UNIVERSITAT DE BARCELONA

Anàlisi bioinformàtica de la base genética de la susceptibilitat al càncer de mama

Núria Bonifaci Cano

ANÀLISI BIOINFORMÀTICA DE LA BASE GENÈTICA DE LA SUSCEPTIBILITAT AL CÀNCER DE MAMA

Núria Bonifaci Cano

2014

ANÀLISI BIOINFORMÀTICA DE LA BASE GENÈTICA DE LA SUSCEPTIBILITAT AL CÀNCER DE MAMA

Memòria presentada per
Núria Bonifaci Cano
per optar al grau de
Doctor per la Universitat de Barcelona

Tesi realitzada sota la direcció del
Dr. **Miquel Àngel Pujana.**
a l'Institut Català d'Oncologia
de l'Institut d'Investigació Biomèdica de Bellvitge
(ICO-IDIBELL)

Tesi adscrita a la Facultat de Medicina, Universitat de Barcelona (UB).
Programa de doctorat en Biomedicina

Director:
Miquel Àngel Pujana

Tutor:
Víctor Moreno Aguado

Doctoranda:
Núria Bonifaci Cano

Agraïments

En primer lloc vull agrair al meu director, Miquel Àngel, la oportunitat de començar aquest projecte i així poder introduir-me en el món de la recerca. Gràcies per la teva dedicació, ajuda i suport. També li vull agrair tot el que m'ha ensenyat i consells que m'ha donat. Finalment li agraeixo molt la seva paciència i comprensió en la etapa final de redacció de la tesi que ha resultat ser molt més llarga del que estava previst.

Voldria agrair al Víctor que acceptés ser el meu tutor i que m'hagi deixat un espai a la UBS. També voldria agrair els valuosos comentaris estadístics que he rebut per part seva.

A la Conxi i a tothom del LRT2 per acollir-me durant el primer any deixant-me un raconet de la poiata per posar el meu ordinador.

Gràcies als membres del nostre grup, Helena, Gris, Laia, Jordi per l'ajuda i consells que he rebut per part seva. Jordi, gràcies per la teva ajuda amb el Python i el Latex. Laia, moltes gràcies pels teus consells, per compartir les teves experiències amb mi i per tot el suport que m'has donat.

Gràcies a tots els membres de la UBS, Xavi, David, Eli, Adriana, Toni, Rebeca, Ferran, Marta, Dani, Nuria per la col·laboració que hem mantingut. Toni, moltes gràcies per respondre totes les meves preguntes y dubtes estadístics que no han sigut pocs i per estar sempre disposat a ajudar. Rebeca, muchas gracias por tus consejos, comentarios y opiniones y sobre todo por los ánimos por WhatsApp en la última etapa de la tesis (això també va per tu Laia).

Finalment, un agraïment molt especial al Toni gràcies per la teva infinita paciència, pels teus consells, per estar sempre al meu costat, per aguantarme aquests últims mesos i per confiar sempre en mi. I, per descomptat, a la Jana i a la Gala, gràcies per la alegria i entusiasme que encomaneu i que fa que tot sigui més fàcil.

Als meus pares
Per vosaltres, Toni, Jana i Gala

Abreviatures i Acrònims

BRCA1	*Breast cancer 1*
BRCA2	*Breast cancer 2*
CGEMS	*Cancer genetic markers of susceptibility*
CIMBA	*Consortium of investigators of modifiers of BRCA1 and BRCA2*
CNV	*Copy number variation*
COGS	*Collaborative oncological gene-environment study*
CV-CD	*Common variant - common disease*
DNA	*Deoxyribonucleic acid*
EPHB1	*Ephrin type-B receptor 1*
eQTLs	*Expression quantitative trait loci*
ERα	*Estrogen eeceptor* α
ESR1	*Estrogen receptor 1*
FA	*Fanconi anemia*
FDR	*False discovery rate*
FRR	*Familial relative risk*
GxG	Interaccions genètiques
GPU	*Graphics processing unit*
GSEA	*Gene set enrichment analysis*
GWAS	*Genome-wide association studies*
GO	*Gene ontology*
HapMap	*Haplotype map*
HER2	Factor de creixement epidèrmic 2
HPRD	*Human protein reference database*
HR	*Hazard ratio*
IC	Intèrval de confiança
kb	Kilobases

KEGG	*Kyoto encyclopedia of genes and genomes*
LD	*Linkage disequilibrium*
MAF	Freqüència de l'al·lel menor
MCF10A	*Human breast epithelial breast line*
MI	*Mutual information*
meQTLs	*Methylation quantitative trait loci*
MORF4L1	*Mortality factor 4 like 1*
NGS	*Next-generation sequencing*
OR	*Odds ratio*
PB	*Procés biològic*
PCC	*Pearson correlation coeficient*
PR	Receptor de Progesterona
RE	*Relative enrichment*
RV-CD	*Rare variant - common disease*
SNP	*Single nucleotide polymorphism*
tagSNP	*Tag single nucleotide polymorphism*
TCGA	*The Cancer Genome Atlas*
WRT	*Wilcoxon rank test*

Continguts

Índex de figures

Índex de taules

1

Introducció

1.1 Etiologia del càncer de mama

El càncer de mama presenta una etiologia complexe en la que hi intervenen una barreja de factors genètics i ambientals. La predisposició genètica heretada a desenvolupar càncer de mama s'observa amb l'existència de famílies amb nombrosos casos. Estudis epidemiològics han demostrat que les dones amb un familiar de primer grau afectat de càncer de mama tenen el doble de risc de desenvolupar la malaltia en comparació amb les dones que no tenen un historial familiar [1, 2]. Paralel·lament, en estudis amb bessons s'ha demostrat que el risc relatiu a patir càncer de mama és més elevat en bessons monozigòtics que en bessons dizigòtics, suggerint que el component predominant en l'agregació familiar de casos és probablement causat per factors genètics més que pel fet de compartir un mateix ambient [3].

La base genètica del càncer de mama ha sigut molt estudiada. La identificació de nous gens de susceptibilitat és de gran interès ja que és una malaltia molt comuna arreu del món i representa un problema molt rellevant de salut pública tant en països desenvolupats com en vies de desenvolupament. Actualment, el càncer de mama és el càncer més freqüent entre les dones arreu del món (amb un risc al llarg de la vida de més del 10 %), essent la primera causa de mort relacionada amb càncer entre les dones [4]. L'estudi de les bases genètiques de la malaltia té principalment dos objectius. El primer seria poder estratificar la població en funció del risc a patir la malaltia i així aplicar una medicina preventiva més efectiva. En segon lloc, la obtenció de coneixement fonamental sobre els processos biològics i vies de senyalització involucrades en la etiologia, el que podria obrir noves oportunitats terapèutiques [5].

1.2 Variants genètiques i malaltia

El genoma de dos individus de la població general difereix en aproximadament el 0,1% de la seva seqüència i aquestes diferències són en part les responsables de les característiques individuals i la susceptibilitat a les malalties [6]. Aquesta variabilitat és deguda a mutacions en la línia germinal que es van donar en la història humana i que, com a conseqüència de diverses forces evolutives i/o poblacionals, apareixen amb una determinada freqüència en les poblacions actuals. En general, si una variant presenta una freqüència major o igual al 1% en la població general es classifica com a **polimorfisme o variant comuna**. Així, les variants genètiques que apareixen en menys de l'1% de la població es classifiquen com a **variants rares**. Llavors, el concepte de **mutació** es limita a variants rares que estan associades funcionalment a un fenotip simptomàtic d'una determinada malaltia. En aquest context, el projecte internacional dels 1.000 genomes [7] proporcionarà el catàleg complert de possibles variants en el genoma humà. Això serà molt útil per a poder caracteritzar les variants genètiques relacionades directament (i.e. mutacions) o indirectament (e.g. en desequilibri de lligament, veure següents seccions) amb les malalties.

1.2.1 Polimorfismes d'un sol nucleòtid

El tipus de polimorfisme més freqüent és el d'una variació en la seqüència de DNA que només afecta una base nucleotídica (*Single Nucleotide Polymorphism*, **SNP**). Es tracta de substitucions d'una base per una altra i, donat que és molt improbable que aquest fet hagi ocorregut més d'una vegada en

la mateixa posició (*locus*) de la seqüència de DNA durant la història de les poblacions humanes, en la majoria dels casos només presenten dues formes (anomenades al·lels): l'al·lel ancestral i l'al·lel mutat en algun moment de la història humana però a priori no associat a malaltia [8]. En una determinada població, cadascun dels dos al·lels d'un SNP té una freqüència que depèn de l'evolució de les poblacions i el possible efecte funcional del canvi genètic. Es denomina **MAF** (de l'anglès *Minor Allele Frequency*) la freqüència de l'al·lel menys comú.

Els SNPs es poden classificar en funció de la regió del genoma on estan situats i d'aquesta manera preveure la seva possible funcionalitat. Un **SNP serà funcional** si altera l'expressió correcte del gen o la funcionalitat del seu producte. D'aquesta manera, els SNPs situats en els exons o en les regions reguladores d'un gen poden associar-se o causar (i.e. mutacions) una malaltia concreta alterant l'estructura o la abundància d'una proteïna específica [9]. Donada la baixa densitat de gens en el genoma humà la majoria dels SNPs són intergènics [10]. Aquest fet però, no descarta que puguin tenir un efecte funcional ja que els SNPs intergènics poden localitzar-se en regions o seqüències reguladores [11].

Donat que en la majoria dels casos els SNPs representen canvis genètics únics, en general estan altament correlacionats amb altres variants que es troben a prop en la seqüència del genoma. D'aquesta manera és possible predir l'al·lel d'un SNP basant-nos en l'al·lel en una altra posició relativament propera en el genoma. Així, els al·lels de diferents SNPs d'una mateixa regió cromosòmica s'hereten de manera conjunta més sovint del que s'esperaria per atzar, el que s'anomena **desequilibri de lligament** (*Linkage Disequilibrium*, LD). Aquest desequilibri però, no està regularment distribuït,

sinó que es troba segmentat per regions amb alta freqüència de recombinació [12–14]. D'aquesta manera, el genoma es pot estructurar en blocs discrets (anomenats **blocs haplotítipcs**) dins dels quals la major part de la variació comuna està correlacionada, presenten elevat LD entre ells i baix LD amb els SNPs dels blocs haplotípics veïns [15]. En un grup d'SNPs amb elevat LD, existeix informació redundant per tant és possible seleccionar un SNP representatiu (**tagSNP**) i utilitzar-lo per inferir la resta d'SNPs. Seleccionant un nombre limitat d'SNPs, cadascun representatiu de la seva regió genòmica o d'elevat LD, es pot capturar la major part de la informació respecte a la variabilitat genètica en una població. Normalment la variabilitat en cada bloc és representada per només quatre o cinc combinacions úniques d'al·lels anomenades haplotips [16]. Des de l'any 2005, gràcies al projecte HapMap (International HapMap Project) [17], es disposa del catàleg dels haplotips del genoma humà i els tagSNPs que els representen per diferents poblacions humanes (Figura 1.1).

1.3 La base genètica del càncer de mama

Podem distingir tres tipus de factors genètics de predisposició o susceptibilitat al càncer (i en concret al càncer de mama, objecte d'aquest treball) en funció del risc relatiu a desenvolupar la malaltia que confereixen als portadors. Les **mutacions d'elevada penetrància**[1] són rares en la població general ($\leq 1/1000$) i estan associades a un increment de $>50\%$ del risc al llarg de la vida; les **mutacions de moderada penetrància** confereixen un increment de risc entre el 20-50% al llarg de la vida; i les **mutacions**

[1] La proporció d'individus amb la mutació que desenvolupen la malaltia associada.

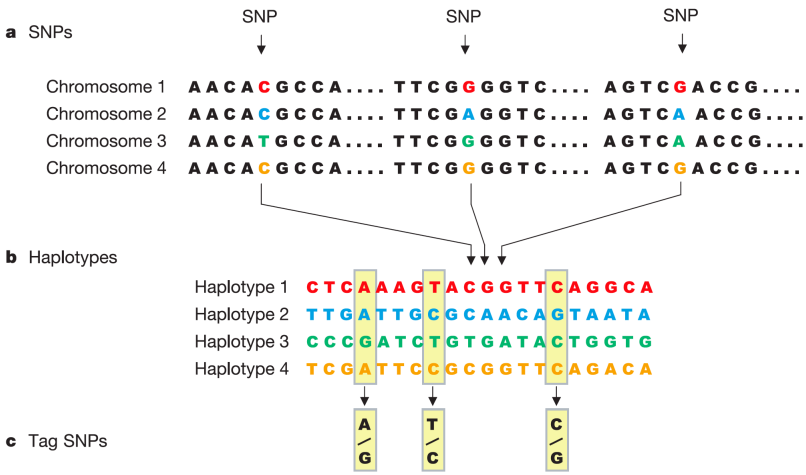


Figura 1.1 – a. Representació de quatre haplotips diferents del mateix fragment de DNA. La seqüència és idèntica excepte en les tres posicions que presenten variació (SNPs). b. Un haplotip està format per una combinació concreta d'al·lels d'SNPs propers. c. El coneixement de només certs SNPs (tagSNPs) és suficient per poder identificar un haplotip en concret (adaptat de *The HapMap Consortium*, 2003 [18]).

de baixa penetrància que poden ser comunes en la població (>5%) i confereixen un increment del risc entre 1%-20% al llarg de la vida [19] (Figura 1.2). **En aquesta tesi, i per simplificar, es parlarà de gens o *loci* d'alta, moderada o baixa penetrància en referència als gens o *loci* que es troben afectats per mutacions d'elevada, moderada o baixa penetrància respectivament**.

La **proporció del risc relatiu familiar** (*familial relative risk*, FRR) és una mesura útil de la contribució dels gens de susceptibilitat a la heretabilitat del càncer, és a dir, la proporció de la variança fenotípica d'aquesta malaltia que es pot atribuir al genotip [20]. Així, s'observa que els gens d'elevada i moderada penetrància identificats fins ara pel càncer de mama expliquen el

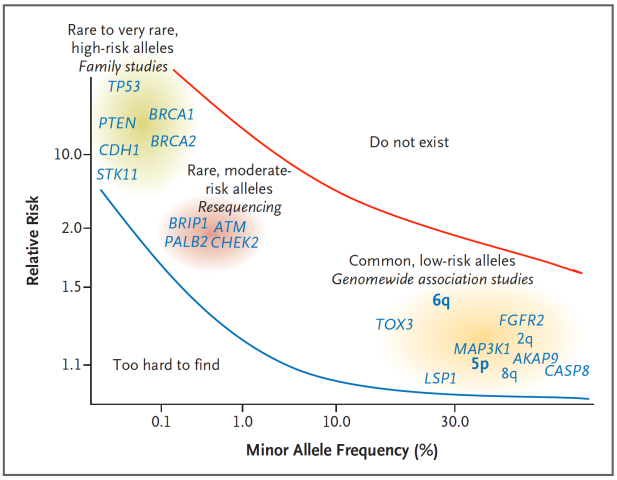


Figura 1.2 – Distribució dels gens de susceptibilitat al càncer de mama en funció del risc que confereixen i de la freqüència de les mutacions/variants en la població (extret de Foulkes, 2008 [19]).

21% del risc familiar (Figura 1.3) i el 15% seria explicat pels 76 gens de baixa penetrància coneguts fins el moment assumint que els riscos que confereixen aquests al·lels es combinen multiplicativament (són independents) [21].

Només una petita fracció del total de casos de càncer de mama són deguts a gens d'alta i moderada penetrància. Com ja s'ha explicat, aquests gens són molt poc freqüents en la població i confereixen un moderat/elevat risc a patir la malaltia al individu portador, per aquest motiu, donen lloc a concentracions familiars de múltiples casos [22]. En més del 90% dels casos però, la susceptibilitat a la malaltia vindria donada per l'efecte de diferents gens de menor penetrància, que actuarien en conjunt i interactuant amb factors ambientals, per conferir un increment en la predisposició a desenvolupar la malaltia [23]. Segons aquest model, els individus amb més risc

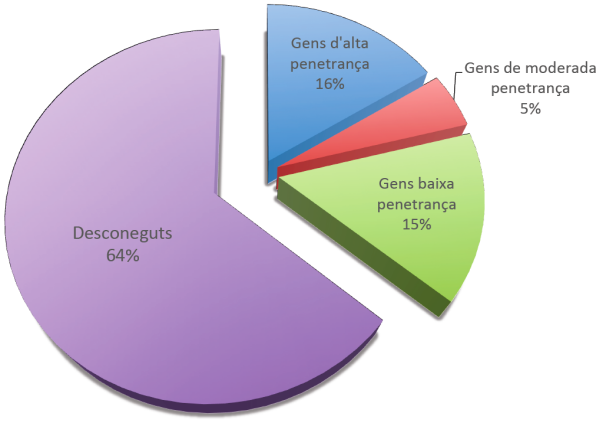


Figura 1.3 – Proporcions estimades del FRR a càncer de mama atribuïdes als diferents tipus de variants genètiques (adaptat de Michailidou et al., 2013 [21]).

de desenvolupar càncer de mama serien aquells portadors de més al·lels de risc (Figura 1.4). En general s'assumeix que els gens de baixa penetrància, no donen lloc a grans concentracions familiars de casos perquè donada la seva penetrància, la probabilitat de desenvolupar la malaltia és relativament baixa.

1.4 Gens de susceptibilitat al càncer de mama

1.4.1 Gens d'alta penetrància

Els principals gens d'alta penetrància donat l'elevat risc que confereixen als portadors de mutacions són *BRCA1* i *BRCA2* (Taula 1.1). Aquests gens

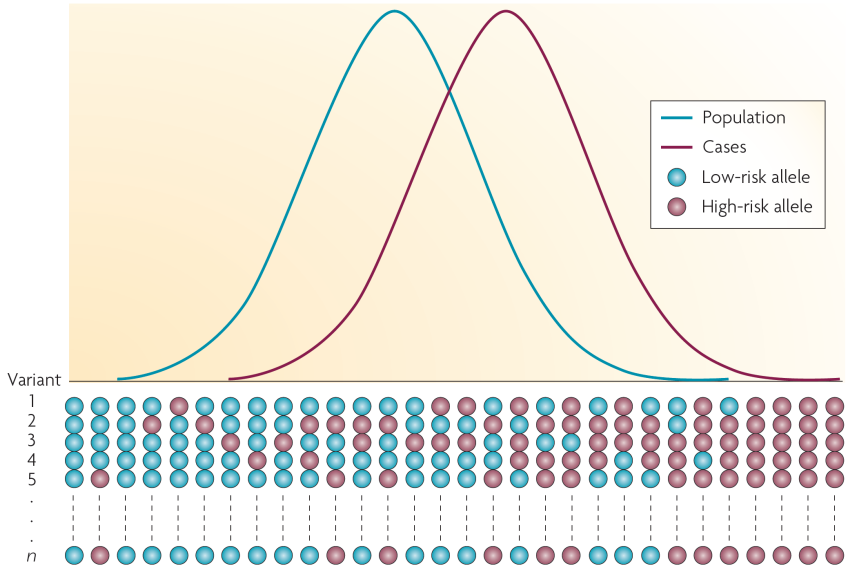


Figura 1.4 – La distribució dels al·lels de risc segueix una distribució normal en casos i controls. La distribució dels casos està desplaçada cap a un nombre major d'al·lels de risc. Els individus que presenten baix risc de desenvolupar càncer, portaran pocs al·lels de risc (en color vermell)(extret de Fletcher & Houlston, 2010 [24]).

van ser identificats a principis dels anys 90 mitjançant anàlisis de lligament genètic i clonació posicional en famílies amb alta agregació de casos [25, 26]. Així, *BRCA1* i *BRCA2* són supressors tumorals i les proteïnes per les que codifiquen es troben implicades en mecanismes de reparació del dany al DNA, entre altres funcions [27]. Tot i que aquests gens estan associats a un risc elevat a desenvolupar càncer de mama, les estimacions d'aquest risc abans dels 70 anys varien del 48% al 87% en portadors de mutacions en *BRCA1* i del 43% al 84% en portadors de mutacions en *BRCA2* [28, 29]. La magnitud del risc en anàlisis poblacionals és menor que la magnitud del risc en anàlisis basats en famílies amb múltiples individus afectats. A

més, el risc de càncer de mama en portadors de mutacions en *BRCA1* i *BRCA2* també varia en funció de la edat al diagnòstic i del tipus de càncer (unilateral vs contralateral) [28, 30, 31]. Aquestes observacions suggereixen l'existència de factors ambientals i genètics que modifiquen el risc a càncer de mama en els portadors de mutacions en gens d'alta penetrància [28, 30]. Altres gens d'alta penetrància i que s'associen a síndromes de càncer serien *TP53* [32], *PTEN* [33] i *STK11/LKB1* [34] (Taula 1.1). Tot i que s'ha continuat la cerca de gens de susceptibilitat, fins ara no s'han identificat més gens d'alta penetrància; això indicaria que o be no existeixen, o si existeixen les mutacions es donen en una freqüència molt baixa, penetrància intermèdia i/o amb models d'herència més complexos.

Taula 1.1 – Gens d'alta penetrància

Gen	Risc relatiu (%)	MAF	Síndrome clínic	Referència
BRCA1	>50	1/800	HBOS	[25, 35]
BRCA2	>50	1/500	HBCOS	[26, 36]
TP53	>50	<1/1000	Li-Fraumeni	[32]
PTEN	>50	<1/1000	Cowden	[33]
STK11/LKB1	>50	<1/1000	Peutz-Jeghers	[34]

1.4.2 Gens de moderada penetrància

Els gens de moderada penetrància (Taula 1.2) s'han identificat mitjançant la reseqüenciació de gens candidats en sèries de casos (habitualment amb agregació familiar) i controls. Tots aquests gens estan implicats en la resposta al dany al DNA, concretament en la reparació de la doble cadena

per recombinació homòloga. D'aquesta manera s'estableix una connexió entre la susceptibilitat al càncer de mama i la anèmia de Fanconi [37–40]. Donat que no es coneixen completament els mecanismes moleculars de la malaltia, és probable que existeixin altres gens d'aquestes característiques (i.e. de moderada penetrància) en relació al risc.

Taula 1.2 – Gens de moderada penetrància

Gen	Risc relatiu (%)	MAF (%)	Referència
ATM	20-50	<0,5	[41]
CHEk2	20-50	0,7	[42]
BRIP1	20-50	0,1	[43]
PALB2	20-50	<0,1	[44, 45]
RAD51C	20-50	0,1	[46]

1.4.3 Gens de baixa penetrància

Fins al moment, s'han identificat 76 variants de baixa penetrància (Taula 1.3) relacionades amb la predisposició al càncer de mama. Això ha sigut possible gràcies a l'anàlisi d'SNPs en estudis d'associació en gens candidats [47] o a nivell de tot el genoma (GWAS, de l'anglès *Genome-Wide Association Studies*) [48–59], i més recentment, a partir d'estudis de metanàlisis a gran escala realitzats pel consorci COGS (de l'anglès *Collaborative Oncological Gene-Environment Study*) [21, 60, 61]. En aquests estudis, en general, s'analitzen associacions entre els SNPs i el risc a càncer de mama. Si l'SNP identificat està situat en una regió codificadora del genoma, s'associarà al/s gen/s més proper. Llavors, però, encara falta identificar la

mutació concreta en el *locus* identificat i el mecanisme molecular alterat que influeix en el desenvolupament del càncer.

La recent disponibilitat del catàleg complert de les variants comunes realitzat pel projecte dels 1.000 genomes [7] serà clau per a la identificació de les variants causals o mutacions que hi han "lligades" a les variants de susceptibilitat identificades inicialment en els estudis d'associació; fins ara, només s'han descrit mutacions en quatre gens de baixa penetrància: *FGFR2* [62, 63], *CCND1* [64], *TOX3* [65] i *TERT* [66].

Taula 1.3 – Gens de baixa penetrància

Gen	Regió	Variant	$OR_{per\ al{\cdot}lel}$(95% IC)	MAF (%)	Referència
SNPs identificats en estudis de gens candidats					
CASP8	2q33	rs1045485	0,88(0,84-0,92)	13	[47]
TGFB1	19q13	rs1982073	1,08(1,04-1,11)	38	[47]
SNPs identificats en GWAS					
Intergènic	1p11	rs11249433	1,14(1,10-1,19)	39	[48]
Intergènic	2q35	rs13387042	1,20(1,14-1,26)	50	[49]
SCL4A7	3p24	rs4973768	1,11(1,08-1,13)	46	[50]
Intergènic	5p12	rs10941679	1,19(1,13-1,26)	24	[51]
TERT	5p15	rs10069690	1,18(1,13-1,25)	27	[52]
MAP3K1	5q11	rs889312	1,13(1,10-1,16)	28	[53]
Intergènic	6q14	rs17530068	1,12(1,08-1,16)	22	[54]
ESR1	6q25	rs3757318	1,30(1,17-1,46)	7	[55]
ESR1	6q25	rs2046210	1,29(1,21-1,37)	35	[56]
Intergènic	8q24	rs13281615	1,08(1,05-1,11)	40	[53]
CDKN2A/B	9p21	rs1011970	1,09(1,04-1,14)	17	[55]
Intergènic	9q31	rs865686	0,90(0,86-0,96)	39	[57]
ANKRD16	10p15	rs2380205	0,94(0,91-0,98)	43	[55]
ZNF365	10q21	rs10995190	0,86(0,82-0,91)	15	[55]

Continua en la pàgina següent.

ve de la pàgina anterior.

Gen	Regió	Variant	$OR_{per\ al·lel}$(95% IC)	MAF (%)	Referència
ZMIZ1	10q22	rs704010	1,07(1,03-1,11)	39	[55]
FGFR2	10q26	rs2981582	1,26(1,23-1,30)	38	[53]
LSP1	11p15	rs3817198	1,07(1,04-1,11)	30	[53]
CCND1	11q13	rs614367	1,15(1,10-1,20)	15	[55]
PTHLH	12p11	rs10771399	0,85(0,83-0,88)	12	[58]
Intergènic	12q24	rs1292011	0,92(0,91-0,94)	41	[58]
RAD51L1	14q24	rs999737	0,89(0,85-0,93)	24	[48]
TOX3	16q12	rs3803662	1,28(1,21-1,35)	27	[53]
COX11	17q22	rs6504950	0,95(0,92-0,97)	27	[50]
MERIT40	19p13	rs8170	0,99(0,93-1,05)	19	[59]
RALY	20q11	rs2284378	1,08(1,05-1,12)	35	[54]
NRIP1	21q21	rs2823093	0,94(0,92-0,96)	27	[58]
SNPs identificats en metanàlisi del COGS					
PTPN22/BCL2L15	1p13	rs11552449	1,07(1,04-1,09)	17	[21]
PEX14	1p36	rs616488	0,94(0,92-0,96)	33	[21]
MDM4	1q32	rs4245739	1,14(1,10-1,18)	26	[61]
LGR6	1q32	rs6678914	1,10(1,06-1,13)	59	[61]
Intergènic	2p24	rs12710696	1,10(1,06-1,13)	36	[61]
Intergènic	2q14	rs4849887	0,91(0,88-0,94)	10	[21]
METAP1D	2q31	rs2016394	0,95(0,93-0,97)	48	[21]
CDCA7	2q31	rs1550623	0,94(0,92-0,87)	16	[21]
DIRC3	2q35	rs16857609	1,08(1,06-1,10)	26	[21]
TGFBR2	3p24	rs12493607	1,06(1,03-1,08)	35	[21]
ITPR1/EGOT	3p26	rs6762644	1,07(1,04-1.09)	40	[21]
TET2	4q24	rs9790517	1.05(1.03-1.08)	23	[21]
ADAM29	4q34	rs6828523	0.90(0.87-0.92)	13	[21]
RAB3C	5q11	rs10472076	1.05(1.03-1.07)	38	[21]
PDE4D	5q11	rs1353747	0,92(0,89-0,95)	10	[21]
EBF1	5q33	rs1432679	1,07(1,05-1,09)	43	[21]
RANBP9	6p23	rs204247	1,05(1,03-1,07)	43	[21]
FOXQ1	6p25	rs11242675	0,94(0,92-0,96)	10	[21]
ARHGEF5/NOBOX	7q35	rs720475	0,94(0,92-0,96)	25	[21]
Intergènic	8p12	rs9693444	1,07(1,05-1,09)	32	[21]
Intergènic	8q21	rs6472903	0,91(0,89-0,93)	18	[21]
HNF4G	8q21	rs2943559	1,13(1,09-1,17)	7	[21]
MIR1208	8q24	rs11780156	1,07(1,04-1,10)	16	[21]
Intergènic	9q31	rs10759243	1,06(1,03-1,08)	39	[21]
MLLT10/DNAJC1	10p12	rs7072776	1,07(1,05-1,09)	29	[21]

Continua en la pàgina següent.

ve de la pàgina anterior.

Gen	Regió	Variant	$OR_{per\ al·lel}$(95% IC)	MAF (%)	Referència
DNAJC1	10p12	rs11814448	1,26(1,18-1,35)	2	[21]
TCF7L2	10q25	rs7904519	1,06(1,04-1,08)	46	[21]
Intergènic	10q26	rs11199914	0,95(0,93-0,97)	32	[21]
OVOL1/CFL1	11q13	rs3903072	0,95(0,93-0,96)	47	[21]
Intergènic	11q24	rs11820646	0,95(0,93-0,97)	41	[21]
Intergènic	12p13	rs12422552	1,05(1,03-1,07)	26	[21]
NTN4	12q22	rs17356907	0,91(0,89-0,93)	30	[21]
BRCA2	13q13	rs11571833	1,26(1,14-1,39)	10	[21]
PAX9/SLC25A21	14q13	rs2236007	0,93(0,91-0,95)	21	[21]
RAD51L1	14q24	rs2588809	1,08(1,05-1,11)	16	[21]
CCDC88C	14q32	rs941764	1,06(1,04-1,09)	34	[21]
MIR1972-2-FTO	16q12	rs17817449	0,93(0,91-0,95)	40	[21]
FTO	16q22	rs11075995	1,07(1,11-1,15)	24	[61]
CDYL2	16q23	rs13329835	1,08(1,05-1,10)	22	[21]
Intergènic	18q11	rs527616	0,95(0,93-0,97)	38	[21]
CHST9	18q11	rs1436904	0,96(0,94-0,98)	40	[21]
SSBP4/ISYNA1/ELL	19p13	rs4808801	0,93(0,91-0,95)	35	[21]
KCNN4/ZNF283	19p13	rs3760982	1,06(1,04-1,08)	46	[21]
EMID1/RHBDD3	22q12	rs132390	1,12(1,07-1,18)	4	[21]
MKL1	22q13	rs6001930	1,12(1,09-1,16)	11	[21]

Modificadors genètics de *BRCA1* i *BRCA2*

Existeixen gens de baixa penetràncIa identificats com a modificadors de risc de càncer de mama i/o ovari en portadors de mutacions en *BRCA1* i *BRCA2* (Taula 1.4). Com ja s'ha explicat anteriorment, els estudis epidemiològics mostren diferents penetràncies per als portadors de *BRCA1* i *BRCA2* en funció del tipus d'estudi realitzat [28–30] i aquests gens modificadors explicarien en part aquestes diferències. Les principals evidències de la existència de modificadors del risc han tingut lloc després de la creació, l'any 2005, del consorci d'investigadors de modificadors de *BRCA1* i *BRCA2* (*Consortium of Investigators of Modifiers of BRCA1 and BRCA2*, CIMBA) [67]. Aquest consorci, inclou més de 60 grups afiliats i ha recollit material genètic i dades clíniques de més de 20.000 dones portadores de mutacions en *BRCA1* o *BRCA2* per tal de tenir poder estadístic suficient per identificar

els gens modificadors [68]. Partint de les dades del CIMBA, els modificadors genètics s'han identificat de tres maneres: (i) mitjançant anàlisis d'SNPs en gens candidats [69, 70]; (ii) a partir d'estudis d'SNPs associats prèviament a càncer de mama en GWAS de població general [71–75]; (iii) a partir de GWAS de portadors de mutacions en *BRCA1* i *BRCA2* [59, 76]; i, finalment (iv) a partir dels anàlisis a gran escala realitzats pels membres del consorci COGS [77, 78].

Taula 1.4 – Modificadors genètics de *BRCA1* i *BRCA2*

			BRCA1		*BRCA2*		
Gen	Regió	Variant	num. portadors	HR*$_{\text{per al·lel}}$(95% IC)	num. portadors	HR$_{\text{per al·lel}}$(95% IC)	Referència
SNPs identificats en estudis de gens candidats							
CASP8	2q33	rs1045485	4844	0,85(0,75-0,97)	2509	1,06(0,88-1,27)	[69]
HMMR	5q34	rs299290	7584	1,09(1,02-1,16)	3965	1,04(0,94-1,16)	[70]
SNPs prèviament identificats en GWAS de població general							
Intergènic	1p11.2	rs11249433	10911	0,97(0,92-1,02)	6250	1,09(1,02-1,17)	[71]
Intergènic	2q35	rs13387042	9031	1,14(1,04-1,25)	5449	1,06(0,98-1,14)	[72]
SCL4A7	3p24	rs4973768	10283	1,03(0,98-1,08)	6153	1,10(1,03-1,18)	[73]
Intergènic	5p12	rs10941679	9691	0,96(0,9-1,02)	5854	1,09(1,01-1,19)	[73]
MAP3K1	5q11	rs889312	6741	0,99(0,93-1,06)	3524	1,12(1,02-1,24)	[74]
ESR1	6q25	rs2046210	10817	1,17(1,11-1,23)	6188	1,06(0,99-1,14)	[71]
CDKN2A/B	9p21	rs1011970	6374	1,03(0,96-1,09)	3807	1,09(1,00-1,18)	[75]
Intergènic	9q31	rs865686	6369	1(0,96-1,05)	3799	0,95(0,89-1,01)	[75]
FGFR2	10q26	rs2981582	6028	1,02(0,95-1,09)	3263	1,32(1,20-1,45)	[74]
LSP1	11p15	rs3817198	8984	1,05(0,99-1,11)	5434	1,16(1,07-1,25)	[72]
PTHLH	12p11	rs10771399	6368	0,87(0,81-0,94)	3798	0,93(0,84-1,04)	[75]
Intergènic	12q24	rs1292011	3755	1(0,94-1,06)	2530	0,94(0,87-1,01)	[75]
TOX3	16q12	rs3803662	6294	1,11(1,03-1,19)	3255	1,15(1,03-1,27)	[74]
SNPs identificats en GWAS de portadors de *BRCA1* o *BRCA2*							
ZNF365	10q21	rs16917302	-	-	4188	0,75(0,66-0,86)	[76]
MERIT40	19p13	rs8170	8363	1,26(1,17-1,35)	2448	0,9(0,77-1,05)	[59]
GMEB2	20q13	rs311499	-	-	4138	0,72(0,61-0,85)	[76]
SNPs identificats en anàlisis del COGS							
MDM4	1q32	rs2290854	14351	1,13(1,08-1,18)	-	-	[77]
Intergènic	1q32	rs6682208	14351	1,12(1,07-1,17)	-	-	[77]
CYP1B1	2p22	rs184577	-	-	8211	0,85(0,79-0,91)	[78]
Intergènic	6p24	rs9348512	-	-	8211	0,85(0,80-0,90)	[78]
TCF7L2	10q25	rs11196174	14351	1,13(1,07-1,18)	-	-	[77]
FGF13	Xq27	rs619373	-	-	8211	1,30(1,17-1,45)	[78]

* HR: Hazard Ratio a partir del model de regressió de Cox

Actualment, s'han identificat vuit variants de baixa penetrància associades al risc a càncer de mama en portadors de *BRCA1* i 14 variants associades al risc a càncer de mama en portadors de *BRCA2* (Taula 1.4). Només dues variants han presentat associació en els dos grups de portadors de mutacions. Aquestes diferències entre portadors de mutacions en *BRCA1* i *BRCA2* en les associacions, podria reflectir les característiques diferents i/o els processos moleculars alterats que es donen entre els tumors de portadors de *BRCA1* i els de portadors de *BRCA2* (veure apartat 1.7).

L'increment de risc que poden oferir aquests gens modificadors, tot i ser de baixa penetrància, pot ser important al combinar-lo amb el risc a càncer de mama que confereixen les mutacions en *BRCA1* i *BRCA2*. Per exemple, com es pot veure en la Figura 1.5, s'estima que el 5% dels portadors de mutacions en *BRCA2* que estan a menys risc (i.e. són homozigots per l'al·lel protector en la majoria dels 14 *loci* de susceptibilitat) tenen un risc entre el 21% - 47% de patir càncer de mama abans dels 80 anys, en comparació al 83% - 100% pel grup del 5% dels portadors de mutacions en *BRCA2* que tenen més risc (són homozigots per l'al·lel de risc en la majoria dels 14 *loci* de susceptibilitat) [78]. En canvi, a nivell de la població general, amb la combinació de tots els *loci* de susceptibilitat coneguts fins ara, s'estima que el 5% de les dones amb més risc només presenten un increment de 2,3 vegades respecte a la mitjana poblacional [21].

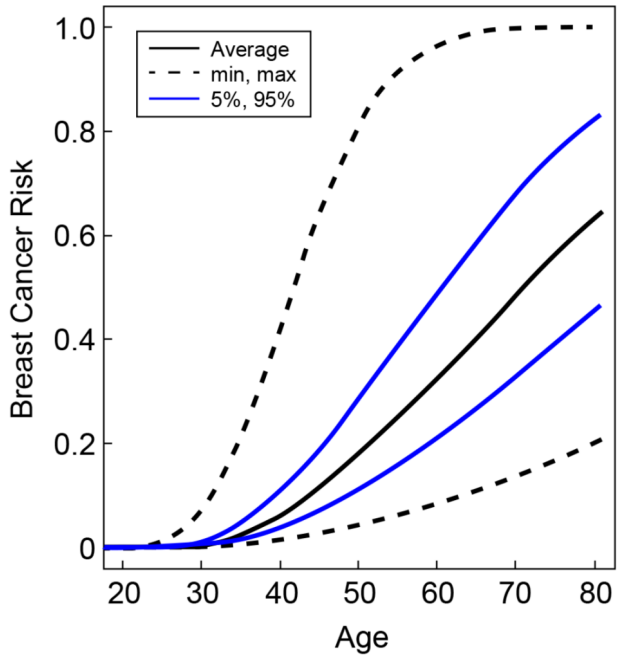


Figura 1.5 – Risc estimat a desenvolupar càncer de mama en els portadors de mutacions en *BRCA2* tenint en compte la combinació independent dels genotips dels SNPs en *FGFR2*, *TOX3*, 12p11, 5q11, *CDKN2A/B*, *LSP1*, 8q24, *ESR1*, *ZNF365*, 3p24, 12q24, 5p12, 11q13, 6p24. Es mostren els riscos en els percentils 5 i 95 de la distribució dels genotips així com els valors mínim, màxim i la mitjana del risc (adaptat de Gaudet et al., 2013 [78]).

1.5 Estudis d'associació genètica

Els estudis d'associació genètica avaluen l'existència de correlacions entre determinats trets, en aquest cas el càncer de mama, i variants genètiques. Els estudis d'associació en malalties complexes habitualment es realitzen a

nivell poblacional, en persones no emparentades, ja que així és més fàcil reclutar sèries mostrals grans. Aquests anàlisis poden presentar un disseny de tipus cas-control o cohort. En aquest últim tipus es selecciona un grup d'individus d'una població, es genotipen i es fa un seguiment durant un període de temps determinat per observar la incidència de la malaltia. Aquest seguiment és costós, i per això, l'estratègia més utilitzada és la primera. En els anàlisis cas-control les variants genètiques són genotipades en un nombre determinat d'individus afectats (casos) i no afectats (controls) per la malaltia. Com ja s'ha dit anteriorment, els polimorfismes genètics més estudiats són els SNPs, ja que, degut a que són molt nombrosos, proporcionen molta resolució i permeten analitzar exhaustivment el genoma humà. A més, el seu caràcter bial·lèlic i la fàcil detecció gràcies al gran desenvolupament de les tecnologies de genotipat dels darreres anys ha facilitat molt la seva detecció a gran escala. En aquests anàlisis es compara les freqüències d'al·lels d'un *locus* determinat entre els casos i els controls aparellats per diferents variables. L'associació existeix quan la distribució dels al·lels difereix entre casos i controls. Aquesta associació evidencia que el *locus* estudiat està relacionat amb la susceptibilitat a la malaltia. D'aquesta manera, una elevada freqüència d'una variant en els individus afectats per la malaltia s'interpreta com que la variant analitzada incrementa el risc de la malaltia en qüestió. I viceversa, la variant en qüestió serà protectora si la freqüència de l'al·lel és menor en els casos.

La mesura de l'efecte de l'al·lel estudiat en aquests estudis és la *odds ratio* (**OR**), definida com la *odds* dels portadors[2] en casos dividida per la dels controls. Un valor de OR de 1 significa que no hi ha associació entre l'al·lel i

[2] *odds* dels portadors: el nombre d'individus portadors de l'al·lel de risc dividit pel nombre d'individus no portadors.

la susceptibilitat a la malaltia. En canvi, diferències estadísticament significatives de la unitat indiquen associació amb el risc. Així, valors de OR > 1 impliquen un increment de risc en els portadors en comparació als no portadors i valors entre $0 <$ OR < 1 impliquen un efecte protector.

Els estudis d'associació tenen limitacions, la més destacada seria la aparició de falsos positius. Existeixen principalment dos aspectes que poden donar lloc a associacions falses: (i) l'estratificació poblacional deguda a un biaix alhora de seleccionar els participants a l'estudi; i (ii) un nivell de significació que no tingui en compte el nombre de testos i models independents realitzats en un estudi concret. Per això, és imprescindible confirmar els resultats en estudis independents [79, 80].

Existeixen principalment dues aproximacions a l'hora d'escollir els polimorfismes a estudiar: la primera es basa en l'estudi de marcadors amb un efecte funcional demostrat o probable (canvi d'activitat de la proteïna o d'expressió del gen, entre altres); en aquests casos el que es cerca és l'**associació directa** del marcador amb el tret estudiat. Això a la pràctica implica que es coneix la funció del gen i el possible paper en el fenotip estudiat. A més, implica que es coneixen variacions genètiques localitzades en aquest gen que produeixen canvis funcionals. Òbviament aquests tipus estudis estan limitats a la disposició prèvia d'aquesta informació.

En la segona aproximació, l'**associació indirecta**, s'explora la variabilitat genètica de la regió genòmica que es vol analitzar sense cap hipòtesi inicial respecte la funcionalitat dels polimorfismes. Normalment, per portar a terme aquests estudis, es genotipen només els tagSNPs reduint molt els costos de genotipat. Per tot això, en els estudis d'associació indirecta, els tagSNPs

genotipats no es considera *a priori* que influenciïn el risc de la malaltia. En lloc d'això, aquests SNPs estarien correlacionats amb l'al·lel causal degut a la presència del desequilibri de lligament, és a dir, servirien de marcadors d'un haplotip que conté la variant funcional.

1.5.1 Estudis d'associació en gens candidats

Aquests estudis, tant d'associació directa com indirecta, analitzen gens dels quals es té coneixement previ o es prediu la seva implicació en el mecanismes moleculars de la malaltia. Els primers gens candidats van ser els que participen en les vies de reparació al dany al DNA perquè la majoria dels gens d'alta i moderada penetrància coneguts pertanyen a aquestes vies (per exemple *BRCA1* [81] i *TP53* [82, 83]). Donada la importància de les hormones i de la història reproductiva en el desenvolupament del càncer de mama, altres gens candidats han sigut els implicats en la biosíntesi i metabolisme hormonal (per exemple *ESR* [84–86] o *CYP2D6* [87, 88]). Tot i el gran nombre d'anàlisis realitzats, la no replicació de les associacions significatives ha sigut una de les característiques dels primers estudis d'associació en gens candidats [89]. Una de les raons possibles per la no replicació de resultats podria ser la manca de poder estadístic d'aquests estudis. Típicament, en els primers estudis en gens candidats, es genotipaven centenars d'individus, el que significava que no hi havia suficient poder estadístic per detectar associacions amb variants de baixa penetrància (tenint en compte la seva contribució al risc que ara coneixem) [80]. No ha sigut fins a la creació de grans consorcis que s'ha obtingut el poder estadístic suficient per validar els resultats amb robustesa en diferents poblacions; així, els primers *loci* identificats van ser *CASP8* i *TGFB1* [47].

1.5.2 *Genome-wide association studies*

Aquests anàlisis exploren el màxim de variació genètica del genoma per identificar les variants associades amb la malaltia, sense establir una presumpció prèvia de la seva localització o funció. Gràcies a la utilització dels tagSNPs això es pot portar a terme amb un gran estalvi en els costos del genotipat. S'estima que en poblacions no africanes, la genotipació de 500.000 tagSNPs és suficient per donar cobertura i identificar la gran majoria dels SNPs amb una MAF major o igual al 5% [17, 90]. També s'estima que existeixen set milions d'SNPs amb una MAF $> 5\%$ en el genoma humà, els quals poden aparèixer en diferents poblacions encara que amb freqüències al·lèliques molt variables [91].

Es requereix un gran nombre de mostres per obtenir un poder estadístic adequat per detectar variants genètiques amb un efecte baix, aproximadament $0{,}70 < \text{OR} < 1{,}50$. Una de les pràctiques més habituals per reduir els costos que això suposa és el disseny en diferents fases. En la primera fase, anomenada de descobriment, es genotipen una proporció relativament petita dels casos i controls (Figura 1.6). En les subsegüents fases, es genotipen mostres addicionals però només per aquells marcadors que han mostrat associacions "significatives" en la primera fase. Degut al gran nombre d'SNPs que es testen en aquests anàlisis, es necessiten uns criteris estadístics molt estrictes per limitar el nombre de falsos positius. En general, es considera un límit de significació de $\text{P}_{\text{valor}} = 10^{-7}$ per GWAS de poblacions no africanes [93]. Com ja s'ha mencionat anteriorment, un cop s'obté una associació significativa, cal identificar la variant funcional concreta o mutació que causa la malaltia. Generalment això significava reseqüenciar la regió genòmica que rodeja les possibles variants de risc i determinar la funció de la variant, és

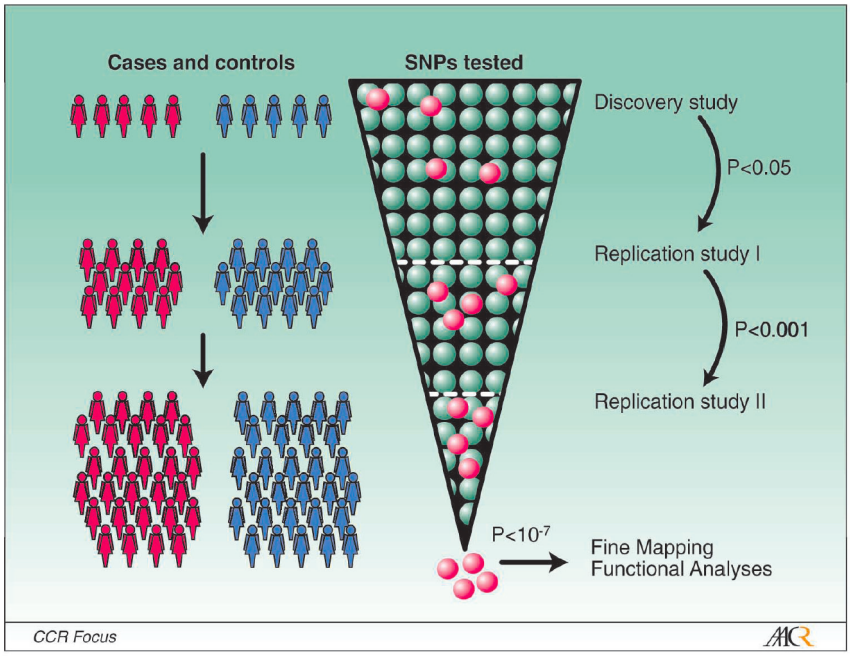


Figura 1.6 – Diferents fases d'un GWAS (extret de Garcia-closas & Chanock, 2008 [92]).

a dir, com afecta a l'estructura i/o funció dels gens (i el seu producte final, les proteïnes). Actualment es poden genotipar i/o imputar totes les variants del *locus* en qüestió gràcies a les dades del projecte dels 1.000 genomes [7] i avaluar els efectes funcionals de les variants candidates més probables com ja s'ha fet amb *CCND1* [64] i *TERT* [66].

