TRABAJO FIN DE MÁSTER

Máster Universitario Oficial en Ciencia de Datos e Ingeniería de Computadores

Epidemiología y detección de biomarcadores en cáncer

Autor:
Daniel Redondo Sánchez

Tutores:
Ignacio Rojas
Luis Javier Herrera
Daniel Castillo

Granada, septiembre de 2020

UNIVERSIDAD DE GRANADA

0. Índice general

Abstract

Abstract en inglés

Resumen

Abstract en español

1. Introducción

1.1. Objetivos del trabajo

En el presente Trabajo Fin de Máster se analiza la epidemiología de los cánceres de hígado y colon-recto y se detectan genes que permiten identificar tumores.

- En el capítulo 1,
- En el capítulo 2,
- En el capítulo 3,
- En el capítulo 4,
- En el capítulo 5,
- En el capítulo 6,

1.2. Cáncer

El cáncer es una enfermedad en la que se produce una división incontrolada de las células [1]. Aunque generalmente se habla del cáncer como una única enfermedad se trata en realidad de un conjunto de enfermedades, existiendo más de 100 tipos distintos de cáncer [2].

El cáncer es una enfermedad genética, esto es, causada por cambios en los genes que controlan las funciones celulares [2]. En general, el proceso de creación del cáncer es complejo y multifactorial: a menudo el causante no es un solo elemento, sino la combinación e interacción de distintos factores ambientales y genéticos [3].

Los factores causantes del cáncer se pueden clasificar principalmente en tres categorías:

1. Factores no modificables. Son elementos que no se pueden cambiar, como la edad o la herencia genética [4, 5].

2. Factores modificables o prevenibles, como el tabaco, el alcohol, la dieta o la exposición a distintos carcinógenos [6].

3. Otros factores. Algunas circunstancias no se corresponden a ninguna de las categorías anteriores ya que algunos de sus aspectos no se pueden cambiar. Es el caso de factores socioeconómicos (como cobertura sanitaria en el lugar de residencia o privación económica) y factores reproductivos u hormonales (como toma de anticonceptivos, lactancia materna o terapia hormonal sustitutiva en mujeres menopáusicas) [5].

A continuación se introducen dos tipos de cáncer con los que se trabajará más adelante: el cáncer de hígado y el cáncer de colon-recto.

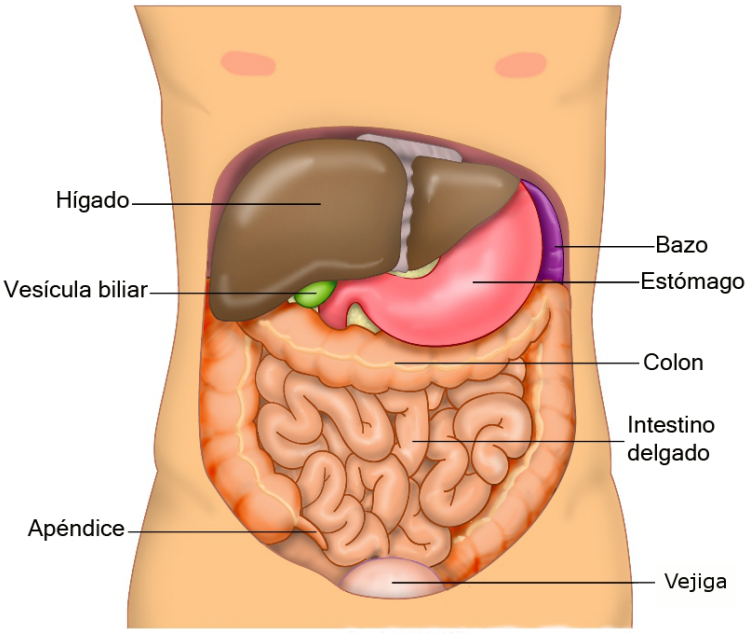
1.2.1. Cáncer de hígado

El cáncer de hígado se corresponde con el código C22 de la Clasificación Internacional de Enfermedades, Décima Revisión, integrando las neoplasias malignas de hígado y vías biliares intrahepáticas [7, 8].

Anatomía y funciones del hígado

El hígado es el órgano interno más grande y pesado del cuerpo humano, está situado en el cuadrante superior derecho del abdomen, debajo de las costillas, y está compuesto principalmente por dos lóbulos [9].

Figura 1. Anatomía del abdomen humano. Ilustración de Ties van Brussel.



Las funciones del hígado son múltiples y diversas. Las principales son procesar, particionar y metabolizar macronutrientes, regular el volumen de sangre, apoyar al sistema inmune, eliminar sustancias químicas como el alcohol y otras drogas y producir bilis para absorber grasas [10]. Es un órgano imprescindible para la vida.

Factores de riesgo

Uno de los factores de riesgo más comunes del cáncer de hígado es la presencia de cirrosis, o sustitución de células sanas de hígado por tejido cicatrizado. La cirrosis puede producirse por varias causas, siendo las más habituales el consumo excesivo de alcohol y la infección con el virus de la hepatitis B o C [11]. Otros factores de riesgo son el tabaco, la obesidad, padecer diabetes tipo II y consumir esteroides anabólicos [11, 12].

La prevención del cáncer de hígado se basa en reducir la exposición a factores de riesgo como el tabaco y el alcohol, y en vacunarse contra la hepatitis B [11].

