Integrative approaches for gene and molecular pathway analysis in cancer

Xavier Solé Acha

INTEGRATIVE APPROACHES FOR GENE AND MOLECULAR PATHWAY ANALYSIS IN CANCER

MÈTODES INTEGRATIUS PER A L'ANÀLISI DE GENS I RUTES MOLECULARS EN CÀNCER

Programa de Doctorat en Genètica
Departament de Ciències Clíniques
Facultat de Medicina
Universitat de Barcelona

XAVIER SOLÉ ACHA

Barcelona, Gener de 2012

INTEGRATIVE APPROACHES FOR GENE AND MOLECULAR PATHWAY ANALYSIS IN CANCER

Memòria presentada per
XAVIER SOLÉ ACHA

Per optar al Grau de
DOCTOR

Tesi doctoral realitzada sota la direcció del Dr. Víctor Raúl Moreno Aguado i del Dr. Miguel Ángel Genestar Pujana a la Unitat de Biomarcadors i Susceptibilitat de l'Institut Català d'Oncologia.

Tesi adscrita al Departament de Ciències Clíniques de la Facultat de Medicina, Universitat de Barcelona. Programa de Genètica, bienni 2001-2003 (RD 778/1998).

Víctor Moreno Aguado	**Xavier Solé Acha**	**Miguel A. Genestar Pujana**
Co-director	Doctorand	Co-director

Barcelona, Gener de 2012

Als meus pares.
Aquesta tesi existeix gràcies a vosaltres.

A mi abuelo Gaspar y a mi tío Ramón.
Por la ilusión que sé que les habría hecho verla.

GENERAL CONTENTS

"You take the blue pill — the story ends,
you wake up in your bed and believe
whatever you want to believe.
You take the red pill — you stay in Wonderland
and I show you how deep the rabbit-hole goes".

— Morpheus to Neo, in "The Matrix".

AGRAÏMENTS

Sempre que he llegit els agraïments de tesis alienes (reconeguem-ho, és la part més llegida de les tesis) he pensat que arribar a aquest moment devia ser quelcom especial. Moment de prendre aire i fer una mirada cap enrere, i alhora inevitablement moment de tancar una etapa i mirar al futur amb il·lusions renovades per encetar la recerca de noves experiències.

A més de ser aquest punt d'inflexió, el moment d'escriure els agraïments d'una tesi és doblement grat. Per un costat, significa que el dur i llarg procés que suposa aquesta tasca gairebé arriba a la seva fi. Per altra banda, permet agrair-ho com cal a totes les persones que al llarg d'aquesta travessia han ajudat i col·laborat amb la causa. Cal dir que aquesta tesi no és en cap cas fruit i mèrit exclusiu del doctorand, sinó que és la suma de moltes hores d'ajuda de moltes persones. Ja per avançat vull donar unes gràcies immenses a TOTES les persones que hi han contribuït. Per descomptat, la responsabilitat de qualsevol error la té exclusivament la persona que es dirigeix a vosaltres.

En primer lloc, vull agrair al Xavier Messeguer l'oportunitat que em va donar de saciar el meu cuquet curiós i poder dur a terme un projecte de bioinformàtica a la Facultat d'Informàtica de Barcelona. Recordo amb molt d'afecte les hores que ens vam passar discutint l'algorisme de MALIG i resolent plegats els problemes que anaven sortint. Gràcies a ell em vaig adonar que m'encantava la bioinformàtica, la recerca, i que volia triar la pastilla vermella per veure fins on arribava el cau del conill. I també gràcies a ell vaig acabar treballant a l'ICO. La Universitat necessita pioners com ell.

En segon lloc, vull agrair al Víctor Moreno la inestimable oportunitat que em va donar al permetre'm treballar en el seu grup a l'Institut Català d'Oncologia, confiant en un informàtic que no havia sentit mai les paraules *microarray* o les sigles PCR. D'allò fa més d'onze anys. Junts vam desentortolligar el complex camí de l'anàlisi d'arrays, experiència

gratificant a estones i desesperant moltes altres. Inevitablement, en onze anys també hem viscut travessies pel desert, però al seu costat he pogut aprendre infinitat de coses. És una persona que sorprèn perquè pots parlar amb ell de qualsevol tema i sempre té alguna cosa interessant a aportar. Gràcies per ser un "jefe" tan poc "jefe", per deixar-me fer i desfer segons el meu criteri i per tot el teu suport i confiança.

Aquesta tesi de cap manera existiria sense el Miguel Ángel Pujana. Ell em va rescatar en una d'aquestes "travessies pel desert" meves, fet que no li podré agrair mai prou. Ell em va introduir en el món de la biologia molecular del càncer, la Systems Biology, les xarxes i els factors de transcripció, món que ja no he abandonat. Infinites gràcies també per obrir-me les portes del Califano Lab a la Universitat de Columbia, i gràcies per ensenyar-me a mirar els p-valors sempre des del cantó positiu! ;-)

Al Gabriel Capellá (Gabi) i al Miguel Ángel Peinado (MAP) els vull agrair la seva col·laboració i el seu suport en tot moment durant els meus primers anys d'estada a l'ICO. Gràcies a tots dos en especial per escoltar-me cada cop que us anava a veure per dir-vos "necessito un tema de tesi"!

La Mònica Grau (jefa!) va ser companya d'aventures i desventures en els meus primers anys a l'ICO. Amb ella vaig veure el que era la vocació i el sacrifici per la ciència. Llàstima que determinades circumstàncies la duguessin a canviar de camp. Estic convençut que es va perdre una bona investigadora, però alhora es va guanyar una gran cooperant! ;-) Gràcies per la teva paciència infinita a l'hora de fer-te entendre per un informàtic i explicar-me què era una sonda, un clon, un primer, la RAP-PCR... tantes coses! Per motius personals ara ens veiem poquet, però l'amistat que ens uneix estic convençut que no desapareixerà mai. Molta sort en les teves aventures, saps que sempre estaré a tir de Skype!

Recordo el dia que em van dir que una noia mallorquina començava la tesi al laboratori, es deia Antònia Obrador (Antu!). Quan la vaig sentir parlar per primera vegada em va fer molta gràcia el seu accent illenc, i això va ser un fet que es va mantenir amb el temps, sempre que xerrava amb ella aprenia expressions mallorquines noves! ;-) L'Antònia ha estat per mi un exemple de perseverància i de lluita per fer les coses ben fetes quan sovint l'entorn ho posava difícil. De vegades diuen que en les situacions adverses saps qui és un amic de veritat, i l'Antònia sempre va ser allà quan la necessitava. Antu saps que sempre tindràs un amic a Barcelona, o allà on jo pari. Et desitjo el millor perquè t'ho mereixes!

Al Juan Ramón (JR) González li agraeixo la seva amistat i el seu bon humor al llarg dels anys que vam compartir a l'ICO. Els dinars amb ell eren els millors!

A l'Elisenda Vendrell li agraeixo també la paciència que va tenir amb mi en les etapes en què moltes de les paraules del lab encara em sonaven a xinès, gràcies!

A la Gemma Tarafa li vull agrair que m'ensenyés que els resultats obtinguts a la recerca depenen en gran part de la persona que la duu a terme.

Al compis de dinar (David, Esther, Ferran, Sareta). Mil gràcies per les estones que passem plegats. Per suportar i aconsellar-me amb les meves històries, pels riures i per les divertides converses que tenim d'infinits temes. Ens uneixen la passió culer i moltes altres coses, però per sobre de tot l'amistat! :-)

El Toni (Tom) Berenguer (Bérenguer) sempre ha estat una persona important per a mi a l'ICO. Des que va trucar preguntant per "l'Institut Oncològic de Catalunya", fins ara que treballem plegats. Quan ell estava en el món de l'epidemiologia jo ja li inflava el cap amb les meves històries dels "semàfors". Sempre he pogut comptar amb ell quan l'he necessitat, tant en l'àmbit professional com en el personal. Gràcies per tot!

A l'Olga López li vull agrair l'ajut que m'ha donat ens aquest darrer tram de la tesi. Gràcies per estar sempre disposada a ajudar-nos de tot cor en tot, gràcies per la teva manera de ser i, sobretot, moltíssimes gràcies per aquest optimisme contagiós i per tenir sempre un somriure a la cara quan et venim a veure. Ens alegres el dia a tots!!

També vull agrair a la resta de companys de l'antic SERC i l'actual UBS tots aquests anys de treballar plegats, i la feina que fan per tirar endavant aquest propòsit que tenim que és la recerca contra el càncer.

I would like to thank Andrea Califano and Adolfo Ferrando for allowing me to stay in their labs at Columbia University and work with them. Although the stay was short, I learnt many things, and I could come back to Barcelona with much more than a picture of the Empire State (thanks a lot Adolfo!).

A l'Anna Crous Bou, moltes gràcies per ajudar-me amb la portada de la tesi. Quan un arriba a aquests moments finals on ja no té gaires forces

s'agraeix que algú et doni un cop de mà amb aquests darrers detalls!

Als meus amics: Jandra, Jenry, MA, Marteta, Miquel, Nat, Neus, Pàwer, Psike, Refa, Truman, Vicky, Willy... gràcies per ser allà amb mi sempre que ho he necessitat. Estar amb vosaltres sempre ha estat un bàlsam quan he necessitat esbargir-me de les meves històries. Hem compartit molts anys de les nostres vides, fins i tot amb alguns de vosaltres hem viscut junts. Sempre diuen que l'amistat no es pot pagar amb diners... doncs jo penso que encara que es pogués jo segur que no podria fer-ho, perquè la vostra senzillament NO TINDRIA PREU!! No canvieu mai!!

Infinites gràcies a la Marta. Ets esencial en la meva vida i, com no podia ser d'una altra manera, has estat essencial en la finalització d'aquesta tesi, tant a nivell pràctic com personal. Gràcies pels ànims que sempre m'has donat, pels consells, per les revisions amb lupa i per moltíssimes coses més. La recerca, l'atzar i LOST ens van posar un davant de l'altre, i només això ja ha compensat els moments durs viscuts tots aquests anys. Gràcies per ser com ets, per fer-me riure, per escoltar-me, per aguantar-me i per fer-me costat des del primer dia. Sóc tremendament afortunat d'haver trobat algú excepcional com tu, i sé que la finalització d'aquesta etapa només és per nosaltres l'excusa per iniciar un projecte en comú que de moment ens durà a Boston... i del que tinc ganes de gaudir dia a dia. Un petó enorme!!

Finalment vull donar les gràcies a la meva família: Carles, Marc, Gigi, Meri, i els petits Jan i Fiona, l'alegria de la casa! Però especialment, aquests agraïments no podien acabar d'una altra manera que dirigint-me als meus pares, als quals aquesta tesi va dedicada. La veritat és que tinc tantes coses que agrair-los que no sé per on començar. Evidentment no només els dec aquesta tesi, a ells els ho dec TOT. Gràcies per haver estat uns pares fantàstics, per cuidar-me, recolzar-me i haver estat allà en tot moment. Gràcies per preocupar-vos per mi SEMPRE, per insistir en què estudiés i en què acabés aquesta tesi. Si algun dia tinc fills, només espero poder ser una petita part dels bons pares que sou vosaltres, amb això em donaré per satisfet. Sé que de vegades a un li costa expressar els sentiments, però vull aprofitar aquestes línies per dir-vos que em sento afortunat de tenir-vos com a pares, i que espero que aquesta tesi us compensi ni que sigui una petita part de tot l'esforç que m'heu dedicat al llarg de la vostra vida. Us estimo!

I, sí, ara puc refermar que arribar a aquest moment és especial! :-)

Barcelona, Gener de 2012.

"Caminante, no hay camino, se hace camino al andar".

— Antonio Machado

PREFACE

Over the last two decades we have been given the privilege of witnessing one of the most relevant breakthroughs in the history of biomedicine: the development and completion of the Human Genome Project. Along with it, large-scale laboratory techniques, every day more powerful and reliable, are now routinely applied in biomedical research. Although there is still a long way to go, this fact has undoubtedly set the seed for a new paradigm in cancer research and the future treatment of patients. We are progressively shifting from a scenario where diagnoses and treatments are mainly based on pathological criteria to a completely personalized one, where every single patient will be diagnosed and treated in a specialized manner according to molecular criteria. Nonetheless, to achieve a fully efficient and personalized cancer medicine, it is essential to obtain an accurate picture of all the molecular processes involved in the development of such a complex pathology as cancer. This picture can only be obtained if we can have a detailed view of a tumor cell's status at a whole genome, epigenome, transcriptome and proteome levels. Once all the information is available, suitable analytical integrative techniques must be applied to detect the molecular alterations that play a driver role in the tumorigenic process. The studies presented in this thesis aim to be examples of such integrative analyses.

This thesis is based on a collection of three articles, which are the result of the work done at the Unit of Biomarkers and Susceptibility at the Catalan Institute of Oncology.

PREFACI

Al llarg de les dues darreres dècades hem tingut el privilegi d'assistir a un dels esdeveniments més importants en la història de la biomedicina: el desenvolupament del Projecte Genoma Humà. Conjuntament amb aquest fet, les tècniques de laboratori a gran escala, cada dia més potents i fiables, s'utilitzen actualment de manera rutinària en el camp de la recerca biomèdica. Tot i que encara queda un gran camí per recórrer, aquests fets indubtablement han sembrat la llavor per desenvolupar un nou paradigma de recerca en càncer, així com per millorar els futurs tractaments dels pacients. Ens estem movent progressivament d'un escenari on els diagnòstics i els tractaments es basen principalment en criteris patològics a un altre completament personalitzat, on cada pacient serà diagnosticat i tractat d'una manera especialitzada en funció de criteris moleculars. Tot i així, per assolir una medicina del càncer totalment eficient i personalitzada és essencial obtenir una imatge acurada de tots els processos moleculars involucrats en el desenvolupament d'aquesta malaltia complexa. Per obtenir aquesta imatge necessitarem, doncs, obtenir informació a gran escala de la cèl·lula tumoral a nivell de genoma, epigenoma, transcriptoma i proteoma. Un cop tota la informació està disponible, s'hauran d'aplicar les tècniques analítiques integratives per detectar les alteracions moleculars que tenen un paper primordial el procés tumoral. Els estudis presentats en aquesta tesi pretenen ser una mostra d'aquests tipus d'anàlisis integratives.

Aquesta tesi es basa en una col·lecció de tres articles que són el resultat de la feina duta a terme a la Unitat de Biomarcadors i Susceptibilitat de l'Institut Català d'Oncologia.

RESUM

En aquest apartat es pot trobar un resum en català del treball presentat. Tot i la reducció en el contingut, s'ha intentat que el resum mantingui una coherència global. Els punts més rellevants de la tesi, com són les hipòtesis, els resums dels articles i les conclusions, s'han mantingut íntegres.

INTRODUCCIÓ

El càncer com a malaltia complexa

El càncer es pot definir com una malaltia caracteritzada per una proliferació cel·lular incontrolada i il·limitada, invasió de teixits adjacents i capacitat de disseminació a òrgans distants [81]. Aquest darrer event és la principal causa de mort en la majoria dels casos [40, 192]. El càncer es pot considerar com un paradigma de les malalties complexes, ja que és el resultat d'una intricada xarxa d'interaccions entre factors genètics i ambientals [25, 103]. Tot i que s'han fet grans avenços en aquest camp, molts dels agents específics que influeixen sobre el risc de desenvolupar la patologia, ja siguin ambientals o particulars d'un individu, encara s'han de determinar [153].

Actualment el càncer és considerat com un problema de salut de primera magnitud a tot el món. Segons dades recents, és una de les principals causes de mortalitat a nivell mundial, responsable de 7.6 milions de defuncions l'any 2008 [65]. Aquest xifra es preveu que arribi als 11 milions l'any 2030[1].

La transformació d'una cèl·lula d'un estat normal a un estat tumoral és un procés complex que comprèn diverses etapes. Habitualment, una lesió precursora és la responsable de desencadenar el desenvolupament del tumor [219]. Un cop el procés s'ha iniciat, aquest serà guiat per una complexa combinació d'interaccions entre factors genètics i ambientals [25]. Així, per tal d'obtenir una visió més precisa de l'etiologia de la patologia, l'epidemiologia del càncer ha de ser analitzada des d'una perspectiva tan ambiental com genètica.

1 WHO Fact Sheet Nº 297, February 2009. http://www.who.int/mediacentre/factsheets/fs297/en/index.html. Consultat Setembre 2011.

L'arribada de l'era genòmica en la darrera dècada ha contribuït de manera rellevant a una millor comprensió del paper essencial de les alteracions genètiques en el càncer. Avui dia és ben conegut que el càncer es deu, en gran part, a una acumulació d'alteracions en una cèl·lula, que seran transmeses a la seva progènie [120]. Tot i que actualment moltes d'aquestes alteracions ja estan caracteritzades [66], estudis recents han demostrat que les mutacions en l'àcid desoxirribonucleic (*deoxyribonucleic acid*, DNA) de les cèl·lules tumorals són molt més freqüents del que inicialment s'havia proposat [188]. Aquestes alteracions poden ser heretades dels nostres avantpassats, anomenades germinals, o poden aparèixer en un moment puntual de la vida d'una cèl·lula, anomenades somàtiques [43]. En les cèl·lules dels tumors s'ha observat que les alteracions somàtiques són molt més freqüents que les germinals [72], i l'efecte de les alteracions també és divers, ja sigui activant processos de proliferació com inhibint funcions d'apoptosi o control del creixement [219]. D'altra banda, les alteracions germinals poden ser classificades en alteracions d'alt, moderat o baix risc, en funció de la seva penetrància, és a dir, del risc que confereixen de desenvolupar la malaltia.

Els elements de l'arquitectura del càncer

Durant el procés de la carcinogènesi, les cèl·lules acumulen un elevat nombre d'alteracions genètiques. Aquests canvis en la seqüència de DNA de les cèl·lules tumorals tenen efectes immediats a nivell de RNA i proteïnes. A més, les alteracions epigenètiques de les cèl·lules també han demostrat tenir un paper essencial en el desenvolupament del càncer. Ambdós tipus d'alteracions interaccionen per modificar els programes transcripcionals i promoure un funcionament anormal de la cèl·lula, que és el responsable final de la carcinogènesi.

Les alteracions que es donen en el DNA al llarg del procés tumoral es poden classificar en funció de la mida de la regió que comprenen. Així, podrem tenir alteracions a gran escala, o aberracions, i alteracions a petita escala, o focals. Els efectes d'aquestes alteracions en el fenotip de la cèl·lula són variables.

Les aberracions cromosòmiques es donen de forma habitual a les cèl·lules tumorals. Aquestes alteracions, que comprenen des de milers fins a fins a milions de parells de bases, van ser les primeres a ser detectades, ja que podien ser vistes amb un simple microscopi. De fet, entre finals del segle XIX i principis del XX ja es van postular les primeres hipòtesis sobre el paper de les alteracions somàtiques en el desenvolupament del càncer [144]. La caracterització de totes aquestes aberracions moleculars és útil per obtenir un millor coneixement dels mecanismes

de desenvolupament del tumor, i alhora per ajudar a dissenyar teràpies més efectives i específiques [53]. Les aberracions cromosòmiques poden ser dividides en reordenaments i desequilibris. Els reordenaments consisteixen en recol·locacions de material genètic, i no provoquen un canvi en la quantitat de DNA de la cèl·lula. D'altra banda, els desequilibris impliquen un guany o pèrdua de material genètic. Tant els desequilibris [15, 32, 56, 74, 75, 92, 125, 145, 146, 175, 180, 191, 193, 230] com els reordenaments [130, 135, 144, 163, 184, 208] sovint s'han trobat associats a múltiples tipus de càncer.

Els canvis en el DNA a petita escala comprenen totes aquelles alteracions a la seva seqüència, i poden afectar des d'un fins a uns pocs milers de nucleòtids. Entre totes aquestes alteracions, els canvis en una sola base, anomenats mutacions, són els més freqüents. Degut a deficiències en els seu sistema de control i reparació de dany genòmic, les cèl·lules tumorals acumulen un gran nombre de mutacions puntuals. Moltes d'aquestes mutacions, anomenades passatgeres, tenen un efecte neutre. Altres, en canvi, poden tenir un gran impacte i contribuir en el procés de desenvolupament tumoral [188]. A causa de la grandària del genoma, aquest tipus de canvis no han pogut ser estudiats en profunditat fins que les tècniques d'anàlisi a gran escala no han estat disponibles per a la comunitat científica.

Les variants en nombre de còpies (*copy number variants*, CNVs) són regions genòmiques per les quals s'observen diferències en nombre de còpies en loci específics entre individus [165]. Les CNVs són la forma més freqüent i complexa de diversitat genètica, que altera el paradigma del 'genoma diploide' que s'havia acceptat històricament. Pel que fa a la seva llargada, habitualment es defineixen com regions més llargues de 1000 parells de bases, i en general cobreixen el rang dels segments de DNA d'escala submicroscòpica. Les cèl·lules tumorals solen mostrar alteracions somàtiques de nombre de còpies [8]. De fet, aquest tipus d'alteracions són considerades actualment com un dels mecanismes destacats de desregulació gènica que contribueixen al desenvolupament dels tumors [99].

L'epigenètica estudia els mecanismes heretables durant la mitosi o la meiosi que estan relacionats amb la regulació de l'expressió gènica, i que no impliquen cap modificació a la seqüència genètica de la cèl·lula [18]. L'epigenètica duu a terme un paper rellevant en processos essencials, com el desenvolupament embrionari [115] o el control dels processos cel·lulars mitjançant la generació de patrons d'expressió gènica específics de cada teixit [231].

Les alteracions epigenètiques en els tumors es poden donar al llarg de tot el genoma. Aquestes comprenen la metilació d'illes CpG, hipometilació global, modificació d'histones o desregulació de l'expressió de RNAs petits, entre d'altres. La combinació d'aquests esdeveniments amb altres aberracions genòmiques confereixen a les cèl·lules tumorals un avantatge selectiu, basat en la inhibició de l'apoptosi, una proliferació incontrolada i un potencial de migració. En els darrers anys múltiples estudis han relacionat abastament tant els canvis en la metilació del DNA [30, 49, 51, 58, 83, 86, 95, 97, 100, 106, 139, 206, 212] com la desregulació de RNAs petits [19, 33, 34, 32, 89, 167, 173, 216] amb el procés de desenvolupament dels tumors.

Integració de dades en càncer

Gràcies a l'aplicació en càncer de les eines d'anàlisi a gran escala, en els darrers anys s'han obtingut ingents quantitats d'informació pel que fa a les alteracions en el DNA (p.ex. mutacions, alteracions en nombre de còpies, aberracions estructurals), RNA (p.ex. canvis en l'expressió gènica, alteracions en el mecanisme de la regulació de la transcripció) i alteracions epigenètiques (p.ex. canvis en metilació, modificacions d'histones, canvis en l'expressió de RNAs petits) involucrades en càncer. Donada la complexitat del procés tumoral, les anàlisis integratives de tots aquests diferents tipus de dades a gran escala s'han postulat com una metodologia essencial per obtenir un millor coneixement de les bases moleculars del càncer. A més, aquestes dades faciliten el desenvolupament de millors mecanismes de classificació de la malaltia, basats en criteris moleculars en comptes de patològics.

Al llarg de la darrera dècada han aparegut rellevants exemples d'anàlisis integratives aplicades a la recerca en càncer. El 2003, Lamb i col·laboradors van combinar dades d'expressió gènica de centenars de tumors humans amb anàlisis de promotors per obtenir els mecanismes d'acció de l'oncogen *CCND1* en un ampli espectre de tumors [107]. Altres estudis també han integrat dades d'expressió gènica a gran escala per múltiples tipus de tumors per detectar mòduls de càncer, és a dir, conjunts de gens que actuen de manera conjunta per dur a terme funcions específiques dins del desenvolupament tumoral [181]. Altres anàlisis recents que integren dades d'expressió de múltiples tipus de tumors han permès detectar noves fusions gèniques relacionades amb el càncer de pròstata [200, 202].

La integració de dades de DNA i RNA ha proporcionat resultats prometedors pel que fa a la detecció de nous gens involucrats en el desenvolupament dels tumors. El 2005, Garraway i col·laboradors van

integrar dades d'expressió gènica, pèrdua d'heterozigositat i variants en el nombre de còpies de múltiples melanomes. Aquesta metodologia els va permetre descobrir que el gen *MITF* és la diana d'una amplificació genòmica prèviament desconeguda en aquesta patologia [73]. Més recentment, un altre estudi que integra dades de nombe de còpies de DNA i expressió gènica ha permès caracteritzar la implicació dels gens *TBC1D16* i *RAB27A* en el desenvolupament del melanoma [5]. Estudis integratius de xarxes també han aconseguit identificar gens amb un paper primordial en el desenvolupament dels tumors [38, 112, 118]. Les anàlisis integratives també han demostrat la seva utilitat en el camp de la farmacogenòmica. Un clar exemple és el projecte *Connectivity Map*, que pretén desenmascarar connexions entre patologies, perturbacions genètiques i els mecanismes d'actuació de les drogues mitjançant la integració analítica de grans conjunts de dades d'expressió gènica [108].

Els primers intents d'aplicar metodologies analítiques integratives a la recerca en càncer van aparèixer després que les primeres plataformes d'anàlisi massiu d'expressió gènica i de nombre de còpies de DNA apareguessin a principis d'aquest segle [185]. Alhora van sorgir les primeres iniciatives per integrar conjunts de dades massives de càncer, com el *Cancer Molecular Analysis Project* del *National Cancer Institute* [28]. Tot i que aquells projectes eren altament ambiciosos i innovadors en aquell moment, les seves conclusions van estar limitades pels tamanys de mostra disponibles als estudis i per la tecnologia emprada, que encara estava en una fase primerenca. Així doncs, no va ser fins el 2005 que aquest tipus de metodologia es va començar a aplicar de forma rutinària en l'estudi del càncer.

El paper destacat dels estudis integratius de dades a gran escala en l'anàlisi molecular del càncer ha estat indubtablement reforçat per la creació de grans consorcis internacionals de recerca. Aquests esforços col·lectius són essencials per superar un dels principals obstacles dels estudis integratius, que és la dificultat d'assolir tamanys de mostra suficients per poder detectar amb una elevada fiabilitat alteracions potencialment relacionades amb el càncer. En aquest sentit, els consorcis *The Cancer Genome Atlas* i l'*International Genome Consortium* són els més rellevants dins del camp de la integració de dades en càncer.

HIPÒTESI DE TREBALL I OBJECTIUS

Com a paradigma de les malalties complexes, el càncer es basa en múltiples interaccions entre un gran nombre de factors moleculars i ambientals. Aquesta complexitat inherent dificulta la comprensió dels mecanismes

de susceptibilitat, aparició i desenvolupament de la malaltia.

Les alteracions cel·lulars que contribueixen al desenvolupament del càncer són diverses, i es poden donar a qualsevol nivell molecular, des del DNA fins a les proteïnes. En el passat, la majoria dels estudis en càncer normalment es centraven en un sol d'aquests nivells. Tot i la seva utilitat, aquest tipus d'enfoc acostumava a proporcionar una visió parcial de la cèl·lula tumoral. Així, en els darrers anys s'ha fet palesa la necessitat de dur a terme anàlisis més completes per obtenir una visió acurada dels processos que confereixen a una cèl·lula fenotípicament normal el potencial de proliferar i envair els teixits que l'envolten. L'aparició i el desenvolupament de tècniques d'anàlisi a gran escala, com els *microarrays* de DNA, han contribuït de manera important a aquesta nova manera d'investigar la malaltia.

La hipòtesi de treball d'aquesta tesi és que la integració d'informació heterogènia a gran escala és essencial per descobrir els mecanismes subjacents en malalties complexes, com el càncer.

Objectius generals

El principal objectiu d'aquesta tesi és aprofundir en el coneixement dels mecanismes moleculars implicats en el càncer mitjançant la integració analítica de dades a gran escala a diferents nivells moleculars (DNA, RNA, proteïnes).

Objectius específics

Cadascun dels tres articles presentat en aquesta tesi conté els seus propis objectius específics, que són esmentats a continuació:

Anàlisi integratiu dels gens mutats en càncer de mama

Un estudi pioner publicat el 2006 [188] va determinar la seqüència de 13203 gens que codifiquen per proteïnes en 11 tumors de mama. L'estudi va obtenir una llista d'uns 700 gens que presentaven mutacions somàtiques. Tot i que algunes d'aquestes mutacions ja havien estat descrites amb anterioritat, la majoria d'elles no havien estat prèviament relacionades amb la patologia. Aquest nou conjunt de gens requeria, doncs, una caracterització i un estudi en més profunditat.

Objectius:

1. Caracteritzar els gens que presenten mutacions somàtiques en càncer de mama, a nivell de DNA, RNA i interactoma, per de-

tectar aquells gens amb més potencial d'estar associats al procés tumoral.

Modelització de l'expressió germinal de MYC i susceptibilitat a càncer

Diversos estudis d'associació han trobat repetidament diferents posicions a la regió 8q24 que confereixen una major susceptibilitat a desenvolupar diversos tumors epitelials, essent el colorectal, el de pròstata i el de mama els més rellevants. Sorprenentment, la regió ha estat caracteritzada com a 'desert gènic', pel que el mecanisme d'acció pel qual aquestes variants de risc actuen és encara desconegut.

Objectius:

1. Aclarir el potencial mecanisme d'acció de les variants de risc localitzades a la regió 8q24 mitjançant la integració de dades genètiques i d'expressió obtingudes de teixit prostàtic.

Convergència biològica dels perfils gènics en càncer

En els darrers anys molts estudis han desenvolupat perfils gènics amb capacitat de predir correctament diferents aspectes clínics (pronòstic, resposta al tractament, probabilitat de desenvolupar metàstasi, etc.). Sorprenentment, el solapament dels gens entre els diferents perfils és molt baix, fins i tot per aquells que estan intentant predir el mateix efecte. Això planteja alguns dubtes sobre les implicacions clíniques i biològiques d'aquests perfils.

Objectius:

1. Descobrir quins són els patrons biològics subjacents en diferents perfils de càncer mitjançant la integració de dades de genòmica, transcriptòmica i interaccions de proteïnes.

RESULTATS

Anàlisi integratiu dels gens mutats en càncer de mama

Resum

El 2006, Sjöblom i col·laboradors varen publicar un estudi pioner, en el qual la major part de la regió codificant del genoma humà (*consensus coding sequences*, CCDS) va ser seqüenciada en 11 tumors de mama i 11 colorectals [188]. Els CCDS representen el conjunt de gens més ben

caracteritzat actualment[2]. Tot i el reduït nombre d'individus seqüenciats, aquest projecte va mostrar per primera vegada la visió més completa del conjunt de mutacions que es produeixen en els tumors de mama i colorectal. Tot i que alguns gens ja havien estat descrits prèviament, la majoria no havien estat mai relacionats amb el procés tumoral. Com a conseqüència d'aquesta acumulació d'alteracions, les interaccions moleculars es reprogramen dins el context de les xarxes cel.lulars altament regulades i interconnectades.

L'objectiu del nostre estudi va ser descriure de manera extensa, i a diferents nivell moleculars (del DNA a les proteïnes), l'estat de potencials candidats a oncogens i gens supressors de tumors en càncer de mama. També es volien predir relacions funcionals no descrites prèviament entre ells i plantejar noves hipòtesis en relació a la seva funció molecular coordinada en el procés neoplàsic.

Per investigar el paper dels gens mutats somàticament en càncer de mama com a potencials supressors de tumors o oncogens, vam estudiar la presència de pèrdua d'heterozigositat (*loss of heterozygosity*, LOH) utilitzant dades de polimorfismes a escala del genoma complet. Els gens mutats van mostrar valors de LOH del 4% fins a un màxim del 76% en el cas de *TP53*. Com era d'esperar, altres gens que van mostrar percentatges alts de LOH van ser *BRCA1* (52%) i *MRE11A* (50%).

Per a una millor comprensió dels resultats de l'anàlisi de LOH, es va dur a terme una anàlisi integrativa d'aquestes dades conjuntament amb dades d'expressió gènica. Aproximadament un 50% dels gens mutats van mostrar expressió diferencial entre el teixit normal i el teixit tumoral. Una avaluació detallada dels resultats va assenyalar 12 gens situats en regions crítiques, és a dir, detectades freqüentment alterades en càncer. Les anàlisis d'expressió diferencial varen reforçar la suposició que 10 d'aquests gens podrien actuar com a supressors tumorals, ja que mostren infraexpressió en els tumors de mama.

Les anàlisis de dosi gènica, aplicades sobre el mateix conjunt de dades que les de LOH, van mostrar valors de nombre de còpies compresos entre 1.60 i 3.37 per tumors de mama de tipus basal i no basal, respectivament. Una avaluació detallada de l'expressió gènica a les zones crítiques amb un nombre de còpies superior a dos va permetre identificar 9 potencial oncogens. Cal destacar que, un d'aquests gens, *GAB1*, havia estat prèviament postulat com un oncogèn involucrat en processos de transformació

2 http://www.ncbi.nlm.nih.gov/projects/CCDS/CcdsBrowse.cgi

cel·lular.

La correlació dels resultats de LOH, nombre de còpies i les anàlisis d'expressió diferencial van permetre identificar quatre grups de gens: amplificació i sobreexpressió d'*ABCB10* i *NUP133* en el cromosoma 1 en tumors basals i luminals A i B; pèrdua i infraexpressió de *COL7A1*, *DNASE1L3*, *FLNB* i *RRP9* en el cromosoma 3, particularment per tumors basals i luminals B; pèrdua i infraexpressió de *MAMDC4*, *GSN*, *NUP214*, *SPTAN1* en el cromosoma 9, per tumors luminals A i B; i, finalment, pèrdua i infraexpressió de *SORL1* i *TECTA* en el cromosoma 11 en tumors basals.

Per avaluar més profundament el nivell d'associació funcional entre els gens somàticament mutats en càncer de mama, es van determinar els patrons de coexpressió gènica en unes dades públiques de 98 tumors primaris de mama. Es van detectar nivells de coexpressió més elevats que els esperats per atzar, fet que apunta a una possible associació funcional entre els gens mutats. Globalment, es van obtenir quatre conjunts de gens altament correlacionats, dos representats pels gens *ETV6-NTRK3*, un altre pel gen *TP53* i el darrer pel gen *RB1*.

Mitjançant un conjunt de dades de 113 pacients de càncer de mama que inclou informació de supervivència es va predir el valor pronòstic dels nivells d'expressió gènica d'aquests gens. Aquestes anàlisis van permetre identificar 4 gens els nivells d'expressió dels quals s'associen amb el temps de supervivència: *ABCA3*, *DBN1*, *SP110*, *SPTAN1*.

Per avaluar potencials associacions funcionals entre les proteïnes derivades d'aquests gens, es va procedir a analitzar la xarxa d'interaccions proteïna-proteïna. Els resultats de les anàlisis van mostrar un alt nivell d'interconnexió entre els gens, fet que apunta cap a una potencial associació a nivell de funcions o rutes moleculars.

Per poder generar un model de xarxa amb informació rellevant a nivell biològic pel procès de desenvolupament del càncer de mama, es van integrar diferents tipus de relacions funcionals identificades mitjançant les anàlisis prèviament mencionades. D'aquesta manera, a la xarxa generada, dos gens estaven connectats quan mostraven valors similars de LOH, nombre de còpies o expressió en els tumors de mama, o bé quan les seves corresponents proteïnes estaven directament connectades a la xarxa d'interacció proteïna-proteïna. L'anàlisi de subregions de la xarxa altament interconnectades va permetre identificar mòduls funcionals enriquits en apoptosi, divisió cel·lular, diferenciació cel·lular, senyalització

dels receptors de proteïnes G, senyalització intracel·lular, regulació de la transcripció, regulació de la traducció i transducció de senyal.

Principals resultats

El gen *DBN1* és un candidat a oncogèn que, quan assoleix elevats nivells d'expressió en tumors de mama, prediu baixos nivells de supervivència en les pacients.

Nivells baixos d'expressió d'*ABCA3* i nivells intermedis o baixos d'expressió de *SPTAN1* podrien estar associats també a una pitjor supervivència en pacients amb tumors de mama. *ABCA3* havia estat prèviament identificat com a un gen regulat pel receptor d'estrògens, fet que reforça la seva potencial implicació en el càncer de mama. *SPTAN1* ha estat relacionat amb el desenvolupament de mecanismes de resistència en càncer d'ovari, fent d'aquest gen una potencial diana terapèutica.

Les anàlisis de rutes moleculars en l'interactoma aporten noves hipòtesis per la identificació de gens potencialment associats amb la supervivència de les pacients. *SPTAN1* interacciona amb *GRIN2D* i *SLC9A2*, ambdós dels quals interaccionen amb el producte proteic del proto-oncogèn *ABL1*. Alhora, l'activació de la kinasa *ABL1* promou la capacitat invasora de les cèl·lules tumorals de mama. Com que nivells baixos d'expressió de *SPTAN1* s'associen a una baixa supervivència, *SPTAN1* podria actuar com un regulador negatiu de l'activitat d'*ABL1*.

Modelització de l'expressió germinal de MYC i susceptibilitat a càncer

Resum

La variació genètica germinal en multiples punts de la regió cromosòmica 8q24 ha estat associada a un risc incrementat de desenvolupar determinats tumors, principalment de mama, pròstata i colorectal. Tot i així, cap de les variants actualment descrites es troba a prop de gens coneguts. Només *MYC* es troba a unes quantes kilobases d'aquesta regió. Com que la variació genètica germinal s'ha associat a l'expressió diferencial de molts gens humans, els efectes fenotípics d'aquest tipus de variants podrien tenir un paper rellevant en els processos de susceptibilitat a malalties amb base genètica.

L'objectiu d'aquest estudi va ser integrar dades genètiques i genòmiques per avaluar l'impacte de les variants germinals de la regió 8q24 i el seu paper en la tumorogènesi.

L'associació entre els genotips i els nivells d'expressió de *MYC* es va avaluar utilitzant dades públiques de polimorfismes i expressió d'individus de *HapMap*. Els resultats obtinguts van ser validats en un altre conjunt de dades de mostres de pròstata d'individus sans.

Les anàlisis de nombre de còpies en la regió de *MYC* van ser dutes a terme en els mateixos individus de *HapMap* descrits prèviament, i també en individus no relacionats de la població espanyola. Aquestes anàlisis van mostrar que la variació en nombre de còpies a la regió de *MYC* no sembla contribuir de manera rellevant al risc de càncer de pròstata ni a la sobreexpressió de *MYC* associada a les variants de la regió 8q24.

Utilitzant dades públiques d'expressió gènica que contenien diferents tipus de cèl·lules prostàtiques, es va dur a terme una anàlisi d'expressió diferencial dels gens de la regió 8q24. Els resultats van confirmar un augment de l'expressió de *MYC* a mesura que l'estat patològic del tumor progressa des de la pròstata sana fins a la metàstasi. L'expressió de *MYC* també va correlacionar positivament amb estadis de Gleason elevats. Aquests resultats suggereixen una relació causal entre la sobreexpressió somàtica de *MYC* i les formes més agressives de càncer de pròstata.

Mitjançant l'ús d'un conjunt de dades de 50 teixits sans i 52 tumors de pròstata, es van estudiar les dianes transcripcionals de *MYC* que confereixen un major risc de patir càncer de pròstata per avaluar la seva potencial associació funcional amb el gen. Aquestes anàlisis van revelar una forta correlació entre els nivells d'expressió de *MYC* i el gen supressor tumoral *KLF6*.

Es van inferir xarxes de regulació transcripcional directa en teixit de pròstata mitjançant l'algorisme *ARACNe*. Es van identificar 88 i 11 dianes putatives de *MYC* i *KLF6*, respectivament. La intersecció d'aquests dos conjunts de gens contenia 25 gens en comú, que era un nombre major de l'esperat per atzar. *MYC* i *KLF6* també estaven directament connectats, i en el promotor de *KLF6* es van predir tres llocs d'unió de *MYC*, el que reforça la seva associació funcional i el seu paper en el desenvolupament de tumors de pròstata.

Per tal de validar els resultats obtinguts es van analitzar dades d'expressió d'un model de transformació activat per *MYC* en cèl·lules epitelials mamàries humanes quiescents, així com d'un model murí de tumors de mama activats pel complex *MMTV-Myc*. La majoria dels 25 gens compartits per *MYC* i *KLF6* es van mostrar diferencialment expressats en ambdós models de ratolí, mentre que *KLF6* va mostrar una forta infraexpressió, així com la seva diana transcripcional *CDH1*.

Principals resultats

Es van identificar reguladors en *cis* dels nivells d'expressió de *MYC* en limfòcits immortalitzats d'individus de *HapMap*. Anàlisis quantitatives de l'expressió de *MYC* en teixit de pròstata sana indica una associació entre la sobreexpressió de *MYC* i variants germinals de risc de càncer de pròstata a la regió 8q24.

La sobreexpressió somàtica de *MYC* correlaciona amb la progressió del càncer de pròstata i amb les formes més agressives del tumor.

Anàlisis d'expressió gènica i la modelització de les xarxes de regulació transcripcional prediuen una associació funcional entre *MYC* i el gen supressor tumoral *KLF6*.

Les anàlisis de la transformació cel·lular i la tumorogènesi guiades per *MYC-Myc* suggereixen un model en el que la sobreexpressió de MYC promou la transformació mitjançant la infraexpressió de *KLF6*. En aquest model, un bucle a través de la infraexpressió d'E-cadherina podria causar la reactivació de *MYC*.

Convergència biològica dels perfils gènics en càncer

Resum

Les anàlisis de perfils d'expressió gènica han permès identificar signatures amb capacitat predictiva en càncer que ofereixen una millora respecte els paràmetres histopatològics o clínics històricament aplicats. D'aquesta manera, els perfils d'expressió s'estan incorporant progressivament a la pràctica clínica i aviat prendran una gran rellevància en la presa de decisions del tractament oncològic. No obstant, l'elevada heterogeneitat entre els perfils d'un determinat tipus de càncer ha plantejat alguns dubtes respecte les seves implicacions clíniques i biològiques. Per tal de clarificar aquestes qüestions es requereix un millor coneixement de les propietats moleculars dels gens que formen els diferents perfils, així com la detecció de les possibles interaccions comunes subjacents en aquests.

L'objectiu d'aquest treball va ser integrar dades de genòmica, transcriptòmica i proteòmica per poder posar de manifest les propietats comunes de 24 signatures generades per estudis independents.

Es van estudiar les propietats comunes a nivell genòmic mitjançant l'anàlisi d'enriquiment de motius d'unió de factors de transcripció a les regions promotores dels gens que conformen les signatures. Com a resultat,

a la majoria de signatures es va detectar una sobrerepresentació de motius dels gens de la família *E2F*, reguladors clau de processos de proliferació i mort cel·lular. L'anàlisi de dades d'immunoprecipitació de cromatina va corroborar el paper rellevant dels programes transcripcionals d'*E2F*. A més d'aquests resultats, també es va detectar una sobrerepresentació de llocs d'unió d'*ESR1* en una major part de les signatures, independentment del seu tipus o condició.

Utilitzant conjunts de dades representatius, es van comparar les correlacions entre els nivells d'expressió dels factors de transcripció sobrerrepresentats i els gens associats amb pronòstic o resposta al tractament amb docetaxel en càncer de mama amb les correlacions entre l'expressió dels mateixos factors de transcripció i gens no diferencialment expressats en aquestes condicions. Com a resultat, es van obtenir correlacions significativament més elevades entre els factors de transcripció i els gens associats amb pronòstic o resposta al tractament.

Utilitzant un conjunt de dades de càncer de mama, es va calcular la correlació promig entre tots els gens de cada possible parell de signatures. Comparant-ho amb un conjunt de 10,000 perfils gènics generats a l'atzar, es va observar un augment significatiu de la coexpressió en aproximadament la meitat de les parelles de signatures analitzades. Aquests resultats suggereixen una associació molecular i funcional entre perfils aparentment dissimilars. També es va detectar una forta correlació amb gens involucrats en els processos de mitosi i mort cel·lular per a la majoria dels perfils.

En base a l'evidència de les relacions entre els diferents perfils a nivell de genoma i transcriptoma, es va hipotetitzar que les proteïnes codificades pels gens de les diferents signatures podien trobar-se més a prop a la xarxa d'interaccions proteïna-proteïna que el que s'esperaria per atzar. Emprant dades d'interaccions proteïna-proteïna experimentalment validades, es va detectar que la majoria de les signatures es trobaven més properes entre elles del que s'esperaria per atzar, així com més properes als gens de mitosi i mort cel·lular.

Tots els resultats prèviament obtinguts van ser validats en dos conjunts de dades independents: un perfil de metàstasi de càncer de mama i un altre de resposta a tractament amb cetuximab en càncer colorectal metastàsic.

Principals resultats

Totes les signatures examinades, excepte dues, van mostrar una sobrerrepresentació significativa d'una o més evidències moleculars associades amb la regulació dels processos de mort i proliferació cel·lular.

Es van observar associacions significatives a nivell de genoma, transcriptoma i proteoma, fet que suggereix l'existència d'un fenotip comú de la cèl·lula cancerosa. La convergència en els processos de mort i proliferació cel·lular dóna suport al paper essencial d'aquests processos en el pronòstic, desenvolupament de metàstasi i resposta al tractament.

Es van identificar associacions funcionals i moleculars amb la resposta immune per diferents tipus i condicions de càncer, fet que complementa la contribució dels processos de proliferació i mort cel·lular.

La comprovació d'aquests resultats en conjunts de dades addicionals corrobora els resultats prèviament descrits.

DISCUSSIÓ

Anàlisi integratiu dels gens mutats en càncer de mama

Les anàlisis genòmiques, trancriptòmiques i proteòmiques derivades dels gens mutats en càncer de mama van permetre identificar aquells marcadors potencialment implicats en el desenvolupament d'aquest tipus de tumor. Així, el gen *DBN1*, un gen involucrat en processos de desenvolupament i diferenciació cel·lular que no s'havia associat mai prèviament a càncer, va ser identificat com un potencial oncogèn. Tot i que fins al moment no s'han descrit més associacions entre aquest gen i el càncer de mama, *DBN1* ha estat recentment associat a limfomes de cèl·lules del mantell [220] i ha mostrat la seva possible utilitat clínica en la predicció de pronòstic en càncer de pulmó [136].

En el nostre estudi, els gens *ABCA3* i *SPTAN1*, prèviament poc caracteritzats, van ser identificats com nous gens associats al pronòstic en càncer de mama. Concretament, nivells baixos d'expressió de *SPTAN1* es van associar amb una menor supervivència. Aquesta hipòtesi ha estat posteriorment postulada en un altre estudi [178], on es mostra que la pèrdua d'*ABCA3* s'associa significativament amb l'afectació ganglionar i la infraexpressió del receptor de la progesterona. Aquest estudi també suggereix que la infraexpressió d'*ABCA3* contribueix a un major risc de recurrència, fet que té un impacte directe en la supervivència de les pacients. Més enllà del nostre treball, *SPTAN1* no s'ha tornat a associar

amb càncer de mama, però prèviament s'havia trobat associat a càncer d'ovari, que també és un tumor hormonal altament relacionat amb el càncer de mama [113].

Modelització de l'expressió germinal de MYC i susceptibilitat a càncer

Per avaluar l'associació entre els diferents genotips i els nivells d'expressió de *MYC* es van utilitzar dades d'expressió i de polimorfismes d'individus de *HapMap*, així com de pacients amb càncer de pròstata. Després de realitzar les anàlisis d'associació, es va observar que l'expressió de *MYC* correlacionava amb variants del polimorfisme rs1447295. Tot i que estudis posteriors han trobat resultats negatius respecte a aquesta associació, aquestes diferències podrien ser motivades per temes de poder estadístic, puresa del teixit o per diferències en la quantificació de l'expressió de *MYC*. Altres estudis han trobat que la variant de risc del polimorfisme rs6983267 es relaciona amb un potenciador de l'expressió de *MYC* durant el desenvolupament primerenc de la pròstata [222], suggerint que les variants de risc podrien estar exercint la seva influència significativament abans de la formació del tumor. Aquests resultats reforcen la utilitat dels estudis combinats genoma-transcriptoma en l'estudi de la susceptibilitat en càncer. En un futur proper s'espera que aquests estudis, que combinen variants de risc amb els gens que regulen, puguin tenir una influència en el diagnòstic i tractament de la malaltia.

Convergència biològica dels perfils gènics en càncer

Tot i que els perfils gènics seran una eina essencial en el futur proper en els procediments de diagnòstic i pronòstic del càncer, encara hi ha algunes questions que s'han de clarificar abans que siguin incorporades a la pràctica clínica de manera rutinària. Una d'aquestes qüestions és l'evident variabilitat de les diferents signatures disenyades per predir un mateix fet, com és el cas del pronòstic en el càncer de mama [22, 41, 46, 134, 147, 211, 221]. Una segona qüestió és l'aparent falta de reproductibilitat dels perfils, és a dir, la seva baixa taxa d'encert quan s'utilitzen per classificar altres individus.

A causa d'aquesta gran controvèrsia en els perfils de càncer, l'objectiu del nostre estudi va ser determinar l'existència d'un possible fenotip de la cèl·lula tumoral associat amb múltiples tipus i condicions de càncer. Per dur a terme aquesta tasca es van comparar múltiples perfils a nivell de genoma, transcriptoma i interactoma. El nostre estudi va identificar propietats moleculars comunes no només en signatures de pronòstic, sinó també en signatures de metàstasi i de resposta a tractament. Aquestes

propietats comunes identificades són les de mort i proliferació cel·lular, així com associacions amb la resposta immune. Alguns estudis previs havien trobat aquest tipus de convergències, però només per perfils de pronòstic en càncer de mama [168, 186, 229], mentre que el nostre estudi inclou un conjunt de signatures més complet i divers.

CONCLUSIONS

En aquesta tesi s'han aplicat anàlisis integratives de dades heterogènies de càncer a gran escala en tres escenaris diferents. En el primer estudi es va caracteritzar un conjunt de gens mutats en càncer de mama per identificar quins d'ells podrien estar més potencialment relacionats amb el procés oncogènic. En el segon treball, es van modelitzar dades genètiques i genòmiques per detectar els processos mecanístics que controlen la modulació del risc en una regió específica del genoma associada a càncer de pròstata i altres tipus de tumors. Finalment, en el darrer article es caracteritzen múltiples perfils de càncer a nivell de genoma, transcriptoma i interactoma per avaluar les seves propietats biològiques i determinar l'existència d'un fenotip putatiu comú en les cèl·lules tumorals.

Les conclusions s'exposen per cadascun dels objectius específics exposats a la part inicial d'aquesta tesi. Finalment, s'exposa una conclusió general a mode de resum.

- *Anàlisi integratiu dels gens mutats en càncer de mama*

  - L'anàlisi integratiu de dades de nombre de còpies de DNA i d'expressió gènica senyala el gen *DBN1* com a candidat a oncogèn. Nivells elevats d'expressió de *DBN1* en tumors respecte a teixit sa prediuen baixos nivells de supervivència en pacients de càncer de mama.

  - Valors baixos d'expressió dels gens *ABCA3* i valors mitjans o baixos d'expressió del gen *SPTAN1* podrien predir una pitjor supervivència en pacients de càncer de mama.

  - L'anàlisi de les interaccions dels gens que formen les diverses rutes moleculars proveeix noves hipòtesis per a la identificació de gens potencialment associats amb la supervivència en càncer. *SPTAN1* interacciona amb *GRIN2D* i *SLC9A2*, i alhora aquestes dues proteïnes interaccionen amb el producte del proto-oncogèn *ABL1*. L'activació de la kinasa *ABL1* promou la invasió en cèl·lules tumorals de mama. Com que nivells baixos d'expressió de *SPTAN1* s'associen a baixa supervivència, *SPTAN1* podria estar actuant com un regulador negatiu de l'activitat d'*ABL1*.

- *Modelització de l'expressió germinal de MYC i susceptibilitat a càncer*

- Les anàlisis quantitatives de l'expressió gènica en teixit prostàtic no tumoral suggereixen una relació entre la sobreexpressió de *MYC* amb la regió 8q24-regió 1 i el risc de càncer de pròstata.

- La sobreexpressió germinal de *MYC* podria promoure la transformació cel·lular de l'epiteli normal i, per extensió, el risc de desenvolupar càncer de pròstata mitjançant la repressió del gen supressor tumoral *KLF6*.

- *Convergència biològica dels perfils gènics en càncer*

- S'han observat associacions entre múltiples perfils de càncer a nivell de genoma, transcriptoma i interactoma, fet que suggereix l'existència d'un fenotip comú a la cèl·lula tumoral que influeix de manera decisiva en aspectes crítics de la neoplàsia.

- La convergència en els processos de mort i proliferació cel·lular destaca el paper essencial d'aquests en el pronòstic, desenvolupament de metàstasi i resposta a tractament.

- Addicionalment, es van detectar associacions moleculars i funcionals amb la resposta immune en diferents tipus i condicions de càncer, fet que complementa la contribució dels processos de mort i proliferació cel·lular.

- *Conclusió general*

- L'aplicació de mètodes analítics integratius a dades genòmiques, transcriptòmiques i d'interacció de proteïnes a gran escala és essencial per assolir un millor aprenentatge del càncer. Mitjançant aquests enfocs basats en la biologia de sistemes, no només comprendrem millor les bases moleculars de la malaltia, sinó que també serem capaços d'identificar nous marcadors de diagnòstic, pronòstic, resposta a tractament i noves dianes terapèutiques, fet que pot tenir un impacte decisiu en la presa de decisions a la clínica en els propers anys.

CONTENTS

LIST OF FIGURES

LIST OF ACRONYMS

aCGH	Array-CGH
ALL	Acute Lymphoblastic Leukemia
AML	Acute Myeloid Leukemia
ARACNe	Algorithm for the Reconstruction of Accurate Cellular NEtworks
CGH	Comparative Genomic Hybridization
CCDS	Consensus CoDing Sequences
CDCV	Common Disease-Common Variant
CGC	Cancer Gene Census
CGP	Cancer Genome Project
CIMP	CpG Island Methylator Phenotype
ChIP	CHromatin ImmunoPrecipitation
ChIP-Seq	ChIP SEQuencing
CLL	Chronic Lymphocytic Leukemia
CMAP	Connectivity MAP
CMAP	Cancer Molecular Analysis Project
CN	Copy Number
CNV	Copy Number Variant
CpG	Cytosine-Phosphate-Guanine
CRC	ColoRectal Cancer
DNA	DeoxyriboNucleic Acid
ENCODE	ENCyclopedia of DNA Elements
eQTL	Expression Quantitative Trait Loci
FAP	Familial Adenomatous Polyposis
GEO	Gene Expression Omnibus
GO	Gene Ontology
GSEA	Gene Set Enrichment Analysis
GWAS	Genome-Wide Association Studies
HBV	Hepatitis B Virus
HGP	Human Genome Project

HIV	Human Immunodeficiency Virus
HMM	Hidden Markov Models
HPV	Human Papilloma Virus
HWE	Hardy-Weinberg Equilibrium
IARC	International Agency for Research on Cancer
ICGC	International Cancer Genome Consortium
LD	Linkage Disequilibrium
LOH	Loss Of Heterozygosity
MAF	Minor Allele Frequency
MeDIP-Seq	MEthylated DNA ImmunoPrecipitation SEQuencing
miRNA	Micro RNA
MMTV	Mouse Mammary Tumor Virus
MR	Master Regulator
mRNA	Messenger RNA
NCI	National Cancer Institute
NHGRI	National Human Genome Research Institute
PCA	Principal Components Analysis
PCC	Pearson Correlation Coefficient
PPI	Protein-Protein Interaction
RNA	RiboNucleic Acid
RNA-Seq	RNA SEQuencing
RT-qPCR	Real-Time Quantitative Polymerase Chain Reaction
SAGE	Serial Analysis of Gene Expression
Small RNA-Seq	Small RNA SEQuencing
SNP	Single Nucleotide Polymorphism
TCGA	The Cancer Genome Atlas
TF	Transcription Factor
TFBS	Transcription Factor Binding Site
UHTS	Ultra-High Throughput Sequencing
UTR	UnTranslated Region
VNTR	Variable Number Tandem Repeat
WCRF	World Cancer Research Fund
WHO	World Health Organization

LIST OF GENES

ABCA3	ATP-binding cassette, sub-family A (ABC1), member 3
ABCB10	ATP-binding cassette, sub-family B (MDR/TAP), member 10
ABL1	c-abl oncogene 1, non-receptor tyrosine kinase
AKT1	v-akt murine thymoma viral oncogene homolog 1
APC	adenomatous polyposis coli
ATM	ataxia telangiectasia mutated
BCR	breakpoint cluster region
BCL2	B-cell CLL/lymphoma 2
BRCA1	breast cancer 1, early onset
BRCA2	breast cancer 2, early onset
BRIP1	BRCA1 interacting protein C-terminal helicase 1
CCND1	cyclin D1
CDH1	cadherin 1, type 1, E-cadherin (epithelial)
CDKN2A	cyclin-dependent kinase inhibitor 2A (melanoma, p16, inhibits CDK4)
CHEK2	checkpoint kinase 2
CNNM4	cyclin M4
COL7A1	collagen, type VII, alpha 1
DBN1	drebrin 1
DNASE1L3	deoxyribonuclease I-like 3
ERBB2	v-erb-b2 erythroblastic leukemia viral oncogene homolog 2, neuro/glioblastoma derived oncogene homolog avian)
ESR1	estrogen receptor 1
ETV6	ets variant 6
FLNB	filamin B, beta
FOXP1	forkhead box P1
GAB1	GRB2-associated binding protein 1
GRIN2D	glutamate receptor, ionotropic, N-methyl D-aspartate 2D
GSN	gelsolin

KLF6	Kruppel-like factor 6
KRAS	v-Ki-ras2 Kirsten rat sarcoma viral oncogene homolog
MAMDC4	MAM domain containing 4
MGMT	O-6-methylguanine-DNA methyltransferase
MITF	microphthalmia-associated transcription factor
MLH1	mutL homolog 1, colon cancer, nonpolyposis type 2 (E. coli)
MRE11A	MRE11 meiotic recombination 11 homolog A (S. cerevisiae)
MSH2	mutS homolog 2, colon cancer, nonpolyposis type 1 (E. coli)
MSH6	mutS homolog 6 (E. coli)
MTOR	mechanistic target of rapamycin (serine/threonine kinase)
MYC	v-myc myelocytomatosis viral oncogene homolog (avian)
NBN	nibrin
NCOA2	nuclear receptor coactivator 2
NUP133	nucleoporin 133kDa
NUP214	nucleoporin 214kDa
NTRK3	neurotrophic tyrosine kinase, receptor, type 3
PALB2	partner and localizer of BRCA2
PGR	progesterone receptor
PIK3R1	phosphoinositide-3-kinase, regulatory subunit 1
PMS2	PMS2 postmeiotic segregation increased 2 (S. cerevisiae)
PTEN	phosphatase and tensin homolog
RAB27A	RAB27A, member RAS oncogene family
RB1	retinoblastoma 1
RRP9	ribosomal RNA processing 9, small subunit (SSU) processome component, homolog (yeast)
RYBP	RING1 and YY1 binding protein
SHQ1	SHQ1 homolog (S. cerevisiae)
SLC9A2	solute carrier family 9 (sodium/hydrogen exchanger), member 2
SORL1	sortilin-related receptor, L(DLR class) A repeats containing

SP110	SP110 nuclear body protein
SPTAN1	spectrin, alpha, non-erythrocytic 1 (alpha-fodrin)
TBC1D16	TBC1 domain family, member 16
TECTA	tectorin alpha
TP53	tumor protein p53

Part I

INTRODUCTION

1

CANCER AS A COMPLEX DISEASE

Cancer can be defined as a group of diseases characterized by an uncontrolled and limitless cellular proliferation, invasion of surrounding tissues and capability of dissemination to distant organs [81]. This latter event is the ultimate cause of death in most cases [40, 192]. Cancer can be considered a paradigmatic complex disease, since it arises from an intricate interaction network of genetic and environmental factors [25, 103]. Although much progress has been done in this field, many of the specific agents that influence the risk of developing the pathology, either environmental or host-specific, have yet to be determined [153]. In this chapter, some facts and figures about cancer are introduced, as well as some key points of its genetic and environmental epidemiology. Finally, a general overview of the neoplastic process can be found in section 1.3.

1.1 CANCER FACTS AND FIGURES

Cancer is a worldwide health problem of first magnitude. According to recently collected data, it is one of the leading causes of mortality across the globe, accounting for 7.6 million deaths in 2008 [65]. Moreover, projected mortality rates estimate this figure will rise up to 11 million deaths in 2030[1].

Accurate statistics on cancer occurrence and outcome are primordial, both for the improvement of research and for a better planning and evaluation of cancer control programmes [148]. Mortality and incidence are, therefore, the two mainly used indexes for this purpose. In 2008, the main types of neoplasias leading to overall cancer mortality in Europe [64] were:

- lung (342,100 deaths - 19.9% of all cancer deaths)
- colon and rectum (212,100 deaths - 12.3% of all cancer deaths)
- breast (129,300 deaths - 7.5% of all cancer deaths)
- stomach (116,600 deaths - 6.8% of all cancer deaths)

1 WHO Fact Sheet Nº 297, February 2009. http://www.who.int/mediacentre/factsheets/fs297/en/index.html. Accessed September 2011.

Regarding incidence, a few sites account for most of the newly diagnosed cases. As stated by Ferlay et al. [64], the most incident localizations in Europe in 2008 were:

- colon and rectum (435,600 cases - 13.6% of new cancer cases per year)
- breast (420,800 cases - 13.1% of new cancer cases per year)
- lung (390,900 cases - 12.2% of new cancer cases per year)
- prostate (382,300 cases - 11.9% of new cancer cases per year)

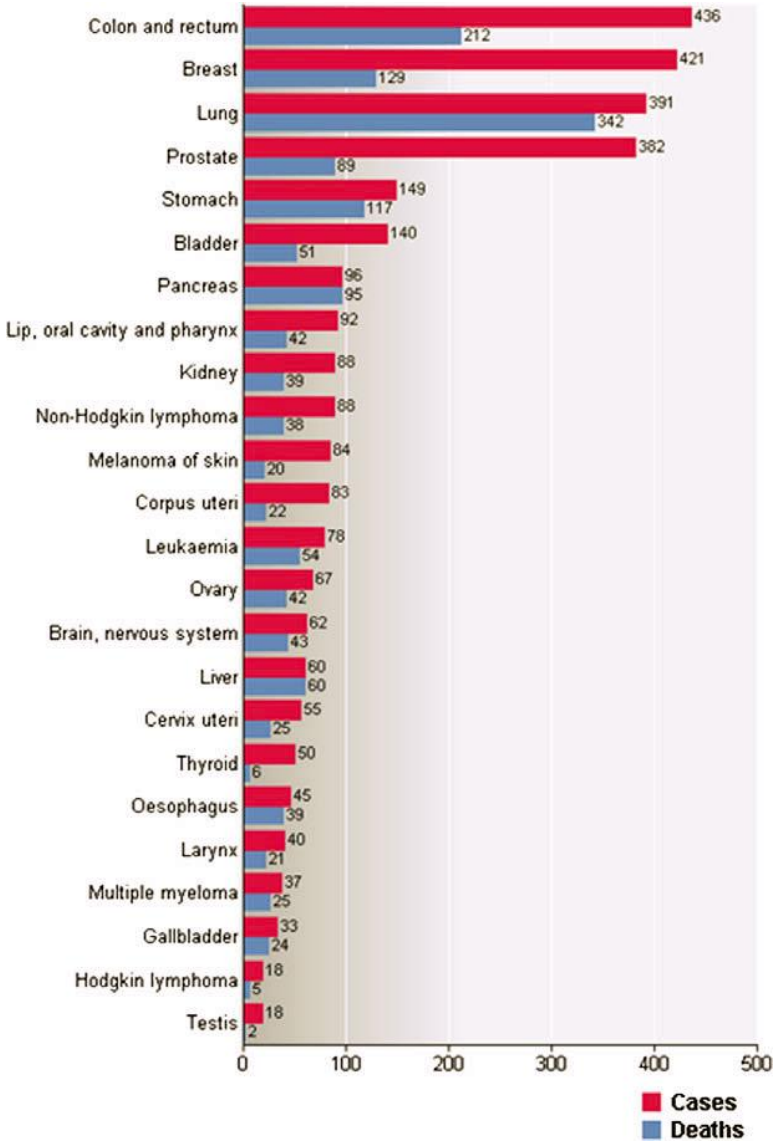


Figure 1: Estimated cancer incidence and mortality for 40 European countries, year 2008. (Adapted from [64].)

From the previous listings, it can be observed that cancer incidence and mortality are not always related. Breast cancer, for instance, accounts for a 13% of the incident cases, but only for a 7.5% of the deaths. This fact is undoubtedly related to well-established screening procedures [27], as well as to improved clinical diagnosis and better targeted therapies recently developed [6]. Contrarily, lung cancer displays a higher mortality than its corresponding incidence, probably due to less effective therapies and a predominantly advanced stage of the disease at the time of diagnosis [150]. This fact can be more clearly seen in Figure 1, where European 2008 estimates of incidence and mortality for different localizations are shown.

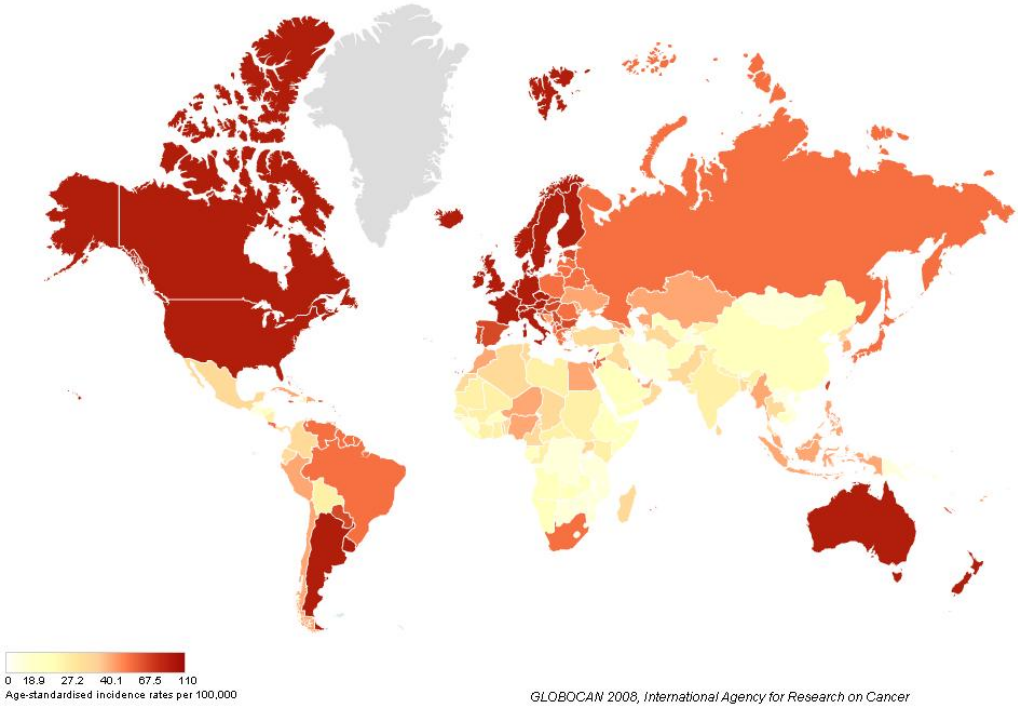


Figure 2: Worldwide distribution of breast cancer incidence. (Source: GLOBOCAN 2008.)

About 63% of all cancer deaths occur in poorly-developed countries [65]. This uneven distribution aggravates even more the burden caused by the disease. Cancer incidence also varies throughout geographic regions. An example of this is displayed in Figure 2, where higher incidence rates of breast cancer can be observed for most Western countries. This variation, which holds true for other tumor sites as well, may be attributed to different genetic background between populations, but more importantly to differential environmental conditions in each geographic region (e. g. carcinogenic agents exposure, cultural and dietary habits) [55].

1.2 CANCER EPIDEMIOLOGY

The transformation of a cell from normal into tumoral status is a progressive, multistage process. This tumorigenic progression usually derives from a precursor lesion, which is responsible for triggering the tumor development [219]. Once the process has been activated, it will be guided by a complex combination of interactions between genetic and environmental factors [25], as depicted in Figure 3. Thus, to obtain a more precise view of the etiology of the disease, cancer epidemiology must be approached both from an environmental and a genetic perspective.

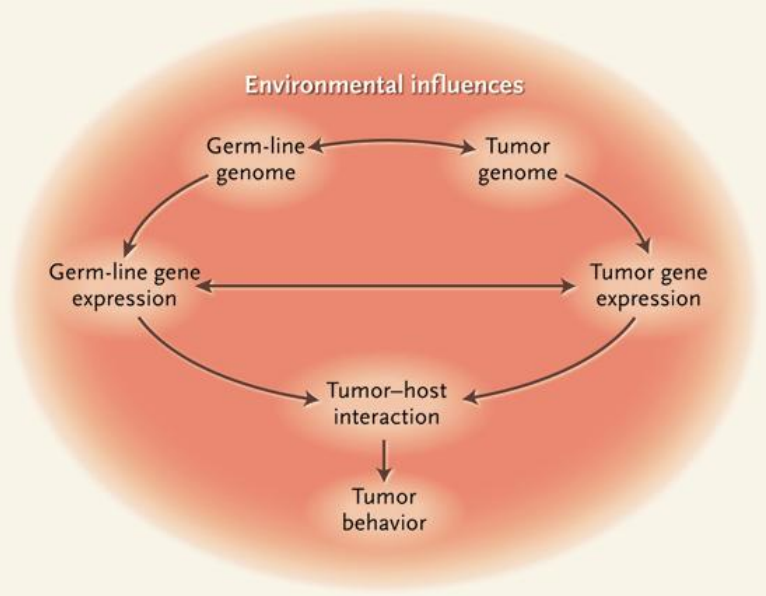


Figure 3: Genetic and environmental factors contribute to the development and progression of complex diseases, such as cancer. This factors not only act alone, but also interact with each other in an entangled manner. (Adapted from [127].)