1.6 Mutacions somàtiques

Fins ara s'ha descrit l'estudi de les variants genètiques que es troben en la línia germinal i que poden predisposar a desenvolupar càncer de mama. No obstant, el càncer és una malaltia genètica causada per l'acumulació de canvis en el genoma de les cèl·lules canceroses [94]. Les **mutacions somàtiques** són aquelles que tenen lloc en el genoma de certes cèl·lules de l'individu; és a dir, no s'hereten dels progenitors. Poden ser degudes tant a errors durant la replicació com induïdes per l'exposició a carcinògens, tant interns com externs. La majoria de les mutacions somàtiques que es van acumulant al llarg de la vida probablement no tenen conseqüències funcionals, però, en alguns casos podrien desregular mecanismes moleculars i així generar el creixement descontrolat de les cèl·lules de l'epiteli mamari (en el cas de càncer de mama) i contribuir a la tumorigènesi. D'aquesta manera, les mutacions somàtiques que es troben en el genoma d'una cèl·lula cancerosa es poden classificar segons les seves conseqüències en el desenvolupament del càncer. Les **mutacions conductores** (*driver mutations*) [94] són les que contribueixen a la progressió tumoral i són seleccionades positivament durant aquest procés perquè proporcionen a les cèl·lules un avantatge en el creixement, proliferació i/o supervivència (Figura 1.7). Els gens que adquireixen aquestes mutacions es defineixen com a "gens de càncer" i, fins al moment, se n'han identificat més de 500 [95] (http://www.sanger.ac.uk/genetics/CGP/Census/), essent el domini proteic quinasa el més freqüentment codificat per aquests gens. En canvi les **mutacions passatgeres** (*passenger mutations*) [94] no

contribueixen al desenvolupament del càncer si no que van ser adquirides abans que la cèl·lula cancerosa adquirís aquest fenotip o durant la expansió tumoral i són retingudes per atzar durant els cicles de divisió cel·lular i expansió tumoral.

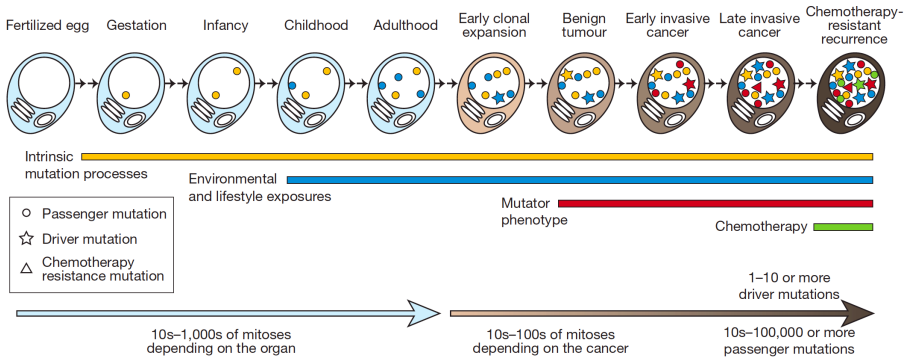


Figura 1.7 – Les mutacions poden ser adquirides en un llinatge de cèl·lules fenotípicament normals. Aquestes mutacions intrínseques poden ser adquirides tan durant la divisió cel·lular com a conseqüència de agents mutàgens exògens. Durant el desenvolupament del càncer altres processos com per exemple els defectes en la reparació del dany al DNA poden contribuir al conjunt de les mutacions. Les *passenger mutations* no tenen cap efecte en la cèl·lula cancerosa, però les *driver mutations* provocaran la expansió clonal del càncer (adaptat de Stratton et al., 2009 [94]).

1.7 Característiques moleculars i histopatològiques del càncer de mama

El càncer de mama és una malaltia heterogènia ja que presenta diversitat molecular i patològica. A nivell patològic, existeixen característiques com la morfologia, el grau i el patró d'expressió dels receptors d'hormona com el

d'estrogen (ER), el de progesterona (PR) i el del factor de creixement epidermic 2 (HER2) que estratifiquen els tumors/casos en grups diferenciats tant biològicament com clínicament al influir en la progressió i la prognosi de la malaltia [96–98]. Per aquest motiu, s'utilitzen aquestes característiques per determinar el tractament més efectiu [99–101]. A nivell molecular, el càncer de mama també presenta molta variabilitat i, en conseqüència, s'han definit com a mínim cinc subtipus principals de tumors a partir dels seus patrons d'expressió gènica (basal, HER-2, luminal A, luminal B i *normal-like*) [102, 103]. Aquests subtipus també estan associats a diferències de pronòstic i de resposta al tractament [104, 105]. Els estudis epidemiològics han demostrat que els factors de risc a càncer de mama, tant els genètics com els ambientals, varien en funció del subtipus de tumor [106–108]. Per exemple, factors de risc ambientals o d'estil de vida, com l'edat de la menàrquia, nul·liparitat, edat al primer fill i obesitat postmenopausia estan associats a tumors que expressen el receptor d'estrogen α (ERα-positius) [108–110]. També els gens de susceptibilitat a càncer de mama presenten diferencies en l'associació en funció de l'expressió dels receptors hormonals. El 90% dels tumors de portadors de mutacions en *BRCA1* no expressen el ER α (ERα-negatius) [111, 112]; en canvi, els tumors de portadors de mutacions en *BRCA2* tendeixen a ser ERα-positius [111]. De manera similar, la majoria dels gens de baixa penetrància identificats fins al moment presenten una associació més forta o exclusiva pels casos amb tumors ERα-positius [21, 92]; amb les excepcions dels *loci* *ESR1* [56], *FTO* [61], *LGR6* [61], *MDM4* [61], *MERIT40* [59], *PEX14* [21], *RALY* [54], *TERT* [52], 2p24 [61] i 6q14 [54] que estan associats a tumors ERα-negatius. Tres *loci* més (*PTHLH*, *TOX3* i *ZNF365* estan associats als dos tipus de tumors [51, 53, 92]). A més, el patró d'associació que presenten els gens de baixa penetrància definit per l'estat de l'ERα en la població general és consistent amb el patró d'associació d'a-

quests gens en els portadors de mutacions en *BRCA1* i *BRCA2* (explicat en l'apartat 1.4.3). És a dir, els gens (i.e. les seves variants) que presenten una associació més forta en casos de la població general que han desenvolupat tumors ERα-positius tendeixen a estar associats al risc en portadors de mutacions en *BRCA2* i, pel contrari, els gens més associats al risc en tumors ERα-negatius tendeixen a presentar una associació similar en els portadors de mutacions en *BRCA1* (Figura 1.8).

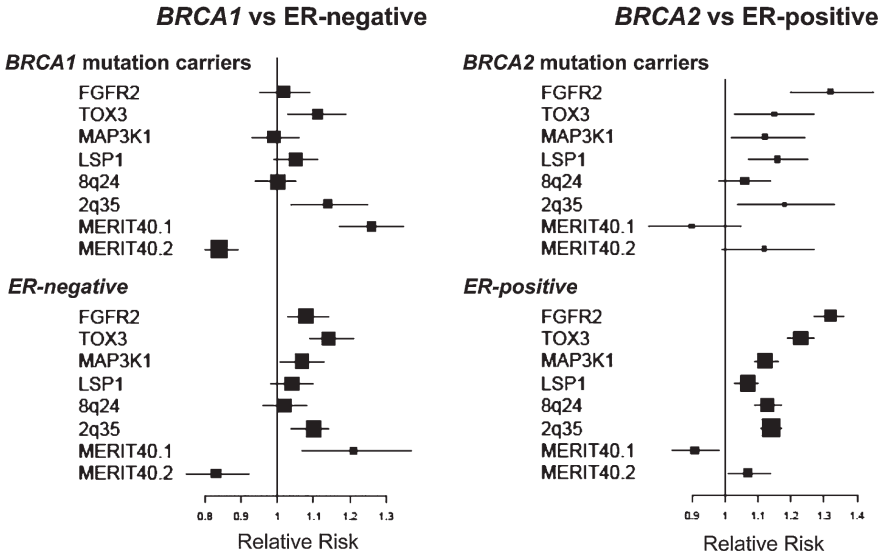


Figura 1.8 – HRs de vuit SNPs en portadors de *BRCA1* i *BRCA2* i les OR de tumors ERα-positius i ERα-negatius en la població general. Els patrons d'associació en portadors de mutacions en *BRCA1* són similars als dels tumors ERα-negatius de la mateixa manera que els patrons d'associació de portadors de mutacions en *BRCA2* s'assemblen als dels tumors ERα-positius (extret de Milne & Antoniou, 2011 [113]).

En conjunt, les diferencies observades en les associacions genètiques en funció del subtipus de tumor suggereixen que els tumors ERα-positius i ERα-negatius s'originen a partir de vies etiològiques parcialment diferents [108].

1.8 El concepte de *Missing heritability*

Només una part del risc a desenvolupar càncer de mama es pot explicar amb els gens identificats fins a l'actualitat (Figura 1.3). Gairebé el 60% del component genètic de la malaltia continua sense identificar i correspon a l'anomenada ***missing heritability*** [114].Tradicionalment hi han hagut dues hipòtesis per explicar la base genètica de les malalties complexes. Aquestes hipòtesis es distingeixen per la freqüència de les mutacions que predisposarien a patir la malaltia en la població. Alguns autors defensen la hipòtesi de la **"Variant Comuna - Malaltia Comuna"** (*Common Variant-Common Disease*, CV-CD) segons la qual les variants al·lèliques comunes, amb una freqüència superior al 1-5% però de baixa penetrància, són les principals contribuïdores a la susceptibilitat a patir malalties comunes [115]. En contraposició, la hipòtesi de la **"Variant Rara - Malaltia Comuna"** (*Rare Variant-Common Disease*, RV-CD) defensa que múltiples variacions rares ($< 1\%$), cadascuna amb una penetrància moderada, són les que contribueixen principalment a la susceptibilitat a patir malalties comunes [116]. Aquestes últimes mutacions, degut al seu efecte deleteri i la conseqüent purificació genètica, serien variants relativament especifiques de cada població.

Com s'ha vist, la susceptibilitat genètica al càncer de mama està definida per un conjunt de mutacions/gens amb diferents nivells de risc i prevalen-

ça en la població que aportarien evidències a les dues hipòtesis. Fins ara, les variants de susceptibilitat rares s'han identificat mitjançant la reseqüenciació de gens candidats. És possible que existeixin altres variants rares de moderada/alta penetrància donat que no es coneixen completament els mecanismes moleculars de la malaltia. Les noves tècniques de seqüenciació massiva, sobretot en casos joves [21], suposarà una nova estratègia per identificar noves variants d'aquestes característiques.

És també possible que existeix una classe intermèdia de variants de risc amb una MAF entre 1-5% i moderada penetrància. Així, no és probable que tinguin efectes suficientment elevats per haver-se detectat per anàlisis de lligament, ni freqüències suficientment elevades per identificar-se a partir de GWAS. En part, per tal de facilitar la detecció d'aquest tipus de variants, es va crear l'any 2.008 el projecte dels 1.000 genomes [7]. L'objectiu és la generació del catàleg complert de les variants amb una MAF $\geq$ 1%; amb aquest catàleg es podrà realitzar una nova generació de GWAS dirigits a aquest tipus de variabilitat.

Els GWAS han sigut els estudis que més variants de susceptibilitat comunes han identificat, però donat que encara queden gens per identificar (i.e. *missing heritability*), s'han proposat diferents estratègies complementaries [22, 117]. L' estratificació dels GWAS per subtipus de tumors podria ajudar en la identificació i caracterització de nous factors de risc [118]; exemples en aquest sentit són les identificacions de variants en el cromosoma 19p13 a través d'un GWAS en dones portadores de *BRCA1* [59] i la identificació de quatre *loci* de susceptibilitat en casos amb tumors de mama ERα-negatius (*FTO*, *LGR6*, *MDM4* i 2p24) [61]. L'estudi en diferents ètnies també pot conduir a la identificació de més *loci* de susceptibilitat, com

per exemple el *locus* 6q25.1 identificat en un GWAS de poblacions asiàtiques [119]. El més probable però, és que ja s'hagin identificat les variants comunes amb major efecte i que només quedin per identificar els al·lels amb menor efecte (OR $<$ 1,1). Per tal de detectar aquests ORs propers a 1 calen GWAS/metanàlisis a gran escala, com per exemple l'estudi realitzat per artritis reumatoide on el nombre d'individus inclosos en el estudi va ser de més de 100.000 entre casos i controls [120]. Cal destacar el metanàlisi més gran fet fins a dia d'avui per càncer de mama on es va detectar un important excés d'SNPs amb associacions significatives (a nivell nominal) sense arribar a l'estricte llindar de significació del GWAS [21]. En aquest estudi, tenint en compte "l'excés de senyals", 9.851 SNPs amb OR $<$ 1,02, junt amb les variants de baixa penetrància conegudes fins ara, explicarien un 28% de la heretabilitat (el que correspon al doble de la heretabilitat explicada per aquest grup de variants actualment). Això podria significar que realment els trets complexes estiguessin afectats per milers de variants d'efecte molt petit, consistent amb el **model infinitesimal** [121]. Segons això, serien centenars o milers de variants comunes les que explicarien el risc a patir càncer en la població. D'aquesta manera, la *missing heritability* es trobaria "amagada" pels estrictes llindars de significació utilitzats en els GWAS. A part de tot això, la *missing heritability* també podria ser deguda a estimacions errònies de la heretabilitat, a la epístasi, a la epigenètica o a variació genòmica estructural com les variants en el nombre de còpia (*copy number variants*, CNVs) [122].

1.9 Biologia de sistemes

Els GWAS inclouen anàlisis estadístics on, majoritàriament, els SNPs/gens s'interpreten individualment en relació a la malaltia. D'aquesta manera, no es té en compte aquells SNPs/gens identificats per sota d'un estricte llindar de significació. Però les malalties complexes no s'originen a partir d'alteracions en gens/proteïnes de forma individual sinó que probablement tenen el seu origen en alteracions coordinades i/o relacionades funcionalment de diversos gens/proteïnes (a més dels factors ambientals). Així, és ben conegut, que els gens i les proteïnes no realitzen les seves funcions de manera aïllada dins de les cèl·lules sinó que interaccionen uns amb els altres formant xarxes complexes. Per aquest motiu, per entendre els mecanismes biològics cal una aproximació global, a nivell de tot el "sistema". Per aquest motiu, per entendre la complexitat de les malalties comunes, podriem dir que cal una aproximació des de la **"biologia de sistemes"**, que vagi més enllà de l'anàlisi dels components individuals i confereixi una perspectiva global i integradora. Des d'aquest punt de vista, els sistemes biològics són xarxes de macromolècules interconnectades (i.e **interactoma**) [123] en les que la majoria de gens i els seus productes realitzen les seves funcions. Per exemple, la xarxa d'interaccions proteïna - proteïna en humans on els nodes representen les proteïnes i les arestes, sense direcció, les interaccions físiques entre elles [124, 125], la **distància** entre dues proteïnes de la xarxa es defineix com el nombre mínim d'arestes que s'han de seguir per tal de connectar les dues proteïnes. Aquesta informació es pot trobar en la base de dades HPRD (de l'anglès **Human Protein Reference Database** [126]). Aquesta xarxa, igual que totes les xarxes biològiques presenta unes propietats determinades. Per exemple, presenten una **escala lliure** [127], la principal

conseqüència d'aquest fet és que la majoria de les proteïnes tenen un baix nombre de connexions mentre que poques proteïnes, anomenades ***hubs*** estan altament connectades [128–130]. Altres proteïnes presenten un elevat coeficient de centralitat o ***betweenness***, és a dir, forma part de del grup de camins més curts que connecta tots els parells de nodes de la xarxa [131]. Altres propietats de les xarxes d'interacció proteïna-proteïna és el ***small-world*** que vol dir que entre dos parells de nodes qualsevol existeix una distància curta [132] i l' arquitectura modular. Els mòduls topològics representen regions de la xarxa altament interconnectades. Assumint que si una proteïna està involucrada en un determinat procés biològic, els seus interactors més directes probablement participaran en el mateix procés (**mòdul funcional**) [133–135].

La biologia de sistemes ha sigut possible gràcies al ràpid progrés en biologia molecular impulsat per l'aparició de les tecnologies d'alt rendiment (*high throughput*) que permeten obtenir grans quantitats de dades; per exemple, la seqüenciació del genoma humà (genòmica), la proteòmica, els perfils d'expressió gènica, les interaccions proteïna-proteïna, etc. Però tot aquest allau de dades biològiques seria impossible d'integrar sense el desenvolupament en paral·lel de la **Bioinformàtica**, la disciplina que permet l'emmagatzematge, gestió i anàlisi de diverses dades biològiques.

Tots aquests recursos mencionats permeten un anàlisi de les dades dels GWAS que possiblement pot reflectir millor l'arquitectura de les malalties complexes. A continuació s'expliquen dues aproximacions.

1.9.1 Processos biològics i vies de senyalització

Els gens/proteïnes estan organitzats en mòduls funcionals (grups d'elements altament interconnectats) en el si de les xarxes complexes per a realitzar la seva funció biològica [128]. Aquestes són les estructures que s'alteren o desregulen en les malalties, de manera que alteracions en diferents gens/proteïnes de la mateixa unitat funcional sovint produeixen el mateix fenotip o malaltia [136]. Per això, la variabilitat genètica que confereix risc a malalties comunes és probable que resulti de l'acumulació dels efectes de variants genètiques dins de processos o funcions específiques [137].

L'objectiu de l'anàlisi de processos biològics és reconèixer si els grups de gens/proteïnes que es coordinen per assolir una tasca específica (procés biològic, PB), estan associats amb el tret d'interès, en aquest cas, el risc a desenvolupar càncer de mama a partir de les dades dels GWAS. Existeix un ampli ventall de grups de gens/proteïnes entre els que es pot destacar els referents a funcions cel·lulars, processos metabòlics, biosíntesi, processament d'informació genètica com reparació del DNA, senyalització cel·lular, resposta immune, característiques del desenvolupament embrionari i factors que condueixen a les malalties humanes, i que es poden trobar en diferents bases de dades com per exemple *Kyoto Encyclopedia of Genes and Genomes* (KEGG) [138], *Biocarta* [139] o *Gene Ontology* (GO) [140].

Els anàlisis de dades de GWAS basats en les anotacions funcionals descrites anteriorment parteixen d'una llista de gens ordenada (rànquing) en relació a l'associació amb el risc a càncer de mama (del gen més associat al menys). Habitualment el gen pren el valor de l'SNP més associat entre tots els que pertanyen al seu *locus*. Aquesta opció però no és òptima ja que provoca

un cert biaix [141] (descrit en següents seccions) i, així, existeixen diversos mètodes per intentar corregir-ho [142–144]. Posteriorment s'observa com estan distribuïts els gens d'un determinat PB en aquest rànquing. Si estan distribuïts uniformement, probablement el PB en qüestió no està relacionat amb la susceptibilitat al càncer de mama. Contrariament, una acumulació dels gens del PB en la part superior del rànquing indicaria una possible relació entre aquest PB i la susceptibilitat a la malaltia. Diferents algoritmes permeten fer aquests anàlisis [145, 146], entre els més coneguts podem trobar el GSEA [147] i el Fatiscan [148].

La significació de l'associació de cada procés es resumeix en funció de les associacions amb la malaltia dels SNPs assignats a cada gen que composa el procés concret. D'aquesta manera, s'incrementa el poder per detectar efectes subtils de diferents SNPs en el mateix grup de gens que es podrien perdre quan s'analitzen individualment [149]. A més, donat que es poden combinar nombrosos gens en un nombre limitat de grups de gens per a l'anàlisi, les comparacions múltiples es redueixen molt. Per una altra banda, els resultats d'aquests anàlisis també donen informació dels processos biològics (i vies de senyalització, no detallades) involucrats en la malaltia.

1.9.2 Interaccions genètiques

És raonable esperar que alguns dels efectes de les variants genètiques que contribueixen a la expressió d'una malaltia complexe interaccionin entre ells [150, 151]. D'aquesta manera, la presència de dues o més variants en concret podria incrementar el risc a patir la malaltia més del que s'esperaria a partir dels seus efectes per separat. Podríem dir, de manera general,

que una interacció (també anomenada epístasi) té lloc quan l'efecte d'una variant és modulat per una altra variant en un altre *locus* [152].

Com ja s'ha explicat anteriorment, les interaccions genètiques s'han suggerit com una de les explicacions per a la *missing heritability* [122]. Així, la seva identificació podria millorar l'exactitud dels models de risc existents i millorar la prevenció del càncer [153, 154].

En organismes model, anàlisis a gran escala han demostrat l'existència d'interaccions genètiques en la explicació de la majoria de fenotips "comuns" [155]. Per exemple, el genoma del llevat *Saccharomyces cerevisiae* conté ∼6000 gens i s'ha estimat l'existència de ∼200.000 interaccions genètiques [156]. En humans però, aquestes metodologies no són aplicables. Just ara es comencen a descriure els primers mètodes en cèl·lules de mamífers [157, 158]. A nivell estadístic es poden analitzar les interaccions epistàtiques de les malalties complexes humanes en els estudis GWAS. En aquest context, la epístasi es refereix específicament a la desviació de l'additivitat d'un fenotip quantitatiu per l'efecte d'una variant genètica o mutació en un *locus* diferent [155].

Les aproximacions basades en la regressió logística són molt utilitzades per calcular les interaccions en dades de GWAS, ja que el model i els paràmetres són fàcilment interpretables; no obstant, existeixen limitacions i s'han desenvolupat molts altres mètodes [159–162].

Cercar exhaustivament les interaccions de totes les parelles d'SNPs d'un GWAS significa realitzar centenars de bilions de testos i això suposa una limitació pel temps de computació que requereix. Actualment però, l'ús

d'unitats de processament gràfic (GPUs, de l'anglès *Graphics Processing Unit*) ho redueixen a un temps raonable [163, 164]. A més de les limitacions computacionals, l'anàlisi d'interaccions a nivell GWAS també presenta serioses limitacions estadístiques. El gran nombre de parelles d'SNPs avaluades implica llindars de significació molt més estrictes per tal d'evitar falsos positius.

Alguns mètodes per reduir el temps de computació i sobretot, per limitar el nombre d'hipòtesis, restringeixen l'anàlisi d'interaccions genètiques a un grup d'SNPs del GWAS. Aquests SNPs candidats poden ser seleccionats simplement en base als seus efectes marginals [165] o en base a diversos algoritmes de filtració de la informació [166–168]. Altres mètodes usen el coneixement biològic per filtrar les dades, d'aquesta manera només es calculen les interaccions entre els marcadors que esperem *a priori* que interactuïn en base al coneixement biològic [169, 170].

Finalment cal destacar que connexió entre els resultats estadístics de les interaccions genètiques a partir dels GWAS i l'efecte biològic no està clara [171]. Existeixen molts reptes al intentar fer inferències sobre la biologia a nivell cel·lular a partir d'un model estadístic que està resumint dades a nivell poblacional [172]. L'epístasi biològica és el resultat d'interaccions físiques entre molècules dins de la xarxa de regulació gènica, processos biològics i/o vies de senyalització, de manera que l'efecte d'un gen/proteïna en un fenotip depèn d'un o més gens/proteïnes. No obstant, desconeixem en gran mesura la rellevància funcional de la majoria d'interaccions moleculars i, encara més, desconeixem gran part de les interaccions moleculars.

El conjunt d'aquestes observacions il·lustra la dificultat d'anàlisi i interpretació de possibles associacions genètiques amb malalties i en concret amb el càncer de mama.

2

Hipòtesi i Objectius

Hipòtesi

Donades les observacions principals descrites anteriorment, que de forma resumida estableixen que:

1. Tot i el gran avenç experimentat en els últims anys respecte al coneixement de la base genètica del risc a càncer de mama, no s'han identificat encara tots els gens implicats en el risc familiar ni poblacional.
2. Es desconeixen les característiques dels gens de susceptibilitat de baixa penetrància així com els processos biològics i vies de senyalització alterats en la malaltia.
3. Les noves tecnologies de gran rendiment han generat una gran quantitat de dades genòmiques i proteòmiques que integrades sota la perspectiva de la biologia de sistemes, permeten anàlisis més globals de la complexitat biològica de la susceptibilitat al càncer.

formulem la següent hipòtesi: **L'anàlisi de grans quantitats de dades biològiques sota la visió de la biologia de sistemes proporciona una aproximació per identificar nous gens de susceptibilitat al càncer de mama, les seves interaccions i els processos biològics en els que participen.**

Objectius

D'acord amb la hipòtesi formulada, els objectius principals de la tesi són els següents:

1. **Identificació dels processos biològics associats al risc de càncer de mama.**
2. **Identificació de candidats a gens de susceptibilitat a càncer de mama de baixa penetrància així com les seves característiques i xarxes moleculars en les que participen, mitjançant l'anàlisi de dades de GWAS i de genòmica funcional i proteòmica.**
3. **Evidenciar l'existència d'interaccions genètiques associades amb el risc a càncer de mama i la seva implicació en mecanismes moleculars de susceptibilitat, a partir de la integració de les interaccions predites de les dades de GWAS i de perfils de coexpressió gènica.**

3

Resum dels resultats

La secció de resultats d'aquesta tesi consta de tres articles publicats i un manuscrit en preparació que es resumeixen a continuació. Per altra banda, en l'Annex I, es detallen altres articles relacionats amb el present treball en els quals la doctoranda ha participat i n'és coautora.

1. **Biological processes, properties and molecular wiring diagrams of candidate low-penetrance breast cancer susceptibility genes.**
   Bonifaci N, Berenguer A, Diez J, Reina O, Medina I, Dopazo J, Moreno V, Pujana MA.
   BMC Med Genomics 2008, 1:62.

2. **Exploring the link between germline and somatic genetic alterations in breast carcinogenesis.**
   Bonifaci N, Górski B, Masojc B, Wokolorczyk D, Jakubowska A, Debniak T, Berenguer A, Serra Musach J, Brunet J, Dopazo J, Narod SA, Lubinski J, Lázaro C, Cybulski C, Pujana MA.
   PLoS One 2010, 5(11) e14078.

3. **Evaluating associations between genetic variants at cancer driver kinase loci and cancer risk in BRCA1/2 mutation carriers.**
   Kuchenbaecker K, **Bonifaci N**, Cuadras D, García N, Peterlongo P, Radice P, Barrowdale D, McGuffog L, Serra-Musach J, Ruiz de Garibay G, Gómez A, Esteller M, Díez O, Balmaña J, Lasa A, Ramón y Cajal T, Miramar MD, de la Hoya M, Caldés T, Pérez-Segura P, Osorio A, Benítez J, Bertran M, Izquierdo A, Teulé A, Feliubadaló L, Darder E, Brunet J, XXXCIMBA, Blanco I, Lázaro C, Chenevix-Trench G, Antoniou AC, Pujana MA.
   Manuscrit en preparació.

4. **Integrating gene expression and epidemiological data for the discovery of genetic interactions associated with cancer risk.**
   Bonifaci N, Colas E, Serra-Musach J, Karbalai N, Brunet J, Gómez A, Esteller M, Fernández-Taboada E, Berenguer A, Reventós J, Müller-Myhsok B, Amundadottir L, Duell EJ, Pujana MA.
   Carcinogenesis 2014, 35(3) 578-585.

3.1 Processos biològics, propietats i xarxes moleculars dels candidats a gens de susceptibilitat a càncer de mama de baixa penetrància.

Els GWAS han ajudat a identificar una part de la base genètica comuna del risc a càncer [173]. Tot i això, encara queda per completar una gran part de la heretabilitat a la malaltia [122]. Les limitacions estadístiques que requereixen els GWAS però, compliquen la identificació de les variants genètiques comunes associades a baix increment del risc [174]. Una estratègia basada en una interpretació més biològica dels resultats dels GWAS, evitant les limitacions estadístiques, pot facilitar la priorització de gens candidats, les seves característiques i els processos biològics en els que participen.

En aquest treball es van avaluar associacions/correlacions entre processos biològics i els resultats del GWAS realitzat per CGEMS (de l'anglès *Cancer Genetic Markers of Susceptibility*) [175], així com de diferents estudis a escala genòmica relacionats amb el càncer de mama. En aquests anàlisis, a partir dels rànquings, es va examinar la distribució asimètrica dels processos biològics, anotats en el nivell 3 de *Gene Ontology* (GO) [140], utilitzant un algoritme d'enriquiment basat en el concepte de particions. Es va obtenir un llistat de gens candidats que van ser validats analitzant la xarxa d'interacció proteïna-proteïna (HPRD [126]) i una xarxa de regulació transcripcional.

Els principals resultats obtinguts en aquest treball van ser:

1. Els processos biològics de *Transport*, *Cell Communication* i *Cell Adhesion* estaven distribuïts asimètricament en el rànquing de resultats del GWAS i, per tant, són possiblement processos rellevants en el risc a càncer de mama.

2. Els processos biològics de *Cell Communication* i/o *Cell Adhesion* també van presentar asimetries en tres dels aspectes de la biologia del càncer de mama analitzades: (i) expressió diferencial entre teixit normal i tumoral [176]; (ii) depleció de *BRCA1* en un model cel·lular no tumorigenic [177]; i (iii) associació entre expressió gènica en tumors i l'edat al diagnòstic [178]. Per tant, aquestes tres condicions serien útils per definir les característiques dels gens que contribueixen al risc a càncer de mama. A partir dels gens en comú d'aquests tres llistats i utilitzant la posició mitjana com a mesura, es va crear un rànquing combinat de gens candidats.

3. Els interactors directes i a un sol pas en l'interactoma de proteïnes (HPRD) dels productes de reconeguts gens de susceptibilitat de baixa penetrància (referents: *CASP8*, *CDH1*, *FGFR2*, *HMMR*, *LSP1*, *RASSF1* i *TGFBR1*) presenten una proporció d'anotacions en *Cell Communication* i *Cell Adhesion* major que la mitjana de la component principal de la xarxa. Això recolza els resultats anteriors a nivell de gens candidats identificats a partir de l'anàlisi de dades de GWAS. Ressaltant els interactors directes i a un pas d'aquests quatre referents més destacats, es pot llavors acotar la llista de candidats més probables.

4. La xarxa de pertorbacions de l'expressió (basada en *expression quantitative trait loci*, eQTLs), on els nodes representen els SNP/*loci* dels 50 primers candidats més els referents i les connexions representen la direcció de l'efecte en l'expressió gènica, presenta una connectivitat més elevada que les xarxes de gens triats a l'atzar tant pel que fa al nombre de nodes com al nombre d'arestes. Això recolza l'associació funcional entre els 50 primers candidats i, a més, la connexió amb gens referents de risc a càncer de mama. Els nous candidats es poden prioritzar en funció de l'alta centralitat en la xarxa (*BCL2*, *BMP1*, *NTRK2*, *PTGER3* o *RUNX2*) o pel fet de connectar dos referents (*DKK3* i *NTRK2*).

Aquest estudi proposa *Cell Communication* i *Cell Adhesion* com a processos biològics pertorbats en el risc a càncer de mama conferit per variants de baixa penetrància i defineix les propietats, interaccions moleculars i els possibles efectes funcionals dels gens i les proteïnes candidates.

BMC Medical Genomics

BioMed Central

Research article

Open Access

Biological processes, properties and molecular wiring diagrams of candidate low-penetrance breast cancer susceptibility genes

Núria Bonifaci[†1], Antoni Berenguer[†1], Javier Díez[1], Oscar Reina[2], Ignacio Medina[3], Joaquín Dopazo[3], Víctor Moreno[1] and Miguel Angel Pujana*[1]

Address: [1]Bioinformatics and Biostatistics Unit, and Translational Research Laboratory, Catalan Institute of Oncology, Bellvitge Biomedical Research Institute (IDIBELL), L'Hospitalet, Barcelona, Spain, [2]Unit of Infections and Cancer, Biomedical Research Centre Network for Epidemiology and Public Health (CIBERESP), Cancer Epidemiology Research Program, Catalan Institute of Oncology, Bellvitge Biomedical Research Institute (IDIBELL), L'Hospitalet, Barcelona, Spain and [3]Department of Bioinformatics, Functional Genomics Node and Biomedical Research Centre Network for Rare Diseases (CIBERER), Centro de Investigación Príncipe Felipe (CIPF), Valencia, Spain

Email: Núria Bonifaci - nbonifaci@iconcologia.net; Antoni Berenguer - aberenguer@iconcologia.net; Javier Díez - jdiez@iconcologia.net; Oscar Reina - oreina@iconcologia.net; Ignacio Medina - imedina@cipf.es; Joaquín Dopazo - jdopazo@cipf.es; Víctor Moreno - v.moreno@iconcologia.net; Miguel Angel Pujana* - mapujana@iconcologia.net

* Corresponding author †Equal contributors

Published: 18 December 2008

BMC Medical Genomics 2008, **1**:62 doi:10.1186/1755-8794-1-62

Received: 20 June 2008
Accepted: 18 December 2008

This article is available from: http://www.biomedcentral.com/1755-8794/1/62

Abstract

Background: Recent advances in whole-genome association studies (WGASs) for human cancer risk are beginning to provide the part lists of low-penetrance susceptibility genes. However, statistical analysis in these studies is complicated by the vast number of genetic variants examined and the weak effects observed, as a result of which constraints must be incorporated into the study design and analytical approach. In this scenario, biological attributes beyond the adjusted statistics generally receive little attention and, more importantly, the fundamental biological characteristics of low-penetrance susceptibility genes have yet to be determined.

Methods: We applied an integrative approach for identifying candidate low-penetrance breast cancer susceptibility genes, their characteristics and molecular networks through the analysis of diverse sources of biological evidence.

Results: First, examination of the distribution of Gene Ontology terms in ordered WGAS results identified asymmetrical distribution of Cell Communication and Cell Death processes linked to risk. Second, analysis of 11 different types of molecular or functional relationships in genomic and proteomic data sets defined the "omic" properties of candidate genes: i/ differential expression in tumors relative to normal tissue; ii/ somatic genomic copy number changes correlating with gene expression levels; iii/ differentially expressed across age at diagnosis; and iv/ expression changes after *BRCA1* perturbation. Finally, network modeling of the effects of variants on germline gene expression showed higher connectivity than expected by chance between novel candidates and with known susceptibility genes, which supports functional relationships and provides mechanistic hypotheses of risk.

Conclusion: This study proposes that cell communication and cell death are major biological processes perturbed in risk of breast cancer conferred by low-penetrance variants, and defines the common omic properties, molecular interactions and possible functional effects of candidate genes and proteins.

Background

Technical and methodological advances in genome-wide assessment of genetic variation have provided tools for detecting low-penetrance susceptibility genes for common human diseases [1]. As a result of this progress, the last year has seen a spectacular increase in the number of published studies in which these types of variants or single nucleotide polymorphisms (SNPs) are detected. Projects such as the National Cancer Institute's Cancer Genetic Markers of Susceptibility (CGEMS) and work carried out by deCODE Genetics and the Breast Cancer Association Consortium have produced partial lists of the risk variants of different cancer types in diverse populations [2-4].

Whole-genome association studies (WGAS) are unbiased, which is highlighted by the fact that they identify unexpected candidate genes that are not strictly involved in *a priori* biological process such as DNA damage response in breast cancer [2-4]. The absence of bias is further revealed by the identification of possible master susceptibility loci for different cancer types, such as the convergence of risk variants at chromosome 8q24 [3,5-12]. The drawback of the agnostic nature of WGAS is the challenging statistical analysis and, thus, the biological interpretation of the results beyond single candidate SNPs and their *P* values. The vast number of variants interrogated means that *P* values below 10^{-7} must be obtained to pass multiple-comparison corrections. Consequently, the number of samples needed to obtain the necessary statistical power is an important limitation, as is the fact that uncontrolled population stratification may introduce false positives. In addition, most variants seem to confer very modest risks in the order of 1.2–1.6 fold, which are hard to detect given the statistical difficulties described above. Indeed, current WGAS results contain thousands of SNPs and, by extension, thousands of candidate genes with unadjusted *P* values of < 0.05. As a result of these complications, the findings cannot be considered true positives until they have been replicated in an independent, preferentially larger-scale study [13,14].

Given these statistical constraints, possible biological interpretations of WGAS results are generally overlooked. In most cases genes are interpreted individually, and a gene ranked below the significance threshold will not be measured or experimentally characterized in relation to the disease or to genes that passed the threshold unless strong evidence is obtained from additional association studies. In this scenario, the fundamental principles of low-penetrance susceptibility genes and/or proteins (genes/proteins) – such as biological processes or pathways, properties and the molecular networks in which they commonly participate – have yet to be defined.

Systems-based interpretation of biological data is a common strategy in many areas of research [15-17]. It is clear that genes and proteins are organized in higher-order structures within complex molecular networks to execute biological functions [18]. The genes/proteins organized in these structures are the indivisible elements that are disrupted or regulated abnormally in disease but alterations of different genes/proteins in the same functional unit often converge in a common disease phenotype [19]. Genetic variability that confers risk of common diseases is also likely to converge at some level in specific processes or functions. Pioneering work by Wang and Bucan [20] has shown that the use of biological labels and microarray data analysis tools can facilitate the interpretation and priorization of candidate genes in WGAS.

Taking breast cancer as a model, we applied an integrative approach for uncovering the biological processes underlying breast cancer susceptibility mediated by low-penetrant alleles, as well as the genes/proteins and their properties and molecular interactions that are critical in cancer risk. Our strategy avoids the statistical constraints of WGAS by providing a method for prioritizing candidate markers based on the identification of common biological processes and characteristics. In addition, we provide hypotheses on the possible molecular mechanisms of risk between novel candidates and known susceptibility genes/proteins.

Methods

WGAS ordered gene lists

The breast cancer pre-computed WGAS data set released by the CGEMS initiative was downloaded from the corresponding public web site on September 2007. To examine biological information in WGAS results, we generated two complementary gene ranks: one according to the lowest *P* value per gene for the genotypic test in a genomic region of +/- 10 kilo bases (kb) at each locus, adjusted for age and hormone therapy [2]; and the other according to the lowest *P* value but also taking into account the direction of the association using the OR of the minor allele homozygotes (ORs of either > 1 or < 1). Assigned SNPs were curated using Ensembl gene annotations. Note that *P* values and ORs are not strictly comparable as they reflect different statistical analyses; the *P* values indicate the significance of an SNP in a logistic regression model, whereas the OR compares the magnitude of association of an allele against major homozygotes. The "one SNP-one gene" simplification was applied to obtain a single representation of each gene in the ranks. This approach might over-estimate large gene loci, and other strategies that account for the number of SNPs per gene, their linkage disequilibrium and allele frequencies could be used to enhance this analysis. The rank based on *P* values was then examined for differential representation of biological processes at one tail (low *P*

values), while the rank based on ORs may differentiate disease-risk mechanisms (OR > 1) from protection mechanisms (OR < 1). By assigning SNPs as described above, a rank of 24,458 unique gene symbols (NCBI build 36.1) was obtained from an initial number of 528,173 SNPs [2]. Note that with *P* values of < 0.05, the original data set contains 26,859 SNPs corresponding to 7,611 genes. The number of unique genes in the OR-based rank was slightly lower (n = 24,135) because some of the SNPs had no data for minor homozygotes. The reference unit in our analyses was either the Entrez gene symbol or the Ensembl identifier (release 49), and other identifiers were converted to these references using BioMart [21]. Inconsistencies or missing values between Entrez and Ensembl identifiers were curated manually.

GO term annotations

The Gene Ontology (GO) [22] annotations were downloaded from Open Biological Ontologies version 1.2, release 200804 (MySQL version). GO terms were assigned to gene symbols after record linkage in which regular expression searches were required. Splicing variants were collapsed for each gene symbol. Genes annotated at Level 4 or lower in the GO hierarchy were assigned to a parent in Level 3, but those also occurring at Level 2 were excluded. This analysis gave 14,659 (~60%) genes annotated (271 terms and a median of 641 genes in each term) from the starting list of 31,591 while 24,458 of the genes were present in the WGAS, of which 11,675 were annotated. The remaining ~40% of genes were unannotated, mainly because they represent uncharacterized genes/proteins or do not contain known biological features. The same procedure was used when evaluating terms at Level 4 giving 1,867 gene sets.

Analysis of rank partitions

We implemented the procedure devised by Al-Shahrour and colleagues [23,24] to examine outputs flexibly (Additional file 1). The implementation was performed in the R language and environment [25] and consisted of the following steps, as defined by the original authors: 1/ the list of gene/protein identifiers was ordered according to a measure of association; 2/ a selected number of partitions p was applied, each of which separated the ordered list into two parts, and used the index in order to force each partition to increase with the same number of genes (we show results for 50 partitions, but we also explored the range between 30 and 50 that was recommended in the original publication [24], which revealed similar results); 3/ for each partition, the frequencies of genes/proteins with a specific GO term annotation were compared using a Fisher's exact test for two-by-two contingency tables; 4/ the previous step was repeated for m terms; 5/ a multi-testing adjustment procedure was applied to *P* values taking into account $p \times m$ tests, using the FDR approach [26] implemented in the *multtest* package [27]; 6/ significant terms were selected and graphics were created in R. In comparison with GSEA, the partitions methodology may be capable of detecting modest differences [24], although it is probably less effective at providing detailed interpretations of the position of these differences. One hundred permutations of gene order in WGAS ranks were examined for possible asymmetries obtained by chance. In addition, in our analyses using partitions, we controlled for possible background bias of annotated and unannotated genes for any term.

GSEA analysis

The GSEA algorithm was applied using the Java implementation [28], with ordered gene lists and annotations from Level 3 and 4 Biological Process GO terms, and the enrichment weighting exponent $p = 1$ (except when examining gene index ranks). The statistical significance (nominal *P* value) of the enrichment score (ES) was calculated in the implementation by permuting the class labels (genes) 1,000 times. Log-transformed *P* values were used in the analysis of WGAS-ordered gene lists.

Analysis of breast cancer-related data sets

Differential expression between normal breast tissue and tumors was assessed at the genome-wide level using the data set provided by Richardson and colleagues [29]. Differences were evaluated using the *t*-statistic across all tumors and also for basal-like or non-basal-like subclasses. No differences were observed in GO term profiles so we used the comparison with all tumors. Genetic alterations in tumor subclasses were evaluated using copy number information from the study of Chin and colleagues [30]. For each SNP-gene position of the WGAS an average copy number was obtained in each tumor class. To calculate correlations between gene expression and copy numbers, we first obtained average gene expression values in tumor classes using all possible probes mapping each gene, and then calculated correlations with copy numbers using the Pearson correlation coefficient (PCC). To evaluate prognosis we used the data set of Chang and colleagues [31], which contains 295 breast tumors. We fitted a Cox regression model to each probe using disease-free survival time information. Models were fitted adjusting for ER tumor status and grade, and likelihood ratio tests were calculated to evaluate the effect of microarray probe values on survival. Genes were then ordered according to hazard ratios and/or *P* values using only the extreme probe results. To evaluate age at diagnosis we used the same data set and fitted a linear model for each probe, adjusting for ER tumor status and grade. Next, we applied the same procedure as that used for the prognosis analysis to obtain a definitive ordered list of genes based on the regression coefficient and the corresponding *P* values. The same data set was used to assess expression differ-

ences between ER-negative and ER-positive breast tumors and for co-expression analyses with benchmark breast cancer genes using the PCC. In addition, we investigated expression perturbations after *BRCA1* depletion in MCF10 cells [32], using fold-changes, and expression perturbations between BRCA1 and sporadic breast tumors (non-hereditary ER-negative and grade 3) using the *t*-test [33]. Finally, we examined gene expression changes in tissue abnormalities precursors of breast cancer, using the *t*-test [34].

Analysis of the human interactome network

The human interactome network was built by combining three previously published data sets, which consist mainly of experimentally verified interactions. The data set based on the Human Protein Reference Database (HPRD) was combined with high-confidence yeast two-hybrid interactions from Rual and colleagues [35] and Stelzl and colleagues [36]. Orthology-based predictions and homodimers were excluded from our analyses. Shortest paths were calculated using only the giant network component and the geodesic formulation given by Freeman [37] using the R programming language [25]. GO term annotations were used as detailed above. Proportions of annotations in direct and one-hop interactors of benchmarks were evaluated in the giant network component using as controls seed proteins annotated with the same terms as the benchmark that was being compared. *P* values were then computed using empirical distributions.

Genetics of gene expression

The Dixon and colleagues data set [38] was down-loaded from the public web site and analyzed focusing on SNPs with lod scores of > 2.3. Variants at $r^2 > 0.8$ were identified using Phase II HapMap release 21a data for individuals with European ancestry. Data is provided for lod scores of > 6 and SNPs-genes in the combined rank, whereas information for variants at lod scores of > 2.3 and $r^2 > 0.8$ is available from the authors. To avoid any bias, the network and simulations only refer to the original SNPs annotated by Dixon and colleagues [38] and exclude variants at $r^2 > 0.8$. Networks were generated in Cytoscape [39] and using the R programming language [25]. SNPs at each gene locus (+/- 10 kb) were collapsed into a single node for network representation.

Results

Biological processes in breast cancer risk

Breast cancer is probably the paradigm of deeply characterized neoplastic process at many molecular levels. The key to this study was the public availability of the landmark WGAS for breast cancer risk released by the CGEMS initiative [2]. We analyzed the results of this WGAS alongside various omic data sets of breast cancer and normal cellular conditions, following a biology-driven strategy based on the asymmetrical representation of biological information in ordered gene lists (Figure 1). The combined rank provides a prioritized list of gene/protein candidates and their interactions in pathology.

To examine the distribution of biological information in WGAS ordered gene lists (see Methods), we compiled Level 3 Biological Process GO term annotations and applied two complementary algorithms: one that uses the "partitions" concept devised by Al-Shahrour and colleagues [23,24] (the implementation of this algorithm is available in Additional file 1); and the Gene Set Enrichment Analysis (GSEA), which evaluates asymmetries based on the Kolmogorov-Smirnov statistic [40]. The first algorithm generates *p* partitions in an ordered gene list and then computes a Fisher's exact test for each of the *p* two-by-two contingency tables to detect asymmetries between the top and the bottom parts of the list. Next, *P* values are corrected based on the false FDR approach [26]. All known genes in the human genome NCBI build 36.1 were included in the examination of WGAS ranks. In our implementation we took into account both annotated and unannotated genes/proteins, which we found to prevent false positives due to background asymmetrical distributions (not shown).

Of the 271 terms in Level 3, asymmetries were identified in the distribution of Transport, Cell Communication and Cell Adhesion processes using the partitions methodology and two possible WGAS ranks (Figure 2a and Methods). To evaluate the significance of these results we performed the same analysis for 100 permutations of gene order. None of the permutations showed significant differences for any of the 271 terms at any partition. In addition, when the GSEA algorithm and our Level 3 annotations were used, the greatest asymmetries were found in the same terms (particularly Cell Adhesion), and smaller differences were observed in other terms including Cell Development and Death (Additional file 2). The consistency of the results suggests that the terms identified represent key biological processes in breast cancer risk conferred by common variants.