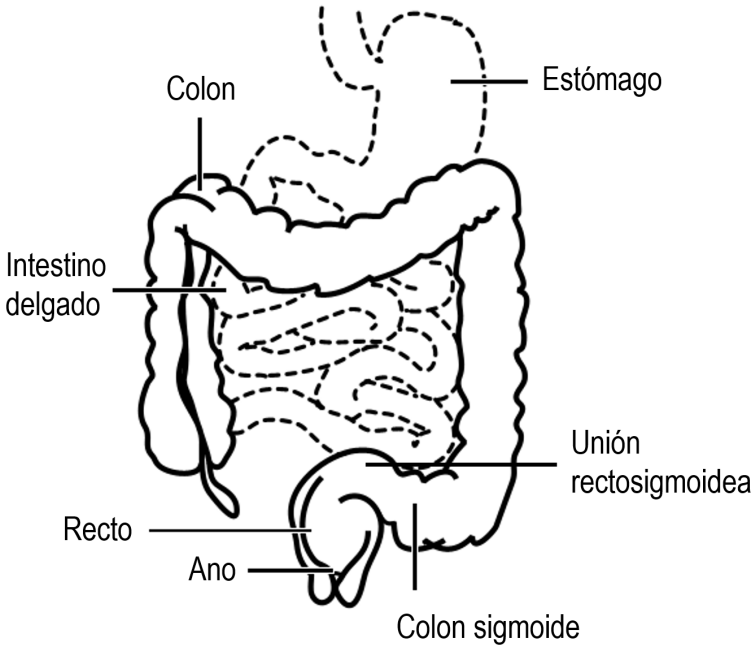
1.2.2. Cáncer de colon-recto

Las neoplasias malignas de colon, recto, unión rectosigmoidea, ano y canal anal (códigos C18-C21 según la Clasificación Internacional de Enfermedades, Décima Revisión [7, 8]) a menudo se estudian agrupadas por tener características muy similares.

Anatomía y funciones del colon-recto

El colon tiene 3 funciones principales: absorción de agua y electrolitos, producción y absorción de vitaminas y movimiento de heces hacia el recto para su eliminación por el ano [13].

Figura 2. Anatomía del intestino humano. Ilustración de Ties van Brussel.



Factores de riesgo

Entre los factores de riesgo del cáncer de colon-recto se puede distinguir entre factores modificables y no modificables.

Entre los factores de riesgo que son modificables destacan el sobrepeso, la inactividad física, las dietas con alto consumo de carnes rojas o procesadas, y el consumo

de tabaco y alcohol [14].

Una edad superior a 50 años, padecer diabetes tipo 2 y tener antecedentes personales o familiares de cáncer de colon-recto, pólipos o enfermedad intestinal inflamatoria, como colitis ulcerosa y enfermedad de Crohn, son algunos de los factores de riesgo no modificables [14]. También existen algunos síndromes hereditarios como el síndrome de Lynch que aumentan las posibilidades de padecer cáncer de colon-recto [15].

Para intentar prevenir el cáncer de colon-recto se deben cambiar aquellos factores que son modificables: realizar ejercicio, mantener una dieta saludable y evitar el consumo de tabaco y alcohol. Además, en los últimos años se están implementando programas de cribado de cáncer de colon-recto para detectar pólipos o diagnosticar el cáncer en etapas iniciales mediante análisis como pruebas de sangre oculta en heces o colonoscopias [16].

1.3. Ciencias -ómicas

Se presenta a continuación una corta introducción a las ciencias -ómicas, con el objetivo de comprender los conceptos que se utilizarán más adelante y ubicarlos dentro de su contexto biológico.

1.3.1. Algunas definiciones básicas

- Los seres vivos están hechos de células. En el núcleo de cada célula se encuentran los cromosomas, estructuras que almacenan el material genético del individuo.

- Los cromosomas están formados por ácido desoxirribonucleico (DNA, por sus siglas en inglés), una molécula que codifica las instrucciones genéticas para la vida.

- Un gen es la región del DNA que codifica una proteína. Las proteínas son cadenas de aminoácidos unidos por enlaces peptídicos (enlaces entre el grupo amino y carboxilo).

- El genoma es la secuencia de nucleótidos que forman el ADN de un individuo.
- El ácido ribonucleico (RNA, por sus siglas en inglés) es un ácido nucleico formado por ribonucleótidos.

1.3.2. Genómica

La genómica es la ciencia que estudia la composición, estructura y función de los genomas. Se dedica por tanto a estudiar cromosomas, mutaciones y variaciones tanto de nucleótidos concretos como de regiones del genoma.

El análisis GWAS (Genome-wide association study) es un ejemplo de análisis genómico.

1.3.3. Transcriptómica

La transcriptómica estudia el transcriptoma, esto es, el conjunto de RNA presente en una célula). El transcriptoma indica el nivel de expresión de genes en un determinado momento.

Los análisis de RNA-Seq y microRNA se enmarcan en el ámbito de la transcriptómica.

1.3.4. Otras ciencias -ómicas

La proteómica es la ciencia que estudia y caracteriza el proteoma (imagen dinámica de todas las proteínas expresadas).

Metabolómica.

1.4. RNA-Seq

Leer [17–19].

Este trabajo se enmarca dentro de la transcriptómica, y está basado en datos obtenidos mediante RNA-Seq, técnica en la que se cuentan distintas lecturas de

cada gen para ver si están sobreexpresados o infraexpresados, para finalmente comparar esas expresiones con una referencia (por ejemplo, enfermos contra sanos). Como sólo se realizan cuentas de la expresión de los genes, el RNA-Seq de un individuo no permite su identificación, por lo que los datos a menudo son accesibles de manera abierta.

2. Epidemiología del cáncer

La Epidemiología se ha definido tradicionalmente como la ciencia que estudia la distribución y los determinantes de la enfermedad en los seres humanos [20]. En una definición más moderna, no limitada exclusivamente a la enfermedad, la Epidemiología se define como el estudio de la aparición y distribución de los estados o acontecimientos relacionados con la salud en poblaciones específicas, incluyendo el estudio de los determinantes de estos estados, y la aplicación de este conocimiento al control de los problemas de la salud [21].

2.1. Indicadores epidemiológicos

Para medir en la población el impacto del cáncer se utilizan principalmente cuatro indicadores:

- **Incidencia** (casos nuevos). Mide el riesgo de presentar cáncer.
- **Mortalidad** (defunciones). Mide el riesgo de morir por cáncer.
- **Supervivencia** (porcentaje de casos vivos). Mide la historia natural del cáncer y efectividad del tratamiento.
- **Prevalencia** (casos nuevos y antiguos, vivos). Mide la carga asistencial de la enfermedad.

