UNIVERSITAT DE BARCELONA

Anàlisi molecular del càncer colorectal mitjançant un enfocament computacional integratiu

David Cordero Romera

Anàlisi molecular del càncer colorectal mitjançant un enfocament computacional integratiu

Tesi presentada per

David Cordero Romera

Per obtenir el títol de doctor/a per la Universitat de Barcelona

Dirigida per:

Víctor Moreno Aguado

Programa de doctorat Medicina
Universitat de Barcelona

2015

A nuestra amiga Ana, *in memoriam*

Por ser como eras y para recordarte que nunca te olvidaremos

Agraïments

En primer lloc, voldria donar les gràcies a la família: el meu pare Luis, la meva mare Carmen i la meva germana Ester. A casa, anys enrere no es guanyaven la vida amb res molt proper al món de la ciència o la recerca, precisament. Malgrat això, els meus pares sempre em van animar perquè estudiés molt, fet que els agraeixo. A més, per casa sempre hi havia disponible alguna revista de divulgació d'interès general com *Muy interesante*, que jo podia llegir, en aquells temps en els quals encara no existia Internet. Això en moltes ocasions em va despertar la curiositat i inquietud per certes coses, potser estranyes per altres persones. Així que moltes gràcies a tots, per haver estat així. Al meu tiet Ignacio per inculcar en mi una vocació científica buscant sempre el perquè de les coses. Portant-me més d'una vegada al museu de la ciència i explicant-me cadascun d'aquells aparells que tenien allà. També explicant-me una demostració i fent-me comprendre el Teorema de Pitàgores, en comptes de simplement anotar-me la fórmula. Ajudant-me durant els llargs i calorosos estius per aprofitar i aprendre coses noves. Així que per tot això, moltíssimes gràcies. Per descomptat, moltes gràcies també per a tota la resta de la família, els meus avis que per sort han pogut veure tot això, els tiets, els meus cosins, la família que per circumstàncies de la vida sempre ha estat lluny, com els meus tiets d'Alemanya i molt especialment a la meva cosina Mónica amb la que sempre he pogut parlar i que sempre m'ha animat.

Als meus amics Dani, Sergio, David, Lourdes, Fran,... que a vegades no entenen ben bé el que estem fent, però que saben que és difícil i que requereix molt d'esforç. Tot i que ja ens fem grans i que ens veiem menys del que ens agradaria, moltes gràcies per ser-hi sempre que us he necessitat per a qualsevol cosa. També volia agrair al meu amic Raúl i a la seva dona Ana, a qui en el seu record està especialment dedicada aquesta tesi, la seva proximitat, la seva comprensió i la seva gran amistat per sempre.

Volia agrair als companys de feina més propers, tot el seu suport incondicional que sempre he rebut. Al cap i a la fi moltes vegades l'èxit o el fracàs de les coses es deuen a l'entorn i no únicament a un mateix. Com són molts aniré per ordre i espero no oblidar-me a ningú, però en tot cas, si em deixo a algú li demano perdó i que sàpiga que ha sigut sense voler. En primer lloc a la Sara i l'Esther, les antigues companyes del SERC (Servei d'Epidemiologia i Registre del Càncer) d'on vam néixer i que ara són molt més que companyes de feina. A la resta dels companys dels dinars, com l'Adrià o el Fran. Als informàtics, com el Xavi Junyent i molt especialment al Ferran, perquè sense la seva gran ajuda i el seu suport durant tant de temps és ben cert que jo mai podria haver acabat tot això. A l'Olga, gràcies per tota la teva ajuda que sempre rebem i per aquesta energia positiva inesgotable que desprens constantment. A la Gemma, gràcies pel taxi dels matins ja que a aquestes hores no és fàcil trobar transport. Bé, bromes a banda, gràcies per tota la resta de coses també, veïna. A

l'Esteve mòltes gràcies pels teus bons consells i per mantenir la porta oberta en tot moment, malgrat la feina que tens. Als companys del Pla Director d'Oncologia, que ara són al pis de sota i per això ens veiem menys: el Ramón, la Laura Esteban, el Xavi, el Jordi, la Pitus i, com no, l'estimadíssima Laura Pareja. També molts records pels que ja no hi són aquí a l'ICO, però que encara els recordo com si hi fossin. Començant per la Raquel Iniesta, l'Oscar Reina, el Toni Berenguer, el David Olivares i altres tants que no menciono per no estendre'm massa. Gràcies també a tota la resta dels companys que m'han ajudat a fer possible tot això. En especial als companys que són actualment o que han passat en algun moment per la Unitat de Biomarcadors i Susceptibilitat de l'Institut Català d'Oncologia.

Volia dedicar un paràgraf especial per aquells que hi són, però que són lluny. És a dir, per a la Marta i el Xavi. Sempre s'han esforçat molt en que la distància no sigui cap problema per tirar endavant tot el treball pendent i portar-lo a bon port. A la Marta la conec des de fa menys temps, però que sàpigues que estic encantat d'haver-te conegut. I per al Xavi un agraïment molt i molt especial per tota la ajuda que sempre m'ha ofert i per ensenyar-me tantes i tantes coses. Ell m'ha fet de guia durant tots aquests anys com si d'un director de tesi es tractés i aprofito aquí per dir-li que sempre comptarà amb el meu suport. Una abraçada molt gran per tots dos!

Al Víctor, el meu director, gràcies per haver estat aquí durant tots aquests anys, per haver-me deixat fer, i finalment, per ajudar-me a tancar aquesta difícil etapa.

Al Gabi, perquè tot i ser a dalt i molt atabalat, la porta del seu despatx sempre ha estat oberta per qualsevol cosa, i per això vull específicament donar-li les gràcies.

En cap cas voldria oblidar-me d'agrair a en Xavier Messeguer tota la seva tasca i la seva guia inicial, ja que gràcies a ell em vaig introduir al mon de la bioinformàtica quan només era informàtic. La curiositat i les ganes de fer alguna cosa útil per a tothom em va portar un bon dia a trucar a la seva porta. Des d'aquell moment no tan sols no m'he penedit ni un segon, sinó que crec que és el millor que vaig poder fer amb el meu projecte de final de carrera. Tan de bo altres docents tinguessin aquesta iniciativa i interès per la recerca, així que per tot això, moltíssimes gràcies Xavier!

Reservo l'últim lloc, però no per això menys important, als agraïments per a la meva família. És a dir, a la meva dona Carmen i a la meva filla Mónica. Ara formen una part tan gran de mi, que ja no entendria la vida sense elles. A la meva dona que me l'estimo molt, sento no haver estat amb tu tot el m'hagués agradat durant aquestes darreres setmanes. Recordo a la meva filla quan em deia *papa te vas a poner a trabajar con el ordenador otra vez*. Moltes gràcies per estar sempre aquí, per tenir tanta paciència, per recolzar-me en tot moment i per donar-me sempre el suport necessari. Us estimo!

Una abraçada enorme per tots i moltes gràcies!!

Prefaci

Durant aquests últims anys hem viscut el gran impacte de l'era post-genòmica, en la qual la seqüència del genoma humà ja és coneguda. En aquesta etapa, per exemple, s'ha pogut demostrar que el DNA no codificant, inicialment considerat com a no funcional, té un rol essencial en el funcionament de les cèl·lules humanes. Com a conseqüència d'aquest nou escenari, així com de la reducció del cost dels experiments a gran escala i la incorporació de la bioinformàtica com a eina essencial, la recerca en biologia molecular i salut pública ha modificat el seu enfocament clàssic per incorporar aquest canvi de paradigma. El càncer, una malaltia ja tractada per civilitzacions presents segles enrere, per primera vegada podrà ser estudiat des d'un punt de vista holístic a nivell cel·lular. Més específicament, en el cas del càncer colorectal, una malaltia altament heterogènia i molecularment molt complexa, seran de gran ajuda tots els avenços possibles en quant a tècniques de laboratori, tecnologia i noves disciplines. Durant els propers anys molt probablement s'incorporaran a la rutina clínica noves classificacions moleculars, que permetran prescriure un tractament més precís en comptes dels agents quimioteràpics utilitzats actualment amb la majoria dels pacients. També s'identificaran biomarcadors més acurats per al diagnòstic precoç i per predir millor el pronòstic de la malaltia. L'aprofundiment en les bases moleculars del càncer colorectal ens durà fins i tot a desenvolupar eines diagnòstiques basades en criteris moleculars, en substitució dels criteris clínics i patològics emprats majoritàriament a dia d'avui.

Amb aquesta tesi el lector podrà endinsar-se en treballs concrets que intenten avançar en aquestes direccions. Els resultats es presenten com a compendi de tres articles científics publicats a revistes científiques internacionals. Tots tres tracten de càncer colorectal, i són el resultat de part del treball realitzat durant els últims anys a la Unitat de Biomarcadors i Susceptibilitat de l'Institut Català d'Oncologia.

Llistat d'acrònims

AFAP	*Attenuated FAP*
ARACNe	*Algorithm for the Reconstruction of Accurate Cellular Networks*
CCR	Càncer ColoRectal
CIMP	*CpG Island Methylator Phenotype*
CIN	*Chromosomal INstability*
CpG	*Cytosine-phosphate-Guanine*
DNA	*Deoxyribo Nucleic Acid*
ELISA	*Enzyme-Linked ImmunoSorbent Assay*
FAP	*Familial Adenomatous Polyposis*
FDR	*False Discovery Rate*
GEO	*Gene Expression Omnibus*
HNPCC	*Hereditary NonPolyposis Colorectal Cancer*
IARC	*International Agency for Research on Cancer*
MAP	*MYH Associated Polyposis*
MARINa	*MAster Regulator INference algorithm*
MINDy	*Modulator Inference by Network Dynamics*
miRNA	*micro RNA*
MR	*Master Regulator*
mRNA	*messenger RNA*
MSI	*MicroSatellite Instability*
PCR	*Polymerase Chain Reaction*
RNA	*RiboNucleic Acid*
ROC	*Receiver Operating Characteristic*
RT-qPCR	*Real-Time quantitative PCR*
TCGA	*The Cancer Genome Atlas*
TF	*Transcription Factor*
TSOF	Test de Sang Oculta en Femta
UICC	*Union for International Cancer Control*

Índex

Introducció

1. El càncer colorectal

1.1 Epidemiologia del càncer colorectal

El càncer colorectal (CCR) a nivell mundial és el tercer càncer més freqüent en homes i el segon en dones. L'any 2012 es va registrar una incidència mundial superior a 1,3 milions de casos nous, amb una distribució molt similar entre homes i dones. Al menys un 55% dels casos es donen a les regions més desenvolupades. A Europa es registra una alta incidència, situant al CCR com la segona neoplàsia més freqüent si es sumen ambdós sexes. L'any 2012 es van registrar més de 345.000 casos nous només als estats membres de la unió Europea [1].

A Catalunya s'estima que la incidència de CCR tindrà un increment d'un 35% amb un total de 8.022 casos nous l'any 2020, en comparació dels 5.903 casos al 2010 [2]. La major part d'aquest increment es deurà a l'envelliment de la població, però també hi contribuiran l'increment de la població i els canvis esperats derivats de l'evolució en els factors de risc del CCR com els patrons de dieta, l'obesitat i la manca d'activitat física. A Catalunya avui més d'un 68% dels nous casos diagnosticats tenen una edat superior als 65 anys, el que indica que aquesta malaltia està relacionada amb l'envelliment. El risc acumulat d'un individu de patir la malaltia al llarg de tota la seva vida és aproximadament del 6%.

La supervivència als 5 anys en països desenvolupats actualment és al voltant del 65% (www.seer.cancer.gov, www.eurocare.it). L'elevada mortalitat del CCR, juntament amb l'augment de la incidència, converteixen aquesta malaltia en un greu problema de salut, ja que és una de les principals causes de mort a les regions desenvolupades. La mortalitat del CCR en gran mesura depèn de l'estadi en el moment del diagnòstic. Però, a diferència de la tendència creixent que s'observa en la incidència, la mortalitat per CCR s'ha mantingut estable o en lleuger descens, degut principalment a la detecció precoç, que permet detectar la malaltia en estadis inicials, i a lleugeres millores en el tractament. En conseqüència, i com es pot veure en la Figura 1, la mortalitat no té una distribució geogràficament uniforme, degut a que les regions menys desenvolupades acumulen una mortalitat més elevada [1].

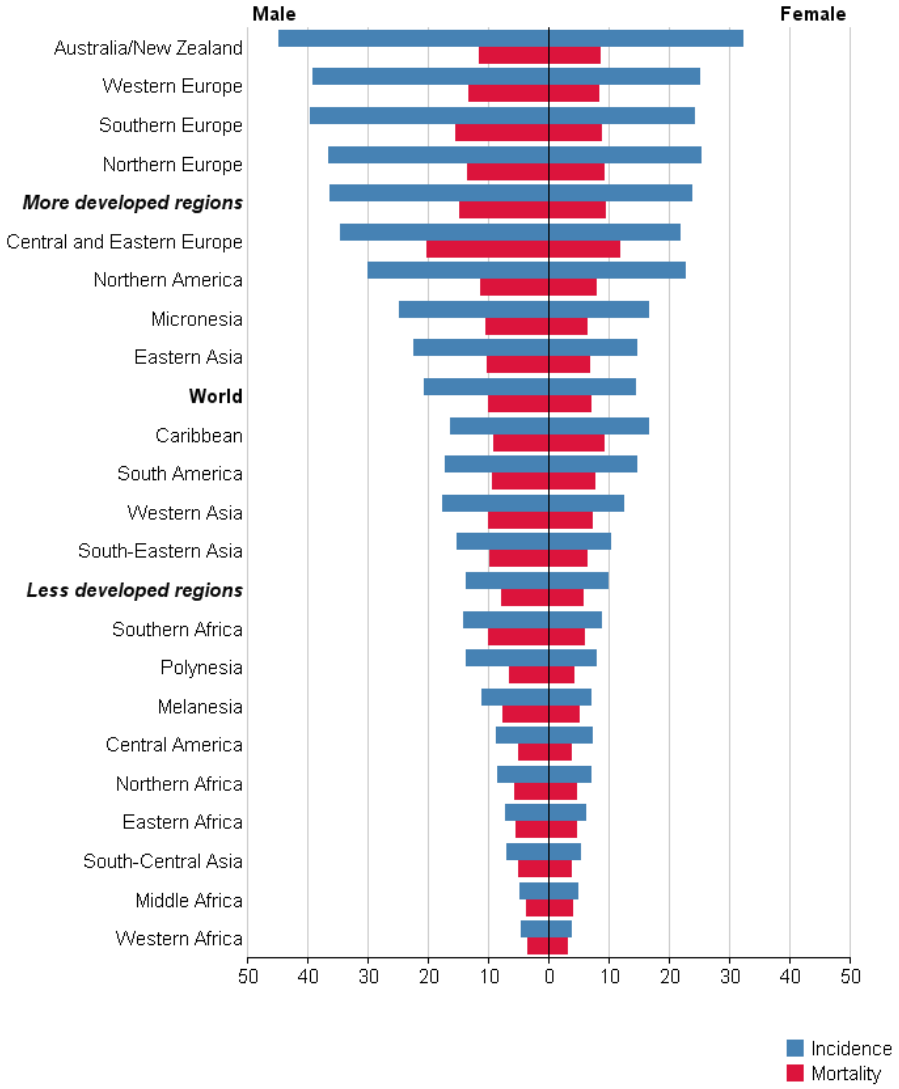

Figura 1. Taxes estimades i estandarditzades per edat (per 100.000 habitants), de la incidència i la mortalitat de càncer colorectal a nivell mundial l'any 2012; GLOBOCAN - IARC (globocan.iarc.fr).

1.2 Etiologia del càncer colorectal

Encara a l'actualitat no es coneixen les causes exactes de l'elevada incidència de CCR als països desenvolupats. En el desenvolupament del CCR, com en la majoria de les altres neoplàsies, intervenen una barreja de factors ambientals i factors genètics, com representa la Figura 2. És ja ben conegut que l'herència de certes mutacions en gens concrets poden elevar el risc de patir CCR fins a prop de 100%. Estudis en bessons han estimat d'heretabilitat del CCR en un 35% [3]. Per altra banda, es coneix que la component ambiental és força important ja que poblacions amb el mateix origen ancestral tenen diferent incidència en funció del grau de desenvolupament de la regió on resideixin.

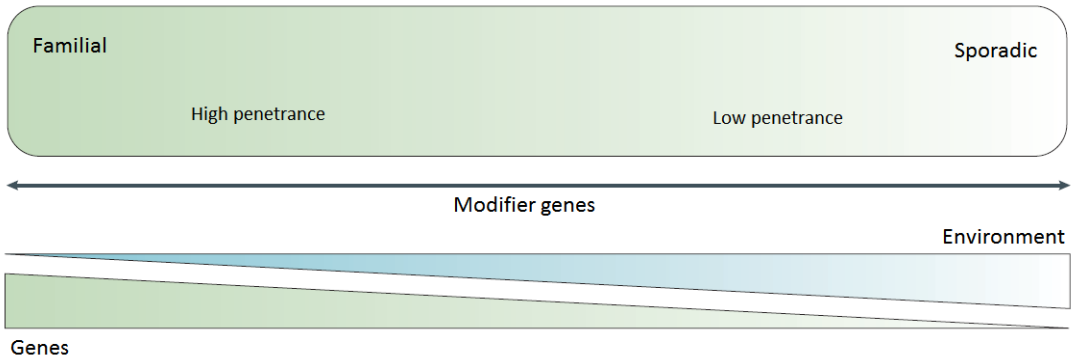

Figura 2. Visió global de la contribució genètica al CCR [4].

Els principals factors de risc per al CCR són, en primer lloc i més important, l'edat avançada. En segon lloc trobem els antecedents patològics d'adenomes, de malalties inflamatòries intestinals, com poden ser la malaltia de Crohn o la colitis ulcerosa, antecedents de diabetis i, òbviament els propis antecedents familiars de CCR. A continuació trobem tota una sèrie de factors ambientals i d'estil de vida suportats per diferents nivells d'evidència científica. Els factors ambientals que poden influenciar el risc de càncer de còlon més estudiats han estat el consum de tabac i d'alcohol, la dieta inadequada (elevat consum de carns vermelles i consum reduït de fruita i verdura), el sedentarisme i l'obesitat. En una menor mesura per al cas del CCR, es podria mencionar també l'exposició a carcinògens ambientals. Per altra banda, el consum crònic de certs fàrmacs podria tenir un cert efecte protector.

1.2.1 Factors ambientals

Entre els factors ambientals relacionats amb el risc de patir CCR, la dieta hi juga un paper clau [5]. S'han observat diferències geogràfiques importants en la incidència de CCR; les regions amb incidències més elevades s'han associat al fet de consumir una dieta més occidental. Nombrosos estudis han demostrat una associació significativa entre els hàbits alimentaris (dieta pobra en fibres, verdures i fruites, i rica en greixos animals i carn processada) i l'aparició de la malaltia [6]. A més a més, les dietes poc saludables amb un alt contingut en greixos i carn processada, normalment van acompanyades d'una reducció al consum de fibra, fuites i verdures, que des de ja fa temps que es coneix que tenen un cert efecte protector [7].

També fa anys que es coneix que el sedentarisme és un factor de risc important per al CCR. Però recentment van apareixent estudis més concrets demostrant que estils de vida sedentaris, com per exemple passar estones prolongades assegut mentre es veu la televisió, estan significativament associats amb un major risc de patir CCR [8]. Contràriament, en nombrosos estudis s'ha vist que la pràctica diària d'activitat física

redueix el risc de patir CCR [9]. Es creu que aquesta reducció en el risc de CCR gràcies a la realització d'activitat física, està provocada per diversos factors. Un d'ells és que l'increment de moviment a l'intestí implica un descens del temps que els factors carcinògens estan en contacte amb la mucosa del còlon. Un altre podria ser que la disminució de l'estat inflamatori sistèmic degut a l'activitat física podria tenir un efecte antineoplàstic.

L'obesitat, que ha ja arribat a assolir proporcions epidèmiques i que recentment ha estat reconeguda com a malaltia, també està relacionada amb el risc de desenvolupar molts tipus de càncer. Es creu que certs mecanismes inflamatoris de l'obesitat, com els perfils anormals de lípids, els nivells de glucosa o certs síndromes metabòlics, poden estar estretament relacionats amb el càncer [10].

El paper d'altres factors de risc associats a l'estil de vida com ara el consum de tabac o alcohol encara són poc concloents i generen certa controvèrsia. Ja fa uns anys van aparèixer tota una sèrie d'estudis en els quals es descrivia una clara associació entre el consum de tabac i el risc de patir CCR [11]. Es postulava que els carcinògens derivats de la combustió del tabac podien arribar a l'intestí gros a través del tracte digestiu o de la circulació sanguínia i que podien actuar de forma negativa en gens relacionats amb el càncer. Estudis més recents mostren un petit increment de risc, però amb diferencies segons la localització del tumor [12]. Els resultats pel que fa al consum de tabac són per tant inconsistents. Alhora, amb la hipòtesi que el risc degut al consum de tabac podria interactuar amb una certa predisposició genètica, es van realitzar estudis d'interacció gen-ambient que no han donat resultats positius. És a dir, cap dels polimorfismes genètics coneguts que estan implicats en el metabolisme dels carcinògens del tabac sembla modificar el risc de CCR degut al consum de tabac [13]. Respecte al consum d'alcohol, s'ha reportat en alguns estudis que el consum excessiu està associat amb un increment de risc de CCR [5], però els efectes d'un consum moderat segueixen sense estar clars [14]. També, el risc difereix segons el subtipus molecular i la localització anatòmica del tumor, per exemple, el recte està associat amb un major risc que el còlon [15]. A més a més, el mecanismes exactes per els quals l'alcohol influeix al risc de patir CCR tampoc estan massa clars [16].

D'altra banda, ja des de fa anys es va observar que l'aspirina i altres antiinflamatoris no esteroïdals, els suplements d'àcid fòlic o de calci, i la teràpia hormonal substitutiva (estròg ens) després de la menopausa, s'associen a un menor risc de patir CCR [17]. Posteriorment es va incorporar a la llista les estatines, que amb la seva interacció en el metabolisme del colesterol, també s'han vist relacionades amb una reducció del risc de CCR [18].

1.2.2 Predisposició genètica al càncer colorectal

Del total de casos de CCR, al voltant del 70-85% es corresponen al que s'anomenen les formes esporàdiques de la malaltia, és a dir, en els que no hi ha antecedents familiars de la malaltia coneguts. El CCR familiar, que representa aproximadament un 15% dels casos, és aquell en que, si bé la freqüència d'aparició en una mateixa família és superior a la que seria en el cas dels esporàdics, tampoc no encaixa exactament amb al patró característic d'una síndrome hereditària. Un cas de CCR es pot considerar familiar quan es presenten dos o més familiars de primer grau afectats. La resta de síndromes hereditaris del CCR causats per gens d'alta penetrància coneguts, representen en conjunt entre un 2-6% de tots els casos de CCR [19], com es pot veure a la Figura 3.

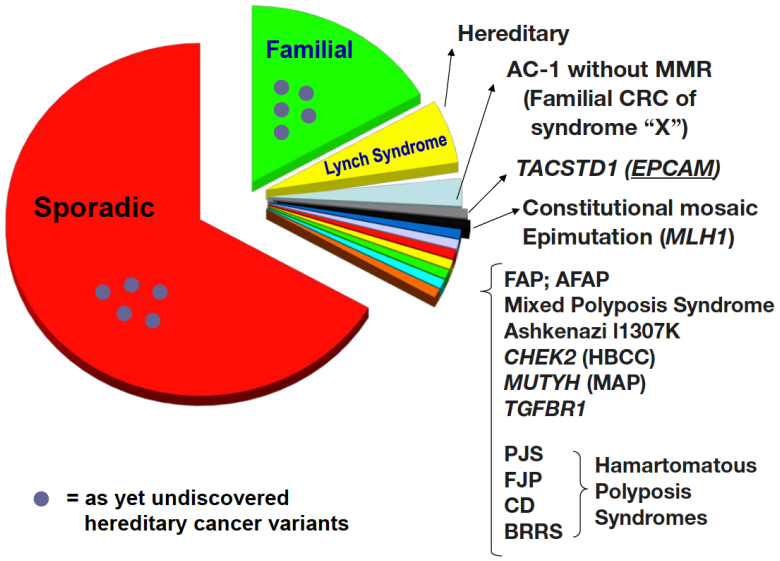

Figura 3. Nombre relatiu de casos de càncer colorectal considerats esporàdics, familiars o deguts a algun síndrome conegut de càncer hereditari [20].

Respecte les formes hereditàries del CCR, aquestes es poden dividir principalment en dos tipus: les associades a poliposi i les no associades a poliposi. Entre les síndromes associades a poliposi, la Poliposi Adenomatosa Familiar clàssica (FAP, de l'anglès *Familial Adenomatous Polyposi*) és la que es troba més freqüentment. Amb menys freqüència d'aparició trobem la Poliposi Adenomatosa Familiar Atenuada (AFAP, de l'anglès *Attenuated FAP*), i també la poliposi associada al gen *MYH* (MAP, de l'anglès *MYH Associated Polyposis*). Les poliposis hamartomatoses són un grup poc freqüent de poliposis, que engloben un conjunt de síndromes de diferents tipus, com la síndrome de Peutz-Jeghers, la poliposi Juvenil Familiar, la síndrome de Cowden i la síndrome de Bannayan-Ruvalcaba-Riley. La síndrome no associada a poliposi més important és la síndrome de Lynch o el CCR hereditari no associat a poliposi (HNPCC, de l'anglès *Hereditary NonPolyposis Colorectal Cancer*) [4].

L'estudi molecular detallat de les formes hereditàries més freqüents (FAP i HNPCC) ha resultat de gran importància per la comprensió i maneig del CCR, permetent conèixer les bases genètiques i els mecanismes implicats en el desenvolupament de la majoria dels casos de CCR esporàdic. S'ha descrit que alguns dels gens que es troben mutats a nivell germinal a la FAP i al HNPCC, també tenen un paper clau en els casos de CCR esporàdic. Aproximadament el 85% dels tumors esporàdics segueixen la mateixa via que els de la FAP i fins a un 15% segueixen la mateixa que els HNPCC [21].

Donat que actualment els casos de CCR hereditari no arriben a explicar fins el 35% de la possible heretabilitat genètica estimada, continua havent-hi oportunitat per descobrir variants addicionals. Els factors de risc genètics establerts fins al moment s'estenen entre dos extrems: les mutacions rares d'alta penetrança que confereixen grans augments de risc als síndromes hereditaris, i les variants comunes que confereixen efectes febles sobre el risc en individus amb o sense antecedents familiars de CCR. Durant els últims anys s'han publicat diversos estudis d'associació genètica a nivell del genoma complert, on s'han identificat nous polimorfismes freqüents que modifiquen lleugerament el risc de patir CCR, fins arribar a més de 40 regions genètiques que s'associen amb efectes febles sobre el CCR esporàdic [22]. Actualment, les variants més rares també s'estan estudiant, però fins al moment els resultats són escassos, probablement degut a la dificultat estadística per identificar associacions amb aquests polimorfismes. Una possible solució a aquest problema seria realitzar estudis amb mides de mostra molt grans. Per últim, altres aspectes genètics com la variació en el nombre de còpies de certes regions genòmiques, també podria ser un factor de susceptibilitat a tenir en compte.

No obstant això, durant el desenvolupament dels treballs d'aquesta tesi, totes les mostres seleccionades provenien de pacients amb CCR esporàdic. Els casos amb història familiar de CCR reportada van ser exclosos.

1.3 Anatomia i histologia del còlon i el recte

El còlon i el recte, juntament amb el cec, l'apèndix i el canal anal, conformen l'última part del sistema digestiu, coneguda com a intestí gros. La seva principal funció és la reabsorció d'aigua, ions minerals i els últims nutrients, abans d'eliminar finalment del cos el material de rebuig. Té una longitud aproximada de 150 centímetres, estenent-se des del final de l'intestí prim fins a l'anus, com es pot veure a la Figura 4. A nivell anatòmic, l'intestí gros comença a la part inferior dreta de l'abdomen, on desemboca l'intestí prim i es diu cec. Des d'aquí el còlon puja fins arribar a la zona del fetge (còlon ascendent), travessa l'abdomen (còlon transvers) i es dirigeix posteriorment cap avall (còlon descendent). Finalment arriba a una zona anomenada sigma que desemboca al recte i aquest a l'anus que s'obre a l'exterior per l'esfínter. Molt sovint també es fa ús d'altres terminologies per fer referencia

a les diferents localitzacions del còlon. Per exemple, el còlon dret (o proximal) seria el comprès des del cec fins a una mica més de la meitat del transvers i el còlon esquerre (o distal) tota la resta fins al recte.

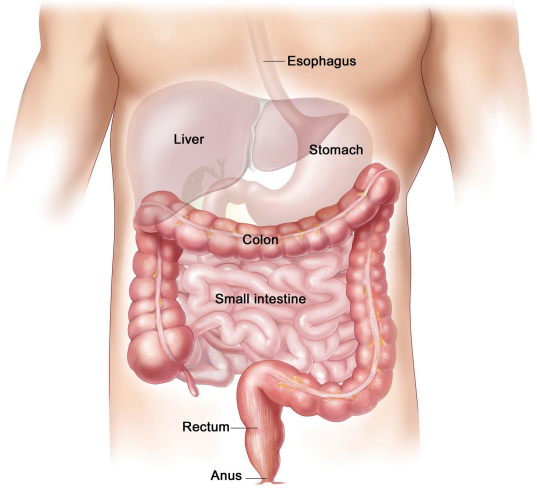

Figura 4. Anatomia gastrointestinal inferior i localització del còlon i el recte.

Existeixen importants diferencies en la vascularització del còlon i el recte segons la seva localització. La part dreta del còlon rep branques de l'artèria mesentèrica superior. Aproximadament a partir de la meitat del còlon transvers, la part esquerra i el recte, reben branques de l'artèria mesentèrica inferior. Les venes porten un curs anàleg al de les artèries i van a confluir a la vena mesentèrica inferior, que s'uneix a l'esplènica i mesentèrica superior per formar la vena porta hepàtica. Aquestes diferencies morfològiques del sistema venós del còlon podrien estar relacionades amb les petites diferències de pronòstic que es troben, degut a una major o menor facilitat per la disseminació a distància [23]. De fet, una altra qüestió que es tracta més endavant i que també genera controvèrsia és si els tumors situats al còlon i al recte, o els situats al còlon dret o còlon esquerre, són molecularment el mateix tipus de tumors [24].

Tant el còlon com el recte estan constituïts per diverses capes de teixit, la més interna i que realment separa el cos humà de l'exterior és l'epiteli i les seves vellositats, que es troben situades a la mucosa. Aquesta es troba envoltada per la submucosa i més externament es troba la capa muscular, que és la que aconsegueix l'avanç del contingut pel tub digestiu gràcies a les seves contraccions. Per últim, aquesta capa muscular es troba recoberta per la serosa, que a grans trets és la capa més externa (Figura 5A).

A l'epiteli es troben les criptes intestinals o criptes de Lieberkühn dissenyades per incrementar la superfície de l' intestí i a on es realitza finalment l'absorció de nutrients.

Degut als constants esforços mecànics i químics que pateixen, aquestes criptes necessiten estar dotades d'un mecanisme de constant renovació. És per això que a cadascuna d'aquestes criptes s'hi pot trobar un petit reservori de cèl·lules mare (*stem cells* en anglès) que renoven constantment l'epiteli mitjançant una sèrie de processos ben ordenats, que inclouen la proliferació, la diferenciació i la migració cap a la superfície de l'epiteli (Figura 5B). Durant el viatge des del fons de la cripta fins a la seva superfície, aquestes cèl·lules mare donen lloc a cèl·lules pluripotents que es van diferenciant en diversos tipus cel·lulars encarregats de les diferents funcions per mantenir l'homeòstasi intestinal. Quan aquestes cèl·lules ja diferenciades, després de 5 dies aproximadament arriben a la superfície de la cripta, activen el seu procés d'apòptosi i moren a l'epiteli. De fet, l'origen del CCR es troba en aquestes cèl·lules mare internes, situades al fons de les criptes de l'epiteli, que es troben en constant renovació i diferenciació.

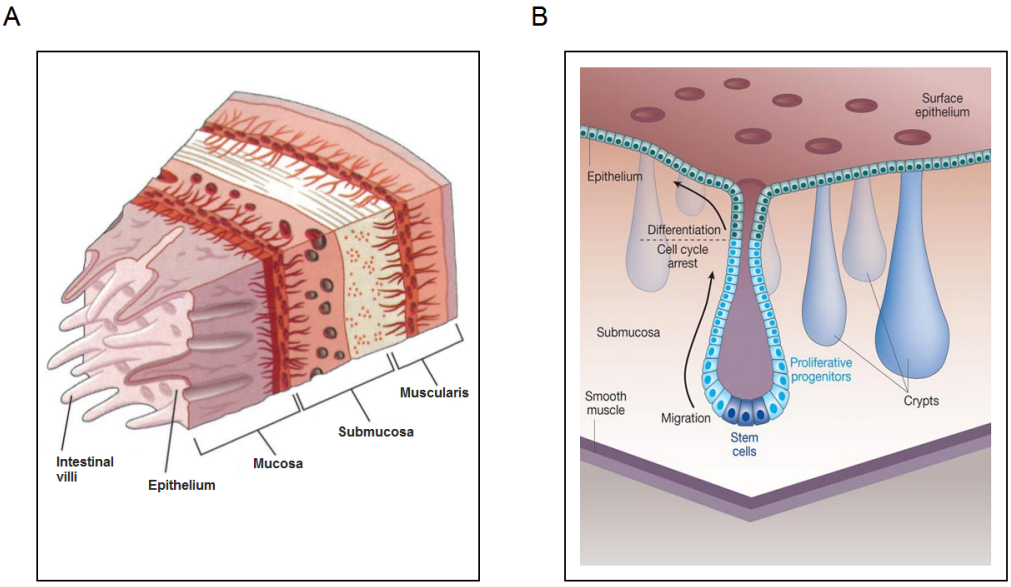

Figura 5. (A) Secció transversal del còlon i les diferents capes que el composen. (B) Anatomia del teixit de l'epiteli i representació esquemàtica d'una cripta del còlon [25].

1.4 Patologia del còlon i el recte

1.4.1 Seqüència clàssica adenoma-carcinoma

Degut a la seva accessibilitat, i a que és relativament fàcil obtenir biòpsies durant les diferents etapes de la progressió de la malaltia, els tumors de CCR són uns dels més estudiats molecularment. Ara fa 25 anys Fearon i Vogelstein van proposar un model de progressió per al CCR, que es basa en un procés de carcinogènesi seqüencial en el qual es van adquirint successives alteracions moleculars, a mesura que van apareixent els estadis més avançats de la malaltia [21]. Aquests estadis serien els següents:

- *Epiteli normal:* mucosa colònica patològicament normal.
- *Focus de criptes aberrants:* aquesta seria la primera lesió que es pot observar a la mucosa colònica i que podria donar lloc a un futur carcinoma.
- *Adenoma primerenc:* majoritàriament són pòlips hiperplàsics de mida inferior a 1 cm.
- *Adenoma intermedi:* pòlips adenomatosos o adenomes tubulars de més d'1 cm però sense presència de carcinoma.
- *Adenoma tardà:* adenomes de mida superior a 1 cm i que presenten focus de carcinoma intraepitelial. És el que també es coneix com a carcinoma *in situ*.
- *Carcinoma:* tumor maligne (càncer) que es desenvolupa a les cèl·lules epitelials i que ja posseeix capacitat invasiva.
- *Metàstasi:* és la colonització tumoral d'òrgans a distància mitjançant una disseminació de cèl·lules que es desprenen del tumor primari.

Un aspecte addicional al qual també els autors fan èmfasi i que després ha estat corroborat en altres treballs posteriors és que, tot i que les alteracions genètiques sovint tenen lloc en una seqüència coneguda, l'acumulació total de canvis, més que el seu ordre, sembla ser un determinant crític de les propietats biològiques de la cèl·lula tumoral [26]. En l'actualitat es creu àmpliament que el procés de carcinogènesi s'origina degut a mutacions somàtiques i una inhibició de supressors del creixement, que dóna lloc a una proliferació cel·lular anormal, la invasió dels teixits adjacents i finalment el risc de metàstasi [27].

En el context del CCR, mutacions en el gen *APC* serien les primeres en iniciar el procés neoplàsic provocant una proliferació anormal a l'epiteli i donant origen a certs focus de criptes aberrants. Posteriorment, algunes d'aquestes es podrien transformar en petites lesions precursores com són els adenomes, créixer i donar lloc a carcinomes. A més, es coneix que mutacions en el gen *KRAS* i d'altres oncogens, poden conferir als adenomes primerencs, la capacitat de creixement i transformació cap a adenomes avançats.

Durant l'ultima dècada, gràcies als nous mètodes computacionals i als esforços integrals de seqüenciació, aquest model de progressió sobre el CCR proposat inicialment per Fearon i Vogelstein a l'any 1990, s'ha pogut anar revisant i completant. Aquests nous coneixements, han revelat un gran nombre de gens mutats somàticament però amb baixa freqüència entre els diferents tumors analitzats, i d'altra banda uns pocs gens alterats recurrentment en un alt percentatge dels tumors. Actualment s'està treballant per identificar acuradament els gens, que quan estan mutats, promouen o condueixen el procés tumoral. A aquestes se les anomena mutacions en gens conductors; la resta en canvi, s'anomenen mutacions

passatgeres, és a dir, mutacions que no confereixen cap avantatge per a un creixement selectiu. De tota manera, és complicat identificar realment quines mutacions somàtiques són conductores i quines simplement són passatgeres, i és que gens amb mutacions conductores, alhora poden contenir mutacions passatgeres. Fins al moment s'ha pogut comprovar que aproximadament un tumor conté entre dos i vuit d'aquests gens amb mutacions conductores. Per altra banda, s'ha comprovat que aquests gens es poden classificar dins d'unes poques vies de senyalització que s'encarreguen de regular tres processos cel·lulars fonamentals: el destí cel·lular, la supervivència i manteniment del genoma.

En el context del CCR, mutacions en el gen *APC* es considerarien les principals mutacions conductores en el desenvolupament dels tumors. La Figura 6 resumeix les principals vies de senyalització que condueixen al desenvolupament tumoral i a les transicions entre les diferents etapes del CCR. Cal tenir en compte que algunes d'aquestes mutacions conductores en múltiples gens que codifiquen components d'aquestes vies, es poden trobar alterats a qualsevol tumor individual. I que cadascun dels tumors no té perquè presentar alteracions en tots aquests gens.

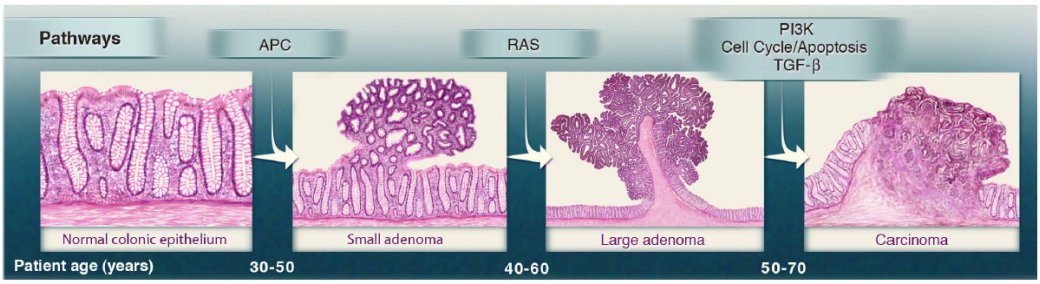

Figura 6. Model de progressió del CCR i principals alteracions genètiques, classificades als intervals de temps durant el qual usualment apareixen [28].

Realment, encara es necessita una millor comprensió d'aquestes vies de senyalització fonamentals, perquè aquests nous coneixements de la biologia molecular del CCR siguin suficients per guiar el desenvolupament d'enfocaments més eficaços per reduir la morbiditat i la mortalitat.

1.4.2 Classificació i estadis del càncer colorectal

S'han definit diverses classificacions del CCR segons la seva extensió. La més simple i la primera que va ser utilitzada és la classificació original de Dukes, publicada a l'any 1932 i modificada uns anys després per Astler i Coller. Posteriorment, a l'any 1986 Hutter i Sobin van desenvolupar un nou sistema d'estadiatge anomenat TNM, que té en compte diversos paràmetres clínics, exploracions físiques, proves radiològiques i d'anatomia patològica, per finalment acabar classificant els tumors en quatre estadis diferents. El sistema TNM es basa en caracteritzar el grau d'invasió i la propagació del tumor, en relació a les diferents capes que composen l'intestí i que estan afectades pel tumor primari (T), al nombre de nòduls limfàtics afectats (N) i a la presència o no de metàstasi a distància (M) [29]. Aquesta classificació ha estat globalment acceptada per la unió internacional contra el càncer (UICC, de l'anglès *Union for International Cancer Control*) com a sistema per descriure l'extensió anatòmica dels tumors, i ha estat objecte de diverses revisions significatives pels experts [30] per intentar reflectir la comprensió actual de l'extensió de la malaltia i el seu rol al pronòstic, utilitzat habitualment per predir la supervivència del pacients [31].

Concretament, durant el transcurs d'aquesta tesi, l'agrupació per estadis dels tumors colorectals es va realitzar amb l'ajuda de la guia autoritzada de la UICC ''TNM Atlas, 6th Edition'' i mitjançant les definicions i la classificació que podem trobar resumida a la Taula 1.

El pronòstic del CCR està estretament relacionat amb l'estadi al diagnòstic [32]. Els tractaments actuals s'adeqüen a l'estadi. Per exemple, la cirurgia és suficient per al tractament del càncer de còlon no disseminat (estadis I-II). La quimioteràpia adjuvant és eficaç quan hi ha disseminació ganglionar (estadi III) i la radioteràpia complementa el tractament en tumors localitzats en el recte [33]. El pronòstic de la malaltia una vegada s'ha disseminat a altres òrgans (estadi IV) és pobre [31] i les alternatives terapèutiques escasses, fet que reforça el gran interès per la detecció precoç del CCR.

T - Tumor primari	
TX	El tumor primari no pot ser avaluat
T0	No hi ha evidència de tumor primari
Tis	Carcinoma *in situ*: intraepitelial o invasió de la *làmina pròpia*
T1	Tumor que envaeix la submucosa
T2	Tumor que envaeix la *muscularis propria*
T3	Tumor que envaeix la subserosa
T4	Tumor que envaeix directament altres òrgans o perfora el peritoneu visceral
T4a	El tumor perfora el peritoneu visceral
T4b	El tumor envaeix directament altres òrgans
N - Nòduls limfàtics regionals	
NX	Els nòduls limfàtics regionals no poden ser avaluats
N0	No hi ha metàstasi en els nòduls limfàtics regionals
N1	Metàstasi en 1-3 nòduls limfàtics regionals
N1a	Metàstasi en 1 nòdul limfàtic regional
N1b	Metàstasi en 2 o 3 nòduls limfàtics regionals
N1c	Dipòsits tumorals a la subserosa, a la membrana pericòlica o al teixit tou perirectal
N2	Metàstasi en 4 o més nòduls limfàtics regionals
N2a	Metàstasi en 4-6 nòduls limfàtics regionals
N2b	Metàstasi en 7 o més nòduls limfàtics regionals
M - Metàstasi a distància	
M0	No hi ha metàstasi a distància
M1	Metàstasi a distància
M1a	Metàstasi confinada a un òrgan
M1b	Metàstasi en més d'un òrgan o al peritoneu

AGRUPACIÓ PER ESTADIS

Stage	**Grups TNM**
0	Tis, N0, M0
I	T1-T2, N0, M0
II A	T3, N0, M0
II B	T4a, N0, M0
II C	T4b, N0, M0
III A	T1-T2, N1, M0; T1, N2a, M0
III B	T1-T2, N2b, M0; T2-T3, N2a, M0; T3-T4a, N1, M0
III C	T3-T4a, N2b, M0; T4b, N1-N2, M0; T4a, N2a, M0
IV A	qualsevol T, qualsevol N, M1a
IV B	qualsevol T, qualsevol N, M1b

Taula 1. Definicions del TNM als carcinomes colorectals, classificació clínica i agrupació per estadis.

2. Estat actual de la prevenció del càncer colorectal

2.1 Prevenció primària i secundària

La reducció en la incidència del CCR pot beneficiar-se d'estratègies preventives, aplicables en diferents moments de la història natural de la malaltia. Per exemple, mitjançant la promoció d'hàbits saludables a través de campanyes d'educació per la salut quan la malaltia no és encara present, amb la implementació de programes de cribratge per a la detecció precoç de la malaltia, o mitjançant l'ús de teràpies més personalitzades segons les característiques del pacient i del propi tumor [34].

La prevenció primària es podria definir, de forma general, com el conjunt d'actes destinats a disminuir la incidència de la malaltia reduint el risc d'aparició de nous casos. Més concretament i parlant de CCR, el que busca és evitar l'inici o retardar al màxim l'aparició de la malaltia eliminant els principals factors etiològics, com poden ser el consum de tabac i alcohol, el sedentarisme, l'obesitat i la dieta inadequada. Normalment la manera de portar-ho a terme és modificant els hàbits poc saludables de la població cap a uns altres més adequats mitjançant accions de promoció de la salut, prevenció de la malaltia i protecció de la salut. Una altra estratègia preventiva per al CCR podria ser la quimioprevenció, que encara no pot ser acceptada a la pràctica mèdica habitual degut a que el seu benefici no ha sigut confirmat mitjançant els assaigs clínics corresponents. A més, certs informes alarmen sobre els riscos cardiovasculars d'alguns dels antiinflamatoris no esteroïdals més tradicionalment utilitzats, fent que es produís un cert estancament en aquest camp [35]. Fins al moment cap mesura de prevenció primària a nivell poblacional ha demostrat ser eficient per reduir-ne la incidència. Encara es necessiten estudis addicionals, per definir millor els agents protectors específics i així posteriorment poder recomanar modificacions concretes dels hàbits o l'estil de vida [36]. Malgrat tot, es manté com a important missatge de salut pública l'interès per modificar els factors de risc coneguts per al CCR, com la reducció en el consum de carn vermella, fer exercici regularment, deixar de fumar i tenir un pes adequat.

D'altra banda, la prevenció secundaria intenta detectar i aplicar tractament a la malaltia durant estadis molt primerencs. És a dir, la intervenció té lloc al principi de la malaltia amb l'objectiu d'impedir o retardar el desenvolupament de la mateixa. Parlant en el context del CCR, la prevenció secundària es converteix principalment en la implantació de programes de cribratge poblacional mitjançant les diferents estratègies i recomanacions existents [37]. Principalment els mètodes utilitzats poden ser dividits en dos grans categories: la visualització de l'intestí gros gràcies a diferents mètodes exploratoris com la colonoscòpia [38], sigmoidoscòpia [39, 40], ènema de bari de doble contrast [41], colonoscòpia virtual

[42] i d'altres, i la detecció de marcadors mitjançant l'anàlisi de mostres biològiques, com pot ser el Test de Sang Oculta en Femta (TSOF) i altres biomarcadors que detallarem més endavant. Per exemple, els programes de cribratge poblacional europeus típicament utilitzen bianualment el TSOF com a prova de cribratge i la colonoscòpia com a confirmació exploratòria en cas de que el TSOF doni un resultat positiu [43]. En altres països les recomanacions són diferents: al Japó per exemple, s'utilitza el TSOF anualment, en canvi als Estats Units, les recomanacions són o bé un únic TSOF anual, sigmoidoscòpia, ènema de bari de doble contrast o colonoscòpia virtual cada 5 anys, o una colonoscòpia convencional cada 10 anys.

Per últim, es considera prevenció terciària el conjunt de totes aquelles actuacions aplicades durant el curs clínic de la malaltia, és a dir, quan els símptomes i els signes són ja presents. Estan principalment destinades a tractar la malaltia amb la menor agressivitat possible i evitant al màxim les toxicitats o limitacions per la qualitat de vida del pacient. El seu principal objectiu seria doncs reduir la prevalença de la malaltia.

2.2 Beneficis de la detecció precoç del càncer colorectal

Les estratègies de detecció precoç de CCR han demostrat ser efectives degut a que són capaces de trobar la malaltia durant els seus estats inicials, quan els tractaments són més efectius. La detecció precoç del CCR amb el TSOF seguit d'una colonoscòpia confirmatòria, ha demostrat una evident reducció a la mortalitat [44-46]. A grans trets existeixen dos tipus de TSOF, els antics tests bioquímics basats en la resina de Guaiac i els nous tests immunològics quantitatius, basats en la detecció específica d'hemoglobina humana en femta.

El test Guaiac consisteix en una prova qualitativa, en que els resultats s'obtenen mitjançant un reactiu que canvia el color en contacte amb la mostra de femta, fet que exposa el resultat a una avaluació subjectiva per part d'un tècnic de laboratori. Diversos estudis realitzats amb el Guaiac van evidenciar una disminució de la mortalitat que oscil·la entre el 15% i el 33% [38, 45, 47]. La franja de sensibilitat que ofereix el Guaiac per la detecció de neolplasies, depèn principalment del test comercial utilitzat i és molt amplia ja que oscil·la des d'un 6,2% fins a un 83,3%. En canvi, l'especificitat és més constant, normalment superior al 80% i aconseguint fins un 98,4% en alguns casos [48]. El gran punt feble del Guaiac és que degut al seu funcionament, que es basa en la detecció de l'activitat peroxidasa a l'hemoglobina, no és específic per l'hemoglobina humana. És a dir, que pot reaccionar amb la presencia de sang a la dieta (carn vermella) i provocar falsos positius. Per reduir aquesta problemàtica es recomana la recollida de diferents mostres espaiades en el temps i una sèrie de restriccions dietètiques uns dies previs a la realització de la prova [49].

En l'actualitat generalment s'empra un test immunològic més sensible i específic per detectar sang oculta en femta. Aquest test es basa en l'ús d'anticossos específics per detectar l'hemoglobina humana [50] i ofereix un resultat semi-quantitatiu. Alhora és un test millor acceptat per la població ja que no és necessari cap tipus de restricció dietètica prèvia, la prova és més fàcil de realitzar i requereix recollir un nombre inferior de mostres. Els resultats publicats fins al moment pel que fa a la sensibilitat del test immunològic són molt variables, incloent resultats amb molt poca sensibilitat (5,4%) i d'altres que arriben gairebé fins al 98%. L'especificitat també varia entre el 77% i el 99% [48, 51]. Diversos estudis han demostrat que per la detecció de lesions neoplàsiques, el test immunològic té una clara superioritat davant del Guaiac [52, 53].

Les possibles avantatges de fer directament colonoscòpia com a proba de cribratge és actualment un tema de debat en el nostre entorn [54, 55] als Estats Units i altres països ja fa temps que s'aplica [56]. La colonoscòpia està considerada com la prova de referència per al diagnòstic de patologies colorectals, degut a que exhibeix els millors resultats, amb un 99% de sensibilitat i amb una especificitat d'aproximadament el 98% per a les lesions grans [57]. En un estudi prospectiu observacional de cohorts, s'ha informat d'una reducció significativa en la incidència i la mortalitat per CCR de fins un 67% i 65%, respectivament [58]. De fet, en estudis recents s'ha observat que la reducció sostinguda de la mortalitat per CCR dels participats als programes de cribratge, està relacionada amb l'efecte de la polipectomia [38]. Però d'altra banda, la colonoscòpia té grans desavantatges, degut a què es tracta d'una prova invasiva i el procediment no està exempt de complicacions. La perforació i sagnat posterior a la polipectomia són els més greus. Típicament, els estudis reporten taxes de complicacions durant la colonoscòpia d'aproximadament un 0,1% [38]. Altres limitacions per l'ús de la colonoscòpia com a prova de detecció, són el seu cost econòmic i la seva baixa acceptació a nivell poblacional [59].

Un repte encara per aconseguir és implementar els programes de cribratge poblacional a tots els territoris perquè tothom hi pugui tenir accés i per tant beneficiar-se'n. També s'hauria de treballar per augmentar la taxa de participació a aquests programes de cribratge [60], potser amb el suport de campanyes d'informació i promoció que aconsegueixin sensibilitzar a la població. D'altra banda, un altre possible aspecte a millorar és la capacitat de detecció del TSOF o directament apostar per la recerca en nous biomarcadors útils, econòmics i que es puguin trobar en teixits no invasius per ser més acceptats per la població diana.