As expected, profile differences were observed between the two defined WGAS ranks, and Cell Adhesion was more clearly asymmetrically distributed in the ordered gene list that takes into account the lowest *P* value per gene locus and the corresponding odds ratio (OR) (Figure 2a, right panels). Cell Communication is visibly asymmetrically distributed in the *P* value based rank, whereas the inclusion of OR criteria suggests the existence of gene subgroups in this process associated with risk. Under-representation of genes involved in Metabolism was also revealed at the top of the rank, which leads us to speculate

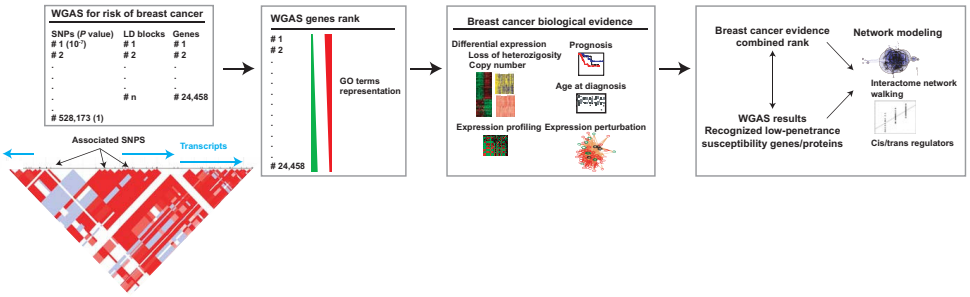


Figure 1

Strategy for candidate gene prioritization in WGAS results. Given a WGAS such as the breast cancer study of the CGEMS initiative [2], ~500,000 SNPs were initially interrogated, which represent a lower number of linkage disequilibrium (LD) blocks in which 24,458 known human genes are distributed. Even when a clear LD block contains several significant SNPs, different genes may be present and molecular and/or functional analyses are required to determine the most likely candidates and their interactions. To obtain this information at the genome-wide level, we propose first to use GO terms to examine the WGAS rank for asymmetries in biological processes. These asymmetries will then be used to guide the analysis of omic data sets relevant to breast cancer biology. Next, higher-level data analyses – protein-protein interactions that may be over-represented for the same processes, and variants in cis/trans affecting germline gene expression levels that lead to hypotheses on the possible functional effects of risk alleles – are performed using a combination of evidences, WGAS results and recognized low-penetrance susceptibility genes/proteins or benchmarks.

that common variants in this process play a protective role.

Fine mapping of processes

Given the asymmetries at Level 3, and taking into account that the gene sets were relatively large, candidate processes were narrowed down using child terms at Level 4. In agreement with results above, terms for Transport, Cell Communication and Cell Adhesion were found to be distributed asymmetrically in both WGAS ranks (Figure 2b). For example, Signal Transduction was a child of Cell Communication and was found to be over-represented at low *P* values. Several recognized low-penetrance susceptibility genes are annotated in this term (*AURKA* [41-45],*CASP8* [46],*LSP1* [3]and *TGFBR1* [47-51]). The child terms for Transport and, in particular, for Cell Adhesion also showed similar asymmetries to those at Level 3 (Figure 2b). Profiles were also found to be consistent with the list ordered by OR/*P* value, with many child terms for Cell Adhesion over-represented at ORs of > 1. These observations corroborate the identification of key processes – in particular Cell Communication and Cell Adhesion – mediating breast cancer risk.

Breast cancer-related properties

To further define the characteristics of candidate susceptibility genes in breast cancer conditions, we examined various sources of biological evidence according to the observed WGAS rank GO asymmetries. Nine types of evidence were examined (Additional file 3):

1/ Differential expression between normal breast tissue and tumors [29] (accounting for different known molecular classes of breast tumors [52]).

2/ Differential expression between normal breast tissue at terminal duct lobular units and hyperplasic units [34].

3/ Correlations between transcript profiles using as benchmarks known genes of low/moderate risk (*ATM*, *AURKA*, *BRIP1*, *CASP8*, *CHEK2*, *FGFR2*, *HMMR*, *LSP1*, *MAP3K1*, *PALB2*, *RASSF1*, *TGFBR1 and TNRC9*), high risk (*BRCA1* and *BRCA2*) and cancer syndromes (*LKB1*, *PTEN* and *TP53*) [53].

4/ Somatic loss of heterozygosity and copy number alterations in tumors [30] (accounting for the different known tumor types).

5/ The correlation between somatic copy number alterations and transcript profiles [30] (again, accounting for the different known tumor types).

6/ The dependence of the estrogen receptor (ER) pathway signaling on differential expression between ER-positive and ER-negative tumors [31,33,54].

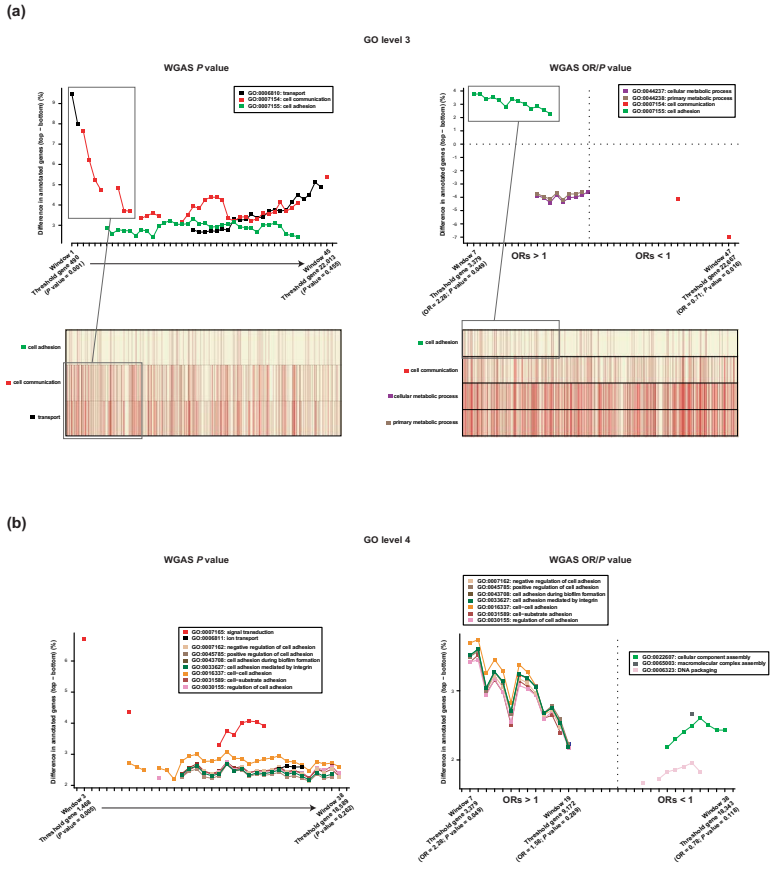


Figure 2

WGAS rank asymmetries for specific biological processes. (a) Graphical representation of over- and/or under-representation of biological processes in partitions of WGAS ranks using Level 3 GO annotations. Top left panel, results of the analysis of the WGAS rank according to the lowest *P* value per gene locus. Differences are always shown from top to bottom, so the top shows over-representation in the GO terms Transport and Cell Communication. Graphics show significant partitions. Bottom left panel, graphical representation of the positions of genes annotated with GO terms distributed asymmetrically in the WGAS *P* value rank. Right panels, results of the analysis of the WGAS rank according to ORs and to *P* values. This analysis seems to better capture the differences in risk (ORs > 1) associated with the over-representation of Cell Adhesion. Under-representation (negative differences when comparing top with bottom parts) of Metabolic processes annotations is also suggested with ORs of > 1. The graphical representation of gene positions shows clear differences between Cell Adhesion and more complex patterns – perhaps with different gene subgroups – for Cell Communication and Metabolic processes. **(b)** Graphical representation of over- and/or under-representation of biological processes in partitions of WGAS ranks using Level 4 GO annotations. Left panel, child terms of Cell Communication (Signal Transduction), Transport (Ion Transport) and Cell Adhesion (rest of terms shown in the inset) are over-represented at *P* values of up to 0.262. Right panel, over-representations in the WGAS OR/*P* value ordered list as shown in the insets.

7/ The association between gene expression and patient prognosis [31,33,54] (adjusting for major confounding variables of ER status and tumor grade).

8/ The association between gene expression in tumors and patient age at diagnosis [31,33,54] (again, adjusting for major confounding variables of ER status and tumor grade).

9/ Expression perturbation in BRCA1 tumors (tumors originating in carriers of germline *BRCA1* mutations) relative to sporadic (non-hereditary) tumors [33], or after depletion of *BRCA1* in a non-tumorigenic cell model [32,55].

These different types of evidence characterize different aspects of breast cancer biology, including the following: the identification of putative tumor suppressors and oncogenes by analyzing differential expression and/or somatic genetic alterations [30]; genes with a role in the early stages of breast tissue transformation [34]; hormone dependencies that may be related to susceptibility, as noted recently for newly identified low-penetrance susceptibility genes [4]; expression perturbations in BRCA1 tumors that may reveal functional relationships with high-penetrance genes/proteins [32,56,57]; and associations with age at diagnosis that may also indicate critical molecular roles in initiating tumorigenesis [57].

Analysis of the evidence described above identified biological processes consistent with existing knowledge in the literature. For example, Cell Division was distributed asymmetrically in genes ranked according to the hazard ratio that measures survival probability (Figure 3), which is consistent with the fact that the potential for cell proliferation can be considered a strong predictor of prognosis or metastasis [58-63].

Of the nine types of evidence described above, three showed similar asymmetries in Cell Adhesion to those observed in the WGAS ranks: differential expression between normal breast tissue and tumors, patient age at diagnosis, and *BRCA1* depletion in MCF10A cells (comparison of BRCA1 and sporadic tumors also revealed similar asymmetries, but it was excluded from the analyses below to avoid duplication). Two of these data sets also showed similar asymmetries for Cell Communication (Figure 4). As mentioned above, permutation analysis of gene ranks did not show asymmetries in any process, which indicates that these evidences are useful for categorizing and defining the omic properties of genes contributing to breast cancer risk.

Asymmetries in these processes were also observed in tumor subclasses when the rank of correlations between somatic genomic alterations and gene expression levels were examined. This was found principally in luminal A tumors (Additional file 4), and although the corresponding combined rank did not vary considerably from those of the three types of evidence described above, it captured as likely candidate genes those involved in ER signaling such as *TFF1* (Additional file 5), which was expected for a hormone-dependent tumor class [52]. This specific evidence for a given subclass can then be used when examining breast cancer subtypes.

Evaluation of a combined evidence rank

Given that three breast cancer conditions showed similar asymmetries in processes to those observed in the WGAS ranks, a combined rank of these conditions might provide a prioritized list of more likely candidates. This analysis was performed using all genes in common between these three omic data sets (n = 8,986) and the final rank was created using the average position (Additional file 6). Although there is not a large "gold standard" of low-penetrance susceptibility genes, some features of the combined rank suggest that it is biologically meaningful in the assessment of genetic risk factors.

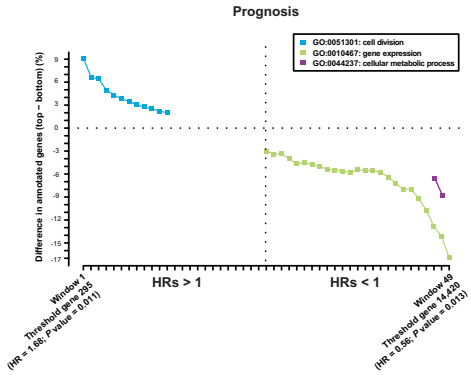


Figure 3

Asymmetries in biological evidence of breast cancer. An example of the results of applying the methodology used to examine the WGAS ranks to a breast cancer biological evidence data set. This analysis identifies the association between gene expression levels and patient survival or prognosis measured by the hazard ratio (HR). The results suggest an association between poor prognosis (HRs > 1) and genes involved in Cell Division, and between good prognosis (HRs < 1) and Gene Expression and Metabolic processes. Importantly, the association between cell proliferation and poor prognosis has been demonstrated in previous studies using different approaches [58-63].

Examination of the 50 top-ranked genes in the combination identified candidate tumor suppressors and/or oncogenes from the literature (*DKK3* [64] and *TFPI2* [65]), genes with variants that confer breast cancer risk (*IGF1* [66]) and, notably, four genes (*PDGFRA*, *PDGFRL*, *MAP3K12* and *NTRK2*) whose products participate in the MAPK signaling pathway, where known susceptibility genes also participate (*FGFR2* and *MAP3K1* [2-4]) (Table 1 shows the results for the 50 top-ranked genes in the combined evidence ranking ordered by their lowest WGAS *P* value). This 50-set also contains genes previously linked to breast cancer prognosis, metastasis or treatment response (*BCL2* [67], *CXCL12* [68-70] and *FBLN1* [71,72]). In addition, consistent with predicted relationships in this set, experimental studies have demonstrated interactions between the corresponding proteins in neoplasia; for example ABTB1 and EGR2 are mediators of PTEN tumor suppressor function [73]. These observations support the hypothesis that the combined rank contains numerous functional and molecular associations of relevance for breast tumorigenesis.

The second position of the combined ranking that takes into account the WGAS results is occupied by the platelet-derived growth factor receptor-like (*PDGFRL*) gene, while the first gene in the combined rank is *PDGFRA* (Table 1). *PDGFRA* is expressed in invasive carcinomas and is associated with aggressiveness [74], and, importantly, *PDGFRL* is mutated in cancer cells [75,76] and maps at chromosome 8p22-p21, where it is thought to map a breast cancer tumor suppressor gene(s) [77-79]. More recently, an integrative approach based on disease-specific pathways has revealed that *PDGFRL* may play a critical role in promoting breast tumorigenesis [80]. Our independent observations of breast cancer risk may lead to the replication of the WGAS findings for these *PDGFR* genes and others shown in Table 1. In this way, evaluation of genes with somatic point mutations in breast tumors as compiled in the COSMIC database (release v36) [81] placed *MAP3K12* at the top of the combined rank (Additional file 7), which reinforces the putative involvement of the MAPK signaling pathway and supports *MAP3K12* as a likely candidate.

Examination and integration of higher-order evidence

Correlations across different biological levels provide better proof of molecular associations and their possible perturbation in disease [16,18,82]. We examined the network of protein-protein interactions (interactome network) of recognized low-penetrance susceptibility gene products (hereafter referred to as benchmarks) for proportions of annotations in Cell Communication and Cell Adhesion. Proportions of annotations were compared between interactors of benchmarks and the average in the giant network component and, to avoid bias, only proteins annotated at

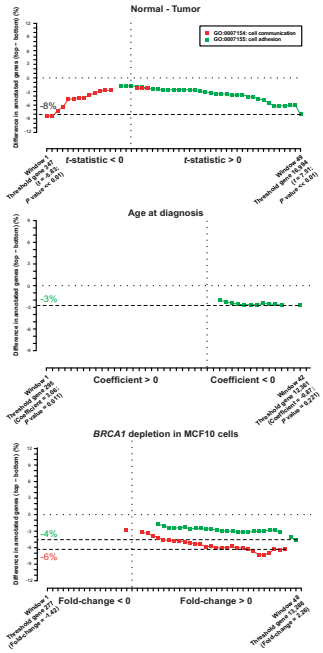


Figure 4

Asymmetries for Cell Communication and Cell Adhesion. Of the biological evidence of breast cancer examined in this study, three cases showed asymmetries in biological processes that are similar to those observed in the WGAS ranks. Three cases showed similar asymmetries for Cell Adhesion: 1/ top panel, differential expression between normal breast tissues and tumors measured using the *t*-statistic, as a result of which genes involved in Cell Adhesion are over-represented at the bottom, which indicates that they are generally under-expressed in tumors, while Cell Communication is under-represented at the top (note that both patterns follow the same direction); 2/ middle panel, association between age at diagnosis and gene expression levels measured using the coefficient from the linear model, so coefficients < 0 indicate association with early age at diagnosis, which is consistent with the expected contribution of genetic effects to breast cancer risk [57]; 3/ bottom panel, fold change in gene expression changes between *BRCA1*-depleted and control-treated MCF10A cells, which indicates possible molecular and/or functional dependencies on processes linked to breast cancer risk [57]. Differences in annotation percentages between top and bottom range from -8% to -4% for the most significant partition at the bottom end. On the basis of these results, all three ranks were inverted and combined for comparison with the WGAS results.

Table 1. Combination of breast cancer biology evidence and evaluation of WGAS results for priorization of candidate low-penetrance susceptibility genes*

	Breast cancer biology evidence				WGAS	
	Combined	Normal-Tumor	Age at diagnosis	BRCA1 depletion	SNP	*P* value
AZGP1	40	551	1156	655	rs2107349	1.70E-05
PDGFRL	31	84	1838	176	rs2720552	0.0004
DKK3 (§)	16	63	1083	598	rs3750936	0.0018
ITPR1	28	318	1514	222	rs7616234	0.0025
ZNF423	6	657	44	572	rs193908	0.0034
BCL2 (§)	11	618	124	782	rs8086644	0.0036
MAP3K12	5	794	68	294	rs8192593	0.0041
SNRPN	46	849	1198	381	rs2732043	0.0049
PDGFRA	1	24	105	120	rs7660560	0.0051
BNC2	41	612	563	1215	rs13290470	0.0067
BMP1	45	1408	883	123	rs7814885	0.0085
LAMA4	43	468	318	1608	rs2072020	0.0104
COL14A1	3	10	141	628	rs6989074	0.0114
F2RL2	14	1298	305	115	rs2455232	0.0119
SYNPO2 (§)	49	355	1287	920	rs10518312	0.0124
ANK2	32	328	1296	512	rs17580458	0.0132
OLFM1	47	1466	454	541	rs4262379	0.0135
VWF	30	676	1060	359	rs2283333	0.0145
PTGER3	8	748	423	171	rs4649932	0.0147
LDB2	2	150	43	528	rs13110049	0.0174
FMO1	38	278	1847	149	rs17642661	0.0275
SSPN	12	49	589	911	rs12827659	0.0282
ANGPTL2	7	360	822	112	rs1768374	0.0288
RUNX1	39	866	1410	6	rs9974986	0.0289
NTRK2 (§)	29	42	794	1225	rs2120266	0.0374
JAM3	50	113	2313	154	rs12287552	0.0499
JAM2	25	539	953	547	rs7283477	0.0526
TBX15	15	767	260	701	rs716217	0.0536
FBLN1	13	276	1035	391	rs17564689	0.0561
DLL1	34	634	381	1140	rs9348305	0.0638
CXCL12 (§)	24	197	742	1050	rs4948878	0.0785
HOXA2	23	927	428	533	rs7811753	0.1236
IGF1	19	60	828	916	rs2971575	0.1251
COL6A1	9	1069	193	155	rs2277814	0.1319
BTD	35	486	1630	64	rs6797119	0.1386
CNOT6L	18	819	47	923	rs1587563	0.1410
EGR2	22	34	99	1753	rs224278	0.1427
TFPI2	44	1153	625	626	rs180283	0.1540
ABTB1	27	1110	204	736	rs782446	0.1570
NPR1	42	497	1046	848	rs7541193	0.1641
PCOLCE	33	673	1050	431	rs11768465	0.1781
LTBP4	21	287	727	853	rs1131620	0.1869
BTNL9	26	135	29	1883	rs7725222	0.2352
MRGPRF	36	464	1042	683	rs1249582	0.2862
PLD1	48	2014	227	292	rs10440040	0.3314
LTC4S	17	1411	358	7	rs6895902	0.3400
SCGB2A2	20	732	790	295	rs10501333	0.3466
PODN	10	212	1087	183	rs1288386	0.4446
BHLHB3	4	563	126	397	rs3809140	0.5892
STAT2	37	1561	345	289	rs2066808	0.6143

***Table showing the 50 top-ranked genes using a combination of three global breast cancer biology evidence and ordered according to their lowest WGAS *P* value.**
Genes in bold type indicate products directly interacting or at one-hop in the interactome network of benchmark proteins (CDH1, FGFR2, HMMR or RASSF1).
(§) Loci with variants possibly affecting the germline expression of a recognized low-penetrance susceptibility gene(s).

Table 1
Combination of breast cancer biology evidence and evaluation of WGAS results for priorization of candidate low-penetrance susceptibility genes

any GO level were considered. Using as network seeds those nodes representing seven benchmark proteins with at least one known interaction in the giant component (CASP8, CDH1, FGFR2, HMMR, LSP1, RASSF1 and TGFBR1), over-representation of Cell Communication and Cell Adhesion was detected in several neighborhoods using the shortest path measure, particularly in direct and one-hop interactors (Figure 5a/b). The benchmark neighborhoods showing the highest over-representation of these processes were those corresponding to CDH1, FGFR2, HMMR and RASSF1 (Additional file 8).

To assess which of these benchmarks shown the maximum information at the interactome level for breast cancer risk, we calculated the probability of showing similar proportions of annotations in the giant component and, to avoid functional bias, used as controls seed proteins with the same annotations at Level 3 as each of the benchmarks being compared. The results of this controlled analysis suggest higher enrichment of the processes in the direct or one-hop interactors of CDH1 and FGFR2 (percentile 87 and 94, respectively) (Figure 5c). This observation suggests the close interactors of these low-penetrance susceptibility gene products as more likely candidates.

The results in the interactome network provide additional information that can be combined discretely with the rank in Table 1. Consequently, annotating this rank for direct and one-hop interactors of CDH1, FGFR2, HMMR and RASSF1 provides a more restricted list of likely candidates. Again, this set contains previously defined candidates such as IGF1 [66] and members of the MAPK signaling pathway such as NTRK2 and PDGFRA, which are found in the one-hop neighborhood of FGFR2.

Functional effects of variants and their evaluation in the combined rank

To determine the possible functional effects of risk variants in candidates, we examined differences in germline expression levels correlating with genetic variation, using the data set of Dixon and colleagues [38] derived from lymphoblastoid cell lines. To search SNPs we used the original data or, in cases which provided no information for an SNP, variants at linkage disequilibrium $r^2 > 0.8$ according to HapMap individuals with European ancestry [83]. In this analysis we not only examined single SNP/gene effects (Additional file 9) but also generated expression-perturbation networks in which nodes are formed by gene loci and edges represent direct or indirect expression effects, possibly mediated by coding and/or regulatory SNPs in candidate genes (see Methods).

Taking as candidates the 50 top-ranked genes from Table 1, we identified many edges between their loci and with benchmarks (Figure 6, left panel). New candidates may then be prioritized based on their high centrality in the network (*BCL2*, *BMP1*, *NTRK2*, *PTGER3* or *RUNX2*) or by the fact that they connect two benchmarks (*DKK3* and *NTRK2* connect *HMMR-TNRC9* and *FGFR2-TGFBR1*, respectively), which suggests a possible risk effect through the expression perturbation of known low-penetrance susceptibility genes.

To evaluate the biological significance of this network, we performed similar analyses with 100 randomly chosen sets of 50 genes and the same benchmarks. The connectivity was higher for the 50 top-tanked genes in the combined rank than for any of the randomly generated networks, both for the number of nodes and the number of edges (Figure 6, right panels). This observation supports the functional association between the 50 top-ranked candidates and, importantly, the association with known genes of breast cancer risk. These results also provide many functional hypotheses of genetic variants in redefined candidates that may influence breast cancer susceptibility. Overall, this integrative study identifies candidate low-penetrance breast cancer susceptibility genes and the corresponding wiring diagram of molecular interactions.

Discussion

This study identifies biological processes that play key roles in breast cancer risk, which are revealed by asymmetrical distributions of GO terms in complete WGAS ranks. Common variants that affect, in particular, the function of genes/proteins in Cell Communication and Cell Adhesion probably confer breast cancer risk to a greater extent than variants in genes associated with different processes. Thus, this study provides a foundation for the analysis of fundamental issues in breast cancer risk conferred by low-penetrant alleles.

The involvement of Cell Communication and Cell Adhesion is intriguing given their long-known contribution to epithelial neoplasia, although typically at the somatic level [84]. Our results may link initial molecular perturbations to subsequent events in cancer progression, which suggests a more continuous path than previously thought between germline and somatic alterations. This hypothesis was highlighted primarily by the identification of risk variants at the *FGFR2* and *MAP3K1* loci – two genes known to be somatically altered in human cancer and whose products are involved in signal transduction among other processes [85,86]. These considerations apply to sporadic breast cancer but may also provide insights into the mechanisms of high-penetrance susceptibility genes since risk variants at low-penetrance loci also contribute to the risk of *BRCA1* and *BRCA2* mutation carriers [87]. Overall, these observations point to a molecular diagram for breast cancer risk that may be more complex

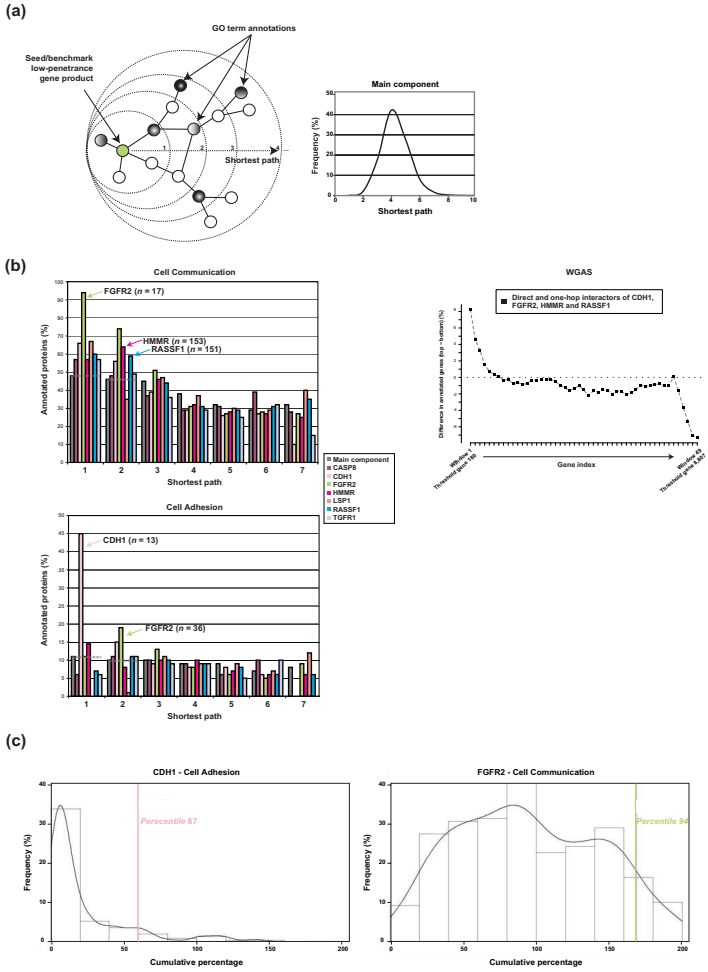


Figure 5

Same biological processes in the interactome network neighborhoods. (a) Left panel, strategy used to examine the interactome network; given a seed or benchmark protein encoded by a recognized low-penetrance susceptibility gene and using a shortest path algorithm, we calculated at each step the percentage of nodes annotated with Cell Communication or Cell Adhesion among proteins annotated with any term (excluding non-annotated proteins). Right panel, distribution of all possible short paths in the giant network component. **(b)** Left panels, results for percentages in short paths of up to seven steps for benchmark proteins. Over-representation in Cell Communication and Cell Adhesion annotations is suggested for CDH1, FGFR2, HMMR and RASSF1 at direct and/or one-hop interactions. Right panel, asymmetrical distribution of CDH1, FGFR2, HMMR and RASSF1 direct and one-hop interactors in the complete WGAS rank. **(c)** Over-representation of processes in the one-hop neighborhood of CDH1 or FGFR2 (vertical lines) using as controls seed proteins with the same Level 3 annotations (curves). The x-axis represents the cumulative percentage up to 200. The CDH1 and FGFR2 percentiles are shown.

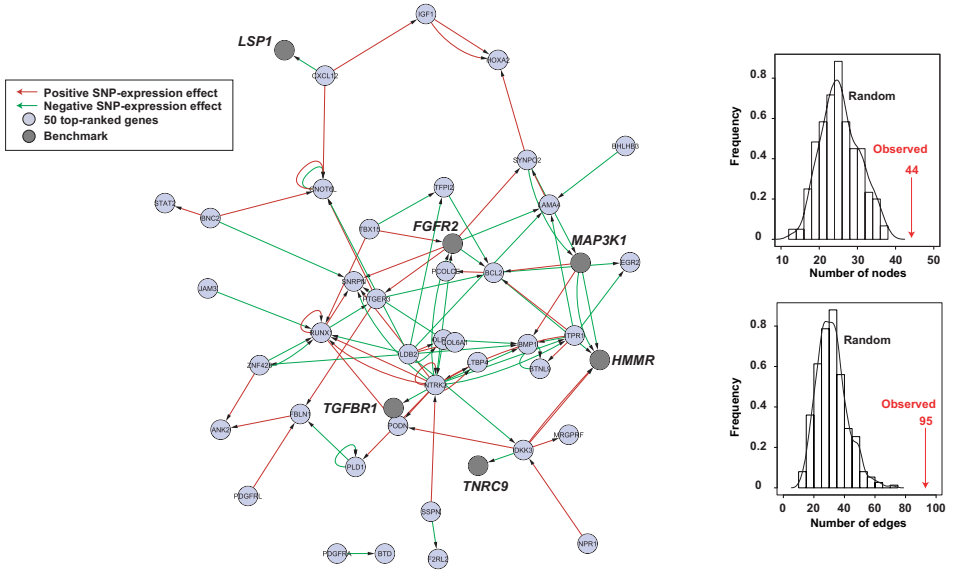


Figure 6

Functional effects and associations between candidates and benchmarks. Left panel, network of transcriptional perturbations mediated by SNPs at gene loci. Nodes represent SNP/gene loci of the 50 top-ranked candidates (Table 1) or of benchmarks, and edges represent the direction of the effect on gene expression, as shown in the inset. To avoid bias, we excluded those SNPs that are not annotated in the original data set of Dixon and colleagues [38]. Right panels, network results of the analysis of 100 randomly chosen sets of 50 genes and the same benchmarks (histograms and curves) compared to the observed values in the left panel (vertical arrows), for connected nodes (top) or edges (bottom).

than previously thought, probably based not only on the alteration of the DNA damage response.

However, the limitations of this study must also be presented. Firstly, methodological constraints might hamper the detection of subtle asymmetries of GO terms. To improve sensitivity, WGAS results could be ordered by combining the effect and magnitude of variants using Bayesian principles. Alternatively, different biological labels could be used – we considered annotations of pathways [88] that did not reveal significant differences (not shown). Secondly, although the application of the average across ordered lists detected genes/proteins known to be involved in breast tumorigenesis (Table 1), more sophisticated methods for combining ranks could improve the detection of susceptibility genes. Finally, this study is limited by the analysis of a single WGAS data set with certain epidemiological specificities [2], thus any candidate highlighted here should be examined in an independent epidemiological study.

Based on the observations from the WGAS ranks, we then examined different breast cancer conditions that could provide further categorization of candidates and reveal the common properties of low-penetrance susceptibility genes. Variants of these genes appear to correlate with transcripts that are differentially expressed in tumors, with somatic copy number changes that correlate with gene expression, differentially expressed across age at diagnosis, and which show changes in expression level after depletion or in the presence of *BRCA1* mutation. Correlations between somatic genomic alterations and gene expression may indicate tumor suppressors or oncogenes, depending on the direction of the correlation [89]. The association with age at diagnosis (identified when adjusting for confounding variables) supports a role in cancer risk, for example differential expression at early age [57]. Finally, changes mediated by *BRCA1* perturbation suggest molecular or functional dependencies with high-penetrance susceptibility genes/proteins [56,57]. This study

suggests that these are frequent features of low-penetrance breast cancer susceptibility genes.

Combination of these evidences provides a comprehensive rank to evaluate WGAS results beyond statistical constraints. This observation is supported by analyses at higher-order molecular levels. Direct and one-hop physical interactors of susceptibility benchmarks are over-represented in the same biological processes as the top of the WGAS ranks. In addition, modeling of a germline transcriptional regulatory network identifies connections with benchmarks but also reveals higher connectivity than randomly expected, which supports that these genes/proteins function in biologically related processes. We propose this integrative study provides the basis for better biological knowledge of the genes/proteins, their omic properties and interactions that mediate the initial steps of breast tumorigenesis. This strategy may be useful for revealing the genes/proteins and their wiring molecular diagrams of susceptibility for other cancer types where WGAS are being carrying out and have vast omic data sets.

Conclusion

This study proposes biological criteria that may facilitate the prioritization of candidate genes in WGAS for breast cancer. The identification of the processes, omic properties and molecular interactions may represent the first step towards a more comprehensive understanding of the molecular mechanisms of risk of breast cancer conferred by of low-penetrance susceptibility genes.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

NB, AB participated in the study design and performed the WGAS and omic data analyses. OR performed the analysis of the regulatory network. JD and VM participated in scientific discussions and helped with the overall interpretation of the data. IM and JD helped in the analyses using the partition algorithm. MAP conceived and designed the study and drafted the manuscript. All authors read and approved the final manuscript.

Additional material

Additional file 1

Algorithm for partition strategy in studying ordered lists.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1755-8794-1-62-S1.txt]

Additional file 2

Results of the GSEA algorithm.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1755-8794-1-62-S2.pdf]

Additional file 3

Analyses of breast cancer data sets.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1755-8794-1-62-S3.xls]

Additional file 4

Asymmetries in the rank of somatic copy number-gene expression correlations in luminal A tumors.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1755-8794-1-62-S4.pdf]

Additional file 5

Combined rank of breast cancer biological evidence including somatic copy number-gene expression correlations in luminal A tumors.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1755-8794-1-62-S5.xls]

Additional file 6

Combined rank of common breast cancer biological evidence.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1755-8794-1-62-S6.xls]

Additional file 7

Combined rank of COSMIC genes.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1755-8794-1-62-S7.xls]

Additional file 8

Direct and one-hop interactors of benchmark low-penetrance susceptibility gene products.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1755-8794-1-62-S8.xls]

Additional file 9

SNPs with possible functional effects at lod score > 6. Information for variants at lod scores > 2.3 is available from the authors.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1755-8794-1-62-S9.xls]

Acknowledgements

The authors are indebted to the CGEMS initiative as well as to many other authors publicly providing omic data sets. This study was supported by the Catalan Institute of Oncology, the "la Caixa" Foundation grant BM 05/254, the Spanish Ministry of Health grants FIS 05/1006 and 06/0545, RCESP C03/09 and RTICCC C03/10, and the Catalan Government DURSI grant

2005SGR00727. MAP is a Ramón y Cajal Researcher with the Spanish Ministry of Education and Science.

References

1. Kruglyak L: **The road to genome-wide association studies.** *Nat Rev Genet* 2008, **9:**314-318.
2. Hunter DJ, Kraft P, Jacobs KB, Cox DG, Yeager M, Hankinson SE, Wacholder S, Wang Z, Welch R, Hutchinson A, Wang J, Yu K, Chatterjee N, Orr N, Willett WC, Colditz GA, Ziegler RG, Berg CD, Buys SS, McCarty CA, Feigelson HS, Calle EE, Thun MJ, Hayes RB, Tucker M, Gerhard DS, Fraumeni JF Jr, Hoover RN, Thomas G, Chanock SJ: **A genome-wide association study identifies alleles in *FGFR2* associated with risk of sporadic postmenopausal breast cancer.** *Nat Genet* 2007, **39:**870-874.
3. 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 C-Y, Wu P-E, Wang H-C, Eccles D, Evans DG, Peto J, Fletcher O, *et al.*: **Genome-wide association study identifies novel breast cancer susceptibility loci.** *Nature* 2007, **447:**1087-1093.
4. Stacey SN, Manolescu A, Sulem P, Thorlacius S, Gudjonsson SA, Jonsson GF, Jakobsdottir M, Bergthorsson JT, Gudmundsson J, Aben KK, Strobbe LJ, Swinkels DW, van Engelenburg KC, Henderson BE, Kolonel LN, Le Marchand L, Millastre E, Andres R, Saez B, Lambea J, Godino J, Polo E, Tres A, Picelli S, Rantala J, Margolin S, Jonsson T, Sigurdsson H, Jonsdottir T, Hrafnkelsson J, *et al.*: **Common variants on chromosome 5p12 confer susceptibility to estrogen receptor-positive breast cancer.** *Nat Genet* 2008, **40(6):**703-706.
5. 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, *et al.*: **Genome-wide association study identifies a second prostate cancer susceptibility variant at 8q24.** *Nat Genet* 2007, **39:**631-637.
6. 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:**638-644.
7. 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, *et al.*: **Genome-wide association study of prostate cancer identifies a second risk locus at 8q24.** *Nat Genet* 2007, **39:**645-649.
8. 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, *et al.*: **A genome-wide association scan of tag SNPs identifies a susceptibility variant for colorectal cancer at 8q24.21.** *Nat Genet* 2007, **39:**984-988.
9. 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 J-F, 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, *et al.*: **Genome-wide association scan identifies a colorectal cancer susceptibility locus on chromosome 8q24.** *Nat Genet* 2007, **39:**989-994.
10. Witte JS: **Multiple prostate cancer risk variants on 8q24.** *Nat Genet* 2007, **39:**579-580.
11. 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 USA* 2006, **103:**14068-14073.
12. 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, *et al.*: **A common variant associated with prostate cancer in European and African populations.** *Nat Genet* 2006, **38:**652-658.
13. Hunter DJ, Thomas G, Hoover RN, Chanock SJ: **Scanning the horizon: what is the future of genome-wide association studies in accelerating discoveries in cancer etiology and prevention?** *Cancer Causes Control* 2007, **18:**479-484.
14. Chanock SJ, Manolio T, Boehnke M, Boerwinkle E, Hunter DJ, Thomas G, Hirschhorn JN, Abecasis G, Altshuler D, Bailey-Wilson JE, Brooks LD, Cardon LR, Daly M, Donnelly P, Fraumeni JF Jr, Freimer NB, Gerhard DS, Gunter C, Guttmacher AE, Guyer MS, Harris EL, Hoh J, Hoover R, Kong CA, Merikangas KR, Morton CC, Palmer LJ, Phimister EG, Rice JP, Roberts J, *et al.*: **Replicating genotype-phenotype associations.** *Nature* 2007, **447:**655-660.
15. Liu ET: **Systems biology, integrative biology, predictive biology.** *Cell* 2005, **121:**505-506.
16. Kitano H: **Systems biology: a brief overview.** *Science* 2002, **295:**1662-1664.
17. Vidal M: **Interactome modeling.** *FEBS Lett* 2005, **579:**1834-1838.
18. Barabasi AL, Oltvai ZN: **Network biology: understanding the cell's functional organization.** *Nat Rev Genet* 2004, **5:**101-113.
19. Loscalzo J, Kohane I, Barabasi AL: **Human disease classification in the postgenomic era: a complex systems approach to human pathobiology.** *Mol Syst Biol* 2007, **3:**124.
20. Wang K, Li M, Bucan M: **Pathway-based approaches for analysis of genomewide association studies.** *Am J Hum Genet* 2007, **81(6):**. Epub ahead of print
21. Durinck S, Moreau Y, Kasprzyk A, Davis S, De Moor B, Brazma A, Huber W: **BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis.** *Bioinformatics* 2005, **21:**3439-3440.
22. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G: **Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.** *Nat Genet* 2000, **25:**25-29.
23. Al-Shahrour F, Arbiza L, Dopazo H, Huerta-Cepas J, Minguez P, Montaner D, Dopazo J: **From genes to functional classes in the study of biological systems.** *BMC Bioinformatics* 2007, **8:**114.
24. Al-Shahrour F, Diaz-Uriarte R, Dopazo J: **Discovering molecular functions significantly related to phenotypes by combining gene expression data and biological information.** *Bioinformatics* 2005, **21:**2988-2993.
25. R Development Core Team: **R: A language and environment for statistical computing.** *ISBN* 2005. 3-900051-07-0
26. Benjamini Y, Hochberg Y: **Controlling the false discovery rate: a practical and powerful approach to multiple testing.** *Journal of the Royal Statistical Society Series B-Statistical Methodology* 1995, **57:**289-300.
27. Pollard KS, Dudoit S, Laan MJ van der: **Multiple testing procedures: R *multtest* package and applications to genomics.** 2004.
28. Subramanian A, Kuehn H, Gould J, Tamayo P, Mesirov JP: **GSEA-P: a desktop application for Gene Set Enrichment Analysis.** *Bioinformatics* 2007, **23:**3251-3253.
29. 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:**121-132.
30. Chin K, DeVries S, Fridlyand J, Spellman PT, Roydasgupta R, Kuo WL, Lapuk A, Neve RM, Qian Z, Ryder T, Chen F, Feiler H, Tokuyasu T, Kingsley C, Dairkee S, Meng Z, Chew K, Pinkel D, Jain A, Ljung BM, Esserman L, Albertson DG, Waldman FM, Gray JW: **Genomic and transcriptional aberrations linked to breast cancer pathophysiologies.** *Cancer Cell* 2006, **10:**529-541.

31. Chang HY, Nuyten DS, Sneddon JB, Hastie T, Tibshirani R, Sorlie T, Dai H, He YD, van't Veer LJ, Bartelink H, Rijn M van de, Brown PO, Vijver MJ van de: **Robustness, scalability, and integration of a wound-response gene expression signature in predicting breast cancer survival.** *Proc Natl Acad Sci USA* 2005, **102:**3738-3743.
32. Furuta S, Wang JM, Wei S, Jeng YM, Jiang X, Gu B, Chen PL, Lee EY, Lee WH: **Removal of BRCA1/CtIP/ZBRK1 repressor complex on ANG1 promoter leads to accelerated mammary tumor growth contributed by prominent vasculature.** *Cancer Cell* 2006, **10:**13-24.
33. van 't Veer LJ, Dai H, Vijver MJ van de, He YD, Hart AA, Mao M, Peterse HL, Kooy K van der, 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:**530-536.
34. Lee S, Medina D, Tsimelzon A, Mohsin SK, Mao S, Wu Y, Allred DC: **Alterations of gene expression in the development of early hyperplastic precursors of breast cancer.** *Am J Pathol* 2007, **171:**252-262.
35. 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, *et al.*: **Towards a proteome-scale map of the human protein-protein interaction network.** *Nature* 2005, **437(7062):**1173-1178.
36. 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:**957-968.
37. Freeman LC: **A set of measures of centrality based on betweenness.** *Sociometry* 1977, **40:**35.
38. Dixon AL, Liang L, Moffatt MF, Chen W, Heath S, Wong KC, Taylor J, Burnett E, Gut I, Farrall M, Lathrop GM, Abecasis GR, Cookson WO: **A genome-wide association study of global gene expression.** *Nat Genet* 2007, **39:**1202-1207.
39. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T: **Cytoscape: a software environment for integrated models of biomolecular interaction networks.** *Genome Res* 2003, **13:**2498-2504.
40. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP: **Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles.** *Proc Natl Acad Sci USA* 2005, **102:**15545-15550.
41. Cox DG, Hankinson SE, Hunter DJ: **Polymorphisms of the *AURKA (STK15/Aurora Kinase)* gene and breast cancer risk (United States).** *Cancer Causes Control* 2006, **17:**81-83.
42. Ewart-Toland A, Dai Q, Gao YT, Nagase H, Dunlop MG, Farrington SM, Barnetson RA, Anton-Culver H, Peel D, Ziogas A, Lin D, Miao X, Sun T, Ostrander EA, Stanford JL, Langlois M, Chan JM, Yuan J, Harris CC, Bowman ED, Clayman GL, Lippman SM, Lee JJ, Zheng W, Balmain A: ***Aurora-A/STK15* T+91A is a general low penetrance cancer susceptibility gene: a meta-analysis of multiple cancer types.** *Carcinogenesis* 2005, **26:**1368-1373.
43. Lo YL, Yu JC, Chen ST, Yang HC, Fann CS, Mau YC, Shen CY: **Breast cancer risk associated with genotypic polymorphism of the mitosis-regulating gene *Aurora-A/STK15/BTAK*.** *Int J Cancer* 2005, **115:**276-283.
44. Sun T, Miao X, Wang J, Tan W, Zhou Y, Yu C, Lin D: **Functional Phe31Ile polymorphism in Aurora A and risk of breast carcinoma.** *Carcinogenesis* 2004, **25:**2225-2230.
45. Ewart-Toland A, Briassouli P, de Koning JP, Mao JH, Yuan J, Chan F, MacCarthy-Morrogh L, Ponder BA, Nagase H, Burn J, Ball S, Almeida M, Linardopoulos S, Balmain A: **Identification of Stk6/STK15 as a candidate low-penetrance tumor-susceptibility gene in mouse and human.** *Nat Genet* 2003, **34:**403-412.
46. Cox A, Dunning AM, Garcia-Closas M, Balasubramanian S, Reed MW, Pooley KA, Scollen S, Baynes C, Ponder BA, Chanock S, Lissowska J, Brinton L, Peplonska B, Southey MC, Hopper JL, McCredie MR, Giles GG, Fletcher O, Johnson N, dos Santos Silva I, Gibson L, Bojesen SE, Nordestgaard BG, Axelsson CK, Torres D, Hamann U, Justenhoven C, Brauch H, Chang-Claude J, Kropp S, *et al.*: **A common coding variant in CASP8 is associated with breast cancer risk.** *Nat Genet* 2007, **39:**352-358.
47. Song B, Margolin S, Skoglund J, Zhou X, Rantala J, Picelli S, Werelius B, Lindblom A: **TGFBR1(*)6A and Int7G24A variants of transforming growth factor-beta receptor I in Swedish familial and sporadic breast cancer.** *Br J Cancer* 2007, **97:**1175-1179.
48. Cox DG, Penney K, Guo Q, Hankinson SE, Hunter DJ: **TGFB1 and TGFBR1 polymorphisms and breast cancer risk in the Nurses' Health Study.** *BMC Cancer* 2007, **7:**175.
49. Chen T, Jackson CR, Link A, Markey MP, Colligan BM, Douglass LE, Pemberton JO, Deddens JA, Graff JR, Carter JH: **Int7G24A variant of transforming growth factor-beta receptor type I is associated with invasive breast cancer.** *Clin Cancer Res* 2006, **12:**392-397.
50. Kaklamani VG, Baddi L, Liu J, Rosman D, Phukan S, Bradley C, Hegarty C, McDaniel B, Rademaker A, Oddoux C, Ostrer H, Michel LS, Huang H, Chen Y, Ahsan H, Offit K, Pasche B: **Combined genetic assessment of transforming growth factor-beta signaling pathway variants may predict breast cancer risk.** *Cancer Res* 2005, **65:**3454-3461.
51. Kaklamani VG, Hou N, Bian Y, Reich J, Offit K, Michel LS, Rubinstein WS, Rademaker A, Pasche B: **TGFBR1*6A and cancer risk: a meta-analysis of seven case-control studies.** *J Clin Oncol* 2003, **21:**3236-3243.
52. Perou CM, Sorlie T, Eisen MB, Rijn M van de, Jeffrey SS, Rees CA, Pollack JR, Ross DT, Johnsen H, Akslen LA, Fluge O, Pergamenschikov A, Williams C, Zhu SX, Lonning PE, Borresen-Dale AL, Brown PO, Botstein D: **Molecular portraits of human breast tumours.** *Nature* 2000, **406:**747-752.
53. Pasche B: **Recent advances in breast cancer genetics.** *Cancer Treat Res* 2008, **141:**1-10.
54. Vijver MJ van de, He YD, van't Veer LJ, Dai H, Hart AA, Voskuil DW, Schreiber GJ, Peterse JL, Roberts C, Marton MJ, Parrish M, Atsma D, Witteveen A, Glas A, Delahaye L, Velde T van der, Bartelink H, Rodenhuis S, Rutgers ET, Friend SH, Bernards R: **A gene-expression signature as a predictor of survival in breast cancer.** *N Engl J Med* 2002, **347:**1999-2009.
55. Furuta S, Jiang X, Gu B, Cheng E, Chen PL, Lee WH: **Depletion of BRCA1 impairs differentiation but enhances proliferation of mammary epithelial cells.** *Proc Natl Acad Sci USA* 2005, **102:**9176-9181.
56. Walker LC, Waddell N, Ten Haaf A, Grimmond S, Spurdle AB: **Use of expression data and the CGEMS genome-wide breast cancer association study to identify genes that may modify risk in BRCA1/2 mutation carriers.** *Breast Cancer Res Treat* 2007.
57. 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, Sole X, Hernandez P, Lazaro C, Nathanson KL, Weber BL, Cusick ME, Hill DE, Offit K, *et al.*: **Network modeling links breast cancer susceptibility and centrosome dysfunction.** *Nat Genet* 2007, **39:**1338-1349.
58. Chang JT, Nevins JR: **GATHER: a systems approach to interpreting genomic signatures.** *Bioinformatics* 2006, **22:**2926-2933.
59. Wennmalm K, Miller LD, Bergh J: **A gene signature in breast cancer.** *N Engl J Med* 2007, **356:**1887-1888. author reply 1887–1888.
60. Yu JX, Sieuwerts AM, Zhang Y, Martens JW, Smid M, Klijn JG, Wang Y, Foekens JA: **Pathway analysis of gene signatures predicting metastasis of node-negative primary breast cancer.** *BMC Cancer* 2007, **7:**182.
61. Shen R, Ghosh D, Chinnaiyan AM: **Prognostic meta-signature of breast cancer developed by two-stage mixture modeling of microarray data.** *BMC Genomics* 2004, **5:**94.
62. Zhang Z, Chen D, Fenstermacher DA: **Integrated analysis of independent gene expression microarray datasets improves the predictability of breast cancer outcome.** *BMC Genomics* 2007, **8:**331.
63. Vuaroqueaux V, Urban P, Labuhn M, Delorenzi M, Wirapati P, Benz CC, Flury R, Dieterich H, Spyratos F, Eppenberger U, Eppenberger-Castori S: **Low *E2F1* transcript levels are a strong determinant of favorable breast cancer outcome.** *Breast Cancer Res* 2007, **9:**R33.
64. Untergasser G, Steurer M, Zimmermann M, Hermann M, Kern J, Amberger A, Gastl G, Gunsilius E: **The Dickkopf-homolog 3 is**

Int J Cancer 2008, 1539-1547.

