UNIVERSITAT DE BARCELONA

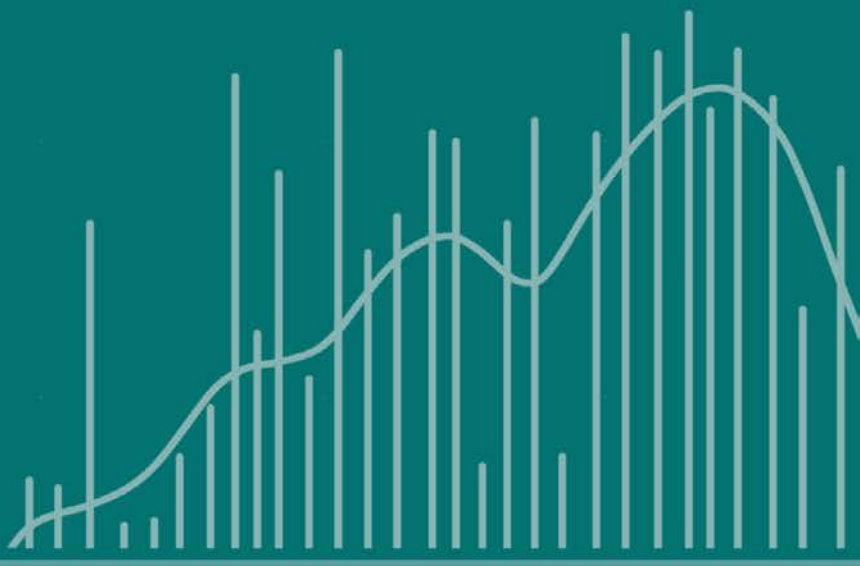
Variants del número de còpia al càncer colorectal: predisposició germinal i perfils genòmics tumorals

Sebastià Franch Expósito

Aquesta tesi doctoral està subjecta a la llicència ***Reconeixement 4.0. Espanya de Creative Commons***.

Esta tesis doctoral está sujeta a la licencia ***Reconocimiento 4.0. España de Creative Commons***.

This doctoral thesis is licensed under the ***Creative Commons Attribution 4.0. Spain License***.



Variants del número de còpia al càncer colorectal: predisposició germinal i perfils genòmics tumorals

Tesi doctoral

Sebastià Franch Expósito

2019

UNIVERSITAT DE BARCELONA

Programa de Doctorat en Medicina i Recerca Translacional

VARIANTS DE NÚMERO DE CÒPIA AL CÀNCER COLORECTAL: PREDISPOSICIÓ GERMINAL I PERFILS GENÒMICS TUMORALS

Memòria per optar al títol de Doctor presentada per

Sebastià Franch Expósito

Tesi doctoral realitzada
al departament d'Oncologia Gastrointestinal i Pancreàtica,
de l'Institut d'Investigacions Biomèdiques August Pi i Sunyer (IDIBAPS),
sota la direcció de

Sergi Castellví Bel
Predisposició Genètica a Càncer Gastrointestinal, IDIBAPS

Juan José Lozano Salvatella
Plataforma de Bioinformàtica, Centro de Investigación Biomédica en Red en Enfermedades Hepáticas y Digestivas (CIBEREHD)

IDIBAPS ciberehd

Sergi Castellví Bel
(co-director)

Juan José Lozano Salvatella
(co-director)

Antoni Catells Garangou
(tutor)

Sebastià Franch Expósito
(candidat)

Barcelona, Abril del 2019

A ma mare, mon pare
i al meu germà.

Do not go gentle into that good night,
Old age should burn and rave at close of day;
Rage, rage against the dying of the light.

Though wise men at their end know dark is right,
Because their words had forked no lightning they
Do not go gentle into that good night.

Good men, the last wave by, crying how bright
Their frail deeds might have danced in a green bay,
Rage, rage against the dying of the light.

Wild men who caught and sang the sun in flight,
And learn, too late, they grieved it on its way,
Do not go gentle into that good night.

Grave men, near death, who see with blinding sight
Blind eyes could blaze like meteors and be gay,
Rage, rage against the dying of the light.

And you, my father, there on the sad height,
Curse, bless, me now with your fierce tears, I pray.
Do not go gentle into that good night.
Rage, rage against the dying of the light.

Dylan Thomas, 1914 - 153

Agraïments

Potser és pel fet d'haver-lo conegut per la meravellosa pel·lícula *Interstellar*, del director Christopher Nolan, però el poema de Dylan Thomas "*Do not go gentle into that good night*" sempre m'ha suposat una certa metàfora en quant a la carrera científica o, pel tema que em porta aquí, la realització d'una tesi doctoral.

Les lluites internes i externes que et sorgeixen comencen arrelades a una gran necessitat inicial (a la pel·lícula, la de buscar un nou planeta habitable per als humans perquè, com no podia ésser d'una altra manera, ens hem carregat el nostre). A la vida real del científic, aquesta necessitat pot ésser la més senzilla curiositat del saber com funciona el món i les regles naturals que el dirigeixen. I allà que ens encaminem, en l'aventura del descobriment científic, a contra corrent i, a vegades, de forma conscientment inconscient, per tal d'expandir una miqueta (ni que sigui molt petita) el coneixement humà de la nostra realitat. És una continua lluita, amb una constant sensació de saber que sempre es pot fer més del que es fa, tot i que el més important és no aturar-se mai. Seguir treballant en les teves passes, petites o grans, encara que suposi adreçar-te per un camí difícil. Perquè, essent sincers, el món de la carrera científica no és un camí fàcil. Per altra banda, aquest encaminar-se continu cap el que encara es desconeix és apassionant.

Però, ja que et disposes a sortir de la teva zona de confort, en direcció a allò desconegut, que sigui acompanyat de la millor manera. I d'això van aquestes lletres. De destacar la companyia, mencionar-la i agrair-la. No per altra cosa aquest apartat es titula "agraïments".

Per començar, agrair moltíssim als meus directors de tesi, Sergi i Juanjo. Segurament fou una aposta arriscada agafar un alumne sense idea de la part de bioinformàtica, però només puc agrair la paciència i el temps que m'heu dedicat i el fet de deixar-me aprendre al meu ritme. Gràcies, Sergi, per tota l'atenció i la dedicació al meu treball, i per els debats i les aportacions que, de ben segur, m'han ajudat a créixer molt. Al Juanjo, per la seva total confiança envers a les meves capacitats ("latents" encara, en el seu moment) i per donar-me cert sentit de responsabilitat com a "representant bioinfo" a la quarta planta (juntament amb el Marcos).

També al Jordi, sobretot aquest últim tram, amb la història del CNApp, per aquesta immersió en la "part somàtica". Gràcies per les converses comunes, ja fossen científiques o de caire més personal. I d'aquí ho enllaço amb la Laia

Bassaganyas també. Moltíssimes gràcies als dos per contagiar-me part de la vostra passió per la ciència pura, la que enllaça amb la curiositat innocent.

Gràcies al Toni pel suport com a tutor d'aquesta tesi i per l'oportunitat de participar de l'equip d'oncologia gastrointestinal i pancreàtica.

A la gent del laboratori, on he corroborat que, indubtablement, la ciència és de les científiques. Gràcies per l'acollida quan (definitivament) vaig pujar a la planta quarta. I per totes les excursions, calçotades, ràftings i demés activitats que hem fet tots junts. La veritat és que no ens ho hem passat gens malament!

A la Maria Vila, amb qui he passat gran part d'aquests anys (els primers) a la planta -1, amb el "clúster de bioinformàtics". Gràcies per iniciar-me "al mundillo". Estic orgullós i content d'haver pogut aprendre coses de tu i compartir el despatxet d'allà baix. Sempre m'enrecordo dels "bucles" de pantalla infinits que vam estar fent aquell dia. Aprofito per recordar al Guerau, Guillaume, Teressa i Marc. La planta -1 era més amistosa amb ells.

A la Clara, gairebé vam coincidir més de viatges del COST que al laboratori (com el de Sevilla, amb la Jenny i l'Isa). Moltes gràcies per ensenyar-me durant el temps que vam estar els dos, i per la paciència i l'ajuda quan vam estar automatitzant la *pipeline* de les variants puntuals. Van ser desenes de correus amb milers de línies d'Excel que em comprovaves súper ràpid sempre. Espero que estigui essent un postdoc súper productiu als Països Baixos.

A la Laia ("la meva postdoc"). Em vas donar llum quan vas arribar! Només pensant amb el que em venia per endavant del CRISPR... Al final "et vaig abandonar" per la "pantalla negra" i, tu vas agrair perdre de vista el meu "freestyle" (a banda d'enamorar-te del Benchling i la seva medusa). M'encantaven (i encara ara) les nostres discussions científiques i debats dels resultats i protocols. Ara ja no et podria seguir el ritme, segurament.

A Evita (¡dinamita!). Siempre me sueltas el "aaay, Sebas, ¡qué paciencia tienes conmigo...!" Pero si alguna cosa ha quedado demostrada, es que la paciencia es cuál moneda de cambio. Muchas gracias por cuidarme y por preocuparte, y por ayudarme siempre que podías. Intento arreglarlo dándote caña en los entrenos de escalada, con Lorena y Ari.

A Marcos. No podría haber tenido un mejor compañero. Muchas gracias por todo: las charlas, las discusiones, los cursos juntos, ¡el crear

aplicaciones Shiny!... El gran ambiente del laboratorio ha sido en su mayor parte gracias a ti: organizando partidillos, animando las porras de futbol y Eurovisión (y ganándolas)... Gracias por confiar en mí y por compartir momentos e historias, tanto personales como otras. Espero que hayas disfrutado de la estancia en San Diego, te la mereces muchísimo, y todo lo bueno que te venga de ahora en adelante. ¡Eres un crack! Aprovecho para mencionar a Mariano: gracias por estas visitas anuales al laboratorio, sobretodo la última (partido River-Boca y las pachanguitas de fútbol).

Al Keyvan, el nostre mestre de les calçotades i la muntanya. Gràcies pels moments d'intercanvi d'idees i opinions i del teu punt de vista, sempre positiu, en totes les coses. Ets d'aquestes persones amb qui agrada parlar durant llargues estones.

A Clàudia, Maria i l'Elena Vila, amb els nostres cafès/té post-dinars, amb grans sortits de xocolata per a re-animar els capvespres. Gràcies pels moments compartits: els "bailoteos" quan hem sortit de marxa i les rialles en qualsevol de les activitats que hem fet junts. Sou unes bellíssimes persones!

A la Jenny ("la jefa"), la Saray, l'Elena Asensio, la Coral, el Manuel, l'Elena Fernández i altra gent que ja no són al laboratori, com la Isa, l'Esther i la Paula. Espero no deixar-me a ningú (i si ho faig, perdó!). Moltes gràcies també, per l'ajuda i els moments. Gràcies a la resta de l'equip OGiP, al "clínics" per la seva ajuda i disponibilitat per atendre els dubtes (Míriam, Sabela, Francesc...)

A n'Irene, segurament la millor part que m'emporto d'aquesta etapa al CEK. Gràcies per deixar-me conèixer-te més, compartir tot això i el que ha de venir. Crec que no puc arribar a ser conscient de lo important que ets per mi. Un "gràcies" continuat que es seguirà allargant. Espero poder ésser tan bona companyia com ho ets tu.

Als amics "de l'escalada", que han estat i són la meva família a Barcelona: Àlex i Raquel (gràcies per tot, no sé què hauria fet sense voltros, i estic molt i molt content de conèixer-vos), Pol i Paula (el mateix us dic), el Joan i l'Ari, la família Maqueda, Toni, etc... Àlex, gràcies per ensenyar-me a estimar l'escalada i per contagiar-me el teu fanatisme. I per portar-me a esquiar amb voltros els primers anys. Són uns records meravellosos els que tenc amb voltros!

Als amics de la universitat que encara tenim contacte: Dani, Naty, Carme, Raquel, Aina, Francina ("rosseta"); i a n'Alfonsina (Barcelona no hauria estat el mateix sense tu). Moltes gràcies per tot! També als del departament de Biologia Fonamental de la UIB: Jordi (gràcies), Dani i Mercedes, i als professors d'institut que em van ensenyar què era la ciència (gràcies Mª Antònia, per fer-la divertida).

Al fet d'haver nascut a Mallorca. Tenim un illa preciosa. Sa nostra estrella polar, sa Serra de Tramuntana, no és comparable en res que hi hagi al món. I a poder dir que sóc pobler, de Sa Pobla, i mig "pollensí", per part de mare. Des nostre Sant Antoni, de ses nostres ensaïmades (que tant vos agraden a tots!), espinagades, panades i robiols; de sa nostra festa a Gràcia i de compartir vila d'origen amb persones com en Toni Celià, a qui després de quasi dos anys d'aquella amistosa entrevista, ja vull considerar un amic, a banda de gran referent científic i exemple a seguir.

Als amics de Mallorca, que són sa família que un tria tenir, i que, per molt temps que passem sense veure'ns, sabem que ens tenim ben a prop: Guillem i Vicky, Saúl i Cintia (i n'Enzo menut) i Pere i Aurora. Tornar a Sa Pobla és sinònim de veure'ns, i això m'encanta. Sou es meu suport. També en Jaume i na Neus (i sa seva menuda, Laia).

I, sobretot, gràcies als meus: pares, germà i padrins. Als padrins, per pensar amb mi sempre, i tenir-me sempre present, tot i que no em poden veure tant com ells voldrien. Al meu germà Lluís, perquè, tot i ésser el meu germà petit, cada vegada em fas sentir més orgullós de ser-ho i m'encanta tenir-te el més a prop possible. A més, sé que et cuiden beníssim (gràcies Laia). I als meus pares: papà i mamà, gràcies per donar-me tot i esforçar-vos sempre per a que jo pogués fer el que volia. Esper que pugueu estar orgullosos des vostre fill, perquè res em fa més feliç que veure's contents i satisfets amb el que estic fent. Vos estim moltíssim.

A tots, moltes gràcies.

Una aferrada gran,

Sebastià.

Índex de continguts

Abreviatures

ACC *Adrenocortical Carcinoma*
BAF *B-Allele Frequency*
BCS *Broad CNA Score*
BLCA *Bladder Urothelial Carcinoma*
pb *parell de bases*
BRCA *Breast Invasive Carcinoma*
CCR Càncer Colorectal
CESC *Cervical Squamous Cell Carcinoma and Endocervical Adenocarcinoma*
CGH *Comparative Genomic Hybridization*
CHOL *Cholangiocarcinoma*
CIMP *Cpg Island Methylator Phenotype*
CIN *Chromosomal Instability*
CMS *Consensus Molecular Subtypes*
CNA *Copy Number Alterations*
CNV *Copy Number Variant*
COAD *Colon Adenocarcinoma*
DGV *The Database Of Genomic Variants*
DLBC *Lymphoid Neoplasm Diffuse Large B-Cell Lymphoma*
DNA *Deoxyribonucleic Acid*
ESCA *Esophageal Carcinoma*
FAP *Familial Adenomatous Polyposis*
FCS *Focal Cna Score*
FISH *Fluorescence In Situ Hybridization*
FOSTES *Fork Stalling and Template Switch*
GATK *Genome Analysis Toolkit*
GBM *Glioblastoma Multiforme*
GCS *Global Cna Score*
GWAS *Genome-Wide Association Studies*
HNSC *Head and Neck Squamous Cell Carcinoma*
HR *Homologous Recombination*
Kb *Kilo bases*
KICH *Kidney Chromophobe*
KIRC *Kidney Renal Clear Cell Carcinoma*
KIRP *Kidney Renal Papillary Cell Carcinoma*
LAML *Acute Myeloid Leukemia*
LGG *Brain Lower Grade Glioma*
LIHC *Liver Hepatocellular Carcinoma*
LOH *Loss Of Heterozygozity*
LUAD *Lung Adenocarcinoma*
LUSC *Lung Squamous Cell Carcinoma*
Mb *Mega bases*
MESO *Mesothelioma*
MLPA *Multiplex Ligation-Dependent Probe Amplification*
MMR *MisMatch Repair*
MSI *Microsatelite Instability*
MSS *Microsatelite Stability*
NAHR *Nonallellic Homologous Recombination*
NGS *Next Generation Sequencing*
NHEJ *Nonhomologous End-Joining*
OV *Ovarian Serous Cystadenocarcinoma*
PAAD *Pancreatic Adenocarcinoma*
PCPG *Pheochromocytoma and Paraganglioma*
PRAD *Prostate Adenocarcinoma*
READ *Rectum Adenocarcinoma*
RNA *Ribonucleic Acid*
SARC *Sarcoma*
SCNA *Somatic Copy Number Alterations*
SKCM *Skin Cutaneous Melanoma*
SNP *Single Nucleotide Polymorphism*
SNV *Single Nucleotide Variant*
STAD *Stomach Adenocarcinoma*
TAD *Topologically Associated Domain*
TCGA *The Cancer Genome Atlas*
TGCT *Testicular Germ Cell Tumors*
THCA *Thyroid Carcinoma*
THYM *Thymoma*
TSG *Tumor Supressor Gene*
UCEC *Uterine Corpus Endometrial Carcinoma*
UCS *Uterine Carcinosarcoma*
UPD *Uniparental Disomy*
UVM *Uveal Melanoma*
WES *Whole Exome Sequencing*
WGS *Whole Genome Sequencing*

Introducció

1 Càncer colorectal

Es diu que la història del càncer comença en paral·lel a la de l'ésser humà. Tot i així, no és fins al papir Edwin Smith, un document mèdic datat de la XVII dinastia de l'antic Egipte, on s'hi descriuen el que es pensa que són les primeres referències al càncer (Breasted & University of Chicago. Oriental Institute., 1930). Més tard, en l'època de la Grècia clàssica, Hipòcrates de Kos, considerat el pare de la medicina, utilitzà la paraula grega *karkinos* (cranc o cranc de riu) a l'hora de descriure els patrons particulars que presentaven els talls de distints tumors sòlids. Però seria el filòsof i enciclopedista romà, Celsus, qui finalment traduiria *karkinos* al terme llatí *cancer*. El mateix Celsus descriuria distints tipus de càncers superficials en la seva enciclopèdia *De Medicina*, continuant l'herència de les descripcions iniciades per Hipòcrates (Celsus, 50AD). Entre els distints tipus de càncer descrits en aquesta obra s'hi troben els d'estómac, de fetge i de còlon (Hajdu, 2011).

Com és ben sabut, el càncer és una malaltia genètica. Millor dit, és un conjunt de malalties que sorgeixen com a conseqüència de canvis en la

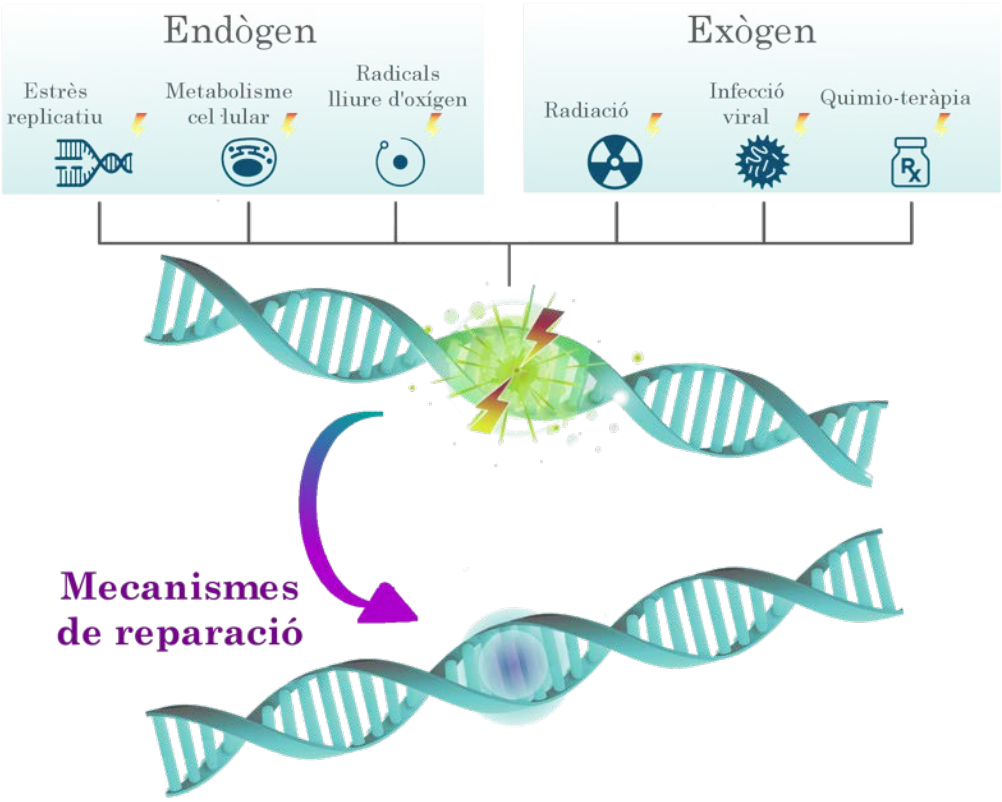


Figura 1. Dany i reparació del DNA.
Els agents endògens i/o exògens poden ocasionar danys al DNA (mutacions genètiques) que normalment són reparades gràcies als mecanismes de reparació.

informació genètica d'una o més cèl·lules, provocant el descontrol del creixement i del funcionament d'aquestes, i donant lloc a la seva expansió clonal. Aquestes mutacions genètiques es poden acumular al llarg de la vida, com a conseqüència d'errades genètiques que apareixen durant les contínues divisions cel·lulars, o per l'exposició a agents ambientals i/o químics que ocasionen dany a la molècula del DNA (**Figura 1**). Els distints mecanismes de reparació s'encarreguen de corregir aquests canvis i recuperar la versió "sana" de la informació genètica, tot i que a vegades poden fallar en la seva funció, facilitant que les mutacions s'acumulin (Loeb & Harris, 2008).

El càncer de còlon i recte, o càncer colorectal, és un d'entre els més de cent tipus de càncer que es coneixen actualment, i inclou tot aquell creixement anormal que apareix al còlon o recte ("What Is Cancer? - National Cancer Institute").

1.1 Epidemiologia

El càncer colorectal -o CCR- és el tercer tipus de càncer amb més incidència mundial quan ambdós sexes es tenen en compte, després del càncer de pulmó i el de mama. S'estima que aproximadament uns 1.900.000 casos hauran estat diagnosticats al 2018, representant la segona causa de mort per càncer amb 880.000 morts. L'Observatori Global del Càncer (GLOBOCAN, pel seu nom en anglès: Global Cancer Observatory), de l'Organització Mundial de la Salut, planteja un increment en la incidència d'aquesta patologia fins als 3 milions de casos per al 2040 (Bray et al., 2018). En termes socioeconòmics, el CCR podria ésser considerat com un marcador de desenvolupament nacional a nivell global, ja que quant més desenvolupament socioeconòmic presenten els països, majors ratis d'incidència de la malaltia es mesuren.

A Espanya, el rati d'incidència és de 80,1 casos per cada 100.000 habitants (**Figura 2**), essent el cinquè país amb més nombre de casos d'entre els països del sud d'Europa, després de Portugal, Eslovènia, Itàlia i Croàcia, i per davant de Sèrbia, Malta, Grècia, Bòsnia i Hercegovina i Macedònia. És el tipus de càncer amb més incidència al país i el segon en mortalitat, després del càncer de pulmó. Les estimacions en quant a incidència apunten a una xifra aproximada de 1.700.000 casos de CCR per a l'any 2040 i més de 25.000 morts (Bray et

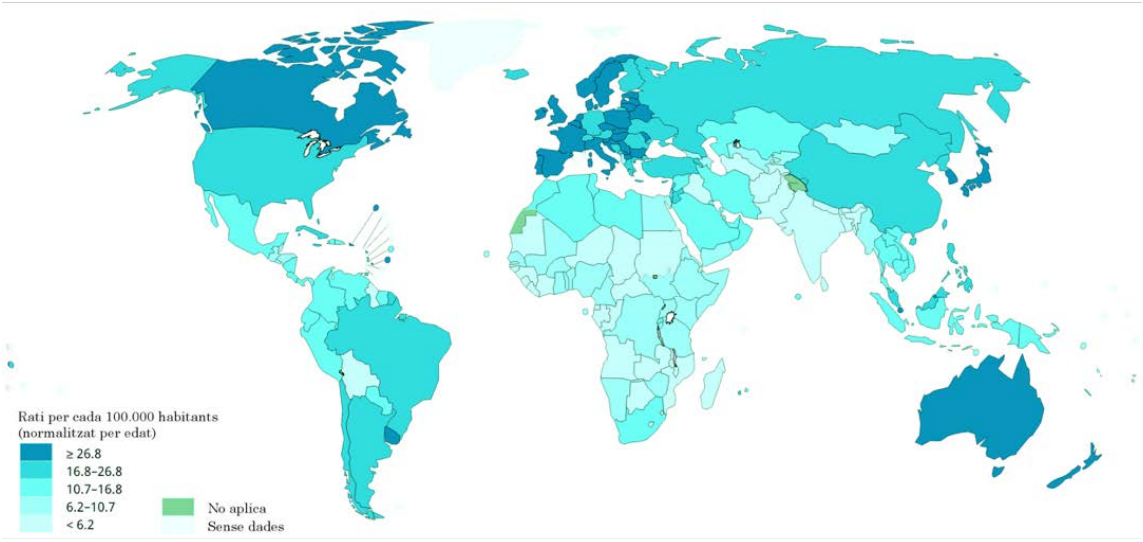


Figura 2. Estimacions de la incidència global del CCR per cada 100.000 habitants al 2018.
La incidència del CCR podria utilitzar-se com un marcador socio-econòmic degut a la correlació que mostra amb l'anomenat estil de vida occidental, més propi dels EUA i l'oest d'Europa i Austràlia. (imatge extreta i modificada de GLOBOCAN 2018 - IARC http://gco.iarc.fr/today)

al., 2018). A nivell regional, les comunitats autònomes amb major nombre de nous casos de CCR al 2017 són Andalusia (5.591), Catalunya (5.449), Madrid (4.423), la Comunitat Valenciana (3.624) i Galícia (2.468) (AECC, 2019).

1.2 Etiologia i factors de risc

El CCR es caracteritza per tenir una etiologia diversa i heterogènia. Al contrari que en alguns altres tipus de càncer, com el càncer de pulmó, el CCR no té cap factor de risc majoritari que abraci un alt percentatge dels casos (Brenner, Kloor, & Pox, 2014). Fins a tres quartes parts dels casos de CCR es classifiquen com a casos esporàdics, els quals no presenten història familiar de la malaltia (Kuipers et al., 2015). L'edat mitjana de diagnòstic del CCR es troba als 67 anys, amb un risc del 3-5% de patir la malaltia al llarg de la vida (Noone et al., 2018). Els factors de risc coneguts a l'hora de patir CCR són, entre d'altres, la història familiar, la malaltia inflamatòria intestinal, el tabac, el consum excessiu d'alcohol, alts ratis de consum de carn vermella i/o processada,

l'obesitat i la diabetis. Per altra banda, entre els factors de protecció establerts davant el CCR amb una reducció de la incidència del 20-30% hi trobem l'activitat física (Boyle, Keegel, Bull, Heyworth, & Fritschi, 2012), la teràpia de reemplaçament hormonal d'estrògens (Lin, Cheung, Lai, & Giovannucci, 2012) i l'aspirina (Rothwell et al., 2011; Bosetti, Rosato, Gallus, Cuzick, & La Vecchia, 2012). Altres factors de protecció, amb efectes més dubtosos, són la dieta rica en vegetals, fruites i cereals, el calci i algunes estatines (Hawk & Levin, 2005; Poynter et al., 2005; Brenner et al., 2014).

1.2.1 Edat

L'edat representa el major factor de risc a patir CCR (Binefa, Rodríguez-Moranta, Teule, & Medina-Hayas, 2014). Assolida l'edat de 50 anys, el CCR és molt més freqüent i la seva incidència incrementa substancialment, fins al punt que aproximadament el 90% de casos de la malaltia es diagnostiquen en pacients majors de 50 anys (**Figura 3**) (Noone et al., 2018). La generació de radicals lliures d'oxigen derivats del metabolisme cel·lular, la secreció de factors oncogènics de cèl·lules senescents, canvis en els patrons de metilació del genoma i, fins i tot, l'acumulació d'errors en la correcció de les mutacions genètiques per part dels sistemes reparadors del DNA fan que augmentin les probabilitats de patir mutacions en gens implicats en la carcinogènesis colorectal,

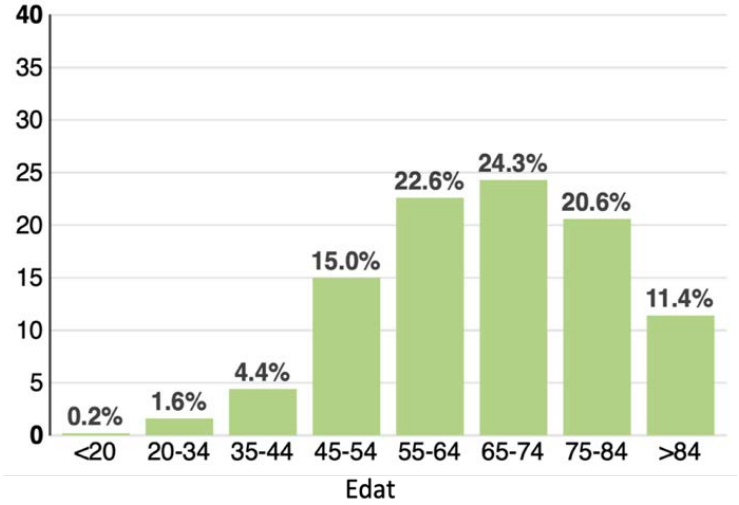


Figura 3. Edat de diagnòstic del CCR.
Els percentatges de diagnòstic segons les franges d'edat de la població. (Imatge extreta i adaptada de https://seer.cancer.gov/statfacts/html/colorect.html)

contribuint al desenvolupament de la malaltia (Martincorena & Campbell, 2015; Tomasetti, Li, & Vogelstein, 2017).

1.2.2 Ambient

La influència dels factors ambientals en el desenvolupament del CCR s'exemplifica per la gran varietat de ratis d'incidència que presenta la malaltia entre països (**Figura 2**). Els països més industrialitzats com els Estats Units, Canadà o Nova Zelanda poden presentar ratis d'incidència de fins a 25 vegades més que els països amb perfils menys industrials (Huxley et al., 2009). De fet, la progressiva adaptació de l'estil de vida occidental per part de països emergents genera un ràpid increment de la incidència de CCR en aquests. Un exemple d'això n'és el Japó que, des d'algunes dècades enrere cap al present, ha vist augmentar fins al 90% la incidència del CCR (Minami, Nishino, Tsubono, Tsuji, & Hisamichi, 2006). Tot això ens dona a entendre la importància dels riscos que comporten l'estil de vida dels països industrialitzats i de les seves característiques socials com la dieta hipercalòrica i rica en greixos saturats, la inactivitat física o els ratis d'obesitat creixents (Gingras & Béliveau, 2011).

Dieta

Encara que alguns estudis han descartat una capacitat protectora *per se* davant el risc a patir CCR, sembla que una dieta rica en fruites i verdures i amb alta aportació de fibra podria arribar a ser beneficiosa (van Duijnhoven et al., 2009). A més, la incorporació de productes vegetals en la dieta també pot aportar un significatiu nombre de vitamines. Entre elles, la vitamina B ha rebut una considerable atenció donada la seva participació en processos moleculars com la reparació, síntesis i metilació del DNA. Per exemple, l'àcid fòlic (vitamina B9) s'ha associat amb la reducció del risc al desenvolupament del CCR o adenomes, tot i que la seva aportació en forma de suplement alimentari pot afectar de forma contrària i afavorir la recurrència d'adenomes avançats (Giovannucci, 2002). Per altra banda, és ben sabut que una ingesta elevada de carn vermella o processada incrementa el risc a patir CCR tant en homes com en dones (Sandhu, White, & McPherson, 2001). Les altes temperatures en què es cuinen aquests productes afavoreixen la formació de fins 17 tipus distints d'amines heterocícliques que són considerades altament mutagèniques. A més, l'alt contingut en grup hemo que aporten també s'ha relacionat amb l'augment del risc a patir CCR (Sugimura, Wakabayashi, Nakagama, & Nagao, 2004).

Inactivitat física, obesitat i diabetis

Pel que fa als índex de massa corporal (IMC), múltiples estudis han relacionat IMC alts amb l'augment de risc de CCR, sobretot en homes (Ning, Wang, & Giovannucci, 2010). Per contra, alts nivells d'activitat física s'han associat amb una reducció de fins el 40% del risc a patir la malaltia (Wolin, Yan, Colditz, & Lee, 2009). La reducció de l'activitat física associada a la industrialització ha incrementat de forma dramàtica la incidència d'afeccions cròniques com malalties cardiovasculars, hipertensió, diabetis tipus 2 o l'obesitat. La sobrecàrrega dels teixits adiposos, associada a la inactivitat física i al sobrepès, pot alliberar altes concentracions de triglicèrids a la sang, provocant el desenvolupament de resistència a la insulina (Chavez & Summers, 2010). Aquesta resistència dels receptors pancreàtics provocarà l'excés d'insulina en sang, afavorint processos de proliferació cel·lular i carcinogènesis. A més, el teixit adipós és reconegut com a un òrgan endocrí actiu que participa del metabolisme hormonal alliberant les molècules funcionals conegudes com adipoquines. Per això, nivells elevats d'aquestes substàncies, conjuntament amb alts continguts lipídics associats a l'obesitat, poden acabar desenvolupant condicions d'inflamació crònica que participin en el desenvolupament del càncer (Calle & Kaaks, 2004). Per altra banda, individus diagnosticats amb colitis ulcerosa o malaltia de Crohn tenen fins a 19 vegades més probabilitats de patir CCR (Gillen, Walmsley, Prior, Andrews, & Allan, 1994).

Alcohol i tabaquisme

De la mateixa manera, el consum elevat d'alcohol i el tabaquisme han estat altament associats en estudis prospectius i de casos i controls a alt risc de patir CCR. Un consum igual o superior a 30g d'alcohol al dia augmenta el risc a patir la malaltia fins 1,24 vegades, en comparació a nivells baixos de consum, i les causes moleculars i/o fisiològiques poden ser múltiples (alts nivells d'àcid fòlic, defectes en la metilació del DNA, defectes del sistema immunitari o, fins i tot, del metabolisme associat al citocrom P450) (Chan & Giovannucci, 2010). En quant al tabac, tot i estar associat a patir neoplàsies en òrgans en contacte directe amb els seus composts carcinogènics, com el pulmó, la laringe o l'esòfag, alguns d'aquests elements nocius poden arribar a la mucosa colorectal pel sistema circulatori, promovent el desenvolupament del càncer (Botteri et al., 2008).

1.2.3 Herència genètica

Fins a un 35% dels casos de CCR venen determinats per l'herència familiar i com a conseqüència de mutacions genètiques germinals que es transmeten a la descendència i de generació en generació. Això condiciona una alta densitat en la presentació de casos de la malaltia dins una mateixa família, fenomen que s'identifica amb el terme "agregació familiar". Aquests casos es coneixen com a CCR hereditari o germinal, diferenciant-se del CCR esporàdic, que no presenta agregació familiar (Lichtenstein et al., 2000).

El risc a desenvolupar el CCR es duplica per aquells individus que compten amb dos o més familiars de primer grau que han patit la malaltia. Entre el 5-10% del total de casos anualment diagnosticats presentarien patrons d'herència mendeliana en forma autosòmica dominant, com a conseqüència de mutacions germinals en gens que ja han estat definits (Henry T. Lynch & de la Chapelle, 2003). El percentatge d'heretabilitat desconeguda del CCR s'atribueix a gens que encara no han estat identificats i que estarien aportant

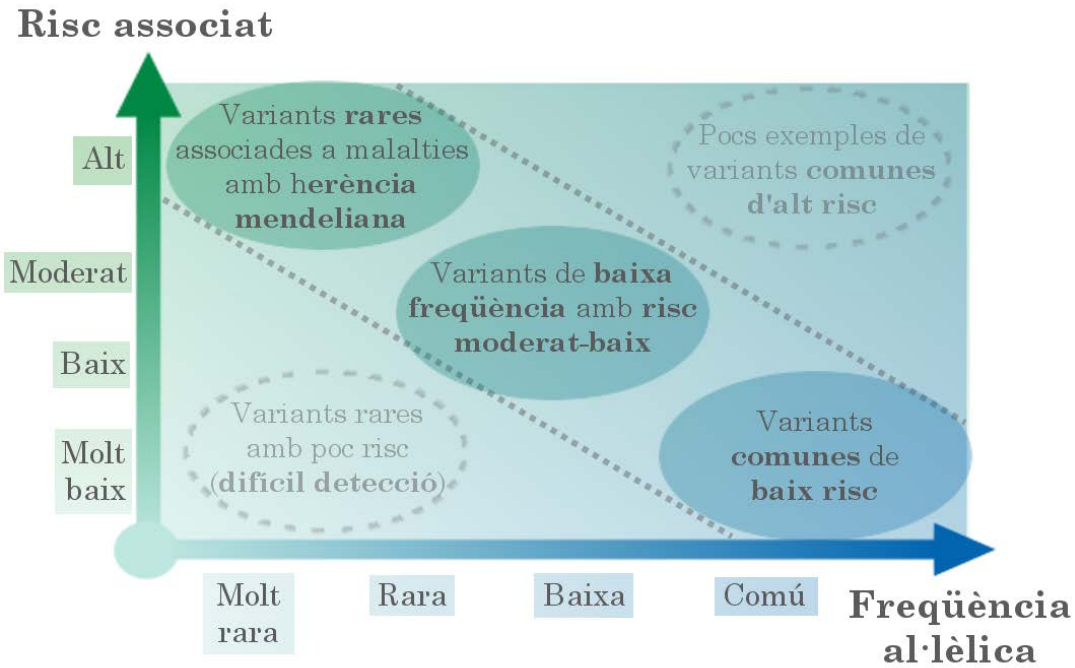


Figura 4. Relació entre freqüència al·lèlica i risc associat de les variants genètiques.
La relació entre dues característiques ens dona la capacitat de penetrància de la variant: les variants d'alta penetrància presenten un risc associat al fenotip d'estudi alt i presenten freqüències poblacionals molt baixes. Per altra banda, les variants amb penetrància baixa es troben comunament en la població general i atorguen risc baix a presentar el fenotip. Les variants que es troben en la diagonal descendent han estat objecte de grans esforços en quan al seu estudi en les darreres dècades. (Adaptada de Manolio T. et.al *Nature* 2009)

susceptibilitat alta o moderada, tot i que també i hauria part de la proporció que s'explicaria per la intervenció de múltiples variants genètiques de baix risc, més comunes en la població, normalment identificades en estudis poblacionals d'associació genètica (Peters, Bien, & Zubair, 2015; Win et al., 2017; Huyghe et al., 2018). Aquestes dues vessants de susceptibilitat defineixen el que es coneix com variants d'alta i baixa penetrància.

La penetrància és la capacitat o potencial genètic d'una variant a l'hora de condicionar un fenotip o provocar una malaltia a la qual se l'associa. En aquest sentit, aquelles variants genètiques que presenten baixa penetrància presenten un risc baix associat a la malaltia en qüestió i són comunes en la població general; mentre que les variants amb altra penetrància representen un risc alt a patir la malaltia, o el fenotip associat, i es presenten de forma rara en la població (**Figura 4**) (McCarthy et al., 2008; Manolio et al., 2009).