2.3 Nous biomarcadors de càncer colorectal

Durant aquests últims anys s'ha treballat intensament per identificar els factors de susceptibilitat genètica que permetin estratificar els individus amb més risc de patir un CCR

[61]. Amb tot i això, l'evidència acumulada fins avui suggereix que la utilitat dels polimorfismes funcionals per predir risc és limitada. En canvi, als mètodes actuals per la detecció precoç del CCR i al cribratge hi ha més marge de millora, tal com hem descrit anteriorment. De fet, actualment s'estan explorant una multitud d'opcions amb aquesta finalitat, com per exemple, la detecció de mutacions en femta, plasma o orina, la detecció de patrons de metilació aberrants, la detecció de certs microRNAs (miRNAs) o proteïnes circulants, etc. [62-64]. A continuació es detallen alguns dels biomarcadors que han tingut un major impacte durant els últims anys.

Detecció de l'antigen carcinoembrionari:

Els biomarcadors basats en la detecció de proteïnes al sèrum mitjançant un assaig per inmunoabsorció lligat a enzims (ELISA, de l'anglès *Enzyme-Linked ImmunoSorbent Assay*), podrien ser l'opció més econòmica i fiable per ser utilitzada a la pràctica clínica diària i també al cribratge poblacional. La detecció de l'antigen carcinoembrionari [65] és un dels marcadors tumorals més utilitzats arreu del món, i especialment per al CCR. L'antigen carcinoembrionari es coneix també com a CEA, en referència al nom de la proteïna que codifica pel gen *CEACAM5*. Encara que porta ja gairebé 30 anys en ús clínic amb un clar valor per al pronòstic, la progressió i per la detecció del CRC, la detecció de l'antigen carcinoembrionari per al cribratge del CCR té un impacte molt limitat. Això es degut principalment a causa de la baixa sensibilitat (aproximadament d'un 35%) i especificitat (entre el 30% i el 80%) [66] que demostra.

Detecció de la metilació del DNA a la Septina 9:

Un marcador molecular trobat en sang i que pot ser interesant, és la detecció de la metilació del l'àcid desoxiribonucleic (DNA, de l'anglès *Deoxyribo Nucleic Acid*) al gen *SEPT9*. Aquest és un candidat molt prometedor per al desenvolupament d'un mètode de detecció molecular no invasiu [67, 68]. Encara que l'assaig de la Septina 9 ha identificat amb èxit un 68% dels càncers de còlon amb una especificitat el 89% [69], el cost de la prova és massa elevat degut a que implica l'extracció del DNA i un assaig quantitatiu per avaluar el nivell de metilació. A més, l'assaig per avaluar la metilació al DNA pot esdevenir en un obstacle per la creació d'una eina de diagnòstic robusta, ja que sovint es requereix d'una etapa d'amplificació mitjançant una reacció en cadena de la polimerasa (PCR, de l'anglès *Polymerase Chain Reaction*).

Detecció de la metilació del DNA a la Vimentina:

Un altre potencial biomarcador per al CCR és la detecció de la metilació aberrant del gen de la Vimentina (*VIM*). Principalment, hi han hagut dues aproximacions diferents. Fa uns anys es va reportar una sensibilitat del 46% i una especificitat del 90% en mostres de femta [70].

Més recentment s'han reportat els resultats d'un estudi en que es troba una associació significativa de la forma metilada de la Vimentina, en mostres d'orina de casos de CCR comparant amb controls sense càncer [71].

Detecció de mutacions al DNA de la femta:

Es coneix que la presencia de DNA humà a la femta és escassa, ja que la petita fracció que s'obté prové en gran mesura de la dieta i de la flora bacteriana [72]. Malgrat això, la genètica molecular del CCR proporciona les bases per l'anàlisi del DNA en femta, degut a que aproximadament un 85% dels tumors colorectals acumulen mutacions als gens *APC*, *TP53* i *KRAS*, i la resta manifesten la pèrdua de gens implicats en la reparació del DNA. De fet, nombrosos estudis previs han reportat resultats satisfactoris en aquest àrea. Per exemple, un d'ells va comparar el resultats del TSOF bioquímic amb un panell de 21 mutacions que van ser analitzades a la femta. La sensibilitat reportada per la detecció del CCR va ser del 13% i 53%, respectivament [73].

Utilitat dels biomarcadors:

Nombrosos potencials candidats a biomarcadors addicionals han sigut proposats, com poden ser la detecció en femta d'algunes proteïnes habitualment presents al plasma [74], panells de gens que es detecten hípermetilats en DNA de femta [75], o anticossos de proteïnes circulants que es poden trobar al sèrum de pacients amb CCR [76]. En CCR també s'han publicat diversos estudis que han treballat en la detecció de miRNAs en teixit no invasiu [63, 77, 78]. Una sensibilitat compresa entre el 62% i el 89% i una especificitat entre el 70% i el 89%, va ser reportada pels diferents autors [79]. I molts altres candidats, que a la majoria dels casos no han progressat més enllà de la proposta [32, 80-84].

Malgrat l'àmplia llista de potencials biomarcadors pel diagnòstic precoç del CCR, molt pocs han arribat a la pràctica clínica. Per la majoria no s'arriba a demostrar la utilitat. Moltes vegades durant el procés de recerca de nous biomarcadors s'utilitzen aproximacions inadequades o incomplertes. Molts altres autors ja han evidenciat les limitacions dels estudis de biomarcadors, que són principalment: la limitada mida de la mostra emprada, el biaix de publicació (moltes vegades només es publiquen resultats significatius però no els resultats nuls) i una validació inadequada dels propis resultats en sèries de mostres independents [32, 80, 85, 86]. Per tot això, durant el procés de recerca de biomarcadors cal definir un protocol amb una mida mostral adequada per dur a terme les validacions corresponents. A més, també existeixen guies estandarditzades amb moltes recomanacions i els criteris necessaris per reportar resultats sobre els estudis de nous biomarcadors. Un exemple són les guies STARD [87] i REMARK [88], que es poden utilitzar com a base per definir mètodes i protocols de manera adequada.

3. Perfils moleculars en càncer colorectal

3.1 Bases moleculars del càncer

Hannahan i Weinberg ja a l'any 2000 van suggerir que la majoria de genotips cancerosos són la manifestació de sis capacitats adquirides al comportament cel·lular que, de forma conjunta, dicten el fenotip maligne [89]. Anys després, en una nova revisió, els mateixos autors van ampliar aquestes sis capacitats funcionals del fenotip tumoral a un total de vuit, més dos característiques habilitades pels tumors, com es pot veure a la Figura 7.

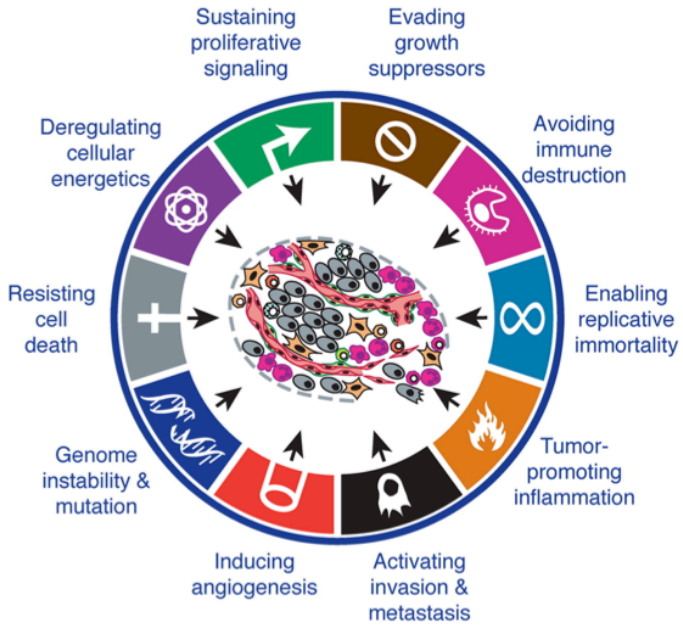

Figura 7. Capacitats cel·lulars adquirides pel càncer [90].

A continuació es llisten amb més detall les capacitats adquirides per les cèl·lules cancerígenes. Actualment aquestes capacitats constitueixen l'eix central de les bases moleculars per a la investigació en càncer:

- *Autosuficiència en senyals de creixement:* les cèl·lules necessiten senyals específics de creixement per passar d'un estat de quiescència a un estat de proliferació. Les cèl·lules tumorals generen aquests senyals de creixement, imitant a les cèl·lules normals i així aconsegueixen reduir la seva dependència del microentorn.

- *Insensibilitat a les senyals d'inhibició del creixement:* als teixits existeixen tota una sèrie de senyals d'inhibició del creixement que s'encarreguen de mantenir l'homeòstasi i l'estat de quiescència de les cèl·lules. Majoritàriament es tracten de

senyals relacionades amb el cicle cel·lular que les cèl·lules tumorals han d'aconseguir evadir per poder proliferar.

- *Evasió de l'apoptosi:* si pensem en les cèl·lules tumorals com en una població, la capacitat per créixer depèn tant de la capacitat proliferativa, com també de la mort cel·lular. El mecanisme més comú de mort cel·lular programada s'anomena apoptosi. Així doncs, l'evasió de l'apoptosi seria una capacitat necessària perquè les poblacions de cèl·lules tumorals puguin créixer.

- *Potencial replicatiu il·limitat:* a més de les interaccions amb l'entorn, les cèl·lules també tenen mecanismes propis per limitar el seu nombre de divisions i aturar el creixement per entrar en senescència. Els telòmers, situats al final de cada cromosoma, es van escurçant a cada cicle de replicació i com que tenen una longitud determinada, el nombre de divisions cel·lulars queda limitat intrínsecament. La majoria dels tumors adquireixen la capacitat de mantenir constant la mida dels seus telòmers i aconseguir així una capacitat de replicació il·limitada.

- *Angiogènesi sostinguda:* les cèl·lules necessiten l'aportament d'oxigen i nutrients a través dels vasos sanguinis. L'angiogènesi és el procés fisiològic que consisteix en la formació de nous vasos sanguinis a partir dels vasos preexistents. Normalment només es donaria en determinades ocasions, com al creixement o a la cicatrització de les ferides. Llavors, les cèl·lules tumorals adquireixen la propietat de promoure l'angiogènesi per així assegurar-se l'obtenció d'oxigen i nutrients.

- *Invasió de teixits i metàstasi:* un dels passos clau durant el procés de progressió tumoral és la invasió de teixits veïns. Les cèl·lules tumorals adquireixen la capacitat de viatjar cap a localitzacions distants o fins i tot altres òrgans, on poder formar noves colònies, canviant el seu microentorn. Aquest procés de colonitzar altres teixits és el que es coneix com a metàstasi, que és la responsable de la majoria de morts per càncer.

- *Desregulació de la energia cel·lular:* la progressió tumoral no està basada únicament en la proliferació cel·lular descontrolada, sinó que també es basa en ajustos del metabolisme de l'energia per tal d'impulsar el creixement i la divisió cel·lular. Concretament les cèl·lules canceroses, fins i tot en la presència d'oxigen, poden reprogramar el seu metabolisme de la glucosa, i per tant la seva producció d'energia, modificant el seu metabolisme de l'energia en gran mesura cap a la glucòlisi.

- *Evasió de la destrucció immune:* un últim aspecte, encara no resolt, és el paper que juga el propi sistema immunitari en la resistència o l'eradicació de les lesions

incipients, la progressió tumoral i la metàstasi. Les cèl·lules i els teixits són constantment monitoritzats per un sistema immunitari que està sempre alerta. Aquest és responsable de reconèixer i eliminar la gran majoria de les cèl·lules canceroses i dels tumors incipients. Sembla que els tumors que s'arriben a manifestar és perquè han aconseguit evitar ser detectats pels diferents mecanismes del sistema immunitari o bé han estat capaços de limitar la destrucció immune.

L'adquisició d'aquestes noves capacitats funcionals de les cèl·lules canceroses és possible gràcies a dues característiques que les promouen: la inestabilitat genòmica en les cèl·lules canceroses, que provoca mutacions aleatòries inclús reordenaments cromosòmics, i l'estat inflamatori a les lesions, que està impulsat per les cèl·lules del sistema immunològic i que serveix per promoure la progressió del tumor a través de diversos mitjans. Addicionalment, els tumors exhibeixen una altra dimensió de complexitat: aconsegueixen contenir un repertori de cèl·lules reclutades, aparentment normals, que contribueixen a l'adquisició d'aquestes noves capacitats, mitjançant la creació del "microentorn tumoral" [90].

Cal entendre que, per descomptat, no a tots els tumors es donen totes aquestes capacitats adquirides alhora, de fet, el que normalment es troba són combinacions d'elles que finalment confereixen al tumor la capacitat de proliferació il·limitada i independència del medi que l'envolta. A més, un mateix tumor pot exhibir diferents capacitats en diversos moments durant el procés de tumorigènesi. És per tota aquesta complexitat que cada vegada és més important conèixer específicament cada tumor a nivell molecular per veure en quina categoria encaixa, conèixer millor el seu pronòstic i poder utilitzar el tractament més adequat. En CCR, així com en molts altres tumors, existeixen ja diferents categories en base al seu tipus molecular, que estudiarem més endavant.

3.2 Tipus moleculars del càncer colorectal

El comportament clínic del CCR depèn de múltiples interaccions a diferents nivells. Els reptes actuals són entendre la bases moleculars de cada tumor per determinar els factors que l'inicien, que condueixen la seva progressió, i que determinen la seva capacitat de resposta o resistència a determinats agents quimioteràpics. Ja fa anys que es coneix que el CCR és una malaltia complexa i heterogènia que engloba diversos fenotips tumorals. Per intentar classificar els tumors colorectals des d'un punt de vista molecular es van proposar les tres categories que es descriuen a continuació [91, 92].

Inestabilitat cromosòmica: és el tipus més comú d'inestabilitat genòmica al CCR. Provoca nombrosos canvis estructurals i de nombre de còpies als cromosomes. És un mecanisme

eficient per aconseguir la pèrdua física d'un gen supressor de tumors, com ara *APC*, *TP53* o *SMAD4*. Habitualment es coneix com CIN (de l'anglès *Chromosomal INstability*).

Defectes a la reparació del DNA: la inactivació dels gens reparadors d'errors al DNA (necessaris per a la reparació d'errors base a base) pot ser heretada, com en la síndrome de Lynch, o adquirida, com en els tumors amb silenciament epigenètic d'un gen que codifica una proteïna de reparació del DNA (gens *MLH1* i *MSH2* principalment). Mutacions a la línia germinal del gen *MSH6* atenuen la predisposició al càncer familiar. La metilació al promotor del gen *MLH1* també inactiva el mecanisme de reparació. Amb aquestes deficiències als mecanismes de reparació del DNA, alguns gens supressors de tumors, com ara *TGFBR2* i *BAX*, poden ser inactivats. Una altra ruta alternativa al CCR està relacionada amb la inactivació de la línia germinal del gen de reparació *MUTYH*. La pèrdua de la funció reparadora del DNA és fàcil de reconèixer pel fenomen associat d'inestabilitat de microsatèl·lits (MSI, de l'anglès *MicroSatellite Instability*).

Metilació aberrant del DNA: el silenciament epigenètic és un altre mecanisme d'inactivació de gens. Normalment, la metilació de la citosina es produeix a les zones de seqüències repetitives del DNA fora dels exons, però en el context del CCR es pot trobar aquesta metilació aberrant dins d'illes CpG (de l'anglès *Cytosine-phosphate-Guanine*) als promotors, per induir silenciament epigenètic de l'expressió gènica. En el CCR esporàdic amb inestabilitat de microsatèl·lits, per exemple, té lloc el silenciament epigenètic de l'expressió gènica del gen *MLH1*. Un fenomen anomenat fenotip metilador d'illes CpG (CIMP, de l'anglès *CpG Island Methylator Phenotype*) apareix en un subgrup de casos de CCR.

Cal entendre que aquestes categories proposades exhibeixen una certa superposició entre elles, com es pot veure a la Figura 8.

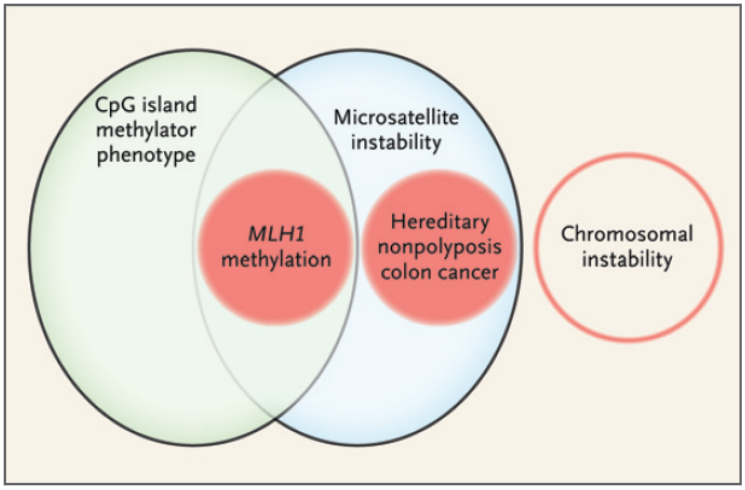

Figura 8. Relacions superposades que defineixen les principals vies de la inestabilitat genòmica en el càncer de còlon: la inestabilitat cromosòmica, la inestabilitat de microsatèl·lits i el fenotip metilador d'illes CpG [91].

Aquestes classificacions moleculars del CCR, han tingut un escàs impacte a la pràctica clínica de la malaltia. Fins ara, l'estat mutacional del gen *KRAS* ha demostrat un benefici en la recomanació del Cetuximab (inti-*EGFR*) per al CCR metastàtic [93]. També és conegut que els tumors amb inestabilitat en microsatèl·lits, encara que tenen en general un millor pronòstic, son resistents al 5-fluorouracil [94-97].

Per això, durant els últims anys s'han anat proposant diverses classificacions moleculars per al CCR basades en perfils d'expressió de l'àcid ribonucleic (RNA, de l'anglès *RiboNucleic Acid*) [98-105]. Recentment s'ha publicat un treball col·laboratiu que ha establert una classificació consens per al CCR en base a les dades genòmiques de 4.000 mostres [106]. Aquesta classificació proposa els següents quatre grups: (1) Immunes amb inestabilitat de microsatèl·lits, tumors hípermutats amb inestabilitat de microsatèl·lits i una forta activació immune; (2) Canònic, tumors epitelials amb una activació marcada de les vies de senyalització WNT i MYC; (3) Metabòlic, també epitelials però amb una desregulació metabòlica evident; i (4) Mesenquimal, tumors amb una destacada activació de TGFB2, invasió de l'estroma i angiogènesi.

És d'esperar que aquesta classificació molecular del CCR sigui útil per tractar els tumors de manera més específica, com ja és habitual a la pràctica clínica del càncer de mama [107, 108]. També amb el temps el coneixement sobre els factors de risc s'hauria d'integrar en aquesta nova classificació, doncs possiblement aquests són diferents per cada subtipus molecular.

Hipòtesi i objectius

El CCR és un greu problema de salut a nivell mundial a causa de la seva elevada incidència i mortalitat. És una malaltia complexa, amb alteracions descrites a múltiples nivells moleculars dins la cèl·lula (DNA, RNA, proteïnes, metabolisme), en la que intervenen una varietat de factors genètics i ambientals. Aquesta diversitat d'alteracions moleculars i de factors etiològics fan del CCR una malaltia altament heterogènia, en la que la caracterització detallada a nivell molecular esdevé essencial per tal d'assolir una comprensió total de la mateixa i així poder millorar els actuals tractaments. L'anàlisi de dades moleculars provinents d'experiments massius mitjançant enfocaments computacionals integratius i noves metodologies bioinformàtiques, ens pot ajudar a caracteritzar els tumors de CCR, a descriure nous mecanismes moleculars, així com a identificar nous potencials biomarcadors útils per al diagnòstic o pronòstic de la malaltia.

Objectius generals

Caracteritzar tumors colorectals mitjançant l'anàlisi computacional integrativa de perfils transcripcionals a gran escala, i identificar nous potencials biomarcadors per al diagnòstic precoç de CCR.

A continuació s'enumeren les hipòtesis i els objectius específics per a cadascun dels treballs que composen aquesta tesi.

- *Gene expression differences between colon and rectum tumor:*

  El CCR és una malaltia complexa i altament heterogènia. Actualment ja s'han començat a proposar noves classificacions per als tumors colorectals basades en criteris moleculars. En relació a la localització del tumor, els estudis en CCR habitualment inclouen els tumors de còlon i de recte de forma conjunta com una única entitat. Sobre aquest supòsit, però, hi ha una certa controvèrsia, ja que des d'un punt de vista clínic els tumors de còlon i de recte són habitualment tractats de forma diferent.

  Objectius

  1) Comparar perfils d'expressió gènica a gran escala de tumors de còlon i de recte, per tal de determinar si existeixen patrons transcripcionals diferencials que ajudin a identificar les possibles especificitats moleculars de cadascuna de les dues localitzacions tumorals.

- *Discovery and validation of new potential biomarkers for early detection of colon càncer:*

  El CCR és un greu problema de salut a causa de la seva incidència i mortalitat. El pronòstic de la malaltia depèn en gran mesura de l'estadi en el moment del diagnòstic. La detecció precoç s'ha mostrat com un mètode eficaç per reduir la mortalitat. Tot i així, les actuals tècniques de cribratge de CCR, basades en el TSOF, encara són lluny d'assolir uns nivells de sensibilitat i especificitat òptims, i tenen una acceptació poblacional excessivament baixa, que disminueix la seva efectivitat. Proves alternatives com la colonoscòpia s'associen a un cost i un risc massa elevats que dificulten la seva aplicació a nivell poblacional. En aquest sentit, la identificació de biomarcadors econòmics, més fiables i menys invasius per al diagnòstic precoç del CCR ha esdevingut una necessitat de salut pública. La detecció precisa en sèrum de proteïnes secretades específicament per cèl·lules de tumors de còlon podria servir com a prova econòmica, no invasiva i més eficaç en el context d'un cribratge poblacional de càncer de còlon.

  Objectius

  1) Identificar nous biomarcadors per al diagnòstic precoç del càncer de còlon mitjançant la integració computacional de perfils transcripcionals a gran escala de tumors i teixit sà de còlon.

  2) Validar la potencial utilitat dels biomarcadors identificats per al diagnòstic precoç del càncer de còlon en mostres de sèrum.

- *Large differences in global transcriptional regulatory programs of normal and tumor colon cells:*

  Una correcta regulació transcripcional és essencial per al funcionament de les cèl·lules. Processos biològics com el desenvolupament i la diferenciació cel·lular, que estan estretament relacionats amb els processos neoplàsics, són executats per cascades de regulació transcripcional. La identificació a nivell global de les pertorbacions reguladores que participen en l'inici i el desenvolupament dels tumors aportarà informació essencial per entendre la patologia tumoral en el context de la biologia de sistemes i identificar les rutes transcripcionals més susceptibles de ser atacades efectivament amb fàrmacs. Actualment ja s'han descrit alteracions transcripcionals específiques en el context del CCR, però es necessiten anàlisis exhaustives i a nivell global per obtenir més informació dels canvis transcripcionals implicats en el desenvolupament tumoral.

Objectius

1) Reconstruir les xarxes de regulació transcripcional de cèl·lules normals i tumorals del còlon mitjançant mètodes computacionals integratius de perfils d'expressió gènica a gran escala.

2) Identificar i caracteritzar funcionalment les alteracions presents en la xarxa de regulació transcripcional de les cèl·lules tumorals en comparació a les cèl·lules de la mucosa normal del còlon.

Materials i mètodes

1. Pacients i mostres

Les dades utilitzades en l'anàlisi combinat del primer treball de la tesi pertanyen a mostres de teixit de còlon i recte provinents de tres estudis independents. Dos d'aquests estudis provenen d'una cerca exhaustiva a repositoris de dades públiques d'Internet. Les dades del tercer provenen d'un estudi de casos i controls de pacients amb CCR dut a terme a Israel, però com es detalla a l'article, actualment aquestes dades també estan públicament disponibles.

Els perfils d'expressió gènica emprats en el segon treball d'aquesta tesi, per realitzar una cerca exhaustiva inicial de candidats a biomarcadors, així com en el tercer treball, per reconstruir les xarxes de regulació transcripcional, provenen de mostres incloses en el projecte COLONOMICS, actualment ja consolidat al nostre grup i que detallem més endavant.

1.1 Repositoris de dades públiques

Des de fa un anys, i cada cop més, les revistes científiques requereixen als autors que totes les dades subjacents a les troballes descrites en el seus manuscrits estiguin totalment disponibles (i sense restriccions) en repositoris de dades públiques. Només existeixen unes rares excepcions relacionades amb aspectes ètics o legals.

Per dur a terme aquesta tasca es recomana l'ús d'algun dels múltiples repositoris de dades públiques existents i accessibles a través d'Internet, que ajuden a complir els estàndards per a la preparació, adequació i emmagatzematge de les dades de cada camp en particular. Pel que fa a les dades genòmiques, com poden ser les dades de *microarrays*, de seqüenciació o dades provinents d'altres tipus d'experiments funcionals, existeixen diversos repositoris disponibles, com per exemple el GEO [109] o l'ArrayExpress [110]. Pel que fa a dades de seqüencies gèniques hi ha disponibles, per exemple el GenBank, l'EMBL, el DDBJ i la miRBase, entre d' altres. Per dades de polimorfismes i variació genòmica estructural disposem del dbSNP i el dbVar, entre d'altres. A més, existeixen molts altres repositoris addicionals per dades d'altres camps.

1.2 El projecte COLONOMICS

El nostre grup té en marxa el projecte COLONOMICS (www.colonomics.org) en el qual es disposa d'informació molecular exhaustiva de 100 casos de càncer de còlon. Concretament aquests pacients van ser diagnosticats com a casos incidents d'adenocarcinoma de còlon i van ser atesos a l'Hospital Universitari de Bellvitge de Barcelona, entre gener de 1996 i desembre de 2007. Els pacients van ser seleccionats definint una sèrie homogènia, tots

amb càncer de còlon esporàdic, estadi II, tumors estables en microsatèl·lits, que no van rebre quimioteràpia prèvia a la cirurgia radical i amb un seguiment d'un mínim de 3 anys respecte l'inici del projecte. Per tots aquests casos es van analitzar les mostres aparellades normal-tumor, és a dir, a més de la mostra de tumor per cadascun d'ells es va analitzar també una mostra de la mucosa adjacent patològicament normal. Aquesta es va obtenir sempre a una distancia mínima de 10 cm del tumor. Addicionalment, es van incloure 50 mucoses de còlon d'individus sense càncer ni lesions precursores, que van ser obtingudes entre febrer i maig de 2010. Aquestes mostres provenen d'una sèrie d'individus que es van sotmetre a colonoscòpies indicades per altres causes, com poden ser anèmia, sagnat, dolor o alteracions del ritme gastrointestinal. Els individus amb historia familiar de CCR reportada, van ser exclosos del projecte. Així doncs, es disposa d'un total de 250 mostres amb una àmplia caracterització molecular i epidemiològica.

Les mostres, moments després de la cirurgia, van ser examinades per anatomia patològica i directament conservades a -80ºC al biobanc del hospital. Posteriorment, mitjançant tinció d'hematoxilina-eosina, es va confirmar que les mostres de tumor tinguessin com a mínim un 75% de cèl·lules tumorals i que les mostres normals no continguessin cèl·lules tumorals. Per cadascuna de les 250 mostres es va extreure el DNA i RNA. Abans de realitzar qualsevol experiment es van aplicar estrictes processos de control de qualitat per evitar possibles errors durant la manipulació i alhora garantir la qualitat de la mostra.

Finalment, pel que fa al RNA, es disposen de dades l'expressió dels RNA missatgers (mRNA, de l'anglès *messenger RNA*) (Human Genome U219 Array Plate d'Affymetrix) i de dades d'expressió de miRNA obtingudes mitjançant la seqüenciació massiva dels RNAs petits (Solid 4 System d'Applied Biosystems), prèviament seleccionats per la seva grandària, compresa entre 18 i 40 nucleòtids. D'altra banda, pel que fa al DNA, es disposa de les dades de variació gènica referents a SNPs i CNVs (Genome-Wide Human SNP Array 6.0 d'Affymetrix), així com les dades dels perfils de metilació d'illes CpG per tot el genoma (Infinium Human Methylation 450K BeadChip d'Illumina). També s'han analitzat els perfils mutacionals de gens candidats i més habitualment implicats en l'etiologia del CCR, mitjançant sondes especifiques i PCRs de forma multiplexada (Biomark HD System de Fluidigm). Addicionalment, per obtenir una informació més detallada sobre l'estat mutacional, es va realitzar la seqüenciació completa de l'exoma d'un total de 84 mostres. Aquestes 84 mostres corresponen a un total de 42 tumors (21 de bon pronòstic i 21 de mal pronòstic, és a dir, corresponents a pacients que van desenvolupar metàstasi) i 42 mostres de teixit patològicament normal aparellat (Figura 9).

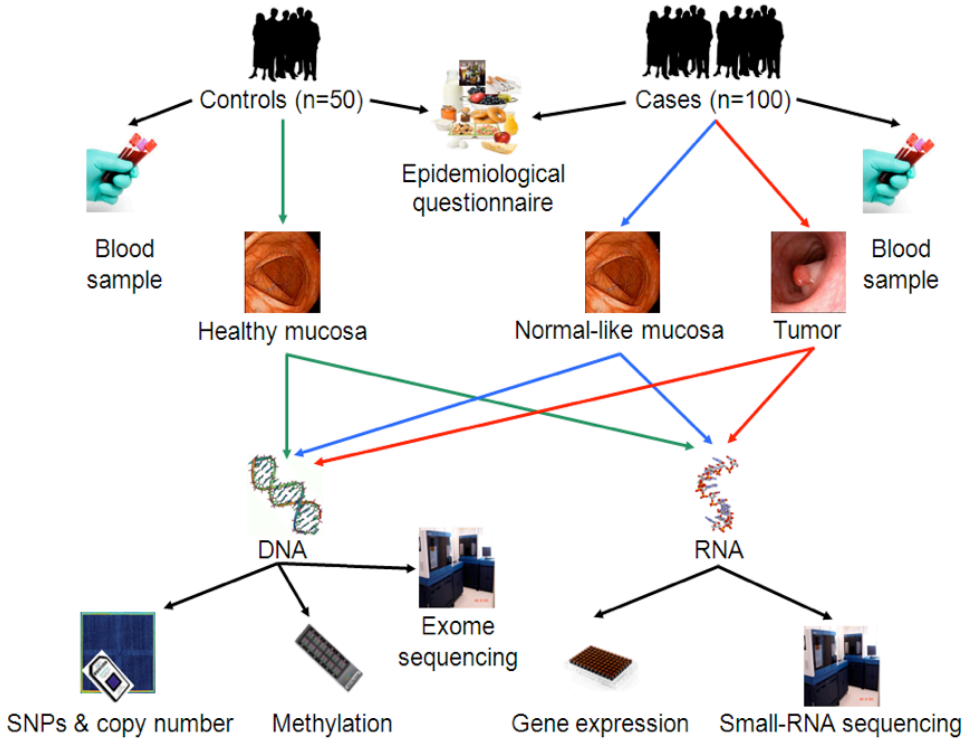

Figura 9. Esquema del disseny experimental del projecte COLONOMICS.

L'objectiu general del projecte COLONOMICS és identificar biomarcadors útils per al diagnòstic i pronòstic del CCR. Més específicament, el que pretén el projecte a mig termini és millorar les proves diagnòstiques per a la detecció del CCR utilitzant biomarcadors que complementin el TSOF. Alhora també poder proporcionar biomarcadors útils per predir la recurrència del CCR en estadi II, degut a que aquests pacients podrien beneficiar-se de tractament amb quimioteràpia adjuvant. Addicionalment, com a objectius a més llarg termini, es pretén que mitjançant l'anàlisi de les dades complexes provinents de la quantitat de mostres del projecte i dels diferents experiments massius realitzats, en combinació amb fonts externes i utilitzant tècniques de bioinformàtica i biologia de sistemes, aquest projecte permeti comprendre millor els mecanismes moleculars subjacents del CCR.

De forma colateral, i com a objectiu intermedi, el projecte té la necessitat de construir una plataforma de validació de biomarcadors amb mostres biològiques d'alta qualitat i amb la seva informació clínica completa. Des de l'inici del projecte s'han estat recollint un gran nombre mostres de casos de CCR, adenomes, mostres de pòlips i controls sans. De gran part de les mostres recollides ja s'han realitzat extraccions de DNA i RNA, sempre preservant els RNAs petits per possibles futures anàlisis de miRNAs. Així doncs el projecte

ha permès construir una plataforma molt completa per a la validació dels múltiples biomarcadors candidats a les diferents fases i estudis del projecte.

Un punt fort del projecte és que les anàlisis massives realitzades a diferents nivells (DNA, RNA, etc.) s'han realitzat sempre sobre les mateixes mostres dels mateixos individus. D'aquesta manera tota aquesta informació es pot relacionar i analitzar de forma conjunta mitjançant mètodes d'anàlisi integratius. Un altre aspecte interessant és que es disposa de mostres del teixit normal adjacent al tumor per cada cas. Aquest teixit aparentment sà ens proporciona una visió de la variació genètica intrínseca dels individus i del comportament de la mucosa patològicament sana de cadascun d'ells. Per últim, un altre aspecte molt positiu del projecte és la disponibilitat d'un conjunt de mostres de mucosa d'individus sans amb dades a tots el nivells, ja que pocs projectes disposen d'aquesta informació tan valuosa i que pot ser clau en l'estudi dels mecanismes moleculars de la malaltia. D'altra banda, una possible limitació del projecte i pensant de forma conjunta en tot el procés de progressió tumoral del CCR, seria el fet de disposar només de casos d'estadi II a la sèrie inicial de mostres.

Els aspectes ètics del projecte es corresponen bàsicament a l'ús de mostres humanes. Els estudis realitzats en aquesta tesi i que estan emmarcats dins del projecte COLONOMICS, es van realitzar sota les directrius i normes ètiques corresponents. Concretament, el comitè ètic d'investigació clínica de l'Hospital Universitari de Bellvitge va aprovar el protocol de l'estudi (PR178/11) . A més, es disposa de consentiment informat escrit de tots els individus que van participar al projecte, per les anàlisis genètiques que s'han realitzat amb les seves mostres.

2. Estratègies computacionals d'anàlisi molecular

Durant els últims anys, en primer lloc amb la revolució que va provocar l'era dels *microarrays* i ara més recentment amb la reducció dels costos econòmics de les tecnologies de seqüenciació massives, l'ús dels experiments a gran escala que generen enormes quantitats de dades, s'ha introduït com una rutina a la pràctica diària dins el camp de la recerca biomèdica.

Actualment és possible avaluar l'expressió gènica, els polimorfismes, la variació estructural o la metilació, entre d'altres, mitjançant experiments econòmics i ràpids de *microarrays*. També, sorprenentment es pot seqüenciar un genoma humà complet en poc temps, a un preu que s'ha reduït dràsticament durant els últims anys i que ara ja és per sota dels 10.000$, com es pot veure a la Figura 10.

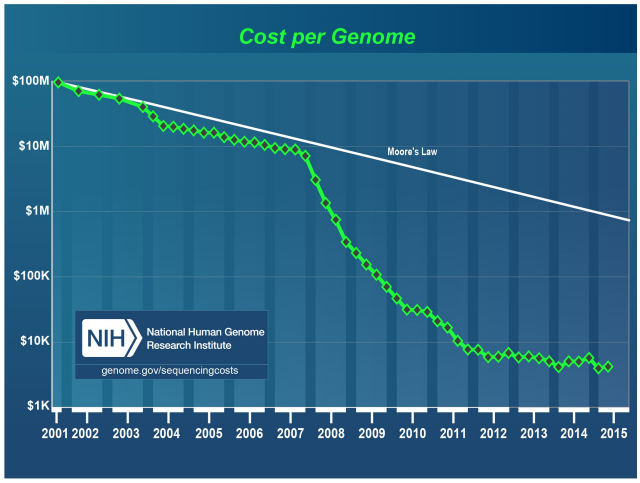

Figura 10. Reducció de costos de la seqüenciació del DNA durant els últims anys, en comparació amb la reducció teòrica que hauria seguit mitjançant la llei de Moore (augment de capacitat i reducció de costos dels computadors).

El principal coll d'ampolla de les noves tecnologies d'anàlisi molecular a gran escala, a banda dels costos econòmics, ha estat el volum i la complexitat del tractament de les dades que generaven. Aquest fet ha requerit generar noves eines bioinformàtiques i l'adaptació de les tècniques d'estadística clàssica. La bioinformàtica, mitjançant enfocaments computacionals integratius, té com a finalitat indagar sobre qüestions biològiques gestionant i processant la gran quantitat de dades genòmiques generades pels experiments a gran escala. Gràcies a aquesta disciplina es pot aprofundir en qüestions complexes, com per exemple la identificació de nous mecanismes moleculars [111]. D'altra banda, la bioestadística aplica noves metodologies estadístiques a una àmplia gamma de qüestions biològiques i participa en la millora dels mètodes de recollida de les dades, disseny d'experiments, anàlisi, interpretació i inferència dels resultats [112].

2.1 Tècniques d'anàlisi d'expressió diferencial

Les tècniques actuals d'anàlisi molecular permeten avaluar de forma massiva un gran nombre de marcadors alhora. Per exemple, els *microarrays* d'expressió gènica quantifiquen en un únic experiment, el valor de pràcticament tots els gens coneguts, com es pot veure a la Figura 11. Seguidament, el repte per a l'investigador és identificar el conjunt de gens rellevants per al seu estudi en particular.

Figura 11. Esquema representatiu del procés d'hibridació i anàlisi de *microarrays* d'expressió gènica.

Després de la hibridació i obtenció de les dades crues, és convenient realitzar un procés de control de qualitat, tot seguint les principals recomanacions del fabricant dels *microarrays* corresponents, per així conèixer la qualitat de les dades generades. Principalment es tracta d'avaluar la qualitat a cadascun del *microarrays*, mitjançant certs aspectes tècnics com la intensitat mitjana o la intensitat de fons i també d'aspectes biològicament relacionats amb la pròpia mostra, com poden ser els nivells de degradació, concentració, etc. Quan alguna mostra no assoleix els nivells mínims de qualitat recomanada, és eliminada de les anàlisis subseqüents.

Una vegada es disposa del conjunt final de mostres que participaran a l'estudi, és necessari realitzar un pas de normalització per igualar les distribucions d'intensitat de les diferents mostres, així es fan comparables entre si evitant possibles biaixos. Per aquesta tasca existeixen diferents mètodes computacionals, publicats prèviament. Per totes les normalitzacions de dades de *microarrays* realitzades durant les anàlisis dels diferents treballs d'aquesta tesi, s'ha utilitzat l'algorisme RMA [113]. Després de la normalització també és poden realitzar alguns passos de control de qualitat addicionals.

A continuació, i amb les mostres que han passat el control de qualitat i han estat normalitzades, es pot treballar en la identificació de gens diferencialment expressats estadísticament significatius. Per això, també existeixen múltiples estratègies, la majoria de les quals estan relacionades amb proves que tenen en compte tant la magnitud del canvi com la variabilitat observada. Cal tenir en compte que degut al gran nombre de tests estadístics independents realitzats, pot aparèixer per atzar, un elevat nombre de falsos positius ocults entre els resultats. Així doncs cal utilitzar un mètode per contrarestar aquest problema de comparacions múltiples. La correcció de Bonferroni és un mètode àmpliament utilitzat però alhora molt restrictiu. Un altra opció també molt utilitzada i alhora menys conservadora, és ajustar la taxa de falsos descobriments (FDR, de l'anglès *False Discovery Rate*) mitjançant algun dels múltiples mètodes prèviament publicats, com per exemple el proposat ja fa anys per Benjamini i Hochberg [114].

2.2 Tècniques d'inferència de xarxes transcripcionals

Una metodologia clàssica per estudiar els sistemes complexos és la modelització mitjançant xarxes amb relacions entre els seus elements i així poder estudiar les seves propietats. Aquestes metodologies es van desenvolupar inicialment per investigadors d'altres àrees com la física i les matemàtiques, però durant l'última dècada el seu ús s'ha anant incorporant a noves àrees del nostre interès com poden ser la biologia de sistemes o la bioinformàtica [115-117]. Per exemple, una de les representacions en xarxa típicament utilitzada és aquella en la que els gens constitueixen els nodes i que poden estar o no, connectats amb altres gens. Les connexions o arestes es defineixen segons propietats que poden compartir els gens. Són típiques les xarxes basades en coexpressió gènica mesurada amb *microarrays*, però també es poden representar altres propietats com per exemple la unió d'una proteïna a una regió reguladora del DNA o la interacció física entre dues proteïnes, entre d'altres. Una de les complexitats d'aquestes metodologies consisteix en identificar exactament quina propietat o propietats són les que es volen representar relacionalment a les arestes de la xarxa. Normalment els nodes que no estan connectats amb cap altre node de la xarxa, no apareixen a la representació final.

A partir de l'ús de dades d'expressió gènica s'han descrit diversos algorismes per intentar identificar xarxes de regulació. Un tipus d'enfocament són els basats en xarxes bayesianes [118-120] que normalment treballen amb grafs acíclics i assumeixen direcció en les relacions. D'altra banda existeixen tota una sèrie de metodologies que utilitzen en una major mesura la teoria general de grafs, fins i tot poden treballar amb grafs cíclics i en la majoria d'ocasions sense assumir direcció en les arestes. Un exemple conegut és l'algorisme ARACNe [121-123], que ha estat l'utilitzat en el tercer treball d'aquesta tesi. En definitiva, aquests mètodes normalment el que pretenen és extreure la informació de les matrius de covariància a partir dels nivells d'expressió dels diferents gens analitzats en un experiment de *microarrays* i mitjançant mètodes computacionals estimar les relacions entre els diferents gens. Posteriorment aquestes relacions es poden visualitzar en grafs amb estructura de xarxa i es poden calcular diverses mesures topològiques, identificar "hubs" (nodes altament connectats), calcular distàncies entre nodes, identificar clústers, etc. També es poden realitzar experiments de simulació per identificar quins nodes són més rellevants per mantenir l'estructura de la xarxa, degut a que aquests nodes podrien ser potencials candidats a dianes terapèutiques. Alhora, les pròpies característiques d'aquestes xarxes ens poden ajudar a classificar els tumors o les seves metàstasis [120] o a estudiar propietats que modifiquen l'expressió de gens [124].

Mitjançant aquestes noves metodologies computacionals es poden adreçar múltiples problemes i qüestions biològiques diferents. Per exemple, la comparació de xarxes definides a partir de diversos fenotips ens pot permetre identificar quins gens són més rellevants entre aquests. Malgrat tot, la teoria i algorismes necessaris per realitzar totes

aquestes comparacions és un tema encara en investigació i desenvolupament. D'altra banda, com que els algorismes d'inferència de xarxes transcripcionals actualment existents, no estan exempts d'un percentatge d'error, es continua treballant a la seva millora. En aquest sentit, s'ha desenvolupat recentment una extensió de l'algorisme ARACNe (hARACNe) específicament dissenyada per tractar les interaccions d'ordre múltiple. Segons els propis autors del mètode, aquesta nova extensió del algorisme millora la qualitat i a la robustesa de la xarxa inferida [125].

Resultats

Els resultats d'aquesta tesi es presenten com a compendi de tres articles científics. Tots ells han estat revisats i publicats en revistes científiques internacionals. A continuació es llisten les referències d'aquests tres articles i tot seguit s'adjunta, per cadascun d'ells, un resum en català i el text complet en anglès en el format final de la revista on van ser publicats.

Article 1:

- Sanz-Pamplona R*, **Cordero D***, Berenguer A, Lejbkowicz F, Rennert H, Salazar R, Biondo S, Sanjuan X, Pujana MA, Rozek L, Giordano TJ, Ben-Izhak O, Cohen HI, Trougouboff P, Bejhar J, Sova Y, Rennert G, Gruber SB, Moreno V. **Gene expression differences between colon and rectum tumors.** Clinical Cancer Research, 17(23):7303-12, 2011.
  **Aquests autors han contribuït igualment en aquest treball.*
  Factor d'impacte de la revista (2011): 8,722
  Quartil de la revista (2011): Q1 en oncologia

Article 2:

- Solé X*, Crous-Bou M*, **Cordero D***, Olivares D, Guinó E, Sanz-Pamplona R, Rodriguez-Moranta F, Sanjuan X, de Oca J, Salazar R, Moreno V. **Discovery and validation of new potential biomarkers for early detection of colon cancer.** PLoS One, 9(9):e106748, 2014.
  **Aquests autors han contribuït igualment en aquest treball.*
  Factor d'impacte de la revista (2014): 3,234
  Quartil de la revista (2014): Q1 en ciències multidisciplinars

Article 3:

- **Cordero D***, Solé X*, Crous-Bou M, Sanz-Pamplona R, Paré-Brunet L, Guinó E, Olivares D, Berenguer A, Santos C, Salazar R, Biondo S, Moreno V. **Large differences in global transcriptional regulatory programs of normal and tumor colon cells.** BMC Cancer, 14(1):708, 2014.
  **Aquests autors han contribuït igualment en aquest treball.*
  Factor d'impacte de la revista (2014): 3,362
  Quartil de la revista (2014): Q2 en oncologia

1. Article 1: *Gene expression differences between colon and rectum tumors.*

1.1 Resum en català

El CCR està considerat com una malaltia complexa i altament heterogènia, que engloba diversos fenotips tumorals. Intentant classificar els tumors colorectals des d'un punt de vista molecular es van proposar tres categories, amb una certa superposició entre elles: els tumors amb inestabilitat cromosòmica, el tumors inestables en micro-satèl·lits i els tumors amb fenotip metilador d'illes CpG. Respecte a la localització del tumor, els estudis en CCR típicament inclouen els tumors de còlon i recte de forma conjunta com una única entitat. Sobre aquest supòsit hi ha una certa controvèrsia, ja que des d'un punt de vista clínic els tumors de còlon i de recte són tractats de forma diferent. De fet, existeixen indicis que suggereixen que agrupar aquestes dues entitats anatòmiques pot resultar en una simplificació excessiva. Per exemple, els càncers rectals mostren una taxa més elevada de recidiva local i metàstasis pulmonars, mentre que els càncers de còlon tenen més tendència a una disseminació hepàtica. Però malgrat tot això, certament encara s'han reportat poques diferències a nivell molecular i epidemiològic entre els tumors de còlon i els de recte.

L'objectiu d'aquest treball va ser comparar els perfils d'expressió gènica entre els tumors de còlon i els de recte, per identificar el grau de similitud que existeix a nivell transcriptòmic entre les diferents localitzacions dels tumors colorectals.

Es va realitzar una anàlisi combinada amb dades de *microarrays* de 460 tumors de còlon i de 100 tumors de recte, provinents de tres estudis independents. Dos d'aquests estudis provenen d'una cerca exhaustiva a repositoris de dades públiques d'Internet, i les dades del tercer provenen d'un estudi de casos i controls de pacients amb CCR dut a terme a Israel. Es va tenir una cura especial amb els criteris d'inclusió, els processos de control de qualitat, la normalització i el control de la heterogeneïtat entre els estudis. Es pretenia que l'ús combinat de diferents conjunts de dades no afectés de cap manera als resultats finals. Els tumors inestables en microsatèl·lits van ser exclosos, ja que es coneix que tenen un perfil d'expressió diferent i una ubicació preferencial al còlon proximal. Les diferències d'expressió gènica van ser avaluades mitjançant models lineals ajustats per edat, sexe i estudi.

Com a resultat, es van trobar només 11 gens diferencialment expressats entre els tumors de còlon i els tumors de recte, després d'ajustar per comparacions múltiples aplicant la correcció de Bonferroni. També ajustant per comparacions múltiples, però mitjançant el mètode menys conservatiu dels q-valors, vam identificar 16 gens amb un FDR del 1% (Figura 12) i 111 gens si acceptem fins un FDR del 5%. A més a més, per a tots els gens

trobats com diferencialment expressats, es va comprovar que les magnituds del canvi eren petites (sempre inferiors a una unitat en escala logarítmica en base 2).

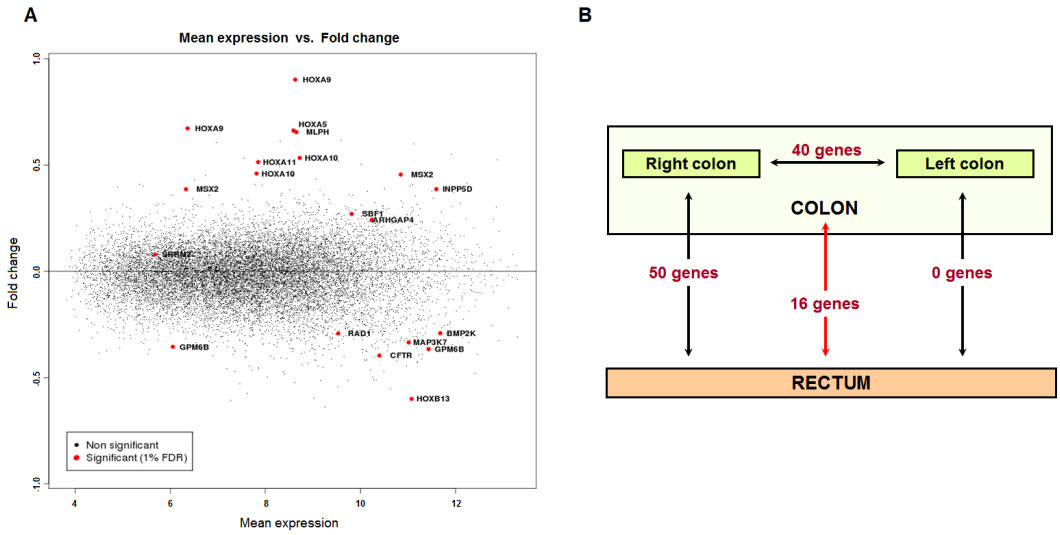

Figura 12. Mitjana d'expressió de cada sonda davant del seu valor de canvi entre els tumors de còlon i els de recte (A). Nombre de gens diferencialment expressats entre cada localització tumoral amb un FDR del 1% (B).

Cal destacar que entre els sis gens més diferencialment expressats, cinc d'ells pertanyen a la família de factors de transcripció (TF, de l'anglès *Transcription Factor*) *HOX*. També entre els gens més diferencialment expressats apareixen funcions com reparació del DNA, activitat de TFs, transport intracel·lular, transducció de senyals i apoptosi. Per identificar si en els gens diferencialment expressats entre els tumors de còlon i els tumors de recte hi estava sobrerepresentada alguna signatura molecular ja coneguda o anteriorment descrita, es va realitzar una anàlisi d'enriquiment mitjançant el mètode computacional GSEA. No vam trobar cap conjunt de gens conegut significativament sobrerepresentat, però el conjunt de gens *HOX* va aparèixer amb la puntuació més alta entre els conjunts de gens d'enriquiment.

Addicionalment es va fer una anàlisi més detallada entre els tumors del còlon dret, els situats al còlon esquerre i els de recte. Els resultats van indicar que les majors diferències es troben entre les localitzacions més oposades (Figura 12B). Però de forma similar als resultats previs, les diferències que es van trobar entre els tumors del còlon dret i els del còlon esquerre eren molt petites. De fet, es van trobar més diferències entre els extrems proximal i distal del còlon que entre el còlon distal i el recte, indicant que no hi ha diferències moleculars suficients que *a priori* justifiqui tractar de forma separada els perfils d'expressió gènica dels tumors de còlon dels de recte.

Com a conclusions d'aquest treball podem afirmar que els tumors colorectals estables en microsatèl·lits exhibeixen perfils d'expressió gènica molt similars, independentment de la seva localització. I que les petites però consistents diferències observades entre els tumors situats al còlon dret, al esquerre i al recte, són impulsades en gran mesura pels gens *HOX*. Aquests resultats trobats poden tenir importants implicacions en el disseny i la interpretació dels futurs estudis de CCR.