Además, se puede examinar la evolución de cada indicador a lo largo del tiempo, hablando así de tendencias de la incidencia, de la mortalidad, de la supervivencia o de la prevalencia.

(Ver si se incluyen tendencias, sería conveniente al menos para incidencia y/o mortalidad)

2.2. Fuentes de información

A nivel mundial, las estadísticas de cáncer las proporciona el *Global Cancer Observatory* (GCO), una plataforma web de la *International Agency for Research on Cancer*, de la Organización Mundial de la Salud [22,23]. El organismo equivalente al GCO a nivel europeo es el *European Cancer Information System* (ECIS), de reciente creación y apoyado por la Comisión Europea [24,25].

Aunque estos organismos proporcionan estadísticas sobre cáncer en España, también existen fuentes a nivel nacional que cuentan con datos más actualizados y con distinta metodología. La Red Española de Registros de Cáncer (REDECAN) publica periódicamente datos sobre incidencia y supervivencia de cáncer en España [26,27], mientras que las estadísticas de mortalidad por cáncer se pueden calcular a partir de las defunciones que publica el Ministerio de Sanidad, Consumo y Bienestar Social del Gobierno de España [28] y la población que proporciona el Instituto Nacional de Estadística [29].

SUPERVIVENCIA: CONCORD y/o EUROCARE.

2.3. Incidencia de cáncer

2.3.1. Metodología

Para medir de manera precisa la incidencia de cáncer en una población es necesaria la existencia de un Registro de Cáncer Poblacional. Estas entidades se dedican a registrar exhaustivamente todos los casos de cáncer diagnosticados en un área geográfica, y sus datos son muy útiles para todo tipo de estudios epidemiológicos. Algunos de estos Registros cubren la población de todo un país (por ejemplo, Canadá) mientras que otros cubren regiones concretas (por ejemplo, la provincia de Granada). Desgraciadamente, muchas áreas geográficas no están cubiertas por un Registro de Cáncer Poblacional. Es el caso de España, en el que sólo el 27 % de la población está cubierta por un Registro de Cáncer Poblacional [30]. Para conocer de manera estimada la incidencia de cáncer en territorios sin Registro de Cáncer Poblacional o proyectar la incidencia a años posteriores se utilizan diversos métodos matemáticos y estadísticos [22–26,30].

Con respecto a las medidas usadas para reportar la incidencia, la más sencilla y

fácil de interpretar es el número nuevo de casos de cáncer, enmarcado siempre en un periodo concreto de tiempo y un área geográfica. A partir del número de casos se puede calcular la tasa bruta (TB), un indicador que tiene en cuenta el tamaño de la población y que se suele calcular por 100.000 habitantes [31].

$$\text{TB} = 100.000 \cdot \frac{\text{Número de casos nuevos}}{\text{Personas-año a riesgo}}$$

Para permitir comparaciones entre distintas poblaciones, o la misma población en momentos distintos, es necesario tener en cuenta la estructura de edad de la población. Para responder a esta motivación se define la tasa estandarizada por edad (ASR por sus siglas en inglés, *Age-Standardised Rate*) como aquella tasa que habría en la población de estudio si tuviese exactamente la misma estructura de edad que una población estándar predefinida [31]. La definición de la tasa estandarizada por edad para 18 grupos de edad quinquenales (0-4 años, 5-9 años, ..., 80-84 años, 85 años y más) es la siguiente:

$$\text{ASR} = \sum_{i=1}^{18} \omega_i \frac{N_i}{P_i}$$

donde N_i y P_i son respectivamente el número de casos incidentes y la población en el i-ésimo grupo de edad, y ω_i es el peso que toma la población de referencia en el grupo i-ésimo, con $\sum_{i=1}^{18} \omega_i = 100.000$. Los valores de ω_i están predefinidos en base a poblaciones estándar, siendo las más utilizadas en nuestro contexto las siguientes:

- Población mundial. Propuesta por primera vez en 1960 [32] y modificada más tarde en 1966 [33], permite realizar comparaciones a nivel mundial.

- Antigua población estándar europea. Propuesta en 1976 [34] basándose en la estructura de edad de varias poblaciones escandinavas, permite comparaciones entre zonas europeas.

- Nueva población estándar europea. En el año 2013, la Oficina Europea de Estadística (EUROSTAT) realiza una revisión de la población estándar europea con el objetivo de que la población refleje fielmente el envejecimiento existente en la población europea [35]. Debido a su novedad, el uso de esta

población aún no está ampliamente extendido en los organismos internacionales [25] y en ocasiones se reportan las dos tasas estandarizadas por las poblaciones estándar antigua y nueva [24].

En la Tabla 1 se muestran los pesos para cada una de las poblaciones de referencia mencionadas anteriormente.

Tabla 1. Pesos de las poblaciones estándar para el cálculo de tasas estandarizadas por edad.

Grupo de edad	Población estándar mundial	Población estándar europea 1976	Población estándar europea 2013
0-4 años	12.000	8.000	5.000
5-9 años	10.000	7.000	5.500
10-14 años	9.000	7.000	5.500
15-19 años	9.000	7.000	5.500
20-24 años	8.000	7.000	6.000
25-29 años	8.000	7.000	6.000
30-34 años	6.000	7.000	6.500
35-39 años	6.000	7.000	7.000
40-44 años	6.000	7.000	7.000
45-49 años	6.000	7.000	7.000
50-54 años	5.000	7.000	7.000
55-59 años	4.000	6.000	6.500
60-64 años	4.000	5.000	6.000
65-69 años	3.000	4.000	5.500
70-74 años	2.000	3.000	5.000
75-79 años	1.000	2.000	4.000
80-84 años	500	1.000	2.500
$\geq$85 años	500	1.000	2.500

