et al.* developed in 2003 a program called PRIMA (Elkon *et al.*, 2003) (part of their EXPANDER (Ulitsky *et al.*, 2010) suite) to identify the presence of binding sites, and used the motifs published in TRANSFAC (Wingender *et al.*, 2000) to determine which transcription factors had binding sites in the upstream region of every gene. According to the documentation in the webpage of "PRIMA – a software for promoter analysis" ('PRIMA: A software for promoter analysis', no date):

> "EXPANDER, on the other hand, executes PRIMA on a fixed set of precomputed fingerprints, which were constructed as follows: A set of about 17,000 human promoter sequences, spanning from 1000 bp upstream the TSS to 200 bp downstream the TSS, was scanned in order to locate putative BSs (hits). The scan was performed for each TF motif (PWM) in TRANSFAC (version 5.4, April '02) that corresponds to a Human TF. The information on the number of hits of each PWM in a promoter is called the fingerprint of that promoter. The fingerprints of all human promoters are supplied with EXPANDER. The human promoter sequences were downloaded from Ensembl (release 13.30)"

The 2004 version of EXPANDER included the data required to produce the matrix used by Hödar *et al.* (2010). It was a matrix with 15476 genes and 432 features per gene (TFBS motifs). Each gene was identified using its NCBI Entrez Locus ID (Table 2.1).

A newer version of the TFBS fingerprint matrix has 19336 genes and identifies 1046 different features in the 3000 bp upstream of each gene, as shown in Table 2.1 (Private communication from Tom Hait, R. Shamir's lab).

Versions of EXPANDER published after 2007 use ENSEMBL ID for each gene. In other words, Shamir Lab changed from the American to the European standard. This change introduced some incompatibilities since both nomenclatures do not match 100%. Some genes have a single ID in one standard and several IDs on the other one.

Moreover, the newer version of EXPANDER used the JASPAR database, since TRANSFAC updates are not available free-of-charge after 2007.

2.4. CLASSIFICATION WITH UPDATED TFBS MATRIX

The researchers at Ron Shamir's lab have updated their predictions of TFBS fingerprints.

They have two updated versions that include 1046 and 1080 features for each gene respectively, using 3000 basepairs upstream as the promoter region.

In this version, however, each gene is identified with EMBL identifiers (from the European Molecular Biology Lab), which do not map 100% to the Entrez identifiers (from the USA National Center for Biotechnology Information) previously used. Therefore, we were unable to correctly identify all the bonafide genes.

Table 2.1: Overview of fingerprint matrices created by Ron Shamir's lab. Each of these matrices can be used to represent genes.

Date	Num. Motifs	Num. Genes	ID method	Name
2003	432	15476	Locus ID	PRIMA
2009	1046	19336	Ensembl ID	EXPANDER
2015	1080	21246	Ensembl ID	EXPANDER

2.5. USING DIFFERENTIAL EXPRESSION DATA TO REPRESENT THE GENES

2.5.1. Getting differential expression data

We used several gene expression data as an alternative to the TFBS fingerprint for gene characterization. The raw data of these experiments is available at NCBI GEO. In the case of RNAseq experiments, the raw reads are available at NCBI SRA.

Following the tissue specificity idea, described in the Introduction chapter, and focusing on colorectal cancer, we chose experiments performed either in colorectal cancer cell lines or in colon organoids.

We chose colorectal cancer as our focusing point since the Wnt signaling pathway is frequently studied in this condition. Most of the validated target genes are studied in colorectal cancer tissues, so more data is available in this context.

Since we selected the bonafide target genes from studies conducted in human colorectal cancer cell lines, we expect that these genes would be differentially expressed in that condition. Therefore we chose the gene expression experiments in colorectal cancer cells and colon organoids.

In colorectal cancer cell lines, Wnt signaling is constitutively active. Silencing the key Wnt components is a frequently used approach for obtaining differential expression.

In these gene expression experiments, different strategies are used to repress or activate the Wnt signaling. We chose the experiments where only direct Wnt stimulation or repression was applied such as APC silencing, recombinant Wnt treatment, genetically modified cells with double negative LEF1.

Finally, to get a complete overview of all genes, we selected the experiments that were performed either with microarray or high-throughput sequencing (RNAseq). We processed microarray produced data in R and the sequencing data in the Galaxy environment.

2.5.2. Protocol for choosing the gene expression experiments

We searched the NCBI GEO database and chose the experiments where Wnt activity is stimulated or repressed through the manipulation of different Wnt components. We searched using keywords such as “WNT”, “beta-catenin”, “colon cancer” or “colorectal cancer”.

We used RNAseq results from the datasets GSE76307, GSE95670, GSE120106, GSE125472,

GSE125578, GSE130236, GSE131575, GSE131948, and microarray data from GSE18560, GSE44097, GSE87429, GSE64889, GSE56896. Table 2.2 shows all the GEO series used for this work.

Table 2.2: List of the data we used in our study.

GSE Accession	Cell Line	Experiment Type	Reference
GSE76307	SW480	high-throughput sequencing	(King *et al.*, 2016)
GSE120106	CCD-18Co	high-throughput sequencing	(Ferrer-Mayorga *et al.*, 2019)
GSE125578	Human colon organoids	high-throughput sequencing	(Michels *et al.*, 2019)
GSE125472	Human colon organoids	high-throughput sequencing	(Michels *et al.*, 2019)
GSE95670	HCT116	high-throughput sequencing	(Ewing *et al.*, 2018)
GSE131575	HT-29	high-throughput sequencing	(Dietinger, Durango, *et al.*, 2020)
GSE131948	SW480 and SW620	high-throughput sequencing	(L. and Marian, no date)
GSE130236	SW480 and SW620	high-throughput sequencing	(Waterman ML, no date)
GSE18560	Ls174T	microarray	(Mokry *et al.*, 2011)
GSE44097	DLD1 and SW480	microarray	(Herbst *et al.*, 2014)
GSE87429	HCT116	microarray	(Ohtsuka *et al.*, 2016)
GSE56896	LS174T	microarray	(Whissell *et al.*, 2014)

2.5.3. Processing microarray data

We processed the collected microarray data in the R environment. We used limma (Ritchie *et al.*, 2015) and several other libraries (*oligo*, *annotationdbi*, *org.hg.db*). The processing protocol is described in Appendix 7.

2.5.4. Processing RNAseq data in the cloud using Galaxy

Galaxy is a web-based platform where one can process biological data without any programming knowledge. There are numerous tools for different tasks in Galaxy, which makes it easy to transform the data for further analysis. We used Galaxy to analyze the RNAseq data that we found in our search (Fig. 2.1).

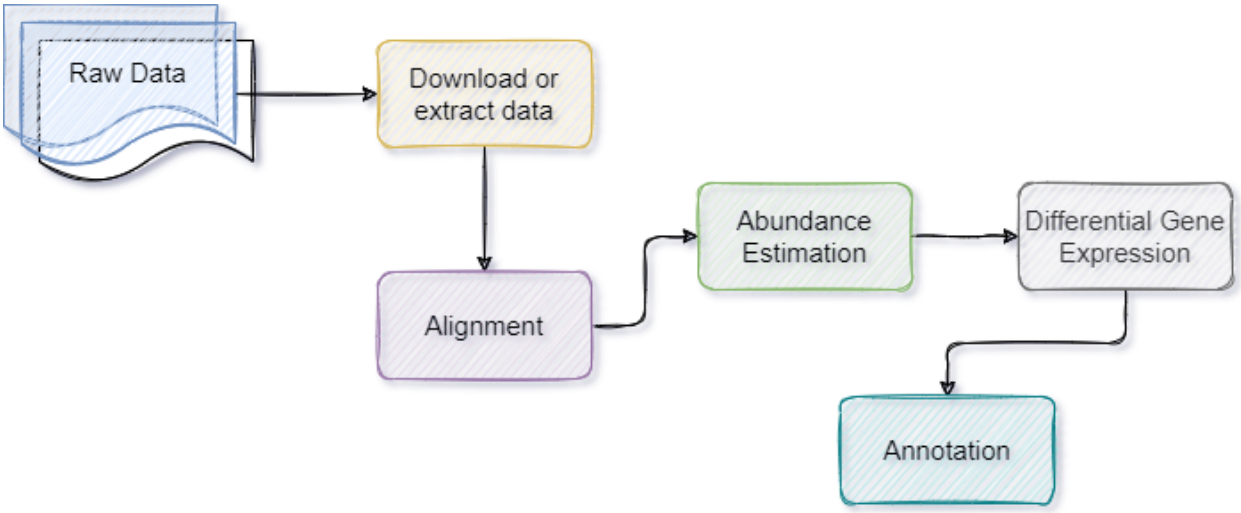

Figure 2.1: Overview of Galaxy workflow.

2.5.4.1. Importing RNAseq data to Galaxy

RNAseq is a next-generation sequencing technology that identifies the nucleotide sequence of each transcript. Every RNA fragment that is sequenced produces a *read*. This information allows us to quantify the number of reads in a given sample, and therefore evaluate the expression of every gene.

NCBI GEO contains information about all gene expression experiments. In the case of RNAseq experiments, the GEO sample has a link to the corresponding Sequence Read Archive (SRA) database entry.

DNA sequencers produce output information in *fastq* file format. This kind of file contains the base sequence of the fragments and a quality score for a base at each position. *Fastq* files are generated by RNAseq experiments and stored in the NCBI repository Sequence Read Archive (SRA).

To analyze RNAseq data in Galaxy, we imported SRA data using the tools described below.

For this step, we use the relevant SRA accession numbers. The NCBI GEO page of RNA-seq experiments has an *SRA Run Selector* link at the end of the page. We accessed experiments' sequenced reads and further information about reads via this link.

We specified the list of SRA accession numbers in a text file using the *paste/fetch* data option in the *upload data* section. Then we used the tool *Faster Download and Extract Reads in FASTQ* (Leinonen, Sugawara and and, 2010). This tool takes as input the list of SRA accessions and produces outputs based on whether the library preparation is single-end or pair-end.

2.5.4.2. Quality control

After we transfer the data from NCBI to Galaxy, the next step is quality control. We checked the quality of the samples using the *FastQC* tool (Andrews *et al.*, 2012). This tool produces a report with plots on several important parameters such as base quality, adapter content, levels of duplication, GC content, and similar.

The sequencing process assigns a quality score for each base, which is related to the probability of an incorrect base calling. It is important to verify that most of the bases have high quality to be sure that the results correspond to the underlying biology, and not to technological artifacts.

2.5.4.3. Trimming

Typically, some of the bases at the beginning and the end of each read have bad quality, and it is better to trim them to increase the validity of the data. There are several tools for trimming and/or filtering low-quality reads. We used *trimmomatic*, a practical tool where one can perform several trimming operations such as removing low-quality bases at the beginning or end of the reads.

We also need to remove the sequences of the adapters used in the sequencing process. Adapters are small sequences that are ligated at the beginning and end of the RNA fragments. They are used so that the machine can recognize the sequences. To remove the adapters, there are several tools such as *cutadapt*, *clip*, and *trim galore*. They are very similar. In these tools, one can choose either default adapters or provide an adapter sequence to be removed. We used the tool *clip* for these tasks.

We processed the experiment GSE87429 with this protocol. Some of the samples in this series

had overrepresented adapters so we removed them. And some of them had low-quality bases at the beginning or end of the reads. We also removed them. The result of this analysis is in the github page: *https://github.com/cemrekefeli/cemre-kefeli-WNT*. The other experiments/sra data were of good quality.

2.5.4.4. Aligning the reads to a reference genome

After quality control, we aligned the reads to a reference genome to determine which part of the genome corresponds to these RNA fragments. There are several tools for the alignment of reads to a reference. We used HISAT2, an alignment tool based on the Burrows-Wheeler transform algorithm that takes *fastq* files as input and produces a BAM output file. HISAT2 in Galaxy provides many reference genomes but it is also possible to upload a reference genome. We chose hg38 which was a built-in genome in HISAT2.

The output of HISAT2 is a file in SAM format.

2.5.4.5. What are SAM and BAM

SAM is short for Sequence Alignment Map, which is a tab-delimited text-based file format that contains aligned sequences mapped to a reference genome. SAM files often start with header rows, which begin with "@". After the headers, we see the information about the aligned reads. Each row shows one aligned read, and for every read there are 11 fields.

The binary equivalent of SAM is called BAM. It is smaller in size and easier to manipulate on the computer compared to the SAM files, but contains the same information. SAM files are human- and computer-readable, BAM files are only computer-readable.

2.5.4.6. The number of transcripts mapped to each gene

The alignment step is followed by the estimation of gene expression by evaluating the number of transcripts mapped to each gene. The tools available for this task in Galaxy are *Stringtie*, *HTSeq-count*, and *featurecounts*. We used *featurecounts* because this tool provides an annotation of the human genome. With the other tools, we would have to provide this annotation independently, in an GTF or GFF3 file.

This tool takes SAM or BAM alignment files as input, and produces a set of tabular files containing the number of reads for each gene. These files have two columns: the first one has the gene identifiers and the second one has the number of counts per gene. In our case,

the gene identifiers are Entrez Ids. These output files can then be used as input to differential gene expression analysis tools, such as *DESeq2*, *edgeR*, and *limma*. We used the tool *limma* for posterior analysis.

2.5.5. Normalization

Raw counts need to be normalized before the analysis. The main reason for normalization is to accurately see the gene expression. Normalization improves the quality of data and reduces the biases, therefore, it ensures the validity of the downstream analysis. Biases might be from known or unknown sources so we need to choose the correct normalization method carefully.

The difference that we initially see between samples may not be due to biology, but instead it may be due to some other factor, such as sequencing depth or gene length. For instance, one sample may have more reads in total than the other samples, so its genes will show greater counts. This does not necessarily correspond to higher gene expression. The difference might be due to the increased number of total reads. Some genes are longer, they tend to have more reads mapped to them, which does not necessarily mean higher gene expression. Therefore samples are normalized, correcting these issues, to be comparable with each other.

There are several types of normalization methods such as; RPM, RPKM, FPKM, TPM, TMM. These methods are designed to address several technical biases such as library size, gene length, or GC content. The details of these methods are beyond the scope of this thesis, and can be seen in Abbas-Aghababazadeh, Li and Fridley (2018). Here we explain the basic idea of TMM, which is the default normalization method for limma and edgeR packages, and is the one we used (Robinson and Oshlack, 2010).

Basically, TMM considers that the total RNA expression in each library will affect the observed count of reads per gene. This total RNA cannot be measured directly, but we can estimate its ratio between conditions if we assume that most genes are not differentially expressed. In that case the between-samples normalization factor can be estimated as the trimmed mean of the M values (that is, the log-ratios of observed read counts). A *trimmed mean* is an average in which a percentage of the highest and lowest values are discarded, assuming that they are outliers.

One interesting feature of TMM is that the normalization is independent of the gene length, thus simplifying the analysis.

2.5.6. Statistical analysis of differential gene expression

For the analysis of the gene expression we used *limma*, a tool that uses linear models to analyze the differential expression in an experiment (Ritchie *et al.*, 2015).

2.5.6.1. Groups and contrasts

Several parameters need to be specified. To identify the experimental conditions, we need to declare one or more factors and their group names (levels). These names will be used to build the contrasts required to compare two or more conditions.

In *limma*, one can either declare a contrast of interest (in which case the contrast matrix will be prepared automatically) or provide an *ad-hoc* contrast matrix. More than one contrast can be added in Galaxy/limma by using the insert contrast button.

2.5.6.2. Filtering low-expressed genes

The next step in *limma* is to filter-out genes with noisy or uninformative counts. Low-expressed genes have higher variability, therefore removing low expressed genes/counts decreases the total variance. And, in our experience, filtering genes with very low counts improves the number of differentially expressed genes compared to the not filtered analysis. Filtering can be done in limma/Galaxy using the Filter Low Counts tab.

Here, there are two ways of filtering.The first option is to consider the total count across all samples and discard the rows/genes where this total is below a fixed threshold. The second option is to discard the rows/genes where several samples are below a threshold. In this case, the minimum number of samples to meet the criteria needs to be specified. Clearly, since counts are always positive numbers, the first filter is more stringent than the second one, because if the total count is small, then each sample is small. On the other hand, several samples can have low counts without resulting in a small total. Alternatively, we can filter genes considering their normalized counts per million (CPM) instead of the raw count. In this case we can only filter with the second criteria.

We filtered by raw counts with a threshold of 10 counts. We discarded every row where the sum of counts across all samples was below 10.

2.5.6.3. Outputs of limma

Limma has two main outputs, a differential expression table that contains statistical information about each gene, and an HTML report that contains the plots and information about the analysis. A differential expression table is created for each contrast. This table contains gene identifiers in the first column, and log fold-change, average expression, t-value, *p*-value, adjusted *p*-value, and B value in the following columns.

We combined all differential expression tables in a single matrix. Since some samples contain different sets of genes, we had to decide what to do with the missing ones. Initially we kept only the genes present in all samples, but we lost a considerable number of genes. Since the R implementation of CART can handle missing values, we included all of the observations, in this case, genes with missing values (Therneau and Atkinson, 2022).

2.5.6.4. Annotating the genes

At this point, we only know the Entrez identifiers of the genes. We retrieved the annotation for the genes with the annotateMyIDs tool. A tabular file that contains a gene identifier in its first column should be provided as input. We should also choose which organism the identifiers belong to. In this case, we chose "human". There are only five built-in organisms for identifiers. The type of the gene identifier should be specified in the Id Type section.

2.5.7. Combining RNAseq and microarray data

We build a matrix combining differential expression data from RNAseq and microarray experiments. This matrix was used to train a new set of classifiers.

This matrix has 19073 rows and 180 columns. The first column contains the ENTREZ identifiers.

2.6. TRAINING THE CLASSIFIER AND PREDICTING NEW GENES

Once we have a list of bonafide genes and a matrix to characterize each gene, we use them to train different sets of machine learning classifiers. As in Hödar *et al.* (2010), we use the classification tree (CART) method. As usual, to train these classifiers we need to present positive and negative examples (Fig. 2.2).

Our classification problem is different from the classical ones.

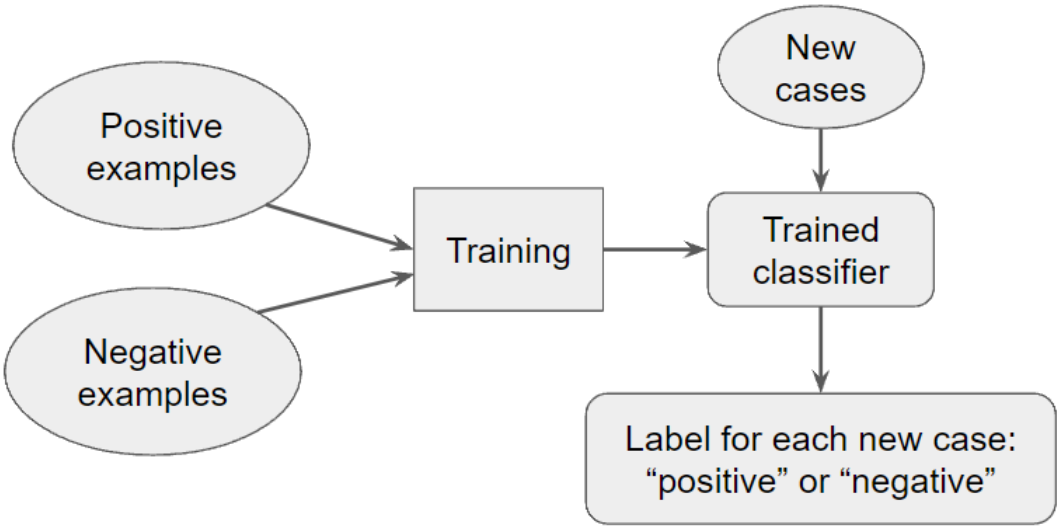

Figure 2.2: Machine learning process. Training is performed with examples from both cases. After training, the classifier or model assigns labels to new instances.

Instead of a set of *positive* and *negative* examples, we only have a subset of the positive examples, which we expect that in reality there are more. The "positive" examples are the bonafide genes that we compiled in our literature search. They are experimentally validated target genes of the Wnt signaling pathway. We expect that there are more target genes and we are looking for them.

Moreover, we do not have real *negative* examples. We only have positive examples and the rest of the genes may or may not be Wnt signaling targets.

To address this issue, we train each classifier using all bonafide genes as "positive" examples, and a *random sample* of non-bonafide genes as "negative" examples (Fig. 2.3). Here we write "positives" and "negatives" in quotation marks to show that they are positives and negatives only for the sake of training, not examples of well-known positives or negatives. Our hypothesis is that there are only a few real target genes. This suggests that a random sample will *probably* include only non-target genes, but sometimes some of them may be real targets.

Since this classification uses random sampling, the results may be spurious. To control this variability, we use a pool of at least 500 (sometimes 1000) independent classifiers, each one with a different selection of random "negative" examples. The idea is to reduce the chance that any particular positive target gene (real but yet unknown, i.e. a novel target) gets used

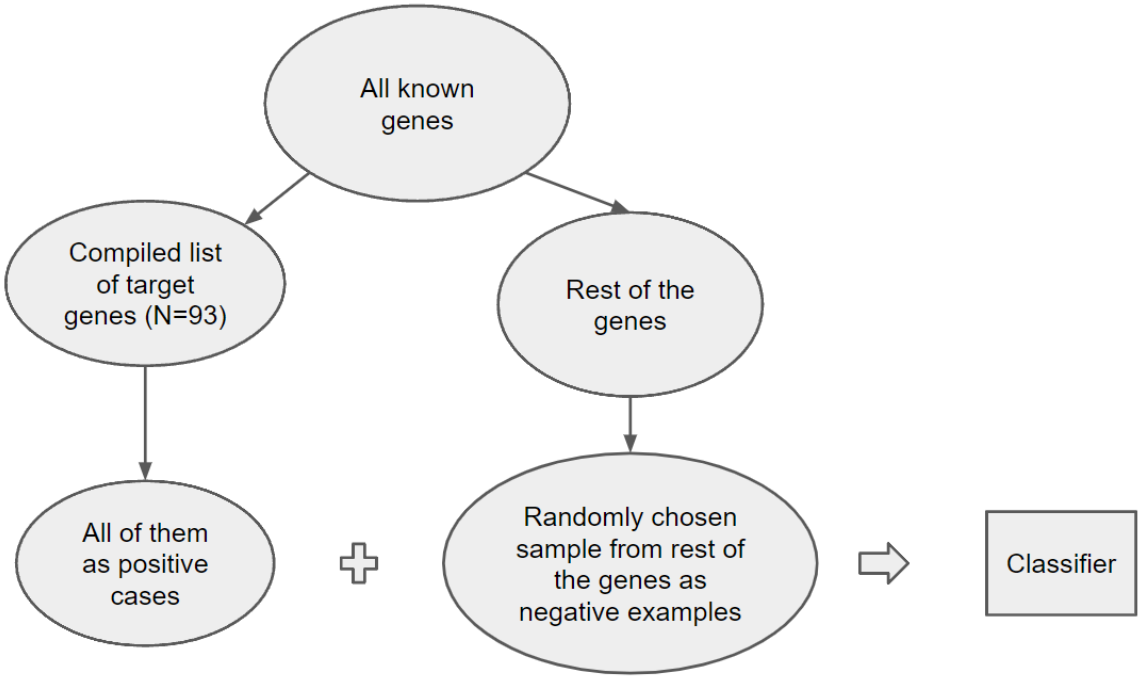

Figure 2.3: Schema of the training process for a single classifier.

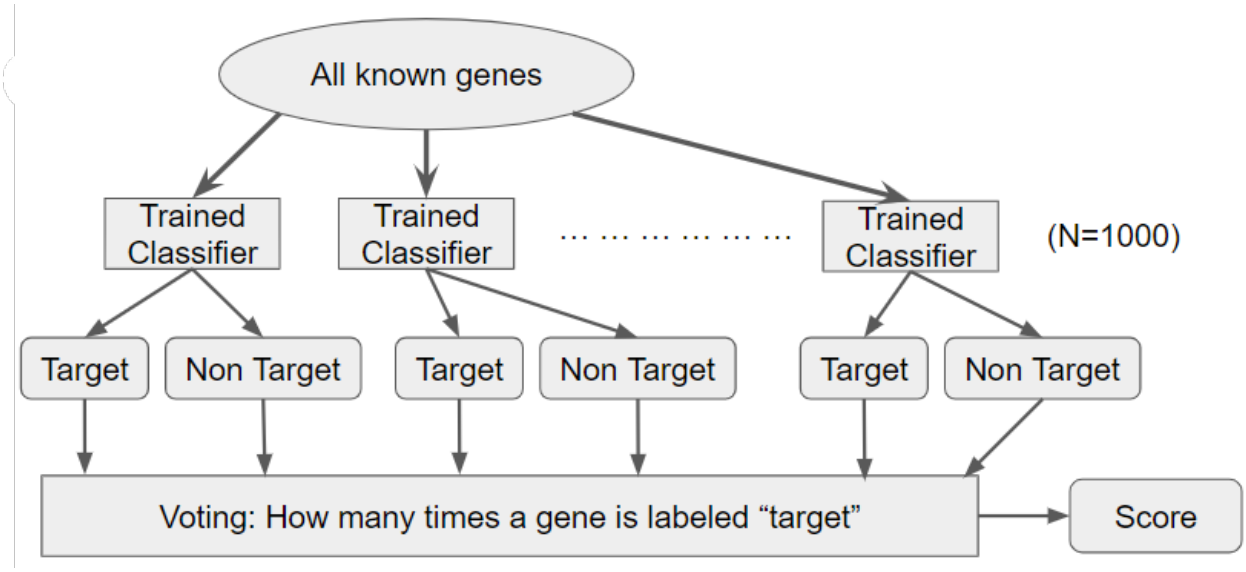

Figure 2.4: Schema of the decision process by the pool of classifiers to find putative target genes. This process is iterated for every combination of sample size and penalty.

many times as a "negative" example. We chose 80 genes as our number of randomly chosen positive targets for all of these training sessions.

2.6.1. Deciding which genes may be targets

Each classifier in the pool decides independently which genes share the characteristics of "positive" examples. At the end of the experiment, each gene gets a score corresponding to the number of classifiers that declared it as a target gene. This score is a number between 0 and 500 (or 1000, depending on the number of classifiers in the pool). Higher scores correspond to higher likelihood of being a target (Fig. 2.4). The final decision is taken by comparing with a score threshold chosen using the Fisher method. The details of this evaluation can be found in Aravena Duarte (2013).