1.2.1 *Genetic factors*

Cancer has been clearly characterized as an essentially genetic disease [219]. Doubtlessly, the advent of the genomic era at the end of the 20th century has remarkably contributed to better comprehend the predominant role that genetic alterations play in cancer. Nowadays it is well known that cancer, for the most part, is caused by an accumulation of mutations in a single cell and its progeny [120]. In fact, many of these alterations have already been uncovered and characterized [66]. Furthermore, recent landmark studies have demonstrated that deoxyribonucleic acid (DNA) mutations are far more frequent events in neoplasms than it was previously thought [188]. These alterations can be inherited from

our parents (i. e. germline) or can emerge at a specific moment of a cell's life (i. e. somatic) [43]. However, frequency of somatic and inherited mutations in cancer-related genes has been found to be notably different. According to Futreal et al., about 90% of cancer causal genes have somatic mutations, while 20% display germline mutations. Furthermore, only 10% of the genes show both type of alterations [72]. The effect of these mutations on the cell is heterogeneous: while many can be neutral, others may promote abnormal cell growth and proliferation, and others can affect processes such as cell aging, DNA repair or apoptosis [219]. Alterations that affect the proper functioning of the cell are called *driver* mutations. Genes that when mutated give the cell a functional gain are called *oncogenes*. Contrarily, genes that produce cellular loss of functionality if they are altered are called *tumor-suppressor genes*. Oncogenes and tumor-suppressor genes are main contributors to the strong proliferation and dissemination capacity of tumoral cells [14].

One of the main focus of interest of cancer genetic epidemiology is the study of genetic variants that contribute to the risk of developing the pathology. Depending on how much this risk is increased, these variants can be divided into high, moderate or low penetrance. Although landmark twin studies have argued that about a third of the variation in cancer risk has a genetic basis [117], cancer Mendelian disorders (i. e. caused by high-penetrance mutations) account only for a rough 5% of the cases [39]. Given the high effect of these mutations, which makes them easier to be detected, it is not unrealistic to believe that probably only a few more recessive rare familial cancer syndromes remain to be uncovered [39]. The unexplained large proportion of genetic factors that may have an impact on the predisposition to cancer are currently attributed to lower penetrance variants, which may in conjunction have a huge influence on cancer susceptibility at the population level [39].

1.2.1.1 *High-risk mutations: cancer inheritance*

It has been clearly observed that relatives of patients with cancer are at a higher risk of developing a tumor at the same site [154]. This clustering of a relatively large number of cancer cases within families may indicate that an inherited mutation in one gene is sufficient to substantially increase risk. During the 1980s and 1990s, linkage and positional cloning analyses led to the identification of high-penetrance cancer susceptibility genes. Indeed, specific germline mutations in different genes have been identified through sequencing of affected family members. Examples include the *TP53* gene and Li-Fraumeni cancer syndrome, which predisposes to childhood sarcoma, brain tumors, as well as early-onset breast and/or ovarian cancer [60, 214]; the *ATM* gene and ataxia telangiectasia, which in-

creases susceptibility to lymphoma, T-cell leukemia or breast cancer [111]; *APC* mutations are linked to the development of colorectal cancer (CRC) in nearly a 100% of the cases [124]; mutations in the mismatch repair machinery genes, which *MLH1*, *MSH2*, *MSH6* and *PMS2* are involved in Lynch Syndrome [157]; *CDKN2A* mutations, which are involved in pancreatic cancer [123]; mutations in the *RB1* gene, which predispose to retinoblastoma, bladder cancer or osteosarcoma [29]. Germline mutations in *BRCA1* and *BRCA2* genes, both of which are involved in DNA repair, have also been identified as increasing susceptibility to breast and ovarian cancers [215]. Although these high-penetrance germline mutations substantially increase the risk of developing a specific cancer, the vast majority of people who develop the disease do not carry these mutations. Overall, the proportion of cancer classified as being attributable to dominantly inherited high-risk genes is estimated to be as much as a 5% of all cancers that occur in the general population, as it was previously stated [39].

1.2.1.2 *Low-risk mutations: cancer susceptibility*

Cancer occurrences that are not apparently caused by any known high-penetrance variant are usually called *sporadic*. However, there is still an unexplained excess of familial risk observed for this type of tumors. This excess has been attributed to the accumulation of an undefined number of low-to-moderate effect alleles. Much laboratory and epidemiological research over the last decade has focused on the identification of these genetic variants, which may have important effect at the population level because of their large number [91, 155]. Although the exact number of low-penetrance alleles contributing to the disease is not clear yet, some approximations can be done, depending on the allele frequencies of the potentially involved variants and their expected relative risk (Fig. 4). Recent advances in the optimization of large-scale technology has allowed to genotype hundreds of thousands of single nucleotide polymorphisms (SNP) for sets of thousands of samples with almost perfect accuracy [62]. This fact has yielded the discovery of many common genetic variants (i. e. allele frequency $\geqslant 5\%$) with low contribution to the genetic risk of developing cancer (i.e. $\leqslant$ 1.5-fold). Actually, by the 1st quarter of 2011 more than 1300 loci have been found to be significantly related to different diseases, with cancer among them[2].

The theory that many common genetic variants with low penetrance are responsible for the inherited susceptibility of developing cancer, known as the common disease-common variant (CDCV) hypothesis [48], has

2 NHGRI Catalog of Genome-Wide Association Studies. http://www.genome.gov/gwastudies/. Accessed September 2011.

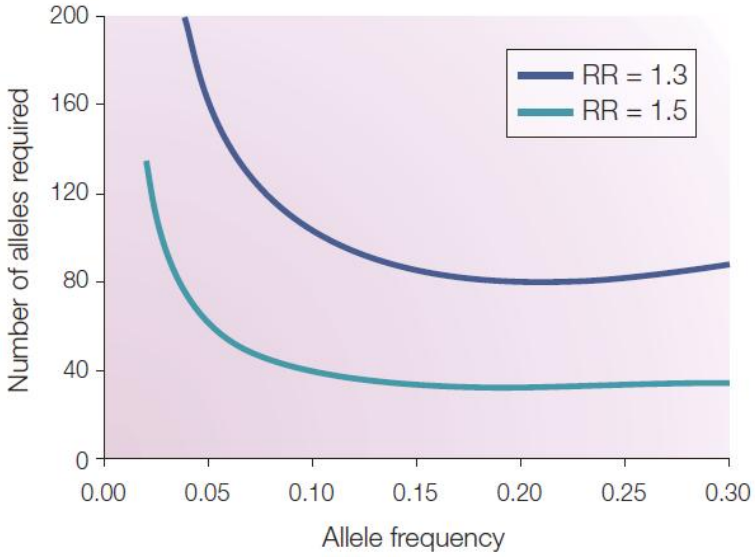


Figure 4: Number of alleles required to explain excess of familial risk, based on a codominant model. The excess risk could be explained by a modest number of common variants or a large number of rare variants. (Adapted from [155].)

been accepted for several years. Nonetheless, there is still a big deal of genetic contribution to cancer that cannot be explained by all the variants currently known. Sample sizes used in the latest genome-wide association studies (GWAS) are mostly unable to detect effects smaller than 1.1-fold, and it is believed that about 50,000 cases and 50,000 controls would be required to detect the expected ~800 very-low-penetrance variants that may contribute to a single cancer type [151]. Although the underlying mechanisms of action of these variants in cancer are unknown, slight differences in the expression of one of the two alleles have been recently detected to be associated with a genetic predisposition to the disease [52]. Well-described examples of these low-to-mid-penetrance rare cancer variants include breast cancer susceptibility genes *ATM*, *CHEK2*, *BRIP1*, *PALB2* and *NBN* [90].

1.2.2 *Environmental factors*

In epidemiology, environmental factors can be defined as those which have an influence on the probability of developing a disease but are not inherited from our ancestors. This factors can be related both to exposures to external agents and lifestyle. In general, the list of carcinogenic agents

provided by the International Agency for Research on Cancer (IARC)[3] can be classified into three main groups:

- Physical carcinogens, such as solar or ionizing radiation.
- Chemical carcinogens, such as asbestos, tobacco smoke, aflatoxin or arsenic-contaminated water.
- Biological carcinogens, such as infections from certain viruses (e. g. human immunodeficiency virus (HIV), hepatitis B virus (HBV), human papilloma virus (HPV)), bacteria (e. g. *Helicobacter pylori*) or parasites (e. g. *Schistosoma*).

Ageing is another fundamental factor for the development of most adult human cancers, since the incidence of the disease rises dramatically with age (Fig. 5). Indeed, age might be seen as the most important risk factor for carcinogenesis. This fact reinforces the idea that cancer is, in most of the cases, a slow-evolving process that requires years of interactions between genetic and environmental factors. The events that enable the development of the malignant process along with aging can be divided into molecular, cellular and physiological [11]. Examples for each one of these categories are:

- Molecular: accumulation of DNA adducts, DNA methylation changes.
- Cellular: senescence of fibroblasts associated with production of tumor growth factors and metalloproteinases that favor metastatic spread; premature senescence associated with loss of apoptosis and development of immortal cells.
- Physiological: decline of the immune system might favor the growth of highly immunogenic tumors; premature senescence of stromal cells associated with increased production of growth factors and metalloproteinases; increased concentration of catabolic cytokines in the circulation, which might lead to muscle loss and oppose the growth of highly proliferative tissues and neoplasias; decline in DNA repair capacity.

According to the World Cancer Research Fund (WCRF) report, major risk factors related to lifestyle comprise food intake, nutrition, physical activity, alcohol consumption and smoking habits.[4].

3 IARC Monographs on the Evaluation of Carcinogenic Risks to Humans. http://monographs.iarc.fr/ENG/Classification/index.php. Accessed September 2011.

4 http://www.dietandcancerreport.org. Accessed September 2011.

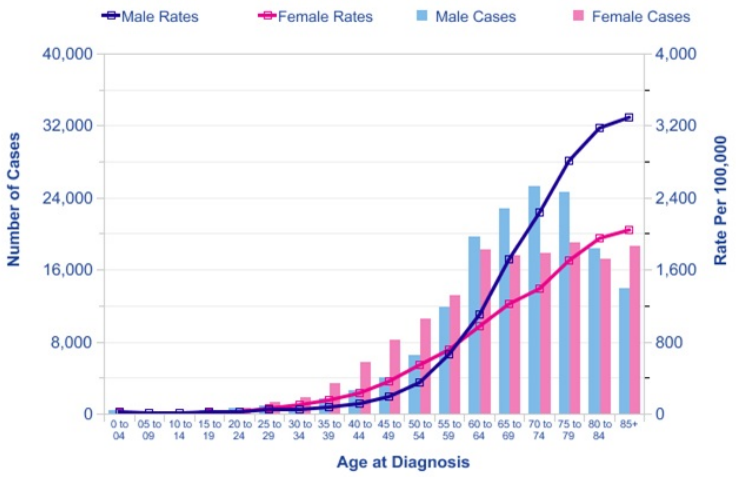


Figure 5: Average number of new cases per year and age-specific incidence rates, UK, 2006-2008. (Source: Cancer Research UK — http://info.cancerresearchuk.org/cancerstats/incidence/age/. Accessed September 2011.)

1.2.3 *Gene-environment interactions*

Genetic and environmental modifiers of the risk of developing complex diseases do not work independently [94]. That is, the effect of environmental exposures on each person may vary depending on their genetic background. This fact, commonly known as gene-environment interactions, was first described at the beginning of the 20th century [25]. However, the influence of genes and environment is not the same for all tumors, with some examples displayed in Figure 6.

Genes \ Environment	Low	High
Low	*Sporadic* (Glioblastoma)	*Induced* (Mesothelioma)
High	*Inherited* (Retinoblastoma)	*Interactive* (Colorectal)

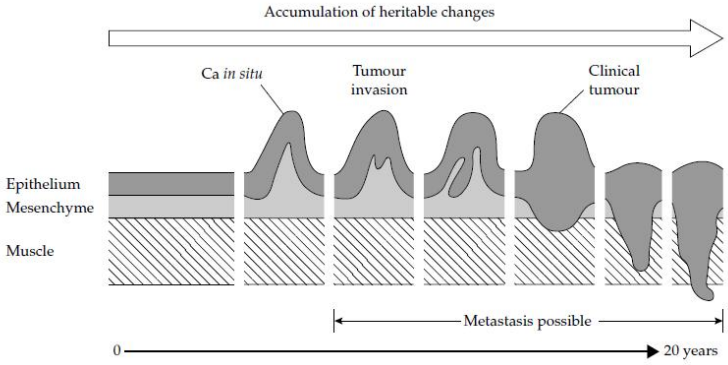
Figure 6: Some types of tumors may be primarily influenced by genes or environment alone, while others can be strongly affected by the interaction of both factors. (Adapted from [122].)

Estimating only the separate contributions of genes and environment to a disease, ignoring their interactions, may lead to an incorrect assessment of the proportion of the disease that is explained by genes, the environment and their joint effect. Despite the big sample sizes required

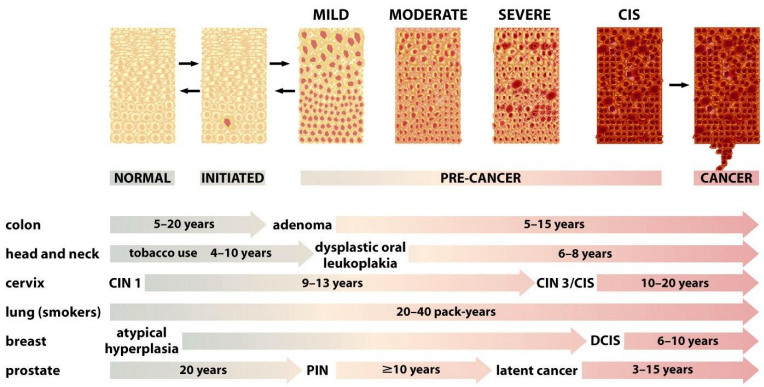
for this type of approaches, well-known gene-environment interactions for different types of cancer can be found in the literature [42, 88, 226].

1.3 THE NEOPLASTIC PROCESS

As it was stated in section 1.2, carcinogenesis is a complex and multi-stage process, which can usually span over decades from the first pre-malignant lesion until an invasive tumor appears (Fig. 7a). Furthermore, instead of being particular of a few tumor sites, this slow progression occurs for many different types of cancer, as exemplified in Figure 7b. This evolving process is driven by a series of sequentially acquired random mutations, as well as epigenetic alterations, that affect genes controlling essential cellular processes such as cell death and proliferation [219].



(a) Graphical representation of the tumor development process, from normal to invasive tumor, via accumulation of heritable changes. (Adapted from *Introduction to the Cellular and Molecular Biology of Cancer, OUP, 2005*.)



(b) Diagram of cancer progression for different types of cancer and their timeline. (Adapted from *The Biology of Cancer, Garland Science, 2007*.)

Figure 7: Schemas of cancer progression.

1.3.1 *Cancer initiation: the two hit paradigm*

In 1971, Knudson discovered that for the oncogenesis of childhood retinoblastoma only two mutations were needed [104]. These *driver* mutations promote the inactivation of a well-characterized tumor suppressor gene, *RB1*. When the two mutational events, known as the "two hits", affect both alleles of the gene (Fig. 8) the tumor starts developing. Furthermore, if an individual inherits one mutation from their parents, then only one more hit will be required to develop the disease. Besides retinoblastoma, it has also been observed that this paradigm not only applies to childhood tumors but also to adults as well in the case of high penetrance mutations. Although other variants of cancer initiation have been lately described (e. g. haploinsufficiency [176] or the three-hit model [182]), what Knudson postulated has been thoroughly accepted as one of the main theories of tumor formation, describing the role of recessive tumour suppressor genes in dominantly inherited cancer susceptibility syndromes [16]. Allelic loss (i. e. loss of heterozygosity, LOH), small-scale transcription-truncating mutations or even gene promoter methylation are some of these major genetic and epigenetic events currently known to contribute to deregulate protein function and trigger the carcinogenic process [203].

A major connotation of the two-hit model is that a disease that follows a recessive inheritance pattern could be transmitted under a high penetrance dominant model if the probability of somatic mutations in the wild type allele was high. In familial cancer, the affected person has only one wild-type copy of the gene in their cells, since they have inherited a mutated allele from one of their parents. A second somatic mutation occurring hereafter in the target tissue inactivates the remaining functional allele inherited from the other parent. Thus, cancer will be more probable in those individuals who carry a heterozygous germline mutation (i. e. those with a predisposition to cancer). The need for two hits explains why not all people belonging to high cancer-risk families develop a malignant pathology: inheritance of just one genetic defect predisposes a person to cancer but does not cause it, since this second event is required. Major examples of tumor suppressors that can trigger cancer after two genetic hits are *APC*, *BRCA1*, *BRCA2*, *MLH1*, *MSH2*, *MSH6*, *TP53* or *PTEN* [203].

1.3.2 *Multi-stage clonal expansion of tumors*

It is now generally accepted that most sporadic solid tumors result from a series of clonal expansions and a multistep process of accumulation of cellular genetic alterations. This model of tumor progression was

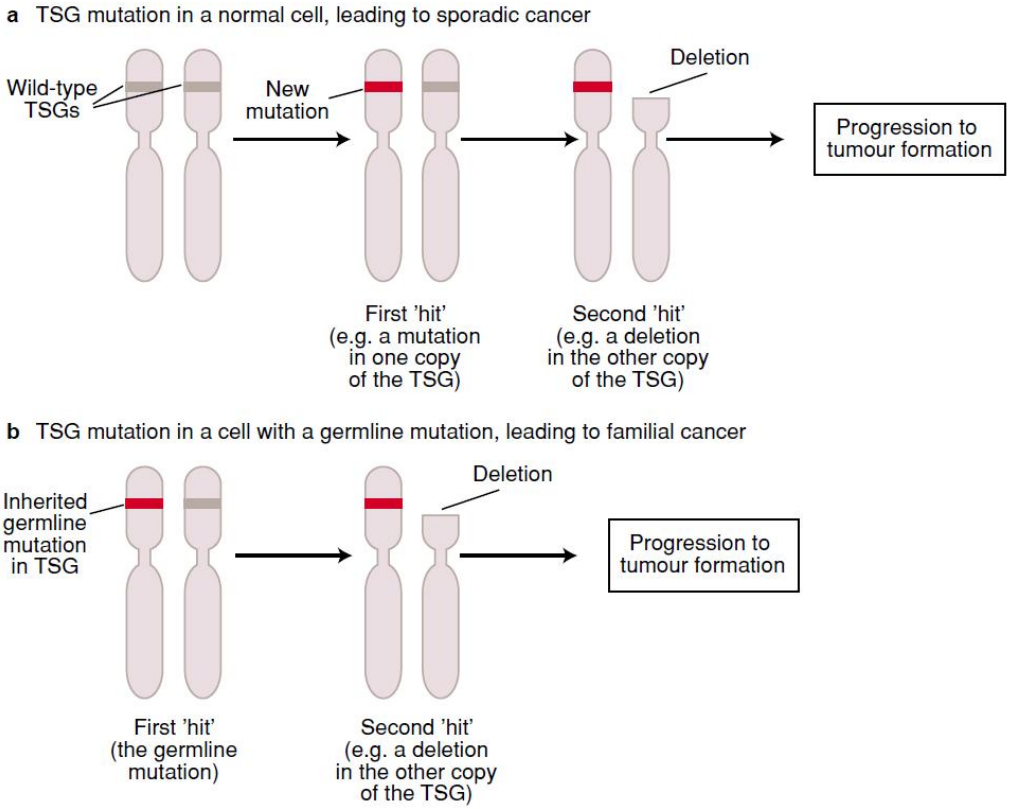


Figure 8: Graphical representation of Knudson's two-hit paradigm of cancer development. If the first hit is already inherited, only one more will be needed during a person's lifetime to trigger cancer initiation. (Extracted from [169].)

proposed by Nowell in 1976 [143]. This author was the first to present genomic instability as a genetic-variation generation mechanism, onto which afterwards the selective pressure would act (Fig. 9). Both oncogene and tumor-suppressor mutations are involved and accumulated in one cell and its direct descendants by a process known as clonal evolution. These inappropriately dividing cells copy their DNA and give identical sets to their offspring. One of these cells or its descendants undergoes a mutation that further enhances its ability to escape normal regulation. Repetition of the process enables one cell to accumulate the mutations it needs to metastasize and colonize other organs. Each tumor cell clone follows its own genetic path as it evolves towards malignancy. The proliferative model proposed for colorectal cancer has become a paradigm of this clonal expansion process [63]. These type of neoplasms progress through different stages ranging from benign adenomas to malignant and invasive carcinomas.

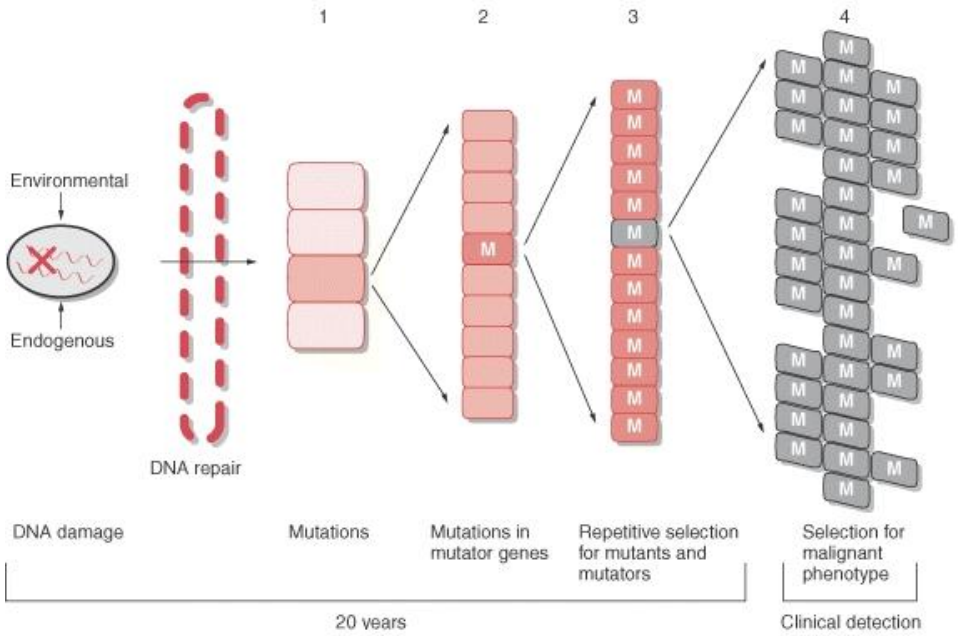


Figure 9: Cancer clonal expansion model. (Extracted from *Holland-Frei Cancer Medicine, 6th edition, BC Decker, 2003*.)

1.4 THE COMPLEXITY OF CANCER CELL PHYSIOLOGY

As it has been stated, cancer is a paradigmatic example of what are known as complex diseases. It involves a large number of alterations in the cell physiology, which ultimately lead to malignant tumors. Moreover, these alterations not only are inherited from our parents, but also are strongly influenced by environmental exposures and lifestyle habits. The ability to invade surrounding tissues and distant organs is the primary cause of death for most cancer patients.

The biological processes that guide the transformation of normal cells into tumoral cells have been a matter of study during many decades. Most of the increase in patient survival rates along the last decades is due to improved intervention protocols. Contrarily, clinical treatment of metastatic tumors still remains a challenge today [47]. Besides, cancer origin still remains quite uncertain, with different models of carcinogenesis still being proposed [218]. Although the molecular processes involved in the tumorigenesis are very specific, there is a vast number of non-specific factors that can trigger the tumor formation (e. g. ionizing radiation, viruses, chronical inflammation, among others). This apparent heterogeneity of tumor triggering events hinders the development of effective methodologies for an effective management of the disease.

The list of physiological alterations of the cell that have been found to be inherently related to the carcinogenic process was described by Hanahan and Weinberg in [81] and [82] (Figure 10). These alterations, common to most of the tumors, comprise:

- Sustaining proliferative signaling.
- Evading growth suppressors.
- Resisting cell death.
- Enabling replicative immortality.
- Inducing angiogenesis.
- Activating invasion and metastasis.
- Deregulating cellular energetics.
- Avoiding immune destruction.

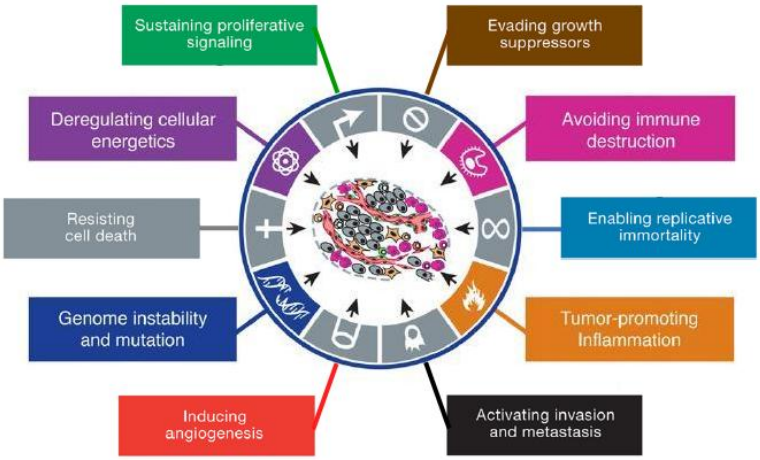


Figure 10: Representation of the eight proposed hallmarks of cancer and two of the potential cellular alterations that enable tumor cells to acquire them. (Adapted from [81, 82]

One of the two possibly enabling characteristics of these hallmarks of cancer are genome instability, which promotes random mutations and alterations in the tumors. The second factor fostering these alterations is the inflammation of pre-malignant and malignant tissues triggered by the immune cells, which may be promoting cancer pathogenesis. The complexity of this system is even higher when a new variable is included: the tumor microenvironment of *normal* cells that help cancer cells to

acquire the required hallmarks.

Given the inherent complexity of the oncogenic process, more comprehensive analytical approaches are required for unveiling the specific alterations related to the disease. Fortunately, profiling techniques have significantly improved over the last decade and studies of gene expression or mutation analysis for small sets of genes have been replaced with projects that provide large-scale genome, epigenome and transcriptome information for multiple individuals. Successful applications of such integrative analyses have been able to identify new cancer-related genes such as *PIK3R1* in glioblastoma [2] or *NCOA2* in prostate cancer [198]. Only when these key cancer-driving events are identified it is feasible to understand how they interact to achieve the above mentioned hallmarks of cancer.

2

ELEMENTS OF CANCER ARCHITECTURE

In chapter 1, cancer has been introduced as a complex disease. It has been shown that during the carcinogenic process, cells accumulate a large number of genetic alterations. These changes in the DNA sequence of tumor cells rapidly spread to the RNA and protein levels. Furthermore, alterations in cellular epigenetic mechanisms have also been shown to play an essential role in cancer development. Both type of alterations interact to alter cellular transcriptional programs and promote abnormal cell functionality, which is the ultimate responsible for tumorigenesis. In this chapter, some of the most relevant types of alterations are detailed, both at the genetic and epigenetic levels. Jointly, these cellular changes comprise what could be seen as the basic elements of the cancer architecture.

2.1 GENETIC ALTERATIONS

DNA can be considered as the *molecule of life*. It can be found in almost every *living* organism known up to date, excluding some viruses and prions [152]. Its role is to encode all the required information to create essential molecules for the correct functioning of the cells, such as ribonucleic acids (RNA) and proteins. Human DNA contains about 3 billion base pairs, structured in 22 somatic chromosomes and one sexual chromosome, with almost every cell having two copies of each set (i. e. diploid cells). During decades it was widely believed that most part of our DNA did not have any relevant purpose, what was colloquially named as *junk DNA* [114]. Nonetheless, with the completion of the Human Genome Project (HGP) and the development of large-scale laboratory techniques (e. g. microarrays, ultra-high throughput sequencing (UHTS)), scientific efforts such as the Encyclopedia of DNA Elements (ENCODE) consortium have been able to demonstrate that most part of the human genome is full of potentially functional elements that are eventually transcribed. However, the role of many of these newly identified entities, which may be involved in cancer development, still remains to be elucidated [23].

Evolution has provided human cells with efficient and sophisticated mechanisms to prevent DNA damage. The cellular DNA repair machinery detects lesions, triggers signals that warn about their presence and promotes the activation of mechanisms that attempt to repair the alterations [84]. Furthermore, if repair mechanisms fail (e. g. because

damage is too high to be fixed), cells are still in possession of the ability to arrest their proliferative cycle or even to activate signalling pathways that will drive themselves to a controlled death. Failure of these control mechanisms could induce a higher number of DNA alterations and could therefore have a deep influence in most molecular mechanisms that regulate essential cellular processes. This commonly occurs in many neoplasms [121]. DNA alterations occurring along the tumorigenic process can be classified into large or small-scale, depending on the size of the affected DNA region, and their effect on the cell phenotype is variable.

2.1.1 *Large-scale DNA variation*

Large cytogenetic aberrations are usually found in cancer cells. These alterations, which may affect regions as large as a complete chromosome, were the first ones to be detected, since they could be seen using standard karyotypic cytogenetic techniques. In fact, between late 19th and early 20th centuries, the first hypotheses about the role of somatic genetic alterations in the development of cancer were postulated [144]. Currently, all these alterations are classified and stored in a publicly available repository[1]. Their characterization is useful for a better understanding of the mechanisms of tumor development, which in turn may ultimately help to design better and more specific therapies [53]. Large chromosomal aberrations can be divided into imbalances and rearrangements. The former imply a change (i. e. gain or loss) in the amount of genetic material, while the latter consist on reallocations of genomic segments.

2.1.1.1 *Chromosome imbalances*

Chromosome imbalances (Fig. 11) can affect regions ranging from millions (e. g. complete chromosomes) to thousands of bases long (e. g. intragenic alterations). Imbalances typically affect a large number of genes, and tumors usually have many of these abnormalities, making it harder to detect which regions are more likely to be involved in the disease. Therefore, many of these alterations still have unknown functional implications [12]. However, as it will be exposed in chapter 3, this complexity can be overcome with the integration of genome-wide analysis of DNA dosage, gene expression levels, and functional genomic techniques.

Genomic gains usually contribute to cancer development by the activation of genes located in the amplified segments [125]. Some of these genes encode proteins that can be specifically targeted by new anticancer agents. In breast cancer, about a third of the cases carry an amplification

1 National Cancer Institute's Cancer Genome Anatomy Project. http://cgap.nci.nih.gov/Chromosomes. Accessed September 2011.

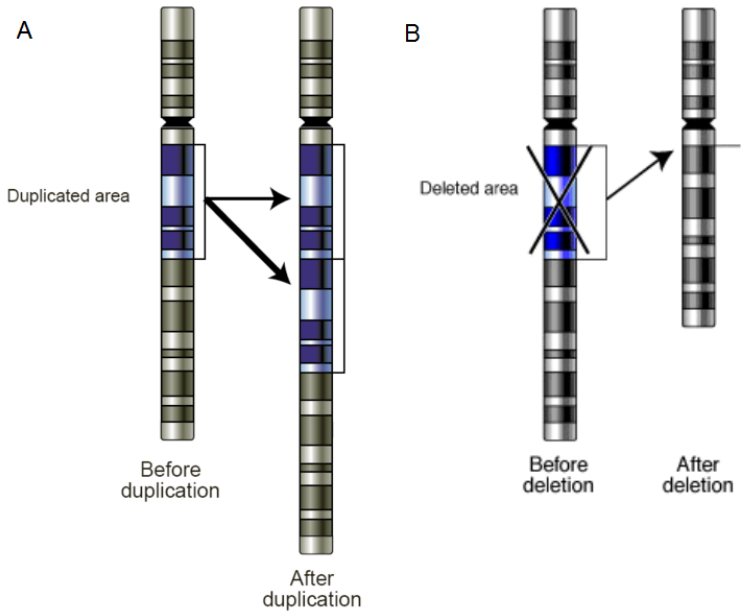


Figure 11: Schematic representation of chromosome imbalances: amplifications (A) and deletions (B).

of the 17q21.1 locus, containing the gene *ERBB2*. Subjects who overexpress *ERBB2* are likely to show a good response to treatment with the monoclonal antibody trastuzumab, in combination with chemotherapy, which displays higher survival rates both in the adjuvant and metastatic settings [92].

The contribution of genomic losses to cancer development is usually through the reduction of the function of specific genes found in the affected chromosomal regions. Important tumor-suppressor genes are affected by chromosomal deletions, such as *RB1* [193], *TP53* [180, 191], *APC* [75] or *PTEN* [56]. Nonetheless, for other loci the critical genes have not been so clearly delimited, such as the chromosome region 1p related to neuroblastoma [145] and region 3p in lung cancer [230]. Deletions not only may affect standard tumor-suppressors, but also haploinsufficient genes [15, 56] and other important regulatory elements, such as non-coding RNAs [74]. Regarding their potential as direct treatment targets, up to date scientists have not been able to develop efficient drugs to compensate the loss of genomic material. However, a deeper knowledge of the functional implications of these abnormalities may help to detect indirect targets that could be clinically relevant [146]. This is the case for genes *AKT1* and *MTOR* when *PTEN* is lost [175].

2.1.1.2 *Loss of heterozygosity (LOH)*

Mutations in tumor suppressor genes are generally recessive. That is, a mutation in one allele is not sufficient to alter the cell functionality, as long as the other allele is still working properly. Consequently, the loss of the second allele is one of the essential events in cancer development. Suppression of this second wild-type allele often involves loss-of-heterozygosity events, which can be considered a specific case chromosome imbalance. Recurrently observed LOH in a certain chromosomal region for a specific type of tumor can serve as an indication of the presence of a tumor suppressor gene in that segment. Although a large number of regions display recurrent LOH for different tumors, only a few have been found to contain tumor suppressor genes. This fact may indicate that there are still a huge number of this type of genes to discover, or may be related to intratumor heterogeneity, contamination by normal cells or other artifacts [204].

2.1.1.3 *Chromosome rearrangements*

Chromosome rearrangements occur when one or more fragments of DNA move to another genomic position, with no overall variation in the amount of genetic material contained in the cell. Rearrangements commonly found in cancer cells are inversions, insertions and translocations (Fig. 12). It has been argued that some of these alterations are triggers of the tumor initiation [163]. These aberrations are usually caused by double strand breaks of the DNA molecules [208]. Although rearrangements have been usually associated with hematological tumors, they have also been recently linked to epithelial cancers as well [130, 184]. Notably, chromosomal rearrangements are not always found to be specific of a single tumor type [135].

In terms of functional consequences, these aberrations typically result in the formation of a new fusion gene product with new or altered activity. A classic example is the Philadelphia chromosome, in which part of the *BCR* gene on chromosome 22 is fused with gene *ABL1*, located in locus 9q34.1. This translocation is present in all cases of chronic myeloid leukemia, among other hematological pathologies [144]. Chromosomal rearrangements have also been linked to deregulation of non-coding RNAs [32].

2.1.2 *Small-scale DNA variation*

Small-scale DNA variation comprises all those alterations in the DNA sequence that affect from one to a few thousand nucleotides. Among them, changes in just one base pair, usually called mutations, are one of

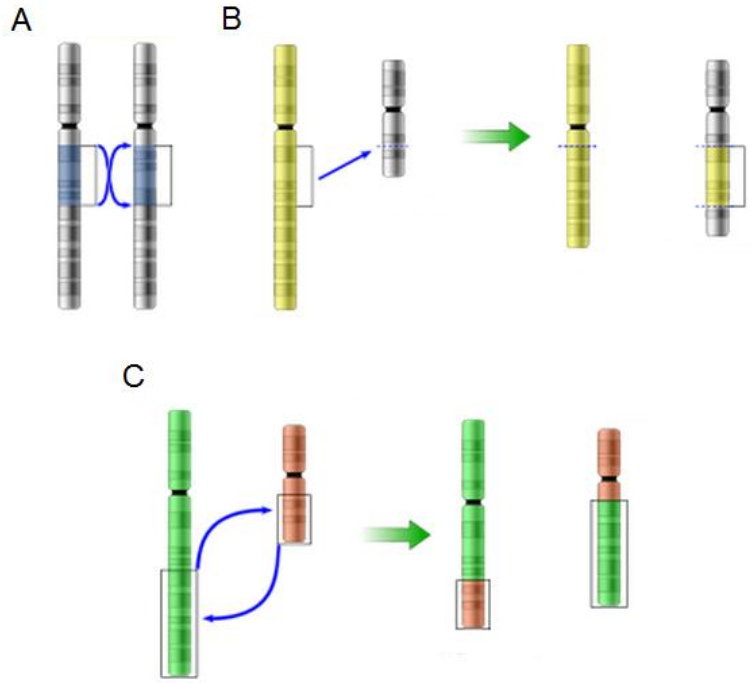


Figure 12: Schematic representation of chromosome rearrangements: inversions (A), insertions (B) and translocations (C).

the most commonly found (Fig. 13). Due to the large size of our genome, these type of variation has not been thoroughly studied until large-scale genotyping techniques have been made available to the scientific community. Variable number tandem repeats (VNTR), which are repetitions of short DNA sequences that differ in number among individuals, are also a source DNA variation at a small scale.

2.1.2.1 *Single-nucleotide mutations*

A mutation is defined as any variation in a DNA sequence, compared to a standard consensus. This implies the existence of a normal allele that is highly prevalent in the population, and the mutation turns it into a rare variant. Evolution is a process along which point mutations are accumulated in our genomes. Thus, after millions of years, DNA mutations have become one of the most common forms of genetic variation across human beings, only exceeded by copy number variants (CNV) [165]. Mutations in protein-coding regions that promote a change in the aminoacid sequence are called missense mutations, while changes that do no alter the aminoacid sequence of the protein are called nonsense mutations. Regarding their effect on individuals, only those changes in the DNA sequence that might have a pathological effect are usually called mutations, while changes with a neutral or unknown effect are usually called

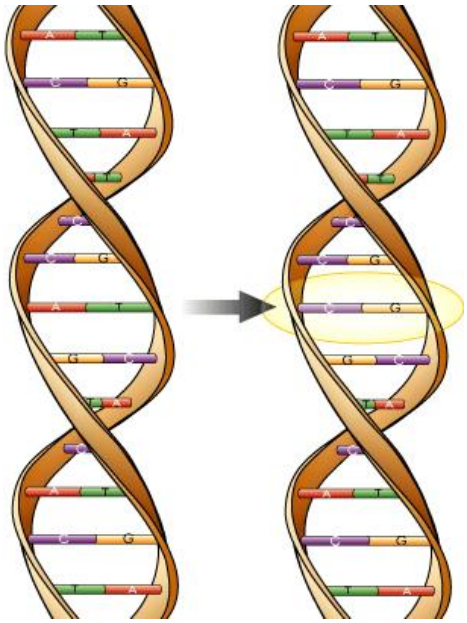


Figure 13: Graphical representation of a single-nucleotide mutation.

variants.

Along their evolving process, due to their defective DNA control and reparation mechanisms, DNA of tumor cells can accumulate a large number of somatic point mutations. While most of them can be categorized as *passenger* (i. e. not related to the carcinogenic process), others may have a deep influence and contribute to the tumor development [188]. The latter ones are usually called *cancer-driving mutations*. Consequently, these mutations must be unveiled and studied to get a better insight into the molecular basis of the neoplasms. This is the basis of the work presented in chapter 5.

2.1.2.2 *Copy number variants*

CNVs are genomic regions for which differences in number of copies of specific DNA small segments are observed across individuals [165]. As it was stated in 2.1.2.1, CNVs are the most frequent and complex form of genetic diversity. This type of variation alters the paradigm of 'diploid genomes' that had been historically accepted. Regarding their size, they are typically defined to be larger than 1,000 bases, and usually involve DNA segments at a submicroscopic scale. These regions may be gained (i. e. more than two copies) or lost (i. e. less than two copies). In much the same way as SNPs, large-scale technology has enabled many groups to determine the association of CNVs with various diseases, including cancer among them.

CNVs have an indirect influence in cancer susceptibility, by varying the gene dosage of cancer-related genes. Some associations in this direction have been found so far (Fig. 14). Interestingly, cancer cells are usually found to display somatically-acquired copy number alterations [8]. Genomic DNA copy number alterations are currently considered a prominent mechanism of gene disruption that contributes to tumor development [99]. Segmental amplification may lead to an increase in gene and protein expression of oncogenes, while deletions may lead to haploinsufficiency or the loss of expression of tumor suppressor genes.

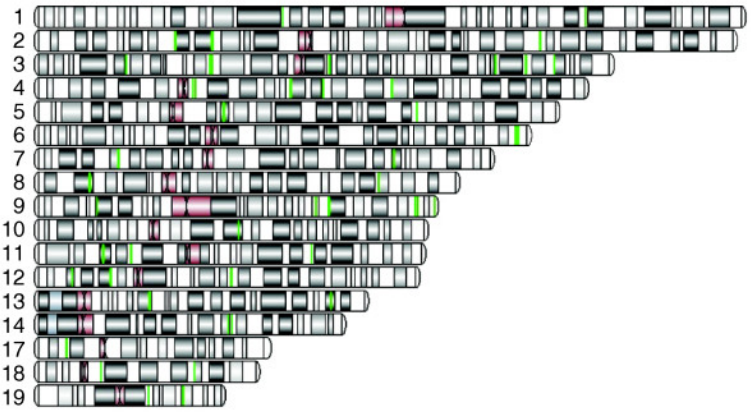


Figure 14: Map of cancer CNVs. Green regions are CNVs that contain a cancer related gene, while centromeric regions are displayed in red. (Extracted from [187].)

2.2 EPIGENETIC ALTERATIONS

Epigenetics deals with mitotically or meiotically inheritable mechanisms related to gene expression regulation, without involving any modification at the genetic sequence level [18]. These mechanisms play a relevant role in essential processes, such as early embryogenesis [115] or regulation of cell fate by promoting tissue-specific gene expression patterns [231]. Moreover, epigenetics is also significantly responsible for biological diversity among individuals, and it explains phenotypically differences among genetically identical subjects [69]. Main epigenetic events occurring in human cells are DNA cytosine methylation at cytosine-phosphate-guanine (CpG) dinucleotide sites, histone modifications and expression of small RNAs, which do not code for any protein product.

Tumors can have genome-wide epigenetic alterations. These comprise promoter CpG island methylation, generalized hypomethylation, loss of imprinting, histone modifications, chromatin looping or deregulation of small RNAs expression, among others. The combination of these events

with other genomic aberrations gives tumor cells a selective advantage, based on inhibition of apoptosis, uncontrolled proliferation and potential of migration (Fig. 15). Therefore, in the following subsections the role of these different types of epigenetic alterations in cancer will be briefly described. Focus will be given to methylation events and transcription of microRNAs (miRNA), since they have been the more comprehensively analyzed at a large scale level in the last years, due to the increased complexity of histone modification analyses in clinical samples [138].

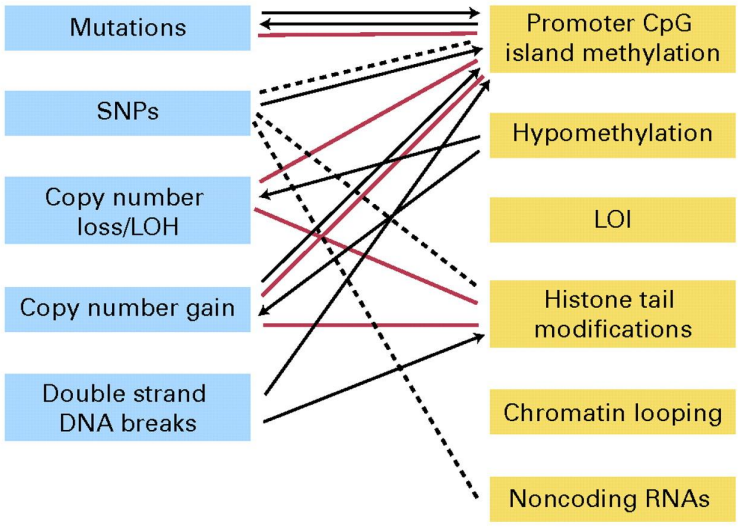


Figure 15: Genetic (left) and epigenetic (right) alterations in cancer interact in multiple combinations and equivalently alter important signalling pathways. Black arrows indicate causative interactions. Red lines correspond to events that can collaborate to regulate gene expression of a specific locus. Dashed lines indicate associations that have been hypothesized/observed but for which a causal relationship has not yet been established. (Adapted from [212].)

2.2.1 *DNA methylation*

Tumor initiation is produced by a generalized deregulation of essential cellular mechanisms, such as proliferation, apoptosis or migration. This deregulation was initially thought to be mainly motivated by genetic alterations, such as mutations, large chromosome aberrations, translocations or genomic rearrangements, among others [219]. However, over the last decade DNA methylation has been identified as an alternative mechanism of transcriptional regulation of cancer genes, in much the same way as DNA alterations do [86]. Global DNA hypomethylation was the first epigenetic alteration detected in tumors [106]. It has been

demonstrated that ubiquitous hypomethylation of DNA promotes genomic instability and confers an increased risk of developing diverse genomic alterations [206]. Thus, generalized hypomethylation has been associated with larger number of genetic alterations in colorectal cancer [97] or glioblastoma [30], among other types of neoplasia. Loss of usually-methylated CpG sites can also lead to enhanced expression levels of specific genes, or even gene-embedded miRNAs, that may trigger the activation of oncogenic processes [51, 58].

The effect of methylation alterations in cancer is not only mediated by the inhibition of tumor-suppressor genes, but also by downregulating oncogene inhibitors. *MLH1* and *MGMT* genes are well-described examples of directly methylated genes involved in CRC development [49, 100]. Besides their direct impact in the tumorigenic process, epigenetic alterations in cancer do not work independently. As it is depicted in Figure 15, strong interactions between genetic and epigenetic events have been found in CRC and other type of tumors [212], which pinpoints the complexity of the neoplastic process and reinforces the need of comprehensive molecular profiling studies to fully characterize the mechanistic processes underlying tumor cells.

Aberrant methylation of CpG sites has been reported to be an early event in cancer development that promotes the accumulation of further genetic and epigenetic alterations [139]. In much the same manner as genetic aberrations, epigenetic patterns greatly differ among different types of tumors, and even among the same tumor type in different subjects [83]. A paradigmatic, although controversial as well due to lack of agreement among researchers, example of a methylation-based tumor is the CpG island methylator phenotype (CIMP) variant of CRC. This type of CRC tumor displays pervasive gene promoter methylation, and is strongly associated to microsatellite instability or defects in DNA mismatch repair genes, as *MLH1*. Interestingly, these methylation-based tumors may have distinctive clinical characteristics, which could be beneficial for future for personalized cancer therapies [95].

2.2.2 *Small RNAs*

MiRNAs are small RNA molecules ranging in size from 19 to 24 nucleotides that do not code for any protein product. They are classified as a subfamily within the category small RNA molecules. Discovered only a decade ago [105], miRNAs have been found to play a determinant role in cell biology, since they potentially target up to one-third of human coding genes [71]. These molecules post-transcriptionally downregulate gene expression by binding to complementary RNA sequences located at the

3′ untranslated region (UTR) of specific messenger RNAs (mRNA) [156]. Therefore, since miRNAs regulate the expression of their targets, high or low levels of these molecules are expected to result in underexpression or overexpression, respectively, of the protein product of the target mRNA.

Human miRNA	Deregulation in cancer
let-7 family (various)	Down-regulated in lung, breast, gastric, ovary, prostate and colon cancer
	Overexpression in AML
miR-10b (2q31.1, intergenic)	Down-regulated in breast cancer. Overexpressed in metastatic breast cancer
miR-15a, miR-16-1 cluster (13q14.3, intron 4 non-coding RNA *DLEU2*).	Down-regulated in CLL, DLBCL, multiple myeloma, pituitary adenoma, prostate and pancreatic cancer
	Up-regulated in nasopharyngeal carcinoma
miR-17, miR-18a, miR-19a, miR-20a, miR-19b-1, miR-17-92 cluster (13q31.3, intron 3 C13orf25)	LOH at miR-17-92 locus in melanoma, ovarian and breast cancer
	Overexpression in lung and colon cancer, lymphoma, multiple myeloma, medulloblastoma
miR-26a (3p22.2)	Down-regulation in hepatocellular carcinomas
	Up-regulation in breast cancer
miR-106b-93-25 cluster (7q22.1)	Overexpression in gastric, colon and prostate cancer, neuroblastoma and multiple myeloma
miR-21 (17q23.1, 3′UTR TMEM49)	Overexpression in glioblastoma, breast, lung, prostate, colon, stomach, esophageal, and cervical cancer, uterine leiomyosarcoma, DLBCL, head and neck cancer
miR-29 family (various)	Down-regulation in CLL, colon, breast, and lung cancer and cholangiocarcinomas
	Up-regulation in breast cancer
miR-34 family (1p36.23, 11q23.1, intergenic	Down-regulated in pancreatic cancer and Burkitt's lymphoma.
	Hypermethylation of miR-34b, c in colon cancer
miR-101 (1p31.3, 9p24.1)	Down-regulation in prostate cancer, hepatocellular carcinoma, and bladder cancer
miR-122a (18q21.31 intergenic)	Down-regulation in hepatocellular carcinoma
miR-124a family (various)	Hypermethylation in colon, breast, gastric and lung cancer, leukemia and lymphoma
miR-125a, miR-125b (various)	Down-regulation in glioblastoma, breast, prostate and ovarian cancer
	Up-regulation in myelodisplastic syndrome and AML
miR-127 (14q32, RTL1 exon)	Hypermethylation in tumor cell lines
miR-143, miR-145 cluster (intergenic 5q32)	Down-regulated in colon adenoma/carcinoma, in breast, lung, and cervical cancer, in B-cell malignancies
miR-155 (21q21.3, exon 3 ncRNA BIC)	Overexpressed in pediatric Burkitt's lymphoma, Hodgkin's lymphoma, primary mediastinal lymphoma, DLBCL, breast, lung, colon, pancreatic cancer
miR-181 family (various)	Overexpressed in breast, pancreas, and prostate cancer
miR-221, miR-222 cluster (Xp11.3, intergenic)	Overexpressed in CLL, thyroid papillary carcinoma, glioblastoma. Down-regulated in AML
miR-200 family (various)	Down-regulated in clear-cell carcinoma, metastatic breast cancer
miR-205 (1q32.2)	Overexpression in NSCLC
	Down-regulated in prostate cancer
miR-372, miR-373 cluster (19q13.41, intergenic)	Overexpression in testicular cancer

Figure 16: List of miRNAs currently known be involved in cancer. Blue lines correspond to oncogenic associations and orange lines correspond to antitumorigenic associations (Adapted from [129].)

MiRNAs were initially found to be involved in cancer using animal models such as *Caenorhabditis elegans* and *Drosophila melanogaster*. Knockout organisms of miRNAs lin-4 or let-7 in *C. elegans* produced altered differentiation behavior [167], while overexpression of Bantam miRNA in flies promoted abnormal cell growth and inhibition of apoptosis [89]. Subsequent studies in mammals confirmed the relationship between miRNAs and tumor development, since it was observed that Dicer knockout murine models caused alterations in the miRNA transcriptional machinery that led to aberrant development and cell differentiation [19]. Regarding humans, recent studies have reported the involvement of both genetic and epigenetic mechanisms in miRNA deregulation that can potentially lead to cancer development [216]. Chromosomal aberrations can lead to the deletion, amplification, or translocation of miRNAs [33, 32]. Moreover, about half of all currently-annotated human miRNA genes have been postulated to be located at cancer-related hot-spots of breakage and rearrangement of the genome [34]. In this sense, tumor-supressor miRNAs miR-15 and miR-16 display extremely low expression levels in about 70% of patients with chronic lymphocytic leukemia (CLL) because of deletions or mutations at the 13q13.4 loci where they are situated [33].

These miRNAs are known to promote apoptosis by targeting pivotal tumor-suppressor gene *BCL2* [173]. Up to date, other miRNAs have been found to be transcriptionally altered in different types of tumors, as depicted in Figure 16.

3

INTEGRATION OF CANCER DATA

In chapter 2 many types of alterations at different molecular levels occurring along the carcinogenic process have been described. In the past, researchers could only have a limited view of all these changes, since it was not technically feasible to obtain a complete overview of a cell's status. These circumstances began to change when DNA microarrays came to the fore in the mid-90s [177]. Since then, large-scale experiments have become a routine in biomedical research. Consequently, the generation of multiple vast datasets, combined with the complete sequence of the human genome [1, 110], has provided researchers with the unquestionably largest amount of biomedical data ever [109].

Human cells are complex biological systems, based on networks of highly-interacting molecular elements. Therefore, the integration of data coming from multiple molecular sources is essential to understand the mechanisms that drive the transformation of a phenotypically normal cell into a highly proliferative, disseminating tumor. However, it must be kept in mind that large-scale data analysis and integration has raised serious challenges in the fields of bioinformatics and biostatistics, which must be carefully addressed to ensure the reliability of the obtained results [207].

This chapter initially focuses on how microarrays and next-generation sequencing technologies have shifted the paradigm of biomedical research from single-gene experiments to genome-wide approaches. Secondly, the need for integrative approaches in cancer studies will be discussed. Finally, some of the most well-known international efforts for cancer data integration will also be exposed, along with the most relevant findings of cancer integrative analysis up to date.

3.1 A NEW STEP IN BIOMEDICAL RESEARCH: THE MICROARRAY ERA

Historically, molecular biology assays used to be extremely expensive and time consuming. Performing experiments at a scale of thousands or even hundreds of measurements was extremely unfeasible, since everything had to be done 'one gene at a time' [164]. Thus, it was not possible to obtain a complete snapshot of the molecular state of cancer cells at a given moment. Fortunately, this situation reverted when first DNA microarrays appeared. Based on the property of DNA complementary strands hybridization, these arrays were able to perform measures for

thousands or even millions of loci in just a single experiment (Fig. 17).

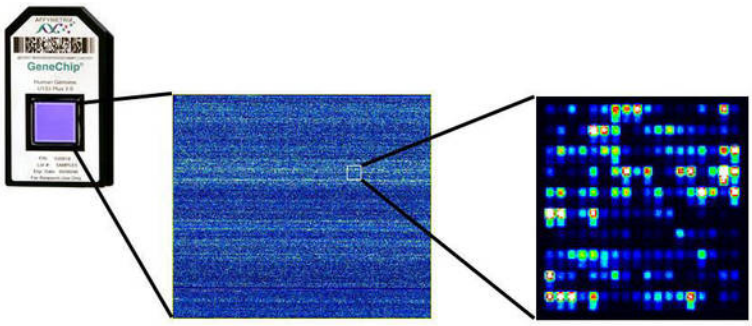


Figure 17: Schematic view of a DNA microarray. DNA microarrays are usually encapsulated in a cartridge to prevent contamination and standardize experimental conditions. Once they are scanned, an image similar to the one at the center of the figure is typically obtained. Bright spots are those where there has been specific hybridization. For gene expression arrays, the intensity of light returned by the fluorescent molecules is used as a proxy for the amount of RNA present in the cell for that gene.

First microarray projects related to cancer were used to compare the expression level of thousands of genes between tumor and healthy control cells [54, 101, 227]. The technique was still evolving, but results seemed promising. In 1999, Golub et al. published the first major study of cancer using microarrays. They were able to obtain a reduced set of genes whose expression levels could classify acute myeloid leukemia (AML) and acute lymphoblastic leukemia (ALL) samples [78]. This work meant a great improvement for the diagnosis of both diseases, since they are often hard to differentiate by standard pathological parameters. Initially, expression arrays were designed to detect gene transcripts without taking into account splice variants, since all the probes in the array were located at the 3′ end of the transcripts. More advanced and complex array platforms are now able to detect differential expression at the exon level and successfully determinate alternate splicing events related to cancer [9, 196].

Although DNA microarrays were first designed for gene expression experiments, this large-scale technology was soon adapted for other types of measurements. For DNA alteration analyses, conventional comparative genomic hybridization (CGH) started being replaced by CGH arrays (aCGH). While original CGH gave a resolution from 1 to 10 Mbp, first aCGH platforms reduced that figure about two orders of magnitude [158, 159]. These arrays allowed a more precise detection of LOH and copy number regions, although they still suffered from high levels of noise that precluded the detection of small-magnitude changes [31].

CGH-arrays later evolved into higher-quality currently used SNP arrays, which are able to genotype up to millions of SNPs for a given individual. Some of them even also contain large sets of non-polymorphic probes to better assess copy number alterations. SNP arrays have been the basis of most cancer GWAS published up to date [96]. For LOH and copy number analysis, the resolution of these arrays can be as much as a few hundred base pairs, allowing the detection of micro-scale alterations [10].

Many transcription factors (TF) act as oncogenes or tumor suppressor genes, binding to DNA to repress or activate the expression of other transcripts. With DNA arrays, it is also possible to obtain a detailed knowledge of a genome's transcriptional regulatory program at a large-scale. Using chromatin immunoprecipitation, DNA sequences bound to a specific protein can be isolated. This set of DNA-protein interacting sequences are then labeled and hybridized into tiling arrays, which consist of overlapping probes designed to densely represent a genomic region of interest, or even the whole genome. Thus, it is possible to obtain a complete map of all the specific binding sites of a oncogenic transcription factor. This technique is known as chromatin immunoprecipitation-on-chip (ChIP-on-chip), and has been successfully applied to study the transcriptional regulatory programs involved in cancer development [85, 116, 126].

Cancer epigenomics, the study of cancer not induced by alterations in the DNA sequence, can also be analyzed at the whole-genome level. Analysis of genome-wide patterns of methylation of cytosines at CpG dinucleotides [44, 223] and large-scale profiling of miRNA expression using arrays [183] have been successfully applied to cancer research with promising results.

3.2 NEXT STEP IN LARGE-SCALE TECHNOLOGY: NEXT GENERATION SEQUENCING

In the very recent years, UHTS has represented a remarkable breakthrough in large-scale technologies. Compared to microarrays, they confer the advantage that they are not any more restricted to analyze a previously-selected set of markers. Moreover, they can provide not only quantitative information (i. e. number of copies of a transcript), but also information about the sequence of the transcript itself.

With next-generation sequencing machines, it is now possible to sequence a complete human genome in a few hours at a reasonable cost. Remarkably, the price of sequencing per base has dramatically decreased during the past years, and now it is possible to sequence a complete with

a reasonable coverage for less than 10,000 USD (Fig. 18).

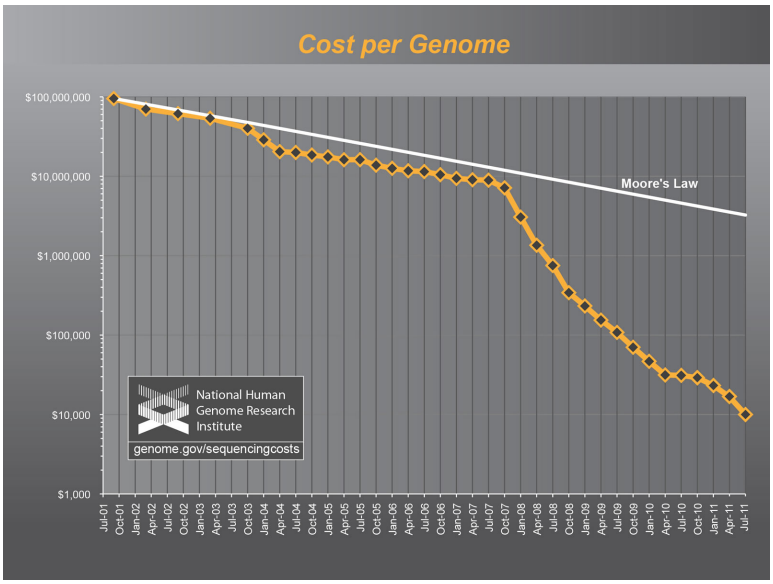


Figure 18: Cost of sequencing a complete genome has dramatically decreased in only 10 years. Even computers Moore's Law, which points to an outstanding level of improvement, falls short behind the trend of the reduction of the cost per genome. The price of sequencing a genome reached the goal of the 10,000 USD last July. (Source: NHGRI — http://www.genome.gov/sequencingcosts/. Accessed September 2011.)

Systematic sequencing of the cancer genome will provide invaluable information about mutations and other types of DNA alterations involved in tumor development [13, 35, 131]. Furthermore, large-scale sequencing studies could also reveal information about variability in mutational frequencies and patterns between tumor types and between cancer patients with apparently similar disease [67]. Therefore, genome-wide DNA sequence analysis not only leads to the discovery of cancer-related genes, but also demonstrates structural differences in mutagenic processes between tumor types and individuals. Successful application of these type of studies have been performed by the Cancer Genome Project (CGP), which is an international consortium aimed to identify somatically acquired sequence variants and mutations to uncover central genes involved in the development of human cancers using high throughput sequencing techniques [80].

Large-scale sequencing technology in cancer not only has been successfully applied to DNA sequencing projects, but also to whole-transcriptome analysis (RNA-Seq) [162], small RNA sequencing (small RNA-seq) [210], chromatin immunoprecipitation sequencing (ChIP-Seq) [87], or whole-

genome methylation analysis (MeDIP-seq) [172]. Although some these sequencing applications are still evolving and need further refinement, this wide range of sequencing techniques will unquestionably foster the development of large cancer integrative projects in the very near future.

3.3 THE NEED FOR INTEGRATIVE ANALYTICAL APPROACHES IN CANCER RESEARCH

As it was stated at the beginning of this chapter, large-scale cancer molecular profiling has already provided an enormous amount of information about DNA (e. g. mutations, copy number alterations, structural aberrations), RNA (e. g. gene expression changes, transcription regulation alteration), protein-protein interactions and epigenetic alterations (e. g. methylation changes, histone modifications, miRNA expression changes) related to the pathology. Therefore, integrative analyses are remarkably contributing to obtain a better comprehension of cancer molecular basis. Moreover, they are helping us to design more accurate disease classification mechanisms, based on molecular assessments rather than pathological criteria.

Efforts of large consortiums, such as the CGP, could lead to the obtention of an almost-complete set of occurring mutations for different types of tumors in the short term [80]. Potential limitations in the sample size of the studies, as well as the non-random distribution of passenger events [21], could hinder the distinction of true driver alterations from those which are not really involved in the tumorigenic process. However, the combination of mutational events with the analysis of copy number alterations could help to detect significantly cancer-related events [195].

Once a large proportion of the alterations significantly involved in a specific tumor are detected, if will be possible to uncover how these alterations interact with each other to promote cell proliferation and migration. This is essential for the study of cancer, because it is widely accepted that specific alterations are not usually likely to be the only culprit for tumor initiation and development. Contrarily, interactions between critical genes and their associated signalling pathways are known play a key role in the progression of the disease [24].

As it was emphasized in the introduction of this chapter, these type of integrating analyses, which aim to merge different sources of data such as gene or miRNA expression, copy number alterations, mutational changes, structural chromosomal aberrations or protein-protein interactions, raise some statistical concerns that must be properly addressed to obtain trustworthy results. One issue of special relevance is the need for

a proper correction for multiple testing. As an example, given the large number of mutations observed in a small set of breast and colorectal tumors [188], it could be expected that the number of potential functional interactions involved in cancer development could be almost unfeasible to be explored [128]. A potential workaround for this problem could be to increase the sample size of the studies to achieve enough statistical power to detect significant interactions. However, since the number of required individuals could be so large, it may be necessary to merge data from different tumor types. In this way, the efforts of the International Cancer Genome Consortium (ICGC) which aims to completely characterize 25,000 tumors at a large-scale level, will be essential [93].

The ultimate identification of cancer driver alterations will make it possible to unveil how they collaborate with each other in the context of the tumor cell to finally attain the central hallmarks of cancer [81, 82]. However, even with such vast amounts of data, this will not be a straightforward task. It should always be kept in mind that tumors are dynamic entities evolving over time due to stressing and changing conditions in their surrounding environment. Therefore, only by placing alterations in signalling pathways, understanding how these pathways functionally support each other, uncovering significantly correlated events across a very large number of samples, and conceptualizing these changes in terms of the hallmarks of cancer we may be able to make sense of the complexity and individual diversity of the altered cancer genome.

Over the last decade there have been remarkable examples of integrative data analysis applied to cancer research. In 2003, Lamb et al. combined gene expression data from hundreds of human tumors with gene promoter analysis to reveal a mechanism of action for the *CCND1* oncogene in a wide range of tumors [107]. Other studies have also integrated large-scale gene expression datasets of multiple tumor types to detect cancer modules, that is, sets of genes that act in concert to carry out a specific tumorigenic function [181]. Moreover, recently developed methods of integration of gene expression data from multiple tumor types have enabled to detect novel gene fusions related to prostate cancer development [200, 202].

The integration of DNA and RNA data have also yielded promising results detecting new genes involved in tumor development. In 2005, Garraway et al. integrated gene expression, LOH and CNV data to uncover *MITF* as the target of a novel melanoma amplification [73]. More recently, another integrative computational framework that integrates chromosomal copy number and gene expression data has been able to detect aberrations that promote cancer progression (Fig. 19). More specifically, it

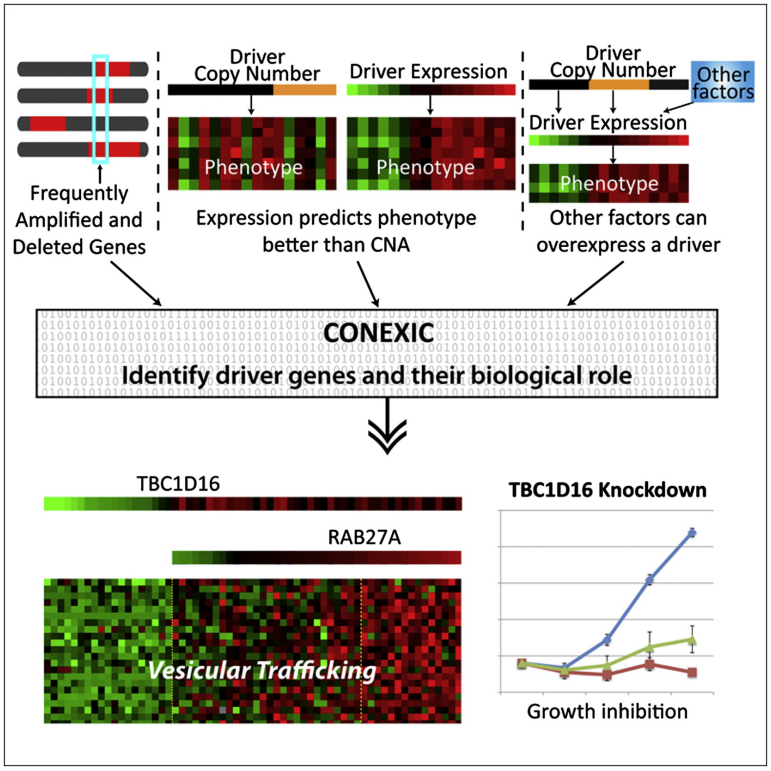


Figure 19: Integrative framework that combines large-scale CNV and gene expression data to uncover new cancer drivers. (Extracted from [5].)

has identified *TBC1D16* and *RAB27A* as two new candidate driver genes in melanoma [5]. Network integrative approaches have also been applied in this context, where there are remarkable examples of integration of gene expression, ChIP-on-chip and protein interaction data to identify master regulator genes involved in the process of the tumor development [38, 112, 118].

Integrative analyses have also proven its usefulness in the field of drug discovery and pharmacogenomics. A remarkable example is The Connectivity Map (CMAP) project, aimed to discover functional connections among diseases, genetic perturbations, and drug action by the analytical integration of large gene expression datasets [108].

3.4 LARGE CANCER DATA INTEGRATION PROJECTS

The field of cancer integrative profiling has experienced a growing interest among the research community over the last 15 years (Fig. 20). First successful attempts of applying this integrating methodology to cancer research came up after initial large-scale gene expression and aCGH platforms became routinely used after year 2000 [185]. Concurrently, pioneer

large efforts to massively integrate cancer data started being developed, such as the National Cancer Institute's (NCI) Cancer Molecular Analysis Project (CMAP) [28]. Although those projects were remarkably ambitious and innovative at that time, their conclusions were limited by the sample size of their studies and the technology used for large-scale profiling, which was under intense development. It was not until 2005 that these type of methodology became widely applied in cancer research. As an example, different integrative analyses applied to prostate cancer progression data were able to identify 8p21.2, 11q13.1 and 10q23 regions to be involved in prostate cancer development, and also were able to classify prostate tumors according to their LOH pattern (reviewed in [201]). In 2010, a more comprehensive integrative analysis of prostate cancer, based on the integration of copy number, mRNA expression and focused exon sequencing data, was able to identify *NCOA2* as a new potential oncogene altered in 11% of the tumors, and they also detected a small and previously unknown region in 3p14 that implicated *FOXP1*, *RYBP* and *SHQ1* as potential cooperative tumor suppressors [198]. Other type of tumors were also studied by primarily integrating mRNA expression and copy number data: oesophageal adenocarcinoma [77], melanoma [5, 17], CRC [36, 140, 166], breast [137] and neurofibromatosis [133]. Although all these previously mentioned projects have been successful in identifying new mechanisms involved in tumorigenesis, in the very following years the accuracy and specificity of these type of analyses is expected to improve remarkably, specially due to better large-scale profiling techniques and to the routine integration of more data sources, such as miRNA expression or methylation profiles.