Hipòtesi de Knudson

El metge i genetista Alfred Knudson (1922 - 2016) formulà el model genètic pel qual el càncer s'assolia com a conseqüència de l'adquisició d'una primera mutació –*first hit*– en un dels dos al·lels del gen responsable de la malaltia, i una segona mutació –*second hit*– a l'altre al·lel, invalidant completament la funció del gen (Knudson, 1971). A nivell somàtic, aquesta acumulació de mutacions es dona al llarg de la vida de l'individu, mentre que a nivell germinal o hereditari la primera mutació ja està present, com a conseqüència de l'herència genètica familiar, facilitant l'aparició primerenca de la malaltia en comparació al càncer o esporàdic (**Figura 5**) (Knudson, 2001).

Aquest model genètic, conegut com "la hipòtesis de Knudson", va conduir de forma indirecta al descobriment dels gens supressors de tumors (TSGs, de l'anglès *tumor supressor genes*), la funció dels quals els atorga un paper regulador o protector davant el desenvolupament tumoral (Knudson, 1971). En el cas del CCR, el TSG per antonomàsia és l'*APC*, implicat en diversos processos com l'adhesió i migració cel·lular, la regulació dels nivells de ß-catenina de la via de senyalització WNT i la segregació dels cromosomes durant la mitosis (Rubinfeld et al., 1993; Su, Vogelstein, & Kinzler, 1993; Kinzler & Vogelstein, 1996; Polakis, 1997; Fodde, Kuipers, et al., 2001).

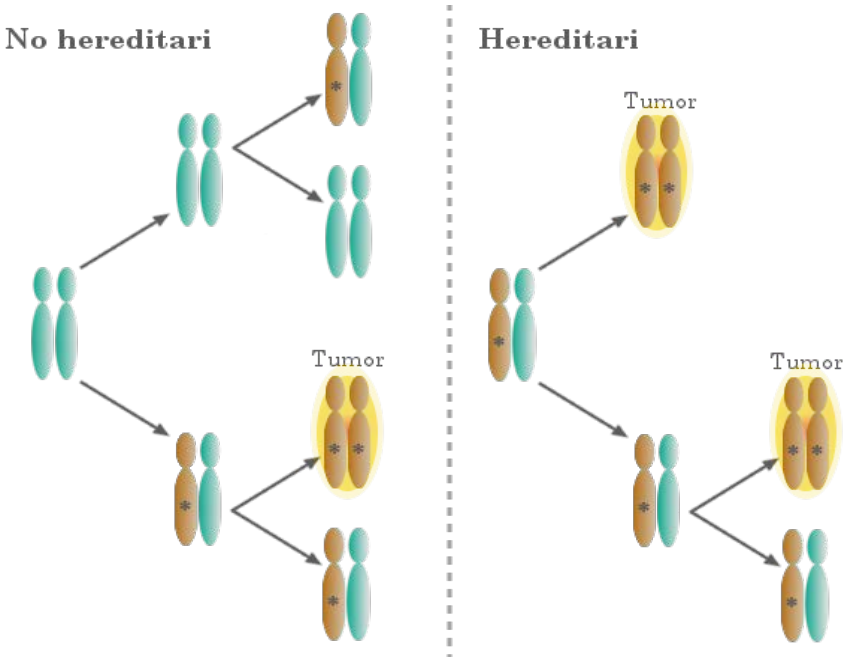


Figura 5. Esquema de la hipòtesis de Knudson.
El teixit normal en el context no hereditari encara no ha assolit ninguna mutació en cap dels dos al·lels del gen supressor de tumors (TSG). En el cas hereditari, totes les cèl·lules del teixit normal presenten un dels dos al·lels del TSG mutat. La inactivació bi-al·lèlica es pot donar en la següent divisió clonal, desenvolupament la malaltia en edats més primerenques. (Adaptada de Knudson A et.al *Nature Reviews Cancer* 2001)

1.3 Carcinogènesis del càncer colorectal

El model d'evolució del càncer més acceptat en la comunitat científica defensa un procés darwinià durant en el que les mutacions atzaroses acumulades al genoma de cèl·lules normals confereixen avantatges per al creixement i manteniment del tumor. Aquestes característiques adquirides es seleccionaran durant l'evolució clonal, producte d'una competència natural entre els distints clons portadors de mutacions diverses (Cairns, 1975; Nowell, 1976).

El genoma es replica i divideix d'una manera altament eficient. De fet, el rati de mutacions com a conseqüència d'aquests processos endògens de replicació i divisió cel·lular són molts baixos (uns 0,77 x 10^{-10} per base genòmica i divisió cel·lular), així com les errades de segregació cromosòmica

(aproximadament, una errada cada 100 divisions) (Burrell, McGranahan, Bartek, & Swanton, 2013). La desregulació dels processos de manteniment del genoma i/o l'exposició a agents exògens mutagènics acaba afavorint l'acumulació d'alteracions genètiques que aporten diversitat genòmica en les diferents cèl·lules proliferatives (Martincorena & Campbell, 2015).

Les cèl·lules canceroses es caracteritzen per presentar aberracions cromosòmiques complexes i re-ordenaments genòmics que poden anar des de variants a nivell de nucleòtid, fins a alteracions genòmiques estructurals que modifiquen la quantitat de material genètic present al nucli. L'estudi dels distints tipus de variants i alteracions que s'acumulen al genoma durant el procés de carcinogènesis té com a objectiu elucidar de les conseqüències funcionals que provoquen els distints tipus d'alteracions acumulades al genoma i com aquestes afavoreixen l'aparició i el creixement tumoral (Stratton, Campbell, & Futreal, 2009).

L'estudi per a la identificació de les alteracions genètiques que aporten predisposició al CCR ha sofert un gran progrés en els últims anys i això ha beneficiat el coneixement de la genètica molecular que participa en la carcinogènesi del CCR (Eric R. Fearon, 1995; Cunningham et al., 2010; Kuipers et al., 2015). Els defectes moleculars com a conseqüència d'aquestes alteracions genètiques poden actuar de dues formes: tant provocant l'augment de la funció (i fins i tot generant noves funcions) de gens que afavoreixen l'aparició del càncer, anomenats "oncogens"; o d'altra banda, mitjançant la pèrdua de la funció de gens encarregats del control de la homeòstasis cel·lular, coneguts com TSGs (Knudson, 2001; Bert Vogelstein et al., 2013). El procés evolutiu a nivell cel·lular conegut com a "selecció clonal" és l'encarregat de seleccionar aquells clons cel·lulars que han assolit les alteracions genètiques més propenses a l'hora d'iniciar i promocionar el procés de carcinogènesi al CCR (Burrell et al., 2013; McGranahan & Swanton, 2017).

L'alteració funcional dels oncogens o dels TSGs pot ocórrer com a conseqüència de mutacions puntuals que afectin l'estructura de la futura proteïna i la seva funció o, per altra banda, degut a re-ordenaments cromosòmics o de segments genòmics que provoquin la desregulació de la seva expressió gènica (Stratton et al., 2009; Garraway & Lander, 2013). En qualsevol cas, aquestes alteracions provoquen la desregulació de les vies de senyalització i mecanismes moleculars de les quals en són partícips els gens afectats. A nivell cel·lular, això pot implicar el creixement descontrolat de les cèl·lules, la inactivació dels seus sistemes de reparació o la capacitat d'evasió de la mort cel·lular, entre (Bert

Vogelstein et al., 2013; Garraway & Lander, 2013; Wheeler & Wang, 2013).

Per tal de plasmar aquests distints processos que les cèl·lules assoleixen a l'hora d'iniciar la progressió tumoral, a l'any 2000, els investigadors Hanahan i Weinberg publicaren el que seria el model lògic per a entendre la diversitat de mecanismes que podien presentar les unitats cel·lulars neoplàsiques (D Hanahan & Weinberg, 2000). En ell, presentaren sis característiques o processos claus (en anglès *hallmarks*) als quals les cèl·lules pre-neoplàsiques podien recórrer per iniciar la carcinogènesi i afavorir la metàstasi. La revisió del seu propi treball, després d'una dècada de grans avenços en el camp de la investigació del càncer, portà als dos investigadors a expandir les característiques principals fins a un total de deu mecanismes comuns que es poden trobar en les cèl·lules tumorals (**Figura 6**) (Douglas Hanahan & Weinberg, 2011).

Figura 6. Hallmarks del càncer.
Els processos moleculars que permeten a les cèl·lules canceroses sobreviure, proliferar i créixer recorrent el procés tumoral. (Extreta i adaptada de Hanhan i Weinberg *Cell* 2011)

Així doncs, els *hallmarks* del càncer es defineixen com a capacitats funcionals adquirides que permeten a les cèl·lules canceroses sobreviure, proliferar i escampar-se pel teixit primari (o cap a altres teixits, durant el

procés que metastàtic). Una de les característiques més importants és el desenvolupament de la inestabilitat genòmica en les cèl·lules canceroses. Gràcies a la generació de mutacions genètiques i alteracions cromosòmiques, es facilita l'adquisició dels demés mecanismes cancerosos, assolint l'avantatge selectiu entre els distints sub-clons cel·lulars per a iniciar el procés tumoral.

1.3.1 Lesions precursores i vies moleculars

El CCR sorgeix per l'acumulació de mutacions genètiques en gens claus que provoquen la progressió del teixit normal - l'epiteli intestinal - , passant pel que s'anomenen lesions precursores, i fins al desenvolupament final del càncer (**Figura 7**). Aquestes lesions precursores es presenten com a masses cel·lulars o de teixit que sobresurten de l'epiteli intestinal cap a la llum de l'òrgan, i es poden dividir en adenomes convencionals o pòlips serrats (Eric R. Fearon, 1995; Davies, Miller, & Coleman, 2005).

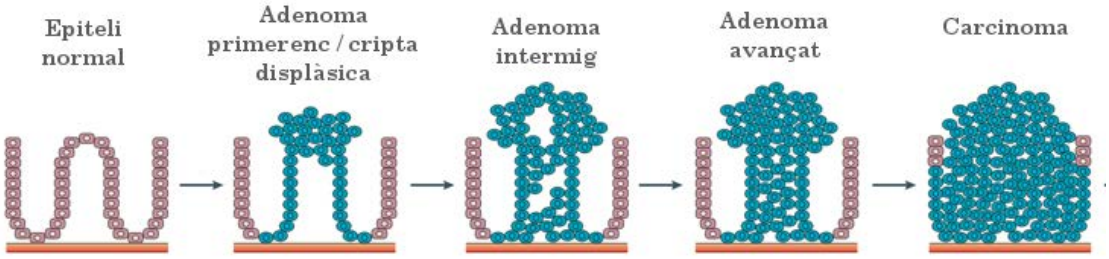


Figura 7. Transició de l'epiteli normal a la formació de pòlips i posterior desenvolupament del càncer de còlon.
Mitjançant l'adquisició de certes mutacions genètiques, l'epiteli normal passa a formar estructures displàsiques a les criptes colòniques (pòlips) que podran iniciar el desenvolupament del càncer colorrectal si segueixen acumulant mutacions en gens claus. (Adaptada de Davies R et.al *Nature Reviews Cancer* 2005)

Els adenomes convencionals són lesions benignes definides per la presència d'epiteli displàsic. La majoria d'adenomes presenten dimensions inferiors a un centímetre i arquitectura tubular (Eric R. Fearon, 2011; Langner, 2015). Històricament, els adenomes han estat considerades les lesions pre-canceroses més importants per a l'esdeveniment del CCR. S'estima que més d'un 50% dels individus

desenvoluparà adenomes al llarg de la seva vida, però només un 6-10% d'aquests adenomes arribaran a formar el càncer i, així i tot, aquest procés cancerós es pot allargar durant anys, o fins i tot dècades (Ahlquist, 2010; Dickinson, Kisiel, Ahlquist, & Grady, 2015).

La via tradicional -o clàssica- de la carcinogènesis colorectal, coneguda com la "seqüència adenoma-carcinoma", representa la gran majoria de CCR. El procés s'inicia per la presència de pòlips benignes en una cripta aberrant, per després formar l'adenoma primerenc, més petit d'un centímetre i amb components histològics tubulars. L'adenoma progressa cap a un adenoma avançat, ja més gran d'un centímetre i amb component vellós, per finalment formar el CCR. El procés ve condicionat per l'acumulació de variants genòmiques que comporten la desregulació de mecanismes moleculars i originen els distints estadis histològics i fenotípics (E R Fearon & Vogelstein, 1990; Kuipers et al., 2015) (**Figura 8**). El principal mecanisme iniciador de la via a nivell genètic és la inactivació del gen *APC* i la conseqüent desregulació de la via de senyalització de WNT (Fodde, Smits, & Clevers, 2001).

Per altra banda, aproximadament un terç del CCR es desenvolupa a partir de pòlips serrats o lesions serrades. Aquest grup de lesions és molt més

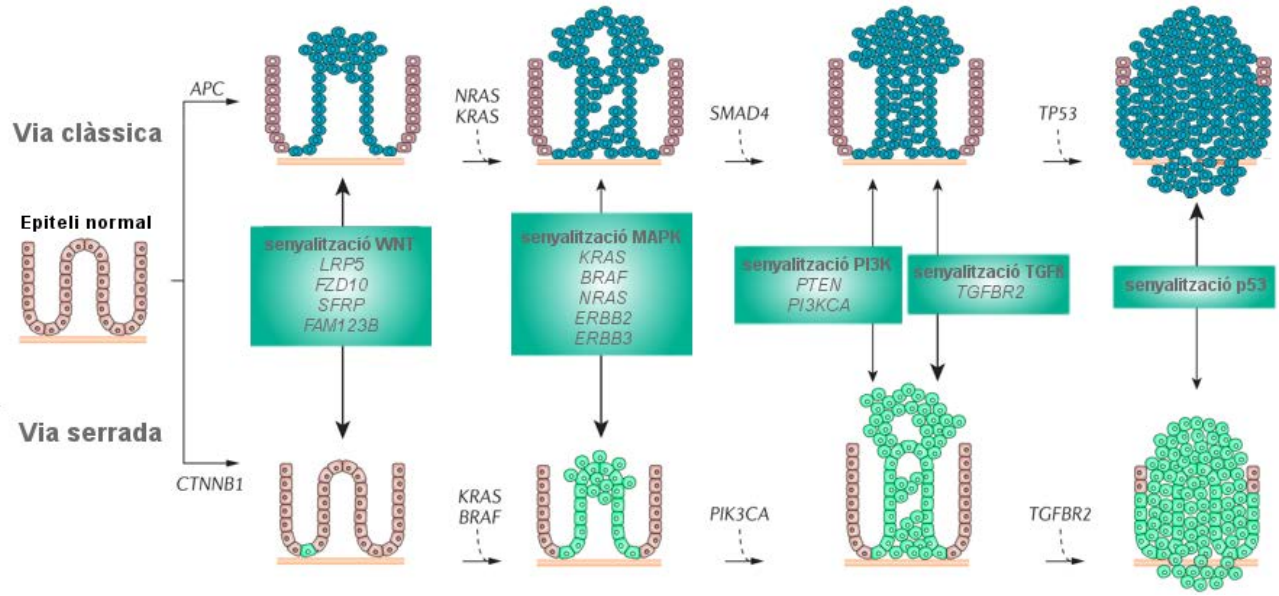


Figura 8. Via clàssica i la via serrada en la seqüència epiteli normal al carcinoma.
La diferenciació entre dos tipus principals de lesions precursores –adenoma convencional o pòlip serrat– donen lloc a la diferenciació de dues vies histològiques i moleculars en la carcinogènesis des de l'epiteli colònic normal cap al CCR. La recurrència en les alteracions genètiques en estudis genòmics ha permès identificar els gens principalment alterats durant la carcinogènesi del CCR i corroborar les vies de senyalització implicades. (Extreta i adaptada de Kuipers E et.al *Nature Reviews* 2015)

heterogeni, però es caracteritza morfològicament per la formació d'estructures en forma serrada al compartiment epitelial (Ahlquist, 2010). S'inclouen dins aquest grup els pòlips hiperplàsics, l'adenoma serrat sèssil i l'adenoma tradicional serrat. Aquests tipus de lesions precursores continuen per la via serrada fins al desenvolupament del CCR, diferenciant-se histològica i molecularment dels adenomes tubulars (**Figura 8**) (Eric R. Fearon, 2011; Langner, 2015).

El gatekeeper de l'epiteli colorectal: el gen *APC*

Al voltant del 70-80% d'adenomes colorectals esporàdics presenten inactivació somàtica del gen *APC* (Rubinfeld et al., 1993; Polakis, 1997; Fodde, Kuipers, et al., 2001). Això suggereix la importància del gen a l'hora d'iniciar els desenvolupament de la majoria d'adenomes, i el fet que els autors Kinzler i Volgestein l'anomenaren el gatekeeper de la seqüència adenoma-carcinoma, és a dir, una espècie de controlador que, d'inactivar-se, desencadenaria l'inici del camí al CCR (Kinzler & Vogelstein, 1996). De fet, seguint el model de Knudson per als TSGs, ambdós al·lels *d'APC* es troben inactivats en adenomes i carcinomes, tant en la majoria de CCR amb component genètic, com els casos de la poliposis familiar adenomatosa, com en la majoria de casos esporàdics (Kuipers et al., 2015; Langner, 2015).

El gen *APC* codifica per una proteïna d'uns 300 kDa, amb el mateix nom, relacionada amb algunes funcions com la regulació de l'adhesió entre cèl·lules, la migració cel·lular, la segregació cromosòmica i l'apoptosi en la cripta colònica. Molecularment, la funció més reconeguda de la proteïna APC és la de reguladora dels nivells de ß-catenina de la via de senyalització cel·lular WNT. A grans trets, la presència d'APC, que forma part d'un complex format per altres proteïnes, afavoreix la fosforilació de la ß-catenina, el que originarà la seva posterior ubiqüitinitzacióí i la seva degradació per part del proteosoma (Fodde, Smits, et al., 2001; Eric R. Fearon, 2011).

Degut a la inactivació genètica del gen *APC*, i la pèrdua de la seva funció a nivell cel·lular, la fosforilació i degradació de la ß-catenina no succeeix. Aquesta s'acumularà al nucli, i gràcies a la interacció amb altres proteïnes, seràs capaç d'accedir a l'interior del nucli, on desplegarà el paper de co-activador transcripcional, afavorint l'activació de l'expressió gènica de distints gens, com els proto-oncogens *c-MYC* i la *ciclina D1*. Amb tot, el programa transcripcional que s'assoleix en la cèl·lula gràcies a la presència de la ß-catenina al nucli és molt pròxim al que caracteritza les cèl·lules mare en les criptes del còlon (Fodde, Smits, et al., 2001; Kuipers et al., 2015).

1.3.2 Vies de la carcinogènesi

En el cas del CCR, el *hallmark* de la inestabilitat genòmica és el procés responsable del major percentatge de casos de CCR. Actualment s'accepten tres vies de carcinogènesi colorectal des del punt de vista de les alteracions genètiques i moleculars: l'anomenada via d'inestabilitat cromosòmica, la via d'inestabilitat de microsatèl·lits, i la del fenotip metilador d'illes CpG (**Figura 9**) (Cunningham et al., 2010; Carethers & Jung, 2015). Tot i que les tres vies han estat molt estudiades, cada una d'elles no és biològicament excloent de les altres, i s'han vist que certs tumors que presenten perfils moleculars específics d'una, també poden presentar característiques de les altres. De fet, la inestabilitat genòmica és protagonista tant en la via d'inestabilitat cromosòmica com en la via d'inestabilitat de microsatèl·lits; mentre que el

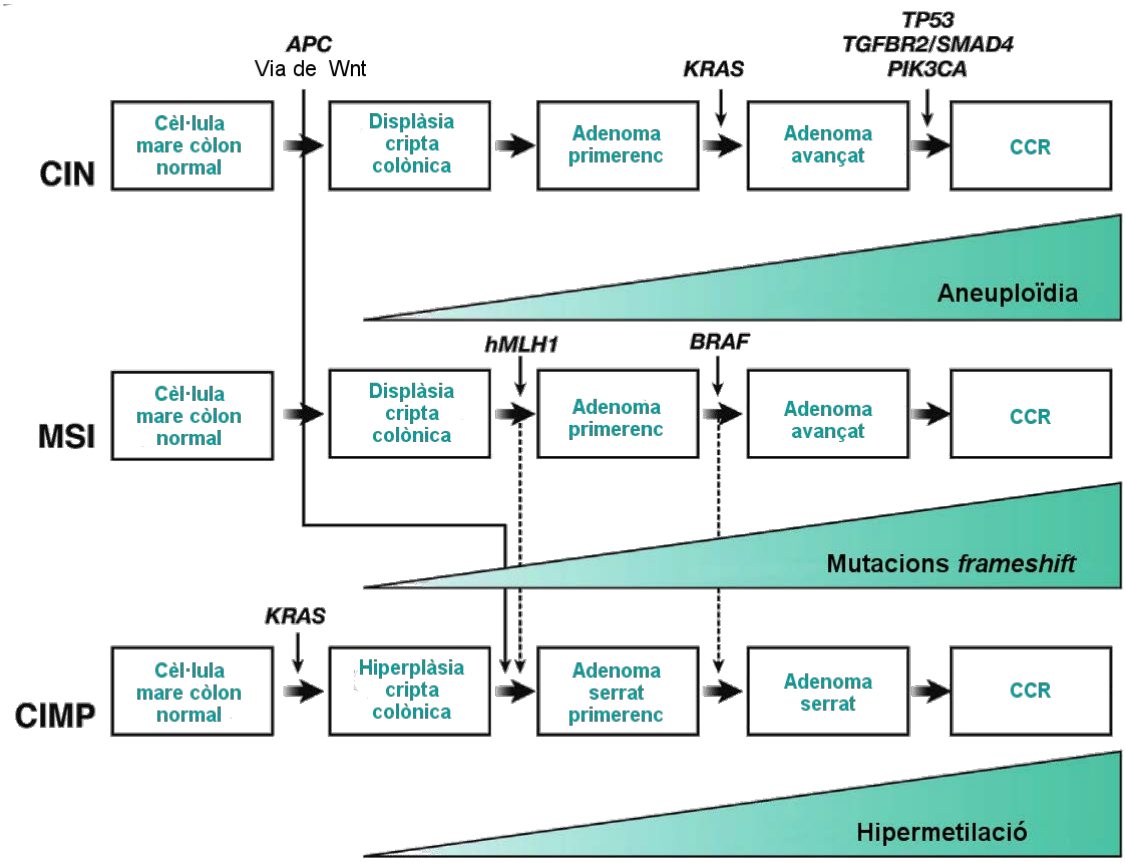


Figura 9. Vies de la carcinogènesi del CCR.
El desenvolupament del CCR pot donar-se entre tres distintes vies moleculars: la inestabilitat cromosòmica, la inestabilitat de microsatèl·lits o la via del fenotip metilador d'illes CpG. Cada via es caracteritza per certes alteracions genètiques i moleculars específiques, tot i que existeix cert solapament entre elles. CIN: *chromosomal instability*; MSI: *microsatelite instability*; CIMP: *CpG islands methylation phenotype*; *hMLH1*: *hipermethylated MLH1* (Extreta i adaptada de Carethers J. and Jung B. *Gastroenterology* 2015)

fenotip metilador és àmpliament present en la via d'inestabilitat de microsatèl·lits (S. D. Markowitz & Bertagnolli, 2009).

1.3.2.1 Inestabilitat cromosòmica

La gran majoria de CCR es desenvolupen com a conseqüència de l'adquisició d'alteracions a nivell cromosòmic, el que es coneix com a via d'inestabilitat cromosòmica (CIN, de l'anglès *chromosomal instability*). Els tumors que segueixen la via CIN es desenvolupen d'acord al model clàssic de progressió adenoma-carcinoma proposat per Vogelstein i Fearon al 1990. En ell, es produeixen nombroses alteracions que acabaran provocant l'activació de determinats oncogens (com *KRAS)* i, per contra, la inactivació de TSGs (com *APC*, *TP53*, *DCC* o *SMAD4*) (E R Fearon & Vogelstein, 1990; Pino & Chung, 2010).

L'esdeveniment que inicia la CIN consisteix en la desregulació de la via de senyalització WNT, afavorint la formació de l'adenoma, normalment com a conseqüència de la pèrdua de la funció del TSG *APC*, com a conseqüència de la inactivació dels seus dos al·lels, essent, en la majoria dels casos, la pèrdua del braç llarg del cromosoma 5 (braç 5q on s'hi troba aquest gen) el *second hit* (Fodde, Kuipers, et al., 2001; Hermsen et al., 2002). Els següents passos més recurrents de la via CIN –que acabaran amb el desenvolupament del càncer– consisteixen en l'acumulació de mutacions al gen *KRAS*, involucrat en la via de senyalització de les MAPK, pèrdues del braç 18q –afectant la via de senyalització de la TGF-ß–, i mutacions al gen *TP53* o pèrdua del braç 17p, on està situat aquest gen (**Figura 9**) (Eric R. Fearon, 2011; Agrawal, Bhattacharya, Manhas, & Sen, 2019).

La CIN està íntimament lligada al concepte d'aneuploïdia (o CIN numèrica), és a dir, aquell estat genòmic de la cèl·lula que difereix de la seva normalitat (euploïdia) –en el cas del genoma humà, la diploïdia–. Aquesta inestabilitat cromosòmica incrementa durant el procés tumoral, al adenocarcinoma, i es manté més o menys estable en estadis metastàtics. Aquest equilibri complex del material genòmic al càncer resulta en desregulacions funcionals de múltiples processos cel·lulars. De fet, moltes de les alteracions assolides, tot i que poden ocórrer de forma atzarosa, són necessàries durant el creixement tumoral. Aquestes implicaran la modificació de certes dianes gèniques que conferiran un avantatge selectiu als clons cel·lulars que les presentin (Matano et al., 2015; Saito et al., 2018). Tot i això, certs estudis han demostrat que els adenomes que presenten certs focus de carcinoma ja han assolit la

majoria d'aberracions cromosòmiques en la seva part histològica benigna. Les alteracions cromosòmiques recurrents que caracteritzen la progressió des de l'adenoma al carcinoma són: els guanys dels braços cromosòmics 8q, 13q i 20q, i les pèrdues de 8p, 15q, 17p i 18q (Ried et al., 1996; Meijer et al., 1998; Hermsen et al., 2002).

1.3.2.2 Inestabilitat de microsatèl·lits

Els microsatèl·lits, també anomenats repeticions curtes en tàndem (STR, de l'anglès *short tandem repeats*), són elements repetitius del DNA distribuïts en tàndem i amb longituds aproximades de 1-6 pb. Aquestes seqüències, i juntament amb altres tipus de seqüències repetitives (minisatèl·lits, satèl·lits), poden arribar a conformar fins el 3% del total del genoma humà (I. H. G. S. Consortium, 2001). La seva estructura repetitiva afavoreix que, durant el procés de replicació del DNA, la polimerasa es desplaci de manera errònia, generant errades d'aparellament entre les cadenes i es formin petits bucles. En condicions normals, aquestes errades seran reparades pel sistema de reparació del DNA *mismatch repair* (MMR, reparació d'errades d'aparellament), on hi participen els gens per les proteïnes *MLH1*, *MLH3*, *MSH2*, *MSH3*, *MSH6* i *PMS2* (Grilley, Holmes, Yashar, & Modrich, 1990; Ionov, Peinado, Malkhosyan, Shibata, & Perucho, 1993; S. Markowitz et al., 1995; Miyaki et al., 1997).

La via d'inestabilitat de microsatèl·lits (MSI, de l'anglès *microsatelite instability*) es caracteritza per la inactivació de qualsevol de les proteïnes que participen del sistema reparador MMR, mentre que la forma activa defineix els tumors com a estables (MSS, de l'anglès *microsatelite stability*) (Grilley et al., 1990; Boland & Goel, 2010). La falta de correcció de les errades d'aparellament condiciona l'acumulació d'errors de tipus inserció o deleció (anomenats *frameshift*, en anglès) que malbaraten la pauta de lectura del DNA i poden arribar a impedir la traducció proteica o, fins i tot, de generar-ne formes no funcionals (**Figura 9**). Així, processos cel·lulars com l'apoptosi, el cicle cel·lular o la reparació del DNA poden veure's afectats degut a l'acumulació de mutacions en regions de microsatèl·lits presents en les zones codificants dels gens que participen d'aquets processos (considerats oncogens o TSGs). La desregulació d'aquells gens que afavoreixen el procés tumoral suposa la selecció positiva dels clons cel·lulars que la presenten, contribuint al fenotip tumoral cel·lular. Els tumors que segueixen la via MSI es caracteritzen per presentar cariotips propers a ser diploides i fenotips híper-mutadors, contràriament a la via d'inestabilitat cromosòmica (S. D. Markowitz & Bertagnolli, 2009).

Al voltant d'un 15% del total de casos de CCR es desenvolupen per la via MSI, i solen presentar un fenotip accentuat d'aquesta inestabilitat –identificada com com MSI-H (*MSI-high*)–, caracteritzats per presentar la inestabilitat en més del 40% de les regions repetitives que s'estudien. Per contra, alguns tumors poden presentar un fenotip intermig entre MSS i MSI, que s'han etiquetat com MSI-L (*MSI-low*). El marcador més utilitzat a l'hora d'identificar MSI és la regió de microsatèl·lits BAT26 (de l'anglès, *big adenine tract 26*), localitzat en un dels introns del gen *MSH2*, donat que és la regió més alterada entre els casos de MSI. Entre el 15-20% dels casos són d'origen hereditari amb síndrome de Lynch, on l'alteració del sistema MMR sorgeix a nivell germinal (Boland & Goel, 2010).

Una de les majors distincions que fan els investigadors alhora de dividir biològicament els CCR és entre aquells tumors amb MSI, normalment localitzats al còlon dret i freqüentment associats al fenotip metilador d'illes CpG i híper-mutat, i els tumors amb MSS però presentant inestabilitat cromosòmica. De fet, la gran majoria (80-85%) dels CCR MSI són tumors esporàdics caracteritzats per la hipermetilació somàtica de la regió promotora del gen *MLH1* (Cunningham et al., 2010).

1.3.2.3 Fenotip metilador d'illes CpG

Les illes CpG són regions del DNA riques en dinucleòtids de citosina-guanina presents en les regions promotores d'aproximadament la meitat dels gens del genoma humà. La metilació d'aquestes regions promotores seria la causa de la inactivació transcripcional dels gens que les contenen (Bird, 1986).

L'anomenada via del fenotip metilador de les illes CpG (CIMP, de l'anglès *CpG island methylator phenotype*) – o via serrada –, s'ha descrit com una de les possibles vies d'inici de la carcinogènesis del CCR, caracteritzant-se per la hipermetilació d'illes CpG en zones promotores que controlen l'expressió gènica (Toyota et al., 1999). Aquesta hipermetilació es produeix d'una forma estocàstica, associada a una inestabilitat epigenètica, que afavorirà la progressió tumoral gràcies a la inactivació de TSGs com *CDKN2A*, *MGMT* i *MLH1* (Dickinson et al., 2015).

En aquest cas, la via s'inicia amb l'aparició de les lesions precursores anomenades pòlips serrats. Es sap que fins un 15-30% dels casos de CCR podrien estar associats a la via CIMP i, tot i que els perfils mutacional dels tumors sobrevinguts per aquesta via difereixen respecte

dels que sorgeixen de la via clàssica, alguns subgrups de tumors poden arribar a mostrar cert solapament amb les formes tumorals CIN i MSI, com a conseqüència de la inactivació promotora per metilació de gens associats a inestabilitat cromosòmica o reparació del DNA (Kambara et al., 2004; S. D. Markowitz & Bertagnolli, 2009).

El fenotip CIMP sol ésser el primer esdeveniment en l'acumulació d'un seguit d'alteracions genètiques, les quals semblen tenir la mutació activadora V600E de l'oncogen *BRAF* com a senyal precursora (Minoo, Moyer, & Jass, 2007). Aquesta mutació és considerada un fort marcador de CCR d'origen esporàdic i de lesions serrades. De fet, existeix una forta associació entre el fenotip CIMP, la mutació en *BRAF* i el MSI, segurament deguda a la inactivació del gen *MLH1*, que aporta l'estat MSI, com a conseqüència de la hipermetilació somàtica del promotor del mateix *MLH1* (*hMLH1*) (**Figura 9**). D'aquesta manera, la majoria de CCR esporàdics via serrada són tumors MSI i amb fenotip CIMP, mentre que els adenomes que segueixen la via clàssica estarien més associats amb tumors MSS (Jass, 2008; Dickinson et al., 2015).

1.4 Caracterització del càncer colorectal

1.4.1 The Cancer Genome Atlas

El projecte *The Cancer Genome Atlas* (TCGA) ha representat un enorme projecte col·laboratiu entre les institucions americanes *National Cancer Institute* (NCI) i *National Human Genome Research Institute* encaminat a la caracterització exhaustiva, a nivell multi-dimensional en quant a l'anàlisi de distintes capes òmiques, de 33 tipus diferents de càncer (The Cancer Genome Atlas, n.d.).

Durant més d'una dècada, el consorci del TCGA ha generat aproximadament 2,5 petabytes de dades per a la caracterització dels 33 tipus de càncer, incloent 10 tipus de càncers rars, basant-se en mostres tumorals aparellades amb la corresponent mostra genòmica germinal del pacient, per als més de 11.000 individus estudiats. Els resultats de l'estudi de les dades del TCGA es troben unificats i disponibles a la comunitat científica mitjançant el programa GDC (en anglès *Genomic Data Commons*) del NCI, que té com a objectiu impulsar i promocionar la medicina de precisió en el camp de

l'oncologia (Grossman et al., 2016). Per exemplificar la gran quantitat de dades disponibles per a la seva exploració, en aquests moments, el repositori presenta informació referent a més de 3 milions de mutacions en uns 22.000 gens.

El consorci de recerca encarregat de la gestió del TCGA ha posat punt i final amb la publicació del *Pan-Cancer Atlas* (Pan-Cancer Atlas, n.d.), una col·lecció d'estudis d'integració de les distintes capes d'informació generades per als distints càncer, posant el focus en tres aspectes principals de l'estudi com la importància dels patrons genòmics tumorals dependents de la cèl·lula originària del tumor, la caracterització dels processos oncogènics involucrats o la desregulació específica de les distintes vies de senyalització molecular (**Figura 10**).

Les tres categories especificades acumulen un sèrie d'un total de 27 articles de gran impacte, publicats en la revista *Cell*, que tracten la temàtica de cada una d'elles arribant i a certes conclusions generals:

- *"Cell-of-Origin Patterns"*: (patrons de cèl·lula originària) la cèl·lula originària del procés tumoral influencia de forma contundent, tot i que no de manera completa o definitiva, la classificació dels tumors desenvolupats, el que condueix a intuir futures implicacions clíniques i interpretatives a l'hora de l'estudi del càncer. S'estableixen nous indicis en la subclassificació dels càncers que inclouen els ginecològics i de mama, els gastrointestinals, esquamosos i els renals. A més, s'ha

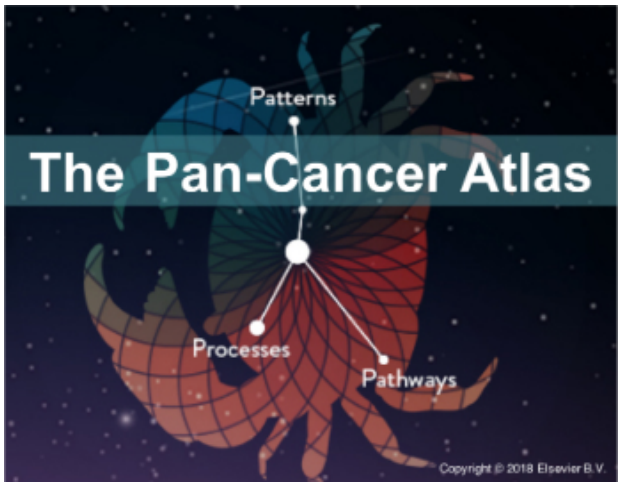


Figura 10. Esquema conceptual de la sèrie *Pan-Cancer Atlas.*
(Extreta de la web del Pan-Cancer Atlas)

identificat característiques de relacionades amb els fenotips de cèl·lules mare associades amb la desdiferenciació oncogènica (Hoadley et al., 2018).

- "*Oncogenic Processes*": (processos oncogènics) visió panoràmica dels processos relacionats amb la oncogènesis tumoral dels distints tipus de càncer. S'identifiquen els mecanismes pels quals les variants genètiques germinals i les mutacions somàtiques col·laboren en la progressió del càncer i s'explora la influència de les mutacions en la senyalització cel·lular i immunològica, facilitant eines per al desenvolupament de nous tractaments i immunoteràpies (Ding et al., 2018).
- "*Signalling Pathways*": (vies de senyalització) la identificació, caracterització i estudi de les vies moleculars de senyalització implicades en els distints tipus de càncers ha aportat patrons de vulnerabilitat que podran aprofitar-se per al desenvolupament de noves teràpies i tractaments personalitzats contra el càncer. Algunes de les més importants inclouen els gens *MYC* i *RAS*, la ubiqüitina, els processos de reparació del DNA, l'*splicing* i el metabolisme cel·lular (Sanchez-Vega et al., 2018).

1.4.2 Caracterització genòmica i molecular del càncer colorectal

El CCR va ésser un dels primers tumors sòlids sotmesos a una extensa caracterització molecular. La introducció de la tècnica de comparació genòmica per hibridació (CGH, de l'anglès *comparative genòmic hybridization*) (Kallioniemi et al., 1992) i, posteriorment, amb la versió que utilitza matrius de sondes al llarg del genoma –la aCHG (de l'anglès *array CGH*)– (Pinkel et al., 1998), va facilitar i millorar l'estudi i anàlisi dels perfils genòmics i la caracterització de les alteracions del número de còpia. L'estudi d'adenomes colorectals i carcinomes identificà les alteracions cromosòmiques característiques d'aquest tipus de càncer: els guanys del cromosoma 7, 8q, 13q i 20q i les pèrdues del 4q, 8p, 17p i 18q (Ried et al., 1996; Meijer et al., 1998; Douglas et al., 2004; Nakao et al., 2004). De fet, l'aplicació d'aquests estudis genòmics va ajudar en la diferenciació entre els casos de CCR amb MSS i els MSI segons els perfils d'alteracions cromosòmiques (Camps et al., 2006).

Al 2012, el consorci del TCGA publicà una caracterització extensiva de 276 mostres de CCR que suplementava estudis genòmics anteriors en aquesta neoplàsia (The Cancer Genome Atlas, 2012), que incloïa mutacions puntuals, canvis en el número de còpia, hipermetilació i nivells d'expressió gènica (Wood et al., 2007; Bass et al., 2011). L'anàlisi de les dades genòmiques de la seqüenciació de l'exoma permeté la classificació de les mostres en dos grups

tenint en compte la seva càrrega mutacional: híper-mutades, per a les mostres amb ratis de mutació majors de 12 mutacions per 10^6 pb i 728 mutacions de canvi de nucleòtid de mitjana; i les mostres no híper-mutades, amb ratis de mutació menors de 8,24 mutacions per 10^6 pb i una mitjana de 58 mutacions de canvi de nucleòtid. Al primer grup, de mostres híper-mutades, s'observà un enriquiment per a tumors amb MSI, fenotip CIMP i inactivació del gen *MLH1*. A més, la distribució de la freqüència de mutacions en gens importants es diferenciava entre els dos grups: en les mostres híper-mutades, els vuit gens més mutats foren *ACVR2A*, *APC*, *TGFBR2*, *BRAF*, *MSH3*, *MSH6*, *SLC9A9* i *TCF7L2*; mentre que per al grup de mostres no híper-mutades, els gens més mutats foren *APC*, *TP53*, *KRAS*, *PIK3CA*, *FBXW7*, *SMAD4*, *TCF7L2* i *NRAS* (The Cancer Genome Atlas, 2012).