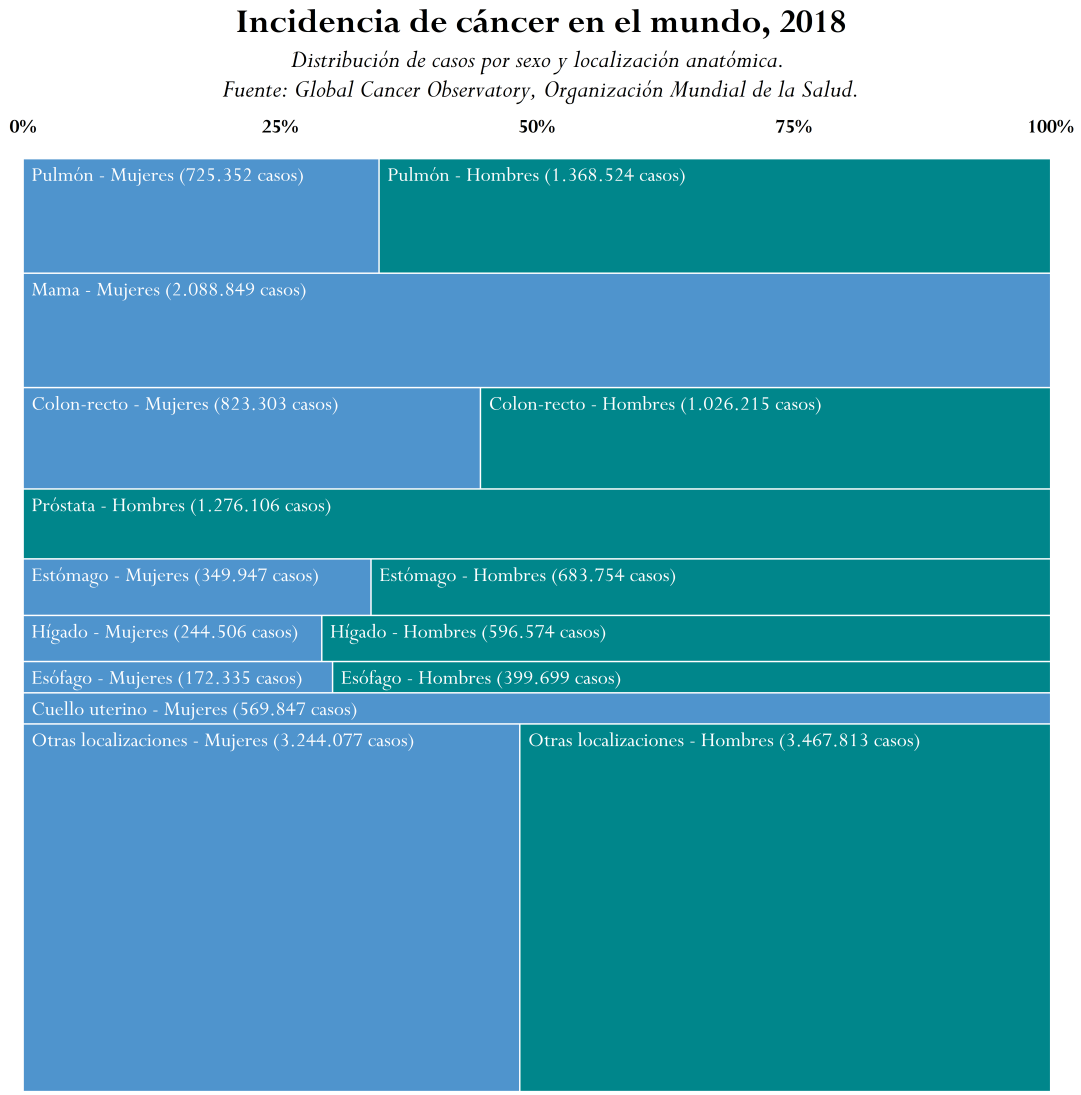
Para utilizar notación internacional, la tasa estandarizada por la población mundial se notará ASR-W (de *World standard population*), la tasa estandarizada por la población europea de 1976 se notará ASR-oE (*old European standard population*) y la de 2013 se notará ASR-nE (*new European standard population*).

2.3.2. Incidencia del total del cáncer excepto piel no melanoma

El cáncer de piel no melanoma se suele excluir al reportar datos de incidencia del total del cáncer, debido a que es muy frecuente y cuenta con buen pronóstico, por lo que no se suele registrar en los Registros de Cáncer Poblacionales [36, 37].

Para dar una perspectiva global del cáncer y sus diferentes tipos, en la Figura X se muestran las localizaciones anatómicas más frecuentes de cáncer en el mundo, así como su distribución por sexos.

Figura X. Gráfico de mosaico con la incidencia estimada de cáncer excepto piel no melanoma en el mundo para el año 2018. Ocho localizaciones anatómicas más frecuentes en ambos sexos. Fuente: *Global Cancer Observatory*, Organización Mundial de la Salud [23].



El cáncer de pulmón es el más frecuente en todo el mundo, seguido por los cánce-

res de mama, colon-recto, próstata, estómago, hígado, esófago y cuello uterino. En la mayoría de las localizaciones anatómicas el cáncer es más frecuente en hombres que en mujeres.

Tabla X. Incidencia del total del cáncer excepto piel no melanoma en 2018, por sexo y población. Número de casos nuevos (N), tasa bruta (TB), tasa estandarizada por la población mundial (ASR-W), tasa estandarizada por la antigua población europea (ASR-oE) y tasa estandarizada por la nueva población europea (ASR-nE).