The prominent role of large-scale data integration in cancer molecular analysis has undoubtedly been reinforced by the recent creation of large research consortiums. As it was previously pointed, these joint efforts are essential to overcome potential pitfalls of integrative studies, specially the difficulty in achieving reasonably good sample sizes to gain statistical power and reliably detect cancer-related alterations. Among others, the most currently relevant established consortiums are the NCI's The Cancer Genome Atlas (TCGA)[1] and the ICGC[2], which is an international cooperative organization with participant countries across the globe.

TCGA was initiated in 2006, after the NCI proposed the investment of $1.5 billion to catalog the genomic changes involved in cancer. Although the initial main goal of the project was to comprehensively characterize brain, ovarian and lung tumors to improve the ability to diagnose, treat and prevent these pathologies, it was later expanded to characterize up

1 http://cancergenome.nih.gov/. Accessed September 2011.
2 http://www.icgc.org/. Accessed September 2011.

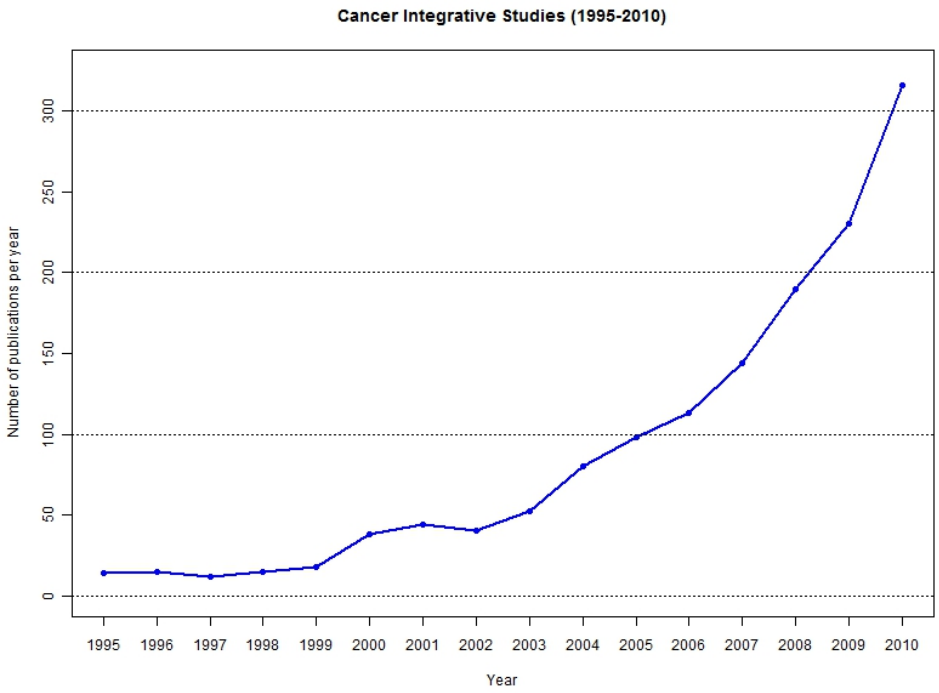


Figure 20: Number of studies citing the terms "Cancer AND (Integrative OR Integrating) AND (profiling OR analysis)" between 1995 and 2010. (Source: US National Library of Medicine, National Institutes of Health.)

to 20 different types of cancer. For each type of tumor it was planned to obtain mRNA expression data, copy number alterations, methylation profiling and miRNA expression levels. Results of the pilot project, based on the study of glioblastoma multiforme, were published two years later [2]. Promising outcomes were presented, including alterations which had never been previously related to the disease. Interestingly, one of the detected mutations in a mismatch-repair gene was found to be related to temozolomide resistance, which may help for further improvement in the clinical management of the disease.

Once TCGA data has passed internal quality control checking it is immediately made available through their data portal. The project has currently finished obtaining data for ovarian carcinoma, while the collection of the remaining type of tumors is expected to be finished soon. Although data repository is still incomplete, the project has already been able to obtain remarkable results for glioblastoma [142, 217] and ovarian cancer [3, 45].

The ICGC is even a bigger effort than the TCGA project. The goal of the project is to coordinate large-scale cancer studies for 50 different

tumors that are clinically relevant all over the world. This will lead to a systematic analysis of more than 25,000 tumor samples at the genomic, epigenomic and transcriptomic levels [93]. As for TCGA, datasets will also be made available as soon as they are obtained, with minimum restrictions. Although due to legal issues regarding sample information sharing policies TCGA and ICGC projects are not legally related, they can be currently considered as a joint collaboratively international effort. However, the ICGC project is still in its initial steps, and preliminary results are yet to arrive. These type of projects will be essential in the following years to reveal the repertoire of oncogenic mutations, uncover traces of the mutagenic influences, define clinically relevant subtypes for prognosis and therapeutic management, and enable the development of new cancer therapies.

Part II

RATIONALE

4

WORKING HYPOTHESIS AND OBJECTIVES

As a paradigmatic complex disease, cancer is based on strong interactions between a vast number of molecular and environmental factors. This inherent complexity, therefore, precludes an easy understanding of the susceptibility, emergence and development of the pathology.

Cellular alterations that contribute to the development of cancer are varied, and can be found at any molecular level, from DNA to proteins. In the past, most cancer studies focused on just one of these entities. Although useful, this kind of approaches usually only leads to a partial view of the tumor cell. Thus, in the last few years it has been clearly understood that more comprehensive analyses are required to obtain an accurate view of the cellular processes that confer to a phenotypically normal cell the ability to proliferate and invade its surrounding tissues. The advent and consolidation of large-scale technologies, such as DNA microarrays, have also been a main contributor to this new manner of tackling the biology of the disease.

The working hypothesis of this thesis is that the integration of diverse, large-scale molecular data is essential to unveil the molecular mechanisms underlying complex diseases, such as cancer.

4.1 GENERAL OBJECTIVES

The main aim of this thesis is to achieve a more complete understanding of the molecular mechanisms of cancer by means of the analytical integration of large-scale data at different molecular levels (i. e. DNA, RNA, proteins).

4.2 SPECIFIC OBJECTIVES

Each one of the three projects presented in this thesis has its own specific objectives, which are stated below.

Integrative analysis of a cancer somatic mutome

A landmark study published by Sjoblom *et al.* [188] determined the sequence of 13,023 protein-coding genes for 11 breast tumors. For breast

cancer, this work obtained a list of ~700 genes harboring somatic mutations. Although some of these genes were previously known, many of them had never been linked before to this pathology. This new set of genes required further study and characterization.
Aim:

1. To characterize genes harboring somatic mutations in breast cancer at the DNA, RNA and protein interactome levels to detect those genes more likely to be associated with the development of the pathology.

MYC germline expression modeling and cancer susceptibility

Many association studies have recurrently found loci in the 8q24 region that confer susceptibility to different epithelial tumors, being colorectal, prostate and breast among the most relevant. Interestingly, the region is known to be a gene desert, with only *MYC* located a few hundred kilobases away.
Aim:

1. To elucidate the potential mechanism of action of the 8q24 region risk variants and their putative role as *MYC* regulators by the integration of genetic and expression data of prostate samples.

Biological convergence of cancer signatures

Many studies have developed gene signatures that can correctly predict different clinical features (e. g. prognosis, response to treatment, probability of developing metastasis, etc.). Surprisingly enough, gene overlap across signatures is poor, even for the same type of tumor. This raises concerns about their biological and clinical implications.
Aim:

1. To unveil common underlying biological properties across different cancer signatures by integrating large-scale genome, transcriptome and protein interaction data.

Part III

RESULTS

PUBLICATIONS

All three publications presented in this thesis have already been published in international peer-reviewed journals. Each one of the articles is preceded by a summary and a brief description of their main findings.

The references are the following:

- Hernández P*, **Solé X***, Valls J*, Moreno V, Capellá G, Urruticoechea A and Pujana MA. Integrative analysis of a cancer somatic mutome. Mol Cancer, 6:13, 2007
  Journal Impact Factor (2007): 3.693
  Journal Quartiles (2007): Q2 Oncology; Q2 Biochemistry & Molecular Biology
  Number of citations (December 2011): 11
  * *Equally contributing authors*

- **Solé X**, Hernández P, de Heredia ML, Armengol L, Rodríguez-Santiago B, Gómez L, Maxwell CA, Aguiló F, Condom E, Abril J, Pérez-Jurado L, Estivill X, Nunes V, Capellá G, Gruber SB, Moreno V and Pujana MA. Genetic and genomic analysis modeling of germline c-myc overexpression and cancer susceptibility. BMC Genomics, 9:12, 2008
  Journal Impact Factor (2008): 3.926
  Journal Quartiles (2008): Q1 Biotechnology & applied microbiology; Q1 Genetics & heredity
  Number of citations (December 2011): 11

- **Solé X**, Bonifaci N, López-Bigas N, Berenguer A, Hernández P, Reina O, Maxwell CA, Aguilar H, Urruticoechea A, de Sanjosé S, Comellas F, Capellá G, Moreno V and Pujana MA. Biological convergence of cancer signatures. PLoS One, 4(2): e4544, 2009
  Journal Impact Factor (2009): 4.351
  Journal Quartiles (2009): Q1 Biology
  Number of citations (December 2011): 5

5

ARTICLE 1: INTEGRATIVE ANALYSIS OF A CANCER SOMATIC MUTOME

5.1 SUMMARY

In 2006, Sjöblom and colleagues published a pioneer study, in which most part of the human consensus coding sequences (CCDS) were sequenced for 11 breast [188]. These CCDS represent the most highly curated gene set currently available[1]. Despite the small number of sequenced samples, this project unveiled for the first time the most complete view of the breast and colorectal cancer mutations (what could be called as the *mutome*). Although some of the altered genes had already been described, most of them had not been previously related to the carcinogenic process. As a consequence of this accumulation of alterations, molecular interactions are re-programmed in the context of highly connected and regulated cellular networks.

The aim of our study was to comprehensively describe the status of candidate breast cancer tumor suppressors and oncogenes at different molecular levels (from gene to proteins), as well as to predict new functional relationships between them and provide new hypotheses regarding their coordinated molecular function in the neoplastic process.

To investigate the potential role in cancer of somatically mutated breast cancer genes, genomic LOH was examined using a whole-genome SNP genotyping dataset. Mutated genes showed LOH ranging from 7% (*CNNM4*) to a maximum of 76% (*TP53*). As it was expected, other genes showing relatively high percentages of LOH in breast tumors were *BRCA1* (52%) and *MRE11A* (50%).

For a comprehensive understanding of LOH results, an integrative analysis of gene expression and SNP data was performed. About a 50% of mutome genes showed differential expression between healthy and tumor tissue samples. Careful examination of LOH identified 20 genes mapping to 12 critical regions. Expression analysis supported the supposition that 10 of these genes may act as tumor suppressors, as they show down-regulation in breast tumors.

1 http://www.ncbi.nlm.nih.gov/projects/CCDS/CcdsBrowse.cgi. Accessed September 2011.

CNV analysis was done on the same set of samples as the LOH analysis. Mutated genes showed copy numbers (CN) ranging from 1.60 to 3.37 across basal-like and non basal-like breast tumors. Examination of gene expression and critical regions with CN > 2 identified nine candidate oncogenes. Notably, one of these genes, *GAB1*, had been previously postulated to act as an oncogene in cellular transformation.

Correlation of LOH, CNV and expression data identified four concordant gene clusters (i. e. close located loci): amplification and overexpression of *ABCB10* and *NUP133* genes at chromosome 1 in basal-like and luminal A and B tumors; loss and down-regulation of *COL7A1*, *DNASE1L3*, *FLNB* and *RRP9* at chromosome 3, particularly in basal-like and luminal B tumors; loss and down-regulation of *MAMDC4*, *GSN*, *NUP214* and *SPTAN1* at chromosome 9, particularly in luminal A and B tumors; loss and down-regulation of *SORL1* and *TECTA* at chromosome 11, particularly in basal-like tumors.

To further determine the level of functional association among somatically mutated breast cancer genes, their co-expression pattern was assessed using a large breast tumor dataset containing 98 primary tumors. A higher level of co-expression than expected by chance was found . Overall, four clusters of strongly correlated genes were observed, which could be classified as *ETV6-NTRK3* (two), *TP53* or *RB1*-related.

Using a dataset containing survival information from 113 patients, the prognostic value of gene expression levels was evaluated. This analysis identified four validated genes whose expression levels predicted disease-free survival: *ABCA3*, *DBN1*, *SP110* and *SPTAN1*.

To evaluate functional associations between proteins, mutome gene products were mapped on the human interactome network. Analysis of the network showed that mutated genes were highly connected, supporting the theory that they may be involved in related molecular pathways or functions.

To generate a network model containing relevant biological information for the breast cancer neoplastic process, different types of functional relationships identified through previously-mentioned genomics (i.e. LOH, CN and expression) and proteomics (i.e. interactome) analyses were integrated. Thus, in the network model two nodes were connected when their corresponding genes showed similar LOH, CN or expression profiles across breast tumors, or when their corresponding encoded gene products were directly connected in the protein interactome network. Analysis of densely connected sub-graphs and their gene ontology (GO)

terms[2] identified functional modules enriched for apoptosis, cell division, cell differentiation, G-protein coupled receptor protein signaling pathway, intracellular signaling cascade, regulation of transcription, regulation of translation and signaling transduction.

5.2 MAIN RESULTS

DBN1 is a candidate oncogene that, when highly expressed in tumors with respect to healthy tissues, predicts poor survival in breast cancer patients.

Low expression ratios of *ABCA3* and low or medium expression ratios of *SPTAN1* may also predict poor survival in breast cancer. *ABCA3* was previously identified as an *ESR1*-regulated gene, which supports its involvement in breast tumorigenesis, and *SPTAN1* was involved in chemotherapy resistance in ovarian cancer, which makes this gene a potential target for cancer treatment.

The interactome analysis of molecular pathways provides new hypotheses for the identification of genes potentially associated with survival outcome. *SPTAN1* interacts with *GRIN2D* and *SLC9A2*, both of which interact with the product of the *ABL1* proto-oncogene. Activated *ABL1* kinase promotes invasion of breast cancer cells. Since low expression ratios of *SPTAN1* predict poor survival, *SPTAN1* could therefore act as a negative regulator of *ABL1* activity.

2 http://www.geneontology.org/. Accessed September 2011.

Molecular Cancer

BioMed Central

Research

Open Access

Integrative analysis of a cancer somatic mutome

Pilar Hernández†, Xavier Solé†, Joan Valls†, Víctor Moreno, Gabriel Capellá, Ander Urruticoechea and Miguel Angel Pujana*

Address: Bioinformatics and Biostatistics Unit, and Translational Research Laboratory, Catalan Institute of Oncology, IDIBELL, L'Hospitalet, Barcelona 08907, Spain

Email: Pilar Hernández - phgutierrez@ico.scs.es; Xavier Solé - xsole@ico.scs.es; Joan Valls - jvalls@ico.scs.es; Víctor Moreno - vmoreno@ico.scs.es; Gabriel Capellá - gcapella@ico.scs.es; Ander Urruticoechea - anderu@ico.scs.es; Miguel Angel Pujana* - mapujana@ico.scs.es

* Corresponding author †Equal contributors

Published: 5 February 2007

Molecular Cancer 2007, **6**:13 doi:10.1186/1476-4598-6-13

Received: 4 December 2006
Accepted: 5 February 2007

This article is available from: http://www.molecular-cancer.com/content/6/1/13

Abstract

Background: The consecutive acquisition of genetic alterations characterizes neoplastic processes. As a consequence of these alterations, molecular interactions are reprogrammed in the context of highly connected and regulated cellular networks. The recent identification of the collection of somatically mutated genes in breast tumors (breast cancer somatic "mutome") allows the comprehensive study of its function and organization in complex networks.

Results: We analyzed functional genomic data (loss of heterozygosity, copy number variation and gene expression in breast tumors) and protein binary interactions from public repositories to identify potential novel components of neoplastic processes, the functional relationships between them, and to examine their coordinated function in breast cancer pathogenesis. This analysis identified candidate tumor suppressors and oncogenes, and new genes whose expression level predicts survival rate in breast cancer patients. Mutome network modeling using different types of pathological and healthy functional relationships unveils functional modules significantly enriched in genes or proteins (genes/proteins) with related biological process Gene Ontology terms and containing known breast cancer-related genes/proteins.

Conclusion: This study presents a comprehensive analysis of the breast somatic mutome, highlighting those genes with a higher probability of playing a determinant role in tumorigenesis and better defining molecular interactions related to the neoplastic process.

Background

Recent landmark work has described the genetic landscape of the breast and colorectal cancer genomes by identifying the collection of somatically mutated genes (cancer somatic mutome) that contributes to the neoplastic process in these cancer types [1]. Most of these genes were not previously identified as linked to human cancer and some of them encode uncharacterized proteins. A larger set of "passenger" mutations or mutations present at a frequency that is too low to determine their relationship with cancer were also identified, prompting further genetic and molecular characterization.

Most biological processes involve groups of genes and proteins that behave in a coordinated way to perform a cellular function [2]. The coordinated task of genes/pro-

teins can be represented by different types of functional relationships (e.g. gene co-expression, genetic interactions, protein binary interactions, protein complex membership) [3]. Network modeling has been used to predict new gene/protein functions and to define pathway components or modulators of particular processes [reviewed in [4-6]]. The application of similar approaches has also identified new genes responsible for human diseases [7,8].

Defining biological processes at the systems-level will help to understand cancer cellular networks. The application of an integrative "omic" approach to the breast cancer somatic mutome is encouraged by the identification of uncharacterized genes/proteins and because the complete wiring diagram of functional associations has yet to be determined. The aim of this study is therefore to comprehensively describe the status of candidate breast cancer tumor suppressors and oncogenes at different molecular levels (from gene to protein), to predict new functional relationships between them and to provide new hypotheses regarding their coordinated molecular function in the neoplastic process. This study is focused on the somatic mutome described by Sjoblom et al. [1], which contains validated (contributing to the neoplastic process) and non-validated (i.e. harboring putative "passenger" mutations or mutations present at a frequency that is too low to determine their relationship with the neoplastic process) gene sets (total 672), combined with previously known somatically mutated breast cancer genes compiled in the COSMIC database [9].

Results

Loss of heterozigosity analysis

To investigate the role of somatically mutated breast cancer genes as classical tumor suppressors or oncogenes, we first examined genomic loss of heterozigosity (LOH) using a whole-genome SNP genotyping data set [10]. This data set has a resolution of one SNP every ~210 genomic kilo-bases and contains information from 42 breast tumors (20 non basal-like, 18 basal-like and 4 BRCA1 tumors) and matched healthy breast tissue samples.

When all breast tumors were considered, mutated genes in the validated set showed LOH ranging from 4% to a maximum of 76% (*TP53*)(Additional file 1). As was expected, other genes showing relatively high percentages of LOH in breast tumors were *BRCA1* (52%) and *MRE11A* (50%). Remarkably, of the validated genes only *CDH5* was previously described in detail as showing LOH [11], which might be explained by the unbiased approach used to identify the breast cancer somatic mutome, or by the inexistence of LOH as a second-hit genetic mechanism common to this set of genes. The detection of ~33% of LOH at the *TMPRSS6* locus supports its role as a tumor suppressor suggested by a previous observation that *TMPRSS6* nucleotide variants conferred a risk of breast cancer [12]. However, LOH should be interpreted with caution as it shows a high correlation with chromosome location (e.g. complete LOH of chromosome 17). LOH results do not significantly vary between basal-like and non basal-like tumor subtypes except for the isodisomy of chromosomes 14, 17 and X [10].

For a comprehensive understanding of LOH results, we integrated gene expression data available for the same healthy and tumor samples used for SNP genotyping, and combined it with a larger expression data set containing basal-like and other tumor subtypes [13] (Fig. 1). Approximately 50% of mutome genes showed differential expression between healthy and tumor tissue samples. Careful examination of LOH identified 20 genes in the validated set mapping to 12 critical regions (relatively close genomic boundaries of LOH). Expression analysis supports the supposition that 10 of these genes may act as tumor suppressors, as they show down-regulation in breast tumors (Fig. 1C, LOH column and down-regulated genes in tumors). In addition to these genes, a few others showed concordant results between LOH and expression analyses but cannot be mapped to critical regions (*CENTG1*, *MAGEE1*, *PRPS1*, *SYNE2* and *TP53*). Although not completely clear from LOH, the integration of expression data also supports the role of *ICAM5* as a tumor suppressor proposed by the identification of nucleotide variants that confer a risk of breast cancer [14]. The present LOH analysis suggests the loss of the *ICAM5* locus in non basal-like tumors (15%) but not in BRCA1 or basal-like (< 5%) tumors, and its expression appears significantly down-regulated in three distinct types of tumors when compared to healthy tissues [luminal A, luminal B and tumors showing human epidermal growth factor receptor 2 positivity (HER-2+) and estrogen receptor negativity (ER-)]. Collectively, the integration of LOH and expression analyses suggests the hypothesis of the existence of at least ~10 tumor suppressor genes in the breast cancer somatic mutome.

Copy number analysis

Using the same data set described above, genes in the validated set showed copy numbers (CNs) ranging from 1.60 to 3.37 across basal-like and non basal-like tumors (Additional file 2). As expected for tumors with relatively higher levels of genomic instability, broader margins of CN variation were observed in BRCA1 tumors, ranging from 0.57 to 3.82. Examination of gene expression and critical regions with CN > 2 identified nine candidate oncogenes (Fig. 1C, CN > 2 column and up-regulated genes in tumors). Notably, one of these genes, *GAB1*, was previously suggested to act as an oncogene in cellular transformation [15]. CN analysis also identified critical regions of

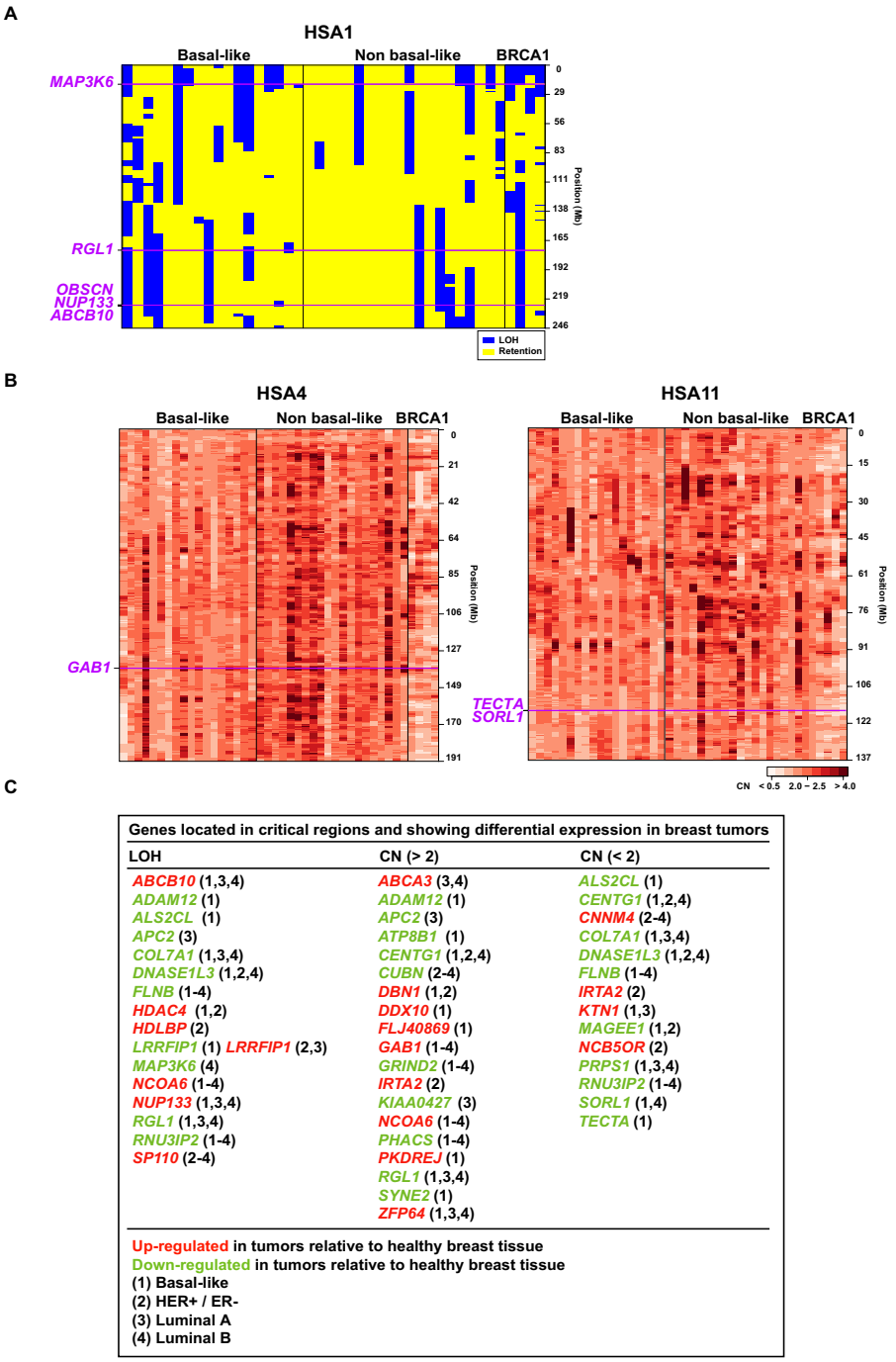


Figure 1

Integration of LOH, CN and expression data to better define candidate tumor suppressors and oncogenes for the breast cancer neoplastic process. Examples of LOH and CN analyses: (A) LOH analysis for HSA1 shows three critical regions (defined by close boundaries of LOH) indicated by pink lines across tumor samples; (B) CN analyses indicate *GAB1* locus genomic amplification in HSA4, and *SORL1* and *TECTA* loci genomic loss in HSA11; and (C) Integration of LOH and CN, and differential expression in tumors relative to healthy tissues indicate candidate tumor suppressors (down-regulated in tumors, green) and oncogenes (up-regulated in tumors, red) in four different types of breast tumors as indicated by numbers in brackets.

genomic loss that were not evident in the LOH analysis, such as the *SORL1* and *TECTA* loci that showed loss and expression down-regulation particularly in basal-like tumors (Fig. 1B and 1C). Thus, eight additional genes showed CN < 2 in a critical region and concordant down-regulation in tumors, which suggests their role as tumor suppressors (Fig. 1C, CN < 2 column and down-regulated genes in tumors).

In addition to the particular genes mentioned above, the correlation of LOH, CN and expression data identified four concordant gene clusters (i.e. close located loci). First, the amplification and over-expression of *ABCB10* and *NUP133* genes at chromosome 1 in basal-like and luminal A and B tumors. Remarkably, the amplification of *ATP-binding cassette (ABC) transporter* genes is commonly found in cancer cell lines as a probable mechanism of drug resistance [16] and nuclear pore (NUP) subunits have been found over-expressed in breast tumors [17]. Second, the loss and down-regulation of *COL7A1*, *DNASE1L3*, *FLNB* and *RNU3IP2* at chromosome 3, particularly in basal-like and luminal B tumors. Third, the loss and down-regulation of *AEGP*, *GSN*, *NUP214* and *SPTAN1* at chromosome 9, particularly in luminal A and B tumors. Finally, the loss and down-regulation of *SORL1* and *TECTA* at chromosome 11, particularly in basal-like tumors. These genomic mutome clusters suggest that, in addition to point mutations, large-scale alterations of these regions might constitute a mechanism contributing to the neoplastic process.

Expression analysis

To further determine the level of functional association among somatically mutated breast cancer genes, we investigated their co-expression in a large breast tumor data set containing 98 primary tumors [18]. A total of 878 probes corresponding to 680 (mutome plus benchmark) genes gave rise to 385,003 pair-wise comparisons. A higher number of these pairs than expected by chance show significant co-expression measured by the Pearson's correlation coefficient (PCC) (15,994 significant pairs applying a false discovery rate (FDR) of 0.01). Considering absolute PCC values, four clusters of high expression correlation were observed (Fig. 2). According to the presence of benchmark genes, co-expression clusters could be classified as *ETV6-NTKR3*, *TP53* or *RB1*-related. Since gene pairs that encode functionally related proteins tend to show higher expression correlation than pairs of unrelated genes, functional associations can be predicted based on profiling comparison. Thus, two genes in the *RB1*-related cluster encode known physical interactors of pRb (*ATF2* and *CUTL1*, included in the non-validated set) [19,20]. Similarly, the presence of *ABCB10* and *NUP133*, and candidate tumor suppressors *LRRFIP1* and *RNU3IP2* in the *RB1*-related cluster, further support their functional association in breast cancer.

Next, we examined whether gene expression levels have prognostic value and how this correlates with genomic and expression alterations in breast tumors. We used a data set containing information from 113 patients [13] and performed Kaplan-Meier analyses using the Cox-Mantel log-rank test. Cox's regression models were adjusted and non-adjusted for tumor grade and ER status. This analysis identified four validated genes whose expression levels predict survival (non-adjusted *P* values < 0.001 and adjusted *P* values < 0.05; genes *ABCA3*, *DBN1*, *SP110* and *SPTAN1* with adjusted hazard ratios (HR) of 0.58, 2.86, 0.59 and 0.20, respectively) (Fig. 3). Two other validated genes were identified with a lower significance level (non-adjusted *P* values < 0.01 and adjusted *P* values < 0.1; *C22orf19* and *RASGRF2* withHR of 2.29 and 0.36, respectively) and 17 genes in the non-validated set show association (adjusted *P* values < 0.05) (Additional file 3). Analysis of an independent data set containing information from 295 patients [21] supports the observation that high expression ratios of *DBN1* predict poor survival (adjusted *P* value of 0.03 and HR of 3.81) and indicates the same tendency as previously noted for low expression ratios of *ABCA3*, *SP110* and *SPTAN1* (non-adjusted HR of 0.31, 0.34 and 0.64, respectively), although this now appears non-significant when adjusted for tumor grade and ER status (adjusted HR of 0.61, 0.25 and 1.19). In the non-validated set, only *WFDC1* expression remained associated with survival in the multivariate analysis of the independent data (adjusted *P* values of 0.001 and 0.03, and HR of 3.99 and 7.63 for two different microarray probes).

Interactome analysis

To evaluate functional associations between proteins, we mapped mutome gene products on the human interactome network [22-24]. Since similar Gene Ontology (GO) annotations are more likely to be present in pairs of interacting proteins than in pairs of unrelated proteins, functional predictions can be formulated based on annotations of neighbor proteins in the network. In particular, the examination of GO annotations provides functional assignment of uncharacterized gene products (Fig. 4A), such as the VEPH1 protein that was identified in a large-scale interactome mapping study of the TGF-beta signaling pathway [25].

An examination of binary protein interactions also highlights the possible need for more detailed mutational analyses of specific cellular components. Thus, an association between the breast and colorectal mutomes identified by Sjoblom et al. [1] is revealed by examining interactions between proteins of the extracellular matrix

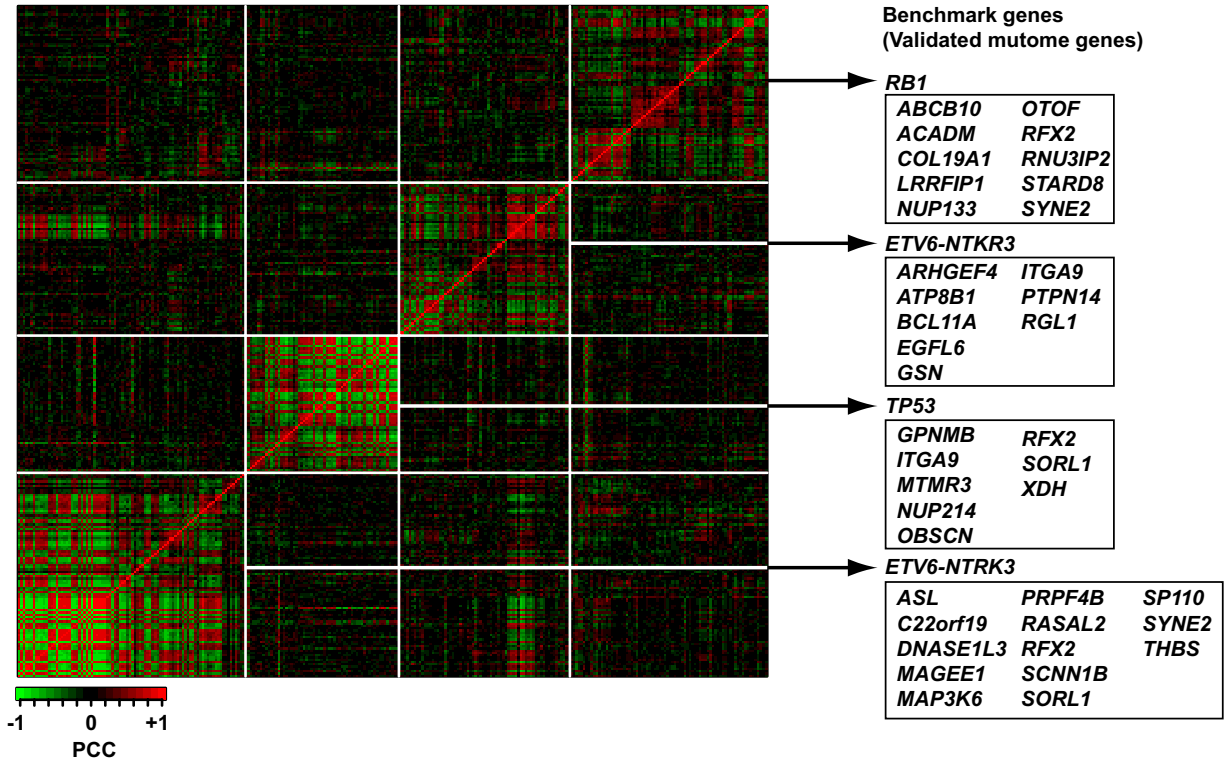


Figure 2
Gene co-expression analysis in breast tumors. Clustering of microarray probes (297 × 297) representing mutome (validated and non-validated) [1] and benchmark (literature) [9] genes according to absolute PCC values. Clusters are named according to the benchmark(s) gene(s) present in each of them (i.e. *RB1*, *ETV6-NTKR3* or *TP53*-related). Boxes contain validated mutome genes present in each cluster. Non-validated gene names are not shown.

and cytoskeleton functional module (Fig. 4B). In this module, four out of nine proteins included were found to be mutated in breast tumors and three were found to be mutated in colorectal tumors by Sjoblom et al. [1].

Next, we investigated the existence of coordinated molecular tasks by examining the level of connectivity between mutome gene products in the interactome network. We compared the size (number of nodes and edges) of the largest component generated by direct interactions between mutome validated proteins and compared it to equivalent randomly selected sets of 100 proteins. The results showed that mutome gene products are highly connected, more so than expected by chance (interactions/node, empirical P value < 0.05), thus supporting the theory that they are involved in related molecular pathways or functions. However, this observation is partially dependent on the presence of p53 and BRCA1, which exhibit extremely high connectivity. Without taking into account p53 and BRCA1, the level of connectivity of the validated mutome is still moderately high with respect to equivalent, randomly selected protein sets (empirical P value < 0.15). These results suggest greater centrality of the breast somatic mutome proteins and are consistent with earlier observations involving previously known human cancer proteins [26].

When only direct interactions are considered between validated and benchmark gene products, examination of the largest network component supports a critical role for three transcription factors or co-activators: MYOD1, NCOA6 and TCF1. These proteins appear included in a module with high connectivity that contains five members of the benchmark set (Fig. 5A). Notably among these genes, *NCOA6* maps to a critical region of CN > 2 (Fig. 1C). This gene was previously identified as amplified in breast tumors [27] and in this study appeared particularly over-expressed in basal-like tumors.

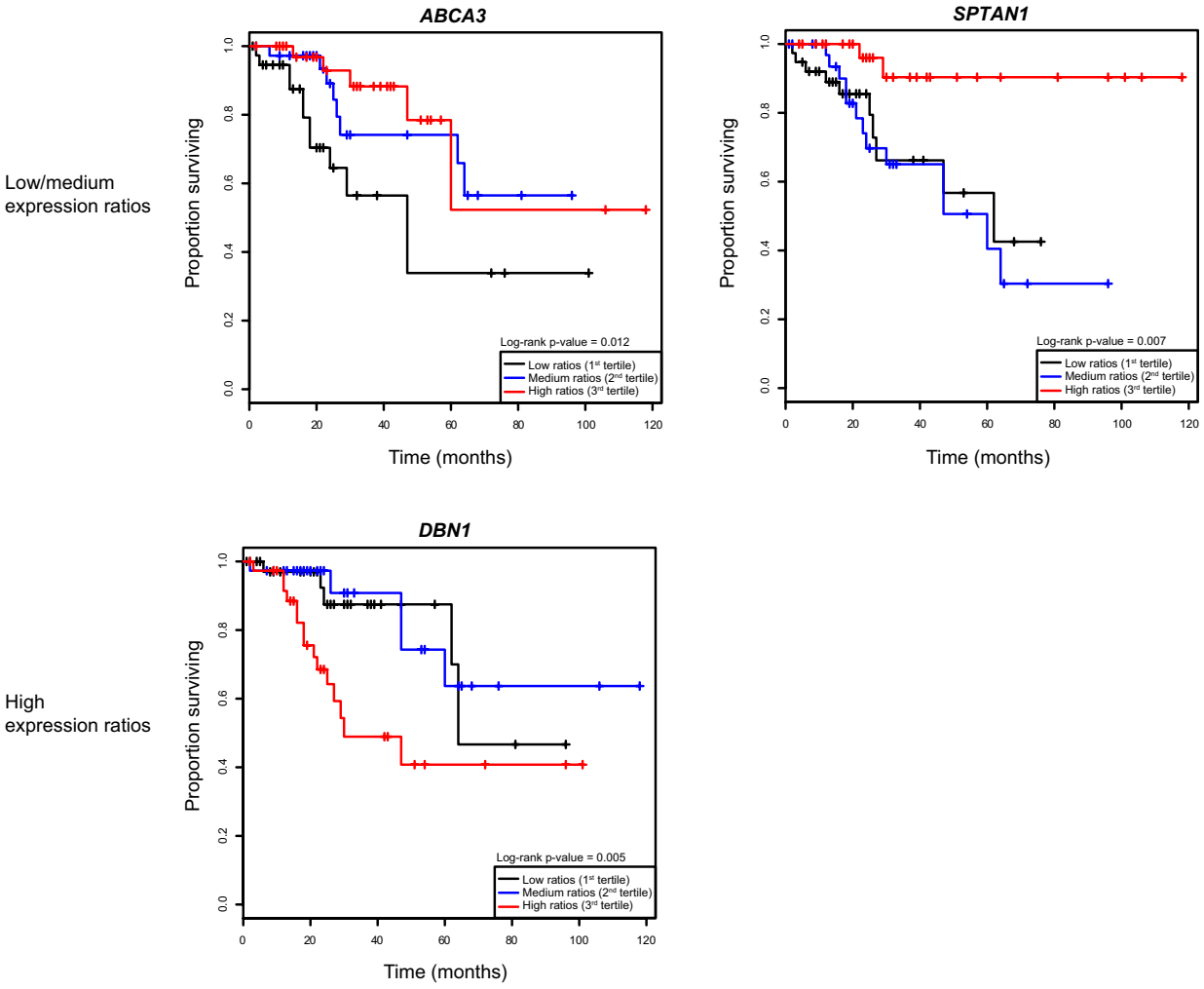


Figure 3
Gene expression analysis and breast cancer survival. Kaplan-Meier survival curves based upon categorized expression in tertiles are shown for three validated genes in the Hu et al. [13] data set.

When non-validated gene products are included in the interactome analysis, a large component with 127 edges and 94 nodes is revealed (Fig. 5B). Eight non-validated gene products occupy critical positions in this component, connecting validated and/or benchmark proteins: BCAR1 (breast cancer anti-estrogen resistance 1) links ADAM12 and GSN, therefore mediating extracellular matrix and cytoskeleton remodeling; and three gene products show a high degree of connectivity (between 5–10 interactions; PIK3R1, PLCG1 and POU2F1), which suggests a central role in the transmission of molecular information within this component. PIK3R1 and PLCG1 are involved in intracellular signaling cascades and their differential regulation is known to be involved in tumorigenesis [28,29], while POU2F1 interacts with several known breast cancer-associated proteins (i.e. BRCA1, BARD1 and PARP1) [30,31]. Together, these observations suggest a coordinated function between validated and non-validated gene products in the breast cancer neoplastic process.

Clustering analysis has previously proved useful for the identification of functionally related genes or proteins [32]. To further examine the higher-level organization of the breast cancer mutome, we identified densely interconnected regions of the interactome harboring a higher proportion of mutome gene products than expected by chance. One such cluster shows enrichment in functional

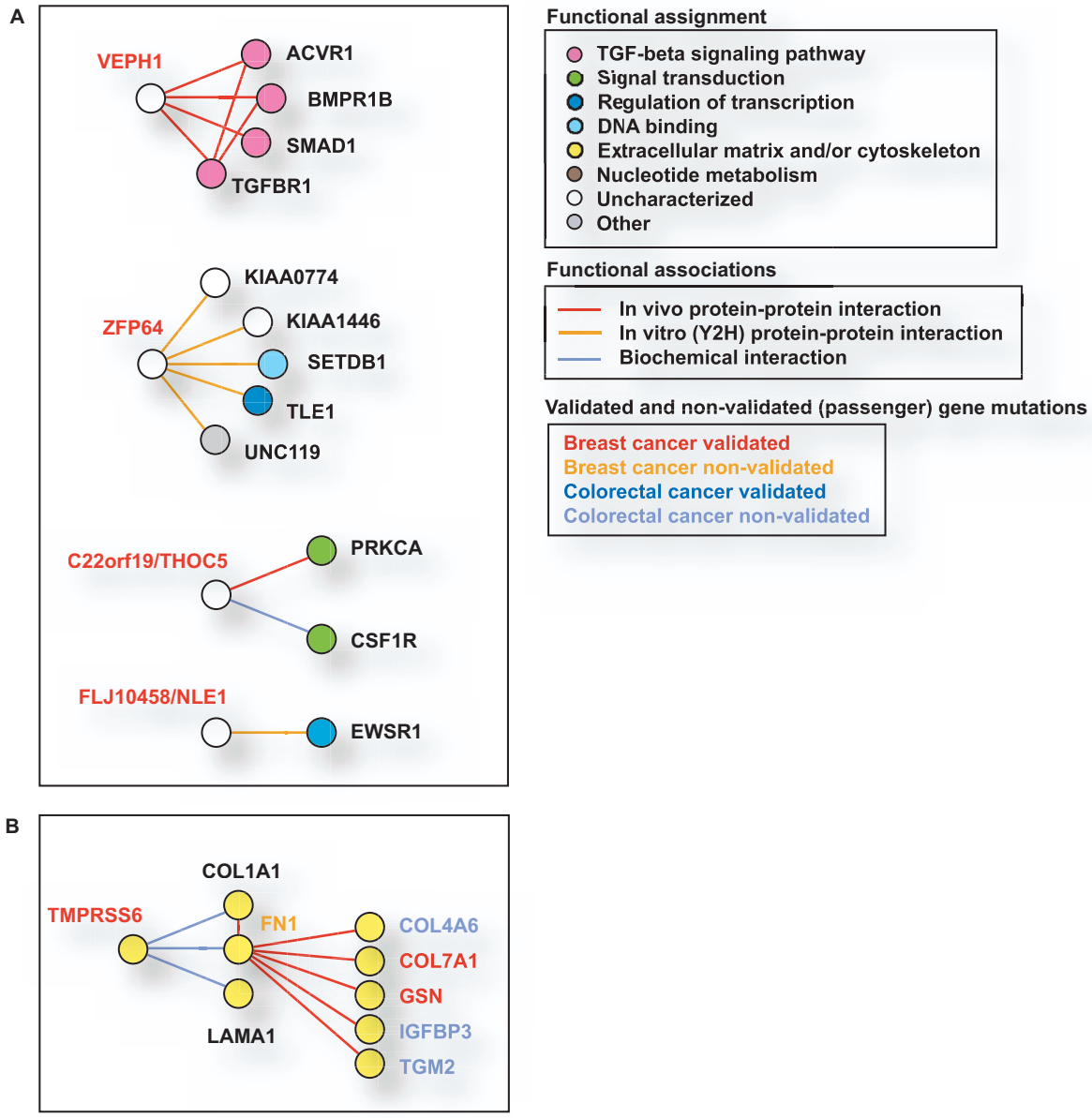


Figure 4
Human interactome network analysis, functional prediction and breast and colorectal cancer mutome association. (A) Predicted interactions for uncharacterized validated mutome gene products. Functional assignment is based on GO term annotations. Protein interactions and node types are indicated as shown in the insets. (B) Breast and colorectal cancer mutome association through extracellular matrix and cytoskeleton constituents.

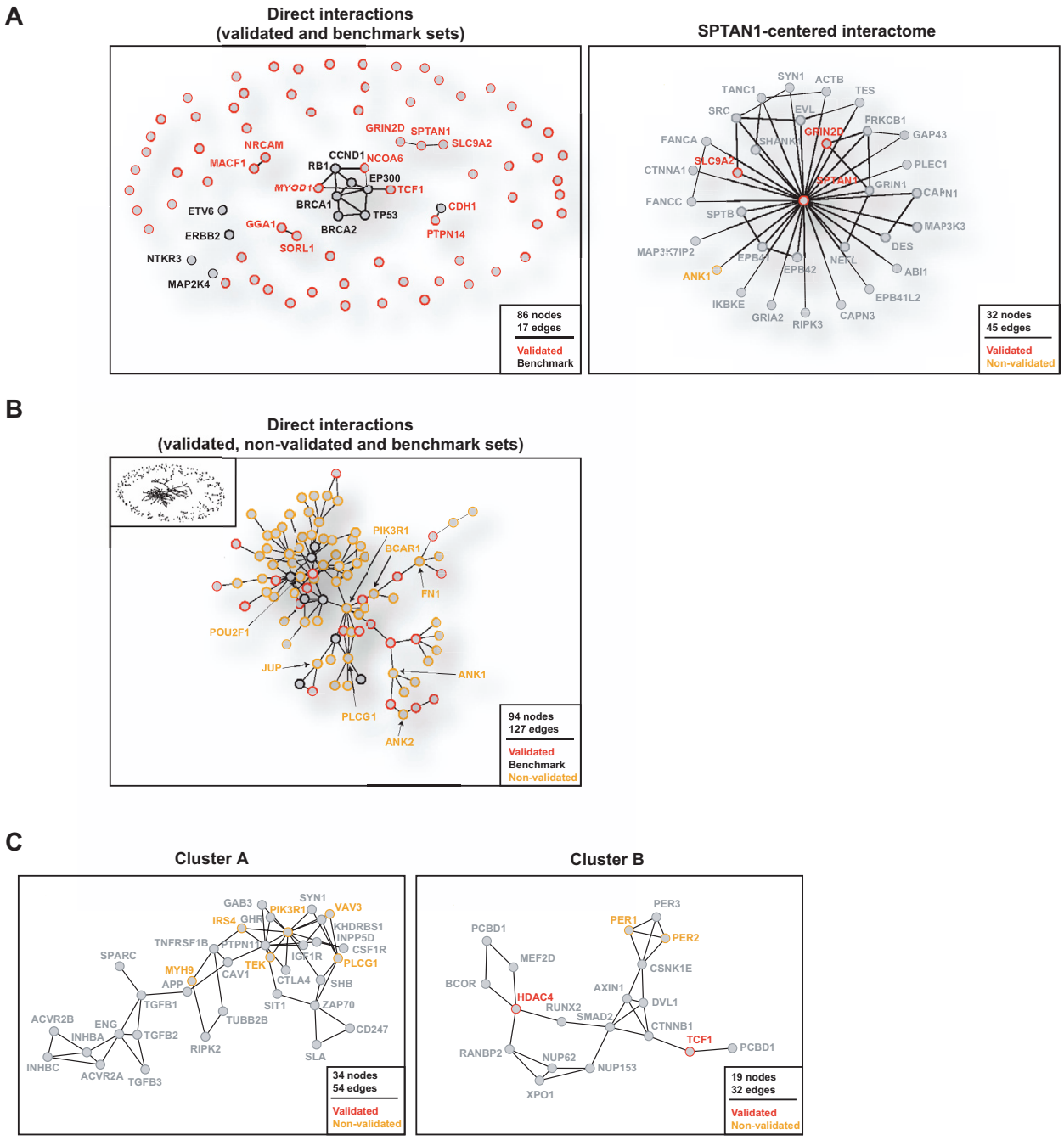


Figure 5

Human interactome network analysis, direct interactions between mutome gene products. (A) Left panel, direct interactions between validated mutome and/or benchmark gene products. Right panel, interactions centered on SPTAN1, whose expression level predicts survival (Fig. 3). Grey nodes represent non-mutome/benchmark proteins. (B) Network generated by direct protein interactions between validated and non-validated mutome and/or benchmark gene products (top left inset). An image of the largest component of this network is shown, with critical nodes that connect benchmark or mutome proteins indicated by arrows. (C) Clusters or densely connected regions in the interactome network that contain more mutome gene products than expected by chance: cluster A shows enrichment in annotations of the TGF-beta and insulin signaling pathways and of DNA transcriptional activity; cluster B shows enrichment for centrosome-related tasks and DNA transcriptional activity.

annotations of the TGF-beta and insulin signaling pathways as well as DNA transcriptional activity (Fig. 5C, cluster A). Another cluster shows enrichment for centrosome-related tasks and DNA transcriptional activity (Fig. 5C, cluster B). Cluster enrichment therefore points to known critical functional modules involved in breast tumorigenesis.

Mutome network modeling

To generate a network model containing relevant biological information for the breast cancer neoplastic process, we integrated different types of functional relationships identified through the genomic (i.e. LOH, CN and expression) and proteomic (i.e. interactome) analyses explained above. Thus, using network modeling we connected two nodes when their corresponding genes showed similar LOH, CN or expression profiles across breast tumors (see Methods), or when their corresponding encoded gene products were directly connected in the interactome network. The breast cancer mutome network contains 648 nodes and 8,371 edges, and shows a high degree of connectivity that further supports the existence of biologically related functions (Fig. 6 and Additional file 4).

Cluster analysis of this network identifies underlying molecular mechanisms of breast cancer. Analysis of densely connected sub-graphs and their GO terms identified functional modules enriched for apoptosis, cell division, cell differentiation, G-protein coupled receptor protein signaling pathway, intracellular signaling cascade, regulation of transcription, regulation of translation and signaling transduction (Fig. 6). Some benchmark genes/proteins can be located in these modules, supporting their role in the neoplastic process. These observations support the theory that the network modeled here represents a framework for a more in-depth experimental study of genes/proteins related to breast cancer somatic alterations.

Discussion

Although issues of specificity and sensitivity in the detection of the mutome will probably be addressed in the future, particularly regarding germline genomic CN variation [33] and the likelihood of detecting sequence changes as presented by Sjoblom et al. [1], by examining functional genomic (LOH, CN and gene expression) data in breast tumors, this study supports newly identified tumor suppressors and oncogenes. Through the examination of protein binary interactions, this study further provides new hypotheses regarding the functional associations of these gene products. Finally, the integration of pathological and healthy functional relationships generated a mutome network model that provides a framework for studying the coordinated molecular function of mutome genes/proteins.

The apparent discrepancy between cancer genomic and expression changes for some genes, such as genomic CN > 2 and expression down-regulation, is not exceptional and has been observed previously [34]. Autoregulation of gene expression, dosage compensation, epistatic modifications, or merely issues such as the sensitivity and specificity of LOH/CN and expression analyses can explain these apparent discrepancies. As is to be expected, the proportion of down-regulated genes is higher in CN < 2 than in CN > 2 regions, while the proportion of up-regulated genes is higher in CN > 2 than in CN < 2 regions (Fig. 1C). Nonetheless, experimental investigation of these genes/proteins is required to demonstrate their role as tumor suppressors or oncogenes.

The integrative study also serves as an indication of new prognosis markers. For the mutome genes, the integrative analysis of genomic copy number and expression data strongly indicates that *DBN1* is a candidate oncogene that, when highly expressed in tumors with respect to healthy tissues, predicts poor survival in breast cancer patients (Fig. 3). Low expression ratios of *ABCA3* and low or medium expression ratios of *SPTAN1* may also predict poor survival. *ABCA3* was previously identified as an ER-regulated gene [35], which supports its involvement in breast tumorigenesis, and *SPTAN1* was involved in chemotherapy resistance in ovarian cancer [36], which makes this gene a potential target for cancer treatment. Finally, the interactome analysis of molecular pathways provides new hypotheses for the identification of genes potentially associated with survival outcome. SPTAN1 interacts with GRIND2 and SLC9A2, both of which interact with the product of the *ABL1* proto-oncogene. Activated ABL1 kinase promotes invasion of breast cancer cells [37]. Since low expression ratios of *SPTAN1* predict poor survival, SPTAN1 could therefore act as a negative regulator of ABL1 activity.

The integration of omic data highlights likely functional candidates of a particular biological process with increased confidence [7,38]. The strategy used here is applicable to other cancer types and would help to identify new tumor suppressor genes and oncogenes and the wiring diagram of functional interactions between them. The analysis of the breast cancer somatic mutome indicates that at least a few of the genes identified by Sjoblom et al. [1] play a key role in the breast cancer neoplastic process. These results will help to focus subsequent experimental characterizations on key gene/protein candidates.

Conclusion

We have presented the first comprehensive omic analysis of a cancer somatic mutome. Our analysis supports the theory that a few of these genes play a key role in the breast cancer neoplastic process. This study also provides

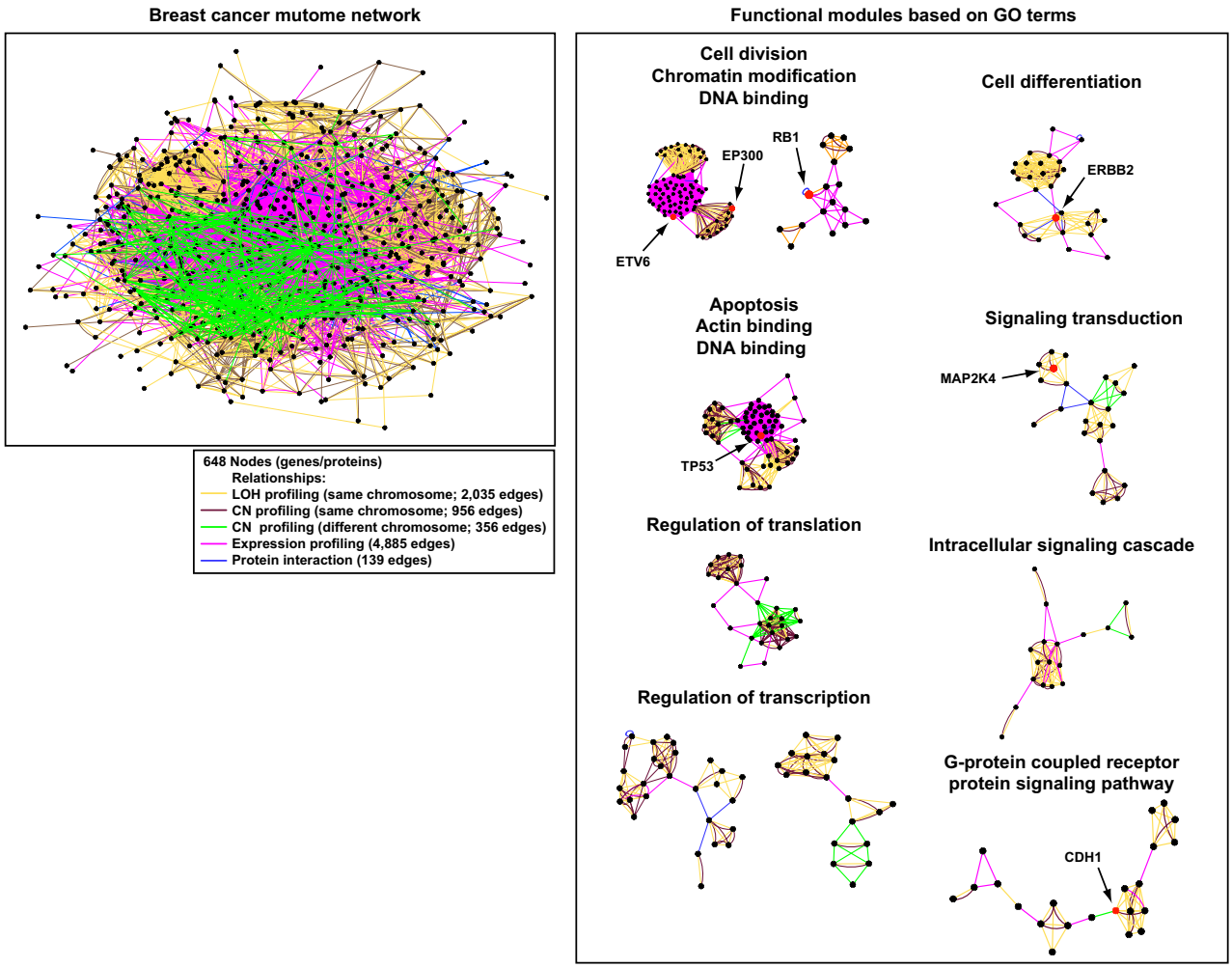


Figure 6
Breast cancer mutome network modeling. Left panel, five functional genomic or proteomic, pathological or healthy-related associations; each one indicated by one of the colored lines shown in the inset was included to generate a mutome network model. Right panel, clusters or densely connected regions in the network that show enrichment in GO terms (functional modules). Benchmark nodes present in these functional modules are marked by arrows.

new hypotheses for the coordinated function of these genes/proteins as tumor suppressors or oncogenes. Network modeling identifies hundreds of new potential pathological associations between the cancer genes/proteins studied. Extensive future research will be carried out by different groups focusing on each of the candidate genes highlighted by Sjoblom et al. [1]. Our study provides a possible framework for the appropriate initial categorization of these genes.

Methods

Genomic data analysis

To analyze LOH and CN alterations in breast tumors, we used the Gene Expression Omnibus (GEO) record GSE3743 [10]. Data were normalized and modelled using dChip software [39]. LOH and CN were obtained after mapping genes in build 35.1 of the NCBI human genome sequence. For each gene and sample we took the closest SNPs to infer LOH and CN. If there was a mismatch in LOH calling for surrounding SNPs, the gene was left as missing for that particular sample. LOH profile correlation and confidence intervals (CI) were computed using Cohen's kappa coefficient of agreement, suitable for categorical data. We then classified genes as showing similar profiling if the lower limit of the CI was greater than 0.6. PCC was used to assess CN profile correlations, setting 0.6 as the lower cut-off. To determine the level of correlation between gene expression and genomic CN variation, we

used PCC and FDR adjusted *P* values. All these analyses were performed using the R statistical software package [40].

Gene expression data analysis

Breast cancer gene expression was analyzed using two large data sets [10,13]. Data from Richardson et al. [10] was down-loaded from the GEO record GSE3744 and analyzed using the limma and affy packages in R. Background correction, normalization and averaging of expression values were computed using the RMA algorithm [41]. Differentially expressed genes were detected after computing an empirical Bayes moderated t-statistic and *P* values adjusted by a FDR of 5%. Data obtained from Hu et al. [13] was previously normalized and analyzed using the *t*-test. To evaluate co-expression, we used the data set of van 't Veer et al. [18], calculated PCCs and significance levels based on the t-distribution. A hierarchical algorithm was used to cluster genes, taking as distance the absolute value of 1-PCC. To evaluate prognosis, we used the Hu et al. data set [13] and fitted a Cox regression model to each gene using the overall survival information. An adjusted model taking into account tumor grade and ER status was also fitted for each gene. Likelihood ratio tests were used to evaluate the effect of gene expression on survival. For genes that appeared significant in both models, expression was categorized into tertiles using Kaplan-Meier curves. For these genes, the (non-parametric) log-rank test was calculated. The replica data set used for survival analysis was that of Chang et al. [21].

Human interactome network and clustering analyses

The human interactome network was built by combining three previously published data sets, which mainly represent experimentally-verified interactions [22-24]. The Gandhi et al. [22] data set contains compiled and filtered protein binary interactions from all currently available databases (HPRD, BIND, DIP, MINT, INTACT and MIPS). High-confidence yeast two-hybrid interactions from Rual et al. [24] and Stelzl et al. [23] were then included. After removing common interactions between the three data sets, the resulting network contained 8,174 nodes and 27,810 edges. The Molecular Complex Detection (MCODE) algorithm [42] was used to detect densely connected regions in the interactome network. To calculate the enrichment of mutome proteins in network clusters, a binomial distribution was used. Enrichment in GO terms was investigated using OntoExpress tools [43] and GENECODIS [44]. To determine the level of connectivity between validated mutome gene products, we compared the number of nodes and interactions in the largest component generated by direct interactions between these proteins (73 of 122 were mapped in the interactome) to the number of nodes and interactions generated by 100 iterations of 73 randomly chosen proteins in the interactome.

Competing interests

The author(s) declare that they have no competing interests.

Authors' contributions

PH compiled and analyzed the expression and interactome data sets. XS compiled and analyzed genomic loss of heterozygosity and copy number data. JV performed the gene co-expression and survival analyses. PH, XS, JV and AU helped to draft the manuscript. VM and GC provided institutional support and participated in scientific discussions. AU and MAP conceived the study. MAP designed and coordinated the study, and wrote the original and final versions of the manuscript. All authors have read and approved the final version of the manuscript

Additional material

Additional File 1

LOH analyses results.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1476-4598-6-13-S1.xls]

Additional File 2

CN analyses results.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1476-4598-6-13-S2.xls]

Additional File 3

Cox regression analyses of non-validated mutome genes using the Hu et al. [13] data set.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1476-4598-6-13-S3.xls]

Additional File 4

Functional relationships in the mutome network model.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1476-4598-6-13-S4.xls]

Acknowledgements

MAP would like to offer his personal thanks to Marc Vidal, for introducing him to and developing his knowledge of the world of complex systems. This work was supported by the Fundació la Caixa (grant BM05-254-00 awarded to MAP), the Catalan Institute of Oncology (PH), the Instituto de Salud Carlos III (RCESP-C03/09 and RTICCC-C03/10) and SAF2003/5821. MAP is a Ramón y Cajal Researcher with the Spanish Ministry of Education and Science.

References

1. Sjoblom T, Jones S, Wood LD, Parsons DW, Lin J, Barber TD, Mandelker D, Leary RJ, Ptak J, Silliman N, Szabo S, Buckhaults P, Farrell C, Meeh P, Markowitz SD, Willis J, Dawson D, Willson JK, Gazdar AF, Hartigan J, Wu L, Liu C, Parmigiani G, Park BH, Bachman KE, Papadopoulos N, Vogelstein B, Kinzler KW, Velculescu VE: **The consensus coding sequences of human breast and colorectal cancers.** *Science* 2006, **314(5797):**268-274.
2. Hartwell LH, Hopfield JJ, Leibler S, Murray AW: **From molecular to modular cell biology.** *Nature* 1999, **402(6761 Suppl):**C47-52.
3. Vidal M: **A biological atlas of functional maps.** *Cell* 2001, **104(3):**333-339.
4. Ge H, Walhout AJ, Vidal M: **Integrating 'omic' information: a bridge between genomics and systems biology.** *Trends Genet* 2003, **19(10):**551-560.
5. Liu ET: **Systems biology, integrative biology, predictive biology.** *Cell* 2005, **121(4):**505-506.
6. Vidal M: **Interactome modeling.** *FEBS Lett* 2005, **579(8):**1834-1838.
7. Mootha VK, Lepage P, Miller K, Bunkenborg J, Reich M, Hjerrild M, Delmonte T, Villeneuve A, Sladek R, Xu F, Mitchell GA, Morin C, Mann M, Hudson TJ, Robinson B, Rioux JD, Lander ES: **Identification of a gene causing human cytochrome c oxidase deficiency by integrative genomics.** *Proc Natl Acad Sci U S A* 2003, **100(2):**605-610.
8. Spinazzola A, Viscomi C, Fernandez-Vizarra E, Carrara F, D'Adamo P, Calvo S, Marsano RM, Donnini C, Weiher H, Strisciuglio P, Parini R, Sarzi E, Chan A, DiMauro S, Rotig A, Gasparini P, Ferrero I, Mootha VK, Tiranti V, Zeviani M: **MPV17 encodes an inner mitochondrial membrane protein and is mutated in infantile hepatic mitochondrial DNA depletion.** *Nat Genet* 2006, **38(5):**570-575.
9. Futreal PA, Coin L, Marshall M, Down T, Hubbard T, Wooster R, Rahman N, Stratton MR: **A census of human cancer genes.** *Nat Rev Cancer* 2004, **4(3):**177-183.
10. Richardson AL, Wang ZC, De Nicolo A, Lu X, Brown M, Miron A, Liao X, Iglehart JD, Livingston DM, Ganesan S: **X chromosomal abnormalities in basal-like human breast cancer.** *Cancer Cell* 2006, **9(2):**121-132.
11. Roylance R, Gorman P, Papior T, Wan YL, Ives M, Watson JE, Collins C, Wortham N, Langford C, Fiegler H, Carter N, Gillett C, Sasieni P, Pinder S, Hanby A, Tomlinson I: **A comprehensive study of chromosome 16q in invasive ductal and lobular breast carcinoma using array CGH.** *Oncogene* 2006, **25(49):**6544-6553.
12. Hartikainen JM, Tuhkanen H, Kataja V, Eskelinen M, Uusitupa M, Kosma VM, Mannermaa A: **Refinement of the 22q12-q13 breast cancer--associated region: evidence of TMPRSS6 as a candidate gene in an eastern Finnish population.** *Clin Cancer Res* 2006, **12(5):**1454-1462.
13. Hu Z, Fan C, Oh DS, Marron JS, He X, Qaqish BF, Livasy C, Carey LA, Reynolds E, Dressler L, Nobel A, Parker J, Ewend MG, Sawyer LR, Wu J, Liu Y, Nanda R, Tretiakova M, Ruiz Orrico A, Dreher D, Palazzo JP, Perreard L, Nelson E, Mone M, Hansen H, Mullins M, Quackenbush JF, Ellis MJ, Olopade OI, Bernard PS, Perou CM: **The molecular portraits of breast tumors are conserved across microarray platforms.** *BMC Genomics* 2006, **7:**96.
14. Kammerer S, Roth RB, Reneland R, Marnellos G, Hoyal CR, Markward NJ, Ebner F, Kiechle M, Schwarz-Boeger U, Griffiths LR, Ulbrich C, Chrobok K, Forster G, Praetorius GM, Meyer P, Rehbock J, Cantor CR, Nelson MR, Braun A: **Large-scale association study identifies ICAM gene region as breast and prostate cancer susceptibility locus.** *Cancer Res* 2004, **64(24):**8906-8910.
15. Holgado-Madruga M, Emlet DR, Moscatello DK, Godwin AK, Wong AJ: **A Grb2-associated docking protein in EGF- and insulin-receptor signalling.** *Nature* 1996, **379(6565):**560-564.
16. Yasui K, Mihara S, Zhao C, Okamoto H, Saito-Ohara F, Tomida A, Funato T, Yokomizo A, Naito S, Imoto I, Tsuruo T, Inazawa J: **Alteration in copy numbers of genes as a mechanism for acquired drug resistance.** *Cancer Res* 2004, **64(4):**1403-1410.
17. Agudo D, Gomez-Esquer F, Martinez-Arribas F, Nunez-Villar MJ, Pollan M, Schneider J: **Nup88 mRNA overexpression is associated with high aggressiveness of breast cancer.** *Int J Cancer* 2004, **109(5):**717-720.
18. van 't Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AA, Mao M, Peterse HL, van der Kooy K, Marton MJ, Witteveen AT, Schreiber GJ, Kerkhoven RM, Roberts C, Linsley PS, Bernards R, Friend SH: **Gene expression profiling predicts clinical outcome of breast cancer.** *Nature* 2002, **415(6871):**530-536.
19. Kim SJ, Wagner S, Liu F, O'Reilly MA, Robbins PD, Green MR: **Retinoblastoma gene product activates expression of the human TGF-beta 2 gene through transcription factor ATF-2.** *Nature* 1992, **358(6384):**331-334.
20. Gupta S, Luong MX, Bleuming SA, Miele A, Luong M, Young D, Knudsen ES, Van Wijnen AJ, Stein JL, Stein GS: **Tumor suppressor pRB functions as a co-repressor of the CCAAT displacement protein (CDP/cut) to regulate cell cycle controlled histone H4 transcription.** *J Cell Physiol* 2003, **196(3):**541-556.
21. Chang HY, Nuyten DS, Sneddon JB, Hastie T, Tibshirani R, Sorlie T, Dai H, He YD, van't Veer LJ, Bartelink H, van de Rijn M, Brown PO, van de Vijver MJ: **Robustness, scalability, and integration of a wound-response gene expression signature in predicting breast cancer survival.** *Proc Natl Acad Sci U S A* 2005, **102(10):**3738-3743.
22. Gandhi TK, Zhong J, Mathivanan S, Karthick L, Chandrika KN, Mohan SS, Sharma S, Pinkert S, Nagaraju S, Periaswamy B, Mishra G, Nandakumar K, Shen B, Deshpande N, Nayak R, Sarker M, Boeke JD, Parmigiani G, Schultz J, Bader JS, Pandey A: **Analysis of the human protein interactome and comparison with yeast, worm and fly interaction datasets.** *Nat Genet* 2006, **38(3):**285-293.
23. Stelzl U, Worm U, Lalowski M, Haenig C, Brembeck FH, Goehler H, Stroedicke M, Zenkner M, Schoenherr A, Koeppen S, Timm J, Mintzlaff S, Abraham C, Bock N, Kietzmann S, Goedde A, Toksoz E, Droege A, Krobitsch S, Korn B, Birchmeier W, Lehrach H, Wanker EE: **A human protein-protein interaction network: a resource for annotating the proteome.** *Cell* 2005, **122(6):**957-968.
24. Rual JF, Venkatesan K, Hao T, Hirozane-Kishikawa T, Dricot A, Li N, Berriz GF, Gibbons FD, Dreze M, Ayivi-Guedehoussou N, Klitgord N, Simon C, Boxem M, Milstein S, Rosenberg J, Goldberg DS, Zhang LV, Wong SL, Franklin G, Li S, Albala JS, Lim J, Fraughton C, Llamosas E, Cevik S, Bex C, Lamesch P, Sikorski RS, Vandenhaute J, Zoghbi HY, Smolyar A, Bosak S, Sequerra R, Doucette-Stamm L, Cusick ME, Hill DE, Roth FP, Vidal M: **Towards a proteome-scale map of the human protein-protein interaction network.** *Nature* 2005, **437(7062):**1173-1178.
25. Barrios-Rodiles M, Brown KR, Ozdamar B, Bose R, Liu Z, Donovan RS, Shinjo F, Liu Y, Dembowy J, Taylor IW, Luga V, Przulj N, Robinson M, Suzuki H, Hayashizaki Y, Jurisica I, Wrana JL: **High-throughput mapping of a dynamic signaling network in mammalian cells.** *Science* 2005, **307(5715):**1621-1625.
26. Jonsson PF, Bates PA: **Global topological features of cancer proteins in the human interactome.** *Bioinformatics* 2006, **22(18):**2291-2297.
27. Guan XY, Xu J, Anzick SL, Zhang H, Trent JM, Meltzer PS: **Hybrid selection of transcribed sequences from microdissected DNA: isolation of genes within amplified region at 20q11-q13.2 in breast cancer.** *Cancer Res* 1996, **56(15):**3446-3450.
28. Crowder RJ, Ellis MJ: **Treating breast cancer through novel inhibitors of the phosphatidylinositol 3'-kinase pathway.** *Breast Cancer Res* 2005, **7(5):**212-214.
29. Arteaga CL, Johnson MD, Todderud G, Coffey RJ, Carpenter G, Page DL: **Elevated content of the tyrosine kinase substrate phospholipase C-gamma 1 in primary human breast carcinomas.** *Proc Natl Acad Sci U S A* 1991, **88(23):**10435-10439.
30. Fan W, Jin S, Tong T, Zhao H, Fan F, Antinore MJ, Rajasekaran B, Wu M, Zhan Q: **BRCA1 regulates GADD45 through its interactions with the OCT-1 and CAAT motifs.** *J Biol Chem* 2002, **277(10):**8061-8067.
31. Nie J, Sakamoto S, Song D, Qu Z, Ota K, Taniguchi T: **Interaction of Oct-1 and automodification domain of poly(ADP-ribose) synthetase.** *FEBS Lett* 1998, **424(1-2):**27-32.
32. Gunsalus KC, Ge H, Schetter AJ, Goldberg DS, Han JD, Hao T, Berriz GF, Bertin N, Huang J, Chuang LS, Li N, Mani R, Hyman AA, Sonnichsen B, Echeverri CJ, Roth FP, Vidal M, Piano F: **Predictive models of molecular machines involved in Caenorhabditis elegans early embryogenesis.** *Nature* 2005, **436(7052):**861-865.
33. Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH, Andrews TD, Fiegler H, Shapero MH, Carson AR, Chen W, Cho EK, Dallaire S, Freeman JL, Gonzalez JR, Gratacos M, Huang J, Kalaitzopoulos D, Komura D, Macdonald JR, Marshall CR, Mei R, Montgomery L, Nishimura K, Okamura K, Shen F, Somerville MJ, Tchinda J, Valsesia A, Woodwark C, Yang F, Zhang J, Zerjal T, Zhang J, Armengol L, Conrad DF, Estivill X, Tyler-Smith C, Carter NP, Aburatani H, Lee C, Jones KW, Scherer

SW, Hurles ME: **Global variation in copy number in the human genome.** *Nature* 2006, **444(7118):**444-454.
34. Pollack JR, Sorlie T, Perou CM, Rees CA, Jeffrey SS, Lonning PE, Tibshirani R, Botstein D, Borresen-Dale AL, Brown PO: **Microarray analysis reveals a major direct role of DNA copy number alteration in the transcriptional program of human breast tumors.** *Proc Natl Acad Sci U S A* 2002, **99(20):**12963-12968.
35. Lin CY, Strom A, Vega VB, Kong SL, Yeo AL, Thomsen JS, Chan WC, Doray B, Bangarusamy DK, Ramasamy A, Vergara LA, Tang S, Chong A, Bajic VB, Miller LD, Gustafsson JA, Liu ET: **Discovery of estrogen receptor alpha target genes and response elements in breast tumor cells.** *Genome Biol* 2004, **5(9):**R66.
36. L'Esperance S, Popa I, Bachvarova M, Plante M, Patten N, Wu L, Tetu B, Bachvarov D: **Gene expression profiling of paired ovarian tumors obtained prior to and following adjuvant chemotherapy: molecular signatures of chemoresistant tumors.** *Int J Oncol* 2006, **29(1):**5-24.
37. Srinivasan D, Plattner R: **Activation of Abl tyrosine kinases promotes invasion of aggressive breast cancer cells.** *Cancer Res* 2006, **66(11):**5648-5655.
38. Dahia PL, Hao K, Rogus J, Colin C, Pujana MA, Ross K, Magoffin D, Aronin N, Cascon A, Hayashida CY, Li C, Toledo SP, Stiles CD: **Novel pheochromocytoma susceptibility loci identified by integrative genomics.** *Cancer Res* 2005, **65(21):**9651-9658.
39. Zhong S, Li C, Wong WH: **ChipInfo: Software for extracting gene annotation and gene ontology information for microarray analysis.** *Nucleic Acids Res* 2003, **31(13):**3483-3486.
40. **Website title [www.r-project.org].** .
41. 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.
42. Bader GD, Hogue CW: **An automated method for finding molecular complexes in large protein interaction networks.** *BMC Bioinformatics* 2003, **4:**2.
43. Draghici S, Khatri P, Martins RP, Ostermeier GC, Krawetz SA: **Global functional profiling of gene expression.** *Genomics* 2003, **81(2):**98-104.
44. Carmona-Saez P, Chagoyen M, Tirado F, Carazo JM, Pascual-Montano A: **GENECODIS: A web-based tool for finding significant concurrent annotations in gene lists.** *Genome Biol* 2007, **8(1):**R3.