Els perfils d'alteracions genòmiques en quant a esdeveniments cromosòmics o sub-cromosòmics també es diferenciaren entre ambdós grups (**Figura 11**). Les mostres híper-mutades es caracteritzaren per perfils molt baixos o inexistents en quant a esdeveniments genòmics de canvi en el número de còpia, mentre que les mostres amb ratis de mutacions puntuals baixos presentaven perfils amb alta càrrega d'alteracions, presentant les alteracions del número de còpia característiques dels perfils genòmics del CCR: guanys dels cromosomes 1q, 7, 8q, 13q i 20q, i les pèrdues de 1p, 4, 8p, 14q, 15q, 17p i 18q. Per altra banda, es detectaren regions més específiques (o focals) en forma de pèrdues o guanys recurrents entre les mostres amb alteracions genòmiques. Entre els gens afectats recurrentment per delecions focals eren *FHIT*, *RBFOX1*, *WWOX* i els supressors tumorals *SMAD4*, *APC*, *PTEN*, *SMAD3* i *TCF7L2* (The Cancer Genome Atlas, 2012).

Pel que fa a les amplificacions focals recurrents, algunes es trobaven en regions afectades per alteracions àmplies guanyades, com l'amplificació de 13q12.13, pròxima al gen *USP12* i adjacent al oncogen candidat CDK8; la 13q12; la 13q22, afectant al gen *KLF5*; i la 20q13.12, que involucra el gen *HNF4A*. En el cromosoma 8, s'identificaren les regions amplificades 8p12 (afectant als gens *WHSC1L1* i prop de *FGFR1*) i 8q24, que contenia el proto-oncogen *MYC*. Els gens *ERBB2* i *IGF2* s'observaren condicionats per les amplificacions focals més recurrents: 17q21.1, en un 4% de les mostres i 11p15.5, en un 7%, respectivament (Camps et al., 2009; The Cancer Genome Atlas, 2012).

En el mateix estudi, l'anàlisi integratiu de les dades genòmiques de mutacions puntuals, alteracions de canvis del número de còpia i expressió gènica va permetre caracteritzar aquelles vies de senyalització molecular alterades en les mostres de CCR (**Figura 12**). Específicament,

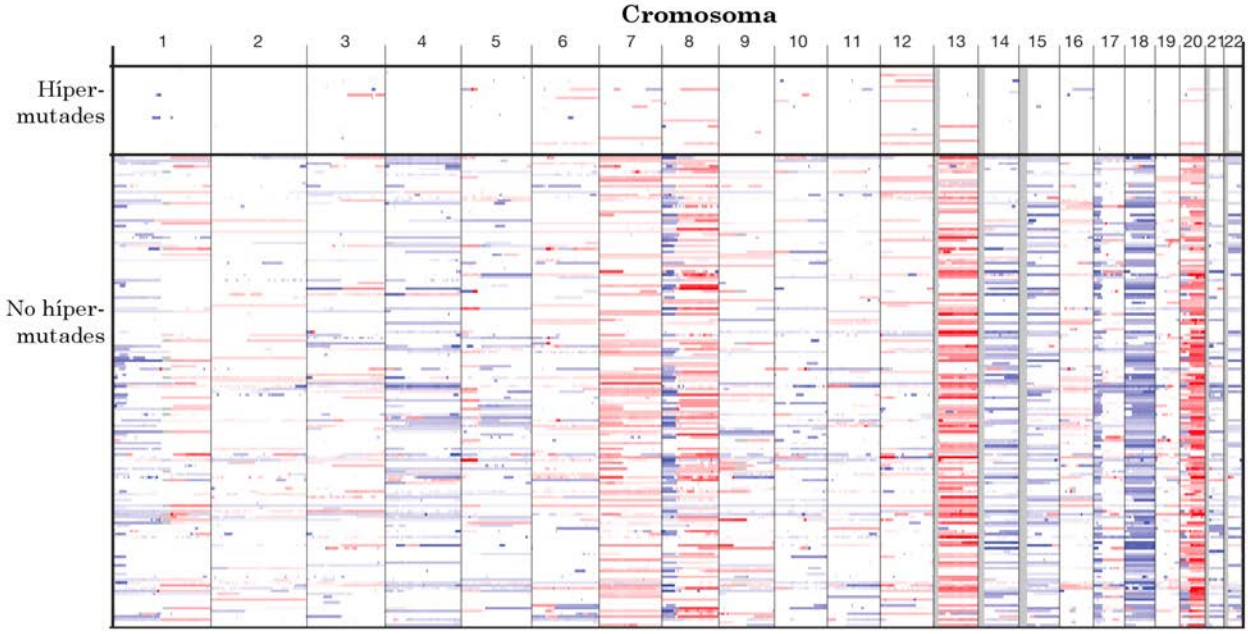


Figura 11. Alteracions genòmiques entre els grups de mostres híper-mutades i les no híper-mutades.
Les alteracions del número de còpia es presenten per als 22 cromosomes autosòmics en forma de guanys (vermell) i pèrdues (blau), per a les mostres híper-mutades (zona superior) i les mostres no híper-mutades (zona inferior). (Extreta de The Cancer Genome Atlas *Nature* 2012)

la via de WNT s'observà alterada en el 93% de les mostres (amb el ~80% dels casos degut a inactivació bi-al·lèlica del gen *APC* o per mutacions activadors de *CTNNB1*, mentre que la sobreexpressió del receptor de la via, *FZD10*, s'observà en ~17% dels casos). Les vies PI3K i RAS-MAPK també es trobaren alterades, amb mutacions exclusives en els gens *PIK3R1* i *PIK3CA* i delecions del gen *PTEN*; i amb el ~55% de tumors híper-mutats que presentaven alteracions als gens *KRAS*, *NRAS* o *BRAF*, amb patrons d'exclusivitat entre ells. Per altra banda, en quan a la via TGF-ß (Massagué, Blain, & Lo, 2000), normalment inactivada al CCR, s'identificaren alteracions genòmiques que implicaven els gens *TGFBR1*, *TGFBR2*, *ACVR2A*, *ACVR1B*, *SMAD2*, *SMAD3* i *SMAD4* al 27% de mostres no híper-mutades i al 87% de mostres híper-mutades. Finalment, la via de p53 s'observà alterada com a conseqüència de mutacions i alteracions genòmiques del gen *TP53*, gairebé totes a nivells bi-al·lèlic, en el 59% de les mostres no híper-mutades, mentre que el gen *ATM* es trobà alterat en el 7% d'aquestes mostres (Brady et al., 2011; The Cancer Genome Atlas, 2012).

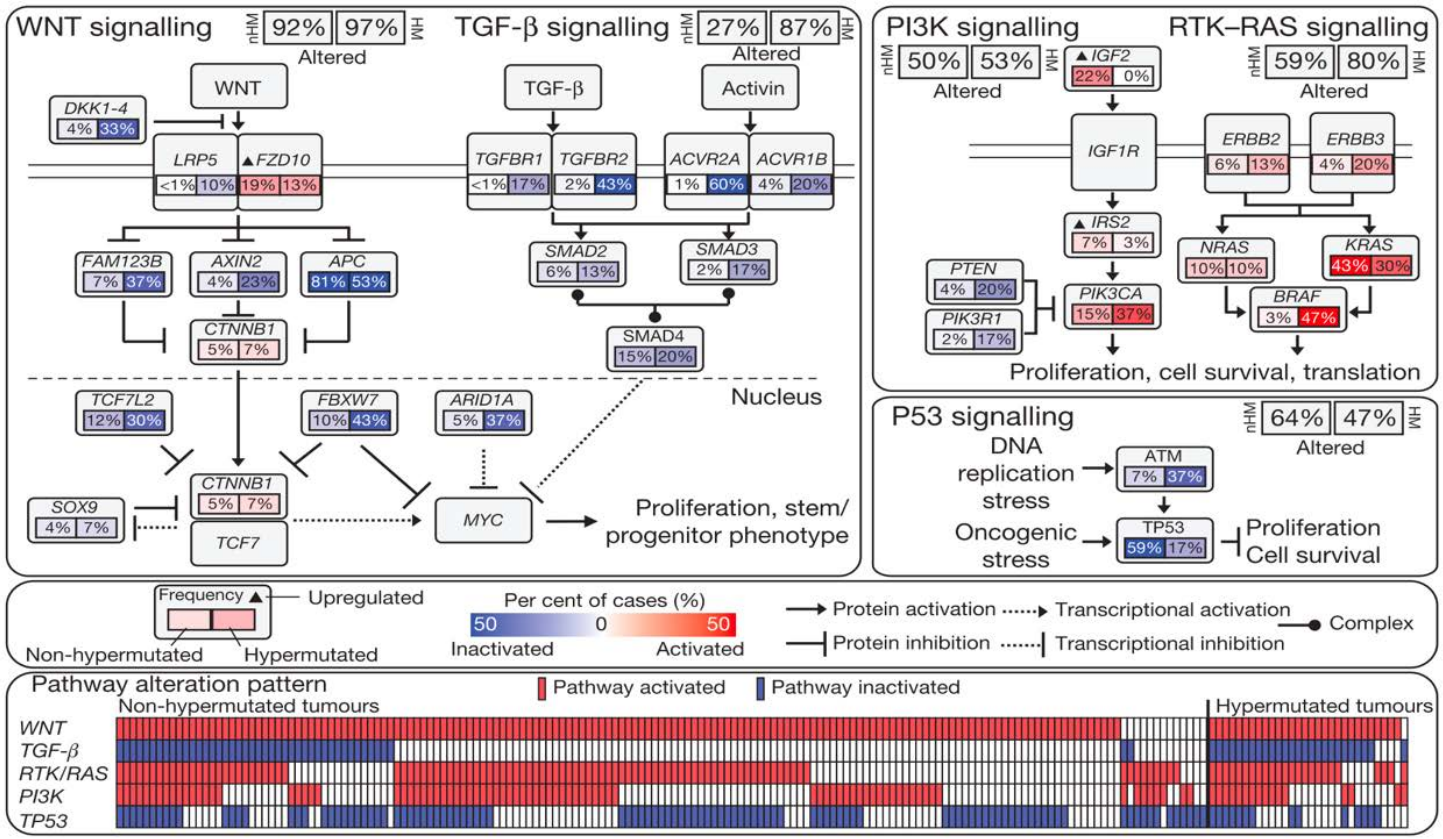


Figura 12. Vies de senyalització cel·lular alterades en el CCR.
Les alteracions genòmiques identificades en l'estudi del TCGA, tant mutacions puntuals com canvis en el número de còpia, provoquen la des-regulació de les vies de senyalització WNT, PI3K, RAS-MAPK, TGF-ß i p53. Al esquemes de les vies de senyalització s'indiquen les freqüències d'alteració (percentatge de mostres) en que es veuen afectats els distints gens, diferenciant entre els grups de mostres híper-mutades (valor de la dreta) i les no híper-mutades (valor de l'esquerre). En la part inferior de la imatge s'hi indiquen les mostres d'ambdós grups i l'estat de cada via de senyalització en la mostra (vermell: activa; blau: inactiva. (Extreta de The Cancer Genome Atlas *Nature* 2012)

1.4.3 The Consensus Molecular Subtypes

Com a conseqüència d'aquesta caracterització genòmica i molecular exhaustiva dels últims anys, facilitada per la disponibilitat de dades biològiques que es generen mitjançant l'ús de les plataformes òmiques, s'han dut a terme diferents intents per a la classificació molecular del CCR. L'últim d'aquests ha estat l'anomenat *The Consensus Molecular Subtypes*, com a conseqüència d'un gran estudi de col·laboració que posà en comú distintes cohorts de tumors CCR amb anotacions clíniques i moleculars disponibles (Guinney et al., 2015; Dienstmann et al., 2017).

La classificació es basa en els perfils d'expressió gènica de les mostres per a les distintes cohorts de CCR analitzades. Posteriorment, els investigadors d'aquest estudi es dedicaren a la caracterització genòmica i molecular dels subtipus generats. Com explica la **Taula 1**, els distints subtipus presentaren mostres amb característiques moleculars específiques que els conferien diferències en quan a la resposta immunitària, l'activació de certes vies de senyalització cel·lular, la desregulació metabòlica o l'adaptació cel·lular i tissular dels tumors (com l'angiogènesi i la infiltració de l'estroma).

Taula 1. Classificació taxonòmica dels subtipus de CMS per al CCR.

CMS1 MSI immune	**CMS2** canònica	**CMS3** metabòlica	**CMS4** mesenquimal
14%	37%	13%	23%
MSI, CIMP elevada, híper-mutació	SCNA elevada	MSI/MSS, SCNA baix i CIMP baix	SCNA elevada
Mutacions a *BRAF*		Mutacions a *KRAS*	
Infiltració i activació immunitària	Activació de WNT i MYC	Desregulació metabòlica	Infiltració de l'estroma, activació via TGF-ß i angiogènesis
Pitjor supervivència després de recurrència			Elevada recurrència i pitjor supervivència

MSI: microsatellite instability; MSS: micrsatellite stability;CIMP: CpG islands methylation phenotype; SCNA: somatic copy number alterations.

Més en detall, el grup CMS1 englobaria gairebé totes les mostres MSI (amb l'excepció d'algunes mostres al grup CMS3), caracteritzades per l'absència d'alteracions del canvi de número de còpia somàtica (SCNA, en

anglès *somatic copy number alterations*), gran càrrega mutacional, amb enriquiment de mostres amb mutacions al gen *BRAF*, amb forta infiltració immunitària i demostrant una pitjor supervivència després de la recurrència tumoral. Per altra banda, les mostres MSS es repartirien entre els demés grups. El CMS2 representa el grup més canònic i majoritari, amb els perfils genòmics d'inestabilitat cromosòmica característics i amb activació de la via WNT, protagonista en la carcinogènesi del CCR, i alteracions al proto-oncogen *MYC*. El grup CMS3 es presenta com el subtipus més heterogeni, amb nivells de SCNA mitjans, i amb enriquiment de mostres amb mutacions a l'oncogen *KRAS* i amb activació metabòlica de sucres i àcids grassos. Per últim, el CMS4 presenta perfils genòmics similars al grup CMS2, infiltració de l'estroma i activació de la via de senyalització del TGF-ß, amb elevada recurrència tumoral i baixa supervivència global.

Els estudis integrats d'aquestes classificacions amb les variables moleculars, genòmiques i clíniques s'encaminen a determinar subgrups de mostres susceptibles a ésser tractades amb teràpies específiques. Tot i així, l'alta heterogeneïtat inter-tumoral al CCR a distints nivells no facilita la tasca. A més, la complexitat augmenta quan la heterogeneïtat es pot estudiar a nivells espacial (intra-tumor) i temporal (en distints punts de la carcinogènesis) (Saito et al., 2018). Per tant, la identificació de futurs subgrups de tumors encara més homogenis passa per tenir en compte aquests principis d'heterogeneïtat, per tal d'avançar tant en la prevenció del desenvolupament del tumor, com en el tractament específic del mateix (McGranahan & Swanton, 2017).

2 Càncer colorectal germinal

El CCR és d'entre les neoplàsies comuns la que presenta més proporció de casos amb agregació familiar. Estudis de parents i de bessons han estimat que un 30-35% dels casos de CCR presenten alguna forma d'herència familiar de la malaltia, o CCR germinal (Eric R. Fearon, 2011). Aproximadament un quart dels casos que presenten agregació familiar per CCR s'identifica entre les anomenades formes hereditàries del CCR. Aquestes síndromes de predisposició al CCR es manifesten a una edat més primerenca en comparació a les formes esporàdiques del CCR (Lichtenstein et al., 2000; Grady, 2003). Això és degut a la càrrega genètica heretada en gens implicats en la carcinogènesis del CCR, la majoria d'ells desenvolupant un paper com a TSGs.

2.1 Hereditari

Aproximadament el 35% dels casos de CCR presenten agregació familiar i s'associen a factors de risc genètics (Jasperson, Tuohy, Neklason, & Burt, 2010). El 5-10% d'aquests casos familiars es troben associats a mutacions d'alta penetrància en gens coneguts que aporten alt risc a desenvolupar el CCR. Una de les majors distincions entre les síndromes hereditàries és la presentació de poliposis, diferenciant-los entre les síndromes polipòsiques i les no-polipòsiques (**Figura 13**) (Burt et al., 2004).

Càncer colorectal no-polipòsic o síndrome de Lynch

El CCR no-polipòsic (HNPCC, de l'anglès *hereditary nonpolyposis colorectal cancer*) o síndrome de Lynch fou una de les primeres síndromes del càncer descrits (H T Lynch & Krush, 1971). Els individus que la pateixen presenten predisposició a distints tipus de càncer, però especialment al CCR i d'endometri (Umar et al., 1994; H T Lynch, Smyrk, & Lynch, 1997). Es caracteritza per desplegar patrons d'herència familiar autosòmica dominant, amb individus afectats de CCR sense presentar poliposis. Per altra banda, aquests càncers presenten infiltració limfocitària augmentada i una diferenciació limfocítica semblant a la de la malaltia de Crohn, entre d'altres aspectes histològics específics. El diagnòstic d'aquesta síndrome es basa en criteris d'agregació familiar coneguts com a criteris d'Amsterdam, posteriorment revisats i actualitzats amb les clíniques Bethesda (**Taula 2**) (Umar et al., 2004).

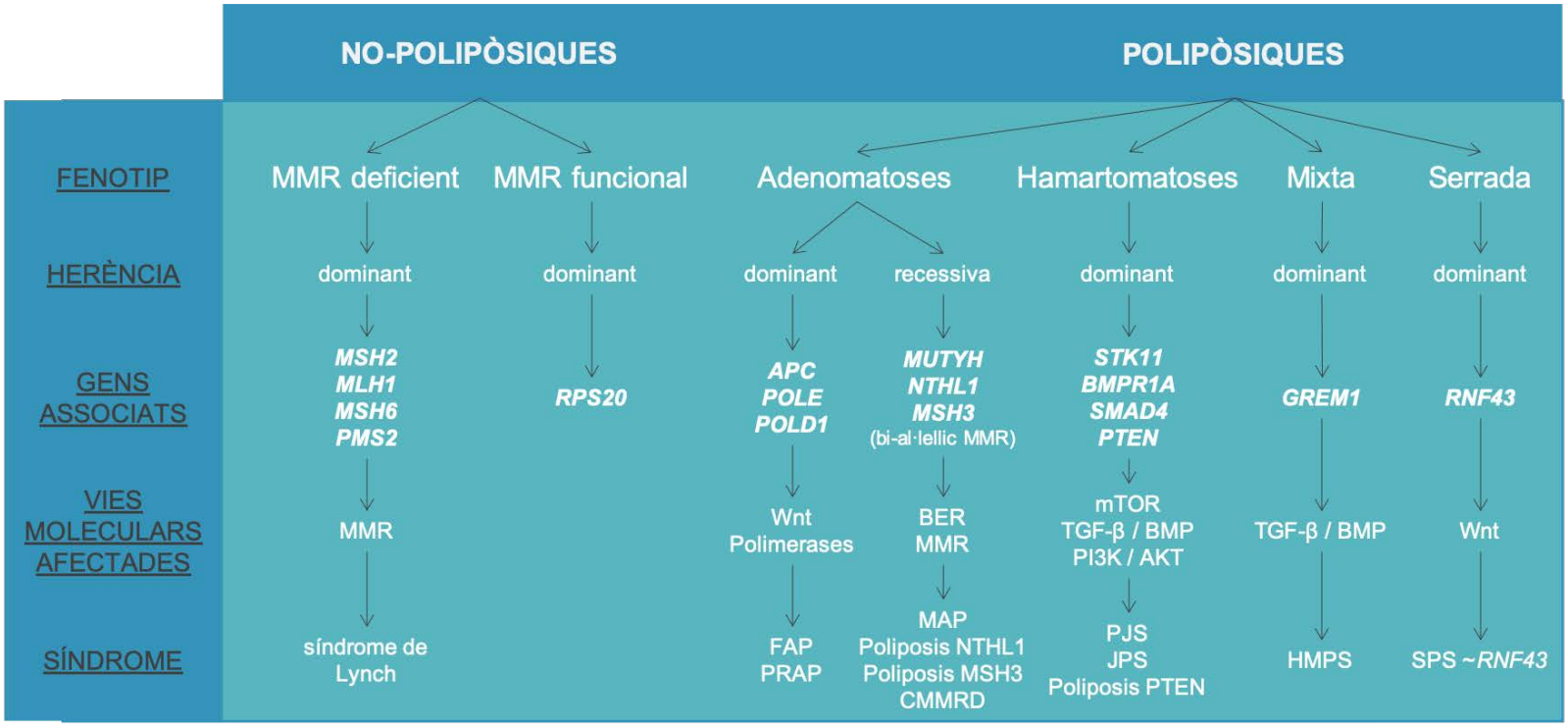


Figura 13. Síndromes hereditàries del CCR: fenotip, herència genètica i vies moleculars afectades.
MMR: mismatch repair; FAP: familial adenomatous polyposis; PPAP: polymerase proofreading-associated polyposis; BER: base-excision repair; MAP: MUTYH-associated polyposis; CMMRD: constitutional mismatch repair deficiency; PJS: Peutz-Jeghers syndrome; JPS: Juvenile Polyposis syndrome; HMPS: hereditary mixed polyposis syndrome. (Adaptada de Valle L et.al *The Journal of Pathology* 2018)

L'estudi de famílies que complien aquests criteris clínics per a la identificació del risc a desenvolupar síndrome de Lynch (Vasen, Watson, J, & Lynch, 1999; H. Lynch et al., 2009), suposà la identificació d'alteracions genòmiques en regions enriquides amb seqüències de DNA repetitiu, anomenades "microsatèl·lits". Aquest tipus de fenotip, lligat al sistema de reparació MMR defectuós, apuntà a alteracions genètiques en gens implicats d'aquest sistema. Així, es sap que els casos de HNPCC es donen a conseqüència de mutacions als gens *MSH2*, *MLH1*, *PMS2* i *MSH6*. Els casos amb mutacions als gens MSH2 i *MLH1* representen el 70% dels HNPCC (Bocker, Rüschoff, & Fishel, 1999; Henry T. Lynch & de la Chapelle, 2003; H. Lynch et al., 2009). Tot i així, fins a un 60% dels casos que compleixen els criteris de diagnòstic per HNPCC no presenten mutacions genètiques en gens de reparació MMR, classificant-los en el que s'anomena síndrome del CCR familiar tipus X (Lindor et al., 2005).

Taula 2. Guies clíniques Bethesda revisades i criteris Amsterdam per a la identificació de pacients amb risc a desenvolupar síndrome de Lynch (o HNPCC)

Guia Bethesda:

- CCR diagnosticat en pacient menor dels 50 anys
- Presència de CCR sincrònic, metacrònic o altres tumors associats a síndrome de Lynch, sense considerar l'edat
- CCR diagnosticat amb MSI-H verificat histològicament en un pacient menor dels 60 anys
- CCR diagnosticat en un o més pacients de primer grau amb tumor associat a la síndrome de Lynch, amb un dels tumors diagnosticat abans dels 50
- CCR diagnosticat en dos o més pacients de primer o segon grau amb tumors associats a síndrome de Lynch, sense considerar l'edat

Criteris d'Amsterdam I i II:

- Un individu diagnosticat de CCR (or tumor extra-colònic associat a la síndrome de Lynch) abans dels 50 anys
- Tres individus afectats, un d'aquests relatiu de primer grau dels altres dos
- Dues generacions successives afectades
- La poliposis adenomatosa familiar descartada
- Tumors examinats i verificats a nivell patològic

Poliposis adenomatosa familiar

La poliposis adenomatosa familiar (FAP, de l'anglès *familial adenomatous polyposis*) és la segona síndrome més comuna implicada en l'herència del CCR, amb una prevalença de un cada 10.000 individus. Es caracteritza per presentar grans quantitats de pòlips benignes, d'aquí el gran risc a desenvolupar CCR, tot i que existeix una forma atenuada d'aquesta on es presenta menor nombre de pòlips adenomatosos (Stoffel et al., 2009). La incidència a patir CCR al llarg de la vida d'un individu és aproximadament del 100% (Burt et al., 2004).Tot i que la FAP és una síndrome autosòmica dominant, fins un 25% dels casos d'aquesta es desenvolupa com a conseqüència de mutacions germinals *de novo* i, per tant, no mostren el patró d'herència. La causa genètica d'aquesta la trobem en l'adquisició de mutacions germinals al gen *APC* (Groden et al., 1991; Kinzler et al., 1991; Lamlum et al., 2000). El 95% d'aquestes mutacions són de canvi de pauta de lectura (*frameshift*, en anglès) o de pèrdua de sentit (en anglès, *nonesense*), que provoquen la pèrdua de l'al·lel mutat per truncament prematur de la síntesis proteica (Knudsen, Bisgaard, & Bülow, 2003). La majoria de casos de FAP evolucionaran per la via CIN, seguint la seqüència adenoma-carcinoma clàssica. Altres presentacions clíniques i tumors extra-colònics s'han associat amb el FAP, com la síndrome de Turcot (on es combinen la poliposis i certs tumors cerebrals) o la síndrome Gardner (presentació d'osteomes de crani i mandíbula o anormalitats dentals i fibromes) (Rustgi, 2007).

Poliposis associada a MUTYH

La poliposis associada a *MUTYH* (MAP, de l'anglès *MUTYH-associated polyposis*), com el seu nom dona a entendre, s'origina com a conseqüència de la inactivació bi-al·lèlica del gen *MUTYH* (Al-Tassan et al., 2002; Sampson, Jones, Dolwani, & Cheadle, 2005; Balaguer et al., 2007). Aquesta síndrome presenta una poliposis que es podria confondre amb la presentada en la forma atenuada de FAP, tot i que també s'han trobat casos amb gairebé cap pòlip i amb inactivació bi-al·lèlica de *MUTYH* (Henry T. Lynch & de la Chapelle, 2003). El gen participa en la via de reparació del DNA per escissió de base (BER, de l'anglès *base-excision repair*), encarregada de prevenir transversions de G:C cap a T:A ocasionades per l'estrès oxidatiu cel·lular (Al-Tassan et al., 2002; Cleary et al., 2009).

Poliposis hamartomatoses

Entre les poliposis hamartomatoses, associades amb un risc alt de desenvolupar CCR i altres neoplàsies, s'hi consideren les síndromes

Peutz-Jeghers (PJS, del seu nom en anglès Peutz-Jeghers *syndrome*) i de poliposis juvenil (JPS, de l'anglès Juvenile Polyposis *syndrome*) (Shepherd, Bussey, & Jass, 1987; Olschwang, Serova-Sinilnikova, Lenoir, & Thomas, 1998). La síndrome de Cowden també es pot considerar dintre d'aquest grup, la qual sorgeix per mutacions al gen *PTEN*, considerat com a TSG i que participa de la via de senyalització PI3K, associada a la supervivència cel·lular (Al-Tassan et al., 2002; Jelsig et al., 2014). El PJS s'associa amb mutacions al gen *STK11*. La seva inactivació comporta la híper-activació de la via cel·lular mTOR, encarregada del control de les reserves energètiques de la cèl·lula i la seva proliferació. Mutacions germinals als TSGs *SMAD4* o *BMPR1A* són responsables del fenotip JPS. Ambdós gens participen de la via de senyalització TGF-ß (Pilarski, 2009).

Poliposis mixta

La síndrome de la poliposis hereditària mixta (HMPS, de l'anglès *hereditary mixed polyposis syndrome*) presenta un fenotip de poliposis inusual amb múltiples i distintes morfologies alhora. Aquesta patologia està associada a la duplicació de la zona reguladora del gen *GREM1*, el qual presenta a un paper antagonista en la via reguladora BMP (Whitelaw et al., 1997; Jaeger et al., 2012).

Síndrome de la poliposis serrada

La síndrome de poliposis serrada (SPS) es caracteritza per la presentació de pòlips serrats distribuïts al llarg del còlon i un gran increment del risc a patir CCR (Young & Jass, 2006; Rosty, Parry, & Young, 2011). Es diagnostica de forma clínica mitjançant els criteris establerts per l'Organització Mundial de la Salut (Snover, Jass, Fenoglio-Preiser, & Batts, 2005):

- Presència d'almenys cinc pòlips serrats proximals a la regió sigma, essent dos d'aquests iguals o més grans de 10 mil·límetres.
- Qualsevol nombre de pòlips serrats proximals al colon sigmoide en un individu amb familiar de primer grau diagnosticat de SPS.
- Més de 20 pòlips serrats de qualsevol mida i distribuïts al llarg del còlon.

Per altra banda, els factors germinals de risc associats a la SPS encara són prou incerts. Diversos estudis han proposat com a possible origen d'aquesta patologia les mutacions en el gen *RNF43* (Taupin et al., 2015; Yan et al., 2017), tot i que d'altres aporten conclusions desfavorables a aquesta associació genètica (Buchanan et al., 2017). El gen *BRAF* s'ha identificat mutat de forma somàtica als pòlips serrats dels pacients diagnosticats de SPS (Beach et al.,

2005), mentre que un percentatge minoritari dels casos es vinculen a mutacions germinals al gen *MUTYH* en pacients que també presenten adenomes tubulars (Boparai et al., 2008).

Poliposis associades a defectes en les polimerases

Als últims anys, l'estudi del CCR familiar ha resultat en la caracterització de les poliposis associades a defectes en polimerases (PPAP, de l'anglès *polymerase proofreading-associated polyposis*) (Briggs & Tomlinson, 2013). Aquestes formes hereditàries consisteixen en mutacions germinals als gens *POLE* i *POLD1* (sobretot al domini exonucleasa d'aquestes unitats funcionals) i que presenten patrons d'herència dominant amb poliposis clàssica o atenuada (Palles et al., 2013; Valle et al., 2014; Esteban-Jurado et al., 2017). Els casos de CCR i altres tumors amb aquests gens afectats presenten càrregues de hípermutació a nivell somàtic, normalment amb fenotips de deficiència de MMR (Rayner et al., 2016).

2.2 Familiar

Aquells casos amb agregació familiar per al CCR però que no presenten mutacions genètiques als gens de les formes hereditàries constitueixen el percentatge amb factors de risc genètics desconeguts de la malaltia. Diversos estudis apunten a que aquesta herència desconeguda podria venir donada per variants genètiques de susceptibilitat que estarien aportant risc alt o moderat per al CCR (Woods et al., 2010; Valle, 2017; Win et al., 2017).

2.2.1 Càncer colorectal familiar de tipus X

Els individus que han desenvolupat CCR i compleixen els criteris d'Amsterdam però, per altra banda, no presenten deficiència en el sistema de reparació del DNA per MMR es classifiquen com "CCR familiar de tipus X" (Lindor et al., 2005). En comparació a les famílies afectades per síndrome de Lynch, les classificades dins aquest tipus de CCR familiar presenten una edat mitja de diagnòstic per al CCR de fins a 10 anys més alta. Per altra banda, aquestes presentarien un risc menor

de presentar la malaltia i sense increment del risc de presentar tumors extra-colònics (Peters et al., 2015).

Des de el punt de vista genètic, sembla ser que aquest és un grup molt heterogeni. Per això i donada la forta agregació familiar que presenten aquests casos, multitud d'estudis per tal d'identificar els possibles factors de risc genètics s'han dut a terme. La hipòtesi rere aquests estudis proposa l'herència de variants d'alta penetrància situades en gens encara per descobrir. En els últims anys, l'escalada en l'ús de plataformes genòmiques de seqüenciació massiva ha permès avenços en l'estudi d'aquestes famílies, sobretot mitjançant l'anàlisi de dades de seqüenciació del genoma i l'exoma. Gràcies a aquests esforços s'han identificat nous gens implicats en la predisposició al CCR, caracteritzant noves formes genètiques de la malaltia, com la forma hereditària PPAP (Briggs & Tomlinson, 2013; Palles et al., 2013).

2.2.2 Gens de penetrància moderada i associats a altres neoplàsies

La significança clínica de les variants de penetrància moderada en la susceptibilitat al CCR encara ara és dubtosa i no s'estableix de forma clara. Al CCR, les mutacions més prevalent en aquest sentit són la p.I130K a *APC*, la c.1100delC i p.I157T al gen *CHEK2*, i les mutacions mono-al·lèliques al gen *MUTYH* (Yurgelun et al., 2015, 2017; Valle, Vilar, Tavtigian, & Stoffel, 2018).

Per altra banda, les estratègies de cribratge tradicional i els estudis per panells gènics en casos de CCR han identificat diversos gens mutats que fins ara havien estat associats a la susceptibilitat a altres tipus de neoplàsies, com el càncer de mama o d'ovari. En aquest sentit, estudis d'entre cents i fins a uns 2000 casos de CCR i controls van poder determinar els gens *ATM*, *BRCA1*, *BRCA2*, *CDKN2A*, *PALB2* i *TP53* com de susceptibilitat moderada per al CCR (Yurgelun et al., 2017; Katona et al., 2018). Entre aquests, els més mutats són *BRCA1* i *BRCA2* (0,7-1,3% dels casos), seguits del gen *ATM* (0,7-0,9%). Tot i aquesta identificació de gens no relacionats amb el CCR i mutats en casos de CCR familiar, les seves conseqüències en quan a l'augment de risc a patir la malaltia o a la vigilància clínica que s'hauria d'oferir als portadors d'aquestes variants i als seus familiars són aspectes encara sense definir. No està clar si aquests indicis responen a un enriquiment poblacional en la prevalença d'aquests tipus de mutacions o, si en realitat, responen a un fenomen de pleiotropisme, és a dir, la capacitat de què un mateix genotip pugui estar involucrat en el desenvolupament de diversos fenotips diferents (Dobbins et al., 2016; Valle et al., 2018).

2.3 Identificació de nous gens de predisposició al càncer colorectal

La predisposició genètica al càncer es ve observant mitjançant l'estudi de l'agregació inusual de casos en famílies des de fa uns quants segles. Les variants que afecten certs gens en la línia germinal confereixen alt i moderat risc a patir el fenotip de la malaltia, i són l'objectiu a identificar en aquests tipus d'estudis. Durant els últims 30 anys, aquests gens de predisposició al càncer s'han anat identificant mitjançant distints tipus d'aproximacions, entre ells les anàlisi de lligament, els estudis d'associació del genoma i les aproximacions dirigides al cribratge de gens candidats.

Abans de l'arribada de les plataformes de seqüenciació genòmica massiva, la identificació dels gens implicats al càncer hereditari es basava en l'estudi d'associació del genoma complet i dels desequilibris de lligament, seguidament d'una caracterització d'aquelles regions d'interès identificades mitjançant la seqüenciació específica. Aquests tipus d'aproximacions van permetre la identificació de la majoria de gens implicats en la predisposició al CCR, donada la seva alta penetrància i la forta agregació familiar que provoquen (**Figura 13**) (Chubb et al., 2016; Valle, 2017; Valle et al., 2019).

Estudi de desequilibri de lligaments

La tècnica d'anàlisi del lligament es basa en l'estudi de la segregació de marcadors genòmics informatius, com regions de microsatèl·lits específiques o SNPs (de l'anglès, *single nucleotide polymorphism*), dins d'una família amb individus afectats per la patologia d'estudi i que presenta agregació per aquesta (Slatkin, 2008). Les regions del genoma que es transmeten en bloc entre els membres afectats, sense tenir en compte cap tipus de recombinació cromosòmica implicada, es seleccionen i es calcula la probabilitat de què aquesta segregació sigui degut a la malaltia (Haseman & Elston, 1972). L'associació d'aquestes regions d'herència amb els patrons familiars de la patologia estudiada confirma el desequilibri de lligament.

Aquest tipus d'estudi són útils per a identificar regions genòmiques d'herència mendeliana d'alta penetrància. De fet, la majoria de gens de predisposició d'alta penetrància, responsables de les formes hereditàries explicades abans, es van identificar seguint aquest model d'estudi (Leppert et al., 1987; Nishisho et al., 1991).

Estudis d'associació del genoma complet

Els estudis d'associació del genoma complet, o GWAS (de l'anglès, *genome-wide association studies*) es recolzen en l'anàlisi del genotipat de mostres problema i controls de milers de SNPs distribuïts al llarg del genoma. Aquelles variants que es troben enriquides al grup de mostres problema es classifiquen com a variants de risc per a la patologia estudiada. En aquest cas, els estudis de GWAS s'utilitzen per a la identificació massiva de variants de baixa penetrància associada al fenotip patològic estudiat, i necessiten de grans nombres de mostres en les cohorts d'estudi per a la seva pràctica (Pritchard, Stephens, & Donnelly, 2000; Risch, 2000).

Els estudis GWAS van permetre la identificació de variants de baixa penetrància afectant gens de vies de senyalització implicades en processos de carcinogènesi i, per tant, associant aquestes vies al desenvolupament del CCR. Alguns exemples serien els gens implicats en la via de TGF-ß (*BMP2*, *BMP4*, *SMAD7*, *CCND2*, *GREM1*) o gens la via de les MAPK (*DUSP10*, *MYO1B*, *MYC*, *CCND2*, *SH2B*) (Tenesa & Dunlop, 2009; Tomlinson et al., 2011).

Curiosament, mitjançant aquest tipus d'estudi, s'han identificat variants de baixa penetrància en gens que també poden presentar variants d'alt risc. Un exemple d'això és el gen *GREM1* (com s'ha explicat, responsable protagonista de la síndrome hereditari HMPS) al qual ja s'hi varen identificar variants de baixa penetrància (Tomlinson et al., 2011), i en els últims anys també s'hi ha caracteritzat mutacions d'alta penetrància per al CCR (Venkatachalam et al., 2011; Jaeger et al., 2012; Rohlin et al., 2016).

NGS i matrius genòmiques

Durant l'última dècada, el ràpid desenvolupament de la seqüenciació de nova generació i les tècniques de caracterització de les variacions del número de còpia al genoma han permès encaminar els estudis per a la identificació de nous gens implicats en la predisposició al càncer, després d'arribar a un punt d'incapacitat amb les tècniques tradicionals descrites abans. En els últims anys, les tècniques més utilitzades per la identificació de noves variants de predisposició han estat la seqüenciació de l'exoma (WES, de l'anglès *whole exoma sequencing*), del genoma (WGS, de l'anglès *whole genome sequencing*) i les tècniques basades en matrius de sondes genòmiques per a la identificació de variants del número de còpia genòmica (Valle, 2017).

Aquestes aproximacions genòmiques s'han aplicat en l'estudi de famílies amb grans nombres de pacients o, també, en cohorts familiars que presenten individus amb edats prematures de diagnòstic del càncer i forta agregació per la malaltia, per tal d'identificar les variants d'alta i moderada penetrància implicades en la predisposició. Un exemple satisfactori d'aquests estudis fou la

recent identificació de les polimerases *POLE* i *POLD1*. Mutacions germinals que afecten als dominis exonucleasa en els gens que codifiquen per aquestes proteïnes se'ls ha pogut atribuir la característica fenotípica que consisteix en perfils d'híper-mutació al tumor, relacionant-les com a nova síndrome d'alt risc al CCR (Briggs & Tomlinson, 2013; Palles et al., 2013). De forma recent, alguns casos de poliposis adenomatoses també s'han associat a formes d'herència recessiva com a conseqüència de inactivacions bi-al·lèliques dels gens *NTHL1*, implicat en la via *base-excision repair*, i *MSH3*, implicat en la via MMR, però no associat a la síndrome de Lynch (**Figura 13**) (Weren, Ligtenberg, et al., 2015; Adam et al., 2016).