Published OnlineFirst October 5, 2011; DOI: 10.1158/1078-0432.CCR-11-1570

Human Cancer Biology

Clinical Cancer Research

Gene Expression Differences between Colon and Rectum Tumors

Rebeca Sanz-Pamplona[1], David Cordero[1], Antonio Berenguer[1], Flavio Lejbkowicz[6], Hedy Rennert[6], Ramon Salazar[2], Sebastiano Biondo[3,5], Xavier Sanjuan[4], Miguel A. Pujana[1], Laura Rozek[12], Thomas J. Giordano[13], Ofer Ben-Izhak[7], Hector I. Cohen[8], Philip Trougouboff[10], Jacob Bejhar[11], Yanina Sova[9], Gad Rennert[6], Stephen B. Gruber[14], and Victor Moreno[1,5]

Abstract

Purpose: Colorectal cancer studies typically include both colon and rectum tumors as a common entity, though this assumption is controversial and only minor differences have been reported at the molecular and epidemiologic level. We conducted a molecular study based on gene expression data of tumors from colon and rectum to assess the degree of similarity between these cancer sites at transcriptomic level.

Experimental Design: A pooled analysis of 460 colon tumors and 100 rectum tumors from four data sets belonging to three independent studies was conducted. Microsatellite instable tumors were excluded as these are known to have a different expression profile and have a preferential proximal colon location. Expression differences were assessed with linear models, and significant genes were identified using adjustment for multiple comparisons.

Results: Minor differences at a gene expression level were found between tumors arising in the proximal colon, distal colon, or rectum. Only several *HOX* genes were found to be associated with tumor location. More differences were found between proximal and distal colon than between distal colon and rectum.

Conclusions: Microsatellite stable colorectal cancers do not show major transcriptomic differences for tumors arising in the colon or rectum. The small but consistent differences observed are largely driven by the *HOX* genes. These results may have important implications in the design and interpretation of studies in colorectal cancer. *Clin Cancer Res; 17(23); 7303–12. ©2011 AACR.*

Introduction

Colorectal cancer (CRC) is considered a heterogeneous complex disease that comprises different tumor phenotypes (1). Attempts to classify tumors from a molecular perspective that identify carcinogenic pathways have proposed 3 categories with some overlap as follows: chromosomal instability (CIN) tumors, microsatellite instability (MSI) tumors, and CpG island methylator phenotype (CIMP) tumors. This taxonomy plays a significant role in determining clinical, pathologic, and biological characteristics of CRC (2).

From a clinical point of view, the colon and rectal cancers are treated as distinct entities. Colon tumors are usually divided as proximal or right sided when originating proximal to the splenic flexure (cecum, ascending colon, and transverse colon) whereas distal tumors arise distal to this site (descending colon and sigmoid colon). Distal colon or left-sided tumors most often appear in the rectum sigmoid flexure, and the distinction of these from rectal tumors is not always easy. Usually, a tumor is considered rectal when arising within 15 cm from the anal sphincter (3, 4). Indeed, accumulating evidences suggest that grouping these anatomically distinct diseases could be a clinical and biological oversimplification: rectal cancers show higher rates of locoregional relapse and lung metastases, whereas colon cancers have a higher tropism for liver spread and a slightly better overall prognosis (5). Moreover, proximal location of colon cancer is a risk factor for development of metachronous CRC (6). Treatment also differs for colon and rectal tumors. Although both colon and rectal cancers benefit from adjuvant chemotherapy, radiation therapy is

Authors' Affiliations: [1]Unit of Biomarkers and Susceptibility and [2]Medical Oncology Service, Catalan Institute of Oncology (ICO), IDIBELL, and CIBERESP, Departments of [3]Digestive Surgery and [4]Pathology, University Hospital Bellvitge, L'Hospitalet de Llobregat; [5]Department of Medical Sciences, School of Medicine, University of Barcelona, Barcelona, Spain; [6]Clalit Health Services National Cancer Control Center and Department of Community Medicine and Epidemiology, Carmel Medical Center and B. Rappaport Faculty of Medicine, Technion-Israel Institute of Technology; [7]Rambam Medical Center; [8]Bnai Zion Medical Center; [9]Carmel Medical Center, Haifa; [10]Western Galilee Medical Center, Nahariya; and [11]Haemek Medical Center, Afula, Israel; and Departments of [12]Environmental Health Sciences, [13]Pathology, and [14]Internal Medicine, Epidemiology, Human Genetics, University of Michigan, Ann Arbor, Michigan

Note: Supplementary data for this article are available at Clinical Cancer Research Online (http://clincancerres.aacrjournals.org/).

R. Sanz-Pamplona and D. Cordero contributed equally to this work.

Corresponding Author: Victor Moreno, Catalan Institute of Oncology, Av. Gran Via 199, L'Hospitalet de Llobregat, Barcelona 08908, Spain. Phone: 34-93-260-71-86; Fax: 34-93-260-71-88; E-mail: v.moreno@iconcologia.net

doi: 10.1158/1078-0432.CCR-11-1570

Published OnlineFirst October 5, 2011; DOI: 10.1158/1078-0432.CCR-11-1570

Translational Relevance

Colorectal cancer (CRC) studies typically include both colon and rectum tumors as a common entity, though this assumption is controversial and only minor differences have been reported at the molecular and epidemiologic level. Here, we report a large sample pool study concluding that only minor differences at a gene expression level exist between microsatellite stable CRCs at different locations. These results have important implications in the design and interpretation of studies in colorectal cancer. For instance, several molecular profiles have been recently proposed to predict prognosis in patients with CRC that combine colon and rectum cases, assuming this hypothesis without the real proof. The conclusions provided by this study will help consolidate the idea that at the molecular level, the minor expression differences identified are more related to anatomic developmental differences than to tumoral mechanisms.

only indicated in locally advanced rectal tumors (7). Epidemiologic risk factors reflect somewhat more controversial distinctions between cancers of the colon and rectum: alcohol intake was significantly positively associated with higher risk in the rectum than in colon tumors (8). Other dietary risk factors differing between colon and rectum tumors have been suggested more inconsistently (9, 10).

At the molecular level, differences in expression of specific genes and proteins (cyclin A2, COX-2, and β-catenin) have been reported (reviewed in ref. 6). Moreover, colon cancers have a higher number of mutations including *KRAS* and *BRAF* mutations. The CIN pathway is far more common in rectal cancers than colon cancers, whereas MSI and CIMP cancers are more likely to be in the right colon. Some of the reported differences in gene expression probably correspond to molecular signatures of MSI, such as the correlation between *CDX2* expression and MSI (11).

Recently, several molecular profiles have been proposed to predict prognosis in patients with CRC (12–15). These studies typically combine colon and rectal cancers, but it is not known whether this combination is appropriate. Expression profiles may inform this choice. If proximal colon, distal colon, and rectal tumors share a common set of expressed transcripts, then it may be reasonable to combine data for prognostic studies, and in fact may inform choices for epidemiologic study designs. The aim of this work was to compare gene expression among CRC subsites in an attempt to identify molecular factors that correspond to differences in the clinical behavior of these tumors.

Materials and Methods

Study population

The Molecular Epidemiology of Colorectal Cancer (MECC) study is a population-based, case–control study that included 2,138 incident CRC cases and 2,049 population controls from Northern Israel (16). A pathology review of the diagnostic slides centralized at the University of Michigan (Ann Arbor, MI) confirmed the eligibility criteria of invasive adenocarcinoma. The study was approved by the Institutional Review Boards at the University of Michigan and Carmel Medical Center in Haifa. Written and informed consent was required for inclusion.

A subset of these patients provided fresh tumor tissue samples that were analyzed for expression in 2 stages as previously described (17). Initially, a subset of 170 tumors was hybridized with the Affymetrix HG-U133A gene array (MECC-A). In a second stage, an additional sample of 232 tumors was hybridized in the HG-U133plus 2.0 gene array (MECC-P2). Of these patients, 4 from the first set and 7 from the second were excluded because they had multiple tumors in the colon and rectum, or the precise location was not provided. Expression data are available in Gene Expression Omnibus (GEO; ref. 18) repository with accession code GSE26682.

In addition of these 2 gene expression data sets (MECC-A and MECC-P2), publicly available expression data with information about subsite were searched in the GEO and ArrayExpress (19) databases. To guarantee a high quality analysis, the inclusion criteria was restricted to studies that had used Affymetrix U133 gene chips, with more than 50 samples, and a minimum number of 10 for each site. Two data sets were identified matching these criteria: GSE14333 included 290 consecutive patients with CRC [colon (n = 250), rectum (n = 39), and missing site (n = 1); ref. 20]. GSE13294 comprised 155 patients with CRC (122 colon, 25 rectum, and 8 missing; ref. 21). In addition, data set GSE9254 was identified, that included 19 normal mucosa samples from different colonic locations: cecum (n = 2), ascending (n = 3), transverse (n = 3), sigmoid (n = 4), and rectum (n = 7; ref. 22).

Quality control and normalization

Prior to data analysis, a careful quality control process following the Affymetrix recommendations was conducted (23). This procedure rejected 122 samples: 27 (16%) from MECC-A, 49 (21%) from MECC-P2, 21 (7%) from GSE14333, and 25 (16%) from GSE13294.

Data normalization were carried out with the *R* statistical software, version 2.9.0 (*R* foundation for statistical computing; http://www.r-project.org) and Bioconductor package (Bioconductor core group; http://www.bioconductor.org). Raw data from the different data sets were normalized together with the Robust Multiarray Average (RMA) method (24). To improve comparability between arrays from different studies, only the common subset of probes from the U133A array (n = 22,283) were selected, and data were renormalized with a quantile method.

MSI

Tumors showing MSI appear more often in right colon and are known to have a marked different expression profile

Published OnlineFirst October 5, 2011; DOI: 10.1158/1078-0432.CCR-11-1570

(25). In an attempt to homogenize the analysis and avoid potential biases due to this condition, MSI tumors were excluded from all data sets. For MECC cases, MSI was analyzed using 7 microsatellite markers that included the National Cancer Institute panel (26). Cases were considered MSI when more than 30% of the markers were instable. A total of 16 cases were excluded from MECC-A and 15 from MECC-P2. A total of 61 MSI samples from data set GSE1324 were also excluded.

MSI status was not available for the public GSE14333 data set but was imputed using a molecular profiling–based approach (details in Supplementary Table S1 and Fig. S1). Out of the 268 samples, 53 (20%) were labeled as MSI and removed for further analysis. These excluded cases might not be a perfect selection of the real subset of MSI tumors, but their clinical characteristics are in agreement with the expectations: more frequent in female and older patients and with preferential location in right colon (Supplementary Table S2).

Differential expression analysis

Prior to the identification of differentially expressed probes, a filter was applied to remove those with low variability ($n = 7{,}509$), which mostly correspond to nonhybridized and saturated probes. The remaining 14,774 probes with SD greater than 0.3 were considered for further analysis. To test for differences in expression between sites, a linear model adjusted for gender, age, and study was fitted to each probe. To account for multiple comparisons, the Bonferroni correction was used. Also the less conservative q value method was used to control the false discovery rate (FDR).

Heterogeneity of expression profiles by tumor site across studies was evaluated for each probe by the linear models described earlier. A test for interaction between cancer site and study was conducted for each probe and, again, the q value method was used to correct the results by multiple comparisons.

Gene set enrichment analysis

The gene set enrichment analysis (GSEA) algorithm (27) was applied to identify enrichment of specific functions in the list of genes preranked according to their P value for the test of differences in expression between subsites. The statistical significance of the enrichment score was calculated by permuting the genes 1,000 times as implemented in the GSEA software.

Classification of colon/rectum samples using differentially expressed genes

For each comparison considered, an agglomerative hierarchical clustering method was used to display the classification ability among site of the corresponding list of differentially expressed probes sets. This discriminating ability was formally tested using a linear discriminant analysis with leave-one-out crossvalidation to estimate the prediction error rate.

Results

Clinical data for the 460 colon tumors and 100 rectum tumors included in the analysis are summarized in Table 1. A principal component analysis (PCA) was done to assess global differences between each data set. The first and second components separated the samples by study, suggesting systematic differences that could not be corrected by careful homogeneous criteria and normalization (Supplementary Fig. S2). The most dissimilar data set was MECC-A, probably due to be the fact that the platform was Affymetrix H-U133 A gene chips instead of H-U133 Plus 2.0 used in the other studies. All pooled analyses were adjusted for study to account for these systematic differences.

Gene expression profiling: colon versus rectum tumors

Linear models adjusted for study, age, and gender identified only 11 of 14,774 differentially expressed probes between colon and rectum after Bonferroni correction. The less conservative q value method identified 20 probes (corresponding to 16 genes, Table 2) when a 1% FDR was used and 131 probes (111 genes) at the 5% FDR. Moreover, among these differentially expressed genes, no one had an absolute $\log_2$ fold change greater than one (Fig. 1A). These results suggest that the magnitude of expression differences

Table 1. Samples description

$n = 560$	Site[a]			Platform	Mean age	Gender[b]		Stage[b]			
	Right	Left	Rectum			Male	Female	I	II	III	IV
MECC-A ($n = 123$)	55 (44.7%)	57 (46.4%)	11 (8.9%)	Affy HG-U133A	72.53	68 (55.3%)	55 (44.7%)	4 (3.4%)	58 (50%)	41 (35.4%)	13 (11.2%)
MECC-P2 ($n = 161$)	58 (36.9%)	59 (37.6%)	40 (25.5%)	Affy U133 Plus 2.0	72.01	87 (54%)	74 (46%)	20 (15.4%)	55 (42.3%)	39 (30%)	16 (12.3%)
GSE14333 ($n = 215$)	79 (37.1%)	100 (46.9%)	34 (16%)	Affy U133 Plus 2.0	65.65	132 (61.4%)	83 (38.6%)	34 (15.8%)	61 (28.4%)	64 (29.8%)	56 (26%)
GSE13294 ($n = 61$)	46 (75.4%)		15 (24.6%)	Affy U133 Plus 2.0	65.43	32 (53.3%)	28 (46.7%)	0 (0%)	46 (75.4%)	7 (11.5%)	8 (13.1%)

[a]Some cases were classified as "colon" with no information about specific subsite.

[b]Number may not add to total due to missing information.

Published OnlineFirst October 5, 2011; DOI: 10.1158/1078-0432.CCR-11-1570

Table 2. Differentially expressed genes between colon and rectum tumors

Probe	Gene	*q* value	Log_2 fold change	Function
209844_at	*HOXB13*	3.65E-06	−0.600	Transcription factor activity
213823_at	*HOXA11*	5.91E-06	0.514	Transcription factor activity
209167_at	*GPM6B*	1.88E-05	−0.355	Cell differentiation
209170_s_at		3.15E-03	−0.366	
214651_s_at	*HOXA9*	2.32E-05	0.902	Transcription factor activity
209905_at		5.08E-05	0.673	
213147_at	*HOXA10*	2.99E-05	0.460	Transcription factor activity
213150_at		2.68E-04	0.534	
213844_at	*HOXA5*	2.40E-04	0.663	Transcription factor activity
39835_at	*SBF1*	3.20E-04	0.270	Protein amino acid dephosphorylation
218211_s_at	*MLPH*	2.52E-03	0.655	Melanosome transport
216629_at	*SRRM2*	2.78E-03	0.079	RNA splicing
205555_s_at	*MSX2*	2.89E-03	0.387	Transcription factor activity
210319_x_at		3.15E-03	0.455	
204461_x_at	*RAD1*	3.15E-03	−0.292	DNA repair
59644_at	*BMP2K*	3.65E-03	−0.291	Protein amino acid phosphorylation
215703_at	*CFTR*	5.60E-03	−0.396	Transmembrane transport
204425_at	*ARHGAP4*	7.13E-03	0.242	Apoptosis
203332_s_at	*INPP5D*	7.47E-03	0.387	Apoptosis
206854_s_at	*MAP3K7*	9.86E-03	−0.335	Signal transduction

among microsatellite stable (MSS) tumors arising in the colon and rectum is quite small.

Functionally, it was noteworthy that 5 of the top 6 genes belonged to the *HOX* family of transcription factors (Table 2). Other top differentially expressed genes displayed assorted functions such as DNA repair, transcription factor activity, intracellular transport, signal transduction, and apoptosis among others. To formally identify enriched biological processes associated with differentially expressed genes, a GSEA was done. Although no significant function was retrieved, the "*HOX* genes" set appeared with the highest gene enrichment score (Supplementary Fig. S3).

Heterogeneity across studies was explored to identify genes that might have differences in some studies but opposite direction in others that might compensate in the pooled analysis. Only 12 probes showed heterogeneity between studies at the 5% FDR and these could not be ascribed to a systematic effect of one specific study (Supplementary Fig. S4). None of these 12 heterogeneous probes corresponded to differentially expressed genes. Therefore, the 4 studies included in our analysis were considered homogeneous about their differences in expression profiles between colon and rectum.

Refining gene expression profiling: right colon versus left colon tumors and right colon versus rectum tumors

To discount the possibility that similar molecular backgrounds in left colon and rectum tumors were masking possible differences between total colon samples and rectum tumors, a more detailed analysis was conducted looking for differences between right colon, left colon, and rectum tumors, when detailed data about cancer site were available ($n = 499$, all data sets except GSE13294).

Similar to previous results, no major differences were detected between right and left colon, reinforcing our impression that MSS colorectal tumors show very similar expression profiles regardless of their site of origin. Ten genes were found to be differentially expressed between right and left colon tumors after Bonferroni correction. The *q* value method only identified 44 probes differentially expressed corresponding to 40 genes at 1% FDR (Table 3) and 174 probes (150 genes) at 5% FDR. Interestingly, the comparison between left colon and rectum did not identify any differentially expressed gene at 1% FDR (only 3 genes were found at FDR 5%). In contrast, 54 probes (50 genes) were differentially expressed between right colon and rectum when a 1% FDR was used (Table 4) and 374 probes (324 genes) at the 5% FDR. From those, 21 probes (18 genes) passed Bonferroni correction (Fig. 1B). Functionally, those genes showed varied functions, highlighting the *HOX* family as in previous analysis.

To assess the ability of these profiles to discriminate cancer samples by location, a linear discriminant analysis model was built. Leave-one-out internal validation showed that only 37% of rectum tumors were correctly classified when using the colon versus rectum signature (Fig. 1C). Better performance was obtained using the right versus left signature, with 77% accuracy both in right and left tumors (Fig. 1D). The best classification was achieved using the right versus rectum tumors profile (with a total accuracy of 86%), indicating that the major differences exist between the most opposite locations (Fig. 1E).

Published OnlineFirst October 5, 2011; DOI: 10.1158/1078-0432.CCR-11-1570

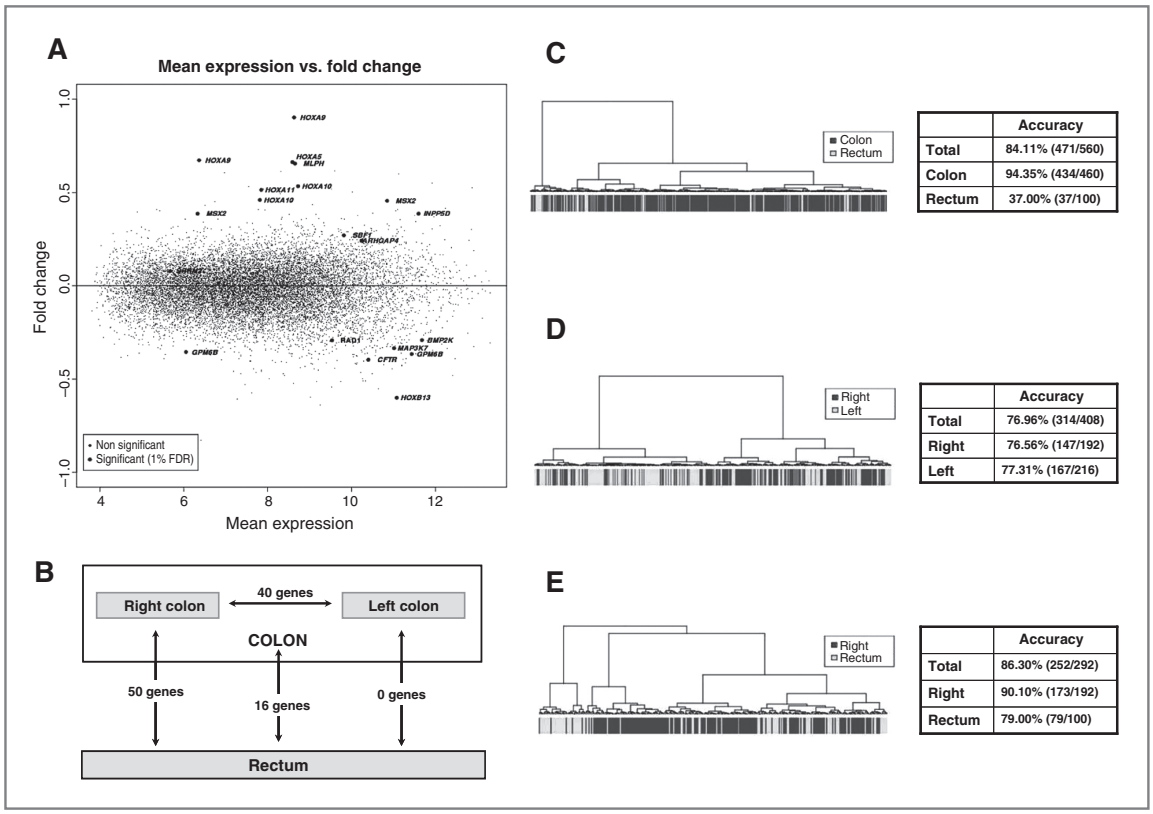

Figure 1. Fold change plot and prediction ability of site-related differentially expressed genes. A, mean expression of each probe set versus its fold change between colon and rectum tumors. B, number of differentially expressed genes between each tumoral location at FDR 1%. Dendrogram illustrating the classification ability of differentially expressed genes among site in colon versus rectum (C), right versus left (D), and right versus rectum (E). Companion tables show the accuracy of each study.

Because classification of rectal tumors is controversial and misclassification could exist between rectal and sigmoid colon tumors, an analysis in which rectal and left-sided colon cancers were pooled and compared with right-sided colon cancer was also conducted. As a result, 46 probes corresponding to 35 genes were found to be differentially expressed after Bonferroni correction. The *q* value method identified 256 probes (202 genes) differentially expressed at 1% FDR (Supplementary Table S3) and 884 probes at 5% FDR. Though this comparison showed a larger number of significant probes, related to the increased sample size of the distal location group, the magnitude of the differences were very small (<10%) and probably not biologically relevant.

HOX genes

Remarkably, *HOX* appeared as the most differentially expressed genes in all transcriptomic comparisons and emerged in the intersection of the lists of differentially expressed genes. In fact, these *HOX* genes were expressed in a gradient in colorectal tumors. The *HOX* genes were more expressed in tumors from the proximal colon, and their expression decreased along more distal locations in the gastrointestinal tract, with the exception of *HOXB13* that showed a reversed pattern (Fig. 2). Genes known to be targets of *HOX* transcription factors (28) were analyzed, but these showed no differences in expression between subsites, indicating that differences observed in *HOX* genes were not affecting a cascade of regulated genes (Supplementary Fig. S5). Also, specific GSEA analysis using *HOX*-related gene sets showed a statistically significant enrichment for genes activated by the chimeric protein NUP98-HOXA9, an aberrant HOX transcription factor and also an enrichment in genes with promoter regions around transcription start site containing the motif that binds with *HOX9* (Supplementary Table S4).

Interestingly, the analysis of expression for *HOX* genes in human normal colorectal mucosa in the data set GES9254 showed the same gradient along the gut than in tumor samples (Supplementary Fig. S6).

Discussion

This pool analysis of 4 data sets from 3 independent studies including a total of 560 samples suggests that there

Published OnlineFirst October 5, 2011; DOI: 10.1158/1078-0432.CCR-11-1570

Table 3. Differentially expressed genes between right and left colon

Probe	Gene	*q* value	Log_2 fold change	Function
206858_s_at	*HOXC6*	2.04E-08	0.868	Transcription factor activity
209844_at	*HOXB13*	1.18E-06	−0.521	Transcription factor activity
219109_at	*SPAG16*	6.11E-05	−0.703	Cell projection
205767_at	*EREG*	1.47E-04	−1.082	Growth factor activity
206307_s_at	*FOXD1*	2.50E-04	0.434	Transcription factor activity
209524_at		2.76E-04	−0.678	
209526_s_at	*HDGFRP3*	6.17E-04	−0.512	Growth factor activity
216693_x_at		6.31E-04	−0.496	
203988_s_at	*FUT8*	1.62E-03	0.308	*N*-glycan processing
205555_s_at	*MSX2*	2.01E-03	0.393	Transcription factor activity
210319_x_at		8.60E-03	0.440	
209752_at	*REG1A*	2.01E-03	1.263	Growth factor activity
217918_at	*DYNLRB1*	3.16E-03	−0.212	Microtubule-based movement
212423_at	*ZCCHC24*	3.63E-03	−0.406	Nucleic acid binding
212419_at		9.56E-03	−0.322	
219228_at	*ZNF331*	3.63E-03	−0.316	Transcription factor activity
219955_at	*L1TD1*	3.82E-03	0.878	Transposase
207457_s_at	*LY6G6D*	4.19E-03	−0.786	—
218094_s_at	*DBNDD2*	4.30E-03	−0.254	Regulation of protein kinase activity
217665_at	—	5.11E-03	−0.247	—
202925_s_at	*PLAGL2*	5.56E-03	−0.334	Transcription factor activity
208948_s_at	*STAU1*	5.56E-03	−0.171	RNA binding
217801_at	*ATP5E*	5.56E-03	−0.138	ATP synthesis
212349_at	*POFUT1*	5.98E-03	−0.252	Notch signaling pathway
204819_at	*FGD1*	6.02E-03	−0.201	Signal transduction
205815_at	*REG3A*	7.19E-03	1.011	Cell proliferation
206340_at	*NR1H4*	7.19E-03	0.177	Transcription factor activity
208979_at	*NCOA6*	7.94E-03	−0.194	Transcription regulation
2019.98_at	*ST6GAL1*	8.51E-03	−0.409	Protein amino acid glycosylation
202673_at	*DPM1*	8.51E-03	−0.239	Protein binding
217718_s_at	*YWHAB*	8.60E-03	−0.138	Signal transduction
204555_s_at	*PPP1R3D*	8.82E-03	−0.260	Protein binding
205463_s_at	*PDGFA*	8.82E-03	−0.323	Growth factor activity
205997_at	*ADAM28*	8.82E-03	0.295	Proteolysis
212234_at	*ASXL1*	8.82E-03	−0.200	Regulation of transcription
212787_at	*YLPM1*	8.82E-03	0.141	Regulation of transcription
213170_at	*GPX7*	8.82E-03	−0.287	Response to oxidative stress
214482_at	*ZBTB25*	8.82E-03	0.131	Transcription factor activity
215210_s_at	*DLST*	8.82E-03	0.238	Tricarboxylic acid cycle
218325_s_at	*DIDO1*	8.82E-03	−0.241	Apoptosis
219108_x_at	*DDX27*	8.82E-03	−0.188	RNA binding
221472_at	*SERINC3*	8.82E-03	−0.190	Protein binding
204015_s_at	*DUSP4*	9.56E-03	0.368	Signal transduction
2031.27_s_at	*SPTLC2*	9.79E-03	0.199	Lipid metabolism

are identifiable expression differences among MSS CRCs that arise in different sites within the large intestine. However, the number of statistically significant differentially expressed genes found between tumor locations was minimal, and the fold change of their expression was within random variation for most cases. With the exception of the *HOX* family, there were no identifiable functional distinctions among the differentially expressed genes. Moreover, the most evident distinctions in expression profiles were those between the right colon and either the left colon or rectum. Expression profiles of MSS rectal cancers and left-sided colon cancers were virtually indistinguishable.

These results imply that anatomic differences are relevant for the clinical management of CRC, but specific molecular

Published OnlineFirst October 5, 2011; DOI: 10.1158/1078-0432.CCR-11-1570

Table 4. Differentially expressed genes between right colon and rectum tumors

Probe	Gene	*q* value	Log_2 fold change	Function
209844_at	*HOXB13*	3.51E-09	−0.856	Transcription factor activity
205555_s_at	*MSX2*	4.30E-05	0.586	Transcription factor activity
210319_x_at		7.11E-05	0.696	
213823_at	*HOXA11*	4.30E-05	0.590	Transcription factor activity
214651_s_at	*HOXA9*	4.30E-05	1.013	Transcription factor activity
209905_at		3.98E-04	0.748	
206858_s_at	*HOXC6*	8.90E-05	1.057	Transcription factor activity
218211_s_at	*MLPH*	9.10E-05	0.856	ROS metabolism
213844_at	*HOXA5*	1.02E-04	0.806	Transcription factor activity
213150_at	*HOXA10*	1.77E-04	0.590	Transcription factor activity
213147_at		6.82E-04	0.509	
3983.5_at	*SBF1*	1.77E-04	0.343	Protein amino acid dephosphorylation
211756_at	*PTHLH*	8.02E-04	−0.167	Hormone activity
206854_s_at	*MAP3K7*	8.77E-04	−0.408	Signal transduction
219109_at	*SPAG16*	9.80E-04	−0.858	Cell projection
214598_at	*CLDN8*	9.93E-04	−0.722	Cell adhesion
209167_at	*GPM6B*	1.15E-03	−0.389	Cell differentiation
204425_at	*ARHGAP4*	1.18E-03	0.334	Apoptosis
36554_at	*ASMTL*	1.36E-03	0.263	Melatonin biosynthesis
204667_at	*FOXA1*	1.43E-03	0.481	Transcription factor activity
204042_at	*WASF3*	1.44E-03	−0.660	Actin binding
203699_s_at	*DIO2*	1.69E-03	−0.281	Hormone biosynthesis
213927_at	*MAP3K9*	1.69E-03	0.130	Signal transduction
211737_x_at	*PTN*	1.92E-03	−0.240	Growth factor activity
209465_x_at		2.34E-03	−0.367	
212840_at	*UBXN7*	2.34E-03	−0.501	Protein binding
210766_s_at	*CSE1L*	2.70E-03	−0.396	Protein transport
215703_at	*CFTR*	2.70E-03	−0.441	Respiratory gaseous exchange
216129_at	*ATP9A*	2.70E-03	−0.458	ATP biosynthesis
212234_at	*ASXL1*	3.21E-03	−0.257	Regulation of transcription
218454_at	*PLBD1*	3.57E-03	−0.375	Lipid degradation
205423_at	*AP1B1*	4.08E-03	0.204	Protein transport
206070_s_at	*EPHA3*	4.59E-03	−0.421	Receptor
203628_at	*IGF1R*	4.83E-03	−0.544	Receptor
202949_s_at	*FHL2*	4.98E-03	0.347	Transcription regulation
221738_at	*KIAA1219*	4.98E-03	−0.229	Signal transduction
202760_s_at	*PALM2*	5.30E-03	−0.503	Regulation of cell shape
219228_at	*ZNF331*	5.30E-03	−0.218	Regulation of transcription
219426_at	*EIF2C3*	6.45E-03	−0.486	RNA binding
214234_s_at	*CYP3A5*	6.64E-03	0.437	Electron carrier activity
218892_at	*DCHS1*	6.64E-03	−0.162	Cell adhesion
222015_at	*CSNK1E*	6.67E-03	0.321	Signal transduction
209195_s_at	*ADCY6*	6.76E-03	0.260	Signal transduction
215078_at	*SOD2*	7.65E-03	−0.363	Removal of superoxide radicals
203671_at	*TPMT*	7.85E-03	−0.238	Metabolism of thiopurine drugs
205767_at	*EREG*	7.85E-03	−1.211	Growth factor activity
221091_at	*INSL5*	7.85E-03	−0.406	Hormone activity
202925_s_at	*PLAGL2*	7.88E-03	−0.395	Transcription factor activity
213242_x_at	*KIAA0284*	8.06E-03	0.327	Microtubule organization
202673_at	*DPM1*	8.45E-03	−0.240	Protein binding
219955_at	*L1TD1*	8.47E-03	1.064	Transposase
201978_s_at	*KIAA0141*	8.75E-03	0.300	—
32069_at	*N4BP1*	8.75E-03	−0.220	Protein binding
211843_x_at	*CYP3A7*	9.25E-03	0.367	Electron carrier activity

Published OnlineFirst October 5, 2011; DOI: 10.1158/1078-0432.CCR-11-1570

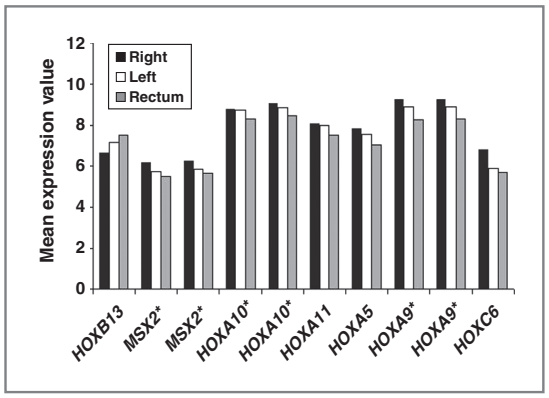

Figure 2. *HOX* genes reverse gradient of expression along colorectal tumor locations. Mean expression value of *HOX* genes in right colon, left colon, and rectum tumors. Genes marked with an asterisk are represented in the microarray by more than one probe set.

profiles of MSS CRC are for the large part quite similar. It is well known that metastases from CRC develop in a stepwise process (29). Rectal cancers usually have a pattern of local recurrence, and retrospective studies show a relevant influence of the surgeon on the prognosis of these patients (30). For colon cancers, the progression pattern is more typically characterized by liver metastases, potentially explained by the fact that superior mesenteric vein drains the right colon whereas neither the left colon nor the rectal vasculature directly drains to liver (29). One might have hypothesized that molecular differences such as DNA repair, apoptosis, or angiogenesis might have distinguished rectal cancers, given the differential efficacy of radiotherapy for rectal cancers. However, our study did not reveal any such clues or signatures. The samples that were analyzed were all tumors collected prior to treatment. Although it is possible that expression profiles that predict response to radiotherapy might exist, our pretreatment data are unable to address this hypothesis. In addition, there is no known evidence of differential radiation sensitivity between colon and rectal cancers. It is only the particular topographic intrapelvic location of the rectum that renders it appropriate for radiotherapy due to the lack of small bowel interaction with the radiation field, which is the limiting factor of the radiotherapy administration in colon cancer (31, 32).

A potential concern of studies that fail to detect differences in expression patterns between tumors is the possibility of insufficient statistical power to detect clinically or biologically meaningful differences due to a small sample size. To address this issue, a pooled analysis has been conducted that included a total of 560 samples, enough to detect differences of 0.5 SD units. In practice, most of the few significant genes identified showed fold changes smaller than 0.6 or a 50% variation in expression, which is usually considered small in microarray expression analyses. Small studies also may show apparent differences that are particular to the selection of cases analyzed. The strength of meta-analyses like the one reported here is that only consistent results remain, and these are easily identified as power is larger and heterogeneity can be explored to identify study specificities. In our analysis, heterogeneity among studies was not a concern as only 12 probes, out of almost 15,000 explored, showed significant heterogeneity and they could not be ascribed to a specific study.

MSI tumors were not included in the analysis due to their known different molecular background (21, 25, 33) and strong association with tumor location. In the case of GSE14333 data set, the researchers did not provide information about MSI status so a simple signature-based imputation was done to exclude putative MSI tumors from the analysis. This procedure had its limitations as its accuracy for MSI was only 85% (Supplementary Table S1). Thus more MSS tumors than necessary may have been excluded, and some MSI cancers from GSE14333 may have been inadvertently included by our simple imputation. This strategy of attempting to eliminate MSI CRCs was preferred to the alternative design that would have resulted in a strong biased estimation or a choice to completely exclude all 215 of the otherwise informative tumors from GSE14333. A choice to exclude these tumors would have further reduced the power to detect any possible existing differences. It is reassuring to note that tumors excluded from the analysis had clinical features related to MSI, such as a predominance of female and older patient that originate in the colon, mainly in the right side (Supplementary Table S2; ref. 34). In addition, an analysis excluding GSE14333 data set was conducted and similar results (still less significant genes) were obtained (Supplementary Table S5).

It is worth mentioning that differences between cancer sites previously reported in some studies may be related to MSI status: Komuro and colleagues found gene expression differences between right- and left-sided CRCs in genes related to MSI such as *MSH2* in right-sided tumors (35). A similar work by Birkenkamp-Demtroder and colleagues also reported differences between 25 MSS and MSI right and left tumors (36). Watanabe and colleagues describes small differences between proximal and distal MSI colorectal tumors (37). These differences are probably related to the combination of MSI and MSS tumors. *CDX2* has been reported to be more expressed in proximal structures than distal (11) but we did not found it as a right side–associated gene. However, if we include in our analysis MSI tumors and look for *CDX2* expression, it appeared as a differentially expressed gene with a q value less than 0.01. So, the significance of *CDX2* is probably due to MSI and not due to tumor location.

Although most of CIMP-positive tumors are MSI and therefore were not included in this analysis, there are some CIMP-positive MSS tumors that preferentially arise in the right colon (2, 38) which could explain some of the larger differences between the tumors arising in the right colon and other tumors. In an attempt to explore this possibility, a gene expression signature that differentiates MSS CIMP$^+$ and MSS CIMP$^-$ colorectal carcinomas was used (39) in a GSEA analysis. This revealed an association between *CIMP*$^+$ genes and right-sided genes (Supplementary Fig. S7) and

Published OnlineFirst October 5, 2011; DOI: 10.1158/1078-0432.CCR-11-1570

suggests that some of the described differences could be related to CIMP phenotype.

Only *HOX* genes were found to be an enriched set associated with colon tumors. These genes (also known as homeobox genes) encode transcription factors that play essential roles in controlling cell growth and differentiation during embryonic and normal tissue development. Many homeobox genes have been reported to be deregulated in a variety of solid tumors including CRC and also to vary between normal mucosa and CRC tissue (40, 41). Interestingly, differences in *HOX* expression between carcinomas from the right colon and left colon have been reported previously (42). In normal human intestinal mucosa, *HOX-A* genes are widely expressed in undifferentiated proliferating cells at the base of the crypts (43). So, we speculate that *HOX* expression in colon tumors could be an amplification of the signal from colon cancer stem cell that drives intestinal cell differentiation. Because *HOX* expression patterns along the gut reflect pivotal roles of these genes in the regional regenerative process of the epithelial cells (44), it is possible that our results simply mirrors the *HOX* expression pattern maintained in tumors as it usually is in the normal mucosa. In fact, we observed the same gradient of expression in normal mucosa along the gut (Supplementary Fig. S6). However, despite our analysis showed no differential expression among genes targeted by *HOX*, enrichment in genes activated by NUP98-HOXA9 was found. This is an aberrant HOXA9 transcription factor that promotes the growth of murine hematopoietic progenitors and blocks their differentiation (45). This result might be related to a possible role of *HOX* genes in CRC right-sided tumor progression that deserves experimentally exploration.

In conclusion, our study strongly suggests that the expression profiles of MSS CRCs do not show major differences for tumors arising in the colon or rectum, and that the small, but consistent differences observed between right sided and left sided/rectal cancers are largely driven by the *HOX* family of genes. Although it is clear that diverse somatic mutations that characterize individual cancers suggest the possibility for targeted therapies to be developed for each individual cancer in each patient, our data show that CRCs, on average, show few differences based on tumor location. This observation could have important clinical implications in terms of prognostic analysis, biomarker discovery, or drug development.

Disclosure of Potential Conflicts of Interest

No potential conflicts of interest were disclosed.

Grant Support

This study was supported by a grant (1R01CA81488) from the National Cancer Institute, the Catalan Institute of Oncology and the Private Foundation of the Biomedical Research Institute of Bellvitge (IDIBELL), the Instituto de Salud Carlos III (grants PI08-1635, PI08-1359, PS09-1037), CIBERESP CB06/02/2005 and the "Acción Transversal del Cancer," the Catalan Government DURSI grant 2009SGR1489, the European Commission grant FP7-COOP-Health-2007-B "HiPerDART," and the AECC (Spanish Association Against Cancer) Scientific Foundation.

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked *advertisement* in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

Received June 21, 2011; revised September 12, 2011; accepted September 13, 2011; published OnlineFirst October 5, 2011.

References

1. Markowitz SD, Bertagnolli MM. Molecular origins of cancer: molecular basis of colorectal cancer. N Engl J Med 2009;361:2449–60.
2. Ogino S, Goel A. Molecular classification and correlates in colorectal cancer. J Mol Diagn 2008;10:13–27.
3. Iacopetta B. Are there two sides to colorectal cancer? Int J Cancer 2002;101:403–8.
4. Bufill JA. Colorectal cancer: evidence for distinct genetic categories based on proximal or distal tumor location. Ann Intern Med 1990;113:779–88.
5. Tan KK, Lopes Gde L Jr, Sim R. How uncommon are isolated lung metastases in colorectal cancer? A review from database of 754 patients over 4 years. J Gastrointest Surg 2009;13:642–8.
6. Li FY, Lai MD. Colorectal cancer, one entity or three. J Zhejiang Univ Sci B 2009;10:219–29.
7. Casillas S, Pelley RJ, Milsom JW. Adjuvant therapy for colorectal cancer: present and future perspectives. Dis Colon Rectum 1997;40:977–92.
8. Hermann S, Rohrmann S, Linseisen J. Lifestyle factors, obesity and the risk of colorectal adenomas in EPIC-Heidelberg. Cancer Causes Control 2009;20:1397–408.
9. Wei EK, Giovannucci E, Wu K, Rosner B, Fuchs CS, Willett WC, et al. Comparison of risk factors for colon and rectal cancer. Int J Cancer 2004;108:433–42.
10. Terry P, Giovannucci E, Michels KB, Bergkvist L, Hansen H, Holmberg L, et al. Fruit, vegetables, dietary fiber, and risk of colorectal cancer. J Natl Cancer Inst 2001;93:525–33.
11. Rozek LS, Lipkin SM, Fearon ER, Hanash S, Giordano TJ, Greenson JK, et al. CDX2 polymorphisms, RNA expression, and risk of colorectal cancer. Cancer Res 2005;65:5488–92.
12. Fritzmann J, Morkel M, Besser D, Budczies J, Kosel F, Brembeck FH, et al. A colorectal cancer expression profile that includes transforming growth factor beta inhibitor BAMBI predicts metastatic potential. Gastroenterology 2009;137:165–75.
13. Yamasaki M, Takemasa I, Komori T, Watanabe S, Sekimoto M, Doki Y, et al. The gene expression profile represents the molecular nature of liver metastasis in colorectal cancer. Int J Oncol 2007;30:129–38.
14. Matsuyama T, Ishikawa T, Mogushi K, Yoshida T, Iida S, Uetake H, et al. MUC12 mRNA expression is an independent marker of prognosis in stage II and stage III colorectal cancer. Int J Cancer 2010;127:2292–9.
15. Salazar R, Roepman P, Capella G, Moreno V, Simon I, Dreezen C, et al. Gene expression signature to improve prognosis prediction of stage II and III colorectal cancer. J Clin Oncol 2011;29:17–24.
16. Poynter JN, Gruber SB, Higgin PD, Almog R, Bonner JD, Rennert HS, et al. Statins and the risk of colorectal cancer. N Engl J Med 2005;352:2184–92.
17. Vilar E, Bartnik CM, Stenzel SL, Raskin L, Ahn J, Moreno V, et al. MRE11 deficiency increases sensitivity to poly(ADP-ribose) polymerase inhibition in microsatellite unstable colorectal cancers. Cancer Res 2011;71:2632–42.
18. Barrett T, Edgar R. Gene expression omnibus: microarray data storage, submission, retrieval, and analysis. Methods Enzymol 2006;411:352–69.
19. Parkinson H, Kapushesky M, Shojatalab M, Abeygunawardena N, Coulson RA, Holloway E, et al. ArrayExpress–a public database of microarray experiments and gene expression profiles. Nucleic Acids Res 2007;35 (Database issue):D747–50.

Published OnlineFirst October 5, 2011; DOI: 10.1158/1078-0432.CCR-11-1570

20. Jorissen RN, Gibbs P, Christie M, Prakash S, Lipton L, Desai J, et al. Metastasis-associated gene expression changes predict poor outcomes in patients with dukes stage B and C colorectal cancer. Clin Cancer Res 2009;15:7642–51.
21. Jorissen RN, Lipton L, Gibbs P, Chapman M, Desai J, Jones IT, et al. DNA copy-number alterations underlie gene expression differences between microsatellite stable and unstable colorectal cancers. Clin Cancer Res 2008;14:8061–9.
22. LaPointe LC, Dunne R, Brown GS, Worthley DL, Molloy PL, Wattchow D, et al. Map of differential transcript expression in the normal human large intestine. Physiol Genomics 2008;33:50–64.
23. Affymetrix, Inc. GeneChip expression analysis—data analysis fundamentals. Santa Clara, CA: Affymetrix, Inc. [cited 2002]. Available from: http://media.affymetrix.com/support/downloads/manuals/data_analysis_fundamentals_manual.pdf.
24. Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, et al. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics 2003;4:249–64.
25. Kim H, Nam SW, Rhee H, Shan Li L, Ju Kang H, Hye Koh K, et al. Different gene expression profiles between microsatellite instability-high and microsatellite stable colorectal carcinomas. Oncogene 2004;23:6218–25.
26. Rozek LS, Herron CM, Greenson JK, Moreno V, Capella G, Rennert G, et al. Smoking, gender, and ethnicity predict somatic BRAF mutations in colorectal cancer. Cancer Epidemiol Biomarkers Prev 2010;19:838–43.
27. Subramanian A, Kuehn H, Gould J, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A 2005;102:15545–50.
28. Svingen T, Tonissen K. Hox transcription factors and their elusive mammalian gene targets. Heredity 2006;97:88–96.
29. Sugarbaker PH. Metastatic inefficiency: the scientific basis for resection of liver metastases from colorectal cancer. J Surg Oncol Suppl 1993;3:158–60.
30. Di Cataldo A, Scilletta B, Latino R, Cocuzza A, Li Destri G. The surgeon as a prognostic factor in the surgical treatment of rectal cancer. Surg Oncol 2007;16 Suppl 1:S53–6.
31. Aleman BM, Bartelink H, Gunderson LL. The current role of radiotherapy in colorectal cancer. Eur J Cancer 1995;31:1333–9.
32. Foroudi F, Tyldesley S, Barbera L, Huang J, Mackillop WJ. An evidence-based estimate of the appropriate radiotherapy utilization rate for colorectal cancer. Int J Radiat Oncol Biol Phys 2003;56:1295–307.
33. Dunican DS, McWilliam P, Tighe O, Parle-McDermott A, Croke DT. Gene expression differences between the microsatellite instability (MIN) and chromosomal instability (CIN) phenotypes in colorectal cancer revealed by high-density cDNA array hybridization. Oncogene 2002;21:3253–7.
34. Kakar S, Burgart LJ, Thibodeau SN, Rabe KG, Petersen GM, Goldberg RM, et al. Frequency of loss of hMLH1 expression in colorectal carcinoma increases with advancing age. Cancer 2003;97:1421–7.
35. Komuro K, Tada M, Tamoto E, Kawakami A, Matsunaga A, Teramoto K, et al. Right- and left-sided colorectal cancers display distinct expression profiles and the anatomical stratification allows a high accuracy prediction of lymph node metastasis. J Surg Res 2005;124:216–24.
36. Birkenkamp-Demtroder K, Olesen SH, Sorensen FB, Laurberg S, Laiho P, Aaltonen LA, et al. Differential gene expression in colon cancer of the caecum versus the sigmoid and rectosigmoid. Gut 2005;54:374–84.
37. Watanabe T, Kobunai T, Toda E, Yamamoto Y, Kanazawa T, Kazama Y, et al. Distal colorectal cancers with microsatellite instability (MSI) display distinct gene expression profiles that are different from proximal MSI cancers. Cancer Res 2006;66:9804–8.
38. Curtin K, Slattery ML, Samowitz WS. CpG island methylation in colorectal cancer: past, present and future. Patholog Res Int 2011;2011:902674.
39. Ferracin M, Gafà R, Miotto E, Veronese A, Pultrone C, Sabbioni S, et al. The methylator phenotype in microsatellite stable colorectal cancers is characterized by a distinct gene expression profile. J Pathol 2008;214:594–602.
40. Shah N, Sukumar S. The Hox genes and their roles in oncogenesis. Nat Rev Cancer 2010;10:361–71.
41. Samuel S, Naora H. Homeobox gene expression in cancer: insights from developmental regulation and deregulation. Eur J Cancer 2005;41:2428–37.
42. Kanai M, Hamada J, Takada M, Asano T, Murakawa K, Takahashi Y, et al. Aberrant expressions of HOX genes in colorectal and hepatocellular carcinomas. Oncol Rep 2010;23:843–51.
43. Freschi G, Taddei A, Bechi P, Faiella A, Gulisano M, Cillo C, et al. Expression of HOX homeobox genes in the adult human colonic mucosa (and colorectal cancer?). Int J Mol Med 2005;16:581–7.
44. Yahagi N, Kosaki R, Ito T, Mitsuhashi T, Shimada H, Tomita M, et al. Position-specific expression of Hox genes along the gastrointestinal tract. Congenit Anom (Kyoto) 2004;44:18–26.
45. Takeda A, Goolsby C, Yaseen NR. NUP98-HOXA9 induces long-term proliferation and blocks differentiation of primary human CD34 +hematopoietic cells. Cancer Res 2006;66:6628–37.

2. Article 2: *Discovery and validation of new potential biomarkers for early detection of colon cancer.*

2.1 Resum en català

El CCR és un problema de salut molt greu a nivell mundial. Cada any es diagnostiquen més d'un milió de nous casos i causa mig milió de morts a tot el món. Cap mesura de prevenció primària ha resultat efectiva per reduir la incidència, però s'ha demostrat que la detecció precoç mitjançant els programes de cribratge en redueix la mortalitat, ja que el pronòstic de la malaltia depèn en gran mesura de l'estadi en el moment del diagnòstic. No obstant, encara hi ha un extens debat sobre quin test hauria de ser utilitzat per al cribratge del CCR. Per exemple, utilitzar com a prova estàndard la colonoscòpia, malgrat la seva alta sensibilitat i especificitat, genera molta controvèrsia entre la comunitat científica i clínica a causa del cost econòmic, la incòmoda preparació intestinal i els riscos associats, com a prova invasiva que és. Per altra banda, el TSOF com l'antic test bioquímic conegut com a Guaiac i el test immunològic basat en la detecció d'hemoglobina humana en femta, són utilitzats en molts programes de cribratge. Tot i així, es coneix que la seva sensibilitat, al voltant del 80%, i la seva especificitat, que es mou entre el 91% i el 94%, són encara millorables. Addicionalment, cal tenir en compte que els assajos basats en femta sempre tenen una acceptació més baixa que altres tipus de proves de cribratge.

Com a conseqüència del context prèviament exposat queda clar, doncs, que la recerca de nous i millors biomarcadors per al diagnòstic precoç del CCR és encara una gran necessitat. D'aquesta manera, en el nostre treball es postula que, la detecció precisa de proteïnes secretades específicament per les cèl·lules tumorals del càncer de còlon en fluids biològics de fàcil obtenció, podria servir com a prova, ja que es podria incorporar d'una forma fàcil i econòmica a la pràctica clínica i seria susceptible de ser més ben acceptada a nivell poblacional. L'objectiu principal d'aquest estudi va ser, identificar nous biomarcadors sèrics i demostrar la seva potencial utilitat per al diagnòstic precoç del càncer de còlon.

Aquest estudi comprenia una sèrie homogènia de 250 mostres de còlon. D'aquestes, 100 mostres provenien de tumors de còlon estables en microsatèl·lits, sense tractament previ a la cirurgia i amb estadi II en el moment del diagnòstic. Per completar aquests casos, també es disposava de 100 mostres més de la mucosa adjacent i patològicament normal, aparellada als tumors. Finalment, 50 mostres provenien de la mucosa del còlon d'individus sans. De totes elles es va mesurar l'expressió gènica mitjançant *microarrays*. Gràcies a les anàlisis d'expressió diferencial ajustats per comparacions múltiples utilitzant la correcció de Bonferroni, uns filtres de variabilitat i uns valors mínims de canvi exigits, es va identificar un conjunt inicial de 505 gens candidats a biomarcadors. Tots ells posseïen resultats significatius i una alta capacitat de discriminació entre els diferents tipus de teixit. Per continuar reduint aquesta llista es van aplicar filtres basats en bases de dades de secreció

de proteïnes, de consistència amb altres dades públiques disponibles, i exigint sempre que els nostres candidats inicialment ja tinguessin una baixa presència a la sang de controls sans, segons la literatura i alguns conjunts dades públiques analitzats.