Sexo	Población	Fuente	N	TB	ASR-W	ASR-oE	ASR-nE
Hombres	Mundo	GCO [23]	8.818.685	229,0	204,7		
	Europa	ECIS [24]	2.059.673	572,9	302,7	436,0	651,7
	España	ECIS [24]	142.353	625,6	309,7	444,7	658,6
Mujeres	Mundo	GCO [23]	8.218.216	217,3	175,6		
	Europa	ECIS [24]	1.851.644	481,8	242,7	332,6	451,2
	España	ECIS [24]	106.647	451,1	218,4	298,5	401,7
Ambos sexos	Mundo	GCO [23]	17.036.901	223,2	187,8		
	Europa	ECIS [24]	3.911.317	525,8	266,7	374,3	531,9
	España	ECIS [24]	249.000	536,7	259,4	363,8	515,3

A nivel nacional, REDECAN ha publicado estimaciones de la incidencia en España más recientes a las mostradas en la Tabla X, correspondientes al año 2020, [26]. En este análisis se estima el número de casos de cáncer excepto piel no melanoma en 277.394 casos (57,8 % en hombres), con una tasa bruta de 588,0 por 100.000 habitantes y tasas estandarizadas de 280,3 (ASR-W), 399,4 (ASR-oE) y 579,8 (ASR-nE).

2.3.3. Incidencia de cáncer de hígado

El cáncer de hígado es el sexto cáncer más frecuente del mundo (Figura X), con más de 840.000 casos nuevos anuales en todo el mundo, 82.000 de ellos en Europa y 6.600 en España. Es un cáncer más frecuente en hombres que en mujeres: por cada caso en mujeres hay 2,4 casos en hombres. En ambos sexos, la ASR-W de España (6,5) es mayor que la de Europa (5,1) aunque mucho menor que la del mundo (9,3).

Tabla X. Incidencia de cáncer de hígado en 2018, por sexo y población. Número de casos nuevos (N), tasa bruta (TB), tasa estandarizada por la población mundial (ASR-W), tasa estandarizada por la antigua población europea (ASR-oE) y tasa estandarizada por la nueva población europea (ASR-nE).

Sexo	Población	Fuente	N	TB	ASR-W	ASR-oE	ASR-nE
Hombres	Mundo	GCO [23]	596.574	15,5	13,9		
	Europa	ECIS [24]	55.825	15,5	8,0	11,7	17,7
	España	ECIS [24]	4.976	21,9	10,9	15,7	22,5
Mujeres	Mundo	GCO [23]	244.506	6,5	4,9		
	Europa	ECIS [24]	26.641	6,9	2,7	4,0	6,3
	España	ECIS [24]	1.654	7,0	2,4	3,6	6,0
Ambos sexos	Mundo	GCO [23]	841.080	11,0	9,3		
	Europa	ECIS [24]	82.466	11,1	5,1	7,4	11,3
	España	ECIS [24]	6.630	14,3	6,5	9,3	13,6

En las estimaciones publicadas por REDECAN para 2020 se estiman en España 6.595 casos de cáncer de hígado (75,4 % en hombres), con una tasa bruta de 14,0 y tasas estandarizadas de 6,5 (ASR-W), 9,4 (ASR-oE) y 13,9 (ASR-nE) [26].

2.3.4. Incidencia de cáncer de colon-recto

El cáncer de colon-recto es el tercer cáncer más frecuente del mundo (Figura X), con más de 1.800.000 casos nuevos anuales en todo el mundo. En Europa y España es el cáncer más frecuente en ambos sexos con 510.000 casos anuales en Europa y 37.000 en España. Es un cáncer ligeramente más frecuente en hombres que en mujeres: por cada caso en mujeres hay 1,2 casos en hombres. En ambos sexos, la ASR-W de España (81,1) es mayor que la de Europa (68,8) y la del mundo (24,2), lo que puede deberse a una mayor exposición a factores de riesgo como el tipo de dieta o la ausencia de un programa de cribado a nivel nacional [38].

Tabla X. Incidencia de cáncer de colon-recto en 2018, por sexo y población. Número de casos nuevos (N), tasa bruta (TB), tasa estandarizada por la población mundial (ASR-W), tasa estandarizada por la antigua población europea (ASR-oE) y tasa estandarizada por la nueva población europea (ASR-nE).

Sexo	Población	Fuente	N	TB	ASR-W	ASR-oE	ASR-nE
Hombres	Mundo	GCO [23]	1.026.215	26,6	23,6		
	Europa	ECIS [24]	275.519	76,6	38,1	56,8	88,9
	España	ECIS [24]	23.013	101,1	45,8	68,5	107,2
Mujeres	Mundo	GCO [23]	823.303	21,8	16,3		
	Europa	ECIS [24]	236.101	61,4	25,2	37,0	56,3
	España	ECIS [24]	14.642	61,9	23,6	34,9	53,5
Ambos sexos	Mundo	GCO [23]	1.849.518	24,2	19,7		
	Europa	ECIS [24]	511.620	68,8	30,8	45,6	70,0
	España	ECIS [24]	37.655	81,1	33,9	50,4	77,5

En las estimaciones publicadas por REDECAN para 2020 se estiman en España 44.231 casos de cáncer de colon-recto (58,9 % en hombres), con una tasa bruta de 93,8 y tasas estandarizadas de 40,0 (ASR-W), 59,5 (ASR-oE) y 91,9 (ASR-nE) [26].

2.4. Mortalidad por cáncer

2.4.1. Metodología

Los indicadores para medir la mortalidad por cáncer son los mismos que para la incidencia, cambiando número de casos por defunciones por cáncer. Es importante destacar que la mortalidad por cáncer es por definición aquella mortalidad que es causada directamente por el cáncer. En este sentido, una persona diagnosticada de cáncer que falleciese por otras causas no puede ser considerada como fallecida por cáncer, sino fallecida con cáncer.

Aunque ECIS [24] reporta mortalidad por cáncer en España para el año 2018, la mortalidad que se presenta a nivel nacional es la que proporciona el Ministerio de Sanidad, Consumo y Bienestar Social [28], al tratarse de datos observados basados en certificados médicos de defunción y no estimaciones, por lo que se consideran datos más fiables.