2.6.2. Hyperparameters

There are some parameters that directly impact the machine learning process. These are called hyperparameters. To find the optimal hyperparameters we trained the classifiers with different hyperparameter values.

The first hyperparameter to define is this sample size. We tested samples of sizes 800, 1000, 1200, 1500, 2000, 3000, 5000, and 8000 in different rounds of training that we call experiments.

Since we have a small set of positive examples to be compared with a larger set of negative examples, we risk that a large number of negative examples will bias the results. To compensate for this, we use a loss function with different *penalizations* (sometimes called *weights*) for misclassification. Thus, we introduce a second hyperparameter: the penalty of misclassification of "positive" examples. We tested with 1, 5, 10, 20, 50, 100, and 200 in several rounds of training.

Now we have to define a criteria to determine which hyperparameters are better. In this case, we have two opposite goals:

- We want to minimize the *false negatives*. In other words, we want to recover all (or most) of the bonafide genes in our predictions.
- We want to find some *false positives* but not too many. These will be the novel target genes.

The first criterion is natural. The bonafide target genes are indeed the target genes. The second criterion is based on the premise of this thesis: there are a few Wnt target genes that have not yet been discovered/characterized. Thus, we need to measure two indices for each combination of hyperparameters: the number of novel targets, and the number of bonafide genes found among the novel targets. This last index is called *recall*.

2.6.3. Evaluation of classification results/avoiding overfitting

To measure the effect of the hyperparameters in these two indices, and to verify that there is not any overfitting, we contrasted pools of classifiers trained using the bonafide genes as "positives" versus pools of classifiers trained with a set of genes chosen randomly as "positives".

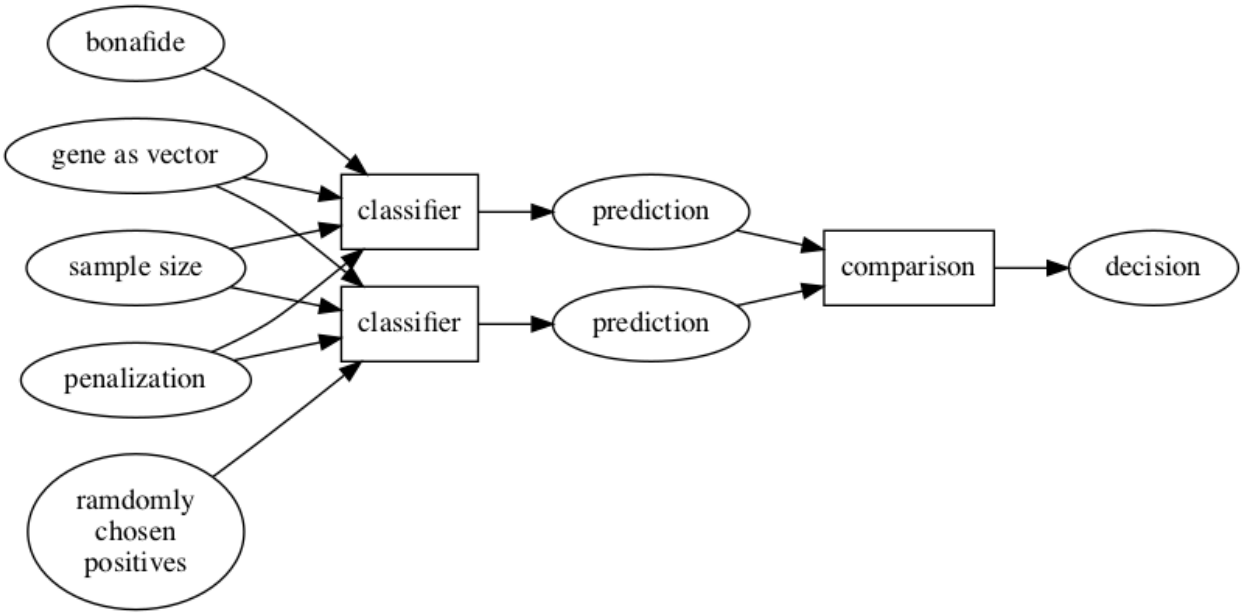

Figure 2.5: Overview of our strategy.

We expect that bonafide target genes should have some common characteristics, the ones that the automatic classifier should learn. On the other hand, any randomly chosen set of genes will *probably* not have any common characteristics so they will not generalize well. This should be reflected in the *novel* and *recall* indices. The good hyperparameters are the ones that satisfy

- *Recall* should be 100% for classifiers trained with bonafide genes. Ideally, all bonafide genes should be recovered.
- *Recall* should be much less than 100% for classifiers trained with random genes. In

other words, we do not want to overfit.

- *Novels* should be more than zero, but not very large. This is for classifiers trained with bonafide genes.
- Bonafides should score close to maximum score, in other words they must be voted as target nearly by all the classifiers.
- The scores of bonafides should not vary much. That being said, the standard deviation of bonafide scores should be close to zero within the same hyperparameters.

We contrasted each bonafide-based classifier against ten random-sample-based classifiers.

2.7. SUMMARY

Our general strategy (Fig. 2.5) for training classifiers can be summarized as follows:

1. Determine a list of bonafide target genes.
    a. Literature search
2. Characterize all genes with a suitable matrix.
    a. Fingerprints of Transcription Factor binding sites
    b. Differential gene expression
3. Choose different values for hyperparameters
    a. sample size
    b. misclassification penalty
4. Train a pool of classifiers using the bonafide list to select “positive examples”, random samples as “negative” examples, and the gene matrix as features.
5. Predict novel target genes using the trained pool of classifiers.
6. Evaluate the prediction quality comparing the results to a negative control trained using randomly selected genes as “positive examples”.
7. Repeat the evaluation for different sample sizes and penalizations.

8. Decide which hyperparameters result in the best outcome.

3. RESULTS

3.1. BONAFIDE TARGET GENES IN THE LITERATURE

Our literature exploration allowed us to identify 93 experimentally validated target genes in the Wnt signaling pathway related to colorectal cancer in humans. Table in Appendix 1 shows our findings. These genes are our reference for training the machine learning classifier. To avoid ambiguity, we call them *bonafide* target genes.

These 93 bonafide target genes have been collected from several sources. Nusse *et al.* reported 33 of them on their website ('Wnt target genes', no date). Another set of 33 genes was reported in the work of Herbst *et al.* (2014). We identified the remaining 29 bonafide targets in our NCBI PubMed search. Each gene has been validated experimentally in the referenced work.

Among these genes, 69 have been reported as direct targets of the pathway, two genes have been reported as indirect targets, and 24 of them have not been reported in either way. Out of 93 genes, 88 of them were reported as upregulated. We found four downregulated genes in the literature. Most of the target genes identified experimentally are up-regulated direct targets. This is surprising since *a priori* we do not have any reason to expect that a signaling pathway triggers only overexpression. This may be due to a bias in the literature, since these cases may be easier to identify experimentally.

3.2. REPLICATION OF HÖDAR *ET AL.* PAPER

To test the classification machine learning tool, we used the same gene characterization and bonafide genes list used in Hödar *et al.* (2010) to build a pool of classifiers expecting to see a result similar to the original paper. This pool of classifiers is trained with the 66 bonafide genes that were known at the time the original paper was written. The genes are represented by a matrix of 432 features according to the work of Ron Shamir's lab (Elkon *et al.* (2003)). These features are transcription factor binding site (TFBS) fingerprints, that is, the number of predicted binding sites in the promoter region. In this case, the authors (Shamir's lab) considered the 1000bp before the transcription start site as the promoter region. There are 15476 genes in this matrix.

In the paper by Hödar *et al.* (2010) authors used a sample size of 8000 and they calculated

the penalty as the sample size divided by the number of bonafide positive examples. Since both of these numbers are fixed, the penalty is always near 120.

In this round of training, we trained pools of 500 classifiers for each hyperparameter combination. We tested sample sizes 1000, 1200, 1500, 3000, and 8000, and penalties of 10, 20, 50, and 100. The classifiers trained with randomly chosen "positives" did not achieve 100% recall for any of the hyperparameter combinations. From this result, we concluded that there is no overfitting. Among the classifiers trained with bonafide "positives", the pools of classifiers have a penalty of 20 achieved 100% recall for all sample sizes (Fig. 3.1 and Fig. 3.2).

We also look at the number of novel genes predicted considering the recall performances across the hyperparameters in detail in Fig. 3.3.

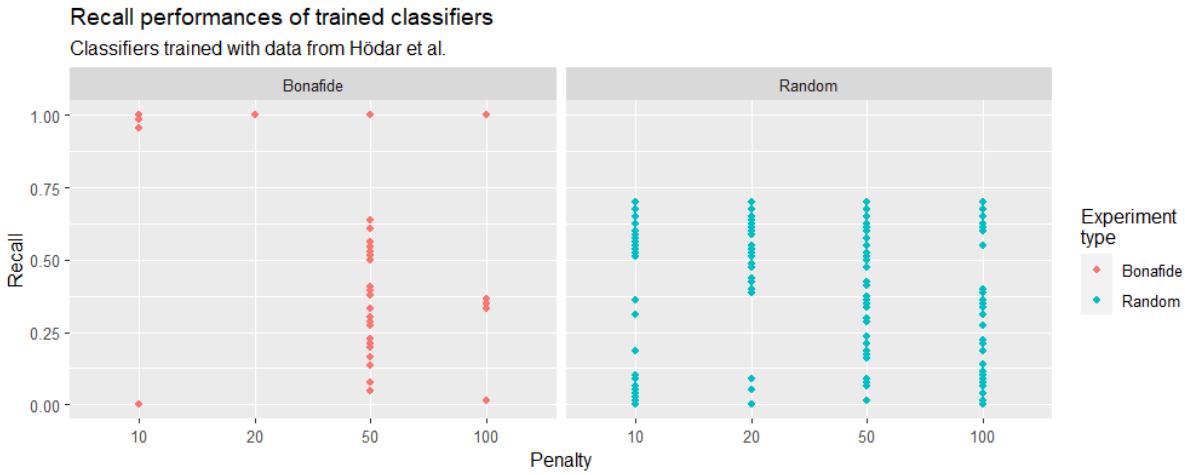

Figure 3.1: Recall performances of the classifiers trained with data from Hödar *et al.* (2010). Recall is the average number of recovered positive cases. For all penalties there are some sample sizes where pools of classifiers recover 100% of bonafide positive examples. Moreover, when the penalty is 20, all sample sizes resulted in 100% recall of the bonafide positive cases. On the other hand, all pools of classifiers trained with randomly chosen positives only achieved less than 75% recall.

For penalty 10 and sample sizes 1200 and 1500, recall is zero. On the other hand, sample sizes 1000, 3000 and 8000 the recall is 100%. This is consistent through all cases but surprising. For penalty 100 we observe that small sample sizes result in zero recall. The only case with 100% recall is sample size 8000, and separation of bonafide and randomly chosen positive is only good in this case. For penalty 20, all sample sizes performed well, because recall is 100% and there is no overfitting. The best cases are either any sample size with penalty 20, or sample sizes 3000 and 8000 for penalty 10 or 50.

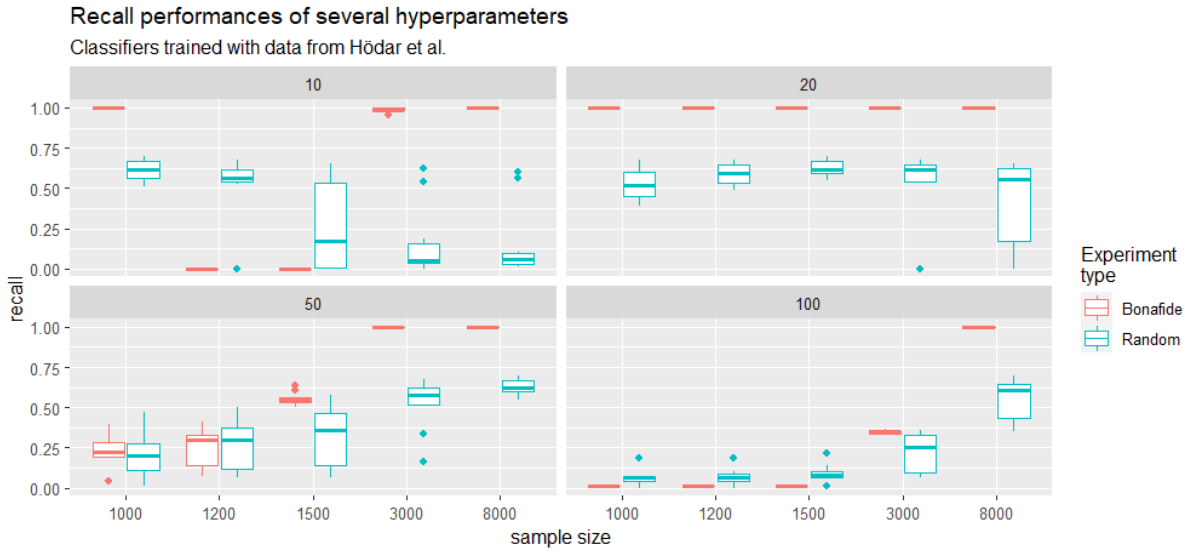

Figure 3.2: Recall performances of several hyperparameters in classifiers trained with data from Hödar *et al.* (2010). The headers of each facet shows the penalization applied during training. There are two cases resulting in zero recall for penalty 10 where sample sizes are 1200 and 1500. For penalty 20 all sample sizes resulted in 100% recall. For penalties 50 and 100, larger sample sizes resulted in 100% recall. The recall performances in these penalties increase with larger sample sizes.

3.2.1. Novel genes

Here we include a brief discussion of some of the novel genes predicted using sample size 8000 and penalty 100, to compare with the original paper. We checked the results of classifiers trained with sample size 8000 and penalty 100 (since these were closest to the hyperparameters used in the paper by Hödar *et al.* (2010)) to see if there are any similar predictions to the original paper. Classifiers trained with these hyperparameters produced 53 genes in total. None of them were shared with the original paper's predicted genes. These genes are listed in the Appendix 2 .

We chose the classifiers trained with sample size 1000 and penalty 20 as best performing of them all. Classifiers trained with these parameters produced 115 novel genes. The WNT2 was included in these predictions. The list of these novel genes included in the Appendix 3 .

3.3. CLASSIFICATION WITH UPDATED BONAFIDE LIST

Since 2010 when Hödar *et al.* (2010) was published, new target genes have been identified. Thus, it is natural to ask if this new information can lead to discoveries.

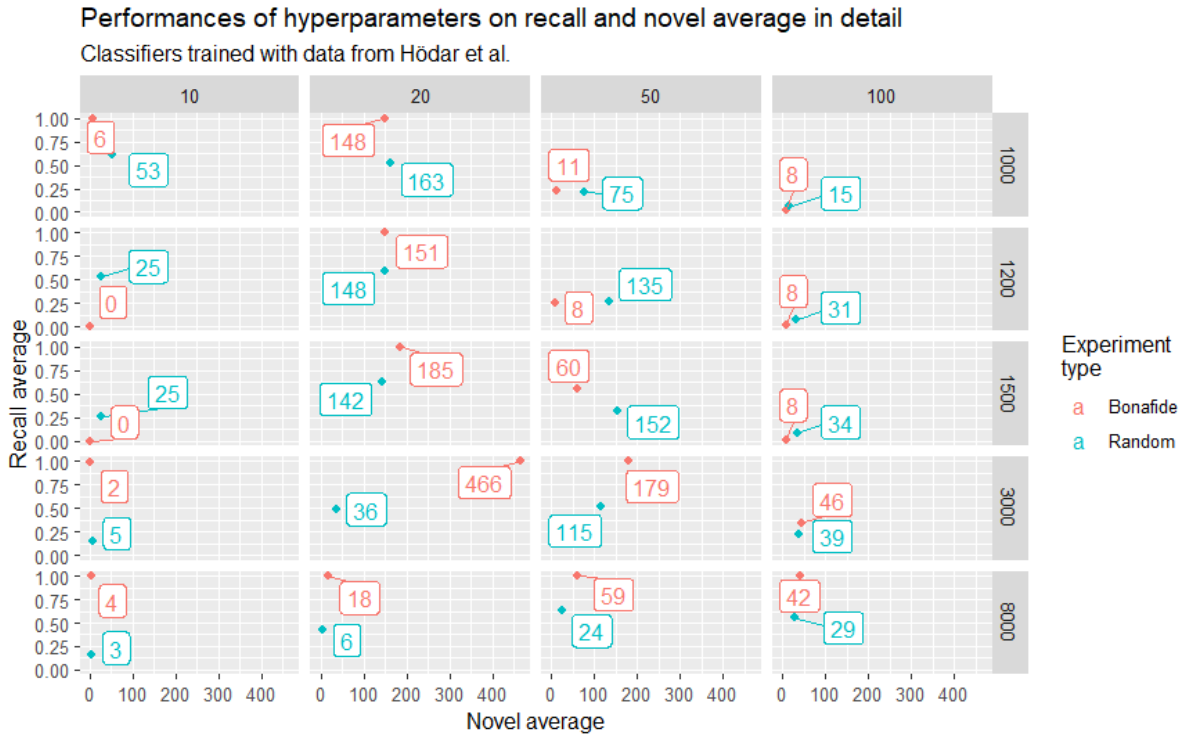

Figure 3.3: Performance of hyperparameters on recall and novel average in classifiers trained with data from Hödar *et al.* (2010). Each facet shows pools of classifiers trained with a set of hyperparameters. The headers of each facet show the penalty applied. The numbers on the right-hand side shows the sample sizes. Penalty 20 is consistently performed well in all sample sizes and produced a decent number of predictions. Sample size 8000 is also performed 100% in all applied penalties.

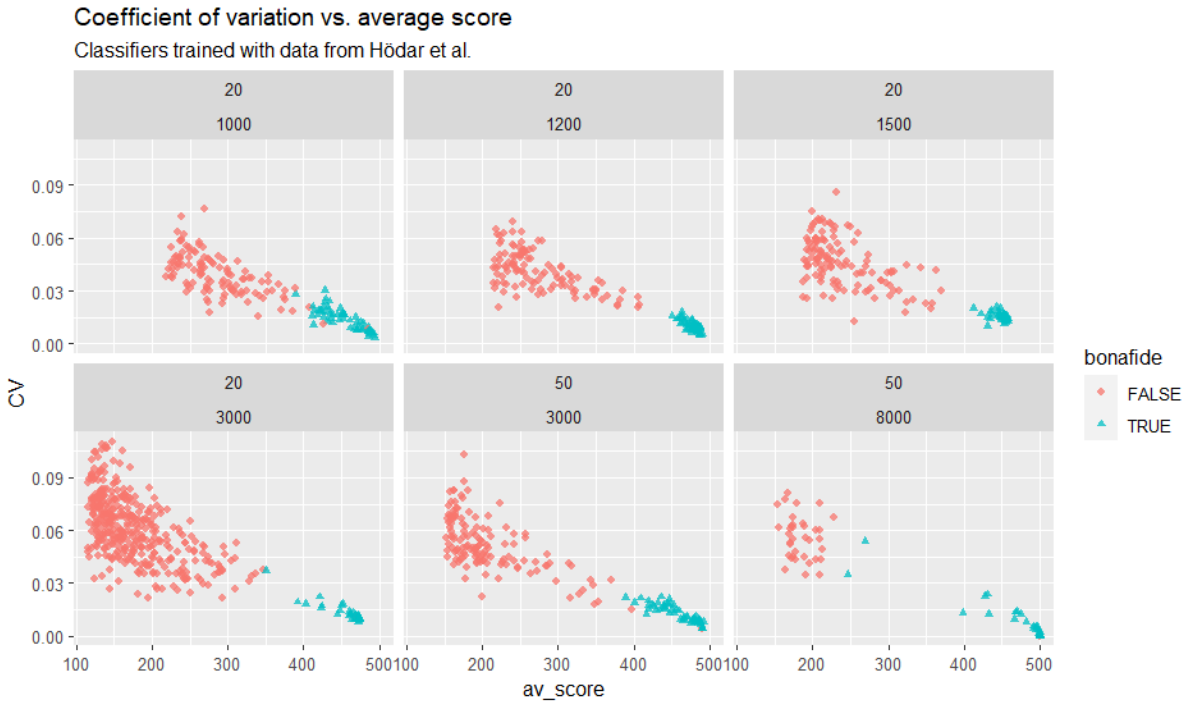

Figure 3.4: Coefficient of variation vs. average score for several hyperparameters in classifiers trained with data from Hödar *et al.* (2010). Each facet shows pools of classifiers trained with one set of hyperparameters and each point indicates a gene. The headers of each facet show penalty and sample size. Sample size 8000 produced significantly low numbers of novel predictions compared to other sample sizes.

We trained a new set of classifiers with the original table of transcription factor binding site fingerprints and 88 of the 93 bonafide targets that we compiled (Table 3.1). There are five genes in our list of bonafide targets that are not represented in the original matrix (432 columns, made in 2003). This is due to changes in the nomenclature of human genes between the time the gene expression experiments were carried on and the time the papers were written.

In this round of training, we trained pools of 1000 classifiers for each hyperparameter combination.

Table 3.1: Bonafide genes that are missing in the TFBS 432 fingerprint matrix.

Entrez ID (Locus ID)	**Symbol**
339287	MSL1
389058	SP5
4513	MT-CO2
55740	ENAH
9232	PTTG1

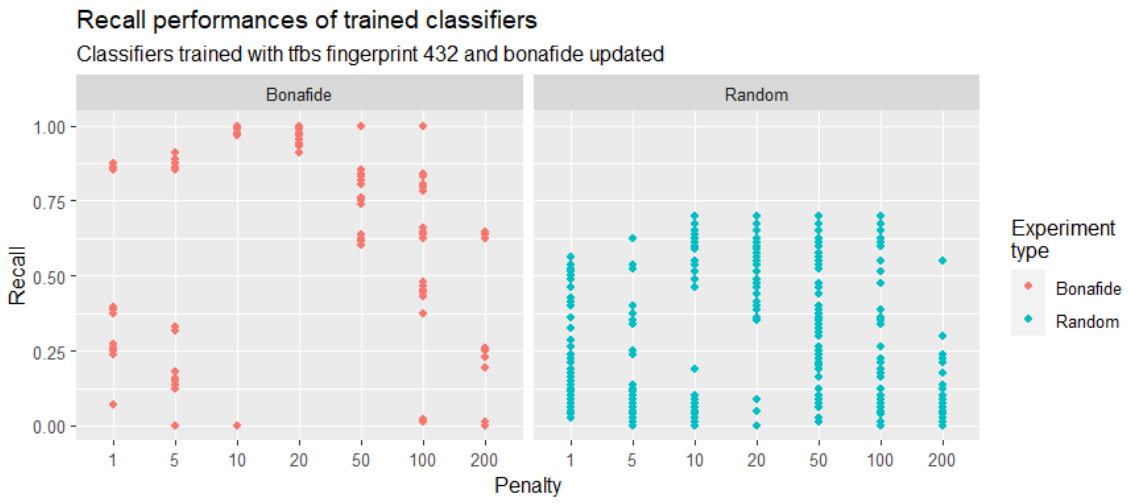

Figure 3.5: Recall performances of the trained classifiers trained with TFBS fingerprint and updated bonafide list. The facets indicate the training for positive cases as bonafide genes or randomly chosen genes. The recall is indicating the recovery of positive cases in each training. Each point corresponds to a pool of classifiers. Some of the classifiers trained with 10, 20, 50 and 1000 achieved 100% recall for bonafide targets as positive cases. Rest of the penalties did not achieve 100% recall for this type of training. For the training with randomly chosen positive cases none of them achieve above the 75% recall.

There are combinations of hyperparameters that result in 100% recall when trained using

bonafide genes as “positive” examples. The recall is under 75% for all the classifiers trained with randomly chosen “positive” genes, as can be seen in Fig. 3.5. A 100% recall is achieved (at least by some of them) only when using penalties 10, 20, 50, and 100. None of the classifiers trained with penalties 1, 5, and 200 achieved 100% of recall. The classifiers tend to perform better with penalties that are around 10 and 20, and there is a decline in this trend over 50.

Although none of the classifiers trained with randomly chosen genes achieve 100% recall, we see a similar trend in these classifiers regarding the penalties and recall performances.

Most of the classifiers trained with bonafide target and penalty 10 performed well, but the ones trained with sample size 1500 performed poorly (Fig. 3.5 and Fig. 3.6).

We chose the hyperparameters with the best results and focused on these to see how many genes are classified as a novel target by these classifiers (Fig. 3.7). We chose classifiers trained with penalties 10, 20, and 50 across all sample sizes. Then we chose some of the best performing cases to see the effect of hyperparameters on average score and variation of the average score (Fig. 3.8, Fig. 3.9 and Fig. 3.10).

For penalty 10, sample sizes 800, 3000 and 8000 resulted in good numbers of novel predictions (not too large not too small). The other sample sizes for this penalty resulted with either too large or too small predictions. In Fig. 3.8 for sample sizes 800 and 3000 all the predictions including bonafide scored under 600. For sample size 8000 some of the bonafides are close to score 1000. But all the novel predictions got a score under 500 and most the bonafide got over 500.

For penalty 20, sample size 8000 produced a number of predictions that is close to our criteria. Sample size 3000 produced a small number of novel predictions contrast to the rest of the sample sizes which were too large. The one that satisfies our criteria was sample size 8000. According to Fig. 3.9, for sample size 8000 all the novel predictions got a score under 500 and all the bonafide positives over 500.

For penalty 50, sample sizes 3000 and 5000 produced a large number of novel predictions. Sample size 8000 satisfies our criteria for the number of novel predictions.