2.3.1 Estudis de gens candidats

En els últims anys s'han dut a terme distints estudis aplicant la seqüenciació massiva per WES en cohorts de pacients afectats per CCR per tal d'identificar nous gens candidats a la predisposició a la malaltia. Diferents aproximacions s'han dut a terme en aquest sentit, com la identificació de noves variants en gens que ja han estat identificats com a responsables de la predisposició, o com l'estudi de nous gens que participen de vies moleculars implicades al desenvolupament de la malaltia (Valle, 2017).

Aquests estudis es dedicaren a presentar llistats de gens candidats rars potencialment relacionats amb la predisposició al CCR (Valle et al., 2019). Els criteris de selecció responien a l'associació depenent de la seva funció gènica, la participació en vies de senyalització cel·lular implicades en el desenvolupament de la patologia, l'enriquiment de variants en cohorts amb pacients diagnosticats a edats primerenques, l'absència en pacients sans i la co-segregació familiar de les variants (C. G. Smith et al., 2013; DeRycke et al., 2013; Gylfe et al., 2013; Esteban-Jurado et al., 2015, 2016; Tanskanen et al., 2015; Chubb et al., 2016; Hahn et al., 2016; Spier et al., 2016; Thutkawkorapin et al., 2016; Yu et al., 2018).

Posteriorment, la importància d'alguns d'aquests gens s'ha anat validant mitjançant estudis funcionals, entre ells els mencionats *POLE*, *POLD1*, *NTHL1* o *MSH3* (Briggs & Tomlinson, 2013; Palles et al., 2013; Weren, Ligtenberg, et al., 2015; Adam et al., 2016). Entre els nous gens candidats amb majors evidències d'associació al CCR hereditari es troben els gens *FAN1* (Seguí et al., 2015), *RPS20* (Nieminen et al., 2014), *LRP6* (de Voer et al., 2016), *BUB1* i *BUB3* (de Voer et al., 2013) i

SEMA4A (Schulz et al., 2014). Els gens *FAN1*, *BUB1* i *BUB3* desenvolupen el seu paper en la resposta al dany del DNA i la inestabilitat genètica, mentre que *LRP6* participa de la via WNT. El gen *RPS20* ha estat un dels últims en demostrar forta associació amb casos de CCR no-polipòsics (**Figura 13**) (Nieminen et al., 2014), mentre que estudis encaminats a avaluar el risc aportat pel gen *SEMA4A*, inicialment fortament associat al CCR hereditari, van fallar en aquesta determinació (Kinnersley et al., 2016). Altres gens candidats identificats són *WRN*, *ERCC6*, *BLM* (de Voer et al., 2015), *SMAD9*, *FOCAD* (Weren, Venkatachalam, et al., 2015), *SETD6* (Martín-Morales et al., 2017) i *BRF1* (Bellido et al., 2018).

3 Variants del número de còpia

La introducció de noves eines i plataformes de seqüenciació massiva durant les últimes dècades han suposat una gran empenta per a la caracterització del genoma humà i el descobriment de la variabilitat genòmica entre els individus. La finalització del projecte del genoma humà va identificar la seva gran diversitat (Venter et al., 2001; Iafrate et al., 2004; Chaisson, Wilson, & Eichler, 2015). Tot i que l'espectre de variacions del genoma va des de canvis de base fins als esdeveniments cromosòmics, gran part de variabilitat entre individus és degut a les variacions estructurals. De fet, aproximadament el 9% del genoma humà semblaria estar afectat per variants del número de còpia (CNVs, de l'anglès *copy number variants*) (I. H. G. S. Consortium, 2001; Venter et al., 2001; Human Genome Sequencing Consortium, 2004; Zarrei, MacDonald, Merico, & Scherer, 2015).

Les CNVs formen part del conjunt de les variants estructurals. Es tracten de variacions genòmiques "no balancejades", és a dir, aquelles que aporten variabilitat quantitativa al genoma, alterant el seu caràcter diploide (Conrad et al., 2010). Per tant, es podrien definir com a seccions del genoma que varien en el número de còpia (duplicacions o delecions) i que aporten diversitat a la població genètica humana (Iafrate et al., 2004; Spielmann, Lupiáñez, & Mundlos, 2018).

Al 2006, Redon i col·laboradors aportaren una caracterització global del genoma humà en quant a la seva variabilitat com conseqüència de les CNVs (Redon et al., 2006). El treball de revisió de Zarrei i col·laboradors, al 2015, suposà una actualització d'aquest mapa global de les CNVs al genoma humà (Zarrei et al., 2015). En ella s'hi analitzen les variants llistades a la base de dades *Database of Genomic Variants* (DGV), la qual, des de la seva creació al 2004, ha anat catalogant les CNVs identificades en els estudis genòmics que s'han anat generat en la comunitat científica. A hores d'ara, la DGV conté una selecció curada de més de 6 milions de CNVs, corresponent a 72 estudis genòmics diferents que analitzen individus sans. En l'anomenada revisió, i com a notes curioses, s'identifica la contribució al catàleg de CNVs del 4,8-9,5% del genoma i, a més, s'observa l'existència de fins a 100 gens sense conseqüències funcionals després de la seva deleció total (Zarrei et al., 2015).

Per altra banda, el concepte d'alteracions del número de còpia (CNAs, de l'anglès *copy number alterations*) va molt lligat al context tumoral (Tang & Amon, 2013). En l'estudi del càncer a nivell somàtic es poden diferenciar les alteracions àmplies (*broad,* en anglès), que

esdevenen en estats cromosòmics d'aneuploïdia (guanys i pèrdues de cromosomes sencers), i les alteracions focals, aquelles que afecten regions molt específiques i localitzades del genoma de les cèl·lules canceroses, que esdevenen en amplificacions o delecions (Weaver & Cleveland, 2006; Holland & Cleveland, 2009; Tang & Amon, 2013; Davoli, Uno, Wooten, & Elledge, 2017).

L'aneuploïdia, o l'adquisició i presentació d'alteracions numèriques i estructurals dels cromosomes, és una característica definitòria del genoma de les cèl·lules malignes als tumors. Aquesta és present en aproximadament el 90% dels tumors sòlids i, com ja s'ha explicat, es considera un dels processos més importants a l'hora de facilitar i impulsar la progressió tumoral (*hallmark* del càncer) (Weaver & Cleveland, 2006; Douglas Hanahan & Weinberg, 2011). Per tant, es parla d'aneuploïdia quan el contingut cromosòmic de la cèl·lula no està balancejat en l'estat diploide característic del genoma humà. De fet, aquest tipus de desviació o aberració numèrica del contingut genòmic ja va ésser associat al càncer ja fa més de 100 anys (Hansemann, 1890; Holland & Cleveland, 2012).

La presentació d'aquest desequilibri pel que fa al contingut genòmic – l'estat aneuploide– de les cèl·lules canceroses és, en part, una conseqüència directa de la inestabilitat cromosòmica, o CIN (Tanaka & Hirota, 2016). Les cèl·lules tumorals adquireixen la CIN durant el procés tumoral per tal d'assolir les alteracions genòmiques que conferiran avantatge selectiu i, així, afavorir el procés de carcinogènesis (Sansregret, Vanhaesebroeck, & Swanton, 2018). És a dir, la anormal adquisició biològica de CNAs, com a conseqüència de l'increment del rati en errades durant la segregació cromosòmica mitòtica, condicionarà l'adquisició d'aneuploïdia en les cèl·lules del càncer. Però, tot i aquesta aparent interacció directa, la relació entre la CIN i l'aneuploïdia no sempre és de forma simple. De fet, en certs casos, l'assoliment i presentació de característiques d'aneuploïdia també pot provocar CIN, segurament com a conseqüència de l'alteració de funcions cel·lulars implicades en la mutagènesis genòmica, com els sistemes de reparació del DNA (Tanaka & Hirota, 2016).

3.1 Mecanismes de generació

3.1.1 Variants del número de còpia i alteracions focals

Els mecanismes de generació de les CNVs i les CNAs focals es poden diferenciar entre aquells associats amb esdeveniments recurrents, caracteritzats per presentar longituds d'alteració comunes en distints desordres genòmics associats i entre individus afectats, i els mecanismes associats a esdeveniments no-recurrents, els quals són altament específics dels individus que els presenten (C. M. B. Carvalho & Lupski, 2016).

Les CNVs o CNAs focals recurrents solen generar-se com a conseqüència de la recombinació homòloga no-al·lèlica (NAHR, de l'anglès, *nonallellic homologous recombination*) i gràcies a la presència de zones de repetició adjacents, mentre que els esdeveniments no-recurrents, caracteritzats per zones de ruptura amb presència de seqüències de micro-homologia i/o petites insercions de seqüències properes, s'associen a processos diferents, com la unió d'extrems no-homòlegs (NHEJ, de l'anglés *nonhomologous end-joining*), o els basats en la replicació del DNA, com la replicació induïda per ruptura (BIR, de

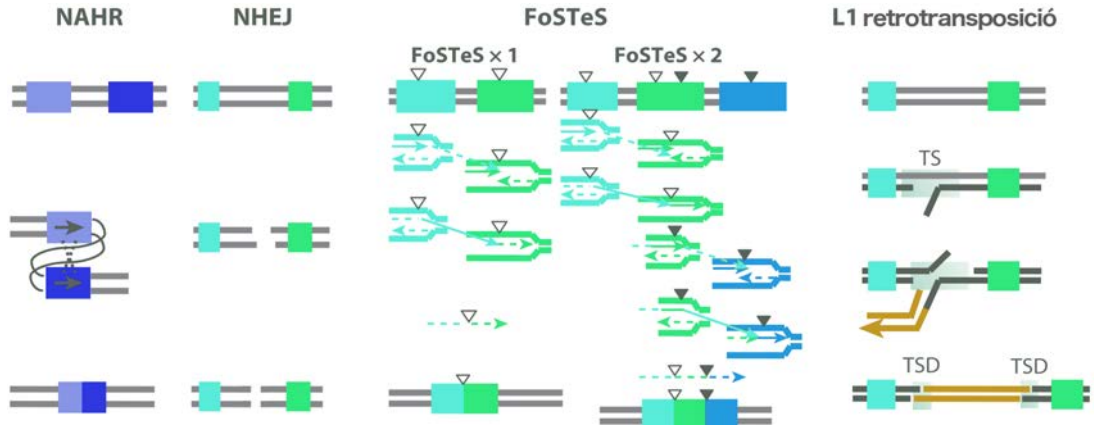


Figura 14. Mecanismes moleculars implicats en la generació de variacions del número de còpia.
Entre els processos responsables de la formació d'esdeveniments no-balancejats en regions localitzades del genoma hi trobem la recombinació homòloga no-al·lèlica (NAHR), la unió d'extrems no homòlegs (NHEJ), el sistema FoSTeS i alguns sitemes de retrotransposició. Els triangles indicats en el procés FoSTeS representen seqüències de micro-homologia. TS: *target site* (punt diana); TSD: *duplicated target site* (punt diana duplicat).(Extreta i adaptada de Zhang F et.al *Annu Rev Hum Genet* 2009)

l'anglès *break-induced replication*), la replicació induïda per ruptura degut a regions de micro-homologia (MMBIR, *microhomology-mediated break-induced replication*) o el sistema FoSTeS (de l'anglès, *fork stalling and template switch*). Alguns sistemes de retrotransposició també s'han implicat (Zhang, Gu, Hurles, & Lupski, 2009) .

Recombinació homòloga no-al·lèlica

El mecanisme del NAHR es pot donar tant en la mitosis com en la meiosis. Es produeix per l'alineament i posterior entrecreuament de dues regions no-al·lèliques amb alta similitud de seqüència (seqüències paràlogues) (Stankiewicz & Lupski, 2002). En seqüències paràlogues del mateix cromosoma, si aquestes presenten una orientació directa, s'afavorirà la formació de duplicacions o delecions, mentre que si les orientacions són inverses entre elles, llavors es poden donar inversions genòmiques (**Figura 14**). Esdeveniments de translocació poden donar-se quan les seqüències alineades no es troben en el mateix cromosoma.

Les seqüències proclius a ser substrats del NAHR són regions de baix contingut en repeticions o duplicacions segmentals majors a 10 Kb i amb graus de similitud d'un 95-97% (Stankiewicz & Lupski, 2002).

Unió d'extrems no-homòlegs

Les ruptures de doble cadena del DNA poden sorgir com a conseqüència de processos de recombinació, radiacions ionitzants o per l'efecte de les espècies reactives de l'oxigen. La seva reparació passa pel mecanisme NHEJ que es caracteritza per el fet de no necessitar substrats d'homologia, a diferència de la NAHR, i pel fet de que, acabat el procés de reparació, és comú que en la regió reparada resti alguna "cicatriu" en forma de pèrdua de material genòmic o, per altra banda, d'addicció d'alguns nucleòtids (**Figura 14**) (Lieber, 2008).

FoSTeS

En els últims anys, la identificació de re-ordenaments genòmics complexes va fer que Lee i col·laboradors proposessin un nou sistema molecular que podria estar implicat també en la generació d'alteracions de tipus estructural: el FoSTes (Zhang, Khajavi, et al., 2009).

Aquest sistema es basa en que, durant la replicació del DNA, la forquilla de replicació pot quedar-se estancada. Això, sumat a possibles micro-homologies entre una forquilla de replicació i una altra, pot afavorir que la replicació del material genètic passi de la primera forquilla a la segona, generant re-ordenaments genòmics (**Figura 14**). A més, aquestes transicions

entre forquilles poden ésser múltiples, generant alteracions més complexes (J. A. Lee, Carvalho, & Lupski, 2007).

Retrotransposició L1

Els elements d'intercalat amplis 1 (L1) cobreixen fins el 16,98% del genoma humà i són els únics transposons autònoms actius actualment (Ostertag & Kazazian Jr, 2001; Goodier & Kazazian, 2008). La transposició d'aquests elements L1 potencialment ocorren mitjançant la mediació d'una molècula de RNA intermediari, probablement transcrita per la RNA polimerasa II, generant la formació d'alteracions estructurals (**Figura 14**) (Kazazian & Moran, 1998).

Replicació induïda per ruptura

Els mecanismes de replicació induïda com els BIR o el MMBIR són processos de recombinació homòloga (HR, de l'anglès *homologous recombination*) encarregats de reparar ruptures de doble-cadena del DNA (Malkova, Ivanov, & Haber, 1996; Ira & Haber, 2002). Es tracten de processos complexos que poden activar-se degut a l'estancament de les forquilles de replicació i s'han relacionat amb la generació d'alteracions genòmiques com les duplicacions segmentals. El MMBIR utilitza seqüències de micro-homologia properes a la zona de ruptura per a iniciar el procés de reparació, mentre que el BIR necessita de regions d'homologia més àmplies per a la seva funcionalitat (C. E. Smith, Llorente, & Symington, 2007; Hastings, Ira, & Lupski, 2009).

3.1.2 Aneuploïdia

Durant el procés de divisió cel·lular, el material genòmic es duplica per a, posteriorment, repartir-se entre les cèl·lules filles. L'aneuploïdia resulta d'errors durant la mitosis, on es dona la segregació de cromosomes entre els nuclis filials (Holland & Cleveland, 2012). Els principals mecanismes que desemboquen en l'estat d'aneuploïdia es descriuen a continuació

Defectes als punts de control de la mitosi

La via de senyalització responsable del control del procés de mitosi és una complexa xarxa de proteïnes que s'encarrega d'assegurar la correcte unió entre els cinetocors i els microtúbuls al fus mitòtic, en la transició de metafase i anafase. L'alteració d'aquest sistema afavoreix

que l'anafase continuï tot i que els cinetocors no es trobin ben ancorats al fus, acabant les dues còpies del cromosoma al mateix nucli filial, provocant l'aneuploïdia (**Figura 15A**) (Rieder, Cole, Khodjakov, & Sluder, 1995; Kops, Foltz, & Cleveland, 2004).

Defectes en la cohesió entre cromàtides

En cas de que ambdues cromàtides germanes, portant el mateix material genètic, es mantinguin fixades entre elles, la segregació de d'aquestes no s'efectuarà correctament i acabaran ambdues en un dels dos nuclis filials. De la mateixa manera, si la cohesió entre cromàtides germanes es perd prematurament, ambdues podrien acabar en el mateix nucli filial (**Figura 15B**) (Barber et al., 2008).

Amplificació de centrosomes

La formació de múltiples pols de segregació ve condicionada per la presència d'un major nombre de centrosomes. En aquests casos pot augmentar la formació d'unions merotèliques, on un cinetocor està simultàniament unit a microtúbuls que provenen de distints centrosomes, truncant el repartiment correcte de les cromàtides germanes (**Figura 15C**) (Pihan et al., 1998; Nigg, 2006).

Estabilitat de la interacció entre cinetocors i microtúbuls

Quan les unions entre els cinetocors i els microtúbuls són incorrectes, ja sigui perquè la unió dels microtúbuls de distints pols es realitza al mateix cinetocor o perquè existeixen múltiples microtúbuls d'un mateix pol enganxats a un cinetocor, existeixen mecanismes de control que espontàniament trunquen la connexió errònia. En les cèl·lules que pateixen la CIN, aquests mecanismes es troben desregulats i, per tant, la correcció de la interacció errònia entre els cinetocors i els microtúbuls no és possible. Això afavoreix la formació d'unions merotèliques, facilitant l'aparició d'aneuploïdia (**Figura 15D**).

Tetraploïdia

Les cèl·lules tetraploides tenen el doble de material genòmic que presentaria una cèl·lula normal diploide (Shackney et al., 1989). La tetraploïdia pot venir donada per errades en la citocinesis, fusió cel·lular, la concatenació de dues rondes de replicació del DNA o per la falta d'activitat de telomerasa (Davoli, Denchi, & de Lange, 2010). Aquesta situació, on la quantitat de material cromosòmic es troba duplicat en una mateixa cèl·lula,

s'ha postulat com un estadi de transició entre un fenotip cel·lular de CIN i l'adquisició de l'aneuploïdia (Olaharski et al., 2006; Davoli & de Lange, 2011).

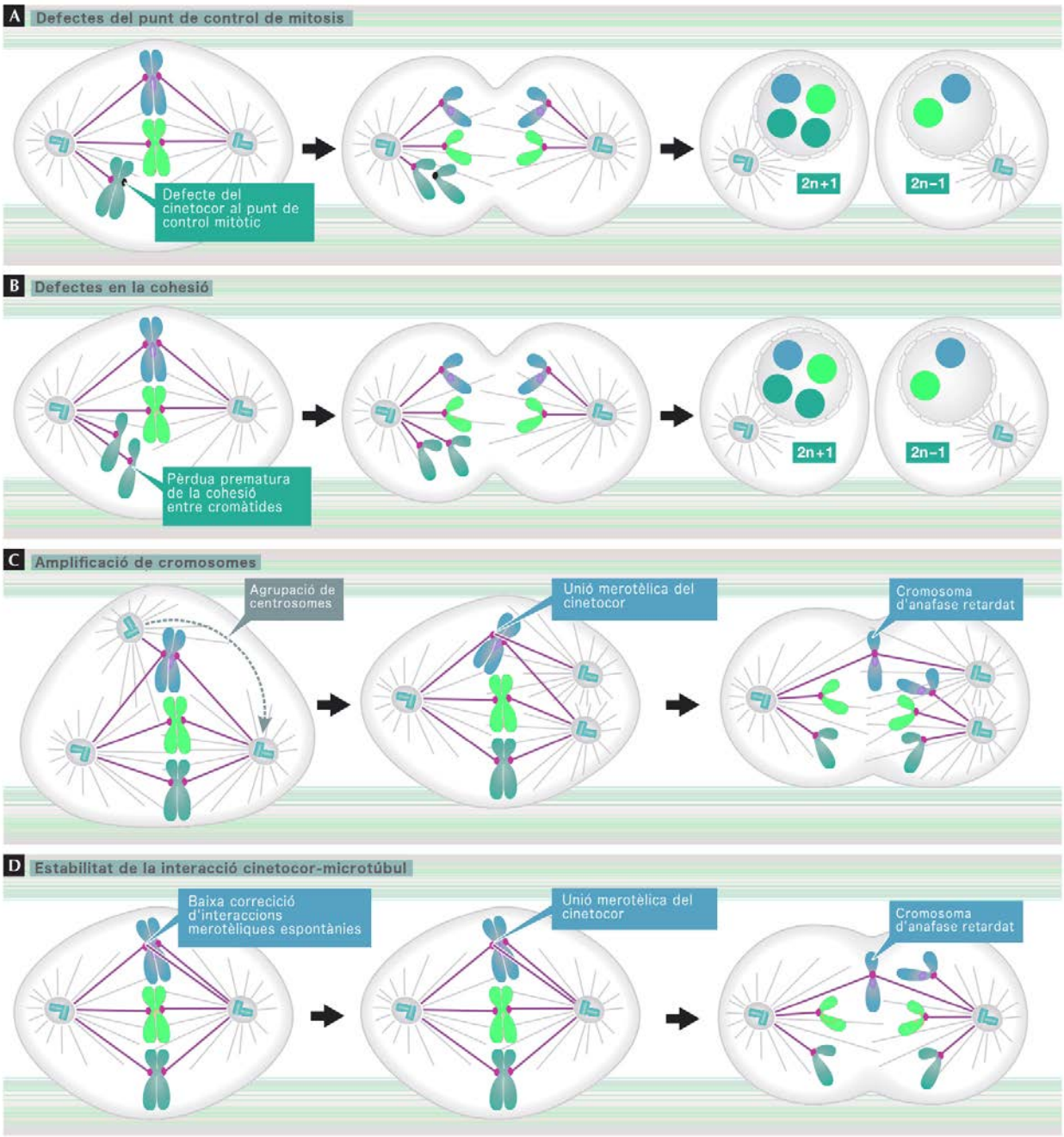


Figura 15. Camins a l'aneuploïdia.
Existeixen distints mecanismes per a l'aparició dels estats d'aneuploïdia en les cèl·lules canceroses: defectes al punts de control de la mitosi, defectes en la cohesió de les cromàtides, amplificació de centrosomes i alteracions en l'estabilitat de la interacció entre cinetocors i microtúbuls. (Extreta i adaptada de Holland and Cleveland et.al *EMBO* 2012)

3.2 Detecció de variants i alteracions del número de còpia

Les primeres imatges dels cromosomes humans es remunten al segle XIX, tot i que fins al 1956 no es va establir el correcte nombre de cromosomes: 22 parelles de cromosomes autosòmics i una parella de cromosomes sexuals (Arnold, 1879; Tjio & Levan, 2010).

El desenvolupament de les tècniques d'observació del material genètic ha anat lligat a la necessitat d'observar el comportament del material genòmic en relació a situacions patològiques de l'ésser humà. Les anomenades tècniques citogenètiques, encaminades a l'observació de grans alteracions genòmiques i cromosòmiques, han donat pas a les eines de biologia molecular d'alta resolució, les quals han permès la identificació de re-ordenaments genòmics més focals (Speicher & Carter, 2005).

Les principals tècniques genòmiques i moleculars per a l'observació de les alteracions estructurals del genoma s'expliquen a continuació.

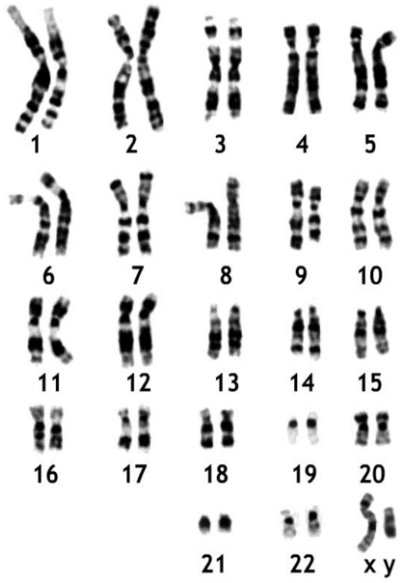


Figura 16. Cariotip complet d'un perfil genòmic masculí normal.
L'observació dels cariotips cromosòmics es basen en la tinció de les bandes G als cromosomes en fase de mitòtica, quan es condensen per a formar les cromàtides.

3.2.1 Citogenètiques

Cariotips

La visualització dels cariotips, és a dir, els perfils genòmics que representen les parelles de cromosomes organitzades segons els seu tamany i basada en la tinció de bandes G cromosòmiques, s'ha vingut utilitzant per a la observació d'aneuploïdies cromosòmiques des de fa molts anys (Caspersson, Zech, & Johansson, 1970). Les 22 parelles de cromosomes autosòmics més els cromosomes sexuals s'observen a partir del material genètic d'una cèl·lula única durant la fase mitòtica, quan els cromosomes es troben condensats formant les cromàtides (**Figura 16**).

Hibridació fluorescent *in situ*

L'aparició de la hibridació *in situ* fluorescent (FISH, de l'anglès *fluorescence in situ hybridization*) a la dècada dels anys 80 representà un gran avançament per al camp de la detecció i estudi de les alteracions cromosòmiques (Bauman, Wiegant, Borst, & van Duijn, 1980). La FISH es basa, de la mateixa manera que ho fan altres tipus de tècniques d'hibridació, en el ús d'una sonda nucleotídica complementària a la regió del DNA que es sotmet a estudi. Aquesta sonda acostuma a estar marcada amb una molècula fluorescent, permetent que els resultats de la hibridació s'avaluïn sota el microscopi de fluorescència (Pinkel et al., 1988).

Amplificació de sondes depenent de lligació

Per a la detecció de regions genòmiques amplificades o delecionades d'una forma més dirigida i regional s'ha vingut utilitzant la tècnica d'amplificació de sondes depenent de lligació, o MLPA (de l'anglès *multiplex ligation-dependent probe amplification*) (Schouten, 2002; Schouten et al., 2002). Aquesta tècnica es caracteritza per la lligació de dos fragments de DNA específics mitjançant primers als extrems que, quan s'uneixen, formen una sonda d'amplificació que produirà la senyal de quantificació. En la situació en què la regió genòmica presenti variació del número de còpia, la senyal quantificada variarà respecte a la mostra control diploide.

3.2.2 Matrius genòmiques

Matriu d'hibridació genòmica comparada

Als anys 90, el desenvolupament de la tècnica d'hibridació genòmica comparada (CGH, de l'anglès *comparative genòmic hibridization*) va revolucionar l'estudi del càncer (Kallioniemi et al., 1992). El principi bàsic d'aquesta es base en la comparació de dues mostres genòmiques, una control i una problema. El rati de quantificació de la senyal de fluorescència en les regions genòmiques estudiades entre les dues mostres determina el número de còpia de la mostra problema. El marcatge de les regions es realitza mitjançant molècules fluorescents per a facilitar la seva quantificació. La CGH va permetre el mapatge d'alteracions genòmiques no-balancejades en tumors sòlids fins a nivells sense precedents, aportant evidències que aquestes contribuïen enormement en la progressió tumoral (Ried et al., 1996).

En l'actualitat, la tècnica ha evolucionat i s'utilitzen microarrays de sondes d'oligonucleòtids distribuïdes a tot el genoma per tal d'estudiar els perfils genòmics complets, amb el que es coneix com matriu de CGH o aCGH (de l'anglès *array CGH*) (**Figura 17**) (B Carvalho, Ouwerkerk, Meijer, & Ylstra, 2004; Feuk, Carson, & Scherer, 2006).

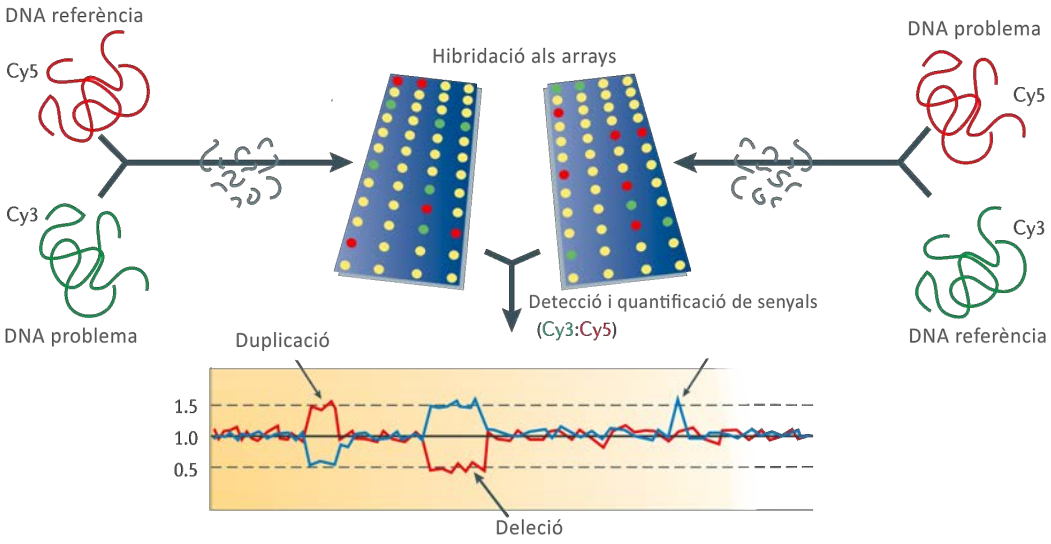


Figura 17. Esquema de la metodologia del aCGH.
Les mostres genòmiques de referencia i problema es marquen mitjançant els fluorocroms (Cy5 i Cy3). Posteriorment, aquestes senyals es quantifiquen, ens normalitzen i es calculen els ratis entre ambdues mostres. Les desviacions entre elles identificaran variacions del número de còpia en la mostra problema. (Extreta i adaptada de Feuk et.al *Nature Reviews Genetics* 2006)

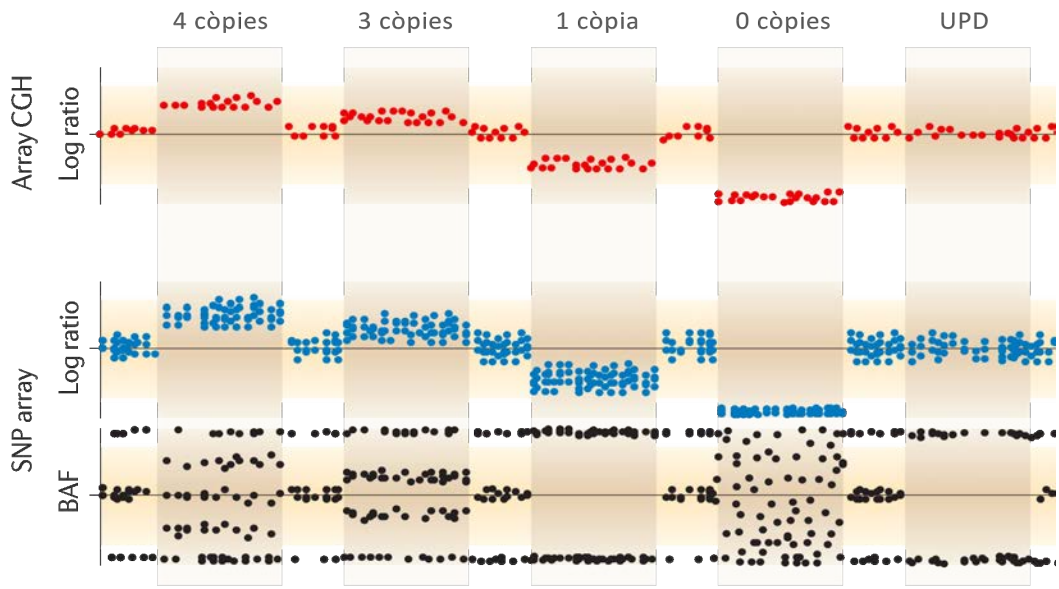


Figura 18. Representació de les dades de aCGH i SNP array.
La tècnica aCGH aporta els valors de log-ratio després de la normalització dels ratis de les senyals de les mostres genòmiques comparades. Per altra banda, la plataforma SNParray aporta la mateixa informació i, de forma paral·lela, també dona informació del BAF (*B-allele frequency*). Aquesta pot ésser utilitzada per a calcular regions amb LOH (*loss of heterozygosity*) i inferir la disomia uniparental (UPD). (Adaptació de Alkan et.al *Nature Reviews Genetics* 2011).

Matriu de SNPs

Les variacions més comunes del genoma humà, i que aporten variabilitat genètica entre individus i poblacions humanes, són els polimorfismes de nucleòtids simples, coneguts com SNPs (de l'anglès, *single nucleotide polymorphisms*) (Brookes, 1999).

En el camp de la investigació de la variabilitat del genoma humà, els SNPs s'han utilitzat per al descobriment i la caracterització d'aquesta variabilitat mitjançant els estudis d'associació del genoma complet (GWAS, de l'anglès *genome-wide association studies*), on aquestes unitats de variació són comparades entre conjunts d'individus controls i conjunts d'individus associats a una malaltia específica. Per altra banda, i més extensivament en la última dècada, la caracterització dels perfils genòmics de CNAs es duu a terme mitjançant les matrius de sondes d'SNPs (SNP *arrays*). Aquestes matriu, a banda de permetre la identificació de CNVs i aneuploïdia, també faciliten l'estudi dels ratis entre els diferents al·lels –o freqüència de l'al·lel B (BAF, en anglès *B-allele frequency*) (**Figura 18**) (Alkan, Coe, & Eichler, 2011). Mitjançant l'anàlisi dels valors de BAF en les regions genòmiques és possible la

identificació de regions amb pèrdua d'heterozigositat (LOH, de l'anglès *loss of heterozygozity*), però sense canvis en el número de còpia (CN-LOH, de l'anglès *copy neutral LOH*), el que s'anomenen regions amb disomia uniparental (UPD, de l'anglès *uniparental disomy*) (Peiffer et al., 2006; González et al., 2011).

3.2.3 La seqüenciació de nova generació

Des del descobriment de l'estructura de la doble hèlix de DNA per part de Watson i Crick (Watson & Crick, 1953), ja fa més de seixanta anys, i passant per l'assoliment i finalització del projecte del genoma humà, al 2003 (Human Genome Sequencing Consortium, 2004), s'han presentat grans i extraordinaris avenços en l'estudi del genoma i les seves patologies, destapant la extraordinària complexitat en què es presenta l'arquitectura genòmica. Els grans avenços en el desenvolupament tecnològic aplicat a la seqüenciació del material genètic han permès la reducció dels costos en aquesta seqüenciació, fent que la diversitat en les tecnologies aplicades en aquest camp s'expandeixi.

Una de les fites més importants en aquest desenvolupament tecnològic ha estat la capacitat de seqüenciació massiva de nova generació mitjançant la tecnologia de "seqüenciació de nova generació", en anglès, *next generation sequencing* (NGS). Durant l'última dècada, la tecnologia NGS ha anat evolucionant moltíssim, millorant la seva capacitat entre 100 i 1000 vegades, i incorporant innovacions revolucionàries que faciliten l'estudi de la ja esmentada complexitat genòmica. La carrera tecnològica i les millores assolides han disminuït els costos de seqüenciació del genoma al voltant dels 1000 dòlars (Check Hayden, 2014), i fins i tot s'han plantejat com eines actuals i futures del diagnòstic clínic (Vandrovcova et al., 2013; Hehir-Kwa, Pfundt, & Veltman, 2015; Pfundt et al., 2017).

La particularitat d'aquesta tecnologia és la generació de seqüències de lectura o *reads*, normalment més curtes de les generades per tecnologies clàssiques com la seqüenciació per Sanger. Tot i que la tecnologia és molt fiable, els ratis d'error en la seqüencia resultant són relativament elevats (aproximadament entre un 0,1 i un 15%, sobretot en *reads* curts). Així doncs, la tecnologia NGS pot ésser dividida en dos tipus: la NGS de *reads* curts i la de *reads* llargs. La primera és la més àmpliament utilitzada degut al seu baix cost i menor temps de generació. Per altra banda, la generació de *reads* llargs té l'avantatge d'augmentar la fiabilitat a l'hora d'identificar regions que presenten alteracions o esdeveniments estructurals grans, ja que la possibilitat de que el *read* generat aquí cobreixi tota la regió conflictiva és més alta (Cirulli & Goldstein, 2010; Chaisson, Huddleston, et al., 2015).

En la tecnologia NGS de *reads* curts hi trobem les categories de seqüenciació per lligació (SBL, de l'anglès *sequencing by ligation*) i la de seqüenciació per síntesis (SBS, de l'anglès *sequencing by synthesis*). En la SBL, una seqüència sonda unida a un fluoroform s'hibrida a un fragment de DNA i es lliga a un oligonucleòtid adjacent per a captar la senyal (Goodwin, McPherson, & McCombie, 2016). L'espectre d'emissió d'aquesta senyal indica quina base nucleotídica ha estat incorporada durant la seqüenciació del fragment. Per altra banda, en la SBS, s'utilitza una polimerasa i una senyal, ja sigui un fluoroform o un canvi en la concentració iònica, determina el nucleòtid incorporat. Tant en una com en l'altra tecnologia, els fragments de DNA s'amplifiquen clonalment en una superfície sòlida per eliminar soroll tecnològic, mentre que la paral·lelització massiva de milions de processos SBL o SBS també es veu facilitada per aquesta amplificació clonal. La plataforma de seqüenciació és capaç de captar els milions de senyals instantànies, seqüenciant milions de molècules de DNA al mateix temps (Shendure & Ji, 2008).

Existeixen distintes aproximacions per a la generació de fragments clonals a partir de la mostra de DNA. Una de les més conegudes és la generació de fragments per fase sòlid (en anglès *solid-phase*), utilitzada en la tecnologia de l'empresa Illumina. En la primera fase es fragmenta el DNA, i en la segona, en el cas de l'alternativa *solid-phase*, aquests fragments s'hibriden per complementarietat a multitud de primers –tant *forward* (5'→3') com *reverse* (3'→5')– lligats

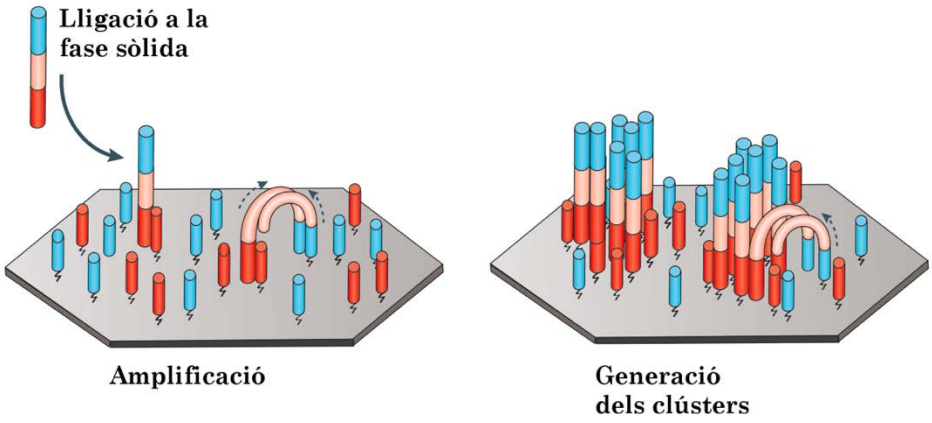


Figura 19. Seqüenciació NGS: lligació de primers a la fase sòlida i hibridació dels fragments de DNA.
Posteriorment a la fragmentació de la mostra de DNA a seqüenciar, els fragments s'hibriden a primers covalentment lligats a la superfície sòlida de la plataforma genòmica en qüestió per a la seva posterior amplificació clonal. (Adaptada de M. Metzker et.al *Nature Reviews* 2010)

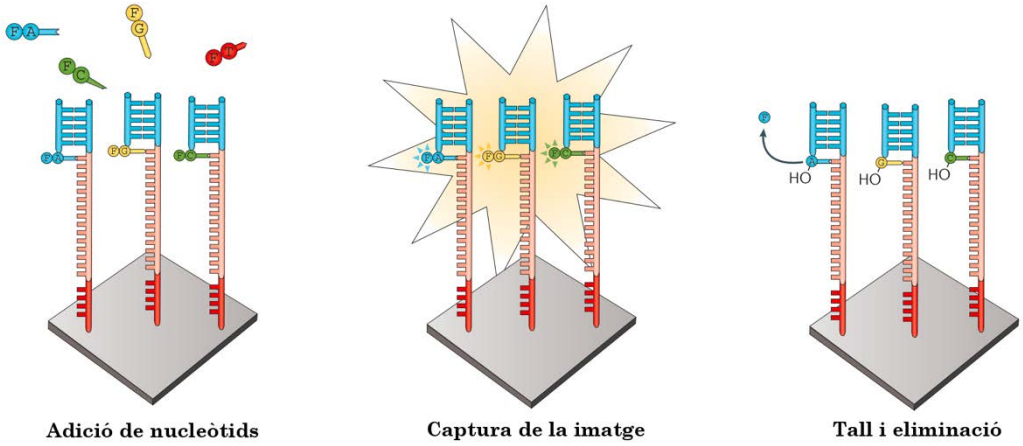


Figura 20. Seqüenciació NGS: incorporació de nucleòtids i lectura de la seqüenciació.
La incorporació de nucleòtids cíclica implica la identificació del nucleòtid incorporat en cada ronda d'addicció. El nucleòtid, marcat amb una molècula fluorescent, emetrà una senyal específica durant la captació de la imatge per a la seva identificació. Finalment, la molècula fluorescent serà tallada i eliminada de la seqüència. (Adaptada de Goodwin et.al *Nature Reviews Genetics* 2016)

covalentment en una superfície sòlida (**Figura 19**). Posteriorment a la lligació, s'inicia la incorporació de nucleòtids de manera cíclica (CTR, de l'anglès *cyclic reversible termination*). Els nucleòtids incorporats es troben modificats per emetre una senyal que els identifiqui durant cada incorporació mitjançant l'estimulació amb els làsers de la plataforma de captació. Prèviament a la presa de la imatge per a captar la senyal del nucleòtid incorporat, es neteja la superfície sòlida dels nucleòtids en suspensió no incorporats. Posteriorment s'efectua el tall de la regió final de la molècula del nucleòtid modificat, que conté el fluorescent, i es torna a netejar la superfície (**Figura 20**).