65. Guo H, Lin Y, Zhang H, Liu J, Zhang N, Li Y, Kong D, Tang Q, Ma D:

BMC Mol Biol 2007, 110.

66. Renehan AG, Harvie M, Howell A:

Endocr Relat Cancer 2006, 273-278.

67. Bachmann HS, Otterbach F, Callies R, Nuckel H, Bau M, Schmid KW, Siffert W, Kimmig R: ***BCL2* promoter polymorphism (938C>A) is associated with a favorable outcome in lymph node negative invasive breast cancer patients.** *Clin Cancer Res* 2007, **13:**5790-5797.
68. Hassan S, Baccarelli A, Salvucci O, Basik M: **Plasma stromal cell-derived factor-1: host derived marker predictive of distant metastasis in breast cancer.** *Clin Cancer Res* 2008, **14:**446-454.
69. Hsu EL, Chen N, Westbrook A, Wang F, Zhang R, Taylor RT, Hankinson O: **CXCR4 and CXCL12 down-regulation: A novel mechanism for the chemoprotection of 3,3'-diindolylmethane for breast and ovarian cancers.** *Cancer Lett* 2008, **265:**113-123.
70. Wendt MK, Cooper AN, Dwinell MB: **Epigenetic silencing of *CXCL12* increases the metastatic potential of mammary carcinoma cells.** *Oncogene* 2008, **27:**1461-1471.
71. Pupa SM, Argraves WS, Forti S, Casalini P, Berno V, Agresti R, Aiello P, Invernizzi A, Baldassari P, Twal WO, Mortarini R, Anichini A, Menard S: **Immunological and pathobiological roles of fibulin-1 in breast cancer.** *Oncogene* 2004, **23:**2153-2160.
72. Greene LM, Twal WO, Duffy MJ, McDermott EW, Hill AD, O'Higgins NJ, McCann AH, Dervan PA, Argraves WS, Gallagher WM: **Elevated expression and altered processing of fibulin-1 protein in human breast cancer.** *Br J Cancer* 2003, **88:**871-878.
73. Unoki M, Nakamura Y: **Growth-suppressive effects of *BPOZ* and *EGR2*, two genes involved in the *PTEN* signaling pathway.** *Oncogene* 2001, **20:**4457-4465.
74. Carvalho I, Milanezi F, Martins A, Reis RM, Schmitt F: **Overexpression of platelet-derived growth factor receptor alpha in breast cancer is associated with tumour progression.** *Breast Cancer Res* 2005, **7:**R788-795.
75. Lerebours F, Olschwang S, Thuille B, Schmitz A, Fouchet P, Buecher B, Martinet N, Galateau F, Thomas G: **Fine deletion mapping of chromosome 8p in non-small-cell lung carcinoma.** *Int J Cancer* 1999, **81:**854-858.
76. Komiya A, Suzuki H, Ueda T, Aida S, Ito N, Shiraishi T, Yatani R, Emi M, Yasuda K, Shimazaki J, Ito H: ***PRLTS* gene alterations in human prostate cancer.** *Jpn J Cancer Res* 1997, **88:**389-393.
77. Seitz S, Werner S, Fischer J, Nothnagel A, Schlag PM, Scherneck S: **Refined deletion mapping in sporadic breast cancer at chromosomal region 8p12-p21 and association with clinicopathological parameters.** *Eur J Cancer* 2000, **36:**1507-1513.
78. Yaremko ML, Kutza C, Lyzak J, Mick R, Recant WM, Westbrook CA: **Loss of heterozygosity from the short arm of chromosome 8 is associated with invasive behavior in breast cancer.** *Genes Chromosomes Cancer* 1996, **16:**189-195.
79. Rennstam K, Ahlstedt-Soini M, Baldetorp B, Bendahl PO, Borg A, Karhu R, Tanner M, Tirkkonen M, Isola J: **Patterns of chromosomal imbalances defines subgroups of breast cancer with distinct clinical features and prognosis. A study of 305 tumors by comparative genomic hybridization.** *Cancer Res* 2003, **63:**8861-8868.
80. Xu M, Kao MC, Nunez-Iglesias J, Nevins JR, West M, Zhou XJ: **An integrative approach to characterize disease-specific pathways and their coordination: a case study in cancer.** *BMC Genomics* 2008, **9(Suppl 1):**S12.
81. Forbes S, Clements J, Dawson E, Bamford S, Webb T, Dogan A, Flanagan A, Teague J, Wooster R, Futreal PA, Stratton MR: **Cosmic 2005.** *Br J Cancer* 2006, **94:**318-322.
82. Kitano H: **Cancer as a robust system: implications for anticancer therapy.** *Nat Rev Cancer* 2004, **4:**227-235.
83. Frazer KA, Ballinger DG, Cox DR, Hinds DA, Stuve LL, Gibbs RA, Belmont JW, Boudreau A, Hardenbol P, Leal SM, Pasternak S, Wheeler DA, Willis TD, Yu F, Yang H, Zeng C, Gao Y, Hu H, Hu W, Li C, Lin W, Liu S, Pan H, Tang X, Wang J, Wang W, Yu J, Zhang B, Zhang Q, Zhao H, *et al.*: **A second generation human haplotype map of over 3.1 million SNPs.** *Nature* 2007, **449:**851-861.
84. Hanahan D, Weinberg RA: **The hallmarks of cancer.** *Cell* 2000, **100:**57-70.
85. Eswarakumar VP, Lax I, Schlessinger J: **Cellular signaling by fibroblast growth factor receptors.** *Cytokine Growth Factor Rev* 2005, **16:**139-149.
86. Cuevas BD, Winter-Vann AM, Johnson NL, Johnson GL: **MEKK1 controls matrix degradation and tumor cell dissemination during metastasis of polyoma middle-T driven mammary cancer.** *Oncogene* 2006, **25:**4998-5010.
87. Antoniou AC, Spurdle AB, Sinilnikova OM, Healey S, Pooley KA, Schmutzler RK, Versmold B, Engel C, Meindl A, Arnold N, Hofmann W, Sutter C, Niederacher D, Deissler H, Caldes T, Kampjarvi K, Nevanlinna H, Simard J, Beesley J, Chen X, Neuhausen SL, Rebbeck TR, Wagner T, Lynch HT, Isaacs C, Weitzel J, Ganz PA, Daly MB, Tomlinson G, Olopade OI, *et al.*: **Common breast cancer-predisposition alleles are associated with breast cancer risk in BRCA1 and BRCA2 mutation carriers.** *Am J Hum Genet* 2008, **82:**937-948.
88. Kanehisa M, Goto S, Hattori M, Aoki-Kinoshita KF, Itoh M, Kawashima S, Katayama T, Araki M, Hirakawa M: **From genomics to chemical genomics: new developments in KEGG.** *Nucleic Acids Res* 2006, **34:**D354-357.
89. Garraway LA, Widlund HR, Rubin MA, Getz G, Berger AJ, Ramaswamy S, Beroukhim R, Milner DA, Granter SR, Du J, Lee C, Wagner SN, Li C, Golub TR, Rimm DL, Meyerson ML, Fisher DE, Sellers WR: **Integrative genomic analyses identify *MITF* as a lineage survival oncogene amplified in malignant melanoma.** *Nature* 2005, **436:**117-122.

Pre-publication history

The pre-publication history for this paper can be accessed here:

http://www.biomedcentral.com/1755-8794/1/62/prepub

3.2 Exploració de la connexió entre alteracions genètiques germinals i somàtiques en la carcinogènesi de mama

El fet que *Cell Communication* i *Cell Adhesion* eren processos ja coneguts per participar en la neoplàsia epitelial a nivell somàtic [179], va suggerir una connexió de les pertorbacions moleculars germinals amb les posteriors alteracions somàtiques presents en la progressió del càncer. Aquesta possible connexió entre la línia germinal i les alteracions somàtiques podria ser destacada amb la identificació de variants de risc en *FGFR2*, *MAP3K1* i *CDKN2A/B* [53, 55], gens que es troben mutats somàticament en càncer [180–183].

Per avaluar aquesta hipòtesi, es va examinar la distribució en el rànquing obtingut del resultat del GWAS de CGEMS [175] (prèviament corregit per evitar el biaix degut a la llargada dels gens) de grups de gens alterats somàticament coneguts i relacionats amb el pronòstic [102, 178, 184–188], la resposta al tractament [189–191] i, finalment, relacionats amb alteracions genètiques i genòmiques en tumors [95, 192–196]. Per tal d'avaluar les prediccions, es van genotipar els 20 primers SNPs de les *driver kinases* que apareixen en el rànquing del GWAS en un estudi de 880 controls i 1.173 casos de càncer de mama de Polònia.

3. Resum dels resultats

Els principals resultats obtinguts en aquest treball van ser:

1. Es va obtenir una distribució asimètrica de les quinases que contribueixen a la progressió tumoral degut a mutacions somàtiques (*driver kinases*) (GSEA $P_{valor} = 0{,}001$; $P_{valor\ ajustat\ FDR} = 0{,}010$). Això suggereix que variacions genètiques comunes en els *loci* que codifiquen per les *driver kinases* poden influir en el risc a càncer de mama.

2. L'SNP rs3732568 en el receptor 1 de la epinefrina tipus-B (*EPHB1*) es va trobar associat a risc de càncer de mama (OR = 0,79; 95% IC: 0,63 - 0,98; $P_{trend} = 0{,}031$), en la mateixa direcció i similar magnitud que en el GWAS de CGEMS [175].

3. En l'anàlisi d'associació amb el risc en edats primerenques de diagnosi (40 anys), es van trobar dues associacions: rs6852678 (*CDKL2*) en el model recessiu (OR = 0,32; 95% IC: 0,10 - 1,00; $P_{valor} = 0{,}044$) i rs10878640 (*DYRK2*) en el model dominant (OR = 2,39; 95% IC: 1,32 - 4,30; $P_{valor} = 0{,}003$). A més, degut a les possibles diferències en funció del nivell d'expressió dels receptors d'estrogen, es van examinar les associacions en els pacients ER-positius i en els ER-negatius. Es va veure que rs3732568 (*EPHB1*) presenta un efecte similar en els dos tipus de tumors i que rs12765929 (*BMPR1A*) i rs9836340 (*EPHA3*) mostren més efecte en el risc de tumors ER-negatius mentre que rs4707795 (*EPHA7*) presenta un efecte diferent segons el tipus de tumor.

4. Es va detectar infraexpressió de *EPHB1* en la hiperplàsia ductal atípica respecte al teixit mamari normal en l'anàlisi de dades transcriptòmiques en la progressió del càncer de mama [197].

En resum, l'anàlisi dels grups de gens en el rànquing del GWAS i la posterior replicació indicaria que les variants comunes en determinats *loci* de les *driver kinases*, particularment en els gens que codifiquen per receptors de EPHs, podria influir en el risc de càncer de mama.

OPEN ACCESS Freely available online

PLoS one

Exploring the Link between Germline and Somatic Genetic Alterations in Breast Carcinogenesis

Núria Bonifaci[1], Bohdan Górski[2], Bartlomiej Masojć[2], Dominika Wokołorczyk[2], Anna Jakubowska[2], Tadeusz Dębniak[2], Antoni Berenguer[1], Jordi Serra Musach[1], Joan Brunet[3], Joaquín Dopazo[4], Steven A. Narod[5], Jan Lubiński[2], Conxi Lázaro[6], Cezary Cybulski[2]*, Miguel Angel Pujana[1,7]*

1 Biomarkers and Susceptibility Unit, Spanish Biomedical Research Centre Network for Epidemiology and Public Health, Catalan Institute of Oncology, L'Institut d'Investigació Biomèdica de Bellvitge (IDIBELL), L'Hospitalet, Barcelona, Spain, **2** Department of Genetics and Pathology, International Hereditary Cancer Center, Pomeranian Medical University, Szczecin, Poland, **3** Hereditary Cancer Programme, Catalan Institute of Oncology, IdIBGi, Girona, Spain, **4** Department of Bioinformatics and Genomics, Centro de Investigación Príncipe Felipe, Functional Genomics Node and Spanish Biomedical Research Centre Network for Rare Diseases, Valencia, Spain, **5** Womens College Research Institute, University of Toronto and Women's College Hospital, Toronto, Ontario, Canada, **6** Hereditary Cancer Programme, Catalan Institute of Oncology, IDIBELL, L'Hospitalet, Barcelona, Spain, **7** Translational Research Laboratory, Catalan Institute of Oncology, IDIBELL, L'Hospitalet, Barcelona, Spain

Abstract

Recent genome-wide association studies (GWASs) have identified candidate genes contributing to cancer risk through low-penetrance mutations. Many of these genes were unexpected and, intriguingly, included well-known players in carcinogenesis at the somatic level. To assess the hypothesis of a germline-somatic link in carcinogenesis, we evaluated the distribution of somatic gene labels within the ordered results of a breast cancer risk GWAS. This analysis suggested frequent influence on risk of genetic variation in loci encoding for "driver kinases" (i.e., kinases encoded by genes that showed higher somatic mutation rates than expected by chance and, therefore, whose deregulation may contribute to cancer development and/or progression). Assessment of these predictions using a population-based case-control study in Poland replicated the association for rs3732568 in *EPHB1* (odds ratio (OR) = 0.79; 95% confidence interval (CI): 0.63–0.98; $P_{trend} = 0.031$). Analyses by early age at diagnosis and by estrogen receptor α (ERα) tumor status indicated potential associations for rs6852678 in *CDKL2* (OR = 0.32, 95% CI: 0.10–1.00; $P_{recessive} = 0.044$) and rs10878640 in *DYRK2* (OR = 2.39, 95% CI: 1.32–4.30; $P_{dominant} = 0.003$), and for rs12765929, rs9836340, rs4707795 in *BMPR1A*, *EPHA3* and *EPHA7*, respectively (ERα tumor status $P_{interaction} < 0.05$). The identification of three novel candidates as *EPH receptor* genes might indicate a link between perturbed compartmentalization of early neoplastic lesions and breast cancer risk and progression. Together, these data may lay the foundations for replication in additional populations and could potentially increase our knowledge of the underlying molecular mechanisms of breast carcinogenesis.

Citation: Bonifaci N, Górski B, Masojć B, Wokołorczyk D, Jakubowska A, et al. (2010) Exploring the Link between Germline and Somatic Genetic Alterations in Breast Carcinogenesis. PLoS ONE 5(11): e14078. doi:10.1371/journal.pone.0014078

Editor: Kelvin Yuen Kwong Chan, The University of Hong Kong, China

Received July 2, 2010; **Accepted** November 2, 2010; **Published** November 22, 2010

Funding: The work was supported by the Spanish Society of Medical Oncology 2009 grant to JB and MAP, the Ramón Areces Foundation XV and FIS (09/02483) grants, and the Biomedical Research Centre Network for Epidemiology and Public Health group 55 to MAP, and the "Roses contra el Càncer" Foundation and RTICCC RD06/0020/0028 grants to JB. MAP was supported by the "Ramón y Cajal" Young Investigator program of the Spanish Ministry of Science and Innovation and NB by an IDIBELL fellowship. 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: cezarycy@sci.pam.szczecin.pl (CC); mapujana@iconcologia.net (MAP)

Introduction

With the advent of technical and methodological advances, several GWASs identifying common genetic variation associated with risk of developing cancer have been completed recently [1]. Thus, initiatives such as the National Cancer Institute's Cancer Genetic Markers of Susceptibility (CGEMS) and efforts carried out by deCODE Genetics and the Breast Cancer Association Consortium have led to the identification of breast cancer risk alleles in single nucleotide polymorphisms (SNPs) replicated across populations [2–6]. Intriguingly, illustrating the unbiased nature of GWASs, most hits have corresponded to *a priori* unexpected candidate genes. In this context, the involvement of biological processes beyond the canonical DNA damage response in breast cancer is further suggested by the observed differential influence of low-penetrance risk alleles among *BRCA1* and *BRCA2* mutation carriers [7–9].

A potential common characteristic of the unexpected low-penetrance susceptibility genes is the previously identified contribution to tumorigenesis, but at the somatic level. Common genetic variation in loci encoding for *FGFR2* and *MAP3K1* influences risk of breast cancer [2,4], and these genes were previously found to be somatically mutated in diverse neoplasias including breast cancer [10,11]. In addition, and central to the understanding of cancer progression, common risk alleles showed differential influence according to ERα tumor status [12], and variation in the locus encoding for ERα, *ESR1*, also influences risk of breast cancer [13,14]. More recently, additional breast cancer susceptibility loci have been described that include *CDKN2A/B* as candidates [15]. While these observations suggest a "germline-somatic" link in breast carcinogenesis, an analogous situation may exist for other neoplasias. Variation in loci encoding for *CDH1* and *SMAD7* influences risk of colorectal cancer [16,17] and, similarly, these genes were previously identified as inactivated or deregulated

in tumors [18–21]. Moreover, deregulated germline expression of a paradigmatic proto-oncogene, *MYC*, may be a common mechanism of tumorigenesis in epithelial tissues [22–25]. However, despite some evidence of a germline-somatic link, as yet there is no explicit evaluation of this hypothesis and its potential usefulness in replication studies. Here we present an examination of this link through analysis of the CGEMS GWAS breast cancer dataset and subsequent assessment of the predictions in a case-control study of incident breast cancer in Poland.

Results

Distribution of somatic gene sets in ordered breast cancer GWAS results

Previously, analysis of the CGEMS GWAS dataset using the lowest genotypic *P* value per gene locus suggested true associations in genes annotated with Gene Ontology (GO) biological process terms linked to somatic events [26,27]. However, since there is a positive correlation between the extension of a given locus and the number of SNPs it may contain (and, therefore, the possibility of significant association results being obtained by chance), an unadjusted GWAS rank is biased at its lowest *P* values for specific processes in which large gene products frequently participate [26,28,29] (Fig. 1*A*). Nevertheless, cancer genes tend to expand across large genomic regions [30], and examination of eight genes likely involved in breast cancer through low-penetrance mutations–*CASP8*, *COX11*, *ESR1*, *FGFR2*, *LSP1*, *MAP3K1*, *RAD51L1* and *TOX3* [2–6,13,14]–showed a trend for larger genomic loci (mean ($\overline{x}$) genomic extension = 211 kilo bases (kb) and standard deviation (σ) = 283 kb; compared to $\overline{x}$ = 66 kb and σ = 128 kb for all annotated genes in the CGEMS GWAS rank).

Having identified caveats to the ranking of GWAS results, we performed 10,000 permutations of case-control status and used the null distribution of *t* statistics from the age-adjusted partial correlation analysis to correct the original rank, which then showed an unbiased distribution (Fig. 1*B*). Prior to the evaluation of somatic sets, analysis of GO biological process terms in the GWAS permutation *P* values rank did not show any significant asymmetry using the Gene Set Enrichment Analysis (GSEA) tool [31] with multiple testing correction by the false discovery rate (FDR) approach [32]. Nonetheless, most processes with nominally significant *P* values were those previously highlighted, which are associated with somatic events [26,27] (Table S1). This observation appears to agree with recently described results of pathway-based analysis of the same GWAS dataset [33].

Next, evaluation of somatic sets related to cancer prognosis and treatment response prediction, and to genetic and genomic alterations (see Materials and Methods), revealed significant asymmetrical distribution of "driver kinases" [34,35]; that is, kinases whose deregulation through frequent somatic mutation contributes to tumor development and/or progression ("driver mutations"). In contrast, "passenger mutations" were defined as essentially neutral and linked to the inherent genetic instability in cancer cells [34,35]. Thus, the driver kinases set was found to be biased towards the top (nominal significant association results) of the GWAS permutation rank (GSEA nominal $P<0.001$; FDR-adjusted $P = 0.010$) (Fig. 1*C* and Table S2). Among the remaining of somatic sets evaluated, only cooperation response genes (CRGs) to oncogenic mutations [36] showed a trend for a distribution similar to that of driver kinases (GSEA nominal $P = 0.080$; FDR-adjusted *P* value = 0.25) (Fig. 1*D*), although the intersection between both sets only contained two genes (Table S2). Therefore, in somatic cancer genes, common genetic variation in driver kinase loci might frequently influence risk of breast cancer.

The set of driver kinases contained a benchmark gene, *FGFR2* [2,4], and a locus recently replicated in an independent study, *BMPR1B* [37]. Nevertheless, a significant bias was still observed following exclusion of these two loci (GSEA nominal $P = 0.001$; FDR-adjusted $P = 0.048$), which suggests that variation at additional driver kinase loci influences risk of breast cancer. Importantly, using the set of non-driver kinases–either the subsequent equivalent set as originally statistically ordered or the total set ($n = 344$) [35]–did not reveal significant bias (GSEA nominal $P = 0.99$ and 0.66, respectively), which reinforces the idea of frequent involvement of driver kinases. However, if only the individual statistical data for each locus were considered, most of the driver kinase loci would perhaps not have been selected for replication in other populations.

Independent association results for common variation in driver kinase loci

Given the possible bias in GWAS rank identified above, we examined the top 20 driver kinase variants in the original rank (Table S3, including details of the CGEMS and results below) in a case-control study of incident breast cancer in Szczecin (Poland), previously used in other replications [38]. Applying genotyping quality controls and Hardy-Weinberg equilibrium analysis, 16 SNPs representing an identical number of driver kinase loci (i.e., a single SNP for each locus and representing the strongest potential statistical association) were examined for their association with risk of breast cancer using 880 controls and 1,173 cases (see Materials and Methods). In this analysis, the rs3732568 variant in the *ephrin type-B receptor 1* (*EPHB1*) locus was found to be associated with risk of breast cancer: OR = 0.79, 95% CI: 0.63–0.98; $P_{trend} = 0.031$ (Table 1). Further evaluation of this association through 10,000 case-control permutations in our study gave a similar significance value, $P_{trend} = 0.034$. Importantly, this association was in the same direction and with similar magnitude to the result in the CGEMS GWAS: age-adjusted OR = 0.78, 95% CI: 0.64–0.94; $P_{trend} = 0.009$.

While deregulated expression or function of EPHs and EPH receptors is thought to play a critical role in the initial stages of epithelial neoplasia [39,40], recent analysis of early breast cancer expression changes suggests a link between disruption of cell adhesion and extracellular matrix pathways, and the risk of developing breast cancer [41]. Analysis of this recent dataset also revealed an early expression change of *EPHB1*, between normal breast tissue and atypical ductal hyperplasia (Fig. 2). This alteration consisted of infra-expression in hyperplasia, akin to its potential role in the compartmentalization of early neoplastic lesions [42]. Together, association studies, early expression changes in carcinogenesis and the regulation of cell adhesion suggest the involvement of *EPHB1* in risk of breast cancer.

Next, given accepted models of inherited breast cancer susceptibility [43], we examined associations with risk at early age of diagnosis (≤40 years old). This analysis indicated two additional potential associations: rs6852678 in *CDKL2*, recessive model OR = 0.32, 95% CI: 0.10–1.00; $P = 0.044$; and rs10878640 in *DYRK2*, dominant model OR = 2.39, 95% CI: 1.32–4.30; $P = 0.003$ (Table 2). Results for rs6852678 appeared to be consistent with CGEMS GWAS analysis; age-adjusted recessive model OR = 0.71, 95% CI: 0.53–0.95; $P = 0.019$; however, the pattern for rs10878640 might be more complex (CGEMS GWAS ORs = 1.05 and 0.68 for heterozygotes and minor allele homozygotes, respectively).

Having potential differences by ERα tumor status, we next examined associations in ERα-positive and -negative breast cancer patients. Thus, rs3732568 in *EPHB1* showed a similar influence on

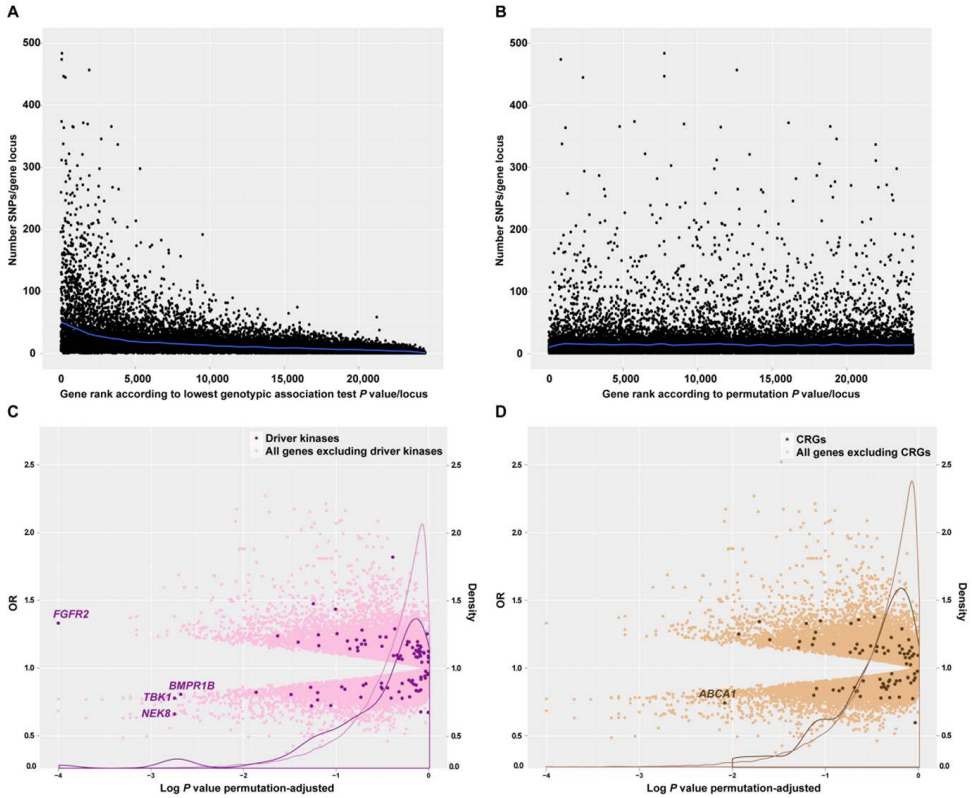


Figure 1. GWAS ranks and distribution of cancer somatic gene sets. *A*, Original GWAS results ranked according to the lowest genotypic association test *P* value per gene locus (unadjusted for genomic extension; taken SNPs in defined genomic window of ±10 kb relative to the first and last exons of a given gene). The Y-axis indicates the number of SNPs per gene locus while the X-axis indicates the lowest association *P* value per gene locus. Bias can be appreciated as the number of SNPs per gene locus increases at lower *P* values. *B*, GWAS results ranked according to the lowest association *P* value per gene locus but adjusted by genomic extension through case-control permutations. Compared to the previous graph, the bias largely disappears. *C*, Following the rank in *B*, the Y-axis indicates odds ratios (ORs) of allele effects and density distributions of gene sets (driver kinases correspond to a light lilac curve; the rest of the genome in the GWAS dataset is shown by a dark lilac curve), while the X-axis indicates the log-transformed association *P* values, previously adjusted by genomic extension. As indicated by the density curves, SNPs mapping to driver kinase loci are relatively more frequent at lower association adjusted *P* values. This observation is supported by GSEA results using the same CGEMS GWAS adjusted rank; nominal $P<0.001$ and FDR-adjusted $P=0.010$ (Table S2). *D*, Similarly to the graph in *C*, distribution of CRGs in the CGEMS GWAS rank adjusted through permutations.
doi:10.1371/journal.pone.0014078.g001

either type of breast cancer (Table 3)–which is consistent with an overall significant association–and rs12765929 in *BMPR1A* and rs9836340 in *EPHA3* showed a potential major impact on the risk of ERα-negative breast cancer (*P* for difference in OR (interaction) by ERα status <0.05), while rs4707795 in *EPHA7* showed a differential effect between ERα-negative versus ERα-positive breast cancer risk ($P_{interaction}=0.007$) (Table 3). None of these additional candidates linked to ERα tumor status, or those linked to an early age of diagnosis above, showed significant expression differences at early stages of breast carcinogenesis as *EPHB1*. On the other hand, the remaining SNPs examined in this study after applying quality controls and Hardy-Weinberg equilibrium analysis (i.e., 10 out of 16), did not show significant associations following CGEMS evidence (Table S3). Together, the gene-set based analysis of GWAS data and the subsequent replication attempt might indicate that common genetic variation in specific driver kinase loci, and particularly in *EPH receptor* genes, influence risk of breast cancer.

Discussion

Evaluation of a germline-somatic link in breast carcinogenesis suggests a role for driver kinases and, perhaps to a lesser extent, genes with a synergistic response to oncogenic mutations. This study might be limited by the assignment of the lowest genotypic *P* value per gene locus within a defined genomic window (i.e.,

Table 1. Association between genetic variation in *EPHB1* and risk of breast cancer in Poland.

***EPHB1*, rs3732568**

	Controls		Cases			
	n	%	*n*	%	OR	95% CI
C/C	693	79.8	891	83.2	1.00	
C/A	165	19.0	172	16.1	0.79	0.62–1.00
A/A	10	1.2	8	0.7	0.60	0.23–1.55
Total	868		1,071			
Trend					0.79	0.63–0.98
					$P_{trend} = 0.031^{\dagger}$	

[†]Adjusted by age.

doi:10.1371/journal.pone.0014078.t001

±10 kb)–thus excluding a large proportion of variation that cannot be assigned to a specific known gene–and by its focus on the additive model of influence of risk alleles when adjusted through case-control permutations. Future analyses taking into account the potential perturbation of germline gene expression by, for example, common variation at distant regulatory regions may improve the identification of susceptibility genes using GWAS complete data. Another limitation in the interpretation of the results presented here may lie in the case-control study designs: the CGEMS addressed breast cancer risk in postmenopausal women, while the Polish study was relatively enriched in early-onset cases. Therefore, studies in additional populations, with diverse designs, are warranted to corroborate the results shown here.

The results of the replication study may be consistent with previously detected somatic genetic alterations and/or functional roles. Somatic mutations in *CDKL2* were nonsense and were only detected in breast and ovarian cancer cell lines or tumors [11,35]. CDKL2 (also known as p56 or KKIAMRE) is the most distant member of the CDC2-related serine/threonine protein kinase family, involved in epidermal growth factor signaling [44], but with a mostly uncharacterized function. *DYRK2* was found to be mutated in breast and central nervous system tumors, in nonsense and missense alterations, respectively [11,35]. The functional role of DYRK2 in the DNA damage response [45] may link to CGEMS GWAS results for *RAD51L1* [3]: loss of DYRK2 function alters the activation of apoptosis in response to DNA damage via ATM [45], which may therefore promote carcinogenesis.

Having revealed potential associations linked to known somatic alterations, the most striking results of this study may concern the identification of risk alleles at three *EPH receptor* loci. EPH-mediated signaling regulates important biological process altered in carcinogenesis, such as cell-to-cell communication, and cell migration and adhesion via the actin cytoskeleton [39,40]. Thus, through RHO and RAS/MAPK activities [46], this signaling pathway has been implicated in the maintenance of epithelial tissue architectures and is therefore thought to act as a tumor suppressor [39,40]. These observations may indicate that, similarly to colorectal tumorigenesis [42], EPH-mediated compartmentalization of early breast tissue neoplastic lesions is critical to prevent the subsequent emergence of carcinoma. Therefore, through a germline expression or functional perturbation, *EPHB1* may contribute to the observed variability in the transition from an *in situ* lesion to an invasive carcinoma [47]. While the associations revealed here warrant further replication in other populations, the existing data could potentially increase current knowledge of the genetic basis and molecular mechanisms of breast carcinogenesis.

Materials and Methods

CGEMS dataset

The National Cancer Institute CGEMS initiative has conducted genome-wide association studies to identify common genetic variants and the corresponding functionally affected genes involved in breast cancer and prostate cancer susceptibility. An initial CGEMS whole genome scan was designed to study the main effect of SNPs on breast cancer risk in postmenopausal women [2]. The study involved 1,145 invasive postmenopausal breast cancer cases and 1,142 matched controls from the Nurses' Health Study nested case-control study [48]. Results of the CGEMS GWAS of breast cancer were obtained upon approval of a Data Access Request.

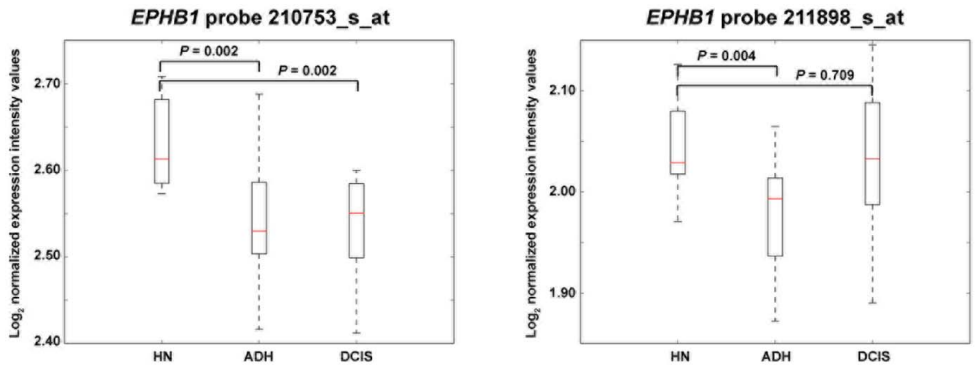


Figure 2. Early change of *EPHB1* expression in breast carcinogenesis. The graphs show expression profiles in histologically normal (HN) breast tissues versus patient-matched atypical ductal hyperplasia (ADH) and ductal carcinoma *in situ* (DCIS) [41]. Results of two *EPHB1* microarray probes (names shown at the top) and the corresponding significance *P* values are shown.

doi:10.1371/journal.pone.0014078.g002

Table 2. Associations between genetic variation in driver kinase loci and risk of breast cancer at ≤40 years of first age at diagnosis.

***CDKL2*, rs6852678**						
	Controls		**Cases**			
	n	**%**	***n***	**%**	**OR**	**95% CI**
C/C	39	51.3	62	51.2	1.00	
C/T	28	36.8	54	44.6	1.21	0.66–2.23
T/T	9	11.8	5	4.1	0.35	0.11–1.12
Total	76		121			
Recessive					0.32	0.10–1.00
					$P_{recessive} = 0.044$	
***DYRK2*, rs10878640**						
	Controls		**Cases**			
	n	**%**	***n***	**%**	**OR**	**95% CI**
G/G	42	56.8	44	35.5	1.00	
G/T	24	32.4	66	53.2	2.62	1.40–4.93
T/T	8	10.8	14	11.3	1.67	0.64–4.39
Total	74		124			
Dominant					2.39	1.32–4.30
					$P_{dominant} = 0.003$	

doi:10.1371/journal.pone.0014078.t002

GWAS rank

In our previous analyses [26,27], ordered CGEMS GWAS results (i.e., ranks) corresponded to the lowest *P* value per gene for the genotypic test in a genomic region of +/−10 kb at each gene locus, defined by the Ensembl human genome release 57. Assigned SNPs were curated using Ensembl gene annotations. We [26] and others [28] noted that such ranks were biased along with the genomic extension–and therefore with the number of SNPs–per gene locus. To adjust for this bias, several statistical strategies are possible [28], including carrying out permutations of the case-control status to correct the significance of the original statistic. In our analysis, considering typed and informative SNPs in each gene locus, we first chose the maximum absolute value of the *t* statistic from the age-adjusted partial correlation in the additive model. Next, 10,000 permutations of the same informative SNPs were performed to create a null distribution for this maximum *t* statistic, which was used to assess its significance corrected by number of SNPs.

GSEA application

The distribution of gene sets in ranked GWAS results was examined using the non-parametric algorithm in the GSEA tool, with default values for all parameters [31] except for the set size when appropriated. In GSEA, a pre-defined gene set is mapped to a rank–in our case genes/loci ordered according to the adjusted association statistic–to assess potential bias using an enrichment score that reflects the degree to which this set is overrepresented at the extremes of the entire ranked list. In the interpretation of the results, caution should be taken when considering sets of different size. In our study, different hypotheses were examined independently (i.e., gene sets linked to prognosis, prediction or genetic/genomic somatic alterations), and *P* values were corrected for multiple testing within each group : 1) genes whose expression in primary breast tumors was associated with patient prognosis and/or metastasis [49–55]; 2) genes whose expression in primary breast tumors was associated with patient therapeutic treatment response [56–59]; 3) genes whose expression levels differed according to ERα breast tumor status or grade [60], or in response to 17β-estradiol [61]; and 4) genes with somatic genetic and/or genomic somatic alterations (Table S2). This last group was made up of five sets : i/ driver kinases (conditional probability of containing driver mutations >0.70, $n = 119$ as defined previously [35], of which 95 were uniquely mapped in the GWAS rank); ii/ CRGs to oncogenic mutations [36]; iii/ cancer gene census, somatically-mutated only [62,63]; iv/ genes affected by somatic chromosomal rearrangements and/or fusions [64]; and v/ amplified and over-expressed cancer genes [65] (Table S2).

Gene expression analysis

Raw expression microarray data on breast cancer progression [41] were downloaded from the Gene Expression Omnibus reference GSE16873 and normalized with robust multiarray average (RMA) [66] and significance analysis was performed using the significance analysis of microarray (SAM) algorithm [67].

Study samples in Poland and association study

A case-control study of unselected invasive breast cancer collected between 1996 and 2003 in Szczecin (Poland) was analyzed. The series included 976 cases of breast cancer unselected for age and an additional group of 367 cases of breast cancer diagnosed at age 50 or below. Therefore, the series was enriched for early-onset cases: mean age of diagnosis was 52.4 years (range 19–88). Subjects were unselected for family history and 15% of cases reported a first- or second-degree relative with breast cancer. The participation rate exceeded 70% among women with breast cancer invited to enroll. Collected information included year of birth, age at diagnosis of breast and/or ovarian cancer, tumor bilaterality, family history (first- and second-degree relatives with breast and/or ovarian cancer) and tumor pathological features in >80% of cases (ERα and progesterone receptor status, and grade). Cases were also examined for *BRCA1* founder mutations in Poland [68] and, if positive,

Table 3. Associations of genetic variation in driver kinase loci and risk of breast cancer by ERα tumor status†.

***BMPR1A*, rs12765929**

	Controls		ERα-negative				ERα-positive			
	n	%	*n*	%	OR	95% CI	*n*	%	OR	95% CI
G/G	514	59.1	189	64.5	1.00		389	58.4	1.00	
G/T	306	35.2	96	32.8	0.87	0.65–1.16	243	36.5	1.07	0.86–1.33
T/T	50	5.7	8	2.7	0.45	0.21–0.98	34	5.1	0.93	0.59–1.48
Total	870		293				666			
Trend					0.79	0.62–1.00			1.02	0.86–1.21
					$P_{trend} = 0.050$				$P_{trend} = 0.81$ $P_{interaction} = 0.024$	

***EPHB1*, rs3732568**

	Controls		ERα-negative				ERα-positive			
	n	%	*n*	%	OR	95% CI	*n*	%	OR	95% CI
C/C	693	79.8	242	82.6	1.00		563	84.9	1.00	
C/A	165	19.0	49	16.7	0.81	0.57–1.16	94	14.2	0.68	0.51–0.90
A/A	10	1.2	2	0.7	0.55	0.12–2.56	6	0.9	0.72	0.26–2.00
Total	868		293				663			
Trend					0.80	0.58–1.11			0.71	0.55–0.91
					$P_{trend} = 0.18$				$P_{trend} = 0.007$ $P_{interaction} = 0.56$	

***EPHA3*, rs9836340**

	Controls		ERα-negative				ERα-positive			
	n	%	*n*	%	OR	95% CI	*n*	%	OR	95% CI
A/A	446	51.3	154	52.4	1.00		356	53.7	1.00	
A/G	341	39.2	99	33.7	0.84	0.63–1.13	251	37.9	0.91	0.74–1.14
G/G	82	9.5	41	13.9	1.43	0.93–2.19	56	8.4	0.85	0.58–1.22
Total	869		294				663			
Recessive					1.53	1.02–2.31			0.88	0.61–1.26
					$P_{recessive} = 0.040$				$P_{recessive} = 0.48$ $P_{interaction} = 0.010$	

***EPHA7*, rs4707795**

	Controls		ERα-negative				ERα-positive			
	n	%	*n*	%	OR	95% CI	*n*	%	OR	95% CI
G/G	618	71.0	204	69.6	1.00		479	71.9	1.00	
G/A	239	27.5	87	29.7	1.18	0.88–1.60	166	24.9	0.92	0.73–1.17
A/A	13	1.5	2	0.7	0.45	0.10–2.06	21	3.2	2.11	1.04–4.28
Total	870		293				666			
Recessive					0.43	0.10–1.96			2.15	1.06–4.37
					$P_{recessive} = 0.28$				$P_{recessive} = 0.034$ $P_{interaction} = 0.007$	

†Adjusted by age.

doi:10.1371/journal.pone.0014078.t003

excluded from the association study ($n = 50$). The control group included cancer-free adult women from the same population (920 women with mean age of diagnosis of 56.7, range 20–91) taken from the healthy adult patients of five family doctors practicing in the Szczecin region. These individuals were selected randomly from the patient lists of the participating doctors. The study was carried out with informed consent of the probands and approved by local ethics committees. Genotypes were obtained using Sequenom iPLEX chemistry at the International Hereditary Cancer Center. Quality controls were of >95% calling for each SNP and >90% of calls per sample. Thus, in the set of 16 SNPs, we observed an average concordance rate of 98.7% of genotype calls using 3.3% replicates. Genotypes of 880 controls and 1,173 cases were effectively analyzed using conditional and unconditional logistic regressions (age adjustment using similar strata size; 20–46, 46–56, 56–66, and 66–91 years old).

Supporting Information

Table S1
Found at: doi:10.1371/journal.pone.0014078.s001 (0.02 MB XLS)

Table S2
Found at: doi:10.1371/journal.pone.0014078.s002 (0.05 MB XLS)

Table S3
Found at: doi:10.1371/journal.pone.0014078.s003 (0.03 MB XLS)

Acknowledgments

The authors are indebted to the CGEMS initiative and to Dr. L.A. Emery and colleagues for making their data available. We also wish to thank the study participants in Poland for their generous contribution.

Author Contributions

Conceived and designed the experiments: CL CC MAP. Performed the experiments: NB BG BM DW CC. Analyzed the data: NB BG BM DW TD. Contributed reagents/materials/analysis tools: NB BG BM DW AJ TD AB JSM JB JD SN JL CL CC. Wrote the paper: MAP.