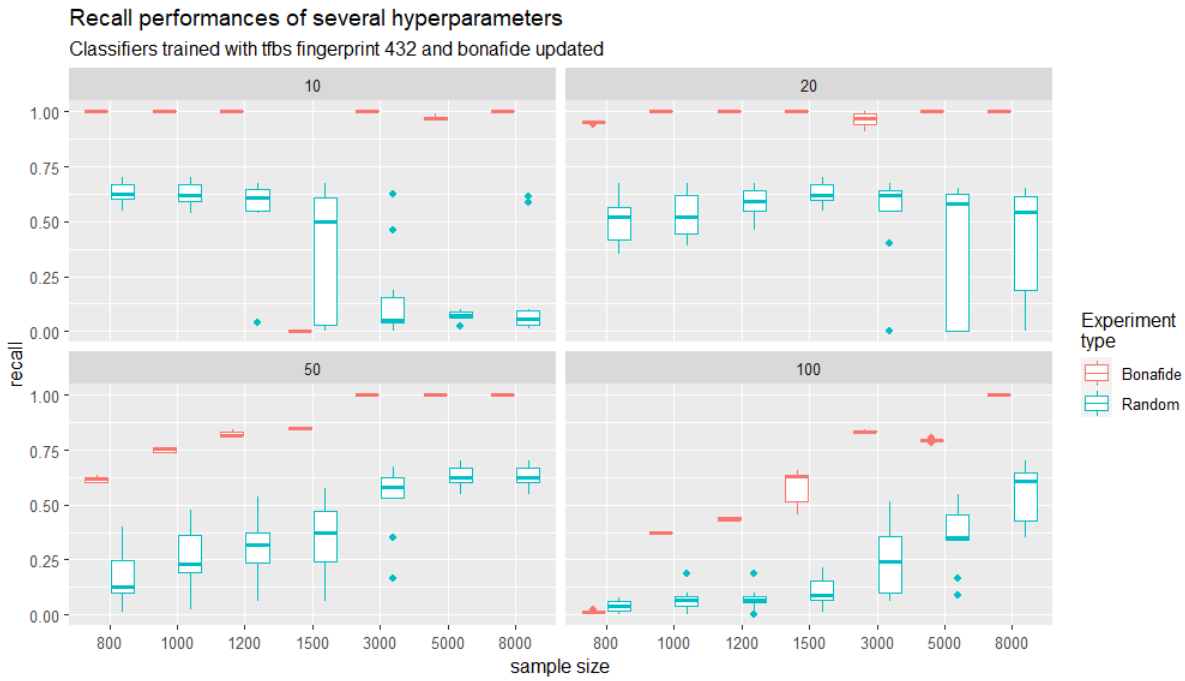

Figure 3.6: Recall performances of the hyperparameters in classifiers trained with TFBS fingerprint and updated bonafide list. Headers on top of each facet indicate the penalty applied. Recall is the percentage of positive cases recovered in each training. None of the pools of classifiers trained with randomly chosen "positives" achieve 100% recall. For penalties 50 and 100, the recall percentage increases with larger sample sizes. For penalties 10 and 20 generally 100% of the bonafides recovered. There is one particular case where sample size 1500 and penalty 10 resulted in 0 recall.

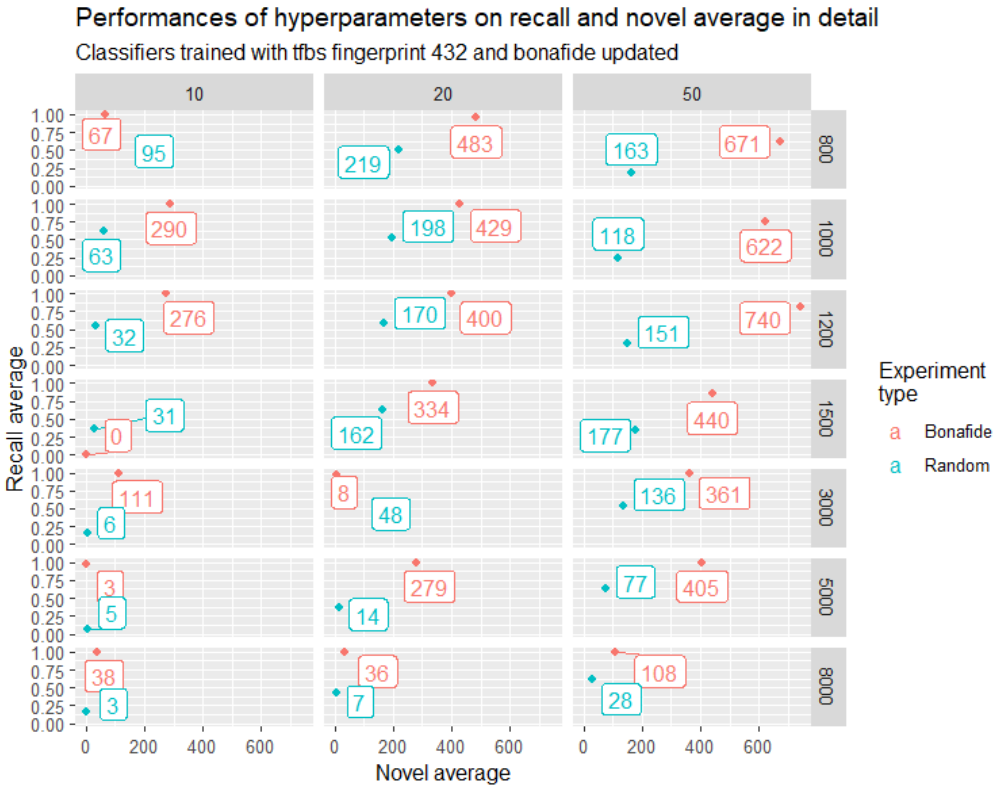

Figure 3.7: Performance of hyperparameters on recall in detail in classifiers trained with TFBS fingerprint and updated bonafide list. The headers on top indicate the penalty applied and the numbers on the left-hand side indicate sample sizes. Each box shows the performance of one penalty and sample size combination. Usually the number of novels produced decreases with larger sample sizes. For sample size 1500 and penalty 10 resulted in 0 recall and produced 0 novel genes.

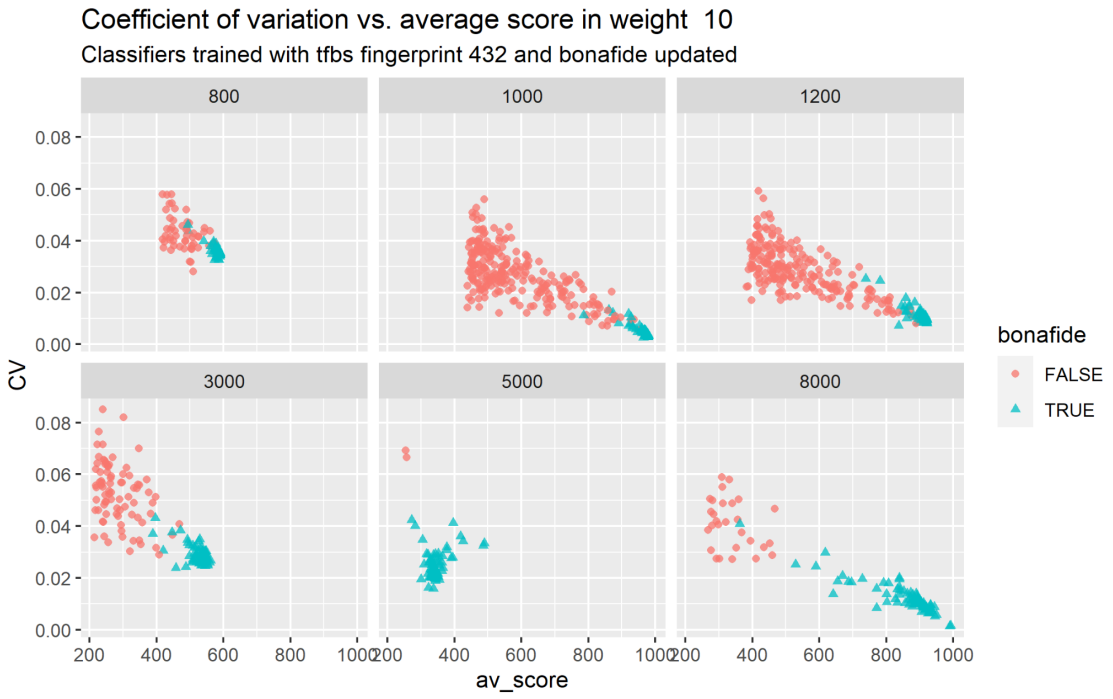

Figure 3.8: Coefficient of variation vs. average score in classifiers trained with weight 10, TFBS fingerprint and updated bonafide list. The headers on top of each facet indicate the sample size. For sample size 800 the scores accumulate around 500. Sample sizes 1000 and 1200 have similar results. Both resulted in high scores for bonafide. For sample size 3000 all predictions are under score 600. For sample size 5000 bonafides are accumulated around 300 with a few novel predictions. For sample size 8000 most bonafides achieved high scores with minor exceptions and novel predictions are mostly around 300 score.

Figure 3.9: Coefficient of variation vs. average score in classifiers trained with weight 20, TFBS fingerprint and updated bonafide list. The headers on top of each facet indicate the sample size. Sample sizes 800, 1000, 1200 and 1500 produced similar results: nearly all the genes scored above 400 with low variance. For sample size 3000 predictions accumulated around score 200 with few predictions. Sample size 5000 produced a lot of novel predictions but most of them have scores under 400. For sample size 8000, all bonafide genes have scores over 600 and all the novel predictions are under 600.

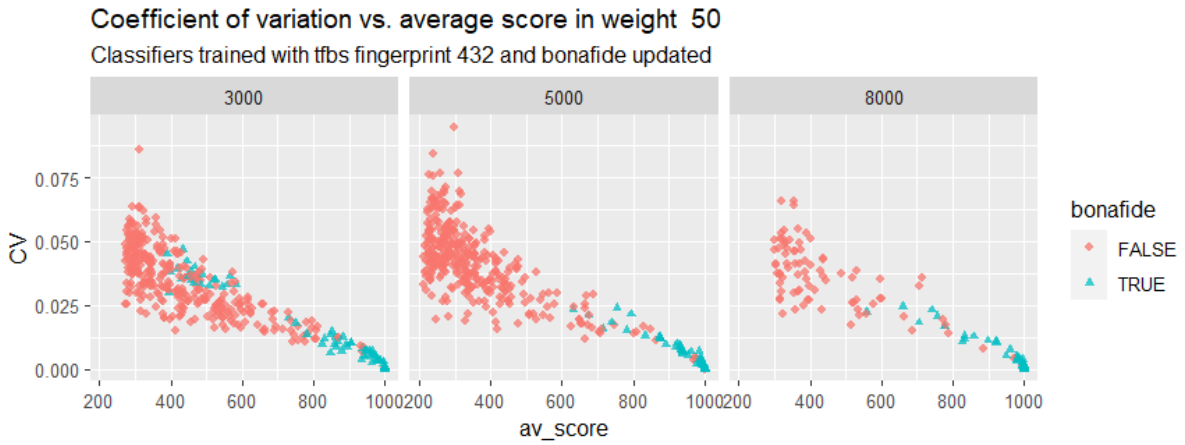

Figure 3.10: Coefficient of variation vs. average score in classifiers trained with weight 50, TFBS fingerprint and updated bonafide list. The headers on top of each facet indicate the sample size. Bonafide target genes are more scattered for sample size 3000 compared to sample sizes 5000 and 8000.

3.3.1. Novel genes

Classifiers trained with hyperparameters sample size 8000 and penalty 50 satisfies our criteria. We identified 85 novel predictions as the result of classifiers trained with these hyperparameters. The five of the highest scoring genes were RAB32, GET4, FANK1, BARX2 and COL4A2. The rest of the predictions include WNT2 and WNT10A. The full list is Appendix 4.

3.4. CLASSIFICATION USING DIFFERENTIAL GENE EXPRESSION DATA

There are two new matrices containing the statistical analysis results of gene expression data. The first one contains only the statistical analysis for RNA-seq samples, and the second one contains both RNA-seq and microarray samples.

After being processed, gene expression data has six columns (variables) for each contrast. The columns are: log fold change, average expression, t statistic, *p*-value, adjusted *p*-value, and B statistic.

3.5. RESULTS FOR TRAINING WITH RNA-SEQ DATA

The first training set with processed gene expression data contains the statistical analysis from eight RNA-seq experiments. This matrix represents 22667 genes under 23 contrasts. When training with this matrix we used sample sizes 1000, 1200, 1500, 3000, and 8000. For the penalty, we used 5, 10, 20, 50, 100, and 200. In this round of training, we trained pools of 1000 classifiers for each hyperparameter combination.

Some of the classifiers trained with randomly chosen positives achieved 100% recall (Fig. 3.11). This suggests that there is an overfitting of some of the cases, especially where the penalty is 10 or 20, as we can see in Fig. 3.12. For penalty 50 there is an overfitting only when sample size is 8000 (Fig. 3.12), but with the rest of the sample sizes there is no overfitting.

We focus on the best results in Fig. 3.13. We choose the cases where penalties 10, 20, and 50 combinations with all the sample sizes since all of these achieved 100% recall. Since cases where the penalty is 20 classifiers showed overfitting, we looked at the penalty 10 and 50 to see how they impacted on the average score (Fig. 3.14 and Fig. 3.15).

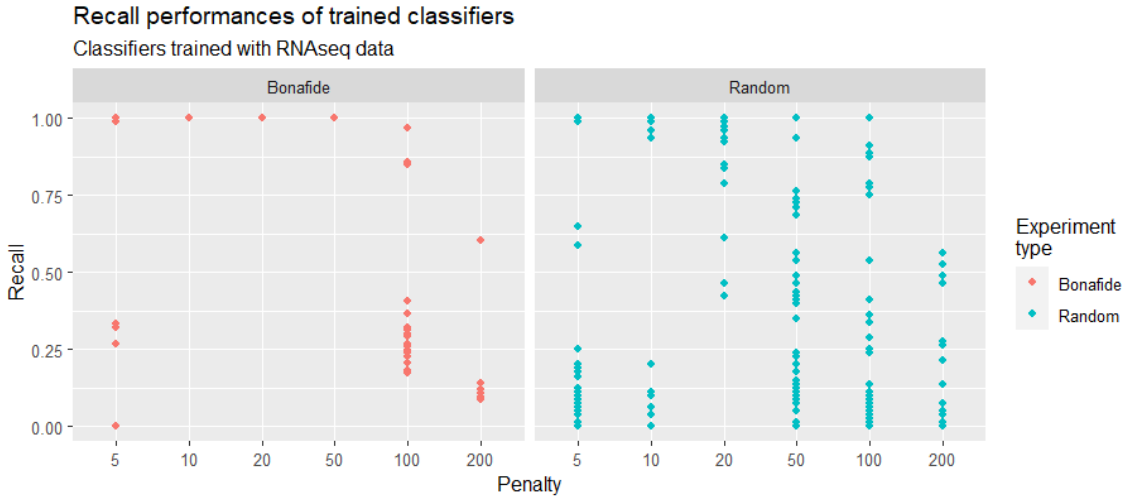

Figure 3.11: Comparison of the recall performances of classifiers trained with RNAseq data. Classifiers trained with gene expression show different behavior compared to classifiers trained with TFBS fingerprints. Some of the classifiers trained with random "positives" achieved 100% recall, suggesting that there is overfitting. For penalties 10, 20 and 50, all classifiers trained with bonafide examples achieved 100% recall. For penalty 200, classifiers did not achieve 100% recall in any case.

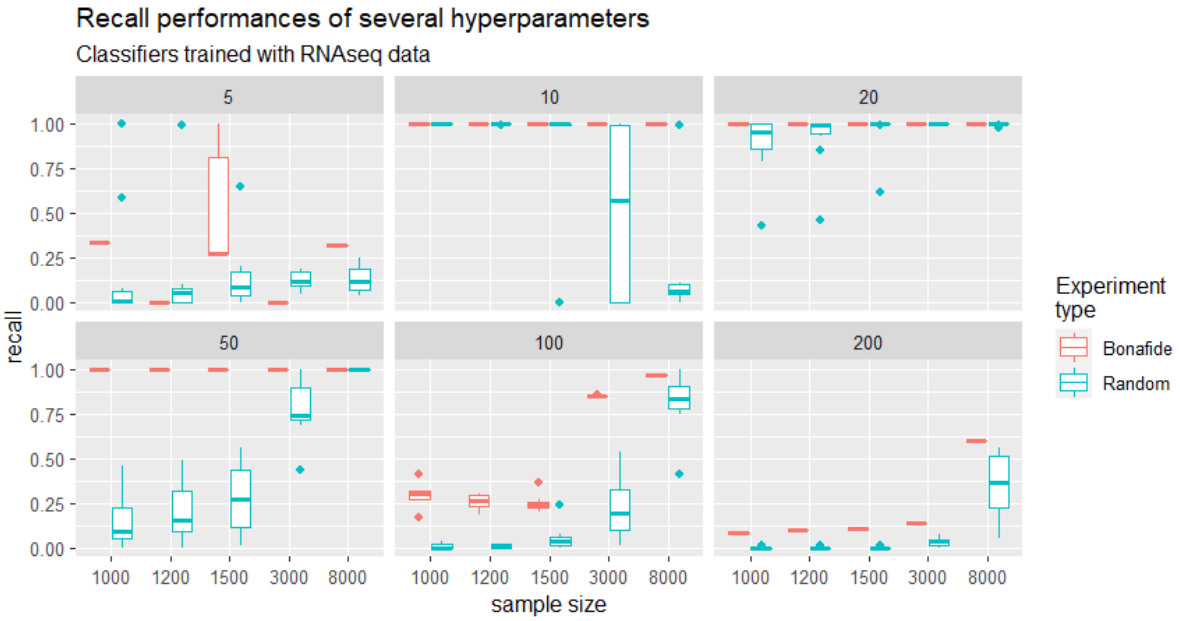

Figure 3.12: Recall performances of classifiers trained with RNAseq data depending on sample size and misclassification penalty. The headers on top indicate the corresponding penalty. For penalties 10, 20, and 50, classifiers trained with bonafide "positives" achieved 100% recall. For penalties 10 and 20, classifiers trained with random positives also achieved 100% recall, suggesting that there is overfitting. On the other hand, classifiers trained with random positives with penalty 50 did not achieve 100% recall when the sample size was under 3000. This suggests that penalty 50 can be used to predict novel target genes.

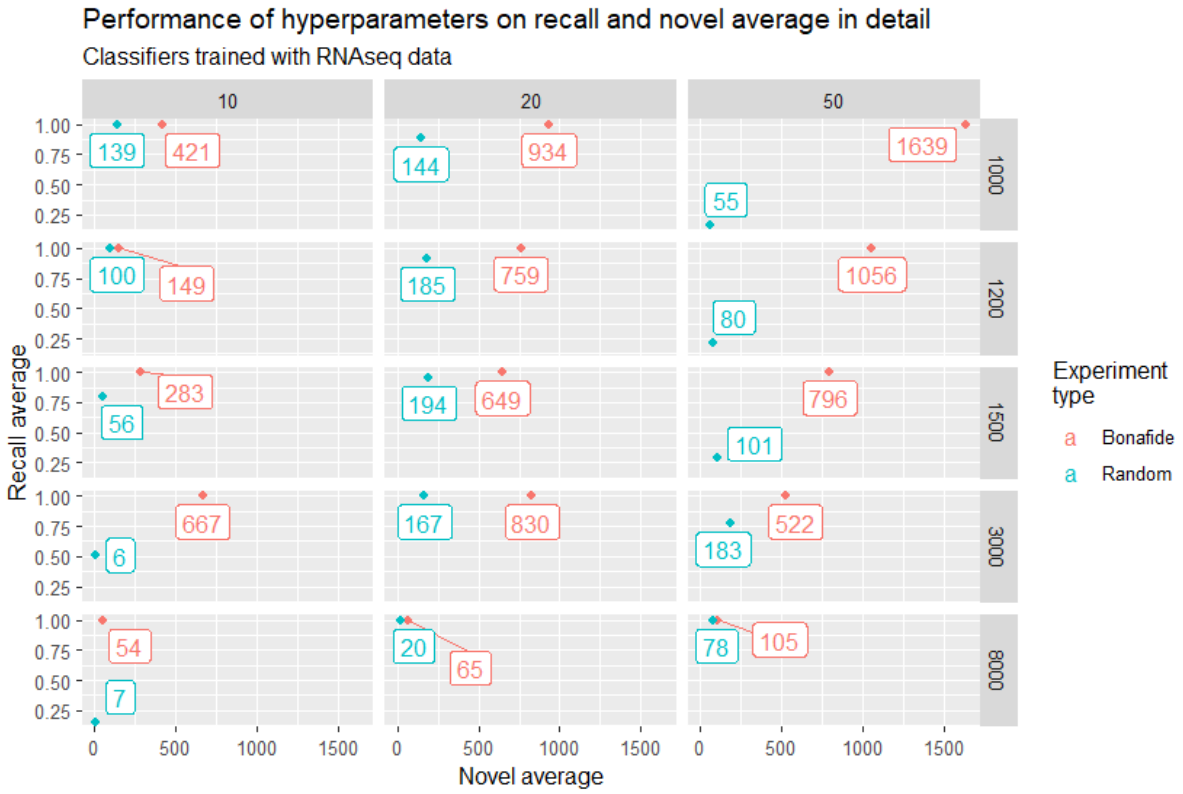

Figure 3.13: Average number of novel targets and recall for classifiers trained with RNAseq data using different sample sizes and penalties. The headers on the top and right of the facets show the penalty and sample size. Each point represents the average of 10 iterations of training with the same data and hyperparameters. The labels show the average number of novel genes. The high values of recall for classifiers trained with random "positives" indicate that penalty 20 results in overfitting. The same is true for penalty 10 with small sample sizes, but there is no evidence of overfitting for sample size 3000 and 8000. Penalty 50 results in no-overfitting most of the training and a large number of novel genes.

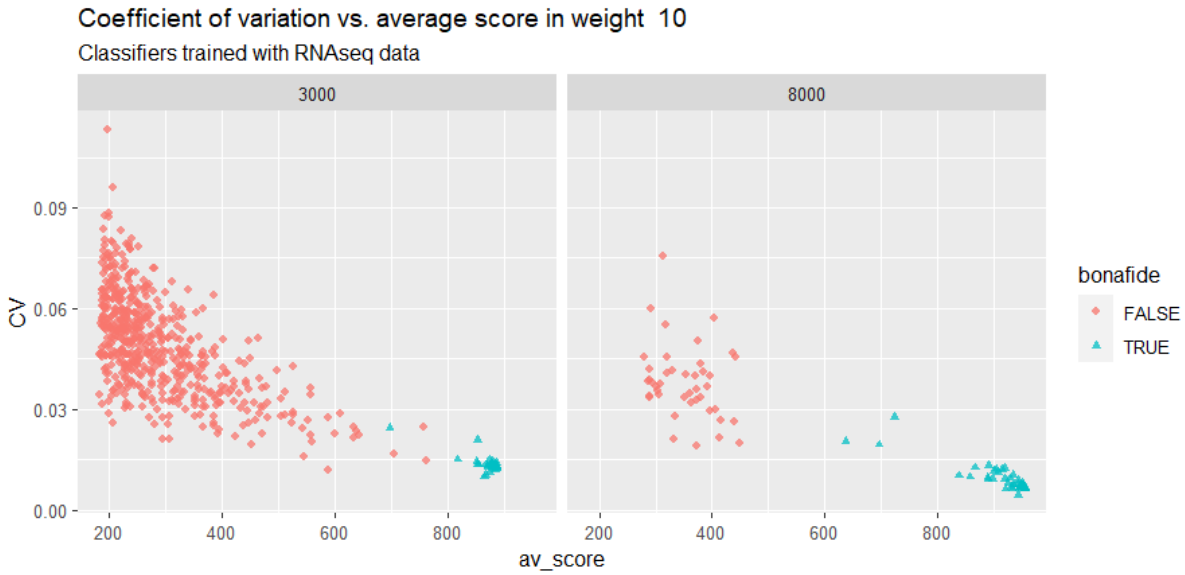

Figure 3.14: Coefficient of variation vs. average score in classifiers trained with weight 10 and RNAseq data. We show only the sample sizes that do not result in overfitting. For sample size 3000, most bonafide genes are scored over 800 but the novel predictions are scored mostly below 500. For sample size 8000 bonafide genes scored around 900 and novel predictions scored between 300 and 400.

3.5.1. Novel genes

We chose the best case as the classifiers trained with sample size 3000 and penalty 50. With these hyperparameters all the bonafide genes are scored above 500. In other words, they are classified as targets by at least 50% of the classifiers. Some of the predicted novel targets scored over 900 when training with these hyperparameters.

This pool of classifiers predicted an average number of 522 novel genes. We focus on novel genes with higher scores. We chose a simple criteria of considering the novel genes with scores greater than 500. In other words, we focused on genes that are classified as targets by at least 50% of the classifiers. This criteria resulted in 144 novel target genes. The highest scoring genes are GLI3, HOXB3, SOX4, TIMP3, and PTCH1. The rest of the list includes CDH1, FN1, VIM, and TBX3. The filtered list is in the Appendix 5. The complete list is available at the github page : *https://github.com/cemrekefeli/cemre-kefeli-WNT*.