La seqüència final dels *read* extreta de les múltiples rondes de seqüenciació i imatge es processen de forma informàtica per a procedir als seu alineament a la versió més actualitzada del genoma de referència. Durant aquest processament informàtic es generen diversos passos de control de qualitat per assegurar que la seqüenciació dels *reads* a resultat satisfactòria i per eliminar aquelles seqüències errònies. Aprofitant l'aparició de les directrius respecte els arxius SAM, el format estàndard d'emmagatzemament de dades de seqüenciació NGS (H. Li et al., 2009), durant els últims anys s'ha assolit cert consens per tal d'estandarditzar el màxim possible el processament de les dades de seqüenciació de la informació genètica mitjançant les guies de bones pràctiques del *Genome Analysis Toolkit* (GATK) (McKenna et al., 2010). La gran quantitat de dades que es generen per cada mostra seqüenciada (aproximadament uns 7Gb d'informació per al WES i uns 100Gb per WGS) fa que l'anàlisi d'aquestes requereixi altes condicions computacionals. A més, els

programes informàtics que treballen amb aquest tipus d'informació s'han desenvolupat i millorat per a perfeccionar el processament al màxim.

3.2.3.1 Eines i algoritmes de detecció en dades NGS

Com s'ha esmentat abans, el consens per al maneig i gestió dels *reads* curts que resultaven de les tecnologies WGS i WES va arribar gràcies a grans esforços com el del GATK (McKenna et al., 2010). Contràriament, la comunitat científica ha vist com als últims anys s'ha generat una llarga llista de mètodes i eines per a la inferència de CNVs mitjançant aquestes dades de seqüenciació massiva. De fet, la consulta actual a la base de dades omicX (https://omictools.com, anteriorment coneguda com OMICtools), una reconeguda base de dades que cataloga aquest tipus de metodologies i eines per a la seva aplicació en l'anàlisi de dades de distints tipus d'òmiques, ens llista 69 eines per a WES i 172 per a WGS (només tenint en compte les categories CNV detection i Somatic CNA). Algunes de les eines més rellevants es troben llistades en la **Taula 3**.

Tot i que existeixen quatre tipus d'aproximacions a l'hora de manegar les dades de seqüenciació genòmica i detectar variants estructurals (*read pair*, *split read*, *read depth* i metodologia d'assemblatge), per a la inferència de CNVs (per tant, duplicacions o delecions), la metodologia més aplicada és la *read depth* o *read count*. Aquesta es basa en el supòsit de que la distribució dels *reads* mapejats al llarg del genoma és uniforme, per tant, el comptatge d'aquests en cada regió serà proporcional al número de còpia de la mateixa regió (Feuk et al., 2006; Yoon, Xuan, Makarov, Ye, & Sebat, 2009).

Taula 3. Eines i mètodes computacionals per la inferència de CNVs en dades de seqüenciació de nova generació.

Nom	NGS	Llenguatge	URL	Ref
CNVkit	WES/WGS	Python	https://cnvkit.readthedocs.io/en/stable/index.html	(Talevich, Shain, Botton, & Bastian, 2016)
ExomeDepth	WES	R	https://cran.r-project.org/web/packages/ExomeDepth/index.html	(Plagnol et al., 2012)
VarScan2	WES/WGS	Java	http://dkoboldt.github.io/varscan/	(Koboldt et al., 2012)
ControlFreeC	WES/WGS	C++	http://boevalab.com/FREEC/	(Boeva et al., 2011)
ExomeCNV	WES	R	https://cran.r-project.org/src/contrib/Archive/ExomeCNV/	(Sathirapongsasuti et al., 2011)
XHMM	WES	C++	https://atgu.mgh.harvard.edu/xhmm/index.shtml	(Fromer et al., 2012)
CoNIFER	WES	Python	http://conifer.sourceforge.net/index.html	(Krumm et al., 2012)
Delly	WGS	C++	https://tobiasrausch.com/delly/	(Rausch et al., 2012)
XCAVATOR	WGS	Perl, bash, R, Fortran	http://sourceforge.net/projects/xcavator/.	(Magi, Pippucci, & Sidore, 2017)
CNVnator	WGS	C++	http://sv.gersteinlab.org/cnvnator	(Abyzov, Urban, Snyder, & Gerstein, 2011)
CNV-seq	WGS	R, perl	http://tiger.dbs.nus.edu.sg/cnv-seq/	(Xie & Tammi, 2009)
Pindel	WGS	C++	http://gmt.genome.wustl.edu/packages/pindel/quick-start.html	(Ye, Schulz, Long, Apweiler, & Ning, 2009)
CONTRA	WES	Python/R	http://contra-cnv.sourceforge.net	(J. Li et al., 2012)

WES: whole exome sequencing; WGS: whole genome sequencing.

En molts casos la inferència de CNVs resultant necessita d'una re-segmentació, per tal d'ajuntar certs segments detectats que possiblement formin part d'esdeveniments majors i pròxims al mateixos valors de canvi. En aquest aspecte, l'algoritme estadístic més utilitzat és la segmentació binària circular (en anglès, *circular binary segmentation* o CBS) (Olshen, Venkatraman, Lucito, & Wigler, 2004), posteriorment implementat en format de paquet d'R: el DNAcopy (Venkatraman & Olshen, 2007).

Durant aquests últims anys, múltiples estudis han intentat realitzar comparatives entre les diferents eines de detecció de CNVs, per tal de destacar les distintes característiques de cada una d'elles, però sense arribar a conclusions factibles per tal de generar consens a l'hora de seleccionar unes poques que puguin utilitzar-se àmpliament (B. Liu et al., 2013; Alkodsi, Louhimo, & Hautaniemi, 2015; Kadalayil et al., 2015; Nam et al., 2016; Zare, Dow, Monteleone, Hosny, & Nabavi, 2017; Trost et al., 2018). De fet, encara hi ha espai per a la millora, tant en termes de tècniques de seqüenciació massiva, com aspectes de la pròpia inferència i detecció de CNVs i altres variants estructurals.

3.2.4 Eines de caracterització de variants i alteracions del número de còpia

L'estudi i anàlisi de les alteracions estructurals i la seva implicació i conseqüències, tant a nivell genòmic com a nivell funcional és prioritari i de màxima importància. Però, tot i que en l'apartat d'inferència i identificació de variants estructurals (incloent CNVs) i CNAs es presenta altament explorat, l'avaluació funcional i de les possibles conseqüències d'aquestes alteracions resta en mans d'especialistes amb competències i capacitats a nivell bioinformàtic i anàlisis de dades. L'alta demanda, tant en temps de dedicació com en habilitats informàtiques i de maneig de dades, suposa un impediment per aquells investigadors que no tenen aquestes capacitats a l'abast.

Algunes eines han sorgit de cara a la interpretació d'aquestes alteracions de número de còpia en aquest sentit. En són exemple les eines bioinformàtiques GISTIC2.0 (Mermel et al., 2011) o el ConVaQ (Larsen, do Canto, Rogatto, & Baumbach, 2018). La primera va encaminada a la identificació d'aquelles regions genòmiques amb alteració del número de còpia que es presenten de manera més recurrent entre les mostres que s'estudien, centrant-se en la mínima regió compartida en aquestes alteracions recurrents i per tal de destacar els gens implicats en elles. Aplicant diversos test de probabilitat i la

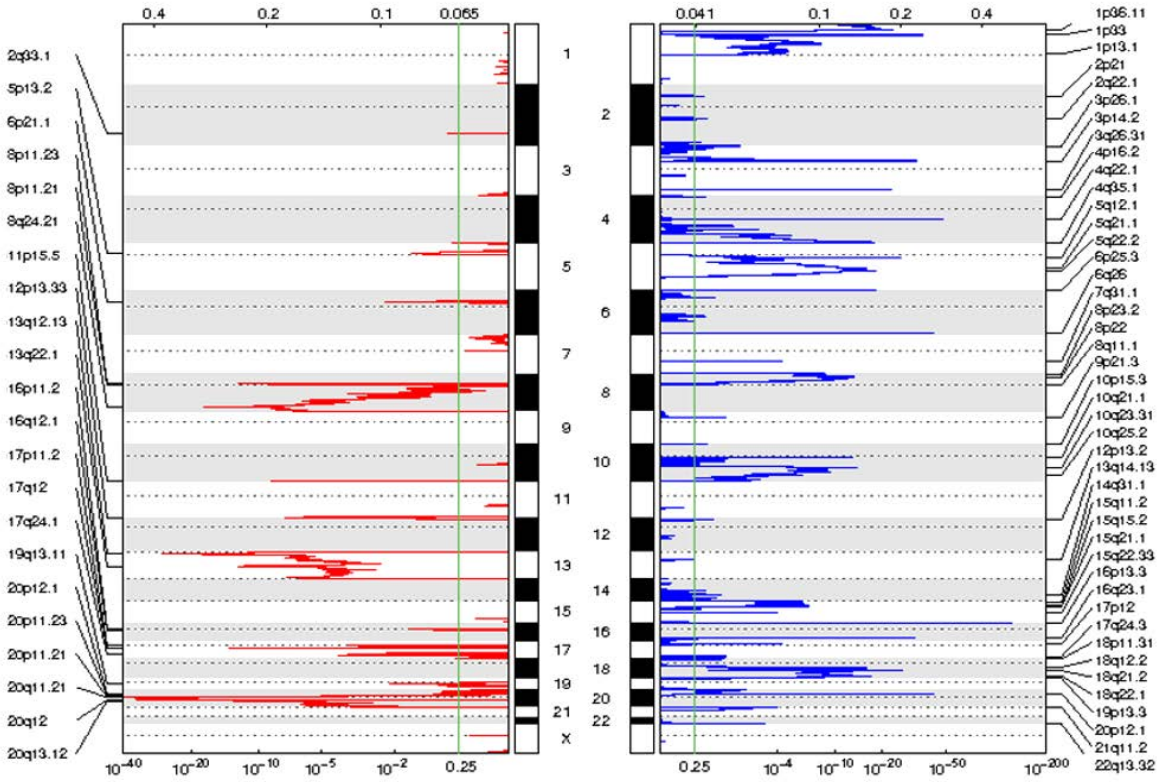


Figura 21. Perfil de CNAs detectat per l'eina GISTIC2.0 en la cohort COAD del TCGA.
Com exemple de perfils generats pel GISTIC2.0 es presenta el perfil produït per aquesta eina d'anàlisi de CNAs genòmiques a partir de les mostres de còlon pertanyents al projecte COAD del TCGA. En vermell els guanys o amplificacions i en blau les pèrdues o delecions (Extret de http//:firebrowse.org).

presentació de valors corregits per la multiplicitat de testos aplicats en la identificació d'aquestes regions recurrents, el programa GISTIC2.0 és capaç d'il·lustrar la significança estadística de les regions estudiades (valor que anomenen *Q-value*, i s'utilitza per a prioritzar les regions de CNAs més representatives de les dades), detectant els gens involucrats en les regions alterades de més recurrència i, per tant, enfocant l'estudi d'identificació de CNAs per tal de ressaltar els gens potencialment implicats en la condició patològica de les mostres (**Figura 21**). De fet, aquesta eina (i la seva versió anterior) (Beroukhim et al., 2007) ha estat la més utilitzada en els estudis de caracterització dels patrons d'alteració genòmica de les cohorts del TCGA. La gran majoria d'estudis que han donat a conèixer els perfils específics de CNAs dels distints tipus de càncer primaris han aplicat el GISTIC2.0, permeten la identificació dels gens implicats en la carcinogènesis d'aquestes neoplàsies (Weir et al., 2007; Network, 2008; Bass et al., 2009; Etemadmoghadam et al.,

2009; Northcott et al., 2009; The Cancer Genome Atlas, 2012; Ciriello et al., 2013; Nik-Zainal et al., 2016).

Per altra banda, l'eina bioinformàtica ConVaQ proporciona una simple entorn d'anàlisi per a la comparació de distints grups de mostres per als quals es té informació dels segments genòmics de les mostres (https://convaq.compbio.sdu.dk). L'eina compara els grups indicats exposant aquelles CNVs descriptives o diferencials entre grups, aportants certs valors estadístics als resultats (Larsen et al., 2018). A banda, l'aplicació també ofereix la possibilitat d'extreure les dades de freqüència d'aquelles regions més recurrents en els grups especificats.

Així doncs, amb aquests pocs exemples, es deixa entreveure l'existència de la necessitat poc explotada per al desenvolupament d'eines bioinformàtiques que facilitin l'estudi de les implicacions funcionals o descriptives associades a les variacions del número de còpia, i que vagin més enllà de la simple inferència o identificació d'aquestes alteracions genòmiques estructurals. Aplicacions d'aquest tipus poden ser claus a l'hora de portar a terme estudis que integrin les CNVs tenint en compte dades molecular i/o clíniques addicionals, per tal de poder descobrir noves implicacions funcionals o de diagnòstic d'aquests esdeveniments genòmics.

3.3 Variants del número de còpia de predisposició al CCR

Part de la herència genètica desconeguda sota les formes rares del CCR familiar pot venir provocada per les variants estructurals de canvi del número de còpia. Les CNVs aporten gran variabilitat genètica al genoma humà, i normalment s'han vist associades a baixos increments en la predisposició a certes patologies. Per exemple, la SNP *rs1944682*, localitzada en una regió de CNV comú -11q11-, es va associar amb un augment en la predisposició al CCR mitjançant estudis de GWAS de casos i controls, on s'avaluava la relació entre les variacions estructurals i el risc a desenvolupar la malaltia (Ceres Fernandez-Rozadilla et al., 2013). Per altra banda, diversos estudis han analitzat la implicació de CNVs en casos familiars de CCR amb diagnòstic primerenc, o associades a formes hereditàries polipòsiques, identificant gens candidats responsables de la predisposició al CCR (**Taula 4**).

Taula 4. Estudis d'identificació de CNVs en CCR familiar.

Estudi	Aproximació	Casos	Variants identificades
Ligtenberg M et.al Nature Genetics 2009	MLPA i aSNP	Famílies amb síndrome de Lynch	Delecions a regió 3' de *EPCAM* provocant metilació de *MSH2*
Venkatachalam R et.al *Gastroenterology* 2010	aSNP i validació per MLPA	CCR primerenc	Duplicació regió 5' del gen *PTPRJ*
Venkatachalam et.al *Int J Cancer* 2011	aSNP i validació per MLPA	41 casos de CCR familiar	Variants en gens *CDH18*, *GREM1* i *BCR*; i en zones promotores de *MIR491* i *MIR646*
Jaeger et.al. *Nature Genetics* 2012	aCGH i SNP-*array*	Família amb HMPS	Duplicació 40 Kb zona promotora *GREM1*
Fernandez- Rozadilla et.al *Clin Genet* 2013	aSNP	32 casos amb CCR primerenc	Deleció 7,32 Mb inclou 27 gens, entre ells *BMPR1A*
de Voer et.al *Gastroenterology* 2013	aSNP	72 casos amb CCR primerenc	Deleció 1,7 Mb al gen *BUB1*
Weren et.al *J Pathol.* 2015	MLPA i *target sequencing*	41 casos amb CCR primerenc	Deleció al gen *FOCAD*
Villacis et.al *Int J Cancer.* 2016	aCGH	45 casos de CCR familiar tipus X	35 variants rares, amb duplicació de *GALNT11* i *KMT2C*
Brea-Fernández et.al *Clin Transl Oncol.* 2017	aSNP i validació per MLPA	27 casos amb CCR primerenc	Delecions rares en *SLIT2* i *AK3*

WES: whole exome sequencing; aCGH: array-comparative genomic hybridization; aSNP: SNP array.

L'estudi de famílies holandeses i xineses amb síndrome de Lynch que presentaven híper-metilació del gen *MSH2*, implicat en la via de MMR, va identificar que el mecanisme de predisposició a la patologia venia donat per la deleció dels últims exons del gen EPCAM, situat en la regió 5' del gen *MSH2* (Ligtenberg et al., 2009). De forma similar, l'estudi de pacients de CCR primerenc que presentaven patrons d'herència recessiva es va associar la presència d'una duplicació de la regió 5' del gen *PTPRJ* amb la hipermetilació

de la regió promotora del mateix gen, fet que causava la seva infra-expressió gènica (Venkatachalam et al., 2010). En un altre estudi, en el que es genotipaven els SNPs de 41 pacients amb CCR familiar, s'identificaren CNVs rares als gens *CDH18*, *GREM1* i *BCR*, i en les zones promotores de *MIR491* i *MIR646* (Venkatachalam et al., 2011). Un any després, el paper del gen *GREM1* com a gen de predisposició al CCR fou corroborat en un estudi de lligament i validació per aCGH, en el que s'estudià una família amb forta agregació per a la patologia i s'identificà una duplicació de 40 Kb en la zona promotora del gen, provocant la seva sobre-expressió (Jaeger et al., 2012). De forma addicional, un altre estudi identificà diverses CNVs rares en individus amb CCR primerenc, entre elles una duplicació de 7,23 Mb que incloïa el gen *BMPR1A*, relacionant-lo amb la predisposició al CCR i el fenotip primerenc (C Fernandez-Rozadilla et al., 2013). El gen *BUB1* també es va relacionar amb el fenotip de predisposició al CCR com a conseqüència de la identificació d'una deleció de 1,7 Mb en la seva regió genètica, relació que fou posteriorment validada mitjançant estudis funcionals (de Voer et al., 2013). De manera similar, la deleció de la regió genòmica del gen *FOCAD* fou identificada en pacients classificats de CCR familiar (Weren, Venkatachalam, et al., 2015). A més, fins a 35 CNVs rares van ser identificades en pacients amb CCR familiar de tipus X, entre les quals s'observà una duplicació que afectava als gens *GALNT1* i *KMT2C* (Villacis et al., 2016). Recentment, estudis de seqüenciació de l'exoma de pacients amb CCR primerenc identificaren delecions germinals rares que afectaven els gens *SLIT2* i *AK3*, proposant la seva implicació en la susceptibilitat genètica al CCR (Brea-Fernandez et al., 2017).

3.4 Conseqüències i implicacions funcionals

Diversos estudis han demostrat la presència de CNAs específiques en distints tipus de tumors en forma d'alteracions de braços cromosòmics o cromosomes sencers (Beroukhim et al., 2010; Ried et al., 2012; Hoadley et al., 2018). Aquestes regions de CNAs condicionen patrons recurrents en els distints tipus de càncer i, de fet, estudis recents han comprovat que aquests s'agreguen depenent de l'origen tissular del càncer (**Figura 22**) (Beroukhim et al., 2010; Taylor et al., 2018). Per exemple, el CCR presenta els guanys específics dels cromosomes 7, 8q, 13 i 20q i les pèrdues de 8p, 17p i 18. Les hipòtesis que s'han originat com a conseqüència d'aquests patrons d'alteracions genòmiques característiques apunten a una desregulació gènica destacable i, de la mateixa manera, específica del tipus tumoral (Camps et al., 2009; Alaei-Mahabadi, Bhadury, Karlsson, Nilsson, & Larsson, 2016). Per tant, les CNAs estarien afectant directament els gens implicats en les regions genòmiques afectades per les alteracions estructurals. Per altra banda, també existeix la possibilitat de que s'estiguin generant mecanismes de desregulació de gens situats en altres indrets del genoma, allunyats de les CNAs (Franke et al., 2016; Spielmann et al., 2018).

3.4.1 Regulació de la dosis gènica

La patogenicitat de les variants estructurals, tant de CNVs com CNAs, tractant-se de variants estructurals "no-balancejades", s'interpreta damunt la base de que la dotació (la dosis) gènica es veurà alterada com a conseqüència de l'esdeveniment. Això és lògic, ja que gairebé tots els gens del genoma humà presenten dues còpies i són sensitius a la seva dosis (és a dir, la insuficiència o sobre-presentació del gen pot suposar efectes nocius per a la cèl·lula o l'organisme). De fet, es calcula que tan sols un 17% dels gens tenen alta probabilitat de presentar intolerància davant variacions en la seva dotació del número de còpies (Lek et al., 2016).

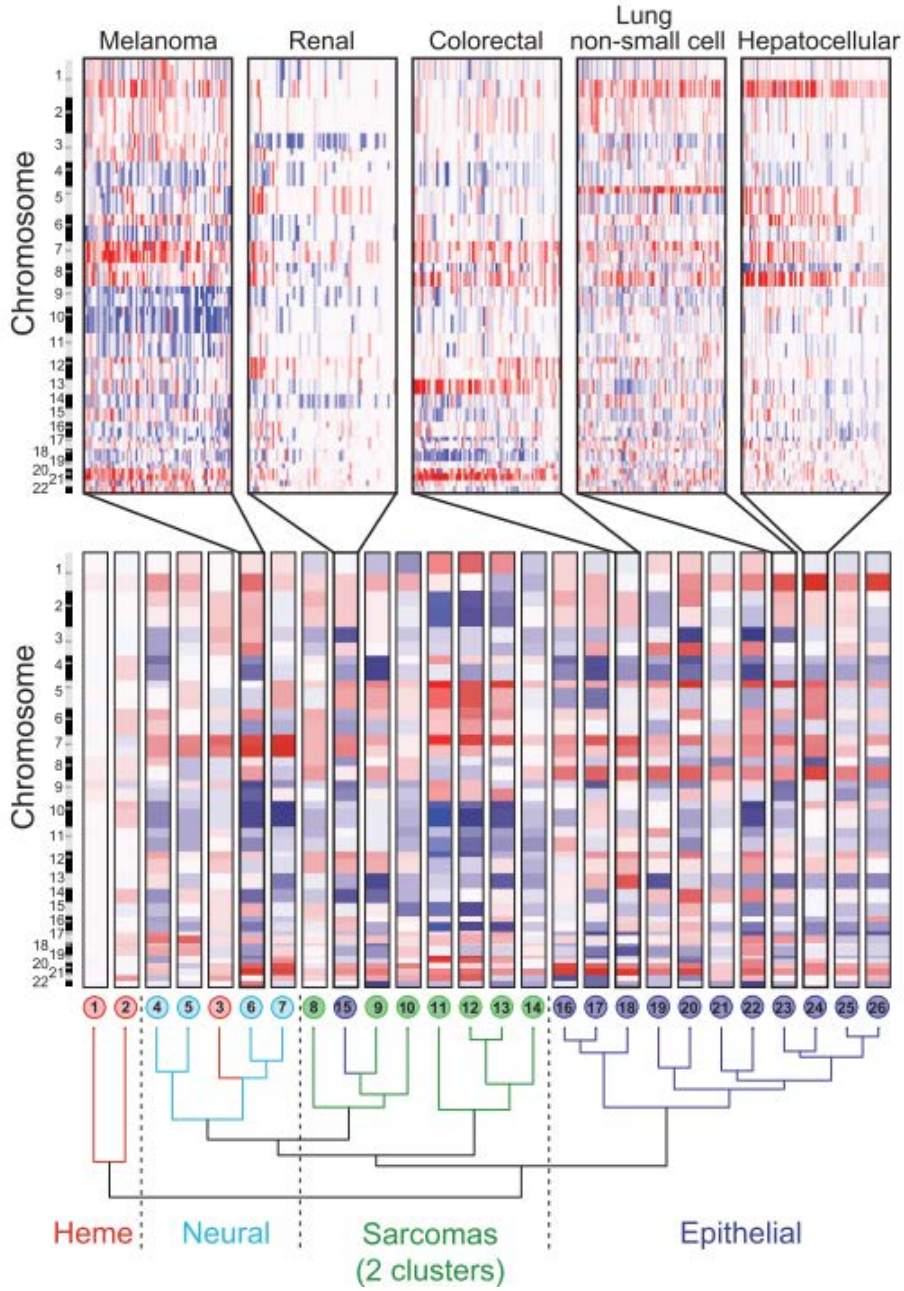


Figura 22. Patrons de CNAs dels braços cromosòmics específics al tipus de càncer i clusterització per tipus de teixit tumoral.
En la zona superior s'observa com els la caracterització de CNAs de les mostres tumorals deixen entreveure els distints patrons característics per als càncers de melanoma, renal, colorectal, de pulmó i hepàtic. En la zona inferior s'observa la clusterització dels patrons específics tumorals dels tipus de càncer segons el llinatge tissular (hematopoiètic, neural, sarcoma o epitelial) d'aquests. Les regions genòmiques delecionades es representen en blau, mentre que les amplificades ewsrepresenten en vermell. (Extreta de Beroukhim R et.al *Nature* 2010)

Estudis de fins a 15 línies cel·lulars distintes de CCR han demostrat una correlació positiva entre les CNAs validades per aCGH i els nivells d'expressió dels gens implicats (Camps et al GCC 2009). De fet, aquesta correlació s'ha confirmat en estudis amb tumors primaris, on l'alteració de l'expressió gènica es presentava en el sentit corresponent al que s'esperaria donat l'alteració implicada (sobre-expressió dels gens en alteracions en forma de guany i infra-expressió per a les pèrdues) (Grade et al., 2006, 2007; Camps et al., 2008). La seqüenciació completa d'aproximadament 600 genomes tumorals ha contribuït a la caracterització de les alteracions en l'expressió gènica com a conseqüència d'alteracions estructurals, entre elles les CNAs (Alaei-Mahabadi et al., 2016).

El fet d'avaluar la correlació entre els nivells d'expressió gènica i les CNAs ha contribuït al descobriment de nous gens relacionats amb processos carcinogènics. Per exemple, l'estudi del guany complet del cromosoma 13, una de les CNAs característiques del CCR, ha facilitat i confirmat l'associació de diversos gens a aquesta neoplàsia, entre ells *CDK8*, *CDX2* i *LNX2* (Firestein et al., 2008; Camps et al., 2013). Un altre exemple seria la regió cromosòmica 20q11-20q13, amplificada en una gran majoria dels tumors de CCR, on s'hi troben els gens *AHCY*, *POFUT1*, *RPN2*, *TH1L* i *PRPF6*, i que es presenten sobre-expressats i amb una clara correlació positiva entre la seva dosis gènica i els seus nivells d'expressió (Loo et al., 2013).

Per altra banda, tot i que aquesta relació entre el re-ordenament estructural que afecta a la dosis gènica en la regió i l'expressió gènica dels gens involucrats sembli lògica, no sempre es respecte. En són exemples alguns dels gens implicats en la regió del braç cromosòmic 20q, guanyat de forma recurrent al CCR (Ried et al., 1996; B Carvalho et al., 2009; The Cancer Genome Atlas, 2012; Ptashkin et al., 2017).

3.4.2 Remodelació espacial del genoma

Donada la gran quantitat de re-ordenaments que es generen en el genoma, és lògic pensar que aquests puguin tenir un efecte en la posició o la dotació gènica -és a dir, la còpia quantitativa- tant dels transcrits genètics com dels seus elements reguladors. Però, a més, si considerem en la remodelació de l'espai nuclear en la cèl·lula i la re-estructuració del material genètic dins aquest nucli, també seria crucial entendre els efectes que poden tenir aquests canvis espacials i els contactes entre el genoma que apareixen i desapareixen (**Figura 23**) (Spielmann et al., 2018).

Mitjançant el desenvolupament de les anomenades tècniques 3C, que són capaces d'interrogar els contactes físics que es donen en una regió específica

del genoma, l'estudi de les alteracions estructurals genòmiques ha entrat en una tercera capa de dificultat. Aquestes metodologies es recolzen en la quantificació de les freqüències en que dues regions interaccionen, independentment de la seva dimensió lineal. A més a més, la *HiC*, tècnica derivada de les 3C, és capaç d'interrogar tots els contactes genòmics a l'hora. D'aquesta manera es pot obtenir un mapa de dues dimensions que ens aporta les freqüències de contacte entre totes les regions del genoma.

Al contrari que les regions promotores, normalment situades a menys d'una kilobase de distància de la unitat gènica que han de regular, els anomenats *enhancers* són regions de regulació a llarga distància, que poden estar situats fins i tot a distàncies majors d'una megabase, i sense afectar altres gens que puguin estar més propers a nivell lineal (T. E. P. Consortium, 2012; Weischenfeldt, Symmons, Spitz, & Korbel, 2013). Aquests *enhancers* interaccionen amb les zones promotores mitjançant la formació de plegaments (en anglès, *loops*) genòmics que aproximen ambdues regions. Aquestes interaccions entre les regions promotores i els *enhancers*, com a conseqüència dels *loops*, afavoreixen una elevada freqüència de contactes entre les regions genòmiques que formen part del mateix plegament, d'aquí que es coneguin com a TADs (en anglès, *topologically associated domains*) (Rao et al., 2014). Els distints TADs del genoma estan separats i aïllats entre ells per unes regions aïllades de qualsevol contacte genòmic (mostren baixes freqüències de contacte fora de les seves pròpies regions), anomenades *boundary regions* (el que podria traduir-se al català com "regions de frontera"). Curiosament, una gran fracció dels TADs que s'han caracteritzat es mostren invariables entre distintes línies cel·lulars, suggerint que podrien tractar-se d'unitats estructurals fonamentals del genoma (Dixon et al., 2012).

El guany o la pèrdua de material genètic involucrat en les interaccions internes dels TADs simplement podran afectar a nivell de dosis gènica dels elements funcionals que es vegin implicats en l'alteració. Per altra banda, les alteracions genòmiques que involucren múltiples TADs tenen la capacitat de generar o destruir els contactes genòmics de les regions afectades, fins a tal punt de generar nous TADs, en el cas de duplicacions, o d'eliminar aquestes unitats de contacte i, potencialment, fusionar-les amb les més pròximes, en el cas de les delecions (**Figura 23**) (Spielmann et al., 2018).

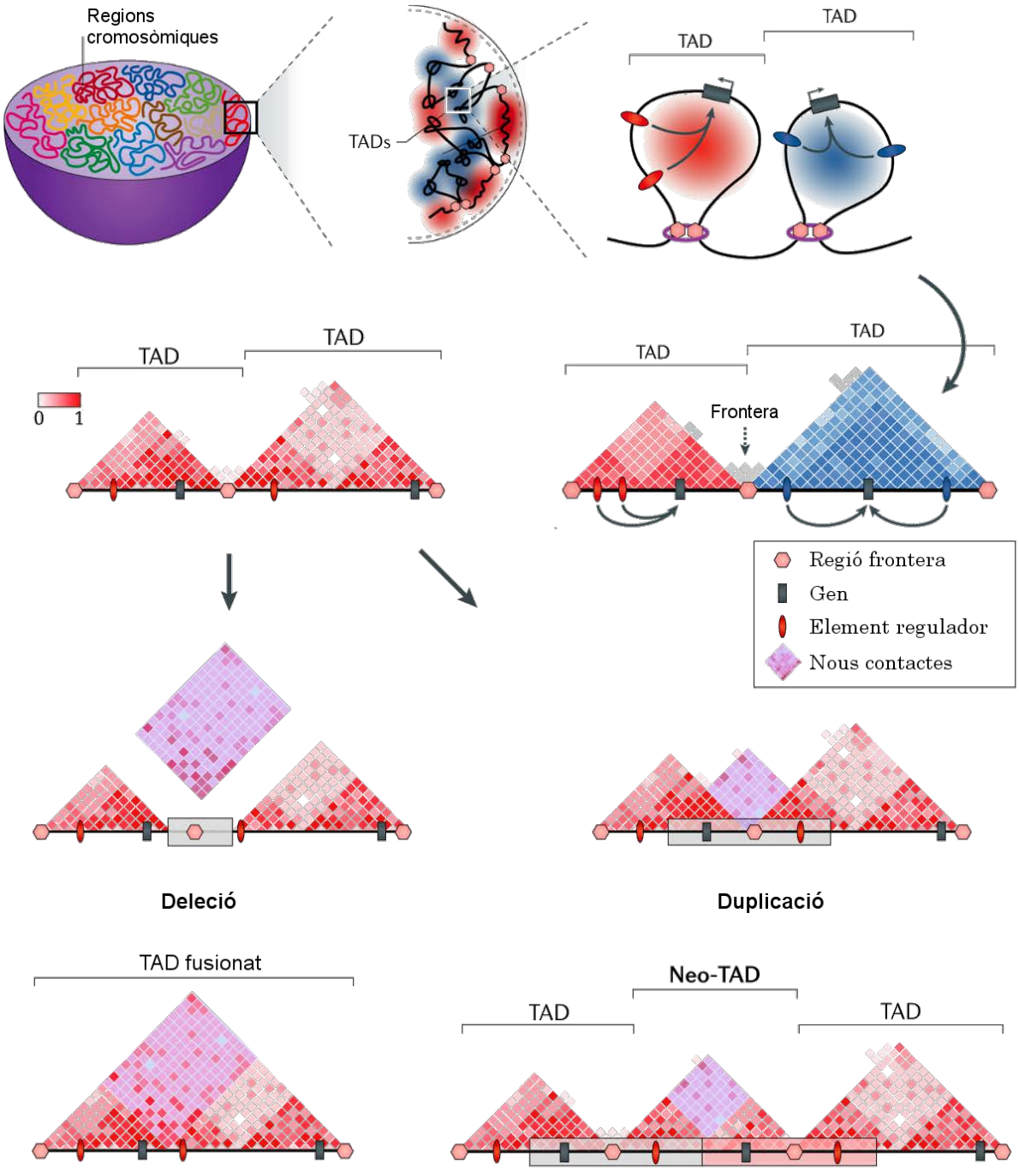


Figura 23. Organització tri-dimensional del genoma, formació i desregulació dels TADs.
Les CNVs poden condicionar l'arquitectura genòmica depenent de la implicació de les regions frontera dels TADs: si les delecions impliquen aquestes regions frontera es produeix la fusió dels TADs adjacents, mentre que si es tracta d'una duplicació, es formarà uan nova regió d'alta densitat de contactes, és a dir, un nou TAD (neo-TAD). (Extreta i adaptada de Spielman D et.al *Nature Review Genetics* 2018)

3.4.3 Signatures de les alteracions de número de còpia

En certa mesura, les mutacions genòmiques, tant puntuals com estructurals, funcionen com una forma de registres de les lesions no reparades que s'han anat acumulant al llarg del temps o, en el cas del càncer, durant l'evolució clonal del tumor. La capacitat d'associar aquests registres mutacionals amb processos exògens o endògens implicats en el dany al DNA, pot suposar un avantatge a l'hora d'identificar la causa o l'origen del desenvolupament tumoral.

En aquest sentit, el recent èxit dels estudis d'inferència de signatures mutacionals puntuals s'ha donat gràcies a la capacitat de quantificar els patrons de canvis trinucleotídics en les mostres tumorals i la seva associació amb característiques moleculars o clíniques de les mostres en qüestió, assolint la funció de bio-marcadors per a la identificació dels processos moleculars exògens o endògens que danyen el material genètic (**Figura 24**) (Alexandrov et al., 2013; Nesic, Wakefield, Kondrashova, Scott, & McNeish, 2018). Per altra banda, en l'estudi dels patrons de CNAs es troba a faltar aquesta multiplicitat genotípica, ja que només es reprodueixen patrons de presència (duplicacions, delecions, translocacions o inversions) o absència (genoma diploide). Davant això, s'estan desenvolupament diversos esforços en la comunitat científica per tal de considerar diferents aproximacions per a la inferència de signatures de CNAs, estudiant distintes característiques dels patrons generats per aquests tipus d'alteracions estructurals: des de la simple subclassificació i quantificació de les CNAs, fins a l'estudi de les "meta-característiques" dels perfils genòmics i la seva distribució estadística (Nik-Zainal et al., 2016; Macintyre et al., 2018).

Un dels primers estudis a l'hora d'inferir signatures mutacional de número de còpia quantificà 32 tipus de re-ordenaments genòmics en 560 mostres de càncer de mama que s'havien seqüenciat per WGS (Nik-Zainal et al., 2016). Aplicant la mateixa metodologia d'inferència utilitzada per a les signatures de mutacions puntuals (Alexandrov et al., 2013), es varen aconseguir identificar sis signatures distintes. Curiosament, tres d'aquestes sis signatures s'associaren amb deficiències en la HR, mentre que signatures enriquides amb delecions associades a regions repetitives i de microhomologies es relacionaren amb deficiència del sistema MMR i HR (Morganella et al., 2016).