References

1. Easton DF, Eeles RA (2008) Genome-wide association studies in cancer. Hum Mol Genet 17: R109–115.
2. Hunter DJ, Kraft P, Jacobs KB, Cox DG, Yeager M, et al. (2007) A genome-wide association study identifies alleles in *FGFR2* associated with risk of sporadic postmenopausal breast cancer. Nat Genet 39: 870–874.
3. Thomas G, Jacobs KB, Kraft P, Yeager M, Wacholder S, et al. (2009) A multistage genome-wide association study in breast cancer identifies two new risk alleles at 1p11.2 and 14q24.1 (*RAD51L1*). Nat Genet 41: 579–584.
4. Easton DF, Pooley KA, Dunning AM, Pharoah PD, Thompson D, et al. (2007) Genome-wide association study identifies novel breast cancer susceptibility loci. Nature 447: 1087–1093.
5. Stacey SN, Manolescu A, Sulem P, Thorlacius S, Gudjonsson SA, et al. (2008) Common variants on chromosome 5p12 confer susceptibility to estrogen receptor-positive breast cancer. Nat Genet 40: 703–706.
6. Ahmed S, Thomas G, Ghoussaini M, Healey CS, Humphreys MK, et al. (2009) Newly discovered breast cancer susceptibility loci on 3p24 and 17q23.2. Nat Genet 41: 585–590.
7. Antoniou AC, Sinilnikova OM, McGuffog L, Healey S, Nevanlinna H, et al. (2009) Common variants in *LSP1*, 2q35 and 8q24 and breast cancer risk for *BRCA1* and *BRCA2* mutation carriers. Hum Mol Genet 18: 4442–4456.
8. Antoniou AC, Spurdle AB, Sinilnikova OM, Healey S, Pooley KA, et al. (2008) Common breast cancer-predisposition alleles are associated with breast cancer risk in *BRCA1* and *BRCA2* mutation carriers. Am J Hum Genet 82: 937–948.
9. Antoniou AC, Sinilnikova OM, Simard J, Leone M, Dumont M, et al. (2007) *RAD51* 135G→C modifies breast cancer risk among *BRCA2* mutation carriers: results from a combined analysis of 19 studies. Am J Hum Genet 81: 1186–1200.
10. Hansen RM, Goriely A, Wall SA, Roberts IS, Wilkie AO (2005) Fibroblast growth factor receptor 2, gain-of-function mutations, and tumourigenesis: investigating a potential link. J Pathol 207: 27–31.
11. Stephens P, Edkins S, Davies H, Greenman C, Cox C, et al. (2005) A screen of the complete protein kinase gene family identifies diverse patterns of somatic mutations in human breast cancer. Nat Genet 37: 590–592.
12. Garcia-Closas M, Chanock S (2008) Genetic susceptibility loci for breast cancer by estrogen receptor status. Clin Cancer Res 14: 8000–8009.
13. Zheng W, Long J, Gao YT, Li C, Zheng Y, et al. (2009) Genome-wide association study identifies a new breast cancer susceptibility locus at 6q25.1. Nat Genet 41: 324–328.
14. Dunning AM, Healey CS, Baynes C, Maia AT, Scollen S, et al. (2009) Association of *ESR1* gene tagging SNPs with breast cancer risk. Hum Mol Genet 18: 1131–1139.
15. Turnbull C, Ahmed S, Morrison J, Pernet D, Renwick A, et al. (2010) Genome-wide association study identifies five new breast cancer susceptibility loci. Nat Genet 42: 504–507.
16. Houlston RS, Webb E, Broderick P, Pittman AM, Di Bernardo MC, et al. (2008) Meta-analysis of genome-wide association data identifies four new susceptibility loci for colorectal cancer. Nat Genet 40: 1426–1435.
17. Broderick P, Carvajal-Carmona L, Pittman AM, Webb E, Howarth K, et al. (2007) A genome-wide association study shows that common alleles of *SMAD7* influence colorectal cancer risk. Nat Genet 39: 1315–1317.
18. Levy L, Hill CS (2006) Alterations in components of the TGF-beta superfamily signaling pathways in human cancer. Cytokine Growth Factor Rev 17: 41–58.
19. Wheeler JM, Kim HC, Efstathiou JA, Ilyas M, Mortensen NJ, et al. (2001) Hypermethylation of the promoter region of the E-cadherin gene (*CDH1*) in sporadic and ulcerative colitis associated colorectal cancer. Gut 48: 367–371.
20. Guilford P, Hopkins J, Harraway J, McLeod M, McLeod N, et al. (1998) E-cadherin germline mutations in familial gastric cancer. Nature 392: 402–405.
21. Richards FM, McKee SA, Rajpar MH, Cole TR, Evans DG, et al. (1999) Germline E-cadherin gene (*CDH1*) mutations predispose to familial gastric cancer and colorectal cancer. Hum Mol Genet 8: 607–610.
22. Pomerantz MM, Ahmadiyeh N, Jia L, Herman P, Verzi MP, et al. (2009) The 8q24 cancer risk variant rs6983267 shows long-range interaction with *MYC* in colorectal cancer. Nat Genet 41: 882–884.
23. Tuupanen S, Turunen M, Lehtonen R, Hallikas O, Vanharanta S, et al. (2009) The common colorectal cancer predisposition SNP rs6983267 at chromosome 8q24 confers potential to enhanced Wnt signaling. Nat Genet 41: 885–890.
24. Solé X, Hernández P, de Heredia ML, Armengol L, Rodríguez-Santiago B, et al. (2008) Genetic and genomic analysis modeling of germline *c-MYC* overexpression and cancer susceptibility. BMC Genomics 9: 12.
25. Sotelo J, Esposito D, Duhagon MA, Banfield K, Mehalko J, et al. (2010) Long-range enhancers on 8q24 regulate c-Myc. Proc Natl Acad Sci U S A 107: 3001–3005.
26. Bonifaci N, Berenguer A, Díez J, Reina O, Medina I, et al. (2008) Biological processes, properties and molecular wiring diagrams of candidate low-penetrance breast cancer susceptibility genes. BMC Med Genomics 1: 62.
27. Medina I, Montaner D, Bonifaci N, Pujana MA, Carbonell J, et al. (2009) Gene set-based analysis of polymorphisms: finding pathways or biological processes associated to traits in genome-wide association studies. Nucleic Acids Res 37: W340–344.
28. Kraft P, Raychaudhuri S (2009) Complex diseases, complex genes: keeping pathways on the right track. Epidemiology 20: 508–511.
29. Stanley SM, Bailey TL, Mattick JS (2006) GONOME: measuring correlations between GO terms and genomic positions. BMC Bioinformatics 7: 94.
30. Furney SJ, Higgins DG, Ouzounis CA, López-Bigas N (2006) Structural and functional properties of genes involved in human cancer. BMC Genomics 7: 3.
31. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, et al. (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A 102: 15545–15550.
32. Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. J Roy Statist Soc Ser B 57: 289–300.
33. Menashe I, Maeder D, Garcia-Closas M, Figueroa JD, Bhattacharjee S, et al. (2010) Pathway analysis of breast cancer genome-wide association study highlights three pathways and one canonical signaling cascade. Cancer Res 70: 4453–4459.
34. Stratton MR, Campbell PJ, Futreal PA (2009) The cancer genome. Nature 458: 719–724.
35. Greenman C, Stephens P, Smith R, Dalgliesh GL, Hunter C, et al. (2007) Patterns of somatic mutation in human cancer genomes. Nature 446: 153–158.
36. McMurray HR, Sampson ER, Compitello G, Kinsey C, Newman L, et al. (2008) Synergistic response to oncogenic mutations defines gene class critical to cancer phenotype. Nature 453: 1112–1116.
37. Saetrom P, Biesinger J, Li SM, Smith D, Thomas LF, et al. (2009) A risk variant in an miR-125b binding site in *BMPR1B* is associated with breast cancer pathogenesis. Cancer Res 69: 7459–7465.
38. Wokolorczyk D, Gliniewicz B, Sikorski A, Zlowocka E, Masojc B, et al. (2008) A range of cancers is associated with the rs6983267 marker on chromosome 8. Cancer Res 68: 9982–9986.
39. Merlos-Suárez A, Batlle E (2008) Eph-ephrin signalling in adult tissues and cancer. Curr Opin Cell Biol 20: 194–200.
40. Vaught D, Brantley-Sieders DM, Chen J (2008) Eph receptors in breast cancer: roles in tumor promotion and tumor suppression. Breast Cancer Res 10: 217.
41. Emery LA, Tripathi A, King C, Kavanah M, Mendez J, et al. (2009) Early dysregulation of cell adhesion and extracellular matrix pathways in breast cancer progression. Am J Pathol 175: 1292–1302.
42. Cortina C, Palomo-Ponce S, Iglesias M, Fernández-Masip JL, Vivancos A, et al. (2007) EphB-ephrin-B interactions suppress colorectal cancer progression by compartmentalizing tumor cells. Nat Genet 39: 1376–1383.
43. Claus EB, Risch NJ, Thompson WD (1990) Using age of onset to distinguish between subforms of breast cancer. Ann Hum Genet 54: 169–177.
44. Taglienti CA, Wysk M, Davis RJ (1996) Molecular cloning of the epidermal growth factor-stimulated protein kinase p56 KKIAMRE. Oncogene 13: 2563–2574.
45. Taira N, Yamamoto H, Yamaguchi T, Miki Y, Yoshida K (2010) ATM augments nuclear stabilization of DYRK2 by inhibiting MDM2 in the apoptotic response to DNA damage. J Biol Chem 285: 4909–4919.
46. Brantley-Sieders DM, Zhuang G, Hicks D, Fang WB, Hwang Y, et al. (2008) The receptor tyrosine kinase EphA2 promotes mammary adenocarcinoma tumorigenesis and metastatic progression in mice by amplifying ErbB2 signaling. J Clin Invest 118: 64–78.
47. Schnitt SJ (2009) The transition from ductal carcinoma in situ to invasive breast cancer: the other side of the coin. Breast Cancer Res 11: 101.

48. Colditz GA, Hankinson SE (2005) The Nurses' Health Study: lifestyle and health among women. Nat Rev Cancer 5: 388–396.
49. 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.
50. Chi JT, Wang Z, Nuyten DS, Rodriguez EH, Schaner ME, et al. (2006) Gene expression programs in response to hypoxia: cell type specificity and prognostic significance in human cancers. PLoS Med 3: e47.
51. 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.
52. 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.
53. Liu R, Wang X, Chen GY, Dalerba P, Gurney A, et al. (2007) The prognostic role of a gene signature from tumorigenic breast-cancer cells. N Engl J Med 356: 217–226.
54. 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.
55. Ramaswamy S, Ross KN, Lander ES, Golub TR (2003) A molecular signature of metastasis in primary solid tumors. Nat Genet 33: 49–54.
56. Ayers M, Symmans WF, Stec J, Damokosh AI, Clark E, et al. (2004) Gene expression profiles predict complete pathologic response to neoadjuvant paclitaxel and fluorouracil, doxorubicin, and cyclophosphamide chemotherapy in breast cancer. J Clin Oncol 22: 2284–2293.
57. 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.
58. Ma XJ, Wang Z, Ryan PD, Isakoff SJ, Barmettler A, et al. (2004) A two-gene expression ratio predicts clinical outcome in breast cancer patients treated with tamoxifen. Cancer Cell 5: 607–616.
59. 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.
60. van de Vijver MJ, He YD, van't Veer LJ, Dai H, Hart AA, et al. (2002) A gene-expression signature as a predictor of survival in breast cancer. N Engl J Med 347: 1999–2009.
61. 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.
62. Futreal PA, Coin L, Marshall M, Down T, Hubbard T, et al. (2004) A census of human cancer genes. Nat Rev Cancer 4: 177–183.
63. Forbes SA, Tang G, Bindal N, Bamford S, Dawson E, et al. (2010) COSMIC (the Catalogue of Somatic Mutations in Cancer): a resource to investigate acquired mutations in human cancer. Nucleic Acids Res 38: D652–657.
64. Stephens PJ, McBride DJ, Lin ML, Varela I, Pleasance ED, et al. (2009) Complex landscapes of somatic rearrangement in human breast cancer genomes. Nature 462: 1005–1010.
65. Santarius T, Shipley J, Brewer D, Stratton MR, Cooper CS (2010) A census of amplified and overexpressed human cancer genes. Nat Rev Cancer 10: 59–64.
66. Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, et al. (2003) Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics 4: 249–264.
67. Tusher VG, Tibshirani R, Chu G (2001) Significance analysis of microarrays applied to the ionizing radiation response. Proc Natl Acad Sci U S A 98: 5116–5121.
68. Gorski B, Byrski T, Huzarski T, Jakubowska A, Menkiszak J, et al. (2000) Founder mutations in the *BRCA1* gene in Polish families with breast-ovarian cancer. Am J Hum Genet 66: 1963–1968.

3.3 Anàlisi de l'associació entre variants genètiques en els *loci* de les *driver kinases* i el risc a càncer en els portadors de mutacions en *BRCA1* i *BRCA2*.

Els resultats del treball anterior van proposar sis *loci* (*driver kinases*) de predisposició al càncer de mama en funció de les associacions observades en població general polonesa [198]. Tenint en compte que els al·lels de baixa penetrància poden actuar com a modificadors de les mutacions d'alta penetrància [199], a continuació es van avaluar les associacions entre les variants en aquests *loci* i el risc a càncer de mama en portadors de mutacions en *BRCA1* o *BRCA2*.

Amb aquest objectiu es van genotipar 95 SNPs dels sis *loci* que codifiquen per les *driver kinases* i es va analitzar l'associació amb el risc a càncer de mama en 15.252 portadors de mutacions en *BRCA1* i 8.211 portadors de mutacions en *BRCA2*. A més, es van analitzar 2.000 variants imputades en el *locus* de *EPHB1*. Finalment, per avaluar la possible alteració de la expressió d'*EPHB1* en tumors de mama, es van realitzar anàlsis immunohistoquímics i es van analitzar dades d'expressió i metilació gènica per detectar la presència de *loci* que provoquen canvis en la expressió i/o metilació dels gens (eQTLs i meQTLs) en el *locus EPHB1*.

Els principals resultats obtinguts en aquest treball van ser:

1. La variant rs7074064 (*BMPR1A*) va presentar associació al risc de càncer de mama en portadors de mutacions en *BRCA1* (hazard ratio (HR) = 1,06; 95% IC: 1,01 - 1,11; P_{valor} = 0,019) i també en portadors de mutacions en *BRCA2* però en la direcció oposada a l'anterior (HR = 0,93; 95% IC: 0,87 - 0,99; P_{valor} = 0,037).

   Nou variants en *EPHB1* van presentar associació al risc de càncer de mama en portadors de mutacions en *BRCA1* . La variant amb l'associació més forta va ser rs4309752 (HR = 1,06; 95% IC: 1,01 - 1,10; P_{valor} = 0,015). Altres variants en *EPHB1* però no lligades a les anteriors, van presentar associació al risc de càncer de mama en portadors de mutacions en *BRCA2*, l'associació més forta va ser per rs16842235 ($r^2 \sim 0$ amb rs4309752) (HR = 1,16; 95% IC: 1,06 - 1,27; P_{valor} = 0,003)

2. Dos SNPs van presentar evidències per a una possible associació amb els casos ERα-negatius: rs7074064 (*BMPR1A*) en portadors de mutacions en *BRCA1* (HR = 1,06; 95% IC: 1,00 - 1,12; P_{valor} = 0,045) i rs16842235 (*EPHB1*) en portadors de mutacions en *BRCA2* (HR = 1,34; 95% IC: 1,09 - 1,66; p = 0,006). No obstant, aquestes estimacions de HR no van ser significativament diferents de les obtingudes amb els casos ERα-positius ($p_{diferencia} > 0{,}15$).

3. Diferents variants de les 2.000 imputades en el *locus* d'*EPHB1* van presentar associació a risc de càncer de mama tant en portadors de mutacions en *BRCA1* com en portadors de mutacions en *BRCA2*. Els resultats més significatius es van obtenir per la variant rs182738811 en

portadors de mutacions en *BRCA1* ((HR) = 1,44; P_{valor} = 7,2 x 10^{-4}) i per dues variants independents (r^2 = 0.04) rs9843661 (HRs = 0,88; P_{valor} = 2,2 x 10^{-4}) i rs115984427 (HRs = 0,78; P_{valor} = 3,6 x 10^{-4}) en portadors de mutacions en *BRCA2*. Les associacions observades en portadors de mutacions en *BRCA1* es troben en diferents blocs de desequilibri de lligament de les associacions observades en portadors de mutacions en *BRCA2*.

4. Els anàlisis immunohistoquímics van mostrar una menor expressió d'*EPHB1* en tots els tumors en relació al teixit mamari normal.

5. L'anàlisi de dades del consorci TCGA (de l'anglès *The Cancer Genome Atlas*) [200] va identificar la variant rs16842235 com un possible meQTL, localitzat a $<$ 2 kb de l'exò 5' de *EPHB1*. Per tant, l'al·lel rs16842235-A podria estar associat a un increment del risc a càncer de mama mitjançant la hipermetilació de *EPHB1*. Aquesta observació és coherent amb la pèrdua d'*EPHB1* en carcinomes invasius tot i que són necessaris més anàlisis per examinar les diferencies d'expressió gènica esperades.

Aquests resultats suggereixen que variants en el *locus EPHB1* podrien estar associades amb el risc a càncer de mama en portadors de mutacions en *BRCA1* i *BRCA2*. Donat que la senyalització per epinefrina regula el desenvolupament de les glàndules mamàries i que la expressió d'EPHB1 caracteritza les cèl·lules embrionàries mamaries, aquest estudi suggeriria la relació entre una alteració de la funció o els nivells d'EPHB1 degut a una disminució de la diferenciació epitelial.

Nota: Aquest estudi serà sotmés a reavaluació en funció dels resultats derivats d'un nou anàlisi COGS col·laboratiu [60] amb la participació del CIMBA [67]. Esperem rebre noves dades del COGS a finals del 2015.

Evaluating associations between genetic variants at cancer driver kinase loci and cancer risk in *BRCA1/2* mutation carriers

Karoline Kuchenbaecker[1], Núria Bonifaci[2], Daniel Cuadras[3], Nadia García[4], Paolo Peterlongo[5], Paolo Radice[5], Daniel Barrowdale[1], Lesley McGuffog[1], Jordi Serra-Musach[2], Gorka Ruiz de Garibay[2], Antonio Gómez[6], Manel Esteller[6-8], Orland Díez[9], Judith Balmaña[10], Adriana Lasa[11], Teresa Ramón y Cajal[11], María-Dolores Miramar[12], Miguel de la Hoya[13], Trinidad Caldés[13], Pedro Pérez-Segura[14], Ana Osorio[15], Javier Benítez[15], Miquel Bertran[16], Àngel Izquierdo[16], Alex Teulé[17], Lídia Feliubadaló[17], Esther Darder[18], Joan Brunet[18], CIMBA, Ignacio Blanco[17], Conxi Lázaro[17], Georgia Chenevix-Trench[19], Antonis C. Antoniou[1,*] and Miguel Angel Pujana[2,*]

[1]Centre for Cancer Genetic Epidemiology, Department of Public Health and Primary Care, University of Cambridge, Cambridge CB1 8RN, UK, [2]Breast Cancer and Systems Biology Unit, Catalan Institute of Oncology (ICO), Bellvitge Institute for Biomedical research (IDIBELL), L'Hospitalet del Llobregat, Barcelona, Catalonia 08908, Spain, [3]Statistics Unit, IDIBELL, L'Hospitalet del Llobregat, Barcelona, Catalonia 08908, Spain, [4]Translational Research Laboratory, ICO, IDIBELL, L'Hospitalet del Llobregat, Barcelona, Catalonia 08908, Spain, [5]Unit of Molecular Bases of Genetic Risk and Genetic Testing, Department of Preventive and Predictive Medicine, Fondazione IRCCS Istituto Nazionale Tumori (INT), Milan, and IFOM Fondazione Istituto FIRC di Oncologia Molecolare, Milan 20133, Italy, [6]Cancer Epigenetics and Biology Program (PEBC), IDIBELL, L'Hospitalet del Llobregat, Barcelona, Catalonia 08909, Spain, [7]Department of Physiological Sciences II, School of Medicine, University of Barcelona, Barcelona, Catalonia 08908, Spain, [8]Catalan Institution for Research and Advanced Studies (ICREA), Barcelona, Catalonia 08010, Spain, [9]Oncogenetics Laboratory, Vall d'Hebron Institute of Oncology (VHIO), University Hospital Vall d'Hebron, Barcelona, Catalonia 08035, Spain, [10]Department of Medical Oncology, VHIO, University Hospital Vall d'Hebron, Barcelona, Catalonia 08035, Spain, [11]Departments of Medical Oncology and Genetics, Hospital Santa Creu i Sant Pau, Barcelona, Catalonia 08025, Spain, [12]Genetics Unit, Department of Biochemistry, University Hospital Miguel Servet, Zaragoza 50009, Spain, [13]Molecular Oncology Laboratory, Hospital Clínico San Carlos, Instituto de Investigación Sanitaria del Hospital San Carlos (IdISSC), Madrid 28040, Spain, [14]Medical Oncology Service, Hospital Clínico San Carlos, Madrid 28040, Spain, [15]Human Cancer Genetics Program and Biomedical Research Center Network for Rare Diseases (CIBERER), Spanish National Cancer Research Centre (CNIO), Madrid 28029, Spain, [16]Department of Medical Oncology, ICO, Hospital Josep Trueta, Girona Biomedical Research Institute (IDIBGi), Girona, Catalonia 17007, Spain, [17]Hereditary Cancer Program, ICO, IDIBELL, L'Hospitalet del Llobregat, Barcelona, Catalonia 08908, Spain, [18]Hereditary Cancer Program, ICO, Hospital Josep Trueta, IDIBGi, Girona, Catalonia 17007, Spain, [19]Queensland Institute of Medical Research, Queensland 4029, Australia.

To whom correspondence should be addressed.
Tel: +44-1223-740163; Fax: +44-1223-740159
Email: aca20@medschl.cam.ac.uk
Correspondence may also be addressed to Miguel Angel Pujana.
Tel: +34-932607463; Fax: +34-932607466
Email: mapujana@iconcologia.net

Abstract

Background: Genes frequently mutated in tumors may also play a role in cancer risk through low penetrance germline mutations. Previous studies have suggested that this "germline-somatic link" is relevant for loci encoding for cancer driver kinases. Following this observation, we assessed the associations between genetic variants at six cancer driver kinase loci and breast cancer risk in *BRCA1* and *BRCA2* mutation carriers.

Methods: Ninety-five genotyped single nucleotide polymorphisms were analyzed for associations with breast cancer risk in 15,252 *BRCA1* and 8,211 *BRCA2* mutation carriers using a retrospective likelihood approach. Two thousand imputed variants at the ephrin type-B receptor 1 (*EPHB1*) gene locus were also analyzed. Immunohistochemical analyses were carried out to assess the alteration of EPHB1 expression and/or cellular localization in tumors. Gene expression and genomic methylation datasets were analyzed to examine the existence of expression and/or methylation quantitative trait loci (eQTLs and meQTLs, respectively) at the *EPHB1* locus.

Results: Among the 95 genotyped variants, there were suggestions of associations for variants in *EPHB1* and breast cancer risk in both settings: rs4309752 in *BRCA1* mutation carriers, hazard ratio (HR) = 1.05, 95% confidence interval (CI) 1.01 – 1.10, p = 0.015; rs16842235 (r^2 < 0.2 with rs4309752) in *BRCA2* mutation carriers, HR = 1.17, 95% confidence interval (CI) 1.06 – 1.29, p = 0.003. The analysis of imputed variants suggested independent (pairwise r^2 < 0.2) association signals: rs182738811 in *BRCA1* mutation carriers (hazard ratio (HR) = 1.44, p = 7.2 x 10^{-4}); and rs9843661 and rs115984427 in *BRCA2* mutation carriers (HRs = 0.88 and 0.78, p = 2.2 x 10^{-4} and 3.6 x 10^{-4}, respectively). Loss of EPHB1 expression was detected in the transition from in situ to invasive carcinomas, and rs16842235 may represent a meQTL for *EPHB1*.

Conclusions: Independent variants in the *EPHB1* locus might be associated with breast cancer risk in *BRCA1* and *BRCA2* mutation carriers. Since ephrin signaling regulates mammary gland development and *EPHB1* expression characterizes mammary stem cells, this study could suggest a link between altered EPHB1 function/levels and carcinogenesis through impaired epithelial differentiation. The results of this study may warrant corroboration in larger series of mutation carriers and case-control studies.

Background

Genome-wide association studies (GWASs) completed in recent years have substantially increased our knowledge of the genetic basis of breast cancer risk [1, 2]. Intriguingly, some of the gene candidates identified in these studies were linked to known molecular alterations at the somatic level that promote carcinogenesis and/or define cancer subtypes: among others, the candidate genes include *CDKN2A/B* [3], which encodes for a tumor suppressor involved in cell cycle regulation [4], *ESR1* [5, 6], which encodes for estrogen receptor α (ERα), and *MYC* [7, 8], which represents a proto-oncogene in many neoplastic conditions [9].

Complementary to analyses directed at identifying significant marginal effects, gene set and ranking-based analyses of GWAS data have suggested a link between germline and somatic molecular alterations. Thus, an excess of association signals was proposed for genes that encode for components of the RAS/RAF/MAPK signaling pathway [10] and for genes that encode for cancer driver kinases (i.e., kinases encoded by genes that show higher somatic mutation rates than expected by chance and, therefore, whose deregulation probably contributes to carcinogenesis) [11]. In the latter study, six gene loci were proposed as candidates, according to the results of an association study in a Polish cohort [11]. On the basis of evidence that low-penetrance alleles may act as modifiers of highly penetrant mutations [12], we have assessed the associations between variants at these cancer driver kinase loci and breast cancer risk in *BRCA1* and/or *BRCA2* (*BRCA1/2*) mutation carriers.

Methods

Study subjects

Sixty-one study centers in the Consortium of Investigators of Modifiers of BRCA1/2 (CIMBA) recruited a total of 15,252 *BRCA1* and 8,211 *BRCA2* mutation carriers to this study (samples selected after quality control process). Most of these individuals were recruited through cancer genetics clinics and enrolled into national or regional studies. The remaining carriers were identified by population-based sampling or community recruitment. Eligibility for inclusion in CIMBA was restricted to female carriers of pathogenic *BRCA1* and *BRCA2* mutations who were 18 years or older at recruitment. Information collected included year of birth, mutation description, self-reported ethnic ancestry, age at last follow-up, age at breast or ovarian cancer diagnosis, and age at bilateral prophylactic mastectomy and oophorectomy. Information about tumor characteristics, including estrogen receptor α (ERα) status, was also collected for 3,458 *BRCA1* and 1,924 *BRCA2* mutation carriers. Related individuals were identified by a unique family identifier.

iCOGS

The design, genotyping and quality controls of the iSelect single nucleotide polymorphism (SNP) array of the Collaborative Oncological Gene-environment Study (iCOGS) have been described recently [13, 14]. The final array design included 211,155 manufactured SNPs, selected primarily on the basis of evidence from GWASs of breast, ovarian and prostate cancer, for fine mapping of known cancer susceptibility loci, and included functional candidate variants of interest [13-17] (also see http://www.nature.com/icogs/primer/cogs-project-and-design-of-the-icogs-array/ and http://ccge.medschl.cam.ac.uk/research/consortia/icogs/). Details of the iCOGS array design have been given elsewhere [13-17]. The genotyping cluster plots for all variants cited in the text were checked manually for quality.

Statistical analyses

The iCOGS array included SNPs in *BMPR1A* (n = 9), *CDKL2* (n = 11), *DYRK2* (n = 9), *EPHA3* (n = 23), *EPHA7* (n = 12) and *EPHB1* (n = 39) loci (defined as ± 20 kilobases (kb) from the genomic structure of each gene), which were analyzed in the present study (**Additional file 1: Table S1**). The associations were further assessed with the imputed

genotypes at *EPHB1* and *EPHA7* using data from the 1,000 Genomes project (March 2012 version [18]). The main analyses focused on evaluating associations between each genotype and breast cancer or ovarian cancer risk separately, in a survival analysis framework. In the breast cancer analysis, the phenotype of each individual was defined by age at breast cancer diagnosis or age at last follow-up. Individuals were monitored until the age of the first breast or ovarian cancer diagnosis, or bilateral prophylactic mastectomy, whichever occurred first, or until age at last observation. Mutation carriers censored at ovarian cancer diagnosis were considered to be unaffected. For the ovarian cancer analysis, the primary endpoint was the age at ovarian cancer diagnosis, and mutation carriers were monitored until the age of ovarian cancer diagnosis, risk-reducing salpingo-oophorectomy or age at last observation. In order to maximize the number of ovarian cancer cases, breast cancer was not considered to be a censoring event in this analysis, and mutation carriers who developed ovarian cancer after breast cancer diagnosis were considered to be affected in the ovarian cancer analysis. To adjust for the non-random sampling of mutation carriers with respect to their disease status, data were analyzed by modeling the retrospective likelihood of the observed genotypes conditional on the disease phenotypes [19]. The associations were assessed using the 1-degree of freedom score test statistic based on this retrospective likelihood. To allow for the non-independence among related individuals, we took into account the correlation between the genotypes using a kinship-adjusted version of the score test statistic [20]. The p values presented are based on the adjusted score test. To estimate the HRs, the effect of each SNP was modeled as either a per-allele or genotype on the log-scale by maximizing the retrospective likelihood. We also evaluated the evidence of heterogeneity in the associations between countries/study centers. Associations with breast and ovarian cancer risks were assessed simultaneously within a competing risk analysis framework [13, 19]. The *BRCA1* mutation classes assessed were: mutations expected to result in a reduced transcript or protein level due to nonsense-mediated RNA decay (class 1); and mutations likely to generate stable proteins with a potential residual or dominant-negative function (class 2). The IMPUTE2 software [21] was used to impute non-genotyped SNPs, with the 1,000 Genomes Phase I integrated variant set v3, March 2012, as the reference panel. The associations of each marker with cancer risk were assessed with a similar score test to that used for the observed SNPs, but based on the posterior genotype probabilities at each imputed marker for each individual. In all analyses, we considered only those SNPs with

an imputation information/accuracy of $r^2 > 0.30$ and a minor allele frequency (MAF) > 0.3%.

GWAS data

The population-based breast cancer GWAS carried out by the CGEMS initiative was designed to identify variants with a significant marginal effect in postmenopausal women [22]. The study involved 1,145 invasive postmenopausal breast cancer cases and 1,142 matched controls from the Nurses' Health Study. The GWAS data were obtained upon approval of a Data Access Request to dbGAP (http://cgems.cancer.gov/data/). Missing genotypes were imputed using the MACH software [23].

Immunohistochemistry

The EPHB1 antibody used in this study was a rabbit polyclonal, catalog number SC-926, from Santa Cruz Biotechnology. The assays were performed on serial paraffin sections (4 μm thick) using the Envision method (Dako). Endogenous peroxidase was blocked by pre-incubation in a solution of 3% H2O2, and blocking was performed in 1X phosphate buffered saline with 5% goat serum and 0.1% Tween 20 (Sigma-Aldrich). Sections were counterstained with hematoxylin and examined with an Olympus BX51 microscope. Each tissue sample and marker was evaluated in at least two independent assays and no substantial intra-tissue differences were observed. For a subset of samples, equivalent sections were processed to include incubation with a non-immune rabbit immunoglobulin control (Sigma-Aldrich), which did not reveal staining in any case. Ten *BRCA2*-mutated and six *BRCA1*-mutated tumor tissues were analyzed in this study.

TCGA data analyses

Data from The Cancer Genome Atlas (TCGA) were downloaded from the corresponding repository (tcga-data.nci.nih.gov/tcga/tcgaDownload) upon approval of a Data Access Request to the database of Genotypes and Phenotypes (dbGaP). The computation of eQTLs and meQTLs was performed as described elsewhere. Briefly, meQTLs were analyzed by integrating TCGA tumor SNP data from the Affymetrix Genome-Wide Human SNP Array 6.0 platform and CpG data from the Infinium HumanMethylation450 platform, and using the multivariate Random Forest Selection Frequency (RFSF) method as previously described [24].

Results

Association study in *BRCA1/2* mutation carriers using genotyped data

This CIMBA study used data generated through the iCOGS array [13, 14]. The analysis was restricted to six loci encoding for cancer driver kinases and previously suggested to be associated with breast cancer risk in the general population [11]. After quality controls, 95 genotyped variants, which represented 29 partially independent SNPs (pairwise $r^2 < 0.80$), were available for analysis. Thus, using a retrospective cohort analytical approach [19], one genotyped variant in *BMPR1A* and nine in *EPHB1* indicated associations with p values < 0.05 and breast cancer risk in *BRCA1* mutation carriers: *BMPR1A* rs7074064, hazard ratio (HR) = 1.06, 95% confidence interval (CI) 1.01 – 1.11, p = 0.019; and the strongest *EPHB1* signal was for rs4309752 HR = 1.06, 95% CI 1.01 – 1.10, p = 0.015. In *BRCA2* mutation carriers, *BMPR1A* rs7074064 also showed a suggestion of association, but in the opposite direction to that observed for *BRCA1* mutation carriers: HR = 0.93, 95% CI 0.87 – 0.99, p = 0.037 (**Additional file 1: Table S1**). Additional variants in *EPHB1* were also suggested to be associated with breast cancer risk in *BRCA2* mutation carriers at p < 0.05, but they were not correlated with the variant that showed some evidence of association for *BRCA1* mutation carriers: the strongest association was for rs16842235 ($r^2 \sim 0$ with rs4309752), HR = 1.16, 95% CI 1.06 – 1.27, p = 0.003 (**Additional file 1: Table S1**). Similar results were obtained in an analysis under a competing risks model that evaluates associations between breast and ovarian cancer risks simultaneously [19]: rs16842235 association with breast cancer risk in *BRCA2* mutation carriers HR = 1.19, 95% CI 1.08 – 1.32, $p = 7.4 \times 10^{-4}$.

Evaluation of associations by tumor ERα status

As genetic modifiers may influence the development of tumors with different molecular characteristics [25, 26], the above observations were assessed by the expression of ERα in tumors. Both *BMPR1A* rs7074064 and *EPHB1* rs16842235 showed evidence of a potential association with ERα-negative status: rs7074064 in *BRCA1* ERα-negative mutation carriers HR = 1.06, 95% CI 1.00 – 1.12, p = 0.045; and rs16842235 in *BRCA2* ERα-negative mutation carriers HR = 1.34, 95% CI 1.09 – 1.66, p = 0.006. Nonetheless,

these HR estimations were not significantly different from the ERα-positive cases ($p_{differences} > 0.15$).

***EPHB1* association study using imputed data**

Following the analysis of the iCOGS genotyped variants, imputed genotypes from the 1,000 Genomes project were used to further evaluate the potential associations at the *EPHB1* locus. This gene locus spans 465 kilo bases (kb) in chromosome 3q22 and shows multiple linkage disequilibrium blocks (> 34 according to HapMap Caucasians data, **Figure 1**); of the genotype variants, 29 had a pairwise $r^2 < 0.8$. A total of 2,000 variants with imputation accuracy $r^2 > 0.30$ were available for analysis, which suggested stronger associations with breast cancer risk for both *BRCA1* and *BRCA2* mutation carriers: rs182738811 (accuracy $r^2 = 0.68$) in *BRCA1* mutation carriers $p = 7.2 \times 10^{-4}$ (**Figure 1A**); and two independent variants, rs9843661 (accuracy $r^2 = 0.51$) and rs115984427 (accuracy $r^2 = 0.64$; pairwise $r^2 = 0.04$), in *BRCA2* mutation carriers $p = 2.2 \times 10^{-4}$ and 3.6×10^{-4}, respectively (**Figure 1B**). The rs9843661 variant is located close to rs16842235, but the two are poorly correlated ($r^2 \sim 0$). In addition, the associations observed in for *BRCA1* and *BRCA2* mutation carriers were in distinct linkage disequilibrium blocks (**Figure 1A,B**). Several additional variants displayed p values < 0.01 (accuracy $r^2 > 0.49$; rs11708725, rs147655817, and rs199779292 for *BRCA1* mutation carriers, **Figure 1A**; and rs59540927 for *BRCA2* mutation carriers, **Figure 1B**).

Potential alteration of EPHB1 expression in breast cancer

The results presented above could suggest that germline genetic alterations in *EPHB1* influence the risk of breast cancer in *BRCA1* and/ or *BRCA2* mutation carriers. Next, to investigate the potential link to carcinogenesis, we performed immunohistochemical analyses in paraffin-embedded tumor tissue from 16 *BRCA1/2* carriers. Relative to normal breast tissue, under-expression of EPHB1 was observed in all tumors (**Figure 2A**). In addition, this alteration was apparent in the transition from an in situ lesion to invasive carcinoma (**Figure 2A**). This observation is akin to the role of EPHB1 in colorectal cancer [27] and its potential tumor suppressor function as a regulator of epithelial tissue architecture [28].

Next, TCGA data [29] were analyzed for the existence of eQTLs and/or meQTLs associated with *EPHB1*. Among the variants described above, only rs16842235 was

represented (including $r^2 > 0.8$) in the TCGA dataset, and this variation was suggested to be associated ($p < 0.05$) with three CpG island probes mapping < 2 kb from the 5'-exon of *EPHB1* (**Figures 2B**). Thus, the rs16842235-A allele suggested to be associated with an increase in breast cancer risk was linked to a relative *EPHB1* hypermethylation. This observation is consistent with the loss of EPHB1 in invasive carcinoma, although additional analyses are necessary to examine the expected gene expression differences. While the TCGA data analysis did not reveal rs16842235 to be an eQTL, a meta-analysis of data from lymphoblastoid cell lines detected genome-wide significance for this variant [30].

Discussion

Since common breast cancer susceptibility alleles identified through association studies in the general population frequently act as modifiers of cancer risk in *BRCA1/2* mutation carriers [12, 31], we examined whether previously suggested associations for six cancer driver kinase loci [11] influence cancer risk in these mutation carriers. Analysis of 95 iCOGS-genotyped variants provided suggestions of associations between *EPHB1* variants and breast cancer risk in both *BRCA1* ($p = 0.015$) and *BRCA2* ($p = 0.003$) mutation carriers. Of the *EPHB1* genotype variants, 29 could be considered independent ($r^2 < 0.8$). Next, analysis of 2,000 imputed variants at this locus indicated potentially stronger associations and independent signals; with a p value threshold < 0.001, one rare variant (rs182738811, MAF = 0.006) might be associated with breast cancer risk in *BRCA1* mutation carriers, and two variants (rs9843661 and rs115984427) with breast cancer risk in *BRCA2* mutation carriers. Additional genotyped and/or imputed variants might be associated with p values < 0.01. However, given the number of statistical analyses performed, these findings require further investigation with larger series of carriers before they can be considered consistent associations. Of particular interest would be rs16842235, which appears to represent an eQTL and/or meQTL. This observation may be relevant if we consider that *EPHB1* expression forms part of a transcriptional program characteristic of mammary stem cells [32, 33] and that ephrin receptor signaling participates in mammary gland development and epithelial differentiation [34, 35].

In the previous study that proposed an excess of association signals for cancer driver kinase loci, the predictions from the analysis of the Cancer Genetic Markers of Susceptibility GWAS [22] were assessed in a Polish cohort by genotyping the strongest GWAS signal in each locus [11]. The CGEMS and Polish study results indicated a consistent nominally significant marginal effect for *EPHB1* rs3732568; however, this variant showed no evidence of association with breast cancer risk for *BRCA1* and/or *BRCA2* mutation carriers in the present study (the p values from the imputation analyses were > 0.15). This apparent discrepancy, together with the potential existence of independent signals according to *BRCA1/2* status, suggests a need for further investigation, including case-control studies.

Conclusion

Candidate genes and the integration of somatic data in the results of breast cancer GWASs have suggested that common genetic variation in cancer driver kinase loci frequently influences cancer risk. Assessment of this hypothesis in *BRCA1/2* mutation carriers suggests that independent variants in the *EPHB1* locus are associated with breast cancer risk in *BRCA1* or *BRCA2* mutation carriers. The expression of EPHB1 appears to be substantially reduced in the transition to an invasive carcinoma in both mutation carriers. The role of ephrin signaling in mammary development and of *EPHB1* in mammary stem cells leads to speculate that altered EPHB1 function/levels could promote carcinogenesis through impaired epithelial differentiation. The results of this study may warrant corroboration in larger series of mutation carriers and case-control studies.

Additional files

Additional file 1: Table S1. Genotyped variants and breast cancer association results in *BRCA1* and *BRCA2* mutation carriers.

Additional file 1: Table S2. Details of acknowledgments and funding support.

Abbreviations

BRCA1: breast cancer 1, early onset gene; *BRCA2*: breast cancer 2, early onset gene; CI: confidence interval; CIMBA: Consortium of Investigators of Modifiers of BRCA1/2; EPHB1: ephrin type-B receptor 1; eQTL: expression quantitative trait locus; ERα: estrogen receptor α; GWAS: genome-wide association study; HR: hazard ratio; iCOGS: iSelect SNP array of the Collaborative Oncological Gene-environment Study; meQTL: methylation quantitative trait locus; SNP: single nucleotide polymorphism; TCGA: The Cancer Genome Atlas.

Competing interests

The authors declare that they have no competing interest.

Acknowledgments

We wish to thank all study participants and clinicians for their valuable contributions to the iCOGS study. We also wish to thank the CGEMS initiative for making available the breast cancer GWAS data. Additional acknowledgments are detailed in **Additional file 2: Table S2**.

Funding

Funding support is detailed in **Additional file 2: Table S2**.

References

1. Varghese JS, Easton DF: **Genome-wide association studies in common cancers--what have we learnt?** *Curr Opin Genet Dev* 2010, **20**(3)**:**201-209.
2. Easton DF, Eeles RA: **Genome-wide association studies in cancer.** *Hum Mol Genet* 2008, **17**(R2)**:**R109-115.
3. Turnbull C, Ahmed S, Morrison J, Pernet D, Renwick A, Maranian M, Seal S, Ghoussaini M, Hines S, Healey CS*, et al*: **Genome-wide association study identifies five new breast cancer susceptibility loci.** *Nat Genet* 2010, **42**(6)**:**504-507.
4. Matheu A, Maraver A, Serrano M: **The Arf/p53 pathway in cancer and aging.** *Cancer Res* 2008, **68**(15)**:**6031-6034.
5. Zheng W, Long J, Gao YT, Li C, Zheng Y, Xiang YB, Wen W, Levy S, Deming SL, Haines JL*, et al*: **Genome-wide association study identifies a new breast cancer susceptibility locus at 6q25.1.** *Nat Genet* 2009, **41**(3)**:**324-328.
6. Dunning AM, Healey CS, Baynes C, Maia AT, Scollen S, Vega A, Rodriguez R, Barbosa-Morais NL, Ponder BA, Low YL*, et al*: **Association of *ESR1* gene tagging SNPs with breast cancer risk.** *Hum Mol Genet* 2009, **18**(6)**:**1131-1139.
7. Pomerantz MM, Ahmadiyeh N, Jia L, Herman P, Verzi MP, Doddapaneni H, Beckwith CA, Chan JA, Hills A, Davis M*, et al*: **The 8q24 cancer risk variant rs6983267 shows long-range interaction with *MYC* in colorectal cancer.** *Nat Genet* 2009, **41**(8)**:**882-884.
8. Easton DF, Pooley KA, Dunning AM, Pharoah PD, Thompson D, Ballinger DG, Struewing JP, Morrison J, Field H, Luben R*, et al*: **Genome-wide association study identifies novel breast cancer susceptibility loci.** *Nature* 2007, **447**(7148)**:**1087-1093.
9. Soucek L, Evan GI: **The ups and downs of Myc biology.** *Curr Opin Genet Dev* 2010, **20**(1)**:**91-95.
10. Menashe I, Maeder D, Garcia-Closas M, Figueroa JD, Bhattacharjee S, Rotunno M, Kraft P, Hunter DJ, Chanock SJ, Rosenberg PS, Chatterjee N: **Pathway analysis of breast cancer genome-wide association study highlights three pathways and one canonical signaling cascade.** *Cancer Res* 2010, **70**(11)**:**4453-4459.
11. Bonifaci N, Gorski B, Masojc B, Wokolorczyk D, Jakubowska A, Debniak T, Berenguer A, Serra Musach J, Brunet J, Dopazo J*, et al*: **Exploring the link between germline and somatic genetic alterations in breast carcinogenesis.** *PLoS One* 2010, **5**(11)**:**e14078.
12. Antoniou AC, Spurdle AB, Sinilnikova OM, Healey S, Pooley KA, Schmutzler RK, Versmold B, Engel C, Meindl A, Arnold N*, et al*: **Common breast cancer-predisposition alleles are associated with breast cancer risk in *BRCA1* and *BRCA2* mutation carriers.** *Am J Hum Genet* 2008, **82**(4)**:**937-948.
13. Couch FJ, Wang X, McGuffog L, Lee A, Olswold C, Kuchenbaecker KB, Soucy P, Fredericksen Z, Barrowdale D, Dennis J*, et al*: **Genome-wide association study in *BRCA1* mutation carriers identifies novel loci associated with breast and ovarian cancer risk.** *PLoS Genet* 2013, **9**(3)**:**e1003212.
14. Gaudet MM, Kuchenbaecker KB, Vijai J, Klein RJ, Kirchhoff T, McGuffog L, Barrowdale D, Dunning AM, Lee A, Dennis J*, et al*: **Identification of a *BRCA2*-specific modifier locus at 6p24 related to breast cancer risk.** *PLoS Genet* 2013, **9**(3)**:**e1003173.

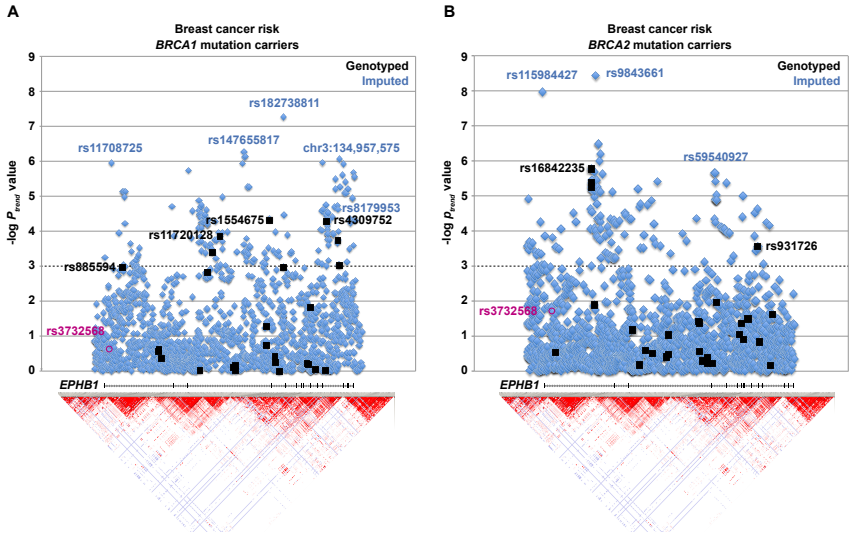
15. Eeles RA, Olama AA, Benlloch S, Saunders EJ, Leongamornlert DA, Tymrakiewicz M, Ghoussaini M, Luccarini C, Dennis J, Jugurnauth-Little S*, et al*: **Identification of 23 new prostate cancer susceptibility loci using the iCOGS custom genotyping array.** *Nat Genet* 2013, **45**(4)**:**385-391, 391e381-382.
16. Pharoah PD, Tsai YY, Ramus SJ, Phelan CM, Goode EL, Lawrenson K, Buckley M, Fridley BL, Tyrer JP, Shen H*, et al*: **GWAS meta-analysis and replication identifies three new susceptibility loci for ovarian cancer.** *Nat Genet* 2013, **45**(4)**:**362-370, 370e361-362.
17. Michailidou K, Hall P, Gonzalez-Neira A, Ghoussaini M, Dennis J, Milne RL, Schmidt MK, Chang-Claude J, Bojesen SE, Bolla MK*, et al*: **Large-scale genotyping identifies 41 new loci associated with breast cancer risk.** *Nat Genet* 2013, **45**(4)**:**353-361, 361e351-352.
18. Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, Handsaker RE, Kang HM, Marth GT, McVean GA: **An integrated map of genetic variation from 1,092 human genomes.** *Nature* 2012, **491**(7422)**:**56-65.
19. Barnes DR, Lee A, Easton DF, Antoniou AC: **Evaluation of association methods for analysing modifiers of disease risk in carriers of high-risk mutations.** *Genet Epidemiol* 2012, **36**(3)**:**274-291.
20. Antoniou AC, Wang X, Fredericksen ZS, McGuffog L, Tarrell R, Sinilnikova OM, Healey S, Morrison J, Kartsonaki C, Lesnick T*, et al*: **A locus on 19p13 modifies risk of breast cancer in *BRCA1* mutation carriers and is associated with hormone receptor-negative breast cancer in the general population.** *Nat Genet* 2010, **42**(10)**:**885-892.
21. Howie B, Marchini J, Stephens M: **Genotype imputation with thousands of genomes.** *G3 (Bethesda)* 2011, **1**(6)**:**457-470.
22. Hunter DJ, Kraft P, Jacobs KB, Cox DG, Yeager M, Hankinson SE, Wacholder S, Wang Z, Welch R, Hutchinson A*, et al*: **A genome-wide association study identifies alleles in *FGFR2* associated with risk of sporadic postmenopausal breast cancer.** *Nat Genet* 2007, **39**(7)**:**870-874.
23. Li Y, Ding J, Abecasis GR: **MACH 1.0: Rapid haplotype reconstruction and missing genotype inference.** *Am J Hum Genet* 2006, **79**S2290.
24. Michaelson JJ, Alberts R, Schughart K, Beyer A: **Data-driven assessment of eQTL mapping methods.** *BMC Genomics* 2010, **11**502.
25. Mulligan AM, Couch FJ, Barrowdale D, Domchek SM, Eccles D, Nevanlinna H, Ramus SJ, Robson M, Sherman M, Spurdle AB*, et al*: **Common breast cancer susceptibility alleles are associated with tumor subtypes in *BRCA1* and *BRCA2* mutation carriers: results from the Consortium of Investigators of Modifiers of BRCA1/2.** *Breast Cancer Res* 2011, **13**(6)**:**R110.
26. Milne RL, Antoniou AC: **Genetic modifiers of cancer risk for *BRCA1* and *BRCA2* mutation carriers.** *Ann Oncol* 2011, **22 Suppl 1**i11-17.
27. Cortina C, Palomo-Ponce S, Iglesias M, Fernández-Masip JL, Vivancos A, Whissell G, Huma M, Peiro N, Gallego L, Jonkheer S*, et al*: **EphB-ephrin-B interactions suppress colorectal cancer progression by compartmentalizing tumor cells.** *Nat Genet* 2007, **39**(11)**:**1376-1383.
28. Vaught D, Brantley-Sieders DM, Chen J: **Eph receptors in breast cancer: roles in tumor promotion and tumor suppression.** *Breast Cancer Res* 2008, **10**(6)**:**217.
29. TCGA: **Comprehensive molecular portraits of human breast tumours.** *Nature* 2012, **490**(7418)**:**61-70.