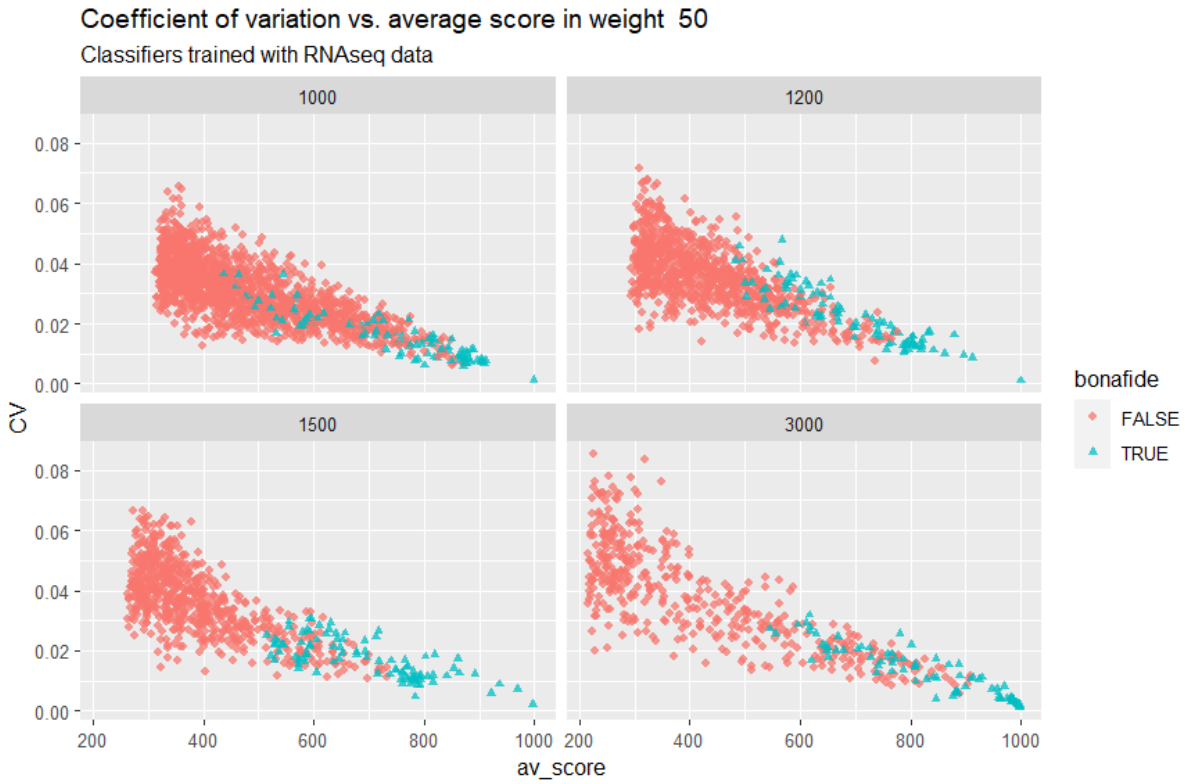

Figure 3.15: Coefficient of variation vs. average score in classifiers trained with RNAseq data and penalty 50. The headers on top of each facet indicates the sample size. We show only the sample sizes that do not result in overfitting. For all sample sizes, most bonafide genes are spread between the scores 500 and 1000. Sample sizes 1000, 1200, and 1500 produce similar results. For sample size 3000 novel genes are more scattered compared to the other sample sizes.

3.6. CLASSIFICATION USING RNA-SEQ AND MICROARRAY DATA COMBINED

We created another matrix similar to the previous one that contains RNA-seq experiments, but this time we combined them with data from microarray experiments. This matrix has 19073 rows and 176 columns, including the first one containing the gene identifiers.

To train with this data, we used sample sizes 1000, 1200, 1500, 3000, and 8000, and penalties 10, 20, 50, and 100. In this round we trained pools of 500 classifiers. The matrix used for this round of training contains 92 of the bonafide genes. The bonafide gene MT-CO2 is not represented in this matrix.

None of the classifiers trained with random positives achieved 100% recall. Classifiers trained with penalties 10, 20, and 50 performed well combined with all sample sizes compared to penalty 100 (Fig. 3.16).

Figure 3.16: Recall performances of the classifiers trained with RNAseq and Microarray data. For penalties 10, 20, and 50 classifiers trained with bonafide positive examples achieved 100% recall. Thus, we discard penalty 100. None of the classifiers trained with random positive examples achieved 100% recall, so there is no sign of overfitting.

3.6.1. Novel genes

The best hyperparameters were sample size 1000 and penalty 50. Classifiers trained with these hyperparameters produced 619 novel predictions. These genes were all present in each shuffling instance. Using the criteria from the previous section we reduced the novel genes to focus by only choosing the ones that get the 50% of the votes. In other words the genes that scored 250 and higher.

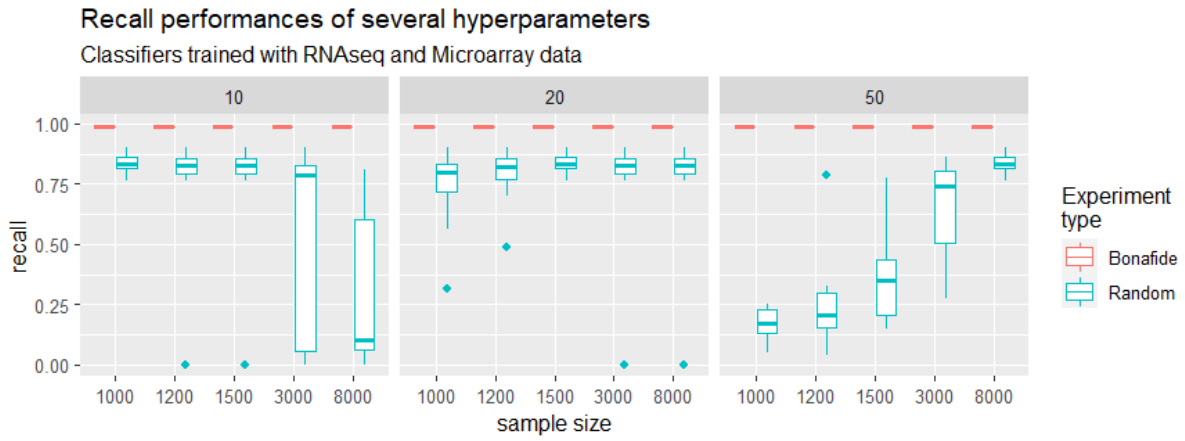

Figure 3.17: Recall performances of the classifiers trained with RNAseq and Microarray data for different sample sizes and penalties. The headers on each facet show the penalty applied. Classifiers trained with bonafide "positives" all achieved 100% recall. Recall is high for most of the classifiers trained with random "positives", except for small sample sizes and penalty 50. For penalty 10 and 20 recall for randomly chosen "positive" cases are very close to 100%. On the other hand for penalty 50 and small sample sizes recall decreases for these cases.

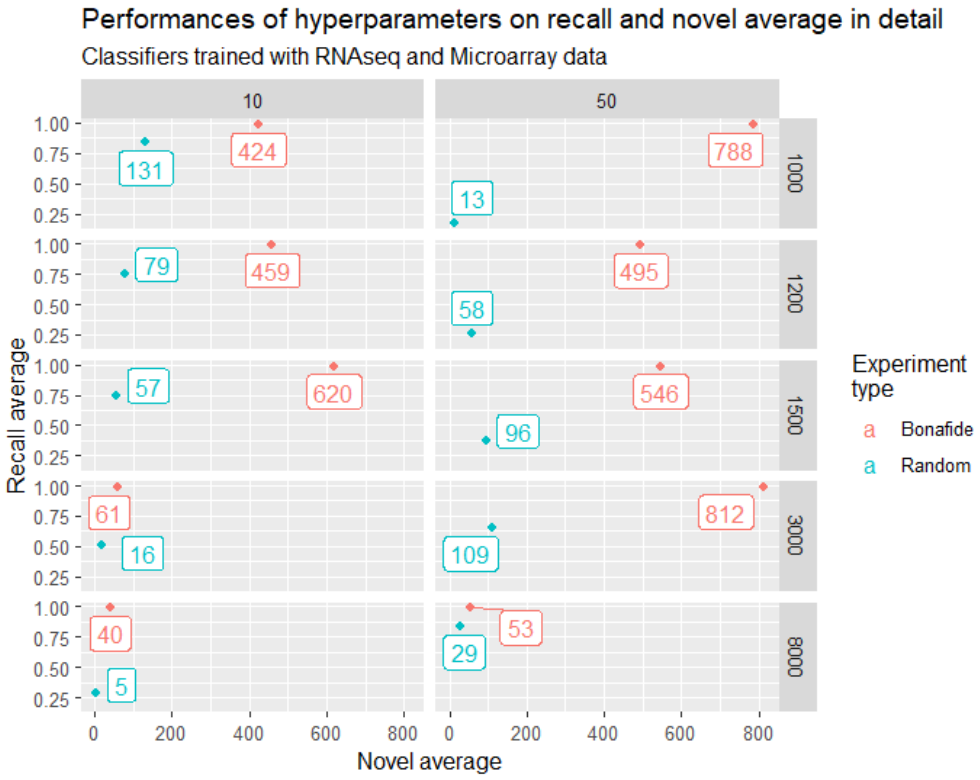

Figure 3.18: Average number of novel targets and recall for classifiers trained with RNAseq and Microarray data using different sample sizes and penalties. The number on top of each facet indicates the penalty and the numbers on the right indicate the sample sizes. Good cases are penalty 10 with sample size 8000, and penalty 50 with sample size between 1000 and 3000.

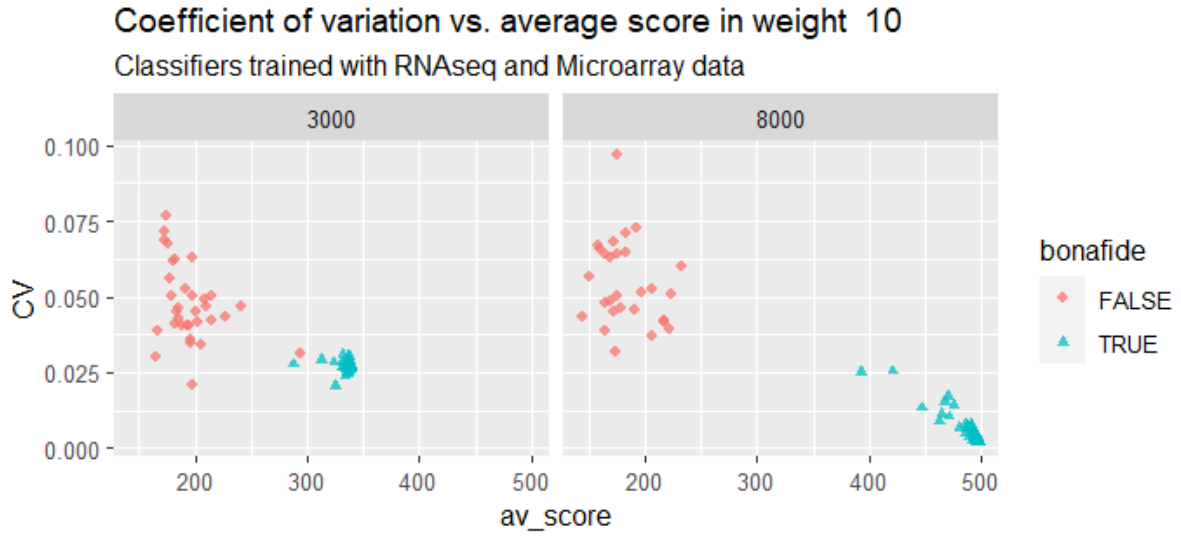

Figure 3.19: Coefficient of variation vs. average score in classifiers trained with RNAseq and Microarray data and penalty 10. For sample size 3000 the bonafides scored between 300 and 350 and novel predictions are around 200. For sample size 8000 bonafide genes scored mostly 500 and novel predictions are around 200.

Figure 3.20: Coefficient of variation vs. average score in classifiers trained with RNAseq and Microarray data and penalty 50. The best cases are 1000 and 1200. Notice that, when bonafide and novel have similar scores, then novel genes have lower variance (higher coherence).

This filtering process resulted in 179 novel predictions. The genes PALD1, UMPS, SAPCD2, PTCH1 and SLC35F2 are the highest scoring genes. The rest of the novel prediction includes TCF7L2, LRP8, SOX4, SOX12 and CDH1. The filtered list is in the Appendix 6. The complete list is available at the github page: *https://github.com/cemrekefeli/cemre-kefeli-WNT*.

3.7. FINAL RESULTS

3.7.1. Some trends of hyperparameters

Sample size 3000 usually produces the largest numbers of novel target genes, but the majority of them have low scores. Also, most of the time this sample size results in the highest coefficient of variation. This might be due to lower scores allowing higher coefficients of variation. Sample size 8000 tends to produce a lower number of novels compared to other sample sizes.

Out of all the penalties, with 5 and 1, the recall has an unexpected behavior. They do not follow a monotone path. In other cases, the recall either increases with the sample size, or decreases with sample size. Pools of classifiers that have penalty 200 do not achieve 100% recall.

In both of the training sessions with TFBS fingerprints, penalty 10 resulted in 100% recall for all sample sizes, except for sample size 1500, which gets 0%. This is surprising because we expected that recall changes monotonically, and this is not the case. Moreover, if for some condition the recall is under 100%, we still expected it to be high. Nevertheless, the recall is 0% in these instances. This suggests that this behavior depends on the data used for training. In particular, the TFBS matrix.

4. DISCUSSION

In this chapter we discuss the meaning of the results presented in the previous chapter. We find that many of the novel target genes we predict have been independently validated by experiments.

At a first glance it may seem paradoxical to have known target genes that are not in our bonafide targets list. This is due to two reasons.

First, although we tried our best in the initial literature search, there is always the possibility that we missed something. We focused our search on the colorectal cancer context. Thus, we missed target genes described in other contexts. Turns out, many of our proposed targets have been described as Wnt targets in other contexts. We will discuss them in the rest of this chapter.

The second reason why we do not have all known target genes in our bonafide list is because we need an independent set of genes for cross validation. This is a standard practice in Machine Learning, used to control the overfitting. In other words, we want to characterize the target genes in a way that can generalize and extrapolate to new cases. This can be tested using some positive cases that are not used in the training.

4.1. WNT-HEDGEHOG CROSSTALK

In our search of novel Wnt target genes, we predicted several lists of candidate target genes as the result of 4 different training sessions/experiments.

Two interesting predictions came from the classifiers that were trained with the differential expression data. These two genes, PTCH1 and GLI3, are the key components of Hedgehog signaling. This may indicate a crosstalk between the Wnt and the Hedgehog signaling pathways, at least in the colorectal cancer context.

Hedgehog signaling is one of the pathways that orchestrate embryonic development. It was discovered in the 1980s by Nusslein-Volhard and Wieschaus alongside the Wnt signaling. Hedgehog signaling starts with the ligand of the same name. Three types of Hedgehog ligand have been discovered to this day: Indian, Desert, and Sonic Hedgehog. Sonic Hedgehog is

the most studied one.

PTCH1 is the receptor for this pathway, and GLI3 is responsible for transcriptional regulation of the Hedgehog target genes where a Hedgehog signal is absent. When the Hedgehog signal is not present, the receptor *Patched-1* (PTCH1) suppresses the Smoothened (SMO), thus the signaling pathway is in the inactive state. When a Hedgehog signal is present, it binds to the receptor PTCH1. This action removes inhibition over the SMO. Then SMO activates downstream players which are GLI transcription factors. These transcription factors can bind directly to DNA and can regulate the target gene expression.

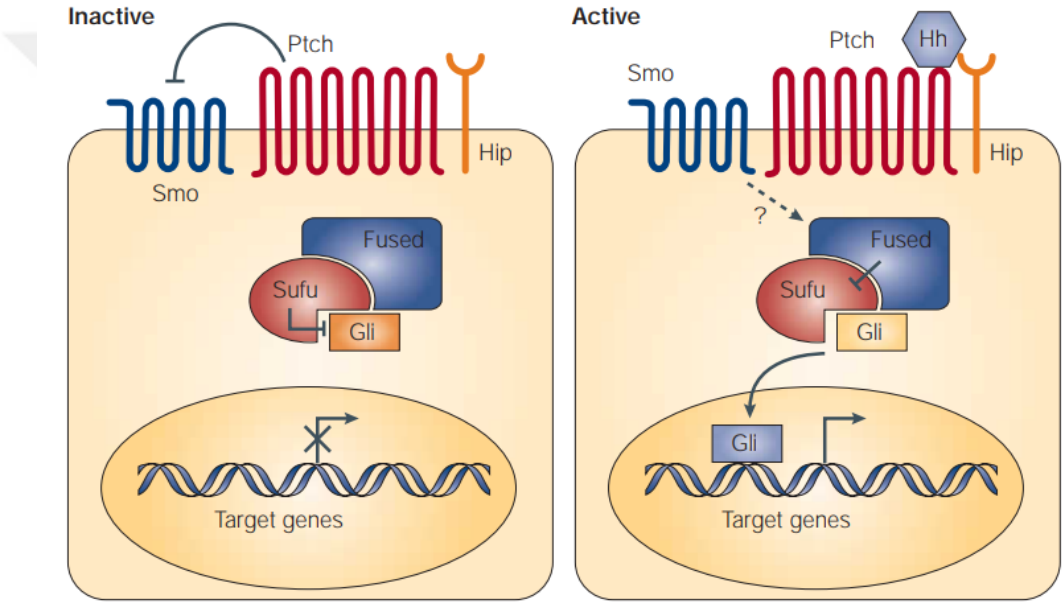

Figure 4.1: Outline of the Hedgehog Signaling. When no Hedgehog signal is present, the signaling is an inactive state. When Hedgehog signaling is present it binds to Ptch receptors. Following events release the inhibition over Gli transcription factors. Then Gli travels to the nucleus and controls the transcription of the target genes of this pathway. Figure is from Pasca di Magliano and Hebrok (2003).

There are three members of the GLI family: GLI1, GLI2, and GLI3. GLI2 and GLI3 possess both transcription activator and repressor domains, whereas GLI1 only has an activator domain. Both GLI2 and GLI3 have two forms. The full-length forms GLI2-FL and GLI3-FL act as activators of target genes. The second truncated version of these transcription factors, GLI2-R and GLI3-R are acting as repressors. They are short forms of the molecule that are cleaved by the ubiquitin ligases.

Some works, such as Alvarez-Medina *et al.* (2008) and Ulloa, Itasaki and Briscoe (2007),

suggested that Wnt signaling is acting on the Shh/Gli pathway by controlling the main repressor of the pathway, GLI3. These works also suggest that GLI3 and β-catenin have a direct interaction. Several works suggested the presence of the TCF/LEF binding motif in the highly conserved non-coding regions of GLI3 (Abbasi *et al.*, 2007; Alvarez-Medina *et al.*, 2008; Ulloa and Martí, 2009). GLI3 is suggested as a key regulator in colorectal cancer (Shen, Zhang and Wang, 2021). It also has been suggested that GLI3 increases the tumorigenicity by regulating the genes related to adherence (Iwasaki *et al.*, 2013).

The gene PTCH1 is the main receptor in both the canonical and the non-canonical Hedgehog signaling. A recent work by Boonekamp *et al.* (2021) (published after we began this study) suggests that PTCH1 is a colon-specific target of the Wnt signaling pathway, which coincides with our results (Boonekamp *et al.*, 2021). The same work suggests other genes from our prediction lists, PALD1 and ZNRF3 as commonly regulated Wnt signaling target genes. This can be seen as an experimental validation of our prediction, since we did not use this information and, nevertheless, got to the same conclusion.

Another study, by Regan *et al.* (2017), suggests that the PTCH1-dependent non-canonical Hedgehog signaling pathway is a positive regulator of the Wnt signaling in colon cancer stem cells (Regan *et al.*, 2017).

Our study supports the hypothesis of an interplay between Wnt and Hedgehog signaling in the context of colorectal cancer. This interaction has been stated in several works (Brink *et al.*, 2004; Madison *et al.*, 2005; Akiyoshi, 2006; Regan *et al.*, 2017).

4.2. EMT AND WNT SIGNALING

We also predicted that the genes CDH1, FN1, and VIM are targets. Although we did not use them as bonafide targets, these genes have been reported as Wnt signaling targets (Gradl, Kühl and Wedlich, 1999; Gilles *et al.*, 2003; Jamora *et al.*, 2003; Berge *et al.*, 2008). We did not include these genes in our bonafide list because they were absent from the literature we reviewed, and they are not reported in colorectal cancer.

These genes are epithelial mesenchymal transition (EMT) markers. During EMT, cells lose epithelial characteristics —such as tight junctions between cells, distinct angular shape, limited motility— and gain mesenchymal characteristics. Cells under this process decrease the number of epithelial markers and increase the number of mesenchymal ones. This transition

also plays a role in cancer formation. Wnt signaling is reported to be one of the initiators of EMT. Several Wnt target genes are also considered as EMT biomarkers, such as claudin, ZEB1, β-catenin, S100A4.

4.3. TBX3 AND OTHER GENES

Another predicted Wnt target gene is TBX3. This gene is a member of the T-box family of transcription factors and TBX3 has many roles in different developmental stages. TBX3 has been previously reported as a direct target gene of the Wnt signaling pathway (Renard *et al.*, 2007). One study suggests that TBX3 is a member of the β-catenin/TCF transcriptional complex, and overexpression of TBX3 aggravates the progression of colorectal cancer (Zimmerli *et al.*, 2020). It is also reported that its expression is increased in colon adenomas (Van der Flier *et al.*, 2007).

Another gene that we predicted as a Wnt target is FOXN3, a tumor suppressor gene that has been reported as downregulated in several cancer-related works (Dai *et al.*, 2016; Xue *et al.*, 2018; He *et al.*, 2019; Zhao *et al.*, 2020). When upregulated, FOXN3 reduces the invasive character of colorectal cancer and also inhibits the interaction of TCF4 and β-catenin (Dai *et al.*, 2016).

We also predicted that SOX4, a member of the Sox transcription factor family, is a Wnt target. The relationship between the Sox transcription factor family and Wnt signaling has been previously studied (Kormish, Sinner and Zorn, 2010). Other members of this family, such as SOX9, SOX17, and SOX2, have been reported as target genes of the Wnt signaling (Blache *et al.*, 2004; Van Raay *et al.*, 2005; Du *et al.*, 2009). Sinner *et al.* (2007) reported that SOX4 increases Wnt signaling activity in the gut and can bind directly to β-catenin and TCF/LEF. It has also been reported in several works that SOX4 is upregulated in colon cancers (Reichling *et al.*, 2005; Van der Flier *et al.*, 2007; Andersen *et al.*, 2009; Lin *et al.*, 2013; Wang *et al.*, 2016). A genome-wide SOX4 occupancy study, which was conducted in prostate cancer, reported some interesting SOX4 targets. The targets of SOX4 include the genes FZD3, FZD5, and FZD8, which are Wnt receptors. The Hedgehog signaling pathway receptor PTCH1 was also identified as a SOX4 target. Some target genes that are included in our bonafide list, such as TNC and ADAM10, have also been reported as SOX4 targets (Scharer *et al.*, 2009).

PMEPA1 has also been reported as upregulated in colon cancer in (Reichling *et al.*, 2005).

We also predicted SLC2A3 also known as GLUT3. It has been shown that GSK3-β inhibitors decrease the expression of GLUT3 (Watanabe *et al.*, 2012). GLUT3 recently proved to be a direct Wnt target in hepatoblastoma cells (Crippa *et al.*, 2017).

One study by Ohsugi *et al.* (2019) suggests that IRF1, one of the targets we predict, is downregulated as a result of the β-catenin knockdown in CRC cell lines. Another study suggests that IRF1 interacts directly with β-catenin and downregulates β-catenin expression in ischemic liver (Yan *et al.*, 2020). IRF1 is also downregulated in colorectal cancer in correlation with the previous study (Xu *et al.*, 2021).

Another gene in our list is TNFRSF19 (also known as Troy), which has been reported as a direct Wnt target gene in several studies (Buttitta *et al.*, 2003; Qiu *et al.*, 2010; Fafilek *et al.*, 2013; Schön *et al.*, 2014). This gene is also reported in studies about colorectal cancer, gastric cancer and stem cell marker (Stange *et al.*, 2013; Wilhelm *et al.*, 2017; Nishioka *et al.*, 2018).

A study from Yochum *et al.* (2007) reported several Wnt signaling components as Wnt targets. One of them, WNT2, is also in one of our prediction lists. Another study by Jung *et al.* (2015) showed that WNT2 is highly expressed in colorectal cancer cells.

In a study by Watanabe *et al.* (2014) they combined ChIP-seq and microarray to discover direct β-catenin targets in SW480 cells. They found 162 directly regulated genes that have responsive β-catenin binding regions, and these genes change their expression as a result of deregulated Wnt signaling. The genes KIZ, GLI3, PALD1, NSMF, TNFRSF19, ZNRF3, ABHD6, CTR9, HSDL1, FAM169A, CCDC186 that we predicted as novel targets were also in this list of 162 genes.

In the study by Van der Flier *et al.* (2007) they performed gene expression profiling experiments in colorectal cancer cell lines through blocking the Wnt signaling. They identified 121 genes that are upregulated in adenomas, in carcinomas, and in both cell lines. Some of the novel target genes that we predicted in our study also reported in this work. The genes HES6, SOX4, DKC1, CD320, CAD, POLR1B, and SLC19A1 upregulated in both adenomas and carcinomas, ZNRF3, RHOBTB3, ARID5B and TBX3 upregulated only in adenomas, and SLC25A19 and PPIF upregulated in carcinomas.

A study by Bottomly *et al.* (2010) used ChIP-seq to identify β-catenin binding motifs in HCT116 cells. They found β-catenin enriched regions in some of the genes that we

also predicted as Wnt signaling target: MPP7, SORBS1, ABLIM1, FOXN3, SMAD3, STAT3, GALNT7, ADAM19, TRIB1, PTCH1, DHRS3, GDE1, TCF7L2, ASAP2, SPATS2L, B4GALT5, ZNRF3, TSPAN5, KANK1, ST6GALNAC4.

A chromatin occupancy of TCF4 colorectal cancer cells study reported genes that are in our list ARHGEF16, DHRS3, LRP8, SLC4A3, CTDSPL, PPP3CA, GALNT7, IFRD1, CAV2, CHMP4C, TRIB1, MPP7, SLK, ABLIM1, BIRC3, TMEM45B, TBX3, TNFRSF19, SMAD3, PPL, CDH1, PER1, MMD, EPHA4, DAP, AHR, FIGNL1, DUSP4, TSNARE1, ALDH1B1, GSN, TFAM, ARID5B, HELLS, SORBS1, SLC35F2, ST14, KLK6, B4GALT5, SLC19A1, ZNRF3, GCAT (Hatzis *et al.*, 2008).

4.4. PERSPECTIVES

In this work we propose lists of potential novel targets of the Wnt Signaling pathway. These predictions are based on the previous knowledge of bona fide target genes of this pathway, and transcriptomic data from colorectal cancer cell lines. Our method is based on a method published by Hödar *et al.* (2010) which we extend to gene expression and complement by defining suitable negative control. We took a novel approach in our study and used transcriptomic data to represent the genes. Previously Hödar *et al.* (2010) used genomic data, in the form of *fingerprints* of transcription factors' binding sites. Also, in the same paper, the training phase used fixed values for the hyperparameters. In this work we used several values for the hyperparameters penalty and sample size, to explore how they impact on the generalization of predictions. Our contribution to the literature is not only biological but also technical in the machine learning domain.