2.4.2. Mortalidad del total del cáncer excepto piel no melanoma

texto

2.4.3. Mortalidad de cáncer de hígado

texto

2.4.4. Mortalidad de cáncer de colon-recto

texto

2.5. Supervivencia de cáncer

Supervivencia se calcula principalmente a partir de inc, mort y tablas de vida población general

2.5.1. Supervivencia del total del cáncer excepto piel no melanoma

texto

2.5.2. Supervivencia de cáncer de hígado

texto

2.5.3. Supervivencia de cáncer de colon-recto

texto

2.6. Prevalencia de cáncer

texto

2.6.1. Prevalencia del total del cáncer excepto piel no melanoma

texto

2.6.2. Prevalencia de cáncer de hígado

texto

2.6.3. Prevalencia de cáncer de colon-recto

texto

3. *Machine learning* aplicado a RNA-Seq

3.1. Selección de características

Ver apuntes asignatura bioinformática [39].

Leer y citar [40].

La selección de genes se ha utilizado ampliamente para intentar predecir el diagnóstico del cáncer, basándose en microarrays [41,42], aunque el uso de microarrays está siendo reemplazado por el uso de RNA-Seq.

3.1.1. Mínima redundancia, máxima relevancia (mRMR)

3.1.2. *Random Forest* (RF)

3.1.3. Asociación de enfermedades (DA)

3.2. Algoritmos de clasificación

3.2.1. Máquinas de soporte vectorial (SVM)

3.2.2. k-vecinos más cercanos (kNN)

4. Detección de biomarcadores en cáncer de hígado y colon-recto

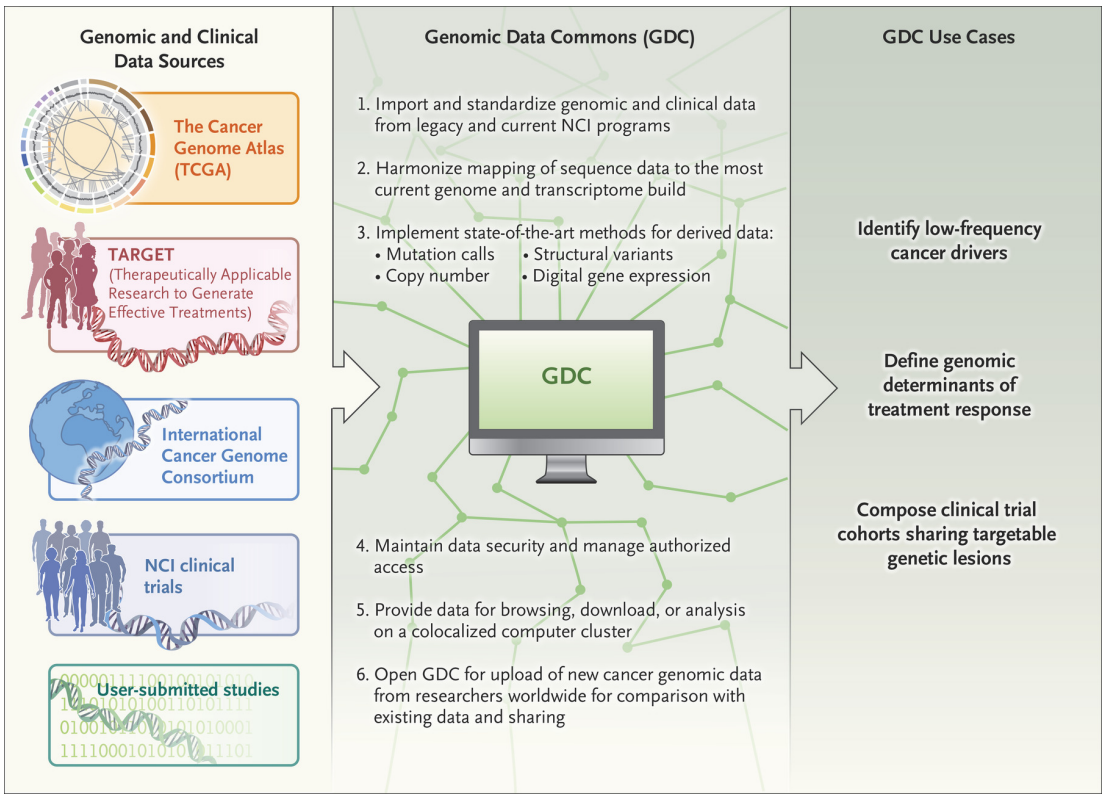
4.1. Introducción

4.2. Metodología

4.2.1. Fuente de datos

La fuente de los datos es GDC (Genomic Data Commons) Portal, una plataforma web sobre cáncer del Instituto Nacional del Cáncer de Estados Unidos (*National Cancer Institute*) [43, 44]. GDC Portal fue desarrollado por el Instituto Nacional del Cáncer de Estados Unidos, la Universidad de Chicago, el Instituto de Ontario para la Investigación del Cáncer y la empresa *Leidos Biomedical Research*, y su principal fortaleza reside en la integración y armonización de diversas fuentes heterogéneas, creando así un sistema de información amplio y robusto [45].

Figura XX. Diagrama de funcionalidad y utilidad de GDC. Extraído de Grossman et al. [45].



GDC Portal, a día 22 de Junio, contenía información sobre unos 84.000 casos, 23.000 genes y más de 3 millones de mutaciones de genes [43]. Los datos de los que dispone son muy variados, y se pueden distinguir en tres grandes categorías:

- Información clínica, como la edad del sujeto, su sexo o el estadio del cáncer del que ha sido diagnosticado.

- Información genética y transcriptómica proveniente de diversos proyectos de investigación.

- Imágenes de tejidos tumorales y sanos.

Algunos de estos datos son abiertos, mientras que para otros es necesario solicitar acceso.

4.2.2. Análisis

Para el análisis se ha utilizado el software estadístico R [46] y el paquete `KnowSeq` (v.1.2.0), librería que ha sido desarrollada por los tutores del presente trabajo, y en la que el autor ha contribuido con algunas nuevas funciones y pequeñas modificaciones [47]. El paquete está además disponible en Bioconductor, una relevante plataforma de código abierto en R para el análisis de datos en genómica y transcriptómica [48].