A continuació, es va portar a terme una validació tècnica i biològica dels 23 candidats finalment seleccionats mitjançant una PCR quantitativa (RT-qPCR, de l'anglès *Real-Time quantitative PCR*). Aquesta validació, que es va realitzar en una sèrie de mostres independent però amb similars característiques, va proporcionar uns bons resultats altament concordants amb els perfils d'expressió observats en els *microarrays*. Finalment, com a prova de concepte es van avaluar en sèrum 9 dels candidats identificats mitjançant kits d'ELISA comercials. La sèrie de mostres en sèrum constava de 80 casos de càncer de còlon, 23 pacients amb adenomes i 77 controls sense càncer. Un dels candidats, la proteïna COL10A1, va mostrar diferències significatives entre els nivells de proteïna en sèrum dels controls, amb els dels pacients amb adenomes (p=0,0083) i amb els sèrums dels casos de càncer de còlon (p=3,2e-6), com es pot veure a la Figura 13. El gen *COL10A1* es correspon amb la subunitat alfa del col·lagen de tipus X i s'expressa normalment durant el procés d'ossificació. En conseqüència, en el nostre estudi observem baixos nivells d'expressió i de proteïna a la mucosa sana del còlon, però elevats en els tumors.

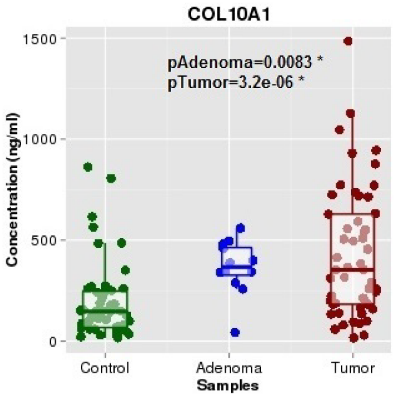

Figura 13. Valors de la concentració en sèrum de la proteïna COL10A1 en controls sense càncer, pacients amb adenomes i en casos de càncer de còlon.

Com a conclusió, es presenta un procés de validació seqüencial en fases, gràcies al qual es va identificar un conjunt de possibles candidats a biomarcadors per la detecció precoç del càncer colorectal. El candidat més prometedor és la detecció de la proteïna COL10A1 en sèrum, la qual permet identificar els adenomes i els càncers invasius amb una bona sensibilitat i especificitat i una àrea sota la corba ROC (de l'anglès *Receiver Operating Characteristic*) de 0,76. Aquests resultats evidencien que els biomarcadors sèrics poden canviar l'escenari actual del cribratge de càncer de còlon en un futur proper: la implementació d'una prova més econòmica i més acceptada a nivell poblacional contribuiria a disminuir el gran impacte en la societat d'aquesta malaltia.

OPEN ACCESS Freely available online

PLOS ONE

Discovery and Validation of New Potential Biomarkers for Early Detection of Colon Cancer

Xavier Solé[1,2☯], Marta Crous-Bou[1,2☯], David Cordero[1,2☯], David Olivares[1,2], Elisabet Guinó[1,2], Rebeca Sanz-Pamplona[1,2], Francisco Rodriguez-Moranta[2,3], Xavier Sanjuan[2,4], Javier de Oca[2,5,6], Ramon Salazar[2,7], Victor Moreno[1,2,5]*

1 Unit of Biomarkers and Susceptibility, Cancer Prevention and Control Program, Catalan Institute of Oncology (ICO) and CIBERESP, Hospitalet de Llobregat, Barcelona, Spain, **2** Colorectal Cancer Group, Bellvitge Biomedical Research Institute (IDIBELL), Hospitalet de Llobregat, Barcelona, Spain, **3** Department of Gastroenterology, University Hospital of Bellvitge, Hospitalet de Llobregat, Barcelona, Spain, **4** Department of Pathology, University Hospital of Bellvitge, Hospitalet de Llobregat, Barcelona, Spain, **5** Department of Clinical Sciences, Faculty of Medicine, University of Barcelona (UB), Barcelona, Spain, **6** Department of General and Digestive Surgery, Colorectal Unit, University Hospital of Bellvitge, Hospitalet de Llobregat, Barcelona, Spain, **7** Department of Medical Oncology, Catalan Institute of Oncology (ICO), Hospitalet de Llobregat, Barcelona, Spain

Abstract

Background: Accurate detection of characteristic proteins secreted by colon cancer tumor cells in biological fluids could serve as a biomarker for the disease. The aim of the present study was to identify and validate new serum biomarkers and demonstrate their potential usefulness for early diagnosis of colon cancer.

Methods: The study was organized in three sequential phases: 1) biomarker discovery, 2) technical and biological validation, and 3) proof of concept to test the potential clinical use of selected biomarkers. A prioritized subset of the differentially-expressed genes between tissue types (50 colon mucosa from cancer-free individuals and 100 normal-tumor pairs from colon cancer patients) was validated and further tested in a series of serum samples from 80 colon cancer cases, 23 patients with adenoma and 77 cancer-free controls.

Results: In the discovery phase, 505 unique candidate biomarkers were identified, with highly significant results and high capacity to discriminate between the different tissue types. After a subsequent prioritization, all tested genes (N = 23) were successfully validated in tissue, and one of them, COL10A1, showed relevant differences in serum protein levels between controls, patients with adenoma (p = 0.0083) and colon cancer cases (p = 3.2e-6).

Conclusion: We present a sequential process for the identification and further validation of biomarkers for early detection of colon cancer that identifies COL10A1 protein levels in serum as a potential diagnostic candidate to detect both adenoma lesions and tumor.

Impact: The use of a cheap serum test for colon cancer screening should improve its participation rates and contribute to decrease the burden of this disease.

Citation: Solé X, Crous-Bou M, Cordero D, Olivares D, Guinó E, et al. (2014) Discovery and Validation of New Potential Biomarkers for Early Detection of Colon Cancer. PLoS ONE 9(9): e106748. doi:10.1371/journal.pone.0106748

Editor: Francisco X. Real, Centro Nacional de Investigaciones Oncológicas (CNIO), Spain

Received April 23, 2014; **Accepted** August 1, 2014; **Published** September 12, 2014

Data Availability: The authors confirm that all data underlying the findings are fully available without restriction. The gene expression dataset is available in the National Center for Biotechnology Information's Gene Expression Omnibus with GEO series accession number GSE44076. All relevant data are within the paper and its Supporting Information files.

Funding: This work was supported by the Catalan Institute of Oncology (ICO) and the Private Foundation of the Biomedical Research Institute of Bellvitge (IDIBELL), the Instituto de Salud Carlos III [grants PI08-1635, PI08-1359, PS09-1037, PI11-1439], CIBERESP CB06/02/2005 and the "Acción Transversal del Cancer", the Catalan Government DURSI [grant 2009SGR1489], the Fundació Privada Olga Torres (FOT), the Spanish Association Against Cancer (AECC) Scientific Foundation, the European Commission [grant FP7-COOP-Health-2007-B HiPerDART], and Xarxa de Bancs de Tumors de Catalunya sponsored by Pla Director d'Oncología de Catalunya (XBTC). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have filed a patent regarding the use of COL10A1 as biomarker for early diagnosis of colon cancer with title: "SERUM BIOMARKER FOR DIAGNOSING COLORECTAL CANCER". European patent number: EP2680003A1 (28.06.2012) and Spain ES2436667A2(03.01.2014). The authors confirm that this does not alter their adherence to all PLOS ONE policies on sharing data and materials.

* Email: v.moreno@iconcologia.net

☯ These authors contributed equally to this work.

Introduction

Colorectal cancer (CRC) is a leading cause of death worldwide, with over one million of new cases and half a million of deaths around the world every year [1]. Five-year relative survival rates are under 50%, but this greatly depends on the stage at the time of diagnosis [2]. No primary preventive measure has proven efficacy in reducing incidence, but early detection through population screening has been found to reduce mortality [3].

Nowadays there is debate about which test should be used for CRC screening. Until further evidence is collected, current European guidelines accept the fecal occult blood test followed by confirmatory colonoscopy, which is therapeutic when resectable adenomas are identified [4]. Most population-based screening programs are using guaiac based fecal occult blood test, which biochemically detects small traces of blood derived from bleeding lesions in feces, or fecal immunological test, which is based on immunodetection of human hemoglobin in feces. These tests have a sensitivity of 80% for CRC and 28% for adenoma >1 cm, and specificities in the range of 91 to 94% [5]. Moreover, patient compliance with stool-based assays tends to be low [6]. The use of colonoscopy as a gold standard for CRC screening is controversial. It has reported higher sensitivity (97%) and specificity (98%) for early detection of CRC, but it also has several pitfalls associated to it: increased economic cost; requirement of highly trained staff; uncomfortable bowel preparation; invasiveness; risk of morbidity and mortality attached to the procedure [7,8]. Moreover, in countries where national colonoscopy screening is available, compliance has often been low [9].

Serum-based markers would be highly attractive for CRC screening since they are minimally invasive and could be integrated in any routine health checkup without the need of additional stool sampling, thereby increasing acceptance among patients. Current molecular biology techniques allow an easier generation of many hypotheses of candidate biomarkers for diagnosis, prognosis or therapeutic response in CRC, but the need for a proper validation has been often reported [10–12]. The underlying hypothesis is that tumor cells of CRC, even in its pre-invasive stages, suffer important genetic alterations that induce release of characteristic proteins or nucleic acids potentially detectable in biological fluids obtained by non-invasive methods such as blood or feces [11]. Detection by molecular biology techniques of these substances will serve as a biomarker of disease to develop diagnostic tests with improved predictive power of current screening tests. Therefore, the aim of the present study was to identify and validate new serum biomarkers and demonstrate their potential usefulness for early diagnosis of colon cancer.

Methods

The biomarker assessment in this study was organized in sequential and consecutive phases for discovery, technical and biological validation, and proof of concept to test the potential clinical use of selected biomarkers. Firstly, gene expression microarray data were analyzed to identify candidate biomarkers in tissue samples from colon cancer cases and cancer-free controls (*Discovery Phase*). Secondly, using an alternative technique based on quantitative real-time PCR (RT-qPCR), a selection of differentially expressed genes was validated in the same set of tissue biopsies (*Technical Validation Phase*), as well as in biopsies obtained from an independent set of patients (*Biological Validation Phase*). Finally, the potential clinical use of the most promising validated candidates was tested in serum samples from colon cancer cases, a small set of adenomas, and cancer-free controls through the detection of the corresponding secreted protein using enzyme-linked immunosorbent assay (ELISA) tests (*Proof of Concept Phase*). Reported STARD guidelines [13] have been the basis for defining our protocol.

Patients and samples

Main characteristics of the subjects included in the present study are shown in Table 1. Colon tumor and paired adjacent (~5–10 cm) pathologically normal mucosa tissue samples used in this study were obtained at the time of surgery from a series of cases with an incident diagnosis of colon adenocarcinoma attending the Bellvitge University Hospital (Barcelona, Spain) between January 1996 and December 2007. Included cases were selected to form a homogenous series of patients with stage II, microsatellite-stable sporadic colon cancer. All patients underwent radical surgery and did not receive chemotherapy prior to surgery. Pathologists confirmed all colon cancer diagnoses and selected fresh tissue samples from tumor and adjacent mucosa taken from the proximal resection margin. A hematoxylin-eosin staining was performed on a slide cut of the tumor specimen to guide the pathologist in selecting an area with at least 75% of tumor cells. The stage grouping followed the authoritative UICC guide "TNM Atlas, 6th Edition". The best approximation to this classification was derived from the information collected at the time of diagnosis for each case.

Tissue samples of colon mucosa from cancer-free controls were obtained through colonoscopy between February and May 2010. A series of consecutive patients who underwent colonoscopy indicated by symptoms (usually anemia, bleeding, gastrointestinal pain or altered rhythm) were invited to participate. Those with negative results (i.e. without colonic lesions) were included in this study. None of them reported family history of cancer.

Finally, serum samples from colon cancer cases and cancer-free controls were selected from an epidemiologic case-control study on gene-environment interactions that has been previously described in detail [14]. All serum samples were collected prior to surgery for cases and just before colonoscopy for controls.

To simplify naming different sample types, here we will use *tumor* (T) when referring to tumor samples from colon cancer patients, *adjacent normal* (A) when referring to pathology normal colon mucosa samples from colon cancer patients, and *cancer free* (F) when referring to colon mucosa samples from cancer-free individuals.

The Clinical Research Ethics Committee of the Bellvitge University Hospital approved the study protocol, and all individuals provided written informed consent to participate and for genetic analyses to be done on their samples.

RNA extraction

Total RNA was isolated from frozen tissue samples using Exiqon miRCURY RNA Isolation Kit (Exiqon A/S, Denmark), according to manufacturer's protocol, and considering all recommended precautions to avoid RNA degradation by RNases. Extracted RNA was quantified by NanoDrop ND-1000 Spectrophotometer (Nanodrop Technologies, Wilmington, DE) and stored at -80°C. The quality of these RNA samples was further checked using RNA 6000 Nano Kit (Agilent Technologies, Santa Clara, CA) following manufacturer's guidelines, and was confirmed by gel electrophoresis. RNA integrity numbers (RIN) showed good quality ([Q1 = 7.5; Median = 8.25; Q3 = 8.9] for tumors, [Q1 = 7; Median = 7.5; Q3 = 8] for adjacent normal and [Q1 = 7.8; Median = 8.3; Q3 = 8.65] for healthy normal). RNA purity was measured with the ratio of absorbance at 260 nm and 280 nm (mean = 1.96, SD = 0.04), with no differences among tissue types.

Discovery series - expression arrays

The discovery series included 100 pairs of tumor and adjacent normal colonic mucosa samples and 50 samples of colonic mucosa from cancer-free individuals (total n = 250). Total RNA extracted from these samples was hybridized onto Affymetrix Human Genome U219 array plates (Affymetrix, Santa Clara, CA) following manufacturer's recommendations. Four samples (two

Table 1. Main characteristics of the subjects included in the study.

	Discovery series	Validation series	
		Tissue	Serum
Cancer-free controls	**n = 50**	**n = 34**	**n = 77**
Gender			
Male	27 (54%)	16 (47%)	39 (51%)
Female	23 (46%)	18 (53%)	38 (49%)
Median age (range in years)	63 (25–88)	62.5 (50–69)	67 (34–89)
Biopsy localization			
Right colon	27 (54%)	26 (76%)	-
Left colon	23 (46%)	8 (24%)	-
Patients with adenoma			**n = 23**
Gender			
Male	-	-	16 (70%)
Female	-	-	7 (30%)
Median age (range in years)	-	-	60 (53–69)
Adenoma localization			
Right colon	-	-	4 (17.39%)
Left colon	-	-	17 (73.91%)
Rectum	-	-	2 (8.70%)
Median size (range in mm)	-	-	12 (4–23)
Histological type			
Tubular	-	-	7 (30.43%)
Tubulovillous	-	-	15 (65.22%)
Not available	-	-	1 (4.35%)
Degree of dysplasia			
High	-	-	8 (34.78%)
Low	-	-	14 (60.87%)
Not available	-	-	1 (4.35%)
Cases	**n = 100**	**n = 70**	**n = 80**
Gender			
Male	72 (72%)	39 (56%)	52 (65%)
Female	28 (28%)	31 (44%)	28 (35%)
Median age (range in years)	71.5 (43–87)	68.5 (41–91)	66 (22–83)
Tumor localization			
Right colon	39 (39%)	35 (50%)	19 (23.75%)
Left colon	61 (61%)	35 (50%)	37 (46.25%)
Rectum	-	-	24 (30%)
Histological grade			
High	6 (6%)	17 (24%)	18 (22.5%)
Low	94 (94%)	53 (76%)	60 (75%)
Not available			2 (2.5%)
Tumor stage			
I	-	-	9 (11.25%)
II	100 (100%)	70 (100%)	27 (33.75%)
III	-	-	34 (42.25%)
IV	-	-	10 (12.5%)
T - Primary tumor			
T1	-	-	9 (11.25%)
T2	-	-	9 (11.25%)
T3	92 (92%)	61 (87%)	46 (57.5%)

Table 1. Cont.

	Discovery series	Validation series	
		Tissue	Serum
T4	8 (8%)	9 (13%)	16 (20%)
N - Regional lymph nodes			
N0	100 (100%)	70 (100%)	44 (55%)
N1	-	-	22 (27.5%)
N2	-	-	14 (17.5%)
M – Distant metastasis			
M0	100 (100%)	70 (100%)	70 (87.5%)
M1	-	-	10 (12.5%)
Mean lymph node yield	19.6	31	28.8
Extramural vascular invasion			
Present	7 (7%)	16 (22.9%)	30 (37.5%)
Absent	93 (93%)	54 (77.1%)	50 (62.5%)

doi:10.1371/journal.pone.0106748.t001

adjacent normal-tumor pairs) were excluded from the dataset after quality control. Thus, a final dataset of 246 arrays was used for subsequent analyses.

Raw data were normalized using the Robust Multiarray Average algorithm [15] implemented in the *affy* package [16] of the Bioconductor suite (http://www.bioconductor.org) [17]. All statistical analysis were done with the R statistical computing software (http://www.r-project.org) [18].

Before the differential expression analysis was performed, low-variant and Y-chromosome transcripts were removed from subsequent analyses. For the remaining probesets, regularized-Student's t-tests were used to detect significant overexpression between adjacent normal (A) or tumor samples (T) and cancer-free mucosa (F). Bonferroni correction was applied to account for multiple hypothesis testing. In order to narrow down the initially obtained lists, candidate probesets were further filtered based on different criteria: low expression levels and low variability in cancer-free mucosa; large average fold change between T/F or A/F; and homogeneity of effects among multiple probes for the same gene, when available. Probesets that passed the filtering criteria were mapped to genes, the units of information used for downstream analyses.

A prioritization procedure was performed to select the best candidate genes for validation using publicly available data [19–24]. Criteria accounted for were related to reproducibility and specificity issues: observed reproducibility of the expression differences; very low levels of expression in blood tissue; and selection of genes with large expression in colon tissue when compared to other tissues according to GeneCards database (http://www.genecards.org) [25], though most genes were expressed in multiple tissues. The gene expression dataset is available in the National Center for Biotechnology Information's Gene Expression Omnibus [26] with GEO series accession number GSE44076 and in the project website (http://www.colonomics.org).

Technical and biological validation – RT-qPCR for expression assessment

Expression levels of selected genes were assessed with RT-qPCR both for the discovery series and for an additional set of 104 samples (70 paired adjacent normal/tumor tissues from colon cancer patients and 34 from cancer-free controls). These samples were collected between January 1996 and June 2011 following the same protocol and stored under the same conditions as the discovery series. cDNA was synthesized from the extracted mRNA with the transcription first strand cDNA synthesis kit (Roche Applied Science, Penzberg, Germany) following standard procedures.

Two sets of primers were designed for each gene, and each set was assayed in duplicate. Three control genes were included in the assay: *ACTB*, *TPT1*, and *UBC*. *ACTB* was chosen based on the extensive previous literature pointing it as a suitable housekeeping gene for gene expression analyses in colon samples [27–29]. *UBC* and *TPT1* were selected based on the high stability of their expression levels across all samples in our array data (Figures S1 and S2). Interestingly, they had also been previously postulated as potentially suitable housekeeping genes for gene expression assays in colon samples [29].

Multiplexed RT-qPCRs assays were done using BioMark Dynamic Array 96×96 Plates (Fluidigm Corporation, San Francisco, CA). Resulting images were analyzed with Fluidigm Biomark software using standard parameters. Raw qPCR data were processed with the HTqPCR package v1.10.0 [30]. Before the assessment of differential expression between different tissue types, the expression matrix was filtered for quality purposes. *UBC* was finally selected as the housekeeping control based on the stability of its threshold cycle values (Figure S2). Mann-Whitney tests were used to compare expression levels between cancer-free and adjacent normal samples and between cancer-free and tumor samples. Each set of primers was analyzed independently, and the set of primers that displayed the highest significant results in the analysis of differential expression was selected as a representative.

Identification of serum biomarkers – proof of concept for screening validity

To test the potential value for early detection, the most promising candidates from the biological validation were assayed in serum samples in a series of 80 colon cancer cases, 23 patients with adenoma and 77 cancer-free controls, all tested in duplicate to increase the precision of the experiment. Ten-milliliter samples of peripheral venous blood were collected from colon cancer cases, patients with adenomas and controls. After centrifuge for 15

minutes at 1000 rpm within 30 minutes of collection, serum was aliquoted and stored at −80°C. Commercial ELISA kits from Life Sciences Inc and R&D Systems, depending on availability, were used according to the manufacturer's instructions to assess serum protein concentrations. All assays employ quantitative sandwich enzyme immunoassay technique. The concentration of target proteins in each sample was calculated from a standard curve run in duplicate in each plate. The scientists examining these serum samples were unaware of the patient's diagnosis. A linear model adjusting for age, gender and potential batch effects was used to assess the statistical significance of the differential protein levels among groups. The association of markers with patient characteristics as age and gender, multiple epidemiological factors and tumor characteristics is shown in Table S1. Since some serum markers showed extreme values for a few subjects, a rank-based test was also performed. The results did not change in a relevant way and the p-values derived from the linear models are reported.

The number of samples used was calculated to attain a 10% precision on sensitivity and specificity estimates for expected values of 75%. This required at least 72 subjects per group. For the discovery series, this number was unbalanced to oversample tumors, which have larger variability and a wide range of candidates had to be analyzed. The validation series in serum was supplemented with a smaller subgroup of adenoma (n = 23) to explore the usefulness of the markers to detect this premalignant lesion.

Results

Biomarker discovery

In this well-selected homogeneous set of samples, differences in mRNA expression measured with Affymetrix HG-U219 array plates were so remarkable that an unsupervised technique (principal components analysis) using the full set of probesets separated almost perfectly the three tissue types (Figure 1a). The first principal component clearly divides tumor samples of colon cancer cases (T) and non-tumor samples. Remarkably, the second principal component also split cancer-free (F) from adjacent normal samples belonging to patients with cancer (A).

From 33,853 probe sets included in the array with high variability, 5,503 were over-expressed in A when compared to F, and 11,229 were over-expressed in T when compared to F ($p < 0.05$, Bonferroni corrected). We have focused specifically on over-expressed genes because these differences are more likely to be detected in serum, and therefore more suitable to be used as diagnostic biomarkers. Interestingly, a remarkable level of overlap (3,101 probe sets, ~56%) was found between these two lists of probe sets, suggesting that adjacent normal mucosa in patients with cancer had already experienced important alterations in gene expression, and rising the importance of using cancer-free mucosa as reference tissue. Global results from differential expression analysis are shown in Figures 1b and 1c.

To prioritize candidates for diagnostic biomarkers, a set of filters were applied to the initial set of differentially expressed probe sets. These filters were based primarily on statistical criteria (i.e. large fold-change between A/F or T/F, low levels of expression and low variability in F samples). These filters yielded a final number of 242 selected probe sets between A/F and 443 between T/F, corresponding to a set of 194 and 352 genes, respectively (Table S2).

This first selection provided a list of 505 unique candidate biomarkers with highly significant expression differences between tumor and cancer-free tissue. Due to the technical difficulties derived from the validation of such a large amount of biomarkers, additional technical and biological criteria were applied to further narrow down the list of potential candidates. This second set of filters were based on assessing the consistency between the different probesets for each gene; confirmation of our results in independent and publicly available gene expression datasets; and null or low expression levels of these genes in blood samples from cancer-free individuals. Moreover, prior knowledge and molecular information for each gene was compiled from the literature and online databases to ensure the selection of the most reliable candidates (i.e. protein secretion, tissue specificity, protein function, previous evidence as a biomarker, among others). A list of the 23 best candidates was finally selected for validation in the next step (Table 2, columns 1–2).

Technical and biological validation

To ensure the reliability of the results obtained from the gene expression arrays, the selection of 23 biomarkers was validated with an alternative technique (RT-qPCR) both in the same set of samples (i.e. technical validation), and also in an independent series with equivalent clinical and epidemiological characteristics (i.e. biological validation).

Figure 2 displays two-way gene and sample clustering of the RT-qPCR expression values both for the results of the technical (Figure 2a) and the biological validation (Figure 2b). Horizontal axes of the heatmaps (i.e. columns) show a clear separation of tumor samples from adjacent normal and cancer-free groups. The vertical axis of genes (i.e. rows) showed two clusters of genes, one for those differentially expressed between A/F and another for the differentially expressed between T/F. These results highly replicate the pattern of expression observed in the arrays in the discovery phase, reinforcing the potential role of these genes as diagnostic biomarkers of colon cancer. A formal comparison of the expression differences between sample types for the technical and biological validation is shown in Table 2, columns 3–4. Although only p-values are displayed in Table 2, the expression levels of all the validated genes behaved consistently throughout the different phases, as depicted in Figure S3.

Proof of concept – identification of serum biomarkers

As a pilot proof-of-concept to demonstrate their potential usefulness as colon cancer early diagnostic biomarkers, a selection of 9 genes were tested in serum using ELISA tests. The prioritization of these candidates was based on an extensive literature review and availability of commercial ELISA kit.

Results for each protein are shown in Figure 3. Remarkably, collagen type X alpha1 (COL10A1) displayed very high concentrations in colon cancer cases and adenomas when compared to controls ($p = 3.2e\text{-}6$ and $p = 0.0083$, respectively). Serum concentrations of COL10A1 in controls, adenomas and colon cancer cases by stage are shown in Figure S4. Interestingly, statistically significant differences were found when controls were compared to each one of the different tumor stages, except Stage I, probably due to the small sample size of this group. The area under the receiver operating characteristic (ROC) curve was 0.75 for cancer and 0.76 when adenoma and colon cancer were considered together (Figure 4), showing potential good classification ability. Matrix metalloproteinase-7 (MMP7) also showed a significant association but was not further considered because it was due to an underexpression in adenomas compared to controls, which represented the opposite sense of differential expression that we were trying to validate. No combination of COL10A1 with any of the other proteins significantly increased the area under the ROC curve (data not shown).

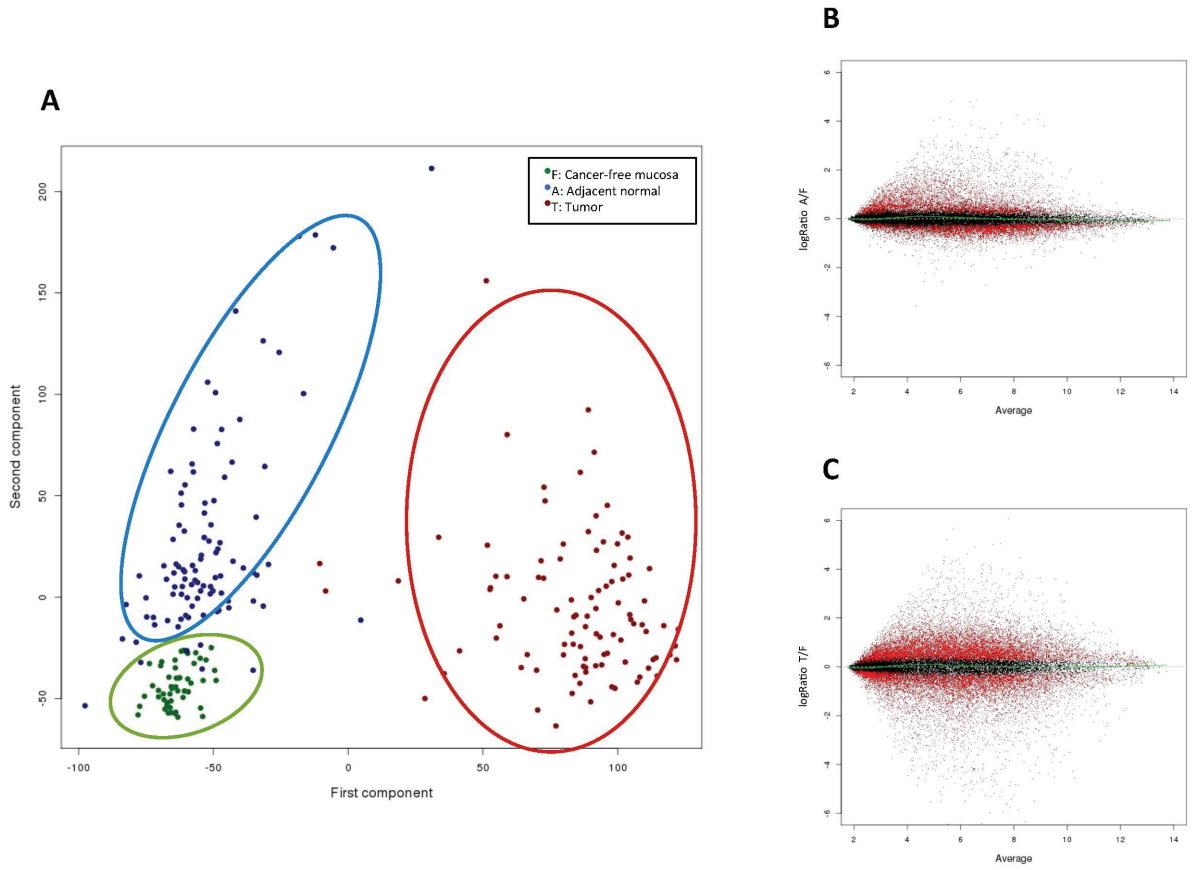

Figure 1. Differences in expression between tissue types in the biomarker discovery series. A. Principal component analysis. **B.** Differentially expressed genes between adjacent normal and cancer-free samples. **C.** Differentially expressed genes between tumor and cancer-free samples.
doi:10.1371/journal.pone.0106748.g001

Discussion

CRC screening with fecal occult blood test has demonstrated efficacy in randomized trials. Nonetheless, the low specificity of the test suggests the need of more accurate alternative diagnostic tests. Sigmoidoscopy, colonoscopy and Computerized Tomography scan (i.e. virtual colonoscopy) are strong alternative candidates, but all have important limitations, mainly regarding costs, possible severe side effects and reduced participation. Participation is an important factor for screening effectiveness, and it is also a generalized observation that screening based on fecal occult blood test has low participation rates [31–33]. Thus, a diagnostic test based on a routine blood test would probably be able to reach a higher percentage of the population, and public health authorities would favor such a test if efficacy and costs were similar to fecal occult blood test. With these premises in mind, we started this study to search for diagnostic biomarkers that can be detected in blood with a simple and affordable ELISA test.

Our study of gene expression in colon tissue has confirmed previous observations that a large number of genes are deregulated in tumor when compared to adjacent normal mucosa. From about 20,000 genes interrogated in the expression array, and after filtering by several restrictive criteria, 505 unique candidate biomarkers have been identified (Table S2), with highly significant results and high capacity to discriminate between paired tumor and adjacent normal samples. A strong feature of our study design is the inclusion of a set of samples from cancer-free controls (n = 50). This has allowed us to identify genes that do not show expression differences between adjacent normal and tumor tissue from colon cancer patients, as well as to confirm that overexpressed genes in tumors do not display high expression levels in cancer-free colon tissue, which could preclude their potential use as biomarkers. We have previously described that gene expression of adjacent normal colon mucosa in a patient with cancer already has been significantly altered when compared to cancer-free colon mucosa [34], which reinforces the need of including tissue from cancer-free individuals in projects aiming to find diagnosis biomarkers for colon cancer.

The large number of candidates identified in the analysis of expression data led us to prioritize which ones were to be selected for further validation. We used a combination of criteria, which included consistency with other publicly available datasets and literature; low or no expression levels in cancer-free mucosa or other tissues; expression predominant to colon cancer tissue; and selection of secretable proteins. Since the identification of serum proteins is expensive and time consuming, we undertook a technical and biological validation of the best candidates before attempting ELISA tests. The technical validation (i.e. in the same

Table 2. Selected genes to be technically and biologically validated.

Gene name	Type*	Discovery fold change	Discovery p-value	Technical validation p-value	Biological validation p-value
COL11A1	T/F	4.01	6.66e-36	2.06e-16	2.20e-5
KIAA1199	T/F	4.91	1.47e-71	5.94e-18	2.86e-13
MMP7	T/F	5.66	1.06e-43	4.06e-21	2.36e-15
CEL	T/F	3.14	1.83e-20	9.34e-13	6.40e-8
GAL	A/F	1.88	5.50e-29	2.65e-20	1.60e-9
MMP3	T/F	4.63	9.72e-35	2.54e-18	1.21e-12
THBS2	T/F	4.23	2.48e-38	6.71e-16	2.0e-3
COL10A1	T/F	2.67	3.08e-20	3.82e-08	5.40e-6
ESM1	T/F	2.47	2.80e-37	2.54e-18	1.05e-12
JUB	T/F	2.76	7.17e-61	2.66e-19	3.74e-08
CST1	T/F	1.92	2.88e-17	2.61e-20	2.35 e-14
MSX2	T/F	3.05	1.10e-36	2.45e-17	4.35e-9
EPHX4	T/F	2.88	5.52e-39	1.34e-20	5.61e-12
TNC	A/F	1.82	2.12e-15	1.66e-11	3.13e-4
CA9	T/F	2.62	7.75e-20	1.08e-11	4.65e-10
CLDN2	T/F	3.10	2.37e-26	9.12e-5	0.013
DPT	A/F	2.91	5.96e-36	6.01e-17	4.02e-6
SFRP2	A/F	4.02	2.29e-50	1.15e-19	2.23e-10
MMP10	T/F	1.92	7.94e-18	1.59e-13	7.65e-12
FAP	T/F	2.60	2.29e-29	1.11e-17	5.13e-12
SRPX2	T/F	3.18	3.83e-35	7.81e-12	1.57e-5
LOC100127888	T/F	1.92	2.56e-17	1.38e-09	9.55e-5
CXCL5	T/F	2.96	6.49e-19	9.6e-3	2.06e-4

*T/F: expression in tumor > expression in cancer-free mucosa; A/F: expression in adjacent normal mucosa > expression in cancer-free mucosa.
The association for all genes are significant after Bonferroni correction, but p-values shown are unadjusted for multiple comparisons.
doi:10.1371/journal.pone.0106748.t002

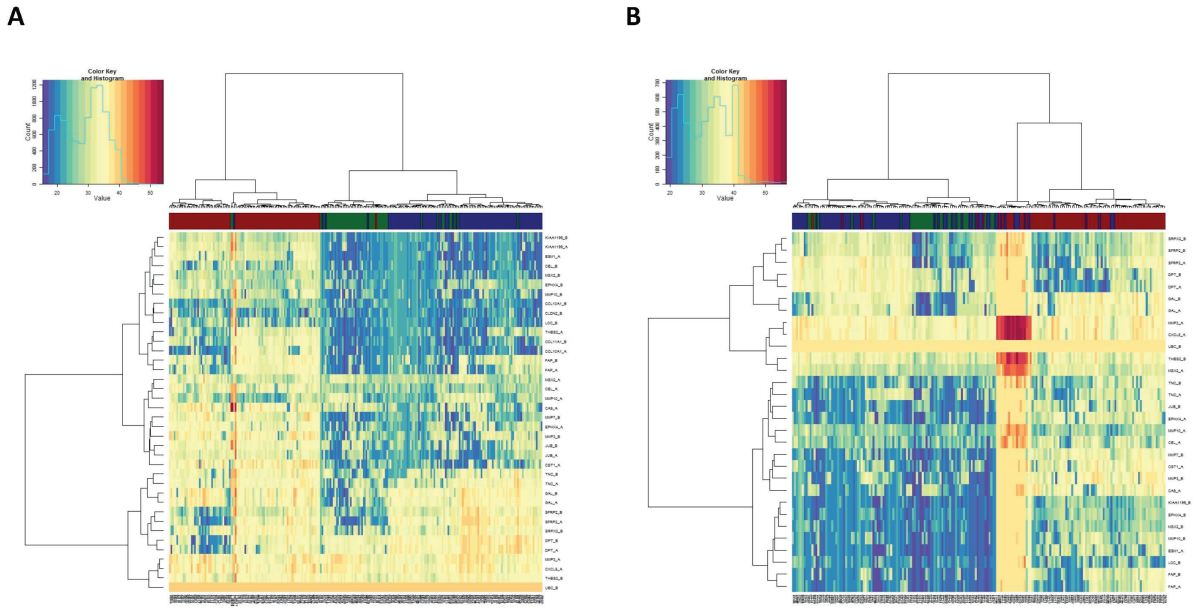

Figure 2. Heatmap of threshold cycle values from technical (A) and biological validation (B). Samples are color-coded on top of the heatmaps based on the tissue type (i.e., cancer-free mucosa = green, adjacent normal tissue from colon cancer patients = blue, tumor tissue = red).
doi:10.1371/journal.pone.0106748.g002

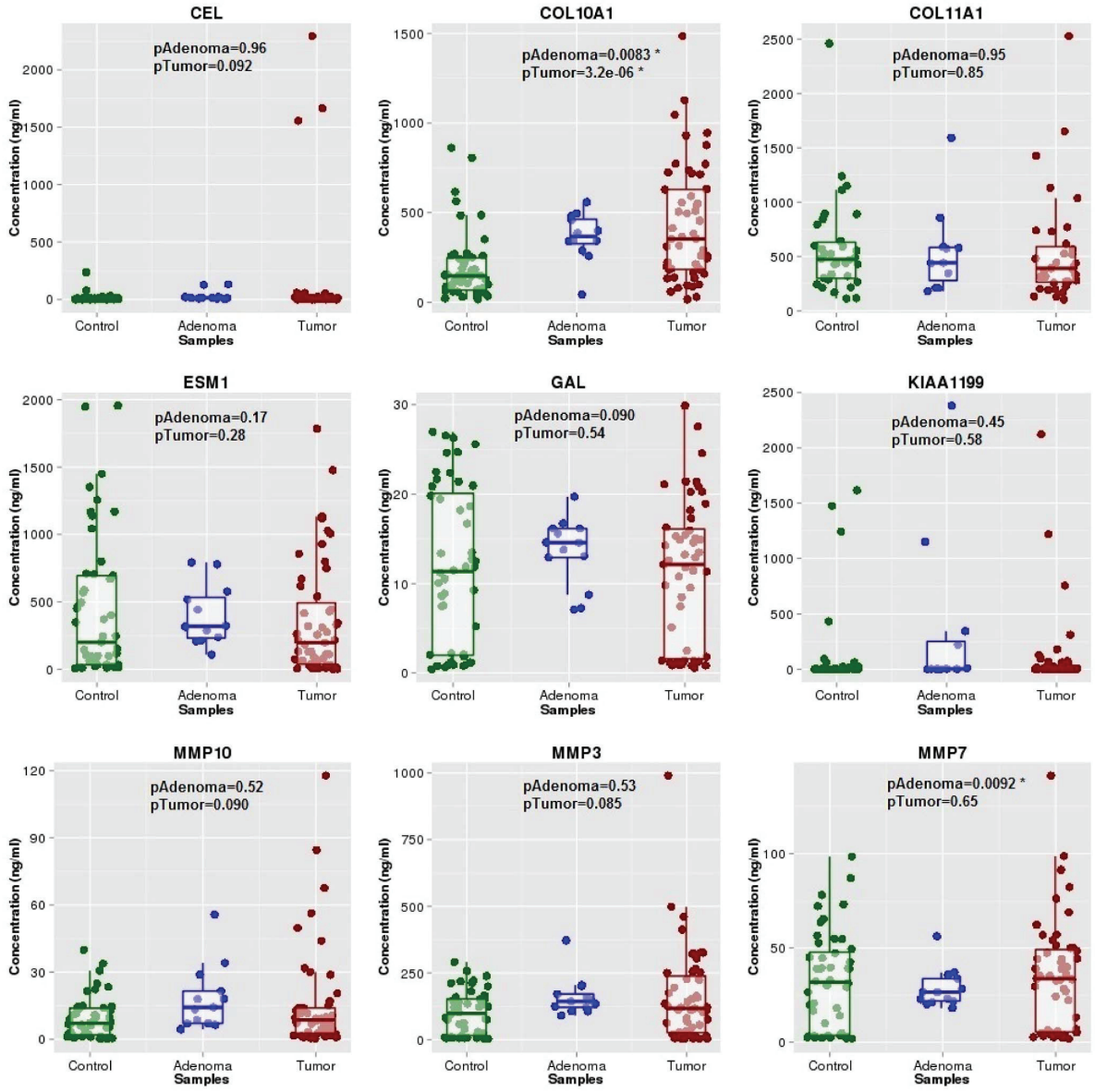

Figure 3. ELISA serum concentrations of each selected protein in cancer-free controls, patients with adenoma and colon cancer cases.
doi:10.1371/journal.pone.0106748.g003

set of samples but with a different technique) showed a remarkable reproducibility of the expression level differences measured by RT-qPCR and microarrays for all tested genes, thus confirming that the expression dataset obtained with Affymetrix HG-U219 microarrays was of outstanding quality and reliably identified expression differences between the different tissue types. Therefore, we expect that the number of false positives in the remaining list (not validated) of significant differentially expressed genes between tissue types to be low. Moreover, the confirmation of the previously identified differences in a biologically independent dataset also highlights the validity of the results obtained with microarrays.

The next step in our sequential validation process was attempting to identify in serum the corresponding proteins for our candidate genes and assess their potential use for early diagnosis. We also included a subgroup of patients with adenoma, since this is also an important target for CRC screening. Using commercial ELISA kits we could assess the protein levels of all the genes prioritized. Remarkably, COL10A1 showed relevant enough differences between controls and colon cancer patients ($p = 3.2 \times 10^{-6}$) to be proposed as a potential diagnostic candidate. MMP7 also showed some differences for adenomas ($p = 0.0092$), but showed an opposite direction to the expected one.

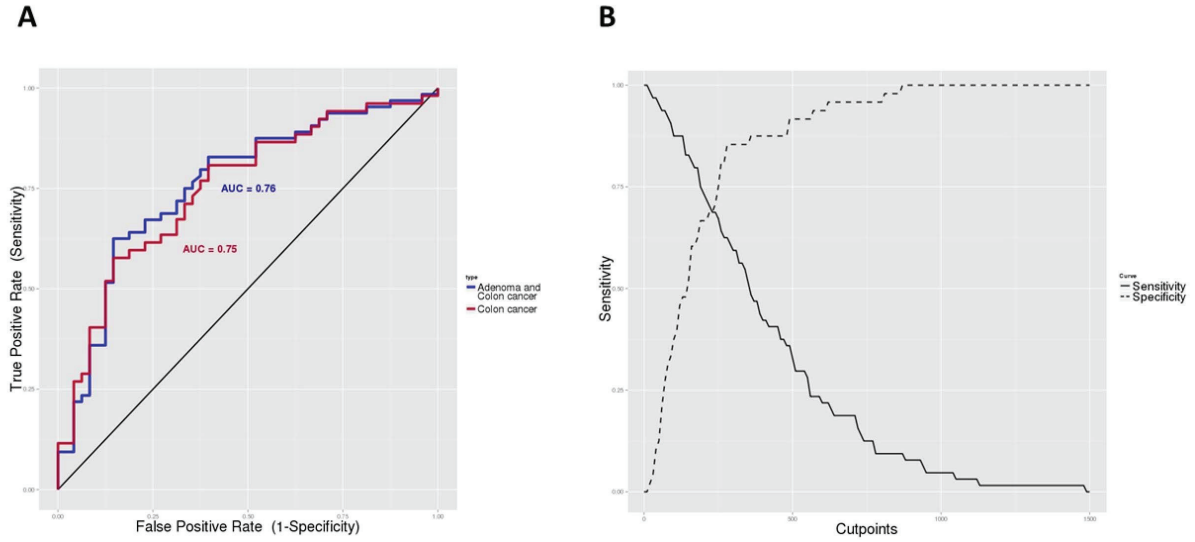

Figure 4. COL10A1 performance as a diagnostic biomarker. A. Receiver operating characteristic curves for both adenomas and colon cancer together (purple) and colon cancer cases only (red). **B.** Different marker cutpoints against the sensitivity and specificity curves.
doi:10.1371/journal.pone.0106748.g004

We have identified the protein COL10A1 which, when detected at high concentrations in blood, may be indicative of the presence of a neoplastic lesion in the colon. This protein was selected after a sequential procedure in which we started exploring whole genome expression data in colon tissue. Elevated serum levels of COL10A1 were observed both for adenoma and colon cancer patients. The area under the ROC curve was 0.76, which makes COL10A1 as a promising diagnostic biomarker. The cutpoint of 280 ng/ml attained 0.63 sensitivity and 0.85 specificity for colon cancer or adenoma (Figure 4). Similar values were obtained for cancer only. A few cancer-free subjects showed high levels of COL10A1 in serum, higher than the average for adenoma, indicating that other processes not related to colorectal lesions can increase COL10A1 levels.

COL10A1 is a short chain collagen mainly expressed by chondrocytes during ossification. Defects in this protein have been related to Schmid-type metaphyseal chondrodysplasia [35]. COL10A1 is not expressed in normal colon epithelium, but is a direct transcriptional target of *RUNX2* [36], a transcription factor that is expressed in cancer cells, and has been related to multiple cancers. The elevated expression of *COL10A1* observed in tumors might be an indirect effect of higher-level regulatory alterations occurring in the tumor. In fact, we have observed a high correlation between RUNX2 and *COL10A1* expression in tumors (Pearson $R = 0.5$, results not shown). Our expression data also identifies high correlation between *COL10A1* and other genes: *SFRP4*, *INHBA*, *TNFSF4* that are involved in cytokine and Wnt signaling [37–40]. Recently *COL10A1* has been found to be overexpressed in diverse tumors related to the vasculature component [41]. However, the expression in other tissues other that cartilage is low, which contributes to the specificity observed in our study. Moreover, a recent study suggests that *COL10A1* expression may be regulated by non-steroidal anti-inflammatory drugs [42], which have protective effect on CRC risk, suggesting that the mechanism of *COL10A1* overexpression might be also related to inflammatory processes. Specifically, we have also explored if COL10A1 levels were related to non-steroidal anti-inflammatory drugs consumption, as indicative of inflammatory conditions, but we could not find a strong association (Table S1).

Since the sample size used for the serum assays in our study is limited, further validation studies are needed to confirm that *COL10A1* is useful for population screening. However, the large differences observed among colon cancer patients and cancer-free controls position this biomarker as a promising candidate. We have adjusted the analyses for age, gender and batch effect to control for potential confounding. This was only relevant for MMP3, which showed a strong association with gender. COL10A1 also shows an association with primary tumor (T) reinforcing the idea that the association could be an additional link with colon cancer and tumor size (Figure S5) more than stage, showing no association. Other potential confounders explored, including multiple epidemiological factors and tumor characteristics, were not associated with serum levels of the different markers (Table S1).

Our findings evidence that serum biomarkers for CRC screening can be identified and may change the scenario in a near future. Other blood molecular markers can also be of interest. Detection of DNA methylated septin 9 (*SEPT9*) gene is a promising candidate in the development of a non-invasive molecular screening method [43,44]. Although the *SEPT9* assay successfully identified 68% of colon cancers at a specificity of 89% [45], the cost of the test is high since it involves DNA extraction and a quantitative DNA methylation assay. Besides, the method of assaying DNA methylation is still a handicap for the creation of a robust diagnostic tool, since a quantitative PCR step is often required. Another biomarker for colon cancer is the fecal detection of aberrant methylation of Vimentin gene (*VIM*). In this case the authors report a sensitivity of 46% for a specificity of 90% [46]. However, biomarkers based on serum proteins detected by conventional ELISA could be cheaper and more reliable to be used in a daily clinical practice and for population screening. The detection of the carcinoembryonic antigen [47] is one of the most widely used tumor markers worldwide, especially in CRC. Although in clinical use for almost 30 years, with clear value for

prognosis and progression detection of CRC, the value of carcinoembryonic antigen in colorectal cancer screening is low mainly due to its low sensitivity (about 35%) and specificity (between 30 and 80%) [48]. Other candidate biomarkers have been proposed by other studies, such as metalloproteinases MMP7 [49] and MMP9 [50]. MMP7 was one of the potential biomarker that appears in our candidate list. In that study the authors reported a 58% of sensitivity and 100% of specificity, and the area under ROC curve was 0.81, but we have not been able to confirm these results with the commercial ELISA kit used. As the authors pointed, further studies are required involving larger numbers of subjects, to confirm these results.

Although many other biomarkers for colon cancer have been previously proposed [51–54], in most cases there is no further progress beyond the proposal [2,6,11,55–57] since none of them may have sufficient sensitivity and specificity to be considered in the current guidelines [13].

It is possible that other genes from the list of candidates identified in the expression analysis also could be useful for early diagnosis, either alone or in combination with *COL10A1*. Therefore, further studies are required to assess the utility of other potential biomarkers for the early detection of colorectal cancer.

In conclusion, after different steps of sequential validation, we have identified a list of candidate biomarkers for early detection of colon cancer. The most promising one is the detection of COL10A1 in serum, which can identify adenoma and invasive cancer with high sensitivity and specificity. The use of a cheap serum test for CRC screening should improve participation and contribute to decrease the burden of this disease.

Supporting Information

Figure S1 Strip charts of expression values for housekeeping genes in discovery series and technical validation series.
(PDF)

Figure S2 Density plots of expression values for housekeeping genes in discovery series and technical validation series.
(PDF)

Figure S3 Expression levels for the discovery series, technical validation and biological validation of the 23 selected genes.
(PDF)

Figure S4 Serum concentration values of COL10A1 in relation to tumor stage.
(PDF)

Figure S5 Serum concentration values of COL10A1 in relation to tumor size.
(PDF)

Table S1 Association of serum markers with patient characteristics, epidemiological factors and tumor characteristics.
(DOCX)

Table S2 Excel file with the complete list of candidate biomarkers identified in the discovery phase. Sheet "T-F": expression in tumor > expression in cancer-free mucosa; Sheet "A-F": expression in adjacent normal mucosa > expression in cancer-free mucosa.
(XLS)

Acknowledgments

The authors would like to thank Carmen Atencia, Ferran Martínez, Pilar Medina, Isabel Padrol and Thais Pereira for their technical assistance.

Author Contributions

Conceived and designed the experiments: XS MCB DC VM. Performed the experiments: DO MCB. Analyzed the data: XS MCB DC EG VM. Contributed reagents/materials/analysis tools: XS MCB DC DO VM. Contributed to the writing of the manuscript: XS MCB DC EG RSP VM. Recruitment of patients and collection of clinical data: FRM XSJ JDO RS.