In this study we evaluated the effect of several values of the training hyperparameters. Once the best hyperparameters were found, we immediately found the proposed novel target genes. Since our initial goal was not focused on finding the best hyperparameters, we tested each hyperparameter with a large number of pools. This could have been accomplished faster by first finding the good hyperparameters with smaller pools, and then doing the final evaluation with larger pools.

We found evidence in the literature that supports the idea that the genes we predicted as Wnt signaling targets are somehow involved in colorectal cancer. In the end, due to the complex nature of the transcriptional control of expression, the reality might be different from our

results. The predicted genes might be direct or indirect targets. We aim to reduce the number of genes that need to be validated experimentally. Our findings need a wet-lab experimental validation to support our hypothesis. This can be accomplished with the experiments we described in Section 1.8..

In our quest of finding novel targets of the Wnt signaling pathway, first we compiled a list of genes reported as targets of this signaling pathway. We chose these genes based on 'Wnt target genes' (no date) and the work of Herbst *et al.* (2014). In doing so, we may have overlooked some of the genes that could be in our list. Thus, a future project may include more bonafide target genes of the Wnt signaling pathway.

We also took a tissue specific approach based on works of many researchers (Song, Wang and Ewing, 2014; Watanabe *et al.*, 2014) and we chose genes reported as targets either in human colon tissue or colorectal cancer cell lines. To represent the target genes, we chose gene expression data from colorectal cancer cell lines. Our approach of focusing on a single tissue can be tested for other tissues, cancers or even developmental stages. Moreover, it should be tested if these novel target genes are tissue specific or universal.

Gene expression profiling is used in biomarker studies. The machine learning method that we used for target gene discovery, combined with differential gene expression data, can also be implemented to candidate biomarker search. In such scenario, instead of bonafide target genes, existing biomarkers can be used as positive examples, and representing genes with differential expression data might be suitable for discovering some co-expression patterns which are widely used in biomarker search.

Different types of data, such as proteomic or metabolomic could also, in principle, be used in the same way.

Although using differential gene expression data resulted in overfitting in some cases, gene expression data produced novel predictions that are more promising, from the biological point of view, compared to the predictions made using TFBS fingerprints.

REFERENCES

Abbas-Aghababazadeh, F., Li, Q. and Fridley, B. L. (2018) 'Comparison of normalization approaches for gene expression studies completed with high-throughput sequencing', *PLoS One*. Public Library of Science (PLoS), 13(10), p. e0206312.

Abbasi, A. A. *et al.* (2007) 'Human GLI3 intragenic conserved non-coding sequences are tissue-specific enhancers', *PLoS ONE*. Edited by J.-N. Volff. Public Library of Science (PLoS), 2(4), p. e366. doi: 10.1371/journal.pone.0000366.

Ai, M. *et al.* (2005) 'Reduced affinity to and inhibition by DKK1 form a common mechanism by which high bone mass-associated missense mutations in LRP5 affect canonical wnt signaling', *Molecular and Cellular Biology*. American Society for Microbiology, 25(12), pp. 4946–4955. doi: 10.1128/mcb.25.12.4946-4955.2005.

Akiyoshi, T. (2006) 'Gli1, downregulated in colorectal cancers, inhibits proliferation of colon cancer cells involving wnt signalling activation', *Gut*. BMJ, 55(7), pp. 991–999. doi: 10.1136/gut.2005.080333.

Alexander, C. *et al.* (2014) 'Wnt signaling interacts with bmp and Edn1 to regulate dorsal-ventral patterning and growth of the craniofacial skeleton', *PLoS Genetics*. Edited by M. C. Mullins. Public Library of Science (PLoS), 10(7), p. e1004479. doi: 10.1371/journal.pgen.1004479.

Alvarez-Medina, R. *et al.* (2008) 'Wnt canonical pathway restricts graded shh/gli patterning activity through the regulation of Gli3 expression', *Development*. The Company of Biologists, 135(2), pp. 237–247. doi: 10.1242/dev.012054.

Andersen, C. L. *et al.* (2009) 'Dysregulation of the transcription factors SOX4, CBFB and SMARCC1 correlates with outcome of colorectal cancer', *British Journal of Cancer*. Springer Science; Business Media LLC, 100(3), pp. 511–523. doi: 10.1038/sj.bjc.6604884.

Andrews, S. *et al.* (2012) 'FastQC'. Babraham, UK: Babraham Institute.

Aono, S. *et al.* (2019) '*B*-catenin/TCF4 complex-mediated induction of the NRF3 (NFE2L3) gene in cancer cells', *Int. J. Mol. Sci.* MDPI AG, 20(13), p. 3344.

Araki, Y. *et al.* (2003) 'Regulation of cyclooxygenase-2 expression by the wnt and ras pathways', *Cancer Research*. American Association for Cancer Research, 63(3), pp. 728–734. Available at: https://cancerres.aacrjournals.org/content/63/3/728.

Aravena Duarte, A. O. (2013) *Probabilistic and constraint based modelling to determine regulation events from heterogeneous biological data*. Theses. Université Rennes 1. Available at: https://tel.archives-ouvertes.fr/tel-00988255.

Babonis, L. S. and Martindale, M. Q. (2017) 'Phylogenetic evidence for the modular evolution of metazoan signalling pathways', *Philosophical Transactions of the Royal Society B: Biological Sciences*. The Royal Society, 372(1713), p. 20150477. doi: 10.1098/rstb.2015.0477.

Baehs, S. *et al.* (2009) 'Dickkopf-4 is frequently down-regulated and inhibits growth of colorectal cancer cells', *Cancer Lett.* Elsevier BV, 276(2), pp. 152–159.

Barker, N. *et al.* (2007) 'Identification of stem cells in small intestine and colon by marker gene Lgr5', *Nature*. Springer Science; Business Media LLC, 449(7165), pp. 1003–1007.

Basu, S. *et al.* (2018) 'The intestinal stem cell regulating gene ASCL2 is required for L1-mediated colon cancer progression', *Cancer Lett.* Elsevier BV, 424, pp. 9–18.

Batlle, E. *et al.* (2002) '*B*-catenin and TCF mediate cell positioning in the intestinal epithelium by controlling the expression of EphB/EphrinB', *Cell*. Elsevier BV, 111(2), pp. 251–263.

Behrens, J. *et al.* (1996) 'Functional interaction of β-catenin with the transcription factor LEF-1', *Nature*. Springer Science; Business Media LLC, 382(6592), pp. 638–642. doi: 10.1038/382638a0.

Beiter, K. *et al.* (2005) 'Beta-catenin regulates the expression of tenascin-c in human colorectal tumors', *Oncogene*. Springer Science; Business Media LLC, 24(55), pp. 8200–8204.

Berge, D. ten *et al.* (2008) 'Wnt signaling mediates self-organization and axis formation in embryoid bodies', *Cell Stem Cell*. Elsevier BV, 3(5), pp. 508–518.

Bhanot, P. *et al.* (1996) 'A new member of the frizzled family from drosophila functions as a wingless receptor', *Nature*. Springer Science; Business Media LLC, 382(6588), pp. 225–230. doi: 10.1038/382225a0.

'Bias–variance tradeoff' (2022) *Wikipedia*. Wikimedia Foundation. Available at: https://en.wikipedia.org/wiki/Bias%E2%80%93variance_tradeoff.

Biotechnology Information, N. C. for (no date) 'Home - PMC - NCBI'. U.S. National Library of Medicine. Available at: https://www.ncbi.nlm.nih.gov/pmc/.

Blache, P. *et al.* (2004) 'SOX9 is an intestine crypt transcription factor, is regulated by the wnt pathway, and represses the CDX2 and MUC2 genes', *J. Cell Biol.* Rockefeller University Press, 166(1), pp. 37–47.

Blanchette, M. *et al.* (2006) 'Genome-wide computational prediction of transcriptional

regulatory modules reveals new insights into human gene expression', *Genome Research*. Cold Spring Harbor Laboratory, 16(5), pp. 656–668. doi: 10.1101/gr.4866006.

Boon, E. M. J. *et al.* (2002) 'Wnt signaling regulates expression of the receptor tyrosine kinase met in colorectal cancer', *Cancer Research*. American Association for Cancer Research, 62(18), pp. 5126–5128. Available at: https://cancerres.aacrjournals.org/content/62/18/5126.

Boonekamp, K. E. *et al.* (2021) 'Identification of novel human Wnt target genes using adult endodermal tissue-derived organoids', *Developmental Biology*. Elsevier BV, 474, pp. 37–47. doi: 10.1016/j.ydbio.2021.01.009.

Bottomly, D. *et al.* (2010) 'Identification of beta-catenin binding regions in colon cancer cells using ChIP-Seq', *Nucleic Acids Res.* Oxford University Press (OUP), 38(17), pp. 5735–5745.

Boyden, L. M. *et al.* (2002) 'High bone density due to a mutation in LDL-receptor–related protein 5', *New England Journal of Medicine*. Massachusetts Medical Society, 346(20), pp. 1513–1521. doi: 10.1056/nejmoa013444.

Brabletz, T. *et al.* (1999) '*B*-catenin regulates the expression of the matrix metalloproteinase-7 in human colorectal cancer', *Am. J. Pathol.* Elsevier BV, 155(4), pp. 1033–1038.

Breiman, L. *et al.* (1984) *Classification and regression trees*. Taylor & Francis. Available at: https://books.google.com.tr/books?id=JwQx-WOmSyQC.

Brink, G. R. van den *et al.* (2004) 'Indian hedgehog is an antagonist of wnt signaling in colonic epithelial cell differentiation', *Nature Genetics*. Springer Science; Business Media LLC, 36(3), pp. 277–282. doi: 10.1038/ng1304.

Bush, B. M. *et al.* (2013) 'The Wnt/β-catenin/T-cell factor 4 pathway up-regulates high-mobility group A1 expression in colon cancer', *Cell Biochem. Funct.* Wiley, 31(3), pp. 228–236.

Buttitta, L. *et al.* (2003) 'Microarray analysis of somitogenesis reveals novel targets of different WNT signaling pathways in the somitic mesoderm', *Dev. Biol.* Elsevier BV, 258(1), pp. 91–104.

Cadigan, K. M. and Nusse, R. (1997) 'Wnt signaling: A common theme in animal development', *Genes & Development*. Cold Spring Harbor Laboratory, 11(24), pp. 3286–3305. doi: 10.1101/gad.11.24.3286.

Campbell, P. M. and Szyf, M. (2003) 'Human DNA methyltransferase gene DNMT1 is regulated by the APC pathway', *Carcinogenesis*. Oxford University Press (OUP), 24(1),

pp. 17–24.

Christensen, J. *et al.* (2013) 'FOXQ1, a novel target of the wnt pathway and a new marker for activation of wnt signaling in solid tumors', *PLoS One*. Public Library of Science (PLoS), 8(3), p. e60051.

Clevers, H. (2006) 'Wnt/β-catenin signaling in development and disease', *Cell*. Elsevier BV, 127(3), pp. 469–480. doi: 10.1016/j.cell.2006.10.018.

Cohen, P. and Goedert, M. (2004) 'GSK3 inhibitors: Development and therapeutic potential', *Nature Reviews Drug Discovery*. Springer Science; Business Media LLC, 3(6), pp. 479–487. doi: 10.1038/nrd1415.

Conacci-Sorrell, M. E. *et al.* (2002) 'Nr-CAM is a target gene of the beta-catenin/LEF-1 pathway in melanoma and colon cancer and its expression enhances motility and confers tumorigenesis', *Genes Dev.* Cold Spring Harbor Laboratory, 16(16), pp. 2058–2072.

Crawford, H. C. *et al.* (1999) 'The metalloproteinase matrilysin is a target of beta-catenin transactivation in intestinal tumors', *Oncogene*. Springer Science; Business Media LLC, 18(18), pp. 2883–2891.

Crippa, S. *et al.* (2017) 'Mutant CTNNB 1 and histological heterogeneity define metabolic subtypes of hepatoblastoma', *EMBO Mol. Med.* EMBO, 9(11), pp. 1589–1604.

Dahal Lamichane, B. *et al.* (2019) 'AGR2 is a target of canonical Wnt/β-catenin signaling and is important for stemness maintenance in colorectal cancer stem cells', *Biochem. Biophys. Res. Commun.* Elsevier BV, 515(4), pp. 600–606.

Dai, Y. *et al.* (2016) 'Loss of FOXN3 in colon cancer activates beta-catenin/TCF signaling and promotes the growth and migration of cancer cells', *Oncotarget*. Impact Journals, LLC, 8(6), pp. 9783–9793. doi: 10.18632/oncotarget.14189.

DasGupta, R. and Fuchs, E. (1999) 'Multiple roles for activated LEF/TCF transcription complexes during hair follicle development and differentiation', *Development*. The Company of Biologists, 126(20), pp. 4557–4568. doi: 10.1242/dev.126.20.4557.

Dehner, M. *et al.* (2008) 'Wnt signaling inhibits forkhead box O3a-induced transcription and apoptosis through up-regulation of serum- and glucocorticoid-inducible kinase 1', *J. Biol. Chem.* Elsevier BV, 283(28), pp. 19201–19210.

Dietinger, V., Durango, C. R. G. de, *et al.* (2020) 'Wnt-driven LARGE2 mediates laminin-adhesive o-glycosylation in human colonic epithelial cells and colorectal cancer', *Cell Communication and Signaling*. Springer Science; Business Media LLC, 18(1). doi: 10.1186/s12964-020-00561-6.

Dietinger, V., García de Durango, C. R., *et al.* (2020) 'Wnt-driven LARGE2 mediates laminin-adhesive o-glycosylation in human colonic epithelial cells and colorectal cancer', *Cell Commun. Signal.* Springer Science; Business Media LLC, 18(1), p. 102.

Dominguez, I., Itoh, K. and Sokol, S. Y. (1995) 'Role of glycogen synthase kinase 3 beta as a negative regulator of dorsoventral axis formation in xenopus embryos.', *Proceedings of the National Academy of Sciences*. Proceedings of the National Academy of Sciences, 92(18), pp. 8498–8502. doi: 10.1073/pnas.92.18.8498.

Domon-Dell, C. and Freund, J.-N. (2002) 'Stimulation of Cdx1 by oncogenic β-catenin/Tcf4 in colon cancer cells; opposite effect of the CDX2 homeoprotein', *FEBS Lett.* Wiley, 518(1-3), pp. 83–87.

Du, Q. *et al.* (2006) 'Regulation of human nitric oxide synthase 2 expression by wnt beta-catenin signaling', *Cancer Res.* American Association for Cancer Research (AACR), 66(14), pp. 7024–7031.

Du, Y.-C. *et al.* (2009) 'Induction and down-regulation of Sox17 and its possible roles during the course of gastrointestinal tumorigenesis', *Gastroenterology*. Elsevier BV, 137(4), pp. 1346–1357.

Elkon, R. *et al.* (2003) 'Genome-wide in silico identification of transcriptional regulators controlling the cell cycle in human cells', *Genome Research*. Cold Spring Harbor Laboratory, 13(5), pp. 773–780. doi: 10.1101/gr.947203.

Embi, N., Rylatt, D. B. and Cohen, P. (2005) 'Glycogen synthase kinase-3 from rabbit skeletal muscle', *Eur. J. Biochem.* Wiley, 107(2), pp. 519–527.

Ewing, R. M. *et al.* (2018) 'Multiproteomic and transcriptomic analysis of oncogenic β-catenin molecular networks', *Journal of Proteome Research*. American Chemical Society (ACS), 17(6), pp. 2216–2225. doi: 10.1021/acs.jproteome.8b00180.

Fafilek, B. *et al.* (2013) 'Troy, a tumor necrosis factor receptor family member, interacts with lgr5 to inhibit wnt signaling in intestinal stem cells', *Gastroenterology*. Elsevier BV, 144(2), pp. 381–391.

Farin, H. F. *et al.* (2016) 'Visualization of a short-range wnt gradient in the intestinal stem-cell niche', *Nature*. Springer Science; Business Media LLC, 530(7590), pp. 340–343. doi: 10.1038/nature16937.

Ferrari, G. V. D. and Inestrosa, N. C. (2000) 'Wnt signaling function in Alzheimer's disease', *Brain Research Reviews*. Elsevier BV, 33(1), pp. 1–12. doi: 10.1016/s0165-0173(00)00021-7.

Ferrer-Mayorga, G. *et al.* (2019) 'Vitamin d and Wnt3A have additive and partially overlapping modulatory effects on gene expression and phenotype in human colon fibroblasts', *Scientific Reports*. Springer Science; Business Media LLC, 9(1). doi: 10.1038/s41598-019-44574-9.

Fevr, T. *et al.* (2007) 'Wnt/β-catenin is essential for intestinal homeostasis and maintenance of intestinal stem cells', *Molecular and Cellular Biology*. American Society for Microbiology, 27(21), pp. 7551–7559. doi: 10.1128/mcb.01034-07.

Filali, M. *et al.* (2002) 'Wnt-3A/beta-catenin signaling induces transcription from the LEF-1 promoter', *J. Biol. Chem.* Elsevier BV, 277(36), pp. 33398–33410.

Fuchs, E. (2007) 'Scratching the surface of skin development', *Nature*. Springer Science; Business Media LLC, 445(7130), pp. 834–842.

Fujita, M. *et al.* (2001) 'Up-regulation of the ectodermal-neural cortex 1 (ENC1) gene, a downstream target of the β-catenin/t-cell factor complex, in colorectal carcinomas', *Cancer Research*. American Association for Cancer Research, 61(21), pp. 7722–7726. Available at: https://cancerres.aacrjournals.org/content/61/21/7722.

Gat, U. *et al.* (1998) 'De novo hair follicle morphogenesis and hair tumors in mice expressing a truncated β-catenin in skin', *Cell*. Elsevier BV, 95(5), pp. 605–614. doi: 10.1016/s0092-8674(00)81631-1.

Gavert, N. *et al.* (2005) 'L1, a novel target of beta-catenin signaling, transforms cells and is expressed at the invasive front of colon cancers', *J. Cell Biol.* Rockefeller University Press, 168(4), pp. 633–642.

Gavert, N. *et al.* (2007) 'Expression of L1-CAM and ADAM10 in human colon cancer cells induces metastasis', *Cancer Res.* American Association for Cancer Research (AACR), 67(16), pp. 7703–7712.

Genderen, C. van *et al.* (1994) 'Development of several organs that require inductive epithelial-mesenchymal interactions is impaired in LEF-1-deficient mice.', *Genes & Development*. Cold Spring Harbor Laboratory, 8(22), pp. 2691–2703. doi: 10.1101/gad.8.22.2691.

Ghiselli, G. *et al.* (2003) 'The cohesin SMC3 is a target the for beta-catenin/TCF4 transactivation pathway', *J. Biol. Chem.* Elsevier BV, 278(22), pp. 20259–20267.

Giannakis, M. *et al.* (2014) 'RNF43 is frequently mutated in colorectal and endometrial cancers', *Nature Genetics*. Springer Science; Business Media LLC, 46(12), pp. 1264–1266. doi: 10.1038/ng.3127.

Giese, K., Cox, J. and Grosschedl, R. (1992) 'The HMG domain of lymphoid enhancer

factor 1 bends DNA and facilitates assembly of functional nucleoprotein structures', *Cell*. Elsevier BV, 69(1), pp. 185–195. doi: 10.1016/0092-8674(92)90129-z.

Gilles, C. *et al.* (2003) 'Transactivation of vimentin by beta-catenin in human breast cancer cells.', *Cancer Research*. American Association for Cancer Research, 63(10), pp. 2658–64. Available at: https://www.hal.inserm.fr/inserm-00149065.

Gonçalves, V., Matos, P. and Jordan, P. (2008) 'The beta-catenin/TCF4 pathway modifies alternative splicing through modulation of SRp20 expression', *RNA*. Cold Spring Harbor Laboratory, 14(12), pp. 2538–2549.

Gong, Y. *et al.* (2001) 'LDL receptor-related protein 5 (LRP5) affects bone accrual and eye development', *Cell*. Elsevier BV, 107(4), pp. 513–523. doi: 10.1016/s0092-8674(01)00571-2.

Gradl, D., Kühl, M. and Wedlich, D. (1999) 'The Wnt/Wg signal transducer beta-catenin controls fibronectin expression', *Mol. Cell. Biol.*, 19(8), pp. 5576–5587.

Groden, J. *et al.* (1991) 'Identification and characterization of the familial adenomatous polyposis coli gene', *Cell*. Elsevier BV, 66(3), pp. 589–600. doi: 10.1016/0092-8674(81)90021-0.

Hatzis, P. *et al.* (2008) 'Genome-wide pattern of TCF7L2/TCF4 chromatin occupancy in colorectal cancer cells', *Mol. Cell. Biol.* American Society for Microbiology, 28(8), pp. 2732–2744.

He, H. *et al.* (2019) 'Novel tumor-suppressor FOXN3 is downregulated in adult acute myeloid leukemia', *Oncology Letters*. Spandidos Publications. doi: 10.3892/ol.2019.10424.

He, T.-C. *et al.* (1998) 'Identification of c- MYC as a target of the APC pathway', *Science*. American Association for the Advancement of Science (AAAS), 281(5382), pp. 1509–1512.

He, T.-C. *et al.* (1999) 'PPARδ is an APC-regulated target of nonsteroidal anti-inflammatory drugs', *Cell*. Elsevier BV, 99(3), pp. 335–345.

He, X. *et al.* (1995) 'Glycogen synthase kinase-3 and dorsoventral patterning in xenopus embryos', *Nature*. Springer Science; Business Media LLC, 374(6523), pp. 617–622. doi: 10.1038/374617a0.

Herbst, A. *et al.* (2014) 'Comprehensive analysis of β-catenin target genes in colorectal carcinoma cell lines with deregulated wnt/β-catenin signaling', *BMC Genomics*. Springer Science; Business Media LLC, 15(1), p. 74. doi: 10.1186/1471-2164-15-74.

Hiendlmeyer, E. *et al.* (2004) '*B*-catenin up-regulates the expression of the urokinase plas-

minogen activator in human colorectal tumors', *Cancer Res.* American Association for Cancer Research (AACR), 64(4), pp. 1209–1214.

Hlubek, F. *et al.* (2001) 'Expression of the invasion factor laminin $\gamma 2$ in colorectal carcinomas is regulated by β-catenin', *Cancer Research*. American Association for Cancer Research, 61(22), pp. 8089–8093. Available at: https://cancerres.aacrjournals.org/content/61/22/8089.

Hlubek, F. *et al.* (2004) 'Beta-catenin activates a coordinated expression of the proinvasive factors laminin-5 gamma2 chain and MT1-MMP in colorectal carcinomas', *Int. J. Cancer*. Wiley, 108(2), pp. 321–326.

Hlubek, F. *et al.* (2006) 'Securin (hPTTG1) expression is regulated by beta-catenin/TCF in human colorectal carcinoma', *Br. J. Cancer*. Springer Science; Business Media LLC, 94(11), pp. 1672–1677.

Hödar, C. *et al.* (2010) 'Genome-wide identification of new wnt/β-catenin target genes in the human genome using CART method', *BMC Genomics*. Springer Science; Business Media LLC, 11(1), p. 348. doi: 10.1186/1471-2164-11-348.

Hoppler, S. and Kavanagh, C. L. (2007) 'Wnt signalling: Variety at the core', *Journal of Cell Science*. The Company of Biologists, 120(3), pp. 385–393. doi: 10.1242/jcs.03363.

Hovanes, K. *et al.* (2001) 'Beta-catenin-sensitive isoforms of lymphoid enhancer factor-1 are selectively expressed in colon cancer', *Nat. Genet.* Springer Science; Business Media LLC, 28(1), pp. 53–57.

Huber, A. H., Nelson, W. J. and Weis, W. I. (1997) 'Three-dimensional structure of the armadillo repeat region of β-catenin', *Cell*. Elsevier BV, 90(5), pp. 871–882. doi: 10.1016/s0092-8674(00)80352-9.

Huelsken, J. *et al.* (2001) 'β-catenin controls hair follicle morphogenesis and stem cell differentiation in the skin', *Cell*. Elsevier BV, 105(4), pp. 533–545. doi: 10.1016/s0092-8674(01)00336-1.

IBM Cloud Education (2021) 'What is underfitting?' https://www.ibm.com/cloud/learn/underfitting.

Inoue, M. *et al.* (2018) 'The stress response gene ATF3 is a direct target of the Wnt/β-catenin pathway and inhibits the invasion and migration of HCT116 human colorectal cancer cells', *PLoS One*. Public Library of Science (PLoS), 13(7), p. e0194160.

Ireland, H. *et al.* (2004) 'Inducible cre-mediated control of gene expression in the murine gastrointestinal tract: Effect of loss of β-catenin', *Gastroenterology*. Elsevier BV, 126(5),

pp. 1236–1246. doi: 10.1053/j.gastro.2004.03.020.

Iseki, H. *et al.* (2010) 'Human arm protein lost in epithelial cancers, on chromosome X 1 (ALEX1) gene is transcriptionally regulated by CREB and wnt/beta-catenin signaling', *Cancer Sci.* Wiley, 101(6), pp. 1361–1366.

Iwasaki, H. *et al.* (2013) 'Hedgehog Gli3 activator signal augments tumorigenicity of colorectal cancer via upregulation of adherence-related genes', *Cancer Sci.* Wiley, 104(3), pp. 328–336.

Jamora, C. *et al.* (2003) 'Links between signal transduction, transcription and adhesion in epithelial bud development', *Nature*. Springer Science; Business Media LLC, 422(6929), pp. 317–322.

Janda, C. Y. *et al.* (2017) 'Surrogate wnt agonists that phenocopy canonical wnt and β-catenin signalling', *Nature*. Springer Science; Business Media LLC, 545(7653), pp. 234–237. doi: 10.1038/nature22306.

Jho, E.-H. *et al.* (2002) 'Wnt/beta-catenin/Tcf signaling induces the transcription of Axin2, a negative regulator of the signaling pathway', *Mol. Cell. Biol.*, 22(4), pp. 1172–1183.

Jubb, A. M. *et al.* (2006) 'Achaete-scute like 2 (ascl2) is a target of wnt signalling and is upregulated in intestinal neoplasia', *Oncogene*. Springer Science; Business Media LLC, 25(24), pp. 3445–3457.