6

ARTICLE 2: GENETIC AND GENOMIC ANALYSIS MODELING OF GERMLINE MYC OVEREXPRESSION AND CANCER SUSCEPTIBILITY

6.1 SUMMARY

Germline genetic variation at multiple loci at 8q24 region has been associated to an increased risk of developing some tumors, mainly in the breast, prostate, colon and rectum. Nevertheless, none of the currently known risk variants map at or relatively close to known genes. Only *MYC* is located a few hundred kilobases away from this region. Since germline genetic variants have been associated with differential expression of many human genes, the phenotypic effects of this type of variation may be important when considering susceptibility to common genetic diseases.

The aim of this study was to integrate genetic and genomic data to assess the impact of 8q24 variants in germline *MYC* expression and its role in tumorigenesis.

The association between genotypes and *MYC* expression levels was done using SNP and expression data publicly available from HapMap samples. The obtained results were further validated using other dataset from healthy prostate samples.

CNV analyses at the *MYC* locus were performed in Caucasian and African HapMap individuals, and also in 322 unrelated individuals from the Spanish general population. These analyses showed that a CNV including *MYC* does not seem to be a major contributor to the risk of prostate cancer and germline *MYC* overexpression associated with 8q24 genotypes.

Using a publicly available expression dataset containing different prostate cellular populations, a differential expression analysis of 8q24 genes was performed. This analysis confirmed the growing expression of *MYC* as the pathological stage of the prostate tumor progresses from normal prostate to metastasis. *MYC* expression was also found to positively correlate with higher Gleason scores. These observations pointed to a causal relationship between somatic *MYC* overexpression and the more aggressive forms of prostate tumors.

Using a dataset containing gene expression data for 50 healthy tissues and 52 prostate tumors, *MYC* targets already known to confer risk of prostate cancer were studied to identify if they could be functionally associated to *MYC*. This was done by examining the similarities between their expression profiles. The analysis revealed a strong correlation between *MYC* and the prostate tumor suppressor *KLF6* gene.

Models of direct transcriptional regulatory networks in prostate tissue were inferred using the ARACNe algorithm. 88 and 111 putative transcriptional targets of *MYC* and *KLF6* in this cell type were identified, respectively. The intersection of these two sets contained 25 genes, which was much a larger number of genes than expected by chance. *MYC* and *KLF6* were also directly connected and the *KLF6* promoter was found to contain three predicted binding sites for *MYC*, supporting their functional association and their role in prostate tumorigenesis.

Expression data derived from a model of *MYC*-driven cellular transformation of quiescent human mammary epithelial cells and from mouse mammary tumor virus (MMTV)-*Myc*-driven mammary tumors in mice were analyzed. Many of the 25 *MYC-KLF6* intersection genes were found to be differentially expressed in both mice models, while *KLF6* displayed a strong down-regulation, as well as its direct transcriptional target *CDH1*.

6.2 MAIN RESULTS

Cis-regulators of germline *MYC* expression in immortalized lymphocytes of HapMap individuals were identified. Quantitative analysis of *MYC* expression in normal prostate tissues suggested an association between *MYC* overexpression and 8q24 variants of prostate cancer risk.

Somatic *MYC* overexpression correlated with prostate cancer progression and more aggressive tumor forms.

Expression profiling analysis and modeling of transcriptional regulatory networks predicted a functional association between *MYC* and the prostate tumor suppressor *KLF6*.

Analysis of *MYC*/*Myc*-driven cell transformation and tumorigenesis substantiated a model in which *MYC* overexpression promotes transformation by down-regulating *KLF6*. In this model, a feedback loop through *CDH1* down-regulation might cause further transactivation of *MYC*.

BMC Genomics

BioMed Central

Research article

Open Access

Genetic and genomic analysis modeling of germline *c-MYC* overexpression and cancer susceptibility

Xavier Solé[1], Pilar Hernández[1], Miguel López de Heredia[2,3], Lluís Armengol[4], Benjamín Rodríguez-Santiago[5,6], Laia Gómez[1], Christopher A Maxwell[1], Fernando Aguiló[7], Enric Condom[8], Jesús Abril[2], Luis Pérez-Jurado[5,6,9], Xavier Estivill[4], Virginia Nunes[2,3,10], Gabriel Capellá[1], Stephen B Gruber[11], Víctor Moreno*[1] and Miguel Angel Pujana*[1]

Address: [1]Bioinformatics and Biostatistics Unit, and Translational Research Laboratory, Catalan Institute of Oncology, IDIBELL, L'Hospitalet, Barcelona, Spain, [2]Medical and Molecular Genetics Center, IDIBELL, L'Hospitalet, Barcelona, Spain, [3]CIBERER-U730, L'Hospitalet, Barcelona, Spain, [4]Genes and Disease Program, Center for Genomic Regulation, Barcelona, Spain, [5]Genetics Unit, Department of Experimental and Health Sciences, Universitat Pompeu Fabra, Barcelona, Spain, [6]CIBERER-U735, Barcelona, Spain, [7]Department of Urology, Bellvitge Hospital University, IDIBELL, L'Hospitalet, Barcelona, Spain, [8]Department of Pathology, Bellvitge Hospital University, IDIBELL, L'Hospitalet, Barcelona, Spain, [9]Program in Molecular Medicine and Genetics, Vall d'Hebron University Hospital, Barcelona, Spain, [10]Genetic Unit, Department of Physiology II, University of Barcelona, Barcelona, Spain and [11]Departments of Epidemiology, Internal Medicine and Human Genetics, University of Michigan, Ann Arbor, Michigan, USA

Email: Xavier Solé - x.sole@iconcologia.net; Pilar Hernández - phgutierrez@iconcologia.net; Miguel López de Heredia - mlopezheredia@idibell.org; Lluís Armengol - lluis.armengol@crg.es; Benjamín Rodríguez-Santiago - benjamin.rodriguez@upf.edu; Laia Gómez - lgomez@iconcologia.net; Christopher A Maxwell - cmaxwell@iconcologia.net; Fernando Aguiló - faguilo@csub.scs.es; Enric Condom - ecm@csub.scs.es; Jesús Abril - jabril@idibell.org; Luis Pérez-Jurado - luis.perez@upf.edu; Xavier Estivill - xavier.estivill@crg.es; Virginia Nunes - vnunes@idibell.org; Gabriel Capellá - gcapella@iconcologia.net; Stephen B Gruber - sgruber@med.umich.edu; Víctor Moreno* - v.moreno@iconcologia.net; Miguel Angel Pujana* - mapujana@iconcologia.net

* Corresponding authors

Published: 11 January 2008

BMC Genomics 2008, **9**:12 doi:10.1186/1471-2164-9-12

Received: 2 August 2007
Accepted: 11 January 2008

This article is available from: http://www.biomedcentral.com/1471-2164/9/12

Abstract

Background: Germline genetic variation is associated with the differential expression of many human genes. The phenotypic effects of this type of variation may be important when considering susceptibility to common genetic diseases. Three regions at 8q24 have recently been identified to independently confer risk of prostate cancer. Variation at 8q24 has also recently been associated with risk of breast and colorectal cancer. However, none of the risk variants map at or relatively close to known genes, with *c-MYC* mapping a few hundred kilobases distally.

Results: This study identifies cis-regulators of germline *c-MYC* expression in immortalized lymphocytes of HapMap individuals. Quantitative analysis of *c-MYC* expression in normal prostate tissues suggests an association between overexpression and variants in Region 1 of prostate cancer risk. Somatic *c-MYC* overexpression correlates with prostate cancer progression and more aggressive tumor forms, which was also a pathological variable associated with Region 1. Expression profiling analysis and modeling of transcriptional regulatory networks predicts a functional association between MYC and the prostate tumor suppressor KLF6. Analysis of MYC/Myc-driven cell transformation and tumorigenesis substantiates a model in which MYC overexpression

promotes transformation by down-regulating *KLF6*. In this model, a feedback loop through E-cadherin down-regulation causes further transactivation of *c-MYC*.

Conclusion: This study proposes that variation at putative 8q24 cis-regulator(s) of transcription can significantly alter germline *c-MYC* expression levels and, thus, contribute to prostate cancer susceptibility by down-regulating the prostate tumor suppressor *KLF6* gene.

Background

Risk of human cancer associated with genetic variation at chromosome 8q24 was first described for prostate cancer in individuals with European ancestry and in African Americans (Risk Region 1) [1,2]. This association was stronger for more aggressive tumor forms [2-4] and for earlier age at diagnosis in African Americans [1,5]. Differences in allele prevalences could account for the higher incidence of prostate cancer in particular populations such as African-Americans [1,2,5]. Subsequently, 8q24 has been associated with risk of prostate cancer by two extra independent regions [6-8] and in risk of breast and colorectal cancer by variation partially overlapping with prostate cancer risk [9-13]. In particular, Haiman *et al.*[12] first noted the existence of common risk variants for breast and colorectal cancer at 8q24. These observations suggest that multiple cancer genes may exist at 8q24 or, alternatively, that risk variants converge on a common biological mechanism [7].

In these studies risk variants did not map to known genes, with few ESTs identified in relatively close proximity. A proposed mechanism includes differences in genomic structure that would make the 8q24 region more prone to subsequent somatic amplification [14]. The *c-MYC* gene is of particular interest in this region because its ectopic expression has been shown to induce prostatic neoplasia [15-17]. Here, we analyze genetic and genomic data to provide evidence of 8q24 cis-regulator(s) of germline *c-MYC* transcription. In addition, genomic data modeling predicts a molecular mechanism linking germline *c-MYC* overexpression and prostate tumorigenesis.

Results

Genetic association scan for germline expression differences

Scanning associations between genetic variation at 8q24 and *c-MYC* gene expression levels in immortalized lymphocytes of HapMap CEU (Utah residents with ancestry from Northern and WesternEurope) and YRI (Yoruba in Ibadan Nigeria) individuals showed the existence of clusters of SNPs with nominal *P* values < 0.05 (Fig. 1). To assess clustering significance, we examined the proportion of significant SNPs in genomic windows 2- or 4-fold the average size of linkage disequilibrium blocks in CEUs or YRIs, respectively (~42 kb corresponding to ~66 SNPs in CEUs and ~36 kb corresponding to ~61 SNPs in YRIs). Twenty thousand permutations were performed to evaluate the significance of the observed clustering. One genomic region in CEUs and three regions in YRIs were identified with high density of significant SNPs (Fig. 1).

Variation at the *c-MYC* locus was observed with a trend in CEUs, which might suggest the existence of cis-regulators in the gene structural elements (blue bar in Fig. 1A). Two variants in this region (rs4645943 C and rs16902364 A) are associated with germline differential expression of *c-MYC*. The allele frequencies of these SNPs were reported to differ between prostate cancer cases and controls in different populations (i.e. 87.7% (cases) and 77.7% (controls) in Hawaiians; 96.3% (cases) and 95.1% (controls) in CEUs for rs4645943 C) [7]. This observation warrants further genetic analysis of the region with regard to prostate cancer risk.

The scan revealed a possible association between variants in Region 1 of prostate cancer risk and differential germline expression of *c-MYC* (Fig. 1B). Several significant SNPs within this region were identified: the most significant were rs7387447, rs10808558 and rs16902176 (*P* values < 0.01). The rs10808558 A allele showed an association with *c-MYC* overexpression in YRIs (expression difference of 0.23 $\log_2$ units, 95% confidence interval (CI) 0.06 – 0.41; $P = 0.007$) and this SNP is in low linkage disequilibrium (LD) with the prostate cancer risk variant rs1447295 ($r^2 = 0.19$). Overall, the scan analysis suggests the existence of 8q24 cis-regulators of germline *c-MYC* transcription in lymphocytes, partially overlapping with Region 1 of prostate cancer risk.

Expression differences in normal prostate tissues

Given the possible association of Region 1 variants with germline *c-MYC* overexpression in immortalized lymphocytes of HapMap individuals, we next examined expression differences in normal prostate tissues. For this analysis we used 54 previously characterized normal prostate tissue samples [18,19] and a real-time qRT-PCR protocol developed for prostate samples [20-22]. Genotyping the prostate cancer-associated rs1447295 variant in these samples identified six heterozygotes harboring the risk allele A (CA genotypes). No significant age differences were found between donors harboring the two different genotypes (CA versus CC; no AA homozygotes were identified). Quantitative RT-PCR study using three gene refer-

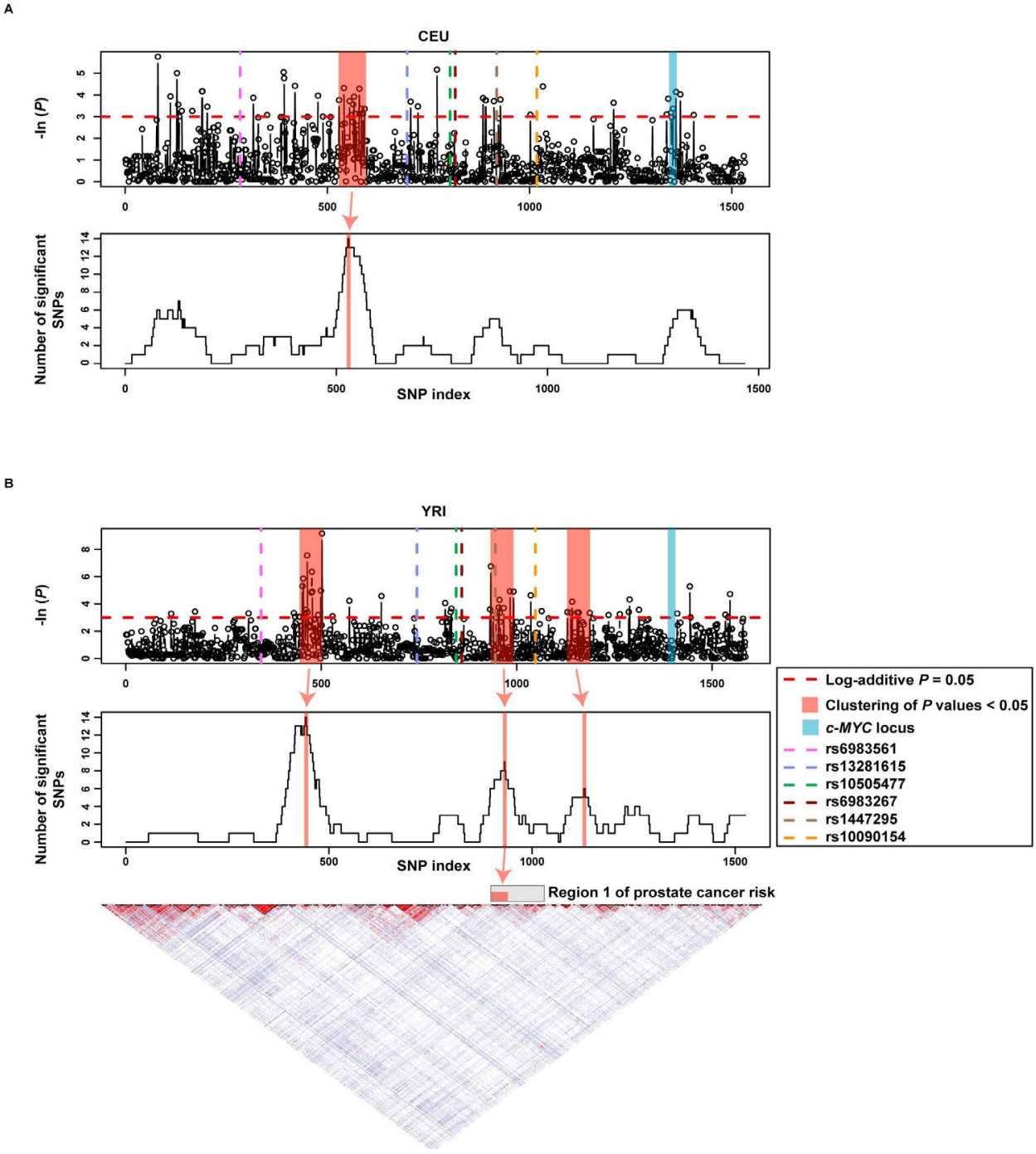


Figure 1
Genetic association scan for germline *c-MYC* differential expression in CEUs and YRIs. (A) Top panel shows results for individual SNPs and bottom panel shows results for significant SNP density in genomic windows of ~42 kb/~66 SNPs in CEUs. The red horizontal dashed line marks the nominal *P* value of 0.05. Variants associated with risk of breast [9], colorectal [10-13] or prostate [1-8, 24] cancer are marked with dashed lines as indicated in the inset. (B) Top panel shows results for individual SNPs and bottom panel shows results for significant SNP density in genomic windows of ~36 kb/~61 SNPs in YRIs. Linkage disequilibrium (*D'/LOD*) plots are shown at the bottom for YRIs. Region 1 of prostate cancer risk is shown.

ences (*18S*, *ALAS1* and *TBP*) identified significant *c-MYC* overexpression in samples harboring the risk allele relative to CC homozygotes (n = 26) (Wilcoxon rank sum test $P = 0.004$) (Fig. 2A). In addition, no evidence of allele-specific amplification in tumors arising in CA individuals was observed (not shown). These results suggest the involvement of germline *c-MYC* overexpression in prostate cancer susceptibility.

Germline copy number variants

As a possible mechanism explaining germline overexpression, we next examined copy number variants (CNVs) at the *c-MYC* locus in CEUs and YRIs, and in 322 unrelated individuals from the Spanish general population using a multiplex ligation-dependent probe amplification (MLPA) assay. This assay identified genomic gains at the *c-MYC* locus at a relatively low frequency in the Spanish general population (< 1%; 2/322) (Additional file 1). However, analysis of rs1447295 genotypes in these individuals did not reveal association with the risk allele and, importantly, none of the CEUs or YRIs showed CNVs with this assay. Therefore, a CNV including *c-MYC* does not seem to be a major contributor to the risk of prostate cancer and germline *c-MYC* overexpression associated with Region 1. Wong *et al.*[23] previously described a CNV including *c-MYC* but only with genomic losses. This observation corroborates the structural complexity of 8q24 and opens the possibility that different genomic configurations are associated with risk alleles in Region 1 or other 8q24 regions.

Gene expression analysis in prostate tumors

Since Region 1 variants were associated with earlier age at diagnosis and high Gleason scores or aggressive tumor forms [1-8,24], we examined the expression level of 8q24 genes in primary prostate tumors and their association with clinical and pathological variables. For these analyses, we used a publicly available expression data set containing different prostate cellular populations isolated using laser-capture microdissection [25].

Comparison of normal versus neoplastic samples showed differential expression of *c-MYC* (Fig. 2B). Specifically, overexpression appears in the more advanced stages of tumorigenesis such as carcinomas and hormone-refractory metastases (*t*-test $P < 10^{-3}$). Tomlins *et al.*[25] previously noted the identification of *c-MYC* in an "overexpressed in progression" signature. The *FAM84B* gene at 8q24 also shows overexpression but mainly at earlier stages ($P = 0.043$ and $P = 0.002$ for intraepithelial neoplasia and carcinomas, respectively), which suggests that *FAM84B* could also be a target of 8q24 somatic amplification. Analysis of Gleason scores showed a trend for *c-MYC* overexpression (ANOVA test $P = 0.056$) (Fig. 2C). Association between *c-MYC* overexpression and high-grade prostate tumors was previously noted by Buttyan *et al.*[26] and Fleming *et al.*[27]. These observations point to a causal relationship between somatic *c-MYC* overexpression and the more aggressive forms of prostate tumors.

Expression profiles and modeling of transcriptional regulatory networks

Transcriptional targets of MYC include many genes that were identified as conferring risk of prostate cancer and/or being somatically mutated in prostate tumors [28,29]. We sought to identify which of these genes, particularly those conferring risk of prostate cancer, could be functionally associated with *c-MYC* by examining the similarity between expression profiles using a data set containing 50 normal tissues and 52 prostate tumors [30]. This analysis revealed strong correlations between *c-MYC* and the prostate tumor suppressor *Kruppel-like factor 6 (KLF6)* gene (Fig. 3A). Correlations were positive for *c-MYC* microarray probes 1973_s_at and 37724_at (Pearson's correlation coefficient (PCC) = 0.65; $P < 10^{-13}$) and negative for 1827_s_at (PCC = -0.71; $P < 10^{-15}$). Extensive alternative splicing of the *c-MYC* mRNA could account for this difference [31].

To determine the molecular consequence of the predicted MYC-KLF6 functional association, we generated models of transcriptional regulatory networks in prostate tissues. Using the ARACNe algorithm [32] and the 102 hybridizations of Singh *et al.*[30], we identified 88 and 111 putative transcriptional targets of MYC and KLF6 in this cell type, respectively (Fig. 3B). The intersection of these two sets contains 25 genes, which is a much larger number of genes than randomly expected using simulations of equivalent gene sets (empirical $P < 0.001$). Importantly, 16 of these genes contain MYC binding sites at their promoters based on TRANSFAC (eukaryotes transcription factors database) matrices [33]. In addition, many known MYC targets [29] were also identified: 22 out of 88 (25%) and 23 out of 111 (20%) of the MYC and KLF6 predicted transcriptional targets, respectively (Fig. 3B). Notably, *c-MYC* and *KLF6* were also directly connected and the *KLF6* promoter contains three predicted binding sites for MYC (not shown). A 5-gene recurrence predictor of prostate cancer [34] contains *KLF6*, three common ARACNe-based predictions between MYC and KLF6 (*FOS*, *JUNB* and *ZFP36*), and *PPFIA3*, which is functionally related to another predicted target of KLF6 (*PPFIBP2*) (Fig. 3B). These observations further support the role of *KLF6*, *c-MYC* and the ARACNe-based predictions in prostate tumorigenesis.

Comparison of ARACNe-based predictions with the Tomlins *et al.* data set [25] identified 13 of the 88 predicted MYC transcriptional targets differentially expressed between normal prostate tissues and androgen-independent metastases (FDR-adjusted *P* values < 0.05). In addi-

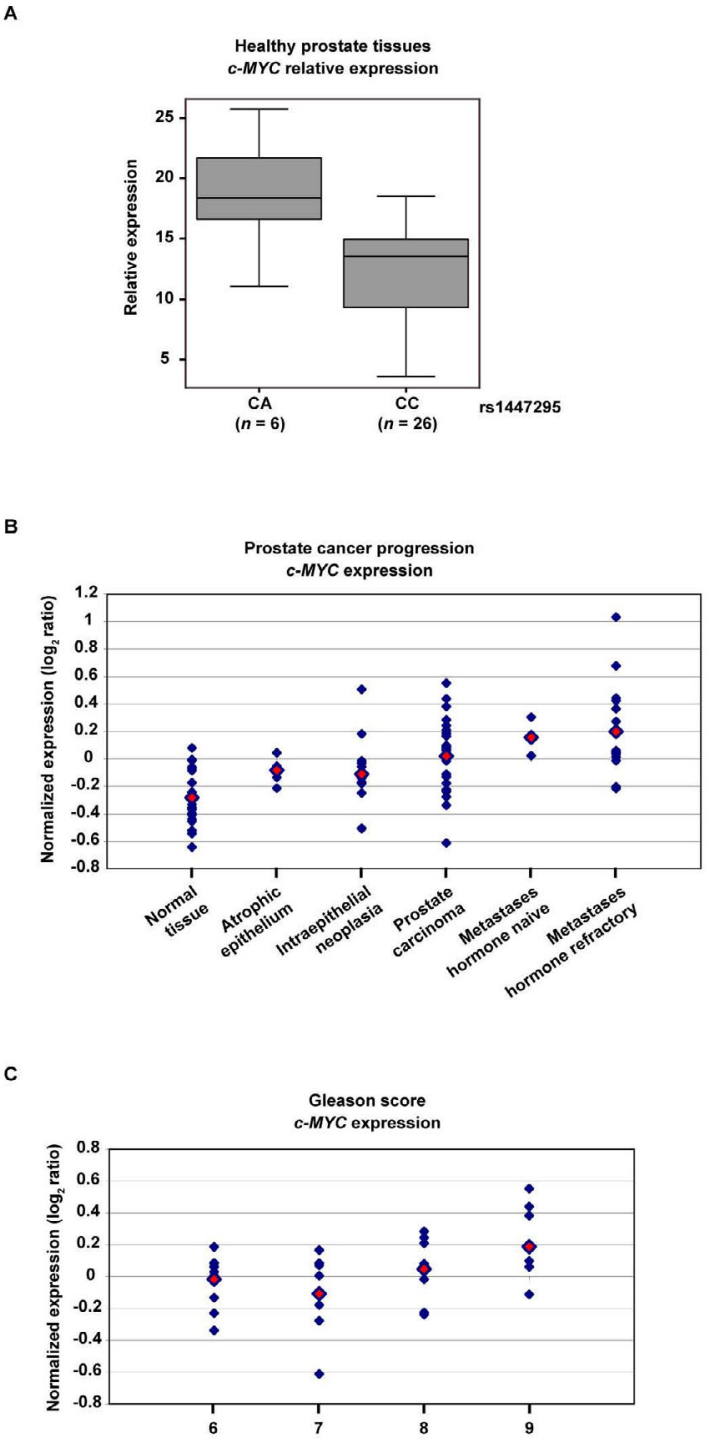


Figure 2
Analysis of *c-MYC* expression in normal and prostate cancer tissues. (A) Relative expression differences of *c-MYC* calculated using three gene references with the following formula: $R = F_{c\text{-}MYC} - (F_{TBP} - F_{ALAS1})$ where $F_{gene\ i} = Ct_{gene\ i} - Ct_{18\ S}$. (B) *c-MYC* expression in prostate cancer progression. Mean expression values are marked by a red solid rhombus. (C) *c-MYC* expression association study with Gleason scores.

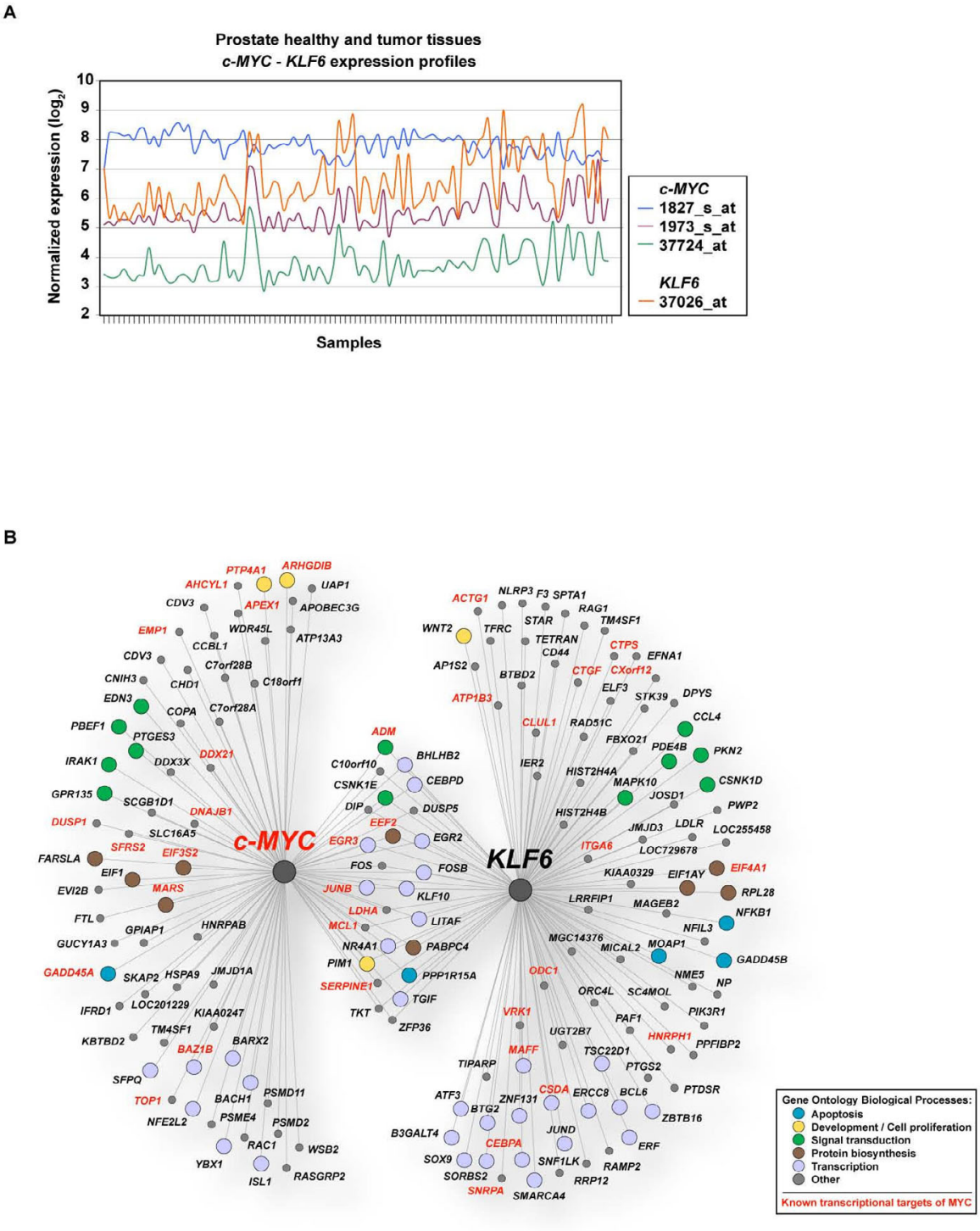


Figure 3
Expression profiling and modeling of transcriptional regulatory networks. (A) Transcriptional profiles of *c-MYC* and *KLF6* in prostate tissues [30] using U95A Affymetrix probes shown in the inset. (B) Integrated transcriptional regulatory networks of MYC and KLF6. Gene function assignment based on GO term annotations and known MYC transcription targets are shown as indicated in the inset.

tion, 20 of the predicted targets were found to be differentially expressed between normal prostate tissues and adenocarcinomas. Notably, ~46–40% of these genes (6/13 and 8/20) were also predicted to be direct transcriptional targets of KLF6 by the ARACNe algorithm, which endorses the putative functional association between MYC and KLF6.

Analysis of MYC/Myc-driven cellular transformation and tumorigenesis

To evaluate the functional significance of the predicted shared MYC/KLF6 transcriptional targets, we examined expression data derived from a model of MYC-driven cellular transformation of quiescent human mammary epithelial cells and from MMTV-Myc-driven mammary tumors in mice [35,36]. Of the 25 predicted common targets, 16 (64%) were found to be differentially expressed in cell transformation of quiescent human mammary epithelial cells (Fig. 4A). This proportion is ~2-fold higher than expected by chance taking into account all genes examined in the microarray platform (χ^2-test $P = 0.004$), which substantiates the identification of true MYC targets. Moreover, 11 of the 16 genes contain MYC binding sites at their promoters. Importantly, *KLF6* was also identified and showed strong down-regulation in this model (*t*-test *P* values < 10^{-3}) (Fig. 4A).

Analysis of MMTV-*Myc*-driven mammary tumors in mice showed consistent results with the analysis of quiescent human mammary epithelial cells. Twelve differentially expressed genes were detected, eight of which coincided with the human genes mentioned above (Fig. 4B). Genes that did not overlap between the two studies showed similar trends, for example the human *TGIF* showed a trend for down-regulation ($P = 0.067$) while it was identified as significant in the study of mice tumors ($P = 0.007$). Importantly, this analysis also revealed *Klf6* down-regulation ($P = 0.003$) (Fig. 4B). Overall, the discovery of *KLF6*/*Klf6* down-regulation in two different models of MYC/Myc-driven cell transformation supports the hypothesis that *c-MYC* germline overexpression could act as a risk factor for prostate cancer by converging on a molecular mechanism such as the functional inactivation of the *KLF6* gene or gene product.

Using the MYC/Myc-driven cell transformation models, we next examined the differential expression of known KLF6 transcriptional targets of relevance to epithelial cancers, E-cadherin (*CDH1* gene) [37] and p21 (*CDKN1A*) [38]. This analysis revealed strong down-regulation of *CDH1* in the transformation of quiescent human mammary epithelial cells (*P* values < 10^{-5}) and a trend in the model of Myc-driven mice tumorigenesis ($P = 0.088$). No significant differences were appreciable for *CDKN1A* or *Cdkn1a*. These observations suggest that KLF6 down-regulation mediated by germline MYC overexpression could promote epithelial neoplasia by down-regulating E-cadherin.

Discussion

Combined analysis of genetic and expression data facilitates the identification of transcriptional regulators acting in any part of the genome [39,40]. Examination of different ethnic groups reinforces the identification of these regulators but also reveals differences between populations [41,42]. Due to their functional and structural complexity, transcriptional regulators are largely undercharacterized. However, it is thought that their genetic variability may be relevant when considering susceptibility to common diseases. Specifically, their causal relationship to cancer is almost unknown since most genetic analyses have been focused on coding regions. Insights into differential germline gene expression and tumorigenesis have been gained mainly from mice models, such as the overexpression of the *RAS* family of genes [43], *Mad2* [44] or *c-MYC* [45,46].

This study analyzed the hypothesis that variation at 8q24 cis-regulator(s) of transcription could significantly alter germline *c-MYC* expression levels and, thus, contribute to cancer susceptibility. Although the genetic scanning analysis performed is susceptible to false positives, the existence of true cis-regulator(s) is suggested by the identification of clusters of significant SNPs. Although larger sample series are required to draw definitive conclusions, the quantitative analysis of geneexpression in normal prostate tissues supports the model of *c-MYC* overexpression associated with Region 1 of prostate cancer risk. Tissue-specific cis-regulator(s) that correlate with additional cancer risk regions at 8q24 may also exist. In a recent study it was noted that tissue specificity is a critical factor in the transcriptional responsiveness of MYC targets [47].

The 8q24 region appears amplified in up to 50% of prostate tumors and *c-MYC* is thought to be the primary target of these amplifications since it is overexpressed in prostate hyperplasia and neoplasia [25]. Ectopic overexpression of *c-MYC*/*c-Myc* is sufficient to immortalize human prostate epithelial cells [17] and has been shown to generate human-like prostate tumors in mice [16]. In addition, *c-MYC* overexpression in prostate cancer cells enables androgen-independent growth [48]. These observations lead to suggestions of a dual role for *c-MYC* in prostate cancer. At early stages it would promote proliferation while at later stages it would facilitate androgen-independent growth [17]. Our study further proposes that germline *c-MYC* overexpression may promote cellular transformation of the normal epithelium and, by extension, risk of prostate cancer by down-regulating the prostate tumor suppressor *KLF6* gene. This model is

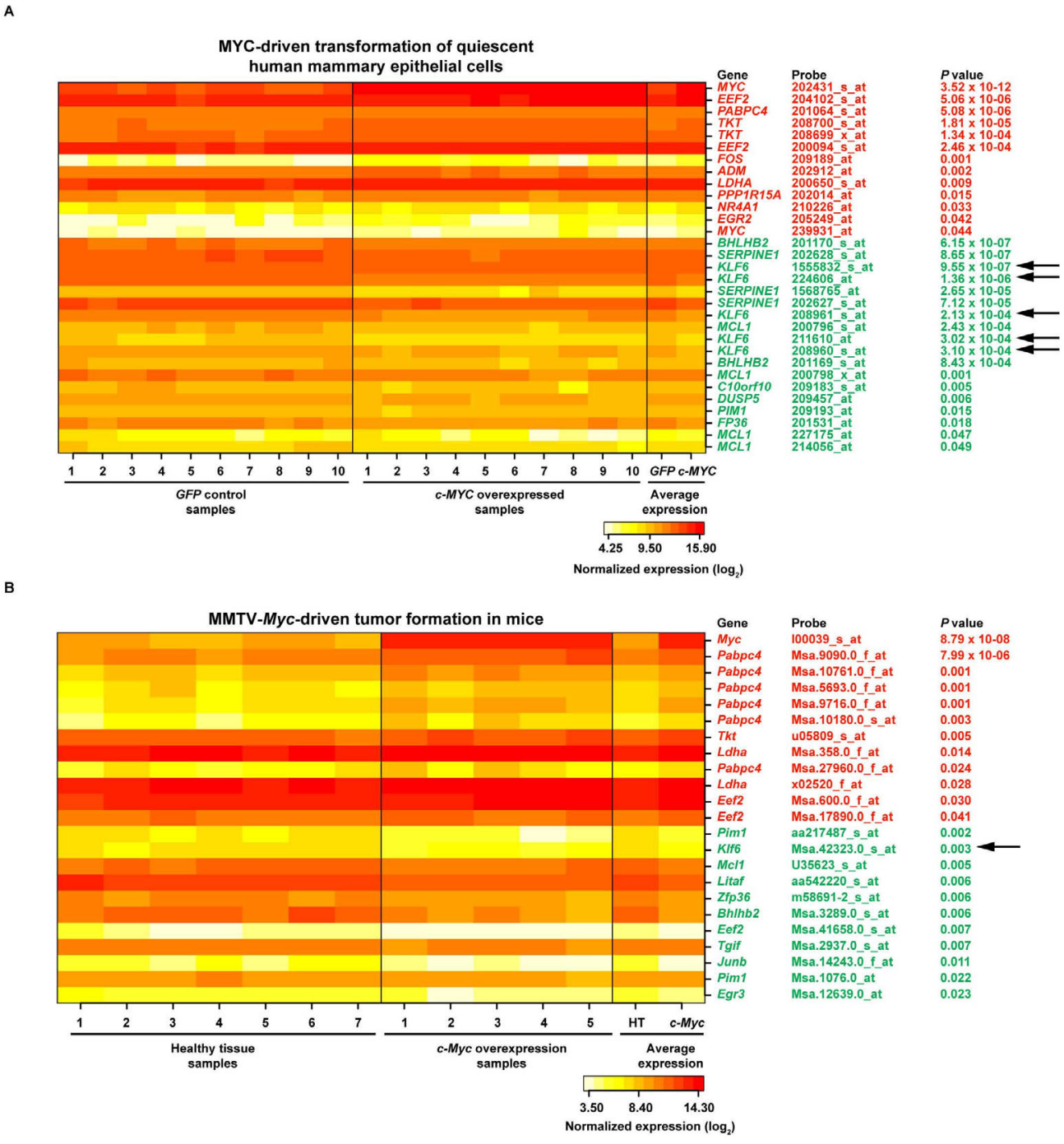


Figure 4
Expression analysis of predicted MYC/KLF6 transcriptional targets in MYC/Myc-driven cell transformation and tumorigenesis. (A) Results of the analysis of quiescent human mammary epithelial cells [36]. (B) Results of the analysis of MMTV-Myc-driven tumors in mice [35]. Genes (red, up-regulated; green, down-regulated), corresponding microarray probes and two-tailed *t*-test *P* values are shown.

hypothetical and mainly based on the application of the ARACNe algorithm, which achieves a reasonable tradeoff between true- and false-positive rates by eliminating the majority of indirect interactions inferred from gene co-expression [49,50]. Experimental corroboration of the predictions generated in this study is therefore needed, particularly in prostate tissues or cell lines.

The *KLF6* gene is inactivated in prostate cancer by loss of heterozygosity and/or by somatic mutations identified in tumors, cell lines and xenografts [51]. Recent evidence has extended the role of *KLF6* inactivation to several other neoplastic processes as esophageal carcinomas [52], glioblastomas [53], head and neck squamous cell carcinomas [54], hepatocellular carcinomas [55], non-small cell lung cancer [56], ovarian carcinomas [57] and particularly, with regard to 8q24 risk variants, to colorectal cancer [58]. A key KLF6 transcriptional target for epithelial neoplasia is E-cadherin (*CDH1* gene), which is a suppressor of cellular invasion [37]. KLF6 directly transactivates the *CDH1* promoter resulting in increased levels of its gene product [37]. *CDH1* is genetically inactivated in many human cancers and shows reduced or absent expression in approximately 50% of prostate tumors [59], playing a critical role in the transition from a noninvasive to an invasive phenotype [60]. Notably, it has recently been proposed that EphB receptors act as tumor suppressors of colorectal cancer, and possibly breast and prostate cancer, through an E-cadherin-mediated mechanism that compartmentalizes tumor cells in the initial stages of tumorigenesis [61]. Loss of E-cadherin can result in β-catenin nuclear localization and, as a result, the up-regulation of LEF/TCF-mediated transcriptional targets such as *c-MYC*[62]. Overall, our study suggests the existence of a transcriptional regulatory circuit that is perturbed in human cancer and which begins with the germline overexpression of *c-MYC*, causing down-regulation of *KLF6* which then reduces the transactivation of *CDH1*, which in turns feeds *c-MYC* expression through β-catenin and LEF/TCF transcriptional complex activation.

Variants at 8q24 have been associated with risk of prostate, breast and colorectal cancer [1-13,24,63]. Although there are different blocks of linkage disequilibrium that harbor risk variants, cancer clustering might suggests the existence of a common molecular mechanism of susceptibility. Expression analyses in normal prostate, breast and colorectal tissues and examination of association with genotypes are needed to determine the convergence on a common mechanism. Nonetheless, tumor tissue specificity may show dependences on specific, although not fully understood, mechanisms of neoplasia. The ectopic overexpression of *MYC*/*Myc* in specific cell types of mice promotes breast or prostate tumorigenesis [16,45,64], while widespread expression produces different types of tumors but with preferential appearance of specific epithelial and non-epithelial origins [46]. Overexpression of *c-MYC* also constitutes an early event after loss of the *APC* tumor suppressor gene that initiates colorectal cancer [62,65]. In addition, recent evidence shows that loss of heterozygosity at the *KLF6* locus contributes to the transition from the compartmentalized carcinoma to the invasive carcinoma, specifically in sporadic colorectal cancer [66,67], which might suggest a link with the mechanism of tumor-cells compartmentalization in the initial stages of tumorigenesis mediated by E-cadherin [61]. Although the predictions generated in this study should be treated with a degree of caution, these observations would agree with the hypothesis of a cancer susceptibility mechanism mediated by *c-MYC* germlineoverexpression.

Conclusion

This study proposes that variation at putative 8q24 cis-regulator(s) of transcription can significantly alter germline *c-MYC* expression levels and, thus, contribute to prostate cancer susceptibility by down-regulating the prostate tumor suppressor *KLF6* gene. We propose a transcriptional regulatory model perturbed in human cancer with a feedback loop for *c-MYC*.

Methods

Genetic association analysis

We analyzed HapMap genotypes and paired expression data recently made available for immortalized lymphocytes from four ethnic groups and including 210 independent individuals in total (60 Utah residents with ancestry from northern and western Europe; 45 Han Chinese in Beijing; 45 Japanese in Tokyo; and 60 Yoruba in Ibadan Nigeria; Gene Expression Omnibus (GEO) record GSE6536) [42]. Transcriptional differences were scanned between the 128 and 129 Mb of chromosome 8, corresponding to ~1,530 SNPs (NCBI build 35). Scans were performed in R with the SNPassoc package [68]. The log-additive effects of alleles were examined. Association of genotypes with the variable response (gene expression level) was calculated by fitting linear equations and P values obtained by assessing the change in deviance against the null model. Association analysis between genotypes, downloaded from the HapMap data release 21a, and gene expression levels were performed using the web-software SNPStats [69]. The *D'*/*LOD* plots were generated using the Haploview software [70].

Microarray gene expression analysis

Using the HapMap lymphocyte expression data [42] and the prostate cancer data of Tomlins *et al.*[25], matrix series were downloaded from GEO references GSE6536 and GSE6099, respectively. Using the Singh *et al.*[30] raw data, background correction, normalization and averaging of expression values were performed with the robust multi-

array average (RMA) algorithm. ARACNe Java [49,50] was used to model the gene expression regulatory networks of *c-MYC* and *KLF6*. In this analysis, data processing inequality (DPI) tolerance was set to 0.20 and the mutual information (MI) threshold was 0.05. Normalized data sets of MYC/Myc-driven cellular transformation and tumorigenesis were downloaded from the GEO records GSE3151 and GSE3158 [35,36]. Gene probes were matched using the NetAffx (Affymetrix) tool and differentially expressed probes were identified by calculating two-tailed *t*-test *P* values.

Genotyping and quantitative RT-PCR analyses

Prostate tissue specimens were collected through the Tumor Bank of the Bellvitge University Hospital and the Catalan Institute of Oncology. Genotyping of rs1447295 was performed by direct sequencing of PCR products of genomic DNA using the following forward and reverse primers, respectively: 5'-GAGTTGCACGCCAGACACTA-3' and 5'-TTTCCCATACCCCATTCTGA-3'. Quantitative RT-PCR analysis of *c-MYC* was performed using a protocol previously developed with the LightCycler™ DNA Master SYBR Green I Kit (Roche Applied Sciences) [20-22] and *c-MYC* primers 5'-CAGCTGCTTAGACGCTGGATT-3' and 5'-GTAGAAATACGGCTGCACCGA-3', and *TBP* primers 5'-GAACCACGGCACTGATTTTC-3' and 5'-CACAGCTCCCCACCATATTC-3'. Relative expression differences were calculated using three gene references (*18S*, *ALAS1* and *TBP*) with the following formula: $R = F_{c\text{-}MYC} - (F_{TBP} - F_{ALAS1})$ where $F_{gene\ i} = Ct_{gene\ i} - Ct_{18\ S}$.

Copy number variant analysis

MLPA assays were performed following the conventional protocol with 150 ng of DNA, overnight ligation and 32 cycles of PCR. Probes for *c-MYC* were 5'-GGGTTCCCTAAGGGTTGGAGGAGGAAC-GAGCTAAAACGGAGCT-3' and 5'P-TTTTTGCCCTGCGTGACCAGATCCTCTAGATT-GGATCTTGCTGGCAC-3'.

Authors' contributions

XS participated in the study design, compiled and analyzed the HapMap data, performed the association analysis and the modeling of transcriptional regulatory networks. PH compiled and analyzed the prostate cancer expression data sets and performed the modeling of transcriptional regulatory networks. FA and EC obtained the tissue specimens. MLH, JA and VN performed the quantitative expression analysis. LA, BRS, LG, LPJ and XE performed the copy number variants analysis. CAM, GC and SBG participated in scientific discussions and helped with the overall interpretation of the data. XS, VM and MAP conceived and designed the study. MAP wrote the original and final versions of the manuscript. All authors read and approved the final version of the manuscript.

Additional material

Additional file 1

MLPA analysis of several cancer loci including c-MYC. *Germline genomic gain at* c-MYC *was identified in a sample (bottom) by comparing relative peak intensities.*

Click here for file

[http://www.biomedcentral.com/content/supplementary/1471-2164-9-12-S1.PNG]

Acknowledgements

The authors are indebted to all those who provided publicly available raw data used in this contribution. We thank Laura González for technical assistance and to three anonymous reviewers for their helpful criticism. Tissue samples were collected through the Tumor Bank of the Bellvitge University Hospital and the Catalan Institute of Oncology, supported by the Tumor Bank Program and the RTICCC C03/10. This work was also supported by the Catalan Institute of Oncology, the "la Caixa" Foundation (BM05-254-00), the ISCIII (FIS-PI06/0545, RCESP-C03/09 and RTICCC-C03/10) and the Spanish Ministry of Education and Science (SAF-2003/5821 and SAF-2005/00166). CAM is supported by a Beatriu de Pinós fellowship from the Agència de Gestió d'Ajuts Universitaris i de Recerca. MAP is a Ramón y Cajal Researcher with the Spanish Ministry of Education and Science.

References

1. Freedman ML, Haiman CA, Patterson N, McDonald GJ, Tandon A, Waliszewska A, Penney K, Steen RG, Ardlie K, John EM, Oakley-Girvan I, Whittemore AS, Cooney KA, Ingles SA, Altshuler D, Henderson BE, Reich D: **Admixture mapping identifies 8q24 as a prostate cancer risk locus in African-American men.** *Proc Natl Acad Sci U S A* 2006, **103(38):**14068-14073.
2. Amundadottir LT, Sulem P, Gudmundsson J, Helgason A, Baker A, Agnarsson BA, Sigurdsson A, Benediktsdottir KR, Cazier JB, Sainz J, Jakobsdottir M, Kostic J, Magnusdottir DN, Ghosh S, Agnarsson K, Birgisdottir B, Le Roux L, Olafsdottir A, Blondal T, Andresdottir M, Gretarsdottir OS, Bergthorsson JT, Gudbjartsson D, Gylfason A, Thorleifsson G, Manolescu A, Kristjansson K, Geirsson G, Isaksson H, Douglas J, Johansson JE, Balter K, Wiklund F, Montie JE, Yu X, Suarez BK, Ober C, Cooney KA, Gronberg H, Catalona WJ, Einarsson GV, Barkardottir RB, Gulcher JR, Kong A, Thorsteinsdottir U, Stefansson K: **A common variant associated with prostate cancer in European and African populations.** *Nat Genet* 2006, **38(6):**652-658.
3. Suuriniemi M, Agalliu I, Schaid DJ, Johanneson B, McDonnell SK, Iwasaki L, Stanford JL, Ostrander EA: **Confirmation of a positive association between prostate cancer risk and a locus at chromosome 8q24.** *Cancer Epidemiol Biomarkers Prev* 2007, **16(4):**809-814.
4. Wang L, McDonnell SK, Slusser JP, Hebbring SJ, Cunningham JM, Jacobsen SJ, Cerhan JR, Blute ML, Schaid DJ, Thibodeau SN: **Two common chromosome 8q24 variants are associated with increased risk for prostate cancer.** *Cancer Res* 2007, **67(7):**2944-2950.
5. Schumacher FR, Feigelson HS, Cox DG, Haiman CA, Albanes D, Buring J, Calle EE, Chanock SJ, Colditz GA, Diver WR, Dunning AM, Freedman ML, Gaziano JM, Giovannucci E, Hankinson SE, Hayes RB, Henderson BE, Hoover RN, Kaaks R, Key T, Kolonel LN, Kraft P, Le Marchand L, Ma J, Pike MC, Riboli E, Stampfer MJ, Stram DO, Thomas G, Thun MJ, Travis R, Virtamo J, Andriole G, Gelmann E, Willett WC, Hunter DJ: **A common 8q24 variant in prostate and breast cancer from a large nested case-control study.** *Cancer Res* 2007, **67(7):**2951-2956.
6. Gudmundsson J, Sulem P, Manolescu A, Amundadottir LT, Gudbjartsson D, Helgason A, Rafnar T, Bergthorsson JT, Agnarsson BA, Baker A, Sigurdsson A, Benediktsdottir KR, Jakobsdottir M, Xu J, Blondal T,

Kostic J, Sun J, Ghosh S, Stacey SN, Mouy M, Saemundsdottir J, Backman VM, Kristjansson K, Tres A, Partin AW, Albers-Akkers MT, Godino-Ivan Marcos J, Walsh PC, Swinkels DW, Navarrete S, Isaacs SD, Aben KK, Graif T, Cashy J, Ruiz-Echarri M, Wiley KE, Suarez BK, Witjes JA, Frigge M, Ober C, Jonsson E, Einarsson GV, Mayordomo JI, Kiemeney LA, Isaacs WB, Catalona WJ, Barkardottir RB, Gulcher JR, Thorsteinsdottir U, Kong A, Stefansson K: **Genome-wide association study identifies a second prostate cancer susceptibility variant at 8q24.** *Nat Genet* 2007, **39(5):**631-637.

7. Haiman CA, Patterson N, Freedman ML, Myers SR, Pike MC, Waliszewska A, Neubauer J, Tandon A, Schirmer C, McDonald GJ, Greenway SC, Stram DO, Le Marchand L, Kolonel LN, Frasco M, Wong D, Pooler LC, Ardlie K, Oakley-Girvan I, Whittemore AS, Cooney KA, John EM, Ingles SA, Altshuler D, Henderson BE, Reich D: **Multiple regions within 8q24 independently affect risk for prostate cancer.** *Nat Genet* 2007, **39(5):**638-644.
8. Yeager M, Orr N, Hayes RB, Jacobs KB, Kraft P, Wacholder S, Minichiello MJ, Fearnhead P, Yu K, Chatterjee N, Wang Z, Welch R, Staats BJ, Calle EE, Feigelson HS, Thun MJ, Rodriguez C, Albanes D, Virtamo J, Weinstein S, Schumacher FR, Giovannucci E, Willett WC, Cancel-Tassin G, Cussenot O, Valeri A, Andriole GL, Gelmann EP, Tucker M, Gerhard DS, Fraumeni JF Jr., Hoover R, Hunter DJ, Chanock SJ, Thomas G: **Genome-wide association study of prostate cancer identifies a second risk locus at 8q24.** *Nat Genet* 2007, **39(5):**645-649.
9. Easton DF, Pooley KA, Dunning AM, Pharoah PDP, Thompson D, Ballinger DG, Struewing JP, Morrison J, Field H, Luben R, Wareham N, Ahmed S, Healey CS, Bowman R, Meyer KB, Haiman CA, Kolonel LK, Henderson BE, Le Marchand L, Brennan P, Sangrajrang S, Gaborieau V, Odefrey F, Shen CY, Wu PE, Wang HC, Eccles D, Evans DG, Peto J, Fletcher O, Johnson N, Seal S, Stratton MR, Rahman N, Chenevix-Trench G, Bojesen SE, Nordestgaard BG, Axelsson CK, Garcia-Closas M, Brinton L, Chanock S, Lissowska J, Peplonska B, Nevanlinna H, Fagerholm R, Eerola H, Kang D, Yoo KY, Noh DY, Ahn SH, Hunter DJ, Hankinson SE, Cox DG, Hall P, Wedren S, Liu J, Low YL, Bogdanova N, Schurmann P, Dork T, Tollenaar RAEM, Jacobi CE, Devilee P, Klijn JGM, Sigurdson AJ, Doody MM, Alexander BH, Zhang J, Cox A, Brock IW, MacPherson G, Reed MWR, Couch FJ, Goode EL, Olson JE, Meijers-Heijboer H, van den Ouweland A, Uitterlinden A, Rivadeneira F, Milne RL, Ribas G, Gonzalez-Neira A, Benitez J, Hopper JL, McCredie M, Southey M, Giles GG, Schroen C, Justenhoven C, Brauch H, Hamann U, Ko YD, Spurdle AB, Beesley J, Chen X, Mannermaa A, Kosma VM, Kataja V, Hartikainen J, Day NE, Cox DR, Ponder BAJ: **Genome-wide association study identifies novel breast cancer susceptibility loci.** *Nature* 2007, **447(7148):**1087-1093.
10. Tomlinson I, Webb E, Carvajal-Carmona L, Broderick P, Kemp Z, Spain S, Penegar S, Chandler I, Gorman M, Wood W, Barclay E, Lubbe S, Martin L, Sellick G, Jaeger E, Hubner R, Wild R, Rowan A, Fielding S, Howarth K, Silver A, Atkin W, Muir K, Logan R, Kerr D, Johnstone E, Sieber O, Gray R, Thomas H, Peto J, Cazier JB, Houlston R: **A genome-wide association scan of tag SNPs identifies a susceptibility variant for colorectal cancer at 8q24.21.** *Nat Genet* 2007, **39:**984-988.
11. Zanke BW, Greenwood CMT, Rangrej J, Kustra R, Tenesa A, Farrington SM, Prendergast J, Olschwang S, Chiang T, Crowdy E, Ferretti V, Laflamme P, Sundararajan S, Roumy S, Olivier JF, Robidoux F, Sladek R, Montpetit A, Campbell P, Bezieau S, O'Shea AM, Zogopoulos G, Cotterchio M, Newcomb P, McLaughlin J, Younghusband B, Green R, Green J, Porteous MEM, Campbell H, Blanche H, Sahbatou M, Tubacher E, Bonaiti-Pellie C, Buecher B, Riboli E, Kury S, Chanock SJ, Potter J, Thomas G, Gallinger S, Hudson TJ, Dunlop MG: **Genome-wide association scan identifies a colorectal cancer susceptibility locus on chromosome 8q24.** *Nat Genet* 2007, **39:**989-994.
12. Haiman CA, Le Marchand L, Yamamato J, Stram DO, Sheng X, Kolonel LN, Wu AH, Reich D, Henderson BE: **A common genetic risk factor for colorectal and prostate cancer.** *Nat Genet* 2007, **39:**954-956.
13. Gruber SB, Moreno V, Rozek LS, Rennert HS, Lejbkowicz F, Bonner JD, Greenson JK, Giordano TJ, Fearon ER, Rennert G: **Genetic Variation in 8q24 Associated with Risk of Colorectal Cancer.** *Cancer Biol Ther* 2007, **6(7):**.
14. Sato K, Qian J, Slezak JM, Lieber MM, Bostwick DG, Bergstralh EJ, Jenkins RB: **Clinical significance of alterations of chromosome 8 in high-grade, advanced, nonmetastatic prostate carcinoma.** *J Natl Cancer Inst* 1999, **91(18):**1574-1580.
15. Williams K, Fernandez S, Stien X, Ishii K, Love HD, Lau YF, Roberts RL, Hayward SW: **Unopposed c-MYC expression in benign prostatic epithelium causes a cancer phenotype.** *Prostate* 2005, **63(4):**369-384.
16. Ellwood-Yen K, Graeber TG, Wongvipat J, Iruela-Arispe ML, Zhang J, Matusik R, Thomas GV, Sawyers CL: **Myc-driven murine prostate cancer shares molecular features with human prostate tumors.** *Cancer Cell* 2003, **4(3):**223-238.
17. Gil J, Kerai P, Lleonart M, Bernard D, Cigudosa JC, Peters G, Carnero A, Beach D: **Immortalization of primary human prostate epithelial cells by c-Myc.** *Cancer Res* 2005, **65(6):**2179-2185.
18. Gomez-Zaera M, Abril J, Gonzalez L, Aguilo F, Condom E, Nadal M, Nunes V: **Identification of somatic and germline mitochondrial DNA sequence variants in prostate cancer patients.** *Mutat Res* 2006, **595(1-2):**42-51.
19. Nadal M, Pera G, Pujadas J, Abril J, Gonzalez L, Aguilo F, Condom E, Gomez-Zaera M, Nunes V: **Aneuploidy of chromosome Y in prostate tumors and seminal vesicles: A possible sign of aging rather than an indicator of carcinogenesis?** *Mol Carcinog* 2007, **46(7):**543-552.
20. Linja MJ, Savinainen KJ, Saramaki OR, Tammela TL, Vessella RL, Visakorpi T: **Amplification and overexpression of androgen receptor gene in hormone-refractory prostate cancer.** *Cancer Res* 2001, **61(9):**3550-3555.
21. Savinainen KJ, Linja MJ, Saramaki OR, Tammela TL, Chang GT, Brinkmann AO, Visakorpi T: **Expression and copy number analysis of TRPS1, EIF3S3 and MYC genes in breast and prostate cancer.** *Br J Cancer* 2004, **90(5):**1041-1046.
22. Ohl F, Jung M, Xu C, Stephan C, Rabien A, Burkhardt M, Nitsche A, Kristiansen G, Loening SA, Radonic A, Jung K: **Gene expression studies in prostate cancer tissue: which reference gene should be selected for normalization?** *J Mol Med* 2005, **83(12):**1014-1024.
23. Wong KK, deLeeuw RJ, Dosanjh NS, Kimm LR, Cheng Z, Horsman DE, MacAulay C, Ng RT, Brown CJ, Eichler EE, Lam WL: **A comprehensive analysis of common copy-number variations in the human genome.** *Am J Hum Genet* 2007, **80(1):**91-104.
24. Severi G, Hayes VM, Padilla EJ, English DR, Southey MC, Sutherland RL, Hopper JL, Giles GG: **The common variant rs1447295 on chromosome 8q24 and prostate cancer risk: results from an Australian population-based case-control study.** *Cancer Epidemiol Biomarkers Prev* 2007, **16(3):**610-612.
25. Tomlins SA, Mehra R, Rhodes DR, Cao X, Wang L, Dhanasekaran SM, Kalyana-Sundaram S, Wei JT, Rubin MA, Pienta KJ, Shah RB, Chinnaiyan AM: **Integrative molecular concept modeling of prostate cancer progression.** *Nat Genet* 2007, **39(1):**41-51.
26. Buttyan R, Sawczuk IS, Benson MC, Siegal JD, Olsson CA: **Enhanced expression of the c-myc protooncogene in high-grade human prostate cancers.** *Prostate* 1987, **11(4):**327-337.
27. Fleming WH, Hamel A, MacDonald R, Ramsey E, Pettigrew NM, Johnston B, Dodd JG, Matusik RJ: **Expression of the c-myc protooncogene in human prostatic carcinoma and benign prostatic hyperplasia.** *Cancer Res* 1986, **46(3):**1535-1538.
28. Futreal PA, Coin L, Marshall M, Down T, Hubbard T, Wooster R, Rahman N, Stratton MR: **A census of human cancer genes.** *Nat Rev Cancer* 2004, **4(3):**177-183.
29. Zeller KI, Jegga AG, Aronow BJ, O'Donnell KA, Dang CV: **An integrated database of genes responsive to the Myc oncogenic transcription factor: identification of direct genomic targets.** *Genome Biol* 2003, **4(10):**R69.
30. Singh D, Febbo PG, Ross K, Jackson DG, Manola J, Ladd C, Tamayo P, Renshaw AA, D'Amico AV, Richie JP, Lander ES, Loda M, Kantoff PW, Golub TR, Sellers WR: **Gene expression correlates of clinical prostate cancer behavior.** *Cancer Cell* 2002, **1(2):**203-209.
31. Bodescot M, Brison O: **Characterization of new human c-myc mRNA species produced by alternative splicing.** *Gene* 1996, **174(1):**115-120.
32. 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.
33. Matys V, Fricke E, Geffers R, Gossling E, Haubrock M, Hehl R, Hornischer K, Karas D, Kel AE, Kel-Margoulis OV, Kloos DU, Land S, Lewicki-Potapov B, Michael H, Munch R, Reuter I, Rotert S, Saxel H, Scheer M, Thiele S, Wingender E: **TRANSFAC: transcriptional**

regulation, from patterns to profiles. *Nucleic Acids Res* 2003, **31(1):**374-378.
34. Glinsky GV, Glinskii AB, Stephenson AJ, Hoffman RM, Gerald WL: **Gene expression profiling predicts clinical outcome of prostate cancer.** *J Clin Invest* 2004, **113(6):**913-923.
35. Bild AH, Yao G, Chang JT, Wang Q, Potti A, Chasse D, Joshi MB, Harpole D, Lancaster JM, Berchuck A, Olson JA Jr., Marks JR, Dressman HK, West M, Nevins JR: **Oncogenic pathway signatures in human cancers as a guide to targeted therapies.** *Nature* 2006, **439(7074):**353-357.
36. Huang E, Ishida S, Pittman J, Dressman H, Bild A, Kloos M, D'Amico M, Pestell RG, West M, Nevins JR: **Gene expression phenotypic models that predict the activity of oncogenic pathways.** *Nat Genet* 2003, **34(2):**226-230.
37. DiFeo A, Narla G, Camacho-Vanegas O, Nishio H, Rose SL, Buller RE, Friedman SL, Walsh MJ, Martignetti JA: **E-cadherin is a novel transcriptional target of the KLF6 tumor suppressor.** *Oncogene* 2006, **25(44):**6026-6031.
38. Narla G, Kremer-Tal S, Matsumoto N, Zhao X, Yao S, Kelley K, Tarocchi M, Friedman SL: **In vivo regulation of p21 by the Kruppel-like factor 6 tumor-suppressor gene in mouse liver and human hepatocellular carcinoma.** *Oncogene* 2007, **26(30):**4428-4434.
39. Cheung VG, Spielman RS: **The genetics of variation in gene expression.** *Nat Genet* 2002, **32 Suppl:**522-525.
40. Buckland PR: **Allele-specific gene expression differences in humans.** *Hum Mol Genet* 2004, **13 Spec No 2:**R255-60.
41. Spielman RS, Bastone LA, Burdick JT, Morley M, Ewens WJ, Cheung VG: **Common genetic variants account for differences in gene expression among ethnic groups.** *Nat Genet* 2007, **39(2):**226-231.
42. Stranger BE, Forrest MS, Dunning M, Ingle CE, Beazley C, Thorne N, Redon R, Bird CP, de Grassi A, Lee C, Tyler-Smith C, Carter N, Scherer SW, Tavare S, Deloukas P, Hurles ME, Dermitzakis ET: **Relative impact of nucleotide and copy number variation on gene expression phenotypes.** *Science* 2007, **315(5813):**848-853.
43. Mangues R, Seidman I, Gordon JW, Pellicer A: **Overexpression of the N-ras proto-oncogene, not somatic mutational activation, associated with malignant tumors in transgenic mice.** *Oncogene* 1992, **7(10):**2073-2076.
44. Sotillo R, Hernando E, Diaz-Rodriguez E, Teruya-Feldstein J, Cordon-Cardo C, Lowe SW, Benezra R: **Mad2 overexpression promotes aneuploidy and tumorigenesis in mice.** *Cancer Cell* 2007, **11(1):**9-23.
45. Stewart TA, Pattengale PK, Leder P: **Spontaneous mammary adenocarcinomas in transgenic mice that carry and express MTV/myc fusion genes.** *Cell* 1984, **38(3):**627-637.
46. Leder A, Pattengale PK, Kuo A, Stewart TA, Leder P: **Consequences of widespread deregulation of the c-myc gene in transgenic mice: multiple neoplasms and normal development.** *Cell* 1986, **45(4):**485-495.
47. Chen Y, Blackwell TW, Chen J, Gao J, Lee AW, States DJ: **Integration of genome and chromatin structure with gene expression profiles to predict c-MYC recognition site binding and function.** *PLoS Comput Biol* 2007, **3(4):**e63.
48. Bernard D, Pourtier-Manzanedo A, Gil J, Beach DH: **Myc confers androgen-independent prostate cancer cell growth.** *J Clin Invest* 2003, **112(11):**1724-1731.
49. 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.
50. Margolin AA, Wang K, Lim WK, Kustagi M, Nemenman I, Califano A: **Reverse engineering cellular networks.** *Nat Protoc* 2006, **1(2):**662-671.
51. Narla G, Heath KE, Reeves HL, Li D, Giono LE, Kimmelman AC, Glucksman MJ, Narla J, Eng FJ, Chan AM, Ferrari AC, Martignetti JA, Friedman SL: **KLF6, a candidate tumor suppressor gene mutated in prostate cancer.** *Science* 2001, **294(5551):**2563-2566.
52. Yamashita K, Upadhyay S, Osada M, Hoque MO, Xiao Y, Mori M, Sato F, Meltzer SJ, Sidransky D: **Pharmacologic unmasking of epigenetically silenced tumor suppressor genes in esophageal squamous cell carcinoma.** *Cancer Cell* 2002, **2(6):**485-495.
53. Jeng YM, Hsu HC: **KLF6, a putative tumor suppressor gene, is mutated in astrocytic gliomas.** *Int J Cancer* 2003, **105(5):**625-629.
54. Teixeira MS, Camacho-Vanegas O, Fernandez Y, Narla G, Difeo A, Lee B, Kalir T, Friedman SL, Schlecht NF, Genden EM, Urken M, Brandwein-Gensler M, Martignetti JA: **KLF6 allelic loss is associated with tumor recurrence and markedly decreased survival in head and neck squamous cell carcinoma.** *Int J Cancer* 2007, **121(9):**1976-1983.
55. Kremer-Tal S, Reeves HL, Narla G, Thung SN, Schwartz M, Difeo A, Katz A, Bruix J, Bioulac-Sage P, Martignetti JA, Friedman SL: **Frequent inactivation of the tumor suppressor Kruppel-like factor 6 (KLF6) in hepatocellular carcinoma.** *Hepatology* 2004, **40(5):**1047-1052.
56. Ito G, Uchiyama M, Kondo M, Mori S, Usami N, Maeda O, Kawabe T, Hasegawa Y, Shimokata K, Sekido Y: **Kruppel-like factor 6 is frequently down-regulated and induces apoptosis in non-small cell lung cancer cells.** *Cancer Res* 2004, **64(11):**3838-3843.
57. DiFeo A, Narla G, Hirshfeld J, Camacho-Vanegas O, Narla J, Rose SL, Kalir T, Yao S, Levine A, Birrer MJ, Bonome T, Friedman SL, Buller RE, Martignetti JA: **Roles of KLF6 and KLF6-SV1 in ovarian cancer progression and intraperitoneal dissemination.** *Clin Cancer Res* 2006, **12(12):**3730-3739.
58. Reeves HL, Narla G, Ogunbiyi O, Haq AI, Katz A, Benzeno S, Hod E, Harpaz N, Goldberg S, Tal-Kremer S, Eng FJ, Arthur MJ, Martignetti JA, Friedman SL: **Kruppel-like factor 6 (KLF6) is a tumor-suppressor gene frequently inactivated in colorectal cancer.** *Gastroenterology* 2004, **126(4):**1090-1103.
59. Richmond PJ, Karayiannakis AJ, Nagafuchi A, Kaisary AV, Pignatelli M: **Aberrant E-cadherin and alpha-catenin expression in prostate cancer: correlation with patient survival.** *Cancer Res* 1997, **57(15):**3189-3193.
60. Isaacs WB, Bova GS, Morton RA, Bussemakers MJ, Brooks JD, Ewing CM: **Molecular biology of prostate cancer progression.** *Cancer Surv* 1995, **23:**19-32.
61. Cortina C, Palomo-Ponce S, Iglesias M, Fernandez-Masip JL, Vivancos A, Whissell G, Huma M, Peiro N, Gallego L, Jonkheer S, Davy A, Lloreta J, Sancho E, Batlle E: **EphB-ephrin-B interactions suppress colorectal cancer progression by compartmentalizing tumor cells.** *Nat Genet* 2007.
62. He TC, Sparks AB, Rago C, Hermeking H, Zawel L, da Costa LT, Morin PJ, Vogelstein B, Kinzler KW: **Identification of c-MYC as a target of the APC pathway.** *Science* 1998, **281(5382):**1509-1512.
63. Witte JS: **Multiple prostate cancer risk variants on 8q24.** *Nat Genet* 2007, **39(5):**579-580.
64. D'Cruz CM, Gunther EJ, Boxer RB, Hartman JL, Sintasath L, Moody SE, Cox JD, Ha SI, Belka GK, Golant A, Cardiff RD, Chodosh LA: **c-MYC induces mammary tumorigenesis by means of a preferred pathway involving spontaneous Kras2 mutations.** *Nat Med* 2001, **7(2):**235-239.
65. van de Wetering M, Sancho E, Verweij C, de Lau W, Oving I, Hurlstone A, van der Horn K, Batlle E, Coudreuse D, Haramis AP, Tjon-Pon-Fong M, Moerer P, van den Born M, Soete G, Pals S, Eilers M, Medema R, Clevers H: **The beta-catenin/TCF-4 complex imposes a crypt progenitor phenotype on colorectal cancer cells.** *Cell* 2002, **111(2):**241-250.
66. Mukai S, Hiyama T, Tanaka S, Yoshihara M, Arihiro K, Chayama K: **Involvement of Kruppel-like factor 6 (KLF6) mutation in the development of nonpolypoid colorectal carcinoma.** *World J Gastroenterol* 2007, **13(29):**3932-3938.
67. Miyaki M, Yamaguchi T, Iijima T, Funata N, Mori T: **Difference in the role of loss of heterozygosity at 10p15 (KLF6 locus) in colorectal carcinogenesis between sporadic and familial adenomatous polyposis and hereditary nonpolyposis colorectal cancer patients.** *Oncology* 2006, **71(1-2):**131-135.
68. Gonzalez JR, Armengol L, Sole X, Guino E, Mercader JM, Estivill X, Moreno V: **SNPassoc: an R package to perform whole genome association studies.** *Bioinformatics* 2007, **23(5):**644-645.
69. Sole X, Guino E, Valls J, Iniesta R, Moreno V: **SNPStats: a web tool for the analysis of association studies.** *Bioinformatics* 2006, **22(15):**1928-1929.
70. Barrett JC, Fry B, Maller J, Daly MJ: **Haploview: analysis and visualization of LD and haplotype maps.** *Bioinformatics* 2005, **21(2):**263-265.