Per altra banda, en l'estudi de Macintyre i col·laboradors, s'identificaren signatures del número de còpia analitzant dades de seqüenciació del genoma de 117 pacients amb càncer d'ovari. Algunes de les signatures es pogueren associar a processos moleculars com dany al

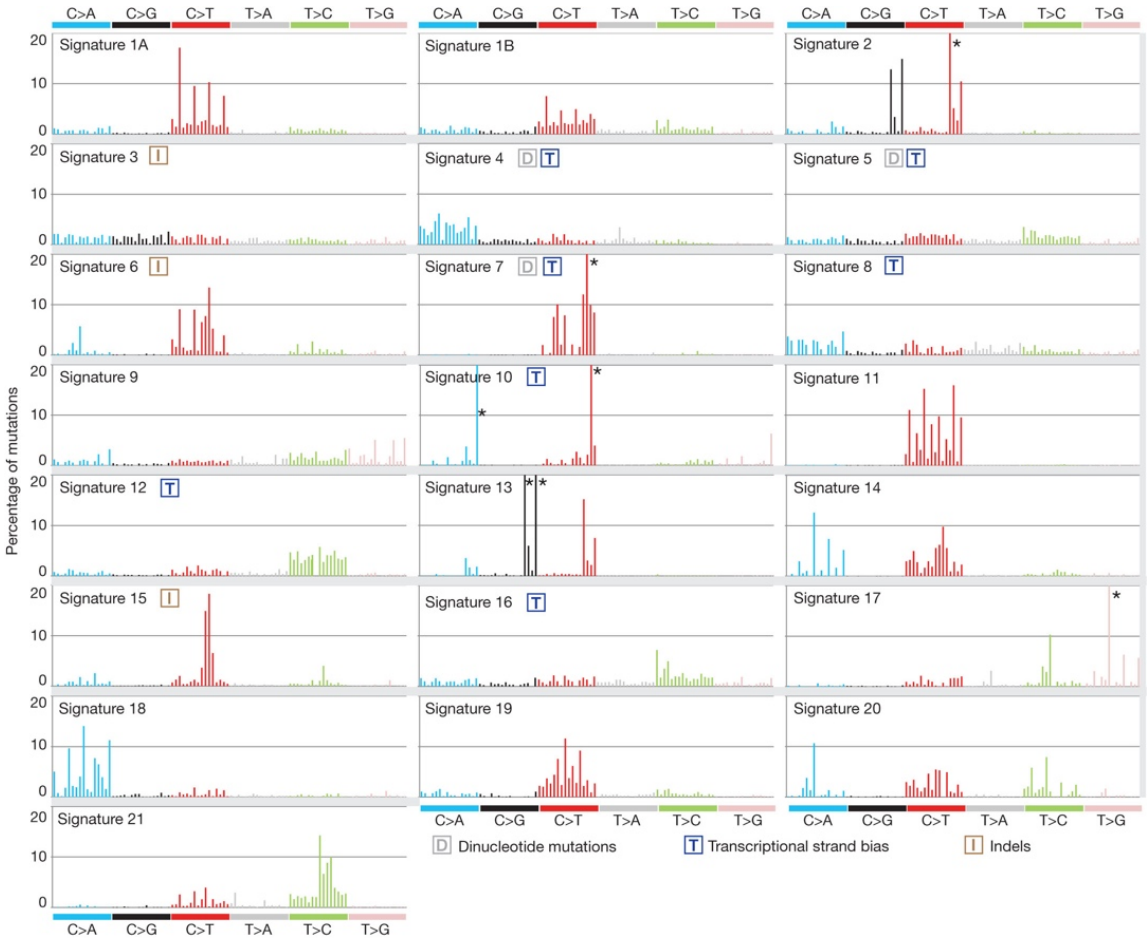


Figura 24. Signatures mutacionals de processos moleculars al genoma del càncer.
Representació gràfica dels percentatges d'aportacions per part de les mutacions puntuals en el context tri-nucleòtid que formen les signatures mutacionals associats a distints processos moleculars implicats amb el càncer. (Alexandrov et.al *Nature* 2013)

DNA, desregulació del cicle cel·lular o deficiència de HR. A més, s'identificà una correlació positiva en quant a la supervivència per a tres de les signatures, mentre que una altra de les signatures fou proposada com a biomarcador de pronòstic positiu i resposta ala immunoteràpia (Macintyre et al., 2018).

De moment, l'estudi de les signatures de CNAs és un camp amb molt marge d'exploració i amb una càrrega computacional important. Les noves tècniques de seqüenciació del genoma complet, amb *reads* més llargs, possiblement puguin facilitar la detecció i caracterització de les variants estructurals i, així, refinar la inferència de les signatures. Aquests patrons estructurals i mutacionals estant demostrant tenir potencial per erigir-se com a biomarcador de supervivència, diagnòstic molecular i resposta a distints tipus de tractaments.

3.4.4 Variants del número de còpia com a bio-marcadors

Als últims anys, l'estudi de dades genòmiques i la seva integració amb dades clíniques i moleculars ha sofert un gran impuls gràcies, sobretot, a la disponibilitat d'aquest tipus de dades en grans cohorts de mostres de càncer. Estudis com el TCGA, que contemplen els objectius de caracteritzar els diferents tipus de càncer a nivell molecular i genòmic, han generat gran quantitat de dades en les distintes òmiques (entre elles, la genòmica, i la transcriptòmica) i han obert les portes per a què els estudis d'integració d'aquestes múltiples capes d'informació es donin cada vegada més.

Una de les avantatges en quant a la realització d'aquests tipus de treballs d'integració es troba en la capacitat d'associar distintes característiques genòmiques o moleculars amb les dades clíniques dels pacients, com per exemple les dades de supervivència o els tractaments anti-cancerígens. Això afavoreix el descobriment de característiques amb capacitat predictiva i de prognosis. Per exemple, el treball de Davoli i col·laboradors va poder identificar l'associació de les CNAs focals amb marcadors de proliferació cel·lular en els distints tipus de càncers i, per contra, la relació de les alteracions més àmplies (*broad)* amb característiques d'evasió del sistema immune (Davoli et al., 2017). A més, l'estudi de la supervivència de distintes cohorts de melanoma, identificaren una major supervivència en aquelles mostres amb alts nivells de mutacions puntuals, en comparació a les mostres amb càrregues altes de CNAs somàtiques.

Els estudis de caracterització genòmica dels distints tipus de càncer han obert el camí per a la identificació de característiques moleculars específiques per les quals es puguin estratificar diferents subtipus de tumors i diferenciar tant la selecció i aplicació dels distints tractaments com la pròpia resposta a aquests. Entre aquestes característiques genòmiques s'hi troben les CNAs (Tang & Amon, 2013), com per exemple la deleció del gen *CDKN2A* en certs melanomes, que condiciona la resposta a inhibidors de ciclines dependents de quinases (Hodis et al., 2012), o l'amplificació del gen *AR* al càncer de pròstata, que senyala la baixa supervivència d'aquests casos i la progressió metastàtica (Grasso et al., 2012). Un altre exemple seria la trisomia del cromosoma 8, aneuploïdia recurrent a la leucèmia mieloide aguda, associada amb baixa supervivència quan coexisteix amb altres tipus d'aberracions cromosòmiques (Wolman, Gundacker, Appelbaum, Slovak, & Southwest Oncology Group, 2002). Així doncs, sembla ser que el paper funcional de les CNAs al càncer passa per la seva capacitat com a

biomarcadors de la progressió en formes d'alt risc i de la capacitat de resposta i/o resistència a tractaments (Singhal et al., 2007; Mekenkamp et al., 2012) o com a predictors de prognosis (Postma, Terwischa, Hermsen, van der Sijp, & Meijer, 2007; Kurashina et al., 2008). Un clar exemple de l'aspecte funcional de les CNAs al càncer és l'amplificació de la regió 9p24.1, la qual conté els gens *PDL1*, *PDL2* i *JAK2*, entre d'altres. La sobre-expressió del gen *PDL1* provocada per aquesta amplificació es correlaciona amb la resposta positiva als tractaments d'immunoteràpia anti-PD-1 en limfoma primari de les cèl·lules B del sistema nerviós central (Nayak et al., 2017). Per tant, la identificació d'aquesta CNA regional en aquest tipus de càncer pot treballar com un biomarcador a la resposta a aquest tractament (Keenan, Burke, & Van Allen, 2019).

Pel que fa al CCR, la seva caracterització dels perfils genòmics i les CNAs implicades durant el procés de carcinogènesis ha permès l'associació d'aquests esdeveniments amb els gens afectats pel canvi de número de còpia i la relació amb les vies moleculars implicades (Diep et al., 2006; The Cancer Genome Atlas, 2012; Beatriz Carvalho et al., 2018). De fet, algunes d'aquestes CNAs es relacionen directament amb teràpies anti-cancerígenes específiques (H. Wang, Liang, Fang, & Xu, 2016). Per exemple, al CCR, l'amplificació de la regió del braç 7p, que conté el gen *EGFR*, implicat en la carcinogènesis per inhibició de la mort cel·lular (Flora et al., 2012), està indicat pel tractament amb cetuximab o panitumubab, dues teràpies a partir d'anticossos monoclonals anti-EGFR (Cunningham et al., 2004; Keating, 2010). Per altra banda, alguns estudis recents han assenyalat que el 25% dels adenomes avançats presenten CNAs associades a la progressió del CCR, mentre que només el 2-4% dels adenomes no avançats també presenten aquests tipus d'alteracions no balancejades (Beatriz Carvalho et al., 2018). Això indicaria que les CNAs associades a la carcinogènesis podrien identificar formes d'adenoma d'alt risc i, per tant, aplicar-se com indicadors per al cribratge preventiu i programes de supervivència.

Hipòtesi i objectius

Hipòtesi general

Aproximadament, el 30-35% dels casos de CCR presenten agregació familiar per la malaltia sense que es conegui la seva causa germinal. Això fa que siguin necessaris estudis d'identificació de nous gens candidats i variants genètiques germinals responsables de la predisposició al CCR en aquests casos.

Gran part de la variabilitat genètica del genoma humà es produeix com a conseqüència de les variacions estructurals del genoma, entre aquestes les CNVs. Aquestes han estat implicades en diversos fenotips cancerígens, principalment a nivell somàtic, en forma de CNAs, però també a nivell germinal, aportant risc a desenvolupar distints tipus de neoplàsies. De fet, algunes CNVs estan implicades amb síndromes de predisposició al CCR, i es pensa que part de la genètica desconeguda als casos familiars d'aquesta neoplàsia podrien venir donats per variants d'aquest tipus.

Durant els últims anys, una de les aproximacions més explorades a l'hora d'estudiar els mecanismes genètics de predisposició al CCR familiar ha estat l'aplicació de plataformes genòmiques de seqüenciació massiva o NGS, com la WES. La gran majoria d'aquests treballs s'han centrat en la priorització i l'estudi de variants puntuals com a potencials mecanismes d'alt risc implicats en la predisposició al CCR, tot i que alguns també han realitzat esforços per identificar CNVs en casos de CCR familiar.

Per altra banda, la implicació de les CNAs en càncer s'ha estudiat des del punt de vista de la seva contribució en la inestabilitat genòmica que apareix durant la carcinogènesi i, com a tal, són considerades *hallmarks* del càncer.

La caracterització i integració d'aquest tipus d'informació genòmica amb altres anotacions moleculars o clíniques pot ajudar a la identificació de CNAs com a potencials bio-marcadors de subgrups de mostres, millorant el camp del diagnòstic, la prognosi o, fins i tot, en la selecció dels tractaments anti-cancerígens adequats segons els perfils genòmics.

Per tal de facilitar aquestes aproximacions són necessàries aplicacions i eines bioinformàtiques que facilitin les tasques de recerca i apropin aquests tipus d'estudis a la comunitat científica en general. La capacitat de conduir aquests estudis de forma automàtica i ràpida afavoriria els estudis d'integració genòmica, facilitant l'exploració d'aquest camp per un major nombre d'investigadors.

Objectiu general

L'objectiu general d'aquesta tesi inclou la identificació de nous gens candidats involucrats en la predisposició al CCR familiar mitjançant la inferència de CNVs en dades de seqüenciació de l'exoma germinal. Per altra banda, desenvolupar una aplicació bioinformàtica que faciliti la integració de perfils genòmics de CNAs amb variables moleculars i/o clíniques d'anotació, per tal d'assenyalar potencials implicacions funcionals d'aquestes.

Objectius específics:

Identificació de nous gens candidats al CCR familiar

1. Identificació i priorització de CNVs rares utilitzant les dades de la seqüenciació completa de l'exoma del DNA germinal d'individus afectats per CCR en les famílies amb forta agregació per la malaltia i sense alteracions genètiques en gens implicats en les síndromes hereditàries de predisposició al CCR.

2. Validació de les CNVs germinal identificades mitjançant tècniques genòmiques d'alta resolució i estudis de incloent familiars addicionals en les famílies portadores.

3. Caracterització de les conseqüències molecular de les variants germinals identificades mitjançant la monitorització dels nivells d'expressió gènica i proteica als pacients portadors.

Aplicació bioinformàtica per a l'estudi de CNAs

4. Implementació del paquet Shiny, de R Studio, per al desenvolupament d'una eina bioinformàtica d'anàlisi de segments genòmics mitjançant el programa d'R, que faciliti la caracterització dels perfils genòmics tumorals segons la càrrega de CNAs somàtiques de les mostres.

5. Implementació de l'eina desenvolupada a l'anàlisi de les dades públiques del projecte *The Cancer Genome Atlas*, intentant validar estudis de caracterització genòmica previs i dels perfils de CNAs somàtiques en els distints tipus tumorals.

6. Identificació dels perfils genòmics tumorals de CNAs en mostres de càncer de còlon i la seva integració amb dades d'anotació molecular disponibles.

7. Anàlisi, mitjançant l'eina desenvolupada, dels segments genòmics inferits de les dades de seqüenciació de l'exoma germinal en la cohort del CCR familiar per a la re-identificació de la regió caracteritzada al primer estudi, i comparativa dels valors de la regió amb mostres de la cohort de càncer de còlon i recte.

Metodologia

Estudi 1

La metodologia descrita a continuació correspon a la presentada a l'article *Rare germline copy number variants in colorectal cancer predisposition characterized by exome sequening analysis*, publicat a la revista *Journal of Genetics and Genomics*, de l'editorial Elsevier, el 20 de gener del 2018.

(doi: 10.1016/j.jgg.2017.12.001)

Selecció familiar i extracció de mostres – projecte FAMCOLON

Donat l'objectiu global de l'estudi d'identificar i caracteritzar aquelles CNVs potencialment implicades en la predisposició al CCR, gran part de la importància recau en la selecció dels individus que s'hauran d'estudiar i d'on s'extrauran les mostres per al seu posterior anàlisi.

Setanta-un pacients d'entre 38 famílies amb forta agregació familiar per a CCR i compatibles amb un patró d'herència autosòmica dominant van ser inclosos en l'estudi 1. A l'hora de seleccionar les famílies i els individus, es tingueren en compte els mateixos criteris de selecció que s'havien aplicat en anteriors treballs del grup (Esteban-Jurado et al., 2015, 2016). Es vigilà la presència d'alteracions en forma de mutacions puntuals o estructurals en els gens hereditaris més coneguts per al CCR. Només es van incloure aquells individus que no presentaren mutacions en el gens *APC*, *MUTYH* i en gens implicats en la via de reparació del DNA *mismatch repair*. Pel que fa al procés d'inclusió de les famílies, s'aplicaren els següents criteris a l'hora de la seva selecció: tres o més individus afectats per CCR en la família, dues o més generacions afectades per la malaltia i de forma consecutiva, i que almenys un dels individus amb CCR hagués estat diagnosticat abans dels 60 anys.

Finalment, la cohort familiar, anomenada FAMCOLON a nivell intern del laboratori, es composava d'una família amb 4 individus seqüenciats, 10 famílies amb 3 membres seqüenciats, 11 famílies amb 2 individus seqüenciats i 16 famílies amb només un individu seqüenciat. Per a la família en que s'identificà la variant més interessant es van poder consultar i obtenir mostres en alguns membres addicionals, facilitant l'estudi i validació dels resultats obtinguts durant l'anàlisi de CNVs. Les institucions i comitès d'ètica dels respectius hospitals van avalar la selecció, inclusió i extracció de mostres dels pacients i es van recollir els consentiments informats de cada un d'ells.

El DNA germinal dels pacients, del qual es seqüenciaria l'exoma, es va aïllar a partir de mostres de sang perifèrica. Per tal d'obtenir mostres d'RNA (incloent els miRNAs) i realitzar estudis d'expressió gènica en alguns dels pacients seleccionats, es van utilitzar els tubs PAXgene Blood miRNA per a l'extracció de sang i el kit d'extracció PAXgene Blood RNA (PreAnalytix, Hombrechtikon, Suïssa). Per alguns dels pacients inclosos en l'estudi es va poder obtenir RNA de mostres de tumor parafinat utilitzant el kit d'extracció RNeasy FFPE (QIAGEN, EUA).

Seqüenciació de l'exoma

La seqüenciació de l'exoma en mostres de DNA germinal va ser realitzada al Centre Nacional d'Anàlisi Genòmic (CNAG, Barcelona), així com també les anàlisis preliminars de les dades de seqüenciació.

En un primer pas, el DNA germinal aïllat de la sang dels pacients es va sotmetre a un control de qualitat per tal d'assegurar la seva viabilitat a l'hora de realitzar la seqüenciació de l'exoma. Les especificacions necessàries per a un resultat positiu en aquest control foren: 3-5 µg de DNA per mostra, a un concentració de 50 – 300 ng/µL mesurada amb l'assaig de detecció PicoGreen (ThermoFisher, EUA) i amb ratis d'absorbància A260/A280 = 1,7 – 2. La integritat de la molècula de DNA, per mostra, fou valorada mitjançant la tècnica d'electroforesis en agarosa.

La seqüenciació de l'exoma complet es realitzà utilitzant la plataforma HiSeq2000 (Illumina, San Diego, Califòrnia) i el kit d'enriquiment SureSelectXT Human All Exon V4 o V5 (Agilent, Santa Clara, Califòrnia). El cisallament inicial de les molècules de DNA es va dur a terme mitjançant un ultrasonicador de la sèrie S2 (Covaris, Massachusetts, EUA). El tamany de les llibreries de fragments obtingudes i la seva concentració foren avaluades mitjançant electroforesis de capil·laritat amb el Bioanalyzer 2100 (Agilent). Adaptadors amb distints identificadors per a cada mostra foren incorporats durant el procés d'enriquiment d'exons, permetent la combinació de diferents senyals de les mostres abans de la seqüenciació. Acabat el procés d'enriquiment, les llibreries indexades foren agrupades i seqüenciades paral·lelament de forma massiva, utilitzant el protocol

d'extrems aparellats amb dos lligands de 75 nucleòtids de llargada per a cada fragment.

Finalitzat el procés de seqüenciació, i seguint les indicacions del GATK (de l'anglès, The Genome Analysis Tool Kit) (McKenna et al., 2010), es realitzaren diversos passos per al pre-processament de les dades. Així, les lectures obtingudes (o *reads*, en anglès) s'alinearen al genoma de referència hg19/GRCh37 mitjançant l'eina bioinformàtica Burrows Wheeler-Aligner (BWA) (H. Li & Durbin, 2009); es marcaren els reads duplicats durant la seqüenciació amb el programa MarkDuplicates, de PICARD (Broad Institute of MIT and Harvard, 2019); es calibraren les avaluacions de la qualitat en la predicció de les bases (Base Quality Score Recalibration, BQSR); i es realinearen localment aquells fragments pròxims a zones de indels mitjançant el programa IndelRealigner, del GATK. Finalment s'obtingueren els arxius de seqüenciació pre-processats en format BAM (The SAM/BAM Format Specification Working Group, 2018) per al seu posterior anàlisi de variants del nombre de còpia.

Detecció, anotació i priorització de variants del número de còpia

Després del pre-processament de les dades i obtinguts els arxius preliminars en format BAM, es procedí a aplicar el fluxe de treball que s'havia dissenyat per tal de detectar les CNVs en els membres de les famílies amb agregació per al CCR (**Figura 25**) Per a la inferència de CNVs en les dades de seqüenciació de l'exoma es van utilitzar dos algoritmes diferents de detecció: el conegut paquet d'R, ExomeDepth (Plagnol et al., 2012), per a la detecció general de les CNVs en el conjunt de les nostres dades, i el programa CoNIFER (de l'anglès *Copy Number Inference From Exome Reads*) (Krumm et al., 2012), un sistema de comandes basat en llenguatge de programació Python que aplica el mètode estadístic SVD (*Singular Value Decomposition* en anglès) per tal d'aïllar les variants de número de còpia més característiques i, d'aquesta manera, obtenir les variants més rares al conjunt de mostres. En ambdues eines s'utilitzaren els paràmetres per defecte. En la inferència obtinguda després d'aplicar CoNIFER fou necessari segmentar les dades mitjançant el paquet d'R DNAcopy (Venkatraman & Olshen, 2007). Finalment, es van seleccionar aquelles CNVs detectades en les dues inferències, reforçant la confiança en la detecció.

Per tal de prioritzar aquelles variants més rares i descartar les més comunes a nivell poblacional, s'anotaren les freqüències de detecció de les CNVs obtingudes en la inferència. Per això, s'utilitzà la base de dades Database of Genomic Variants (DGV) (MacDonald, Ziman, Yuen, Feuk, & Scherer, 2014), un catàleg revisat periòdicament que conté informació de múltiples estudis de variants estructurals en genomes de pacients control, i on s'aporta informació dels mateixos estudis i la plataforma d'identificació de les variants, a més de les freqüències de detecció d'aquestes. Per altra banda, es va poder accedir a 500 mostres d'individus control d'arreu d'Espanya processades mitjançant la tècnica genòmica de SNP arrays que provenien de l'estudi prospectiu i multicèntric EPICOLON (Ceres Fernandez-Rozadilla et al., 2013). Aquestes dades foren analitzades mitjançant les Affymetrix Power Tools (APT) (Affymetrix - Thermofisher, n.d.), per tal de caracteritzar els

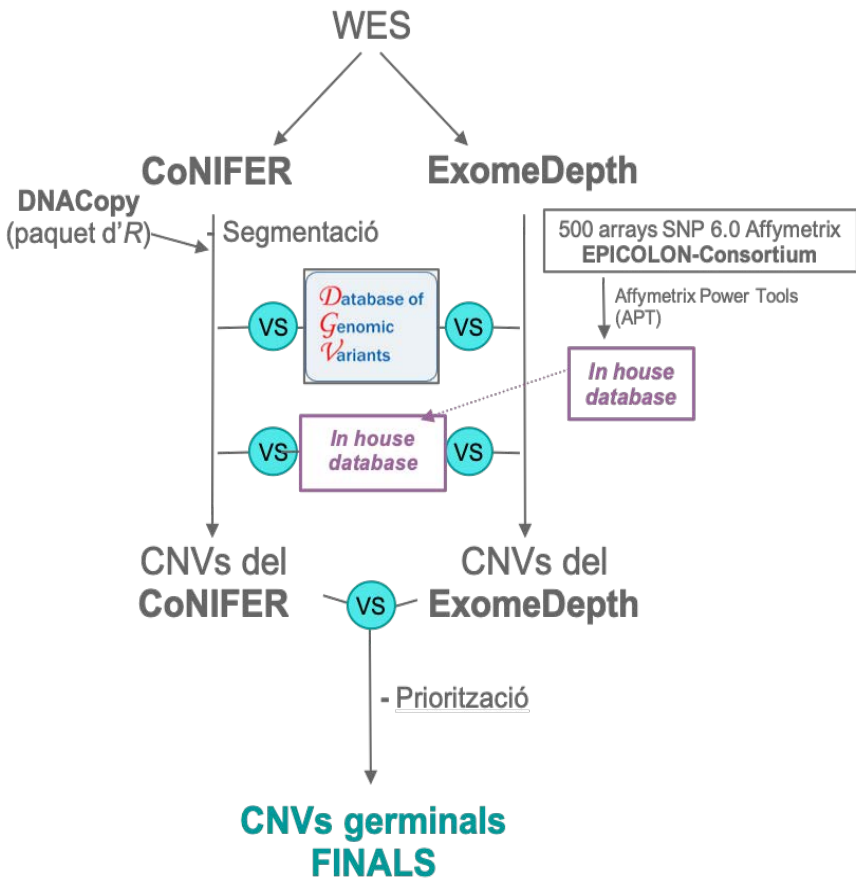


Figura 25. Flux de treball automatitzat per a la inferència, anotació i priorització de CNVs germinals en les dades WES familiars.
L'esquema de treball per a la identificació de CNVs a partir de les dades de seqüenciació de l'exoma dels individus seqüenciats de les famílies amb agregació per CCR.

perfils de CNVs per a cada mostra. Posteriorment, es calcularen les freqüències de detecció per a cada variant dins el conjunt de les 500 mostres. Aquestes dades de freqüència foren utilitzades per anotar les CNVs resultants de la inferència en dades de l'exoma.

Per a l'anotació de les freqüències poblacionals en les variants inferides es considerà un 20% de la variant com a marge d'error en els extrems de les CNVs. D'aquesta manera, es recollien les freqüències d'aquelles variants detectades tant en el DGV com en les 500 mostres de l'EPICOLON que es solapaven amb cada una d'aquestes variants, permetent una errada en la posició dels extrems de fins ± 20% de divergència en les coordenades genòmiques (**Figura 26**).

Seguidament, es prioritzaren les variants comunes entre els individus de la mateixa família i, aprofitant les diverses anotacions i característiques

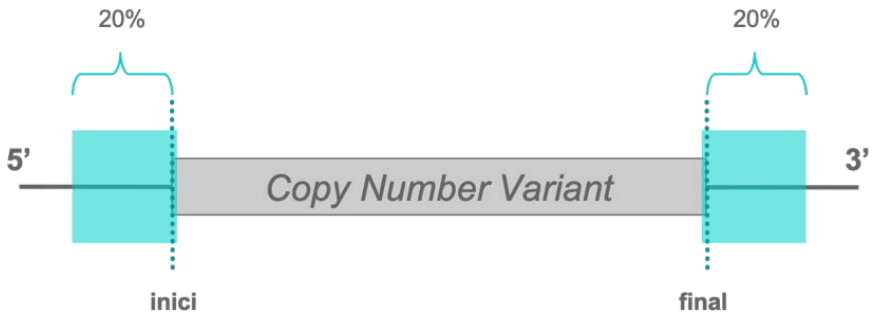


Figura 26. Il·lustració esquemàtica de l'expansió d'extrems de CNVs per a l'anotació en les bases de dades.
Esquema que il·lustra com s'expandeixen els marges que delimiten les CNVs candidates, resultat de les inferències mitjançant CoNIFER i ExomeDepth, per tal de realitzar l'anotació de la freqüència poblacional i de les bases de dades del DGV i de les 500 mostres de l'EPICOLON.

disponibles de les CNVs obtingudes durant la inferència, es van prioritzar aquelles CNVs que complissin els següents criteris: haver-se detectat per ambdues eines (ExomeDepth i CoNIFER), haver-se trobat en cap o menys de 10 individus en les bases de dades indicades i presentar una longitud superior a les 50 Kb.

Caracterització genòmica de la variant de número de còpia detectada

La validació i caracterització genòmica de la duplicació del cromosoma 1 detectada en la família 7 de la cohort FAMCOLON es va dur a terme mitjançant tècniques genòmiques. Les múltiples tècniques utilitzades anaven encaminades a intentar definir, de la manera més exacte i amb la màxima resolució genòmica possible, la longitud i la posició de la variant, per tal de donar pas a la interpretació de les possibles conseqüències funcionals de la mateixa.

Hibridació genòmica comparada en microarray

S'utilitzaren 2 µg de DNA genòmic de les mostres de la família portadora de la variant i de les mostres comercials del kit de la tècnica (Promega, EUA). Les mostres foren incubades durant dues hores amb els enzims de digestió *AluI* i *RsaI*. Posteriorment, les mostres problema i les mostres referència foren marcades, durant dues hores, amb Cy3-dUTP i Cy5-dUTP, respectivament, mitjançant el SureTag Complete DNA Labeling kit (Agilent, EUA). Els nucleòtids no incorporats es rentaren i les mostres, diferencialment marcades amb Cy3 i Cy5, es combinaren en quantitats equivalents seguint la lectura d'incorporació de nucleòtids realitzada amb el Nanodrop ND-1000 (Life Technologies, Carlsbad, EUA). Les matrius d'oligonucleòtids SurePrint G3 Human CGH Microarray 4x180K ò SurePrint G3 Human CGH Microarray 1x1M (Agilent, EUA) es varen hibridar a les mostres amb una incubació a 65ºC durant 24 hores. Les matrius es rentaren, s'escanejaren mitjançant el G2565BA *laser scanner* (Agilent, EUA) i les imatges es processaren amb el programa Feature Extraction™. Finalment, l'anàlisi i la visualització dels resultats es realitzaren aplicant el programa comercial Nexus Copy Number (BioDiscovery, EUA) i, paral·lelament, aplicant els paquets d'R del BioConductor Limma (Ritchie et al., 2015) i DNAcopy (Venkatraman & Olshen, 2007).

Seqüenciació completa del genoma

La seqüenciació completa del genoma (WGS) aportà major resolució del perfil genòmic de la mostra en qüestió i va permetre caracteritzar la totalitat de la informació genètica present en la molècula de DNA. La WGS va ser realitzada al Centre Nacional d'Anàlisi Genòmic

(CNAG, Barcelona), i la preparació de la llibreria d'extrems aparellats es dugué a terme mitjançant el TruSeq™DNA Sample Preparation Kit v2 (Illumina, EUA) i el KAPA Library Preparation kit (Kapa Biosystems, EUA). Els extrems de 2 μg de DNA genòmic fragmentat foren reparats, adenilats i lligats amb adaptadors específics d'Illumina. Per tal d'obtenir fragments de DNA entre 220 i 550 pb s'utilitzaren boles magnètiques de AMPure XP (Beckman Coulter, EUA). La llibreria final fou quantificada mitjançant el Library Quantification Kit (Kapa Biosystems, EUA) i seqüenciada utilitzant *paired-end* mode HiSeq 4000 SBS kit (Illumina, EUA) per a generar 2 x 151 pb de *reads* i obtenint més de 160 Gb amb una cobertura per base de 50x. El programa Real Time Analysis (RTA 2.7.6) s'aplicà a l'hora d'analitzar les dades crues, la inferència de la seqüència genòmica i l'avaluació de la seva qualitat. L'eina bioinformàtica CASAVA es va fer servir per a generar els arxius FASTQ ("Illumina CASAVA-1.8 FASTQ Filter," 2011).

El mapatge i alineament de les seqüències i l'anotació de les variants es va realitzar aplicant l'eina GEM -de PICARD (Broad Institute of MIT and Harvard, 2019)-, el GATK v.3.6, SnpEff i SnpSift (McKenna et al., 2010; Cingolani, Patel, et al., 2012; Cingolani, Platts, et al., 2012; Marco-Sola, Sammeth, Guigó, & Ribeca, 2012). Finalment, s'aplicà el programa DELLY, utilitzant els paràmetres per defecte, per a la inferència de variants estructurals en la seqüenciació del genoma complet, deixant fora de l'anàlisi les zones centromèriques i telomèriques (Rausch et al., 2012).

Seqüenciació Sanger

Aprofitant la identificació de variants estructurals en les dades de WGS es van dissenyar primers per tal d'elucidar la zona del ruptura de la duplicació del cromosoma 1 detectada:

FORWARD: 5'-TTAAGGACTGTGGCTTTTGC-3'

REVERSE: 5'-GAGTTCAGGGCAGACAGAAA-3'

El producte de PCR obtingut utilitzant la mostra de DNA d'un dels individus portadors de la duplicació del cromosoma 1 fou seqüenciat mitjançant la tècnica de *Sanger sequencing* (GATC Biotech, Alemanya).

Bases de dades d'informació gènica

Distintes bases de dades van ésser utilitzades i consultades per a l'anotació i posterior priorització de les CNVs inferides de les dades de seqüenciació del exoma, així com també per obtenir informació sobre els distints gens afectats per aquestes.

Database of Genomic Variants

La *Database of Genomic Variants* (DGV, http://dgv.tcga.ca/) és un repositori públic i accessible que conté informació de les variants genòmiques estructurals identificades en mostres d'individus control sans (MacDonald et al., 2014). El catàleg conté més de 70 estudis que aporten informació d'aproximadament 6.400.000 variants estructurals amb un rang de longitud genòmica des de les 50 pb fins a més d'una megabase. Les entrades individuals per a cada variant identificada al repositori aporten informació útil anotada com el tipus de variant (duplicacions o delecions), la seva freqüència (segons l'estudi que ha identificat aquella variant), la tecnologia aplicada i el mètode de detecció, i el propi estudi on ha estat identificada. A més, la pàgina web dona l'opció de descàrrega del catàleg global, facilitant la integració d'aquesta informació en els estudis propis.

TARGETSCAN

La pàgina TARGETSCAN (http://www.targetscan.org) allotja un programa de predicció de dianes genètiques dels distints miRNA humans (Agarwal, Bell, Nam, & Bartel, 2015). Les prediccions es generen mitjançant l'estudi del grau de complementarietat entre les seqüències dels mRNAs humans i les regions "llavor" (*seed* en anglès) dels miRNAs (Lewis, Shih, Jones-Rhoades, Bartel, & Burge, 2003).

UCSC Genome Browser

Al juny de l'any 2000, la Universitat de Califòrnia – Santa Cruz (UCSC) i els demés membres de consorci internacional del projecte del genoma humà van completar el primer esbós de l'ensamblatge del genoma humà, restant de forma totalment oberta a la comunitat científica des de llavors. Poc després, es publicà el visualitzador del

genoma més conegut i usat avui en dia: el UCSC *Genome Browser* (Kent et al., 2002; Casper et al., 2017).

Durant les posteriors dues dècades, aquest visualitzador ha anat acumulant les dades d'ensamblatge de les noves i millorades versions del genoma humà i altres vertebrats, així com també múltiple i diversa informació de la resta d'òmiques de la biologia, com transcriptòmica i epigenòmica, a més d'altres tipus d'anotació genètica i proteica. Tota aquesta quantitat de dades es troba processada i disponible per a la seva visualització i/o descàrrega en la pàgina https://genome.ucsc.edu, així com diverses eines computacionals que faciliten l'estudi de les òmiques als investigadors de la comunitat científica.

En aquest estudi, el UCSC *Genome Browser* s'ha consultat per a la visualització genòmica de les regions identificades com CNVs en la cohort de WES i per a la comprovació dels gens que s'incloïen en aquestes.

National Center for Biotechnology Information

La creixent importància de la informació computeritzada en el món de la recerca biomèdica ocasionà la formació, al 1998, de la institució americana National Center for Biotechnology Information (NCBI; en català, centre nacional per a la informació biotecnològica). La missió principal d'aquesta és la de desenvolupar, allotjar i oferir noves tecnologies de la informació que serveixen d'ajuda als investigadors en el seu procés de descobriment i comprensió dels processos fonamentals, genètics i moleculars, tant patològics com en la salut.

Gran quantitat d'informació bibliogràfica, pràctica, educativa, d'assessorament o de desenvolupament de nova tecnologia es troba a disposició de la comunitat científica en la seva pàgina web: https://www.ncbi.nlm.nih.gov.

La base de dades del NCBI s'ha consultat per tal de recavar informació sobre els gens inclosos en les regions identificades amb CNVs en aquest estudi, en quan a la seva funció i altres variables d'anotació.

Validació molecular

Matriu d' expressió gènica

Es realitzà un estudi extensiu dels nivells d'expressió dels gens en mostres de sang d'alguns individus portadors i mostres de CCR esporàdic i mostres d'individus sans.

Les mostres d'RNA extretes es rentaren amb RNeasy MinElute Cleanup Kit (Qiageni, EUA) i s'hibridaren a la matriu Whole-Genome DASL HT BeadChip (Illumina, EUA). Les diferències relatives (en anglès, *fold-change*) en l'expressió gènica foren posteriorment analitzades amb les eines BeadArray Reader, el programa BeadStudio i finalment processades aplicant la normalització per quantils del paquet informàtic Lumi, del programa estadístic R (Du, Kibbe, & Lin, 2008).

Real Time quantitative PCR

La quantificació dels nivells d'expressió dels mRNA i els miRNA afectats per la duplicació fou avaluada en mostres de sang i de talls de teixit parafinat mitjançant la tècnica de *Real-Time quantitative PCR* (RT-qPCR).

S'utilitzaren 400 ng d'RNA per a la retro-transcripció amb el High-Capacity cDNA Reverse Transcription Kit, per a les mostres provinents de sang, i la TaqMan MicroRNA reverse transcription kits per a les mostres de teixits parafinats. Per a cada gen a quantificar s'utilitzaren les TaqMan assays (Applied Biosystems, EUA) específiques. Per al control endògen de la tècnica, s'utilitzà el gen GAPDH en la quantificació dels mRNA en sang, i el gen MIR16 per a mostres de parafina. Abans de l'extracció de sang s'introduí un control exògen, cel-MIR39, per a controlar l'extracció i quantificació dels miRNA. L'expressió relativa per a cada gen avaluat fou calculada per a cada mostra, obtenint valors de -ΔCt , de la següent manera:

$$-\Delta Ct = -[Ct_{gen\ problema} - Ct_{control\ endògen}]$$

Els gràfics per a la representació dels resultats es realitzaren utilitzant el programa estadístic SPSS (SPSS, EUA).

Immunohistoquímica

La tècnica inmunohistoquímica en talls de teixit va permetre avaluar la presència o absència de les proteïnes a estudiar. Els talls histològics es van escalfar a 37ºC, es prepararen amb xilè (tres rentats) per tal d'eliminar les restes de parafina del teixit i s'hidrataren mitjançant una gradació negativa d'alcohols. La recuperació d'antígens es realitzà amb buffer de citrat 10 mM. L'activitat peroxidasa es bloquejà amb H_2O_2 al 3%. Els talls es tractaren durant 30 minuts amb bloquejant proteic sense sèrum (DakoCytomation, Dinamarca) i s'incubaren amb els respectius anticossos de detecció. Posteriorment, els talls s'incubaren durant una hora, a 37ºC, amb anticòs secundari. Les característiques dels anticossos primaris i secundaris per a les proteïnes estudiades s'especifiquen a la **Taula 5**.

Taula 5. Anticossos primaris i secundaris per als estudis de immunohistoquímica de les proteïnes TTF2 i TMEM158.

	Proteïna	Nom	Incubació	Dilució
AC Primari	TTF2	Monoclonal Rabbit Anti-TTF2 (AbCam, Regne Unit)	20 hores	1/250
	TMEM158	Polyclonal Rabbit anti-TMEM158 (AbCam, Regne Unit)	14 hores	1/250
AC Secundari	(ambdues)	Goat anti-Mouse/Rabbit ChemMate DAKO Envision /HRP (DakoCytomation, EUA)	1 hora (37ºC)	

El color de les tincions s'aconseguí amb diaminobenzidina -DAB- (DakoCytomation, EUA). Per als contrasts en la tinció s'aplicaren l'hematoxilina i la eosina. Finalment, els talls es muntaren i fixaren utilitzant el DPX i s'observaren amb el microscopi Olympus System Microscope Model BX41 (Olympus, Japó).

Estudi 2

L'estudi 2 ha consistit en el desenvolupament, disseny i implementació d'una eina bioinformàtica, CNApp, per facilitar l'anàlisi de CNAs en estudis d'integració amb anotacions i característiques moleculars o clíniques, i així afavorir projectes de translacionalitat entre la recerca bàsica i la clínica amb l'objectiu d'avaluar l'aneuploïdia en tumors o altres tipus de mostres.

Llenguatge de programació i interfície de desenvolupament

El desenvolupament de l'eina bioinformàtica s'ha dut a terme mitjançant la implementació del llenguatge de programació R (R Core Team, 2018), històricament utilitzat en el camp de la bioestadística. Aquest programa conté un enorme catàleg de paquets i llibreries enfocats a l'estudi i anàlisi de dades biomèdiques. Com a conseqüència de la gran explosió en aquest camp, gràcies al seu caràcter open *source*, i a un manteniment basat en la comunitat d'usuaris, aquest s'ha consolidat com un dels llenguatges més aplicats en l'estudi de les òmiques biomèdiques. La versió del llenguatge utilitzat per al desenvolupament de l'eina CNApp ha estat la R version 3.4.2 (2017-09-28) -- "Short Summer".

El paquet d'R Shiny (versió 1.1.0), del projecte R-Studio (https://www.rstudio.com/), s'ha aplicat a l'hora de generar la interfície d'usuari (Chang, Cheng, Allaire, Xie, & McPherson, 2018). Aquest paquet representa el pont entre el llenguatge en R i el que s'utilitza per al desenvolupament de pàgines web, el llenguatge HTML. Shiny aporta diverses funcions i utilitats que permeten traduir R a HTML, facilitant la implementació i transformació d'anàlisis bioinformàtics en aplicacions amb aparença de pàgina web.