Jung, H.-C. and Kim, K. (2005) 'Identification of MYCBP as a beta-catenin/LEF-1 target using DNA microarray analysis', *Life Sci.* Elsevier BV, 77(11), pp. 1249–1262.

Jung, Y.-S. *et al.* (2015) 'Wnt2 complements Wnt/β-catenin signaling in colorectal cancer', *Oncotarget*. Impact Journals, LLC, 6(35), pp. 37257–37268.

Kiecker, C. and Niehrs, C. (2001) 'A morphogen gradient of wnt/β-catenin signalling regulates anteroposterior neural patterning in xenopus', *Development*. The Company of Biologists, 128(21), pp. 4189–4201. doi: 10.1242/dev.128.21.4189.

Kilańczyk, E. *et al.* (2012) 'S100A6 is transcriptionally regulated by β-catenin and interacts with a novel target, lamin A/C, in colorectal cancer cells', *Cell Calcium*. Elsevier BV, 51(6), pp. 470–477.

Kim, H. J. *et al.* (2018) 'DBC1 regulates Wnt/β-catenin-mediated expression of MACC1, a key regulator of cancer progression, in colon cancer', *Cell Death Dis.* Springer Science; Business Media LLC, 9(8), p. 831.

Kim, J. T. *et al.* (2019) 'Regulation of ketogenic enzyme HMGCS2 by Wnt/β-catenin/PPARγ pathway in intestinal cells', *Cells*. MDPI AG, 8(9), p. 1106.

Kim, J.-H. *et al.* (2017) 'Roles of wnt target genes in the journey of cancer stem cells', *International Journal of Molecular Sciences*. MDPI AG, 18(8), p. 1604. doi: 10.3390/ijms18081604.

Kim, J.-S. *et al.* (2002) 'Oncogenic β-catenin is required for bone morphogenetic protein 4 expression in human cancer cells', *Cancer Research*. American Association for Cancer Research, 62(10), pp. 2744–2748. Available at: https://cancerres.aacrjournals.org/content/62/10/2744.

Kim, P. J. *et al.* (2003) 'Survivin and molecular pathogenesis of colorectal cancer', *Lancet*. Elsevier BV, 362(9379), pp. 205–209.

Kim, T. H. *et al.* (2005) 'Beta-catenin activates the growth factor endothelin-1 in colon cancer cells', *Oncogene*. Springer Science; Business Media LLC, 24(4), pp. 597–604.

King, L. E. *et al.* (2016) 'Differential RNA-seq analysis comparing APC-defective and APC-restored SW480 colorectal cancer cells', *Genomics Data*. Elsevier BV, 7, pp. 293–296. doi: 10.1016/j.gdata.2016.02.001.

Kinzler, K. W. *et al.* (1991) 'Identification of FAP locus genes from chromosome 5q21', *Science*. American Association for the Advancement of Science (AAAS), 253(5020), pp. 661–665. doi: 10.1126/science.1651562.

Kirmizis, A., Bartley, S. M. and Farnham, P. J. (2003) 'Identification of the polycomb group protein SU(z)12 as a potential molecular target for human cancer Therapy1', *Molecular Cancer Therapeutics*. American Association for Cancer Research, 2(1), pp. 113–121. Available at: https://mct.aacrjournals.org/content/2/1/113.

Klein, P. S. and Melton, D. A. (1996) 'A molecular mechanism for the effect of lithium on development.', *Proceedings of the National Academy of Sciences*. Proceedings of the National Academy of Sciences, 93(16), pp. 8455–8459. doi: 10.1073/pnas.93.16.8455.

Koh, T. J. *et al.* (2000) 'Gastrin is a target of the beta-catenin/TCF-4 growth-signaling pathway in a model of intestinal polyposis', *J. Clin. Invest.* American Society for Clinical Investigation, 106(4), pp. 533–539.

Kolligs, F. T. *et al.* (2002) 'ITF-2, a downstream target of the Wnt/TCF pathway, is activated in human cancers with beta-catenin defects and promotes neoplastic transformation', *Cancer Cell*, 1(2), pp. 145–155.

Konsavage, W. M., Jr *et al.* (2012) 'Wnt/β-catenin signaling regulates yes-associated protein (YAP) gene expression in colorectal carcinoma cells', *J. Biol. Chem.* Elsevier BV, 287(15), pp. 11730–11739.

Koo, B.-K. *et al.* (2012) 'Tumour suppressor RNF43 is a stem-cell E3 ligase that induces endocytosis of wnt receptors', *Nature*. Springer Science; Business Media LLC, 488(7413), pp. 665–669. doi: 10.1038/nature11308.

Korinek, V. *et al.* (1998) 'Depletion of epithelial stem-cell compartments in the small intestine of mice lacking tcf-4', *Nature Genetics*. Springer Science; Business Media LLC, 19(4), pp. 379–383. doi: 10.1038/1270.

Kormish, J. D., Sinner, D. and Zorn, A. M. (2010) 'Interactions between SOX factors and wnt/beta-catenin signaling in development and disease', *Dev. Dyn.* Wiley, 239(1), pp. 56–68.

Kratochwil, K. *et al.* (1996) 'Lef1 expression is activated by BMP-4 and regulates inductive tissue interactions in tooth and hair development.', *Genes & Development*. Cold Spring Harbor Laboratory, 10(11), pp. 1382–1394. doi: 10.1101/gad.10.11.1382.

Kuhnert, F. *et al.* (2003) 'Essential requirement for wnt signaling in proliferation of adult small intestine and colon revealed by adenoviral expression of dickkopf-1', *Proceedings of the National Academy of Sciences*. Proceedings of the National Academy of Sciences, 101(1), pp. 266–271. doi: 10.1073/pnas.2536800100.

Kumar, R. *et al.* (2018) 'XPNPEP3 is a novel transcriptional target of canonical Wnt/β-catenin signaling', *Genes Chromosomes Cancer*, 57(6), pp. 304–310.

L., Y. and Marian, W. (no date) 'Wnt signaling inhibition in human colon cancer cell lines (SW480 and SW620) drives changes in gene programs for cell adhesion, mobility, and cell junction'. Available at: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE131948.

Lapham, A. *et al.* (2009) 'The bcl-w promoter is activated by beta-catenin/TCF4 in human colorectal carcinoma cells', *Gene*. Elsevier BV, 432(1-2), pp. 112–117.

Leinonen, R., Sugawara, H. and and, M. S. (2010) 'The sequence read archive', *Nucleic Acids Research*. Oxford University Press (OUP), 39(Database), pp. D19–D21. doi: 10.1093/nar/gkq1019.

Leow, C. C. *et al.* (2004) 'Hath1, down-regulated in colon adenocarcinomas, inhibits proliferation and tumorigenesis of colon cancer cells', *Cancer Res.* American Association for Cancer Research (AACR), 64(17), pp. 6050–6057.

Li, C. *et al.* (2021) 'Regulation of SIRT2 by Wnt/β-catenin signaling pathway in colorectal cancer cells', *Biochim. Biophys. Acta Mol. Cell Res.* Elsevier BV, 1868(4), p. 118966.

Li, Y. *et al.* (2011) 'HEF1, a novel target of wnt signaling, promotes colonic cell migration

and cancer progression', *Oncogene*. Springer Science; Business Media LLC, 30(23), pp. 2633–2643.

Lickert, H. *et al.* (2000) 'Wnt/(beta)-catenin signaling regulates the expression of the homeobox gene Cdx1 in embryonic intestine', *Development*. The Company of Biologists, 127(17), pp. 3805–3813.

Lin, C.-M. *et al.* (2013) 'Clinical and prognostic implications of transcription factor SOX4 in patients with colon cancer', *PLoS ONE*. Edited by I. V. Lebedeva. Public Library of Science (PLoS), 8(6), p. e67128. doi: 10.1371/journal.pone.0067128.

Liu, C. *et al.* (2002) 'Control of β-catenin phosphorylation/degradation by a dual-kinase mechanism', *Cell*. Elsevier BV, 108(6), pp. 837–847. doi: 10.1016/s0092-8674(02)00685-2.

Liu, H. and Motoda, H. (2013) *Feature selection for knowledge discovery and data mining*. New York, NY: Springer (The springer international series in engineering and computer science).

Logan, C. Y. and Nusse, R. (2004) 'The Wnt signaling pathway in development and disease', *Annual Review of Cell and Developmental Biology*. Annual Reviews, 20(1), pp. 781–810. doi: 10.1146/annurev.cellbio.20.010403.113126.

Lustig, B. *et al.* (2002) 'Negative feedback loop of wnt signaling through upregulation of conductin/axin2 in colorectal and liver tumors', *Mol. Cell. Biol.* American Society for Microbiology, 22(4), pp. 1184–1193.

Madison, B. B. *et al.* (2005) 'Epithelial hedgehog signals pattern the intestinal crypt-villus axis', *Development*. The Company of Biologists, 132(2), pp. 279–289. doi: 10.1242/dev.01576.

Malliri, A. *et al.* (2006) 'The rac activator Tiam1 is a wnt-responsive gene that modifies intestinal tumor development', *J. Biol. Chem.* Elsevier BV, 281(1), pp. 543–548.

Mann, B. *et al.* (1999) 'Target genes of beta-catenin-t cell-factor/lymphoid-enhancer-factor signaling in human colorectal carcinomas', *Proc. Natl. Acad. Sci. U. S. A.*, 96(4), pp. 1603–1608.

Markowitz, S. D. and Bertagnolli, M. M. (2009) 'Molecular origins of cancer: Molecular basis of colorectal cancer', *N. Engl. J. Med.* New England Journal of Medicine (NEJM/MMS), 361(25), pp. 2449–2460.

McCrea, P. D., Turck, C. W. and Gumbiner, B. (1991) 'A homolog of the armadillo protein in drosophila (plakoglobin) associated with e-cadherin', *Science*. American Association

for the Advancement of Science (AAAS), 254(5036), pp. 1359–1361. doi: 10.1126/science.1962194.

Medina-Rivera, A. *et al.* (2011) 'Theoretical and empirical quality assessment of transcription factor-binding motifs', *Nucleic Acids Research*. Centro de Ciencias Genomicas, Universidad Nacional Autónoma de México. Av. Universidad s/n. Cuernavaca, Col. Chamilpa, Morelos 62210, Mexico. amedina@lcg.unam.mx: Oxford University Press, 39(3), pp. 808–824.

Miao, Y. *et al.* (2020) 'Next-generation surrogate wnts support organoid growth and deconvolute frizzled pleiotropy in vivo', *Cell Stem Cell*. Elsevier BV, 27(5), pp. 840–851.e6. doi: 10.1016/j.stem.2020.07.020.

Michels, B. E. *et al.* (2019) 'Human colon organoids reveal distinct physiologic and oncogenic wnt responses', *Journal of Experimental Medicine*. Rockefeller University Press, 216(3), pp. 704–720. doi: 10.1084/jem.20180823.

Millar, S. E. (2002) 'Molecular mechanisms regulating hair follicle development', *Journal of Investigative Dermatology*. Elsevier BV, 118(2), pp. 216–225. doi: 10.1046/j.0022-202x.2001.01670.x.

Mir, R. *et al.* (2016) 'Wnt/β-catenin signaling regulated SATB1 promotes colorectal cancer tumorigenesis and progression', *Oncogene*. Springer Science; Business Media LLC, 35(13), pp. 1679–1691.

Miwa, N. *et al.* (2001) 'Involvement of claudin-1 in the beta-catenin/tcf signaling pathway and its frequent upregulation in human colorectal cancers', *Oncol. Res.*, 12(11-12), pp. 469–476.

Mokry, M. *et al.* (2011) 'Integrated genome-wide analysis of transcription factor occupancy, RNA polymerase II binding and steady-state RNA levels identify differentially regulated functional gene classes', *Nucleic Acids Research*. Oxford University Press (OUP), 40(1), pp. 148–158. doi: 10.1093/nar/gkr720.

Molenaar, M. *et al.* (1996) 'XTcf-3 transcription factor mediates β-catenin-induced axis formation in xenopus embryos', *Cell*. Elsevier BV, 86(3), pp. 391–399. doi: 10.1016/s0092-8674(00)80112-9.

Muncan, V. *et al.* (2006) 'Rapid loss of intestinal crypts upon conditional deletion of the wnt/tcf-4 target gene c- myc', *Molecular and Cellular Biology*. American Society for Microbiology, 26(22), pp. 8418–8426. doi: 10.1128/mcb.00821-06.

Murphy, M. *et al.* (2016) 'TCF7L1 modulates colorectal cancer growth by inhibiting expres-

sion of the tumor-suppressor gene EPHB3', *Sci. Rep.*, 6, p. 28299.

Najafov, A. *et al.* (2012) 'MENA is a transcriptional target of the wnt/beta-catenin pathway', *PLoS One*. Public Library of Science (PLoS), 7(5), p. e37013.

Nakamura, Y. *et al.* (2016) 'Tissue- and stage-specific wnt target gene expression is controlled subsequent to β-catenin recruitment', *Development*. The Company of Biologists, 143(11), pp. 1914–1925.

Niehrs, C. (2012) 'The complex world of WNT receptor signalling', *Nature Reviews Molecular Cell Biology*. Springer Science; Business Media LLC, 13(12), pp. 767–779. doi: 10.1038/nrm3470.

Niell, N. *et al.* (2018) 'The humanPKP2/plakophilin-2 gene is induced by Wnt/β-catenin in normal and colon cancer-associated fibroblasts', *Int. J. Cancer*. Wiley, 142(4), pp. 792–804.

Niida, A. *et al.* (2004) 'DKK1, a negative regulator of wnt signaling, is a target of the beta-catenin/TCF pathway', *Oncogene*. Springer Science; Business Media LLC, 23(52), pp. 8520–8526.

Nishioka, M. *et al.* (2018) 'TROY is a promising prognostic biomarker in patients with colorectal cancer', *Oncol. Lett.* Spandidos Publications.

Nishisho, I. *et al.* (1991) 'Mutations of chromosome 5q21 genes in FAP and colorectal cancer patients', *Science*. American Association for the Advancement of Science (AAAS), 253(5020), pp. 665–669. doi: 10.1126/science.1651563.

Noramly, S., Freeman, A. and Morgan, B. A. (1999) 'Beta-catenin signaling can initiate feather bud development', *Development*. The Company of Biologists, 126(16), pp. 3509–3521. doi: 10.1242/dev.126.16.3509.

Nusse, R. *et al.* (1984) 'Mode of proviral activation of a putative mammary oncogene (int-1) on mouse chromosome 15', *Nature*. Springer Science; Business Media LLC, 307(5947), pp. 131–136. doi: 10.1038/307131a0.

Nusse, R. *et al.* (1991) 'A new nomenclature for int-1 and related genes: The wnt gene family', *Cell*. Elsevier BV, 64(2), p. 231. doi: 10.1016/0092-8674(91)90633-a.

Nusse, R. (2012) 'Wnt signaling', *Cold Spring Harbor Perspectives in Biology*. Cold Spring Harbor Laboratory, 4(5), pp. a011163–a011163. doi: 10.1101/cshperspect.a011163.

Nusse, R. and Varmus, H. E. (1982) 'Many tumors induced by the mouse mammary tumor virus contain a provirus integrated in the same region of the host genome', *Cell*. Elsevier BV, 31(1), pp. 99–109. doi: 10.1016/0092-8674(82)90409-3.

Nüsslein-Volhard, C. and Wieschaus, E. (1980) 'Mutations affecting segment number and polarity in drosophila', *Nature*. Springer Science; Business Media LLC, 287(5785), pp. 795–801. doi: 10.1038/287795a0.

Ohsugi, T. *et al.* (2019) 'Anti-apoptotic effect by the suppression of IRF1 as a downstream of Wnt/β-catenin signaling in colorectal cancer cells', *Oncogene*. Springer Science; Business Media LLC, 38(32), pp. 6051–6064.

Ohtsuka, M. *et al.* (2016) 'H19 noncoding RNA, an independent prognostic factor, regulates essential rb-E2F and CDK8-β-catenin signaling in colorectal cancer', *EBioMedicine*. Elsevier BV, 13, pp. 113–124. doi: 10.1016/j.ebiom.2016.10.026.

Oosterwegel, M. A. *et al.* (1991) 'TCF-1, a t cell-specific transcription factor of the HMG box family, interacts with sequence motifs in the TCRβ and TCRδ enhancers', *International Immunology*. Oxford University Press (OUP), 3(11), pp. 1189–1192. doi: 10.1093/intimm/3.11.1189.

Pasca di Magliano, M. and Hebrok, M. (2003) 'Hedgehog signalling in cancer formation and maintenance', *Nat. Rev. Cancer*. Springer Science; Business Media LLC, 3(12), pp. 903–911.

Pate, K. T. *et al.* (2014) 'Wnt signaling directs a metabolic program of glycolysis and angiogenesis in colon cancer', *EMBO J.* EMBO, 33(13), pp. 1454–1473.

Peifer, M. *et al.* (1994) 'Wingless signal and zeste-white 3 kinase trigger opposing changes in the intracellular distribution of armadillo', *Development*. The Company of Biologists, 120(2), pp. 369–380. doi: 10.1242/dev.120.2.369.

Peignon, G. *et al.* (2011) 'Complex interplay between β-catenin signalling and notch effectors in intestinal tumorigenesis', *Gut*. BMJ, 60(2), pp. 166–176.

Pendás-Franco, N. *et al.* (2008) 'DICKKOPF-4 is induced by TCF/beta-catenin and upregulated in human colon cancer, promotes tumour cell invasion and angiogenesis and is repressed by 1alpha,25-dihydroxyvitamin D3', *Oncogene*. Springer Science; Business Media LLC, 27(32), pp. 4467–4477.

Petersen, C. P. and Reddien, P. W. (2009) 'Wnt signaling and the polarity of the primary body axis', *Cell*. Elsevier BV, 139(6), pp. 1056–1068. doi: 10.1016/j.cell.2009.11.035.

Pinto, D. *et al.* (2003) 'Canonical wnt signals are essential for homeostasis of the intestinal epithelium', *Genes & Development*. Cold Spring Harbor Laboratory, 17(14), pp. 1709–1713. doi: 10.1101/gad.267103.

Pinto, D. and Clevers, H. (2005) 'Wnt, stem cells and cancer in the intestine', *Biology of the*

Cell. Wiley, 97(3), pp. 185–196. doi: 10.1042/bc20040094.

Polakis, P. (2000) 'Wnt signaling and cancer', *Genes & Development*. Cold Spring Harbor Laboratory, 14(15), pp. 1837–1851. doi: 10.1101/gad.14.15.1837.

'PRIMA: A software for promoter analysis' (no date). Available at: http://acgt.cs.tau.ac.il/prima/PRIMA.htm.

Qi, J. *et al.* (2016) 'New Wnt/β-catenin target genes promote experimental metastasis and migration of colorectal cancer cells through different signals', *Gut*, 65(10), pp. 1690–1701.

Qiu, W. *et al.* (2010) 'Tumor necrosis factor receptor superfamily member 19 (TNFRSF19) regulates differentiation fate of human mesenchymal (stromal) stem cells through canonical wnt signaling and C/EBP', *J. Biol. Chem.* Elsevier BV, 285(19), pp. 14438–14449.

Ramakrishnan, A.-B. and Cadigan, K. M. (2017) 'Wnt target genes and where to find them', *F1000Research*. F1000 Research Ltd, 6, p. 746. doi: 10.12688/f1000research.11034.1.

Ray, R. *et al.* (2003) 'Up-regulation of fibroblast growth factor-binding protein, by β-catenin during colon carcinogenesis', *Cancer Research*. American Association for Cancer Research, 63(23), pp. 8085–8089. Available at: https://cancerres.aacrjournals.org/content/63/23/8085.

Regan, J. L. *et al.* (2017) 'Non-canonical hedgehog signaling is a positive regulator of the WNT pathway and is required for the survival of colon cancer stem cells', *Cell Reports*. Elsevier BV, 21(10), pp. 2813–2828. doi: 10.1016/j.celrep.2017.11.025.

Reichling, T. *et al.* (2005) 'Transcriptional profiles of intestinal tumors in ApcMin mice are unique from those of embryonic intestine and identify novel gene targets dysregulated in human colorectal tumors', *Cancer Research*. American Association for Cancer Research, 65(1), pp. 166–176. Available at: https://cancerres.aacrjournals.org/content/65/1/166.

Renard, C.-A. *et al.* (2007) 'Tbx3 is a downstream target of the wnt/β-catenin pathway and a critical mediator of β-catenin survival functions in liver cancer', *Cancer Research*. American Association for Cancer Research (AACR), 67(3), pp. 901–910. doi: 10.1158/0008-5472.can-06-2344.

Riggleman, B., Wieschaus, E. and Schedl, P. (1989) 'Molecular analysis of the armadillo locus: Uniformly distributed transcripts and a protein with novel internal repeats are associated with a drosophila segment polarity gene.', *Genes & Development*. Cold Spring Harbor Laboratory, 3(1), pp. 96–113. doi: 10.1101/gad.3.1.96.

Rijsewijk, F. *et al.* (1987) 'The drosophila homology of the mouse mammary oncogene int-1 is identical to the segment polarity gene wingless', *Cell*. Elsevier BV, 50(4), pp. 649–657.

doi: 10.1016/0092-8674(87)90038-9.

Ritchie, M. E. *et al.* (2015) 'limma powers differential expression analyses for RNA-sequencing and microarray studies', *Nucleic Acids Research*, 43(7), pp. e47–e47. doi: 10.1093/nar/gkv007.

Robinson, M. D. and Oshlack, A. (2010) 'A scaling normalization method for differential expression analysis of RNA-seq data', *Genome Biol.* Springer Nature, 11(3), p. R25.

Rockman, S. P. *et al.* (2001) 'Id2 is a target of the beta-catenin/t cell factor pathway in colon carcinoma', *J. Biol. Chem.* Elsevier BV, 276(48), pp. 45113–45119.

Rodilla, V. *et al.* (2009) 'Jagged1 is the pathological link between wnt and notch pathways in colorectal cancer', *Proc. Natl. Acad. Sci. U. S. A.* Proceedings of the National Academy of Sciences, 106(15), pp. 6315–6320.

Roose, J. *et al.* (1999) 'Synergy between tumor suppressor APC and the β-catenin-Tcf4 target Tcf1', *Science*. American Association for the Advancement of Science (AAAS), 285(5435), pp. 1923–1926.

Rubinfeld, B. *et al.* (1993) 'Association of the APC gene product with β-catenin', *Science*. American Association for the Advancement of Science (AAAS), 262(5140), pp. 1731–1734. doi: 10.1126/science.8259518.

Rubinfeld, B. *et al.* (1996) 'Binding of GSK3β to the APC-β-catenin complex and regulation of complex assembly', *Science*. American Association for the Advancement of Science (AAAS), 272(5264), pp. 1023–1026. doi: 10.1126/science.272.5264.1023.

Samuel, A. L. (1959) 'Some studies in machine learning using the game of checkers', *IBM Journal of Research and Development*, 3(3), pp. 210–229. doi: 10.1147/rd.33.0210.

Sánchez-Tilló, E. *et al.* (2011) 'β-catenin/TCF4 complex induces the epithelial-to-mesenchymal transition (EMT)-activator ZEB1 to regulate tumor invasiveness', *Proc. Natl. Acad. Sci. U. S. A.* Proceedings of the National Academy of Sciences, 108(48), pp. 19204–19209.

Scharer, C. D. *et al.* (2009) 'Genome-wide promoter analysis of the SOX4 transcriptional network in prostate cancer cells', *Cancer Research*. American Association for Cancer Research (AACR), 69(2), pp. 709–717. doi: 10.1158/0008-5472.can-08-3415.

Schön, S. *et al.* (2014) 'β-catenin regulates NF-κB activity via TNFRSF19 in colorectal cancer cells', *Int. J. Cancer*. Wiley, 135(8), pp. 1800–1811.

Sekiya, T. *et al.* (2004) 'Identification of BMP and activin membrane-bound inhibitor (BAMBI), an inhibitor of transforming growth factor-beta signaling, as a target of the

beta-catenin pathway in colorectal tumor cells', *J. Biol. Chem.* Elsevier BV, 279(8), pp. 6840–6846.

Seshagiri, S. *et al.* (2012) 'Recurrent r-spondin fusions in colon cancer', *Nature*. Springer Science; Business Media LLC, 488(7413), pp. 660–664.

Sharma, R. P. and Chopra, V. L. (1976) 'Effect of the wingless (wg1) mutation on wing and haltere development in drosophila melanogaster', *Developmental Biology*. Elsevier BV, 48(2), pp. 461–465. doi: 10.1016/0012-1606(76)90108-1.

Shen, M., Zhang, Z. and Wang, P. (2021) 'GLI3 promotes invasion and predicts poor prognosis in colorectal cancer', *Biomed Res. Int.* Hindawi Limited, 2021, p. 8889986.

Shimokawa, T. *et al.* (2003) 'Involvement of the FGF18 gene in colorectal carcinogenesis, as a novel downstream target of the β-catenin/t-cell factor complex', *Cancer Research*. American Association for Cancer Research, 63(19), pp. 6116–6120. Available at: https://cancerres.aacrjournals.org/content/63/19/6116.

Shtutman, M. *et al.* (1999) 'The cyclin D1 gene is a target of the beta-catenin/LEF-1 pathway', *Proc. Natl. Acad. Sci. U. S. A.*, 96(10), pp. 5522–5527.

Siegfried, E., Chou, T.-B. and Perrimon, N. (1992) 'Wingless signaling acts through zeste-white 3, the drosophila homolog of glycogen synthase kinase-3, to regulate engrailed and establish cell fate', *Cell*. Elsevier BV, 71(7), pp. 1167–1179. doi: 10.1016/s0092-8674(05)80065-0.

Sikder, M. O. F. *et al.* (2020) 'SLC6A14, a Na+/Cl–coupled amino acid transporter, functions as a tumor promoter in colon and is a target for wnt signaling', *Biochem. J.* Portland Press Ltd., 477(8), pp. 1409–1425.

Singh, L. N. and Hannenhalli, S. (2008) 'Functional diversification of paralogous transcription factors via divergence in DNA binding site motif and in expression', *PLoS ONE*. Edited by E. T. Dermitzakis. Public Library of Science (PLoS), 3(6), p. e2345. doi: 10.1371/journal.pone.0002345.