7

ARTICLE 3: BIOLOGICAL CONVERGENCE OF CANCER SIGNATURES

7.1 SUMMARY

Gene expression profiling has identified cancer prognostic and predictive signatures with higher performance than conventional histopathological or clinical parameters. Consequently, signatures are being incorporated into clinical practice and will soon influence everyday decisions in oncology. However, the slight overlap in the gene identity between signatures for the same cancer type or condition raises questions about their biological and clinical implications. To clarify these issues, better understanding of the molecular properties and the detection of possible interactions underlying apparently dissimilar signatures is needed.

The aim of this study was to integrate genomics, transcriptomics and proteomics data to unveil potential relationships among cancer signatures of 24 independent studies.

Common properties at the genome level were evaluated by probing the relative enrichment in predicted transcription factor binding sites (TFBS) motifs at the promoters of signature genes. In these analyses the top-ranked motifs across several signatures were from the E2F family, which is a key regulator of cell proliferation and death processes. When analyzing experimental data from chromatin immunoprecipitation assays of TFs, the major role of E2F transcriptional programs was corroborated. Furthermore, significant over-representation of *ESR1* gene binding sites and *ESR1*-mediated transcriptional regulation was identified for most of the signatures, irrespective of their type or condition.

Using cancer representative datasets, Pearson correlation coefficients (PCC) between over-represented TFs in cancer signatures and genes associated with breast cancer prognosis or with the response to docetaxel treatment in breast cancer patients were computed. The correlations were compared to PCCs between the same TFs and genes non-differentially expressed in these conditions. As a result, higher absolute PCCs between TFs and genes associated with prognosis or treatment response were identified in all cases for genes and/or microarray probes.

Using a breast cancer dataset and the average PCC across all microarray probe pairs between any two signatures, significant co-expression was identified in approximately half of the analyses when compared to 10,000 equivalent, randomly selected gene sets. These results support the existence of functional and molecular associations between many apparently dissimilar signatures, despite the fact that the dataset used had evident technical and biological specificities. A strong correlation with genes involved in mitosis or cell death GO categories was also observed for most signatures.

Given the evidence of signatures relationship at the genome and transcriptome levels, it was hypothesized that proteins encoded by apparently dissimilar signatures could be significantly closer in the interactome network. Using a dataset consisting mainly of experimentally identified protein-protein interactions, it was seen that most cancer signatures were more closely located than expected by chance, and also close to cell death and mitosis genes.

All the previously described results were validated in two independent datasets: one of a lung metastasis signature of breast cancer and the other of a signature of response to cetuximab in metastatic CRC patients.

7.2 MAIN RESULTS

22 out of 24 signatures examined showed significant over-representation of one or more of the molecular evidences associated with the regulation of cell proliferation and death.

Significant associations were consistently observed across genomics, transcriptomics and proteomics layers, suggesting the existence of a common cancer cell phenotype. Convergence on cell proliferation and death supports the pivotal involvement of these processes in prognosis, metastasis and treatment response.

Functional and molecular associations were identified with the immune response in different cancer types and conditions that complement the contribution of cell proliferation and death.

Examination of additional, independent, cancer datasets corroborated the previously found observations.

OPEN ACCESS Freely available online

PLoS one

Biological Convergence of Cancer Signatures

Xavier Solé[1], Núria Bonifaci[1,2], Núria López-Bigas[3], Antoni Berenguer[1], Pilar Hernández[2], Oscar Reina[4], Christopher A. Maxwell[2], Helena Aguilar[2], Ander Urruticoechea[2], Silvia de Sanjosé[4], Francesc Comellas[5], Gabriel Capellá[2], Víctor Moreno[1], Miguel Angel Pujana[1,2]*

1 Bioinformatics and Biostatistics Unit, Catalan Institute of Oncology, IDIBELL, L'Hospitalet, Barcelona, Spain, **2** Translational Research Laboratory, Catalan Institute of Oncology, IDIBELL, L'Hospitalet, Barcelona, Spain, **3** Research Unit on Biomedical Informatics of IMIM/UPF, Barcelona Biomedical Research Park, Barcelona, Spain, **4** Unit of Infections and Cancer, CIBERESP, Epidemiology Research of Cancer Program, Catalan Institute of Oncology, IDIBELL, L'Hospitalet, Barcelona, Spain, **5** Department of Applied Mathematics IV, Technical University of Catalonia, Castelldefels, Barcelona, Spain

Abstract

Gene expression profiling has identified cancer prognostic and predictive signatures with superior performance to conventional histopathological or clinical parameters. Consequently, signatures are being incorporated into clinical practice and will soon influence everyday decisions in oncology. However, the slight overlap in the gene identity between signatures for the same cancer type or condition raises questions about their biological and clinical implications. To clarify these issues, better understanding of the molecular properties and possible interactions underlying apparently dissimilar signatures is needed. Here, we evaluated whether the signatures of 24 independent studies are related at the genome, transcriptome or proteome levels. Significant associations were consistently observed across these molecular layers, which suggest the existence of a common cancer cell phenotype. Convergence on cell proliferation and death supports the pivotal involvement of these processes in prognosis, metastasis and treatment response. In addition, functional and molecular associations were identified with the immune response in different cancer types and conditions that complement the contribution of cell proliferation and death. Examination of additional, independent, cancer datasets corroborated our observations. This study proposes a comprehensive strategy for interpreting cancer signatures that reveals common design principles and systems-level properties.

Citation: Solé X, Bonifaci N, López-Bigas N, Berenguer A, Hernández P, et al. (2009) Biological Convergence of Cancer Signatures. PLoS ONE 4(2): e4544. doi:10.1371/journal.pone.0004544

Editor: Gustavo Stolovitzky, IBM Thomas J. Watson Research Center, United States of America

Received October 7, 2008; **Accepted** January 16, 2009; **Published** February 20, 2009

Funding: The la Caixa Foundation grant BM 05/254 (MAP) and the Spanish Ministry of Health grants FIS 05/1006 (VM) and 06/0545 (MAP), RCESP C03/09 and RTICCC C03/10 (VM and GC). CAM is supported by a Beatriu de Pinos fellowship from the AGAUR agency of the Catalan Government, HA is supported by a postdoctoral fellowship of the Spanish Ministry of Health and MAP is a Ramon y Cajal Researcher with the Spanish Ministry of Education and Science. 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: mapujana@ico.scs.es.

Introduction

Recent years have seen the description of a large number of gene expression profiles or signatures with clinical value for the accurate prognostic or predictive characterization of cancer patients or tumors. Breast cancer is probably the paradigm of such studies, with at least three different signatures currently being tested in clinical trials and commercially available for routine clinical practice in oncology [1,2]. However, the lack of overlap in the selected genes has raised fundamental questions about their biological and clinical implications [3,4]. This situation is not unique to breast cancer prognosis, and the description of new expression profiles suggests that it is common to other cancer types or conditions_i.e. metastases and treatments [5]. Reasons to this paradox may be methodological disparities [6] and statistical constraints created by the large number of genes examined with respect to the relatively small number of samples profiled [7–9]. Importantly, a recent study by Perou and colleagues [10] established the common prognostic value of some breast cancer signatures, despite the lack of overlap in gene identities. This observation confirmed the clinical relevance of the signatures and suggested that they may efficiently capture a common tumor cell phenotype(s) [11]. This putative common phenotype for breast cancer and for other neoplasias must be defined if we are to better understand the significance of signatures.

Some of the early descriptions of signatures noted the presence of specific biological processes over-represented in the corresponding gene lists. Among these processes, individual genes involved in the cell cycle and apoptosis were highlighted (e.g. [12,13]). More recent evidence points to specific genes that are globally associated with breast cancer prognosis and related to cell proliferation among other processes or pathways [14–21]. However, it is still unclear how this evidence characterizes different molecular levels and how the levels integrate into a systems-level model containing gene and/or protein interactions for breast cancer and for human cancer in general. Here, we used an integrative approach to determine the existence of a putative common tumor cell phenotype(s) associated with different cancer types and conditions. The study identified common molecular properties and network interactions associated with cell proliferation and death, and revealed associations with the immune response. Our results highlight the importance of studying signatures from a systems-level perpective.

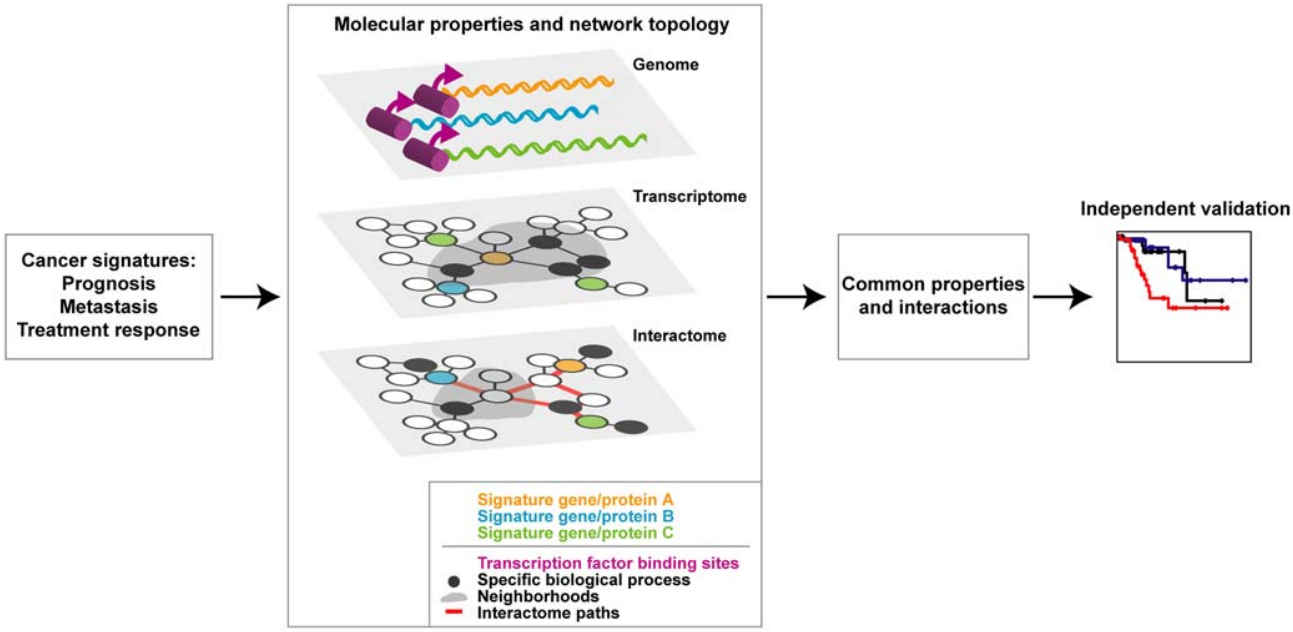


Figure 1. Integrative analysis of cancer signatures. Strategy for defining the common properties and interactions between signatures at the genome, transcriptome and proteome levels, and validation in independent datasets.
doi:10.1371/journal.pone.0004544.g001

Results

Genomic properties: E2Fs and the estrogen receptor (ER)

To identify common properties among cancer signatures we compiled the literature gene lists from 24 studies (Table S1). These represent 19 prognostic signatures, two signatures focused mainly on metastasis, and seven predictive treatment response signatures. All signatures used corresponded to validated sets of genes at the same level. We first examined the molecular properties or network topology characteristics of genes and/or proteins in these signatures at the genome, transcriptome and proteome levels. Next, the identified properties and network associations were corroborated in independent expression datasets of different cancer types and conditions (Fig. 1).

Properties at the genome level were evaluated by probing the relative enrichment in predicted transcription factor binding site motifs at the promoters of signature genes (see Methods). In these analyses the top-ranked motifs across several signatures were from the E2F family. Significant over-representation of E2F motifs was identified in ~45% (13/28) of the signatures tested, including prognostic (bladder, breast and central nervous system (CNS) cancers, and three multi-cancer signatures) and predictive signatures (docetaxel in breast tumors, EGFR tyrosine kinase inhibitors (TKIs) in lung tumors and pemetrexed in advanced solid tumors) (false discovery rate (FDR)-adjusted P values<0.05) (Fig. 2*A*). In contrast, only one signature (the immune response prognostic signature in estrogen receptor (ER)-negative breast cancer [22]) showed under-representation of E2F motifs. This observation will be discussed in the following sections.

To evaluate motif predictions in the promoter sequences of signature genes, we examined experimental data from chromatin immunoprecipitation assays of transcription factors [23,24]. This analysis corroborated the major role of E2F transcriptional programs. Approximately 65% of signatures showed significant over-representation of E2F1-AP2 and/or E2F4 binding sites (Fig. 2*B*). The strongest over-representations were detected in prognostic_particularly breast cancer_and predictive treatment response signatures for E2F1-AP2 sites. Nevertheless, specificities were also suggested for the immune response, which showed under-representation of E2F1-AP2, and for predictive signatures that did not show differential representation of E2F4 in any case.

The E2Fs are key regulators of cell proliferation and death [25,26], and common deregulation of E2F-mediated transcriptional programs is a hallmark of cancer transcriptomes [27]. The link with the potential for cell proliferation was further evaluated by examining transcripts with periodic expression through the cell cycle [28], which indicates a direct or indirect role in phase(s) of cell division, and by analyzing ER functional genomic data [29]. Significant over-representation of periodically expressed genes was observed in ~45% of the signatures, most of which were prognostic signatures for different cancer types (Fig. 2*C*). Detailed examination of cell cycle phases showed specific over-representation of genes with an expression peak at G2 and G2/M, which is in agreement with their role in cell division (data not shown). In addition, consistent with the link between cell proliferation and the ER signaling pathway [30], significant over-representation of ER binding sites and/or ER-mediated transcriptional regulation was identified in most of the signatures (~90%), irrespective of their type or condition (Fig. 2*D*). This high overlap with ER regulation probably reflects an strong association with cell proliferation beyond cancer hormone-dependencies.

Overall, all except two of the signatures examined here showed significant over-representation of one or more of the molecular evidences associated with the regulation of cell proliferation and death. The exceptions were the immune response signature, which may reflect the involvement of different biological processes, and the B-cell lymphoma prognosis signature, which may be explained by the statistical power needed to detect differences in the smallest gene set examined ($n = 19$). Similarities for these signatures at additional molecular levels will be presented in the following sections.

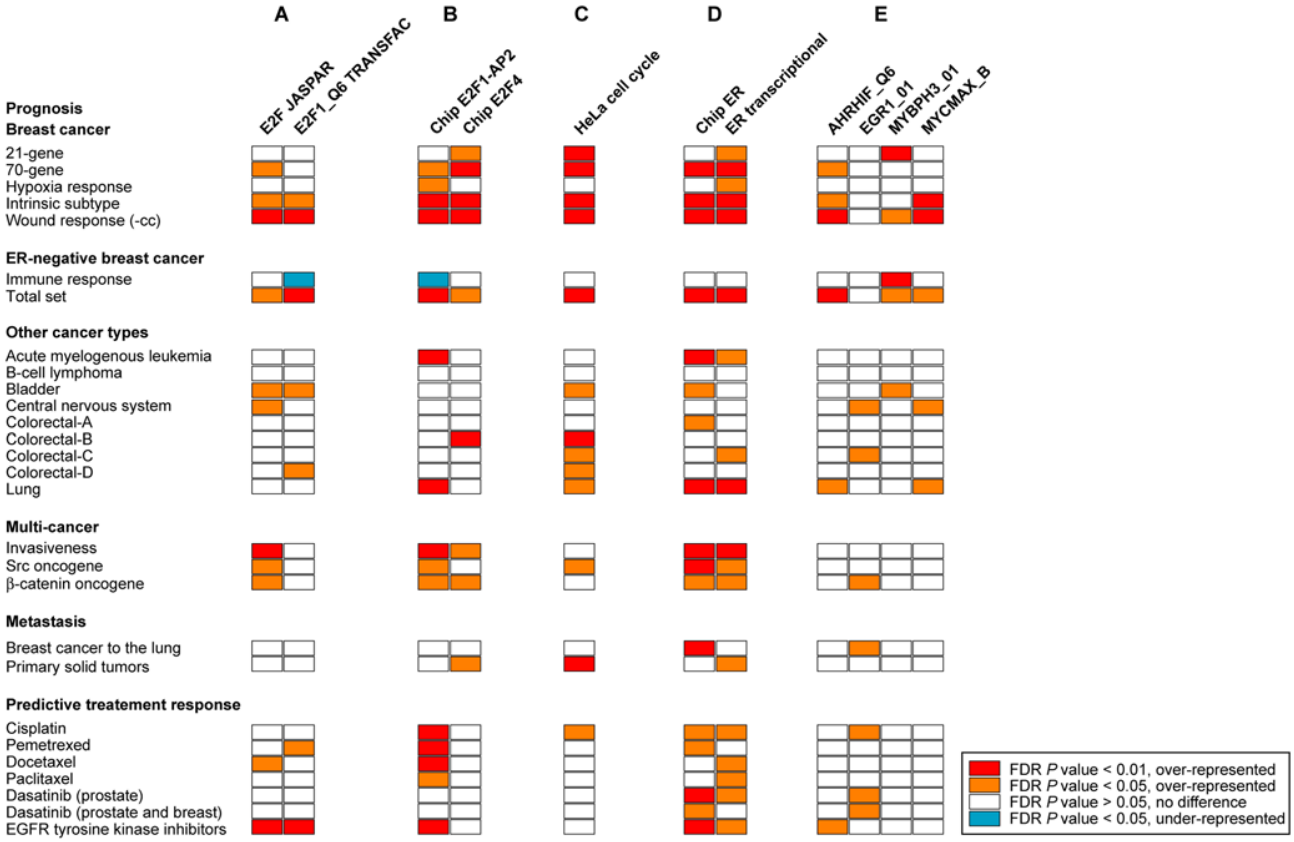
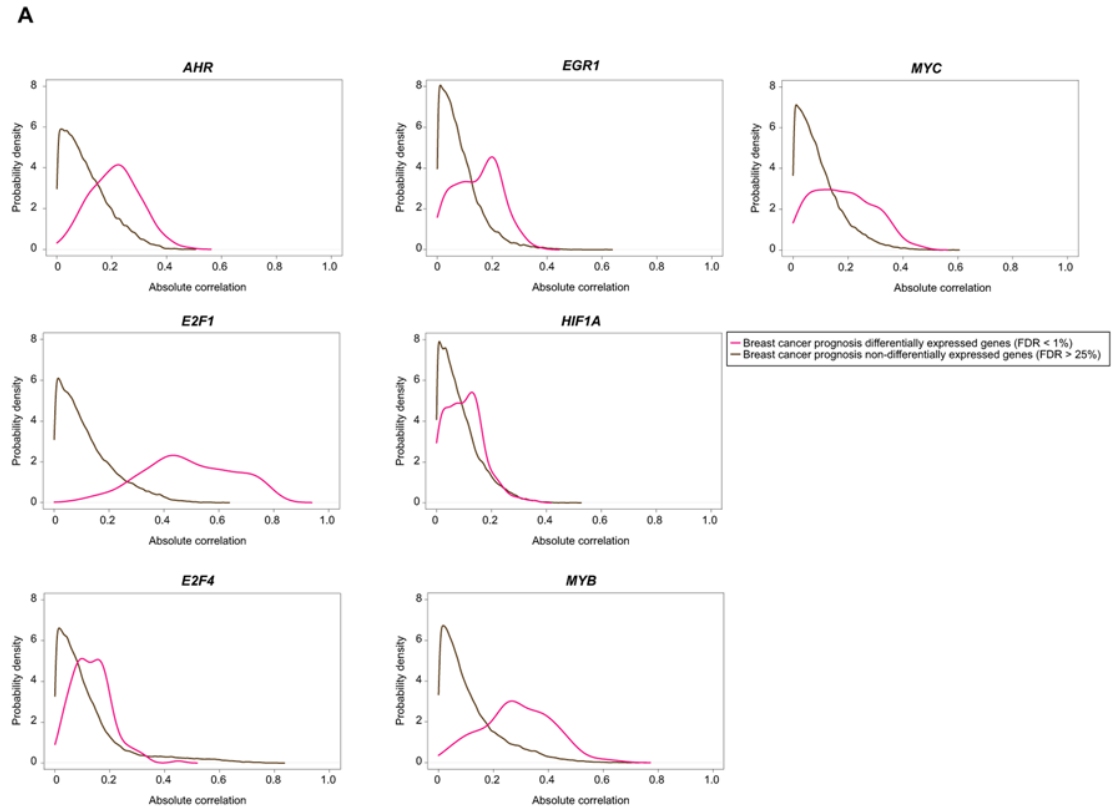


Figure 2. Genomic and transcriptomic properties of cancer signatures associated with the potential for cell proliferation and repressed cell death. *A*, representation of E2F motifs based on JASPAR and TRANSFAC matrices and the Poisson distribution, with *P* values adjusted using the FDR approach for analyses-columns. Values are shown as detailed in the inset: red/orange indicates significant over-representation and blue indicates significant under-representation. The E2F1_Q6 motif represents the putative action of E2F1 and MYC. *B*, representation of E2F1-AP2 and E2F4 binding sites from chromatin immunoprecipitation (chip) assays using the same statistical methodology as described above. The E2F4 data correspond to the joint analysis of cell cycle phases [23]. *C*, representation of genes with periodic expression through the cell cycle. *D*, representation of ER transcriptional regulation from chromatin immunoprecipitation assays or transcriptional changes in MCF7 cells. *E*, representation of additional promoter motifs using TRANSFAC matrices. The wound response signature without cell cycle-associated genes is indicated by the suffix "(-cc)", and the "total set" signature of ER-negative breast cancer contains the immune response plus other biological processes such as the cell cycle. The dasatinib predictive signature is divided into two sets for the effect in prostate and breast cancer respectively. The colorectal prognostic signatures are as defined in Table S1.
doi:10.1371/journal.pone.0004544.g002

Additional programs of cell proliferation, death and metastasis

In an examination for additional mechanisms of transcriptional regulation of signatures, motifs of AHR, EGR1, MYB and MYC were found to be over-represented in a second term. These over-representations were not as widespread as for E2Fs or ER, which suggests that they play only a minor role, but different cancer types and conditions were included: an EGR1 motif was found to be over-represented in CNS and colorectal cancers and the β-catenin multi-cancer prognostic signatures, the breast cancer lung metastasis signature and the predictive signatures of cisplatin and dasatinib (FDR-adjusted P values<0.05) (Fig. 2*E*). In agreement with these observations, we found the lung metastasis signature to contain 22% (12/54) of the genes predicted elsewhere to be EGR1 transcriptional targets [31–34] and the wound response was previously shown to be coordinated with *MYC* amplification [35]. In addition, over-representation of an AHR motif is consistent with its association with ER to regulate cell proliferation [36].

Next, the significance of motif representations was evaluated by analyzing gene expression correlations in representative cancer datasets. Thus, we computed correlations using the Pearson correlation coefficient (PCC) between the seven transcription factors presented above and genes associated with breast cancer prognosis [12] or with the response to docetaxel treatment in breast cancer [37], and compared them with genes non-differentially expressed in these conditions. Higher absolute PCCs between transcription factors and genes associated with prognosis or treatment response were identified in all cases for genes and/or microarray probes (Mann-Whitney (MW) test P values<0.001) (Fig. 3). The prognosis dataset contained a single representative microarray probe for each transcription factor, therefore all of them showed significant differences (Fig. 3*A*). The treatment response dataset contained several probes for some factors, which were evaluated individually to identify technical or biological differences. In this dataset, *AHR*, *EGR1* and *HIF1A* were each represented by a single probe and all of them showed significantly higher correlations with response (Fig. 3*B*). *E2F1*, *E2F4*, *MYC* and *MYB* had more than one probe each, with discordant results in some cases but with average PCCs significantly associated with response in three of them


A
AHR
EGR1
MYC
E2F1
HIF1A
E2F4
MYB
Probability density
Absolute correlation
Breast cancer prognosis differentially expressed genes (FDR < 1%)
Breast cancer prognosis non-differentially expressed genes (FDR > 25%)

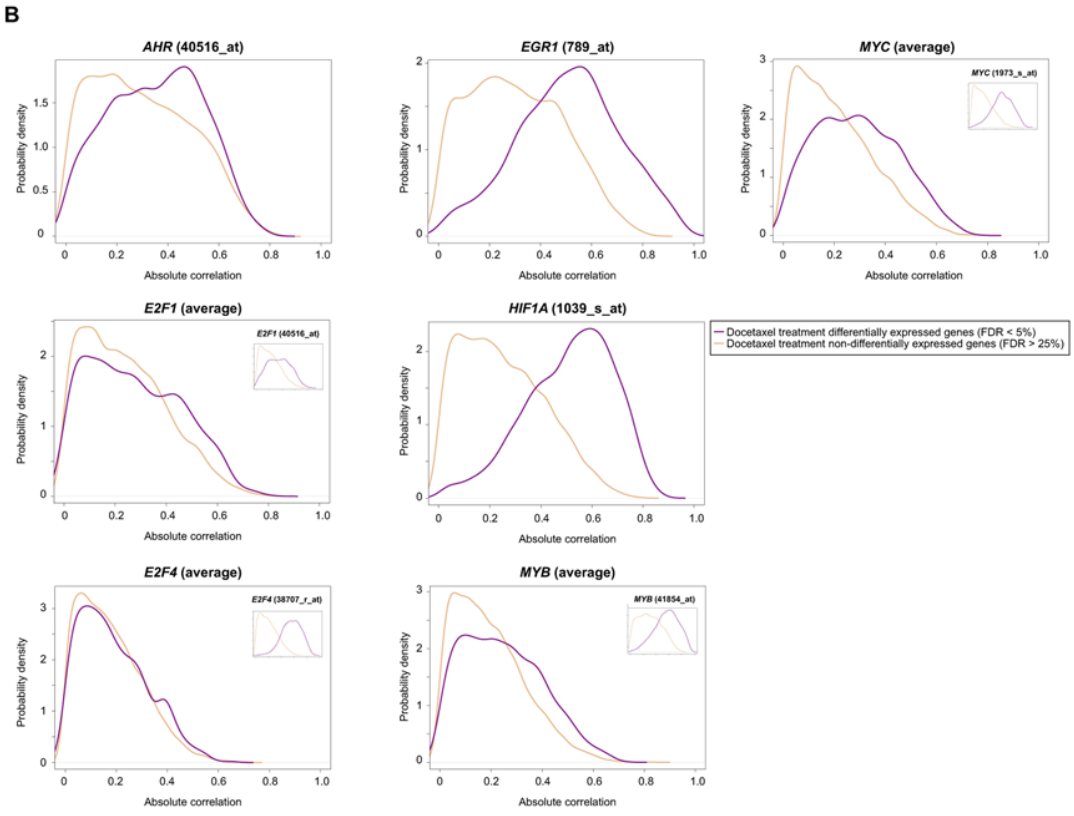

B
AHR (40516_at)
EGR1 (789_at)
MYC (average)
MYC (1973_s_at)
E2F1 (average)
E2F1 (40516_at)
HIF1A (1039_s_at)
E2F4 (average)
E2F4 (38707_r_at)
MYB (average)
MYB (41854_at)
Probability density
Absolute correlation
Docetaxel treatment differentially expressed genes (FDR < 5%)
Docetaxel treatment non-differentially expressed genes (FDR > 25%)

Figure 3. Expression correlations with defined transcription factors. *A*, expression correlations between seven transcription factors_gene names shown at the top of each graph_and genes differentially expressed for breast cancer prognosis measured by metastasis events up to 5 years (pink curves) relative to non-differentially expressed genes in this condition (brown curves). The graphs show absolute PCC values. *B*, same analysis for differentially expressed genes after docetaxel treatment of breast cancer patients relative to non-differentially expressed genes in this condition. Results for *E2F1*, *E2F4*, *MYB* and *MYC* are for average values of all microarray probes representing each factor, whereas the insets show the results for individual probes with significant differences.
doi:10.1371/journal.pone.0004544.g003

(Fig. 3*B*), whereas *E2F4* remained unclassifiable as two probes were significantly correlated and two were not (data not shown).

To further evaluate these observations, we computed correlations between the seven transcription factors and 5,000 randomly selected sets equivalent to the size of the differentially expressed gene sets above. Higher PCCs were observed for most transcription factors in both cancer conditions, with the exception of *E2F4* in prognosis and treatment response (*P* values obtained using the empirical distribution of random PCCs (empirical *P* values) were of 0.16 and 0.11, respectively). Overall, the identification of significant correlations with at least six of the seven factors tested supports the motif predictions and suggests the existence of common transcriptional programs that converge on cell proliferation and death, as well as metastasis as revealed by EGR1 [33].

Analysis of motifs and expression correlations also revealed an association between the apparently dissimilar immune response set and different prognosis signatures. Although it under-represented E2F1 motifs, the immune response shared over-representation of a MYB motif with the 21-gene, wound response breast cancer, and bladder cancer prognostic signatures (Fig. 2*E*). Over-representation of this factor in the immune response is consistent with its role in hematopoiesis [38], and its over-representation in other signatures is consistent with the emerging involvement of the immune response in the prognosis of different cancer types [39]. The high correlations observed in Fig. 3*A* between *MYB* and genes globally associated with breast cancer prognosis (i.e. not limited by the ER status) support this hypothesis. Associations of this signature at other molecular levels will be presented in the following sections.

Transcriptomic correlations between signatures

Given the identification of common transcriptional programs, global expression correlations between signatures should be higher than expected by chance. Using a breast cancer dataset [40] and the average PCC across all microarray probe pairs between any two signatures, significant co-expression was identified in approximately half of the analyses when compared to 10,000 equivalent, randomly selected gene sets (empirical *P* values$<$0.05) (Fig. 4*A*). These results support the existence of functional and molecular associations between many apparently dissimilar signatures, despite the fact that the dataset used had evident technical and biological specificities. Furthermore, the immune response signature showed significant co-expression with 15 of the signatures studied (data not shown), which also supports convergence on this process.

To further test the link to cell proliferation and death at the transcriptomic level, and excluding *a priori* information on expression levels or profiles that could bias the analysis, we examined correlations with gene sets selected using only the criteria for the Gene Ontology (GO) terms Cell Death and Mitosis. These sets were exclusively defined by selecting Entrez genes annotated with those terms, and then used in comparisons in the same way as any other signature. Using 10,000 equivalent random sets, absolute correlations between these GO sets and the signatures were found to be significantly higher in $\geq$12 comparisons (Fig. 4*B*, *left*). The Cell Death set was significantly correlated with five signatures and the Mitosis set was significantly correlated with 11 signatures of different cancer types or conditions. Importantly, differences in the GO sets relative to random were of the same magnitude as comparisons between signatures (Fig. 4*C*).

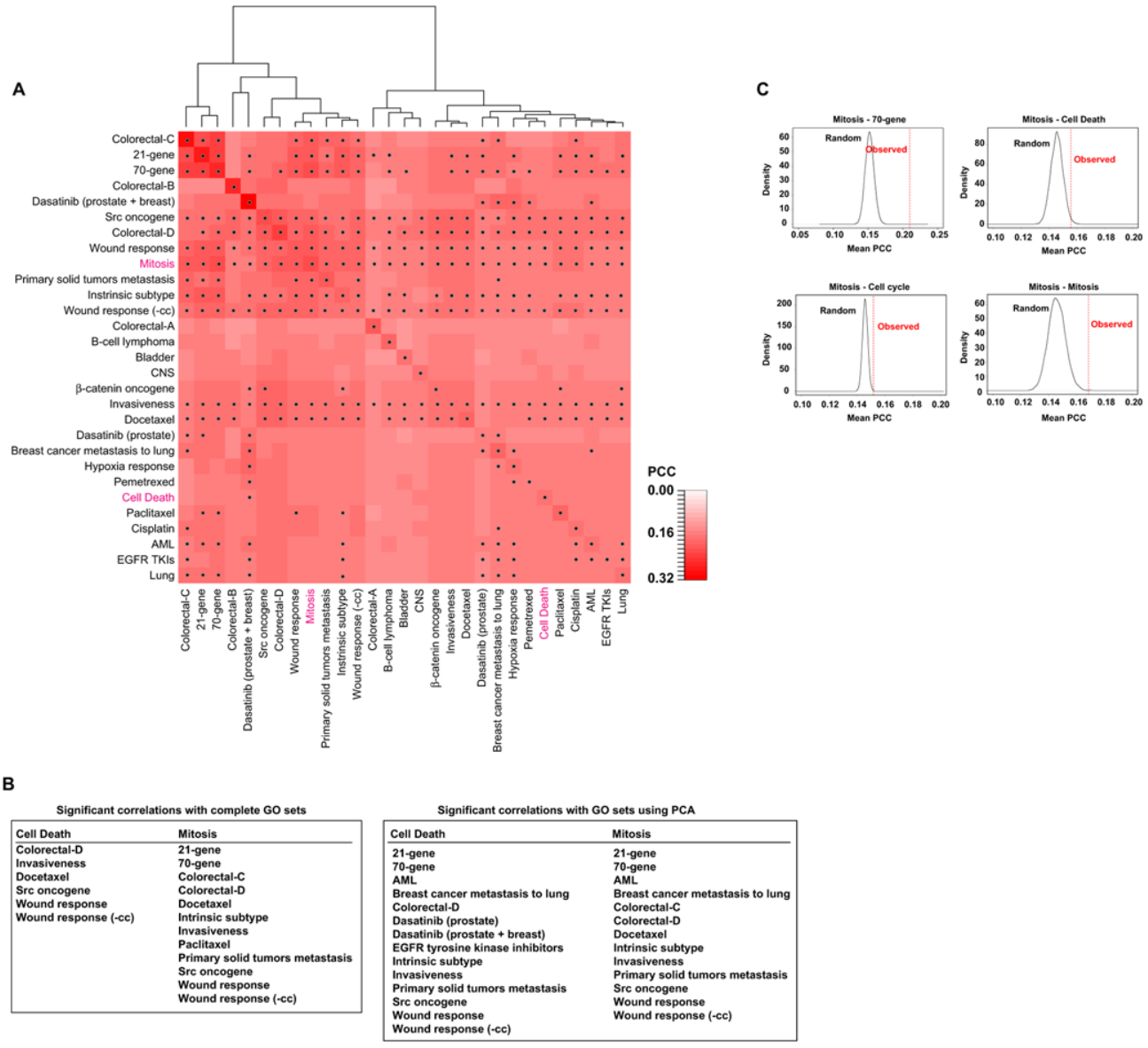
This analysis suggested that measuring the expression levels of genes known to participate in specific biological processes is likely to be of prognostic or predictive value in different situations. However, the analysis was constrained by the possible presence of non-informative expression or sub-sets of genes with different behavior within the GO sets. Thus, reducing the dimensionality of Cell Death and Mitosis sets using a principal component analysis that captured $\sim$80% of the variance raised the number of significant correlations to 12 and 14 sets, respectively (Fig. 4*B*, *right*); these numbers corresponded to a total of $\sim$60% of the signatures examined, irrespective of their type or condition.

Interactome network associations

Functional relationships between proteins can be identified as direct interactions, complex memberships or relatively close connections in the network of protein-protein interactions or interactome network. Given the evidence at the genomic and transcriptomic levels presented above, we hypothesized that proteins encoded by apparently dissimilar signatures will be more closely located in the interactome network than expected by chance. For this analysis we used a dataset consisting mainly of experimentally identified protein-protein interactions, excluding homodimers and orthology-based predictions, and calculated the shortest path between any two nodes or proteins in the giant network component (i.e., the component containing the largest number of connected proteins) [41].

All signature comparisons showed shortest path distributions skewed toward smaller values than expected from the giant component (Fig. 5). Statistical evaluation using the non-parametric MW test identified significant differences with respect to the giant component distribution in 90% of comparisons. The smallest shortest paths were identified for the 21-gene prognostic, and dasatinib and EGFR TKI predictive signatures, although the results may be subject to bias because these sets contain several proteins that are widely studied in the literature and therefore have high network centrality.

To further evaluate these differences, we randomly selected 1,000 sets of 50 proteins with similar average degree centrality to the signatures and obtained their shortest path distributions. Most of the cancer signatures were more closely located than expected by chance and also close to the Cell Death and Mitosis complete sets (empirical *P* values$<$0.05 marked with dots in Fig. 5*A and B*). According to these observations, examination of GO annotations in the direct and one-hop neighborhoods of signatures identified significant over-representation of Cell Cycle or Cell Death terms or their children in all cases (FDR-adjusted *P* values$<$0.05) (GO term details not shown), which reinforces the hypothesis that the signature gene products are molecularly and functionally associated with these processes.



B

Significant correlations with complete GO sets

Cell Death	Mitosis
Colorectal-D	21-gene
Invasiveness	70-gene
Docetaxel	Colorectal-C
Src oncogene	Colorectal-D
Wound response	Docetaxel
Wound response (-cc)	Intrinsic subtype
	Invasiveness
	Paclitaxel
	Primary solid tumors metastasis
	Src oncogene
	Wound response
	Wound response (-cc)

Significant correlations with GO sets using PCA

Cell Death	Mitosis
21-gene	21-gene
70-gene	70-gene
AML	AML
Breast cancer metastasis to lung	Breast cancer metastasis to lung
Colorectal-D	Colorectal-C
Dasatinib (prostate)	Colorectal-D
Dasatinib (prostate + breast)	Docetaxel
EGFR tyrosine kinase inhibitors	Intrinsic subtype
Intrinsic subtype	Invasiveness
Invasiveness	Primary solid tumors metastasis
Primary solid tumors metastasis	Src oncogene
Src oncogene	Wound response
Wound response	Wound response (-cc)
Wound response (-cc)	

Figure 4. Transcriptomic correlations between signatures and with defined biological processes. *A*, heat map of average PCCs between cancer signatures in a breast cancer gene expression dataset [40]. Significant co-expression (empirical *P* values$<$0.05) is indicated by dots. Note that the matrix is not symmetrical because the results were dependent on the size of each gene set; therefore, the larger gene sets (e.g. wound response or invasiveness) showed significant co-expression with many other signatures, perhaps partly due to the fact that they had greater statistical power with which to detect them. Each dot corresponds to the comparison between a signature on the left (simulated set) and a signature at the bottom. The Cell Death and Mitosis sets are highlighted in pink. *B*, left panel, list of signatures that showed significant correlation with the Cell Death or Mitosis complete GO sets. Right panel, list of signatures that showed significant correlation with the Cell Death or Mitosis sets, but only using their principal components. *C*, observed (discontinuous red line) versus expected (black curve for 10,000 randomly selected sets) average PCCs between the Mitosis set and the 70-gene set, the Cell Death set, or genes with periodic expression through the cell cycle.
doi:10.1371/journal.pone.0004544.g004

Next, signatures were depicted as nodes in a network in which the length of the edges is proportional to the average shortest path to the Cell Death and Mitosis sets (Figure 5*C*, *left*). In this network, most signatures were found close to these central processes when compared to 100 random sets with equivalent degree centrality (Figure 5*C*, *right*). Distant signatures represented modest associations at the different molecular levels examined above, such as the prognostic signatures for B-cell lymphoma, colorectal cancer and hypoxia response. These observations suggested correlation across different molecular levels. Thus, negative correlations for all signatures were observed between PCC co-expression values and interactome shortest path distances (average $r = -0.31$ and $\sigma = 0.16$; Mantel test P value $= 0.059$), which is consistent with functional relationships [42–45]. Consequently, higher co-expression between signatures partially correlated with smaller shortest paths between them in the interactome network. These observations highlight the importance of the integrative study, which revealed previously unidentified relationships in gene lists.

The immune response signature was also located close to the Cell Death and Mitosis sets (MW test P values$<$0.001) (Figure

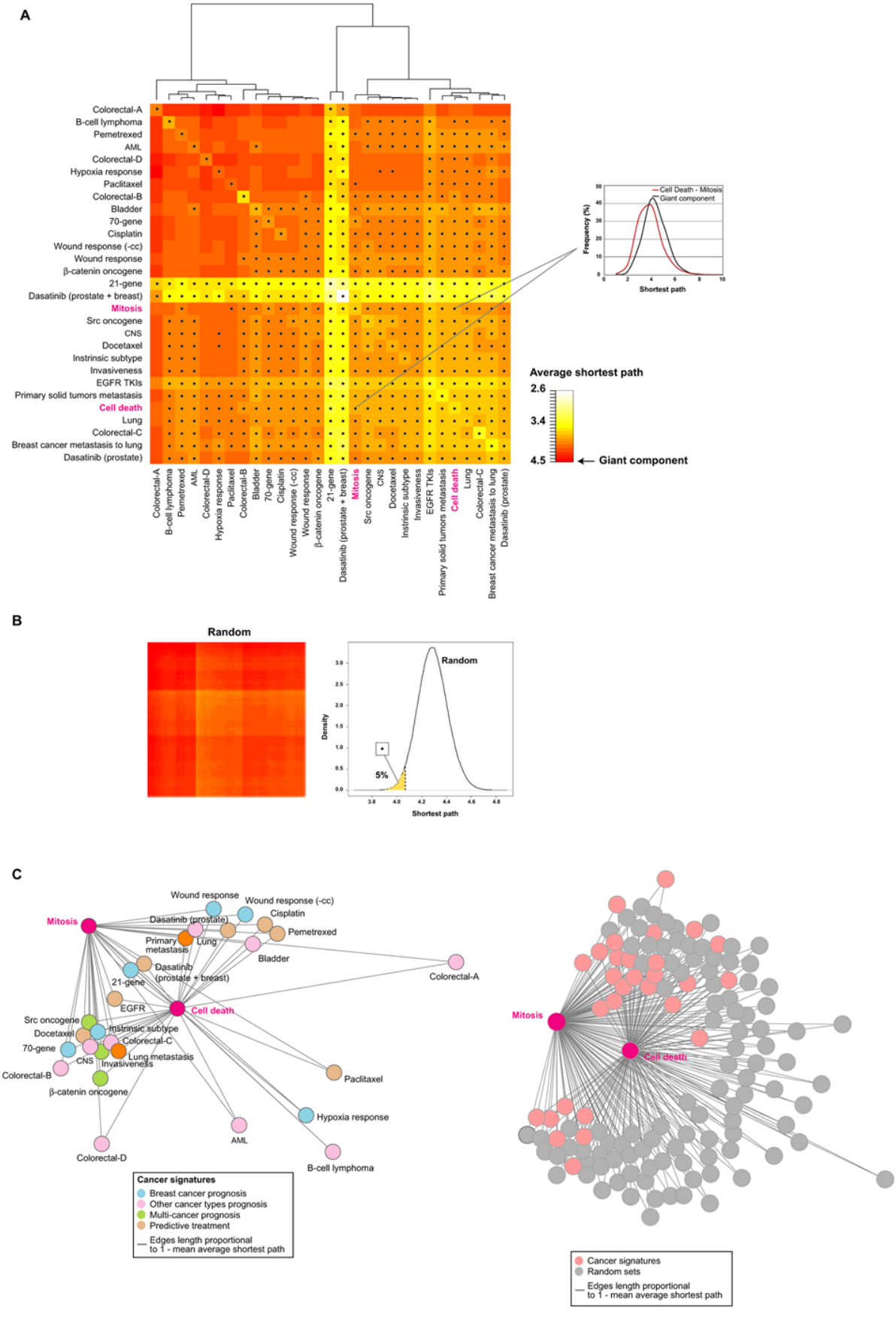

A
Colorectal-A
B-cell lymphoma
Pemetrexed
AML
Colorectal-D
Hypoxia response
Paclitaxel
Colorectal-B
Bladder
70-gene
Cisplatin
Wound response (-cc)
Wound response
β-catenin oncogene
21-gene
Dasatinib (prostate + breast)
Mitosis
Src oncogene
CNS
Docetaxel
Instrinsic subtype
Invasiveness
EGFR TKIs
Primary solid tumors metastasis
Cell death
Lung
Colorectal-C
Breast cancer metastasis to lung
Dasatinib (prostate)
Cell Death - Mitosis
Giant component
Frequency (%)
Shortest path
Average shortest path
2.6
3.4
4.5
Giant component
B
Random
Density
5%
Shortest path
C
Wound response
Wound response (-cc)
Cisplatin
Pemetrexed
Mitosis
Dasatinib (prostate)
Primary metastasis
Lung
Bladder
Colorectal-A
Dasatinib (prostate + breast)
21-gene
EGFR
Cell death
Src oncogene
Instrinsic subtype
Docetaxel
Colorectal-C
70-gene
CNS
Lung metastasis
Invasiveness
Colorectal-B
β-catenin oncogene
Paclitaxel
Hypoxia response
AML
Colorectal-D
B-cell lymphoma
Cancer signatures
Breast cancer prognosis
Other cancer types prognosis
Multi-cancer prognosis
Predictive treatment
Edges length proportional to 1 - mean average shortest path
Mitosis
Cell death
Cancer signatures
Random sets
Edges length proportional to 1 - mean average shortest path

Figure 5. Proximity between gene products of signatures in the interactome network. *A*, heat map of average shortest paths between proteins encoded by signatures. This analysis was performed using only the giant network component. An example of shortest path differences with respect to the giant component is shown in the right panel for the comparison between the complete Cell Death and Mitosis GO sets. *B*, heat map of comparisons of 1,000 randomly selected 50-protein sets in the giant component. Right panel, density plot of average shortest path in randomly selected sets: the 5% lower values are highlighted, which correspond to an average shortest path <4.09. Comparisons between signatures below this empirical cut-off are shown by dots in *A*. *C*, left panel, network representation of average shortest paths between Cell Death and Mitosis and cancer signatures as shown in the inset. Edges lengths are proportional to the average shortest path values. Right panel, network representation of average shortest paths between Cell Death and Mitosis and cancer signatures or randomly selected protein sets with equivalent degree centrality.
doi:10.1371/journal.pone.0004544.g005

S1*A*). Consequently, examination of the proportion of GO annotations in the one-hop neighborhood of this signature identified over-representations of terms related to cell proliferation and death, while the direct interactors only showed over-representation of terms associated with the immune system (Figure S1*B*). Thus, although the gene products with prognostic value for ER-negative breast cancer are not directly connected to the common processes identified above, they are significantly associated in a second term, as well as transcriptionally co-expressed and co-regulated with many signatures.

Evaluation of properties and interactions in independent datasets

The observations described above were evaluated in two independently generated signatures of cancer conditions. A recent study described a lung metastasis signature of breast cancer using a different methodological approach [46]. We found ~70% (15/21) of the genes in this signature to contain E2F TRANSFAC motifs and ~60% (13/21) to be targets of E2F1-AP2 and/or ER. In addition, significant correlations with eight prognostic signatures were identified, seven of them of breast cancer (empirical *P* values<0.001) (results of the analyses of this signature are detailed in Table S2). The correlation with Mitosis was higher than expected (empirical $P<0.001$), while the correlation with Cell Death was non-significant (empirical $P=0.18$). Finally, gene products in this signature showed smaller average shortest paths than expected with 21 of the 28 signatures, including Cell Death, Mitosis and the lung metastasis signature presented previously [31] (empirical *P* values<0.05).

To further corroborate our observations, we selected a different neoplastic condition from the recent literature: metastatic colorectal cancer treated with the EGFR inhibitor cetuximab (Erbitux®) [47]. Previous studies suggest that EGFR mutations are associated with the response to TKIs but not to cetuximab [48,49]. We evaluated our observations by examining the distribution of gene annotations in the rank of hazard ratios (HRs) that measures the response to cetuximab treatment by progression-free survival. In this analysis, cell proliferation and the immune response were identified as the processes with the greatest effect on the response (Fig. 6). Importantly, the set of genes whose high expression most strongly associate to response was for a wound-like phenotype that was previously shown to provide prognosis value for breast, lung and gastric cancer [50]. The next associated high-expression sets were for doxorubicin treatment in gastric cancer, breast cancer prognosis (70-gene) and prognosis of different cancer types not examined in this study (hepatocellular carcinoma and multiple myeloma prognosis). Moreover, high-expression of E2F1, hypoxia and MYC targets was also associated with the response with similar strength (Fig. 6*A*). Collectively, these observations endorse the biological convergence of signatures.

The analysis of the cetuximab dataset also revealed a complementary behavior of cell proliferation and the immune response consistent with the representation of E2F1 motifs shown above. Patients with high expression of cell proliferation-related genes and low expression of immune response-related genes responded to treatment (Fig. 6*B*), whilst there were no patients with high expression values of both processes. Hence, a strong anti-correlation was observed between genes annotated with the GO term Immune Response and genes annotated with Mitosis ($r=-0.79$) (Fig. 6*C*). This observation leads to speculate that these processes play balancing roles in prognosis and treatment response. Good responders to cetuximab may show strong dependence on a "cell proliferation-on" molecular program, while non-responders could be sensitive to immune system-based therapy.

Discussion

Despite the low degree of overlap in terms of gene identity, apparently dissimilar cancer signatures converge on specific biological processes. Convergence is defined by significant molecular and functional associations between genes and/or proteins: i/ predicted promoter motifs; ii/ experimentally identified DNA binding sites; iii/ cell cycle-periodic profiles; iv/ ER-mediated transcriptional regulation; v/ co-expression with defined transcription factors; vi/ co-expression between signatures and with specific GO gene sets; and, vii/ close proximity in the interactome network and neighborhood over-representation in these same GO terms. Consequently, this study suggests the existence of common design principles in a system-level cellular model—illustrated by transcriptome-interactome correlations—not only of prognostic signatures but also of metastasis and treatment response signatures. Overall, the integrative study highlights the importance of analyzing signatures beyond gene names, which provides a better global understanding by revealing previously unidentified properties and associations.

Biological convergence has important implications for the interpretation of signatures. Given a single gene whose transcript levels are associated with differences in patient outcome, this observation should be interpreted *a priori* in the context of cell proliferation, death or the immune response processes. For example, *BRCA1* and *BRCA2* have different cellular functions, with a degree of overlap, but each of them is present in several prognostic and predictive signatures, probably because their transcript levels reproduce precisely the potential for cell proliferation. This potential is defined by the presence of genes with periodic expression through the cell cycle, and other analyses at the genome, transcriptome and proteome levels shown here provide strong evidence of common properties and interactions. Therefore, further conclusions concerning gene functions such as DNA repair and its role in prognosis should be considered, controlling for the possible confounding effect of biological convergence.

From a mechanistic point of view, this study indicates the existence of a cancer cell phenotype that decisively influences critical aspects of neoplasia. This observation follows on from the long-known global importance of the potential for cell prolifera-

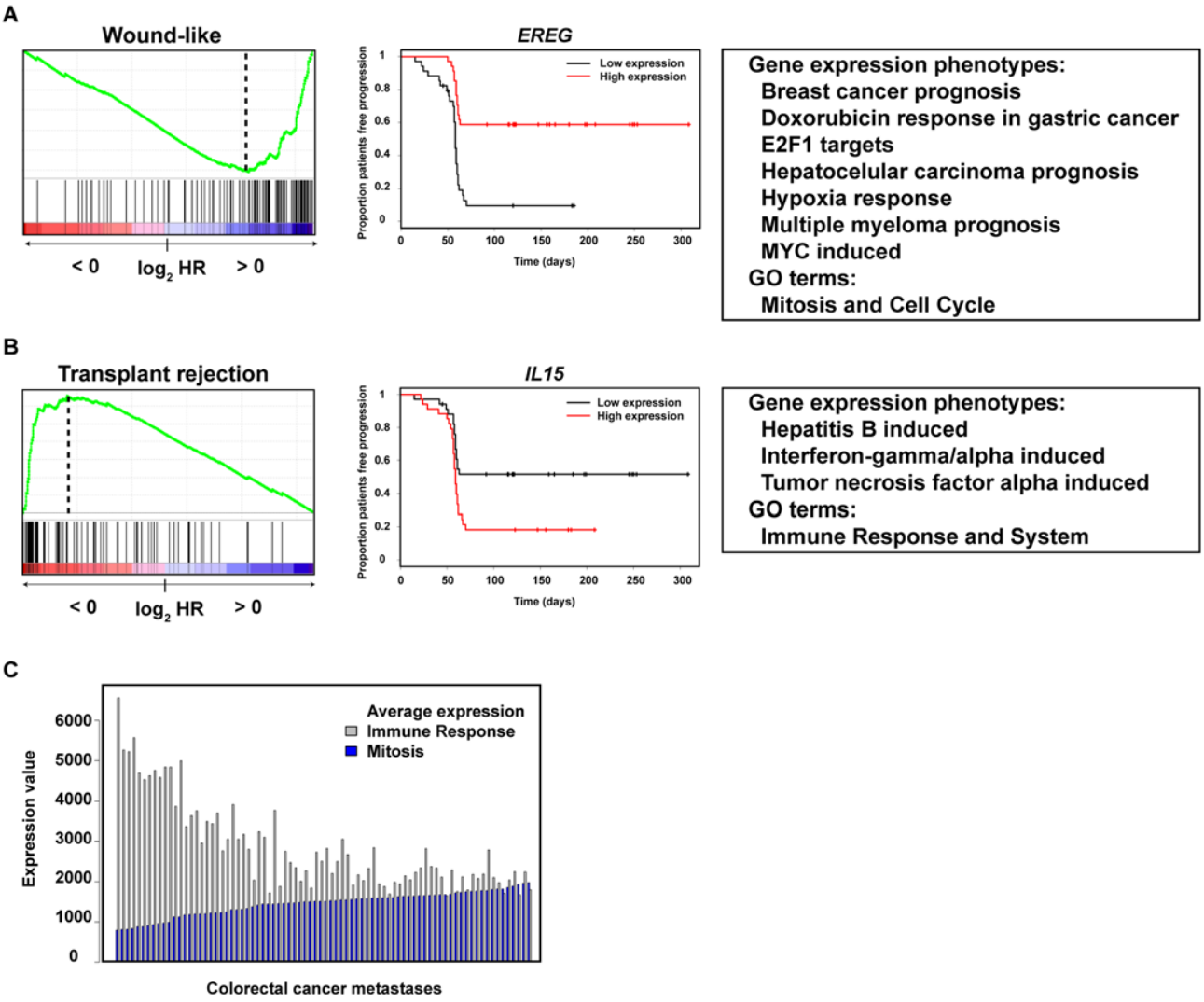


Figure 6. Asymmetric distribution of gene annotations in the response to cetuximab treatment. *A*, left panel, GSEA results for the strongest associated phenotype with high-expression genes predicting treatment response ($\log_2$ HR>0). Central panel, expression analysis plot of the extreme gene expression (*EREG*), which was also noted in the original publication [47]. Right panel, additional phenotypic and GO term sets with high-expression genes associated to treatment response at FDR *Q* values<1%. *B*, left panel, GSEA results for the strongest associated phenotype with low-expression genes predicting treatment response ($\log_2$ HR<0). Central panel, expression analysis plot of the extreme gene expression (*IL15*). Right panel, additional phenotypic and GO term sets with low-expression genes associated to treatment response at FDR *Q* values<1%. *C*, Histogram plot of average expression values of genes annotated with the Immune Response or Mitosis across samples in the cetuximab dataset. Average GO set expression values show a negative correlation with ordered metastatic samples.
doi:10.1371/journal.pone.0004544.g006

tion and repressed cell death in tumorigenesis [51], while reinforcing the emerging role of the immune response in prognosis and prediction [39]. However, while this study provides the first evidence of convergence of prognostic, metastasis and predictive signatures in these processes, other processes or signaling pathways are probably represented and specificities may exist. For instance, the potential for metastasis also depends on the activity of processes such as extracellular matrix remodeling. Similar systems-level analyses of a larger number of metastasis signatures may reveal properties masked here by the restriction of the study to mainly prognostic and predictive sets. Nonetheless, some prognostic or predictive sets are not independent of the potential for developing metastasis [10]. Future research may reveal a more complex molecular wiring diagram of the processes participating in cancer signatures.

Materials and Methods

Cancer signatures

We compiled 28 signatures from 24 studies, comprising 19 prognostic signatures, two signatures focused mainly on metastasis, and seven predictive treatment response signatures, as detailed in Table S1. Note that the 21-gene breast cancer prognosis signature was originally described as a predictive set for tamoxifen treatment [52] and the intrinsic subtype signature [53] corresponds to a validated set taken from the original report [13]. We also examined the wound response prognosis signature without including the initially identified cell cycle-associated genes [40] and the predictive signature for dasatinib treatment response subdivided for prostate and breast cancer [54]. Gene names or microarray probes were taken from the original publications and

mapped to Entrez GeneIDs using the BioMart and Bioconductor [55] tools and by manual curation of each signature.

Genomic analyses

Transcription factor (TF) motifs in promoter sequences 1 kilobase (kb) upstream of the transcription start site were predicted using MatScan [56] and position weight matrices from JASPAR [57] and TRANSFAC [58] (111 and 625 motifs, respectively). Probabilities were calculated using the Poisson distribution as an approximation to the binomial as follows $\left\{f(k,\lambda)=\frac{\lambda^k \cdot e^{-\lambda}}{k!}\right\}$ (where $\lambda = n \cdot p$, $p =$ proportion of genes with a defined motif that are part of the signature and $n =$ total number of genes with this motif in the genome). Promoter sequences (−1 kb) of Ensembl protein-encoding gene entries ($n \approx 18{,}800$) were used as a common reference for the motif analyses. Corrections for multiple comparisons were computed using the false discovery rate (FDR) approach [59]. Chromatin immunoprecipitation data and periodically expressed genes were taken from the respective references [23,24,28] or from the relevant repositories [29] and examined using the same methodology. The ER binding sites identified by chromatin immunoprecipitation assays were assigned to a single GeneID based on the closest known gene locus (5′-end) in the May 2004 version of the human genome in the UCSC Genome Browser.

Transcriptomic analyses

Transcriptional targets of the ER signaling pathway were examined using preprocessed and normalized data [29]. Correlations of transcription factors were performed by defining differentially expressed genes at FDR<1% in breast cancer prognosis measured with metastasis events within 5 years [12], which correspond to 179 microarray probes, or by defining differentially expressed genes at FDR<5% in docetaxel treatment response [37], which correspond to 1,525 probes. Differences in PCC distributions were assessed using the Mann-Whitney (MW) and Kolmogorov Smirnov non-parametric tests, with similar results. Average correlations in pairwise signature comparisons were calculated using all probes in the signature gene lists and compared to equivalent probe sets randomly selected from the same breast cancer dataset [40]. Dimensionality was reduced by applying a principal component analysis (PCA) until ~80% of the variance in gene expression was captured, which represented <25 genes in the Cell Death or Mitosis Gene Ontology (GO) sets (originally containing 58 and 117 genes, respectively). For the analysis of cetuximab treatment response, we computed a Cox proportional hazards model for each microarray probe, using the progression-free survival as the time variable, and dividing the sample set into two equally-sized groups according to the expression level of the corresponding probe (low versus high). Ranks according to the log-hazard ratio were used as input lists for the Gene Set Enrichment Analysis (GSEA) [60]. The GSEA was run for all curated and GO datasets found in MSigDB database. We used default values for all the parameters except for the median probe instead of the max probe as the collapse method when multiple probe sets map to the same gene. The evaluation of correlation between the Immune Response ($n = 311$) and Mitosis GO sets in the dataset of cetuximab treatment response was performed averaging expression values of both gene sets in each metastasis sample. The R programming language was used for analyses and graphics.

Interactome analyses

The human interactome network was built by combining three previously published datasets consisting mainly of experimentally verified interactions [41]. The dataset based on the Human Protein Reference Database (HPRD) contains compiled and filtered binary protein interactions from available databases. High-confidence yeast two-hybrid interactions were then incorporated and orthology-based predictions and homo-dimers were excluded to avoid specific bias. Proteins with no assigned GeneID were also excluded from our analyses. The numbers of proteins or nodes and interactions or edges in the complete dataset were 8,519 and 35,492, respectively. The percentage of signature gene products mapped in this dataset ranged between 40 and 85. Shortest paths were calculated using only the giant network component and the geodesic formulation given by Freeman in the R programming language [41]. Differences in the distributions of shortest paths were assessed using the MW test. Empirical simulations using 50-protein sets were selected as the average size of cancer signatures, using only nodes from the giant component with average degree centrality equivalent to the signatures. The average degree of signatures, excluding three outliers that contain widely studied genes (21-gene, dasatinib prostate and breast, and EGFR TKIs), was 7.48, while the average degree of 1,000 random sets was 7.53. To evaluate the relationship between gene co-expression and interactome distances, a correlation coefficient was calculated between average PCCs in each signature-pair and the corresponding average short path in the giant network component, which was then evaluated to the null hypothesis of no-correlation between the two measures using the Mantel test. The representation of GO terms in neighborhoods was assessed using the shortest path measure and the hypergeometric distribution and FDR P value adjustment, taking as a reference all proteins in the giant component and excluding signature proteins in each case. The Onto-Express tool was used for this analysis [61].