30. Liang L, Morar N, Dixon AL, Lathrop GM, Abecasis GR, Moffatt MF, Cookson WO: **A cross-platform analysis of 14,177 expression quantitative trait loci derived from lymphoblastoid cell lines.** *Genome Res* 2013, **23**(4)**:**716-726.
31. Barnes DR, Antoniou AC: **Unravelling modifiers of breast and ovarian cancer risk for *BRCA1* and *BRCA2* mutation carriers: update on genetic modifiers.** *J Intern Med* 2012, **271**(4)**:**331-343.
32. Lim E, Wu D, Pal B, Bouras T, Asselin-Labat ML, Vaillant F, Yagita H, Lindeman GJ, Smyth GK, Visvader JE: **Transcriptome analyses of mouse and human mammary cell subpopulations reveal multiple conserved genes and pathways.** *Breast Cancer Res* 2010, **12**(2)**:**R21.
33. Kendrick H, Regan JL, Magnay FA, Grigoriadis A, Mitsopoulos C, Zvelebil M, Smalley MJ: **Transcriptome analysis of mammary epithelial subpopulations identifies novel determinants of lineage commitment and cell fate.** *BMC Genomics* 2008, **9**591.
34. Kaenel P, Mosimann M, Andres AC: **The multifaceted roles of Eph/ephrin signaling in breast cancer.** *Cell Adh Migr* 2012, **6**(2)**:**138-147.
35. Kaenel P, Antonijevic M, Richter S, Kuchler S, Sutter N, Wotzkow C, Strange R, Andres AC: **Deregulated ephrin-B2 signaling in mammary epithelial cells alters the stem cell compartment and interferes with the epithelial differentiation pathway.** *Int J Oncol* 2012, **40**(2)**:**357-369.

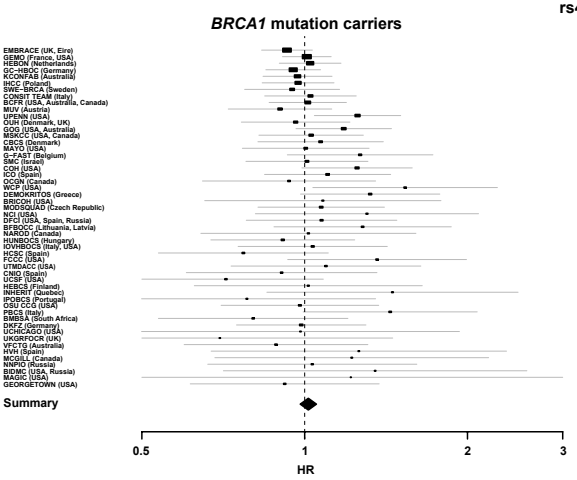
Figure legends

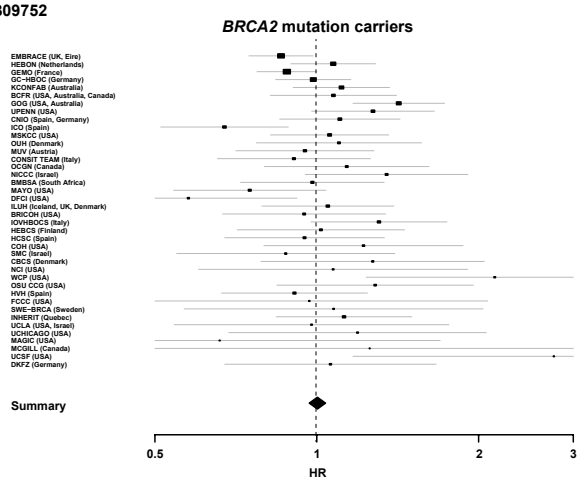
Figure 1 Genetic variation in *EPHB1* and risk of breast cancer in *BRCA1/2* mutation carriers. (**A**) Graph depicting the *EPHB1* association results ($-\log_{10}$ p value) for breast cancer risk in *BRCA1* mutation carriers. Results are shown for the genotyped (black squares) and imputed (blue diamonds) variants. The position and association result for rs3732568 is marked by a pink circle. The horizontal dashed line corresponds to the defined association threshold ($p < 0.001$). The *EPHB1* genomic structure and the linkage disequilibrium pattern from HapMap Caucasian individuals are shown at the bottom. (**B**) *EPHB1* association results for breast cancer risk in *BRCA2* mutation carriers.

Figure 2 Alteration of EPHB1/*EPHB1* in breast carcinogenesis. (**A**) Immunohistochemical results of EPHB1 in normal breast tissue (left panel; the inset shows basal staining in normal acini), ductal in situ lesions (top middle and right panels) and invasive carcinomas (bottom middle and right panels) from *BRCA1* (middle) and *BRCA2* (right) mutation carriers. The scale bar represents 100 μm. (**B**) Box plots depicting the meQTL at rs16842235. The Y-axis and X-axis show the methylation level (β values) and rs16842235 genotypes, respectively.

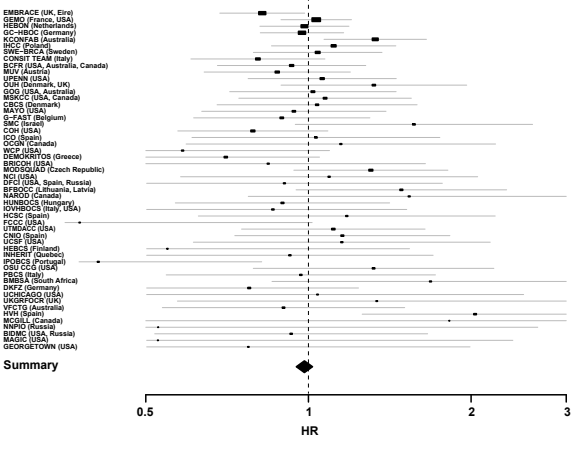

A
Breast cancer risk
BRCA1 mutation carriers
Genotyped
Imputed
-log Pexcl value
rs182738811
rs11708725
rs147655817
chr3:134,957,575
rs8179953
rs1554675
rs4309752
rs11720128
rs885594
rs3732568
EPHB1
B
Breast cancer risk
BRCA2 mutation carriers
Genotyped
Imputed
-log Pexcl value
rs115984427
rs9843661
rs59540927
rs16842235
rs931726
rs3732568
EPHB1

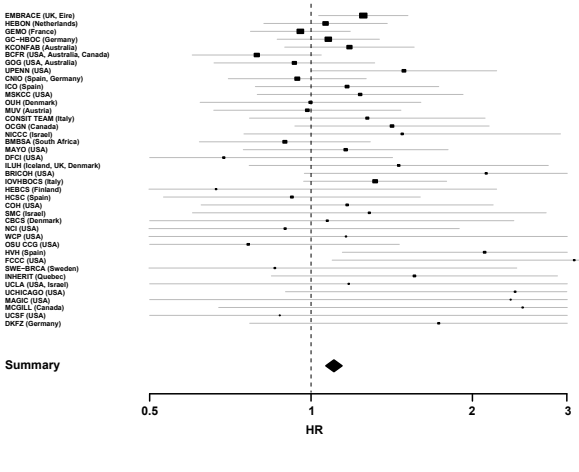
rs4309752





rs16842235





3.4 Integració de dades d'expressió gènica i dades epidemiològiques per a la identificació d'interaccions genètiques associades al risc a càncer

S'han identificat dotzenes de variants genètiques comunes associades al risc a càncer a partir dels GWAS [201]. Tot i això, aquestes variants només expliquen una fracció de la etiologia de la malaltia [21]. La *missing heritability* s'ha atribuït a diferents factors, entre ells l'existència de interaccions genètiques (GxG) [121, 202].

En humans, l'anàlisi de GxG a nivell de tot el genoma s'ha limitat principalment al càlcul entre parelles d'SNPs significatius en els GWAS. Anàlisis més sistemàtics presenten limitacions estadístiques degut al gran nombre de parelles a analitzar [162]. No obstant, el coneixement obtingut a partir dels extensos anàlisis experimentals de les GxG en organismes model ens poden ajudar en aquesta tasca [155]. Estudis en organismes model han demostrat que les GxG donen informació o correlacionen significativament amb altres tipus de relacions moleculars i/o funcionals entre gens i/o proteïnes [156]. Per això, una estratègia integrativa a nivell de tot el genoma pot ajudar a identificar GxG associades al risc de càncer.

En una primera fase, es van calcular les totes les GxG possibles en les dades del GWAS de CGEMS [175] aplicant un algoritme de dos passos [163]. En el primer pas, es va calcular la diferència entre els coeficients de correlació de Pearson (PCCs) entre casos i controls per tots els SNPs localitzats en gens (els intergènics van quedar exclosos al no poder associar-se a proteïnes i funcions). En el segon pas, es va calcular el coeficient d'interacció utilitzant

la regressió logística de les 410.000 parelles d'snps amb P_{valor} inferior a 10^{-5} obtingudes en el pas anterior. Després d'eliminar les parelles en desequilibri de lligament, associar els SNPs al gen més proper en una regió de ± 10 kilobases i seleccionar les interaccions amb P_{valor} inferior a 10^{-6}, vam identificar 39.417 parelles de gens. A continuació, per definir la rellevància d'aquestes GxG predites que estarien associades amb el risc a càncer de mama en la població general, es van realitzar diferents anàlisis; per exemple, estudiar la coincidència amb parelles de gens que es coexpressen en tumors. A partir del rànquing de les parelles de gens en relació al seu valor de coexpressió, es van triar diferents percentatges de parelles començant sempre pel capdamunt de la llista (i.e. més coexpressades). Així es van obtenir diferents intervals (del 0,1% al 30%) amb les parelles més correlacionades i es van comparar (és a dir, determinar el nombre de coincidències) amb prediccions de GxG. Els resultats es van representar en funció de "l'enriquiment relatiu" (RE, de l'anglès *Relative Enrichment*) de les GxG respecte a les parelles de gens coexpressats.

Els principals resultats obtinguts en aquest treball van ser:

1. No es va trobar una coincidència significativa entre les GxG predites i les interaccions proteïna-proteïna de la base de dades *Human Protein Reference Database* (HPRD).

2. Es va trobar un RE significatiu de les GxG predites amb les parelles de gens coexpressades en tumors de mama [203] utilitzant mesures de informació mútua (*Mutual Information*, MI) però no utilitzant com a mesura de coexpressió el coeficient de correlació de Pearson (*Pearson Correlation Coeficient*, PCC). Concretament, en el interval 0,5% de les

parelles de gens amb més MI, van coincidir 205 parelles amb les GxG predites (RE del 16%; $P_{empíric} = 0{,}015$). Al eliminar d'aquest interval les parelles de l'interval 0,5% del rànquing PCC, es va observar un cert increment en l'enriquiment (43 parelles en comú, RE del 43%; $P_{empíric} = 0{,}009$).

3. Els enriquiments en l'interval 0,1% de les parelles amb els valors de MI més alts calculat només pels tumors ERα-positius (RE del 36%; $P_{empíric} = 0{,}041$) van ser superiors al mateix interval de parelles calculat només pels tumors ERα-negatius (RE del -5%; $P_{empíric} = 0{,}56$).

4. No es va trobar solapament significatiu en cap interval del rànquing de MI calculat a partir de les dades d'una altra neoplàsia epitel·lial com és el càncer de colon [204].

5. De les 205 parelles ja mencionades, 173 presentaven anotacions GO en els dos membres. Els processos biològics més freqüents en les GxG predites corresponen a parelles de gens involucrades en metabolisme i biosíntesi.

6. Es va destacar la interacció entre rs2289263 (*SMAD3*) i rs4686980 (*LPP*) per ser *SMAD3* previament un candidat funcional associat al risc de càncer de mama [205]. El producte d'aquest gen actua com a transductor de senyal i regulador de la transcripció *downstream* de *TGFβ1* [206]. Per aquest motiu, es va avaluar la participació en la carcinogènesi d'aquesta interacció mitjançant alteracions cel·lulars a partir de la reducció de l'expressió de *LPP* i modelant la senyal de *TGFβ1* en una línia cel·lular epitelial no tumoral (MCF10A). Els resultats obtinguts van demostrar que la pertorbació simultània de *LPP* i *TGFβ1* produia un augment significatiu de la proliferació cel·lular.

7. Utilitzant una estratègia a l'inrevés de la descrita anteriorment, es van triar parelles de gens per presentar els valors més alts de MI calculats a partir de dades d'expressió gènica de tumors pancreàtics [207] i es va calcular les seves GxG en les dades d'un GWAS de càncer de pàncreas [208, 209]. Així, es va observar un solapament significatiu (més de l'esperat per l'atzar) entre les parelles de gens coexpressats i les GxG predites. Concretament, a partir del rànquing de MI, es van definir 14 finestres de 20 parelles de gens cadascuna (del valor més alt de MI al valor 5.000). A continuació, en cada finestra, es van calcular les GxG utilitzant totes les parelles d'SNPs unlinked ($r^2 < 0{,}2$) i es va avaluar el nombre d'associacions significatives ($P_{\text{Regressió logística}} < 0{,}05$). Aquest anàlisi va revelar més GxG de les esperades per atzar, un 5,99% de mitjana ($P_{\textit{Wilcoxon Rank Test}} = 0{,}003$). El mateix anàlisi però amb les parelles de gens amb menor MI no va resultar significatiu ($P_{\textit{Wilcoxon Rank Test}} = 0{,}35$).

Els resultats obtinguts suggereixen que evidències basades en patrons de coexpressió gènica complexe (i definida en funció del tipus de càncer) poden ser utilitzats per predir GxG associades al risc a càncer.

Carcinogenesis Advance Access published January 2, 2014
Carcinogenesis vol.00 no.00 p.1 of 8, 2013
doi:10.1093/carcin/bgt403
Advance Access publication December 2, 2013

Integrating gene expression and epidemiological data for the discovery of genetic interactions associated with cancer risk

Núria Bonifaci[1], Eva Colas[2], Jordi Serra-Musach[1,3], Nazanin Karbalai[4,5], Joan Brunet[3], Antonio Gómez[6], Manel Esteller[6–8], Enrique Fernández-Taboada[1], Antoni Berenguer[9], Jaume Reventós[2,10], Bertram Müller–Myhsok[4,5,11], Laufey Amundadottir[12], Eric J.Duell[13] and Miquel Àngel Pujana[1,*]

[1]Breast Cancer and Systems Biology Unit, Translational Research Laboratory, Catalan Institute of Oncology (ICO), Bellvitge Institute for Biomedical Research (IDIBELL), L'Hospitalet del Llobregat, Barcelona 08908, Catalonia, Spain, [2]Research Unit in Biomedicine and Translational and Pediatric Oncology, Vall d'Hebron Research Institute and Hospital, and Autonomous University of Barcelona, Barcelona 08035, Catalonia, Spain, [3]Hereditary Cancer Programme, ICO, Girona Biomedical Research Institute (IDIBGI), Girona 17007, Catalonia, Spain, [4]Statistical Genetics, Max Planck Institute of Psychiatry, Munich 80804, Germany, [5]Munich Cluster for Systems Neurology (SyNergy), Munich 80804, Germany, [6]Cancer Epigenetics and Biology Program (PEBC), IDIBELL, L'Hospitalet del Llobregat, Barcelona 08908, Catalonia, Spain, [7]Department of Physiological Sciences II, School of Medicine, University of Barcelona, Barcelona 08908, Catalonia, Spain, [8]Catalan Institution for Research and Advanced Studies (ICREA), Barcelona 08010, Catalonia, Spain, [9]Unit of Biomarkers and Susceptibility, ICO, IDIBELL, L'Hospitalet de Llobregat, Barcelona 08908, Catalonia, Spain, [10]Basic Sciences Department, International University of Catalonia, Barcelona 08017, Catalonia, Spain, [11]Munich Cluster for Systems Neurology (SyNergy), Munich 80336, Germany, [12]Laboratory of Translational Genomics, Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, Bethesda, MD 20892, USA and [13]Unit of Nutrition, Environment and Cancer, Cancer Epidemiology Program, ICO, IDIBELL, L'Hospitalet del Llobregat, Barcelona 08908, Catalonia, Spain

*To whom correspondence should be addressed. Tel: +34 932607463; Fax: +34 932607466;
Email: mapujana@iconcologia.net

Dozens of common genetic variants associated with cancer risk have been identified through genome-wide association studies (GWASs). However, these variants only explain a modest fraction of the heritability of disease. The missing heritability has been attributed to several factors, among them the existence of genetic interactions (G × G). Systematic screens for G × G in model organisms have revealed their fundamental influence in complex phenotypes. In this scenario, G × G overlap significantly with other types of gene and/or protein relationships. Here, by integrating predicted G × G from GWAS data and complex- and context-defined gene coexpression profiles, we provide evidence for G × G associated with cancer risk. G × G predicted from a breast cancer GWAS dataset identified significant overlaps [relative enrichments (REs) of 8–36%, empirical *P* values < 0.05 to 10^{-4}] with complex (non-linear) gene coexpression in breast tumors. The use of gene or protein data not specific for breast cancer did not reveal overlaps. According to the predicted G × G, experimental assays demonstrated functional interplay between lipoma-preferred partner and transforming growth factor-β signaling in the MCF10A non-tumorigenic mammary epithelial cell model. Next, integration of pancreatic tumor gene expression profiles with pancreatic cancer G × G predicted from a GWAS corroborated the observations made for breast cancer risk (REs of 25–59%). The method presented here can potentially support the identification of genetic interactions associated with cancer risk, providing novel mechanistic hypotheses for carcinogenesis.

Abbreviations: ERα, estrogen receptor α; G × G, genetic interaction; GO, Gene Ontology; GWAS, genome-wide association study; LD, linkage disequilibrium; LPP, lipoma-preferred partner; LR, logistic regression; MI, mutual information; PCC, Pearson's correlation coefficient; RE, relative enrichment; shRNA, short hairpin RNA; SNP, single nucleotide polymorphism; TGFβ, transforming growth factor-β; WRT, Wilcoxon rank test.

Introduction

Several genome-wide association studies (GWASs) have been completed that delineate the common genetic basis of cancer risk (1). The gene candidates identified in these studies have considerably expanded the biological knowledge of cancer etiology. These advances are being followed up by projects that aim to identify the corresponding genetic mutations and to improve cancer risk prediction. However, in most cases, the results of GWASs (in addition to complementary candidate approaches) have not yet identified the bulk of disease risk heritability. For example, to date, 79 low-penetrance loci have been identified in breast cancer, but together they account for only a modest percentage (~15%) of the familial relative risk. If moderately and highly penetrant mutations/genes are included, ~50% of the familial relative risk remains unexplained (2,3).

Numerous factors or modeling approaches can explain the problem of 'missing heritability' (4–6). Notably, the recent meta-analysis of several breast cancer GWASs revealed an excess of significant association signals (not reaching genome-wide significance) that suggests that >1000 loci are involved in susceptibility, each of which exerts a very small effect (7). This modeling did not take into account genetic interactions (G × G), which have also been suggested to explain part of the missing heritability (4–6). In this regard, the identification of interactions could potentially improve the accuracy of risk models and improve cancer prevention (8,9). Several methods have been developed for exhaustive searching of statistical interactions in data from GWASs (10,11). These analyses (limited to two locus interactions) are time consuming but computationally achievable. However, the vast number of loci pairs raises the issue of multiple testing, which limits the identification of true interactions based only on statistical terms. In addition, the translation of the statistical findings to biological interactions or models is unclear (12) and potentially complex (13).

Systematic analyses in model organisms have shown that, in many cases, a given phenotype is explained not simply by additive allele effects but also by G × G (or epistasis in statistical terms; i.e. deviation from additivity for a quantitative phenotype by the effect of a genetic variant or mutation in a different locus) (14). Importantly, studies in yeast with ~6000 annotated genes have predicted the existence of hundreds of thousands of G × G (15). It could therefore be hypothesized that G × G are of similar biological relevance in humans (9). Their relevance is also based on the identification of synthetic lethal interactions for specific mutations in cancer (16). However, the methodology to systematically screen for mammalian G × G has only recently been described (17,18). On the basis of previous evidence that G × G inform about other types of molecular or functional relationships between genes and/or proteins (genes/proteins), we hypothesize that a genome-wide integrative strategy could help to discover G × G associated with cancer risk.

Materials and methods

Genetic data and G × G analysis

The National Cancer Institute has conducted GWASs to identify common genetic variants and the corresponding candidate genes associated with cancer risk, which included breast (19) and pancreatic (20,21) cancer. For breast cancer, the initial GWAS by the Cancer Genetic Markers of Susceptibility initiative was designed to identify variants with a significant marginal effect in postmenopausal women. The study involved 1145 invasive postmenopausal breast cancer cases and 1142 matched controls from the Nurses' Health Study. The GWAS data was obtained upon approval of a Data Access Request to dbGAP (http://cgems.cancer.gov/data/). Missing genotypes were imputed using the MACH software (22). The GWAS data for pancreatic cancer was also obtained upon approval of a Data Access Request to dbGAP and analyzed for specific variants (i.e. variants selected by their identifier and/or location in

a specific gene locus, rather than randomly selected from a given gene rank), gene pair bins (bins defined from gene pairs ranked according to their complex coexpression in tumors; each bin corresponds to 20 gene pairs, starting from the highest coexpression value), and G × G using the dbGAP provided PLINK (http://pngu.mgh.harvard.edu/purcell/plink/) (23) file. This file was used to prevent potential differences in reprocessing the original GWAS data and contained 914 cases and 1027 controls. The whole genome screen for breast cancer G × G was carried out using EPIBLASTER (24). A two-stage analytical process was implemented: first, all possible pairwise single nucleotide polymorphism (SNP) combinations (considering only allelic informative and gene-centered mapped SNPs) were assessed for the Pearson's correlation coefficient (PCC) difference between cases and controls; second, the likelihood ratio test of the logistic regression (LR) was applied to those subsets of SNP pairs deemed significant in the previous stage. Using simulated and real datasets, the method was shown previously to conduct a search for G × G that was unbiased to the marginal loci effects and captured most of the real G × G (24). The quality controls for the use of SNP data in this analysis were: minor allele frequency > 0.05 and *P* value cutoff of 10^{-5} for the Hardy–Weinberg equilibrium test. For linking SNPs to genes, each SNP was assigned to a specific gene locus if the variant mapped to a region ±10 kb from the corresponding genomic structure (first and last exon), using the ENSEMBL human genome release 57. Since the analysis required unambiguous SNP gene correspondences, the SNPs that overlapped with two or more gene loci were excluded. The pairs of SNPs with some evidence of linkage disequilibrium (LD; $r^2 > 0.2$) were also excluded from the analysis.

Gene and protein data analyses

Whole genome expression data for primary breast (25–27) (NKI-295 dataset and Gene Expression Omnibus reference GSE2034), colorectal (Gene Expression Omnibus reference GSE14333, ref. 28) and pancreatic (Gene Expression Omnibus reference GSE36924, ref. 29) tumors was analyzed using the preprocessed and normalized values. The NKI-295 breast tumors dataset contained 69 estrogen receptor α (ERα)-negative and 226 ERα-positive tumors. The colorectal and pancreatic datasets included 290 and 91 tumors, respectively. The PCCs were computed in R software and the mutual information (MI) was estimated using the ARACNE approach that applies a Gaussian kernel estimator (30). No pruning of MI-based edges (directed at specifically identify transcription factor–target interactions) was performed. Release #7 of the Human Protein Reference Database (31) was used, which contains 9461 proteins and 37 081 interactions that mainly represent experimentally demonstrated interactions compiled through literature curation. The high-confidence interactions dataset was derived from the integration of diverse data and contained 7401 proteins and 20 614 interactions (32).

Gene Ontology analyses

The Gene Ontology (GO) Biological Processes term annotations were downloaded from the Open Biological Ontologies release 2012/06 (MySQL version). GO terms were assigned to gene symbols after record linkage in which regular expression searches were required. Genes annotated at level 5 or lower in the hierarchy were assigned to level 4, but those also occurring at level 3 were excluded. Homodimers and gene pairs where both members share a GO annotation were also excluded. Only those term pairs with a frequency of ≥15% in the test set were evaluated. The test sets were 173 gene pairs from the 205 predicted G × G in breast cancer, and 82 gene pairs from the significant ($P_{LR} < 0.05$) G × G identified in the 14 highest-ranked bins for pancreatic cancer. Significance was assessed by comparing the observed frequency of each term–term interaction in the test set with the null distribution obtained by randomly selecting equivalent gene pair sets (1000 sets of similar gene pair size to the test set) from the top 0.25% MI values in the breast cancer setting, or from the top 10 000 gene pairs (according to their MI values) in the pancreatic cancer setting.

Cell culture and short hairpin RNAs

The MCF10A cells were obtained from the American Type Culture Collection, cultured in HuMEC (Invitrogen) media supplemented (hereafter 'supplemented media', in contrast to 'non-supplemented') with HuMEC Supplement and Bovine Pituitary Extract (Life Technologies) and used with <10 passages (from the initial American Type Culture Collection vial) for all assays. The short hairpin RNA (shRNA) used for depletion of lipoma-preferred partner (*LPP*) expression was the validated MISSION catalog TRCN0000301082 (Sigma–Aldrich). The lentiviral packaging, envelope, control and green fluorescent protein expression plasmids (psPAX2, pMD2.G, non-hairpin-pLKO.1, scrambled-pLKO.1 and pWPT-GFP) were purchased from Addgene. Production and collection of lentiviral particles followed a modified Addgene protocol. Initial viral titers > 5×10^5/ml were confirmed by Lenti-X GoStix (Clontech) and supernatants were then concentrated by ultracentrifugation or Lenti-X Concentrator (Clontech) and stored at −80°C. Concentrated viral supernatants were titrated for optimal inhibition of the target. Cells were infected with viral supernatants in the presence of 8 μg/ml polybrene and, after 48 h, incubated with puromycin to select stable populations of MCF10A control (shRNA control) or LPP-depleted cells (shRNA–*LPP*).

Proliferation, wound-healing and spheroids assays

Cells (5×10^3) transduced and selected for shRNA control or shRNA–*LPP* were plated in triplicate in 96-well plates with complete medium. After 24 h, adherent cells were cultured with supplemented and non-supplemented media in the presence or absence of transforming growth factor-β1 (TGFβ1) (100 pM in 1x cell culture media). Cell proliferation was measured at the time of replacing media at 24, 48 and 72 h, and using the CellTiter 96® AQueous One Solution Cell Proliferation Assay (Promega). For wound-healing assays, cells transduced and selected for shRNA control or shRNA–*LPP* were plated at confluence in duplicate on 24-well plates and incubated overnight. A straight line was then gently performed at the bottom of the dish. Cells were washed and incubated in non-supplemented media. After 24 h, transmission images were captured for each cell line using a FSX100 microscope (Olympus). Images were analyzed using the ImageJ software (Wright Cell Imaging Facility). Initial and final wound area (mm^2) were the variables used to calculate the wound closure percentage. For spheroids assays, cells (5×10^3) transduced and selected for shRNA control or shRNA–*LPP* were plated in 20 μl drops of supplemented media and allowed to grow in suspension. After 4 days, transmission images were captured line using a FSX100 microscope (Olympus). This experiment was performed with a minimum of 20 drops per cell line.

Western blotting and antibodies

Whole cell extracts from cultures transduced and selected for shRNA control or shRNA–*LPP*, in the presence or absence of supplemented media and TGFβ1, were prepared and used for western blotting as described elsewhere (33). The primary antibodies used for blotting were rabbit anti-ITGA5 (dilution 1:500, #AB1949; Millipore), anti-ITGB1 (dilution 1:2500, #610467; BD Pharmingen), anti-LPP (dilution 1:100, #0032-05; immunoGlobe) and anti-TUBA (dilution 1:2000, #2125; Cell Signaling).

Results

G × G overlap with complex context-defined gene coexpression

To compute G × G from a GWAS, we used data from the Cancer Genetic Markers of Susceptibility initiative in breast cancer (19) and applied a two-step analytical process (24). This strategy was implemented because it allows for fast and exhaustive computation of G × G. The first step consisted in calculating the difference in PCCs between controls and cases across all informative SNP pairs (Figure 1). The SNPs were those that were informative but also mapped to an annotated gene locus (i.e. a known gene). In the second step, significant SNP pairs ($P_{\text{PCC difference}} < 10^{-5}$) were analyzed by LR. Thus, ~390 000 SNP pairs with no evidence of LD and corresponding to ~339 000 gene pairs were analyzed at this stage. Next, from the distribution of P_{LR} values, a threshold $P_{LR} \leq 10^{-6}$ was defined, which yielded a set of 39 417 gene pairs (Figure 1). This set of predicted G × G (i.e. G × G potentially associated with risk of breast cancer in the general population) was subsequently assessed for overlap with known gene/protein relationships (Figure 1).

Integrative analyses in yeast have shown that genome-wide experimentally identified G × G overlap with protein–protein interactions to a degree that is significantly higher than expected by chance, of 10–20% of the known protein–protein interactions (34). We therefore examined whether the predicted G × G overlap with human protein–protein interactions. No significant overlap relative to what would be expected at random was identified using either a compiled dataset from the literature (31) or a high-confidence subset (32): indeed, only nine literature-compiled protein–protein interactions were found to be in common with the predicted G × G (data not shown).

Next, as G × G may also overlap significantly with gene coexpression (34), expression data from a large series of breast tumors (25) was analyzed. First, a standard measure of coexpression (i.e. PCC) was computed. In this analysis, all possible microarray probe pairs were evaluated and the maximum PCC value was then selected for each gene pair. By ranking the gene pairs according to their PCCs and assessing the top bins (starting at the top 0.1% of PCCs, which corresponded to 200 744 gene pairs), no significant overlap was observed

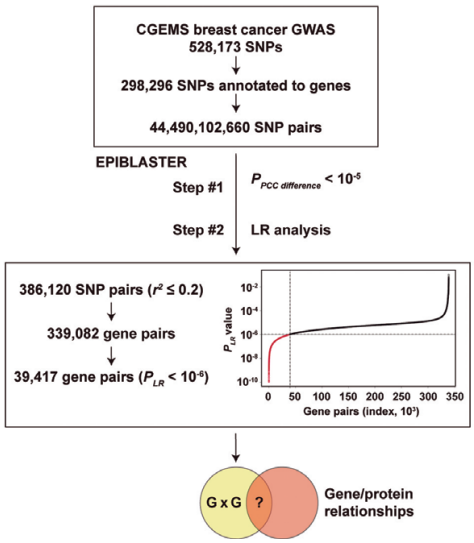


Fig. 1. Analytical strategy for the identification of G × G associated with breast cancer risk. The numbers of SNPs in the original GWAS dataset and in subsequent steps are shown. Also shown are the number of gene pairs at relevant analytical steps and the distribution of P_{LR} values. The selection of 39 417 gene pairs representing predicted G × G is subsequently evaluated for their overlap with other gene/protein relationships.

with the predicted G × G ($P_{empirical}$ = 0.73 using 1000 permutations of the 39 417 G × G); indeed, an opposite trend (i.e. under enrichment) was observed for the first bin (Figure 2A).

As G × G represent complex genetic relationships, we next used the MI measure to assess the overlap with non-linear expression relationships. Notably, using the same breast tumor dataset, the top 0.1 and 0.5% of MI-ranked gene pairs overlapped significantly with the predicted G × G (Figure 2B and Supplementary Table S1, available at *Carcinogenesis* Online). In the top 0.1%, 54 gene pairs were in common, whereas in the top 0.5%, 205 gene pairs were in common, which corresponded to relative enrichments (REs) of 32% (1.32 ratio relative to random, $P_{empirical}$ = 0.019) and 16% (1.16 ratio, $P_{empirical}$ = 0.015), respectively (Figure 2B and Supplementary Table S1, available at *Carcinogenesis* Online). When the top 0.5% PCCs were excluded from these MI bins, a suggestion of higher enrichment was observed in the first bin: RE of 43%, $P_{empirical}$ = 0.009, which corresponded to 43 gene pairs in common (Figure 2C and Supplementary Table S1, available at *Carcinogenesis* Online).

Robustness of the overlap between G × G and context-defined gene coexpression

As we observed an asymptotic distribution of P_{LR} values (Figure 1) and it was computationally unfeasible to analyze ranks of billions of gene pairs, a threshold $P_{LR} \leq 10^{-6}$ was initially used. Nonetheless, a significant enrichment was also revealed at a higher threshold ($P_{LR} \leq 10^{-5}$): with the top 0.5% MI-ranked gene pairs, 1322 were found in common with the predicted G × G, which corresponded to a RE of 16% and $P_{empirical} = 1.1 \times 10^{-4}$. At this threshold, the top 0.5% PCC-ranked gene pairs showed some suggestion of enrichment, although with a lower magnitude: 1002 pairs in common, RE of 5% and $P_{empirical}$ = 0.043. Moreover, the top 5% of MI-ranked gene pairs was also found to be significantly enriched, but, as expected, with a lower magnitude than the top 0.5%: 10 044 pairs in common, RE = 8% and $P_{empirical} = 1.0 \times 10^{-4}$.

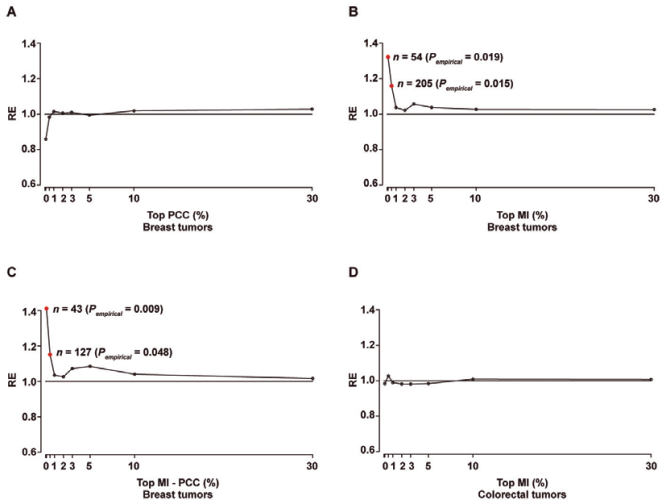


Fig. 2. Overlap between predicted breast cancer G × G and gene coexpression. (**A**) Overlap assessment with gene pairs ranked according to the highest PCCs computed from breast tumors (from top 0.1% to top 30% of pairs). The *y*-axis shows the REs. (**B**) Overlap assessment with gene pairs ranked according to the highest MIs computed from breast tumors (from top 0.1% to top 30% of pairs). The significant bins with the number of overlapping gene pairs are marked (red dots). (**C**) Overlap assessment with gene pairs ranked according to the highest MIs and subtracting those gene pairs included in the top 0.5% of PCCs from breast tumors. (**D**) Overlap assessment with gene pairs ranked according to the highest MIs computed from colorectal tumors.

The Cancer Genetic Markers of Susceptibility study was centered on sporadic postmenopausal breast cancer, hence most of the enrolled cases had developed tumors that were ERα positive (19). Consequently, the explained risk was mainly for ERα-positive and not ERα-negative breast cancer. The MIs were therefore computed separately for ERα-positive and ERα-negative tumors and then examined for their overlap with the predicted G × G ($P_{LR} \leq 10^{-6}$). Having defined the top 0.1% of MI-ranked gene pairs in ERα-positive and ERα-negative tumors, the REs were 36 and −5%, $P_{empirical} = 0.041$ and 0.56, respectively. This enrichment for ERα-positive cases, which was slightly (but not significantly) higher than observed in the full tumor dataset (36 versus 32%), corresponded to 34 gene pairs in common.

Using another large breast cancer expression dataset (27), a similar enrichment to the above was revealed for the top MI-ranked gene pairs: with the $P_{LR} \leq 10^{-6}$ threshold, the REs of the top 0.5 and 5% MI-ranked gene pairs were 15 and 13%, $P_{empirical} = 0.081$ and 0.045, respectively. Although the RE estimation for the top 0.1% was similar (12%), it was not significant ($P_{empirical} = 0.27$) probably because the number of gene pairs contained in this set was relatively low ($n = 175\,614$). Taking the 205 G × G predicted from the analysis of the first breast cancer dataset, the overlap with this second dataset was: 10 pairs at the top 0.1% of MIs; 24 at the top 0.5%; 30 at the top 1% and 43 at the top 5% (Supplementary Table S2, available at *Carcinogenesis* Online). Although this level of overlap was significant ($P_{hypergeometric} = 2.4 \times 10^{-16}$), the difference for the microarray platforms used in these studies may contribute to a substantial proportion of false-negative pairs.

Next, the overlap was assessed for complex gene coexpression in a distinct epithelial neoplasm, colorectal cancer (28). Using an expression dataset of a similar size to the breast cancer studies, no evidence of overlap was obtained at any MI threshold and with $P_{LR} \leq 10^{-6}$: the RE estimation for the top 0.1% of MI-ranked gene pairs, which also contained a similar number of gene pairs to the above study, was −1% (Figure 2D and Supplementary Table S1, available at *Carcinogenesis* Online). Therefore, G × G associated with risk of a given cancer type might only be predicted on the basis of complex gene expression relationships in the specific condition.

Biological processes in the G × G associated with breast cancer risk

No enrichment in significant marginal effects was observed in the set of SNPs involved in the predicted G × G (Supplementary Table S3, available at *Carcinogenesis* Online). However, the 205 gene pairs included four candidate genes identified in GWASs for breast cancer: *FTO*, *ITPR1*, *PDE4D* and *TCF7L2* (Supplementary Table S3, available at *Carcinogenesis* Online). It was predicted that variants in *ITPR1* and *BNC2* interact to confer increased risk (Z score = 4.90, $P_{LR} = 7.32 \times 10^{-7}$), and variation in *BNC2* has previously been associated with ovarian cancer in a GWAS (35). Interestingly, these variants in *BNC2* are not in LD based on HapMap Caucasians data ($r^2 = 0.01$); the interacting variant (rs717267) is at <2 kb from the 3′-exon of *BNC2*, whereas the marginal effect was detected in the 5′-region (35).

Next, an analysis of GO Biological Processes term annotations was performed to define the functional profile of the predicted G × G. Of the 205 gene pairs depicted above, 173 contained annotations for both members. Subsequently, using 1000 randomly selected equivalent gene pair sets, a network of significant (false discovery rate < 1%) term interactions was obtained (Figure 3). In this network, node size was proportional to the number of genes annotated with the corresponding term. Thus, the most frequent terms in the predicted G × G corresponded to interacting genes involved in metabolic or biosynthetic processes (Figure 3).

Biological insight from the predicted G × G

G × G have the potential to uncover functional relationships within or between biological processes and/or signaling pathways (36). In addition to the GWAS-based candidates mentioned above, the 205 gene set contained functional candidates previously linked to breast cancer risk. Among these, and also included in the top 5% MI-ranked pairs from the second breast cancer dataset, variation in *SMAD3* has been associated with breast cancer risk in *BRCA2* mutation carriers (37). SMAD3—together with other SMAD family members—is a critical signal transducer and transcriptional regulator downstream of TGFβR1 (38). In our study, an interaction was predicted between *SMAD3* rs2289263, which is not in LD with the risk variant ($r^2 < 0.2$), and *LPP* rs4686980 (Supplementary Table S3, available at *Carcinogenesis* Online). LPP is a nucleocytoplasmic protein involved in cell adhesion and motility, and transcriptional regulation (39,40). Therefore, to assess the prediction for breast carcinogenesis, cellular alterations upon depletion of LPP and/or modeling of TGFβ1 signaling were assessed using the non-tumorigenic MCF10A mammary epithelial cell line. Notably, simultaneous depletion of LPP and incubation with TGFβ1 (without the presence of other stimuli) increased cellular proliferation relative to the corresponding single perturbations (48 h timepoint, two-tailed t-test P values < 0.01; Figure 4A). In the control assays, TGFβ1 produced an antiproliferative effect as shown by a diminished proliferation rate (Figure 4A). Regarding the potential invasiveness, depletion of LPP impaired the formation of cellular spheroids (Figure 4B). Consistent with this observation, depletion of LPP increased cell migratory capacity in a wound-healing assay (Figure 4C). Moreover, the expression of the integrin receptors β1 and α5 was modulated and, in particular, α5 increased significantly upon simultaneous depletion of LPP and incubation with TGFβ1 (Figure 4D). Therefore, while depletion of LPP alone may provide a protumorigenic phenotype by increasing migration and impairing

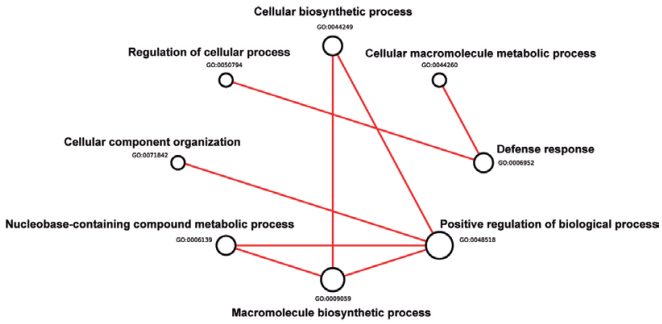


Fig. 3. Network of GO biological process terms linked to breast cancer G × G. The nodes represent GO terms (identifiers are shown) and an edge links two terms if the term–term interaction is overrepresented (false discovery rate < 1%) in the predicted G × G set.

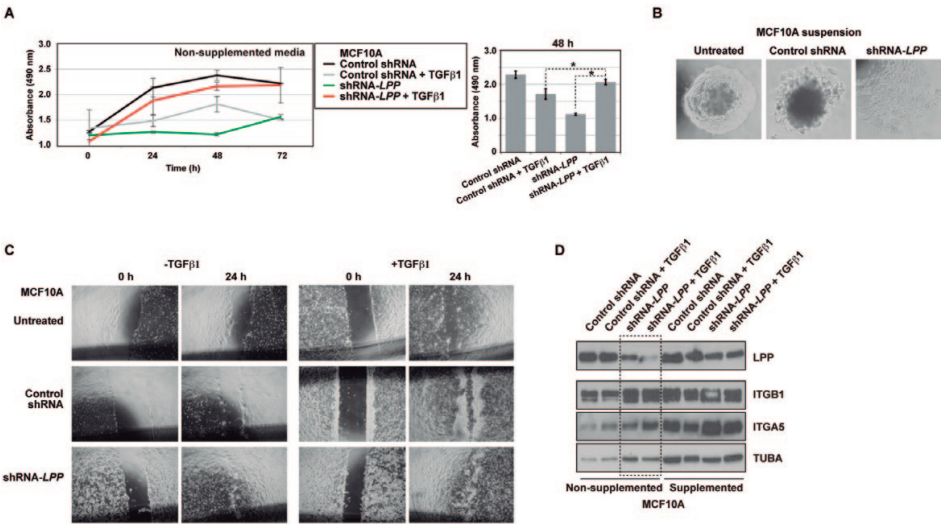


Fig. 4. Functional interplay between LPP and TGFβ signaling in a non-tumorigenic mammary epithelial cell model. (**A**) (Left panel) Graph showing the relative proliferation rates (from 0 to 72 h) of MCF10A cells in unperturbed cultures (shRNA control) or with LPP depletion and/or incubation with TGFβ1 (as shown in the inset). The cultures were not supplemented. (Right panel) Results at 48 h. (**B**) Representative images of MCF10A cells growing in suspension. Results are shown for non-transduced cells, and cells transduced with a control shRNA or directed against *LPP* expression (shRNA–*LPP*). (**C**) Representative images of wound-healing-scratch assays. Results are shown following scratch (0–24 h) of MCF10A cultures grown without/with TGFβ1, non-transduced, transduced with an shRNA control or directed against *LPP* expression. (**D**) Western blots results for LPP, the integrin receptors ITGB1 and ITGA5, and loading control (α-tubulin, TUBA) from extracts of cells transduced with an shRNA control or directed against *LPP* expression, without/with TGFβ1 as indicated. The conditions (non-supplemented) that corresponded to depletion of LPP without/with TGFβ1 are marked.

differentiation, simultaneous activation of TGFβ1 signaling substantially enhances cellular proliferation, which provides a mechanistic hypothesis for the predicted G × G.

Using complex gene coexpression evidence to predict G × G

The results above suggest that evidence based on complex- and context-defined gene coexpression patterns can be used to predict G × G associated with cancer risk. Thus, in a reverse strategy, we first analyzed gene expression profiles in pancreatic tumors (29) and then integrated the results with data from a pancreatic cancer GWAS (20,21). With the gene pairs ranked according to their MIs, 14 bins were defined, each of which contained 20 pairs (from the highest MI value to the 5000th value). Next, G × G using all unlinked ($r^2 < 0.2$) SNPs in a pair were computed and the number of significant associations was evaluated at two thresholds: $P_{LR} < 0.05$ and $P_{LR} < 0.01$. This analysis revealed more G × G than expected by chance: an average of 5.99% ($P_{\text{Wilcoxon rank test (WRT)}} = 0.003$ for the null hypothesis of ≤5% across the 14 bins) and an average of 1.41% ($P_{WRT} = 0.028$ for the null hypothesis of ≤1% across the 14 bins) of the SNP pairs showed $P_{LR} < 0.05$ and $P_{LR} < 0.01$, respectively (Figure 5).

Two control analyses were carried out to assess the identification of excess of G × G nominally significant for pancreatic cancer risk. An analogous bin analysis was carried out, but in this case the lowest 5000 MIs were used (i.e. non-significant gene coexpression). The results of this analysis did not detect significant G × G over the thresholds: $P_{WRT} = 0.35$ and 0.83 for the 5 and 1% thresholds, respectively. In addition, the REs between the 14 top and bottom bins were 25 and 59% for the $P_{LR} < 0.05$ and $P_{LR} < 0.01$, respectively, which appeared to be consistent with the enrichments shown above for breast cancer. Conversely, no enrichment was identified when using the 14 top bins but basing the ranking exclusively on PCCs: $P_{WRT} = 0.64$ and 0.39 for the 5 and 1% thresholds, respectively.

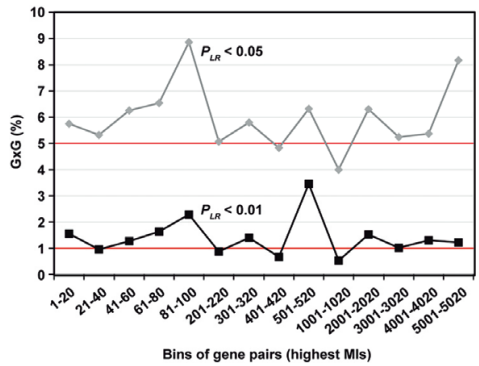


Fig. 5. Distribution of significant pancreatic cancer G × G across gene pairs ranked according to MIs from pancreatic tumors. The *y*-axis indicates the proportion of significant G × G at $P_{LR} < 0.05$ (gray line) and at $P_{LR} < 0.01$ (black line). The *x*-axis shows the 14 bins, starting with the highest MIs (i.e. 1–20 gene pairs). The red lines indicate the proportion thresholds for the corresponding P_{LR} values.