Otros paquetes de R con versiones y referencias! Por relevancia citar paquetes como edgeR y limma

4.3. Resultados: cáncer de hígado

4.3.1. Características clínicas de los pacientes

4.3.2. Detección de biomarcadores

Validación cruzada

Validación en test

4.4. Resultados: cáncer de colon-recto

4.4.1. Características clínicas de los pacientes

4.4.2. Detección de biomarcadores

Validación cruzada

Validación en test

4.5. Conclusiones

Interpretar resultados con cautela: ver pág. 65 de [19] (referencias 77-79).

5. biomaRcadores: una aplicación web interactiva para detección de biomarcadores

5.1. Desarrollo de la aplicación

`Shiny`, versión, documentación breve sobre Shiny, ... y código en anexo 2

5.2. Utilidades de la aplicación

Capturas de pantalla con ejemplos. Quizá grabar vídeo mostrando la aplicación (subir GIF a README).

6. Conclusiones y líneas abiertas de trabajo

Bibliografía

[1] American Cancer Society. What is Cancer? Disponible en: https://www.cancer.org/cancer/cancer-basics/what-is-cancer.html [Consultado 18/06/2020].

[2] National Cancer Institute. What is Cancer? Disponible en: https://www.cancer.gov/about-cancer/understanding/what-is-cancer [Consultado 18/06/2020].

[3] Lucia Migliore and Fabio Coppedè. Genetic and environmental factors in cancer pathogenesis. *Mutation Research*, 512:135–153, 2012.

[4] World Health Organization. *World Cancer Report 2014*. 2014.

[5] World Health Organization. *World Cancer Report. Cancer research for cancer prevention.* 2020.

[6] V. J. Cogliano, R. Baan, K. Straif, Y. Grosse, B. Lauby-Secretan, F. El Ghissassi, V. Bouvard, L. Benbrahim-Tallaa, N. Guha, C. Freeman, L. Galichet, and C. P. Wild. Preventable Exposures Associated With Human Cancers. *JNCI Journal of the National Cancer Institute*, 103(24):1827–1839, 2011.

[7] World Health Organization (WHO). *ICD-10: International Statistical Classification of diseases and related health problems: 10th revision.* 1990.

[8] Ministerio de Sanidad Consumo y Bienestar Social. Edición electrónica de la CIE-10-ES Diagnósticos. Disponible en: https://eciemaps.mscbs.gob.es/ecieMaps/browser/index_10_mc.html [Consultado 21/06/2020].

[9] Sherif R. Z. Abdel-Misih and Mark Bloomston. Liver Anatomy. *Surgical Clinics of North America*, 90(4):643–653, 2010.

[10] Elijah Trefts, Maureen Gannon, and David H. Wasserman. The liver. *Current Biology*, 27(21):R1147–R1151, 2017.

[11] American Cancer Society. Liver Cancer Risk Factors. Disponible en: https://www.cancer.org/cancer/liver-cancer/causes-risks-prevention/risk-factors.html [Consultado 18/06/2020], 2019.

[12] Jorge A. Marrero, Robert J. Fontana, Sherry Fu, Hari S. Conjeevaram, Grace L. Su, and Anna S. Lok. Alcohol, tobacco and obesity are synergistic risk factors for hepatocellular carcinoma. *Journal of Hepatology*, 42(2):218–224, 2005.

[13] Laura L. Azzouz and Sandeep Sharma. *Physiology, Large Intestine*. 2020.

[14] American Cancer Society. Colorectal Cancer Risk Factors. Disponible en: https://www.cancer.org/cancer/colon-rectal-cancer/causes-risks-prevention/risk-factors.html [Consultado 20/06/2020].

[15] Henry T. Lynch and Albert de la Chapelle. Hereditary Colorectal Cancer. *New England Journal of Medicine*, 348(10):919–932, 2003.

[16] B. Levin, D. A. Lieberman, B. McFarland, R. A. Smith, D. Brooks, K. S. Andrews, C. Dash, F. M. Giardiello, S. Glick, T. R. Levin, P. Pickhardt, D. K. Rex, A. Thorson, and S. J. Winawer. Screening and Surveillance for the Early Detection of Colorectal Cancer and Adenomatous Polyps, 2008: A Joint Guideline from the American Cancer Society, the US Multi-Society Task Force on Colorectal Cancer, and the American College of Radiology. *CA: A Cancer Journal for Clinicians*, 58(3):130–160, 2008.

[17] Rory Stark, Marta Grzelak, and James Hadfield. RNA sequencing: the teenage years. *Nature Reviews Genetics*, 20(11):631–656, 2019.

[18] Marcel C. Van Verk, Richard Hickman, Corné M.J. Pieterse, and Saskia C.M. Van Wees. RNA-Seq: revelation of the messengers. *Trends in Plant Science*, 18(4):175–179, apr 2013.

[19] Daniel Castillo Secilla. *Integration of heterogeneous gene expression sources in human cancer pathologies, employing high performance computing and machine learning techniques*. PhD thesis, 2020.

[20] B MacMahon and TF Pugh. *Epidemiology: Principles and Methods*. 1970.

[21] Miquel Porta, editor. *A Dictonary of Epidemiology*. Fifth edit edition, 2008.

[22] Freddie Bray, Jacques Ferlay, Isabelle Soerjomataram, Rebecca L. Siegel, Lindsey A. Torre, and Ahmedin Jemal. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians*, 68(6):394–424, 2018.

[23] International Agency for Research on Cancer and World Health Organization. Global Cancer Observatory, Cancer Today. Disponible en: https://gco.iarc.fr/today/home [Consultado 21/06/2020].

[24] European Comission. ECIS - European Cancer Information System. Disponible en: https://ecis.jrc.ec.europa.eu [Consultado 21/06/2020].

[25] J. Ferlay, M. Colombet, I. Soerjomataram, T. Dyba, G. Randi, M. Bettio, A. Gavin, O. Visser, and F. Bray. Cancer incidence and mortality patterns in Europe: Estimates for 40 countries and 25 major cancers in 2018. *European Journal of Cancer*, 103:356–387, 2018.