Supporting Information

Figure S1 Topological associations of the immune response signature in the interactome network. A, left panel, shortest path distributions between the immune response and the Cell Death and Mitosis sets (yellow and green curves, respectively) relative to the giant component (black curve). Right panel, strategy for evaluating differences in proportions of GO annotations in the direct and one-hop interactome network neighborhoods. B, over-represented GO terms in the direct and one-hop neighborhoods of the immune response signature.
Found at: doi:10.1371/journal.pone.0004544.s001 (2.22 MB EPS)

Table S1
Found at: doi:10.1371/journal.pone.0004544.s002 (0.03 MB XLS)

Table S2
Found at: doi:10.1371/journal.pone.0004544.s003 (0.03 MB XLS)

Acknowledgments

We thank Javier Díez for helpful comments and for contributing to discussions on this study.

Author Contributions

Conceived and designed the experiments: XS NLB MAP. Performed the experiments: XS NB NLB AB. Analyzed the data: XS NB NLB AB PH OR FC. Contributed reagents/materials/analysis tools: CAM HA AU SdS FC GC VM MAP. Wrote the paper: MAP.

References

1. Nuyten DS, van de Vijver MJ (2008) Using microarray analysis as a prognostic and predictive tool in oncology: focus on breast cancer and normal tissue toxicity. Semin Radiat Oncol 18: 105–114.
2. Morris SR, Carey LA (2007) Gene expression profiling in breast cancer. Curr Opin Oncol 19: 547–551.
3. Michiels S, Koscielny S, Hill C (2007) Interpretation of microarray data in cancer. Br J Cancer 96: 1155–1158.
4. Eden P, Ritz C, Rose C, Ferno M, Peterson C (2004) "Good Old" clinical markers have similar power in breast cancer prognosis as microarray gene expression profilers. Eur J Cancer 40: 1837–1841.
5. Lin YH, Friederichs J, Black MA, Mages J, Rosenberg R, et al. (2007) Multiple gene expression classifiers from different array platforms predict poor prognosis of colorectal cancer. Clin Cancer Res 13: 498–507.
6. Sherlock G (2005) Of fish and chips. Nat Methods 2: 329–330.
7. Son CG, Bilke S, Davis S, Greer BT, Wei JS, et al. (2005) Database of mRNA gene expression profiles of multiple human organs. Genome Res 15: 443–450.
8. Ein-Dor L, Zuk O, Domany E (2006) Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer. Proc Natl Acad Sci U S A 103: 5923–5928.
9. Roepman P, Kemmeren P, Wessels LF, Slootweg PJ, Holstege FC (2006) Multiple robust signatures for detecting lymph node metastasis in head and neck cancer. Cancer Res 66: 2361–2366.
10. Fan C, Oh DS, Wessels L, Weigelt B, Nuyten DS, et al. (2006) Concordance among gene-expression-based predictors for breast cancer. N Engl J Med 355: 560–569.
11. Massague J (2007) Sorting out breast-cancer gene signatures. N Engl J Med 356: 294–297.
12. van 't Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AA, et al. (2002) Gene expression profiling predicts clinical outcome of breast cancer. Nature 415: 530–536.
13. Perou CM, Sorlie T, Eisen MB, van de Rijn M, Jeffrey SS, et al. (2000) Molecular portraits of human breast tumours. Nature 406: 747–752.
14. Chang JT, Nevins JR (2006) GATHER: a systems approach to interpreting genomic signatures. Bioinformatics 22: 2926–2933.
15. Wennmalm K, Miller LD, Bergh J (2007) A gene signature in breast cancer. N Engl J Med 356: 1887–1888; author reply 1887–1888.
16. Yu JX, Sieuwerts AM, Zhang Y, Martens JW, Smid M, et al. (2007) Pathway analysis of gene signatures predicting metastasis of node-negative primary breast cancer. BMC Cancer 7: 182.
17. Shen R, Ghosh D, Chinnaiyan AM (2004) Prognostic meta-signature of breast cancer developed by two-stage mixture modeling of microarray data. BMC Genomics 5: 94.
18. Zhang Z, Chen D, Fenstermacher DA (2007) Integrated analysis of independent gene expression microarray datasets improves the predictability of breast cancer outcome. BMC Genomics 8: 331.
19. Vuaroqueaux V, Urban P, Labuhn M, Delorenzi M, Wirapati P, et al. (2007) Low *E2F1* transcript levels are a strong determinant of favorable breast cancer outcome. Breast Cancer Res 9: R33.
20. Hernandez P, Sole X, Valls J, Moreno V, Capella G, et al. (2007) Integrative analysis of a cancer somatic mutome. Mol Cancer 6: 13.
21. Shen R, Chinnaiyan AM, Ghosh D (2008) Pathway analysis reveals functional convergence of gene expression profiles in breast cancer. BMC Med Genomics 1: 28.
22. Teschendorff AE, Miremadi A, Pinder SE, Ellis IO, Caldas C (2007) An immune response gene expression module identifies a good prognosis subtype in estrogen receptor negative breast cancer. Genome Biol 8: R157.
23. Balciunaite E, Spektor A, Lents NH, Cam H, Te Riele H, et al. (2005) Pocket protein complexes are recruited to distinct targets in quiescent and proliferating cells. Mol Cell Biol 25: 8166–8178.
24. Jin VX, Rabinovich A, Squazzo SL, Green R, Farnham PJ (2006) A computational genomics approach to identify cis-regulatory modules from chromatin immunoprecipitation microarray data–a case study using E2F1. Genome Res 16: 1585–1595.
25. Hallstrom TC, Mori S, Nevins JR (2008) An E2F1-dependent gene expression program that determines the balance between proliferation and cell death. Cancer Cell 13: 11–22.
26. Du W, Pogoriler J (2006) Retinoblastoma family genes. Oncogene 25: 5190–5200.
27. Rhodes DR, Kalyana-Sundaram S, Mahavisno V, Barrette TR, Ghosh D, et al. (2005) Mining for regulatory programs in the cancer transcriptome. Nat Genet 37: 579–583.
28. Whitfield ML, Sherlock G, Saldanha AJ, Murray JI, Ball CA, et al. (2002) Identification of genes periodically expressed in the human cell cycle and their expression in tumors. Mol Biol Cell 13: 1977–2000.
29. Carroll JS, Meyer CA, Song J, Li W, Geistlinger TR, et al. (2006) Genome-wide analysis of estrogen receptor binding sites. Nat Genet 38: 1289–1297.
30. Butt AJ, Sutherland RL, Musgrove EA (2007) Live or let die: oestrogen regulation of survival signalling in endocrine response. Breast Cancer Res 9: 306.
31. Minn AJ, Gupta GP, Siegel PM, Bos PD, Shu W, et al. (2005) Genes that mediate breast cancer metastasis to lung. Nature 436: 518–524.
32. Krones-Herzig A, Mittal S, Yule K, Liang H, English C, et al. (2005) Early growth response 1 acts as a tumor suppressor in vivo and in vitro via regulation of p53. Cancer Res 65: 5133–5143.
33. Fahmy RG, Dass CR, Sun LQ, Chesterman CN, Khachigian LM (2003) Transcription factor Egr-1 supports FGF-dependent angiogenesis during neovascularization and tumor growth. Nat Med 9: 1026–1032.
34. Ishikawa H, Shozu M, Okada M, Inukai M, Zhang B, et al. (2007) Early growth response gene-1 plays a pivotal role in down-regulation of a cohort of genes in uterine leiomyoma. J Mol Endocrinol 39: 333–341.
35. Adler AS, Lin M, Horlings H, Nuyten DS, van de Vijver MJ, et al. (2006) Genetic regulators of large-scale transcriptional signatures in cancer. Nat Genet 38: 421–430.
36. Pliskova M, Vondracek J, Vojtesek B, Kozubik A, Machala M (2005) Deregulation of cell proliferation by polycyclic aromatic hydrocarbons in human breast carcinoma MCF-7 cells reflects both genotoxic and nongenotoxic events. Toxicol Sci 83: 246–256.
37. Chang JC, Wooten EC, Tsimelzon A, Hilsenbeck SG, Gutierrez MC, et al. (2005) Patterns of resistance and incomplete response to docetaxel by gene expression profiling in breast cancer patients. J Clin Oncol 23: 1169–1177.
38. Greig KT, Carotta S, Nutt SL (2008) Critical roles for c-Myb in hematopoietic progenitor cells. Semin Immunol.
39. Ambs S, Marincola FM, Thurin M (2008) Profiling of immune response to guide cancer diagnosis, prognosis, and prediction of therapy. Cancer Res 68: 4031–4033.
40. Chang HY, Nuyten DS, Sneddon JB, Hastie T, Tibshirani R, et al. (2005) Robustness, scalability, and integration of a wound-response gene expression signature in predicting breast cancer survival. Proc Natl Acad Sci U S A 102: 3738–3743.
41. Hernandez P, Huerta-Cepas J, Montaner D, Al-Shahrour F, Valls J, et al. (2007) Evidence for systems-level molecular mechanisms of tumorigenesis. BMC Genomics 8: 185.
42. Ge H, Liu Z, Church GM, Vidal M (2001) Correlation between transcriptome and interactome mapping data from Saccharomyces cerevisiae. Nat Genet 29: 482–486.
43. Grigoriev A (2001) A relationship between gene expression and protein interactions on the proteome scale: analysis of the bacteriophage T7 and the yeast Saccharomyces cerevisiae. Nucleic Acids Res 29: 3513–3519.
44. Jansen R, Greenbaum D, Gerstein M (2002) Relating whole-genome expression data with protein-protein interactions. Genome Res 12: 37–46.
45. Kemmeren P, van Berkum NL, Vilo J, Bijma T, Donders R, et al. (2002) Protein interaction verification and functional annotation by integrated analysis of genome-scale data. Mol Cell 9: 1133–1143.
46. Landemaine T, Jackson A, Bellahcene A, Rucci N, Sin S, et al. (2008) A Six-Gene Signature Predicting Breast Cancer Lung Metastasis. Cancer Res 68: 6092–6099.
47. Khambata-Ford S, Garrett CR, Meropol NJ, Basik M, Harbison CT, et al. (2007) Expression of epiregulin and amphiregulin and K-ras mutation status predict disease control in metastatic colorectal cancer patients treated with cetuximab. J Clin Oncol 25: 3230–3237.
48. Janne PA, Engelman JA, Johnson BE (2005) Epidermal growth factor receptor mutations in non-small-cell lung cancer: implications for treatment and tumor biology. J Clin Oncol 23: 3227–3234.
49. Tsuchihashi Z, Khambata-Ford S, Hanna N, Janne PA (2005) Responsiveness to cetuximab without mutations in EGFR. N Engl J Med 353: 208–209.
50. Chang HY, Sneddon JB, Alizadeh AA, Sood R, West RB, et al. (2004) Gene expression signature of fibroblast serum response predicts human cancer progression: similarities between tumors and wounds. PLoS Biol 2: E7.
51. Hanahan D, Weinberg RA (2000) The hallmarks of cancer. Cell 100: 57–70.
52. Paik S, Shak S, Tang G, Kim C, Baker J, et al. (2004) A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. N Engl J Med 351: 2817–2826.
53. Hu Z, Fan C, Oh DS, Marron JS, He X, et al. (2006) The molecular portraits of breast tumors are conserved across microarray platforms. BMC Genomics 7: 96.
54. Wang XD, Reeves K, Luo FR, Xu LA, Lee F, et al. (2007) Identification of candidate predictive and surrogate molecular markers for dasatinib in prostate cancer: rationale for patient selection and efficacy monitoring. Genome Biol 8: R255.
55. Durinck S, Moreau Y, Kasprzyk A, Davis S, De Moor B, et al. (2005) BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. Bioinformatics 21: 3439–3440.
56. Blanco E, Messeguer X, Smith TF, Guigo R (2006) Transcription factor map alignment of promoter regions. PLoS Comput Biol 2: e49.
57. Bryne JC, Valen E, Tang MH, Marstrand T, Winther O, et al. (2007) JASPAR, the open access database of transcription factor-binding profiles: new content and tools in the 2008 update. Nucleic Acids Res 36: D102–106.

58. Matys V, Fricke E, Geffers R, Gossling E, Haubrock M, et al. (2003) TRANSFAC: transcriptional regulation, from patterns to profiles. Nucleic Acids Res 31: 374–378.
59. Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society Series B-Statistical Methodology 57: 289–300.
60. Subramanian A, Kuehn H, Gould J, Tamayo P, Mesirov JP (2007) GSEA-P: a desktop application for Gene Set Enrichment Analysis. Bioinformatics 23: 3251–3253.
61. Khatri P, Voichita C, Kattan K, Ansari N, Khatri A, et al. (2007) Onto-Tools: new additions and improvements in 2006. Nucleic Acids Res 35: W206–211.

Part IV

CLOSING

8

DISCUSSION

Each one of the core articles of this thesis includes its own section of discussion, where the main findings of each study are described, as well as their consistency with previously published results or the underlying biological mechanisms that may explain them. Therefore, in this chapter we will briefly expand some issues not explained in detail in the articles. We also present a brief summary of the findings in relationship to the working hypothesis of each study, and we contrast our results with other studies subsequently published.

8.1 INTEGRATIVE ANALYSIS OF BREAST CANCER SOMATIC MUTOME

In 2006, the landmark paper of Sjöblom et al. represented a milestone in the field of cancer genomic profiling [188]. This work aimed to detect the complete set of mutated genes in two series of breast and colorectal tumors. Most impressively, it was all done before the advent of UHTS technology. Although the contribution of their work vas invaluable, both sets of mutated genes were obtained using relatively small sample sizes (11 breast and 11 colorectal tumors), and some concerns were raised about the statistical power of the study to detect the true mutated genes [68, 76, 171]. Indeed, some of the detected mutations could be false positives and larger sample sizes would be required to detect genes mutated at low rates, but the fact that all previously known cancer-mutated genes were already detected in their study, as well as the proper two-stage experimental design, point to accept the results as reliable and worth for further examination [149].

Sjöblom et al. obtained quite large lists of mutated genes that they considered as validated (137 for breast cancer, 105 for colorectal cancer). Furthermore, when they combined the results of the two stages of their analysis obtained even larger lists (673 for breast cancer, 519 for colorectal cancer). However these expanded lists cannot be considered as completely validated, since they may contain a higher proportion of genes carrying passenger mutations. Therefore, the aim of our work was to use an integrative analytical approach to characterize the set of mutated genes and detect those that are more prone to be indeed related to cancer. To simplify the analysis, we focused only on breast cancer genes.

The genomic, transcriptomic and proteomic analysis of breast cancer mutated genes helped to identify those markers potentially involved in breast cancer carcinogenesis or prognosis. Thus, *DBN1*, a gene involved in cell differentiation and development which had not been related to cancer at that time, was identified as a potential oncogene. Although no further links with *DBN1* to breast cancer have been described, this gene has been recently found to be involved in mantel cell lymphomas [220] and has clinical utility in predicting prognosis in lung cancer patients [136].

In our paper, poorly-characterized genes *ABCA3* and *SPTAN1* were postulated as new genes potentially associated with breast cancer prognosis. More specifically, low levels of *ABCA3* and mid-to-low levels of *SPTAN1* were found to predict poor survival. Interestingly, this fact has recently been also postulated in another work [178], where they find that loss of *ABCA3* is significantly associated with positive nodal status and negative progesterone receptor expression. Moreover, they postulate that underexpression of *ABCA3* contributed to a higher risk for tumor recurrence, which has a direct impact on survival rates. Besides our work, *SPTAN1* has not been directly linked to breast cancer, but it had also been previously found to be involved in ovarian cancer, which is another hormonal-induced tumor tightly related with breast cancer [113].

This work demonstrates how the integration of *omic* data can unveil potential functional candidates of a particular biological process with increased confidence. The strategy used here is applicable to other cancer types and would help to identify new tumor suppressor genes and oncogenes and the wiring diagram of functional interactions between them. This type of analysis not only can be applied to genes displaying point mutations in tumor cells, but also with genes harboring copy number alterations, which can also have an essential role in the tumor development [199].

Besides our own approach, different studies have followed other kind of integrating methodologies, based on network analysis [50] or functional associations [20]. Although it may be argued that our analysis lacks specificity regarding the role of the different mutations in the carcinogenic process, our results will undoubtedly help to focus subsequent experimental characterizations on key gene/protein candidates.

8.2 GENETIC AND GENOMIC ANALYSIS MODELING OF GERMLINE MYC OVEREXPRESSION AND CANCER SUSCEPTIBILITY

In the last years, genetic loci on chromosome region 8q24 have been recurrently found to be conferring an increased risk of developing prostate

cancer, regardless the ethnic origin of the individuals [7, 70]. Subsequent studies involving larger cohorts confirmed these findings [179]. Moreover, these region has been also found to be involved in the susceptibility of developing tumors at other localizations, such as colon and rectum [232, 205], bladder [102] or breast [26]. All these loci cluster in three linkage disequilibrium (LD)-independent genomic regions spanning ~500kb. However, the 8q24 region does not contain any currently known protein-coding elements, and only *MYC* is located a few hundred kilobases away. Therefore, the biological mechanisms underlying these associations remain unclear.

During the last years it has already been argued that gene expression is under genetic control [170] and, more importantly, that transcripts that are associated with risk variants are potential candidates for mediating the effect of the deleterious alleles on the disease [59]. Therefore our study tried to determine if any of the previously reported germline variants could be modulating *MYC* transcript abundance. This approach was based on the possibility of the presence of *MYC* enhancers or regulators in the region, which has been recently confirmed for the region of the rs6983267 SNP [4, 160, 190, 228]. This enhancer has also been demonstrated to confer an increased risk of developing colorectal cancer by activating the Wnt signaling pathway [209].

To assess this association between genotypes and *MYC* RNA expression levels we used gene expression and SNP data from HapMap individuals, as well as public SNP and expression data from prostate samples. After performing the expression quantitative trait loci (eQTL) analysis, we observed that *MYC* expression was correlated with variants of the SNP rs1447295. Although later studies have found negative results [161] on this association, these differences may be attributed to issues of statistical power, tissue purity or differences in *MYC* RNA quantification. Thus, supporting these conclusions may require further insight in the following future. Interestingly, a recent study has demonstrated the risk variant of the rs6983267 SNP is related to an enhancer of *MYC* expression during early prostate organogenesis [222], suggesting that risk variants might even influence prostate cancer risk significantly before tumor formation. Overall, these results reinforce the usefulness of the eQTL approaches in the field of cancer susceptibility analysis.

There are different possibilities that could explain why 8q24 risk variants do not show a consistent association with *MYC* transcript abundance. In most studies, *MYC* transcript levels are usually measured using expression microarrays, which is a technique that has difficulties in measuring small expression differences and is affected by multiple sources of vari-

ability [194]. Thus, changes in *MYC* expression may be too slight to be detected by this technique. This could be overcome by using more sensible gene expression measuring techniques, such as real-time quantitative polymerase chain reaction (RT-qPCR). Moreover, the association may also be cell-specific, or be present only a at certain time point of the cell development. The association could even be dependent on the activation of a specific pathway. Therefore, further studies are required to elucidate this issue. Understanding the functionality of these SNPs and their potential role in *MYC* regulation could have a relevant impact both at the biological and the clinical level, since it would enable us to advance in our knowledge of the molecular basis of carcinogenesis. Hopefully, all the knowledge obtained from eQTL studies, which combine risk variants with the target genes they regulate, may have an influence on the future treatment of the disease.

Using gene regulatory network inference, our study also unveiled a previously unknown transcriptional interaction between *MYC* and tumor-suppressor *KLF6* in prostate tissue. Moreover, *MYC* and *KLF6* were found to share a large number of transcriptional neighbors, pointing to a strong functional relationship between these two genes, that should be further characterized. An interaction between *MYC* and *KLF6* has also been lately characterized in gastric cancer, reinforcing our previously found relationship between these two genes. [174].

8.3 BIOLOGICAL CONVERGENCE OF CANCER SIGNATURES

Oncologists have traditionally used pathological information to characterize tumors and predict their metastatic potential or their likely response to a specific treatment. However, it seems certain that the clinicopathological attributes of a tumor are not informative enough, since many tumors that are apparently similar at the pathological level do display divergent evolution once they are diagnosed. Indeed, this is evident for CRC, where a 20-30% of non-disseminated stage-II tumors relapse and develop metastasis [98]. If we could correctly detect these high-risk patients they could benefit from more intensive surveillance, while the remaining 70-80% could be mainly treated with surgery and avoid side effects of chemotherapy or radiotherapy. Thus, these pathologically-alike tumors must have differences at the molecular level, making it absolutely necessary to include new tools of diagnosis and prognosis in current clinical routine.

Molecular classification of tumors has recently began to be applied in clinical diagnosis and prognosis procedures. However, these type of molecular characterizations are usually based on just a single or a very

limited number of biomarkers, without incorporating information at the whole genome or transcriptome level. A well-known example of this type of tumor classification occurs for breast cancer, a heterogeneous pathology whose tumors are classified and treated according to the expression levels of *ESR1*, *PGR* and *ERBB2* genes. Nonetheless, it has been argued that these variables are not sufficient for achieving a completely individualized therapy, since some patients still remain unclassified and others who fall within the same molecular group are found to differ in their prognosis [225]. Recently, more in-depth studies have been able to define more specific classifiers using genome-scale gene expression datasets. More interestingly, these new classifications of breast tumors could have potential clinical implications [141, 189]. Therefore, better and more accurate profiling is still required to correctly classify each tumor and prescript a more efficient and less harmful treatment to the patient.

A cancer signature, or profile, could be defined as a set of markers with the potential ability to discern or predict a specific phenotypic aspect of a patient, such as presence or absence of a disease, prognosis or response to a specific treatment, among others. Taking profit of large-scale techniques such as microarrays, in the last few years a large number of cancer signatures have been proposed. Although most of them are based on gene expression data, examples of copy number [233], miRNA [183] and methylation [37] signatures have also appeared recently and promising results in this fields are expected in the mid-term [79, 119]. In this section of the discussion, however, we will specifically focus on gene expression profiles, which are the most abundant and have been a matter of study for a longer time than their copy number, methylation or miRNA counterparts.

Although genetic profiling will definitely be an essential tool for cancer diagnosis and prognosis assessment procedures in the very near future, there are still some concerns that must be addressed before incorporating it into the clinical practice. The first one is related to the little or null overlap in the genes of different profiles that have been designed to predict the same (or very similar) outcome. As a paradigmatic example, this is evident for most of the breast cancer prognosis signatures already published [22, 41, 46, 134, 147, 211, 221]. Although there is not a generalized consensus for this fact, there could be several reasons that may explain this heterogeneity among cancer profiles:

- Technical differences: technology of gene expression monitoring used (RT-qPCR, microarrays, serial analysis of gene expression (SAGE)); genes included in the platform; type of gene-expression measuring (3' arrays vs. whole-transcript arrays).

- Statistical differences: feature selection and classification algorithms; validation strategies; statistical overfitting.
- Biological differences: type of experimental design; number and characteristics of samples used; presence of sample replicates; control of potential batch effects.

A second potential pitfall for the future applicability of gene signatures into the clinic is the controversy about their lack of reproducibility. That is, their poor performance when they are used to classify different individuals than the ones that were used to build the profile [132]. Although this issue may be directly influenced by the technical sources of bias formerly explained, we must be aware that applying a signature over a totally independent set of samples is not a straightforward task. Both datasets may be subject to strong sources of bias, and inaccurate comparisons could lead to misleading results. Moreover, this lower performance has not been observed by some studies comparing that compare different profiles [61]. Therefore, we can state that cancer profiling seems a promising field that can remarkably improve currently used diagnostic and prognostic procedures, provided an accurate experimental design both in the development and the testing stage is used [197](Figure 21).

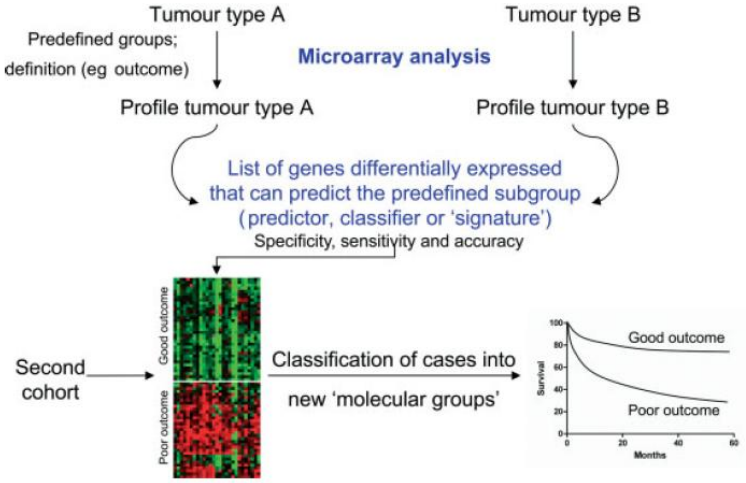


Figure 21: In the experimental design of cancer profiling studies, after identifying the differentially expressed genes between two a-priori predefined sample classes a profile is defined to predict the class membership of a new sample accurately. (Adapted from [224].)

Other approaches, based on systems biology and regulatory networks inference [118], argue that the reason for such unstable and study-dependent cancer profiles are due to the *passenger* role of the genes that comprise them. That is, they are not real *drivers* of the studied outcome (e. g. prognosis). These authors state that the most differentially

expressed genes are usually downstream in the cascade of transcriptional interactions, and due to co-factors and other potential interactors of the transcriptional cascade these downstream genes become unstable. Thus, rather than looking for differentially expressed genes between two phenotypes of interest, they suggest looking for the regulators (they call them *master regulators* (MR)) that are causally responsible for the implementation of the observed differential expression patterns. After applying this methodology for the comparison of two well-known cancer signatures, the authors observed that a common set of MRs for both signatures existed, which also displayed a good classification performance.

Given this great deal of controversy about cancer signatures, the aim of our work was to determine the existence of a putative common tumor cell phenotype associated with different cancer types and conditions, by the study and comparison of the signatures at the genome, transcriptome and interactome levels. Interestingly, our study suggested the existence of common design principles in a system level cellular model, illustrated by transcriptome-interactome correlations, not only of prognostic signatures but also of metastasis and treatment response signatures. More specifically, our work identified common molecular properties and network interactions associated with cell proliferation and death, as well as associations with the immune response. Some previously published studies had also found profile convergence, but only for breast cancer prognosis signatures [168, 186, 229], while our study included a more comprehensive and diverse set of gene expression profiles.

Some concerns about the statistical methodology or the experimental design applied in profiling studies have been recently raised [57]. Regarding our work, Drier et al. [57] have also argued that there is no such biological convergence in cancer signatures, but their conclusions are partial, since they only focus on two signatures of breast cancer prognosis [213, 221]. Furthermore, they are unable to technically criticize our study, and just state that our results may not be reflecting a real convergence of the signatures but only their prognosis potential. However, all their analysis only focuses on the two above mentioned breast cancer signatures, and they do not take into account that our analysis not only included prognosis signatures, but also metastasis and treatment response, which are obviously linked to prognosis but are not strictly the same.

We must also notice that, while our study provided the first evidence of convergence of prognostic, metastasis and predictive signatures in these processes, other processes or signaling pathways are probably represented and diverse specificities may exist. Nonetheless, our work

confirms the usefulness of systems biology approaches for the biological untangling of the complex molecular wiring diagram of the processes participating in cancer signatures.

8.4 STRONG AND WEAK POINTS OF THE STUDIES

The three studies presented in this thesis are based on the analytical integration of different types of large-scale data to unveil diverse aspects about the underlying molecular basis of cancer: a functional characterization of cancer mutated genes; searching for *MYC* cis-regulatory elements related to prostate cancer in 8q24 gene-desert region; and the characterization of a large set of cancer profiles to uncover their potential biological convergence. These type of integrative analyses have been proven to be essential to reach a deeper understanding of the initiation and evolution of the tumorigenic process.

A strong point of this thesis is the multiple and diverse types of data that have been analyzed. This includes DNA data used for assessing SNP association, LOH, copy number or TFBS enrichment analysis. RNA expression data have been used to assess differential gene expression, gene-gene correlations, unsupervised clusterings, survival analysis, functional enrichments (gene set enrichment analysis (GSEA)) and transcriptional regulatory networks inference (algorithm for the reconstruction of accurate cellular networks (ARACNe)). Protein-protein interaction (PPI) data have been used to perform network analysis. Furthermore, DNA (SNP, LOH and copy number) and RNA (gene expression) data have been linked to evaluate the impact of DNA alterations in the transcriptional activity of affected genes. Besides the comprehensive analytical approach, statistical and bioinformatics analysis have been carefully performed. Therefore, we have always taken into account the large number of statistical tests performed and thus have adjusted p-values for multiple testing.

Another interesting point about our proposed work is that our data sources are mainly based on public repositories, such as Gene Expression Omnibus (GEO)[1] or ArrayExpress[2]. This fact demonstrates the great utility of these public databases, which were designed to allow researchers to access and further exploit data previously used for other studies. Public sharing of all generated data is essential to obtain the most out of it, and therefore currently it has become indispensable for publishing results on international peer-reviewed journals.

1 http://www.ncbi.nlm.nih.gov/geo/.
2 http://www.ebi.ac.uk/arrayexpress/.

The use of public data, however, also brings some drawbacks that must not be neglected. In public repositories the quality of the uploaded data is generally not assessed, so we might encounter datasets of low quality which can add a great amount of statistical noise to our analysis. Furthermore, sometimes there is also a lack of information about the experimental design used, or even the information about the hybridized samples is very reduced or missing. If not taken into account, all these facts could preclude us to obtain reliable results from our analysis. Thus, we have always carefully checked the quality of our data, and have used only datasets that have provided the required information to perform our analyses. We are also aware that some of the datasets we have used for data analysis are relatively small, and that has a direct impact in our statistical power to detect biologically relevant differences.

Another limitation of the studies presented in this thesis is the lack of epigenetic data. It is notorious the role of epigenetics in cancer, and the integration of methylation and miRNA expression data with mRNA expression and copy number variation data would unquestionably have yielded more complete results. However, at the time we were conducting the analyses, large-scale miRNA or methylation platforms were still not widely spread, and publicly accessible datasets were very limited and reduced in sample size.

Finally, we would also like to comment on the heterogeneity in the type of cancer studied in each of the three articles. The first work is focused in breast cancer, while the second is focused on prostate cancer and the third can be seen as a multi-cancer analysis, since it includes a diverse set of cancer profiles. The main reason for this lack of specificity is due to the eventual availability of public data on the matter of study. We are aware that the study of a single type of cancer would have offered a more detailed view of one of these pathologies. However, our main aim for this thesis was to demonstrate the usefulness of integrative approaches for the study of complex diseases, such as cancer, more than dealing with a specific tumor localization.

8.5 FUTURE DIRECTIONS: THE COLONOMICS PROJECT

The limitations stated in the previous section about using public data, as well as our awareness about the importance of the integration of multiple types of data for the study of cancer, led our group to embark in an ambitious project that could overcome most if not even all of them. This project was born about a year and a half ago, and it is called COLO-

NOMICS[3].

The COLONOMICS project main objective is to find biomarkers of diagnosis and prognosis for early stage colon cancer. Moreover, its secondary objective is to reach a better understanding of the complex molecular basis underlying this highly-incident tumor. To achieve these goals, we have collected a curated set of homogeneous samples, consisting of paired normal-tumor frozen tissue from 100 stage-II colon cancer from previously untreated patients, who underwent radical surgery. Besides, fresh normal colonic mucosa was also obtained from 50 individuals who appeared to be completely healthy after undergoing a routine colonoscopy. We have also collected blood samples from all the subjects participating in the study for further experiments. Among all the cancer cases, we have 21 patients who relapsed and developed metastasis, with a minimum follow-up of 3 years was required to be included in the study. Overall, the number of samples included in the study is 250 (Fig. 22). We have also collected complete epidemiological information (i. e. anthropometrical measures, dietary habits, tobacco and alcohol consumption, physical activity, former medical prescriptions, family history of cancer). From healthy controls a reduced set of epidemiological information was also obtained. This information can be useful to complement our findings and control for any potential confounding variables.

For each one of these 250 samples, we have extracted DNA and RNA, and have obtained multiple biological information at a genome-wide level. Before performing any experiment we have applied strict quality control procedures on the samples to avoid possible errors during sample manipulation and ensure sample quality. Regarding DNA, we have collected SNP and copy number data (Genome-Wide Human SNP Array 6.0), as well as genome-wide CpG methylation data (Illumina 450K Infinium Methylation BeadChip). As for RNA, mRNA expression (Affymetrix Human Genome U219 Array Plate) and quantitative small-RNA sequencing (Applied Biosystems SOLiD 4) have also been obtained. For a more detailed information about mutational status of the samples, paired normal-tumor complete exome sequencing for 41 of the 100 cases (21 developing metastasis and 20 who did not relapse) is currently undergoing. Furthermore, paired normal-tumor exome sequencing will be soon obtained for 41 cases (21 developing metastasis and 20 not relapsing).

Up to date, we have been working expression data with promising results. We are currently testing potential early-diagnosis candidates as the ones shown in Figure 23. We have also modeled the transcriptional

3 http://www.colonomics.org/.

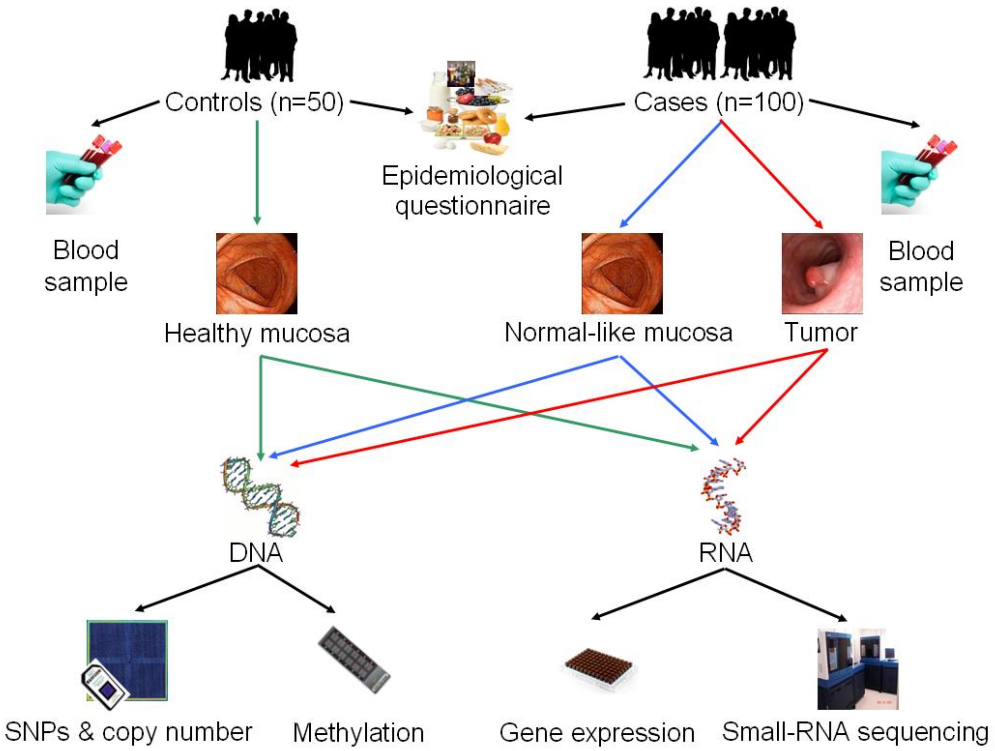


Figure 22: Experimental design of the COLONOMICS project.

networks of normal and tumor sample from cases, obtaining striking differences in both networks (Fig. 24). Regarding prognosis, it seems that gene expression data is not informative enough to accurate predict the metastatic potential of an individual, so we are planning to incorporate copy number, miRNA and methylation information to the predictive models.

With the COLONOMICS project we will hopefully be able to overcome all the limitations of the current studies stated in the previous section. We have a homogeneous, high-quality and curated set of samples, with complete epidemiological and clinical information. Using an accurate experimental design that avoids any potential biases, we have obtained large scale data for all the set of samples, both at the genetic and epigenetic level. Moreover, the consistency of the study is assured, since it is completely focused on colon cancer.

8.6 POTENTIAL IMPACT IN ONCOLOGY: TRANSFERENCE OF KNOWLEDGE INTO THE CLINICAL PRACTICE

In this thesis it has been demonstrated how the application of integrative analytical methods can help researchers to obtain a deeper knowledge of complex diseases, which in our case have been paradigmatically rep-

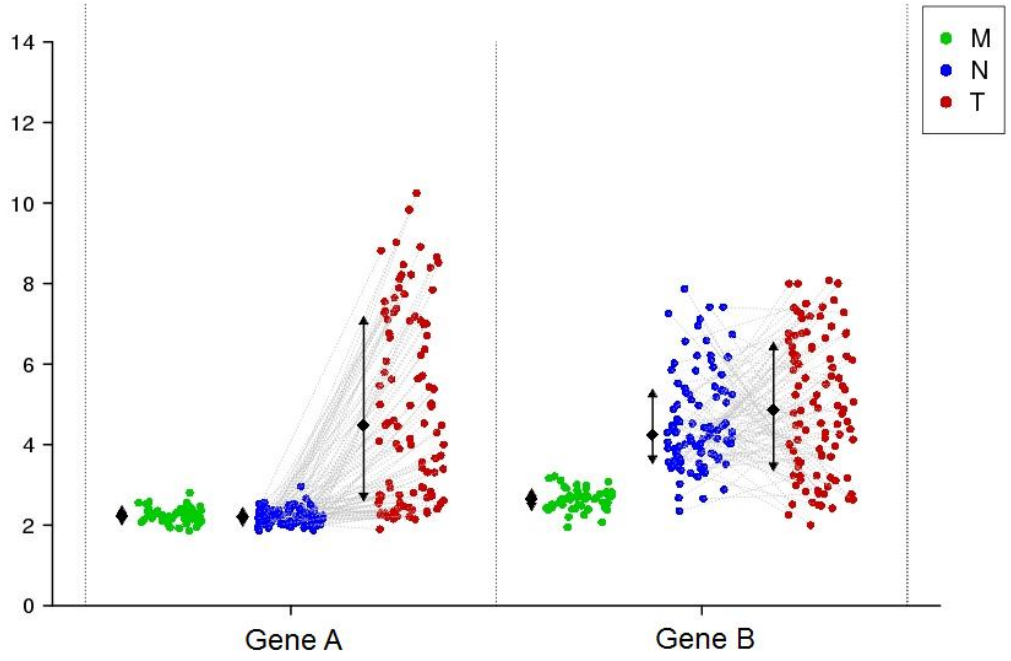


Figure 23: Potential diagnosis biomarkers for early stage colon cancer. Messenger RNA expression levels of two genes, labeled A and B, for healthy mucosa (M) and paired normal (N) and tumor (T) tissue from colon cancer cases. The y-axis represents the intensity value obtained from the microarray hybridizations. Notice that Gene A is only expressed in tumors, while Gene B is highly expressed in both normal and tumor tissue of the cases. Both genes display very low expression levels in healthy individuals. If the proteins resulting from these genes are secreted to the main bloodstream, these two genes could be potential suitable candidates for colon cancer early diagnosis.

resented by cancer. Nonetheless, this deeper understanding of cancer unquestionably needs to be translated into the clinical practice if we want to improve our current management of the pathology.

Systems biology approaches can be helpful in detecting new genes likely involved in the carcinogenic process, or new biomarkers potentially related to cancer diagnosis, prognosis or response to treatment. This is the case of the study where we have analyzed the set of genes mutated in breast tumors. Our integrative analysis was able to prioritize the list of mutated genes and detected *DBN1* as a potential new oncogene which, as well as *SPTAN3*, seemed to be associated with breast cancer survival. Further studies would be required to functionally characterize their biological role in breast cancer tumorigenesis, or to assess their possible utility for breast cancer profiling.

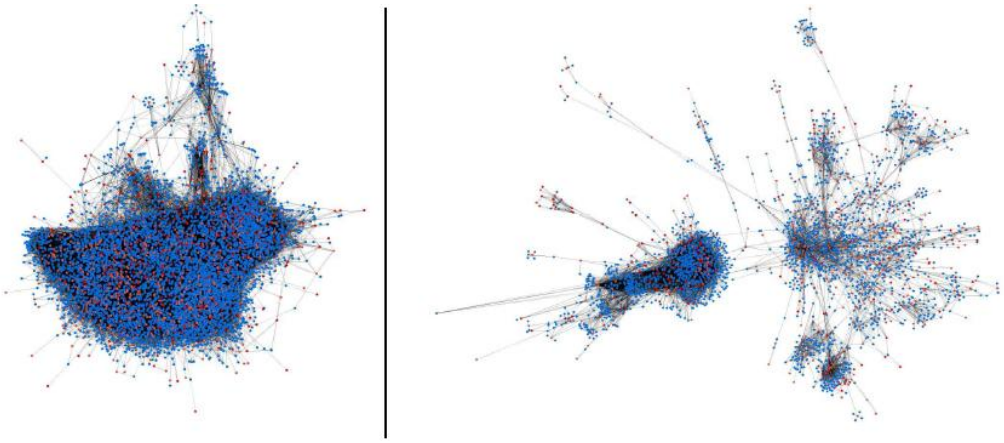


Figure 24: Direct transcriptional regulatory networks of normal (left) and tumor (right) tissue from the same set of individuals. Networks were built with the ARACNe algorithm. Notice the huge differences in number of genes and interactions between both networks. This fact points to a complete deregulation of the transcriptional machinery in colon tumors, which had not been described before.

Integrative methodologies can also be useful to unveil cellular mechanistic processes of cancer susceptibility. By combining gene expression and genotype data it is possible to detect new functional susceptibility variants and study the processes through which they mediate cancer risk. This is what we have presented in the second study, which uncovers a long-range cis-regulatory variant of *MYC* related to prostate cancer susceptibility. Knowledge of all these variants can have a direct clinical impact in procedures of cancer risk assessment. Furthermore, by understanding the mechanistic processes underlying these risk variants it might be possible to modulate the risk before the tumor appears.

Due to the need of more accurate and personalized tools for clinical diagnosis and prognosis assessment, cancer profiling has been a widely studied topic of research in the last years. Once these new molecular-based tests become available, clinicians will no longer be limited to current pathological-based classification criteria. However, the apparent heterogeneity in the signatures designed up to date precludes their imminent application into the clinical routine. Therefore, understanding the relationship between different cancer profiles, which is what we have presented in the third article, can be helpful to understand how signatures biologically relate to each other and eventually can lead researchers to design new reliable and reproducible profiles that could ultimately help clinicians to apply reliable and personalized treatment procedures.

As a complex disease, clinical management of cancer entails so many difficulties that it makes completely necessary to thoroughly comprehend the basis of the disease from different points of view. Since cancer is based on multiple alterations at many molecular levels, only by screening tumor cells at a large scale level and integrating the resulting information we will be able to design better biomarkers for diagnosis and prognosis, or find new potentially therapeutic targets. Therefore, these new systems-paradigm can hopefully help to improve our lives and the lives of those people who will avoid suffering from disease in the mid-term future.

9

CONCLUSIONS

In this thesis we have applied integrative analyses of multiple and heterogeneous large-scale cancer datasets in three different scenarios. The first study aimed to characterize a set of breast cancer mutated genes to identify which of them could be more likely to be related to the oncogenic process. In the second work, we modeled genetic and genomic data to detect and unveil the biological mechanistic processes underlying cancer risk modulation for a specific genomic region associated to different prostate cancer and other type of tumors. Finally, in the third study we characterized multiple cancer signatures at the genome, transcriptome and interactome levels to assess their biological properties and search for a putative common cancer cell phenotype.

Conclusions are exposed for each one of the specific objectives raised at the beginning of this thesis. Finally, a general summarizing conclusion is also exposed.

- Integrative analysis and characterization of breast cancer mutated genes

- The integrative analysis of genomic copy number and expression data strongly indicates that *DBN1* is a candidate oncogene that, when highly expressed in tumors with respect to healthy tissues, predicts poor survival in breast cancer patients.
- Low expression ratios of *ABCA3* and low or medium expression ratios of *SPTAN1* may also predict poor survival in breast cancer patients.
- The interactome analysis of molecular pathways provides new hypotheses for the identification of genes potentially associated with survival outcome. *SPTAN1* interacts with *GRIN2D* and *SLC9A2*, both of which interact with the product of the *ABL1* proto-oncogene. Activated *ABL1* kinase promotes invasion of breast cancer cells. Since low expression ratios of *SPTAN1* predict poor survival, *SPTAN1* could therefore act as a negative regulator of *ABL1* activity.

- Genetic and genomic analysis modeling of germline MYC overexpression and cancer susceptibility

- Quantitative analysis of gene expression in normal prostate tissues supports the model of *MYC* overexpression associated with 8q24-region 1 of prostate cancer risk.
- Germline *MYC* overexpression may promote cellular transformation of the normal epithelium and, by extension, risk of prostate cancer by down-regulating the prostate tumor suppressor *KLF6* gene.

- Biological convergence of cancer signatures

- Significant associations for multiple cancer signatures have been consistently observed across genome, transcriptome and interactome layers, pointing to the existence of a common cancer cell phenotype that decisively influences critical aspects of neoplasia.
- Convergence on cell proliferation and death supports the pivotal involvement of these processes in prognosis, metastasis and treatment response.
- Functional and molecular associations have been identified with the immune response in different cancer types and conditions that complement the contribution of cell proliferation and death.

- General conclusion

- The application of integrative analytical methods to large-scale genomic, transcriptomic and protein interactome data is essential for a better understanding of cancer. Through this systems approach, not only we will better comprehend the molecular basis of the disease, but also we will be able to identify new biomarkers of diagnosis, prognosis, response to treatment and new drug targets, which will could have an ultimate impact in the clinical management of the disease in the following years.

BIBLIOGRAPHY

[1] Finishing the euchromatic sequence of the human genome. *Nature*, 431(7011):931–45, 2004.

[2] Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature*, 455(7216):1061–8, 2008.

[3] Integrated genomic analyses of ovarian carcinoma. *Nature*, 474 (7353):609–15, 2011.

[4] N. Ahmadiyeh, M. M. Pomerantz, C. Grisanzio, P. Herman, L. Jia, V. Almendro, H. H. He, M. Brown, X. S. Liu, M. Davis, J. L. Caswell, C. A. Beckwith, A. Hills, L. Macconaill, G. A. Coetzee, M. M. Regan, and M. L. Freedman. 8q24 prostate, breast, and colon cancer risk loci show tissue-specific long-range interaction with myc. *Proc Natl Acad Sci U S A*, 107(21):9742–6, 2010.

[5] U. D. Akavia, O. Litvin, J. Kim, F. Sanchez-Garcia, D. Kotliar, H. C. Causton, P. Pochanard, E. Mozes, L. A. Garraway, and D. Pe'er. An integrated approach to uncover drivers of cancer. *Cell*, 143(6): 1005–17, 2010.

[6] R. H. Alvarez, V. Valero, and G. N. Hortobagyi. Emerging targeted therapies for breast cancer. *J Clin Oncol*, 28(20):3366–79, 2010.

[7] L. T. Amundadottir, P. Sulem, J. Gudmundsson, A. Helgason, A. Baker, B. A. Agnarsson, A. Sigurdsson, K. R. Benediktsdottir, J. B. Cazier, J. Sainz, M. Jakobsdottir, J. Kostic, D. N. Magnusdottir, S. Ghosh, K. Agnarsson, B. Birgisdottir, L. Le Roux, A. Olafsdottir, T. Blondal, M. Andresdottir, O. S. Gretarsdottir, J. T. Bergthorsson, D. Gudbjartsson, A. Gylfason, G. Thorleifsson, A. Manolescu, K. Kristjansson, G. Geirsson, H. Isaksson, J. Douglas, J. E. Johansson, K. Balter, F. Wiklund, J. E. Montie, X. Yu, B. K. Suarez, C. Ober, K. A. Cooney, H. Gronberg, W. J. Catalona, G. V. Einarsson, R. B. Barkardottir, J. R. Gulcher, A. Kong, U. Thorsteinsdottir, and K. Stefansson. A common variant associated with prostate cancer in european and african populations. *Nat Genet*, 38(6):652–8, 2006.

[8] F. Andre, B. Job, P. Dessen, A. Tordai, S. Michiels, C. Liedtke, C. Richon, K. Yan, B. Wang, G. Vassal, S. Delaloge, G. N. Hortobagyi, W. F. Symmans, V. Lazar, and L. Pusztai. Molecular characterization of breast cancer with high-resolution oligonucleotide comparative genomic hybridization array. *Clin Cancer Res*, 15(2):441–51, 2009.

[9] F. Andre, S. Michiels, P. Dessen, V. Scott, V. Suciu, C. Uzan, V. Lazar, L. Lacroix, G. Vassal, M. Spielmann, P. Vielh, and S. Delaloge. Exonic expression profiling of breast cancer and benign lesions: a retrospective analysis. *Lancet Oncol*, 10(4):381–90, 2009.

[10] M. D. Bacolod, G. S. Schemmann, S. F. Giardina, P. Paty, D. A. Notterman, and F. Barany. Emerging paradigms in cancer genetics: some important findings from high-density single nucleotide polymorphism array studies. *Cancer Res*, 69(3):723–7, 2009.

[11] L. Balducci and W. B. Ershler. Cancer and ageing: a nexus at several levels. *Nat Rev Cancer*, 5(8):655–62, 2005.

[12] B. R. Balsara and J. R. Testa. Chromosomal imbalances in human lung cancer. *Oncogene*, 21(45):6877–83, 2002.

[13] A. J. Bass, M. S. Lawrence, L. E. Brace, A. H. Ramos, Y. Drier, K. Cibulskis, C. Sougnez, D. Voet, G. Saksena, A. Sivachenko, R. Jing, M. Parkin, T. Pugh, R. G. Verhaak, N. Stransky, A. T. Boutin, J. Barretina, D. B. Solit, E. Vakiani, W. Shao, Y. Mishina, M. Warmuth, J. Jimenez, D. Y. Chiang, S. Signoretti, W. G. Kaelin, N. Spardy, W. C. Hahn, Y. Hoshida, S. Ogino, R. A. Depinho, L. Chin, L. A. Garraway, C. S. Fuchs, J. Baselga, J. Tabernero, S. Gabriel, E. S. Lander, G. Getz, and M. Meyerson. Genomic sequencing of colorectal adenocarcinomas identifies a recurrent vti1a-tcf7l2 fusion. *Nat Genet*, 2011.

[14] D. W. Bell. Our changing view of the genomic landscape of cancer. *J Pathol*, 220(2):231–43, 2010.

[15] A. H. Berger and P. P. Pandolfi. Haplo-insufficiency: a driving force in cancer. *J Pathol*, 223(2):137–46, 2011.

[16] A. H. Berger, A. G. Knudson, and P. P. Pandolfi. A continuum model for tumour suppression. *Nature*, 476(7359):163–9, 2011.

[17] M. F. Berger, J. Z. Levin, K. Vijayendran, A. Sivachenko, X. Adiconis, J. Maguire, L. A. Johnson, J. Robinson, R. G. Verhaak, C. Sougnez, R. C. Onofrio, L. Ziaugra, K. Cibulskis, E. Laine, J. Barretina, W. Winckler, D. E. Fisher, G. Getz, M. Meyerson, D. B. Jaffe, S. B. Gabriel, E. S. Lander, R. Dummer, A. Gnirke, C. Nusbaum, and L. A. Garraway. Integrative analysis of the melanoma transcriptome. *Genome Res*, 20(4):413–27, 2010.

[18] S. L. Berger, T. Kouzarides, R. Shiekhattar, and A. Shilatifard. An operational definition of epigenetics. *Genes Dev*, 23(7):781–3, 2009.

[19] E. Bernstein, S. Y. Kim, M. A. Carmell, E. P. Murchison, H. Alcorn, M. Z. Li, A. A. Mills, S. J. Elledge, K. V. Anderson, and G. J. Hannon. Dicer is essential for mouse development. *Nat Genet*, 35(3):215–7, 2003.

[20] M. Bessarabova, O. Pustovalova, W. Shi, T. Serebriyskaya, A. Ishkin, K. Polyak, V. E. Velculescu, T. Nikolskaya, and Y. Nikolsky. Functional synergies yet distinct modulators affected by genetic alterations in common human cancers. *Cancer Res*, 71(10):3471–81, 2011.

[21] G. R. Bignell, C. D. Greenman, H. Davies, A. P. Butler, S. Edkins, J. M. Andrews, G. Buck, L. Chen, D. Beare, C. Latimer, S. Widaa, J. Hinton, C. Fahey, B. Fu, S. Swamy, G. L. Dalgliesh, B. T. Teh, P. Deloukas, F. Yang, P. J. Campbell, P. A. Futreal, and M. R. Stratton. Signatures of mutation and selection in the cancer genome. *Nature*, 463(7283):893–8, 2010.

[22] A. H. Bild, G. Yao, J. T. Chang, Q. Wang, A. Potti, D. Chasse, M. B. Joshi, D. Harpole, J. M. Lancaster, A. Berchuck, Jr. Olson, J. A., J. R. Marks, H. K. Dressman, M. West, and J. R. Nevins. Oncogenic pathway signatures in human cancers as a guide to targeted therapies. *Nature*, 439(7074):353–7, 2006.

[23] E. Birney, J. A. Stamatoyannopoulos, A. Dutta, R. Guigo, T. R. Gingeras, E. H. Margulies, Z. Weng, M. Snyder, E. T. Dermitzakis, R. E. Thurman, M. S. Kuehn, C. M. Taylor, S. Neph, C. M. Koch, S. Asthana, A. Malhotra, I. Adzhubei, J. A. Greenbaum, R. M. Andrews, P. Flicek, P. J. Boyle, H. Cao, N. P. Carter, G. K. Clelland, S. Davis, N. Day, P. Dhami, S. C. Dillon, M. O. Dorschner, H. Fiegler, P. G. Giresi, J. Goldy, M. Hawrylycz, A. Haydock, R. Humbert, K. D. James, B. E. Johnson, E. M. Johnson, T. T. Frum, E. R. Rosenzweig, N. Karnani, K. Lee, G. C. Lefebvre, P. A. Navas, F. Neri, S. C. Parker, P. J. Sabo, R. Sandstrom, A. Shafer, D. Vetrie, M. Weaver, S. Wilcox, M. Yu, F. S. Collins, J. Dekker, J. D. Lieb, T. D. Tullius, G. E. Crawford, S. Sunyaev, W. S. Noble, I. Dunham, F. Denoeud, A. Reymond, P. Kapranov, J. Rozowsky, D. Zheng, R. Castelo, A. Frankish, J. Harrow, S. Ghosh, A. Sandelin, I. L. Hofacker, R. Baertsch, D. Keefe, S. Dike, J. Cheng, H. A. Hirsch, E. A. Sekinger, J. Lagarde, J. F. Abril, A. Shahab, C. Flamm, C. Fried, J. Hackermuller, J. Hertel, M. Lindemeyer, K. Missal, A. Tanzer, S. Washietl, J. Korbel, O. Emanuelsson, J. S. Pedersen, N. Holroyd, R. Taylor, D. Swarbreck, N. Matthews, M. C. Dickson, D. J. Thomas, M. T. Weirauch, J. Gilbert, et al. Identification and analysis of functional elements in 1% of the human genome by the encode pilot project. *Nature*, 447(7146):799–816, 2007.

[24] M. Bredel, D. M. Scholtens, G. R. Harsh, C. Bredel, J. P. Chandler, J. J. Renfrow, A. K. Yadav, H. Vogel, A. C. Scheck, R. Tibshirani, and B. I. Sikic. A network model of a cooperative genetic landscape in brain tumors. *JAMA*, 302(3):261–75, 2009.

[25] P. Brennan. Gene-environment interaction and aetiology of cancer: what does it mean and how can we measure it? *Carcinogenesis*, 23 (3):381–7, 2002.

[26] A. Broeks, M. K. Schmidt, M. E. Sherman, F. J. Couch, J. L. Hopper, G. S. Dite, C. Apicella, L. D. Smith, F. Hammet, M. C. Southey, L. J. Van 't Veer, R. de Groot, V. T. Smit, P. A. Fasching, M. W. Beckmann, S. Jud, A. B. Ekici, A. Hartmann, A. Hein, R. Schulz-Wendtland, B. Burwinkel, F. Marme, A. Schneeweiss, H. P. Sinn, C. Sohn, S. Tchatchou, S. E. Bojesen, B. G. Nordestgaard, H. Flyger, D. D. Orsted, D. Kaur-Knudsen, R. L. Milne, J. I. Perez, P. Zamora, P. M. Rodriguez, J. Benitez, H. Brauch, C. Justenhoven, Y. D. Ko, U. Hamann, H. P. Fischer, T. Bruning, B. Pesch, J. Chang-Claude, S. Wang-Gohrke, M. Bremer, J. H. Karstens, P. Hillemanns, T. Dork, H. A. Nevanlinna, T. Heikkinen, P. Heikkila, C. Blomqvist, K. Aittomaki, K. Aaltonen, A. Lindblom, S. Margolin, A. Mannermaa, V. M. Kosma, J. M. Kauppinen, V. Kataja, P. Auvinen, M. Eskelinen, Y. Soini, G. Chenevix-Trench, A. B. Spurdle, J. Beesley, X. Chen, H. Holland, D. Lambrechts, B. Claes, T. Vandorpe, P. Neven, H. Wildiers, D. Flesch-Janys, R. Hein, T. Loning, M. Kosel, Z. S. Fredericksen, X. Wang, G. G. Giles, L. Baglietto, G. Severi, C. McLean, C. A. Haiman, B. E. Henderson, L. Le Marchand, L. N. Kolonel, G. Grenaker Alnaes, V. Kristensen, A. L. Borresen-Dale, D. J. Hunter, S. E. Hankinson, I. L. Andrulis, A. Marie Mulligan, F. P. O'Malley, P. Devilee, P. E. Huijts, R. A. Tollenaar, C. J. Van Asperen, et al. Low penetrance breast cancer susceptibility loci are associated with specific breast tumor subtypes: findings from the breast cancer association consortium. *Hum Mol Genet*, 20(16):3289–3303, 2011.

[27] M. Brooks. Breast cancer screening and biomarkers. *Methods Mol Biol*, 472:307–21, 2009.

[28] K. H. Buetow, R. D. Klausner, H. Fine, R. Kaplan, D. S. Singer, and R. L. Strausberg. Cancer molecular analysis project: weaving a rich cancer research tapestry. *Cancer Cell*, 1(4):315–8, 2002.

[29] D. L. Burkhart and J. Sage. Cellular mechanisms of tumour suppression by the retinoblastoma gene. *Nat Rev Cancer*, 8(9):671–82, 2008.

[30] B. Cadieux, T. T. Ching, S. R. VandenBerg, and J. F. Costello. Genome-wide hypomethylation in human glioblastomas associated

with specific copy number alteration, methylenetetrahydrofolate reductase allele status, and increased proliferation. *Cancer Res*, 66 (17):8469–76, 2006.

[31] W. W. Cai, J. H. Mao, C. W. Chow, S. Damani, A. Balmain, and A. Bradley. Genome-wide detection of chromosomal imbalances in tumors using bac microarrays. *Nat Biotechnol*, 20(4):393–6, 2002.

[32] G. A. Calin and C. M. Croce. Chromosomal rearrangements and micrornas: a new cancer link with clinical implications. *J Clin Invest*, 117(8):2059–66, 2007.

[33] G. A. Calin, C. D. Dumitru, M. Shimizu, R. Bichi, S. Zupo, E. Noch, H. Aldler, S. Rattan, M. Keating, K. Rai, L. Rassenti, T. Kipps, M. Negrini, F. Bullrich, and C. M. Croce. Frequent deletions and down-regulation of micro- rna genes mir15 and mir16 at 13q14 in chronic lymphocytic leukemia. *Proc Natl Acad Sci U S A*, 99(24): 15524–9, 2002.

[34] G. A. Calin, C. Sevignani, C. D. Dumitru, T. Hyslop, E. Noch, S. Yendamuri, M. Shimizu, S. Rattan, F. Bullrich, M. Negrini, and C. M. Croce. Human microrna genes are frequently located at fragile sites and genomic regions involved in cancers. *Proc Natl Acad Sci U S A*, 101(9):2999–3004, 2004.

[35] P. J. Campbell, P. J. Stephens, E. D. Pleasance, S. O'Meara, H. Li, T. Santarius, L. A. Stebbings, C. Leroy, S. Edkins, C. Hardy, J. W. Teague, A. Menzies, I. Goodhead, D. J. Turner, C. M. Clee, M. A. Quail, A. Cox, C. Brown, R. Durbin, M. E. Hurles, P. A. Edwards, G. R. Bignell, M. R. Stratton, and P. A. Futreal. Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. *Nat Genet*, 40(6):722–9, 2008.

[36] J. Camps, Q. T. Nguyen, H. M. Padilla-Nash, T. Knutsen, N. E. McNeil, D. Wangsa, A. B. Hummon, M. Grade, T. Ried, and M. J. Difilippantonio. Integrative genomics reveals mechanisms of copy number alterations responsible for transcriptional deregulation in colorectal cancer. *Genes Chromosomes Cancer*, 48(11):1002–17, 2009.

[37] F. J. Carmona and M. Esteller. Moving closer to a prognostic dna methylation signature in colon cancer. *Clin Cancer Res*, 17(6):1215–7, 2011.

[38] M. S. Carro, W. K. Lim, M. J. Alvarez, R. J. Bollo, X. Zhao, E. Y. Snyder, E. P. Sulman, S. L. Anne, F. Doetsch, H. Colman, A. Lasorella, K. Aldape, A. Califano, and A. Iavarone. The transcriptional

network for mesenchymal transformation of brain tumours. *Nature*, 463(7279):318–25, 2010.

[39] J. B. Cazier and I. Tomlinson. General lessons from large-scale studies to identify human cancer predisposition genes. *J Pathol*, 220 (2):255–62, 2010.

[40] A. F. Chambers, A. C. Groom, and I. C. MacDonald. Dissemination and growth of cancer cells in metastatic sites. *Nat Rev Cancer*, 2(8): 563–72, 2002.

[41] H. Y. Chang, D. S. Nuyten, J. B. Sneddon, T. Hastie, R. Tibshirani, T. Sorlie, H. Dai, Y. D. He, L. J. van't Veer, H. Bartelink, M. van de Rijn, P. O. Brown, and M. J. van de Vijver. Robustness, scalability, and integration of a wound-response gene expression signature in predicting breast cancer survival. *Proc Natl Acad Sci U S A*, 102(10): 3738–43, 2005.

[42] J. Chen, E. Giovannucci, K. Kelsey, E. B. Rimm, M. J. Stampfer, G. A. Colditz, D. Spiegelman, W. C. Willett, and D. J. Hunter. A methylenetetrahydrofolate reductase polymorphism and the risk of colorectal cancer. *Cancer Res*, 56(21):4862–4, 1996.

[43] Y. C. Chen and D. J. Hunter. Molecular epidemiology of cancer. *CA Cancer J Clin*, 55(1):45–54; quiz 57, 2005.

[44] H. H. Cheung, T. L. Lee, A. J. Davis, D. H. Taft, O. M. Rennert, and W. Y. Chan. Genome-wide dna methylation profiling reveals novel epigenetically regulated genes and non-coding rnas in human testicular cancer. *Br J Cancer*, 102(2):419–27, 2010.

[45] H. W. Cheung, G. S. Cowley, B. A. Weir, J. S. Boehm, S. Rusin, J. A. Scott, A. East, L. D. Ali, P. H. Lizotte, T. C. Wong, G. Jiang, J. Hsiao, C. H. Mermel, G. Getz, J. Barretina, S. Gopal, P. Tamayo, J. Gould, A. Tsherniak, N. Stransky, B. Luo, Y. Ren, R. Drapkin, S. N. Bhatia, J. P. Mesirov, L. A. Garraway, M. Meyerson, E. S. Lander, D. E. Root, and W. C. Hahn. Systematic investigation of genetic vulnerabilities across cancer cell lines reveals lineage-specific dependencies in ovarian cancer. *Proc Natl Acad Sci U S A*, 108(30):12372–7, 2011.

[46] J. T. Chi, Z. Wang, D. S. Nuyten, E. H. Rodriguez, M. E. Schaner, A. Salim, Y. Wang, G. B. Kristensen, A. Helland, A. L. Borresen-Dale, A. Giaccia, M. T. Longaker, T. Hastie, G. P. Yang, M. J. van de Vijver, and P. O. Brown. Gene expression programs in response to hypoxia: cell type specificity and prognostic significance in human cancers. *PLoS Med*, 3(3):e47, 2006.

[47] A. C. Chiang and J. Massague. Molecular basis of metastasis. *N Engl J Med*, 359(26):2814–23, 2008.

[48] F. S. Collins, M. S. Guyer, and A. Charkravarti. Variations on a theme: cataloging human dna sequence variation. *Science*, 278 (5343):1580–1, 1997.

[49] J. F. Costello, B. W. Futscher, K. Tano, D. M. Graunke, and R. O. Pieper. Graded methylation in the promoter and body of the o6-methylguanine dna methyltransferase (mgmt) gene correlates with mgmt expression in human glioma cells. *J Biol Chem*, 269(25): 17228–37, 1994.

[50] Q. Cui. A network of cancer genes with co-occurring and anti-co-occurring mutations. *PLoS One*, 5(10), 2010.

[51] V. Davalos and M. Esteller. Microrнas and cancer epigenetics: a macrorevolution. *Curr Opin Oncol*, 22(1):35–45, 2010.

[52] A. de la Chapelle. Genetic predisposition to human disease: allele-specific expression and low-penetrance regulatory loci. *Oncogene*, 28(38):3345–8, 2009.

[53] Y. Y. Degenhardt, R. Wooster, R. W. McCombie, R. Lucito, and S. Powers. High-content analysis of cancer genome dna alterations. *Curr Opin Genet Dev*, 18(1):68–72, 2008.

[54] J. DeRisi, L. Penland, P. O. Brown, M. L. Bittner, P. S. Meltzer, M. Ray, Y. Chen, Y. A. Su, and J. M. Trent. Use of a cdna microarray to analyse gene expression patterns in human cancer. *Nat Genet*, 14 (4):457–60, 1996.

[55] R. Doll and R. Peto. The causes of cancer: quantitative estimates of avoidable risks of cancer in the united states today. *J Natl Cancer Inst*, 66(6):1191–308, 1981.

[56] J. T. Dong. Chromosomal deletions and tumor suppressor genes in prostate cancer. *Cancer Metastasis Rev*, 20(3-4):173–93, 2001.

[57] Y. Drier and E. Domany. Do two machine-learning based prognostic signatures for breast cancer capture the same biological processes? *PLoS One*, 6(3):e17795, 2011.

[58] M. Ehrlich. Dna methylation in cancer: too much, but also too little. *Oncogene*, 21(35):5400–13, 2002.

[59] V. Emilsson, G. Thorleifsson, B. Zhang, A. S. Leonardson, F. Zink, J. Zhu, S. Carlson, A. Helgason, G. B. Walters, S. Gunnarsdottir,

M. Mouy, V. Steinthorsdottir, G. H. Eiriksdottir, G. Bjornsdottir, I. Reynisdottir, D. Gudbjartsson, A. Helgadottir, A. Jonasdottir, U. Styrkarsdottir, S. Gretarsdottir, K. P. Magnusson, H. Stefansson, R. Fossdal, K. Kristjansson, H. G. Gislason, T. Stefansson, B. G. Leifsson, U. Thorsteinsdottir, J. R. Lamb, J. R. Gulcher, M. L. Reitman, A. Kong, E. E. Schadt, and K. Stefansson. Genetics of gene expression and its effect on disease. *Nature*, 452(7186):423–8, 2008.

[60] S. C. Evans and G. Lozano. The li-fraumeni syndrome: an inherited susceptibility to cancer. *Mol Med Today*, 3(9):390–5, 1997.

[61] C. Fan, D. S. Oh, L. Wessels, B. Weigelt, D. S. Nuyten, A. B. Nobel, L. J. van't Veer, and C. M. Perou. Concordance among gene-expression-based predictors for breast cancer. *N Engl J Med*, 355(6): 560–9, 2006.

[62] J. B. Fan, M. S. Chee, and K. L. Gunderson. Highly parallel genomic assays. *Nat Rev Genet*, 7(8):632–44, 2006.

[63] E. R. Fearon and B. Vogelstein. A genetic model for colorectal tumorigenesis. *Cell*, 61(5):759–67, 1990.

[64] J. Ferlay, D. M. Parkin, and E. Steliarova-Foucher. Estimates of cancer incidence and mortality in europe in 2008. *Eur J Cancer*, 46 (4):765–81, 2010.

[65] J. Ferlay, H. R. Shin, F. Bray, D. Forman, C. Mathers, and D. M. Parkin. Estimates of worldwide burden of cancer in 2008: Globocan 2008. *Int J Cancer*, 2010.

[66] S. Forbes, J. Clements, E. Dawson, S. Bamford, T. Webb, A. Dogan, A. Flanagan, J. Teague, R. Wooster, P. A. Futreal, and M. R. Stratton. Cosmic 2005. *Br J Cancer*, 94(2):318–22, 2006.

[67] S. A. Forbes, G. Bhamra, S. Bamford, E. Dawson, C. Kok, J. Clements, A. Menzies, J. W. Teague, P. A. Futreal, and M. R. Stratton. The catalogue of somatic mutations in cancer (cosmic). *Curr Protoc Hum Genet*, Chapter 10:Unit 10 11, 2008.