The significant G × G for pancreatic cancer risk in the top MI-based ranked bin included the following gene pairs: *AQP8-FGG*, *LOC402251-LOC442270*, *DLX5-MFAP5* and *ABCC8-PCP4* (Supplementary Table S4, available at *Carcinogenesis* Online). Two of these genes, *DLX5* and *PCP4*, have been functionally linked to

axon biology (41,42). Notably, alteration of genes annotated in the axon guidance pathway contributes to pancreatic carcinogenesis (29). Next, GO term enrichment analyses indicated (false discovery rate < 1%) frequent involvement of metabolic and biosynthetic processes, but also indicated mechanistic differences relative to breast cancer G × G (i.e. the involvement of genes in developmental processes; Figure 6).

Discussion

The identification of human G × G has the potential to add fundamental knowledge to our understanding of the genetic basis, molecular mechanisms and biological processes/signaling pathways involved in carcinogenesis (5,8,9,15). Although there are several well-established analytical strategies, the large (and continually increasing) number of genetic variants makes genome-wide G × G analyses highly time consuming, and it remains difficult to interpret the results from a biological perspective. This study introduces an integrative genomics strategy that can potentially support the identification of statistically significant G × G associated with cancer risk. In designing this study, it seemed reasonable to assume that there are G × G associated with cancer risk and, critically, that they can be identified by integrating different types of gene/protein relationships. The degree of overlap between genome-wide gene/protein relationships has been evaluated, and clearly established, in diverse studies in model organisms. Although human conditions should not be an exception, the lack of large-scale human G × G datasets has hampered similar integrative analyses. In addition, while experimental methodologies to systematically identify mammalian G × G have recently been developed (17,18), the results of our study suggest that context-specific studies must also be carried out. Thus, G × G associated with cancer risk may only be confidently identified when gene/protein relationships related to the specific cancer type/subtype are analyzed. In this regard, the lack of overlap with protein–protein interactions may be due to the fact that this type of evidence is typically not tissue- or cell type-specific. In addition, the human protein–protein interactions known to date do not represent the complete space of interactions occurring in cells. In fact, the gene expression analysis probably covers a larger fraction of all potential gene pairs.

The results of this study may lead to the genetic analysis of specific G × G in breast and pancreatic cancer. Although the enrichments shown may be considered relatively low (maximum of 36% for breast cancer and 59% for pancreatic cancer), the integration of additional gene/protein relationships could potentially improve the predictions. From the evaluation of the overlap for the 205 gene pairs between the two breast cancer expression datasets analyzed, it could be presumed that the conclusions of this study are limited by the characteristics of each dataset. In addition, the study may be limited by the relatively small sample size of the GWAS datasets analyzed and by the required assumption that a given SNP pair corresponds to a unique gene pair defined by the genomic location of the SNPs; however, it is frequently observed that the functional effects of low-penetrance mutations can implicate genes located dozens or hundreds of kilobases away (43). Integration of independent GWAS data could help to provide better G × G predictions. Moreover, the EPIBLASTER algorithm might not capture all possible forms of epistasis described in the literature (24).

The predicted G × G including candidate genes previously linked to cancer risk may help to further delineate the mechanisms of carcinogenesis. Furthermore, since some of the proposed G × G involve non-correlated variants relative to the marginal effect, they can potentially unveil mutations linked to differential effects. Following on from the predicted G × G between *LPP* and *SMAD3*, the identification of a signaling interplay between LPP and TGFβ1 in a non-tumorigenic mammary model provides a mechanistic hypothesis centered on altered epithelial cell proliferation and differentiation (38). LPP has been found to be highly expressed in normal luminal mammary cells (44), which are typically ERα positive, and coexpressed in breast tumors with a regulator of mammary cell differentiation (44,45). In this scenario, there is evidence for an expression quantitative trait locus in rs2289263 for *SMAD3* (46), which could provide a hypothesis for the interaction with *LPP*; there is no published evidence for an expression quantitative trait locus in rs4686980 (or for rs28615981 in LD) but these *LPP* variants appear to map within a c-FOS binding region identified by chromatin immunoprecipitation in the ENCODE project (47). SMAD3/4 and c-FOS have been shown to cooperate in promoting TGFβ signaling (48) and, therefore, this cooperation might regulate LPP function/levels. Importantly, a recent study has identified LPP as a key regulator of TGFβ-induced migration and invasion in HER2-overexpressing breast cancer (49). Our study expands on this observation by proposing that perturbation of LPP–TGFβ signaling promotes the initial stage of breast carcinogenesis.

At the level of the biological processes overrepresented in the predicted G × G sets for breast and pancreatic cancer risks, the common identification of metabolic and biosynthetic processes might be explained by their role in buffering phenotypic variability (36,50). Other identified processes, such as defense response for breast cancer risk and cell development for pancreatic cancer risk, might be related to tissue specificity and could be used to integrate additional gene/protein relationships for prioritizing G × G. Together, our study proposes a method that may help to further decipher the genetic basis of cancer risk.

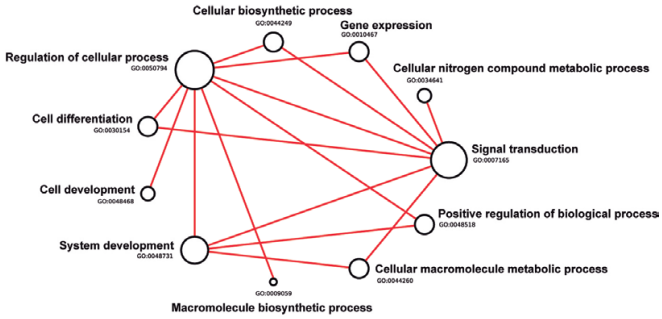


Fig. 6. GO Biological Process terms linked to predicted pancreatic cancer G × G. The nodes represent GO terms (identifiers are shown) and an edge links two terms if the term–term interaction is overrepresented (false discovery rate < 1%) in the predicted G × G set. This test set corresponded to the significant G × G represented in Figure 5 (top 14 gene pair bins, $P_{LR} < 0.05$).

Conclusions

Here, based on the premise that genes/proteins act coordinately across biological levels, we undertook an integrative study in order to predict G × G associated with cancer risk. Our study was centered on breast and pancreatic cancer and the results show that G × G associated with risk may be partially supported on the basis of complex gene coexpression in the specific cancer type. The requirement of complex (i.e. non-linear) coexpression in a defined cancer setting is consistent with the intricate nature of epistasis and the molecular specificities of carcinogenesis. The predicted G × G provide novel hypotheses for the functional interplay between biological processes in carcinogenesis. The knowledge generated by this study may stimulate new research toward a better understanding of the genetic basis of cancer risk.

Supplementary material

Supplementary Tables S1–S4 can be found at http://carcin.oxfordjournals.org/

Funding

Fundación Eugenio Rodríguez Pascual (2012-13); Fundación Ramón Areces (XV Enfermedades Raras); Fundación Roses Contra el Càncer (2012); Generalitat de Catalunya (2009-SGR283); Red Temática de Investigación Colaborativa en Cáncer (12/0036/0008); Spanish Ministry of Health, Instituto Salud Carlos III, Fondo de Investigación Sanitaria (12/01528).

Acknowledgements

We wish to thank all study participants for their valuable contribution and the National Cancer Institute NIH initiatives for making available the genome-wide association studies data.

Conflict of Interest Statement: None declared.

References

1. Easton,D.F. *et al.* (2008) Genome-wide association studies in cancer. *Hum. Mol. Genet.*, **17**, R109–R115.
2. Couch,F.J. *et al.*; kConFab Investigators; SWE-BRCA; Ontario Cancer Genetics Network; HEBON; EMBRACE; GEMO Study Collaborators; BCFR; CIMBA. (2013) Genome-wide association study in BRCA1 mutation carriers identifies novel loci associated with breast and ovarian cancer risk. *PLoS Genet.*, **9**, e1003212.
3. Gaudet,M.M. *et al.*; KConFab Investigators; Ontario Cancer Genetics Network; HEBON; EMBRACE; GEMO Study Collaborators; GENICA Network. (2013) Identification of a BRCA2-specific modifier locus at 6p24 related to breast cancer risk. *PLoS Genet.*, **9**, e1003173.
4. Gibson,G. (2011) Rare and common variants: twenty arguments. *Nat. Rev. Genet.*, **13**, 135–145.
5. Quigley,D. *et al.* (2009) Systems genetics analysis of cancer susceptibility: from mouse models to humans. *Nat. Rev. Genet.*, **10**, 651–657.
6. Pan,Q. *et al.* (2013) Epistasis, complexity, and multifactor dimensionality reduction. *Methods Mol. Biol.*, **1019**, 465–477.
7. Michailidou,K. *et al.*; Breast and Ovarian Cancer Susceptibility Collaboration; Hereditary Breast and Ovarian Cancer Research Group Netherlands (HEBON); kConFab Investigators; Australian Ovarian Cancer Study Group; GENICA (Gene Environment Interaction and Breast Cancer in Germany) Network. (2013) Large-scale genotyping identifies 41 new loci associated with breast cancer risk. *Nat. Genet.*, **45**, 353–61, 361e1.
8. Moore,J.H. *et al.* (2009) Epistasis and its implications for personal genetics. *Am. J. Hum. Genet.*, **85**, 309–320.
9. Maxwell,C.A. *et al.* (2008) Genetic interactions: the missing links for a better understanding of cancer susceptibility, progression and treatment. *Mol. Cancer*, **7**, 4.
10. Cordell,H.J. (2009) Detecting gene-gene interactions that underlie human diseases. *Nat. Rev. Genet.*, **10**, 392–404.
11. Moore,J.H. *et al.* (2010) Bioinformatics challenges for genome-wide association studies. *Bioinformatics*, **26**, 445–455.
12. Greenland,S. (2009) Interactions in epidemiology: relevance, identification, and estimation. *Epidemiology*, **20**, 14–17.
13. Drees,B.L. *et al.* (2005) Derivation of genetic interaction networks from quantitative phenotype data. *Genome Biol.*, **6**, R38.
14. Hartman,J.L. 4th *et al.* (2001) Principles for the buffering of genetic variation. *Science*, **291**, 1001–1004.
15. Boone,C. *et al.* (2007) Exploring genetic interactions and networks with yeast. *Nat. Rev. Genet.*, **8**, 437–449.
16. Bernards,R. (2012) A missing link in genotype-directed cancer therapy. *Cell*, **151**, 465–468.
17. Roguev,A. *et al.* (2013) Quantitative genetic-interaction mapping in mammalian cells. *Nat. Methods*, **10**, 432–437.
18. Laufer,C. *et al.* (2013) Mapping genetic interactions in human cancer cells with RNAi and multiparametric phenotyping. *Nat. Methods*, **10**, 427–431.
19. Hunter,D.J. *et al.* (2007) A genome-wide association study identifies alleles in FGFR2 associated with risk of sporadic postmenopausal breast cancer. *Nat. Genet.*, **39**, 870–874.
20. Amundadottir,L. *et al.* (2009) Genome-wide association study identifies variants in the ABO locus associated with susceptibility to pancreatic cancer. *Nat. Genet.*, **41**, 986–990.
21. Petersen,G.M. *et al.* (2010) A genome-wide association study identifies pancreatic cancer susceptibility loci on chromosomes 13q22.1, 1q32.1 and 5p15.33. *Nat. Genet.*, **42**, 224–228.
22. Li,Y. *et al.* (2006) MACH 1.0: rapid haplotype reconstruction and missing genotype inference. *Am. J. Hum. Genet.*, **79**, S2290.
23. Purcell,S. *et al.* (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.*, **81**, 559–575.
24. Kam-Thong,T. *et al.* (2011) EPIBLASTER-fast exhaustive two-locus epistasis detection strategy using graphical processing units. *Eur. J. Hum. Genet.*, **19**, 465–471.
25. van de Vijver,M.J. *et al.* (2002) A gene-expression signature as a predictor of survival in breast cancer. *N. Engl. J. Med.*, **347**, 1999–2009.
26. Chang,H.Y. *et al.* (2005) Robustness, scalability, and integration of a wound-response gene expression signature in predicting breast cancer survival. *Proc. Natl Acad. Sci. USA*, **102**, 3738–3743.
27. Wang,Y. *et al.* (2005) Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *Lancet*, **365**, 671–679.
28. Jorissen,R.N. *et al.* (2009) Metastasis-associated gene expression changes predict poor outcomes in patients with Dukes stage B and C colorectal cancer. *Clin. Cancer Res.*, **15**, 7642–7651.
29. Biankin,A.V. *et al.*; Australian Pancreatic Cancer Genome Initiative. (2012) Pancreatic cancer genomes reveal aberrations in axon guidance pathway genes. *Nature*, **491**, 399–405.
30. Margolin,A.A. *et al.* (2006) ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics*, **7** (suppl. 1), S7.
31. Prasad,T.S. *et al.* (2009) Human protein reference database and human proteinpedia as discovery tools for systems biology. *Methods Mol. Biol.*, **577**, 67–79.
32. Wang,X. *et al.* (2012) Three-dimensional reconstruction of protein networks provides insight into human genetic disease. *Nat. Biotechnol.*, **30**, 159–164.
33. Colas,E. *et al.* (2012) ETV5 cooperates with LPP as a sensor of extracellular signals and promotes EMT in endometrial carcinomas. *Oncogene*, **31**, 4778–4788.
34. Costanzo,M. *et al.* (2010) The genetic landscape of a cell. *Science*, **327**, 425–431.
35. Song,H. *et al.*; Australian Cancer (Ovarian) Study; Australian Ovarian Cancer Study Group; Ovarian Cancer Association Consortium. (2009) A genome-wide association study identifies a new ovarian cancer susceptibility locus on 9p22.2. *Nat. Genet.*, **41**, 996–1000.
36. Segrè,D. *et al.* (2005) Modular epistasis in yeast metabolism. *Nat. Genet.*, **37**, 77–83.
37. Walker,L.C. *et al.*; kConFab; GEMO Study Collaborators; HEBON; ModSQuaD; EMBRACE; SWE-BRCA. (2010) Evidence for SMAD3 as a modifier of breast cancer risk in BRCA2 mutation carriers. *Breast Cancer Res.*, **12**, R102.
38. Massagué,J. *et al.* (2012) TGF-β control of stem cell differentiation genes. *FEBS Lett.*, **586**, 1953–1958.
39. Majesky,M.W. (2006) Organizing motility: LIM domains, LPP, and smooth muscle migration. *Circ. Res.*, **98**, 306–308.
40. Petit,M.M. *et al.* (2005) The tumor suppressor Scrib interacts with the zyxin-related protein LPP, which shuttles between cell adhesion sites and the nucleus. *BMC Cell Biol.*, **6**, 1.
41. Harashima,S. *et al.* (2011) Purkinje cell protein 4 positively regulates neurite outgrowth and neurotransmitter release. *J. Neurosci. Res.*, **89**, 1519–1530.

42. Long,J.E. *et al.* (2003) DLX5 regulates development of peripheral and central components of the olfactory system. *J. Neurosci.*, **23**, 568–578.
43. Pomerantz,M.M. *et al.* (2009) The 8q24 cancer risk variant rs6983267 shows long-range interaction with MYC in colorectal cancer. *Nat. Genet.*, **41**, 882–884.
44. Guo,B. *et al.* (2006) The LIM domain protein LPP is a coactivator for the ETS domain transcription factor PEA3. *Mol. Cell. Biol.*, **26**, 4529–4538.
45. Kurpios,N.A. *et al.* (2009) The Pea3 Ets transcription factor regulates differentiation of multipotent progenitor cells during mammary gland development. *Dev. Biol.*, **325**, 106–121.
46. Grundberg,E. *et al.*; Multiple Tissue Human Expression Resource (MuTHER) Consortium. (2012) Mapping cis- and trans-regulatory effects across multiple tissues in twins. *Nat. Genet.*, **44**, 1084–1089.
47. ENCODE (2011) A user's guide to the encyclopedia of DNA elements (ENCODE). *PLoS Biol.*, **9**, e1001046.
48. Zhang,Y. *et al.* (1998) Smad3 and Smad4 cooperate with c-Jun/c-Fos to mediate TGF-beta-induced transcription. *Nature*, **394**, 909–913.
49. Ngan,E. *et al.* (2013) A complex containing LPP and α-actinin mediates TGFβ-induced migration and invasion of ErbB2-expressing breast cancer cells. *J. Cell Sci.*, **126**(Pt 9), 1981–1991.
50. Snitkin,E.S. *et al.* (2011) Epistatic interaction maps relative to multiple metabolic phenotypes. *PLoS Genet.*, **7**, e1001294.

Received July 8, 2013; revised November 7, 2013; accepted November 27, 2013

4

Discussió global dels resultats

El càncer de mama és un exemple de malaltia complexe on participen tant factors genètics com ambientals en la seva etiologia. Entendre la base genètica del càncer de mama en profunditat podria permetre entendre millor la malaltia i desenvolupar millors eines de diagnòstic, pronòstic i/o tractament.

Al iniciar aquesta tesi es coneixia el 25% del risc familiar [210] i la hipòtesi més acceptada per explicar la part del risc restant era la CV-CD [115]. Segons aquesta hipòtesi, serien variants de baixa penetrància les que influirien en la proporció de la susceptibilitat a la malaltia que quedava per explicar. La esperança de poder identificar aquests gens de susceptibilitat estava en els estudis de GWAS que començaven a emergir. S'havien realitzat dos GWAS [49, 53] de càncer de mama que havien identificat set al·lels de risc (que en conjunt explicaven aproximadament el 4% del risc familiar). Però degut als estrictes llindars de significació, ja mencionats, que s'han d'aplicar a aquests estudis, és fa molt difícil detectar el senyal d'al·lels que tenen un efecte petit [174]. En aquesta tesi, per tal d'evitar aquesta limitació, vam analitzar els resultats d'un dels primers estudis GWAS de mama publicats [175] sota una perspectiva de la biologia de sistemes, aplicant estratègies basades en conjunts de gens/proteïnes funcionalment relacionats per identificar nous gens de baixa penetrància, els processos biològics en els que participen i els mecanismes moleculars mitjançant els quals confereixen el risc així com, les seves interaccions.

Per aquest motiu, aquest estudi està limitat a l'anàlisi de les dades d'un únic GWAS que presenta certes especificitats epidemiològiques. També cal destacar que degut a que l'estratègia utilitzada es centra en els gens i les seves anotacions funcionals, es van excloure dels anàlisis els SNPs que no es podien assignar amb certa probabilitat a un gen, és a dir, que no estaven

en la regió genòmica que codifica per un gen o a més de 10 kilobases (kb) d'aquesta. Tot i que, com ja hem explicat, aquests SNPs també poden ser funcionals (e.g. l'efecte d'SNPs a la regió 8q24 i el risc en diversos tipus de càncer [53, 211–214]) i per tant, podrien participar en la susceptibilitat a la malaltia.

El primer objectiu d'aquesta tesi va ser identificar els processos biològics possiblement importants en el risc de càncer de mama. Per portar a terme aquests anàlisis necessitàvem un rànquing de gens que vam obtenir assignant a cadascun dels gens del genoma el P_{valor} menor del GWAS entre tots els SNPs anotats en el *locus* concret. Aquest procés introdueix un biaix ja que existeix una correlació positiva entre la longitud d'un determinat gen i el nombre de SNPs que pot contenir (per tant, es pot obtenir una associació significativa simplement per atzar quan més extensió té un gen) [141, 215]. Un rànquing no ajustat, tendirà a obtenir com a significatius els processos biològics en els que participen productes de gens llargs [144, 216]. Cal destacar, però, que els gens de càncer tendeixen a abarcar grans regions cromosòmiques [217]. Al examinar vuit gens de susceptibilitat a càncer de mama de baixa penetrància (*CASP8*, *COX11*, *ESR1*, *FGFR2*, *LSP1*, *MAP3K1*, *RAD51L1* i *TOX3*) van presentar una tendència a ser més extensos (211 kb d'extensió genòmica mitjana (x) i una desviació estàndard (s) de 283 kb) en comparació amb tots els gens del rànquing (x = 66 kb i s = 128 kb). Existeixen diferents mètodes per analitzar els processos biològics en les dades dels GWAS [218] però aquests mètodes, poden donar lloc a resultats diferents analitzant les mateixes dades [219]. Nosaltres vam implementar un algoritme basat en el programa FatiScan [148, 220] per a poder examinar els resultats més flexiblement. Aquest mètode permet detectar diferències més modestes que el GSEA, un altre mètode molt comú [148].

D'aquests anàlisis vam inferir que variants comunes que afecten, en particular, la funció de gens/proteïnes de *Cell Communication* i *Cell Adhesion* probablement tendeixen a influir en el risc més que gens en altres processos. Aquest resultat es va confirmar amb l'anàlisi a nivell de proteïnes. Els interactors físics directes i a un pas dels productes dels gens "referents" de susceptibilitat en la xarxa de l'interactoma estaven sobrerepresentats en els processos biològics de *Cell Communication* i *cell adhesion*. Les proteïnes que interaccionen físicament és una prova de la seva participació en els mateixos processos biològics [128] i manifesta una relació funcional entre els gens que les codifiquen, per tant, alteracions en aquests poden donar lloc a la malaltia [128]. Alguns gens de susceptibilitat de baixa penetrança identificats posteriorment a aquest estudi pertanyen a aquests processos biològics; per exemple, *CDKN2A* [55] pertany a *Cell Adhesion* i 10 gens dels identificats en els anàlisis del COGS pertanyen a *Cell Communication* [21].

A més, la implicació de *Cell Communication* i *Cell Adhesion* va ser interessant donat que ja se sabia la seva contribució a aquesta neoplàsia epitelial, encara que a nivell somàtic [179]. D'aquesta manera, els nostres resultats podrien associar les pertorbacions moleculars inicials amb la posterior progressió del càncer, el que suggeriria un camí més continu del que prèviament s'havia pensat entre la línia germinal i les alteracions somàtiques. En aquest sentit, aquests resultats coincideixen amb els processos identificats en un estudi independent analitzant el mateix GWAS [221].

Continuant amb el segon objectiu d'aquesta tesi, vam revelar les propietats dels gens de susceptibilitat de baixa penetrància basant-nos en les observacions dels rànquings realitzats en el GWAS de CGEMS, i de diferents condicions de càncer de mama. D'aquesta manera, variants en els gens de

baixa penetrància que correlacionarien amb gens expressats diferencialment en teixit normal i tumoral revelaria supressors tumorals o oncogens, segons la direcció del canvi d'expressió [222]. També es va revelar la participació d'aquests gens en l'inici de la tumorigènesi i per tant en el risc, al correlacionar amb la expressió diferencial en edats primerenques de diagnòstic [223]. Finalment, van suggerir la dependència molecular o funcional de gens d'alta penetrància o les seves proteïnes al presentar canvis en l'expressió al pertorbar *BRCA1* [223, 224].

Al combinar les evidències que presentaven una asimetria similar a la observada en el GWAS vam obtenir un llistat ordenat dels candidats més probables que contenia associacions moleculars i funcionals rellevants en la tumorigènesi de mama. Entre els 50 primers gens del rànquing vem trobar supressors tumorals i/o oncogens descrits anteriorment en la literatura (*DKK3* [225] i *TFPI2* [226], gens amb variants que confereixen risc a càncer de mama (*IGF1* [227]). Calia destacar també la identificació de quatre gens (*NTRK2*, *MAP3K12*, *PDGFRA* i *PDGFRL*) les proteïnes dels quals participen en la via de senyalització de *MAPK* on també hi participaven gens de susceptibilitat coneguts (*FGFR2* i *MAP3K1* [51, 53, 175]). Entre aquests 50 gens també hi havien gens previament relacionats amb el pronòstic, metàstasi o resposta al tractament del càncer de mama (*BCL2* [228], *CXCL12* [229–231] i *FBLN1* [232, 233]). A més, esudis experimentals han demostrat que *ABTB1* i *EGR2* són mitjancers del supressor tumoral *PTEN* [234]. Finalment, com a prova adicional de la seva rellevància funcional, la xarxa de regulació de la transcripció realitzada entre els gens candidats i els gens referents presentava una connectivitat més alta de l'esperada per atzar, el que confirmava que aquests gens/proteïnes funcionen en processos biològics relacionats i que estan associats amb gens de risc coneguts. Cal destacar que

un dels gens candidats *ITPR1* de la llista, ha estat identificat en el metanàlisi de GWAS publicat l'any passat [21]. Per prioritzar els gens candidats d'aquesta llista, es podria escollir els interactors directes i a un pas dels referents que tenen més veïns assignats als processos de *Cell Communication* i *Cell Adhesion* (*CDH1* i *FGFR2*) en la xarxa del interactoma. També es podria triar els gens de la xarxa de regulació de la transcripció que presenten més centralitat (*BCL2*, *BMP1*, *NTRK2*, *PTGER3*, *RUNX2*) o que connecten dos gens referents (*DKK3*, *NTRK2*) perquè el seu efecte podria ser l'alteració de diversos gens de baixa penetrança.

En aquest punt, gràcies al fet de disposar ja dels genotips de l'estudi GWAS de CGEMS [175], vam ajustar el rànquing de gens original per minimitzar el biaix comentat anteriorment. Així, per corregir el rànquing original, vam utilitzar la distribució nul·la de l'estadístic t de la correlació de Pearson (ajustada per l'edat) després de realitzar 10.000 permutacions. Cal dir que l'anàlisi dels processos biològics del GO en aquest nou rànquing no va mostrar cap asimetria que passés la correcció per comparacions múltiples (*Cell Communication* $P_{\text{valor nominal}} = 0{,}42$ i *Cell Adhesion* $P_{\text{valor nominal}} = 0{,}37$). Tot i això, la majoria dels processos amb $P_{\text{valor nominal}}$ significatiu estaven relacionats amb alteracions somàtiques (per exemple *Rho protein signal transduction* $P_{\text{valor nominal}} = 0{,}005$).

A continuació, vam portar a terme una avaluació més detallada de la relació entre la línia germinal i la somàtica en la carcinogènesi de mama. Tot i que existien algunes evidències de la existència de la connexió entre la línia germinal i la somàtica, no s'havia avaluat aquesta hipòtesi explícitament. Aquesta possible connexió entre la línia germinal i les alteracions somàtiques podria ser destacada amb la identificació de variants de risc en

CDKN2A/B, *FGFR2* i *MAP3K1* i actualment s'ha vist que gran part dels gens de baixa penetrància presenten també alteracions somàtiques en diverses neoplàsies inclosa la de mama [21]. Per això vam decidir examinar aquesta connexió analitzant la possible relació de diferents grups de gens coneguts per presentar alteracions somàtiques i la susceptibilitat al càncer de mama en les dades del GWAS de CGEMS. Els resultats obtinguts van suggerir que variants genètiques germinals en gens que codifiquen per les *driver kinases* [94, 192] podrien influir en el risc a càncer de mama. Recentment s'ha identificat una variant de risc a aquesta malaltia en el gen *TGFBR2* [21] que pertany a les *driver kinases*. El grup de les *driver kinases* està format per gens que codifiquen per proteïnes quinasa que presenten *driver mutations*, és a dir, mutacions que contribueixen a la transformació d'una cèl·lula normal a una cèl·lula cancerosa [94]. El que suggeriria, una connexió entre la línia germinal i les alteracions somàtiques en càncer de mama a través de les *driver kinases*.

El posterior anàlisi de les *driver kinases* en un estudi cas-control de mama de població polaca va destacar sis nous candidats a *loci* de susceptibilitat coherents amb alteracions genètiques en la línia somàtica i/o funcional prèviament citades. En el cas de *CDKL2* s'han descrit mutacions sense sentit en línies cel·lulars de càncer de mama i ovari i també en tumors [192, 235]. CDKL2 (també coneguda per p56 or KKIAMRE) és el membre més distant de la família de les *CDC2-related serine/threonine protein kinase*, involucrada en el factor de senyalització del creixement epidèrmic [236]. Respecte a altres *driver kinases* identificades, *DYRK2* presenta mutacions *nonsense* en tumors de mama [235] i amb mutacions *missense* en tumors del sistema nerviós central [192]. La funció de DYRK2 en la resposta al dany al DNA [237] podria coincidir amb els resultats de RAD51L1[48] del GWAS

de CGEMS: la pèrdua de la funció de DYRK2 alteraria la activació de l'apoptosi en resposta al dany a ADN via ATM [237], la qual cosa promouria la carcinogènesis. D'aquest resultat, cal destacar la identificació d'al·lels de risc en tres *loci* que codifiquen per receptors d'epinefrina (EPH) ja que, la via de senyalització mitjançant EPH regula processos importants que estan alterats en la carcinogènesi com la comunicació cèl·lula a cèl·lula, la migració cel·lular i l'adhesió via el citoesquelet d'actina [238, 239]. Així, aquesta via de senyalització, a través de les activitats de RHO i RAS/MAPK [240], està implicada en el manteniment de l'arquitectura del teixit epitelial de manera que podria actuar com un supressor tumoral [238, 239]. Aquestes observacions podrien indicar, de manera similar a la tumorigenesi colorectal [241], que una compartimentació primerenca, a través de EPH, de les lesions neoplàsiques en el teixit mamari és crític per prevenir la posterior aparició del carcinoma. Per tant, a partir d'una pertorbació en la línia germinal de la expressió o de la funció, EPHB1 podria contribuir a la variabilitat observada en la transició d'una lesió *in situ* a un carcinoma invasiu [242].

Per tal d'aprofundir en la possible participació dels sis *loci* de les *driver kinases*, i especialment el *locus EPHB1* en el risc de càncer de mama, vam examinar si aquests *loci* influenciaven en el risc a càncer de mama en portadors de mutacions en *BRCA1* i *BRCA2*. Cal recordar que els gens de baixa penetrància identificats en estudis d'associació de població general sovint actuen com a modificadors del risc a càncer en aquests portadors de mutacions [77, 199].

L'anàlisi de les 95 variants genotipades en els sis *loci* candidats van suggerir l'associació entre variants en *EPHB1* i el risc a càncer de mama tant en portadors de mutacions en *BRCA1* com en portadors de mutacions en *BRCA2*.

Caldria destacar la variant rs16842235 ja que sembla actuar com a eQTL i/o meQTL. Aquesta observació podria ser rellevant si consideréssim que la expressió d'*EPHB1* forma part d'un programa transcripcional característic de les cèl·lules mamàries embrionàries [243, 244] i que el receptor de senyal epinefrina participa en el desenvolupament de la glàndula mamària i en la diferenciació epitelial [245, 246].

A continuació, l'anàlisi de 2.000 variants imputades del *locus EPHB1* va mostrar possibles associacions més fortes i senyals en blocs de desequilibri de lligament diferents en portadors de de *BRCA1* i *BRCA2*. Per contra, la variant rs3732568 del *locus* (*EPHB1*), que va presentar associació en el estudi cas-control de Polònia, no va presentar evidències d'associació amb el risc a càncer de mama en portadors de mutacions en *BRCA1* i *BRCA2* (P_{valor} en l'anàlisi d'imputacions $> 0{,}15$). No obstant, donat l'elevat nombre d'anàlisis estadístics realitzats, aquests resultats s'han de comprovar en un nombre més gran de portadors abans de considerar les associacions certes. El nostre grup es troba a l'espera de rebre noves dades genètiques del consorci COGS-CIMBA [60, 67] per la validació d'aquests resultats.

En el tercer objectiu d'aquesta tesi, es van analitzar les interaccions genètiques en les dades del GWAS ja que aquestes s'han proposat com un dels factors que podrien explicar la *missing heritability* [122]. Existeixen diferents estratègies per analitzar les GxG a nivell de genoma [159, 247]. El nostre grup va decidir aplicar l'algoritme EPIBLASTER [163] perquè permetia analitzar exhaustivament totes les parelles d'SNPs que mapaven a un gen del GWAS en un temps relativament curt gràcies a la utilització de GPUs. Per evitar els estrictes llindars de significació que porten associats aquests anàlisis (pel fet de testar moltes parelles d'SNPs) i aprofitant el co-

neixement acumulat a partir dels anàlisis de les GxG a gran escala realitzats en organismes model, vam portar a terme una estratègia d'integració de dades de gens i proteïnes per ajudar a la identificació de les GxG associades al risc a càncer. En llevats s'ha demostrat, a nivell genòmic, que les GxG solapen amb les interaccions proteïna-proteïna en un 10-20% i que també solapen significativament amb dades de coexpressió gènica [156, 248]. A més, una estratègia integradora ajudaria a interpretar biològicament prediccions purament estadístiques de GxG. En humans, es desconeixen les GxG que descriuen les relacions funcionals entre gens més enllà de les interaccions dels seus productes.

Els resultats del nostre estudi van suggerir que les GxG associades a risc de càncer es podien identificar amb més seguretat quan s'integren amb relacions/interaccions entre gens/proteïnes en un subtipus de càncer rellevant. En aquest sentit, la manca de solapament amb les interaccions proteïna-proteïna podia ser deguda al fet que aquestes dades no són en general específiques per un teixit o un tipus de cèl·lula. També hem de tenir en compte que les interaccions proteïna-proteïna humanes que es coneixen actualment no representen totes les interaccions que tenen lloc en les cèl·lules. Actualment la base de dades HPRD conté 30.047 proteïnes i 41.327 interaccions proteïna-proteïna [126], el que podria representar menys del 50% de les interaccions que possiblement existeixen [249]. Per una altra banda, encara que els enriquiments mostrats es poden considerar relativament baixos (un màxim del 36%), la integració addicional de dades d'interaccions entre proteïnes específiques del tipus cel·lular podria millorar les prediccions. També es podrien integrar amb dades de GxG experimentals com les descrites recentment en subtipus cel·lulars concrets [157, 158].

Entre les 205 GxG predites s'hi inclouen gens candidats prèviament associats a risc de càncer i que poden ajudar a acabar de delinear els mecanismes fonamentals de la carcinogènesi. A més, a partir de la interacció destacada entre *LPP* i *SMAD3*, vam identificar que en un model cel·lular epitelial de la mama no tumorigènic, MCF10A, LPP interactuava amb TGFβ1 en la via de senyalització, proporcionant així una hipòtesi funcional per a l'increment de risc observat: l'alteració de la diferenciació epitelial [206].

Actualment, set anys després dels primers GWAS [49, 53] s'han identificat 76 gens de baixa penetrància; tanmateix, degut al petit efecte individual que confereix cadascun, en conjunt aquests explicarien només el 15% del risc familiar. Si hi sumem el 21% que confereixen els gens d'alta i moderada penetrància, tenim que tots els gens de susceptibilitat coneguts expliquen el 36% de la heretabilitat. A partir dels resultats el metanàlisi de càncer més gran realitzat fins el moment, els autors suggereixen que existeixen molts més *loci* (potser alguns milers), la majoria amb $0{,}95 < \text{ORs} < 1{,}05$ que contribueixen a la susceptibilitat. El conjunt d'aquests SNPs explicaria aproximadament un altre 14% del risc familiar. Aquests resultats donen suport a la hipòtesi del model infinitesimal [121] per explicar la *missing heritability* en càncer de mama. Segons aquesta hipòtesi, centenars o milers d'al·lels comuns, amb efectes relativament molt petits, contribuirien a la susceptibilitat a la malaltia. Els GWAS portats a terme fins ara, haurien detectat els al·lels amb un efecte major, el reste dels al·lels de susceptibilitat, no poden ser detectats degut als estrictes llindars de significació utilitzats en els aquests estudis [121, 250]. En aquest context, les GxG tampoc poden ser detectades amb robustesa degut als estrictes llindars de significació requerits. Aquest fet, justifica la nostra estratègia basada en la biologia de sistemes, d'anàlisi dels resultats del GWAS en lloc de centrar-nos en els

SNPs/gens individualment. D'aquesta manera, a partir de la llista ordenada de tots els SNPs/gens en relació a l'associació amb el risc a càncer de mama hem buscat la distribució de grups de gens relacionats funcionalment millorant el poder per detectar associacions i ajudant a la interpretació biològica dels resultats. En resum, el conjunt dels nostres estudis han intentat contribuir al coneixement de la base genètica i molecular que influeix en el risc de desenvolupar càncer de mama. Ho hem fet des d'una perspectiva de biologia de sistemes, proposant nous gens candidats i les seves interaccions funcionals/moleculars (dels productes) associades al risc. Tanmateix, aquests estudis es troben en evolució i avaluació.

5

Conclusions

1. Variants comunes en els gens de *Cell Communication* i *Cell Adhesion* podrien influir en el risc a càncer de mama.

2. Variants comunes en determinats *loci* de les *driver kinases*, particularment en els gens que codifiquen per receptors d'EPHs, podrien influir en el risc de càncer de mama.

3. Variants comunes en el *locus EPHB1* podrien estar associades amb el risc a càncer de mama en portadors de mutacions en *BRCA1* i *BRCA2*.

4. Les GxG associades al risc de càncer podrien ser predites en part mitjançant els patrons complexes de coexpressió gènica específics del tipus de càncer.

5. Les GxG podrien contribuir a explicar part de la *missing heritability* i inclourien freqüentment gens relacionats amb processos fonamentals en diferents tipus cel·lulars, com són el metabolisme i la biosíntesi de molècules.

Bibliografia

[1] Goldgar, D., Easton, D., Cannon-Albright, L. & Skolnick, M. H. Systematic Population-Based Assessment of Cancer Risk in First-Degree Relatives of Cancer Probands. Journal of the National Cancer Institute 86, 1600–1608 (1994). 3

[2] Hemminki, K. & Vaittinen, P. Familial breast cancer in the family-cancer database. International Journal of Cancer 77, 386–391 (1998). 3

[3] Peto, J. & Mack, T. High constant incidence in twins and other relatives of women with breast cancer. Nature Genetics 26, 411–414 (2000). 3

[4] Ferlay, J. et al. Estimates of worldwide burden of cancer in 2008: GLOBOCAN 2008. International journal of cancer. 2893–2917 (2010). 3

[5] Visscher, P. M., Brown, M. A., McCarthy, M. I. & Yang, J. Five Years of GWAS Discovery. The American Journal of Human Genetics 90, 7–24 (2012). 3

[6] Reich, D. E. et al. Human genome sequence variation and the influence of gene history, mutation and recombination. Nature Genetics 32, 135–142 (2002). 4

[7] Consortium, T. . G. P. An integrated map of genetic variation from 1,092 human genomes. Nature 491, 56–65 (2012). 4, 13, 23, 29

[8] Brookes, A. The essence of SNPs. Gene 234, 177–186 (1999). 5

[9] Hoogendoorn, B. et al. Functional analysis of polymorphisms in the promoter regions of genes on 22q11. Human Mutation 24, 35–42 (2004). 5

[10] Sachidanandam, R. et al. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. Nature 409, 928–933 (2001). 5

[11] Cooper, G. M. & Shendure, J. Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data. Nature reviews. Genetics 12, 628–640 (2011). 5

[12] Daly, M. J., Rioux, J. D., Schaffner, S. F., Hudson, T. J. & Lander, E. S. High-resolution haplotype structure in the human genome. Nature Genetics 29, 229–232 (2001). 6

[13] Jeffreys, A. J., Kauppi, L. & Neumann, R. Intensely punctate meiotic recombination in the class ii region of the major histocompatibility complex. Nature Genetics 29, 217–222 (2001).

[14] Patil, N. et al. Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. Science 294, 1719–1723 (2001). 6

[15] Reich, D. E. et al. Linkage disequilibrium in the human genome. Nature 411, 199–204 (2001). 6

[16] Pääbo, S. The mosaic that is our genome. Nature 421, 409–412 (2003). 6

[17] Consortium, T. I. H. A haplotype map of the human genome. Nature 437, 1299–1320 (2005). 6, 22

[18] Consortium, T. I. H. The international hapmap project. Nature 426, 789–796 (2003). 7

[19] Foulkes, W. D. Inherited susceptibility to common cancers. The New England journal of medicine 359, 2143–2153 (2008). 7, 8

[20] Visscher, P. M., Hill, W. G. & Wray, N. R. Heritability in the genomics era -concepts and misconceptions. Nature reviews Genetics 9, 255–266 (2008). 7

[21] Michailidou, K. et al. Large-scale genotyping identifies 41 new loci associated with breast cancer risk. Nature Genetics 45, 353–361 (2013). 8, 9, 12, 14, 15, 17, 26, 29, 30, 103, 119, 121, 122

[22] Turnbull, C. & Rahman, N. Genetic predisposition to breast cancer: past, present, and future. Annual review of genomics and human genetics 9, 321–345 (2008). 8, 29

[23] Pharoah, P. D. P. et al. Polygenic susceptibility to breast cancer and implications for prevention. Nature Genetics 31, 33–36 (2002). 8

[24] Fletcher, O. & Houlston, R. S. Architecture of inherited susceptibility to common cancer. Nature reviews Cancer 10, 353–361 (2010). 10

[25] Futreal, P. et al. Brca1 mutations in primary breast and ovarian carcinomas. Science 266, 120–122 (1994). 10, 11

[26] Wooster, R. et al. Localization of a breast cancer susceptibility gene, brca2, to chromosome 13q12-13. Science 265, 2088–2090 (1994). 10, 11

[27] Foulkes, W. D. & Shuen, A. Y. In brief: Brca1 and brca2. The Journal of Pathology 230, 347–349 (2013). 10

[28] Antoniou, A. et al. Average risks of breast and ovarian cancer associated with BRCA1 or BRCA2 mutations detected in case Series unselected for family history: a combined analysis of 22 studies. American journal of human genetics 72, 1117–1130 (2003). 10, 11, 15

[29] Ford, D. et al. Genetic Heterogeneity and Penetrance Analysis of the BRCA1 and BRCA2 Genes in Breast Cancer Families. Am. J. Hum. Genet. 62, 676–689 (1998). 10

[30] Begg, C. B. et al. Variation of breast cancer risk among brca1/2 carriers. JAMA: The Journal of the American Medical Association 299, 194–201 (2008). 11, 15

[31] Simchoni, S. et al. Familial clustering of site-specific cancer risks associated with brca1 and brca2 mutations in the ashkenazi jewish population. Proceedings of the National Academy of Sciences of the United States of America 103, 3770–3774 (2006). 11

[32] Malkin, D. et al. Germ line p53 mutations in a familial syndrome of breast cancer, sarcomas, and other neoplasms. Science 250, 1233–1238 (1990). 11

[33] Nelen, M. et al. Localization of the gene for cowden disease to chromosome 10q22-23. Nature Genetics 13, 114–116 (1996). 11

[34] Hemminki, A. et al. A serine/threonine kinase gene defective in peutz-jeghers syndrome. Nature 391, 184–187 (1998). 11

[35] Miki, Y. et al. A strong candidate for the breast and ovarian cancer susceptibility gene brca1. Science 266, 66–71 (1994). 11

[36] Wooster, R. et al. Identification of the breast cancer susceptibility gene BRCA2. Nature 378, 789–792 (1995). 11

[37] Xia, B. et al. Fanconi anemia is associated with a defect in the BRCA2 partner PALB2. Nature Genetics 39, 159–161 (2007). 12

[38] Reid, S. et al. Biallelic mutations in PALB2 cause Fanconi anemia subtype FA-N and predispose to childhood cancer. Nature Genetics 39, 162–164 (2007).

[39] Wang, W. Emergence of a DNA-damage response network consisting of Fanconi anaemia and BRCA proteins. Nature Reviews Genetics 8, 735–748 (2007).

[40] Vaz, F. et al. Mutation of the RAD51C gene in a Fanconi anemia-like disorder. Nature Genetics 42, 406–409 (2010). 12

[41] Renwick, A. et al. Atm mutations that cause ataxia-telangiectasia are breast cancer susceptibility alleles. Nature Genetics 38, 873–875 (2006). 12

[42] Meijers-Heijboer, H. et al. Low-penetrance susceptibility to breast cancer due to chek2(*)1100delc in noncarriers of brca1 or brca2 mutations. Nature Genetics 31, 55–9 (2002). 12

[43] Seal, S. et al. Truncating mutations in the fanconi anemia j gene brip1 are low-penetrance breast cancer susceptibility alleles. Nature Genetics 38, 1239–41 (2006). 12

[44] Rahman, N. et al. Palb2, which encodes a brca2-interacting protein, is a breast cancer susceptibility gene. Nature Genetics 39, 165–167 (2007). 12

[45] Erkko, H. et al. A recurrent mutation in PALB2 in Finnish cancer families. Nature 7133, 316–319 (2007). 12

[46] Meindl, A. et al. Germline mutations in breast and ovarian cancer pedigrees establish RAD51C as a human cancer susceptibility gene. Nature Genetics 42, 410–414 (2010). 12

[47] Cox, A. et al. A common coding variant in CASP8 is associated with breast cancer risk. Nature genetics 39, 352–358 (2007). 12, 13, 21

[48] Thomas, G. et al. A multistage genome-wide association study in breast cancer identifies two new risk alleles at 1p11.2 and 14q24.1 (RAD51L1). Nature Genetics 41, 579–584 (2009). 12, 13, 14, 122

[49] Stacey, S. N. et al. Common variants on chromosomes 2q35 and 16q12 confer susceptibility to estrogen receptor-positive breast cancer. Nature Genetics 39, 865–869 (2007). 13, 117, 126

[50] Ahmed, S. et al. Newly discovered breast cancer susceptibility loci on 3p24 and 17q23.2. Nat Genet 41, 585–90 (2009). 13, 14

[51] Stacey, S. N. et al. Common variants on chromosome 5p12 confer susceptibility to estrogen receptor-positive breast cancer. Nature Genetics 40, 703–706 (2008). 13, 26, 120

[52] Haiman, C. A. et al. A common variant at the TERT-CLPTM1L locus is associated with estrogen receptor-negative breast cancer. Nature genetics 43, 1210–1214 (2011). 13, 26

[53] Easton, D. F. et al. Genome-wide association study identifies novel breast cancer susceptibility loci. Nature 447, 1087–1093 (2007). 13, 14, 26, 69, 117, 118, 120, 126

[54] Siddiq, A. et al. A meta-analysis of genome-wide association studies of breast cancer identifies two novel susceptibility loci at 6q14 and 20q11. Human Molecular Genetics (2012). 13, 14, 26

[55] Turnbull, C. et al. Genome-wide association study identifies five new breast cancer susceptibility loci. Nat Genet 42, 504–507 (2010). 13, 14, 69, 119

[56] Zheng, W. et al. Genome-wide association study identifies a new breast cancer susceptibility locus at 6q25.1. Nature Genetics 41, 324–328 (2009). 13, 26

[57] Fletcher, O. et al. Novel breast cancer susceptibility locus at 9q31.2: Results of a genome-wide association study. Journal of the National Cancer Institute 103, 425–435 (2011). 13

[58] Ghoussaini, M. et al. Genome-wide association analysis identifies three new breast cancer susceptibility loci. Nat Genet 44, 312–318 (2012). 14

[59] Antoniou, A. C. et al. A locus on 19p13 modifies risk of breast cancer in BRCA1 mutation carriers and is associated with hormone receptor-negative breast cancer in the general population. Nature genetics 42, 885–892 (2010). 12, 14, 16, 26, 29

[60] Bahcall, O. G. iCOGS collection provides a collaborative model. Nature Genetics 45, 343 (2013). 12, 84, 124

[61] Garcia-Closas, M. et al. Genome-wide association studies identify four ER negative-specific breast cancer risk loci. Nature Genetics 45, 392–398 (2013). 12, 14, 15, 26, 29

[62] Udler, M. S. et al. Fgfr2 variants and breast cancer risk: fine-scale mapping using african american studies and analysis of chromatin conformation. Human Molecular Genetics 18, 1692–1703 (2009). 13

[63] Huijts, P. et al. Allele-specific regulation of fgfr2 expression is cell type-dependent and may increase breast cancer risk through a paracrine stimulus involving fgf10. Breast Cancer Research 13, R72 (2011). 13

[64] French, J. D. et al. Functional Variants at the 11q13 Risk Locus for Breast Cancer Regulate Cyclin D1 Expression through Long-Range Enhancers. Am J Hum Genet 92, 489–503 (2013). 13, 23

[65] Cowper-SalÂ·lari, R. et al. Breast cancer risk-associated snps modulate the affinity of chromatin for foxa1 and alter gene expression. Nature Genetics 44, 1191–1198 (2012). 13

[66] Bojesen, S. E. et al. Multiple independent variants at the TERT locus are associated with telomere length and risks of breast and ovarian cancer. Nature Genetics 45, 371–384 (2013). 13, 23

[67] Chenevix-Trench, G. et al. An international initiative to identify genetic modifiers of cancer risk in BRCA1 and BRCA2 mutation carriers: the Consortium of Investigators of Modifiers of BRCA1 and BRCA2 (CIMBA). Breast Cancer Research 2, 104 (2007). 15, 84, 124

[68] Antoniou, A. C. et al. RAD51 135G–>C Modifies Breast Cancer Risk among BRCA2 Mutation Carriers: Results from a Combined Analysis of 19 Studies. Am J Hum Genet 81, 1186–1200 (2007). 16

[69] Engel, C. et al. Association of the variants casp8 d302h and casp10 v410i with breast and ovarian cancer risk in brca1 and brca2 mutation carriers. Cancer Epidemiology Biomarkers Prevention 19, 2859–2868 (2010). 16

[70] Maxwell, C. A. et al. Interplay between BRCA1 and RHAMM Regulates Epithelial Apicobasal Polarization and May Influence Risk of Breast Cancer. PLoS biology 9 (2011). 16

[71] Antoniou, A. C. et al. Common alleles at 6q25.1 and 1p11.2 are associated with breast cancer risk for brca1 and brca2 mutation carriers. Human Molecular Genetics 20, 3304–3321 (2011). 16

[72] Antoniou, A. C. et al. Common variants in lsp1, 2q35 and 8q24 and breast cancer risk for brca1 and brca2 mutation carriers. Human Molecular Genetics 18, 4442–4456 (2009). 16

[73] Antoniou, A. C. et al. Common breast cancer susceptibility alleles and the risk of breast cancer for brca1 and brca2 mutation carriers: Implications for risk prediction. Cancer Research 70, 9742–9754 (2010). 16

[74] Antoniou, A. C. C. et al. Common Breast Cancer-Predisposition Alleles Are Associated with Breast Cancer Risk in BRCA1 and BRCA2 Mutation Carriers. Am J Hum Genet (2008). 16

[75] Antoniou, A. C. et al. Common variants at 12p11, 12q24, 9p21, 9q31.2 and in ZNF365 are associated with breast cancer risk for BRCA1 and/or BRCA2 mutation carriers. Breast cancer research : BCR 14, R33+ (2012). 16

[76] Gaudet, M. M. et al. Common Genetic Variants and Modification of Penetrance of BRCA2-Associated Breast Cancer. PLoS Genet 6, e1001183+ (2010). 16

[77] Couch, F. J. et al. Genome-Wide Association Study in BRCA1 Mutation Carriers Identifies Novel Loci Associated with Breast and Ovarian Cancer Risk. PLoS genetics 9 (2013). 16, 123

[78] Gaudet, M. M. et al. Identification of a BRCA2-specific modifier locus at 6p24 related to breast cancer risk. PLoS genetics 9 (2013). 16, 17, 18

[79] Hunter, D., Thomas, G., Hoover, R. & Chanock, S. Scanning the horizon: What is the future of genome-wide association studies in accelerating discoveries in cancer etiology and prevention. Cancer Causes and Control 18, 479–484 (2007). 20

[80] Chanock, S. J. et al. Replicating genotype-phenotype associations. Nature 447, 655–660 (2007). 20, 21

[81] Dunning, A. M. et al. Common brca1 variants and susceptibility to breast and ovarian cancer in the general population. Human Molecular Genetics 6, 285–289 (1997). 21

[82] Kawajiri, K., Nakachi, K., Imai, K., Watanabe, J. & Hayashi, S.-I. Germ line polymorphisms of p53 and cyp1a1 genes involved in human lung cancer. Carcinogenesis 14, 1085–1089 (1993). 21

[83] Sjalander, A. et al. p53 polymorphisms and haplotypes in breast cancer. Carcinogenesis 17, 1313–1316 (1996). 21

[84] Anderson, T. et al. Oestrogen receptor (esr) polymorphisms and breast cancer susceptibility. Human Genetics 94, 665–670 (1994). 21

[85] Iwase, H. et al. Sequence variants of the estrogen receptor (er) gene found in breast cancer patients with er negative and progesterone receptor positive tumors. Cancer Letters 108, 179 – 184 (1996).