[26] Red Española de Registros de Cáncer (REDECAN). Estimaciones de la incidencia del cáncer en España, 2020. Disponible en: https://funca.cat/redecan/redecan.org/es/Informe_incidencia_REDECAN_2020.pdf [Consultado 18/06/2020].

[27] Marcela Guevara, Amaia Molinuevo, Diego Salmerón, Rafael Marcos-Gragera, María Dolores Chirlaque, José Ramón Quirós, Araceli Alemán, Dolores Rojas, Consol Sabater, Matilde Chico, Rosario Jiménez, Arantza López de Munain, Visitación de Castro, María José Sánchez, Josefina Perucha, Carmen Sánchez-Contador, Jaume Galceran, Eva Ardanaz, and Nerea Larrañaga. Supervivencia de Cáncer en España, 2002-2013. Disponible en: https://funca.cat/redecan/redecan.org/es/Informe_Supervivencia_REDECAN_2020.pdf [Consultado 26/06/2020], 2019.

[28] Ministerio de Sanidad Consumo y Bienestar Social. Estadísticas de defunciones según la causa de muerte. Disponible en: https://pestadistico.inteligenciadegestion.mscbs.es/ [Consultado 21/06/2020].

[29] Instituto Nacional de Estadística (INE). Estadísticas de cifras de población. Disponible en: http://ine.es/ [Consultado 21/06/2020].

[30] Daniel Redondo-Sánchez. Modelización Matemática de la Estimación de Incidencia de Cáncer. 2019.

[31] IARC. *Registros de Cáncer: Principios y Métodos*. 1995.

[32] Segi M. Cancer mortality for selected sites in 24 countries (1950–57). *Sendai, Japan: Department of Public Health, Tohoku University of Medicine.*, 1960.

[33] Waterhouse PAH Doll R, Payne P. Cancer incidence in five continents, Volume I. *Geneva: Union Internationale Contre le Cancer.*, 1966.

[34] JAH Waterhouse, CS Muir, P Correa, and J Powell. Cancer incidence in five continents, Volume III. *Lyon: IARC*, page 3:456, 1976.

[35] EUROSTAT. Revision of the European standard population: Report of the Eurostat's task force. Technical report, Luxembourg: European Union., 2013.

[36] Randy Gordon. Skin Cancer: An Overview of Epidemiology and Risk Factors. *Seminars in Oncology Nursing*, 29(3):160–169, aug 2013.

[37] Vishal Madan, John T Lear, and Rolf-Markus Szeimies. Non-melanoma skin cancer. *The Lancet*, 375(9715):673–685, feb 2010.

[38] Miroslav Zavoral, Stepan Suchanek, Filip Zavada, Ladislav Dusek, Jan Muzik, Bohumil Seifert, and Premysl Fric. Colorectal cancer screening in Europe. *World Journal of Gastroenterology*, 15(47):5907, 2009.

[39] Ignacio Rojas, Luis Javier Herrera Maldonado, and Daniel Castillo Secilla. Apuntes de la asignatura "Biología Computacional con Big Data omics e Ingeniería Biomédica". 2020.

[40] Eric P Xing, Michael I Jordan, and Richard M Karp. Feature Selection for High-Dimensional Genomic Microarray Data.

[41] Zne-Jung Lee. An integrated algorithm for gene selection and classification applied to microarray data of ovarian cancer. *Artificial Intelligence in Medicine*, 42(1):81–93, jan 2008.

[42] Rosalia Maglietta, Annarita D'Addabbo, Ada Piepoli, Francesco Perri, Sabino Liuni, Graziano Pesole, and Nicola Ancona. Selection of relevant genes in cancer diagnosis based on their prediction accuracy. *Artificial Intelligence in Medicine*, 40(1):29–44, may 2007.

[43] National Cancer Institute and National Institutes of Health. GDC Portal. Disponible en: https://portal.gdc.cancer.gov/ [Consultado 22/06/2020].

[44] National Cancer Institute. National Cancer Institute. Disponible en: https://www.cancer.gov [Consultado 22/06/2020].

[45] Robert L. Grossman, Allison P. Heath, Vincent Ferretti, Harold E. Varmus, Douglas R. Lowy, Warren A. Kibbe, and Louis M. Staudt. Toward a Shared Vision for Cancer Genomic Data. *New England Journal of Medicine*, 375(12):1109–1112, 2016.

[46] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2020.

[47] Daniel Castillo-Secilla, Juan Manuel Galvez, Francisco Carrillo-Perez, Marta Verona-Almeida, Francisco Manuel Ortuno, Luis Javier Herrera, and Ignacio Rojas. *KnowSeq: KnowSeq R/Bioc package: Beyond the traditional Transcriptomic pipeline*, 2020.

[48] Robert C Gentleman, Vincent J Carey, Douglas M Bates, Ben Bolstad, Marcel Dettling, Sandrine Dudoit, Byron Ellis, Laurent Gautier, Yongchao Ge, Jeff Gentry, Kurt Hornik, Torsten Hothorn, Wolfgang Huber, Stefano Iacus, Rafael Irizarry, Friedrich Leisch, Cheng Li, Martin Maechler, Anthony J Rossini, Gunther Sawitzki, Colin Smith, Gordon Smyth, Luke Tierney, Jean Y H Yang, and Jianhua Zhang. Bioconductor: open software development for computational biology and bioinformatics. *Genome biology*, 5(10):R80, 2004.

Anexo I: Código de análisis en R

Anexo I: Código de análisis en R

Funciones

Caja XX. Definición de `funcion`.

```
# Ejemplo de comentario
parametro <- 24000

texto <- "texto"
```

Anexo II: Código de aplicación web

Anexo II: Código de aplicación web

Funciones

Caja XX. Definición de `funcion`.

```
# Ejemplo de comentario
parametro <- 24000

texto <- "texto"
```