[68] W. F. Forrest and G. Cavet. Comment on "the consensus coding sequences of human breast and colorectal cancers". *Science*, 317 (5844):1500; author reply 1500, 2007.

[69] M. F. Fraga, E. Ballestar, M. F. Paz, S. Ropero, F. Setien, M. L. Ballestar, D. Heine-Suner, J. C. Cigudosa, M. Urioste, J. Benitez, M. Boix-Chornet, A. Sanchez-Aguilera, C. Ling, E. Carlsson, P. Poulsen, A. Vaag, Z. Stephan, T. D. Spector, Y. Z. Wu, C. Plass, and M. Esteller. Epigenetic differences arise during the lifetime

of monozygotic twins. *Proc Natl Acad Sci U S A*, 102(30):10604–9, 2005.

[70] M. L. Freedman, C. A. Haiman, N. Patterson, G. J. McDonald, A. Tandon, A. Waliszewska, K. Penney, R. G. Steen, K. Ardlie, E. M. John, I. Oakley-Girvan, A. S. Whittemore, K. A. Cooney, S. A. Ingles, D. Altshuler, B. E. Henderson, and D. Reich. Admixture mapping identifies 8q24 as a prostate cancer risk locus in african-american men. *Proc Natl Acad Sci U S A*, 103(38):14068–73, 2006.

[71] S. Fujita and H. Iba. Putative promoter regions of mirna genes involved in evolutionarily conserved regulatory systems among vertebrates. *Bioinformatics*, 24(3):303–8, 2008.

[72] P. A. Futreal, L. Coin, M. Marshall, T. Down, T. Hubbard, R. Wooster, N. Rahman, and M. R. Stratton. A census of human cancer genes. *Nat Rev Cancer*, 4(3):177–83, 2004.

[73] L. A. Garraway, H. R. Widlund, M. A. Rubin, G. Getz, A. J. Berger, S. Ramaswamy, R. Beroukhim, D. A. Milner, S. R. Granter, J. Du, C. Lee, S. N. Wagner, C. Li, T. R. Golub, D. L. Rimm, M. L. Meyerson, D. E. Fisher, and W. R. Sellers. Integrative genomic analyses identify mitf as a lineage survival oncogene amplified in malignant melanoma. *Nature*, 436(7047):117–22, 2005.

[74] R. Garzon, G. A. Calin, and C. M. Croce. Micrornas in cancer. *Annu Rev Med*, 60:167–79, 2009.

[75] H. Gerdes, Q. Chen, A. H. Elahi, A. Sircar, E. Goldberg, D. Winawer, C. Urmacher, S. J. Winawer, and S. C. Jhanwar. Recurrent deletions involving chromosomes 1, 5, 17, and 18 in colorectal carcinoma: possible role in biological and clinical behavior of tumors. *Anticancer Res*, 15(1):13–24, 1995.

[76] G. Getz, H. Hofling, J. P. Mesirov, T. R. Golub, M. Meyerson, R. Tibshirani, and E. S. Lander. Comment on "the consensus coding sequences of human breast and colorectal cancers". *Science*, 317 (5844):1500, 2007.

[77] X. Y. Goh, J. R. Rees, A. L. Paterson, S. F. Chin, J. C. Marioni, V. Save, M. O'Donovan, P. P. Eijk, D. Alderson, B. Ylstra, C. Caldas, and R. C. Fitzgerald. Integrative analysis of array-comparative genomic hybridisation and matched gene expression profiling data reveals novel genes with prognostic significance in oesophageal adenocarcinoma. *Gut*, page In press, 2011.

[78] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, and E. S. Lander. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, 286(5439):531–7, 1999.

[79] W. M. Grady and M. Tewari. The next thing in prognostic molecular markers: microrna signatures of cancer. *Gut*, 59(6):706–8, 2010.

[80] C. Greenman, P. Stephens, R. Smith, G. L. Dalgliesh, C. Hunter, G. Bignell, H. Davies, J. Teague, A. Butler, C. Stevens, S. Edkins, S. O'Meara, I. Vastrik, E. E. Schmidt, T. Avis, S. Barthorpe, G. Bhamra, G. Buck, B. Choudhury, J. Clements, J. Cole, E. Dicks, S. Forbes, K. Gray, K. Halliday, R. Harrison, K. Hills, J. Hinton, A. Jenkinson, D. Jones, A. Menzies, T. Mironenko, J. Perry, K. Raine, D. Richardson, R. Shepherd, A. Small, C. Tofts, J. Varian, T. Webb, S. West, S. Widaa, A. Yates, D. P. Cahill, D. N. Louis, P. Goldstraw, A. G. Nicholson, F. Brasseur, L. Looijenga, B. L. Weber, Y. E. Chiew, A. DeFazio, M. F. Greaves, A. R. Green, P. Campbell, E. Birney, D. F. Easton, G. Chenevix-Trench, M. H. Tan, S. K. Khoo, B. T. Teh, S. T. Yuen, S. Y. Leung, R. Wooster, P. A. Futreal, and M. R. Stratton. Patterns of somatic mutation in human cancer genomes. *Nature*, 446(7132):153–8, 2007.

[81] D. Hanahan and R. A. Weinberg. The hallmarks of cancer. *Cell*, 100 (1):57–70, 2000.

[82] D. Hanahan and R. A. Weinberg. Hallmarks of cancer: the next generation. *Cell*, 144(5):646–74, 2011.

[83] K. D. Hansen, W. Timp, H. C. Bravo, S. Sabunciyan, B. Langmead, O. G. McDonald, B. Wen, H. Wu, Y. Liu, D. Diep, E. Briem, K. Zhang, R. A. Irizarry, and A. P. Feinberg. Increased methylation variation in epigenetic domains across cancer types. *Nat Genet*, 43(8):768–75, 2011.

[84] J. W. Harper and S. J. Elledge. The dna damage response: ten years after. *Mol Cell*, 28(5):739–45, 2007.

[85] P. Hatzis, L. G. van der Flier, M. A. van Driel, V. Guryev, F. Nielsen, S. Denissov, I. J. Nijman, J. Koster, E. E. Santo, W. Welboren, R. Versteeg, E. Cuppen, M. van de Wetering, H. Clevers, and H. G. Stunnenberg. Genome-wide pattern of tcf7l2/tcf4 chromatin occupancy in colorectal cancer cells. *Mol Cell Biol*, 28(8):2732–44, 2008.

[86] J. G. Herman and S. B. Baylin. Gene silencing in cancer in association with promoter hypermethylation. *N Engl J Med*, 349(21): 2042–54, 2003.

[87] M. Heuser, H. Yun, T. Berg, E. Yung, B. Argiropoulos, F. Kuchenbauer, G. Park, I. Hamwi, L. Palmqvist, C. K. Lai, M. Leung, G. Lin, A. Chaturvedi, B. K. Thakur, M. Iwasaki, M. Bilenky, N. Thiessen, G. Robertson, M. Hirst, D. Kent, N. K. Wilson, B. Gottgens, C. Eaves, M. L. Cleary, M. Marra, A. Ganser, and R. K. Humphries. Cell of origin in aml: susceptibility to mn1-induced transformation is regulated by the meis1/abdb-like hox protein complex. *Cancer Cell*, 20 (1):39–52, 2011.

[88] W. A. High and W. A. Robinson. Genetic mutations involved in melanoma: a summary of our current understanding. *Adv Dermatol*, 23:61–79, 2007.

[89] D. R. Hipfner, K. Weigmann, and S. M. Cohen. The bantam gene regulates drosophila growth. *Genetics*, 161(4):1527–37, 2002.

[90] A. Hollestelle, M. Wasielewski, J. W. Martens, and M. Schutte. Discovering moderate-risk breast cancer susceptibility genes. *Curr Opin Genet Dev*, 20(3):268–76, 2010.

[91] R. S. Houlston and J. Peto. The search for low-penetrance cancer susceptibility alleles. *Oncogene*, 23(38):6471–6, 2004.

[92] C. A. Hudis. Trastuzumab–mechanism of action and use in clinical practice. *N Engl J Med*, 357(1):39–51, 2007.

[93] T. J. Hudson, W. Anderson, A. Artez, A. D. Barker, C. Bell, R. R. Bernabe, M. K. Bhan, F. Calvo, I. Eerola, D. S. Gerhard, A. Guttmacher, M. Guyer, F. M. Hemsley, J. L. Jennings, D. Kerr, P. Klatt, P. Kolar, J. Kusada, D. P. Lane, F. Laplace, L. Youyong, G. Nettekoven, B. Ozenberger, J. Peterson, T. S. Rao, J. Remacle, A. J. Schafer, T. Shibata, M. R. Stratton, J. G. Vockley, K. Watanabe, H. Yang, M. M. Yuen, B. M. Knoppers, M. Bobrow, A. Cambon-Thomsen, L. G. Dressler, S. O. Dyke, Y. Joly, K. Kato, K. L. Kennedy, P. Nicolas, M. J. Parker, E. Rial-Sebbag, C. M. Romeo-Casabona, K. M. Shaw, S. Wallace, G. L. Wiesner, N. Zeps, P. Lichter, A. V. Biankin, C. Chabannon, L. Chin, B. Clement, E. de Alava, F. Degos, M. L. Ferguson, P. Geary, D. N. Hayes, A. L. Johns, A. Kasprzyk, H. Nakagawa, R. Penny, M. A. Piris, R. Sarin, A. Scarpa, M. van de Vijver, P. A. Futreal, H. Aburatani, M. Bayes, D. D. Botwell, P. J. Campbell, X. Estivill, S. M. Grimmond, I. Gut, M. Hirst, C. Lopez-Otin, P. Majumder, M. Marra, J. D. McPherson, Z. Ning, X. S. Puente, Y. Ruan, H. G. Stunnenberg, H. Swerdlow, V. E. Velculescu, R. K. Wilson, H. H. Xue, L. Yang, P. T. Spellman, G. D. Bader, P. C. Boutros, P. Flicek, G. Getz, R. Guigo, G. Guo, D. Haussler, S. Heath, T. J. Hubbard, T. Jiang, et al. International network of cancer genome projects. *Nature*, 464(7291):993–8, 2010.

[94] D. J. Hunter. Gene-environment interactions in human diseases. *Nat Rev Genet*, 6(4):287–98, 2005.

[95] B. Iacopetta, K. Kawakami, and T. Watanabe. Predicting clinical outcome of 5-fluorouracil-based chemotherapy for colon cancer patients: is the cpg island methylator phenotype the 5-fluorouracil-responsive subgroup? *Int J Clin Oncol*, 13(6):498–503, 2008.

[96] J. P. Ioannidis, P. Castaldi, and E. Evangelou. A compendium of genome-wide associations for cancer: critical synopsis and reappraisal. *J Natl Cancer Inst*, 102(12):846–58, 2010.

[97] J. P. Issa. Colon cancer: it's cin or cimp. *Clin Cancer Res*, 14(19): 5939–40, 2008.

[98] P. G. Johnston. Stage ii colorectal cancer: to treat or not to treat. *Oncologist*, 10(5):332–4, 2005.

[99] A. Kallioniemi. Cgh microarrays and cancer. *Curr Opin Biotechnol*, 19(1):36–40, 2008.

[100] M. F. Kane, M. Loda, G. M. Gaida, J. Lipman, R. Mishra, H. Goldman, J. M. Jessup, and R. Kolodner. Methylation of the hmlh1 promoter correlates with lack of expression of hmlh1 in sporadic colon tumors and mismatch repair-defective human tumor cell lines. *Cancer Res*, 57(5):808–11, 1997.

[101] J. Khan, R. Simon, M. Bittner, Y. Chen, S. B. Leighton, T. Pohida, P. D. Smith, Y. Jiang, G. C. Gooden, J. M. Trent, and P. S. Meltzer. Gene expression profiling of alveolar rhabdomyosarcoma with cdna microarrays. *Cancer Res*, 58(22):5009–13, 1998.

[102] L. A. Kiemeney, S. Thorlacius, P. Sulem, F. Geller, K. K. Aben, S. N. Stacey, J. Gudmundsson, M. Jakobsdottir, J. T. Bergthorsson, A. Sigurdsson, T. Blondal, J. A. Witjes, S. H. Vermeulen, C. A. Hulsbergen-van de Kaa, D. W. Swinkels, M. Ploeg, E. B. Cornel, H. Vergunst, T. E. Thorgeirsson, D. Gudbjartsson, S. A. Gudjonsson, G. Thorleifsson, K. T. Kristinsson, M. Mouy, S. Snorradottir, D. Placidi, M. Campagna, C. Arici, K. Koppova, E. Gurzau, P. Rudnai, E. Kellen, S. Polidoro, S. Guarrera, C. Sacerdote, M. Sanchez, B. Saez, G. Valdivia, C. Ryk, P. de Verdier, A. Lindblom, K. Golka, D. T. Bishop, M. A. Knowles, S. Nikulasson, V. Petursdottir, E. Jonsson, G. Geirsson, B. Kristjansson, J. I. Mayordomo, G. Steineck, S. Porru, F. Buntinx, M. P. Zeegers, T. Fletcher, R. Kumar, G. Matullo, P. Vineis, A. E. Kiltie, J. R. Gulcher, U. Thorsteinsdottir, A. Kong, T. Rafnar, and K. Stefansson. Sequence variant on 8q24 confers

susceptibility to urinary bladder cancer. *Nat Genet*, 40(11):1307–12, 2008.

[103] S. S. Knox. From 'omics' to complex disease: a systems biology approach to gene-environment interactions in cancer. *Cancer Cell Int*, 10:11, 2010.

[104] Jr. Knudson, A. G. Mutation and cancer: statistical study of retinoblastoma. *Proc Natl Acad Sci U S A*, 68(4):820–3, 1971.

[105] M. Lagos-Quintana, R. Rauhut, W. Lendeckel, and T. Tuschl. Identification of novel genes coding for small expressed rnas. *Science*, 294(5543):853–8, 2001.

[106] P. W. Laird and R. Jaenisch. Dna methylation and cancer. *Hum Mol Genet*, 3 Spec No:1487–95, 1994.

[107] J. Lamb, S. Ramaswamy, H. L. Ford, B. Contreras, R. V. Martinez, F. S. Kittrell, C. A. Zahnow, N. Patterson, T. R. Golub, and M. E. Ewen. A mechanism of cyclin d1 action encoded in the patterns of gene expression in human cancer. *Cell*, 114(3):323–34, 2003.

[108] J. Lamb, E. D. Crawford, D. Peck, J. W. Modell, I. C. Blat, M. J. Wrobel, J. Lerner, J. P. Brunet, A. Subramanian, K. N. Ross, M. Reich, H. Hieronymus, G. Wei, S. A. Armstrong, S. J. Haggarty, P. A. Clemons, R. Wei, S. A. Carr, E. S. Lander, and T. R. Golub. The connectivity map: using gene-expression signatures to connect small molecules, genes, and disease. *Science*, 313(5795):1929–35, 2006.

[109] E. S. Lander. Initial impact of the sequencing of the human genome. *Nature*, 470(7333):187–97, 2011.

[110] E. S. Lander, L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle, W. FitzHugh, R. Funke, D. Gage, K. Harris, A. Heaford, J. Howland, L. Kann, J. Lehoczky, R. LeVine, P. McEwan, K. McKernan, J. Meldrim, J. P. Mesirov, C. Miranda, W. Morris, J. Naylor, C. Raymond, M. Rosetti, R. Santos, A. Sheridan, C. Sougnez, N. Stange-Thomann, N. Stojanovic, A. Subramanian, D. Wyman, J. Rogers, J. Sulston, R. Ainscough, S. Beck, D. Bentley, J. Burton, C. Clee, N. Carter, A. Coulson, R. Deadman, P. Deloukas, A. Dunham, I. Dunham, R. Durbin, L. French, D. Grafham, S. Gregory, T. Hubbard, S. Humphray, A. Hunt, M. Jones, C. Lloyd, A. McMurray, L. Matthews, S. Mercer, S. Milne, J. C. Mullikin, A. Mungall, R. Plumb, M. Ross, R. Shownkeen, S. Sims, R. H. Waterston, R. K. Wilson, L. W. Hillier, J. D.

McPherson, M. A. Marra, E. R. Mardis, L. A. Fulton, A. T. Chinwalla, K. H. Pepin, W. R. Gish, S. L. Chissoe, M. C. Wendl, K. D. Delehaunty, T. L. Miner, A. Delehaunty, J. B. Kramer, L. L. Cook, R. S. Fulton, D. L. Johnson, P. J. Minx, S. W. Clifton, T. Hawkins, E. Branscomb, P. Predki, P. Richardson, S. Wenning, T. Slezak, N. Doggett, J. F. Cheng, A. Olsen, S. Lucas, C. Elkin, E. Uberbacher, M. Frazier, et al. Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921, 2001.

[111] M. F. Lavin. Ataxia-telangiectasia: from a rare disorder to a paradigm for cell signalling and cancer. *Nat Rev Mol Cell Biol*, 9(10):759–69, 2008.

[112] C. Lefebvre, P. Rajbhandari, M. J. Alvarez, P. Bandaru, W. K. Lim, M. Sato, K. Wang, P. Sumazin, M. Kustagi, B. C. Bisikirska, K. Basso, P. Beltrao, N. Krogan, J. Gautier, R. Dalla-Favera, and A. Califano. A human b-cell interactome identifies myb and foxm1 as master regulators of proliferation in germinal centers. *Mol Syst Biol*, 6:377, 2010.

[113] S. L'Esperance, I. Popa, M. Bachvarova, M. Plante, N. Patten, L. Wu, B. Tetu, and D. Bachvarov. Gene expression profiling of paired ovarian tumors obtained prior to and following adjuvant chemotherapy: molecular signatures of chemoresistant tumors. *Int J Oncol*, 29(1): 5–24, 2006.

[114] R. Lewin. "computer genome" is full of junk dna. *Science*, 232(4750): 577–8, 1986.

[115] E. Li. Chromatin modification and epigenetic reprogramming in mammalian development. *Nat Rev Genet*, 3(9):662–73, 2002.

[116] W. Li, C. A. Meyer, and X. S. Liu. A hidden markov model for analyzing chip-chip experiments on genome tiling arrays and its application to p53 binding sequences. *Bioinformatics*, 21 Suppl 1: i274–82, 2005.

[117] P. Lichtenstein, N. V. Holm, P. K. Verkasalo, A. Iliadou, J. Kaprio, M. Koskenvuo, E. Pukkala, A. Skytthe, and K. Hemminki. Environmental and heritable factors in the causation of cancer–analyses of cohorts of twins from sweden, denmark, and finland. *N Engl J Med*, 343(2):78–85, 2000.

[118] W. K. Lim, E. Lyashenko, and A. Califano. Master regulators used as breast cancer metastasis classifier. *Pac Symp Biocomput*, pages 504–15, 2009.

[119] S. C. Lima, H. Hernandez-Vargas, and Z. Herceg. Epigenetic signatures in cancer: Implications for the control of cancer in the clinic. *Curr Opin Mol Ther*, 12(3):316–24, 2010.

[120] L. A. Loeb. Mutator phenotype may be required for multistage carcinogenesis. *Cancer Res*, 51(12):3075–9, 1991.

[121] S. W. Lowe and A. W. Lin. Apoptosis in cancer. *Carcinogenesis*, 21 (3):485–95, 2000.

[122] L. Luzzatto. The mechanisms of neoplastic transformation. *Eur J Cancer*, 37 Suppl 8:S114–7, 2001.

[123] H. T. Lynch, R. E. Brand, C. A. Deters, T. G. Shaw, and J. F. Lynch. Hereditary pancreatic cancer. *Pancreatology*, 1(5):466–71, 2001.

[124] H. T. Lynch, J. F. Lynch, P. M. Lynch, and T. Attard. Hereditary colorectal cancer syndromes: molecular genetics, genetic counseling, diagnosis and management. *Fam Cancer*, 7(1):27–39, 2008.

[125] C. Mackintosh, J. L. Ordonez, D. J. Garcia-Dominguez, V. Sevillano, A. Llombart-Bosch, K. Szuhai, K. Scotlandi, M. Alberghini, R. Sciot, F. Sinnaeve, P. C. Hogendoorn, P. Picci, S. Knuutila, U. Dirksen, M. Debiec-Rychter, K. L. Schaefer, and E. de Alava. 1q gain and cdt2 overexpression underlie an aggressive and highly proliferative form of ewing sarcoma. *Oncogene*, 2011.

[126] A. A. Margolin, T. Palomero, P. Sumazin, A. Califano, A. A. Ferrando, and G. Stolovitzky. Chip-on-chip significance analysis reveals large-scale binding and regulation by human transcription factor oncogenes. *Proc Natl Acad Sci U S A*, 106(1):244–9, 2009.

[127] S. D. Markowitz and M. M. Bertagnolli. Molecular origins of cancer: Molecular basis of colorectal cancer. *N Engl J Med*, 361(25):2449–60, 2009.

[128] C. A. Maxwell, V. Moreno, X. Sole, L. Gomez, P. Hernandez, A. Urruticoechea, and M. A. Pujana. Genetic interactions: the missing links for a better understanding of cancer susceptibility, progression and treatment. *Mol Cancer*, 7:4, 2008.

[129] S. A. Melo and M. Esteller. Dysregulation of microrna in cancer: playing with fire. *FEBS Lett*, 585(13):2087–99, 2011.

[130] M. Meyerson. Cancer: broken genes in solid tumours. *Nature*, 448 (7153):545–6, 2007.

[131] M. Meyerson, S. Gabriel, and G. Getz. Advances in understanding cancer genomes through second-generation sequencing. *Nat Rev Genet*, 11(10):685–96, 2010.

[132] S. Michiels, S. Koscielny, and C. Hill. Prediction of cancer outcome with microarrays: a multiple random validation strategy. *Lancet*, 365(9458):488–92, 2005.

[133] S. J. Miller, W. J. Jessen, T. Mehta, A. Hardiman, E. Sites, S. Kaiser, A. G. Jegga, H. Li, M. Upadhyaya, M. Giovannini, D. Muir, M. R. Wallace, E. Lopez, E. Serra, G. P. Nielsen, C. Lazaro, A. Stemmer-Rachamimov, G. Page, B. J. Aronow, and N. Ratner. Integrative genomic analyses of neurofibromatosis tumours identify sox9 as a biomarker and survival gene. *EMBO Mol Med*, 1(4):236–48, 2009.

[134] A. J. Minn, G. P. Gupta, P. M. Siegel, P. D. Bos, W. Shu, D. D. Giri, A. Viale, A. B. Olshen, W. L. Gerald, and J. Massague. Genes that mediate breast cancer metastasis to lung. *Nature*, 436(7050):518–24, 2005.

[135] F. Mitelman, B. Johansson, and F. Mertens. The impact of translocations and gene fusions on cancer causation. *Nat Rev Cancer*, 7(4): 233–45, 2007.

[136] R. Mitra, J. Lee, J. Jo, M. Milani, J. N. McClintick, H. J. Edenberg, K. A. Kesler, K. M. Rieger, S. Badve, O. W. Cummings, A. Mohiuddin, D. G. Thomas, X. Luo, B. E. Juliar, L. Li, C. Mesaros, I. A. Blair, A. Srirangam, R. A. Kratzke, C. J. McDonald, J. Kim, and D. A. Potter. Prediction of postoperative recurrence-free survival in non-small cell lung cancer by using an internationally validated gene expression model. *Clin Cancer Res*, 17(9):2934–46, 2011.

[137] R. Natrajan, B. Weigelt, A. Mackay, F. C. Geyer, A. Grigoriadis, D. S. Tan, C. Jones, C. J. Lord, R. Vatcheva, S. M. Rodriguez-Pinilla, J. Palacios, A. Ashworth, and J. S. Reis-Filho. An integrative genomic and transcriptomic analysis reveals molecular pathways and networks regulated by copy number aberrations in basal-like, her2 and luminal cancers. *Breast Cancer Res Treat*, 121(3):575–89, 2010.

[138] T. Neff and S. A. Armstrong. Chromatin maps, histone modifications and leukemia. *Leukemia*, 23(7):1243–51, 2009.

[139] K. P. Nephew and T. H. Huang. Epigenetic gene silencing in cancer initiation and progression. *Cancer Lett*, 190(2):125–33, 2003.

[140] R. K. Nibbe, M. Koyuturk, and M. R. Chance. An integrative -omics approach to identify functional sub-networks in human colorectal cancer. *PLoS Comput Biol*, 6(1):e1000639, 2010.

[141] M. Nicolau, A. J. Levine, and G. Carlsson. Topology based data analysis identifies a subgroup of breast cancers with a unique mutational profile and excellent survival. *Proc Natl Acad Sci U S A*, 108(17):7265–70, 2011.

[142] H. Noushmehr, D. J. Weisenberger, K. Diefes, H. S. Phillips, K. Pujara, B. P. Berman, F. Pan, C. E. Pelloski, E. P. Sulman, K. P. Bhat, R. G. Verhaak, K. A. Hoadley, D. N. Hayes, C. M. Perou, H. K. Schmidt, L. Ding, R. K. Wilson, D. Van Den Berg, H. Shen, H. Bengtsson, P. Neuvial, L. M. Cope, J. Buckley, J. G. Herman, S. B. Baylin, P. W. Laird, and K. Aldape. Identification of a cpg island methylator phenotype that defines a distinct subgroup of glioma. *Cancer Cell*, 17(5):510–22, 2010.

[143] P. C. Nowell. The clonal evolution of tumor cell populations. *Science*, 194(4260):23–8, 1976.

[144] P. C. Nowell. Discovery of the philadelphia chromosome: a personal perspective. *J Clin Invest*, 117(8):2033–5, 2007.

[145] E. R. Okawa, T. Gotoh, J. Manne, J. Igarashi, T. Fujita, K. A. Silverman, H. Xhao, Y. P. Mosse, P. S. White, and G. M. Brodeur. Expression and sequence analysis of candidates for the 1p36.31 tumor suppressor gene deleted in neuroblastomas. *Oncogene*, 27(6): 803–10, 2008.

[146] E. Padron, R. Komrokji, and A. F. List. Biology and treatment of the 5q- syndrome. *Expert Rev Hematol*, 4(1):61–9, 2011.

[147] S. Paik, S. Shak, G. Tang, C. Kim, J. Baker, M. Cronin, F. L. Baehner, M. G. Walker, D. Watson, T. Park, W. Hiller, E. R. Fisher, D. L. Wickerham, J. Bryant, and N. Wolmark. A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. *N Engl J Med*, 351(27):2817–26, 2004.

[148] D. M. Parkin. The evolution of the population-based cancer registry. *Nat Rev Cancer*, 6(8):603–12, 2006.

[149] G. Parmigiani, J. Lin, S. M. Boca, T. Sjöblom, S. Jones, L. D. Wood, D. W. Parsons, T. Barber, P. Buckhaults, S. D. Markowitz, B. H. Park, K. E. Bachman, N. Papadopoulos, B. Vogelstein, K. W. Kinzler, and V. E. Velculescu. Response to comments on "the consensus coding sequences of human breast and colorectal cancers". *Science*, 317 (5844):1500, 2007.

[150] U. Pastorino. Lung cancer: diagnosis and surgery. *Eur J Cancer*, 37 Suppl 7:S75–90, 2001.

[151] Y. Pawitan, K. C. Seng, and P. K. Magnusson. How many genetic variants remain to be discovered? *PLoS One*, 4(12):e7969, 2009.

[152] H. Penzlin. The riddle of "life," a biologist's critical view. *Naturwissenschaften*, 96(1):1–23, 2009.

[153] J. Peto. Cancer epidemiology in the last century and the next decade. *Nature*, 411(6835):390–5, 2001.

[154] J. Peto and R. S. Houlston. Genetics and the common cancers. *Eur J Cancer*, 37 Suppl 8:S88–96, 2001.

[155] P. D. Pharoah, A. M. Dunning, B. A. Ponder, and D. F. Easton. Association studies for finding cancer-susceptibility genetic variants. *Nat Rev Cancer*, 4(11):850–60, 2004.

[156] R. S. Pillai, S. N. Bhattacharyya, C. G. Artus, T. Zoller, N. Cougot, E. Basyuk, E. Bertrand, and W. Filipowicz. Inhibition of translational initiation by let-7 microrna in human cells. *Science*, 309(5740): 1573–6, 2005.

[157] M. Pineda, S. Gonzalez, C. Lazaro, I. Blanco, and G. Capella. Detection of genetic alterations in hereditary colorectal cancer screening. *Mutat Res*, 693(1-2):19–31, 2010.

[158] D. Pinkel, R. Segraves, D. Sudar, S. Clark, I. Poole, D. Kowbel, C. Collins, W. L. Kuo, C. Chen, Y. Zhai, S. H. Dairkee, B. M. Ljung, J. W. Gray, and D. G. Albertson. High resolution analysis of dna copy number variation using comparative genomic hybridization to microarrays. *Nat Genet*, 20(2):207–11, 1998.

[159] J. R. Pollack, C. M. Perou, A. A. Alizadeh, M. B. Eisen, A. Pergamenschikov, C. F. Williams, S. S. Jeffrey, D. Botstein, and P. O. Brown. Genome-wide analysis of dna copy-number changes using cdna microarrays. *Nat Genet*, 23(1):41–6, 1999.

[160] M. M. Pomerantz, N. Ahmadiyeh, L. Jia, P. Herman, M. P. Verzi, H. Doddapaneni, C. A. Beckwith, J. A. Chan, A. Hills, M. Davis, K. Yao, S. M. Kehoe, H. J. Lenz, C. A. Haiman, C. Yan, B. E. Henderson, B. Frenkel, J. Barretina, A. Bass, J. Tabernero, J. Baselga, M. M. Regan, J. R. Manak, R. Shivdasani, G. A. Coetzee, and M. L. Freedman. The 8q24 cancer risk variant rs6983267 shows long-range interaction with myc in colorectal cancer. *Nat Genet*, 41(8):882–4, 2009.

[161] M. M. Pomerantz, C. A. Beckwith, M. M. Regan, S. K. Wyman, G. Petrovics, Y. Chen, D. J. Hawksworth, F. R. Schumacher, L. Mucci, K. L. Penney, M. J. Stampfer, J. A. Chan, K. G. Ardlie, B. R. Fritz,

R. K. Parkin, D. W. Lin, M. Dyke, P. Herman, S. Lee, W. K. Oh, P. W. Kantoff, M. Tewari, D. G. McLeod, S. Srivastava, and M. L. Freedman. Evaluation of the 8q24 prostate cancer risk locus and myc expression. *Cancer Res*, 69(13):5568–74, 2009.

[162] J. R. Prensner, M. K. Iyer, O. A. Balbin, S. M. Dhanasekaran, Q. Cao, J. C. Brenner, B. Laxman, I. A. Asangani, C. S. Grasso, H. D. Kominsky, X. Cao, X. Jing, X. Wang, J. Siddiqui, J. T. Wei, D. Robinson, H. K. Iyer, N. Palanisamy, C. A. Maher, and A. M. Chinnaiyan. Transcriptome sequencing across a prostate cancer cohort identifies pcat-1, an unannotated lincrna implicated in disease progression. *Nat Biotechnol*, 29(8):742–9, 2011.

[163] I. R. Radford. Chromosomal rearrangement as the basis for human tumourigenesis. *Int J Radiat Biol*, 80(8):543–57, 2004.

[164] S. Ramaswamy. Translating cancer genomics into clinical oncology. *N Engl J Med*, 350(18):1814–6, 2004.

[165] R. Redon, S. Ishikawa, K. R. Fitch, L. Feuk, G. H. Perry, T. D. Andrews, H. Fiegler, M. H. Shapero, A. R. Carson, W. Chen, E. K. Cho, S. Dallaire, J. L. Freeman, J. R. Gonzalez, M. Gratacos, J. Huang, D. Kalaitzopoulos, D. Komura, J. R. MacDonald, C. R. Marshall, R. Mei, L. Montgomery, K. Nishimura, K. Okamura, F. Shen, M. J. Somerville, J. Tchinda, A. Valsesia, C. Woodwark, F. Yang, J. Zhang, T. Zerjal, L. Armengol, D. F. Conrad, X. Estivill, C. Tyler-Smith, N. P. Carter, H. Aburatani, C. Lee, K. W. Jones, S. W. Scherer, and M. E. Hurles. Global variation in copy number in the human genome. *Nature*, 444(7118):444–54, 2006.

[166] J. F. Reid, M. Gariboldi, V. Sokolova, P. Capobianco, A. Lampis, F. Perrone, S. Signoroni, A. Costa, E. Leo, S. Pilotti, and M. A. Pierotti. Integrative approach for prioritizing cancer genes in sporadic colon cancer. *Genes Chromosomes Cancer*, 48(11):953–62, 2009.

[167] B. J. Reinhart, F. J. Slack, M. Basson, A. E. Pasquinelli, J. C. Bettinger, A. E. Rougvie, H. R. Horvitz, and G. Ruvkun. The 21-nucleotide let-7 rna regulates developmental timing in caenorhabditis elegans. *Nature*, 403(6772):901–6, 2000.

[168] F. Reyal, M. H. van Vliet, N. J. Armstrong, H. M. Horlings, K. E. de Visser, M. Kok, A. E. Teschendorff, S. Mook, L. van 't Veer, C. Caldas, R. J. Salmon, M. J. van de Vijver, and L. F. Wessels. A comprehensive analysis of prognostic signatures reveals the high predictive capacity of the proliferation, immune response and rna splicing modules in breast cancer. *Breast Cancer Res*, 10(6):R93, 2008.

[169] F. M. Richards. Molecular pathology of von hippellindau disease and the vhl tumour suppressor gene. *Expert Rev Mol Med*, 2001: 1–27, 2001.

[170] M. V. Rockman and L. Kruglyak. Genetics of global gene expression. *Nat Rev Genet*, 7(11):862–72, 2006.

[171] A. F. Rubin and P. Green. Comment on "the consensus coding sequences of human breast and colorectal cancers". *Science*, 317 (5844):1500, 2007.

[172] Y. Ruike, Y. Imanaka, F. Sato, K. Shimizu, and G. Tsujimoto. Genome-wide analysis of aberrant methylation in human breast cancer cells using methyl-dna immunoprecipitation combined with high-throughput sequencing. *BMC Genomics*, 11:137, 2010.

[173] M. Sanchez-Beato, A. Sanchez-Aguilera, and M. A. Piris. Cell cycle deregulation in b-cell lymphomas. *Blood*, 101(4):1220–35, 2003.

[174] J. Sangodkar, J. Shi, A. DiFeo, R. Schwartz, R. Bromberg, A. Choudhri, K. McClinch, R. Hatami, E. Scheer, S. Kremer-Tal, J. A. Martignetti, A. Hui, W. K. Leung, S. L. Friedman, and G. Narla. Functional role of the klf6 tumour suppressor gene in gastric cancer. *Eur J Cancer*, 45(4):666–76, 2009.

[175] I. Sansal and W. R. Sellers. The biology and clinical relevance of the pten tumor suppressor pathway. *J Clin Oncol*, 22(14):2954–63, 2004.

[176] M. Santarosa and A. Ashworth. Haploinsufficiency for tumour suppressor genes: when you don't need to go all the way. *Biochim Biophys Acta*, 1654(2):105–22, 2004.

[177] M. Schena, D. Shalon, R. W. Davis, and P. O. Brown. Quantitative monitoring of gene expression patterns with a complementary dna microarray. *Science*, 270(5235):467–70, 1995.

[178] S. Schimanski, P. J. Wild, O. Treeck, F. Horn, A. Sigruener, C. Rudolph, H. Blaszyk, M. Klinkhammer-Schalke, O. Ortmann, A. Hartmann, and G. Schmitz. Expression of the lipid transporters abca3 and abca1 is diminished in human breast cancer tissue. *Horm Metab Res*, 42(2):102–9, 2010.

[179] F. R. Schumacher, H. S. Feigelson, D. G. Cox, C. A. Haiman, D. Albanes, J. Buring, E. E. Calle, S. J. Chanock, G. A. Colditz, W. R. Diver, A. M. Dunning, M. L. Freedman, J. M. Gaziano, E. Giovannucci, S. E. Hankinson, R. B. Hayes, B. E. Henderson, R. N. Hoover, R. Kaaks, T. Key, L. N. Kolonel, P. Kraft, L. Le Marchand, J. Ma,

M. C. Pike, E. Riboli, M. J. Stampfer, D. O. Stram, G. Thomas, M. J. Thun, R. Travis, J. Virtamo, G. Andriole, E. Gelmann, W. C. Willett, and D. J. Hunter. A common 8q24 variant in prostate and breast cancer from a large nested case-control study. *Cancer Res*, 67(7): 2951–6, 2007.

[180] S. Seemann, D. Maurici, M. Olivier, C. C. de Fromentel, and P. Hainaut. The tumor suppressor gene tp53: implications for cancer management and therapy. *Crit Rev Clin Lab Sci*, 41(5-6):551–83, 2004.

[181] E. Segal, N. Friedman, D. Koller, and A. Regev. A module map showing conditional activity of expression modules in cancer. *Nat Genet*, 36(10):1090–8, 2004.

[182] S. Segditsas, A. J. Rowan, K. Howarth, A. Jones, S. Leedham, N. A. Wright, P. Gorman, W. Chambers, E. Domingo, R. R. Roylance, E. J. Sawyer, O. M. Sieber, and I. P. Tomlinson. Apc and the three-hit hypothesis. *Oncogene*, 28(1):146–55, 2009.

[183] M. F. Segura, I. Belitskaya-Levy, A. E. Rose, J. Zakrzewski, A. Gaziel, D. Hanniford, F. Darvishian, R. S. Berman, R. L. Shapiro, A. C. Pavlick, I. Osman, and E. Hernando. Melanoma microrna signature predicts post-recurrence survival. *Clin Cancer Res*, 16(5):1577–86, 2010.

[184] D. R. Shaffer and P. P. Pandolfi. Breaking the rules of cancer. *Nat Med*, 12(1):14–5, 2006.

[185] Jr. Shaughnessy, J. D. and B. Barlogie. Integrating cytogenetics and gene expression profiling in the molecular analysis of multiple myeloma. *Int J Hematol*, 76 Suppl 2:59–64, 2002.

[186] R. Shen, A. M. Chinnaiyan, and D. Ghosh. Pathway analysis reveals functional convergence of gene expression profiles in breast cancer. *BMC Med Genomics*, 1:28, 2008.

[187] A. Shlien and D. Malkin. Copy number variations and cancer. *Genome Med*, 1(6):62, 2009.

[188] T. Sjoblom, S. Jones, L. D. Wood, D. W. Parsons, J. Lin, T. D. Barber, D. Mandelker, R. J. Leary, J. Ptak, N. Silliman, S. Szabo, P. Buckhaults, C. Farrell, P. Meeh, S. D. Markowitz, J. Willis, D. Dawson, J. K. Willson, A. F. Gazdar, J. Hartigan, L. Wu, C. Liu, G. Parmigiani, B. H. Park, K. E. Bachman, N. Papadopoulos, B. Vogelstein, K. W. Kinzler, and V. E. Velculescu. The consensus coding sequences of human breast and colorectal cancers. *Science*, 314(5797):268–74, 2006.

[189] T. Sorlie, C. M. Perou, R. Tibshirani, T. Aas, S. Geisler, H. Johnsen, T. Hastie, M. B. Eisen, M. van de Rijn, S. S. Jeffrey, T. Thorsen, H. Quist, J. C. Matese, P. O. Brown, D. Botstein, P. Eystein Lonning, and A. L. Borresen-Dale. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc Natl Acad Sci U S A*, 98(19):10869–74, 2001.

[190] J. Sotelo, D. Esposito, M. A. Duhagon, K. Banfield, J. Mehalko, H. Liao, R. M. Stephens, T. J. Harris, D. J. Munroe, and X. Wu. Long-range enhancers on 8q24 regulate c-myc. *Proc Natl Acad Sci U S A*, 107(7):3001–5, 2010.

[191] T. Soussi and C. Beroud. Assessing tp53 status in human tumours to evaluate clinical outcome. *Nat Rev Cancer*, 1(3):233–40, 2001.

[192] M. B. Sporn. The war on cancer. *Lancet*, 347(9012):1377–81, 1996.

[193] A. Stahl, N. Levy, T. Wadzynska, J. M. Sussan, D. Jourdan-Fonta, and J. B. Saracco. The genetics of retinoblastoma. *Ann Genet*, 37(4): 172–8, 1994.

[194] C. Steinhoff and M. Vingron. Normalization and quantification of differential expression in gene expression microarrays. *Brief Bioinform*, 7(2):166–77, 2006.

[195] M. R. Stratton. Exploring the genomes of cancer cells: progress and promise. *Science*, 331(6024):1553–8, 2011.

[196] S. Subbaram, M. Kuentzel, D. Frank, C. M. Dipersio, and S. V. Chittur. Determination of alternate splicing events using the affymetrix exon 1.0 st arrays. *Methods Mol Biol*, 632:63–72, 2010.

[197] J. Subramanian and R. Simon. What should physicians look for in evaluating prognostic gene-expression signatures? *Nat Rev Clin Oncol*, 7(6):327–34, 2010.

[198] B. S. Taylor, N. Schultz, H. Hieronymus, A. Gopalan, Y. Xiao, B. S. Carver, V. K. Arora, P. Kaushik, E. Cerami, B. Reva, Y. Antipin, N. Mitsiades, T. Landers, I. Dolgalev, J. E. Major, M. Wilson, N. D. Socci, A. E. Lash, A. Heguy, J. A. Eastham, H. I. Scher, V. E. Reuter, P. T. Scardino, C. Sander, C. L. Sawyers, and W. L. Gerald. Integrative genomic profiling of human prostate cancer. *Cancer Cell*, 18(1): 11–22, 2010.

[199] A. E. Teschendorff and C. Caldas. The breast cancer somatic 'mutaome': tackling the complexity. *Breast Cancer Res*, 11(2):301, 2009.

[200] S. A. Tomlins, D. R. Rhodes, S. Perner, S. M. Dhanasekaran, R. Mehra, X. W. Sun, S. Varambally, X. Cao, J. Tchinda, R. Kuefer, C. Lee, J. E. Montie, R. B. Shah, K. J. Pienta, M. A. Rubin, and A. M. Chinnaiyan. Recurrent fusion of tmprss2 and ets transcription factor genes in prostate cancer. *Science*, 310(5748):644–8, 2005.

[201] S. A. Tomlins, M. A. Rubin, and A. M. Chinnaiyan. Integrative biology of prostate cancer progression. *Annu Rev Pathol*, 1:243–71, 2006.

[202] S. A. Tomlins, B. Laxman, S. M. Dhanasekaran, B. E. Helgeson, X. Cao, D. S. Morris, A. Menon, X. Jing, Q. Cao, B. Han, J. Yu, L. Wang, J. E. Montie, M. A. Rubin, K. J. Pienta, D. Roulston, R. B. Shah, S. Varambally, R. Mehra, and A. M. Chinnaiyan. Distinct classes of chromosomal rearrangements create oncogenic ets gene fusions in prostate cancer. *Nature*, 448(7153):595–9, 2007.

[203] I. P. Tomlinson, R. Roylance, and R. S. Houlston. Two hits revisited again. *J Med Genet*, 38(2):81–5, 2001.

[204] I. P. Tomlinson, M. B. Lambros, and R. R. Roylance. Loss of heterozygosity analysis: practically and conceptually flawed? *Genes Chromosomes Cancer*, 34(4):349–53, 2002.

[205] I. P. Tomlinson, E. Webb, L. Carvajal-Carmona, P. Broderick, K. Howarth, A. M. Pittman, S. Spain, S. Lubbe, A. Walther, K. Sullivan, E. Jaeger, S. Fielding, A. Rowan, J. Vijayakrishnan, E. Domingo, I. Chandler, Z. Kemp, M. Qureshi, S. M. Farrington, A. Tenesa, J. G. Prendergast, R. A. Barnetson, S. Penegar, E. Barclay, W. Wood, L. Martin, M. Gorman, H. Thomas, J. Peto, D. T. Bishop, R. Gray, E. R. Maher, A. Lucassen, D. Kerr, D. G. Evans, C. Schafmayer, S. Buch, H. Volzke, J. Hampe, S. Schreiber, U. John, T. Koessler, P. Pharoah, T. van Wezel, H. Morreau, J. T. Wijnen, J. L. Hopper, M. C. Southey, G. G. Giles, G. Severi, S. Castellvi-Bel, C. Ruiz-Ponte, A. Carracedo, A. Castells, A. Forsti, K. Hemminki, P. Vodicka, A. Naccarati, L. Lipton, J. W. Ho, K. K. Cheng, P. C. Sham, J. Luk, J. A. Agundez, J. M. Ladero, M. de la Hoya, T. Caldes, I. Niittymaki, S. Tuupanen, A. Karhu, L. Aaltonen, J. B. Cazier, H. Campbell, M. G. Dunlop, and R. S. Houlston. A genome-wide association study identifies colorectal cancer susceptibility loci on chromosomes 10p14 and 8q23.3. *Nat Genet*, 40(5):623–30, 2008.

[206] M. Toyota and H. Suzuki. Epigenetic drivers of genetic alterations. *Adv Genet*, 70:309–23, 2010.

[207] O. G. Troyanskaya. Putting microarrays in a context: integrated analysis of diverse biological data. *Brief Bioinform*, 6(1):34–43, 2005.

[208] A. G. Tsai and M. R. Lieber. Mechanisms of chromosomal rearrangement in the human genome. *BMC Genomics*, 11 Suppl 1:S1, 2010.

[209] S. Tuupanen, M. Turunen, R. Lehtonen, O. Hallikas, S. Vanharanta, T. Kivioja, M. Bjorklund, G. Wei, J. Yan, I. Niittymaki, J. P. Mecklin, H. Jarvinen, A. Ristimaki, M. Di-Bernardo, P. East, L. Carvajal-Carmona, R. S. Houlston, I. Tomlinson, K. Palin, E. Ukkonen, A. Karhu, J. Taipale, and L. A. Aaltonen. The common colorectal cancer predisposition snp rs6983267 at chromosome 8q24 confers potential to enhanced wnt signaling. *Nat Genet*, 41(8):885–90, 2009.

[210] S. Ugras, E. R. Brill, A. Jacobsen, M. Hafner, N. Socci, P. L. Decarolis, R. Khanin, R. B. O'Connor, A. Mihailovic, B. S. Taylor, R. Sheridan, J. Gimble, A. Viale, A. Crago, C. R. Antonescu, C. Sander, T. Tuschl, and S. Singer. Small rna sequencing and functional characterization reveals microrna-143 tumor suppressor activity in liposarcoma. *Cancer Res*, page In press, 2011.

[211] M. J. van de Vijver, Y. D. He, L. J. van't Veer, H. Dai, A. A. Hart, D. W. Voskuil, G. J. Schreiber, J. L. Peterse, C. Roberts, M. J. Marton, M. Parrish, D. Atsma, A. Witteveen, A. Glas, L. Delahaye, T. van der Velde, H. Bartelink, S. Rodenhuis, E. T. Rutgers, S. H. Friend, and R. Bernards. A gene-expression signature as a predictor of survival in breast cancer. *N Engl J Med*, 347(25):1999–2009, 2002.

[212] M. van Engeland, S. Derks, K. M. Smits, G. A. Meijer, and J. G. Herman. Colorectal cancer epigenetics: complex simplicity. *J Clin Oncol*, 29(10):1382–91, 2011.

[213] L. J. van 't Veer, H. Dai, M. J. van de Vijver, Y. D. He, A. A. Hart, M. Mao, H. L. Peterse, K. van der Kooy, M. J. Marton, A. T. Witteveen, G. J. Schreiber, R. M. Kerkhoven, C. Roberts, P. S. Linsley, R. Bernards, and S. H. Friend. Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, 415(6871):530–6, 2002.

[214] J. M. Varley. Germline tp53 mutations and li-fraumeni syndrome. *Hum Mutat*, 21(3):313–20, 2003.

[215] A. R. Venkitaraman. Cancer susceptibility and the functions of brca1 and brca2. *Cell*, 108(2):171–82, 2002.

[216] A. Ventura and T. Jacks. Micrornas and cancer: short rnas go a long way. *Cell*, 136(4):586–91, 2009.

[217] R. G. Verhaak, K. A. Hoadley, E. Purdom, V. Wang, Y. Qi, M. D. Wilkerson, C. R. Miller, L. Ding, T. Golub, J. P. Mesirov, G. Alexe,

M. Lawrence, M. O'Kelly, P. Tamayo, B. A. Weir, S. Gabriel, W. Winckler, S. Gupta, L. Jakkula, H. S. Feiler, J. G. Hodgson, C. D. James, J. N. Sarkaria, C. Brennan, A. Kahn, P. T. Spellman, R. K. Wilson, T. P. Speed, J. W. Gray, M. Meyerson, G. Getz, C. M. Perou, and D. N. Hayes. Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in pdgfra, idh1, egfr, and nf1. *Cancer Cell*, 17(1):98–110, 2010.

[218] P. Vineis, A. Schatzkin, and J. D. Potter. Models of carcinogenesis: an overview. *Carcinogenesis*, 31(10):1703–9, 2010.

[219] B. Vogelstein and K. W. Kinzler. Cancer genes and the pathways they control. *Nat Med*, 10(8):789–99, 2004.

[220] X. Wang, S. Bjorklund, A. M. Wasik, A. Grandien, P. Andersson, E. Kimby, K. Dahlman-Wright, C. Zhao, B. Christensson, and B. Sander. Gene expression profiling and chromatin immunoprecipitation identify dbn1, setmar and hig2 as direct targets of sox11 in mantle cell lymphoma. *PLoS One*, 5(11):e14085, 2010.

[221] Y. Wang, J. G. Klijn, Y. Zhang, A. M. Sieuwerts, M. P. Look, F. Yang, D. Talantov, M. Timmermans, M. E. Meijer-van Gelder, J. Yu, T. Jatkoe, E. M. Berns, D. Atkins, and J. A. Foekens. Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *Lancet*, 365(9460):671–9, 2005.

[222] N. F. Wasserman, I. Aneas, and M. A. Nobrega. An 8q24 gene desert variant associated with prostate cancer risk confers differential in vivo activity to a myc enhancer. *Genome Res*, 20(9):1191–7, 2010.

[223] M. Weber, J. J. Davies, D. Wittig, E. J. Oakeley, M. Haase, W. L. Lam, and D. Schubeler. Chromosome-wide and promoter-specific analyses identify sites of differential dna methylation in normal and transformed human cells. *Nat Genet*, 37(8):853–62, 2005.

[224] B. Weigelt, F. L. Baehner, and J. S. Reis-Filho. The contribution of gene expression profiling to breast cancer classification, prognostication and prediction: a retrospective of the last decade. *J Pathol*, 220(2):263–80, 2010.

[225] B. Weigelt, A. Mackay, R. A'Hern, R. Natrajan, D. S. Tan, M. Dowsett, A. Ashworth, and J. S. Reis-Filho. Breast cancer molecular profiling with single sample predictors: a retrospective analysis. *Lancet Oncol*, 11(4):339–49, 2010.

[226] M. R. Welfare, J. Cooper, M. F. Bassendine, and A. K. Daly. Relationship between acetylator status, smoking, and diet and colorectal

cancer risk in the north-east of england. *Carcinogenesis*, 18(7):1351–4, 1997.

[227] S. M. Welford, J. Gregg, E. Chen, D. Garrison, P. H. Sorensen, C. T. Denny, and S. F. Nelson. Detection of differentially expressed genes in primary tumor tissues using representational differences analysis coupled to microarray hybridization. *Nucleic Acids Res*, 26 (12):3059–65, 1998.

[228] J. B. Wright, S. J. Brown, and M. D. Cole. Upregulation of c-myc in cis through a large chromatin loop linked to a cancer risk-associated single-nucleotide polymorphism in colorectal cancer cells. *Mol Cell Biol*, 30(6):1411–20, 2010.

[229] J. X. Yu, A. M. Sieuwerts, Y. Zhang, J. W. Martens, M. Smid, J. G. Klijn, Y. Wang, and J. A. Foekens. Pathway analysis of gene signatures predicting metastasis of node-negative primary breast cancer. *BMC Cancer*, 7:182, 2007.

[230] E. R. Zabarovsky, M. I. Lerman, and J. D. Minna. Tumor suppressor genes on chromosome 3p involved in the pathogenesis of lung and other cancers. *Oncogene*, 21(45):6915–35, 2002.

[231] H. Zaehres and H. R. Scholer. Induction of pluripotency: from mouse to human. *Cell*, 131(5):834–5, 2007.

[232] B. W. Zanke, C. M. Greenwood, J. Rangrej, R. Kustra, A. Tenesa, S. M. Farrington, J. Prendergast, S. Olschwang, T. Chiang, E. Crowdy, V. Ferretti, P. Laflamme, S. Sundararajan, S. Roumy, J. F. Olivier, F. Robidoux, R. Sladek, A. Montpetit, P. Campbell, S. Bezieau, A. M. O'Shea, G. Zogopoulos, M. Cotterchio, P. Newcomb, J. McLaughlin, B. Younghusband, R. Green, J. Green, M. E. Porteous, H. Campbell, H. Blanche, M. Sahbatou, E. Tubacher, C. Bonaiti-Pellie, B. Buecher, E. Riboli, S. Kury, S. J. Chanock, J. Potter, G. Thomas, S. Gallinger, T. J. Hudson, and M. G. Dunlop. Genome-wide association scan identifies a colorectal cancer susceptibility locus on chromosome 8q24. *Nat Genet*, 39(8):989–94, 2007.

[233] Y. Zhang, J. W. Martens, J. X. Yu, J. Jiang, A. M. Sieuwerts, M. Smid, J. G. Klijn, Y. Wang, and J. A. Foekens. Copy number alterations that predict metastatic capability of human breast cancer. *Cancer Res*, 69(9):3795–801, 2009.

Part V

ADDENDA

A

CURRICULUM VITAE AND OTHER CONTRIBUTED PUBLICATIONS

In the following pages, you will find a brief biosketch that summarizes my academic and research experience, as well as the first page of all the scientific articles in which I have collaborated.

During more than ten years of research experience, not only I have dedicated to my thesis, but also collaborated with many researchers within and outside my institution. Undoubtedly, this collaboration has been doubly fruitful to me. On one hand, it has enabled me to participate in many different studies which have eventually been published in international peer-reviewed journals. On the other hand, and most importantly, it has allowed me to get in touch with many different aspects of cancer, medicine, biology, genetics, systems biology, bioinformatics and statistics. I must say that, by dealing with this, I have perceived the complexity of the problem we are facing, and I have realized that only through strong collaboration and integration of multidisciplinary efforts we will be able to succeed in this overwhelming task that is defeating cancer. Therefore, I can only be extremely grateful to all the researchers that have allowed me to learn something from them along this stage of my research career.

I hope all of them enjoyed our collaboration as much as I did.

BIOGRAPHICAL SKETCH

LAST NAME, FIRST NAME Solé Acha, Xavier	POSITION TITLE PhD student
DATE AND PLACE OF BIRTH November 6th, 1977. Barcelona (Spain).	

EDUCATION/TRAINING *(Begin with baccalaureate or other initial professional education, such as nursing, and include postdoctoral training.)*

INSTITUTION AND LOCATION	DEGREE *(if applicable)*	YEAR	FIELD OF STUDY
Technical University of Catalonia, Barcelona	B.Sc.	1995-2001	Computer Science
University of Barcelona, Barcelona	DEA	2001-2003	Genetics

A. Research interests

During these years, my main research interests have focused on the bioinformatics integration of heterogeneous, large-scale molecular data (gene expression, SNPs, copy number, methylation…) to study the underlying basis of complex diseases, such as cancer. I am also interested in applying Systems Biology and network analysis to assess how the different elements of the cell regulate each other and understand the impact of this regulation in the cell's behavior.

B. Positions and Employment

- 2001 - 2006:
  Research Technician. Unit of Biomarkers and Susceptibility, Catalan Institute of Oncology, Barcelona, Spain.

- 2006 - present:
  PhD student. Unit of Biomarkers and Susceptibility, Catalan Institute of Oncology, Barcelona, Spain.

- February 2008 - May 2008:
  Visiting scientist at Dr. Andrea Califano's and Dr. Adolfo A. Ferrando's Labs, Columbia University.

C. Awards and Honors

- UICC Yamagiwa-Yoshida Memorial International Study Grant (2008)

D. Peer-reviewed publications

1. Della Gatta G, Palomero T, Perez-Garcia A, Ambesi-Impiombato A, Bansal M, Carpenter ZW, De Keersmaecker K, **Solé 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. **Nature Medicine** *In press.*

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

3. Balart J, Pueyo G, de Llobet LI, Baro M, **Solé X**, Marin S, Casanovas O, Mesia R, Capella G. The use of caspase inhibitors in pulsed-field gel electrophoresis may improve the estimation of radiation-induced DNA repair and apoptosis. **Radiat Oncol.** 6(6), 2011.

4. De Keersmaecker K, Real PJ, Gatta GD, Palomero T, Sulis ML, Tosello V, Van Vlierberghe P, Barnes K, Castillo M, **Solé X**, Hadler M, Lenz J, Aplan PD, Kelliher M, Kee BL, Pandolfi PP, Kappes D, Gounari F, Petrie H, Van der Meulen J, Speleman F, Paietta E, Racevskis J, Wiernik PH, Rowe JM, Soulier J, Avran D, Cavé H, Dastugue N, Raimondi S, Meijerink JP, Cordon-Cardo C, Califano A, Ferrando AA. The TLX1 oncogene drives aneuploidy in T cell transformation. **Nature Medicine** 16(11):1321-7, 2010.

5. Aguilar H, **Solé X**, Bonifaci N, Serra-Musach J, Islam A, López-Bigas N, Méndez-Pertuz M, Beijersbergen RL, Lázaro C, Urruticoechea A, Pujana MA. Biological reprogramming in acquired resistance to endocrine therapy of breast cancer. **Oncogene** 29(45):6071-83, 2010.

6. Fernández-Ramires R, **Solé X**, De Cecco L, Llort G, Cazorla A, Bonifaci N, Garcia MJ, Caldés T, Blanco I, Gariboldi M, Pierotti MA, Pujana MA, Benítez J, Osorio A. Gene expression profiling integrated into network modelling reveals heterogeneity in the mechanisms of BRCA1 tumorigenesis. **Br J Cancer** 101(8):1469-80, 2009.

7. **Solé X**, Bonifaci N, López-Bigas N, Berenguer A, Hernández P, Reina O, Maxwell CA, Aguilar H, Urruticoechea A, de Sanjosé S, Comellas F, Capellá G, Moreno V, Pujana MA. Biological convergence of cancer signatures. **PLoS ONE** 4(2):e4544, 2009.

8. **Solé X**, Hernández P, de Heredia ML, Armengol L, Rodríguez-Santiago B, Gómez L, Maxwell CA, Aguiló F, Condom E, Abril J, Pérez-Jurado L, Estivill X, Nunes V, Capellá G, Gruber SB, Moreno V, Pujana MA. Genetic and genomic analysis modeling of germline c-MYC overexpression and cancer susceptibility. **BMC Genomics**, Jan 10, 2008.

9. Maxwell CA, Moreno V, **Solé X**, Gómez L, Hernández P, Urruticoechea A, Pujana MA. Genetic interactions: the missing links for a better understanding of cancer susceptibility, progression and treatment. **Molecular Cancer**, Jan 10, 2008.

10. Pujana MA, Han JD, Starita LM, Stevens KN, Tewari M, Ahn JS, Rennert G, Moreno V, Kirchhoff T, Gold B, Assmann V, Elshamy WM, Rual JF, Levine D, Rozek LS, Gelman RS, Gunsalus KC, Greenberg RA, Sobhian B, Bertin N, Venkatesan K, Ayivi-Guedehoussou N, **Solé X**, Hernández P, Lázaro C, Nathanson KL, Weber BL, Cusick ME, Hill DE, Offit K, Livingston DM, Gruber SB, Parvin JD, Vidal M. Network modeling links breast cancer susceptibility and centrosome dysfunction. **Nature Genetics** 39(11):1338-49, 2007.

11. Urruticoechea A, Aguilar H, **Solé X**, Capellà G, Martin LA, Dowsett M, Germà-Lluch JR. Pre-clinical validation of early molecular markers of sensitivity to aromatase inhibitors in a mouse model of post-menopausal hormone-sensitive breast cancer. **Breast Cancer Res Treat.** 109(3):463-70, 2007.

12. Valls J, Grau M, **Solé X**, Hernández P, Montaner D, Dopazo J, Peinado MA, Capellá G, Moreno V, Pujana MA.CLEAR-test: combining inference for differential expression and variability in microarray data analysis. **Journal of Biomedical Informatics** 41(1):33-45, 2007.

13. Vendrell E, Ribas M, Valls J, **Solé X**, Grau M, Moreno V, Capellà G, Peinado MA. Genomic and transcriptomic prognostic factors in R0 Dukes B and C colorectal cancer patients. **International Journal of Oncology** 30(5):1099-107, 2007.

14. Hernández P*, **Solé X***, Valls J*, Moreno V, Capellá G, Urruticoechea A, Pujana MA. Integrative analysis of a cancer somatic mutome. **Mol Cancer**. 2007 Feb 5;6:13

15. González JR, Armengol L, **Solé X**, Guinó E, Mercader JM, Estivill X, Moreno V. SNPassoc: an R package to perform whole genome association studies. **Bioinformatics** 23(5):644-5, 2007.

16. Pujana MA, Ruiz A, Badenas C, Puig-Butille JA, Nadal M, Stark M, Gómez L, Valls J, **Solé X**, Hernández P, Cerrato C, Madrigal I, de Cid R, Aguilar H, Capellá G, Cal S, James MR, Walker GJ, Malvehy J, Milà M, Hayward NK, Estivill X, Puig S. Molecular characterization of a t(9;12)(p21;q13) balanced chromosome translocation in combination with integrative genomics analysis identifies C9orf14 as a candidate tumor-suppressor. **Genes Chromosomes & Cancer** 46(2):155-62, 2007.

17. **Solé X**, Guinó E, Valls J, Iniesta R, Moreno V. SNPStats: a web tool for the analysis of association studies. **Bioinformatics** 2006; 22(15):1928-9.

18. Grau M*, **Solé X***, Obrador A, Tarafa G, Vendrell E, Valls J, Moreno V, Peinado MA, Capellá G. Validation of RNA arbitrarily primed PCR probes hybridized to glass cDNA microarrays: application to the analysis of limited samples. **Clin Chem.** 2005 51(1):93-101.

19. Frigola J, **Solé X**, Paz MF, Moreno V, Esteller M, Capellà G, Peinado MA. Differential DNA hypermethylation and hypomethylation signatures in colorectal cancer. **Hum Mol Genet.** 2005 15;14(2):319-26.

20. Moreno V, **Solé X**. Use of DNA chips (microarrays) in medicine: technical foundations and basic procedures for statistical analysis of results. **Med Clin (Barc)**. 2004;122 Suppl 1:73-9.

* Contributed equally to the manuscript

E. Contributed book chapters

1. **Solé X**, Gonzalez JR, Moreno V. Analysis of Population-Based Genetic Association Studies Applied to Cancer Susceptibility and Prognosis. T Pham (ed.) Computational Biology: Issues and Applications in Oncology, Applied Bioinformatics and Biostatistics in Cancer Research, Chapter 7. Springer Science+Business Media, 2009.

F. Research stays abroad

- February 2008 – May 2008:

  Collaboration with Dr. Andrea Califano and Dr. Adolfo A. Ferrando at the Center for Computational Biology and Bioinformatics, Columbia University (NY, USA). In this collaboration, we applied Systems Biology techniques to reverse-engineer the transcriptional networks of different leukemia, breast and colorectal tumor cells. For the leukemia analyses, the results of this collaboration have been published in two articles in *Nature Medicine*.

G. Research Support

- FIS PI11/01439 2012/2014 279.234 €
  Fondo de Investigaciones de la Seguridad Social, Instituto de Salud Carlos III, Spanish Ministry of Health
  Project: COLONOMICS - Integration of molecular basis of colorectal cancer for the detection and validation of new Biomarkers.
  Role: Co-investigator.

We have the COLONOMICS project in which, for a series of 100 cases and 50 controls have molecular expression data, SNPs / CNVs, methylation and miRNAs. The analysis of COLONOMICS has provided us with candidate biomarkers that we propose to validate in this project following a sequential approach in several series of patients. The project aims are: 1) Establish a biomarker validation platform with high quality biological samples and comprehensive clinical information that allow the validation of multiple markers in different phases. 2) Validate the already identified biomarkers and to demonstrate its usefulness in early diagnosis or prognosis prediction. 3) To characterize the mechanisms by which some genes are revealed useful as diagnostic or prognostic biomarkers. 4) To assess the mutational profile of tumors using exome deep sequencing and identify new prognostic biomarkers. 5) To advance the knowledge of the molecular mechanisms of cancer, correlating the mutational profiles with those of expression, chromosomal alterations, and regulatory elements (miRNA and methylation) and identify potential therapeutic targets or chemoprevention strategies.

- AATRM 209/2010 2010-12 € 145.800
  Project: Genetic determinants of toxicity to chemotherapy in colorectal cancer: a GWAS study.
  Role: Co-investigator.
  This study aims to identify genetic determinants for toxicity of chemotherapy in CRC. A GWAS has been performed in 259 subjects treated with 5FU, oxaliplatin or irinotecan for whom detailed toxicity was recorded. Candidate SNPs identified will be validated in an independent sample of 800 subjects.

- FP7-HEALTH-2007-B / 223378 2009-12 € 569,400
  Project: HiPerDART - Development of High Performance Diagnostic Array Replication Technology.
  Role: Co-investigator.
  The HiPerDART project aims to develop, using a novel printing technology called Supramolecular NanoStamping (SuNS), a higher standard of clinical microarray technology that promises to dramatically improve the reliability of medical tests. This technology will be tested in a clinical application to provide a prognostic and predictive tool based on a comprehensive set of genomic markers selected using state of the art bioinformatics techniques. Patients diagnosed with stage II colorectal cancer could benefit from a molecular staging prognostic tool to indicate the need of adjuvant chemotherapy.

- FIS PI081635 2009-11 € 241,395
  Fondo de Investigaciones de la Seguridad Social, Instituto de Salud Carlos III, Spanish Ministry of Health.
  Project: Identification of novel candidate genes using gene regulatory networks.
  Role: Co-investigator.
  A new methodology is proponed to identify new genetic markers of susceptibility and/or prognosis in colorectal cancer. It is based on the analysis of complex data from SNPs, CNV and expression, combined with external sources through techniques of systems biology. Paired normal mucosa and tumor samples from 100 individuals will be analyzed for expression with microarrays. In the same subjects, genetic variability (SNPs and CNV) be analyzed using 500K chips and, in tumors, allelic imbalances with aCGH. Gene expression data, variation data, and other external data sources available as gene ontology, pathways and literature will be combined using network theory to identify relationships between genes.

- FIS PI081359 2009-11 € 99,220
  Fondo de Investigaciones de la Seguridad Social, Instituto de Salud Carlos III, Spanish Ministry of Health.
  Project: ENTERICOS - disinfection by-products and other environmental, genetic and molecular determinants of colorectal cancer.
  Role: Co-investigator.

ENTERICOS is a case-control study to evaluate the risk of colorectal cancer associated to exposure to disinfection by-products (DBP) and the interaction with genetic factors measured as polymorphisms in genes related to metabolism, DNA repair or inflammatory response.

H. Patents

- Inventors: Peinado MA, Risques RA, Vendrell E, Capellà G, Grau M, Obrador A, Tarafa G, Moreno V, **Solé X**, Rosell E, Piulats J.
  Title: Genetic analysis of biological samples in arrayed expanded representations of their nucleic acids
  Application #: 02797953.3
  Priority country: Spain
  Year: 2002
  Holder: MEFA - Merck Farma y Química, S.A.

I. Oral communications in international meetings

- October 2011: *6th Annual DREAM on Reverse Engineering Challenges, 7th Annual RECOMB Satellite on Systems Biology, 8th Annual RECOMB Satellite on Regulatory Genomics& 1st IDIBELL Conference on Cancer Informatics (RICCI).* **Masters Regulators of Metastasis in Early Stage Colon Cancer**.

J. Other

List of most relevant meetings, courses and seminars attended

- June 2009: *17th Annual International Conference on Intelligent Systems for Molecular Biology (ISMB) & 8th European Conference on Computational Biology (ECCB)*, Stockholm (Sweden).
- Nov. 2006: *7th Annual Spanish Bioinformatics Conference*, Zaragoza (Spain).
- Dec. 2004: *5th Annual Spanish Bioinformatics Conference*, Barcelona (Spain).
- Oct. 2004: Special AACR conference on *Advances in Proteomics in Cancer Research*, Miami (USA).
- May 2003: Course on *Computational and Statistical Aspects of Microarray Analysis*, Università degli Studi, Milano (Italy).
- Apr. 2003: Seminar on *Cancer: molecular targets for novel therapies*, Lilly Foundation, Madrid (Spain).
- Jun. 2002: Course on *Design and Analysis of DNA Microarray Experiments*, Pompeu Fabra University, Barcelona (Spain).
- Apr. 2002: Workshop on *Bioinformatics and Computational Biology*, BBVA Foundation, Madrid (Spain).
- Apr. 2002: *DNA microarrays 2002* meeting, Madrid (Spain).
- Apr. 2001: *DNA microarrays 2001* meeting, Madrid (Spain).
- Apr. 2001: *3rd meeting on Microarray Databases, Standards and Ontologies (MGED3)*, Stanford University, Palo Alto (USA).
- Nov. 2000: *Data Mining* workshop, Mathematics and Statistics College (FME), Technical University of Catalonia (UPC), Barcelona (Spain).