References

1. Ferlay J, Parkin DM, Steliarova-Foucher E (2010) Estimates of cancer incidence and mortality in Europe in 2008. Eur J Cancer 46: 765–781.
2. Kim HJ, Yu MH, Kim H, Byun J, Lee C (2008) Noninvasive molecular biomarkers for the detection of colorectal cancer. BMB reports 41: 685–692.
3. Burt RW (2010) Colorectal cancer screening. Current opinion in gastroenterology 26: 466–470.
4. von Karsa L, Patnick J, Segnan N, Atkin W, Halloran S, et al. (2013) European guidelines for quality assurance in colorectal cancer screening and diagnosis: overview and introduction to the full supplement publication. Endoscopy 45: 51–59.
5. Parra-Blanco A, Gimeno-Garcia AZ, Quintero E, Nicolas D, Moreno SG, et al. (2010) Diagnostic accuracy of immunochemical versus guaiac faecal occult blood tests for colorectal cancer screening. Journal of gastroenterology 45: 703–712.
6. Pawa N, Arulampalam T, Norton JD (2011) Screening for colorectal cancer: established and emerging modalities. Nature reviews Gastroenterology & hepatology 8: 711–722.
7. Regula J, Rupinski M, Kraszewska E, Polkowski M, Pachlewski J, et al. (2006) Colonoscopy in colorectal-cancer screening for detection of advanced neoplasia. The New England journal of medicine 355: 1863–1872.
8. Ransohoff DF (2009) How much does colonoscopy reduce colon cancer mortality? Annals of internal medicine 150: 50–52.
9. Quintero E, Castells A, Bujanda L, Cubiella J, Salas D, et al. (2012) Colonoscopy versus fecal immunochemical testing in colorectal-cancer screening. The New England journal of medicine 366: 697–706.
10. Newton K, Hill J (2010) 5-FU and mismatch repair deficient colorectal cancer: is it time to consider a change in practice? Colorectal disease: the official journal of the Association of Coloproctology of Great Britain and Ireland 12: 706–707.
11. Ahlquist DA (2010) Molecular detection of colorectal neoplasia. Gastroenterology 138: 2127–2139.
12. Miller S, Steele S (2012) Novel molecular screening approaches in colorectal cancer. Journal of surgical oncology 105: 459–467.
13. Bossuyt PM, Reitsma JB (2003) The STARD initiative. Lancet 361: 71.
14. Moreno V, Gemignani F, Landi S, Gioia-Patricola L, Chabrier A, et al. (2006) Polymorphisms in genes of nucleotide and base excision repair: risk and prognosis of colorectal cancer. Clin Cancer Res 12: 2101–2108.
15. Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, et al. (2003) Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics 4: 249–264.
16. Gautier L, Cope L, Bolstad BM, Irizarry RA (2004) affy—analysis of Affymetrix GeneChip data at the probe level. Bioinformatics 20: 307–315.
17. Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, et al. (2004) Bioconductor: open software development for computational biology and bioinformatics. Genome biology 5: R80.
18. Team RDC (2012) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria ISBN 3-900051-07-0.
19. Huang RS, Duan S, Bleibel WK, Kistner EO, Zhang W, et al. (2007) A genome-wide approach to identify genetic variants that contribute to etoposide-induced cytotoxicity. Proceedings of the National Academy of Sciences of the United States of America 104: 9758–9763.
20. Stranger BE, Forrest MS, Dunning M, Ingle CE, Beazley C, et al. (2007) Relative impact of nucleotide and copy number variation on gene expression phenotypes. Science 315: 848–853.
21. Stranger BE, Nica AC, Forrest MS, Dimas A, Bird CP, et al. (2007) Population genomics of human gene expression. Nature genetics 39: 1217–1224.
22. Zhou B, Xu W, Herndon D, Tompkins R, Davis R, et al. (2010) Analysis of factorial time-course microarrays with application to a clinical study of burn injury. Proceedings of the National Academy of Sciences of the United States of America 107: 9923–9928.
23. Hong Y, Downey T, Eu KW, Koh PK, Cheah PY (2010) A 'metastasis-prone' signature for early-stage mismatch-repair proficient sporadic colorectal cancer patients and its implications for possible therapeutics. Clinical & experimental metastasis 27: 83–90.

24. LaPointe LC, Dunne R, Brown GS, Worthley DL, Molloy PL, et al. (2008) Map of differential transcript expression in the normal human large intestine. Physiological genomics 33: 50–64.
25. Safran M, Dalah I, Alexander J, Rosen N, Iny Stein T, et al. (2010) GeneCards Version 3: the human gene integrator. Database: the journal of biological databases and curation 2010: baq020.
26. Edgar R, Domrachev M, Lash AE (2002) Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic acids research 30: 207–210.
27. Kheirelseid EA, Chang KH, Newell J, Kerin MJ, Miller N (2010) Identification of endogenous control genes for normalisation of real-time quantitative PCR data in colorectal cancer. BMC molecular biology 11: 12.
28. Dydensborg AB, Herring E, Auclair J, Tremblay E, Beaulieu JF (2006) Normalizing genes for quantitative RT-PCR in differentiating human intestinal epithelial cells and adenocarcinomas of the colon. American journal of physiology Gastrointestinal and liver physiology 290: G1067–1074.
29. Andersen CL, Jensen JL, Orntoft TF (2004) Normalization of real-time quantitative reverse transcription-PCR data: a model-based variance estimation approach to identify genes suited for normalization, applied to bladder and colon cancer data sets. Cancer research 64: 5245–5250.
30. Dvinge H, Bertone P (2009) HTqPCR: high-throughput analysis and visualization of quantitative real-time PCR data in R. Bioinformatics 25: 3325–3326.
31. Hewitson P, Glasziou P, Watson E, Towler B, Irwig L (2008) Cochrane systematic review of colorectal cancer screening using the fecal occult blood test (hemoccult): an update. The American journal of gastroenterology 103: 1541–1549.
32. Hol L, van Leerdam ME, van Ballegooijen M, van Vuuren AJ, van Dekken H, et al. (2010) Screening for colorectal cancer: randomised trial comparing guaiac-based and immunochemical faecal occult blood testing and flexible sigmoidoscopy. Gut 59: 62–68.
33. Peris M, Espinas JA, Munoz L, Navarro M, Binefa G, et al. (2007) Lessons learnt from a population-based pilot programme for colorectal cancer screening in Catalonia (Spain). Journal of medical screening 14: 81–86.
34. Sanz-Pamplona R, Berenguer A, Cordero D, Mollevi DG, Crous-Bou M, et al. (2014) Aberrant gene expression in mucosa adjacent to tumor reveals a molecular crosstalk in colon cancer. Mol Cancer 13: 46.
35. Warman ML, Abbott M, Apte SS, Hefferon T, McIntosh I, et al. (1993) A type X collagen mutation causes Schmid metaphyseal chondrodysplasia. Nature genetics 5: 79–82.
36. Zheng Q, Zhou G, Morello R, Chen Y, Garcia-Rojas X, et al. (2003) Type X collagen gene regulation by Runx2 contributes directly to its hypertrophic chondrocyte-specific expression in vivo. The Journal of cell biology 162: 833–842.
37. Feng Han Q, Zhao W, Bentel J, Shearwood AM, Zeps N, et al. (2006) Expression of sFRP-4 and beta-catenin in human colorectal carcinoma. Cancer letters 231: 129–137.
38. Kim H, Watkinson J, Varadan V, Anastassiou D (2010) Multi-cancer computational analysis reveals invasion-associated variant of desmoplastic reaction involving INHBA, THBS2 and COL11A1. BMC medical genomics 3: 51.
39. Lascorz J, Forsti A, Chen B, Buch S, Steinke V, et al. (2010) Genome-wide association study for colorectal cancer identifies risk polymorphisms in German familial cases and implicates MAPK signalling pathways in disease susceptibility. Carcinogenesis 31: 1612–1619.
40. Gough MJ, Ruby CE, Redmond WL, Dhungel B, Brown A, et al. (2008) OX40 agonist therapy enhances CD8 infiltration and decreases immune suppression in the tumor. Cancer research 68: 5206–5215.
41. Chapman KB, Prendes MJ, Sternberg H, Kidd JL, Funk WD, et al. (2012) COL10A1 expression is elevated in diverse solid tumor types and is associated with tumor vasculature. Future oncology 8: 1031–1040.
42. Almaawi A, Wang HT, Ciobanu O, Rowas SA, Rampersad S, et al. (2013) Effect of acetaminophen and nonsteroidal anti-inflammatory drugs on gene expression of mesenchymal stem cells. Tissue engineering Part A 19: 1039–1046.
43. Grutzmann R, Molnar B, Pilarsky C, Habermann JK, Schlag PM, et al. (2008) Sensitive detection of colorectal cancer in peripheral blood by septin 9 DNA methylation assay. PLoS One 3: e3759.
44. Lofton-Day C, Model F, Devos T, Tetzner R, Distler J, et al. (2008) DNA methylation biomarkers for blood-based colorectal cancer screening. Clin Chem 54: 414–423.
45. deVos T, Tetzner R, Model F, Weiss G, Schuster M, et al. (2009) Circulating methylated SEPT9 DNA in plasma is a biomarker for colorectal cancer. Clinical chemistry 55: 1337–1346.
46. Chen WD, Han ZJ, Skoletsky J, Olson J, Sah J, et al. (2005) Detection in fecal DNA of colon cancer-specific methylation of the nonexpressed vimentin gene. J Natl Cancer Inst 97: 1124–1132.
47. Korner H, Soreide K, Stokkeland PJ, Soreide JA (2007) Diagnostic accuracy of serum-carcinoembryonic antigen in recurrent colorectal cancer: a receiver operating characteristic curve analysis. Ann Surg Oncol 14: 417–423.
48. Hundt S, Haug U, Brenner H (2007) Blood markers for early detection of colorectal cancer: a systematic review. Cancer epidemiology, biomarkers & prevention: a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology 16: 1935–1953.
49. Bujanda L, Sarasqueta C, Cosme A, Hijona E, Enriquez-Navascues JM, et al. (2013) Evaluation of alpha 1-antitrypsin and the levels of mRNA expression of matrix metalloproteinase 7, urokinase type plasminogen activator receptor and COX-2 for the diagnosis of colorectal cancer. PLoS One 8: e51810.
50. Jung K (2008) Is serum matrix metalloproteinase 9 a useful biomarker in detection of colorectal cancer? Considering pre-analytical interference that may influence diagnostic accuracy. Br J Cancer 99: 553–554; author reply 555.
51. Alvarez-Chaver P, Otero-Estevez O, Paez de la Cadena M, Rodriguez-Berrocal FJ, Martinez-Zorzano VS (2014) Proteomics for discovery of candidate colorectal cancer biomarkers. World J Gastroenterol 20: 3804–3824.
52. Binefa G, Rodriguez-Moranta F, Teule A, Medina-Hayas M (2014) Colorectal cancer: From prevention to personalized medicine. World J Gastroenterol 20: 6786–6808.
53. Newton KF, Newman W, Hill J (2012) Review of biomarkers in colorectal cancer. Colorectal Dis 14: 3–17.
54. Tanaka T, Tanaka M, Ishigamori R (2010) Biomarkers for colorectal cancer. Int J Mol Sci 11: 3209–3225.
55. Tjalsma H (2010) Identification of biomarkers for colorectal cancer through proteomics-based approaches. Expert review of proteomics 7: 879–895.
56. Diamandis EP (2010) Cancer biomarkers: can we turn recent failures into success? Journal of the National Cancer Institute 102: 1462–1467.
57. Hanash SM, Baik CS, Kallioniemi O (2011) Emerging molecular biomarkers—blood-based strategies to detect and monitor cancer. Nature reviews Clinical oncology 8: 142–150.

3. Article 3: *Large differences in global transcriptional regulatory programs of normal and tumor colon cells.*

3.1 Resum en català

La regulació transcripcional té un rol molt important per al correcte funcionament de les cèl·lules. Entre d'altres, s'encarrega de mantenir estats cel·lulars concrets, evitar desordres metabòlics i assegurar l'homeòstasi cel·lular. Per exemple, processos biològics com el desenvolupament i la diferenciació cel·lular, que estan estretament relacionats amb els processos neoplàsics, són clarament conduïts per la regulació gènica. Així, la identificació global de pertorbacions reguladores que participen en l'inici i el desenvolupament dels tumors és un dels grans reptes per a la biologia del càncer. Actualment ja existeixen alteracions específiques que han sigut descrites i anotades, però es necessiten anàlisis exhaustives i a nivell global per obtenir més informació dels canvis transcripcionals implicats en el desenvolupament tumoral.

Els objectius d'aquest estudi van ser, en primer lloc, reconstruir dues xarxes de regulació transcripcional directa a partir de dades de expressió gènica obtingudes mitjançant *microarrays*. Aquestes dades provenen de 100 mostres de teixit tumoral de còlon de pacients amb estadi II, tumors estables en microsatèl·lits i les seves 100 mucoses aparellades adjacents al tumor, però amb un resultat patològicament normal. En segon lloc, es volia dur a terme un anàlisi exhaustiu i a nivell global dels canvis observats en aquestes dues xarxes de regulació, per caracteritzar les diferències als programes de regulació transcripcional de les cèl·lules normals i tumorals del còlon.

Primer de tot es van reconstruir cadascuna de les dues xarxes transcripcionals mitjançant l'algorisme ARACNe, utilitzant remostreig i consolidant les 1000 rèpliques computades com a xarxes consens. Per donar més fiabilitat a l'estudi es va realitzar una validació *in silico* d'aquestes dues xarxes, mitjançant dades experimentals provinents d'una base de dades pública. I malgrat totes les dificultats existents per realitzar aquest tipus de validacions a gran escala amb dades humanes, es va poder trobar un nivell de concordança molt raonable al 38% de les interaccions.

Com a primer resultat es va observar que la xarxa de cèl·lules patològicament normals contenia 1.177 TFs, 5.466 gens diana i 61.226 interaccions transcripcionals directes. En canvi, a la xarxa tumoral es va observar una gran pèrdua d'aquestes interaccions transcripcionals (concretament una reducció del 81%), un menor nombre de TFs (reducció del 47%) i menys gens diana (reducció del 60%), com es pot veure a la Figura 14. A més, sorprenentment es va observar que el silenciament de gens no era un factor determinant en aquesta pèrdua d'activitat reguladora, ja que l'expressió mitjana del gens corresponents a les interaccions perdudes estava essencialment conservada.

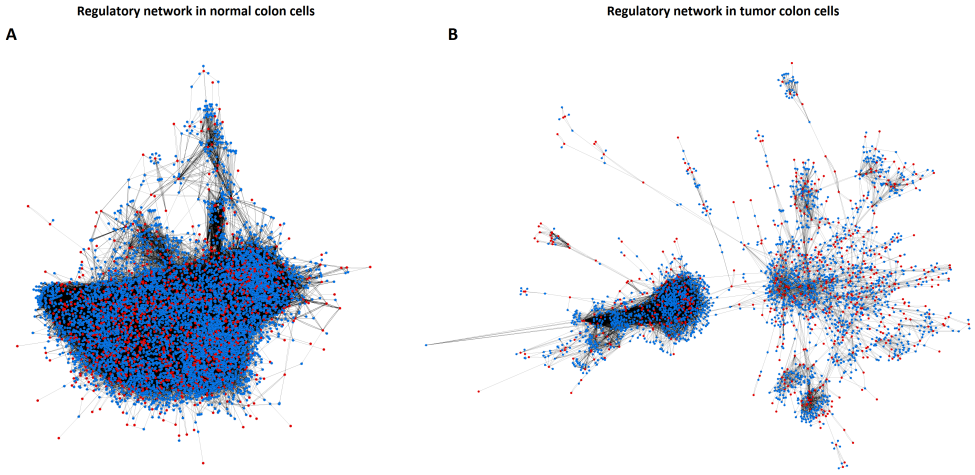

Figura 14. Representació de les xarxes de regulació transcripcional en cèl·lules patològicament normals (A) i tumorals (B). Els nodes vermells corresponen al TFs i els blaus a la resta de gens.

Per altra banda, es va trobar un conjunt específic de 91 TFs que van augmentar notablement la seva connectivitat a la xarxa del tumor. Aquests gens van revelar un programa regulador de la transcripció emergent, específic dels tumors i amb enriquiments funcionals significatius estretament relacionats amb el CCR. A més, un anàlisi de clústers realitzat a la xarxa tumoral va identificar també subxarxes enriquides per vies normalment relacionades amb el càncer (resposta immune, la via de senyalització Wnt, replicació del DNA, adhesió cel·lular, apoptosi i reparació del DNA, entre d'altres). També múltiples vies de metabolisme mostraven agrupació diferencial entre la xarxa tumoral i la normal.

A mode de conclusions, es pot dir que la inferència de les xarxes de regulació transcripcional a nivell del genoma complet, ens ha permès detectar una pèrdua massiva de regulació transcripcional a les cèl·lules tumorals del còlon, no descrita anteriorment. I que a més a més, aquesta pèrdua d'activitat reguladora no és produïda en la majoria dels casos per un silenciament massiu dels gens directament implicats. També s'han descrit alguns TFs concrets que augmenten la seva activitat reguladora als tumors de còlon, tenint així un paper central i molt rellevant a la xarxa. Aquestes troballes permeten una millor comprensió dels programes de regulació transcripcional alterats en el càncer de còlon i podrien ser una nova metodologia molt valuosa per identificar mecanismes amb un paper rellevant en el camp del diagnòstic, el pronòstic i el tractament del càncer de còlon.

Cordero et al. BMC Cancer 2014, 14:708
http://www.biomedcentral.com/1471-2407/14/708

BMC Cancer

RESEARCH ARTICLE Open Access

Large differences in global transcriptional regulatory programs of normal and tumor colon cells

David Cordero[1,2,3†], Xavier Solé[1,2,3†], Marta Crous-Bou[1,2,3], Rebeca Sanz-Pamplona[1,2,3], Laia Paré-Brunet[1,2,3], Elisabet Guinó[1,2,3], David Olivares[1,2,3], Antonio Berenguer[1,2,3], Cristina Santos[2,4], Ramón Salazar[2,4], Sebastiano Biondo[5,6] and Víctor Moreno[1,2,3,6*]

Abstract

Background: Dysregulation of transcriptional programs leads to cell malfunctioning and can have an impact in cancer development. Our study aims to characterize global differences between transcriptional regulatory programs of normal and tumor cells of the colon.

Methods: Affymetrix Human Genome U219 expression arrays were used to assess gene expression in 100 samples of colon tumor and their paired adjacent normal mucosa. Transcriptional networks were reconstructed using ARACNe algorithm using 1,000 bootstrap replicates consolidated into a consensus network. Networks were compared regarding topology parameters and identified well-connected clusters. Functional enrichment was performed with SIGORA method. ENCODE ChIP-Seq data curated in the *hmChIP* database was used for *in silico* validation of the most prominent transcription factors.

Results: The normal network contained 1,177 transcription factors, 5,466 target genes and 61,226 transcriptional interactions. A large loss of transcriptional interactions in the tumor network was observed (11,585; 81% reduction), which also contained fewer transcription factors (621; 47% reduction) and target genes (2,190; 60% reduction) than the normal network. Gene silencing was not a main determinant of this loss of regulatory activity, since the average gene expression was essentially conserved. Also, 91 transcription factors increased their connectivity in the tumor network. These genes revealed a tumor-specific emergent transcriptional regulatory program with significant functional enrichment related to colorectal cancer pathway. In addition, the analysis of clusters again identified subnetworks in the tumors enriched for cancer related pathways (immune response, Wnt signaling, DNA replication, cell adherence, apoptosis, DNA repair, among others). Also multiple metabolism pathways show differential clustering between the tumor and normal network.

Conclusions: These findings will allow a better understanding of the transcriptional regulatory programs altered in colon cancer and could be an invaluable methodology to identify potential hubs with a relevant role in the field of cancer diagnosis, prognosis and therapy.

Keywords: Colon cancer, Gene expression, Gene regulatory networks, Transcription factors, Transcriptional interactions

* Correspondence: v.moreno@iconcologia.net
†Equal contributors
[1]Unit of Biomarkers and Susceptibility, Cancer Prevention and Control Program, Catalan Institute of Oncology (ICO), Av Gran Via 199-203, E-08907 L'Hospitalet de Llobregat, Barcelona, Spain
[2]Colorectal Cancer Group, Bellvitge Biomedical Research Institute (IDIBELL), Barcelona, Spain
Full list of author information is available at the end of the article

Background

Transcriptional regulation has an essential role for proper cell functioning. Gene regulatory programs establish and maintain specific cell states [1], ensure cell homeostasis and avoid metabolic disorders [2]. Genetic regulatory information encoded in DNA binding sites, such as enhancers and promoters, is interpreted by a network of transcription factors (TFs) [3]. Epigenetic events like DNA methylation or histone modifications are regulators of transcription [4,5] and non-coding RNAs such as siRNAs and miRNAs are also involved in gene expression regulation at the post-transcriptional level [6].

Identification of global regulatory perturbations that actively participate in the initiation and maintenance of the tumor state is one of the major challenges in cancer biology [7]. Important processes intimately related to the neoplastic process, such as development and cell differentiation, are widely mediated by gene regulation [8]. Dysregulation of signaling pathways has also been related with tumor growth and cancer progression [9]. Although specific tumor genetic alterations are well described and annotated [10], comprehensive studies are required to obtain more information about the transcriptional programs involved in tumor development. Thus, a global analysis of regulatory network perturbations still remains a fundamental challenge for cancer biology [7].

Recent bioinformatics developments make use of large-scale gene expression datasets to infer genome-wide gene regulatory networks (GRN) [11]. Although not as accurate as methods based on experimental procedures and usually requiring subsequent validation, this approach to computationally-infer regulatory networks can be useful to predict *in-vivo* functions of specific cell types [12]. Diverse methodological approaches to infer GRNs have been proposed, such as regression-based methods, correlation, information-theoretic approaches and Bayesian networks [13]. Among all those, the ARACNe algorithm for the reconstruction of GRNs has been successfully applied to reverse-engineer large-scale transcriptional networks in B-cell leukemia [14,15], neuroblastoma [16], T cell acute lymphoblastic leukemia [17] and prostate cancer [18]. These methodologies have also been applied to analyze and compare GRNs of several human tissues [19]. However, there are a limited number of studies about gene regulatory network inference in colon cancer cells, and these analyses were restricted to a small number of genes or used small sample sizes for the inference [20-23].

The aim of our study is to infer GRNs from transcriptional data obtained for a large sample of stage II colon tumor cells and paired adjacent pathologically normal mucosa, as well as to perform a comprehensive analysis of the changes in the transcriptional regulatory programs related to the tumor phenotype.

Methods

Patients and samples

One hundred patients with an incident diagnosis of colon cancer who were visited at the Bellvitge University Hospital (Barcelona, Spain) between January 1996 and December 2000 were included in the study. Cases were selected to define a homogenous series of patients with stage II, microsatellite-stable, pathology confirmed adenocarcinoma of the colon. All patients underwent radical surgery and had no signs of tumor cells when margins were examined. Fresh samples were collected and frozen by the pathologist from the surgical specimen. Adjacent mucosa was obtained from the proximal margin and was at least 10 cm distant from the tumor lesion. The Clinical Research Ethics Committee (CEIC) of the Bellvitge Hospital approved the study protocol, and all individuals provided written informed consent to participate and for genetic analyses to be done on their samples. The approval number is PR178/11. Additional information about the study and patient samples can be found at http://www.colonomics.org.

Gene expression dataset

Total RNA was isolated from tissue samples of tumor and normal adjacent mucosa using Exiqon's miRCURY™ RNA Isolation Kit (Exiqon, Denmark), according to manufacturer's protocol. Extracted RNA was quantified by NanoDrop® ND-1000 Spectrophotometer (Nanodrop technologies, Wilmington, DE) and stored at −80°C. RNA quality was assessed with RNA 6000 Nano Assay (Agilent Technologies, Santa Clara, CA) following manufacturer's recommendations and was further confirmed by gel electrophoresis. RNA integrity numbers showed good quality (mean = 8.1 for tumors, and 7.5 for adjacent normal). RNA purity was measured with the ratio of absorbance at 260 nm and 280 nm (mean = 1.96, sd = 0.04), with no differences among tissue types.

RNA samples were hybridized onto the Affymetrix Human Genome U219 96-Array Plate platform (Affymetrix, Santa Clara, CA) following Affymetrix standard procedures. Annotation of the array was based on hg19 genome version. A blocked experimental design was implemented to avoid biases due to potential plate effects (i.e. all plates contained the same proportion of normal and tumor samples). After evaluating the quality of the 200 CEL files using Affymetrix standard quality parameters (e.g. level of background noise, labeling and hybridization efficiency, and RNA degradation), 4 arrays (two normal-tumor pairs) were excluded. Therefore, a final dataset of 196 arrays was used for subsequent analyses. Raw data were normalized together using the Robust Multi-array Average (RMA) algorithm [24] implemented in the *affy* package [25] of the Bioconductor suite (http://bioconductor.org). All other analyses were done with R 2.15.1 statistical

computing suite (http://www.R-project.org). A model-based clustering was applied to the full expression dataset in order to detect and remove non-expressed and saturated probe-sets from further analyses.

The complete gene expression dataset was uploaded to the National Center for Biotechnology Information's Gene Expression Omnibus Database with GEO series accession number GSE44076.

Transcription factor selection

The list of TFs used was built by merging two different sources of information. The first one was the manually-curated compilation of human TFs reported by [26]. More specifically, 1,391 TFs classified in Supplementary Information S3 as 'a', 'b' or 'other' were chosen. In order to generate a broader set of putative TF genes, the collection of curated TFs was complemented with an additional set of 1,415 genes that were associated with specific GO terms related to transcription. In particular, genes associated with GO terms (GO:0045449 - Regulation of transcription, GO:0030528 - Transcription regulator activity and GO:0001071 - Nucleic acid binding transcription factor activity) were chosen. The GO database release used was 2011-03-19 accessed from AmiGO version 1.8 [27]. This yielded a set of 2,806 unique TFs, which were represented by 7,811 Affymetrix probe-sets in the expression array that was used.

Inference, representation and analysis of transcriptional regulatory networks

Transcriptional regulatory networks were built using the ARACNe algorithm [15]. Prior to the ARACNe analysis, simulations were performed to model the optimal kernel width that allowed a proper mutual information (MI) estimation in our dataset. The null distribution of the MI was also empirically determined by simulation analysis in order to be able to further identify those significant correlations between TFs and their putative target genes. The significance p-value used as a threshold was 1e-07. ARACNe2 algorithm was run with DPI tolerance set to 0 to remove potential indirect transcriptional interactions from both networks. Remaining parameters were used with their default values. For each network, 1000 bootstrap replicates were performed and summarized to obtain more robust and accurate consensus networks. Only the giant connected component of both networks was considered for downstream analyses. Network visualization, descriptive, simple parameters estimation and figures were performed with Cytoscape software version 2.8.2 [28]. Directed graphs were used to describe networks, in which a regulatory relationship between a TF and a target gene was represented by a directed edge (i.e. arrow) between these two connected nodes, being the origin of the edge the TF. Comprehensive network topology analyses, along with the estimation of complex parameters, were carried out with the Network Analyzer Cytoscape plugin [29]. KEGG pathway enrichment analysis was performed with the SIGORA R package version 0.9.2 and default parameter values [30]. In the analysis of lost edges, a gene was considered to become silenced in the tumor if its average expression level was smaller than 4 and the $\log_2$ fold change between the tumor and the normal expression values was smaller than -1 (i.e. a 2-fold change decrease in the tumor). The analysis of network clusters was performed with the MINE Cytoscape plugin [31]. Only clusters with more than 10 nodes were considered for detailed analysis. Somatic mutation data were obtained from the COSMIC database [10] using the following parameters: large intestine (tissue), all (subtissue), carcinoma (histology), all (subhistology). Only genes with a mutation frequency greater than 5% were considered for further analysis.

In-silico network validation

Gene annotation, (e.g. Ensembl gene id, chromosome, strand, start and end position) was retrieved through *BiomaRt* R/Bioconductor package [32]. For each gene, genomic sequence around the transcription start site +/- 1 kb according to hg18 coordinates was obtained with the *BSgenome* R/Bioconductor package version 1.24.0. The validation analysis was performed using the *hmChIP* database, which contains ChIP-Seq and ChIP-on-chip data from ENCODE experiments that represent more than 10,000,000 protein-DNA interactions [33]. Only the interactions of TFs with at least more than 20 target genes in the normal tissue network were considered for validation. For each TF, the *hmChIP* database was queried providing a list of genomic regions corresponding to the regulatory sequences of their targets in the normal tissue network. Results were rank ordered based on the degree of overlap between the uploaded genomic regions and the peak lists collected by *hmChIP* database from ChIP-Seq and ChIP-on-chip ENCODE datasets. Enrichment ratios and significance p-values for the overlaps were provided by *hmChIP* tool. Benjamini and Hochberg false discovery rates were also reported by the tool to account for multiple testing.

Results

Massive loss of regulatory activity in tumor cells

A large loss of transcriptional interactions was found in the tumor regulatory network (Figure 1, Table 1). The tumor regulatory network contained 47% fewer TFs than the network of normal cells (621 vs. 1,177), as well as 60% fewer target genes (2,190 vs. 5,466). Most nodes disappeared in the tumor network because their expression was completely unrelated to other nodes. Furthermore, the number of direct transcriptional interactions was

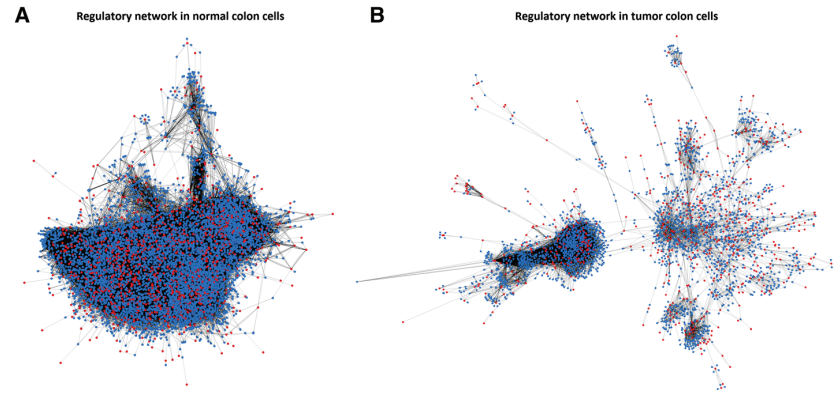

Figure 1 Normal and tumor regulatory networks. Inference and representation of normal **(A)** and tumor **(B)** regulatory networks. Both networks were inferred using microarray expression data from paired normal and tumor colon tissue obtained from the same set of individuals. Red nodes correspond to TFs and blue to non-TFs. Notice that a TF may also be the target gene of another TF. A global loss of transcriptional interactions in the tumor regulatory network is observed.

reduced by 81% (11,585 in the tumor network vs. 61,226 in adjacent normal cells).

Notably, although the node overlap between both networks is large (81% of the tumor nodes are found in the normal network), only 19% of the interactions present in the tumor network are found in the normal network (Figure 2). To visualize both entire networks with Cytoscape [28] or another platform the network representations can be found online (Additional file 1). Additionally, specific TFs and their target genes (or vice versa) can explored in the project website (http://www.colonomics.org/regulatory-networks).

The vast majority of lost edges (76%) show a large decrease in MI but relatively small changes in gene expression (absolute $\log_2$ fold change < 1, Figure 3). This suggests that decreased connectivity in the tumor network was more related to transcriptional dysregulation than to gene silencing. Lost edges in the tumor network were classified into four groups according to their change in MI and gene expression change (Figure 4). Panels A-C contains examples of loss of interaction by either silencing of the TF and/or the target. These groups comprise a small proportion of lost edges (A: 80, 0.2%; B: 1,105, 2.1%; C: 923, 1.7%). Panel D shows a loss of interaction due to a decrease in the correlation, without evidence of TF or target silencing. Remarkably, most of the lost edges in the tumor network (50,882, 96%) belong to pattern D, where the loss of regulatory activity does not depend on major changes in average gene expression levels.

Table 1 Networks descriptive parameters and topological features

	Normal network	Tumor network	Ratio Tumor/Normal
Descriptive parameters			
Nodes	6,643	2,811	0.42
Transcription factors	1,177	621	0.53
Target genes	5,466	2,190	0.40
Edges	61,226	11,585	0.19
Main topological features			
Network diameter	12	17	1.42
Proportion of shortest paths	14%	4%	0.29
Characteristic path length	4.0	5.0	1.25
Average number of neighbors	16.9	7.6	0.45
Multi-edge node pairs	5,204	976	0.19

Loss of robustness in the tumor network was suggested by the comparison of the topological features of both networks, as shown in Table 1. Firstly, a larger distance between nodes in the tumor network was observed for different parameters, such as an increased network diameter, the characteristic path length or the decrease in average shortest paths. Secondly, a lower connectivity in the tumor network was identified according to the values of parameters related to neighborhood, such as the decrease in average number of neighbors and multi-edge node pairs. Furthermore, a characteristic of the tumor network not found on the normal was the existence of a small subset of low connected TFs with a remarkable contribution to minimal shortest paths (closeness centrality, see figure in Additional file 2). Although no significant functional enrichment was found for this set of TFs, these genes may have the potential ability to further disrupt the tumor network by breaking it into multiple disconnected components if some of their incoming our outgoing interactions are further lost. For a full set of figures

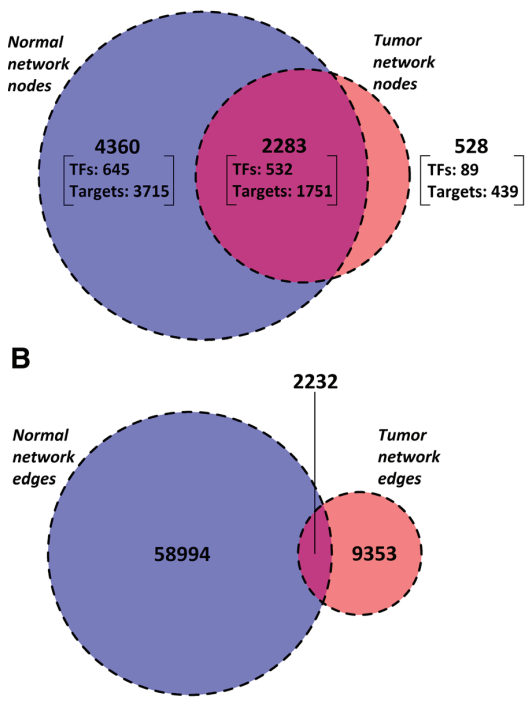

Figure 2 Summary network nodes and edges overlap between normal and tumor networks. Node **(A)** and edge **(B)** overlap between normal and tumor networks. Blue circles correspond to the normal network, red circles correspond to the tumor network, and purple areas correspond to intersections between both networks. Notice the small edge overlap between both networks (19%) even though a large part of the nodes (81%) in the tumor network are present in the normal network.

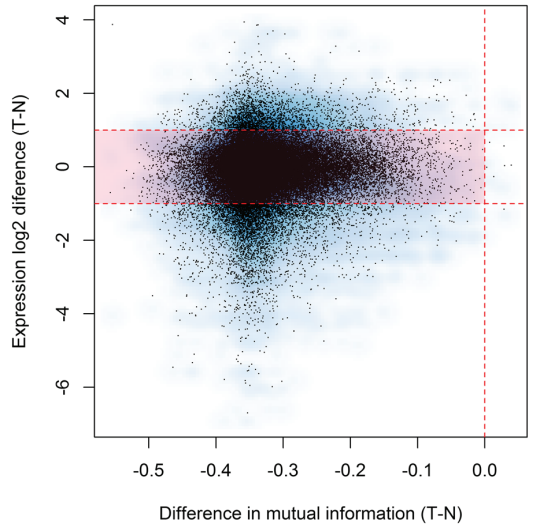

Figure 3 Changes in mutual information vs. expression changes. Each dot corresponds to a lost edge in the tumor network. X-axis represents the difference in mutual information (T-N), while the y-axis contains the expression difference between tumor and normal for either the TF or the target gene of that edge. Thus, every edge is represented by two dots in the plot. The area colored in red, where most of the dots fall, corresponds to lost interactions in the tumor ($\Delta MI < -0.25$) in which there is no transcriptional silencing neither of the TF nor the target gene. The fact that most of the edges (~96%) fall in that region suggests that genetic or epigenetic silencing is not involved in this massive loss of transcriptional regulation in tumor cells.

of other topological features comparing the networks see Additional file 2.

Gain of regulatory activity in tumor cells

Although the tumor network shows a large loss of transcriptional interactions, there are also specific TFs that largely increase their number of target interactions in the tumor network. A total of 91 TFs with increased activity (i.e. out-degree) and 235 up-regulated (i.e. in-degree) target genes were identified in the tumor network. The analysis of gained edges suggests a stronger role of the TFs compared to the targets. Specifically, the 91 TFs with increased activity revealed 2,224 new edges in the tumor network (24 on average, median = 12) while the 235 up-regulated targets only comprise 1,292 new transcriptional interactions (5 on average, median = 4). TFs and target genes that most increase their connectivity in the tumor network are shown in Table 2 (see complete lists in Additional file 3). KEGG pathways [34] enrichment analysis of this set of genes using the SIGORA method [30] revealed that the *Colorectal cancer* pathway (map05210) was significantly overrepresented among these TFs with increased activity (p-value = 8.9e-9). This pathway includes well-known cancer-related genes such as *FOS*, *TGFB3* and *TGFB1* that increased connectivity in the tumor network. In order to evaluate if this gain of regulatory activity in colon tumor cells may be related to somatic mutations we studied the degree distribution (as indicator of regulatory activity) for TFs and target genes, classified as frequently mutated (if present in COSMIC database) or not [10]. We have found that regulatory activity is independent of mutations for TFs. However, target genes included in COSMIC database showed a significant larger regulatory control than other non-mutated genes in tumors (mean in-degree 4.5 in non mutated and 7.7 in mutated, $p = 0.000021$). These differences were not observed in the normal network (mean in-degree 11.3 in non mutated and 12.6 in mutated, $p = 0.16$), indicating that mutated genes tend to loose less regulation or even increase it, since these differences were also true for targets that increased connectivity. Examples of mutated target genes that increase connectivity are *CDH11*, *CFH*, *COL3A1*, *COL6A3* and *COL5A2* (complete list in Additional file 4). These genes are

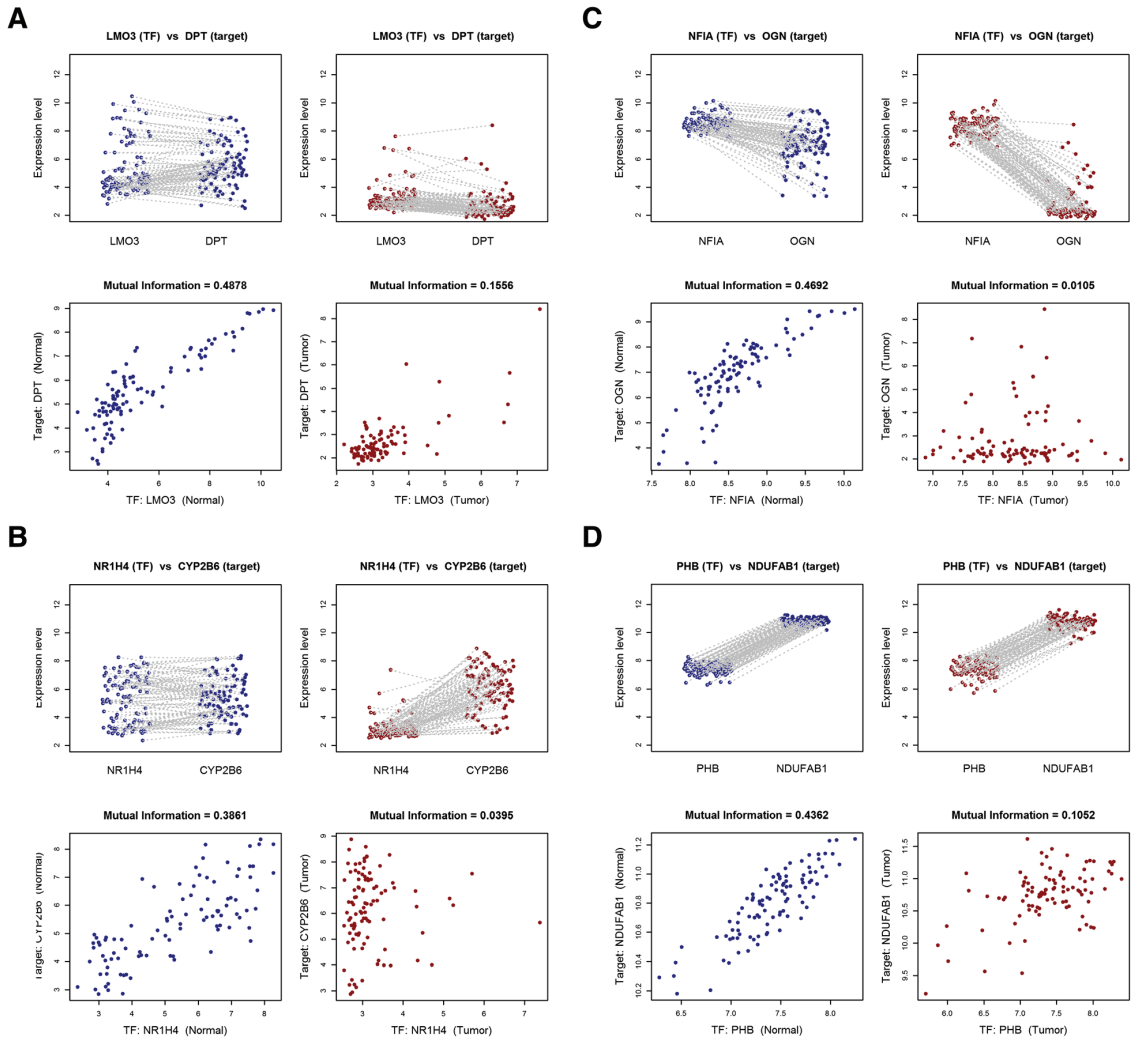

Figure 4 Classification of lost edges. The figure illustrates four examples of loss of correlation in tumor network edges. For each subfigure **(4A-4D)** the upper left plot shows the paired expression values of the TF (left) and the target gene (right) across normal samples. Similarly, the upper right plot contains the expression values across tumor samples. Lower plots show the correlation between the TF (x-axis) and the target gene (y-axis) expression for the normal samples (left) and the tumor samples (right). Blue dots correspond to expression values in normal adjacent mucosa samples and red dots correspond to expression values in tumor samples. **A)** Loss of transcriptional interaction mediated by silencing of both the TF and the target gene simultaneously. This category comprises 0.2% of lost edges (n = 80). **B)** Loss of transcriptional interaction mediated by silencing of the target gene only. This category comprises 2.1% of lost edges (n = 1,105). **C)** Loss of transcriptional interaction mediated by silencing of the TF only. This category comprises 1.7% of lost edges (n = 923). **D)** Loss of transcriptional interaction with no TF or the target gene silencing. About 96% of lost edges in the tumor network (n = 50,882) fall into this last category.

mutated with frequency greater than 5% and show in the tumor network a large increment of regulatory activity.

In-silico network validation with experimental data

Public ChIP-Seq and ChIP-on-chip datasets mainly from the ENCODE project [35] and compiled in the *hmChIP* database were used [33]. In order to avoid biases derived from tumor-specific interactions, only TFs from our normal regulatory network with available datasets from ChIP-Seq or ChIP-on-chip experiments were initially selected for validation. TFs with less than 20 targets in the normal network or showing less than 500 peaks in *hmChIP* database were filtered out to avoid focusing on tissue-specific regulations. Finally 16 TFs and their 1,443 putative target genes were selected for validation. Remarkably, though

Table 2 Nodes with increased activity

TFs that most increase their activity in tumors				
Transcription factor	**Targets in Normal network**	**Targets in Tumor network**	**Gained interactions**	**Ratio T/N**
SNAI2	1	119	118	119.0
MMP14	10	121	111	12.1
AEBP1	103	186	83	1.8
BASP1	43	123	80	2.9
HCLS1	91	170	79	1.9
TFEC	6	84	78	14.0
DKK3	41	112	71	2.7
COL1A1	62	131	69	2.1
CD86	74	141	67	1.9
MAFB	125	189	64	1.5
NOTCH3	18	82	64	4.6
GLI2	37	100	63	2.7
TGFB1	1	61	60	61
GREM1	14	70	56	5.0
HOPX	46	102	56	2.2
Most up-regulated targets in Tumors				
Target gene	**Targets in-degree in Normal**	**Targets in-degree in Tumor**	**Gained interactions**	**Ratio T/N**
NNMT	3	32	29	10.7
CDH11	1	24	23	24.0
RAB31	20	42	22	2.1
MXRA8	3	23	20	7.7
RFTN1	8	28	20	3.5
CFH	3	20	17	6.7
COL3A1	14	31	17	2.2
EMILIN1	12	28	16	2.3
ENTPD1	12	28	16	2.3
MRC2	7	23	16	3.3
STAU1	1	17	16	17.0
AXL	9	24	15	2.7
OLFML2B	10	25	15	2.5
VCAM1	1	15	14	15.0
COL6A3	12	25	13	2.1

The table lists the top 15 TFs and target genes that most increase their activity in the tumor network, sorted by the number of gained interactions. Only nodes that appeared in both networks were considered. See complete lists in Additional file 3.

the experimental datasets were not restricted to colon tissue, 6 out of the 16 TFs (38%) showed significant overrepresentation (enrichment ratio > 1). One additional TF showed marginally significant overrepresentation in the experimental data collected in the *hmChIP* database, as shown in Table 3. This result reinforces the robustness of our inferred networks, which seem to be reasonably capturing transcriptional relationships between TFs and their target genes.

Functional analysis of node clusters

It is known that functionally related genes tend to cluster together in network-defined biological systems (e.g. protein-protein interaction, transcriptional, or co-expression networks). Therefore, we aimed to detect clusters of genes in both the normal and tumor network to identify tumor-specific highly interconnected sub-networks, potentially enriched in relevant biological pathways. The network cluster analysis revealed 42 clusters in the normal network with more than 10 nodes. These included 953 highly interconnected genes. The tumor network included 29 clusters with 871 nodes. The distribution of nodes among clusters was similar for both networks. The list of clusters and enriched pathways (identified by SIGORA method) can be found in Additional file 5. Although most of the clusters in the tumor network were enriched in functions already present in the normal network, some clusters showed tumor-specific significant enrichments in functions with a potential role in tumor development (Table 4). More specifically, clusters 3 and 19 showed an overrepresentation of immune response pathways (e.g., Chemokine signaling pathway, Toll-like receptor signaling pathway, Cytokine-cytokine receptor interaction), and cluster 4 showed enrichment in Wnt signaling proteins. Other clusters, such as 11 and 18, also included significant enrichment of potentially relevant processes such as cell proliferation (e.g. MAPK pathway) or apoptosis, respectively.

Discussion

In this study we have reverse-engineered the transcriptional regulatory networks of both pathologically normal and tumor colon cells obtained from the same set of patients. Using a large-scale gene expression microarray dataset, the ARACNe algorithm was applied to both tissue types independently. ARACNe gives preference to identify direct transcriptional regulatory interactions between TFs and their target genes. When both networks are compared, the most outstanding feature is the considerable loss of transcriptional interactions found in tumor cells (81%), with a global significant decrease in TFs (47%), target genes (60%). The fact that both normal and tumor samples belong to the same set of individuals, as well as the carefully performed experimental design to prevent biases between tissue types, strongly suggests that these large differences between networks are mainly due to the tumor phenotype.

Most of the TFs and target genes involved in disrupted interactions in the tumor network still maintain their expression levels, while only a minor proportion of lost edges may be explained by a complete loss of expression

Table 3 *In-silico* network validation

Transcription factor (Gene Symbol)	# Targets (In normal network)	# Peaks (In hmChIP DB)	Enrichment ratio	p-value	FDR
TCF4	408	46,018	**1.82**	2.0e-07	3.7e-06
NR3C1	246	24,967	0.60	0.12	-
PBX3	186	39,691	0.40	0.0063	0.019
HNF4A	103	32,083	**2.71**	0.00027	0.0016
TCF12	67	54,191	**3.33**	2.0e-06	1.8e-05
RBL2	55	16,395	**2.33**	0.0050	0.018
SUZ12	50	8,742	0.62	0.12	-
ESRRA	42	3,284	1.50	0.37	-
FOXP2	42	44,482	**2.00**	0.043	0.11
MAX	41	16,467	1.80	0.12	-
CDX2	40	24,460	1.38	0.38	-
SRF	39	35,784	1.91	0.052	0.12
STAT1	35	2,804	**3.20**	0.00097	0.0044
FOXA1	32	21,540	0.55	0.062	0.12
NFYB	31	4,630	1.20	1	-
RAD21	26	33,302	1.40	0.50	-

Results provided by hmChIP tool containing ChIP-Seq and ChIP-chip ENCODE experiments [33]. TFs are ordered according to the number of target genes in the normal network. Cells with enrichment ratio in bold highlight significantly overrepresented TFs.

of one or both interactors. This expression silencing may be attributed either to genomic (e.g. DNA deletions, somatic mutations in promoter regions that hinder TF binding, transcript-truncating alterations, etc.) or epigenomic mechanisms (e.g. miRNA-associated transcript degradation, promoter hypermethylation, alterations in chromatin activation and repression marks, etc). On the other hand, disrupted interactions involving TFs and target genes that maintain expression levels in normal and tumor cells may be attributed to multiple reasons: presence or absence of a third-party molecule that could be acting as a post-translational modulator of the TF activity (i.e. phosphorylation, acetylation, ubiquitination) [36], alteration of key co-factors [1], or alterations in promoter regions that could create new TF-binding sites in target genes [37,38]. The small set of genes involved in the loss of interactions through TFs or target gene silencing (~4%) is more likely to belong to currently known altered colon cancer pathways as the Wnt signaling and others, due to apparent under-expression. However, the vast majority of lost edges would not be easy to identify just by exploring the expression values of their TFs or targets genes. We think new and interesting undescribed mechanisms for molecular biology of colon cancer might be related to this gene deregulation without average gene expression change. A potential limitation may be the tumor cellular heterogeneity that could also be contributing to the observed loss of connectivity. While normal mucosa is a relatively homogeneous tissue among subjects, tumors are more heterogeneous, with diverse predominant cellular clones (epithelial, stromal and derived from the immune system). This could result in an apparent global loss of correlation if diverse transcriptional networks were mixed in the tumor.

The network of tumor cells also showed the emergence of a new set of transcriptional interactions that may have an essential role in tumor development and the acquisition of new cellular abilities. Recent studies have demonstrated that the activation of a small regulatory module is necessary and sufficient to initiate and maintain an aberrant phenotypic state in brain tumors [16]. Therefore, network inference approaches could prove effectively useful to uncover new modules and the master regulators that orchestrate malignant transformation. Among the TFs ranked at the top of the list of increased connectivity, our analysis identified colorectal cancer related genes: two oncogenes (*MAFB* [39] and *GLI2* [40]), proliferation-related genes (*NOTCH3* [41] and *TGFB1* [42]), epithelial-mesenchymal transition (*SNAI2* [43]) and the Wnt signaling genes *SFRP4*, *TWIST1*, *SMARCA4* and *DKK3*, potentially involved in colorectal cancer angiogenesis [44]. One remarkable gene with increased activity in the tumor network was *GREM1*. This gene encodes a member of the bone morphogenic protein antagonist family and may play a role in regulating organogenesis, body patterning and tissue differentiation. Interestingly, *GREM1* has been previously related with a locus strongly associated with increased colorectal cancer risk [45]. Moreover, increased expression of *GREM1* has also been recently found in

Table 4 Emergent network clusters in Tumors

Tumor cluster*	Number of genes	Pathway	Adjusted P-value[$]
1	120	Vascular smooth muscle contraction	1.1e-09
2	112	GnRH signaling pathway	5.9e-04
2	112	Staphylococcus aureus infection	4.8e-02
3	70	Chemokine signaling pathway	8.1e-08
3	70	Toll-like receptor signaling pathway	3.1e-07
3	70	Ether lipid metabolism	9.5e-04
4	51	Glycosphingolipid biosynthesis - ganglio series	1.6e-03
4	51	Wnt signaling pathway	1.7e-03
4	51	GnRH signaling pathway	1.3e-02
5	70	Adherens junction	1.9e-04
5	70	Chemokine signaling pathway	4.1e-02
7	44	Tight junction	5.6e-05
7	44	Tryptophan metabolism	2.4e-04
7	44	Glycosaminoglycan biosynthesis - chondroitin sulfate	4.7e-04
8	27	Adherens junction	4.0e-03
9	16	Protein digestion and absorption	4.4e-07
9	16	Adherens junction	5.9e-03
11	16	MAPK signaling pathway	2.1e-15
11	16	Prion diseases	2.4e-03
13	24	Beta-Alanine metabolism	4.4e-04
13	24	NOD-like receptor signaling pathway	9.8e-03
16	32	Glycosaminoglycan biosynthesis - chondroitin sulfate	4.5e-08
18	14	Apoptosis	2.2e-06
18	14	Nucleotide excision repair	1.0e-03
19	14	Cytokine-cytokine receptor interaction	1.4e-02
21	13	Butanoate metabolism	5.6e-05
21	13	Amino sugar and nucleotide sugar metabolism	3.4e-03
22	12	Glutathione metabolism	3.4e-04
23	18	DNA replication	6.7e-06
25	32	Vascular smooth muscle contraction	3.7e-06
28	12	DNA replication	9.6e-05

*Only clusters with significant enriched functions in tumors not already present in normal are shown.
[$]P-value for functional enrichment derived from SIGORA method.

colorectal polyps [46], as well as in the dysplasia to carcinoma transition in colon tumors [47]. Therefore our results suggest that *GREM1* may be mediating its tumorigenic effect by the activation of a large transcriptional program. Furthermore, encouraging results were obtained in the study of the relationship of somatic mutations in colorectal tumors in the set of relevant genes identified through our network approach. Though frequent mutation was independent of regulatory activity for TFs, we observed an association for target genes, with larger regulatory activity among mutated genes. Though this was a correlation analysis using external data from COSMIC database (we do not know if our tumors were actually mutated), it is suggestive that mutated genes trigger a regulatory control in the tumor. The presence of mutations combined with the alteration in their transcriptional regulatory connectivity postulate these genes as strong candidates to be involved in the pathogenesis of colon cancer, and even other type of tumors.