[86] Southey, M. C. et al. Estrogen receptor polymorphism at codon 325 and risk of breast cancer in women before age forty. Journal of the National Cancer Institute 90, 532–536 (1998). 21

[87] Huober, J., Bertram, B., Petru, E., Kaufmann, M. & Schmahl, D. Metabolism of debrisoquine and susceptibility to breast cancer. Breast Cancer Res Treat 18, 43–48 (1991). 21

[88] Ladero, J. et al. Polymorphic oxidation of debrisoquine in women with breast cancer. Oncology 48, 107–110 (1991). 21

[89] Hirschhorn, J. N., Lohmueller, K., Byrne, E. & Hirschhorn, K. A comprehensive review of genetic association studies. Genetics in medicine 4, 45–61 (2002). 21

[90] Consortium, T. I. H. A second generation human haplotype map of over 3.1 million snps. Nature 449, 851–861 (2007). 22

[91] Kruglyak, L. & Nickerson, D. A. Variation is the spice of life. Nat Genet 27, 234–236 (2001). 22

[92] Garcia-Closas, M. & Chanock, S. Genetic susceptibility loci for breast cancer by estrogen receptor status. Clinical cancer research 14, 8000–8009 (2008). 23, 26

[93] Risch, N. & Merikangas, K. The future of genetic studies of complex human diseases. Science 273, 1516–1517 (1996). 22

[94] Stratton, M., Campbell, P. & Futreal, P. The cancer genome. Nature 458, 719–724 (2009). 24, 25, 122

[95] Futreal, P. A. et al. A census of human cancer genes. Nature Reviews Cancer 4, 177–183 (2004). 24, 69

[96] Mallon, E. et al. The basic pathology of human breast cancer. J Mammary Gland Biol Neoplasia 2, 139–63 (2000). 26

[97] H, T. Individualization of breast cancer based on histopathological features and molecular alterations. Breast Cancer 2, 121–132 (2008).

[98] Cianfrocca, M. & Goldstein, L. J. Prognostic and predictive factors in early-stage breast cancer. The oncologist 9, 606–616 (2004). 26

[99] Utsumi, T., Kobayashi, N. & Hanada, H. Recent perspectives of endocrine therapy for breast cancer. Breast Cancer 14, 194–199 (2007). 26

[100] Borg, A. et al. Her-2/neu amplification predicts poor survival in node-positive breast cancer. Cancer Research 50, 4332–4337 (1990).

[101] Blows, F. M. et al. Subtyping of Breast Cancer by Immunohistochemistry to Investigate a Relationship between Subtype and Short and Long Term Survival: A Collaborative Analysis of Data for 10,159 Cases from 12 Studies. PLoS Med 7, e1000279+ (2010). 26

[102] Perou, C. M. et al. Molecular portraits of human breast tumours. Nature 406, 747–752 (2000). 26, 69

[103] Sørlie, T. et al. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. Proceedings of the National Academy of Sciences 98, 10869–10874 (2001). 26

[104] Korde, L. et al. Gene expression pathway analysis to predict response to neoadjuvant docetaxel and capecitabine for breast cancer. Breast Cancer Research and Treatment 119, 685–699 (2010). 26

[105] Williams, P. D. et al. Concordant gene expression signatures predict clinical outcomes of cancer patients undergoing systemic therapy. Cancer Research 69, 8302–8309 (2009). 26

[106] Anderson, W. F., Chu, K. C., Chang, S. & Sherman, M. E. Comparison of age-specific incidence rate patterns for different histopathologic types of breast carcinoma. Cancer Epidemiology Biomarkers Prevention 13, 1128–1135 (2004). 26

[107] Anderson, W., Jatoi, I. & Devesa, S. Distinct breast cancer incidence and prognostic patterns in the nci's seer program: suggesting a possible link between etiology and outcome. Breast Cancer Research and Treatment 90, 127–137 (2005).

[108] Althuis, M. D. et al. Etiology of hormone receptor-defined breast cancer: A systematic review of the literature. Cancer Epidemiology Biomarkers Prevention 13, 1558–1568 (2004). 26, 28

[109] Ma, H., Bernstein, L., Pike, M. & Ursin, G. Reproductive factors and breast cancer risk according to joint estrogen and progesterone receptor status: a meta-analysis of epidemiological studies. Breast Cancer Research 8, R43 (2006).

[110] Reeves, G. K., Beral, V., Green, J., Gathani, T. & Bull, D. Hormonal therapy for menopause and breast-cancer risk by histological type: a cohort study and meta-analysis. The Lancet Oncology 7, 910 – 918 (2006). 26

[111] Lakhani, S. R. et al. The pathology of familial breast cancer: Predictive value of immunohistochemical markers estrogen receptor, progesterone receptor, her-2, and p53 in patients with mutations in brca1 and brca2. Journal of Clinical Oncology 20, 2310–2318 (2002). 26

[112] Lakhani, S. R. et al. Multifactorial analysis of differences between sporadic breast cancers and cancers involving brca1 and brca2 mutations. Journal of the National Cancer Institute 90, 1138–1145 (1998). 26

[113] Milne, R. L. & Antoniou, A. C. Genetic modifiers of cancer risk for brca1 and brca2 mutation carriers. Annals of Oncology 22, i11–i17 (2011). 27

[114] Maher, B. Personal genomes: The case of the missing heritability. Nature 456, 18–21 (2008). 28

[115] Reich, D. E. & Lander, E. S. On the allelic spectrum of human disease. 17, 502–510 (2001). 28, 117

[116] Pritchard, J. K. Are rare variants responsible for susceptibility to complex diseases? American journal of human genetics 69, 124–137 (2001). 28

[117] Mavaddat, N., Antoniou, A., Easton, D. & Garcia-Closas, M. Genetic susceptibility to breast cancer. Molecular oncology 4, 174–191 (2010). 29

[118] Mavaddat, N., Dunning, A. M., Ponder, B. A., Easton, D. F. & Pharoah, P. D. Common genetic variation in candidate genes and susceptibility to subtypes of breast cancer. Cancer Epidemiology Biomarkers & Prevention 18, 255–259 (2009). 29

[119] Zheng, W. et al. Genome-wide association study identifies a new breast cancer susceptibility locus at 6q25.1. Nature Genetics 41, 324–328 (2009). 30

[120] Okada, Y. et al. Genetics of rheumatoid arthritis contributes to biology and drug discovery. Nature 506, 376–381 (2014). 30

[121] Gibson, G. Rare and common variants: twenty arguments. Nature Reveiws Genetics 13, 135–145 (2012). 30, 103, 126

[122] Manolio, T. A. et al. Finding the missing heritability of complex diseases. Nature 461, 747–753 (2009). 30, 35, 49, 124

[123] Cusick, M. E., Klitgord, N., Vidal, M. & Hill, D. E. Interactome: gateway into systems biology. Human Molecular Genetics 14, R171–R181 (2005). 31

[124] Stelzl, U. et al. A Human Protein-Protein Interaction Network: A Resource for Annotating the Proteome. Cell 122, 957–968 (2005). 31

[125] Rual, J.-F. et al. Towards a proteome-scale map of the human protein-protein interaction network. Nature 437, 1173–1178 (2005). 31

[126] Keshava Prasad, T. S. et al. Human protein reference database-2009 update. Nucleic Acids Research 37, D767–D772 (2009). 31, 49, 125

[127] Barabási, A.-L. & Albert, R. Emergence of Scaling in Random Networks. Science 286, 509–512 (1999). 31

[128] Barabasi, A. & Oltvai, Z. Network biology: understanding the cell's functional organization. Nature Review Genetics 5, 101–113 (2004). 32, 33, 119

[129] Yook, S.-H., Oltvai, Z. N. & Barabási, A.-L. Functional and topological characterization of protein interaction networks. Proteomics 4, 928–942 (2004).

[130] Rachlin, J., Cohen, D. D., Cantor, C. & Kasif, S. Biological context networks: a mosaic view of the interactome. Molecular Systems Biology 2 (2006). 32

[131] Joy, M. P. P., Brock, A., Ingber, D. E. & Huang, S. High-betweenness proteins in the yeast protein interaction network. Journal of biomedicine & biotechnology 2005, 96–103 (2005). 32

[132] Watts, D. J. & Strogatz, S. H. Collective dynamics of "small-world" networks. Nature 6684, 440–442 (1998). 32

[133] Hartwell, L., Hopfield, J., Leibler, S. & Murray, A. From molecular to modular cell biology. Nature 402(6761 Suppl), C47–52 (1999). 32

[134] Gagneur, J., Krause, R., Bouwmeester, T. & Casari, G. Modular decomposition of protein-protein interaction networks. Genome Biol 5 (2004).

[135] Spirin, V. & Mirny, L. A. Protein complexes and functional modules in molecular networks. Proceedings of the National Academy of Sciences 100, 12123–12128 (2003). 32

[136] Brunner, H. G. & van Driel, M. A. From syndrome families to functional genomics. Nature Reviews Genetics 5, 545–551 (2004). 33

[137] Loscalzo, J., Kohane, I. & Barabasi, A.-L. Human disease classification in the postgenomic era: A complex systems approach to human pathobiology. Mol Syst Biol 3 (2007). 33

[138] Kanehisa, M. & Goto, S. Kegg: Kyoto encyclopedia of genes and genomes. Nucleic Acids Research 28, 27–30 (2000). 33

[139] Nishimura, D. A view from the web biocarta. Biotech Software & Internet Report 2, 117–120 (2001). 33

[140] Ashburner, M. et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nature genetics 25, 25–29 (2000). 33, 49

[141] Mirina, A., Atzmon, G., Ye, K. & Bergman, A. Gene size matters. PloS one 7, e49093+ (2012). 34, 118

[142] Holmans, P. et al. Gene ontology analysis of gwa study data sets provides insights into the biology of bipolar disorder. The American Journal of Human Genetics 85, 13 – 24 (2009). 34

[143] Yu, K. et al. Pathway analysis by adaptive combination of P-values. Genetic epidemiology 33, 700–709 (2009).

[144] Wang, K., Li, M. & Bucan, M. Pathway-Based Approaches for Analysis of Genomewide Association Studies. Am J Hum Genet 81, 1278–1283 (2007). 34, 118

[145] Holmans, P. 7 - statistical methods for pathway analysis of genome-wide data for association with complex genetic traits. In Dunlap, J. C. & Moore, J. H. (eds.) Computational Methods for Genetics of Complex Traits, vol. 72 of Advances in Genetics, 141 – 179 (Academic Press, 2010). 34

[146] Wang, K., Li, M. & Hakonarson, H. Analysing biological pathways in genome-wide association studies. Nat Rev Genet 11, 843–854 (2010). 34

[147] Subramanian, A. et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences of the United States of America 102, 15545–15550 (2005). 34

[148] Al-Shahrour, F., Díaz-Uriarte, R. & Dopazo, J. Discovering molecular functions significantly related to phenotypes by combining gene expression data and biological information. Bioinformatics 21, 2988–2993 (2005). 34, 118

[149] Holmans, P. Statistical methods for pathway analysis of genome-wide data for association with complex genetic traits. In Dunlap, J. C. & Moore, J. H. (eds.) Computational Methods for Genetics of Complex Traits, vol. 72, 141 – 179 (Academic Press, 2010). 34

[150] Altshuler, D., Daly, M. & ES., L. Genetic mapping in human disease. Science (New York, N.Y.) 322, 881–888 (2008). 34

[151] Benfey, P. N. & Mitchell-Olds, T. From Genotype to Phenotype: Systems Biology Meets Natural Variation. Science 320, 495–497 (2008). 34

[152] Cordell, H. J. Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans. Human Molecular Genetics 11, 2463–2468 (2002). 35

[153] Moore, J. H. & Williams, S. M. Epistasis and its implications for personal genetics. The American Journal of Human Genetics 85, 309 – 320 (2009). 35

[154] Maxwell, C. et al. Genetic interactions: the missing links for a better understanding of cancer susceptibility, progression and treatment. Molecular Cancer 7, 1–10 (2008). 35

[155] Hartman, J. L., Garvik, B. & Hartwell, L. Principles for the Buffering of Genetic Variation. Science 291, 1001–1004 (2001). 35, 103

[156] Boone, C., Bussey, H. & Andrews, B. J. Exploring genetic interactions and networks with yeast. Nature Reviews Genetics 8, 437–449 (2007). 35, 103, 125

[157] Roguev, A. et al. Quantitative genetic-interaction mapping in mammalian cells. Nature Methods 10, 432–437 (2013). 35, 125

[158] Laufer, C., Fischer, B., Billmann, M., Huber, W. & Boutros, M. Mapping genetic interactions in human cancer cells with RNAi and multiparametric phenotyping. Nat Meth 10, 427–431 (2013). 35, 125

[159] Cordell, H. J. Detecting gene-gene interactions that underlie human diseases. Nature reviews. Genetics 10, 392–404 (2009). 35, 124

[160] Musani, S. K. et al. Detection of GenexGene Interactions in Genome-Wide Association Studies of Human Population Data. Human Heredity 63, 67–84 (2007).

[161] Thornton-Wells, T. A., Moore, J. H. & Haines, J. L. Genetics, statistics and human disease: analytical retooling for complexity. Trends in genetics : TIG 20, 640–647 (2004).

[162] Moore, J. H., Asselbergs, F. W. & Williams, S. M. Bioinformatics challenges for genome-wide association studies. Bioinformatics 26, 445–455 (2010). 35, 103

[163] Kam-Thong, T. et al. EPIBLASTER-fast exhaustive two-locus epistasis detection strategy using graphical processing units. Eur J Hum Genet 19, 465–471 (2011). 36, 103, 124

[164] Yung, L. S., Yang, C., Wan, X. & Yu, W. Gboost: a gpu-based tool for detecting gene-gene interactions in genome-wide case control studies. Bioinformatics 27, 1309–1310 (2011). 36

[165] Evans, D. M., Marchini, J., Morris, A. P. & Cardon, L. R. Two-Stage Two-Locus Models in Genome-Wide Association. PLoS Genet 2, e157+ (2006). 36

[166] Dong, C. et al. Exploration of gene-gene interaction effects using entropy-based methods. European journal of human genetics : EJHG 16, 229–235 (2008). 36

[167] Moore, J. & White, B. Tuning ReliefF for Genome-Wide Genetic Analysis. 166–175 (2007).

[168] Greene, C., Penrod, N., Kiralis, J. & Moore, J. Spatially Uniform ReliefF (SURF) for computationally-efficient filtering of gene-gene interactions. BioData Mining 2, 5+ (2009). 36

[169] Pattin, K. & Moore, J. Exploiting the proteome to improve the genome-wide genetic analysis of epistasis in common human diseases. Human Genetics 124, 19–29 (2008). 36

[170] Emily, M., Mailund, T., Hein, J., Schauser, L. & Schierup, M. H. Using biological networks to search for interacting loci in genome-wide association studies. European Journal of Human Genetics 17, 1231–1240 (2009). 36

[171] Greenland, S. Interactions in epidemiology: relevance, identification, and estimation. Epidemiology 20, 14–17 (2009). 36

[172] Moore, J. H. & Williams, S. M. Traversing the conceptual divide between biological and statistical epistasis: systems biology and a more modern synthesis. BioEssays 27, 637–646 (2005). 36

[173] Hindorff, L. A. et al. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proceedings of the National Academy of Sciences of the United States of America 106, 9362–9367 (2009). 49

[174] Houlston, R. S. & Peto, J. The search for low-penetrance cancer susceptibility alleles. Oncogene 23, 6471–6476 (2004). 49, 117

[175] Hunter, D. J. et al. A genome-wide association study identifies alleles in FGFR2 associated with risk of sporadic postmenopausal breast cancer. Nature genetics 39, 870–874 (2007). 49, 69, 70, 103, 117, 120, 121

[176] Richardson, A. L. et al. X chromosomal abnormalities in basal-like human breast cancer. Cancer Cell 9, 121–132 (2006). 50

[177] Furuta, S. et al. Removal of brca1/ctip/zbrk1 repressor complex on ang1 promoter leads to accelerated mammary tumor growth contributed by prominent vasculature. Cancer Cell 10, 13 – 24 (2006). 50

[178] Chang, H. Y. et al. Robustness, scalability, and integration of a wound-response gene expression signature in predicting breast cancer survival. Proceedings of the National Academy of Sciences of the United States of America 102, 3738–3743 (2005). 50, 69

[179] Hanahan, D. & Weinberg, R. A. The Hallmarks of Cancer. Cell 100, 57–70 (2000). 69, 119

[180] Eswarakumar, V. P., Lax, I. & Schlessinger, J. Cellular signaling by fibroblast growth factor receptors. Cytokine & growth factor reviews 16, 139–149 (2005). 69

[181] Cuevas, B. D., Winter-Vann, A. M., Johnson, N. L. & Johnson, G. L. MEKK1 controls matrix degradation and tumor cell dissemination during metastasis of polyoma middle-T driven mammary cancer. Oncogene 25, 4998–5010 (2006).

[182] Kamb, A., Gruis, N. & Weaver-Feldhaus, J. A cell cycle regulator potentially involved in genesis of many tumor types. Science (1994).

[183] Caldon, C. E., Daly, R. J., Sutherland, R. L. & Musgrove, E. A. Cell cycle control in breast cancer cells. Journal of Cellular Biochemistry 97, 261–274 (2006). 69

[184] van't Veer, L. J. et al. Gene expression profiling predicts clinical outcome of breast cancer. Nature 415, 530–536 (2002). 69

[185] Chi, J. T. et al. Gene expression programs in response to hypoxia: cell type specificity and prognostic significance in human cancers. PLoS Med 3 (2006).

[186] Liu, R. et al. The Prognostic Role of a Gene Signature from Tumorigenic Breast-Cancer Cells. N Engl J Med 356, 217–226 (2007).

[187] Minn, A. J. et al. Genes that mediate breast cancer metastasis to lung. Nature 436, 518–524 (2005).

[188] Ramaswamy, S., Ross, K. N., Lander, E. S. & Golub, T. R. A molecular signature of metastasis in primary solid tumors. Nature Genetics 33, 49–54 (2002). 69

[189] Ayers, M. et al. Gene expression profiles predict complete pathologic response to neoadjuvant paclitaxel and fluorouracil, doxorubicin, and cyclophosphamide chemotherapy in breast cancer. Journal of clinical oncology 22, 2284–2293 (2004). 69

[190] Ma, X. J. et al. A two-gene expression ratio predicts clinical outcome in breast cancer patients treated with tamoxifen. Cancer cell 5, 607–616 (2004).

[191] Chang, J. C. et al. Patterns of resistance and incomplete response to docetaxel by gene expression profiling in breast cancer patients. Journal of Clinical Oncology 23, 1169–1177 (2005). 69

[192] Greenman, C. et al. Patterns of somatic mutation in human cancer genomes. Nature 446, 153–158 (2007). 69, 122

[193] McMurray, H. R. et al. Synergistic response to oncogenic mutations defines gene class critical to cancer phenotype. Nature 453, 1112–1116 (2008).

[194] Forbes, S. A. et al. COSMIC (the Catalogue of Somatic Mutations in Cancer): a resource to investigate acquired mutations in human cancer. Nucleic acids research 38 (2010).

[195] Stephens, P. J. et al. Complex landscapes of somatic rearrangement in human breast cancer genomes. Nature 462, 1005–1010 (2009).

[196] Santarius, T., Shipley, J., Brewer, D., Stratton, M. R. & Cooper, C. S. A census of amplified and overexpressed human cancer genes. Nat Rev Cancer 10, 59–64 (2010). 69

[197] Emery, L. A. et al. Early dysregulation of cell adhesion and extracellular matrix pathways in breast cancer progression. The American Journal of Pathology 175, 1292 – 1302 (2009). 70

[198] Bonifaci, N. et al. Exploring the link between germline and somatic genetic alterations in breast carcinogenesis. PLoS ONE 5, e14078 (2010). 81

[199] Antoniou, A. C. C. et al. Common Breast Cancer-Predisposition Alleles Are Associated with Breast Cancer Risk in BRCA1 and BRCA2 Mutation Carriers. Am J Hum Genet 82, 937–948 (2008). 81, 123

[200] TCGA. Comprehensive molecular portraits of human breast tumours. Nature 490, 61–70 (2012). 83

[201] Varghese, J. S. S. & Easton, D. F. Genome-wide association studies in common cancers–what have we learnt? Current opinion in genetics & development 20, 201–209 (2010). 103

[202] Quigley, D. & Balmain, A. Systems genetics analysis of cancer susceptibility: from mouse models to humans. Nat Rev Genet 10, 651–657 (2009). 103

[203] van de Vijver, M. J. et al. A Gene-Expression Signature as a Predictor of Survival in Breast Cancer. N Engl J Med 347, 1999–2009 (2002). 104

[204] Jorissen, R. N. et al. Metastasis-associated gene expression changes predict poor outcomes in patients with dukes stage b and c colorectal cancer. Clinical Cancer Research 15, 7642–7651 (2009). 105

[205] Walker, L. C. et al. Evidence for SMAD3 as a modifier of breast cancer risk in BRCA2 mutation carriers. Breast cancer research : BCR 12, R102+ (2010). 105

[206] Massagué, J. & Xi, Q. Tgf-β control of stem cell differentiation genes. FEBS Letters 586, 1953 – 1958 (2012). 105, 126

[207] Biankin, A. V. et al. Pancreatic cancer genomes reveal aberrations in axon guidance pathway genes. Nature 491, 399–405 (2012). 106

[208] Amundadottir, L. et al. Genome-wide association study identifies variants in the abo locus associated with susceptibility to pancreatic cancer. Nature Genetics 41, 986–990 (2009). 106

[209] Petersen, G. M. et al. A genome-wide association study identifies pancreatic cancer susceptibility loci on chromosomes 13q22.1, 1q32.1 and 5p15.33. Nature genetics 42, 224–228 (2010). 106

[210] Stratton, M. & Rahman, N. The emerging landscape of breast cancer susceptibility. Nature genetics 40, 17–22 (2008). 117

[211] Tomlinson, I. et al. A genome-wide association scan of tag SNPs identifies a susceptibility variant for colorectal cancer at 8q24.21. Nature Genetics 39, 984–988 (2007). 118

[212] Gudmundsson, J. et al. Genome-wide association study identifies a second prostate cancer susceptibility variant at 8q24. Nature Genetics 39, 631–637 (2007).

[213] Kiemeney, L. A. A. et al. Sequence variant on 8q24 confers susceptibility to urinary bladder cancer. Nature genetics (2008).

[214] Shete, S. et al. Genome-wide association study identifies five susceptibility loci for glioma. Nature genetics 41, 899–904 (2009). 118

[215] Hong, M.-G., Pawitan, Y., Magnusson, P. & Prince, J. Strategies and issues in the detection of pathway enrichment in genome-wide association studies. Human Genetics 126, 289–301 (2009). 118

[216] Cantor, R. M., Lange, K. & Sinsheimer, J. S. Prioritizing GWAS Results: A Review of Statistical Methods and Recommendations for Their Application. The American Journal of Human Genetics 86, 6–22 (2010). 118

[217] Furney, S., Higgins, D., Ouzounis, C. & Lopez-Bigas, N. Structural and functional properties of genes involved in human cancer. BMC Genomics 7 (2006). 118

[218] Ramanan, V. K., Shen, L., Moore, J. H. & Saykin, A. J. Pathway analysis of genomic data: concepts, methods, and prospects for future development. Trends in Genetics 28, 323–332 (2012). 118

[219] Fehringer, G. et al. Comparison of pathway analysis approaches using lung cancer GWAS data sets. PloS one 7 (2012). 118

[220] Al-Shahrour, F. et al. From genes to functional classes in the study of biological systems. BMC Bioinformatics 8, 114 (2007). 118

[221] Menashe, I. et al. Pathway analysis of breast cancer genome-wide association study highlights three pathways and one canonical signaling cascade. Cancer Research 70, 4453–4459 (2010). 119

[222] Garraway, L. A. et al. Integrative genomic analyses identify MITF as a lineage survival oncogene amplified in malignant melanoma. Nature 436, 117–122 (2005). 120

[223] Pujana, M. A. et al. Network modeling links breast cancer susceptibility and centrosome dysfunction. Nat Genet 39, 1338–1349 (2007). 120

[224] Walker, L., Waddell, N., Ten Haaf, A., Grimmond, S. & Spurdle, A. Use of expression data and the cgems genome-wide breast cancer association study to identify genes that may modify risk in brca1/2 mutation carriers. Breast Cancer Research and Treatment 112, 229–236 (2008). 120

[225] Untergasser, G. et al. The dickkopf-homolog 3 is expressed in tumor endothelial cells and supports capillary formation. International Journal of Cancer 122, 1539–1547 (2008). 120

[226] Guo, H. et al. Tissue factor pathway inhibitor-2 was repressed by cpg hypermethylation through inhibition of klf6 binding in highly invasive breast cancer cells. BMC Molecular Biology 8, 1–11 (2007). 120

[227] Renehan, A. G., Harvie, M. & Howell, A. Insulin-like growth factor (igf)-i, igf binding protein-3, and breast cancer risk: eight years on. Endocrine-Related Cancer 13, 273–278 (2006). 120

[228] Bachmann, H. S. et al. The aa genotype of the regulatory bcl2 promoter polymorphism (-938c>a) is associated with a favorable outcome in lymph node-negative invasive breast cancer patients. Clinical Cancer Research 13, 5790–5797 (2007). 120

[229] Hassan, S., Baccarelli, A., Salvucci, O. & Basik, M. Plasma stromal cell-derived factor-1: Host derived marker predictive of distant metastasis in breast cancer. Clinical Cancer Research 14, 446–454 (2008). 120

[230] Hsu, E. L. et al. Cxcr4 and cxcl12 down-regulation: A novel mechanism for the chemoprotection of 3,3′-diindolylmethane for breast and ovarian cancers. Cancer Letters 265, 113 – 123 (2008).

[231] Wendt, M. K., Cooper, A. N. & Dwinell, M. B. Epigenetic silencing of cxcl12 increases the metastatic potential of mammary carcinoma cells. Oncogene 27, 1461 – 1471 (2007). 120

[232] Pupa, S. M. et al. Immunological and pathobiological roles of fibulin-1 in breast cancer. Oncogene 23, 2153 – 2160 (2004). 120

[233] Greene, L. M. et al. Elevated expression and altered processing of fibulin-1 protein in human breast cancer. Br J Cancer 88, 871 – 878 (2003). 120

[234] Unoki, M. & Nakamura, Y. Growth-suppressive effects of BPOZ and EGR2, two genes involved in the PTEN signaling pathway. Oncogene 20, 4457–4465 (2001). 120

[235] Stephens, P. et al. A screen of the complete protein kinase gene family identifies diverse patterns of somatic mutations in human breast cancer. Nature Genetics 37, 590–592 (2005). 122

[236] Taglienti, C., Wysk, M. & Davis, R. Molecular cloning of the epidermal growth factor-stimulated protein kinase p56 kkiamre. Oncogene 13, 2563–2574 (1996). 122

[237] Taira, N., Yamamoto, H., Yamaguchi, T., Miki, Y. & Yoshida, K. Atm augments nuclear stabilization of dyrk2 by inhibiting mdm2 in the apoptotic

response to dna damage. Journal of Biological Chemistry 285, 4909–4919 (2010). 122, 123

[238] Merlos-Suárez, A. & Batlle, E. Eph-ephrin signalling in adult tissues and cancer. current Opinion in Cell Biology 20, 194–200 (2008). 123

[239] Vaught, D., Brantley-Sieders, D. & Chen, J. Eph receptors in breast cancer: roles in tumor promotion and tumor suppression. Breast Cancer Research 10, 1–8 (2008). 123

[240] Brantley-Sieders, D. M. et al. The receptor tyrosine kinase epha2 promotes mammary adenocarcinoma tumorigenesis and metastatic progression in mice by amplifying erbb2 signaling. The Journal of Clinical Investigation 118, 64–78 (2008). 123

[241] Cortina, C. et al. Ephb-ephrin-b interactions suppress colorectal cancer progression by compartmentalizing tumor cells. Nature Genetics 39, 1376 – 1383 (2007). 123

[242] Schnitt, S. The transition from ductal carcinoma in situ to invasive breast cancer: the other side of the coin. Breast Cancer Research 11, 1–2 (2009). 123

[243] Lim, E. et al. Transcriptome analyses of mouse and human mammary cell subpopulations reveal multiple conserved genes and pathways. Breast Cancer Research 12 (2010). 124

[244] Kendrick, H. et al. Transcriptome analysis of mammary epithelial subpopulations identifies novel determinants of lineage commitment and cell fate. BMC Genomics 9 (2008). 124

[245] Kaenel, P., Mosimann, M. & Andres, A. The multifaceted roles of eph/ephrin signaling in breast cancer. Cell Adhesion & Migration 6 (2012). 124

[246] Kaenel, P. et al. Deregulated ephrin-b2 signaling in mammary epithelial cells alters the stem cell compartment and interferes with the epithelial differentiation pathway. International Journal of Oncology 40, 357–369 (2012). 124

[247] Van Steen, K. Travelling the world of gene-gene interactions. Briefings in Bioinformatics 13, 1–19 (2012). 124

[248] Costanzo, M. et al. The Genetic Landscape of a Cell. Science 327, 425–431 (2010). 125

[249] Hart, G. T., Ramani, A. & Marcotte, E. How complete are current yeast and human protein-interaction networks? Genome Biology 7, 120 (2006). 125

[250] Goldstein, D. B. Common genetic variation and human traits. New England Journal of Medicine 360, 1696–1698 (2009). 126

Annex I

Altres publicacions

1. **VAV3 mediates resistance to breast cancer endocrine therapy.**
   Aguilar, H., A. Urruticoechea, P. Halonen, K. Kiyotani, T. Mushiroda, X. Barril, J. Serra-Musach, A. Islam, L. Caizzi, L. Di Croce, E. Nevedomskaya, W. Zwart, J. Bostner, E. Karlsson, G. Perez Tenorio, T. Fornander, D. Sgroi, R. Garcia Mata, M. Jansen, N. Garcia, **N. Bonifaci**, F. Climent, M. Soler, A. Rodriguez-Vida, M. Gil, J. Brunet, G. Martrat, L. Gomez-Baldo, A. Extremera, and A. Figueras.

   Breast Cancer Research (2014) 16(3), R53.

2. **Integrated genomic and prospective clinical studies show the importance of modular pleiotropy for disease susceptibility, diagnosis and treatment.**
   Gustafsson, M., M. Edstrom, D. Gawel, C. Nestor, H. Wang, H. Zhang, F. Barrenas, J. Tojo, I. Kockum, T. Olsson, J. Serra-Musach, **N. Bonifaci**, M. Pujana, J. Ernerudh, and M. Benson.

   Genome Medicine (2014) 6(2), 17.

3. **Evaluation of PAX3 genetic variants and nevus number.**
   Ogbah, Z., C. Badenas, M. Harland, J. A. Puig-Butille, F. Elliot, **N. Bonifaci**, E. Guino, J. Randerson-Moor, M. Chan, M. M. Iles, D. Glass, A. A. Brown, C. Carrera, I. Kolm, V. Bataille, T. D. Spector, J. Malvehy, J. Newton-Bishop, M. A. Pujana, T. Bishop, and S. Puig.

   Pigment Cell & Melanoma Research (2013) 26(5), 666-676.

4. **Serum 25-hydroxyvitamin D3 levels and vitamin D receptor variants in melanoma patients from the Mediterranean area of Barcelona.**
   Ogbah, Z., L. Visa, C. Badenas, J. Rios, J. A. Puig-Butille, **N. Bonifaci**, E. Guino, J. M. Auge, I. Kolm, C. Carrera, M. A. Pujana, J. Malvehy, and S. Puig.

   BMC Medical Genetics (2013) 14(1), 26.

5. **Interplay between BRCA1 and RHAMM Regulates Epithelial Apicobasal Polarization and May Influence Risk of Breast Cancer.**
   Maxwell, C. A., J. Benítez, L. Gómez-Baldó, A. Osorio, **N. Bonifaci**, R. Fernández-Ramires, S. V. Costes, E. Guinó, H. Chen, G. J. R. Evans, P. Mohan, I. Català, A. Petit, H. Aguilar, A. Villanueva, A. Aytes, J. Serra-Musach,

G. Rennert, F. Lejbkowicz, P. Peterlongo, S. Manoukian, B. Peissel, C. B. Ripamonti, B. Bonanni, A. Viel, A. Allavena, L. Bernard, P. Radice, E. Friedman, B. Kaufman, Y. Laitman, M. Dubrovsky, R. Milgrom, A. Jakubowska, C. Cybulski, B. Gorski, K. Jaworska, K. Durda, G. Sukiennicki, J. Lubinski, Y. Y. Shugart, S. M. Domchek, R. Letrero, B. L. Weber, F. B. L. Hogervorst, M. A. Rookus, J. M. Collee, P. Devilee, M. J. Ligtenberg, R. B. van der Luijt, C. M. Aalfs, Q. Waisfisz, J. Wijnen, C. E. P. van Roozendaal, D. F. Easton, S. Peock, M. Cook, C. Oliver, D. Frost, P. Harrington, D. G. Evans, F. Lalloo, R. Eeles, L. Izatt, C. Chu, D. Eccles, F. Douglas, C. Brewer, H. Nevanlinna, T. Heikkinen, F. J. Couch, N. M. Lindor, X. Wang, A. K. Godwin, M. A. Caligo, G. Lombardi, N. Loman, P. Karlsson, H. Ehrencrona, A. von Wachenfeldt, R. Bjork Barkardottir, U. Hamann, M. U. Rashid, A. Lasa, T. Caldés, R. Andrés, M. Schmitt, V. Assmann, K. Stevens, K. Offit, J. Curado, H. Tilgner, R. Guigó, G. Aiza, J. Brunet, J. Castellsagué, G. Martrat, A. Urruticoechea, I. Blanco, L. Tihomirova, D. E. Goldgar, S. Buys, E. M. John, A. Miron, M. Southey, M. B. Daly, R. K. Schmutzler, B. Wappenschmidt, A. Meindl, N. Arnold, H. Deissler, R. Varon-Mateeva, C. Sutter, D. Niederacher, E. Imyamitov, O. M. Sinilnikova, D. Stoppa-Lyonne, S. Mazoyer, C. Verny-Pierre, L. Castera, A. de Pauw, Y.-J. Bignon, N. Uhrhammer, J.-P. Peyrat, P. Vennin, S. Fert Ferrer, M.-A. Collonge-Rame, I. Mortemousque, A. B. Spurdle, J. Beesley, X. Chen, S. Healey, M. H. Barcellos-Hoff, M. Vidal, S. B. Gruber, C. Lázaro, G. Capellá, L. McGuffog, K. L. Nathanson, A. C. Antoniou, G. Chenevix-Trench, M. C. Fleisch, V. Moreno, and M. A. Pujana.

PLoS Biology (2011) 9(11), e1001199.

6. **Evidence for a link between TNFRSF11A and risk of breast cancer.**

   Bonifaci, N, M. Palafox, P. Pellegrini, A. Osorio, J. Benítez, P. Peterlongo, S. Manoukian, B. Peissel, D. Zaffaroni, G. Roversi, M. Barile, A. Viel, F. Mariette, L. Bernard, P. Radice, B. Kaufman, Y. Laitman, R. Milgrom, E. Friedman, M. Sáez, F. Climent, M. Soler, O. Diez, J. Balmaña, A. Lasa, T. Ramón y Cajal, M.-D. Miramar, M. de la Hoya, P. Pérez-Segura, T. Caldés, V. Moreno, A. Urruticoechea, J. Brunet, C. Lázaro, I. Blanco, M. Pujana, and E. González-Suárez.

   Breast Cancer Research and Treatment (2011) 129(3), 947-954.

7. **Exploring the link between MORF4L1 and risk of breast cancer.**

   G. Martrat, C. Maxwell, E. Tominaga, M. Porta-de-la Riva, **N. Bonifaci**, L. Gomez-Baldo, M. Bogliolo, C. Lazaro, I. Blanco, J. Brunet, H. Aguilar, J. Fernandez-Rodriguez, S. Seal, A. Renwick, N. Rahman, J. Kuhl, K. Neveling, D. Schindler, M. Ramirez, M. Castella, G. Hernandez, D. Easton, S. Peock, M. Cook, C. Oliver, D. Frost, R. Platte, D. G. Evans, and F. Lalloo.

Breast Cancer Research (2011) 13(2), R40.

8. **Biological reprogramming in acquired resistance to endocrine therapy of breast cancer.**
   Aguilar, H, X Solé, **N Bonifaci**, J Serra-Musach, A Islam, N López-Bigas, M Méndez-Pertuz, R. L. Beijersbergen, C Lázaro, A Urruticoechea, and M. A. Pujana.

   Oncogene (2010) 11;29(45):6071-83.

9. **TACC3-TSC2 maintains nuclear envelope structure and controls cell division.**
   Gomez-Baldo, L., S. Schmidt, C. A. Maxwell, **N. Bonifaci**, T. Gabaldon, P. O. Vidalain, W. Senapedis, A. Kletke, M. Rosing, A. Barnekow, R. Rottapel, G. Capella, M. Vidal, A. Astrinidis, R. P. Piekorz, and M. A. Pujana

   Cell Cycle (2010) 9(6), 1143-1155.

10. **Gene expression profiling integrated into network modelling reveals heterogeneity in the mechanisms of BRCA1 tumorigenesis.**
    Fernandez-Ramires, R, X Sole, L De Cecco, G Llort, A Cazorla, **N Bonifaci**, M. J. Garcia, T Caldes, I Blanco, M Gariboldi, M. A. Pierotti, M. A. Pujana, J Benitez, and A Osorio.

    British Journal of Cancer (2009) (8), 1469-1480.

11. **Gene set-based analysis of polymorphisms: finding pathways or biological processes associated to traits in genome-wide association studies.**
    Medina, I., D. Montaner, **N. Bonifaci**, M. A. Pujana, J. Carbonell, J. Tarraga, F. Al-Shahrour, and J. Dopazo.

    Nucleic Acids Research (2009) 37(supp-2), 340-344.

12. **Biological Convergence of Cancer Signatures.**
    Solé, X., **N. Bonifaci**, N. López-Bigas, A. Berenguer, P. Hernández, O. Reina, C. A. Maxwell, H. Aguilar, A. Urruticoechea, S. de Sanjosé, F. Comellas, G. Capellá, V. Moreno, and M. A. Pujana

    PLoS ONE (2009) 4(2), e4544.

Annex II

Informe del director

IDIBELL
Institut d'Investigació Biomèdica de Bellvitge

Informe del director sobre la tesi doctoral

Nom del director de la tesi
MIQUEL ANGEL PUJANA
Universitat o organisme del director
INSTITUT CATALÀ D'ONCOLOGIA, IDIBELL

Nom del tutor de la tesis
VÍCTOR MORENO AGUADO

Departament o Institut de la inscripció de la tesi	Estudi o programa de doctorat
FACULTAT MEDICINA UB	**BIOMEDICINA**

Títol de la tesi presentada / Título de la tesis presentada / Title of the thesis
ANÀLISI BIOINFORMÀTICA DE LA BASE GENÈTICA DE LA SUSCEPTIBILITAT AL CÀNCER DE MAMA

Nom i cognoms del doctorand
NÚRIA BONIFACI CANO

1. ÉS APTA LA TESI PRESENTADA PER AL TRÀMIT DE LECTURA I DEFENSA PÚBLICA? / ¿ES APTA LA TESIS PRESENTADA PARA EL TRAMITE DE LECTURA Y DEFENSA PUBLICA? / IS THE THESIS SUITABLE TO BE READ AND DEFENDED PUBLICLY?

☐ NO **X SI / Y ES**

2. INFORME RAONAT

Com a director de la tesi doctoral realitzada i presentada per la Llda. en Biologia i Màster en "Bioinformatics for Health Sciences" per la UB i UPF, respectivament, **Núria Bonifaci Cano** vull expressar la meva conformitat i recolzament per la defensa del treball de tesi realitzat amb títol: '**ANÀLISI BIOINFORMÀTICA DE LA BASE GENÈTICA DE LA SUSCEPTIBILITAT AL CÀNCER DE MAMA"**.

Aquesta tesi doctoral ha estat realitzada per la Núria Bonifaci sota la meva direcció en el grup de Càncer de Mama i Biologia de Sistemes del Institut Català d'Oncologia (ICO), en l'Institut d'Investigació Biomèdica de Bellvitge (IDIBELL). Des de fa més de nou anys el nostre grup treballa en l'estudi dels factors de susceptibilitat genètica a malalties complexes com el càncer i, especialment el càncer de mama. Els treballs científics que donen empara a aquesta tesi van ser liderats (sota la meva supervisió) per la doctoranda. Es tracta dels següents 4 treballs, 3 publicats i un altre proper a revisió per publicació:

Bonifaci N*, A Berenguer*, J Díez, O Reina, I Medina, J Dopazo, V Moreno and MA Pujana. Biological processes, properties and molecular wiring diagrams of candidate low-penetrance breast cancer susceptibility genes. ***BMC Medical Genomics*** 1(1):62.
2008 Clave: Article
Factor d'impacte: 3,9
Quartil: Q2

N Bonifaci, B Górski, B Masojć, D Wokołorczyk, A Jakubowska, T Dębniak, A Berenguer, J Serra Musach, J Dopazo, SA Narod, J Lubinski, C Lázaro, C Cybulski and MA Pujana. Cancer-driving kinases may commonly influence risk of breast cancer. ***PLoS ONE*** 5(11):e14078.
2010 Clave: Article
Factor d'impacte: 4,3
Quartil: Q1

Bonifaci N, Colas E, Serra-Musach J, Karbalai N, Brunet J, Gómez A, Esteller M, Fernández-Taboada E, Berenguer A, Reventós J, Müller-Myhsok B, Amundadottir L, Duell EJ, MA Pujana. Integrating gene expression and epidemiological data for the discovery of genetic interactions associated with cancer risk. ***Carcinogenesis***. Jan 2.
2014 Clave: Article
Factor d'impacte: 5.6
Quartil: Q1

N Bonifaci, et al, por el consorcio CIMBA. Evaluating associations between genetic variants at cancer driver kinase loci and cancer risk in BRCA1/2 mutation carriers.
En preparació. Article.

En tots els treballs, la doctoranda Núria Bonifaci ha contribuït de forma principal en el disseny de l'estudi, anàlisi i interpretació de les dades, i escriptura dels manuscrits. El seu treball s'ha basat en la realització i implementació de noves estratègies bioinformàtiques per la identificació de les variants i gens que influencien el risc de càncer de mama en la població general. Tanmateix, en el primer dels treballs la doctoranda comparteix primera autoria amb un investigador de categoria "tècnic", qui va ajudar en la recopilació i anàlisis de les dades genètiques.

En la meva opinió, la Núria Bonifaci presenta un treball de recerca científica original, demostrada en diverses publicacions. Els resultats obtinguts per la Núria obren, a més, noves vies de recerca com són la identificació de les variants causals de les associacions trobades en el seu treball i l'estudi dels mecanismes moleculars responsables d'aquestes associacions. Aquests estudis són importants per avançar el coneixement de les bases moleculars de la carcinogènesi de mama, la seva prevenció i el seu tractament.

DATA / FECHA / DATE: **l'Hospitalet de Llobregat (Barcelona), 29 de Setembre del 2014**
SIGNATURA / FIRMA /SIGNATURE:

Miquel Àngel Pujana, Ph.D.
Breast Cancer and Systems Biology Unit, Catalan Institute of Oncology, IDIBELL, Barcelona
Phone: +34-932607463; E-mail: mapujana@iconcologia.net