Entrada de dades al CNApp

CNApp utilitza arxius amb dades de segments genòmics que poden provenir de distintes plataformes d'identificació d'alteracions de número de còpia al DNA (per exemple SNP-arrays, aCGH, WES o WGS). S'accepten les

construccions del genoma humà hg19 i hg38. Són necessaris indicadors per a les columnes de la matriu de dades de l'usuari (amb l'etiqueta corresponent): nom de la mostra *(ID)*, cromosoma *(chr)*, punt d'inici *(start)* i final *(end)* de les coordenades genòmiques del segment, i valor de *log2ratio* del número de còpia que expressi l'amplitud del canvi *(seg.mean)*. Si es tenen a l'abast, es recomana afegir els valors de puresa de la mostra (*purity*) i de BAF per a cada segment (*BAF*), ja que poden millorar la precisió de la re-segmentació de les CNAs i aportar informació de possibles esdeveniments en forma de *copy number neutral loss-of-heterozigosity* (*CN-LOH*). Les variables d'anotació per a realitzar els estudis d'integració de dades amb característiques clíniques o moleculars poden ésser afegides en la matriu de dades inicial (especificant la variable de columna en cada una de les línies de segment de cada una de les mostres) o, per altra banda, carregant un arxiu independent que contengui els noms de les mostres i les característiques d'anotació associades.

Dades públiques utilitzades

Per al desenvolupament i el posterior anàlisi de l'aplicació CNApp s'utilitzaren dades públiques de mostres genòmiques.

Pan-cancer TCGA

Les 10.635 mostres *Level 3* del consorci TCGA, corresponent a dades genòmiques segmentades de SNP6.0 d'Affymetrix, es descarregaren del repositori públic Genomic Data Commons (National Cancer Institute, NIH) (Grossman et al., 2016). La **Taula 6** presenta el llistat dels 33 tipus de càncer inclosos en la cohort, el teixit primari de la neoplàsia, el tipus de tumor i el nombre de mostres.

Taula 6. Llistat dels 33 projectes del TCGA per als distints tipus de càncer analitzats en la cohort de *pan-cancer*.

Projecte	Nom complet	Teixit primari	Tipus de tumor	N mostres
ACC	*Adrenocortical Carcinoma*	Glàndula adrenal	Endocrí	90
BLCA	*Bladder Urothelial Carcinoma*	Bufeta	Urològic	414
BRCA	*Breast Invasive Carcinoma*	Mama	Mama	1101
CESC	*Cervical Squamous Cell Carcinoma and Endocervical Adenocarcinoma*	Cèrvix	Ginecològic	295
CHOL	*Cholangiocarcinoma*	Conducte biliar	Gastrointestinal	36
COAD	*Colon Adenocarcinoma*	Còlon	Gastrointestinal	462
DLBC	*Lymphoid Neoplasm Diffuse Large B-cell Lymphoma*	Nòduls limfàtics	Hematològic	48
ESCA	*Esophageal Carcinoma*	Esófag	Gastrointestinal	184
GBM	*Glioblastoma Multiforme*	Cervell	Sistema nerviós central	594
HNSC	*Head and Neck Squamous Cell Carcinoma*	Cap i coll	Cap i coll	517
KICH	*Kidney Chromophobe*	Ronyó	Urològic	66
KIRC	*Kidney Renal Clear Cell Carcinoma*	Ronyó	Urològic	534
KIRP	*Kidney Renal Papillary Cell Carcinoma*	Ronyó	Endocrí	290
LAML	*Acute Myeloid Leukemia*	Mèdul·la òssia	Hematològic	143
LGG	*Brain Lower Grade Glioma*	Cervell	Sistema nerviós central	514
LIHC	*Liver Hepatocellular Carcinoma*	Fetge	Gastrointestinal	375
LUAD	*Lung Adenocarcinoma*	Pulmó	Toràcic	530
LUSC	*Lung Squamous Cell Carcinoma*	Pulmó	Toràcic	503
MESO	*Mesothelioma*	Pleura	Toràcic	87
OV	*Ovarian Serous Cystadenocarcinoma*	Ovari	Ginecològic	568
PAAD	*Pancreatic Adenocarcinoma*	Pàncreas	Gastrointestinal	184
PCPG	*Pheochromocytoma and Paraganglioma*	Glàndula adrenal	Endocrí	181
PRAD	*Prostate Adenocarcinoma*	Pròstata	Urològic	501
READ	*Rectum Adenocarcinoma*	Colorectal	Gastrointestinal	164
SARC	*Sarcoma*	Teixit bla	Teixit bla	260
SKCM	*Skin Cutaneous Melanoma*	Pell	Pell	104
STAD	*Stomach Adenocarcinoma*	Estómac	Gastrointestinal	442
TGCT	*Testicular Germ Cell Tumors*	Testicles	Urològic	139
THCA	*Thyroid Carcinoma*	Tiroides	Endocrí	505
THYM	*Thymoma*	Timo	Hematològic	124
UCEC	*Uterine Corpus Endometrial Carcinoma*	Úter	Ginecològic	544
UCS	*Uterine Carcinosarcoma*	Úter	Ginecològic	56
UVM	*Uveal Melanoma*	Ull	Cap i coll	80

Cohort de càncer de còlon

D'entre les 10.635 mostres de la cohort del pan-cancer TCGA es tenia informació del subtipus per a la classificació dels grups CMS i altres anotacions moleculars per a 309 mostres del projecte COAD (colon adenocarcinoma) (Guinney et al., 2015). Les anotacions moleculars consistien en l'estat dels microsatèl·lits (MSI o MSS), el genotip mutat o normal de la mostra en quant als gens *KRAS* i *BRAF* (Mut o WT), i el fenotip CIMP (high, low o negative). La Taula 7 presenta el nombre de mostres en cada subgrup de CMS segons el tipus d'anotació.

Taula 7. Distribució de de les variables d'anotació als subtipus de CMS en la cohort de 309 tumors de còlon del TCGA-COAD.

	N	Microsatèl·lits			*KRAS*			*BRAF*			CIMP			
		MSI	MSS	NA	Mut	WT	NA	Mut	WT	NA	High	Low	Neg	NA
CMS1	64	58	6	0	7	58	0	14	50	0	45	11	8	0
CMS2	112	0	109	0	16	96	0	0	112	0	2	21	87	2
CMS3	51	9	39	3	11	40	0	0	51	0	8	21	21	1
CMS4	82	5	71	6	6	75	1	2	79	1	3	11	68	6

N: número de mostres; MSI: *microsatelite instability*; MSS: *microsatelite stability*; *Mut*: genotip mutat; *WT*: genotip normal (*wild type*); CIMP: *CpG islandès methylator phenotype*; NA: no disponible *(non-available*).

Ajustament de l'amplitud de canvi per puresa

L'heterogeneïtat dels tumors pot expressar-se en quant als distints perfils genòmics que puguin presentar les diferents cèl·lules aberrants que formen el propi tumor, però també a la fracció de cèl·lules normals i tumorals que conté la mostra del teixit tumoral. Aquesta última part pot condicionar la inferència de l'amplitud de les CNAs. Per tal d'homogeneïtzar l'anàlisi, CNApp permet realitzar una correcció dels valors d'amplitud tenint en compte les estimacions de la puresa per mostra, entenent la puresa com a fracció total de cèl·lules tumorals del tumor estudiat, facilitant posteriors comparacions entre mostres. Els valors d'amplitud -valors de seg.mean- (n) per a cada mostra (x), quan

l'estimació de puresa està disponible (*r*), es re-calculen (*N*) de la següent manera:

$$N(x) = 2^{n(x)+(r(x)-1)}$$

Quan no es proveeix l'estimació de la puresa, CNApp assumeix valors de r = 1 (100% de puresa) per mostra, per tant, els valors de *seg.mean* no varien dels originals. Per altra banda, el valor mínim de puresa permès és el de r = 0,4 (40%), mentre que valors més baixos es fixen a aquest mínim. Aquest mateix valor mínim de puresa es va aplicar a l'hora de calcular l'amplitud de pèrdua màxima (*C*), agafant com a referència la màxima pèrdua heterozigota possible (de dues còpies genòmiques a una còpia), representant el grau més alt de pèrdua permès a l'aplicació:

$$C = log_2\left(\frac{1}{2} \cdot 0.4\right) = -2.32$$

Càlcul dels *CNA scores*: BCS, FCS i GCS

Els segments analitzats a l'aplicació - ja siguin els segments originals carregats per l'usuari o els resultants del procés de re-segmentació - es classifiquen entre esdeveniments cromosòmics (*chromosomal*), de braç (*arm-level*) o focals (focal), depenent de la seva longitud relativa al tamany del cromosoma o del braç cromosòmic que estiguin afectant. Així doncs, i segons els paràmetres per defecte en l'aplicació (modificables per l'usuari), aquells segments que afecten com a mínim el 90% del cromosoma es classifiquen com *chromosomal*; aquells que afecten almenys 50% del braç cromosòmic es classifiquen com *arm-level*; i els que afecten menys del 50% del braç s'etiqueten com focal.

Posteriorment, els valors d'amplitud de canvi *(seg.mean)* dels segments genòmics s'utilitzen per aportar un pes numèric específic als distints esdeveniments per tal de calcular els *CNA scores*. Per això, s'estableixen llindars de valors d'amplitud de canvi d'acord a justificacions biològiques en termes de guanys o pèrdues del nombre de còpia genòmica, definint rangs de puntuació. La **Taula 8** especifica els valors del llindars definits per defecte en l'aplicació.

Per altra banda, la inclusió dels valors de BAF en les dades d'origen permet la identificació d'alteracions en forma de CN-LOH en les mostres

Taula 8. Valors de tall per als valors de canvi aplicats als segments i les respectives equivalències en grau d'esdeveniment, número de còpia genòmica i pes que s'atorga al propi segment.

Valors de tall (Log2ratio)	Grau d'esdeveniment	Número de còpia	Pes numèric (*A*)
1	Guany d'alt grau	≥ 4 copies	3
0,58	Guany de grau mig	[3 – 4) copies	2
0,2	Guany de baix grau	[2,3 – 3] copies	1
-0,2	Pèrdua de baix grau	(1 – 1,7] copies	1
-1	Pèrdua de grau mig	(0,6 – 1] copies	2
-1,74	Pèrdua d'alt grau	≤ 0,6 copies	3
[-0,2 – 0,2]	Pèrdua d'heterozigositat – sense canvi de número de còpia (CN-LOH)	[1,7 – 2,3] copies BAF ≥ 0,25	2

CN-LOH: *copy neutrol loss of* heterozygozity; BAF: *B-allelic freqency*.

Així, les CNAs classificades com *broad* (tant les cromosòmiques com les de nivell de braç cromosòmic) s'utilitzen per al càlcul del *broad CNA score*, o BCS, tenint en compte el pes aportat segons el rang d'amplitud de canvi del número de còpia atorgat (**Taula 8**). De la mateixa manera, els segments classificats com a focals s'utilitzaran per a calcular el focal *CNA score*, o FCS.

El càlcul del BCS per a cada mostra respon al següent: per a un total de *N* segments classificats com a *chromosomal* o com *arm-level* en una mostra específica *(x)*, es considera el sumatori de les puntuacions dels segments *(A)* de la mostra, essent i cada respectiu segment:

$$BCS(x) = \sum_{i=1}^{N} A_i$$

El càlcul del FCS es realitza de la mateixa manera que en el BCS, però tenint en compte el pes de la longitud relativa *(L)* del segment focal

en relació al braç cromosòmic que estigui afectant i segons els paràmetres especificats en la **Taula 9**:

$$FCS(x) = \sum_{i=1}^{N} A_i \cdot L_i$$

Taula 9. Percentatges d'afectació dels braços cromosòmics que suposen la puntuació *L* del segment.

% cobertura braç cromosòmic	Puntuació de longitud relativa (*L*)
≤5%	1
>5% to ≤15%	2
>15% to ≤30%	3
>30%	4

Finalment, el càlcul del global *CNA score* (GCS), per a cada mostra es realitza de la següent manera: per a un mostra específica *(x)*, es considera la suma dels valors de BCS i FCS normalitzats per aquella mostra, on *meanBCS* i *meanFCS* representen els valors de mitjana extrets entre el total de mostres de BCS i FCS, respectivament, i *sdBCS* i *sdFCS* expressen els valors de la desviació estàndard de BCS i FCS:

$$normBCS(x) = \frac{BCS(x) - meanBCS}{sdBCS}$$

$$normFCS(x) = \frac{FCS(x) - meanFCS}{sdFCS}$$

$$GCS(x) = normBCS(x) + normFCS(x)$$

Correlació entre els valors dels *CNA scores* i la fracció alterada del genoma

Per tal de valorar la capacitat dels *CNA scores* a l'hora d'estudiar els nivells d'alteracions de número de còpia en els perfils genòmics de les mostres, es dugueren a terme estudis de correlació, aplicant el test d'Spearman, entre els valors dels distints *CNA scores* (BCS, FCS i GCS) i la fracció alterada del genoma. Els valors de BCS i FCS foren correlacionats específicament amb la fracció alterada del genoma per esdeveniments classificats com *broad* (alteracions cromosòmiques i de braç) i focals, respectivament; mentre que els valors de GCS es correlacionaren amb la fracció alterada global del genoma, calculada considerant tant esdeveniments *broad* com focals.

Així, s'analitzaren les 10.635 mostres del TCGA, les quals representaven fins un total de 33 tipus distints de càncer. En un primer pas, s'aplicà la re-segmentació i el càlcul de *CNA scores* per a les mostres. Posteriorment, es procedí a calcular les fraccions alterades del genoma. La fracció alterada global del genoma *(altFract)* es calculà, per a cada mostra *(x)*, considerant el sumatori de totes les longituds *(l)* de les CNAs, dividit entre la longitud total del genoma humà *(hgLength)*:

$$altFract(x) = \frac{\sum_{i=1}^{N} l_i}{hgLength}$$

Les fraccions alterades del genoma per als esdeveniment *broad* i els focals es calcularen, per a cada mostra (*x*), considerant els sumatori de les longituds (*l*) de les *CNAs broad* o focals, respectivament, i dividit entre la longitud total del genoma humà (*hgLength*):

$$Broad\ altFract\ (x) = \frac{\sum_{i=1}^{N} l_{i(Broad)}}{hgLength}$$

$$Focal\ altFract\ (x) = \frac{\sum_{i=1}^{N} l_{i(Focal)}}{hgLength}$$

Correlació entre els distints *CNA scores*

La relació entre els valors dels distints *CNA scores* calculats per l'aplicació (BCS, FCS i GCS) en les 10.635 mostres del TCGA va ésser estudiada mitjançant el test d'associació estadística d'Spearman. Així doncs, es realitzaren les pertinents comparacions entre els *scores* BCS i FCS, BCS i GCS, i FCS i GCS, per a estudiar quins graus de correlació hi havia entre els seus valors.

Per altra banda, per intentar aprofundir en la relació entre el BCS i el FCS, es procedí a realitzar una correlació per a cada un dels valors de BCS. Concretament, i per a cada un d'aquests valors de BCS (K), es seleccionaven aleatòriament 500 mostres que presentaven valors de BCS = $K \pm 5$ i es correlacionaven els seus valors de BCS amb els valors de FCS. En total, s'obtingueren 27 punts de correlació que presenten la dinàmica de relació entre els dos *CNA scores* a mesura que el valors de BCS augmenten.

Estudis d'associació entre els *CNA scores* i variables d'anotació

Per a l'estudi d'associació estadística entre els valors dels distints *CNA scores* i les variables d'anotació carregades per l'usuari, diferents tests s'apliquen de forma automàtica per tal de valorar possibles relacions entre aquestes característiques. Depenent del tipus de variable (categòrica o numèrica) s'apliquen els tests indicats en la **Taula 10**. Els *P*-valors estadístics resultants es presenten en format tabulat a la interfície de l'aplicació, per tal de facilitar la valoració de les potencials relacions entre les distintes variables i els *CNA scores*.

Taula 10. Distints tipus de tests estadístics aplicats segons les característiques de les variables d'anotació.

Tipus de variable		Paramètrica	No paramètrica
Categòrica	n = 2	T-test d'Student	Test de Wilcoxon
	n > 2	ANOVA	ANOVA: Test Kruskal
Numèrica		Correlació de Pearson	Correlació de Spearman

n = quantitat de grups de mostres definida per la variable

Càlcul de les finestres genòmiques

En la secció de *Region profiling* del CNApp el perfil de segments genòmics per a cada mostra es transforma en perfils de finestres genòmiques preestablertes per tal de facilitar la comparació entre mostres i d'estudiar perfils i regions genòmiques específiques entre distints grups de mostres associats a variables d'anotació aportades per l'usuari. Diverses possibilitats de finestres genòmiques estan disponibles en l'aplicació: braços cromosòmics (*arms*), mitjos braços (*half-arms*), citobandes (*cytobands*), sub-citobandes (*sub-cytobands*) i distintes mides de finestres entre 40 Mb fins a 1 Mb (40 Mb, 20 Mb, 10 Mb, 5 Mb i 1 Mb). Per tal de generar aquestes finestres, tant en la versió del genoma hg19 com en la del hg38, es descarregaren els arxius de citobandes del repositori UCSC (Casper et al., 2017), els quals s'utilitzaren com a motlle per a generar les demés finestres genòmiques.

Per tant, els segments genòmics per a cada mostra s'utilitzen per a generar els perfils de finestres genòmiques, depenent del tipus de finestra seleccionada. Les coordenades genòmiques de cada segment són interrogades per tal de seleccionar aquells segments que cauen dins la regió de la finestra genòmica o, per contra, si ocupa la finestra de forma completa. Posteriorment, cada un dels valors d'amplitud de canvi del número de còpia *(S)* dels segments seleccionats *(t)*, s'utilitzen per a calcular la mitjana numèrica *(W)* de la finestra genòmica *(i)*, ponderada

segons la longitud del segment *(l)* relativa a la dimensió de la regió *(L)* de la següent manera:

$$W(i) = \sum_{t=1}^{n} S_t \cdot \frac{l_t}{L(i)}$$

Estudi de regions genòmiques descriptives

Per tal d'estudiar regions genòmiques específiques que puguin estar associades a certes característiques moleculars o clínques, CNApp compara els graus d'alteració de les regions genòmiques, aprofitant els perfils de finestres genòmiques generats, entre grups de mostres definits per les variables d'anotació de l'usuari. S'utilitzen dues aproximacions distintes per aquest estudi: el primer aplica el T-test d'Student per a valorar el grau d'alteració numèrica de cada una de les regions en els distints grups; la segona examina el tipus d'esdeveniment (guany, pèrdua o normalitat) per a cada regió i en cada un dels grups de mostres i aplica el test de Fisher. En ambdues aproximacions s'obtenen els P-valors (*P-value*) per tal de valorar estadísticament les diferències entre els grups. A més, també es poden obtenir els P-valors ajustats (*Adj p-value*) per *false discovery rate* (FDR) aplicant el mètode Benjamini-Hochberg (BH).

Heatmaps de correlació i clusterització

CNApp utilitza visualitzacions en versió heatmap per tal de valorar visualment els estudis presentats. Opcionalment, es poden generar heatmaps de correlació mitjançant els mètodes de Pearson, Spearman i Kendall. A més, es generen heatmaps de clusterització que apliquen el mètode de *hierarchical clustering* per tal de relacionar patrons numèrics entre les mostres.

Models de classificació basats en *machine-learning*

L'aplicació utilitza el paquet d'R randomForest (Liaw & Wiener, 2002) per a calcular models de classificació mitjançant una variable que defineixi grups de mostres i una o múltiples variables amb potencial capacitat de classificar les mostres entre els grups definits. El model es reprodueix 50 vegades i utilitzant subconjunts de mostres aleatòries d'entrenament i de classificació. Per tal de què el model es pugui calcular de forma satisfactòria, s'han de complir certs paràmetres i condicions per defecte:

- La divisió del nombre total de mostres *(N)* entre el nombre de grups *(G)* definits per la variable de grup ha de ser més gran que el nombre de mostres (n) que conté el grup amb menys mostres:

$$P = \frac{N}{G} \;\; ; \;\; P > n$$

- Si aquesta condició no es compleix, el valor de *P* es fixa al 75% del valor d'*n*:

$$\text{si } P \leq n \;\; \text{llavors} \;\; P = n \cdot 0.75$$

- El terme *P* ha de ser més gran que 1 i *N* ha de ser igual o més gran que 20:

$$P > 1 \;\; \text{or} \;\; N \geq 20$$

- Si la variable classificatòria és categòrica, aquesta no ha de tenir més valors únics *(Z)* que el nombre de grups definits *(G)* per la variable de grup:

$$Z < G$$

- El subconjunt de mostres d'entrenament *(T)* es calcula per a cada grup *(g)* dels grups definits *(G)* extraient nombre de mostres *(P)* de cada grup *(g)*:

$$t\,(g) = P \text{ samples from } g$$

$$T = \sum^{G} t_{(g)}$$

Posteriorment al càlcul del model, es genera una matriu de contingència (**Taula 11**) que conté valors de predicció i de referència per tal de calcular valors de precisió global i d'especificitat i sensitivitat per a cada grup.

Taula 11. Exemple de matriu de contingència per al càlcul dels valors d'eficiència, sensitivitat i especificitat del model.

		Referència	
		Group d'interès	(la resta de grups...)
Predicció	Grup d'interès	*A*	*B*
	(la resta de grups...)	*C*	*D*

$$\text{Eficiència} = \frac{A + D}{A + B + C + D}$$

$$\text{Sensitivitat} = \frac{A}{A + C} \qquad \text{Especificitat} = \frac{D}{B + D}$$

Resultats

Estudi 1

Els resultats descrits a continuació corresponen als presentats en l'article *Rare germline copy number variants in colorectal cancer predisposition characterized by exome sequening analysis*, publicat a la revista *Journal of Genetics and Genomics*, de l'editorial Elsevier, el 20 de gener del 2018.

(doi: 10.1016/j.jgg.2017.12.001)

Inferència i priorització de les variants del número de còpia

Les dades de seqüenciació de l'exoma complet foren analitzades mitjançant CoNIFER i ExomeDepth per tal d'inferir les CNVs presents en els respectius individus de les famílies afectades pel CCR familiar. Els resultats de CoNIFER, posteriorment al procés de segmentació aplicant el paquet d'R DNAcopy, presentaren 288 variants (201 duplicacions i 87 delecions) i amb longituds que anaven des de les 133 pb fins a les 1.840.278 pb. Per altra banda, ExomeDepth cridà un total de 3.700 CNVs (1.760 duplicacions i 1.940 delecions), amb un rang de longituds per aquestes variants des de les 36 pb fins 1.378.737 pb. Com a mesura de referència, la variant més curta als resultats d'ExomeDepth (36 pb) afectava un exó; mentre que la CNV més gran, amb 1.840.278 pb i cridada per CoNIFER, incloïa fins a 12 gens.

Amb l'objectiu d'augmentar la confiança en quant a la inferència de les CNV a partir de dades de WES, es compararen els resultats de CoNIFER i ExomeDepth i es prioritzaren aquelles variants que haguessin sigut cridades per ambdues eines. D'aquesta manera, el nombre de variants es va reduir fins a 21, incloent 16 duplicacions i cinc delecions. Seguidament, s'examinaren les variants per tal de que complissin els filtres de priorització en quant a freqüència poblacional (<10 vegades en les bases de dades d'individus control), llargària de la variant (≥50 Kb) i presència en una família amb dos o més membres afectats. Aplicats aquests criteris, la priorització de les CNV resultà en 14 variants finals (**Taula12**).

Taula 12. CNVs candidates inferides mitjançant CoNIFER i ExomeDepth.
Les CNVs predites per cada una de les eines d'inferència es presenten i referides al genoma humà GRCh37/hg19. La llargària de les variants s'especifica, així com també la presència en les bases de dades de *the Database of Genomic Variants* (DGV) i la generada durant l'estudi 1 (EPICOLON). Aquelles variants presents en les bases de dades menys de 10 vegades s'assenyalen en negreta.

	ID familiar	Eina	Posició genòmica	Llargària (pb)	Tipus	Gen	DGV[$]	EPICOLON[&]
1	Ind13	CoNIFER	chr1:1407164-1431582	24418	deleció	*ATAD3C, ATAD3B*	0	0
		ExomeDepth	chr1:1387426-1431197	43771			53	0
2	**fam7**	**CoNIFER**	**chr1:117602949-117753547**	**150598**	**duplicació**	***TTF2, MIR942, TRIM45, VTCN1, MAN1A2***	**0**	**0**
		ExomeDepth	**chr1:117602949-117963271**	**360302**			**1**	**0**
3	Ind6	CoNIFER	chr1:048721845-248790429	68584	deleció	*OR2T29, OR2T34, OR2T10, OR2T11, OR2T35, OR2T27*	493	20
		ExomeDepth	chr1:248737102-248814185	77083			1025	36
4	**Ind3**	**CoNIFER**	**chr11:20691136-21597001**	**905865**	**duplicació**	***NELL1***	**0**	**0**
		ExomeDepth	**chr11:21555920-21596568**	**40648**			**0**	**0**
5	**Ind5**	**CoNIFER**	**chr11:65306274-65325430**	**19156**	**duplicació**	***LTBP3***	**0**	**0**
		ExomeDepth	**chr11:65320896-65321374**	**478**			**0**	**0**
6	fam_nova_4	CoNIFER	chr14:19988700-20374803	386103	duplicació	*POTEM, OR11H2, OR4Q3, OR4M1, OR4N2, OR4K2, OR4K5, OR4K1*	59	38
		ExomeDepth	chr14:20215587-20404761	189174			1158	86
7	Ind10	CoNIFER	chr14:19988700-20374803	386103	duplicació	*POTEM, OR11H2, OR4Q3, OR4M1, OR4N2, OR4K2, OR4K5, OR4K1*	59	38
		ExomeDepth	chr14:20215587-20404761	189174			1158	86

8	**Ind9**	**CoNIFER**	**chr15:43891761-43941039**	**49278**	**deleció**	***CKMT1B, STRC, CATSPER2***	**4**	**0**
		ExomeDepth	**chr15:43888606-43897597**	**8991**			**0**	**0**
9	**fam2**	**CoNIFER**	**chr16:4999972-5135588**	**135616**	**duplicació**	***SEC14L5, NAGPA, ALG1, C16orf89, ALG1, FAM86A***	**1**	**1**
		ExomeDepth	**chr16:4986985-5127535**	**140550**			**1**	**1**
10	Ind4	CoNIFER	chr16:55758786-55808838	50052	duplicació	*CES1P2, CES1*	0	0
		ExomeDepth	chr16:55844429-55866967	22538			114	7
11	**Ind16**	**CoNIFER**	**chr19:4197998-4224805**	**26807**	**duplicació**	***ANKRD24***	**0**	**0**
		ExomeDepth	**chr19:4219588-4224502**	**4914**			**2**	**0**
12	**Ind3**	**CoNIFER**	**chr2:228678570-228789026**	**110456**	**deleció**	***CCL20, DAW1, SPHKAP***	**0**	**0**
		ExomeDepth	**chr2:228678628-228860410**	**181782**			**0**	**0**
13	**fam_nova_2**	**CoNIFER**	**chr4:69313200-69363316**	**50116**	**deleció**	***TMPRSS11E, UGT2B17***	**0**	**1**
		ExomeDepth	**chr4:69403343-69434202**	**30859**			**0**	**0**
14	Ind5	CoNIFER	chr5:795722-851101	55379	duplicació	*ZDHHC11*	11	9
		ExomeDepth	chr5:766813-822010	55197			80	2
15	**Ind9**	**CoNIFER**	**chr5:140529839-140555957**	**26118**	**duplicació**	***PCDHB6, PCDHB17, PCDHB7, PCDHB8***	**0**	**1**
		ExomeDepth	**chr5:140552417-140560021**	**7604**			**9**	**0**
16	Ind5	CoNIFER	chr6:32546549-32634441	87892	duplicació	*HLA-DRB5, HLA-DRB6, HLA-DRB1, HLA-DQA1, HLA-DQB1*	3	0
		ExomeDepth	chr6:32485516-32634384	148868			17	16
17	**Ind8**	**CoNIFER**	**chr6:32546549-32634441**	**87892**	**duplicació**		**3**	**0**

		ExomeDepth	**chr6:32551886-32634384**	**82498**		***HLA-DRB5, HLA-DRB6, HLA-DRB1, HLA-DQA1, HLA-DQB1***	**3**	**0**
18	**Ind3**	**CoNIFER**	**chr7:84624869-84751247**	**126378**	**duplicació**	***SEMA3D***	**0**	**0**
		ExomeDepth	**chr7:84685033-84727281**	**42248**			**0**	**0**
19	**Ind14**	**CoNIFER**	**chr9:27109441-27230165**	**120724**	**duplicació**	***TEK***	**0**	**0**
		ExomeDepth	**chr9:27218775-27229230**	**10455**			**0**	**1**
20	**fam_nova_1**	**CoNIFER**	**chr9:117085336-117088757**	**3421**	**duplicació**	***ORM1***	**0**	**0**
		ExomeDepth	**chr9:117085414-117087432**	**2018**			**0**	**0**
21	**Ind9**	**CoNIFER**	**chr9:117085336-117088757**	**3421**	**duplicació**	***ORM1***	**0**	**0**
		ExomeDepth	**chr9:117085943-117087432**	**1489**			**0**	**0**

ID, identificació. $: La columna DGV indica el nombre d'observacions de la variant a la base de dades *the Database of Genomic Variants* (DGV; http://dgv.tcag.ca/dgv/app/home). &: La columna EPICOLON DGV indica el nombre d'observacions de la variant a la base de dades generada amb les mostres de EPICOLON.

Per tal d'enfocar-se en un estudi de caràcter més funcional en quant a aquestes CNVs prioritzades i de les possibles conseqüències fenotípiques que podrien presentar, es recavà informació per als gens inclosos en cada una. Valorades les funcions dels gens, es va poder concloure la falta de relació potencial amb el CCR entre la majoria de les CNVs prioritzades, amb l'excepció d'una duplicació al cromosoma 1, identificada en ambdós portadors d'una de les famílies estudiades (**Figura 27**). Aquesta duplicació incloïa el gen *VTCN1*, el qual ja s'havia vist sobreexpressat en diversos tipus de càncer, entre ells el CCR (Peng et al., 2015).

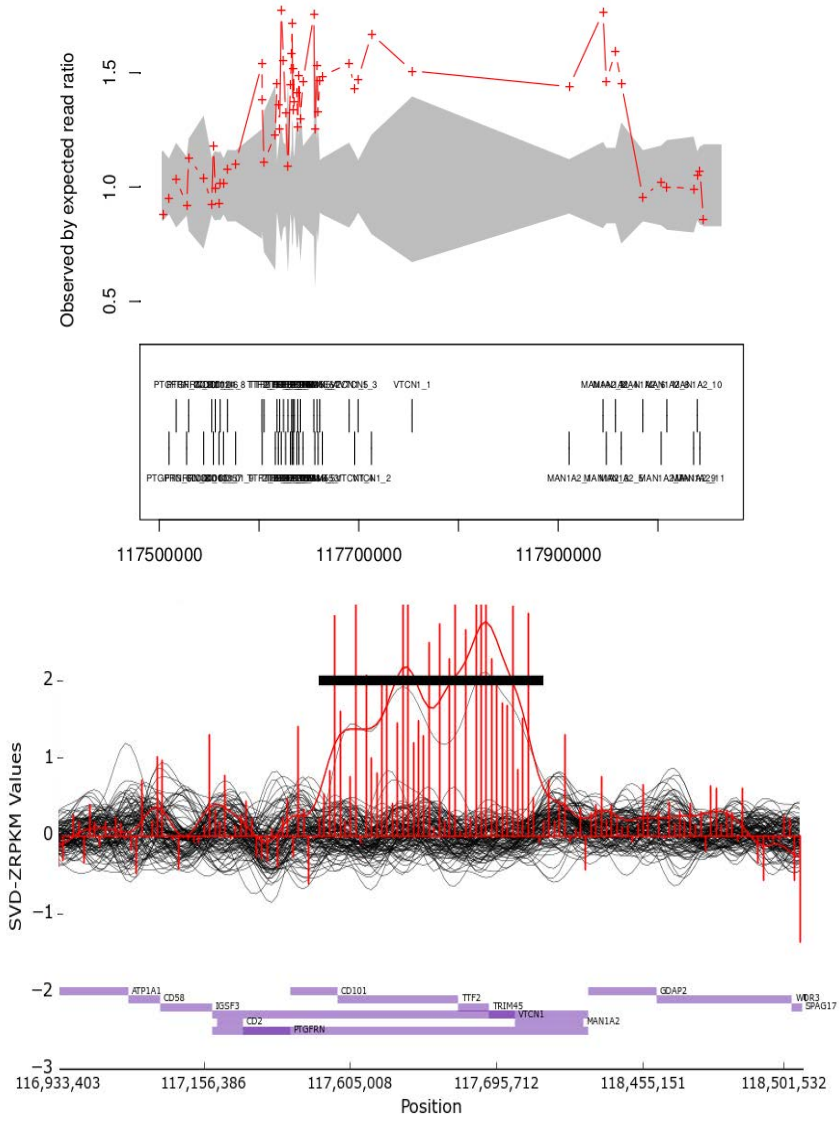


Figura 27. Representacions de la duplicació del cromosoma 1 per ExomeDepth i CoNIFER.
La gràfica superior, exportada mitjançant el paquet ExomeDepth; la inferior, mitjançant l'eina CoNIFER.

Duplicació al cromosoma 1

La duplicació d'una regió del cromosoma 1 identificada en la família 7 del nostre conjunt de famílies va ser la CNV més rellevant d'entre els nostres resultats de la inferència en dades de seqüenciació de l'exoma. Amb una extensió d'aproximadament 400 Kb, la duplicació afectava els gens *TTF2* (de l'anglès: *transcription termination factor 2*), *MIR942* (*microRNA 942*), *TRIM45* (*tripartite motif containing 45*), *VTCN1* (*V-set domaing containing T-cell activation inhibidor 1*) i part de *MAN1A2* (*mannosidase α class 1A member 2*. Consultades les bases de dades de CNVs descrites en aquest treball, es va veure que la variant en qüestió només havia estat identificada una vegada en la DGV, mentre que en el catàleg intern de les 500 mostres del consorci EPICOLON no s'havia detectat mai.

Es procedí a la detecció de la duplicació del cromosoma 1 en mostres de DNA de distints individus de la família portadora, entre ells els dos individus dels quals s'havia seqüenciat l'exoma, mitjançant array de CGH. D'aquesta manera, a banda de confirmar la realitat de la variant, també s'obtenia el patró de segregació de la duplicació en la família. Es validà la CNV en la regió genòmica chr1: 117588023-117991825 (404 Kb) (**Figura 28**), afectant els gens *TTF2*, *MIR942*, *TRIM45*, *VTCN1* i l'extrem 5' del gen *MAN1A2*. La duplicació es va confirmar en els individus amb l'exoma seqüenciat II1 i II2 i s'identificà també en l'individu III1 (fill de l'individu II1 i sense estar afectat per CCR). Els familiars II3 (germà dels individus amb l'exoma seqüenciat), I2 (mare i diagnosticada per CCR als 84 anys) i III2 (filla del II2 i diagnosticada d'un limfoma als 26 anys) resultaren ésser no-portadors de la variant.

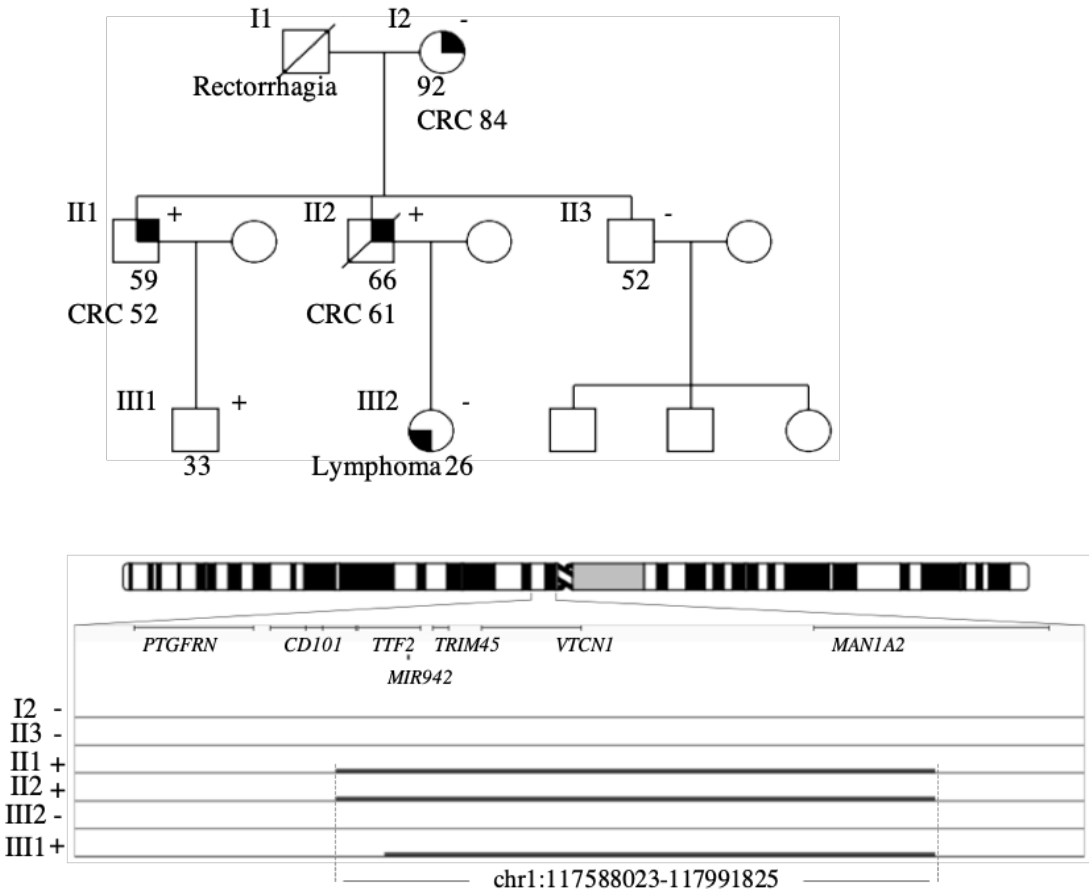


Figura 28. Estudi de la segregació de la duplicació del cromosoma 1 en la família 7.

L'esquema de l'arbre familiar de la família 7 es presenta en la part superior, on es presenten els resultats per a la segregació de la duplicació del cromosoma 1. S'identifiquen els familiars portadors (+) i no portadors (-) de la variant. Els requadres (sexe masculí) i els cercles (sexe femení) amb el quart superior dret negre corresponen als diagnosticats amb CCR, mentre que el quart inferior esquerre negre identifica l'individu diagnosticat amb limfoma. L'edat del pacient al moment de l'estudi s'identifica a sota de la representació individual, així com l'edat de diagnòstic i la patologia associada. En la part inferior es mostra una imatge esquemàtica de la regió del cromosoma 1 afectada per la duplicació. Els gens de la regió es mostren anotats, així com la regió identificada en cada un dels individus estudiats.