The analysis of network clusters has identified relevant sub-networks of highly connected genes specific of tumors. The regulatory network of normal cells is large and compact. Only 42 clusters have been identified with more than 10 genes. These clusters only account for 14% of the network genes, indicating that there is extensive regulation, but relatively low modularity. The tumor cell, however, has revealed 29 clusters that include 30% of their genes. This is consistent with a more modular organization of the regulatory machinery, which is also evident from the network representation (Figure 1). The functional analysis of these clusters has shown significant enrichment of known tumor-specific pathways: immune response, Wnt signaling, DNA replication, cell adherence, apoptosis, DNA repair, among others (Table 4). Some specific metabolism pathways appear also specifically captured by this analysis of sub-networks, which may be candidate for intervention: glycosphingolipid biosynthesis, tryptophan metabolism, glycosaminoglycan biosynthesis (chondroitin sulfate), beta-alanine metabolism, butanoate metabolism, glutathione metabolism. Obviously, all these functions are present in the normal cell, but they seem enhanced at the transcriptional level in the tumor, in such a way that a large cluster of related genes appear as a relevant entity. In this analysis we have generally focused on the gain of activity in the tumor network rather than on the lost interactions, given the massive loss of tumor network interactions that difficult to detect enriched functions. Despite this intrinsic limitation, we want to emphasize that the transcriptional loss found may influence the emergence of new functionality in the tumor cells. This finding may have a potential impact on the future of cancer molecular biology at level of further experiments and their corresponding biological interpretations.

The inference of GRNs has already been successfully applied to other malignances such as leukemia [14], breast cancer [48,49] or ovarian tumors [50], with relevant findings regarding breast cancer metastasis prognostic markers

or prioritization of druggable gene targets for ovarian cancer. In colorectal cancer some researchers have also explored the reconstruction of GRNs, but with limited approaches to one transcription factor [23] or only tumor tissue [21,22]. To our knowledge, this is the first study in colon cancer that has simultaneously inferred networks for both tumor and adjacent normal cells obtained from the same set of individuals with a consistent methodology that makes both networks totally comparable.

We are aware that computational approaches of network reverse-engineering may suffer from intrinsic limitations. Therefore, we attempted a validation of the network to reinforce the validity of our study. An initial attempt to *in-silico* identify expected TF binding sites in targets was rejected because of the limited number and relative quality of the available TF positional weight matrices both in JASPAR [51] and TRANSFAC Public [52] databases. Other approach to validate the inferred regulatory networks would be to replicate our results in another colon cancer dataset. This has not been possible due to the lack of proper datasets to replicate the findings. The ARACNe's authors emphasize in their papers that a hundred samples is the minimum sample size required to infer transcriptional networks with proper accuracy and they specifically discourage users to apply their algorithm on small datasets [15,53]. The TCGA project [54] only provides 23 normal-tumor colon pairs available and we were unable to find a dataset with a more than 50 samples available after an exhaustive search in the most comprehensive public gene expression databases (GEO and ArrayExpress). Over the last decade, ChIP-on-chip and especially ChIP-Seq assays have become gold standard techniques for large-scale protein-DNA interaction identification. Therefore, ChIP-Seq and ChIP-on-chip datasets from the ENCODE project were used to validate interactions inferred by ARACNe. Since we restricted the potential set of TFs to be validated to those that had more than 20 interactions in the normal network and more than 500 experimentally observed peaks, only a very small part of the network could be tested. However, the obtained results were encouraging since 6 of the 16 tested TFs showed a good level of agreement. The large differences between the number of experimentally detected peaks and the number of inferred target genes for each one of the TFs may suggest a high rate of false negative interactions in our inferred networks, though it is not easy to interpret ChiP data, that provides may peaks that are not necessarily related to direct transcription interactions [55]. Failure in the validation of some TFs might also be partially influenced by the failure of the algorithm to completely remove indirect associations from the network due to high order interactions. In this direction, an extension of the ARACNe algorithm (hARACNe) specifically designed to deal with n-order interactions has been recently released, showing a significant increase in the quality and robustness of the inferred network [56]. Network deconvolution solutions over correlation-based networks have also proven to be successful for this purpose [57]. Due that the large heterogeneity of cell line tissues explored in the ENCODE project, we positively consider the overall observed level of agreement (38%), which is in the same range as previous studies found for other inferred transcriptional networks [14].

Conclusion

The inference of direct transcriptional networks at the whole-genome level has allowed us to detect a predominant loss of transcriptional activity in colon tumor cells, which has not been described before to the best of our knowledge. However, some specific TFs and biological processes related to colon cancer also increased the connectivity and became hubs in the dysregulated tumor network. These findings will allow a better comprehension of the transcriptional regulatory programs altered in colon cancer and could be an invaluable methodology to identify potential hubs with a relevant role in the field of cancer diagnosis, prognosis and therapy.

Additional files

Additional file 1: The two networks representation.

Additional file 2: Complex networks parameters.

Additional file 3 Full list of nodes that increase their activity.

Additional file 4: Genes with altered activity and mutations in COSMIC database.

Additional file 5: Clusters enrichment analysis.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

Conceived the study: DC, XS, VM. Performed analysis: DC, XS, MCB, RSP, LPB, EG, DO, AB, VM. Recruited patients: CS, RS, SB. Wrote the manuscript: DC, XS, VM. Discussed the manuscript: MCB, RSP, LPB, EG, DO, AB, CS, RS, SB. All authors read and approved the final manuscript.

Acknowledgments

The authors would like to thank Ferran Martínez, Adrià Closa, Carmen Atencia, Pilar Medina and Isabel Padrol for their technical assistance. This work was supported by the Catalan Institute of Oncology and the Private Foundation of the Biomedical Research Institute of Bellvitge (IDIBELL), the Instituto de Salud Carlos III grants PI08-1635, PS09-1037, PI11-1439, PIE13/00022 and CIBERESP CB06/02/2005 and the "Acción Transversal del Cancer", the European Commission grant FP7-COOP-Health-2007-B HiPerDART, the Catalan Government DURSI grant 2009SGR1489, the Fundación Privada Olga Torres (FOT), and the AECC (Spanish Association Against Cancer) Scientific Foundation. The "Xarxa de Bancs de Tumors de Catalunya" sponsored by "Pla Director d'Oncologia de Catalunya (XBTC)" helped with sample collection.

Author details

[1]Unit of Biomarkers and Susceptibility, Cancer Prevention and Control Program, Catalan Institute of Oncology (ICO), Av Gran Via 199-203,

E-08907 L'Hospitalet de Llobregat, Barcelona, Spain. [2]Colorectal Cancer Group, Bellvitge Biomedical Research Institute (IDIBELL), Barcelona, Spain. [3]Biomedical Research Centre Network for Epidemiology and Public Health (CIBERESP), Barcelona, Spain. [4]Department of Medical Oncology, Catalan Institute of Oncology (ICO), Barcelona, Spain. [5]Department of General and Digestive Surgery, Colorectal Unit, Bellvitge University Hospital (HUB - IDIBELL), Barcelona, Spain. [6]Department of Clinical Sciences, School of Medicine. University of Barcelona (UB), Barcelona, Spain.

Received: 30 April 2014 Accepted: 17 September 2014
Published: 24 September 2014

References

1. Lee TI, Young RA: **Transcriptional regulation and its misregulation in disease.** *Cell* 2013, **152**(6):1237–1251.
2. Desvergne B, Michalik L, Wahli W: **Transcriptional regulation of metabolism.** *Physiol Rev* 2006, **86**(2):465–514.
3. Kadonaga JT: **Regulation of RNA polymerase II transcription by sequence-specific DNA binding factors.** *Cell* 2004, **116**(2):247–257.
4. Bannister AJ, Kouzarides T: **Regulation of chromatin by histone modifications.** *Cell Res* 2011, **21**(3):381–395.
5. Choy MK, Movassagh M, Goh HG, Bennett MR, Down TA, Foo RS: **Genome-wide conserved consensus transcription factor binding motifs are hyper-methylated.** *BMC Genomics* 2010, **11:**519.
6. Lu J, Clark AG: **Impact of microRNA regulation on variation in human gene expression.** *Genome Res* 2012, **22**(7):1243–1254.
7. Goodarzi H, Elemento O, Tavazoie S: **Revealing global regulatory perturbations across human cancers.** *Mol Cell* 2009, **36**(5):900–911.
8. Ben-Tabou de-Leon S, Davidson EH: **Gene regulation: gene control network in development.** *Annu Rev Biophys Biomol Struct* 2007, **36:**191.
9. Anastas JN, Moon RT: **WNT signalling pathways as therapeutic targets in cancer.** *Nat Rev Cancer* 2013, **13**(1):11–26.
10. Forbes SA, Bindal N, Bamford S, Cole C, Kok CY, Beare D, Jia M, Shepherd R, Leung K, Menzies A, Teague JW, Campbell PJ, Stratton MR, Futreal PA: **COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer.** *Nucleic Acids Res* 2011, **39**(Database issue):D945–950.
11. Bansal M, Belcastro V, Ambesi-Impiombato A, di Bernardo D: **How to infer gene networks from expression profiles.** *Mol Syst Biol* 2007, **3:**78.
12. Deng Y, Johnson DR, Guan X, Ang CY, Ai J, Perkins EJ: **In vitro gene regulatory networks predict in vivo function of liver.** *BMC Syst Biol* 2010, **4:**153.
13. Marbach D, Costello JC, Kuffner R, Vega NM, Prill RJ, Camacho DM, Allison KR, Kellis M, Collins JJ, Stolovitzky G: **Wisdom of crowds for robust gene network inference.** *Nat Methods* 2012, **9**(8):796–804.
14. Basso K, Margolin AA, Stolovitzky G, Klein U, Dalla-Favera R, Califano A: **Reverse engineering of regulatory networks in human B cells.** *Nat Genet* 2005, **37**(4):382–390.
15. Margolin AA, Nemenman I, Basso K, Wiggins C, Stolovitzky G, Dalla Favera R, Califano A: **ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context.** *BMC bioinformatics* 2006, **7 Suppl 1:**S7.
16. Carro MS, Lim WK, Alvarez MJ, Bollo RJ, Zhao X, Snyder EY, Sulman EP, Anne SL, Doetsch F, Colman H, Lasorella A, Aldape K, Califano A, Iavarone A: **The transcriptional network for mesenchymal transformation of brain tumours.** *Nature* 2010, **463**(7279):318–325.
17. Della Gatta G, Palomero T, Perez-Garcia A, Ambesi-Impiombato A, Bansal M, Carpenter ZW, De Keersmaecker K, Sole X, Xu L, Paietta E, Racevskis J, Wiernik PH, Rowe JM, Meijerink JP, Califano A, Ferrando AA: **Reverse engineering of TLX oncogenic transcriptional networks identifies RUNX1 as tumor suppressor in T-ALL.** *Nat Med* 2012, **18**(3):436–440.
18. Aytes A, Mitrofanova A, Lefebvre C, Alvarez MJ, Castillo-Martin M, Zheng T, Eastham JA, Gopalan A, Pienta KJ, Shen MM, Califano A, Abate-Shen C: **Cross-species regulatory network analysis identifies a synergistic interaction between FOXM1 and CENPF that drives prostate cancer malignancy.** *Cancer Cell* 2014, **25**(5):638–651.
19. Li J, Hua X, Haubrock M, Wang J, Wingender E: **The architecture of the gene regulatory networks of different tissues.** *Bioinformatics* 2012, **28**(18):i509–i514.
20. Fu J, Tang W, Du P, Wang G, Chen W, Li J, Zhu Y, Gao J, Cui L: **Identifying microRNA-mRNA regulatory network in colorectal cancer by a combination of expression profile and bioinformatics analysis.** *BMC Syst Biol* 2012, **6:**68.
21. Vineetha S, Chandra Shekara Bhat C, Idicula SM: **Gene regulatory network from microarray data of colon cancer patients using TSK-type recurrent neural fuzzy network.** *Gene* 2012, **506**(2):408–416.
22. Wang X, Gotoh O: **Inference of cancer-specific gene regulatory networks using soft computing rules.** *Gene Regul Syst Biol* 2010, **4:**19–34.
23. Weltmeier F, Borlak J: **A high resolution genome-wide scan of HNF4alpha recognition sites infers a regulatory gene network in colon cancer.** *PLoS One* 2011, **6**(7):e21667.
24. Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, Speed TP: **Exploration, normalization, and summaries of high density oligonucleotide array probe level data.** *Biostatistics* 2003, **4**(2):249–264.
25. Gautier L, Cope L, Bolstad BM, Irizarry RA: **affy–analysis of Affymetrix GeneChip data at the probe level.** *Bioinformatics* 2004, **20**(3):307–315.
26. Vaquerizas JM, Kummerfeld SK, Teichmann SA, Luscombe NM: **A census of human transcription factors: function, expression and evolution.** *Nat Rev Genet* 2009, **10**(4):252–263.
27. Carbon S, Ireland A, Mungall CJ, Shu S, Marshall B, Lewis S: **AmiGO: online access to ontology and annotation data.** *Bioinformatics* 2009, **25**(2):288–289.
28. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T: **Cytoscape: a software environment for integrated models of biomolecular interaction networks.** *Genome Res* 2003, **13**(11):2498–2504.
29. Doncheva NT, Assenov Y, Domingues FS, Albrecht M: **Topological analysis and interactive visualization of biological networks and protein structures.** *Nat Protoc* 2012, **7**(4):670–685.
30. Foroushani AB, Brinkman FS, Lynn DJ: **Pathway-GPS and SIGORA: identifying relevant pathways based on the over-representation of their gene-pair signatures.** *PeerJ* 2013, **1:**e229.
31. Rhrissorrakrai K, Gunsalus KC: **MINE: Module identification in networks.** *BMC bioinformatics* 2011, **12:**192.
32. Durinck S, Spellman PT, Birney E, Huber W: **Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt.** *Nat Protoc* 2009, **4**(8):1184–1191.
33. Chen L, Wu G, Ji H: **hmChIP: a database and web server for exploring publicly available human and mouse ChIP-seq and ChIP-chip data.** *Bioinformatics* 2011, **27**(10):1447–1448.
34. Kanehisa M, Goto S, Sato Y, Furumichi M, Tanabe M: **KEGG for integration and interpretation of large-scale molecular data sets.** *Nucleic Acids Res* 2012, **40**(Database issue):D109–114.
35. Encode Project Consortium, Bernstein BE, Birney E, Dunham I, Green ED, Gunter C, Snyder M: **An integrated encyclopedia of DNA elements in the human genome.** *Nature* 2012, **489**(7414):57–74.
36. Wang K, Saito M, Bisikirska BC, Alvarez MJ, Lim WK, Rajbhandari P, Shen Q, Nemenman I, Basso K, Margolin AA, Klein U, Dalla-Favera R, Califano A: **Genome-wide identification of post-translational modulators of transcription factor activity in human B cells.** *Nat Biotechnol* 2009, **27**(9):829–839.
37. Horn S, Figl A, Rachakonda PS, Fischer C, Sucker A, Gast A, Kadel S, Moll I, Nagore E, Hemminki K, Schadendorf D, Kumar R: **TERT promoter mutations in familial and sporadic melanoma.** *Science* 2013, **339**(6122):959–961.
38. Huang FW, Hodis E, Xu MJ, Kryukov GV, Chin L, Garraway LA: **Highly recurrent TERT promoter mutations in human melanoma.** *Science* 2013, **339**(6122):957–959.
39. Suzuki A, Iida S, Kato-Uranishi M, Tajima E, Zhan F, Hanamura I, Huang Y, Ogura T, Takahashi S, Ueda R, Barlogie B, Shaughnessy J Jr, Esumi H: **ARK5 is transcriptionally regulated by the Large-MAF family and mediates IGF-1-induced cell invasion in multiple myeloma: ARK5 as a new molecular determinant of malignant multiple myeloma.** *Oncogene* 2005, **24**(46):6936–6944.
40. Ruiz i Altaba A: **Hedgehog signaling and the Gli code in stem cells, cancer, and metastases.** *Sci Signal* 2011, **4**(200):pt9.
41. Katoh M: **Notch signaling in gastrointestinal tract (review).** *Int J Oncol* 2007, **30**(1):247–251.
42. Biasi F, Tessitore L, Zanetti D, Cutrin JC, Zingaro B, Chiarpotto E, Zarkovic N, Serviddio G, Poli G: **Associated changes of lipid peroxidation and transforming growth factor beta1 levels in human colon cancer during tumour progression.** *Gut* 2002, **50**(3):361–367.
43. Wang Y, Ngo VN, Marani M, Yang Y, Wright G, Staudt LM, Downward J: **Critical role for transcriptional repressor Snail2 in transformation by**

oncogenic RAS in colorectal carcinoma cells. *Oncogene* 2010, **29**(33):4658–4670.
44. Zitt M, Untergasser G, Amberger A, Moser P, Stadlmann S, Muller HM, Muhlmann G, Perathoner A, Margreiter R, Gunsilius E, Ofner D: **Dickkopf-3 as a new potential marker for neoangiogenesis in colorectal cancer: expression in cancer tissue and adjacent non-cancerous tissue.** *Dis Markers* 2008, **24**(2):101–109.
45. Jaeger E, Webb E, Howarth K, Carvajal-Carmona L, Rowan A, Broderick P, Walther A, Spain S, Pittman A, Kemp Z, Sullivan K, Heinimann K, Lubbe S, Domingo E, Barclay E, Martin L, Gorman M, Chandler I, Vijayakrishnan J, Wood W, Papaemmanuil E, Penegar S, Qureshi M, Farrington S, Tenesa A, Cazier JB, Kerr D, Gray R, Peto J, Dunlop M, *et al*: **Common genetic variants at the CRAC1 (HMPS) locus on chromosome 15q13.3 influence colorectal cancer risk.** *Nat Genet* 2008, **40**(1):26–28.
46. Jaeger E, Leedham S, Lewis A, Segditsas S, Becker M, Cuadrado PR, Davis H, Kaur K, Heinimann K, Howarth K, East J, Taylor J, Thomas H, Tomlinson I: **Hereditary mixed polyposis syndrome is caused by a 40-kb upstream duplication that leads to increased and ectopic expression of the BMP antagonist GREM1.** *Nat Genet* 2012, **44**(6):699–703.
47. Galamb O, Wichmann B, Sipos F, Spisak S, Krenacs T, Toth K, Leiszter K, Kalmar A, Tulassay Z, Molnar B: **Dysplasia-carcinoma transition specific transcripts in colonic biopsy samples.** *PLoS One* 2012, **7**(11):e48547.
48. Ahmad FK, Deris S, Othman NH: **The inference of breast cancer metastasis through gene regulatory networks.** *J Biomed Inform* 2012, **45**(2):350–362.
49. Demicheli R, Coradini D: **Gene regulatory networks: a new conceptual framework to analyse breast cancer behaviour.** *Ann Oncol* 2011, **22**(6):1259–1265.
50. Madhamshettiwar PB, Maetschke SR, Davis MJ, Reverter A, Ragan MA: **Gene regulatory network inference: evaluation and application to ovarian cancer allows the prioritization of drug targets.** *Genome Med* 2012, **4**(5):41.
51. Sandelin A, Alkema W, Engstrom P, Wasserman WW, Lenhard B: **JASPAR: an open-access database for eukaryotic transcription factor binding profiles.** *Nucleic Acids Res* 2004, **32**(Database issue):D91–94.
52. Matys V, Kel-Margoulis OV, Fricke E, Liebich I, Land S, Barre-Dirrie A, Reuter I, Chekmenev D, Krull M, Hornischer K, Voss N, Stegmaier P, Lewicki-Potapov B, Saxel H, Kel AE, Wingender E: **TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes.** *Nucleic Acids Res* 2006, **34**(Database issue):D108–110.
53. Margolin AA, Wang K, Lim WK, Kustagi M, Nemenman I, Califano A: **Reverse engineering cellular networks.** *Nat Protoc* 2006, **1**(2):662–671.
54. Cancer Genome Atlas Network: **Comprehensive molecular characterization of human colon and rectal cancer.** *Nature* 2012, **487**(7407):330–337.
55. Levitsky VG, Kulakovskiy IV, Ershov NI, Oschepkov DY, Makeev VJ, Hodgman TC, Merkulova TI: **Application of experimentally verified transcription factor binding sites models for computational analysis of ChIP-Seq data.** *BMC Genomics* 2014, **15**(1):80.
56. Jang IS, Margolin A, Califano A: **hARACNe: improving the accuracy of regulatory model reverse engineering via higher-order data processing inequality tests.** *Interface Focus* 2013, **3**(4):20130011.
57. Feizi S, Marbach D, Medard M, Kellis M: **Network deconvolution as a general method to distinguish direct dependencies in networks.** *Nat Biotechnol* 2013, **31**(8):726–733.

doi:10.1186/1471-2407-14-708
Cite this article as: Cordero *et al.*: Large differences in global transcriptional regulatory programs of normal and tumor colon cells. *BMC Cancer* 2014 **14**:708.

Discussió

En aquest apartat es presenta, en primer lloc, una discussió individual actualitzada de cadascun dels tres articles que composen el nucli d'aquesta tesi. Per cada treball s'han destacat els principals resultats en relació a les hipòtesis de treball. Cada article també s'ha situat en el context de les publicacions relacionades que l'han succeït. Més endavant, es presenta també un apartat sobre els plans de futur i les direccions dels nous projectes relacionats amb aquests tres treballs. Per últim, s'inclou un apartat sobre el impacte en el CCR que pot tenir tot el treball desenvolupat en aquesta tesi.

1. Discussió de l'article 1: *Gene expression differences between colon and rectum tumors.*

Un tumor es considera de recte quan es troba a una distància inferior a 15 cm de l'esfínter anal. Els estudis de recerca en CCR sovint tracten els tumors de còlon i de recte de manera conjunta, com una única entitat. Tot i així, des d'un punt de vista clínic, els tumors de còlon i de recte són habitualment considerats de manera diferenciada en alguns aspectes [24]. Tots dos tipus de tumors es beneficien de quimioteràpia adjuvant, mentre que la radioteràpia només està indicada en els tumors de recte localment avançats [33]. Per altra banda, el càncer de recte mostra una major taxa de recaiguda local i de metàstasi pulmonar, mentre que el càncer de còlon té una major probabilitat de disseminació hepàtica i un pronòstic lleugerament més favorable [82]. Tot i les particularitats d'ambdues localitzacions tumorals, únicament s'han descrit un nombre reduït de diferències menors a nivell epidemiològic i molecular. El consum d'alcohol s'associa a un major risc de desenvolupar càncer de recte que de còlon [15]. Altres factors de risc associats a la dieta, en canvi, han suggerit una major inconsistència [126]. A nivell molecular, només s'han pogut identificar diferències en l'expressió d'un conjunt molt reduït de gens (*TP53*, *COX2*) i proteïnes (CTNNB1) [127].

La nostra anàlisi combinada de quatre conjunts de dades provinents de quatre estudis independents, comprenent un total de 560 mostres, suggereix que hi ha certes diferències d'expressió identificables entre els CCR estables en microsatèl·lits que sorgeixen en diferents localitzacions de l'intestí gros. El nombre de gens identificats com a diferencialment expressats entre les diferents localitzacions tumorals va ser baix, però notablement la majoria d'aquests pertanyen a la família de gens *HOX*. A nivell de les diverses localitzacions, les diferencies més evidents en els perfils d'expressió es van trobar entre el còlon dret i el còlon esquerre o el recte. De fet, els perfils d'expressió dels tumors del còlon esquerre i del recte mostren patrons transcripcionals molt semblants.

Una de les principals fortaleses del nostre treball és la integració analítica de 560 perfils d'expressió obtinguts gràcies als repositoris públics de dades genòmiques. Donada la magnitud de les diferències identificades pel nostre estudi, que és el que aglutina un major

nombre de mostres analitzades, postulem que la majoria de treballs previs d'aquest tipus possiblement no aconsegueixen detectar diferències significatives en els patrons d'expressió de les diferents localitzacions tumorals a causa d'una manca de poder estadístic per la seva menor mida mostral. Un altre problema que també pot aparèixer als estudis efectuats amb un nombre petit de mostres és que les diferències identificades siguin particulars de la selecció dels casos analitzats i no reflecteixin la realitat del motiu estudiat. En aquest sentit, el fet d'haver dut a terme una anàlisi agrupada amb un total de 560 mostres ens proporciona prou capacitat per detectar diferències de 0,5 unitats de desviació estàndard. Generalment aquestes diferències es consideren petites per a les anàlisis clàssiques de *microarrays* d'expressió. Un altre aspecte amb el qual es va tenir molta cura és el control de la possible heterogeneïtat existent entre els diferents estudis inclosos. Això ens va permetre identificar les especificitats de cadascun dels estudis i minimitzar l'impacte de possibles biaixos causats per aquest fet. Segons les nostres anàlisis, l'heterogeneïtat entre els diferents estudis inclosos era baixa, ja que només 12 sondes de les gairebé 15.000 explorades van mostrar una heterogeneïtat significativa. A més, aquesta no podia atribuir-se a un estudi específic, ni tampoc cap d'aquestes 12 sondes que mostraven heterogeneïtat es corresponien amb els gens identificats com a diferencialment expressats als resultats. Per tant, vam considerar els quatre estudis inclosos a les nostres anàlisis com una sèrie homogènia per buscar diferències entre els perfils d'expressió de còlon i els de recte. A més, les dades de *microarrays* de totes les mostres finalment incloses prèviament van ser sotmeses a uns estrictes controls de qualitat tant tècnics (intensitat mitjana, intensitat de fons, etc.) com biològics (degradació, concentració, etc.), que van fer d'aquestes mostres un excel·lent conjunt d'anàlisi.

Una de les principals debilitats de l'estudi és el fet de no disposar d'anàlisis per tumors amb inestabilitat de microsatèl·lits. És conegut que els tumors amb inestabilitat de microsatèl·lits exhibeixen un patró d'expressió gènica diferent [128]. Addicionalment, des de fa temps també es coneix que tenen una forta associació amb la localització del tumor [129]. Per tant, aquestes dues premisses impliquen que les anàlisis dels tumors amb inestabilitat de microsatèl·lits s'haurien de tractar amb molta cura o realitzar-les de manera independent, per evitar biaixos. Malgrat tot, finalment es va optar per no incloure aquest tipus de tumors a les anàlisis del nostre treball ja que, com hem explicat anteriorment, el nombre de mostres disponibles seria insuficient per poder proporcionar resultats amb precisió. Una altra mancança del treball és el fet de no haver pogut disposar d'altres tipus de dades amb un nombre de mostres suficient i de bona qualitat per haver realitzat unes anàlisis d'aquestes característiques. Potser hagués estat interessant contrastar els nostres resultats amb anàlisis de dades epigenètiques o dades de proteòmica. A nivell epigenètic existeixen diversos treballs que han reportat diferències més que evidents entre les diferents localitzacions de l'intestí, a la mucosa sana del còlon [130], en CCR [99] i inclús en adenomes [131]. De fet, aquest últim article va citar el nostre treball perquè trobava a

nivell epigenètic en adenomes colorectals, resultats similars als nostres respecte als gens *HOX*.

Després de la publicació de l'article hem continuat fent revisió periòdica dels possibles treballs i de les possibles opinions d'altres autors relacionades amb aquest tema. Fins a dia d'avui, el nostre treball ha acumulat 12 cites d'altres articles indexats al Pubmed, que parlen de diversos aspectes. Per exemple, un estudi compara els perfils proteics a les diferents localitzacions del còlon i del recte, en teixit sa, en adenomes i en pacients amb síndrome de Lynch [132]. Un altre estudi que intenta caracteritzar les diferències a nivell mutacional entre el còlon i el recte [133]. Un meta-anàlisi que avalua la infiltració immune en múltiples teixits i tumors [134]. Diversos articles relacionats amb el pronòstic del CCR en funció de variables clíniques com pot ser la localització del tumor [135] o la presència de mutacions específiques [136]. També, diversos articles relacionats amb el risc de patir CCR [137, 138]. Addicionalment, hi ha tot un conjunt d'altres articles que referencien el nostre treball per donar suport al fet d'extrapolar els resultats trobats en perfils d'expressió de còlon al recte o per utilitzar els perfils d'expressió del còlon i del recte de manera conjunta.

Un fet important és que, 8 mesos després de la publicació del nostre article, el consorci responsable del projecte TCGA va publicar a la revista Nature l'article referent al CCR [99]. En aquest article els autors també afirmen que segons les seves anàlisis, els adenocarcinomes no-hipermutats de còlon i de recte són pràcticament indistingibles a nivell genòmic. Quan els autors parlen d'adenocarcinomes hipermutats es refereixen principalment a tumors que presenten inestabilitat en microsatèl·lits, que nosaltres també vam excloure de les nostres anàlisis. Així, amb la publicació de l'article del TCGA, la direcció dels nostres resultats va rebre una aclaparadora corroboració.

De forma general, aquests resultats impliquen que, potser, les diferències anatòmiques són rellevants per al maneig clínic del CCR, però els perfils d'expressió dels tumors estables en microsatèl·lits del còlon i del recte són molt similars. En canvi, hi ha indicis que a altres nivells, com per exemple l'epigenètica o la proteòmica, els resultats poden ser diferents. Malgrat tot, nosaltres pensem que els resultats obtinguts poden tenir importants implicacions en el disseny i la interpretació dels futurs estudis de CCR.

2. Discussió de l'article 2: *Discovery and validation of new potential biomarkers for early detection of colon cancer.*

El pronòstic del CCR està estretament relacionat amb l'estadi en el moment del diagnòstic [32], pel que la detecció precoç és actualment la manera més efectiva de reduir la mortalitat [38]. Sobre l'ús de la colonoscòpia com a prova de cribratge hi ha una certa controvèrsia, a causa dels seus costos, els riscos associats i la seva baixa acceptació [139]. El

TSOF ha demostrat una reducció en la mortalitat [47], tot i que la seva sensibilitat i especificitat són encara millorables. Com a conseqüència, la identificació de nous biomarcadors per al diagnòstic precoç del càncer de còlon que puguin ser utilitzats com a prova no invasiva i econòmica es manté encara com un objectiu rellevant dintre de la salut pública. Per exemple, la detecció precisa de proteïnes secretades específicament per les cèl·lules tumorals del càncer de còlon en fluids biològics de fàcil obtenció com el sèrum, seria ideal. El problema és que la identificació d'aquestes proteïnes de forma massiva té un cost econòmic molt elevat i pot portar molt de temps, ja que les tècniques de proteòmica són costoses i encara no permeten avaluar totes les proteïnes simultàniament. En el treball presentat es va proposar una estratègia de validació seqüencial mitjançant la qual es poden trobar nous candidats a biomarcadors de forma assequible.

El nostre treball d'expressió gènica al teixit de còlon ha confirmat estudis anteriors en que un gran nombre de gens estan desregulats al tumor, en comparació amb la seva mucosa normal adjacent. A partir dels més de 20.000 gens interrogats a la matriu d'expressió i després dels diferents filtres emprats amb diversos criteris restrictius, com es pot veure a la secció de mètodes de l'article, han estat identificats 505 candidats a biomarcadors. Tots aquests tenen resultats estadísticament significatius i una alta capacitat per discriminar entre el tumor i les mostres normals adjacents. Com encara disposàvem d'un nombre elevat de candidats, per prioritzar es va utilitzar un conjunt de filtres basats en la consistència existent amb altres publicacions i conjunts de dades públiques, la baixa expressió en altres teixits i que fossin gens amb proteïnes potencialment secretables. Com a conseqüència de l'elevat cost de la identificació de proteïnes en sèrum, sobre aquest conjunt de candidats més acotat es va dur a terme una validació tècnica, per així reduir el màxim el nombre de falsos positius. Els resultats obtinguts mitjançant una altra tècnica de quantificació del mRNA (RT-qPCR), van mostrar en tots els gens avaluats una bona reproductibilitat dels nivells d'expressió mesurats inicialment mitjançant *microarrays*. A continuació, vam testar nou dels millors candidats en sèrum mitjançant kits comercials d'ELISAs. En aquest punt vam incloure també un conjunt de sèrums de pacients amb adenomes, ja que aquestes lesions precursores són molt importants en el context del cribratge de CCR. Els nivells de la proteïna COL10A1 en sèrum mostraven diferències clarament significatives entre els casos de càncer de còlon i els controls ($p=3,2e-06$). L'àrea sota la corba ROC va ser de 0,76, que fa que el COL10A1 sigui un biomarcador de diagnòstic prometedor. Per exemple, en el punt de tall de 280 ng/mL arribem a obtenir una sensibilitat de 0,63 i una especificitat de 0,85, per a la detecció del càncer de còlon o dels adenomes. També la proteïna MMP7, prèviament reportada per altres autors [140], mostrava en els nostres resultats algunes diferències en els sèrums dels adenomes ($p=0,0092$), però no en els casos.

Un punt fort del disseny del nostre estudi és la inclusió d'un conjunt de mostres de mucosa sana de còlon, provinent de controls sense càncer ($n=50$). Això ens ha permès identificar

gens addicionals, que no mostraven diferències d'expressió entre el teixit tumoral i el seu normal adjacent, en pacients amb càncer de còlon. Gràcies a la inclusió d'aquestes mostres també podem confirmar que els gens sobreexpressats en els tumors no mostrin alts nivells d'expressió en el teixit de còlon lliure de càncer, fet que podria limitar el seu potencial ús com a biomarcadors. El nostre grup ja va descriure prèviament que l'expressió gènica de la mucosa del còlon patològicament normal, però adjacent a un tumor en un pacient amb CCR, està alterada de manera significativa en comparació amb la mucosa del còlon d'un pacient sense càncer [141]. Aquest fet reforça la necessitat d'incloure mostres de teixit d'individus sense càncer en projectes destinats a trobar biomarcadors de diagnòstic de càncer de còlon.

D'altra banda, una de les principals limitacions d'aquest estudi ha estat el fet de no poder haver pogut testar més candidats en sèrum, així com també emprar un nombre més elevat de mostres per cada candidat. Com el treball ha seguit un procés de selecció seqüencial, en resultats intermedis disposàvem d'un conjunt més ampli de candidats prometedors que, de moment, no han estat validats. Aquest fet està principalment relacionat amb els costos en temps i material que això suposa. Per dur a terme les ELISAs utilitzant sèrum com a material de partida hi ha diferents aproximacions. La més costosa, i pràcticament inviable per a un projecte d'identificació de biomarcadors candidats, seria la síntesi del propi anticòs per testar posteriorment la presència o no d'una determinada proteïna. Altres alternatives consisteixen en posar a punt un anticòs comercial en assajos a mida per tal que siguin quantificables, o directament utilitzar kits comercials creats específicament per quantificar una proteïna d'interès. Totes aquestes diferents alternatives tenen les seves avantatges i inconvenients, però en tot cas sempre segueixen una relació inversa entre un procés llarg o costós econòmicament.

Un cop el treball va ser publicat, i com a conseqüència de la identificació d'un potencial biomarcador per al diagnòstic precoç del càncer de còlon o dels adenomes, el nostre grup d'investigació, juntament amb la institució, va tramitar una sol·licitud de patent per aquest descobriment. Tot seguit, vam començar a recollir de manera prospectiva mostres de sèrums per poder disposar d'una sèrie molt més amplia on testar amb un major poder estadístic el potencial candidat trobat. La sèrie final estava formada per un total de 480 mostres que passaven tots els controls de qualitat requerits. Aquestes contretament es distribuïen en 198 sèrums de controls sense càncer, 68 sèrums de pacients amb adenomes i 214 sèrums de pacients amb tumors de còlon. Per a cadascuna de les mostres vam procedir a realitzar la quantificació de la proteïna COL10A1 en les mateixes condicions en que ho vam fer a l'article (ELISA). El resultat d'aquestes anàlisis va confirmar que els sèrums dels adenomes tenien una concentració mitjana significativament diferent ($p=0,023$) a la dels controls sense càncer. També els casos amb tumors al còlon mostraven una concentració mitjana significativament diferent ($p=0,0024$) comparada amb la dels sèrums dels controls sense càncer. Tot i així, com es pot veure a la Figura 15A, la dispersió i la superposició dels

valors entre els diferents grups havia augmentat comparat amb els resultats inicials. En conseqüència, la diferència de mitjanes i el poder de classificació s'havia reduït, com es pot veure a la Figura 15B, que presenta un àrea sota la corba ROC de 0,58 per a la detecció del càncer de còlon o dels adenomes.

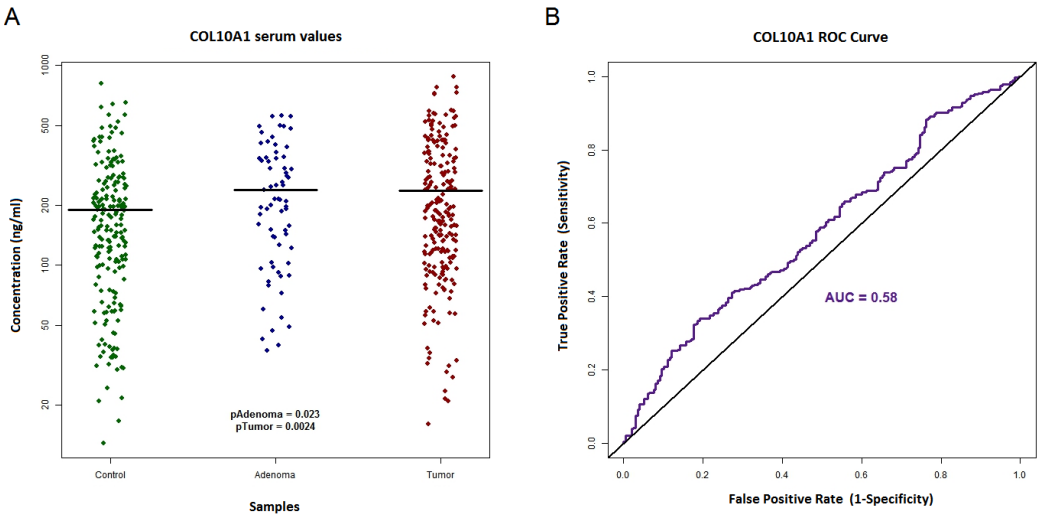

Figura 15. (A) Valors de la concentració en sèrum de la proteïna COL10A1 en una sèrie ampliada de mostres (n=480). (B) Corba ROC de la capacitat predictiva de la proteïna COL10A1 per a la detecció d'adenomes o casos de càncer de còlon.

A l'ampliar considerablement el nombre de mostres dels nostres experiments, la magnitud de les diferències entre mitjanes ha disminuït. Tota la resta de condicions de l'experiment han estat replicades sense cap alteració, des del procediment experimental per quantificar la proteïna, fins l'investigador que ha dut a terme tots els experiments. Com a mesura adicional, totes les mostres utilitzades procedeixen del mateix biobanc, és a dir, han sigut emmagatzemades de la mateixa manera i processades seguint els mateixos protocols. Aquí nosaltres pensem que el que ha passat és que hem observat un efecte de regressió cap a la mitjana, fenomen que ja fa molt de temps que està descrit [142] i que en el disseny i interpretació d'experiments científics s'hauria de tenir sempre en compte per evitar fer inferències incorrectes [143]. El fenomen consisteix en que, teòricament, si observem una variable extrema a una primera mesura, molt probablement aquesta observació estarà més a prop de la mitjana en una segona mesura i paradoxalment, si va ser extrema a la segona mesura, haurà d'haver estat més a prop de la mitjana a la primera mesura. D'aquesta manera, durant la validació inicial dels nivells de COL10A1 en un nombre més reduït de mostres vam observar unes majors diferències entre grups possiblement a causa d'una combinació de dos factors: la capacitat real del propi marcador per discriminar entre casos i controls, i una major influència de l'atzar com a conseqüència d'una mida mostral insuficient.

Els resultats d'aquest treball, en els quals s'han presentat diverses proteïnes candidates a biomarcador, posen de manifest que la metodologia desenvolupada és prometedora, però que cal seguir aprofundint per identificar un biomarcador amb una adequada capacitat diagnòstica a nivell poblacional. Això permetria implementar una prova econòmica en sèrum, per al cribratge del càncer de còlon. A més, creiem que d'aquesta manera es podrien millorar les taxes d'acceptabilitat i de participació per contribuir així a disminuir la gran càrrega poblacional d'aquesta malaltia.

3. Discussió de l'article 3: *Large differences in global transcriptional regulatory programs of normal and tumor colon cells.*

La regulació transcripcional té un rol molt important en el correcte funcionament de les cèl·lules. S'encarrega de mantenir estats cel·lulars concrets, evitar desordres metabòlics i assegurar l'homeòstasi cel·lular. La desregulació dels programes transcripcionals pot causar una gran quantitat de malalties, entre elles càncer [144]. Actualment ja hi ha alteracions específiques que han sigut descrites i anotades, però continua sent un gran repte per a la biologia del càncer la identificació global de pertorbacions reguladores que participen en la iniciació i el desenvolupament dels tumors [145].

En aquest estudi hem reconstruït, mitjançant enginyeria inversa, les dues xarxes de regulació transcripcional de cèl·lules tumorals del còlon i de cèl·lules patològicament normals adjacents al tumor, obtingudes del mateix conjunt de pacients. Es va utilitzar un conjunt de dades d'expressió gènica obtingues mitjançant *microarrays*, sobre el qual es va aplicar, de manera independent a tots dos tipus de teixits, l'algorisme ARACNe. Quan es van comparar les dues xarxes, la característica més sorprenent trobada va ser la considerable pèrdua d'interaccions a les cèl·lules tumorals (81%), amb un descens significatiu de TFs (47%), i de gens diana (60%). A més, la majoria dels TFs i dels gens diana implicats a les interaccions desaparegudes en la xarxa dels tumors mantenen els seus nivells d'expressió, mentre que només una petita proporció de las arestes perdudes (~4%) poden ser explicades per una completa pèrdua d'expressió en algun dels dos participants. Aquest silenciament de l'expressió podria ser atribuït a mecanismes genètics o epigenètics, com per exemple alteracions al DNA o a l'estructura de la cromatina, mutacions somàtiques o metilació a les regions promotores, etc. D'altra banda, les interaccions perdudes que pertanyen a TFs i gens diana que mantenen els nivells d'expressió a les cèl·lules tumorals, poden ser explicades per altres raons. Per exemple, mitjançant terceres molècules que podrien estar actuant com a moduladors de l'activitat del TF [146] o com a conseqüència de determinades alteracions en regions promotores que podrien crear nous llocs d'unió de TFs [147]. Addicionalment, la xarxa de cèl·lules tumorals mostra un conjunt emergent d'interaccions transcripcionals que probablement tingui un rol important en el

desenvolupament tumoral i en l'adquisició de noves capacitats cel·lulars. Estudis previs han demostrat que l'activació de petits mòduls de regulació són necessaris i suficients per iniciar i mantenir un estat fenotípicament aberrant en tumors cerebrals [121]. Aquesta aproximació computacional mitjançant la reconstrucció de les xarxes de regulació transcripcional, doncs, podria ser especialment útil per identificar nous mòduls que dirigeixin o inicïin la transformació maligna. En els resultats del nostre treball s'ofereix una llista completa dels gens en els que s'ha trobat incrementada la seva connectivitat en els tumors. Entre ells es poden trobar oncògens, gens relacionats amb la proliferació cel·lular, la transició epiteli-mesènquima, l'angiogènesi i gens de la via de Wnt, potencialment implicats en CCR. El gen *GREM1*, que és un dels que exhibeix un considerable increment d'activitat a la xarxa dels tumors, es coneix que juga un rol important en la supervivència de l'estroma que envolta els tumors i en la proliferació d'alguns tipus de càncers [148]. En CCR està relacionat per exemple amb un locus associat amb un augment de risc [149] i també s'havia descrit prèviament un augment de la seva expressió en pòlips colorectals [150] i en la transició adenoma-carcinoma dels tumors de còlon [151].

Un punt fort d'aquest estudi és el fet de que tant les mostres normals con les tumorals pertanyin al mateix conjunt d'individus, és a dir, que siguin mostres aparellades. De fet, segons el nostre coneixement, és el primer estudi d'aquestes característiques en CCR. Addicionalment, s'ha tingut molta cura amb el disseny experimental per evitar biaixos entre tipus de teixits. Així que per tot això, es suggereix fortament que aquestes grans diferències trobades entre les dues xarxes es deuen principalment al fenotip tumoral. Un altre punt fort d'aquest estudi és que l'algorisme ARACNe ha sigut executat seguint exhaustivament totes les recomanacions dels seus autors, per així aconseguir uns resultats el més acurats possible. Abans d'executar l'algorisme es van realitzar simulacions per optimitzar els valors de certs paràmetres necessaris, en comptes de deixar el seu valor per defecte. Per exemple, es va optimitzar l'amplada del nucli per poder disposar d'una millor estimació d'informació mútua per al nostre conjunt dades, i també es va determinar la distribució nul·la de la informació mútua, per identificar millor les correlacions significatives entre TFs i els seus possibles gens diana. Addicionalment, per obtenir més robustesa es van computar per a cadascuna de les xarxes 1000 repeticions amb remostreig, que posteriorment van ser resumides com a xarxes consens. Finalment, es va dur a terme una validació *in silico*, mitjançant dades experimentals provinents d'una base de dades pública, a la qual es va trobar un nivell de concordança molt raonable.

D'altra banda, una limitació potencial de l'estudi pot ser el soroll introduït a les anàlisis a causa de l'heterogeneïtat cel·lular present a l'interior dels tumors. Aquest fet inherent podria ser un dels que contribueix, entre d'altres, a la pèrdua de connectivitat observada a la xarxa dels tumors. Mentre que la mucosa patològicament normal és un teixit relativament homogeni entre individus, els tumors són molt més heterogenis degut a la diversitat de clons cel·lulars que contenen [152]. Això podria resultar en una pèrdua

aparent de correlacions globals, si diverses xarxes transcripcionals estan realment barrejades en els tumors. Un altre punt feble de l'estudi pot ser el nombre de mostres utilitzat per construir cadascuna de les dues xarxes. Els propis autors de l'algorisme ARACNe assenyalen en els seus treballs que la mida mínima per inferir les xarxes de regulació transcripcional amb una precisió adequada és aproximadament un centenar de mostres [123]. Amb la mida de mostra que nosaltres hem utilitzat estem dins de les seves recomanacions, però som conscients que estem situats a la banda inferior de la franja recomanada. És a dir, que si haguéssim pogut disposar d'un conjunt de mostres més ampli, possiblement les xarxes tindrien encara una mica més de precisió. Val a dir que aconseguir un conjunt ampli de mostres aparellades d'una sèrie homogènia d'aquest tipus de pacients, passant tots els criteris de qualitat i recomanacions, no és gens fàcil.

La inferència de xarxes de regulació transcripcional ja havia sigut satisfactòriament aplicada a altres lesions malignes com la leucèmia [121], el càncer de mama [153, 154] o en tumors d'ovari [155]. Gràcies a aquesta metodologia es van poder trobar nous marcadors de pronòstic en metàstasi de càncer de mama i es van prioritzar gens per ser candidats específics de nous tractaments pel càncer d'ovari. Actualment sembla que aquestes metodologies d'anàlisi global han sigut ben acceptades i cada vegada més, el seu ús es va incrementant. Per exemple, actualment l'algorisme ARACNe ja ha estat citat en més de 400 treballs científics segons la base de dades Pubmed. Algunes de les publicacions més actuals treballen temes tan interessants com poden ser la identificació de nous gens relacionats amb la formació de tumors [156]; relacionar les vies gèniques amb els programes transcripcionals involucrats en el càncer de mama [157]; o identificar potencials candidats terapèutics en càncer de pròstata metastàtic [158]. En canvi, en CCR altres autors han explorat la reconstrucció de les xarxes de regulació, però amb enfocaments limitats a un únic TF [159] o només en teixit tumoral [160, 161]. Fins on sabem, aquest és el primer estudi de càncer de còlon en el que s'han inferit simultàniament les dues xarxes del tumor i de les cèl·lules normals adjacents, obtingudes a partir del mateix conjunt d'individus i amb una metodologia consistent que les fa totalment comparables. Creiem que gràcies a metodologies computacionals d'anàlisi global com aquesta, es poden trobar nous i interessants mecanismes de la biologia molecular del CCR que encara no han estat descrits. Per exemple, la gran majoria de les arestes perdudes a la xarxa dels tumors no haguessin pogut ser identificades explorant només els valors d'expressió dels gens, ja que no s'observen grans canvis en l'expressió mitjana dels gens directament implicats. També recentment s'ha publicat que el TF *GREM1* és capaç d'iniciar el desenvolupament tumoral a dins de les criptes [162]. Per tant, els nostres resultats trobats mitjançant una metodologia computacional, que ja indicaven que l'efecte del TF *GREM1*, entre d'altres, estava intervenint en el procés tumoral mitjançant l'activació d'un gran programa transcripcional, sembla que estan sent corroborats per altres autors.

Gràcies a tot això pensem que les nostres troballes aporten una millor comprensió dels programes de regulació transcripcional alterats en càncer de còlon i que podrien ser una nova metodologia molt valuosa per identificar mecanismes amb un paper rellevant en el camp del diagnòstic, el pronòstic i el tractament del càncer de còlon. Amb la filosofia de compartir al màxim aquests resultats amb la resta de la comunitat científica, l'article publicat és d'accés obert i en el material suplementari d'aquest es poden trobar les dues xarxes de regulació transcripcional resultants en diversos formats, per facilitar-ne el seu ús. També en el web del projecte es pot trobar una eina que permet a l'usuari veure les interaccions concretes d'un gen d'interès consultant de forma interactiva les dues xarxes a la vegada (www.colonomics.org/regulatory-networks).

4. Plans de futur

Es pot considerar que la temàtica general sobre la qual tracta aquesta tesi, és a dir el CCR, és un tema de gran interès per la recerca. Això és així principalment perquè es tracta d'una malaltia que representa un greu problema de salut a nivell mundial per la seva elevada mortalitat i al constant increment en la incidència. Dit això, potser seria bo poder donar continuïtat a totes les línies de recerca possibles, per així aprofundir al màxim en un ampli ventall d'enfocaments diferents. Malgrat tot, sempre hi ha treballs amb una major continuïtat que d'altres. Tot seguit s'intenta detallar els plans de futur dels treballs que han format part d'aquesta tesi.

Respecte al primer article, en el qual es van comparar els perfils d'expressió gènica entre els tumors de còlon i els tumors de recte per identificar el grau de similitud que existeix entre ells, a partir d'ara treballarem per intentar caracteritzar els tumors de les diferents localitzacions del còlon, però amb dades a altre nivell, com poden ser dades de proteòmica o dades epigenètiques (metilació del DNA o perfils d'expressió de miRNAs). Fins ara, en el nostre article només hem avaluat les diferències existents als perfils d'expressió gènica dels tumors situats a les diferents localitzacions del còlon. Tot i així, cada vegada existeix més evidència, com hem explicat prèviament a la discussió actualitzada d'aquest article, que poden existir diferències moleculars a altres nivells, com per exemple a nivell epigenètic o proteòmic. Posteriorment, tot aquest nou coneixement adquirit a nivell molecular del CCR es podrà utilitzar per classificar i tractar els tumors de manera més específica. De fet, aquesta pràctica clínica ja és habitual en el tractament d'altres tipus de neoplàsies, com per exemple el càncer de mama [107, 108].

Respecte el segon article, en el qual es volia identificar nous biomarcadors sèrics i demostrar la seva potencial utilitat en el diagnòstic precoç del càncer de còlon, finalment el millor marcador trobat no ha demostrat un poder de classificació superior al de les altres alternatives actualment ja existents. Tot i així, a causa del marge de millora encara existent

en els actuals TSOF utilitzats en la majoria dels programes de cribratge, vam decidir continuar treballant en aquesta àrea. Concretament, actualment tenim dues línies de recerca en marxa relacionades amb la identificació de nous biomarcadors pel diagnòstic precoç del CCR.