Sinner, D. *et al.* (2007) 'Sox17 and Sox4 differentially regulate beta-catenin/t-cell factor activity and proliferation of colon carcinoma cells', *Mol. Cell. Biol.* American Society for Microbiology, 27(22), pp. 7802–7815.

Söderholm, S. and Cantù, C. (2021) 'The WNT/β-catenin dependent transcription: A tissue-specific business', *Wiley Interdiscip. Rev. Syst. Biol. Med.*, 13(3), p. e1511.

Sokol, S. Y. (1999) 'Wnt signaling and dorso-ventral axis specification in vertebrates', *Current Opinion in Genetics & Development*. Elsevier BV, 9(4), pp. 405–410. doi: 10.1016/s0959-

437x(99)80061-6.

Song, J., Wang, Z. and Ewing, R. M. (2014) 'Integrated analysis of the wnt responsive proteome in human cells reveals diverse and cell-type specific networks', *Mol. Biosyst.* Royal Society of Chemistry (RSC), 10(1), pp. 45–53.

Spears, E. and Neufeld, K. L. (2011) 'Novel double-negative feedback loop between adenomatous polyposis coli and Musashi1 in colon epithelia', *J. Biol. Chem.* Elsevier BV, 286(7), pp. 4946–4950.

Sprowl-Tanio, S. *et al.* (2016) 'Lactate/pyruvate transporter MCT-1 is a direct wnt target that confers sensitivity to 3-bromopyruvate in colon cancer', *Cancer Metab.* Springer Science; Business Media LLC, 4(1), p. 20.

Stange, D. E. *et al.* (2013) 'Differentiated troy+ chief cells act as reserve stem cells to generate all lineages of the stomach epithelium', *Cell*. Elsevier BV, 155(2), pp. 357–368.

Stein, U. *et al.* (2006) 'The metastasis-associated gene S100A4 is a novel target of beta-catenin/t-cell factor signaling in colon cancer', *Gastroenterology*. Elsevier BV, 131(5), pp. 1486–1500.

Sung, H. *et al.* (2021) 'Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries', *CA: A Cancer Journal for Clinicians*. Wiley, 71(3), pp. 209–249. doi: 10.3322/caac.21660.

Takahashi, M., Tsunoda, T., *et al.* (2002) 'Identification of membrane-type matrix metalloproteinase-1 as a target of the beta-catenin/Tcf4 complex in human colorectal cancers', *Oncogene*. Springer Science; Business Media LLC, 21(38), pp. 5861–5867.

Takahashi, M., Fujita, M., *et al.* (2002) 'Isolation of a novel human gene, APCDD1, as a direct target of the β-catenin/t-cell factor 4 complex with probable involvement in colorectal carcinogenesis', *Cancer Research*. American Association for Cancer Research, 62(20), pp. 5651–5656. Available at: https://cancerres.aacrjournals.org/content/62/20/5651.

Takahashi, M. *et al.* (2005) 'Identification of SP5 as a downstream gene of the β-catenin/Tcf pathway and its enhanced expression in human colon cancer', *Int J Oncol*. Laboratory of Molecular Medicine, The University of Tokyo, Minato-ku, Tokyo, Japan null, 27(6), pp. 1483–1487. doi: 10.3892/ijo.27.6.1483.

Takahashi, N. *et al.* (2014) 'Identification of two wnt-responsive elements in the intron of RING finger protein 43 (RNF43) gene', *PLoS One*. Public Library of Science (PLoS), 9(1), p. e86582.

Tao, Y. *et al.* (2019) 'Tailored tetravalent antibodies potently and specifically activate

wnt/frizzled pathways in cells, organoids and mice', *eLife*. eLife Sciences Publications, Ltd, 8. doi: 10.7554/elife.46134.

Tetsu, O. and McCormick, F. (1999) 'Beta-catenin regulates expression of cyclin D1 in colon carcinoma cells', *Nature*. Springer Science; Business Media LLC, 398(6726), pp. 422–426.

Therneau, T. and Atkinson, B. (2022) *Rpart: Recursive partitioning and regression trees*. Available at: https://CRAN.R-project.org/package=rpart.

Ulitsky, I. *et al.* (2010) 'Expander: From expression microarrays to networks and functions', *Nat. Protoc.* Springer Science; Business Media LLC, 5(2), pp. 303–322.

Ulloa, F., Itasaki, N. and Briscoe, J. (2007) 'Inhibitory Gli3 activity negatively regulates wnt/β-catenin signaling', *Current Biology*. Elsevier BV, 17(6), pp. 545–550. doi: 10.1016/j.cub.2007.01.062.

Ulloa, F. and Martí, E. (2009) 'Wnt won the war: Antagonistic role of wnt over shh controls dorso-ventral patterning of the vertebrate neural tube', *Developmental Dynamics*. Wiley, pp. NA–NA. doi: 10.1002/dvdy.22058.

Ungerbäck, J. *et al.* (2011) 'The notch-2 gene is regulated by wnt signaling in cultured colorectal cancer cells', *PLoS One*. Public Library of Science (PLoS), 6(3), p. e17957.

Van der Flier, L. G. *et al.* (2007) 'The intestinal Wnt/TCF signature', *Gastroenterology*. Elsevier BV, 132(2), pp. 628–632.

Van Raay, T. J. *et al.* (2005) 'Frizzled 5 signaling governs the neural potential of progenitors in the developing xenopus retina', *Neuron*. Elsevier BV, 46(1), pp. 23–36.

Vignjevic, D. *et al.* (2007) 'Fascin, a novel target of beta-catenin-TCF signaling, is expressed at the invasive front of human colon cancer', *Cancer Res.* American Association for Cancer Research (AACR), 67(14), pp. 6844–6853.

Vlad, A. *et al.* (2008) 'The first five years of the wnt targetome', *Cellular Signalling*. Elsevier BV, 20(5), pp. 795–802. doi: 10.1016/j.cellsig.2007.10.031.

Wang, B. *et al.* (2016) 'Increased expression of SOX4 is associated with colorectal cancer progression', *Tumor Biology*. Springer Science; Business Media LLC, 37(7), pp. 9131–9137. doi: 10.1007/s13277-015-4756-5.

Wang, Q. *et al.* (2018) 'Deptor is a novel target of wnt/β-catenin/c-myc and contributes to colorectal cancer cell growth', *Cancer Res.* American Association for Cancer Research (AACR), 78(12), pp. 3163–3175.

Wang, X., Yang, Y. and Huycke, M. M. (2017) 'Commensal-infected macrophages induce

dedifferentiation and reprogramming of epithelial cells during colorectal carcinogenesis', *Oncotarget*, 8(60), pp. 102176–102190.

Wasserman, W. W. and Sandelin, A. (2004) 'Applied bioinformatics for the identification of regulatory elements', *Nat. Rev. Genet.* Springer Science; Business Media LLC, 5(4), pp. 276–287.

Wassermann, S. *et al.* (2009) 'p16INK4a is a beta-catenin target gene and indicates low survival in human colorectal tumors', *Gastroenterology*. Elsevier BV, 136(1), pp. 196–205.e2.

Watanabe, K. *et al.* (2014) 'Integrative ChIP-seq/microarray analysis identifies a CTNNB1 target signature enriched in intestinal stem cells and colon cancer', *PLoS One*. Public Library of Science (PLoS), 9(3), p. e92317.

Watanabe, M. *et al.* (2012) 'Selective growth inhibition by glycogen synthase kinase-3 inhibitors in tumorigenic HeLa hybrid cells is mediated through NF-κB-dependent GLUT3 expression', *Oncogenesis*. Springer Science; Business Media LLC, 1(7), p. e21.

Waterman ML, E. R., Mortazavi A (no date) 'Disrupting ß-catenin dependent wnt signaling activates an invasive gene program predictive of colon cancer progression'. Available at: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE130236.

Wesenbeeck, L. V. *et al.* (2003) 'Six novel missense mutations in the LDL receptor-related protein 5 (LRP5) gene in different conditions with an increased bone density', *The American Journal of Human Genetics*. Elsevier BV, 72(3), pp. 763–771. doi: 10.1086/368277.

Wetering, M. van de *et al.* (2002) 'The β-catenin/TCF-4 complex imposes a crypt progenitor phenotype on colorectal cancer cells', *Cell*. Elsevier BV, 111(2), pp. 241–250.

Whissell, G. *et al.* (2014) 'The transcription factor GATA6 enables self-renewal of colon adenoma stem cells by repressing BMP gene expression', *Nature Cell Biology*. Springer Science; Business Media LLC, 16(7), pp. 695–707. doi: 10.1038/ncb2992.

Widelitz, R. B. *et al.* (2000) 'β-catenin in epithelial morphogenesis: Conversion of part of avian foot scales into feather buds with a mutated β-catenin', *Developmental Biology*. Elsevier BV, 219(1), pp. 98–114. doi: 10.1006/dbio.1999.9580.

Wielenga, V. J. *et al.* (1999) 'Expression of CD44 in apc and tcf mutant mice implies regulation by the WNT pathway', *Am. J. Pathol.* Elsevier BV, 154(2), pp. 515–523.

Wilhelm, F. *et al.* (2017) 'Troy is expressed in human stomach mucosa and a novel putative prognostic marker of intestinal type gastric cancer', *Oncotarget*. Impact Journals, LLC, 8(31), pp. 50557–50569.

Willert, J. *et al.* (2002) 'A transcriptional response to wnt protein in human embryonic carcinoma cells', *BMC Dev. Biol.*, 2, p. 8.

Wilson, P. A. and Hemmati-Brivanlou, A. (1995) 'Induction of epidermis and inhibition of neural fate by bmp-4', *Nature*. Springer Science; Business Media LLC, 376(6538), pp. 331–333. doi: 10.1038/376331a0.

Wilson, S. *et al.* (2001) 'The status of wnt signalling regulates neural and epidermal fates in the chick embryo', *Nature*. Springer Science; Business Media LLC, 411(6835), pp. 325–330. doi: 10.1038/35077115.

Wingender, E. *et al.* (2000) 'TRANSFAC: an integrated system for gene expression regulation', *Nucleic Acids Research*, 28(1), pp. 316–319. doi: 10.1093/nar/28.1.316.

'Wnt target genes' (no date) *Wnt Target genes. The Wnt Homepage*. Available at: https://web.stanford.edu/group/nusselab/cgi-bin/wnt/target_genes (Accessed: 26 May 2022).

Wong, M. H., Rubinfeld, B. and Gordon, J. I. (1998) 'Effects of forced expression of an NH2-terminal truncated β-catenin on mouse intestinal epithelial homeostasis', *Journal of Cell Biology*. Rockefeller University Press, 141(3), pp. 765–777. doi: 10.1083/jcb.141.3.765.

Xu, X. *et al.* (2021) 'IRF1 regulates the progression of colorectal cancer via interferon-induced proteins', *Int. J. Mol. Med.*, 47(6).

Xue, W. *et al.* (2018) 'FOXN3 is downregulated in osteosarcoma and transcriptionally regulates SIRT6, and suppresses migration and invasion in osteosarcoma', *Oncology Reports*. Spandidos Publications. doi: 10.3892/or.2018.6878.

Yamada, T. *et al.* (2000) 'Transactivation of the multidrug resistance 1 gene by t-cell factor 4/β-catenin complex in early colorectal carcinogenesis', *Cancer Research*. American Association for Cancer Research, 60(17), pp. 4761–4766. Available at: https://cancerres.aacrjournals.org/content/60/17/4761.

Yan, B. *et al.* (2020) 'Interferon regulatory factor-1 (IRF1) activates autophagy to promote liver ischemia/reperfusion injury by inhibiting β-catenin in mice', *PLoS One*. Public Library of Science (PLoS), 15(11), p. e0239119.

Yan, D. *et al.* (2001) 'Elevated expression of axin2 and hnkd mRNA provides evidence that wnt/beta -catenin signaling is activated in human colon tumors', *Proc. Natl. Acad. Sci. U. S. A.* Proceedings of the National Academy of Sciences, 98(26), pp. 14973–14978.

Yochum, G. S. *et al.* (2007) 'Serial analysis of chromatin occupancy identifies beta-catenin target genes in colorectal carcinoma cells', *Proc. Natl. Acad. Sci. U. S. A.* Proceedings of the National Academy of Sciences, 104(9), pp. 3324–3329.

Yu, T. *et al.* (2012) 'Regulation of the potential marker for intestinal cells, Bmi1, by β-catenin and the zinc finger protein KLF4: Implications for colon cancer', *J. Biol. Chem.* Elsevier BV, 287(6), pp. 3760–3768.

Zhang, T. *et al.* (2001) 'Evidence that APC regulates survivin expression', *Cancer Research*. American Association for Cancer Research, 61(24), pp. 8664–8667. Available at: https://cancerres.aacrjournals.org/content/61/24/8664.

Zhang, X., Gaspard, J. P. and Chung, D. C. (2001) 'Regulation of vascular endothelial growth factor by the wnt and k-ras pathways in colonic neoplasia', *Cancer Research*. American Association for Cancer Research, 61(16), pp. 6050–6054. Available at: https://cancerres.aacrjournals.org/content/61/16/6050.

Zhao, C. *et al.* (2020) 'FOXN3 suppresses the growth and invasion of papillary thyroid cancer through the inactivation of wnt/β-catenin pathway', *Molecular and Cellular Endocrinology*. Elsevier BV, 515, p. 110925. doi: 10.1016/j.mce.2020.110925.

Zhou, P. *et al.* (1995) 'Lymphoid enhancer factor 1 directs hair follicle patterning and epithelial cell fate.', *Genes & Development*. Cold Spring Harbor Laboratory, 9(6), pp. 700–713. doi: 10.1101/gad.9.6.700.

Zhu, C. *et al.* (2017) 'Identification of FERM domain-containing protein 5 as a novel target of β-catenin/TCF7L2 complex', *Cancer Sci.* Wiley, 108(4), pp. 612–619.

Zimmerli, D. *et al.* (2020) 'TBX3 acts as tissue-specific component of the Wnt/β-catenin transcriptional complex', *Elife*. eLife Sciences Publications, Ltd, 9.

APENDICES

Appendix 1 Bonafide target genes

Table 1.1: List of bonafide target genes as described in the literature. Each gene is identified by their NCBI locus id. If the corresponding paper reports the target as *direct* we indicate it with "yes". When the paper does not report the gene as a direct or indirect target, the cell is empty.

Locus Id	Gene Name	Direct	up/down	Reference
5243	ABCB1	yes	up	(Yamada *et al.*, 2000)
102	ADAM10		up	(Gavert *et al.*, 2007)
10551	AGR2	yes	up	(Dahal Lamichane *et al.*, 2019)
147495	APCDD1	yes	up	(Takahashi, Fujita, *et al.*, 2002)
51309	ARMCX1		up	(Iseki *et al.*, 2010)
430	ASCL2	yes	up	(Jubb *et al.*, 2006; Basu *et al.*, 2018)
467	ATF3	yes	up	(Inoue *et al.*, 2018)
474	ATOH1		down	(Leow *et al.*, 2004)
8313	AXIN2	yes	up	(Yan *et al.*, 2001; Lustig *et al.*, 2002; Jho *et al.*, 2002)
25805	BAMBI	yes		(Sekiya *et al.*, 2004)
599	BCL2L2	yes	up	(Lapham *et al.*, 2009)
332	BIRC5	yes	up	(Zhang *et al.*, 2001; Kim *et al.*, 2003)
648	BMI1		up	(Yu *et al.*, 2012)
652	BMP4		up	(Kim *et al.*, 2002)
23246	BOP1	yes	up	(Qi *et al.*, 2016)
595	CCND1	yes	up	(Tetsu and McCormick, 1999; Shtutman *et al.*, 1999)
960	CD44		up	(Wielenga *et al.*, 1999)
1029	CDKN2A	yes	up	(Wassermann *et al.*, 2009)

Continued on next page

Locus Id	Gene Name	Direct	up/down	Reference
1044	CDX1	yes	up	(Lickert *et al.*, 2000; Domon-Dell and Freund, 2002)
1164	CKS2	yes	up	(Qi *et al.*, 2016)
9076	CLDN1	yes	up	(Miwa *et al.*, 2001)
9201	DCLK1		up	(Wang, Yang and Huycke, 2017)
64798	DEPTOR	yes	up	(Wang *et al.*, 2018)
22943	DKK1	yes	up	(Niida *et al.*, 2004)
27121	DKK4		up	(Pendás-Franco *et al.*, 2008; Baehs *et al.*, 2009)
1786	DNMT1		up	(Campbell and Szyf, 2003)
1906	EDN1		up	(Kim *et al.*, 2005)
1947	EFNB1		up/down	(Batlle *et al.*, 2002)
55740	ENAH	yes	up	(Najafov *et al.*, 2012)
8507	ENC1	yes	up	(Fujita *et al.*, 2001)
2048	EPHB2		up	(Batlle *et al.*, 2002)
2049	EPHB3	yes	up	(Batlle *et al.*, 2002; Murphy *et al.*, 2016)
8817	FGF18	yes	up	(Shimokawa *et al.*, 2003)
9982	FGFBP1	yes	up	(Ray *et al.*, 2003)
8061	FOSL1	yes	up	(Mann *et al.*, 1999)
94234	FOXQ1	yes	up	(Christensen *et al.*, 2013)
84978	FRMD5	yes	up	(Zhu *et al.*, 2017)
6624	FSCN1	yes	up	(Vignjevic *et al.*, 2007)
2520	GAST		up	(Koh *et al.*, 2000)
3280	HES1	yes	up	(Peignon *et al.*, 2011)
3159	HMGA1	yes	up	(Bush *et al.*, 2013)
3158	HMGCS2		down	(Kim *et al.*, 2019)
3398	ID2	yes	up	(Rockman *et al.*, 2001; Willert *et al.*, 2002)
182	JAG1		up	(Rodilla *et al.*, 2009)
3725	JUN	yes	up	(Mann *et al.*, 1999)

Continued on next page

Locus Id	Gene Name	Direct	up/down	Reference
3897	L1CAM		up	(Gavert *et al.*, 2005)
3918	LAMC2	yes	up	(Hlubek *et al.*, 2001)
120071	LARGE2	yes	up	(Dietinger, García de Durango, *et al.*, 2020)
51176	LEF1	yes	up	(Hovanes *et al.*, 2001; Filali *et al.*, 2002)
8549	LGR5		up	(Barker *et al.*, 2007)
346389	MACC1	yes	up	(Kim *et al.*, 2018)
4233	MET		up	(Boon *et al.*, 2002)
4312	MMP1	yes	up	(Takahashi, Tsunoda, *et al.*, 2002)
4323	MMP14	yes	up	(Hlubek *et al.*, 2004)
4316	MMP7	yes	up	(Brabletz *et al.*, 1999; Crawford *et al.*, 1999)
339287	MSL1		up	(Spears and Neufeld, 2011)
4513	MT-CO2	yes	up	(Araki *et al.*, 2003)
4602	MYB		up	(Wetering *et al.*, 2002)
4609	MYC	yes	up	(He *et al.*, 1998)
26292	MYCBP	yes	up	(Jung and Kim, 2005)
4739	NEDD9	yes	up	(Li *et al.*, 2011)
9603	NFE2L3	maybe	up	(Aono *et al.*, 2019)
4783	NFIL3	yes	up	(Qi *et al.*, 2016)
4843	NOS2	yes	up	(Du *et al.*, 2006)
4853	NOTCH2	yes	up	(Ungerbäck *et al.*, 2011)
4897	NRCAM	yes	up	(Conacci-Sorrell *et al.*, 2002)
5163	PDK1	yes	up	(Pate *et al.*, 2014)
5318	PKP2	yes	up	(Niell *et al.*, 2018)
5328	PLAU	yes	up	(Hiendlmeyer *et al.*, 2004)
5329	PLAUR		up	(Mann *et al.*, 1999)
5467	PPARD	yes	up	(He *et al.*, 1999)
9232	PTTG1	yes	up	(Hlubek *et al.*, 2006)

Continued on next page

Locus Id	Gene Name	Direct	up/down	Reference
54894	RNF43	yes	up	(Takahashi *et al.*, 2014)
6275	S100A4	yes	up	(Stein *et al.*, 2006)
6277	S100A6		up	(Kilańczyk *et al.*, 2012)
6304	SATB1	yes	up	(Mir *et al.*, 2016)
6446	SGK1	yes	up	(Dehner *et al.*, 2008)
22933	SIRT2	yes	down	(Li *et al.*, 2021)
6566	SLC16A1	yes	up	(Sprowl-Tanio *et al.*, 2016)
11254	SLC6A14	yes	up	(Sikder *et al.*, 2020)
9126	SMC3	yes	up	(Ghiselli *et al.*, 2003)
6662	SOX9		up	(Blache *et al.*, 2004)
389058	SP5	yes	up	(Takahashi *et al.*, 2005)
6428	SRSF3	no	up	(Gonçalves, Matos and Jordan, 2008)
23512	SUZ12	yes	up	(Kirmizis, Bartley and Farnham, 2003)
6925	TCF4	yes	up	(Kolligs *et al.*, 2002)
6932	TCF7	yes	up	(Roose *et al.*, 1999)
7074	TIAM1			(Malliri *et al.*, 2006)
3371	TNC	yes	up	(Beiter *et al.*, 2005)
7422	VEGFA	yes	up	(Zhang, Gaspard and Chung, 2001)
63929	XPNPEP3	yes	up	(Kumar *et al.*, 2018)
10413	YAP1	yes	up	(Konsavage *et al.*, 2012)
6935	ZEB1	yes	up	(Sánchez-Tilló *et al.*, 2011)

Appendix 2 Prediction list 1

Table 2.1: The list of predicted genes by the classifiers trained with TFBS matrix 432 and 66 bonafide genes, and hyperparameters sample size 8000 and penalty 100. The frequency shows how many times a gene is predicted as target in 10 shuffling instances.

ENTREZID	SYMBOL	Freq
10981	RAB32	10
145447	ABHD12B	10
158801	NKAPP1	10
221150	SKA3	10
223117	SEMA3D	10
2308	FOXO1	10
256714	MAP7D2	10
26053	AUTS2	10
27164	SALL3	10
284382	ACTL9	10
29986	SLC39A2	10
346	APOC4	10
4099	MAG	10
51655	RASD1	10
5292	PIM1	10
5345	SERPINF2	10
56288	PARD3	10
56659	KCNK13	10
5830	PEX5	10
6473	SHOX	10
64784	CRTC3	10
65977	PLEKHA3	10
7274	TTPA	10
7472	WNT2	10
7915	ALDH5A1	10

Continued on next page

ENTREZID	SYMBOL	Freq
79693	YRDC	10
80139	ZNF703	10
84874	ZNF514	10
8491	MAP4K3	10
85285	KRTAP4-1	10
8538	BARX2	10
8911	CACNA1I	10
9775	EIF4A3	10
121274	ZNF641	9
2561	GABRB2	9
30968	STOML2	9
7122	CLDN5	9
257019	FRMD3	8
55129	ANO10	8
7429	VIL1	8
27294	DHDH	7
9665	MARF1	4
348327	ZNF530	3
54681	P4HTM	2
54965	PIGX	2
10771	ZMYND11	1
1487	CTBP1	1
55277	FGGY	1
57787	MARK4	1
708	C1QBP	1
81532	MOB2	1
84300	UQCC2	1
8941	CDK5R2	1

Appendix 3 Prediction list 2

Table 3.1: List of predicted genes by the classifiers trained with TFBS matrix 432 and 66 bonafide genes, and hyperparameters sample size 1000 and penalty 20.

ENTREZID	SYMBOL	AV_SCORE
10981	RAB32	486.2
2192	FBLN1	425.8
8491	MAP4K3	407.4
84874	ZNF514	389.3
256714	MAP7D2	385.3
51608	GET4	375.4
145447	ABHD12B	375.4
3742	KCNA6	373.8
51655	RASD1	370.9
7915	ALDH5A1	367.9
10166	SLC25A15	358.9
151449	GDF7	354.3
80764	THAP7	353.6
25958	NA	352.7
6553	SLC9A5	347
56659	KCNK13	345.4
708	C1QBP	341.9
84232	MAF1	339.9
8538	BARX2	337.5
221491	SMIM29	332.2
56288	PARD3	328.3
64122	FN3K	326.6
54965	PIGX	324.4
4088	SMAD3	323.7
7407	VARS1	322.1
10135	NAMPT	319.1

Continued on next page

ENTREZID	SYMBOL	AV_SCORE
51601	LIPT1	318.6
55686	MREG	317.8
25945	NECTIN3	313.7
54512	EXOSC4	312.2
79693	YRDC	312.1
256302	NATD1	307.3
79685	SAP30L	306.8
9675	TTI1	304.8
4543	MTNR1A	304.6
1284	COL4A2	302.7
51305	KCNK9	301.3
92736	OTOP2	300.7
4854	NOTCH3	300.1
149951	COMMD7	299
26040	SETBP1	295.7
463	ZFHX3	294.3
3068	HDGF	294.3
55539	KCNQ1DN	294.1
84417	ECRG4	294
283149	BCL9L	293.5
55670	PEX26	292.3
6314	ATXN7	289.9
58490	RPRD1B	287.7
23759	PPIL2	281.4
158801	NKAPP1	279.5
29927	SEC61A1	278.3
152189	CMTM8	276.8
83655	LINC00208	276.7
6532	SLC6A4	276.2
166614	DCLK2	274.1
201627	DENND6A	273.8

Continued on next page

ENTREZID	SYMBOL	AV_SCORE
7472	WNT2	273.6
79073	TMEM109	273.3
81631	MAP1LC3B	273.2
6446	SGK1	272.5
5311	PKD2	271.9
7274	TTPA	269.7
6812	STXBP1	269
824	CAPN2	268.8
64409	GALNT17	268.8
29081	METTL5	268
10529	NEBL	266.4
6597	SMARCA4	265.8
5912	RAP2B	265
2869	GRK5	264.7
5917	RARS1	264.6
9955	HS3ST3A1	264.1
3958	LGALS3	263.9
2899	GRIK3	263.5
57118	CAMK1D	262.2
79706	PRKRIP1	261.9
145501	ISM2	256.9
80139	ZNF703	255.6
51315	KRCC1	255
10263	CDK2AP2	254.1
6199	RPS6KB2	250.8
57554	LRRC7	250.3
29091	STXBP6	249.7
29775	CARD10	249.1
3768	KCNJ12	248.1
10294	DNAJA2	247.5
51256	TBC1D7	245.2

Continued on next page

ENTREZID	SYMBOL	AV_SCORE
5825	ABCD3	245
147015	DHRS13	245
83549	UCK1	241.9
257019	FRMD3	240.1
8976	WASL	240
79070	POGLUT2	239.6
79856	SNX22	239.3
4716	NDUFB10	239
1949	EFNB3	237.7
85364	ZCCHC3	237.6
5791	PTPRE	237
113730	KLHDC7B	236.2
8745	ADAM23	235.2
81566	CSRNP2	234
4330	MN1	233.8
115950	ZNF653	233.6
64837	KLC2	232.3
51126	NAA20	230.8
5292	PIM1	228.4
115426	UHRF2	228.2
6575	SLC20A2	227.5
9863	MAGI2	227.3
5581	PRKCE	226.1
6578	SLCO2A1	226
83643	CCDC3	225.2
23401	FRAT2	222.5
30968	STOML2	219.2

Appendix 4 Prediction list 3

Table 4.1: List of predicted genes by the classifiers trained with TFBS matrix 432 and 93 bonafide genes, and hyperparameters sample size 8000 and penalty 50.