Languages

- **Spanish:** mother tongue.
- **Catalan:** mother tongue.
- **English:** Fluent level of reading, writing and speaking. *Certificate of Proficiency in English* diploma by the University of Cambridge.

Balart et al. Radiation Oncology 2011, 6:6
http://www.ro-journal.com/content/6/1/6

RADIATION ONCOLOGY

METHODOLOGY Open Access

The use of caspase inhibitors in pulsed-field gel electrophoresis may improve the estimation of radiation-induced DNA repair and apoptosis

Josep Balart[1,5*], Gemma Pueyo[1], Lara I de Llobet[1], Marta Baro[1], Xavi Sole[2], Susanna Marin[3], Oriol Casanovas[1], Ricard Mesia[4], Gabriel Capella[1]

Abstract

Background: Radiation-induced DNA double-strand break (DSB) repair can be tested by using pulsed-field gel electrophoresis (PFGE) in agarose-encapsulated cells. However, previous studies have reported that this assay is impaired by the spontaneous DNA breakage in this medium. We investigated the mechanisms of this fragmentation with the principal aim of eliminating it in order to improve the estimation of radiation-induced DNA repair.

Methods: Samples from cancer cell cultures or xenografted tumours were encapsulated in agarose plugs. The cell plugs were then irradiated, incubated to allow them to repair, and evaluated by PFGE, caspase-3, and histone H2AX activation (γH2AX). In addition, apoptosis inhibition was evaluated through chemical caspase inhibitors.

Results: We confirmed that spontaneous DNA fragmentation was associated with the process of encapsulation, regardless of whether cells were irradiated or not. This DNA fragmentation was also correlated to apoptosis activation in a fraction of the cells encapsulated in agarose, while non-apoptotic cell fraction could rejoin DNA fragments as was measured by γH2AX decrease and PFGE data. We were able to eliminate interference of apoptosis by applying specific caspase inhibitors, and improve the estimation of DNA repair, and apoptosis itself.

Conclusions: The estimation of radiation-induced DNA repair by PFGE may be improved by the use of apoptosis inhibitors. The ability to simultaneously determine DNA repair and apoptosis, which are involved in cell fate, provides new insights for using the PFGE methodology as functional assay.

Background

The use of pulsed-field gel electrophoresis (PFGE) is widespread in the evaluation of DNA fragmentation caused by double-strand breaks (DSBs) following ionizing radiation [1-4]. The DNA-DSBs may result in the formation of small (often acentric) chromosomal fragments. Following this initial damage, cells activate DNA repair mechanisms to prevent catastrophic mitosis and cell death due to the loss of acentric DNA fragments [5]. The principle of PFGE methodology is that the release of DNA from cells correlates adequately with the intensity of DNA fragmentation [6]. The estimation of DNA repair by PFGE is based on the diminution of DNA released from cells as the length of the DNA fragments increases through the process of rejoining. Thus, a decrease in the ratio of DNA extracted from the cells over a period of time can be used as an evaluation of DNA repair [7].

In the PFGE technique, cells are encapsulated in agarose to form cell-plugs, thus preventing physic damage of the cells while facilitating their manipulation and placement into agarose gels where electrophoresis will take place. Usually in laboratory routine, cells are encapsulated after a period of repair which is allowed to occur in physiological conditions such as either cell cultures or xenografts. Thus, extraction ratios depend exclusively on induced and repaired DNA damage. While the desired strategy is to encapsulate cells after the period

* Correspondence: jbalart@iconcologia.net
[1]Translational Research Laboratory - IDIBELL, Institut Català d'Oncologia, L'Hospitalet de Llobregat, Spain
Full list of author information is available at the end of the article

nature medicine

The *TLX1* oncogene drives aneuploidy in T cell transformation

Kim De Keersmaecker[1–3,28], Pedro J Real[1,27,28], Giusy Della Gatta[1,28], Teresa Palomero[1,4], Maria Luisa Sulis[1,5], Valeria Tosello[1], Pieter Van Vlierberghe[1], Kelly Barnes[1], Mireia Castillo[4], Xavier Sole[6,7], Michael Hadler[1], Jack Lenz[8], Peter D Aplan[9], Michelle Kelliher[10], Barbara L Kee[11], Pier Paolo Pandolfi[12], Dietmar Kappes[13], Fotini Gounari[14], Howard Petrie[15], Joni Van der Meulen[16], Frank Speleman[16], Elisabeth Paietta[17,18], Janis Racevskis[17,18], Peter H Wiernik[17,18], Jacob M Rowe[19], Jean Soulier[20,21], David Avran[20,21], Hélène Cavé[22], Nicole Dastugue[23], Susana Raimondi[24], Jules P P Meijerink[25], Carlos Cordon-Cardo[4], Andrea Califano[1,26] & Adolfo A Ferrando[1,4,5]

The *TLX1* oncogene (encoding the transcription factor T cell leukemia homeobox protein-1) has a major role in the pathogenesis of T cell acute lymphoblastic leukemia (T-ALL). However, the specific mechanisms of T cell transformation downstream of *TLX1* remain to be elucidated. Here we show that transgenic expression of human *TLX1* in mice induces T-ALL with frequent deletions and mutations in *Bcl11b* (encoding B cell leukemia/lymphoma-11B) and identify the presence of recurrent mutations and deletions in *BCL11B* in 16% of human T-ALLs. Most notably, mouse *TLX1* tumors were typically aneuploid and showed a marked defect in the activation of the mitotic checkpoint. Mechanistically, *TLX1* directly downregulates the expression of *CHEK1* (encoding CHK1 checkpoint homolog) and additional mitotic control genes and induces loss of the mitotic checkpoint in nontransformed preleukemic thymocytes. These results identify a previously unrecognized mechanism contributing to chromosomal missegregation and aneuploidy active at the earliest stages of tumor development in the pathogenesis of cancer.

T-ALL is an aggressive hematologic tumor resulting from the malignant transformation of T cell progenitors. The *TLX1* oncogene is translocated and aberrantly expressed in 5–10% of pediatric and up to 30% of adult T-ALL cases[1–4]. In addition, *TLX3*, a closely related TLX family member, is overexpressed as a result of the t(5;14)(q35;q32) translocation in about 25% of pediatric T-ALLs and in 5% of adult T-ALL cases[5]. *TLX1* expression defines a distinct molecular group of T-ALL characterized by a differentiation block at the early cortical stage of thymocyte development[2] and a favorable prognosis[1,2,6]. Moreover, *TLX1* and *TLX3* leukemias seem to constitute a distinct oncogenic group with specific genetic alterations rarely found in non–TLX-induced T-ALLs, including the rearrangement of the *NUP214-ABL1* oncogene[7] (a fusion of the gene encoding 214-kDa nucleoporin and c-abl oncogene-1, non–receptor tyrosine kinase) and mutations in the *WT1* (encoding Wilms tumor-1 homolog)[8] and *PHF6* (encoding PHD finger protein-6)[9] tumor suppressor genes. However, little is known about the specific mechanisms that mediate T cell transformation downstream of *TLX1*. To address this question, we have used an integrative genomic approach to characterize the transcriptional programs and oncogenic pathways active in human and mouse *TLX1*-induced leukemia.

RESULTS

T-ALL development in *TLX1*-transgenic mice

To investigate the mechanisms of T cell transformation driven by *TLX1*, we generated p56^{Lck}-*TLX1* transgenic mice in which the *Lck* proximal promoter drives expression of *TLX1* in T cell progenitors[10,11]. *TLX1*-transgenic mice from three founder lines showed accelerated

[1]Institute for Cancer Genetics, Columbia University, New York, New York, USA. [2]Department of Molecular and Developmental Genetics, VIB, Leuven, Belgium. [3]Center for Human Genetics, K.U. Leuven, Leuven, Belgium. [4]Department of Pathology, Columbia University Medical Center, New York, New York, USA. [5]Department of Pediatrics, Columbia University Medical Center, New York, New York, USA. [6]Biomarkers and Susceptibility Unit, Catalan Institute of Oncology, Institut d'Investigació Biomèdica de Bellvitge, L'Hospitalet, Barcelona, Spain. [7]Biomedical Research Centre Network for Epidemiology and Public Health, Catalan Institute of Oncology, IDIBELL, L'Hospitalet, Barcelona, Spain. [8]Department of Molecular Genetics, Albert Einstein College of Medicine, Bronx, New York, USA. [9]The Genetics Branch, Center for Cancer Research, National Cancer Institute, Bethesda, Maryland, USA. [10]Department of Cancer Biology, University of Massachusetts Medical School, Worcester, Massachusetts, USA. [11]Department of Pathology, University of Chicago, Chicago, Illinois, USA. [12]Departments of Medicine and Pathology, Beth Israel Deaconess Cancer Center, Harvard Medical School, Boston, MA, USA. [13]Fox Chase Cancer Center, Philadelphia, Pennsylvania, USA. [14]Department of Medicine, University of Chicago, Chicago, Illinois, USA. [15]Department of Cancer Biology, The Scripps Research Institute, Jupiter, Florida, USA. [16]Center for Medical Genetics, Ghent University Hospital, Ghent, Belgium. [17]Montefiore Medical Center–North Division, New York, New York, USA. [18]New York Medical College, New York, New York, USA. [19]Rambam Medical Center and Technion, Israel Institute of Technology, Haifa, Israel. [20]Assistance publique—Hôpitaux de Paris Hematology Laboratory and Institut National de la Santé et de la Recherche Médicale U944, Hôpital Saint-Louis, Paris, France. [21]Université Paris 7-Denis Diderot, Institut Universitaire d'Hematology, Hôpital Saint-Louis, Paris, France. [22]Assistance publique—Hôpitaux de Paris, Hôpital Robert Debré, Département de Génétique, Université Paris 7-Denis Diderot, Paris, France. [23]Laboratoire d'Hématologie, Hôpital Purpan, Toulouse, France. [24]Department of Pathology, St Jude Children's Research Hospital, Memphis, Tennessee, USA. [25]Department of Pediatric Oncology/Hematology, Erasmus MC-Sophia Children's Hospital, Rotterdam, The Netherlands. [26]Joint Centers for Systems Biology, Columbia University, New York, New York, USA. [27]Current address: Andalusian Stem Cell Bank, Centro de Investigacion Biomedica, Granada, Spain. [28]These authors contributed equally to this work. Correspondence should be addressed to A.A.F. (af2196@columbia.edu).

Received 31 March; accepted 21 September; published online 24 October 2010; doi:10.1038/nm.2246

Oncogene (2010), 1–13

www.nature.com/onc

ONCOGENOMICS

Biological reprogramming in acquired resistance to endocrine therapy of breast cancer

H Aguilar[1], X Solé[2,3], N Bonifaci[2,3], J Serra-Musach[2,3], A Islam[4], N López-Bigas[4], M Méndez-Pertuz[5], RL Beijersbergen[6], C Lázaro[7], A Urruticoechea[1] and MA Pujana[1,2,3]

[1]Translational Research Laboratory, Catalan Institute of Oncology, Bellvitge Institute for Biomedical Research (IDIBELL), L'Hospitalet, Barcelona, Spain; [2]Biomarkers and Susceptibility Unit, Catalan Institute of Oncology, IDIBELL, L'Hospitalet, Barcelona, Spain; [3]Biomedical Research Centre Network for Epidemiology and Public Health, Catalan Institute of Oncology, IDIBELL, L'Hospitalet, Barcelona, Spain; [4]Research Unit on Biomedical Informatics, Department of Experimental and Health Science, Pompeu Fabra University, Barcelona Biomedical Research Park, Barcelona, Spain; [5]Epithelial Carcinogenesis Group, Molecular Pathology Programme, Spanish National Cancer Research Centre, Madrid, Spain; [6]Division of Molecular Carcinogenesis, Center for Biomedical Genetics and Cancer Genomics Center, Netherlands Cancer Institute, Amsterdam, The Netherlands and [7]Molecular Diagnostics Unit, Hereditary Cancer Program, Catalan Institute of Oncology, IDIBELL, L'Hospitalet, Barcelona, Spain

Endocrine therapies targeting the proliferative effect of 17β-estradiol through estrogen receptor α (ERα) are the most effective systemic treatment of ERα-positive breast cancer. However, most breast tumors initially responsive to these therapies develop resistance through molecular mechanisms that are not yet fully understood. The long-term estrogen-deprived (LTED) MCF7 cell model has been proposed to recapitulate acquired resistance to aromatase inhibitors in postmenopausal women. To elucidate this resistance, genomic, transcriptomic and molecular data were integrated into the time course of MCF7–LTED adaptation. Dynamic and widespread genomic changes were observed, including amplification of the *ESR1* locus consequently linked to an increase in ERα. Dynamic transcriptomic profiles were also observed that correlated significantly with genomic changes and were predicted to be influenced by transcription factors known to be involved in acquired resistance or cell proliferation (for example, interferon regulatory transcription factor 1 and E2F1, respectively) but, notably, not by canonical ERα transcriptional function. Consistently, at the molecular level, activation of growth factor signaling pathways by EGFR/ERBB/AKT and a switch from phospho-Ser118 (pS118)- to pS167-ERα were observed during MCF7–LTED adaptation. Evaluation of relevant clinical settings identified significant associations between MCF7–LTED and breast tumor transcriptome profiles that characterize ERα-negative status, early response to letrozole and tamoxifen, and recurrence after tamoxifen treatment. In accordance with these profiles, MCF7–LTED cells showed increased sensitivity to inhibition of FGFR-mediated signaling with PD173074. This study provides mechanistic insight into acquired resistance to endocrine therapies of breast cancer and highlights a potential therapeutic strategy.
Oncogene advance online publication, 16 August 2010; doi:10.1038/onc.2010.333

Keywords: aromatase inhibition; breast cancer; estrogen receptor; fibroblast growth factor receptor; long-term estrogen-deprived; MCF7

Introduction

Endocrine therapies are the most effective systemic treatment of estrogen receptor α (ERα)-positive breast cancer, and over two-thirds of patients are considered to present with this kind of disease (EBCTCG, 1998; Chlebowski *et al*., 2002; Winer *et al*., 2002). Two major strategies mediate the efficacy of these therapies. Drugs directed at ERα, mainly tamoxifen and fulvestrant, impede its binding to 17β-estradiol (17βE2) and, as a result, canonical ERα-dependent transcriptional regulation (Dowsett *et al*., 2005b). In contrast, the activity of aromatase inhibitor (AIs), which are the most effective treatment of breast cancer in postmenopausal women (the largest group of patients), is based on almost complete deprivation of estrogen production (Geisler *et al*., 2002). However, although endocrine therapies are initially effective, resistance occurs both in the form of tumor relapse after excision during adjuvant treatment and as a near-universal event when tumors cannot be excised. Importantly, acquired resistance is not commonly associated with the conversion to ERα-negative of previous ERα-positive breast tumors. Nevertheless, changes in ERα expression have been found in some series (Johnston *et al*., 1995).

Current literature supports the hypothesis that acquired resistance is mainly mediated by molecular events that—particularly in the case of resistance to AIs—lead to constitutive activation of ERα and growth factor signaling pathway cross-talk (Clarke *et al*., 2003;

Correspondence: Dr Urruticoechea or Dr MA Pujana, Translational Research Laboratory, Catalan Institute of Oncology, Bellvitge Institute for Biomedical Research, Gran Vía 199-203, L'Hospitalet, Barcelona 08907, Spain.
E-mails: anderu@iconcologia.net or mapujana@iconcologia.net
Received 16 February 2010; revised 22 June 2010; accepted 28 June 2010

Chapter 7
Analysis of Population-Based Genetic Association Studies Applied to Cancer Susceptibility and Prognosis

Xavier Solé, Juan Ramón González, and Víctor Moreno

Abstract Along hundreds of thousands of years, genetic variation has been the keystone for human evolution and adaptation to the surrounding environment. Although this fact has supposed a great progress for the species, mutations in our DNA sequence may also lead to an increased risk of developing some diseases with an underlying genetic basis, such as cancer. Among different genetic epidemiology branches, population-based association studies are one of the tools that can help us decipher which of these mutations are involved in the appearance or progression of the disease. This chapter aims to be a didactic but thorough review for those who are interested in genetic association studies and its analytical methodology. It will mainly focus on SNP-array analysis techniques, covering issues such as quality control, assessment of association with disease, gene–gene and gene–environment interactions, haplotype analysis, and genome-wide association studies. In the last part, some of the existing bioinformatics tools that perform the exposed analyses will be reviewed.

7.1 Genetic Variation and Its Implication in Cancer

The implication of genes in cancer has long been suspected because this disease shows familial aggregation, in some instances remarkably. The study of cancer cells shows extensive genomic alterations, ranging from mutations in target genes – known as oncogenes and tumor suppressor genes – to large chromosomal aberrations. These alterations are supposed to be triggered by initial events that accumulate and confer the cancer cells proliferation advantage and escape to control of DNA damage. Alterations are acquired during the carcinogenesis process and are called somatic alterations. However, individuals that carry alterations in germ line are

X. Solé (✉)
Biostatistics and Bioinformatics Unit, Catalan Institute of Oncology – IDIBELL, Av. Gran Via s/n Km 2.7, 08907 L'Hospitalet de Llobregat, Barcelona, Spain
e-mail: x.sole@iconcologia.net

T. Pham (ed.), *Computational Biology: Issues and Applications in Oncology*, Applied Bioinformatics and Biostatistics in Cancer Research,
DOI 10.1007/978-1-4419-0811-7_7,

British Journal of Cancer (2009) 101, 1469 – 1480

www.bjcancer.com

Gene expression profiling integrated into network modelling reveals heterogeneity in the mechanisms of BRCA1 tumorigenesis

R Fernández-Ramires[1], X Solé[2], L De Cecco[3,4], G Llort[5], A Cazorla[6], N Bonifaci[2], MJ Garcia[1], T Caldés[7], I Blanco[5], M Gariboldi[3,4], MA Pierotti[3,4], MA Pujana[*,2], J Benítez[1] and A Osorio[*,1]

[1]Human Genetics Group Human Cancer Genetics Program, Spanish National Cancer Center (CNIO) and CIBERER, Melchor Fernández Almagro, 3, Madrid 28029, Spain; [2]Biostatistics and Bioinformatics Unit, and Translational Research Laboratory, Catalan Institute of Oncology, IDIBELL, L'Hospitalet, Av. Gran Via s/n km, 2,7, Barcelona 08907, Spain; [3]Fondazione Istituto Nazionale dei Tumori, Milan, Italy; [4]Fondazione Istituto FIRC Oncologia Molecolare, Via Giacomo Venezian 1, Milan, Milano 20133, Italy; [5]Genetic Counseling Unit, Catalan Institute of Oncology, IDIBELL, L'Hospitalet, Av. Gran Via s/n km, 2,7, Barcelona 08907, Spain; [6]Department of Pathology, Fundación Jiménez Díaz, Avda. Reyes Católicos, 2 – 28040 Madrid, Spain; [7]Clinical Oncology Laboratory, Hospital Clínico San Carlos, Profesor Martin Lagos s/n E-28040, Madrid, Spain

BACKGROUND: Gene expression profiling has distinguished sporadic breast tumour classes with genetic and clinical differences. Less is known about the molecular classification of familial breast tumours, which are generally considered to be less heterogeneous. Here, we describe molecular signatures that define BRCA1 subclasses depending on the expression of the gene encoding for oestrogen receptor, *ESR1*.

METHODS: For this purpose, we have used the Oncochip v2, a cancer-related cDNA microarray to analyze 14 BRCA1-associated breast tumours.

RESULTS: Signatures were found to be molecularly associated with different biological processes and transcriptional regulatory programs. The signature of *ESR1*-positive tumours was mainly linked to cell proliferation and regulated by ER, whereas the signature of *ESR1*-negative tumours was mainly linked to the immune response and possibly regulated by transcription factors of the REL/NFκB family. These signatures were then verified in an independent series of familial and sporadic breast tumours, which revealed a possible prognostic value for each subclass. Over-expression of immune response genes seems to be a common feature of ER-negative sporadic and familial breast cancer and may be associated with good prognosis. Interestingly, the *ESR1*-negative tumours were substratified into two groups presenting slight differences in the magnitude of the expression of immune response transcripts and REL/NFκB transcription factors, which could be dependent on the type of *BRCA1* germline mutation.

CONCLUSION: This study reveals the molecular complexity of BRCA1 breast tumours, which are found to display similarities to sporadic tumours, and suggests possible prognostic implications.

British Journal of Cancer (2009) **101,** 1469 – 1480. doi:10.1038/sj.bjc.6605275 www.bjcancer.com

Keywords: gene expression profiling; BRCA1-associated tumours; prognosis

Breast cancer is a complex disease, encompassed by different clinically and molecularly stratified entities. In 2000, Perou and colleagues demonstrated that tumour phenotypic diversity correlates with differences in global gene expression patterns, which in turn reflect aspects of the biological behaviour of the tumours (Perou *et al*, 2000). This study and subsequent ones (Sorlie *et al*, 2001; van't Veer *et al*, 2002; Bertucci *et al*, 2006) provide detailed analysis of correlations with histopathological and clinical characteristics.

The level of expression of the oestrogen receptor (ER) is a key feature that divides breast tumours into two main clusters. ER-positive tumours include the luminal A and luminal B subclasses showing different prognosis (Perou *et al*, 2000). Tumours with very low or no detectable expression of ER can be classified into HER2/ErbB2-positive, normal breast-like and basal-like (Perou *et al*, 2000; Sorlie *et al*, 2001). The first subclass is characterised by over-expression of *ERBB2* and other genes at the 17q22 amplicon. Normal breast-like tumours show high heterogeneity, with expression of genes related to the adipose tissue and other nonepithelial cells (Sorlie *et al*, 2001). Finally, the basal-like subclass is known to be negative for HER2/ErbB2, ER and the progesterone receptor (PR), and characterised by the expression of genes from the basal epithelium with high frequency of *TP53* mutations (Sorlie *et al*, 2001; Foulkes *et al*, 2004; Bertucci *et al*, 2006; Turner and Reis-Filho, 2006; Yehiely *et al*, 2006; Adelaide *et al*, 2007; Jumppanen *et al*, 2007). Basal-like tumours account for up to 15% of all breast cancers and the clinical handling of this subclass is a major challenge, once they do not respond to conventional targeted therapies.

Similar features in familial breast cancer are less clearly understood, partially due to the fact that very few studies have been published regarding expression profiling of the corresponding breast tumours. This lack of information probably

*Correspondence: Dr A Osorio; E-mail: aosorio@cnio.es
or Dr MA Pujana; E-mail: mapujana@iconcologia.net
Data deposition footnote: The data were deposited in the GEO database under the accession number [GSE12350]
Received 27 February 2009; revised 17 July 2009; accepted 27 July 2009

Molecular Cancer

BioMed Central

Commentary

Open Access

Genetic interactions: the missing links for a better understanding of cancer susceptibility, progression and treatment

Christopher A Maxwell, Víctor Moreno, Xavier Solé, Laia Gómez, Pilar Hernández, Ander Urruticoechea and Miguel Angel Pujana*

Address: Bioinformatics and Biostatistics Unit, and Translational Research Laboratory, Catalan Institute of Oncology, IDIBELL, Gran Vía km 2.7, L'Hospitalet 08907, Barcelona, Spain

Email: Christopher A Maxwell - cmaxwell@iconcologia.net; Víctor Moreno - v.moreno@iconcologia.net; Xavier Solé - x.sole@iconcologia.net; Laia Gómez - lgomez@iconcologia.net; Pilar Hernández - phgutierrez@iconcologia.net; Ander Urruticoechea - anderu@iconcologia.net; Miguel Angel Pujana* - mapujana@iconcologia.net

* Corresponding author

Published: 10 January 2008

Molecular Cancer 2008, **7**:4 doi:10.1186/1476-4598-7-4

Received: 26 October 2007
Accepted: 10 January 2008

This article is available from: http://www.molecular-cancer.com/content/7/1/4

Abstract

It is increasingly clear that complex networks of relationships between genes and/or proteins govern neoplastic processes. Our understanding of these networks is expanded by the use of functional genomic and proteomic approaches in addition to computational modeling. Concurrently, whole-genome association scans and mutational screens of cancer genomes identify novel cancer genes. Together, these analyses have vastly increased our knowledge of cancer, in terms of both "part lists" and their functional associations. However, genetic interactions have hitherto only been studied in depth in model organisms and remain largely unknown for human systems. Here, we discuss the importance and potential benefits of identifying genetic interactions at the human genome level for creating a better understanding of cancer susceptibility and progression and developing novel effective anticancer therapies. We examine gene expression profiles in the presence and absence of co-amplification of the 8q24 and 20q13 chromosomal regions in breast tumors to illustrate the molecular consequences and complexity of genetic interactions and their role in tumorigenesis. Finally, we highlight current strategies for targeting tumor dependencies and outline potential matrix screening designs for uncovering molecular vulnerabilities in cancer cells.

Background

Most of the current knowledge of cancer susceptibility, progression and treatment has been generated by traditional approaches, in which small numbers of genes or proteins are characterized in depth to study the molecular mechanisms of neoplastic processes. With the advent of large-scale functional genomic and proteomic ("omic") methodologies, additional mechanistic insights into neoplasia have been uncovered. Whole-genome association studies for cancer risk variants and somatic mutation screening projects have completed their initial phases and will provide the "part lists" of cancer genes, both at the germline [1] and the somatic levels [2]. Transcript analyses have identified expression profiles that provide accurate prognoses for cancer patients [3]. Systematic mapping of protein-protein interactions is currently being carried out in what are referred to as 'interactome' mapping projects. This research will elucidate the wiring diagram of

nature genetics

Network modeling links breast cancer susceptibility and centrosome dysfunction

Miguel Angel Pujana[1,2,16,17], Jing-Dong J Han[1,2,16,17], Lea M Starita[3,16,17], Kristen N Stevens[4,17], Muneesh Tewari[1,2,16], Jin Sook Ahn[1,2], Gad Rennert[5], Víctor Moreno[6,7], Tomas Kirchhoff[8], Bert Gold[9], Volker Assmann[10], Wael M ElShamy[2], Jean-François Rual[1,2], Douglas Levine[8], Laura S Rozek[6], Rebecca S Gelman[11], Kristin C Gunsalus[12], Roger A Greenberg[2], Bijan Sobhian[2], Nicolas Bertin[1,2], Kavitha Venkatesan[1,2], Nono Ayivi-Guedehoussou[1,2,16], Xavier Solé[7], Pilar Hernández[13], Conxi Lázaro[13], Katherine L Nathanson[14], Barbara L Weber[14], Michael E Cusick[1,2], David E Hill[1,2], Kenneth Offit[8], David M Livingston[2], Stephen B Gruber[4,6,15], Jeffrey D Parvin[3,16] & Marc Vidal[1,2]

Many cancer-associated genes remain to be identified to clarify the underlying molecular mechanisms of cancer susceptibility and progression. Better understanding is also required of how mutations in cancer genes affect their products in the context of complex cellular networks. Here we have used a network modeling strategy to identify genes potentially associated with higher risk of breast cancer. Starting with four known genes encoding tumor suppressors of breast cancer, we combined gene expression profiling with functional genomic and proteomic (or 'omic') data from various species to generate a network containing 118 genes linked by 866 potential functional associations. This network shows higher connectivity than expected by chance, suggesting that its components function in biologically related pathways. One of the components of the network is *HMMR*, encoding a centrosome subunit, for which we demonstrate previously unknown functional associations with the breast cancer–associated gene *BRCA1*. Two case-control studies of incident breast cancer indicate that the *HMMR* locus is associated with higher risk of breast cancer in humans. Our network modeling strategy should be useful for the discovery of additional cancer-associated genes.

Combinations of mutated and/or aberrantly expressed tumor suppressor genes and oncogenes, or 'cancer genes', are thought to be responsible for most steps of cancer progression. Although fundamental principles have emerged from the study of known cancer genes and their products, many questions remain unanswered. Notably, most cancer genes remain to be identified[1]. In addition, it is becoming increasingly clear that most genes and their products interact in complex cellular networks, the properties of which might be altered in cancer cells as compared with their unaffected counterparts[2]. Achieving a deeper understanding of cancer molecular mechanisms

[1]Center for Cancer Systems Biology (CCSB) and [2]Department of Cancer Biology, Dana-Farber Cancer Institute and Department of Genetics, Harvard Medical School, 44 Binney St., Boston, Massachusetts 02115, USA. [3]Department of Pathology, Brigham and Women's Hospital and Harvard Medical School, 77 Louis Pasteur Ave., Boston, Massachusetts 02115, USA. [4]Department of Epidemiology, University of Michigan, 109 Zina Pitcher Pl., Ann Arbor, Michigan 48109, USA. [5]CHS National Cancer Control Center, Department of Community Medicine and Epidemiology, Carmel Medical Center and Bruce Rappaport Faculty of Medicine, Technion, Haifa 34362, Israel. [6]Department of Internal Medicine, University of Michigan, 109 Zina Pitcher Pl., Ann Arbor, Michigan 48109, USA. [7]Department of Epidemiology and Cancer Registry, and Translational Research Laboratory, Catalan Institute of Oncology, IDIBELL, Gran Vía km 2.7, L'Hospitalet, Barcelona 08907, Spain. [8]Clinical Genetics Service, Department of Medicine, Memorial Sloan-Kettering Cancer Center, 1275 York Ave., New York, New York 10021, USA. [9]National Cancer Institute, Human Genetics Section, Laboratory of Genomic Diversity, Frederick, Maryland 21702, USA. [10]Center for Experimental Medicine, Institute of Tumor Biology, University Hospital Hamburg–Eppendorf, Martinistrasse 52, Hamburg 20246, Germany. [11]Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute and Department of Biostatistics, Harvard School of Public Health, 44 Binney St., Boston, Massachusetts 02115, USA. [12]Center for Comparative Functional Genomics, Department of Biology, New York University, 100 Washington Square East, New York, New York 10003, USA. [13]Translational Research Laboratory, Catalan Institute of Oncology, IDIBELL, Gran Vía km 2.7, L'Hospitalet, Barcelona 08907, Spain. [14]Abramson Family Cancer Research Institute, University of Pennsylvania School of Medicine, 421 Curie Blvd., Philadelphia, Pennsylvania 19104, USA. [15]Department of Human Genetics, University of Michigan, 109 Zina Pitcher Pl., Ann Arbor, Michigan 48109, USA. [16]Present addresses: Bioinformatics and Biostatistics Unit, Translational Research Laboratory, Catalan Institute of Oncology, IDIBELL, Gran Vía km 2.7, L'Hospitalet, Barcelona 08907, Spain (M.A.P.); Institute of Genetics and Developmental Biology, Chinese Academy of Sciences, Datun Rd., Beijing 100101, China (J.-D.J.H.); Department of Genome Sciences, University of Washington, 1705 NE Pacific St., Seattle, Washington 98195, USA (L.M.S.); Human Biology Division, Fred Hutchinson Cancer Research Center, 1100 Fairview Ave. North, Seattle, Washington 98109, USA (M.T.); Harvard School of Public Health, Boston, Massachusetts 02115, USA (N.A.-G.); Department of Biomedical Informatics, Ohio State University Medical Center, 460 West 12th Ave., Columbus, Ohio 43210, USA (J.D.P.). [17]These authors contributed equally to this work. Correspondence should be addressed to M.V. (marc_vidal@dfci.harvard.edu), J.D.P. (jeffrey.parvin@osumc.edu) or S.B.G. (sgruber@med.umich.edu).

Received 31 March; accepted 2 August; published online 7 October 2007; doi:10.1038/ng.2007.2

Breast Cancer Res Treat (2008) 109:463–470
DOI 10.1007/s10549-007-9676-7

PRECLINICAL STUDY

Pre-clinical validation of early molecular markers of sensitivity to aromatase inhibitors in a mouse model of post-menopausal hormone-sensitive breast cancer

Ander Urruticoechea · Helena Aguilar · Xavier Solé · Gabriel Capellà · Lesley-Ann Martin · Mitch Dowsett · Josep Ramon Germà-Lluch

Received: 21 June 2007 / Accepted: 26 June 2007 / Published online: 19 July 2007

Abstract *Introduction* Changes in breast cancer cell biology following hormonal treatment have been claimed as promising predictor markers of clinical benefit even outperforming clinical response. From previous work we selected 10 genes showing both a well known regulation by oestrogen and a high level of early transcriptional regulation following therapy with aromatase inhibitors. Here we use an animal breast cancer model to explore the feasibility of the determination of their expression in minimally invasive samples and to further assess the magnitude of their regulation by letrozole. *Animal and methods* Aromatase inhibitor sensitive breast cancer tumours were grown in athymic mice under supplement with androstenedione. Following initial tumour growth animals were assigned to a control group or to receive letrozole at two different dosages. Fine needle aspirates were obtained at the moment of treatment assignation and one week later. Expression of the following genes at both time points was determined: Ki-67, Cyclin D1, pS2, Trefoil Factor 3, PDZ domain containing 1, Ubiquitin-conjugating enzyme E2C, Stanniocalcin 2, Topoisomerase 2 alfa, MAN1A1 and FAS. *Results* Fine needles aspirates were found to be a feasible and reproducible technique for RNA extraction. Trefoil Factor 3, pS2, Cyclin D1 and Stanniocalcin 2 were significantly downregulated by letrozole. Among them pS2 appears to be most sensitive to aromatase inhibitor treatment even differentiating sub-optimal from optimal letrozole dosage. *Discussion* We present pre-clinical evidence to justify the exploration in clinical trials of pS2, Trefoil factor 3, Cyclin D1 and Stanniocalcin as dynamic markers of oestrogen-driven pathway activation.

Keywords Animal model · Aromatase inhibitors · Biomarkers · Breast cancer · Endocrine treatment · Oestrogen receptor

Ander Urruticoechea and Helena Aguilar contributed equally to this work.

A. Urruticoechea (✉) · H. Aguilar · G. Capellà · J. R. Germà-Lluch
Translational Research Laboratory, Institut Català d'Oncologia, IDIBELL, Gran via s/n, km 2.7, L'Hospitalet de Llobregat, Barcelona 08907, Spain
e-mail: anderu@iconcologia.net

X. Solé
Bioinformatics Unit, Cancer Epidemiology and Registry Service, Institut Català d'Oncologia, IDIBELL, L'Hospitalet de Llobregat, Barcelona, Spain

L.-A. Martin
Chester Beatty Laboratories, Breakthrough Toby Robins Breast Cancer Research Centre, Institute of Cancer Research, Mary-Jean Mitchell Green Building, 237, Fulham Road, London, UK

M. Dowsett
Academic Department of Biochemistry, Royal Marsden Hospital, London, UK

Introduction

Breast glandular epithelium grows and differentiates under the stimulus of oestradiol. Once breast cancer develops from epithelial progenitors oestradiol deprivation can result in tumour regression. The nuclear oestrogen receptor α (ER) is the most important predictor of benefit derived from hormonal treatments [1] and despite multiple reports on novel determinants of hormone-sensitivity no other marker has been introduced into routine practice. Yet 40–50% of ER positive tumours do not respond to the best hormonal treatment strategy (i.e. aromatase inhibitors in the postmenopausal woman) [2].

Available online at www.sciencedirect.com
ScienceDirect
Journal of Biomedical Informatics 41 (2008) 33–45
Journal of Biomedical Informatics
www.elsevier.com/locate/yjbin

CLEAR-test: Combining inference for differential expression and variability in microarray data analysis

Joan Valls [a], Mònica Grau [a], Xavier Solé [a], Pilar Hernández [a], David Montaner [b], Joaquín Dopazo [b], Miguel A. Peinado [c], Gabriel Capellá [a], Víctor Moreno [a], Miguel Angel Pujana [a,*]

[a] *Bioinformatics and Biostatistics Unit, and Translational Research Laboratory, Catalan Institute of Oncology, IDIBELL, L'Hospitalet, 08907 Barcelona, Spain*
[b] *Department of Bioinformatics, CIPF, 46013 Valencia, Spain*
[c] *Cancer Research Institute, IDIBELL, L'Hospitalet, 08907 Barcelona, Spain*

Received 11 August 2006
Available online 17 May 2007

Abstract

A common goal of microarray experiments is to detect genes that are differentially expressed under distinct experimental conditions. Several statistical tests have been proposed to determine whether the observed changes in gene expression are significant. The *t*-test assigns a score to each gene on the basis of changes in its expression relative to its estimated variability, in such a way that genes with a higher score (in absolute values) are more likely to be significant. Most variants of the *t*-test use the complete set of genes to influence the variance estimate for each single gene. However, no inference is made in terms of the variability itself. Here, we highlight the problem of low observed variances in the *t*-test, when genes with relatively small changes are declared differentially expressed. Alternatively, the *z*-test could be used although, unlike the *t*-test, it can declare differentially expressed genes with high observed variances. To overcome this, we propose to combine the *z*-test, which focuses on large changes, with a χ^2 test to evaluate variability. We call this procedure CLEAR-test and we provide a combined *p*-value that offers a compromise between both aspects. Analysis of three publicly available microarray datasets reveals the greater performance of the CLEAR-test relative to the *t*-test and alternative methods. Finally, empirical and simulated data analyses demonstrate the greater reproducibility and statistical power of the CLEAR-test and *z*-test with respect to current alternative methods. In addition, the CLEAR-test improves the *z*-test by capturing reproducible genes with high variability.

Keywords: Microarrays; Differential expression; Gene expression

1. Background

In recent years, functional genomic studies based on microarray gene expression analysis have emerged as a powerful strategy through which to decipher cellular processes, pathways or pathology. The vast number of microarray studies published to date provides an opportunity to develop new systems-level, integrated approaches to the understanding of biological processes. However, one of the limitations of microarray-based studies is the validation of the results, that is the set of genes that are declared differentially expressed between distinct experimental conditions. The success of this validation is influenced not only by the use of standardized protocols [1–3] but also by the statistical method chosen to determine significance. A variety of methods are currently available and concepts such as statistical sophistication and biological interpretation must be taken into account in order to select one or other [4–6].

Once the raw data have been pre-processed and normalized, a statistical method is needed to assess the evidence

* Corresponding author. Fax: +34 93 260 74 66.
E-mail address: mapujana@ico.scs.es (M.A. Pujana).

1532-0464/$ - see front matter © 2007 Elsevier Inc. All rights reserved.
doi:10.1016/j.jbi.2007.05.005

Genomic and transcriptomic prognostic factors in R0 Dukes B and C colorectal cancer patients

ELISENDA VENDRELL[1], MARIA RIBAS[1], JOAN VALLS[2], XAVIER SOLÉ[2], MONICA GRAU[2,3], VICTOR MORENO[2], GABRIEL CAPELLÀ[2] and MIGUEL A. PEINADO[1]

[1]Institut d'Investigació Biomèdica de Bellvitge (IDIBELL), [2]IDIBELL-Institut Català d'Oncologia, L'Hospitalet, Barcelona, Catalonia, Spain

Abstract. The advent of various 'omic' technologies has increased expectations in the field of biomarkers. In an attempt to clarify how different strategies may contribute to improving prognostic classification and to identify new predictors of patient outcome we analyzed genomic and transcriptomic profiles in a series of R0 Dukes B and C colorectal carcinomas. We have compared the predictive capability of each approach against conventional clinicopathological and molecular parameters. At a genomic level, gains at 11q including amplification at 11q13 were an indicator of poorer outcome. In transcriptomic analyses we identified 68 genes whose expression levels correlated with survival ($p<0.01$) and included overexpression of WASF1, NFE2L2, and MMP9, and underexpression of ITGAL, TSC2, and SDF2. Gene expression levels paralleled chromosomal changes only in 56% of the genes, suggesting that, as a general trend, the direct effect of chromosomal copy number changes on gene expression levels is minimal. Classification of tumors by genomic and transcriptomic signatures resulted in non-overlapping subgroups and was not of prognostic value. We conclude that genomic and transcriptomic profiling of colorectal carcinomas may contribute as novel prognostic markers, but it does not improve outcome prediction when global profiles or signatures are considered.

Introduction

Colorectal cancer is the third most common type of neoplasia in both men and women and the second-leading cause of death by cancer in occidental countries (1). The extent of tumor bowel wall infiltration and lymph node metastases, both included in Dukes' stage and TNM classification systems, are the most important prognostic factors in colorectal cancer (2). Nevertheless, traditional morphologic criteria based upon pathologist's evaluation are accurate for predicting recurrence only in 50-75% of the patients with non-metastatic invasive colon carcinoma. Therefore there is a need for additional, less subjective, independent factors to better predict outcome.

Multiple genetic aberrations are required for tumor initiation and progression of colorectal cancer, which is one of the best studied systems of multistage human carcinogenesis. Besides the advances in the understanding of the molecular factors involved in this process, the heterogeneity and complexity of the disease make it difficult to apply molecular information to predict the evolution of an individual patient's disease (3). A major challenge is to integrate information that can describe this complexity so as to facilitate an understanding of the disease mechanisms as well as to guide the development and application of therapies (4). The advent of various 'omic' technologies has increased expectations in the field of biomarkers, but they have not yet produced widely applicable approaches in prognostic assessment and patient treatment. Four levels of analyses can be considered: genomic, transcriptomic, epigenomic, and proteomic, the first two being the most often applied due to the availability of appropriate methodologies.

Chromosomal aberrations in the form of aneuploidy and structural rearrangements are early markers and probably the most prevalent genetic alteration in colorectal carcinogenesis (5,6). Recurrent chromosomal abnormalities often clustered in association patterns are also observed and may be used to classify colorectal cancers (6,7). Furthermore, a subset of colorectal tumors with few or no chromosomal alterations are characterized by ubiquitous somatic mutations at repeated sequences (6-8). These tumors represent a distinctive pathway of tumor progression in which defects in the DNA mismatch repair machinery underlie the genetic instability expressed as an exacerbated microsatellite instability (MSI).

Conventional G-banding cytogenetics has been instrumental in the identification of the chromosomal alterations associated with malignancy and has provided potential prognostic markers in colorectal cancer (9). The availability of comparative genomic hybridization (CGH) (10) as an alternative to classic cytogenetics has facilitated karyotyping and nowadays is the

Correspondence to: Dr Miguel A. Peinado, Institut d'Investigació Biomèdica de Bellvitge (IDIBELL), Av. Granvia km 2.7, 08907 L'Hospitalet, Barcelona, Catalonia, Spain
E-mail: mpeinado@idibell.org

Present address: [3]Microarrays Unit, Centre de Regulació Genòmica, Barcelona, Catalonia, Spain

Key words: colorectal cancer, prognostic factor, chromosomal alterations, transcriptomic profiles

BIOINFORMATICS APPLICATIONS NOTE Vol. 23 no. 5 2007, pages 644–645
doi:10.1093/bioinformatics/btm025

Genetics and population analysis

SNPassoc: an R package to perform whole genome association studies

Juan R. González[1,*], Lluís Armengol[1], Xavier Solé[2], Elisabet Guinó[2], Josep M. Mercader[1], Xavier Estivill[1] and Víctor Moreno[2,*]

[1]Genes and Disease Program, Centre for Genomic Regulation and [2]Unit of Biostatistics and Bioinformatics, Epidemiology Service, IDIBELL, Catalan Institute of Oncology, Barcelona, Spain

Received on September 27, 2006; revised and accepted January 22, 2007

Advance Access publication January 31, 2007

Associate Editor: Keith Crandall

ABSTRACT

Summary: The popularization of large-scale genotyping projects has led to the widespread adoption of genetic association studies as the tool of choice in the search for single nucleotide polymorphisms (SNPs) underlying susceptibility to complex diseases. Although the analysis of individual SNPs is a relatively trivial task, when the number is large and multiple genetic models need to be explored it becomes necessary a tool to automate the analyses. In order to address this issue, we developed SNPassoc, an R package to carry out most common analyses in whole genome association studies. These analyses include descriptive statistics and exploratory analysis of missing values, calculation of Hardy–Weinberg equilibrium, analysis of association based on generalized linear models (either for quantitative or binary traits), and analysis of multiple SNPs (haplotype and epistasis analysis).
Availability: Package SNPassoc is available at CRAN from http://cran.r-project.org
Contact: juanramon.gonzalez@crg.es or v.moreno@iconcologia.net
Supplementary information: A tutorial is available on *Bioinformatics* online and in http://davinci.crg.es/estivill_lab/snpassoc

1 INTRODUCTION

Whole genome association studies, in which a dense set of SNPs across the genome is genotyped, are a novel approach to assess the role of genetic variation in disease. To increase the efficiency of this approach, multistage designs have been proposed (Hirschhorn and Daly, 2005). In the first step, thousands of SNPs are tested for association with the disease. In a second and possibly third step, additional detailed studies are performed, in which only a few hundred SNPs, those with a putative association found in the first step, are genotyped.

Although analysis of a single or a small number of SNPs is a relatively simple task to conduct (Solé *et al*., 2006), the statistical analysis of large-scale studies is challenging. In this article, we present *SNPassoc*, an R package (http://www.r-project.org) designed to analyze genome-wide association studies. *SNPassoc* contains tools for data manipulation, exploratory data analysis with graphics, and assessment of genetic association for both quantitative and binary traits. For the analysis of a small selection of SNPs, the package also provides tools to analyze interactions between SNPs or haplotypes and other SNPs or environmental variables. This note presents an overview of the package but a detailed tutorial is provided in the Supplementary Material.

*To whom correspondence should be addressed.

1.1 Data manipulation and descriptive analysis

SNPassoc uses the object-oriented features of R ('classes and methods') to ease data manipulation, analysis and plots. Variables coding for SNP genotypes are defined with the function `snp`, which takes care of formatting and assigns class 'snp'. The recommended format delimits each allele with a character (i.e. '/'), but two-letter formats or any three codes are also allowed. Objects of class 'snp' can be explored using the generic R functions `print`, `summary` and `plot`. The summary of a 'snp' object shows genotype and allele frequencies, missing values and a test for compliance with Hardy–Weinberg equilibrium. By default, the reference category is the genotype homozygous for the most frequent allele. This may be changed using the method `reorder`.

If the user has a large collection of SNPs coded similarly, the function `setupSNP` prepares the data automatically. Information about chromosome and genomic positions, if given, is used later to classify or sort SNPs in tables and plots. The function `setupSNP` returns a packed object that can be explored and analyzed with a series of functions. For example, the generic function `summary` provides a table with a systematic descriptive analysis, including allele frequencies, percentage of missing values and the test for Hardy–Weinberg equilibrium. This test may also be obtained using the function `tableHWE`, which uses an exact test of Hardy–Weinberg equilibrium as described in Wigginton *et al*. (2005). The function `plotMissing` provides a visual representation of missing values in samples and SNPs (Fig. 2 in the Supplementary Material). Objects with class 'setupSNP' can be manipulated after their creation: variables can be added or deleted and subsets of SNPs can be selected for specific analysis.

1.2 Whole genome association studies

After initial inspection of the data, analysis of association can be performed using the function `WGassociation`, which requires an object of class 'setupSNP'. To demonstrate how to perform this analysis using a real dataset, we have downloaded individual genotypes from the HapMap project (http://www.hapmap.org) and randomly selected close to 10 000 SNPs distributed across the 22 autosomes. We compare the genotype frequencies for all SNPs from this dataset between the European (CEU) and African (YRI) populations. The dataset and the genomic information are loaded typing `data(HapMap)`. The required object of class 'setupSNP' is created executing:

```
myDat<-setupSNP(HapMap, colSNPs=3:9809,
      sort=TRUE, info=SNPs.pos,
      sep='' '')
```

GENES, CHROMOSOMES & CANCER 46:155–162 (2007)

Molecular Characterization of a t(9;12)(p21;q13) Balanced Chromosome Translocation in Combination with Integrative Genomics Analysis Identifies *C9orf14* as a Candidate Tumor-Suppressor

Miguel Angel Pujana,[1]* Anna Ruiz,[2†] Cèlia Badenas,[3] Josep-Anton Puig-Butille,[3,4] Marga Nadal,[2] Mitchell Stark,[5] Laia Gómez,[1] Joan Valls,[1] Xavier Solé,[1] Pilar Hernández,[1] Celia Cerrato,[6] Irene Madrigal,[3] Rafael de Cid,[6] Helena Aguilar,[1] Gabriel Capellá,[1] Santiago Cal,[7] Michael R. James,[5] Graeme J. Walker,[5] Josep Malvehy,[4] Montserrat Milà,[3] Nicholas K. Hayward,[5] Xavier Estivill,[6] and Susana Puig[4]

[1]Bioinformatics and Biostatistics Unit, Translational Research Laboratory, Catalan Institute of Oncology, IDIBELL, L'Hospitalet, Barcelona, Spain
[2]Cancer Research Institute, IDIBELL, L'Hospitalet, Barcelona, Spain
[3]Biochemistry and Molecular Genetics Service, Hospital Clínic, Barcelona, Spain
[4]Department of Dermatology, Melanoma Unit, Hospital Clínic, Barcelona, Spain
[5]Queensland Institute of Medical Research, Brisbane, Australia
[6]Genes and Disease Program, Center for Genomic Regulation (CRG), University Pompeu Fabra (UPF), Barcelona Biomedical Research Park, Barcelona, Spain
[7]Departamento de Bioquímica y Biología Molecular, Facultad de Medicina, Instituto Universitario de Oncología, Universidad de Oviedo, Oviedo, Spain

A large number of nevi (LNN) is a high risk phenotypic trait for developing cutaneous malignant melanoma (CMM). In this study, the breakpoints of a t(9;12)(p21;q13) balanced chromosome translocation were finely mapped in a family with LNN and CMM. Molecular characterization of the 9p21 breakpoint identified a novel gene *C9orf14* expressed in melanocytes disrupted by the translocation. Integrative analysis of functional genomics data was applied to determine the role of *C9orf14* in CMM development. An analysis of genome-wide DNA copy number alterations in melanoma tumors revealed the loss of the *C9orf14* locus, located proximal to *CDKN2A*, in approximately one-fourth of tumors. Analysis of gene expression data in cancer cell lines and melanoma tumors suggests a loss of *C9orf14* expression in melanoma tumorigenesis. Taken together, our results indicate that *C9orf14* is a candidate tumor-suppressor for nevus development and late stage melanoma at 9p21, a region frequently deleted in different types of human cancers. This article contains Supplementary Material available at http://www.interscience.wiley.com/jpages/1045-2257/suppmat.

INTRODUCTION

Cutaneous malignant melanoma (CMM) is a potentially fatal type of skin cancer with increasing incidence and mortality world wide (Rigel, 1996; Jemal et al., 2003). A major etiological factor in the development of CMM is sunlight exposure (Pho et al., 2006). In addition, epidemiological studies have revealed that of a number of phenotypic traits, the highest risk of developing CMM is conferred by the presence of a large number of nevi (LNN) (Swerdlow and Green, 1987; Grob et al., 1990; Bataille et al., 1996; Briollais et al., 2000).

As with CMM, sunlight exposure is also the major etiological factor in nevus ontogenesis. However, the association between sunlight exposure, CMM, and nevus development is relatively complex. In addition to environmental factors, there are known and unidentified genetic factors that contribute to both phenotypes either independently or in association. Thus, total nevi counts and nevus density show familial aggregation (Goldgar et al., 1991; Duffy et al., 1992). Familial aggrega-

Supported by: Fondo de Investigaciones Sanitarias (FIS) of the Spanish Ministry of Health; Grant numbers: 1546-01 and 0019-03; Instituto de Salud Carlos III; Grant number: V2003-REDC03/03-07; U.S. National Cancer Institute; Grant number: CA83115; Fundació "la Caixa"; Grant number: BM05-254-00; Department of Universities, Research and Information Society; Health Department of the Generalitat de Catalunya; The Ramón y Cajal Program of the Spanish Ministry of Education and Science.

*Correspondence to: Miguel Angel Pujana, Ph.D., Bioinformatics and Biostatistics Unit, Translational Research Laboratory, Catalan Institute of Oncology, IDIBELL, L'Hospitalet, 08907 Barcelona, Spain. E-mail: mapujana@ico.scs.es

†Present address: Department of Molecular Haematology, Institute of Child Health, WC1N 1EH London, UK.

Received 15 June 2006; Accepted 2 October 2006

DOI 10.1002/gcc.20396

Published online 10 November 2006 in Wiley InterScience (www.interscience.wiley.com).

BIOINFORMATICS APPLICATIONS NOTE Vol. 22 no. 15 2006, pages 1928–1929
doi:10.1093/bioinformatics/btl268

Genetics and population analysis

SNPStats: a web tool for the analysis of association studies

Xavier Solé[1], Elisabet Guinó[1], Joan Valls[1,2], Raquel Iniesta[1] and Víctor Moreno[1,2,*]

[1]Catalan Institute of Oncology, IDIBELL, Epidemiology and Cancer Registry, L'Hospitalet, Barcelona, Spain and [2]Autonomous University of Barcelona, Laboratory of Biostatistics and Epidemiology, Bellaterra, Barcelona, Spain

Received on March 6, 2006; revised on May 16, 2006; accepted on May 18, 2006
Advance Access publication May 23, 2006
Associate Editor: Charlie Hodgman

ABSTRACT

Summary: A web-based application has been designed from a genetic epidemiology point of view to analyze association studies. Main capabilities include descriptive analysis, test for Hardy–Weinberg equilibrium and linkage disequilibrium. Analysis of association is based on linear or logistic regression according to the response variable (quantitative or binary disease status, respectively). Analysis of single SNPs: multiple inheritance models (co-dominant, dominant, recessive, over-dominant and log-additive), and analysis of interactions (gene–gene or gene–environment). Analysis of multiple SNPs: haplotype frequency estimation, analysis of association of haplotypes with the response, including analysis of interactions.

Availability: http://bioinfo.iconcologia.net/SNPstats. Source code for local installation is available under GNU license.

Contact: v.moreno@iconcologia.net

Supplementary Information: Figures with a sample run are available on *Bioinformatics* online. A detailed online tutorial is available within the application.

The analysis of association between genetic polymorphisms and diseases allows identifying susceptibility genes (Cordell and Clayton, 2005). The proper analysis of these studies can be performed with general purpose statistical packages, but the researcher usually needs the assistance of additional software to perform specific analysis, like haplotype estimation, and results from different packages are difficult to integrate.

We present a free web-based tool to help researchers in the analysis of association studies based on SNPs or biallelic markers. Both the selection of analysis and the output have been designed from a genetic epidemiology perspective. This application can also be used for learning purposes. We have written (in Spanish) an analysis guide with detailed explanations (Iniesta *et al*., 2005). A similar extensive help in English can also be found on the website.

The software is used following three steps, with the possibility of performing multiple analyses in one session. The steps are as follows.

(1) *Data entry*. Raw data in tabular form can be pasted in a window or uploaded from a text file. Variables can be named and the user can choose the field delimiter and the missing value code (Supplementary Figure 1). SNPs should be coded as genotypes with each allele separated by a slash (e.g. 'T/T', 'T/C', 'C/C').

*To whom correspondence should be addressed.

(2) *Data processing*. A list with the variables read by the application is presented with an initial suggestion about the type: quantitative, categorical or SNP, which can be modified (Supplementary Figure 2). The user is prompted to select those needed for the analysis and to specify which one is the response, which may be binary (disease status) or quantitative. For categorical variables, including SNPs, the user can reorder the categories. The first one will be treated as reference category in the analysis. The application assumes that the main interest is the analysis of the SNPs in relation to the response. Other variables selected with type quantitative or categorical will be added to the regression models for analysis as covariates and treated as potential confounders.

(3) *Analyses customization*. The third step requests the selection of the desired statistical analyses that will be described later in this article (Supplementary Figure 3).

Regarding the statistical analysis, the association with disease is modeled depending on the response variable. If binary, the application assumes an unmatched case–control design and unconditional logistic regression models are used. If the response is quantitative, then a unique population is assumed and linear regression models are used to assess the proportion of variation in the response explained by the SNPs.

The association for each SNP is analyzed in turn and adjusted for the selected covariates. If more than one SNP are selected, then the application assumes that haplotype analysis is appropriate. Haplotype frequencies are estimated using the implementation of the EM algorithm coded into the *haplo.stats* package (Sinnwell and Schaid, 2005, http://mayoresearch.mayo.edu/mayo/research/biostat/schaid.cfm). Association between haplotypes and disease appropriately accounts for the uncertainty in the estimation of haplotypes for individuals with multiple heterozygous when phase is unknown or when missing values are present (Schaid *et al*., 2002). Individuals with missing values in the response, in all SNPs or in any covariate are excluded from analysis.

The software main page can be found online at http://bioinfo.iconcologia.net/SNPstats. The application uses PHP server programming language to build the input forms, upload data, call the statistical analysis procedures and process the output. The statistical analyses are performed in a batch call to the R package (R Development Core Team, 2005, http://www.R-project.org). The contributed packages *genetics* (Warnes and Leisch, 2005) and *haplo.stats* (Sinnwell and Schaid, 2005, http://mayoresearch.mayo.edu/mayo/research/biostat/schaid.cfm) are called to perform some of the analysis. Anonymous use is guaranteed and data are

Clinical Chemistry 51:1
93–101 (2005)

Cancer Diagnostics

Validation of RNA Arbitrarily Primed PCR Probes Hybridized to Glass cDNA Microarrays: Application to the Analysis of Limited Samples

Mònica Grau,[1†] Xavier Solé,[1†] Antònia Obrador,[1] Gemma Tarafa,[1] Elisenda Vendrell,[2] Joan Valls,[1] Victor Moreno,[1] Miquel A. Peinado,[2] and Gabriel Capellá[1*]

Background: **The applicability of microarray-based transcriptome massive analysis is often limited by the need for large amounts of high-quality RNA. RNA arbitrarily primed PCR (RAP-PCR) is an unbiased fingerprinting PCR technique that reduces both the amount of initial material needed and the complexity of the transcriptome. The aim of this study was to evaluate the feasibility of using hybridization of RAP-PCR products as transcriptome representations to analyze differential gene expression in a microarray platform.**

Methods: **RAP-PCR products obtained from samples with limited availability of biological material, such as experimental metastases, were hybridized to conventional cDNA microarrays. We performed replicates of self-self hybridizations of RAP-PCR products and mathematical modeling to assess reproducibility and sources of variation.**

Results: **Gene/slide interaction (47.3%) and the PCR reaction (33.8%) accounted for the majority of the variability. From these observations, we designed a protocol using two pools of three independent RAP-PCR reactions coming from two independent reverse transcription reactions hybridized in duplicate and evaluated them in the analyses of paired xenograft-metastases samples. Using this approach, we found that *HER2* and *MMP7* may be down-regulated during distal dissemination of colorectal tumors.**

Conclusion: **RAP-PCR glass array hybridization can be used for transcriptome analysis of small samples.**

The advent of techniques for the massive analysis of cell transcriptomes by use of microarrays has allowed the description of molecular portraits of biological specimens, including tumor biopsies. Tumor gene expression profiles appear to be useful tools for tumor classification and may reveal individual markers with diagnostic or prognostic applications. Nevertheless, routine application of gene expression profiles is often precluded by the demanding conditions of this type of assay, including the need for large amounts of RNA and the difficulties in performing global validation studies.

Several strategies have been developed to allow the analysis of small samples in which the amount of available RNA (<1 μg of total RNA) is insufficient for massive gene expression studies *(1–4)*. Trenkle et al. *(1)* proposed the use of nonstoichiometric reduced-complexity probes for hybridization to cDNA arrays and noted the fitness of the RNA arbitrarily primed PCR (RAP-PCR)[3] method. RAP-PCR is an unbiased fingerprinting PCR that samples a reproducible subset of message population based on the best matches with arbitrary primers *(5)*. It allows the construction of a probe with reduced complexity, which increases the representation of rare messages, and uses small amounts of total RNA (10–100 ng) or mRNA (0.1–1 ng).

Although different RAP-PCR fingerprints give hybrid-

Institut d'Investigació Biomèdica de Bellvitge (IDIBELL)-Institut Català d'Oncologia, [1] Translational Research Laboratory, Unit of Biostatistics and Bioinformatics, Cancer Epidemiology Department, and [2] IDIBELL-Institut de Recerca Oncològica Molecular Oncology Center, L'Hospitalet de Llobregat, Barcelona, Spain.

†These authors contributed equally to this work.

*Address correspondence to this author at: Institut Català d'Oncologia, Laboratori de Recerca Translacional, Av. Gran Via s/n, Km 2.7, 08907 L'Hospitalet de Llobregat, Barcelona, Spain. Fax 34-93-2607466; e-mail gcapella@ico.scs.es.

Received April 29, 2004; accepted September 24, 2004.

Previously published online at DOI: 10.1373/clinchem.2004.036236

[3] Nonstandard abbreviations: RAP-PCR, RNA arbitrarily primed PCR; SSC, standard saline citrate; and qRT-PCR, quantitative reverse transcription-PCR.

Human Molecular Genetics, 2005, Vol. 14, No. 2 **319–326**
doi:10.1093/hmg/ddi028
Advance Access published on December 1, 2004

Differential DNA hypermethylation and hypomethylation signatures in colorectal cancer

Jordi Frigola[1], Xavier Solé[2], Maria F. Paz[3], Victor Moreno[2], Manel Esteller[3], Gabriel Capellà[2] and Miguel A. Peinado[1,*]

[1]IDIBELL-Institut de Recerca Oncològica and [2]IDIBELL-Institut Català d'Oncologia, L'Hospitalet, Barcelona, Spain and [3]Cancer Epigenetics Laboratory, Spanish National Cancer Center (CNIO), Madrid, Spain

Received August 27, 2004; Revised October 28, 2004; Accepted November 23, 2004

Cancer cells are characterized by a generalized disruption of the DNA methylation pattern involving an overall decrease in the level of 5-methylcytosine together with regional hypermethylation of particular CpG islands. The extent of both DNA hypomethylation and hypermethylation in the tumor cell is likely to reflect distinctive biological and clinical features, although no studies have addressed its concurrent analysis until now. DNA methylation profiles in sporadic colorectal carcinomas, synchronous adenoma–carcinoma pairs and their matching normal mucosa were analyzed by using the amplification of inter-methylated sites (AIMS) method. A total of 208 AIMS generated sequences were tagged and evaluated for differential methylation. Global indices of hypermethylation and hypomethylation were calculated. All tumors displayed altered patterns of DNA methylation in reference to normal tissue. On average, 24% of the tagged sequences were differentially methylated in the tumor in regard to the normal pair with an overall prevalence of hypomethylations to hypermethylations. Carcinomas exhibited higher levels of hypermethylation than did adenomas but similar levels of hypomethylation. Indices of hypomethylation and hypermethylation showed independent correlations with patient's sex, tumor staging and specific gene hypermethylation. Hierarchical cluster analysis revealed two main patterns of DNA methylation that were associated to particular mutational spectra in the *K-ras* and the *p53* genes and alternative correlates of hypomethylation and hypermethylation with survival. We conclude that DNA hypermethylation and hypomethylation are independent processes and appear to play different roles in colorectal tumor progression. Subgroups of colorectal tumors show specific genetic and epigenetic signatures and display distinctive correlates with overall survival.

INTRODUCTION

Colorectal cancer is one of the best-studied systems of multistage human carcinogenesis. Epigenetic modification of DNA in the form of hypomethylation was included in early Vogelstein's tumor progression model together with a series of genetic alterations (1). DNA methylation is a post-replication modification predominantly found in cytosines of the dinucleotide CpG that is infrarepresented throughout the genome except at small regions named CpG islands (2). The pattern of DNA methylation in a given cell appears to be associated with the stability of gene expression states (3).

The biological significance of DNA hypomethylation, an early and common feature in colorectal cancer (4), is poorly understood (5). A relationship between global hypomethylation and genetic instability has been postulated (5,6). More recently, the attention of investigators has shifted to the study of cancer-associated regional hypermethylation at specific CpG islands and its association to transcriptional silencing (7,8) and loss of imprinting (9). Inspite of the large number of studies that have investigated cancer-associated hypermethylation in selected CpG islands, the obtention of global estimates of genome hypermethylation has been seldomly addressed (3,10,11).

Therefore, the roles of cumulated hypermethylation and hypomethylation in colorectal cancer progression and outcome are still unknown. By application of a methylome fingerprinting technique (amplification of inter-methylated sites, AIMS) (12), we have obtained information on the methylation status of more than 200 selected sequences in a

*To whom correspondence should be addressed at: IDIBELL-Institut de Recerca Oncològica, Hospital Duran i Reynals, Granvia km 2.7, 08907 L'Hospitalet, Barcelona, Spain. Tel: +34 932607464; Fax: +34 932607466; Email: mpeinado@iro.es

Uso de chips de ADN *(microarrays)* en medicina: fundamentos técnicos y procedimientos básicos para el análisis estadístico de resultados

Víctor Moreno y Xavier Solé

Unidad de Bioestadística y Bioinformática. Servicio de Epidemiología y Registro del Cáncer. Instituto Catalán de Oncología. Hospital Duran i Reynals. L'Hospitalet de Llobregat. Barcelona. España.

La tecnología de *microarrays* de ADN permite realizar análisis genéticos sobre miles de genes simultáneamente. El análisis de estos experimentos supone un reto desde el punto de vista estadístico, ya que los métodos clásicos de análisis deben adaptarse a la enorme multiplicidad de hipótesis que se prueban. Además, la gran variabilidad observada en los experimentos y su elevado coste exigen un diseño cuidadoso. En esta revisión se explicará con detalle qué es un microarray de ADNc, cómo funciona y cuáles son sus principales usos. Seguidamente, se abordarán aspectos estadísticos del diseño experimental y de los diferentes apartados del análisis de un *microarray*, desde el procesamiento de la imagen y control de calidad de los datos hasta los tests para identificar genes de interés. Por último se comentarán diferentes técnicas de análisis multivariante que se pueden utilizar para analizar patrones en la expresión de los genes.

Palabras clave: Microarray de ADN. Análisis estadístico. Diseño de experimentos.

Use of DNA chips (microarrays) in medicine: technical foundations and basic procedures for statistical analysis of results

DNA microarray technology allows the assessment of genetic analyses on thousands of genes simultaneously. The statistical analyses of these experiments are challenging since a high number of multiple hypotheses are tested and classical statistical methods need to adapt to this situation. Furthermore, the great variability observed in the experiments and their high cost of them needs a careful design. In this review we will explain what is a cDNA microarray, how it works and its potential uses. Later we will deal with statistical issues of design and analysis, from the image processing and data quality control, to the statistical test of hypothesis to detect interesting genes. Finally we will comment on multivariate methods to detect patterns in gene expression.

Key words: DNA microarray. Statistical analysis. Experimental design.

Introducción

El genoma de los seres vivos es el conjunto de genes que se encuentran distribuidos en cromosomas. Los genes, a su vez, son secuencias de ADN que contienen toda la información necesaria para sintetizar las proteínas, moléculas esenciales para la vida que realizan prácticamente todas las funciones celulares. Cuando un gen se «activa» para dar lugar a su proteína correspondiente, diremos que ese gen se está expresando en esa célula. Es conocido que anomalías en la expresión de los genes pueden llevar a disfunciones celulares, provocando graves enfermedades como el cáncer, entre muchas otras. Los genes que tengan su expresión alterada en un tejido tumoral respecto a un tejido sano del mismo órgano, por ejemplo, serán claros candidatos a tener alguna implicación en el proceso neoplásico. Por lo tanto, la identificación de los genes desregulados es un paso importante para conocer las bases moleculares de muchas enfermedades de carácter genético.

Desde mediados de los años noventa existe la técnica de los *microarrays* de ADN, que permite monitorizar simultáneamente el nivel de expresión de miles de genes en un conjunto de células. Sin embargo, la potencia que nos ofrece esta herramienta implica nuevos retos en lo que se refiere al análisis estadístico. Los datos que se generan con *microarrays*, aparte de tener un gran volumen, se caracterizan por ser altamente variables, por lo que serán básicos tanto el análisis estadístico como el diseño experimental que se plantee para solucionar las diferentes cuestiones biológicas que nos propongamos.

En este trabajo explicaremos primero con más detalle qué es un *microarray* y cómo funciona, para después tratar sobre cuáles son sus principales usos. Seguidamente hablaremos de los diferentes diseños experimentales que se pueden utilizar, y pasaremos a tratar las diversas partes que componen el análisis de un *microarray*, desde el procesamiento de la imagen y control de calidad de los datos hasta el tratamiento estadístico para identificar genes de interés. Finalmente, hablaremos sobre las diferentes técnicas de análisis multivariante que se pueden utilizar para extraer el máximo conocimiento de nuestros datos. La figura 1 muestra un esquema con los aspectos más relevantes de un protocolo de experimentos con *microarrays*.

¿Qué es un *microarray* de ADN y cómo funciona?

Los *microarrays* de ADN son una herramienta que permite realizar análisis genéticos diversos basados en la miniaturización de procesos biológicos. La primera aplicación de esta tecnología fue para medir simultáneamente el nivel de expresión de miles de genes[1]. Las mejoras tecnológicas han perfeccionado la calidad y han ampliado el espectro de aplicaciones, de manera que los *microarrays* se han consolidado como herramientas útiles en investigación genética con aplicaciones en medicina[2,3]. El funcionamiento de los *microarrays* de expresión se basa en la capacidad de las moléculas complementarias de ADN de hibridar entre sí. Pequeñas cantidades de ADN, correspondientes a diversos genes

Correspondencia: Dr. V. Moreno.
Unidad de Bioestadística y Bioinformática. Servicio de Epidemiología y Registro del Cáncer.
Instituto Catalán de Oncología. Hospital Duran i Reynals.
Gran Vía, km 2,7.
08907 L'Hospitalet de Llobregat. Barcelona. España.
Correo electrónico: v.moreno@iconcologia.catsalut.net