La primera d'aquestes pretén donar continuïtat de forma directa a l'article presentat en aquesta tesi. Consisteix, en primer lloc i de forma més senzilla, en continuar provant a la sèrie de sèrums ampliada de la que disposem actualment, candidats addicionals que inicialment no vam poder provar a causa de limitacions relacionades amb els costos econòmics. Una altra opció que hem estat avaluant és la possibilitat de realitzar un cribratge mitjançant *microarrays* d'anticossos. Actualment aquestes tecnologies encara tenen unes fortes limitacions en relació al nombre de proteïnes disponibles per avaluar i que es poden testar de forma simultània en un mateix experiment. Malgrat això, cada vegada es poden trobar més anticossos disponibles per aquestes plataformes i els seus costos progressivament es van reduint. Actualment ja s'han publicat alguns treballs amb resultats satisfactoris en altres tipus de tumors en els quals s'han utilitzat *microarrays* d'anticossos [163, 164].

La segona línia de treball que estem duent a terme al camp dels biomarcadors sèrics en CCR és una mica diferent en quant al tipus de molècula utilitzada però té la mateixa finalitat, que consisteix en identificar biomarcadors pel diagnòstic de la malaltia. Concretament, en aquest treball estem avaluant el possible ús dels miRNAs com a potencials candidats per a la detecció precoç del CCR. Actualment es coneix que molts miRNAs estan presents fora de les cèl·lules, de forma estable circulant a la sang perifèrica, i que potencialment podrien servir com a biomarcadors no invasius per a la detecció del càncer [165]. A més, en altres tipus de tumors com els de mama [166, 167], pròstata [168], pulmó [169] i càncer de pàncrees [170, 171], determinats miRNAs han estat ja proposats com a candidats per a biomarcadors basats en sang. En el cas del CCR, diversos estudis previs han treballat en la detecció dins d'un context no invasiu [63, 77, 78]. Aquests estudis reporten una sensibilitat compresa entre el 62% i el 89% i una especificitat entre el 70% i el 89% [79]. En base a això, els objectius principals d'aquest treball són posar a punt una metodologia seqüencial per identificar un conjunt prometedor de miRNAs circulants en sèrum i avaluar la seva utilitat en el diagnòstic precoç del CCR. Actualment, el treball no està finalitzat, però fins al moment no hem pogut trobar cap miRNA que proporcioni una bona capacitat de classificació entre els sèrums dels controls sans i els sèrums dels casos de CCR o dels adenomes. Com es pot veure a la Figura 16, que conté uns exemples representatius del miRNAs avaluats, existeixen grans diferències en els seus valors d'expressió en teixit, entre els casos i els controls, tant a la sèrie inicial com a la sèrie independent de validació en teixit. Tot i així, la concentració trobada als sèrums no té la capacitat per discriminar entre els diferents grups d'interès, en alguns casos perquè sembla que el miRNA no es troba

present (Figura 16A) i en altres casos perquè no segueix el mateix patró que presentava prèviament al teixit (Figura 16B).

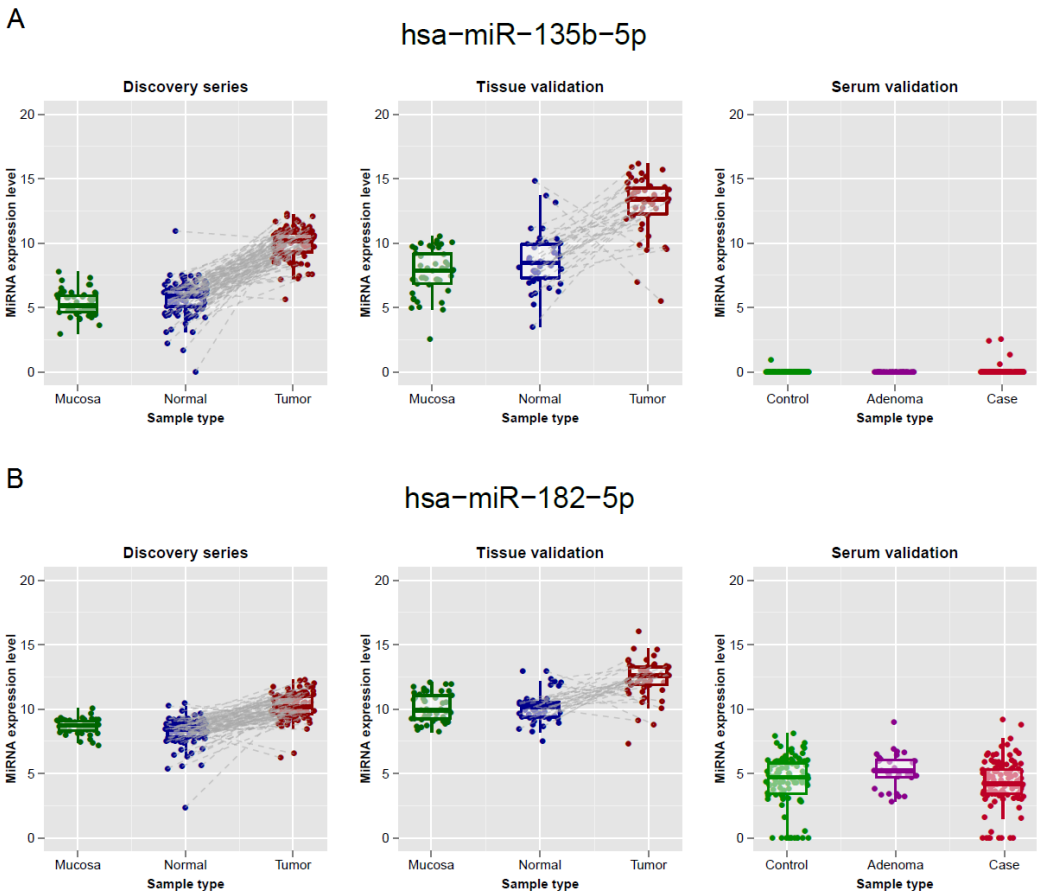

Figura 16. Valors d'expressió al teixit en la sèrie inicial de cerca, en la sèrie de validació i valors de la concentració en sèrum, pel miRNA hsa-miR-135b-5p (A) i el miRNA hsa-miR-182-5p (B).

Inicialment, gràcies als resultats satisfactoris que vam obtenir en el treball de biomarcadors basats en la detecció de proteïnes en sèrum, aquí vam procedir d'una forma molt similar, seguint un procés seqüencial d'identificació i validació de biomarcadors. En aquest treball de miRNAs s'han utilitzat més de 650 mostres, si es tenen en compte les emprades a les tres sèries. Aquest nombre de mostres ens ha de proporcionar un poder estadístic suficient com per poder identificar les possibles diferències existents entre els valors dels sèrums als casos i als controls. També inicialment s'han utilitzat mostres de mucosa sana d'individus sense càncer, que ens han permès novament reduir els possibles falsos positius. Hem estat molt curosos amb tots els aspectes referents al control de qualitat i manipulació de les mostres. Per exemple, totes aquelles mostres que presentaven una mínima presencia d'hemòlisi, no han estat incloses en les anàlisis. També han estat fortament treballats tots els aspectes que podrien provocar una potencial confusió o biaix, com les covariables conegudes, la informació clínica del individus, les característiques dels tumors i més fonts

de possibles efectes tècnics. Un factor que podria haver intervingut en l'obtenció dels resultats negatius observats fins el moment podria estar relacionat amb l'elecció del material biològic utilitzat als nostres experiments. Es va optar per determinar les concentracions de miRNAs en sèrum ja que algun estudi previ havia reportat la presència de majors concentracions de miRNAs en sèrum respecte a mostres de plasma [172]. Un altre factor a tenir en compte en aquest estudi està relacionat amb el mecanisme biològic de secreció dels miRNAs. En el nostre treball nosaltres vam aïllar la fracció de miRNAs que circulaven lliurement en sang. Tot i així, estudis previs han demostrat que és possible detectar els miRNAs de forma estable al sèrum ja que aquests són empaquetats en exosomes, i així es protegeixen de la degradació. Per exemple, en una línia cel·lular de càncer gàstric metastàtic, s'ha demostrat que la família de miRNAs let-7 és secretada de forma selectiva a l'entorn extracel·lular mitjançant exosomes [173]. Com a conseqüència, seria possible re-evaluar els nostres candidats identificats en teixit de còlon mitjançant la quantificació de l'expressió dels miRNAs circulants continguts a la fracció corresponent als exosomes, com s'ha descrit recentment en algun estudi [174]. Malgrat el potencial impacte que pot tenir aquesta línia de recerca en un futur proper, hem de remarcar que treballs previs d'altres autors basats en miRNAs han reportat resultats discordants en biomarcadors sèrics pel càncer de còlon [78]. Això obligarà a dur a terme validacions exhaustives en sèries independents i de mida mostral suficient per tal d'assegurar la fiabilitat dels resultats obtinguts.

En el tercer article, es va dur a terme una anàlisi per identificar els canvis globals entre els programes de regulació transcripcional de les cèl·lules normals i tumorals del còlon mitjançant la reconstrucció de les dues xarxes de regulació transcripcional. La inferència d'aquestes xarxes és un procés computacionalment molt intensiu, però que alhora serveix de punt de partida per aplicar tota una sèrie de metodologies d'anàlisi computacional sobre les mateixes. Així, els mateixos creadors de l'algorisme ARACNe, que s'ha utilitzat per reconstruir les xarxes de regulació transcripcional del nostre treball, van proposar també un altre algorisme per inferir els reguladors claus (MR, de l'anglès *Master Regulator*) d'un fenotip d'interès [175], anomenat MARINa. Donats dos fenotips, A i B, i una xarxa de regulació transcripcional, aquest mètode intenta identificar els TFs que podrien estar induint la transició d'A cap a B. Aquest conjunt de TFs serien els MRs de la transició entre els dos fenotips. Aquesta metodologia actualment ja ha sigut aplicada amb èxit en altres tipus de tumors. Per exemple, en glioblastomes s'ha utilitzat per identificar els mòduls transcripcionals que activen l'expressió de gens relacionats amb la transformació epiteli-mesènquima [176]. En càncer de mama, per construir una classificació basada en mòduls que sigui més robusta i reproduïble que els propis marcadors de forma individual [177]. Finalment, també s'ha aplicat en limfòcits B per identificar els MRs relacionats amb la seva proliferació [178].

Una de les gran fortaleses que tenen les anàlisis de MRs és que permeten identificar marcadors robustos per a un determinat fenotip d'interès. Per exemple, en molts casos és difícil trobar marcadors que prediguin amb precisió el pronòstic d'una malaltia i que sovint siguin reproduïbles quan s'avaluen en altres conjunts de dades independents. Si pensem en el pronòstic del càncer com el nostre fenotip d'interès i volem construir una signatura que sigui capaç de predir-ho de forma robusta i reproduïble, l'enfocament mitjançant l'anàlisi dels MRs ens pot ajudar a inferir marcadors robustos, com es pot veure a la Figura 17.

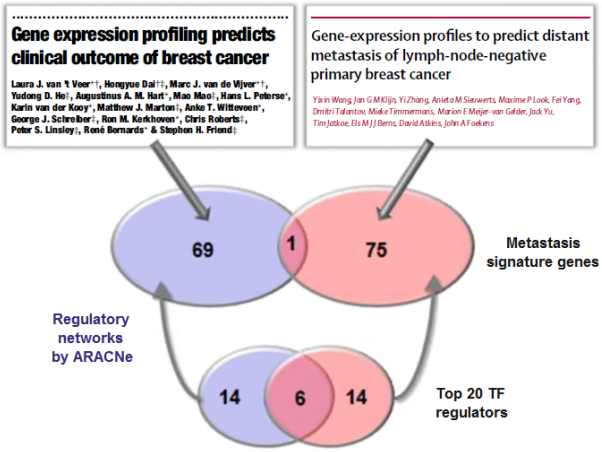

Figura 17. Representació gràfica del nombre de gens que coincideixen entre dues signatures per predir metàstasi en càncer de mama i dels gens que coincideixen entre els 20 primers TFs que regulen el fenotip metastàtic al dos conjunts de dades d'expressió [175].

En aquest punt, nosaltres ja disposem de la xarxa de regulació transcripcional reconstruïda, gràcies a tot el treball realitzat al tercer article presentat com a resultat d'aquesta tesi. A més, dins del marc del projecte COLONOMICS, actualment ja consolidat al nostre grup, es disposa d'informació addicional sobre les sèries de mostres utilitzades per reconstruir les xarxes de regulació. És a dir, es podria disposar i utilitzar informació clínica, epidemiològica, mutacional, etc., amb un gran interès per la recerca en CCR. Per exemple, considerant com a fenotip d'interès el pronòstic dels pacients dels quals prové la mostra original, disponible gràcies a la informació recollida de manera prospectiva al projecte, podem aplicar l'anàlisi dels MRs per intentar inferir marcadors de pronòstic robustos. Al nostre grup prèviament hem descrit les enormes dificultats que existeixen per trobar un conjunt de marcadors robustos i reproduïbles per predir el pronòstic del CCR mitjançant perfils d'expressió gènica [179]. És per això que pensem que aquests nous enfocaments computacionals i integratius ens poden ser de gran utilitat per avançar en aquest camp.

Finalment, si així es desitgés, es podria donar encara més continuïtat a aquesta línia de recerca mitjançant l'aplicació d'enfocaments analítics addicionals. En els darrers anys s'han anat desenvolupant noves metodologies computacionals que permeten estudiar

mecanismes regulatoris cel·lulars més complexos. D'aquesta manera, existeixen alternatives per identificar potencials moduladors de la regulació transcripcional els TFs i els seus gens diana [146]. L'algorisme anomenat MINDy, té com a objectiu trobar de forma computacional cofactors que actuïn com a moduladors, fent inferència sobre les xarxes de regulació prèviament reconstruïdes. Una possible extensió del nostre treball, doncs, seria intentar aplicar aquesta metodologia a les nostres xarxes de regulació transcripcional prèviament reconstruïdes. Addicionalment, es podrien complementar aquestes dades d'expressió amb altres dades disponibles, com els perfils d'expressió de miRNAs provinents també del projecte COLONOMICS.

5. Impacte al càncer colorectal

Fins el moment, la prevenció primària en CCR s'ha mostrat majoritàriament poc efectiva. Tot i que és primordial seguir fent recerca en aquest sentit, les mesures preventives possiblement més efectives a curt termini passen per potenciar la prevenció secundària i terciària. Aquest és l'objectiu d'un dels treballs que formen aquesta tesi, on s'intenta identificar nous biomarcadors sèrics que millorin el diagnòstic precoç de la malaltia. Els altres dos articles se centren en aprofundir en les bases moleculars del CCR per tal de poder classificar i tractar els tumors d'una manera més acurada. A continuació es desenvolupen els aspectes concrets que poden tenir un impacte clar en el CCR i que fan referència al treball desenvolupat en aquesta tesi.

El primer article, que compara els perfils d'expressió gènica entre els tumors de còlon i els tumors de recte per identificar el grau de similitud que existeix entre ells, té un impacte directe a nivell del coneixement molecular de la malaltia. Actualment, aquest treball ja ha acumulat 12 cites d'altres articles indexats al Pubmed. Tanmateix, el projecte TCGA va corroborar recentment els resultats trobats prèviament per nosaltres. Aquest impacte es reflectirà clarament als futurs treballs que incloguin perfils d'expressió gènica de tumors de còlon i de recte. De fet, una de les possibles vies actuals per millorar el pronòstic de la malaltia és caracteritzar amb precisió els tumors a nivell molecular per així poder prescriure el tractament més adequat per a cada subtipus. Aquesta pràctica clínica actualment ja s'està duent a terme en altres tumors, com pot ser per exemple el càncer de mama. De la mateixa manera, doncs, els resultats del nostre treball han contribuït a aprofundir en el coneixement molecular del CCR.

Al segon article, en el qual es volia identificar nous biomarcadors sèrics i demostrar la seva potencial utilitat per al diagnòstic precoç del càncer de còlon. El millor marcador identificat en el treball no ha demostrat un poder de classificació superior al de les alternatives actualment existents, pel que l'impacte real a curt termini és limitat. No obstant això, es coneix que encara hi ha un marge de millora per als actuals TSOF emprats a la majoria dels

programes de cribratge. En el nostre grup continuarem treballant com s'ha explicat a l'apartat anterior sobre els plans de futur, per intentar trobar un biomarcador útil per al diagnòstic precoç del CCR.

Al tercer article, es va dur a terme una anàlisi per trobar els canvis globals entre els programes de regulació transcripcional de les cèl·lules normals i tumorals del còlon, mitjançant la reconstrucció de les dues xarxes de regulació transcripcional. La desregulació trobada als tumors de còlon a nivell global és un fenomen prèviament no descrit que pot tenir un impacte directe al coneixement molecular del CCR. A més, el treball també recull una metodologia per identificar nous gens amb un paper rellevant en el desenvolupament del CCR, fet que també pot tenir un impacte als paradigmes clàssics d'anàlisi computacional. Finalment, l'aplicació de nous enfocaments d'anàlisi computacional pot aportar valuosa informació amb un potencial impacte en la identificació de nous mecanismes moleculars associats amb la malaltia, que alhora contribueixin a millorar el seu diagnòstic, pronòstic i tractament.

Conclusions

En aquesta tesi s'han aplicat metodologies d'anàlisi computacional per identificar nous mecanismes moleculars subjacents a la patologia del CCR que proporcionin coneixements sobre la malaltia, així com per intentar trobar biomarcadors per al seu diagnòstic precoç. En el primer treball s'han comparat els perfils d'expressió gènica de tumors de còlon i tumors de recte, per determinar els seus patrons comuns i diferencials a nivell transcripcional. En el segon treball s'han identificat possibles nous biomarcadors sèrics de CCR, i s'ha testat la seva potencial utilitat per al diagnòstic precoç del càncer de còlon. En el tercer treball s'ha dut a terme una anàlisi de les alteracions globals en els programes de regulació transcripcional de les cèl·lules tumorals del còlon, mitjançant tècniques computacionals de reconstrucció de xarxes de regulació transcripcional a partir de dades d'expressió gènica.

A continuació es troben enumerades les conclusions específiques per cadascun dels treballs que formen part d'aquesta tesi.

- *Gene expression differences between colon and rectum tumors:*

    - Els tumors colorectals estables en microsatèl·lits exhibeixen perfils d'expressió gènica molt similars, independentment de la seva localització.

    - Les petites diferències observades entre els tumors situats al còlon dret, al còlon esquerre i al recte, són impulsades en gran mesura pels gens *HOX*.

- *Discovery and validation of new potential biomarkers for early detection of colon càncer:*

    - L'anàlisi computacional de perfils transcripcionals a gran escala ha permès identificar un conjunt de possibles candidats a biomarcadors per la detecció precoç del càncer de còlon.

    - Un posterior procés de validació seqüencial ha assenyalat la proteïna del gen *COL10A1* com un candidat prometedor pel diagnòstic de càncer de còlon. La determinació de la concentració de la proteïna COL10A1 en sèrum permet identificar els adenomes i els càncers amb una bona sensibilitat i especificitat.

- *Large differences in global transcriptional regulatory programs of normal and tumor colon cells:*

    - La inferència de les xarxes de regulació transcripcional a nivell del genoma complet, ens ha permès detectar una pèrdua massiva de regulació transcripcional a les cèl·lules tumorals del còlon, no descrita anteriorment.

- Aquesta pèrdua d'activitat reguladora trobada als tumors de còlon no es deu en general al silenciament dels gens directament implicats, sinó a la supressió de la pròpia regulació transcripcional, probablement causada per alteracions genètiques o epigenètiques que impedeixen una correcta unió entre el seu factor de transcripció i el seu gen diana.

- En aquest context de pèrdua global de regulació transcripcional, els tumors de còlon augmenten l'activitat reguladora de certs factors de transcripció concrets, amb un paper central i molt rellevant a la xarxa com *SNAI2*, *TGFB1* i *GREM1*, entre d'altres.

Bibliografia

1. Ferlay J, Soerjomataram I, Dikshit R, Eser S, Mathers C, Rebelo M, Parkin DM, Forman D, Bray F: Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012. *International journal of cancer Journal international du cancer* 2015, 136(5):E359-386.

2. Ribes J, Esteban L, Cleries R, Galceran J, Marcos-Gragera R, Gispert R, Ameijide A, Vilardell ML, Borras J, Puigdefabregas A, Buxo M, Freitas A, Izquierdo A, Borras JM: Cancer incidence and mortality projections up to 2020 in Catalonia by means of Bayesian models. *Clin Transl Oncol* 2014, 16(8):714-724.

3. Lichtenstein P, Holm NV, Verkasalo PK, Iliadou A, Kaprio J, Koskenvuo M, Pukkala E, Skytthe A, Hemminki K: Environmental and heritable factors in the causation of cancer--analyses of cohorts of twins from Sweden, Denmark, and Finland. *The New England journal of medicine* 2000, 343(2):78-85.

4. de la Chapelle A: Genetic predisposition to colorectal cancer. *Nature reviews Cancer* 2004, 4(10):769-780.

5. Huxley RR, Ansary-Moghaddam A, Clifton P, Czernichow S, Parr CL, Woodward M: The impact of dietary and lifestyle risk factors on risk of colorectal cancer: a quantitative overview of the epidemiological evidence. *International journal of cancer Journal international du cancer* 2009, 125(1):171-180.

6. Gonzalez CA: Nutrition and cancer: the current epidemiological evidence. *The British journal of nutrition* 2006, 96 Suppl 1:S42-45.

7. Potter JD: Colorectal cancer: molecules and populations. *Journal of the National Cancer Institute* 1999, 91(11):916-932.

8. Cao Y, Keum NN, Chan AT, Fuchs CS, Wu K, Giovannucci EL: Television watching and risk of colorectal adenoma. *British journal of cancer* 2015, 112(5):934-942.

9. Kruk J, Czerniak U: Physical activity and its relation to cancer risk: updating the evidence. *Asian Pacific journal of cancer prevention : APJCP* 2013, 14(7):3993-4003.

10. Garai J, Uddo RB, Mohler MC, Pelligrino N, Scribner R, Sothern MS, Zabaleta J: At the crossroad between obesity and gastric cancer. *Methods in molecular biology* 2015, 1238:689-707.

11. Giovannucci E: An updated review of the epidemiological evidence that cigarette smoking increases risk of colorectal cancer. *Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology* 2001, 10(7):725-731.

12. Liang PS, Chen TY, Giovannucci E: Cigarette smoking and colorectal cancer incidence and mortality: systematic review and meta-analysis. *International journal of cancer Journal international du cancer* 2009, 124(10):2406-2415.

13. Raimondi S, Botteri E, Iodice S, Lowenfels AB, Maisonneuve P: Gene-smoking interaction on colorectal adenoma and cancer risk: review and meta-analysis. *Mutation research* 2009, 670(1-2):6-14.

14. Park JY, Mitrou PN, Dahm CC, Luben RN, Wareham NJ, Khaw KT, Rodwell SA: Baseline alcohol consumption, type of alcoholic beverage and risk of colorectal cancer in the European Prospective Investigation into Cancer and Nutrition-Norfolk study. *Cancer epidemiology* 2009, 33(5):347-354.

15. Hermann S, Rohrmann S, Linseisen J: Lifestyle factors, obesity and the risk of colorectal adenomas in EPIC-Heidelberg. *Cancer causes & control : CCC* 2009, 20(8):1397-1408.

16. Poschl G, Seitz HK: Alcohol and cancer. *Alcohol and alcoholism* 2004, 39(3):155-165.

17. Janne PA, Mayer RJ: Chemoprevention of colorectal cancer. *The New England journal of medicine* 2000, 342(26):1960-1968.

18. Poynter JN, Gruber SB, Higgins PD, Almog R, Bonner JD, Rennert HS, Low M, Greenson JK, Rennert G: Statins and the risk of colorectal cancer. *The New England journal of medicine* 2005, 352(21):2184-2192.

19. Valle L: Genetic predisposition to colorectal cancer: where we stand and future perspectives. *World journal of gastroenterology* 2014, 20(29):9828-9849.

20. Lynch HT, Shaw TG: Practical genetics of colorectal cancer. *Chin Clin Oncol* 2013, 2(2):12.

21. Fearon ER, Vogelstein B: A genetic model for colorectal tumorigenesis. *Cell* 1990, 61(5):759-767.

22. Peters U, Bien S, Zubair N: Genetic architecture of colorectal cancer. *Gut* 2015, 64(10):1623-1636.

23. Puppa G, Maisonneuve P, Sonzogni A, Masullo M, Capelli P, Chilosi M, Menestrina F, Viale G, Pelosi G: Pathological assessment of pericolonic tumor deposits in advanced colonic carcinoma: relevance to prognosis and tumor staging. *Modern pathology : an official journal of the United States and Canadian Academy of Pathology, Inc* 2007, 20(8):843-855.

24. Iacopetta B: Are there two sides to colorectal cancer? *International journal of cancer Journal international du cancer* 2002, 101(5):403-408.

25. Reya T, Clevers H: Wnt signalling in stem cells and cancer. *Nature* 2005, 434(7035):843-850.

26. Fearon ER, Jones PA: Progressing toward a molecular description of colorectal cancer development. *FASEB J* 1992, 6(10):2783-2790.

27. Brucher BL, Jamall IS: Epistemology of the origin of cancer: a new paradigm. *BMC cancer* 2014, 14:331.

28. Vogelstein B, Papadopoulos N, Velculescu VE, Zhou S, Diaz LA, Jr., Kinzler KW: Cancer genome landscapes. *Science* 2013, 339(6127):1546-1558.

29. Beahrs OH: Staging of cancer of the colon and rectum. *Cancer* 1992, 70(5 Suppl):1393-1396.

30. Webber C, Gospodarowicz M, Sobin LH, Wittekind C, Greene FL, Mason MD, Compton C, Brierley J, Groome PA: Improving the TNM classification: findings from a 10-year continuous literature review. *International journal of cancer Journal international du cancer* 2014, 135(2):371-378.

31. Gunderson LL, Jessup JM, Sargent DJ, Greene FL, Stewart AK: Revised TN categorization for colon cancer based on national survival outcomes data. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology* 2010, 28(2):264-271.

32. Kim HJ, Yu MH, Kim H, Byun J, Lee C: Noninvasive molecular biomarkers for the detection of colorectal cancer. *BMB reports* 2008, 41(10):685-692.

33. Casillas S, Pelley RJ, Milsom JW: Adjuvant therapy for colorectal cancer: present and future perspectives. *Diseases of the colon and rectum* 1997, 40(8):977-992.

34. Binefa G, Rodriguez-Moranta F, Teule A, Medina-Hayas M: Colorectal cancer: from prevention to personalized medicine. *World journal of gastroenterology* 2014, 20(22):6786-6808.

35. Das D, Arber N, Jankowski JA: Chemoprevention of colorectal cancer. *Digestion* 2007, 76(1):51-67.

36. Crosara Teixeira M, Braghiroli MI, Sabbaga J, Hoff PM: Primary prevention of colorectal cancer: myth or reality? *World journal of gastroenterology* 2014, 20(41):15060-15069.

37. Walsh JM, Terdiman JP: Colorectal cancer screening: scientific review. *Jama* 2003, 289(10):1288-1296.

38. Shaukat A, Mongin SJ, Geisser MS, Lederle FA, Bond JH, Mandel JS, Church TR: Long-term mortality after screening for colorectal cancer. *The New England journal of medicine* 2013, 369(12):1106-1114.

39. Farraye FA, Wallace M: Clinical significance of small polyps found during screening with flexible sigmoidoscopy. *Gastrointestinal endoscopy clinics of North America* 2002, 12(1):41-51.

40. Imperiale TF, Wagner DR, Lin CY, Larkin GN, Rogge JD, Ransohoff DF: Risk of advanced proximal neoplasms in asymptomatic adults according to the distal colorectal findings. *The New England journal of medicine* 2000, 343(3):169-174.

41. Winawer SJ, Stewart ET, Zauber AG, Bond JH, Ansel H, Waye JD, Hall D, Hamlin JA, Schapiro M, O'Brien MJ, Sternberg SS, Gottlieb LS: A comparison of colonoscopy and double-contrast barium enema for surveillance after polypectomy. National Polyp Study Work Group. *The New England journal of medicine* 2000, 342(24):1766-1772.

42. Rosman AS, Korsten MA: Meta-analysis comparing CT colonography, air contrast barium enema, and colonoscopy. *The American journal of medicine* 2007, 120(3):203-210 e204.

43. Zavoral M, Suchanek S, Zavada F, Dusek L, Muzik J, Seifert B, Fric P: Colorectal cancer screening in Europe. *World journal of gastroenterology* 2009, 15(47):5907-5915.

44. Heresbach D, Manfredi S, D'Halluin P N, Bretagne JF, Branger B: Review in depth and meta-analysis of controlled trials on colorectal cancer screening by faecal occult blood test. *European journal of gastroenterology & hepatology* 2006, 18(4):427-433.

45. Faivre J, Dancourt V, Lejeune C, Tazi MA, Lamour J, Gerard D, Dassonville F, Bonithon-Kopp C: Reduction in colorectal cancer mortality by fecal occult blood screening in a French controlled study. *Gastroenterology* 2004, 126(7):1674-1680.

46. Scholefield JH, Moss S, Sufi F, Mangham CM, Hardcastle JD: Effect of faecal occult blood screening on mortality from colorectal cancer: results from a randomised controlled trial. *Gut* 2002, 50(6):840-844.

47. Hewitson P, Glasziou P, Watson E, Towler B, Irwig L: Cochrane systematic review of colorectal cancer screening using the fecal occult blood test (hemoccult): an update. *The American journal of gastroenterology* 2008, 103(6):1541-1549.

48. Burch JA, Soares-Weiser K, St John DJ, Duffy S, Smith S, Kleijnen J, Westwood M: Diagnostic accuracy of faecal occult blood tests used in screening for colorectal cancer: a systematic review. *Journal of medical screening* 2007, 14(3):132-137.

49. Pignone M, Campbell MK, Carr C, Phillips C: Meta-analysis of dietary restriction during fecal occult blood testing. *Effective clinical practice : ECP* 2001, 4(4):150-156.

50. van Dam L, Kuipers EJ, van Leerdam ME: Performance improvements of stool-based screening tests. *Best practice & research Clinical gastroenterology* 2010, 24(4):479-492.

51. Vilkin A, Rozen P, Levi Z, Waked A, Maoz E, Birkenfeld S, Niv Y: Performance characteristics and evaluation of an automated-developed and quantitative, immunochemical, fecal occult blood screening test. *The American journal of gastroenterology* 2005, 100(11):2519-2525.

52. Guittet L, Bouvier V, Mariotte N, Vallee JP, Arsene D, Boutreux S, Tichet J, Launoy G: Comparison of a guaiac based and an immunochemical faecal occult blood test in screening for colorectal cancer in a general average risk population. *Gut* 2007, 56(2):210-214.

53. Fraser CG, Matthew CM, Mowat NA, Wilson JA, Carey FA, Steele RJ: Immunochemical testing of individuals positive for guaiac faecal occult blood test in a screening programme for colorectal cancer: an observational study. *The Lancet Oncology* 2006, 7(2):127-131.

54. Quintero E, Castells A, Bujanda L, Cubiella J, Salas D, Lanas A, Andreu M, Carballo F, Morillas JD, Hernandez C, Jover R, Montalvo I, Arenas J, Laredo E, Hernandez V, Iglesias F, Cid E, Zubizarreta R, Sala T, Ponce M, Andres M, Teruel G, Peris A, Roncales MP, Polo-Tomas M, Bessa X, Ferrer-Armengou O, Grau J, Serradesanferm A, Ono A *et al*: Colonoscopy versus fecal immunochemical testing in colorectal-cancer screening. *The New England journal of medicine* 2012, 366(8):697-706.

55. Kaminski MF, Bretthauer M, Zauber AG, Kuipers EJ, Adami HO, van Ballegooijen M, Regula J, van Leerdam M, Stefansson T, Pahlman L, Dekker E, Hernan MA, Garborg K, Hoff G: The NordICC

Study: rationale and design of a randomized trial on colonoscopy screening for colorectal cancer. *Endoscopy* 2012, 44(7):695-702.

56. Schreuders EH, Ruco A, Rabeneck L, Schoen RE, Sung JJ, Young GP, Kuipers EJ: Colorectal cancer screening: a global overview of existing programmes. *Gut* 2015, 64(10):1637-1649.

57. Graser A, Stieber P, Nagel D, Schafer C, Horst D, Becker CR, Nikolaou K, Lottes A, Geisbusch S, Kramer H, Wagner AC, Diepolder H, Schirra J, Roth HJ, Seidel D, Goke B, Reiser MF, Kolligs FT: Comparison of CT colonography, colonoscopy, sigmoidoscopy and faecal occult blood tests for the detection of advanced adenoma in an average risk population. *Gut* 2009, 58(2):241-248.

58. Kahi CJ, Imperiale TF, Juliar BE, Rex DK: Effect of screening colonoscopy on colorectal cancer incidence and mortality. *Clinical gastroenterology and hepatology : the official clinical practice journal of the American Gastroenterological Association* 2009, 7(7):770-775; quiz 711.

59. Segnan N, Senore C, Andreoni B, Azzoni A, Bisanti L, Cardelli A, Castiglione G, Crosta C, Ederle A, Fantin A, Ferrari A, Fracchia M, Ferrero F, Gasperoni S, Recchia S, Risio M, Rubeca T, Saracco G, Zappa M, Group-Italy SW: Comparing attendance and detection rate of colonoscopy with sigmoidoscopy and FIT for colorectal cancer screening. *Gastroenterology* 2007, 132(7):2304-2312.

60. Klabunde C, Blom J, Bulliard JL, Garcia M, Hagoel L, Mai V, Patnick J, Rozjabek H, Senore C, Tornberg S: Participation rates for organized colorectal cancer screening programmes: an international comparison. *Journal of medical screening* 2015, 22(3):119-126.

61. Schumacher FR, Schmit SL, Jiao S, Edlund CK, Wang H, Zhang B, Hsu L, Huang SC, Fischer CP, Harju JF, Idos GE, Lejbkowicz F, Manion FJ, McDonnell K, McNeil CE, Melas M, Rennert HS, Shi W, Thomas DC, Van Den Berg DJ, Hutter CM, Aragaki AK, Butterbach K, Caan BJ, Carlson CS, Chanock SJ, Curtis KR, Fuchs CS, Gala M, Giocannucci EL *et al*: Genome-wide association study of colorectal cancer identifies six new susceptibility loci. *Nature communications* 2015, 6:7138.

62. Davies RJ, Miller R, Coleman N: Colorectal cancer screening: prospects for molecular stool analysis. *Nature reviews Cancer* 2005, 5(3):199-209.

63. Huang Z, Huang D, Ni S, Peng Z, Sheng W, Du X: Plasma microRNAs are promising novel biomarkers for early detection of colorectal cancer. *International journal of cancer Journal international du cancer* 2010, 127(1):118-126.

64. Tjalsma H: Identification of biomarkers for colorectal cancer through proteomics-based approaches. *Expert review of proteomics* 2010, 7(6):879-895.

65. Korner H, Soreide K, Stokkeland PJ, Soreide JA: Diagnostic accuracy of serum-carcinoembryonic antigen in recurrent colorectal cancer: a receiver operating characteristic curve analysis. *Annals of surgical oncology* 2007, 14(2):417-423.

66. Hundt S, Haug U, Brenner H: Blood markers for early detection of colorectal cancer: a systematic review. *Cancer epidemiology, biomarkers & prevention : a publication of the American*

Association for Cancer Research, cosponsored by the American Society of Preventive Oncology 2007, 16(10):1935-1953.

67. Grutzmann R, Molnar B, Pilarsky C, Habermann JK, Schlag PM, Saeger HD, Miehlke S, Stolz T, Model F, Roblick UJ, Bruch HP, Koch R, Liebenberg V, Devos T, Song X, Day RH, Sledziewski AZ, Lofton-Day C: Sensitive detection of colorectal cancer in peripheral blood by septin 9 DNA methylation assay. *PloS one* 2008, 3(11):e3759.

68. Lofton-Day C, Model F, Devos T, Tetzner R, Distler J, Schuster M, Song X, Lesche R, Liebenberg V, Ebert M, Molnar B, Grutzmann R, Pilarsky C, Sledziewski A: DNA methylation biomarkers for blood-based colorectal cancer screening. *Clinical chemistry* 2008, 54(2):414-423.

69. deVos T, Tetzner R, Model F, Weiss G, Schuster M, Distler J, Steiger KV, Grutzmann R, Pilarsky C, Habermann JK, Fleshner PR, Oubre BM, Day R, Sledziewski AZ, Lofton-Day C: Circulating methylated SEPT9 DNA in plasma is a biomarker for colorectal cancer. *Clinical chemistry* 2009, 55(7):1337-1346.

70. Chen WD, Han ZJ, Skoletsky J, Olson J, Sah J, Myeroff L, Platzer P, Lu S, Dawson D, Willis J, Pretlow TP, Lutterbaugh J, Kasturi L, Willson JK, Rao JS, Shuber A, Markowitz SD: Detection in fecal DNA of colon cancer-specific methylation of the nonexpressed vimentin gene. *Journal of the National Cancer Institute* 2005, 97(15):1124-1132.

71. Song BP, Jain S, Lin SY, Chen Q, Block TM, Song W, Brenner DE, Su YH: Detection of hypermethylated vimentin in urine of patients with colorectal cancer. *The Journal of molecular diagnostics : JMD* 2012, 14(2):112-119.

72. Klaassen CH, Jeunink MA, Prinsen CF, Ruers TJ, Tan AC, Strobbe LJ, Thunnissen FB: Quantification of human DNA in feces as a diagnostic test for the presence of colorectal cancer. *Clinical chemistry* 2003, 49(7):1185-1187.

73. Imperiale TF, Ransohoff DF, Itzkowitz SH, Turnbull BA, Ross ME, Colorectal Cancer Study G: Fecal DNA versus fecal occult blood for colorectal-cancer screening in an average-risk population. *The New England journal of medicine* 2004, 351(26):2704-2714.

74. Osborn NK, Ahlquist DA: Stool screening for colorectal cancer: molecular approaches. *Gastroenterology* 2005, 128(1):192-206.

75. Azuara D, Rodriguez-Moranta F, de Oca J, Sanjuan X, Guardiola J, Lobaton T, Wang A, Boadas J, Piqueras M, Monfort D, Galter S, Esteller M, Moreno V, Capella G: Novel methylation panel for the early detection of neoplasia in high-risk ulcerative colitis and Crohn's colitis patients. *Inflammatory bowel diseases* 2013, 19(1):165-173.

76. Babel I, Barderas R, Diaz-Uriarte R, Moreno V, Suarez A, Fernandez-Acenero MJ, Salazar R, Capella G, Casal JI: Identification of MST1/STK4 and SULF1 proteins as autoantibody targets for the diagnosis of colorectal cancer by using phage microarrays. *Molecular & cellular proteomics : MCP* 2011, 10(3):M110 001784.

77. Zheng G, Du L, Yang X, Zhang X, Wang L, Yang Y, Li J, Wang C: Serum microRNA panel as biomarkers for early diagnosis of colorectal adenocarcinoma. *British journal of cancer* 2014, 111(10):1985-1992.

78. Hofsli E, Sjursen W, Prestvik WS, Johansen J, Rye M, Trano G, Wasmuth HH, Hatlevoll I, Thommesen L: Identification of serum microRNA profiles in colon cancer. *British journal of cancer* 2013, 108(8):1712-1719.

79. Dong Y, Wu WK, Wu CW, Sung JJ, Yu J, Ng SS: MicroRNA dysregulation in colorectal cancer: a clinical perspective. *British journal of cancer* 2011, 104(6):893-898.

80. Ahlquist DA: Molecular detection of colorectal neoplasia. *Gastroenterology* 2010, 138(6):2127-2139.

81. Diamandis EP: Cancer biomarkers: can we turn recent failures into success? *Journal of the National Cancer Institute* 2010, 102(19):1462-1467.

82. Tan KK, Lopes Gde L, Jr., Sim R: How uncommon are isolated lung metastases in colorectal cancer? A review from database of 754 patients over 4 years. *Journal of gastrointestinal surgery : official journal of the Society for Surgery of the Alimentary Tract* 2009, 13(4):642-648.

83. Hanash SM, Baik CS, Kallioniemi O: Emerging molecular biomarkers--blood-based strategies to detect and monitor cancer. *Nature reviews Clinical oncology* 2011, 8(3):142-150.

84. Pawa N, Arulampalam T, Norton JD: Screening for colorectal cancer: established and emerging modalities. *Nature reviews Gastroenterology & hepatology* 2011, 8(12):711-722.

85. Ransohoff DF: Rules of evidence for cancer molecular-marker discovery and validation. *Nature reviews Cancer* 2004, 4(4):309-314.

86. Simon R: Roadmap for developing and validating therapeutically relevant genomic classifiers. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology* 2005, 23(29):7332-7341.

87. Ochodo EA, Bossuyt PM: Reporting the accuracy of diagnostic tests: the STARD initiative 10 years on. *Clinical chemistry* 2013, 59(6):917-919.

88. McShane LM, Altman DG, Sauerbrei W, Taube SE, Gion M, Clark GM, Statistics Subcommittee of the NCIEWGoCD: Reporting recommendations for tumor marker prognostic studies (REMARK). *Journal of the National Cancer Institute* 2005, 97(16):1180-1184.

89. Hanahan D, Weinberg RA: The hallmarks of cancer. *Cell* 2000, 100(1):57-70.

90. Hanahan D, Weinberg RA: Hallmarks of cancer: the next generation. *Cell* 2011, 144(5):646-674.

91. Markowitz SD, Bertagnolli MM: Molecular origins of cancer: Molecular basis of colorectal cancer. *The New England journal of medicine* 2009, 361(25):2449-2460.

92. Ogino S, Goel A: Molecular classification and correlates in colorectal cancer. *The Journal of molecular diagnostics : JMD* 2008, 10(1):13-27.

93. Lievre A, Bachet JB, Le Corre D, Boige V, Landi B, Emile JF, Cote JF, Tomasic G, Penna C, Ducreux M, Rougier P, Penault-Llorca F, Laurent-Puig P: KRAS mutation status is predictive of response to cetuximab therapy in colorectal cancer. *Cancer research* 2006, 66(8):3992-3995.

94. Sinicrope FA, Sargent DJ: Molecular pathways: microsatellite instability in colorectal cancer: prognostic, predictive, and therapeutic implications. *Clinical cancer research : an official journal of the American Association for Cancer Research* 2012, 18(6):1506-1512.

95. Sargent DJ, Marsoni S, Monges G, Thibodeau SN, Labianca R, Hamilton SR, French AJ, Kabat B, Foster NR, Torri V, Ribic C, Grothey A, Moore M, Zaniboni A, Seitz JF, Sinicrope F, Gallinger S: Defective mismatch repair as a predictive marker for lack of efficacy of fluorouracil-based adjuvant therapy in colon cancer. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology* 2010, 28(20):3219-3226.

96. Jover R, Zapater P, Castells A, Llor X, Andreu M, Cubiella J, Pinol V, Xicola RM, Bujanda L, Rene JM, Clofent J, Bessa X, Morillas JD, Nicolas-Perez D, Paya A, Alenda C, Gastrointestinal Oncology Group of the Spanish Gastroenterological A: Mismatch repair status in the prediction of benefit from adjuvant fluorouracil chemotherapy in colorectal cancer. *Gut* 2006, 55(6):848-855.

97. Ribic CM, Sargent DJ, Moore MJ, Thibodeau SN, French AJ, Goldberg RM, Hamilton SR, Laurent-Puig P, Gryfe R, Shepherd LE, Tu D, Redston M, Gallinger S: Tumor microsatellite-instability status as a predictor of benefit from fluorouracil-based adjuvant chemotherapy for colon cancer. *The New England journal of medicine* 2003, 349(3):247-257.

98. Perez-Villamil B, Romera-Lopez A, Hernandez-Prieto S, Lopez-Campos G, Calles A, Lopez-Asenjo JA, Sanz-Ortega J, Fernandez-Perez C, Sastre J, Alfonso R, Caldes T, Martin-Sanchez F, Diaz-Rubio E: Colon cancer molecular subtypes identified by expression profiling and associated to stroma, mucinous type and different clinical behavior. *BMC cancer* 2012, 12:260.

99. Cancer Genome Atlas N: Comprehensive molecular characterization of human colon and rectal cancer. *Nature* 2012, 487(7407):330-337.

100. Schlicker A, Beran G, Chresta CM, McWalter G, Pritchard A, Weston S, Runswick S, Davenport S, Heathcote K, Castro DA, Orphanides G, French T, Wessels LF: Subtypes of primary colorectal tumors correlate with response to targeted treatment in colorectal cell lines. *BMC medical genomics* 2012, 5:66.

101. De Sousa EMF, Wang X, Jansen M, Fessler E, Trinh A, de Rooij LP, de Jong JH, de Boer OJ, van Leersum R, Bijlsma MF, Rodermond H, van der Heijden M, van Noesel CJ, Tuynman JB, Dekker E, Markowetz F, Medema JP, Vermeulen L: Poor-prognosis colon cancer is defined by a molecularly distinct subtype and develops from serrated precursor lesions. *Nature medicine* 2013, 19(5):614-618.

102. Sadanandam A, Lyssiotis CA, Homicsko K, Collisson EA, Gibb WJ, Wullschleger S, Ostos LC, Lannon WA, Grotzinger C, Del Rio M, Lhermitte B, Olshen AB, Wiedenmann B, Cantley LC, Gray JW, Hanahan D: A colorectal cancer classification system that associates cellular phenotype and responses to therapy. *Nature medicine* 2013, 19(5):619-625.

103. Marisa L, de Reynies A, Duval A, Selves J, Gaub MP, Vescovo L, Etienne-Grimaldi MC, Schiappa R, Guenot D, Ayadi M, Kirzin S, Chazal M, Flejou JF, Benchimol D, Berger A, Lagarde A, Pencreach E, Piard F, Elias D, Parc Y, Olschwang S, Milano G, Laurent-Puig P, Boige V: Gene expression classification of colon cancer into molecular subtypes: characterization, validation, and prognostic value. *PLoS Med* 2013, 10(5):e1001453.

104. Budinska E, Popovici V, Tejpar S, D'Ario G, Lapique N, Sikora KO, Di Narzo AF, Yan P, Hodgson JG, Weinrich S, Bosman F, Roth A, Delorenzi M: Gene expression patterns unveil a new level of molecular heterogeneity in colorectal cancer. *The Journal of pathology* 2013, 231(1):63-76.

105. Roepman P, Schlicker A, Tabernero J, Majewski I, Tian S, Moreno V, Snel MH, Chresta CM, Rosenberg R, Nitsche U, Macarulla T, Capella G, Salazar R, Orphanides G, Wessels LF, Bernards R, Simon IM: Colorectal cancer intrinsic subtypes predict chemotherapy benefit, deficient mismatch repair and epithelial-to-mesenchymal transition. *International journal of cancer Journal international du cancer* 2014, 134(3):552-562.

106. Guinney J, Dienstmann R, Wang X, de Reynies A, Schlicker A, Soneson C, Marisa L, Roepman P, Nyamundanda G, Angelino P, Bot BM, Morris JS, Simon IM, Gerster S, Fessler E, De Sousa EMF, Missiaglia E, Ramay H, Barras D, Homicsko K, Maru D, Manyam GC, Broom B, Boige V, Perez-Villamil B, Laderas T, Salazar R, Gray JW, Hanahan D, Tabernero J *et al*: The consensus molecular subtypes of colorectal cancer. *Nature medicine* 2015.

107. Cancer Genome Atlas N: Comprehensive molecular portraits of human breast tumours. *Nature* 2012, 490(7418):61-70.

108. Rouzier R, Perou CM, Symmans WF, Ibrahim N, Cristofanilli M, Anderson K, Hess KR, Stec J, Ayers M, Wagner P, Morandi P, Fan C, Rabiul I, Ross JS, Hortobagyi GN, Pusztai L: Breast cancer molecular subtypes respond differently to preoperative chemotherapy. *Clinical cancer research : an official journal of the American Association for Cancer Research* 2005, 11(16):5678-5685.

109. Edgar R, Domrachev M, Lash AE: Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic acids research* 2002, 30(1):207-210.

110. Kolesnikov N, Hastings E, Keays M, Melnichuk O, Tang YA, Williams E, Dylag M, Kurbatova N, Brandizi M, Burdett T, Megy K, Pilicheva E, Rustici G, Tikhonov A, Parkinson H, Petryszak R, Sarkans U, Brazma A: ArrayExpress update--simplifying data submissions. *Nucleic acids research* 2015, 43(Database issue):D1113-1116.

111. Singh R, Yang H, Dalziel B, Asarnow D, Murad W, Foote D, Gormley M, Stillman J, Fisher S: Towards human-computer synergetic analysis of large-scale biological data. *BMC bioinformatics* 2013, 14 Suppl 14:S10.

112. Manja V, Lakshminrusimha S: Principles of Use of Biostatistics in Research. *NeoReviews* 2014, 15(4):e133-e150.

113. Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, Speed TP: Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* 2003, 4(2):249-264.

114. Hochberg Y, Benjamini Y: More powerful procedures for multiple significance testing. *Statistics in medicine* 1990, 9(7):811-818.

115. Barabasi AL, Oltvai ZN: Network biology: understanding the cell's functional organization. *Nature reviews Genetics* 2004, 5(2):101-113.

116. Friedman N: Inferring cellular networks using probabilistic graphical models. *Science* 2004, 303(5659):799-805.

117. Ergun A, Lawrence CA, Kohanski MA, Brennan TA, Collins JJ: A network biology approach to prostate cancer. *Molecular systems biology* 2007, 3:82.

118. Kim K, Yang W, Lee KS, Bang H, Jang K, Kim SC, Yang JO, Park S, Park K, Choi JK: Global transcription network incorporating distal regulator binding reveals selective cooperation of cancer drivers and risk genes. *Nucleic acids research* 2015, 43(12):5716-5729.

119. Ogami K, Yamaguchi R, Imoto S, Tamada Y, Araki H, Print C, Miyano S: Computational gene network analysis reveals TNF-induced angiogenesis. *BMC systems biology* 2012, 6 Suppl 2:S12.

120. Chuang HY, Lee E, Liu YT, Lee D, Ideker T: Network-based classification of breast cancer metastasis. *Molecular systems biology* 2007, 3:140.

121. Basso K, Margolin AA, Stolovitzky G, Klein U, Dalla-Favera R, Califano A: Reverse engineering of regulatory networks in human B cells. *Nature genetics* 2005, 37(4):382-390.

122. Margolin AA, Nemenman I, Basso K, Wiggins C, Stolovitzky G, Dalla Favera R, Califano A: ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC bioinformatics* 2006, 7 Suppl 1:S7.

123. Margolin AA, Wang K, Lim WK, Kustagi M, Nemenman I, Califano A: Reverse engineering cellular networks. *Nature protocols* 2006, 1(2):662-671.

124. Walker LC, Waddell N, Ten Haaf A, kConFab I, Grimmond S, Spurdle AB: Use of expression data and the CGEMS genome-wide breast cancer association study to identify genes that may modify risk in BRCA1/2 mutation carriers. *Breast cancer research and treatment* 2008, 112(2):229-236.

125. Jang IS, Margolin A, Califano A: hARACNe: improving the accuracy of regulatory model reverse engineering via higher-order data processing inequality tests. *Interface focus* 2013, 3(4):20130011.

126. Wei EK, Giovannucci E, Wu K, Rosner B, Fuchs CS, Willett WC, Colditz GA: Comparison of risk factors for colon and rectal cancer. *International journal of cancer Journal international du cancer* 2004, 108(3):433-442.

127. Li FY, Lai MD: Colorectal cancer, one entity or three. *Journal of Zhejiang University Science B* 2009, 10(3):219-229.

128. Kim H, Nam SW, Rhee H, Shan Li L, Ju Kang H, Hye Koh K, Kyu Kim N, Song J, Tak-Bun Liu E, Kim H: Different gene expression profiles between microsatellite instability-high and microsatellite stable colorectal carcinomas. *Oncogene* 2004, 23(37):6218-6225.