SYMBOL	ENTREZID	AV_SCORE
RAB32	10981	998.2
FANK1	92565	997.8
GET4	51608	997.8
BARX2	8538	997.8
COL4A2	1284	997.8
TBC1D2	55357	996.5
PMS2P5	5383	971.3
RPS23	6228	885.2
GCSH	2653	789.1
ZC3H18	124245	781.8
MAGI2	9863	774.2
TRAF2	7186	712
WNT2	7472	705.4
SAP30L	79685	686
SPCS1	28972	664.9
PARD3	56288	600.1
EXT2	2132	595.7
KLF10	7071	595.6
RRAGD	58528	576.2
RASAL2	9462	556.1
MAP7D2	256714	539.3
EIF2B2	8892	535.1
MN1	4330	530.9
PIGX	54965	530.7
DCLK2	166614	522.7
ZFHX3	463	518.9

Continued on next page

SYMBOL	ENTREZID	AV_SCORE
C1QBP	708	518.4
ATXN7	6314	515.5
SKAP1	8631	488.6
ZNF514	84874	442.8
SMIM29	221491	437.7
ELK4	2005	435.4
RAC3	5881	435.1
ARSL	415	431.2
CNN2	1265	428.2
ZNF653	115950	413.2
SLC6A4	6532	402
EXOSC4	54512	401.5
UBE2N	7334	398.2
YRDC	79693	398.1
DPY19L2	283417	391.7
LRFN5	145581	390.8
JMJD1C	221037	390.2
RNF24	11237	386.1
SETDB1	9869	381
ACP5	54	372.7
SMAD3	4088	371.3
TRMT6	51605	370
NDUFAF6	137682	368.2
PPIC	5480	366.9
NUP43	348995	366.5
SOX14	8403	365.7
SUN5	140732	359.4
NDUFAB1	4706	356.7
SLC2A11	66035	356.5
GAS6	2621	356.1
RPL29	6159	354.6

Continued on next page

SYMBOL	ENTREZID	AV_SCORE
IER2	9592	353.9
AMZ2	51321	353.3
SLFN13	146857	343.2
RPTOR	57521	338.9
SH2D3A	10045	336
RSPH10B	222967	334.5
PLB1	151056	332.7
CHST3	9469	332.2
PEBP1	5037	326.5
PEA15	8682	325.3
KLHDC7B	113730	324.9
TMEM245	23731	324.6
CIRBP-AS1	148046	324.5
MARVELD1	83742	319.7
NIPA1	123606	319.3
ZNF703	80139	317.9
HMCES	56941	317.3
FYCO1	79443	316.5
CCDC60	160777	316.4
WNT10A	80326	315.9
CUBN	8029	315.6
MAP1LC3B	81631	314.3
RBL2	5934	312.8
RPS19	6223	306.2
MAN2A2	4122	304.4
CCN1	3491	303.6
VAMP5	10791	298.7
RASL10A	10633	298.7

Appendix 5 Prediction list 4

Table 5.1: List of predicted genes by the classifiers trained with RNAseq data and 93 bonafide genes, and hyperparameters sample size 3000 and penalty 50.

ENTREZID	SYMBOL	AV_SCORE
2737	GLI3	908.9
3213	HOXB3	907.6
6659	SOX4	903.4
7078	TIMP3	888.7
5727	PTCH1	880.9
10221	TRIB1	873.4
10581	IFITM2	862.7
4088	SMAD3	859.4
55890	GPRC5C	850.6
119559	SFXN4	837.8
25778	DSTYK	834.6
2273	FHL1	826.5
27143	PALD1	817.2
51573	GDE1	816.4
51701	NLK	810.9
26022	TMEM98	809.4
85014	TMEM141	808.9
10160	FARP1	803
56937	PMEPA1	799
7126	TNFAIP1	791.8
5880	RAC2	790.5
2963	GTF2F2	785.9
84679	SLC9A7	784.2
284023	RNF227	779.2
10370	CITED2	779
23216	TBC1D1	771.9

Continued on next page

ENTREZID	SYMBOL	AV_SCORE
8536	CAMK1	768
90861	JPT2	766.7
1948	EFNB2	766.2
64283	ARHGEF28	766
3983	ABLIM1	765.3
9882	TBC1D4	759.5
3980	LIG3	755.9
27090	ST6GALNAC4	748.8
81539	SLC38A1	748.5
5144	PDE4D	747.6
3490	IGFBP7	746.3
11182	SLC2A6	744.8
3340	NDST1	738.7
3685	ITGAV	737.3
55568	GALNT10	736.5
22836	RHOBTB3	734.5
3659	IRF1	731.9
7436	VLDLR	731.9
100499466	LINC00674	730.6
64780	MICAL1	719.8
51809	GALNT7	710.8
8879	SGPL1	710.4
57664	PLEKHA4	709.8
27237	ARHGEF16	707
10892	MALT1	704.7
157285	PRAG1	704.4
200058	NA	698.6
55604	CARMIL1	698.4
85301	COL27A1	694.4
23363	OBSL1	692.8
23531	MMD	692.7

Continued on next page

ENTREZID	SYMBOL	AV_SCORE
5187	PER1	690.9
2948	GSTM4	689.7
5971	RELB	689.6
29843	SENP1	687.3
143098	MPP7	684.1
999	CDH1	682.7
6515	SLC2A3	682.4
100507246	SNHG16	682.3
6533	SLC6A6	681.6
7431	VIM	681.4
5818	NECTIN1	679
3801	KIFC3	675.3
3475	IFRD1	666.8
83693	HSDL1	665.6
92421	CHMP4C	662
26012	NSMF	660.9
10217	CTDSPL	659.8
22979	EFR3B	659.3
64795	RMND5A	658.6
3673	ITGA2	658.4
351	APP	656.1
92610	TIFA	653.7
6508	SLC4A3	648.2
56905	C15orf39	647.2
60489	APOBEC3G	642.6
55332	DRAM1	642
5264	PHYH	639.8
51491	NOP16	639.4
9249	DHRS3	637.2
57493	HEG1	635
79763	ISOC2	633.5

Continued on next page

ENTREZID	SYMBOL	AV_SCORE
3995	FADS3	626.3
10123	ARL4C	624.4
330	BIRC3	623.4
64097	EPB41L4A	612.9
9646	CTR9	605.4
3855	KRT7	604.7
26010	SPATS2L	595.4
54434	SSH1	591.9
9111	NMI	590.5
5947	RBP1	586.1
2886	GRB7	584.1
56652	TWNK	582.2
55857	KIZ	581.7
7804	LRP8	577.6
27230	SERP1	576.4
144811	LACC1	576.1
2706	GJB2	575.8
10940	POP1	572.6
51393	TRPV2	571
6774	STAT3	569.8
2335	FN1	562.1
147657	ZNF480	561.6
7351	UCP2	561.4
84612	PARD6B	561.3
9748	SLK	560.7
5530	PPP3CA	559.8
4053	LTBP2	556.3
64065	PERP	555.4
26049	FAM169A	551.1
7461	CLIP2	545.8
5699	PSMB10	543.9

Continued on next page

ENTREZID	SYMBOL	AV_SCORE
2644	GCHFR	543.8
858	CAV2	537.4
120224	TMEM45B	536.5
3300	DNAJB2	536.4
57406	ABHD6	536.1
5125	PCSK5	536.1
60386	SLC25A19	533.2
55080	TAPBPL	532.6
28962	OSTM1	530
347735	SERINC2	529
23650	TRIM29	523.7
4129	MAOB	521.6
10105	PPIF	519.4
91010	FMNL3	516.1
144501	KRT80	511.3
6926	TBX3	511.1
7185	TRAF1	507.8
56917	MEIS3	506.8
5493	PPL	505.7
64135	IFIH1	503.8
23082	PPRC1	503.8
1827	RCAN1	503.1
684	BST2	502.5
1520	CTSS	501.2
3489	IGFBP6	500.1

Appendix 6 Prediction list 5

Table 6.1: List of predicted genes by the classifiers trained with RNAseq and Microarray data combined and 93 bonafide genes, and hyperparameters sample size 1000 and penalty 50.

ENTREZID	SYMBOL	AV_SCORE
27143	PALD1	367.7
7372	UMPS	359.8
89958	SAPCD2	357.2
5727	PTCH1	353.7
54733	SLC35F2	351.3
5613	PRKX	348.9
84133	ZNRF3	348.2
196	AHR	345.2
6934	TCF7L2	344.4
51646	YPEL5	342
7804	LRP8	338.1
84159	ARID5B	337.1
9334	B4GALT5	333.1
1952	CELSR2	333.1
6533	SLC6A6	333
51141	INSIG2	331.9
6659	SOX4	331
3268	AGFG2	330.2
55088	CCDC186	329.8
55502	HES6	329.7
4173	MCM4	329.3
9277	WDR46	329.2
3983	ABLIM1	326.3
1736	DKC1	326.3
84678	KDM2B	325.6
51491	NOP16	325

Continued on next page

ENTREZID	SYMBOL	AV_SCORE
84033	OBSCN	324.8
51131	PHF11	324.4
57478	USP31	321.4
219	ALDH1B1	320.8
728554	LOC728554	320.3
9882	TBC1D4	320.2
5985	RFC5	319.4
55890	GPRC5C	319
55540	IL17RB	317.9
8214	DGCR6	317.5
10221	TRIB1	317.5
51659	GINS2	315.3
388	RHOB	311.7
2956	MSH6	311.2
5198	PFAS	310.8
2934	GSN	310.6
3070	HELLS	310.6
23636	NUP62	310.6
4174	MCM5	310.3
7019	TFAM	307.8
100499467	LINC00673	307.5
26010	SPATS2L	306.9
10892	MALT1	306.5
51293	CD320	305.9
5264	PHYH	305.4
83641	FAM107B	304.1
5144	PDE4D	302.8
80227	PAAF1	302
64285	RHBDF1	300.8
118987	PDZD8	300.6
23150	FRMD4B	299.9

Continued on next page

ENTREZID	SYMBOL	AV_SCORE
1434	CSE1L	299.4
90233	ZNF551	299.2
55210	ATAD3A	298.9
23464	GCAT	298.8
28227	PPP2R3B	298.4
7039	TGFA	297.8
27090	ST6GALNAC4	297.7
84172	POLR1B	295.3
56652	TWNK	295.2
51537	MTFP1	294.5
26503	SLC17A5	293.8
5205	ATP8B1	293.4
6573	SLC19A1	292
1163	CKS1B	291.6
64283	ARHGEF28	290.4
128272	ARHGEF19	289.8
109	ADCY3	289.1
8914	TIMELESS	289.1
203427	SLC25A43	288.8
10580	SORBS1	288.2
6390	SDHB	287.1
203062	TSNARE1	286.9
4175	MCM6	286.5
58472	SQOR	285.8
156	GRK2	284.6
114876	OSBPL1A	284.6
79019	CENPM	284.4
440081	DDX12P	284.2
23531	MMD	283.6
2150	F2RL1	282.1
90861	JPT2	282.1

Continued on next page

ENTREZID	SYMBOL	AV_SCORE
6692	SPINT1	282
84612	PARD6B	281.4
1528	CYB5A	281.2
54892	NCAPG2	281.1
119559	SFXN4	281.1
1723	DHODH	280.9
26135	SERBP1	280
8293	SERF1A	279.9
10459	MAD2L2	279.6
11112	HIBADH	277.7
64795	RMND5A	277.1
28231	SLCO4A1	276.8
26278	SACS	276.7
55367	PIDD1	276.5
4221	MEN1	276.2
26012	NSMF	275.9
2707	GJB3	275.3
3300	DNAJB2	274.8
9654	TTLL4	274.8
84321	THOC3	274.7
11198	SUPT16H	274.6
7517	XRCC3	274.2
23165	NUP205	274
9837	GINS1	273.7
51154	MRTO4	273.1
23338	JADE2	272.8
55611	OTUB1	272.6
388115	CCDC9B	272
64780	MICAL1	271.8
284023	RNF227	271.5
790	CAD	271.3

Continued on next page

ENTREZID	SYMBOL	AV_SCORE
8602	NOP14	271.3
79646	PANK3	271.1
1846	DUSP4	269.4
129293	TRABD2A	269
65008	MRPL1	268.9
10370	CITED2	267.9
92370	PXYLP1	267.9
7414	VCL	267.8
2648	KAT2A	267.3
8476	CDC42BPA	266.9
10217	CTDSPL	266.1
6502	SKP2	266.1
55038	CDCA4	265.6
835	CASP2	265.1
2043	EPHA4	264.8
79004	CUEDC2	264.3
29946	SERTAD3	264.2
27229	TUBGCP4	262.2
4249	MGAT5	261.9
55651	NHP2	261.8
3659	IRF1	261.1
5653	KLK6	261.1
26469	PTPN18	260.6
2091	FBL	260
25851	TECPR1	259.8
3146	HMGB1	259.7
54498	SMOX	259.6
1611	DAP	258.3
6666	SOX12	258.3
84679	SLC9A7	257.8
5721	PSME2	257.5

Continued on next page

ENTREZID	SYMBOL	AV_SCORE
142	PARP1	257.4
9748	SLK	257.4
7083	TK1	257.4
84886	C1orf198	257
3340	NDST1	256.4
63979	FIGNL1	256.2
29843	SENP1	256.2
56257	MEPCE	256
2068	ERCC2	255.9
100134229	KDM7A-DT	255.9
613	BCR	255.8
7298	TYMS	255.7
6768	ST14	255.2
22836	RHOBTB3	254.7
56937	PMEPA1	254.1
999	CDH1	253.8
23082	PPRC1	253.6
7301	TYRO3	253.3
9950	GOLGA5	252.4
55080	TAPBPL	252.3
23650	TRIM29	252.3
3485	IGFBP2	252.2
8270	LAGE3	251.4
51433	ANAPC5	251.3
8853	ASAP2	251.1
8879	SGPL1	250.7
144501	KRT80	250.6
157285	PRAG1	250.3
60386	SLC25A19	250.1

Appendix 7 Array Analysis Protocol

CEL files are raw data files generated by Affymetrix microarray software. They contain information about the probes used in array. Before doing a gene expression analysis, first, we must process the CEL file to extract the expression values for each feature

We will use the GEO series GSE87429 as an example. We analysed rest of the data we chose for our study in the same manner.

The experiment was conducted with platform GPL17586 and has four samples. Platforms are commercial gene chips that contain oligonucleotides for DNA to hybridize. There are variety of different microarray platform and it is important to know the type of platform used in an experiment before hand in order to conduct the further analyses.

First, we download CEL files from GEO (https://www.ncbi.nlm.nih.gov/geo/) using experiment's accession number. The CEL files are usually in archive format (tar or zip) we have to extract the CEL files first. Files need to be in working directory or we can set our working directory to CEL files' location.

To read CEL files we need Bioconductor libraries oligo, oligoClasses and affyio.

First we'll get the list of CEL files in directory.

```
cel.files <- list.celfiles()
cel.files
## [1] "GSM2331244_A_GAC-MO-A_HTA-2_0_.CEL"
## [2] "GSM2331245_B_GAC-MO-B_HTA-2_0_.CEL"
## [3] "GSM2331246_E_GAC-MO-E_HTA-2_0_.CEL"
## [4] "GSM2331247_F_GAC-MO-F_HTA-2_0_.CEL"
```

Then we'll create a vector for sample names. Order of the sample name vector must correspond to the CEL file list so that file and sample names match correctly. It is not necessary to create sample names but sometimes the files names can be lengthy so creating sample names may come in handy in these situations.

```
sample.names <- c("control1", "treatment1", "control2", "treatment2")
```

Now we'll read the files passing list and sample names to `read.celfile` function and store in `affy_raw_data`.

```
affy_raw_data <- read.celfiles(cel.files, sampleNames = sample.names)
## Reading in : GSM2331244_A_GAC-MO-A_HTA-2_0_.CEL
## Reading in : GSM2331245_B_GAC-MO-B_HTA-2_0_.CEL
## Reading in : GSM2331246_E_GAC-MO-E_HTA-2_0_.CEL
## Reading in : GSM2331247_F_GAC-MO-F_HTA-2_0_.CEL
```

Now we'll perform rma normalization and save the expression data file in txt format.

```
rma_normalized <- rma(affy_raw_data)
## Background correcting
## Normalizing
## Calculating Expression
```

Now we'll retrieve the normalized expression values for each probe.

```
expr <- data.frame(exprs(rma_normalized))
```

Table 7.1: Normalized gene expression

	control1	treatment1	control2	treatment2
2824546_st	12.29690	12.22003	12.36024	12.32230
2824549_st	11.88990	11.75083	11.87327	11.91384
2824551_st	12.11241	11.96338	12.11936	12.03423
2824554_st	11.97461	11.76430	11.85788	11.87959
2827992_st	11.58672	11.47740	11.54053	11.43947
2827995_st	11.15864	11.21040	11.25329	11.12399

The expression set is loaded. Now we'll perform differential expression analysis.

Appendix 7.1 Differential expression

To see the differentially expressed genes we will use the limma package. Limma builds linear models for microarrays and other transcriptomic data.

```
library(limma)
```

We need to specify the sample groups as factor.

```
sample.groups <- factor(c("C", "T", "C", "T"),
                        levels = c("C", "T"))
```

Now we'll create design matrix.

```
design.mat <- model.matrix(~ 0 + ~sample.groups)
colnames(design.mat) = c("C", "T")
```

Using the design matrix a contrast matrix is built which will be used while building the linear model.

```
contrast.matrix <- makeContrasts(C-T, levels=design.mat)
```

Using the normalized expression data, design matrix and contrast matrix we built linear model and fit the model.

```
fit <- lmFit(rma_normalized, design.mat)
fit2 = contrasts.fit(fit, contrast.matrix)
fit2 = eBayes(fit2)
```

We will store the differential expression table using the `topTable` function

```
table <- topTable(fit2, number = Inf, coef = 1)
head(table) %>% kable(caption = "Differential expression table")
```

Table 7.2: Differential expression table

	logFC	AveExpr	t	P.Value	adj.P.Val	B
TC05000550.hg.1	-2.664223	7.489702	-33.70718	0e+00	0.0008736	8.382083
TC05002579.hg.1	-2.659631	8.554089	-32.97093	0e+00	0.0008736	8.324418
TC06000135.hg.1	2.414019	8.405235	29.93135	0e+00	0.0010717	8.052560
TC04000478.hg.1	-2.096694	5.713626	-27.82441	1e-07	0.0012729	7.826071
TC07000810.hg.1	-2.232153	7.622238	-26.22526	1e-07	0.0013466	7.628687
TC15002249.hg.1	-1.929155	6.206933	-25.85557	1e-07	0.0013466	7.579506

We have probe names in the first column. Now we'll annotate them by finding which probe corresponding to which gene.

Appendix 7.2 Gene Annotations

To do the annotation first we will store the row names as a column.

```
table <- tibble::rownames_to_column(table, "ProbeId")
head(table)  %>% kable(caption = "Differential expression table")
```

Table 7.3: Differential expression table

ProbeId	logFC	AveExpr	t	P.Value	adj.P.Val	B
TC05000550.hg.1	-2.664223	7.489702	-33.70718	0e+00	0.0008736	8.382083
TC05002579.hg.1	-2.659631	8.554089	-32.97093	0e+00	0.0008736	8.324418
TC06000135.hg.1	2.414019	8.405235	29.93135	0e+00	0.0010717	8.052560
TC04000478.hg.1	-2.096694	5.713626	-27.82441	1e-07	0.0012729	7.826071
TC07000810.hg.1	-2.232153	7.622238	-26.22526	1e-07	0.0013466	7.628687
TC15002249.hg.1	-1.929155	6.206933	-25.85557	1e-07	0.0013466	7.579506

For annotation we will use `select` function from AnnotationDbi library.

```
library(AnnotationDbi)
```

Affymetrix hta20 type of microarray was used for this experiment. We need annotation data

package for this platform.

```
library(hta20transcriptcluster.db)
```

We can use `keytypes` function to see what kind information we can call.

```
keytypes(hta20transcriptcluster.db)
```

```
##  [1] "ACCNUM"       "ALIAS"        "ENSEMBL"      "ENSEMBLPROT"
##  [5] "ENSEMBLTRANS" "ENTREZID"     "ENZYME"       "EVIDENCE"
##  [9] "EVIDENCEALL"  "GENENAME"     "GENETYPE"     "GO"
## [13] "GOALL"        "IPI"          "MAP"          "OMIM"
## [17] "ONTOLOGY"     "ONTOLOGYALL"  "PATH"         "PFAM"
## [21] "PMID"         "PROBEID"      "PROSITE"      "REFSEQ"
## [25] "SYMBOL"       "UCSCKG"       "UNIPROT"
```

Since all necessary data loaded now we will do annotation and store in `annotations`

```
annotations <- AnnotationDbi::select(hta20transcriptcluster.db,
      keys = table$ProbeId, c("SYMBOL", "ENTREZID"), "PROBEID")
rows_to_show <- sample(1:nrow(annotations), 15, replace=FALSE) %>%
          sort()
annotations[rows_to_show,] %>%
          kable(caption = "Annotations of the Probe Ids")
```

Table 7.4: Annotations of the Probe Ids

	PROBEID	SYMBOL	ENTREZID
2999	TC16001237.hg.1	MTSS2	92154
4766	TC18000538.hg.1	ATP8B1	5205
9801	TC05001507.hg.1	WDR41	55255
15208	TC12000375.hg.1	NA	NA
18500	TC01005784.hg.1	NA	NA
21629	TC12000086.hg.1	NA	NA
23179	TC03002658.hg.1	SLC7A14-AS1	101928583
33844	TC12001421.hg.1	NA	NA
35068	TC05002595.hg.1	NA	NA
50257	TC01000181.hg.1	PDPN	10630
55506	TC08001863.hg.1	NA	NA
55886	TC02000911.hg.1	NA	NA
56238	TC19000507.hg.1	ZNF790-AS1	284408
57294	TC16000843.hg.1	NA	NA
57704	TC02005060.hg.1	CALM2	805

```
dim(annotations)
```

```
## [1] 72388     3
```

Now we'll drop the na values.

```
dropped <- tidyr::drop_na(annotations)
dropped[sample(1:nrow(dropped), 15,replace=FALSE),] %>% kable()
```

	PROBEID	SYMBOL	ENTREZID
7019	TC01001842.hg.1	DEGS1	8560
29086	TC01003659.hg.1	KCNT2	343450
13295	TC0X000390.hg.1	LINC00891	441501
377	TC05001723.hg.1	PPIC	5480
653	TC0X000120.hg.1	POLA1	5422
19210	TC14000971.hg.1	TGM1	7051
22759	TC01001106.hg.1	NBPF12	149013
3380	TC12001361.hg.1	SINHCAF	58516
6332	TC06002160.hg.1	MIR3145	100423001
15689	TC12001835.hg.1	USP44	84101
16892	TC10001713.hg.1	ATE1	11101
10259	TC05000231.hg.1	GZMA	3001
179	TC10001341.hg.1	HERC4	26091
16404	TC08000003.hg.1	FBXO25	26260
15244	TC14000706.hg.1	SNORD114-26	767606

So we'll drop *NA* values which are transcripts that hybridized with probes that do not have an annotation. We will join annotation data and expression data using dplyr function `inner_join`.

```
library(dplyr)
joined <- inner_join(dropped, table, by = c("PROBEID" = "ProbeId"))
joined[sample(1:nrow(joined), 15,replace=FALSE),] %>% kable()
```

SYMBOL	ENTREZID	logFC	AveExp	t	P.Value	adj.P.Val	B
YME1L1	10730	-0.5741	9.696	-7.1021	0.0003	0.0181	0.8693
MIR4657	100616393	0.4554	6.162	2.9779	0.0231	0.2004	-3.7338
TMEM242	729515	-0.1462	5.872	-1.9533	0.0960	0.4414	-5.1867
TMEM70	54968	0.0422	7.480	0.5652	0.5913	0.9198	-6.7150
DBH	1621	0.0229	6.294	0.2788	0.7892	0.9712	-6.8493
CA4	762	-0.0181	5.374	-0.2205	0.8323	0.9792	-6.8658
RCOR1	23186	0.1627	7.792	1.9115	0.1018	0.4543	-5.2446
CDC25A	993	0.6956	6.747	9.2301	0.0000	0.0096	2.4388
LEKR1	389170	0.1364	4.258	1.6548	0.1463	0.5509	-5.5919
LINC01007	100506527	0.0421	5.033	0.4918	0.6394	0.9366	-6.7576
DLGAP1-AS1	649446	0.3168	6.894	2.9310	0.0246	0.2074	-3.7997
MEST	4232	0.0623	7.428	0.7739	0.4668	0.8628	-6.5658
CYP21A2	1589	-0.0152	5.917	-0.1805	0.8623	0.9828	-6.8750
CRYAA2	102724652	-0.0555	6.779	-0.6223	0.5554	0.9040	-6.6782

```
sum(duplicated(joined$ENTREZID))
## [1] 7402
```

There are multiple probes for the same transcript, so we have multiple entry (observation) for the same gene and we'll remove them.

```
subs <- subset(joined, !duplicated(joined$ENTREZID))
sum(duplicated(subs$ENTREZID))
## [1] 0
```