129. Lothe RA, Peltomaki P, Meling GI, Aaltonen LA, Nystrom-Lahti M, Pylkkanen L, Heimdal K, Andersen TI, Moller P, Rognum TO, et al.: Genomic instability in colorectal cancer: relationship to clinicopathological variables and family history. *Cancer research* 1993, 53(24):5849-5852.

130. Kaz AM, Wong CJ, Dzieciatkowski S, Luo Y, Schoen RE, Grady WM: Patterns of DNA methylation in the normal colon vary by anatomical location, gender, and age. *Epigenetics* 2014, 9(4):492-502.

131. Koestler DC, Li J, Baron JA, Tsongalis GJ, Butterly LF, Goodrich M, Lesseur C, Karagas MR, Marsit CJ, Moore JH, Andrew AS, Srivastava A: Distinct patterns of DNA methylation in conventional adenomas involving the right and left colon. *Modern pathology : an official journal of the United States and Canadian Academy of Pathology, Inc* 2014, 27(1):145-155.

132. Wei C, Chen J, Pande M, Lynch PM, Frazier ML: A pilot study comparing protein expression in different segments of the normal colon and rectum and in normal colon versus adenoma in patients with Lynch syndrome. *Journal of cancer research and clinical oncology* 2013, 139(7):1241-1250.

133. Russo AL, Borger DR, Szymonifka J, Ryan DP, Wo JY, Blaszkowsky LS, Kwak EL, Allen JN, Wadlow RC, Zhu AX, Murphy JE, Faris JE, Dias-Santagata D, Haigis KM, Ellisen LW, Iafrate AJ, Hong TS: Mutational analysis and clinical correlation of metastatic colorectal cancer. *Cancer* 2014, 120(10):1482-1490.

134. Stoll G, Bindea G, Mlecnik B, Galon J, Zitvogel L, Kroemer G: Meta-analysis of organ-specific differences in the structure of the immune infiltrate in major malignancies. *Oncotarget* 2015, 6(14):11894-11909.

135. Pentheroudakis G, Raptou G, Kotoula V, Wirtz RM, Vrettou E, Karavasilis V, Gourgioti G, Gakou C, Syrigos KN, Bournakis E, Rallis G, Varthalitis I, Galani E, Lazaridis G, Papaxoinis G, Pectasides D, Aravantinos G, Makatsoris T, Kalogeras KT, Fountzilas G: Immune response gene expression in colorectal cancer carries distinct prognostic implications according to tissue, stage and site: a prospective retrospective translational study in the context of a hellenic cooperative oncology group randomised trial. *PloS one* 2015, 10(5):e0124612.

136. Chen J, Guo F, Shi X, Zhang L, Zhang A, Jin H, He Y: BRAF V600E mutation and KRAS codon 13 mutations predict poor survival in Chinese colorectal cancer patients. *BMC cancer* 2014, 14:802.

137. Song M, Hu FB, Spiegelman D, Chan AT, Wu K, Ogino S, Fuchs CS, Willett WC, Giovannucci EL: Adulthood Weight Change and Risk of Colorectal Cancer in the Nurses' Health Study and Health Professionals Follow-up Study. *Cancer prevention research* 2015, 8(7):620-627.

138. Schmit SL, Schumacher FR, Edlund CK, Conti DV, Raskin L, Lejbkowicz F, Pinchev M, Rennert HS, Jenkins MA, Hopper JL, Buchanan DD, Lindor NM, Le Marchand L, Gallinger S, Haile RW, Newcomb PA, Huang SC, Rennert G, Casey G, Gruber SB: A novel colorectal cancer risk locus at 4q32.2 identified from an international genome-wide association study. *Carcinogenesis* 2014, 35(11):2512-2519.

139. Regula J, Rupinski M, Kraszewska E, Polkowski M, Pachlewski J, Orlowska J, Nowacki MP, Butruk E: Colonoscopy in colorectal-cancer screening for detection of advanced neoplasia. *The New England journal of medicine* 2006, 355(18):1863-1872.

140. Bujanda L, Sarasqueta C, Cosme A, Hijona E, Enriquez-Navascues JM, Placer C, Villarreal E, Herreros-Villanueva M, Giraldez MD, Gironella M, Balaguer F, Castells A: Evaluation of alpha 1-antitrypsin and the levels of mRNA expression of matrix metalloproteinase 7, urokinase type plasminogen activator receptor and COX-2 for the diagnosis of colorectal cancer. *PloS one* 2013, 8(1):e51810.

141. Sanz-Pamplona R, Berenguer A, Cordero D, Mollevi DG, Crous-Bou M, Sole X, Pare-Brunet L, Guino E, Salazar R, Santos C, de Oca J, Sanjuan X, Rodriguez-Moranta F, Moreno V: Aberrant gene expression in mucosa adjacent to tumor reveals a molecular crosstalk in colon cancer. *Molecular cancer* 2014, 13:46.

142. Stigler SM: Regression towards the mean, historically considered. *Statistical methods in medical research* 1997, 6(2):103-114.

143. Chiolero A, Paradis G, Rich B, Hanley JA: Assessing the Relationship between the Baseline Value of a Continuous Variable and Subsequent Change Over Time. *Frontiers in public health* 2013, 1:29.

144. Lee TI, Young RA: Transcriptional regulation and its misregulation in disease. *Cell* 2013, 152(6):1237-1251.

145. Goodarzi H, Elemento O, Tavazoie S: Revealing global regulatory perturbations across human cancers. *Molecular cell* 2009, 36(5):900-911.

146. Wang K, Saito M, Bisikirska BC, Alvarez MJ, Lim WK, Rajbhandari P, Shen Q, Nemenman I, Basso K, Margolin AA, Klein U, Dalla-Favera R, Califano A: Genome-wide identification of post-translational modulators of transcription factor activity in human B cells. *Nature biotechnology* 2009, 27(9):829-839.

147. Huang FW, Hodis E, Xu MJ, Kryukov GV, Chin L, Garraway LA: Highly recurrent TERT promoter mutations in human melanoma. *Science* 2013, 339(6122):957-959.

148. Sneddon JB, Zhen HH, Montgomery K, van de Rijn M, Tward AD, West R, Gladstone H, Chang HY, Morganroth GS, Oro AE, Brown PO: Bone morphogenetic protein antagonist gremlin 1 is widely expressed by cancer-associated stromal cells and can promote tumor cell proliferation. *Proceedings of the National Academy of Sciences of the United States of America* 2006, 103(40):14842-14847.

149. Jaeger E, Webb E, Howarth K, Carvajal-Carmona L, Rowan A, Broderick P, Walther A, Spain S, Pittman A, Kemp Z, Sullivan K, Heinimann K, Lubbe S, Domingo E, Barclay E, Martin L, Gorman M, Chandler I, Vijayakrishnan J, Wood W, Papaemmanuil E, Penegar S, Qureshi M, Consortium C, Farrington S, Tenesa A, Cazier JB, Kerr D, Gray R, Peto J *et al*: Common genetic variants at the CRAC1 (HMPS) locus on chromosome 15q13.3 influence colorectal cancer risk. *Nature genetics* 2008, 40(1):26-28.

150. Jaeger E, Leedham S, Lewis A, Segditsas S, Becker M, Cuadrado PR, Davis H, Kaur K, Heinimann K, Howarth K, East J, Taylor J, Thomas H, Tomlinson I: Hereditary mixed polyposis syndrome is caused by a 40-kb upstream duplication that leads to increased and ectopic expression of the BMP antagonist GREM1. *Nature genetics* 2012, 44(6):699-703.

151. Galamb O, Wichmann B, Sipos F, Spisak S, Krenacs T, Toth K, Leiszter K, Kalmar A, Tulassay Z, Molnar B: Dysplasia-carcinoma transition specific transcripts in colonic biopsy samples. *PloS one* 2012, 7(11):e48547.

152. Greaves M, Maley CC: Clonal evolution in cancer. *Nature* 2012, 481(7381):306-313.

153. Ahmad FK, Deris S, Othman NH: The inference of breast cancer metastasis through gene regulatory networks. *Journal of biomedical informatics* 2012, 45(2):350-362.

154. Demicheli R, Coradini D: Gene regulatory networks: a new conceptual framework to analyse breast cancer behaviour. *Annals of oncology : official journal of the European Society for Medical Oncology / ESMO* 2011, 22(6):1259-1265.

155. Madhamshettiwar PB, Maetschke SR, Davis MJ, Reverter A, Ragan MA: Gene regulatory network inference: evaluation and application to ovarian cancer allows the prioritization of drug targets. *Genome medicine* 2012, 4(5):41.

156. Logsdon BA, Gentles AJ, Miller CP, Blau CA, Becker PS, Lee SI: Sparse expression bases in cancer reveal tumor drivers. *Nucleic acids research* 2015, 43(3):1332-1344.

157. Osmanbeyoglu HU, Pelossof R, Bromberg JF, Leslie CS: Linking signaling pathways to transcriptional programs in breast cancer. *Genome research* 2014, 24(11):1869-1880.

158. Chiang YT, Wang K, Fazli L, Qi RZ, Gleave ME, Collins CC, Gout PW, Wang Y: GATA2 as a potential metastasis-driving gene in prostate cancer. *Oncotarget* 2014, 5(2):451-461.

159. Weltmeier F, Borlak J: A high resolution genome-wide scan of HNF4alpha recognition sites infers a regulatory gene network in colon cancer. *PloS one* 2011, 6(7):e21667.

160. Vineetha S, Chandra Shekara Bhat C, Idicula SM: Gene regulatory network from microarray data of colon cancer patients using TSK-type recurrent neural fuzzy network. *Gene* 2012, 506(2):408-416.

161. Wang X, Gotoh O: Inference of cancer-specific gene regulatory networks using soft computing rules. *Gene regulation and systems biology* 2010, 4:19-34.

162. Davis H, Irshad S, Bansal M, Rafferty H, Boitsova T, Bardella C, Jaeger E, Lewis A, Freeman-Mills L, Giner FC, Rodenas-Cuadrado P, Mallappa S, Clark S, Thomas H, Jeffery R, Poulsom R, Rodriguez-Justo M, Novelli M, Chetty R, Silver A, Sansom OJ, Greten FR, Wang LM, East JE, Tomlinson I, Leedham SJ: Aberrant epithelial GREM1 expression initiates colonic tumorigenesis from cells outside the stem cell niche. *Nature medicine* 2015, 21(1):62-70.

163. Jiang W, Huang R, Duan C, Fu L, Xi Y, Yang Y, Yang WM, Yang D, Yang DH, Huang RP: Identification of five serum protein markers for detection of ovarian cancer by antibody arrays. *PloS one* 2013, 8(10):e76795.

164. Scheel C, Eaton EN, Li SH, Chaffer CL, Reinhardt F, Kah KJ, Bell G, Guo W, Rubin J, Richardson AL, Weinberg RA: Paracrine and autocrine signals induce and maintain mesenchymal and stem cell states in the breast. *Cell* 2011, 145(6):926-940.

165. Chen X, Ba Y, Ma L, Cai X, Yin Y, Wang K, Guo J, Zhang Y, Chen J, Guo X, Li Q, Li X, Wang W, Zhang Y, Wang J, Jiang X, Xiang Y, Xu C, Zheng P, Zhang J, Li R, Zhang H, Shang X, Gong T, Ning G, Wang J, Zen K, Zhang J, Zhang CY: Characterization of microRNAs in serum: a novel class of biomarkers for diagnosis of cancer and other diseases. *Cell research* 2008, 18(10):997-1006.

166. Shaker O, Maher M, Nassar Y, Morcos G, Gad Z: Role of microRNAs -29b-2, -155, -197 and -205 as diagnostic biomarkers in serum of breast cancer females. *Gene* 2015, 560(1):77-82.

167. Sochor M, Basova P, Pesta M, Dusilkova N, Bartos J, Burda P, Pospisil V, Stopka T: Oncogenic microRNAs: miR-155, miR-19a, miR-181b, and miR-24 enable monitoring of early breast cancer in serum. *BMC cancer* 2014, 14:448.

168. Mitchell PS, Parkin RK, Kroh EM, Fritz BR, Wyman SK, Pogosova-Agadjanyan EL, Peterson A, Noteboom J, O'Briant KC, Allen A, Lin DW, Urban N, Drescher CW, Knudsen BS, Stirewalt DL, Gentleman R, Vessella RL, Nelson PS, Martin DB, Tewari M: Circulating microRNAs as stable blood-based markers for cancer detection. *Proceedings of the National Academy of Sciences of the United States of America* 2008, 105(30):10513-10518.

169. Zheng D, Haddadin S, Wang Y, Gu LQ, Perry MC, Freter CE, Wang MX: Plasma microRNAs as novel biomarkers for early detection of lung cancer. *International journal of clinical and experimental pathology* 2011, 4(6):575-586.

170. Wang WS, Liu LX, Li GP, Chen Y, Li CY, Jin DY, Wang XL: Combined serum CA19-9 and miR-27a-3p in peripheral blood mononuclear cells to diagnose pancreatic cancer. *Cancer prevention research* 2013, 6(4):331-338.

171. Liu J, Gao J, Du Y, Li Z, Ren Y, Gu J, Wang X, Gong Y, Wang W, Kong X: Combination of plasma microRNAs with serum CA19-9 for early detection of pancreatic cancer. *International journal of cancer Journal international du cancer* 2012, 131(3):683-691.

172. Wang K, Yuan Y, Cho JH, McClarty S, Baxter D, Galas DJ: Comparing the MicroRNA spectrum between serum and plasma. *PloS one* 2012, 7(7):e41561.

173. Ohshima K, Inoue K, Fujiwara A, Hatakeyama K, Kanto K, Watanabe Y, Muramatsu K, Fukuda Y, Ogura S, Yamaguchi K, Mochizuki T: Let-7 microRNA family is selectively secreted into the extracellular environment via exosomes in a metastatic gastric cancer cell line. *PloS one* 2010, 5(10):e13247.

174. Ogata-Kawata H, Izumiya M, Kurioka D, Honma Y, Yamada Y, Furuta K, Gunji T, Ohta H, Okamoto H, Sonoda H, Watanabe M, Nakagama H, Yokota J, Kohno T, Tsuchiya N: Circulating exosomal microRNAs as biomarkers of colon cancer. *PloS one* 2014, 9(4):e92921.

175. Lim WK, Lyashenko E, Califano A: Master regulators used as breast cancer metastasis classifier. *Pacific Symposium on Biocomputing Pacific Symposium on Biocomputing* 2009:504-515.

176. Carro MS, Lim WK, Alvarez MJ, Bollo RJ, Zhao X, Snyder EY, Sulman EP, Anne SL, Doetsch F, Colman H, Lasorella A, Aldape K, Califano A, Iavarone A: The transcriptional network for mesenchymal transformation of brain tumours. *Nature* 2010, 463(7279):318-325.

177. Zhang Y, Xuan J, Clarke R, Ressom HW: Module-based breast cancer classification. *International journal of data mining and bioinformatics* 2013, 7(3):284-302.

178. Lefebvre C, Rajbhandari P, Alvarez MJ, Bandaru P, Lim WK, Sato M, Wang K, Sumazin P, Kustagi M, Bisikirska BC, Basso K, Beltrao P, Krogan N, Gautier J, Dalla-Favera R, Califano A: A human B-cell interactome identifies MYB and FOXM1 as master regulators of proliferation in germinal centers. *Molecular systems biology* 2010, 6:377.

179. Sanz-Pamplona R, Berenguer A, Cordero D, Riccadonna S, Sole X, Crous-Bou M, Guino E, Sanjuan X, Biondo S, Soriano A, Jurman G, Capella G, Furlanello C, Moreno V: Clinical value of prognosis gene expression signatures in colorectal cancer: a systematic review. *PloS one* 2012, 7(11):e48877.

Annexos

1. Contribució en altres articles

A continuació es detalla tota una sèrie de publicacions en les que he contribuït com a coautor, fruit de la col·laboració al llarg dels anys amb altres investigadors. Totes elles s'emmarquen dins les línies de recerca de la Unitat de Biomarcadors i Susceptibilitat de l'Institut Català d'Oncologia, i estan relacionades amb el CCR.

La llista complerta de tots els articles es mostra en ordre cronològic invers, i posteriorment la primera pàgina de cadascun d'ells en el format final de la revista on va ser publicat.

- *Llista de publicacions:*

1. Sanz-Pamplona R, Lopez-Doriga A, Paré-Brunet L, Lázaro K, Bellido F, Alonso MH, Aussó S, Guinó E, Beltrán S, Castro-Giner F, Gut M, Sanjuan X, Closa A, **Cordero D**, Morón-Duran FD, Soriano A, Salazar R, Valle L, Moreno V. **Exome Sequencing Reveals AMER1 as a Frequently Mutated Gene in Colorectal Cancer.** Clinical cancer research: an official journal of the American Association for Cancer Research. 2015.

2. Closa A, **Cordero D**, Sanz-Pamplona R, Solé X, Crous-Bou M, Paré-Brunet L, Berenguer A, Guino E, Lopez-Doriga A, Guardiola J, Biondo S, Salazar R, Moreno V. **Identification of candidate susceptibility genes for colorectal cancer through eQTL analysis.** Carcinogenesis. 2014.

3. Sanz-Pamplona R, Berenguer A, **Cordero D**, Molleví DG, Crous-Bou M, Sole X, Paré-Brunet L, Guino E, Salazar R, Santos C, de Oca J, Sanjuan X, Rodriguez-Moranta F, Moreno V. **Aberrant gene expression in mucosa adjacent to tumor reveals a molecular crosstalk in colon cancer.** Mol Cancer. 2014;13(1):46.

4. Crous-Bou M, Rennert G, Cuadras D, Salazar R, **Cordero D**, Saltz Rennert H, Lejbkowicz F, Kopelovich L, Monroe Lipkin S, Bernard Gruber S, Moreno V. **Polymorphisms in alcohol metabolism genes ADH1B and ALDH2, alcohol consumption and colorectal cancer.** PLoS One. 2013;8(11):e80158.

5. Sanz-Pamplona R, Berenguer A, **Cordero D**, Riccadonna S, Sole X, Crous-Bou M, Guinó E, Sanjuan X, Biondo S, Soriano A, Jurman G, Capella G, Furlanello C, Moreno V. **Clinical value of prognosis gene expression signatures in colorectal cancer: a systematic review.** PLoS One. 2012;7(11):e48877.

6. Sanz-Pamplona R, Berenguer A, Sole X, **Cordero D**, Crous-Bou M, Serra-Musach J, Guinó E, Pujana MÁ, Moreno V. **Tools for protein-protein interaction network analysis in cancer research.** Clin Transl Oncol. 2012;14(1):3-14.

7. Obrador-Hevia A, Chin SF, Gonzalez S, Rees J, Vilardell F, Greenson JK, **Cordero D**, Moreno V, Caldas C, Capellá G. **Oncogenic KRAS is not necessary for Wnt signalling activation in APC-associated FAP adenomas.** J Pathol. 2010;221(1):57-67.

Clinical Cancer Research

Exome Sequencing Reveals *AMER1* as a Frequently Mutated Gene in Colorectal Cancer

Rebeca Sanz-Pamplona[1], Adriana Lopez-Doriga[1], Laia Paré-Brunet[1], Kira Lázaro[1], Fernando Bellido[2], M. Henar Alonso[1], Susanna Aussó[1], Elisabet Guinó[1], Sergi Beltrán[3], Francesc Castro-Giner[3], Marta Gut[3], Xavier Sanjuan[4], Adria Closa[1], David Cordero[1], Francisco D. Morón-Duran[1], Antonio Soriano[5], Ramón Salazar[6,7], Laura Valle[2], and Victor Moreno[1,8]

Abstract

Purpose: Somatic mutations occur at early stages of adenoma and accumulate throughout colorectal cancer progression. The aim of this study was to characterize the mutational landscape of stage II tumors and to search for novel recurrent mutations likely implicated in colorectal cancer tumorigenesis.

Experimental Design: The exomic DNA of 42 stage II, microsatellite-stable colon tumors and their paired mucosae were sequenced. Other molecular data available in the discovery dataset [gene expression, methylation, and copy number variations (CNV)] were used to further characterize these tumors. Additional datasets comprising 553 colorectal cancer samples were used to validate the discovered mutations.

Results: As a result, 4,886 somatic single-nucleotide variants (SNV) were found. Almost all SNVs were private changes, with few mutations shared by more than one tumor, thus revealing tumor-specific mutational landscapes. Nevertheless, these diverse mutations converged into common cellular pathways, such as cell cycle or apoptosis. Among this mutational heterogeneity, variants resulting in early stop codons in the *AMER1* (also known as *FAM123B* or *WTX*) gene emerged as recurrent mutations in colorectal cancer. Losses of *AMER1* by other mechanisms apart from mutations such as methylation and copy number aberrations were also found. Tumors lacking this tumor suppressor gene exhibited a mesenchymal phenotype characterized by inhibition of the canonical Wnt pathway.

Conclusion: *In silico* and experimental validation in independent datasets confirmed the existence of functional mutations in *AMER1* in approximately 10% of analyzed colorectal cancer tumors. Moreover, these tumors exhibited a characteristic phenotype. *Clin Cancer Res; 1–10. ©2015 AACR.*

Introduction

Colorectal cancer is the third most common cancer and the second leading cause of cancer death in the world (1). The classic adenoma-to-carcinoma model postulates that colorectal cancer tumorigenesis proceeds through a progressive accumulation of genetic alterations in oncogenes and tumor suppressors genes (2). However, colorectal cancer is currently considered a heterogeneous disease. While tumors fitting into the classic progression model (or chromosomal instability model, CIN) are the most frequent, other tumor phenotypes have been described, such as microsatellite instability (MSI) and CpG island methylator phenotypes (CIMP; ref. 3). Recent studies based on high-throughput technologies have addressed the issue of colorectal cancer molecular complexity, revealing high level of heterogeneity among tumors (4).

Among other biologic mechanisms, it is widely accepted that somatic mutations lead to tumor development in colorectal cancer. It is postulated that most mutations within a tumor are undamaging byproducts of tumorigenesis (passenger mutations) whereas only a few are responsible for driving the initiation and progression of the tumor (driver mutations; ref. 5). In colorectal cancer, a number of mutations have been proposed as drivers, such as those in the *KRAS* and *BRAF* oncogenes, or in the tumor suppressor genes *APC* and *TP53* (6). However, the seminal study by Wood and colleagues revealed that the mutational landscapes of colorectal cancer genomes are composed of a few frequently mutated genes across patients, "mountains," but are dominated by a much larger number of infrequently mutated genes, "hills" (7). Although still controversial, these rarely mutated genes may also contribute to tumor development, thus accounting for intertumor variability (8).

[1]Unit of Biomarkers and Susceptibility, Catalan Institute of Oncology (ICO), Bellvitge Biomedical Research Institute (IDIBELL) and CIBERESP, L'Hospitalet de Llobregat, Barcelona, Spain. [2]Hereditary Cancer Program, Catalan Institute of Oncology (ICO), Bellvitge Biomedical Research Institute (IDIBELL), L'Hospitalet de Llobregat, Barcelona, Spain. [3]Centre Nacional d'Anàlisi Genòmica (CNAG), Barcelona, Spain. [4]Pathology Service, University Hospital Bellvitge (HUB–IDIBELL), L'Hospitalet de Llobregat, Barcelona, Spain. [5]Gastroenterology Service, University Hospital Bellvitge (HUB–IDIBELL), L'Hospitalet de Llobregat, Barcelona, Spain. [6]Department of Medical Oncology, Catalan Institute of Oncology (ICO), Bellvitge Biomedical Research Institute (IDIBELL), L'Hospitalet de Llobregat, Barcelona, Spain. [7]Translational Research Laboratory, Catalan Institute of Oncology (ICO), Bellvitge Biomedical Research Institute (IDIBELL), L'Hospitalet de Llobregat, Barcelona, Spain. [8]Department of Clinical Sciences, Faculty of Medicine, University of Barcelona (UB), Barcelona, Spain.

Note: Supplementary data for this article are available at Clinical Cancer Research Online (http://clincancerres.aacrjournals.org/).

Corresponding Author: Victor Moreno, Catalan Institute of Oncology, Avenue Gran Vía 199-203. 08908, L'Hospitalet de Llobregat, Barcelona 08908, Spain. Phone: 34-932607186; Fax: 34932607188; E-mail: v.moreno@iconcologia.net

doi: 10.1158/1078-0432.CCR-15-0159

Carcinogenesis Advance Access published May 21, 2014
Carcinogenesis vol.00 no.00 p.1 of 8, 2014
doi:10.1093/carcin/bgu092
Advance Access publication April 23, 2014

Identification of candidate susceptibility genes for colorectal cancer through eQTL analysis

Adria Closa[1,2,†], David Cordero[1,2,†], Rebeca Sanz-Pamplona[1,2], Xavier Solé[1,2], Marta Crous-Bou[1,2], Laia Paré-Brunet[1,2], Antoni Berenguer[1,2], Elisabet Guino[1,2], Adriana Lopez-Doriga[1,2], Jordi Guardiola[5], Sebastiano Biondo[3,6], Ramon Salazar[4] and Victor Moreno[1,2,3,*]

[1]Cancer Prevention and Control Program, Catalan Institute of Oncology, and Consortium for Biomedical Research on Epidemiology and Public Health (CIBERESP), Barcelona E08907, Spain, [2]Colorectal Cancer Group, Bellvitge Biomedical Research Institute (IDIBELL), Barcelona E08907, Spain, [3]Department of Clinical Sciences, Faculty of Medicine, University of Barcelona, Barcelona E08907, Spain, [4]Medical Oncology Service, Catalan Institute of Oncology, Barcelona E08907, Spain, [5]Gastroenterology Service, Bellvitge University Hospital, Barcelona E08907, Spain and [6]General and Digestive Surgery Service, Bellvitge University Hospital, Barcelona E08907, Spain

*To whom correspondence should be addressed. Tel: +34 932 607 186; Fax: +34 932 607 188;
Email: v.moreno@iconcologia.net

In this study, we aim to identify the genes responsible for colorectal cancer risk behind the loci identified in genome-wide association studies (GWAS). These genes may be candidate targets for developing new strategies for prevention or therapy. We analyzed the association of genotypes for 26 GWAS single nucleotide polymorphisms (SNPs) with the expression of genes within a 2 Mb region (*cis*-eQTLs). Affymetrix Human Genome U219 expression arrays were used to assess gene expression in two series of samples, one of healthy colonic mucosa (n = 47) and other of normal mucosa adjacent to colon cancer (n = 97, total 144). Paired tumor tissues (n = 97) were also analyzed but did not provide additional findings. Partial Pearson correlation (r), adjusted for sample type, was used for the analysis. We have found Bonferroni-significant *cis*-eQTLs in three loci: rs3802842 in 11q23.1 associated to *C11orf53*, *COLCA1* (*C11orf92*) and *COLCA2* (*C11orf93*; r = 0.60); rs7136702 in 12q13.12 associated to *DIP2B* (r = 0.63) and rs5934683 in Xp22.3 associated to *SHROOM2* and *GPR143* (r = 0.47). For loci in chromosomes 11 and 12, we have found other SNPs in linkage disequilibrium that are more strongly associated with the expression of the identified genes and are better functional candidates: rs7130173 for 11q23.1 (r = 0.66) and rs61927768 for 12q13.12 (r = 0.86). These SNPs are located in DNA regions that may harbor enhancers or transcription factor binding sites. The analysis of *trans*-eQTLs has identified additional genes in these loci that may have common regulatory mechanisms as shown by the analysis of protein–protein interaction networks.

Introduction

Genome-wide association studies (GWAS) have been successful in identifying susceptibility loci for cancer and other diseases, but no progress has been made regarding the functional mechanisms underlying the associations. In colorectal cancer (CRC), 26 susceptibility single nucleotide polymorphisms (SNPs) in 23 different loci have been identified in GWAS to date (Supplementary Table 1, available at *Carcinogenesis* Online) (1–12). Most of them are located in intergenic positions and the genes responsible for the risk modification are unknown. The identification of these genes is important because they may be considered targets for developing new strategies for prevention or therapy (13).

The combination between high throughput genotyping and gene expression profiling technologies allows studying genome-wide associations between genetic polymorphisms and gene expression levels, known as expression quantitative trait loci (eQTL). The identification of eQTL has been proposed as a method to find genes underlying the associations with disease risk (14). The eQTL analysis also has been proposed as a tool to improve the power of GWAS (15) or to engineer genetic-gene expression networks and discover new mechanisms or pathways related to disease (16).

Most analyses of eQTL have used lymphoblastoid cell lines (14), which may not be optimal when the interest is in explaining risk in specific target tissues. Global eQTL analyses of diverse tissues have been done in liver (17), kidney (18) and brain (19), among others. The Genotype-Tissue Expression (GTEx) project (20) aims to create a comprehensive public atlas of gene expression and regulation across multiple human tissues (http://www.broadinstitute.org/gtex). Regarding colon cancer, the interest of analyzing eQTL for GWAS SNPs has been recognized (21) and some of the articles reporting GWAS SNPs have analyzed expression levels in reduced sets of tumors or lymphoblastoid cell lines to document a potential functional role of the SNPs (1,3,5,9). More recently, Loo *et al*. have found interesting associations using expression data assessed in colonic mucosa, either from tumor or normal mucosa adjacent to tumor, though the limited sample size provided low power to identify small associations (22).

In this study, we analyze *cis*- and *trans*-eQTL for GWAS SNPs to identify candidate genes responsible for CRC susceptibility. We combine two series of samples, one of healthy colonic mucosa and other of normal mucosa adjacent to colon cancer. In parallel, we have also analyzed the effect in tumor mucosa, but these data are more difficult to interpret because the expression in tumors is more heterogeneous and is highly altered by diverse mechanisms.

Materials and methods

Subjects and samples

Colon tumor and paired adjacent normal mucosa tissue samples used in this study were selected from a series of cases with a new diagnosis of colon adenocarcinoma attending the University Hospital of Bellvitge in Barcelona between January 1996 and December 2000. Patients included were diagnosed of stage II, microsatellite stable colon cancer, were surgically treated and had not received adjuvant chemotherapy. Adjacent mucosa was obtained from the proximal surgical margins and was at least 10 cm distant from the tumor lesion. Healthy colon mucosa samples were obtained during colonoscopy between February and May 2010. These samples come from a series of unselected patients who underwent a colonoscopy indicated for screening or suspicion of colonic pathology but no colonic lesions were observed. Biopsies were obtained from left and right colon. For this study, we selected randomly approximately half from each site (Supplementary Table 2, available at *Carcinogenesis* Online).

All subjects provided written informed consent to participate in the study and the ethics committee of the hospital cleared the protocol. Additional information about the study can be found at http://www.colonomics.org.

The eQTL analysis was focused on expression data assessed in normal mucosa. Though we initially selected 100 patients and 50 healthy controls, the final sample size after quality control of the data was 144: 47 from healthy donors and 97 adjacent normal mucosa from patients. Gene expression in tumors (n = 97) was also analyzed, and the results compared with those of normal mucosa. Also, for completeness and because we have previously demonstrated in these same samples that the expression in some genes is different between adjacent normal and healthy mucosa (23), we have performed the analyses separated for each tissue (Supplementary File 1, available at *Carcinogenesis* Online).

Abbreviations: CRC, colorectal cancer; GWAS, genome-wide association studies; LD, linkage disequilibrium; PPIN, protein–protein interaction network; SNP, single nucleotide polymorphism.

[†]These authors contributed equally to this work

Sanz-Pamplona et al. Molecular Cancer 2014, 13:46
http://www.molecular-cancer.com/content/13/1/46

MOLECULAR CANCER

RESEARCH Open Access

Aberrant gene expression in mucosa adjacent to tumor reveals a molecular crosstalk in colon cancer

Rebeca Sanz-Pamplona[1], Antoni Berenguer[1], David Cordero[1], David G Molleví[2], Marta Crous-Bou[1], Xavier Sole[1], Laia Paré-Brunet[1], Elisabet Guino[1], Ramón Salazar[2,3], Cristina Santos[2,3], Javier de Oca[4,5], Xavier Sanjuan[6], Francisco Rodriguez-Moranta[7] and Victor Moreno[1,5*]

Abstract

Background: A colorectal tumor is not an isolated entity growing in a restricted location of the body. The patient's gut environment constitutes the framework where the tumor evolves and this relationship promotes and includes a complex and tight correlation of the tumor with inflammation, blood vessels formation, nutrition, and gut microbiome composition. The tumor influence in the environment could both promote an anti-tumor or a pro-tumor response.

Methods: A set of 98 paired adjacent mucosa and tumor tissues from colorectal cancer (CRC) patients and 50 colon mucosa from healthy donors (246 samples in total) were included in this work. RNA extracted from each sample was hybridized in Affymetrix chips Human Genome U219. Functional relationships between genes were inferred by means of systems biology using both transcriptional regulation networks (ARACNe algorithm) and protein-protein interaction networks (BIANA software).

Results: Here we report a transcriptomic analysis revealing a number of genes activated in adjacent mucosa from CRC patients, not activated in mucosa from healthy donors. A functional analysis of these genes suggested that this active reaction of the adjacent mucosa was related to the presence of the tumor. Transcriptional and protein-interaction networks were used to further elucidate this response of normal gut in front of the tumor, revealing a crosstalk between proteins secreted by the tumor and receptors activated in the adjacent colon tissue; and vice versa. Remarkably, Slit family of proteins activated ROBO receptors in tumor whereas tumor-secreted proteins transduced a cellular signal finally activating AP-1 in adjacent tissue.

Conclusions: The systems-level approach provides new insights into the micro-ecology of colorectal tumorogenesis. Disrupting this intricate molecular network of cell-cell communication and pro-inflammatory microenvironment could be a therapeutic target in CRC patients.

Keywords: Colorectal cancer, Network, Microenvironment, Molecular crosstalk, Systems biology

* Correspondence: v.moreno@iconcologia.net
[1]Unit of Biomarkers and Susceptibility, Catalan Institute of Oncology (ICO), Bellvitge Biomedical Research Institute (IDIBELL) and CIBERESP, L'Hospitalet de Llobregat, Barcelona, Spain
[5]Department of Clinical Sciences, Faculty of Medicine, University of Barcelona (UB), Av. Gran Vía 199-203, 08908 L'Hospitalet de Llobregat, Barcelona, Spain
Full list of author information is available at the end of the article

OPEN ACCESS Freely available online

PLOS ONE

Polymorphisms in Alcohol Metabolism Genes *ADH1B* and *ALDH2*, Alcohol Consumption and Colorectal Cancer

Marta Crous-Bou[1,2], Gad Rennert[3,4], Daniel Cuadras[2], Ramon Salazar[2,5], David Cordero[1,2], Hedy Saltz Rennert[3,4], Flavio Lejbkowicz[3,4], Levy Kopelovich[6], Steven Monroe Lipkin[7], Stephen Bernard Gruber[8,9]*, Victor Moreno[1,2,10]*

1 Cancer Prevention and Control Program, Catalan Institute of Oncology, Barcelona, Spain, **2** Colorectal Cancer Group, Bellvitge Biomedical Research Institute and Consorcio de Investigación Biomédica de Epidemiología y Salud Pública (CIBERESP), Barcelona, Spain, **3** Clalit Health Services, National Cancer Control Center, Department of Community Medicine and Epidemiology, Technion-Israel Institute of Technology, Haifa, Israel, **4** B. Rappaport Faculty, Medicine Carmel Medical Center, Technion-Israel Institute of Technology, Haifa, Israel, **5** Medical Oncology Service, Catalan Institute of Oncology, Barcelona, Spain, **6** Division of Cancer Prevention, National Cancer Institute, Rockville, Maryland, United States of America, **7** Department of Medicine, Weill Cornell Medical College, New York, New York, United States of America, **8** Department of Internal Medicine, Epidemiology and Human Genetics, University of Michigan Medical School, Ann Arbor, Michigan, United States of America, **9** University of Southern California Norris Comprehensive Cancer Center, Los Angeles, California, United States of America, **10** Department of Clinical Sciences, Faculty of Medicine, University of Barcelona, Barcelona, Spain

Abstract

Background: Colorectal cancer (CRC) is a leading cause of cancer death worldwide. Epidemiological risk factors for CRC included alcohol intake, which is mainly metabolized to acetaldehyde by alcohol dehydrogenase and further oxidized to acetate by aldehyde dehydrogenase; consequently, the role of genes in the alcohol metabolism pathways is of particular interest. The aim of this study is to analyze the association between SNPs in *ADH1B* and *ALDH2* genes and CRC risk, and also the main effect of alcohol consumption on CRC risk in the study population.

Methodology/Principal Findings: SNPs from *ADH1B* and *ALDH2* genes, included in alcohol metabolism pathway, were genotyped in 1694 CRC cases and 1851 matched controls from the Molecular Epidemiology of Colorectal Cancer study. Information on clinicopathological characteristics, lifestyle and dietary habits were also obtained. Logistic regression and association analysis were conducted. A positive association between alcohol consumption and CRC risk was observed in male participants from the Molecular Epidemiology of Colorectal Cancer study (MECC) study (OR = 1.47; 95%CI = 1.18-1.81). Moreover, the SNPs rs1229984 in *ADH1B* gene was found to be associated with CRC risk: under the recessive model, the OR was 1.75 for A/A genotype (95%CI = 1.21-2.52; p-value = 0.0025). A path analysis based on structural equation modeling showed a direct effect of *ADH1B* gene polymorphisms on colorectal carcinogenesis and also an indirect effect mediated through alcohol consumption.

Conclusions/Significance: Genetic polymorphisms in the alcohol metabolism pathways have a potential role in colorectal carcinogenesis, probably due to the differences in the ethanol metabolism and acetaldehyde oxidation of these enzyme variants.

Citation: Crous-Bou M, Rennert G, Cuadras D, Salazar R, Cordero D, et al. (2013) Polymorphisms in Alcohol Metabolism Genes *ADH1B* and *ALDH2*, Alcohol Consumption and Colorectal Cancer. PLoS ONE 8(11): e80158. doi:10.1371/journal.pone.0080158

Editor: Rui Medeiros, IPO, Inst Port Oncology, Portugal

Received June 13, 2013; **Accepted** September 30, 2013; **Published** November 25, 2013

Funding: This work was supported by a grant from the National Cancer Institute [N01-CN43308 (SML), NCI R01-CA81488 (SBG)], University of Michigan's Cancer Center Support Grant [5 P30 CA46592]. Also the Catalan Institute of Oncology and the Private Foundation of the Biomedical Research Institute of Bellvitge (IDIBELL), the Instituto de Salud Carlos III [grants PI08-1635, PI08-1359, PS09-1037], CIBERESP CB06/02/2005 and the "Acción Transversal del Cancer", the Catalan Government DURSI [grant 2009SGR1489], and the AECC (Spanish Association Against Cancer) Scientific Foundation. USC Norris Cancer Center Support Grant [NCI P30 CA014089 (SBG). No authors have financial conflict of interest with this manuscript. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: sgruber@usc.edu (SBG); v.moreno@iconcologia.net (VM)

Introduction

Colorectal cancer (CRC) is a leading cause of death worldwide, with over one million new cases and half a million deaths around the world every year [1,2]. Risk factors for CRC include advanced age, medical history of benign adenomatous polyps and inflammatory bowel diseases, family history of CRC, low intake of vegetables and fruits and high intake of dietary fat (particularly animal fat) and processed meat [3,4,5,6]. Chronic consumption of non-steroidal anti-inflammatory drugs, hormone replacement therapy and statins are protective [7,8]. The role of other lifestyle factors such as tobacco smoking [9,10] or alcohol consumption [11,12,13,14,15,16] remains inconclusive. Alcohol consumption has been reported to be associated with modest increased risks of CRC in some studies [17], but cancer risk may differ by tumor molecular subtype and anatomical site.

Although the mechanism by which alcohol influences CRC risk also remains not well understood [18], different hypothesis have

OPEN ACCESS Freely available online

PLOS ONE

Clinical Value of Prognosis Gene Expression Signatures in Colorectal Cancer: A Systematic Review

Rebeca Sanz-Pamplona[1☯], Antoni Berenguer[1☯], David Cordero[1], Samantha Riccadonna[2], Xavier Solé[1], Marta Crous-Bou[1], Elisabet Guinó[1], Xavier Sanjuan[3], Sebastiano Biondo[4,5], Antonio Soriano[6], Giuseppe Jurman[2], Gabriel Capella[7], Cesare Furlanello[2], Victor Moreno[1,5]*

1 Unit of Biomarkers and Susceptibility (UBS), Catalan Institute of Oncology (ICO), Bellvitge Biomedical Research Institute (IDIBELL), and CIBERESP, L'Hospitalet de Llobregat, Barcelona, Spain, **2** Predictive Models for Biomedicine & Environment (PMBE), Fondazione Bruno Kessler (FBK), Trento, Italy, **3** Pathology Service, University Hospital Bellvitge (HUB – IDIBELL), L'Hospitalet de Llobregat, Barcelona, Spain, **4** General and Digestive Surgery Service, University Hospital Bellvitge (HUB – IDIBELL), L'Hospitalet de Llobregat, Barcelona, Spain, **5** Department of Clinical Sciences, Faculty of Medicine, University of Barcelona (UB), Barcelona, Spain, **6** Gastroenterology Service, University Hospital Bellvitge (HUB – IDIBELL), L'Hospitalet de Llobregat, Barcelona, Spain, **7** Hereditary Cancer Program, Catalan Institute of Oncology (ICO - IDIBELL), L'Hospitalet de Llobregat, Barcelona, Spain

Abstract

Introduction: The traditional staging system is inadequate to identify those patients with stage II colorectal cancer (CRC) at high risk of recurrence or with stage III CRC at low risk. A number of gene expression signatures to predict CRC prognosis have been proposed, but none is routinely used in the clinic. The aim of this work was to assess the prediction ability and potential clinical usefulness of these signatures in a series of independent datasets.

Methods: A literature review identified 31 gene expression signatures that used gene expression data to predict prognosis in CRC tissue. The search was based on the PubMed database and was restricted to papers published from January 2004 to December 2011. Eleven CRC gene expression datasets with outcome information were identified and downloaded from public repositories. Random Forest classifier was used to build predictors from the gene lists. Matthews correlation coefficient was chosen as a measure of classification accuracy and its associated p-value was used to assess association with prognosis. For clinical usefulness evaluation, positive and negative post-tests probabilities were computed in stage II and III samples.

Results: Five gene signatures showed significant association with prognosis and provided reasonable prediction accuracy in their own training datasets. Nevertheless, all signatures showed low reproducibility in independent data. Stratified analyses by stage or microsatellite instability status showed significant association but limited discrimination ability, especially in stage II tumors. From a clinical perspective, the most predictive signatures showed a minor but significant improvement over the classical staging system.

Conclusions: The published signatures show low prediction accuracy but moderate clinical usefulness. Although gene expression data may inform prognosis, better strategies for signature validation are needed to encourage their widespread use in the clinic.

Citation: Sanz-Pamplona R, Berenguer A, Cordero D, Riccadonna S, Solé X, et al. (2012) Clinical Value of Prognosis Gene Expression Signatures in Colorectal Cancer: A Systematic Review. PLoS ONE 7(11): e48877. doi:10.1371/journal.pone.0048877

Editor: Wayne A. Phillips, Peter MacCallum Cancer Centre, Australia

Received May 30, 2012; **Accepted** October 2, 2012; **Published** November 7, 2012

Funding: This study was supported by the European Commission grant FP7-COOP-Health-2007-B HiPerDART. Also the Instituto de Salud Carlos III grants (FIS PI08/1635, FIS PI08/1359 and FIS PI09-01037), CIBERESP CB07/02/2005, the Spanish Association Against Cancer (AECC) Scientific Foundation, and the Catalan Government DURSI grant 2009SGR1489. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: v.moreno@iconcologia.net

☯ These authors contributed equally to this work.

Introduction

Colorectal cancer (CRC) is the third most common cancer worldwide and the second leading cause of cancer death. During the last decades, incidence has been increasing, while mortality has slowly been decreasing [1]. A remarkable feature of CRC is the difference in prognosis of the early and late stages of the disease: stage I and II have moderate risk of relapse after surgical resection, whereas patients with stage III have a higher chance of recurrence [2]. Recognized clinical risk factors for recurrence are emergency presentation, poorly differentiated tumor, depth of tumor invasion, and adjacent organ involvement (T4) [3–5]. However, these factors are insufficient to identify those patients with stage II CRC at high risk of recurrence and posterior metastasis or those patients with stage III CRC at low risk [6], leading to potential under-treatment or over-treatment [3].

Colon cancer metastasis is a tightly regulated process that requires aberrations in gene expression allowing cancer cells to

Clin Transl Oncol (2012) 14:3-14
DOI 10.1007/s12094-012-0755-9

EDUCATIONAL SERIES Blue Series

ADVANCES IN TRANSLATIONAL ONCOLOGY

Tools for protein-protein interaction network analysis in cancer research

Rebeca Sanz-Pamplona · Antoni Berenguer · Xavier Sole · David Cordero · Marta Crous-Bou · Jordi Serra-Musach · Elisabet Guinó · Miguel Ángel Pujana · Víctor Moreno

Received: 14 July 2011 / Accepted: 20 August 2011

Abstract As cancer is a complex disease, the representation of a malignant cell as a protein-protein interaction network (PPIN) and its subsequent analysis can provide insight into the behaviour of cancer cells and lead to the discovery of new biomarkers. The aim of this review is to help life-science researchers without previous computer programming skills to extract meaningful biological information from such networks, taking advantage of easy-to-use, public bioinformatics tools. It is structured in four parts: the first section describes the pipeline of consecutive steps from network construction to biological hypothesis generation. The second part provides a repository of public, user-friendly tools for network construction, visualisation and analysis. Two different and complementary approaches of network analysis are presented: the topological approach studies the network as a whole by means of structural graph theory, whereas the global approach divides the PPIN into sub-graphs, or modules. In section three, some concepts and tools regarding heterogeneous molecular data integration through a PPIN are described. Finally, the fourth part is an example of how to extract meaningful biological information from a colorectal cancer PPIN using some of the described tools.

Keywords Cancer · Systems biology · Protein-protein interaction network · Public bioinformatics tools · Biomarker discovery

R. Sanz-Pamplona · A. Berenguer · X. Sole · D. Cordero · M. Crous-Bou · J. Serra-Musach · E. Guinó · M.A. Pujana · V. Moreno (✉)
Unit of Biomarkers and Susceptibility
Catalan Institute of Oncology (ICO)
Bellvitge Institute for Biomedical Research (IDIBELL)
Biomedical Research Centre Network for Epidemiology and Public Health (CIBERESP)
Av. Gran Vía, 199
ES-08908 L'Hospitalet de Llobregat, Barcelona, Spain
e-mail: v.moreno@iconcologia.net

V. Moreno
Department of Clinical Sciences
Faculty of Medicine
University of Barcelona
Barcelona, Spain

Introduction

Cancer is a complex disease in which many proteins, genes and molecular processes are implicated [1]. Genes and proteins do not work independently, but are organised into co-regulated units that perform a common biological function. It is the alteration of these functional elements that leads to the development of a particular cancer phenotype (i.e., drug response or disease outcome) and, consequently, their study cannot be tackled from the classical one-gene approach. A systems biology approach, the analysis of the molecular relationship between the implicated genes and proteins as a whole, is required to understand the disease phenotype [2–4].

In this scenario, cancer systems medicine emerges as a translational extension of systems biology that meets the clinical information and the *-omics* disciplines for the classification and diagnosis of cancer subtypes, the prognosis of patient outcomes, the prediction of treatment responses and the identification of perturbation targets for drug development [5, 6].

Proteins interact with each other within a cell, and those interactions can be represented by a network, defined as an abstract representation of nodes or vertices (i.e., proteins)

Journal of Pathology
J Pathol (2010)
Published online in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/path.2685

ORIGINAL PAPER

Oncogenic *KRAS* is not necessary for Wnt signalling activation in APC-associated FAP adenomas

Antònia Obrador-Hevia,[1] Suet-Feung Chin,[2] Sara González,[3] Jonathan Rees,[4] Felip Vilardell,[3] Joel K Greenson,[5] David Cordero,[3] Víctor Moreno,[3] Carlos Caldas[2] and Gabriel Capellá[3]*

[1] *Cancer Cell Biology Group, Institut Universitari d'Investigació en Ciències de la Salut (IUNICS)—Universitat de les Illes Balears, Mallorca, Illes Balears, Spain*
[2] *CRUK Cambridge Research Institute. Li Ka Shing Centre. Robinson Way, Cambridge CB2 0RE, UK*
[3] *Laboratori de Recerca Translacional, Departament de Prevenció i Control del Càncer, Servei d'Epidemiologia i Registre del Càncer, IDIBELL—Institut Català d'Oncologia, L'Hospitalet de Llobregat, Barcelona, Spain*
[4] *Department of Colorectal Surgery, Gloucestershire Royal Hospital, Gloucestershire Hospitals NHS Foundation Trust, Great Western Road, Gloucester, UK*
[5] *Department of Pathology, University of Michigan Heath System, Ann Arbor, USA*

**Correspondence to: Gabriel Capellá Laboratori de Recerca Translacional, Institut Català d'Oncologia, Barcelona, Spain*
e-mail: gcapella@iconcologia.net

Abstract

Recent studies have suggested that APC loss alone may be insufficient to promote aberrant Wnt/β-catenin signalling. Our aim was to comprehensively characterize Wnt signalling components in a set of APC-associated familial adenomatous polyposis (FAP) tumours. Sixty adenomas from six FAP patients with known pathogenic *APC* mutations were included. Somatic *APC* and *KRAS* mutations, β-catenin immunostaining, and qRT-PCR of *APC, MYC, AXIN2* and *SFRP1* were analysed. Array-comparative genomic hybridization (aCGH) was also assessed in 26 FAP adenomas and 24 paired adenoma–carcinoma samples. A somatic *APC* alteration was present in 15 adenomas (LOH in 11 and four point mutations). *KRAS* mutations were detected in 10% of the cases. *APC* mRNA was overexpressed in adenomas. *MYC* and *AXIN2* were also overexpressed, with significant intra-case heterogeneity. Increased cytoplasmic and/or nuclear β-catenin staining was seen in 94% and 80% of the adenomas. β-Catenin nuclear staining was strongly associated with *MYC* levels (*p* value 0.03) but not with *KRAS* mutations. Copy number aberrations were rare. However, the recurrent chromosome changes observed more frequently contained Wnt pathway genes (*p* value 0.012). Based on β-catenin staining and Wnt pathway target genes alterations the Wnt pathway appears to be constitutively activated in all APC-FAP tumours, with alterations occurring both upstream and downstream of APC. Wnt aberrations are present at both the DNA and the RNA level. Somatic profiling of APC-FAP tumours provides new insights into the role of APC in tumourigenesis.

Keywords: colorectal cancer; familial adenomatous polyposis; APC; genomic profiling; Wnt signalling

Received 22 July 2009; Revised 29 December 2009; Accepted 2 January 2010

No conflicts of interest were declared.

Introduction

Classical familial adenomatous polyposis (FAP) is most often caused by truncating germline mutations typically located in the central region of the *Adenomatous Polyposis Coli* (*APC*) tumour suppressor gene [1]. In some cases, missense mutations can also occur [2,3]. *APC* is also somatically mutated in sporadic colorectal cancer (CRC) at a high frequency. *APC* mutations can be detected in aberrant crypt foci, suggesting that the loss of APC function represents an initiating event in CRC [4]. Inactivation of both *APC* alleles is both necessary and sufficient to promote adenoma growth [5].

Molecular analyses of FAP-associated tumours have provided deep insights into tumourigenesis. Biallelic mutation of the *APC* gene is a hallmark of the colorectal, duodenal, and desmoid tumours that develop in FAP patients. The site of the 'first hit' in the *APC* tumour suppressor gene determines the type of the 'second hit'. Mutations near codon 1300 [codons 1285–1378; in the mutation cluster region (MCR)] [6] are associated with loss of heterozygosity (LOH), with no loss of genetic material [7]. More recently, putative 'third hits', mostly copy number gains or deletions, have been reported [8]. Combined profiling of mouse and human adenomas has allowed the identification of new direct and indirect target genes such as *BUB1, MAD2L1*, and *CD44*, which are associated with APC-driven tumour progression [9–14].

The APC protein plays an integral role in the Wnt signalling pathway, as it binds and down-regulates β-catenin [15] by the formation of a protein complex