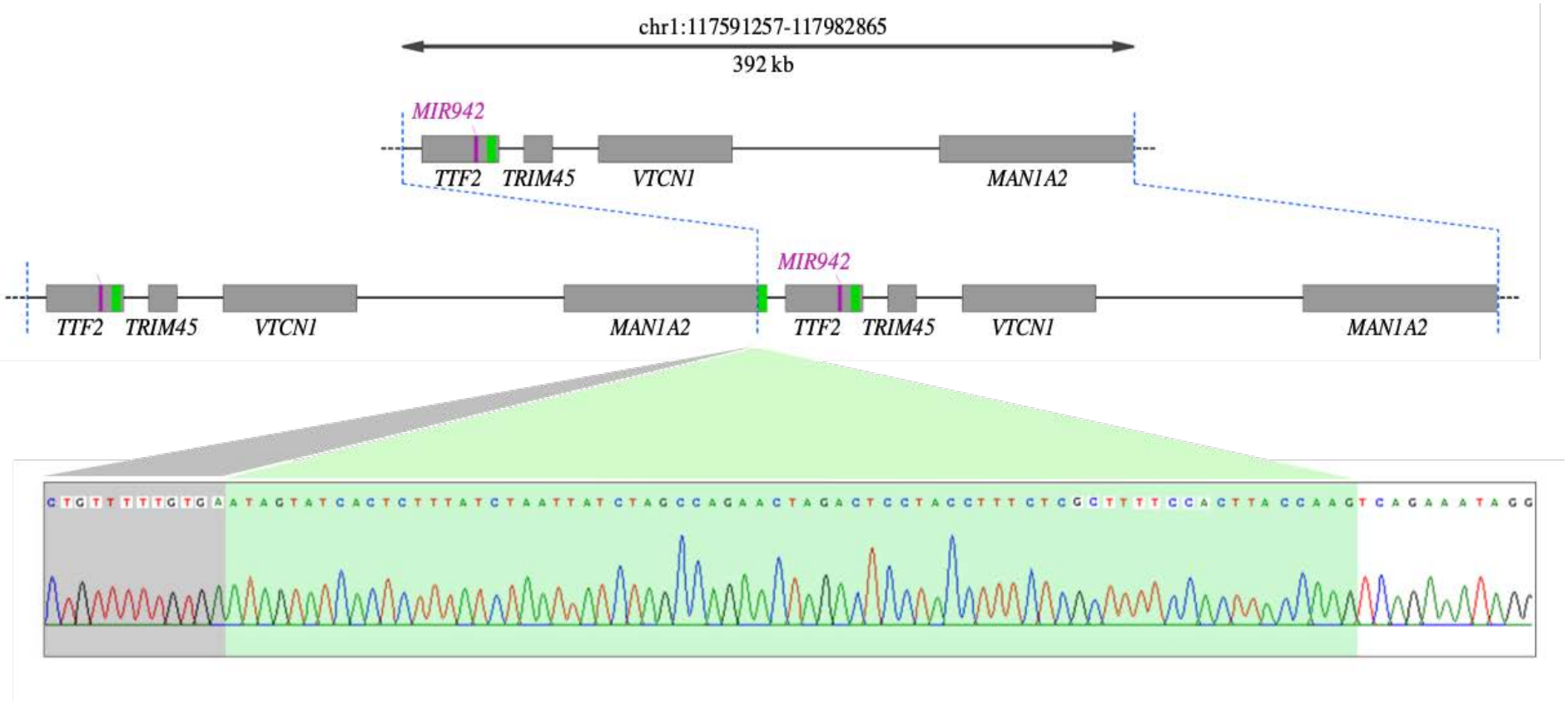


Figura 29. Representació gràfica de la duplicació del cromosoma 1 identificada en la família 7.
Els gens *TTF2*, *TRIM45*, *VTCN1* i *MAN1A2* es presenten com a caixes grises. La regió genòmica del *MIR942* s'identifica amb el color lila, mentre que la regió verda fa referencia a la inserció de 72 pb caracteritzada. Els resultats de la seqüenciació per Sanger s'indiquen a la caixa inferior. Els distints nucleòtids s'identifiquen segons el codi de colors: adenina (verd), timina (vermell), guanina (negre) i citosina (blau). Les microhomologies identificades en la zona de ruptura de la duplicació s'indiquen amb quadrats blancs sobre de la base nitrogenada pertinent.

Amb l'objectiu de millorar la caracterització genòmica de la duplicació del cromosoma 1 i eliminar la possibilitat d'existència d'altres variants, tant puntuals com estructurals, que poguessin relacionar-se en la predisposició al CCR en la família, es seqüencià el genoma complet de l'individu I2, un dels portadors de la variant. Les dades de WGS s'analitzaren per a la identificació de variants estructurals i no es detectaren variants amb afectació per als gens relacionats amb la carcinogènesis al CCR. Les noves coordenades genòmiques per a la duplicació, més acurades que en les versions anteriors, s'aprofitaren per a dissenyar els primers utilitzats en la seqüenciació per Sanger posterior. Finalment, els resultats d'aquesta seqüenciació permeteren esbrinar les coordenades genòmiques de la variant, situada en la regió p13.1-p12 del cromosoma 1, i amb una longitud de 391.608 pb (chr1: 117591257- 117982865; aproximadament 392 Kb). A banda, s'identificà una inserció de 72 pb a la zona de ruptura de la variant que provenia de l'últim intró del gen *TTF2* (chr1:117642853-117642924), i es detectaren regions de micro-homologia entre els extrems d'aquesta regió inserida i els extrems 5' i 3' de la duplicació caracteritzada (**Figura 29**).

Per altra banda, la consulta en la base de dades epigenètiques WashU Epigenome Browser (http://epigenomegateway.wustl.edu) de les coordenades genòmiques finals de la duplicació del cromosoma 1 va revelar que aquesta regió podria estar afectant dos dominis associats topològicament (TAD, de l'anglès *topologically associated domain*) (**Figura 30**).

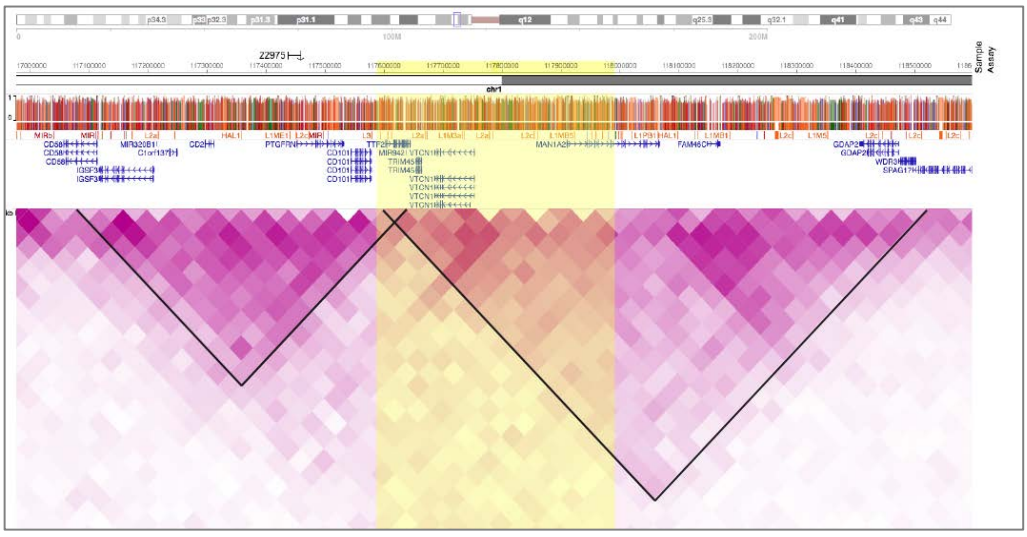


Figura 30. Contactes genòmics en la regió genòmica de la duplicació del cromosoma 1.
Mitjançant la consulta de dades d'interaccions genòmiques de HiC-seq s'identifica que la regió de la duplicació del cromosoma 1 estaria involucrant dos dominis associats topològicament. (Extret de http://epigenomegateway.wustl.edu/browser/))

Caracterització molecular dels gens implicats en la duplicació

Per tal de caracteritzar els efectes funcionals de la duplicació identificada, el següent pas fou estudiar les conseqüències moleculars d'aquesta variant a sobre dels gens implicats en aquesta regió. Primer de tot, es va voler monitoritzar els nivells d'expressió gènica en mostres de RNA provinents de sang d'un individu portador de la duplicació (II2) i un individu no-portador (I2). Els resultats de la matriu d'estudi de tot el genoma mostraren una clara sobreexpressió del gen *TTF2* a l'individu portador en comparació a les mostres no-portadores de la variant

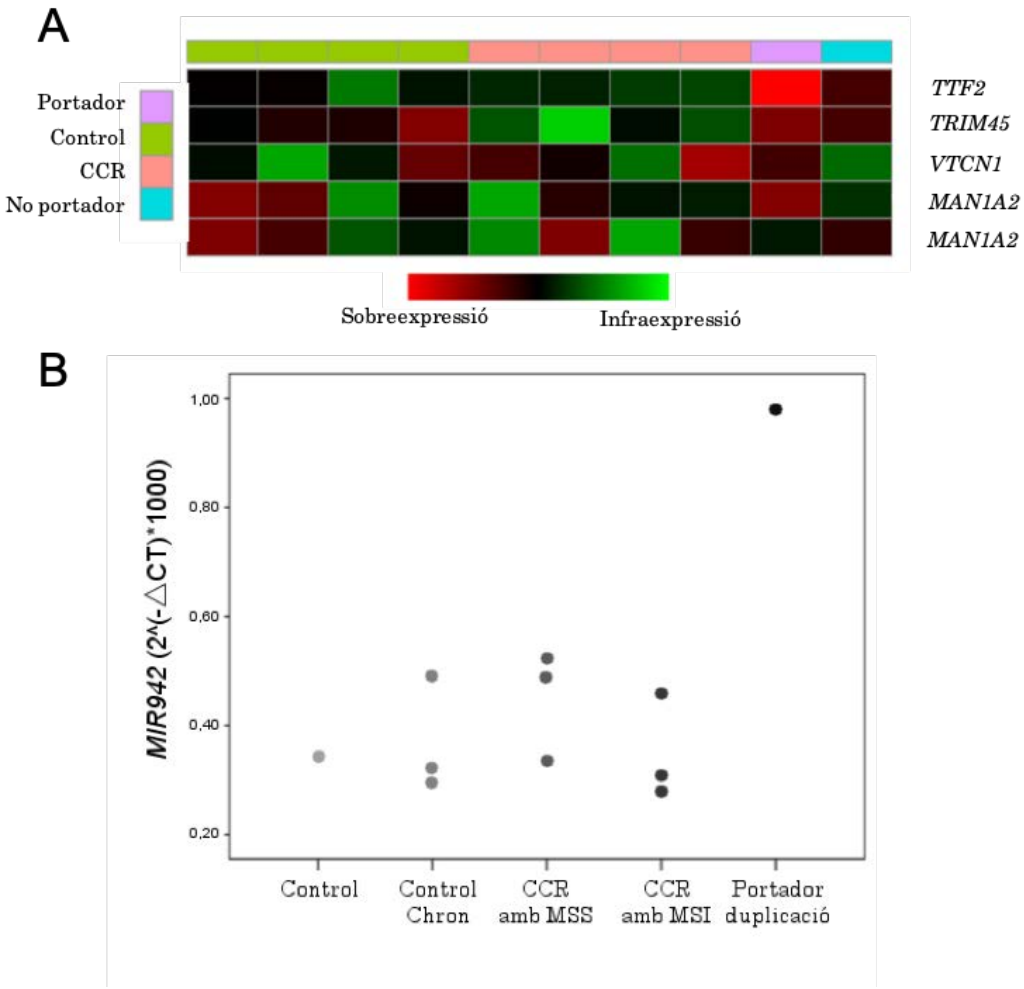


Figura 31. Resultats dels estudis d'expressió gènica.
(A) Heatmap dels resultats de l'anàlisi de les dades de la matriu d'expressió en mostres sanguínies d'un dels individus portadors de la duplicació 1 (portador), individu de la família 7 no portadora de la duplicació (no portador), i altres individus control amb distints fenotips. (B) Valors d'expressió de l'experiment de RT-qPCR per a dinstints individus, entre ells, un dels individus portadors de la duplicació del cromosoma 1. (CCR: individus amb CCR somàtic; Control: individus sans).

(individu I2 -amb CCR esporàdic-, mostres de CCR no-portadores i mostres control no-portadores) (**Figura 31A**). Contràriament, els demés gens implicats en la variant (*TRIM45*, *VTCN1* i *MAN1A2*) no mostraren expressió diferencial entre la mostra portadora i les mostres no-portadores. Aquesta sobreexpressió de *TTF2* es confirmà utilitzant la tècnica de RT-qPCR (**Figura 31B**). A més, en aquest cas també es pogué comprovar que els nivells d'expressió del miRNA situat a l'interior de l'intró 18 del gen *TTF2*, el *MIR942*, eren alts en la mostra de teixit tumoral de l'individu portador de la duplicació en comparació als nivells d'expressió mostres no-portadores de la duplicació.

Provada la sobreexpressió del gen *MIR942*, es va consultar la base de dades TARGETSCAN per extreure'n la predicció de dianes gèniques del miRNA. Per altra banda, es seleccionaren els gens diferencialment infraexpressats resultat de la comparació entre mostres d'individus portadors i no-portadors de la duplicació en l'estudi d'expressió gènica de tot el genoma. El llistat de gens predits com a potencials dianes del miRNA-942 es creuà amb el llistat dels gens infraexpressats en la sang dels portadors de la variant. D'aquest creuament s'obtingueren vuit gens (**Taula 13**) i, d'entre aquests, el gen *TMEM158* (també anomenat *RIS1*) destacà de forma important, degut a que estudis anteriors havien postulat el gen com a diana de la via mutagènica en carcinogènesis del CCR (Iglesias et al., 2006). A més, la implicació del gen *TMEM158* en la senescència cel·lular induïda per Ras i la seva localització genòmica en la citobanda 3p21.3, una petita regió del cromosoma 3 anomenada CER1 (de l'anglès, *common eliminated region 1*), semblarien recolzar el potencial paper d'aquest gen com a supressor tumoral (Hesson, Cooper, & Latif, 2007).

Estudi dels nivells proteics

Amb la finalitat de corroborar la regulació a la baixa de *TMEM158* i validar la regulació a l'alça de *TTF2* es dugueren a terme experiments d'immunohistoquímica en mostres de teixit colònic controls i de teixit tumoral d'un dels portadors de la duplicació del cromosoma 1 (**Figura 32**). Els resultats presentaren tincions elevades per a la proteïna TTF2 en les mostres de l'individu portador, tant en teixit de la paret muscular fina del còlon, com en teixit tumoral, quan es comparava amb les mostres de l'individu control, indicant una major presència de la proteïna. En el cas de TMEM158, els resultats de la immunohistoquímica evidenciaren una reducció dels nivells

d'aquesta proteïna en les mostres de teixit de l'individu portador de la duplicació, quan es comparaven amb els talls de teixit control de mucosa normal.

Taula 13. Gens diana predits per a *MIR942* i diferencialment infra-expressats al pacient portador de la duplicació del cromosoma 1.

Gen	Nom del gen	Posició genòmica	Funció gènica
SLC22A15	*solute carrier family 22 member 15*	1p13.1	Transport de components a nivell fisiològic, com toxines, hormones, neurotransmissors i metabòlits cel·lulars.
LMOD1	*leiomodin 1*	1q32.1	Increments en l'expressió de *LMOD1* podrien estar relacionats amb la síndrome de Graves, malaltia oftalmològica associada a deficiència de la tiroide.
LRRC18	*leucine rich repeat containing 18*	10q11.23	(Sense caracterització)
KCNMA1	*potassium calcium-activated channel subfamily M alpha 1*	10q22.3	Els canals "MaxiK" poden formar-se a partir de dues sub-unitats: la subunitat codificada pel gen *KCNMA1* i la versió "beta" de la mateixa subunitat.
TEPP	*testis, prostate and placenta expressed*	16q21	L'expressió selectiva del gen *TEPP* als testicles, pròstata i la placenta, així com la seva conservació en distintes espècies indiquen una possible implicació en l'aparell reproductiu.
TMEM158	*transmembrane protein 158*	3p21.31	La transcripció d'aquest gen es veu induïda com a resposta de la via d'activació Ras, mentre que no ho fa en altres circumstàncies de senescència cel·lular.
LARP1	*ribonucleoprotein domain family member1*	5q33.2	La subfamília proteica LARP desplega una important funció en la transcripció i maduració de les unitats mRNA.
TNXB	*tenascin XB*	6p21.33-p21.32	Les unitats tenascines aporten activitat anti-adhesiva, contràriament a les fibronectines. Potencialment implicades en la maduració de la matriu en processos de curació de ferides. La seva deficiència està associada a la síndrome teixit connectiu anomenat Ehlers-Danlos.

Prediccions dels gens diana per a *MIR942* extretes de la base de dades TARGETSCAN (http://www.targetscan.org)

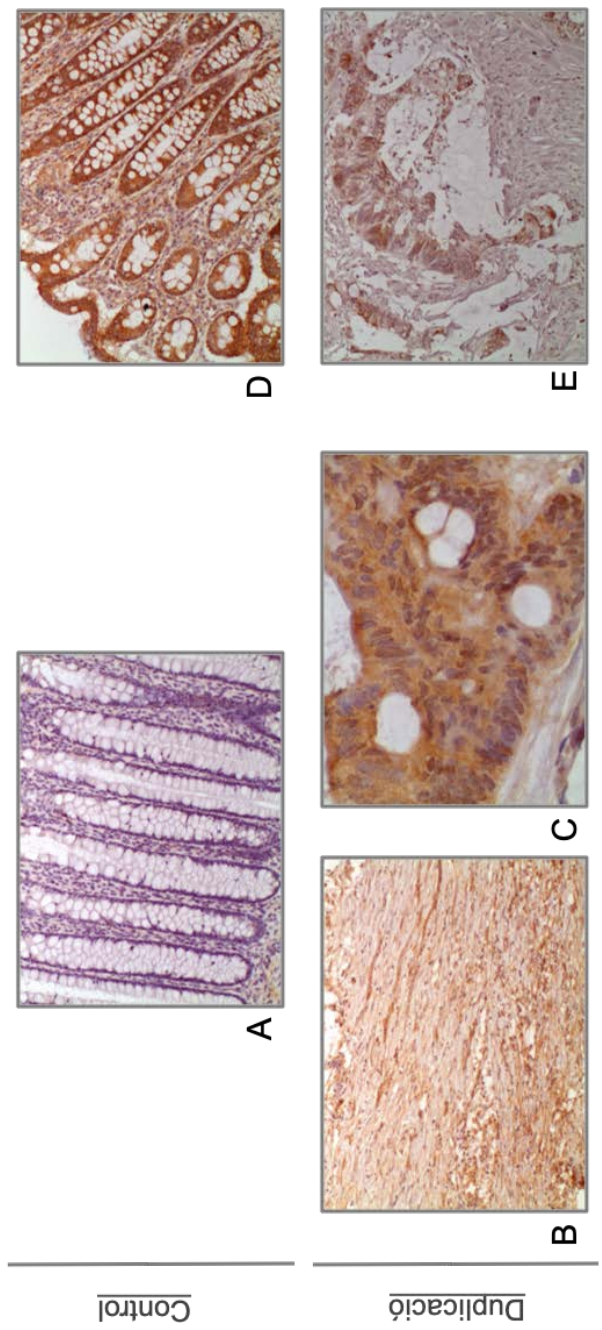


Figura 32. Resultats dels estudis proteòics per immunohistoquímica per a TTF2 i TMEM158.
S'observa un augment de la tinció (marró) per a la proteina TTF2 en les mostres tissulars del portador de la duplicació, tant en la paret muscular colònica normal (B) com en la secció tumoral (C), quan es comparen a la imatge del teixit control (A). Per altra banda, s'observa una disminució de la tinció de la proteina TMEM158 en el tall tumoral del pacient portador (E), comparat amb la tinció del teixit normal (D).

Estudi 2

Els resultats descrits a continuació corresponen al desenvolupament de l'eina bioinformàtica CNApp i la seva aplicació en dades genòmiques públiques.

Una versió preliminar de l'article on s'explica l'eina i la seva implementació es troba disponible al repositori web bioRxiv amb el títol *CNApp: a web-based tool for integrative analysis of genomic copy number alterations in cancer*.

(doi: https://doi.org/10.1101/479667).

Esquema, implementació i flux d'anàlisi

CNApp és una eina bioinformàtica d'accés lliure (http://bioinfo.ciberehd.org/CNApp). A més, l'eina es pot descarregar per al seu ús local des del compte personal de GitHub (https://github.com/ait5/CNApp), on es troben les instruccions necessàries per a la seva instal·lació, així com tots els paquets i codis de programació implementats en ella.

L'estructura de CNApp consta de tres grans seccions: *Re-Seg&Score*, on es realitza una re-segmentació opcional de les dades, es quantifiquen les càrregues d'alteracions focals o àmplies (*broad*, en anglès), a més de la càrrega de CNAs global, i s'estudien associacions entre aquestes quantificacions i lesvariables d'anotació per mostra; *Region profile*, on es transformen les dades re-segmentades en perfils genòmics complets utilitzant finestres genòmiques definides per l'usuari, afavorint la comparació i correlació entre les mostres i, per altra banda, s'estudien les regions descriptives per a grups de mostres especificats per les variables d'anotació; i *Classifier model*, on s'apliquen models de *machine-learning* per a predir noves classificacions entre les mostres mitjançant múltiples variables (**Figura 33**).

Tot i que les distintes parts del CNApp es troben connectades, permetent que els resultats d'una s'utilitzin a la següent, cada una d'aquetes parts pot ésser utilitzada de forma individual, explotant les dades originals carregades per l'usuari. CNApp aporta resultats visuals en forma de gràfiques interactives descarregables en alta qualitat per a la seva utilització en publicacions. A més, alguns dels resultats també es poden descarregar en format tabulat per a anàlisis més extensius.

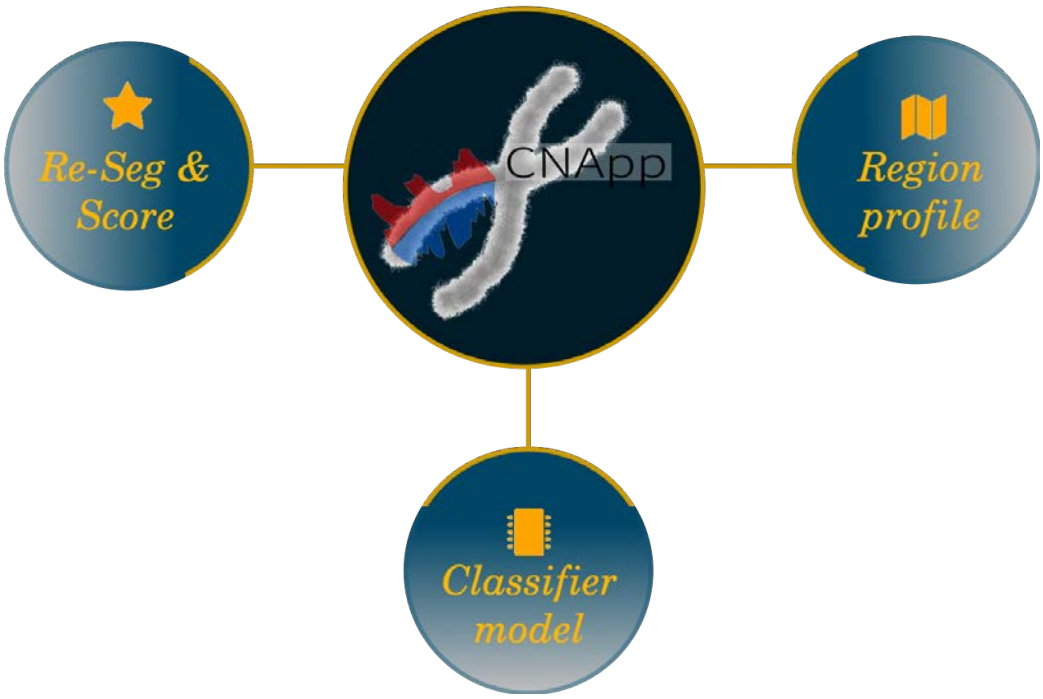


Figura 33. Il·lustració esquemàtica de les seccions del CNApp.
L'aplicació desenvolupada consta de tres seccions principals: *Re-Seg&Score*, *Region profile* i *Classifier model*.

Re-Seg&Score: re-segmentació, càlcul dels *CNA scores* i associació de variables

La secció *Re-Seg&Score* consta de diverses funcionalitats: re-segmentació opcional dels segments genòmics; càlcul de puntuacions -*CNA scores*- per a la quantificació de la càrrega en quant a alteracions focals, d'ampli espectre (*broad*) o a nivell global per mostra; i l'associació estadística de les variables d'anotació amb els distints *CNA scores*.

Re-segmentació

La re-segmentació opcional de les dades genòmiques va encaminada a corregir possibles diferències de base tècnica als perfils genòmics de les mostres (per exemple, mostres sobre-segmentades) i, d'aquesta manera, aconseguir una caracterització dels perfils de CNAs més acurada. Els paràmetres que s'utilitzen per a dur a terme aquesta re-segmentació són els següents (amb els valors per defecte especificats entre parèntesis): longitud mínima dels segments (100 Kb), màxima distància entre els segments a re-segmentar (1 Mb), desviació màxima d'amplitud de canvi -és a dir, valors de *seg.mean*- entre dos segments a re-segmentar (0,16), desviació mínima d'amplitud de canvi entre un segment i l'eix zero central (0,16), i màxima desviació del valor de BAF entre dos segments (0,1). Aquests paràmetres poden ésser manipulats per part de l'usuari per tal de realitzar una re-segmentació personalitzada. A més, si s'introdueixen valors de puresa per a cada mostra, el programa és capaç de corregir els valors d'amplitud del canvi dels segments genòmics (*seg.mean*) de forma depenent d'aquesta puresa. Per contra, si no s'aporten els valors de puresa i/o de BAF, CNApp només considera els valors de canvi (*seg.mean*) a l'hora de calcular la re-segmentació.

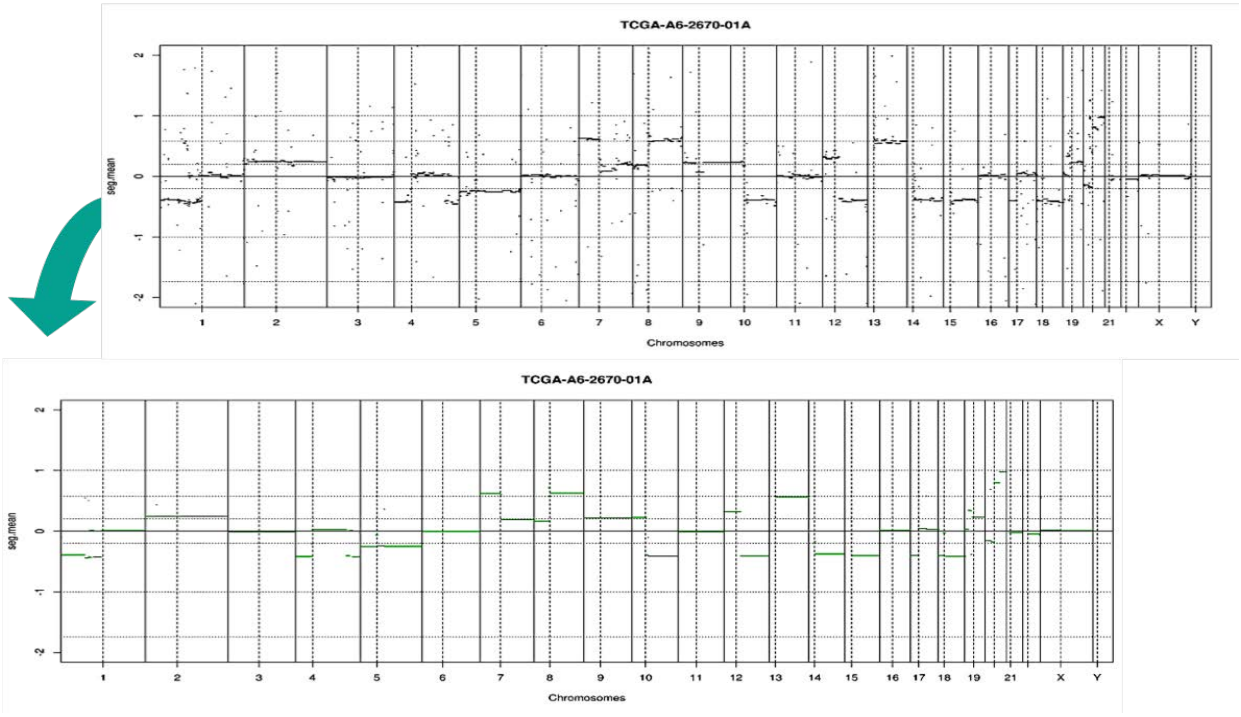


Figura 34. Exemple de re-segmentació de perfil genòmic de CNAs. La sobre-segmentació de la mostra del TCGA (*Affymetrix 6.0 SNP array*) es veu corregida, en gran mesura, gràcies al procés de re-segmentació de la secció *Re-Seg & Score*. Perfil original en la imatge superior (segments genòmics de color negre); perfil re-segmentat en la imatge inferior (nous segments en verd).

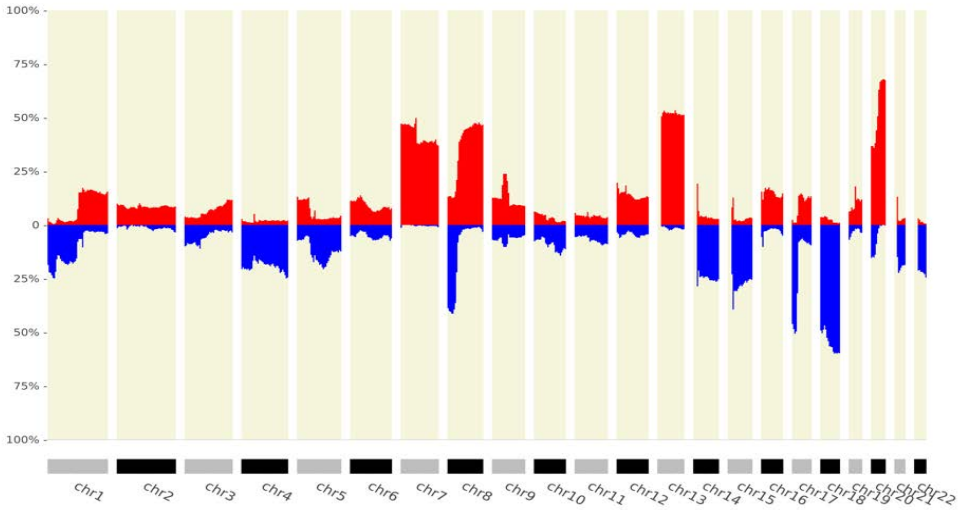


Figura 35. Perfil de regions alterades recurrents de la cohort de COAD del TCGA.
CNApp utilitza els segments genòmics introduïts per l'usuari o els resultants de la re-segmentació per a calcular la freqüència en que les regions apareixen alterades. En la imatge apareixen les regions recurrentment alterades per a la cohort del TCGA de COAD (*colon adenocarcinoma*).

Els resultats de la re-segmentació es poden descarregar en format tabulat. A nivell visual, l'aplicació genera imatges amb els perfils de segments genòmics abans i després de la re-segmentació per a cada mostra (**Figura 34**). A més, en el mateix apartat dels resultats de la re-segmentació, es poden calcular les freqüències de les mínimes regions genòmiques compartides entre les mostres de l'usuari (ja sigui entre totes les mostres o entre grups definits per les variables d'anotació indexades) i generar perfils gràfics amb les freqüències de les alteracions recurrents en tot el genoma (**Figura 35**).

Càlcul dels *CNA Scores*: FCS, BCS i GCS

Les dades de segmentació genòmica resultat de la re-segmentació -o, les dades de segmentació original si l'usuari ha saltat aquest procés- s'utilitzen per a quantificar la càrrega de CNAs en cada una de les mostres. Així, es calculen tres puntuacions distintes o *CNA scores* per a les alteracions focals, les *broad* i a nivell global: FCS, BCS i GCS, respectivament.

Els valors de BCS, FCS i GCS, juntament amb les variables d'anotació per a cada mostra, es poden descarregar en format tabulat. Visualment, l'aplicació genera gràfiques de *boxplots* per a representar la

distribució dels *CNA scores*, ja sigui a nivell global (entre totes les mostres), o a nivell de variable. En el cas de distingir entre grups de mostres, s'avaluen les possibles diferències significatives entre les distribucions dels valors dels *CNA scores* entre els grups de mostres aplicant la prova estadística T de Student.

Associacions estadístiques amb variables d'anotació

Calculats els valors de FCS, BCS i GCS per mostra, CNApp realitza testos estadístics entre aquests i les variables d'anotació introduïdes per l'usuari. Tenint en compte el format de cada una de les variables (si són numèriques o categòriques, si es tracten de variables paramètriques o no paramètriques), l'eina aplica distints tipus d'assaigs estadístics per valorar la relació entre les càrregues d'alteracions de número de còpia i les característiques clíniques i moleculars indexades (**Taula 10**). Per tal d'avaluar aquestes relacions, CNApp reprodueix els P-valors estadístics entre els distints valors.

Region profile: perfils de regions genòmiques

Region profile aprofita les dades de segments genòmics per a generar perfils complets del genoma definits per finestres genòmiques preestablertes, per tal de facilitar la comparació directa entre mostres. En cas que s'hagi dut a terme la re-segmentació de *Re-Seg&Score*, l'aplicació permet seleccionar només aquells segments classificats com *broad* (per tant, els etiquetats com cromosòmics -*chromosomal*- o de nivell de braços -*arm-level*-), o aquells només classificats com focals; calculant les mitjanes per finestra només amb el tipus de segments seleccionats i permetent encaminar estudis de CNAs més específics. Els llindars d'amplitud de canvi definits pels valors de guany i pèrdua de baix grau (0,2 i -0,2) s'utilitzen per a classificar aquestes mitjanes com a guanys o pèrdues, de la mateixa manera que per a calcular les freqüències per a cada finestra entre les mostres (aquests llindars poden ésser modificats per part de l'usuari).

Per tal de representar aquests perfils genòmics per regions, CNApp mostra un *heatmap* de colors -vermell per als guanys i blau per a les pèrdues-, acompanyat de gràfiques d'anotació per a les variables clíniques i/o moleculars escollides, i on es poden incloure els *CNA scores* si ja han estat calculats. En aquesta secció també existeix la possibilitat de realitzar una anàlisi de

clusterització d'aquests perfils genòmics. A més, l'eina quantifica el nombre de regions guanyades o perdudes per tal d'avaluar quines d'aquestes regions són les més recurrents entre les mostres i, per altra banda, calcular el nombre de regions perdudes o guanyades en cada mostra; tot presentant-ho mitjançant gràfiques de barres d'acumulació. També es poden obtenir *heatmaps* de correlació de les mostres i de clusterització d'aquestes mateixes correlacions.

Regions genòmiques descriptives

Per altra banda, CNApp presenta una altra funcionalitat per esbrinar regions genòmiques específiques relacionades amb grups de mostres depenent de variables d'anotació. Aprofitant els perfils genòmics de regions calculats i la facilitat de comparar mostra a mostra aquests perfils, l'eina avalua quines regions presenten valors diferencialment alterats entre grups de mostres (definits per les variables d'anotació de l'usuari) per tal d'esbrinar l'especificitat de certes regions a l'hora de diferenciar aquests grups. L'eina representa l'anàlisi mitjançant un *heatmap* amb corba de colors segons el P-valor estadístic en la regió, indicant el grau de significança en què aquella regió es troba diferencialment alterada entre dos grups de mostres comparats. Aquest estudi es fa de dues maneres: una d'elles avalua directament els valors de la regió entre dos grups de mostres (T-test d'Student), mentre que l'altra quantifica els guanys i les pèrdues d'aquella regió i avalua aquestes quantificacions entre els dos grups (test de Fisher). Es poden representar gràfiques de caixes (*boxplots*, en anglès) i de barres, per a una de les regions seleccionades, en el primer tipus d'anàlisi o en el segon, respectivament. A més, els gens inclosos en aquesta regió seleccionada poden ésser consultats i descarregats en format tabulat.

Classifier model: prediccions de models de classificació

La secció *Classifier model* permet a l'usuari de generar models classificatoris basats en algoritmes d'aprenentatge automàtic (en anglès, *machine learning*) mitjançant la selecció d'una variable que defineixi grups de mostres i una altra (o múltiples) de la qual es vol estudiar si té capacitat per a classificar eficientment les mostres.

Per a dur a terme això, CNApp aplica diverses funcions del paquet d'R randomForest (Barradas et al., 2002; Hesson et al., 2007). La construcció dels models es realitza 50 vegades, canviat el conjunt de dades d'entrenament en cada iteració i de forma aleatòria. Posteriorment, s'extreuen freqüències de les prediccions per mostra. Per defecte, només es poden utilitzar variables d'anotació de les dades carregades per l'usuari, tant com variables per a la definició de grups o com variables de classificació. En cas de que les seccions *Re-Seg&Score* i/o *Region profile* ja s'hagin implementat, les variables generades per CNApp en aquestes seccions també es poden utilitzar per a la generació de models de predicció (com les regions genòmiques o els *CNA scores*). Valors de precisió, sensibilitat i especificitat per a l'avaluació de la capacitat dels models generats són presentats com a resultat. Els resultats de les prediccions s'ofereixen en format de taula descarregable i es generen gràfics de barres expressant les diferències entre les mostres reals i les mostres predites.

Caracterització genòmica de 10.635 mostres de tumor

Per tal de valorar la capacitat i l'aplicabilitat de l'eina bioinformàtica CNApp es varen analitzar les 10.635 mostres tumorals disponibles al repositori web del projecte *The Cancer Genome Atlas* (TCGA), que incloïen 33 tipus de càncers primaris. Els perfils de CNAs per aquestes mostres, corresponents a dades de la plataforma genòmica Affymetrix SNP 6.0 array i analitzades mitjançant el paquet d'R DNAcopy, es varen descarregar i analitzar aplicant les seccions *Re-Seg&Score* i *Region profile* del CNApp, amb tots els paràmetres per defecte.

Valors dels *CNA scores* i correlació amb la fracció alterada del genoma

Els valors de *CNA scores* obtinguts d'aquesta anàlisi per a cada una de les mostres varen ésser correlacionats amb els corresponents valors de fracció del genoma alterat, per tal de provar la seva fiabilitat a l'hora d'avaluar la càrrega de CNAs que presenten les mostres. Així, els valors de BCS [0 – 44] i FCS [5 – 2.466] demostraren altíssims nivells de correlació amb les fraccions alterades del genoma per alteracions d'ampli rang (*broad*) i focals, respectivament. Concretament, la correlació per al BCS fou del 0,957 i del 0,938 per al FCS (**Figura 36A-B**). De la mateixa manera, valors de GCS també correlacionaren de forma important amb la fracció alterada global del genoma, presentant un valor de 0,963 de correlació (**Figura 36C**).

Per altra banda, els valors de *CNA scores* es correlacionaren entre ells per intentar esbrinar quin tipus de relació presentaven i, alhora, elucidar alguna dinàmica entre els tipus d'alteracions que aquests *CNA scores* representen. Així, el valor de la correlació entre BCS i FCS fou del 0,59, de 0,90 entre BCS i GCS, i de 0,85 entre FCS i GCS. A més, es va avaluar de manera més extensiva i específica la correlació entre BCS i FCS mitjançant una correlació per rangs de valors de BCS, centrant rangs de 10 valors per a cada valor únic de BCS. D'aquesta manera, es va observar que aquelles mostres presentant valors baixos de BCS presentaven valors de correlació positiva alts entre alteracions *broad* i focals, mentre que mostres amb valors més baixos de BCS perdien aquesta correlació (**Figura 37**).

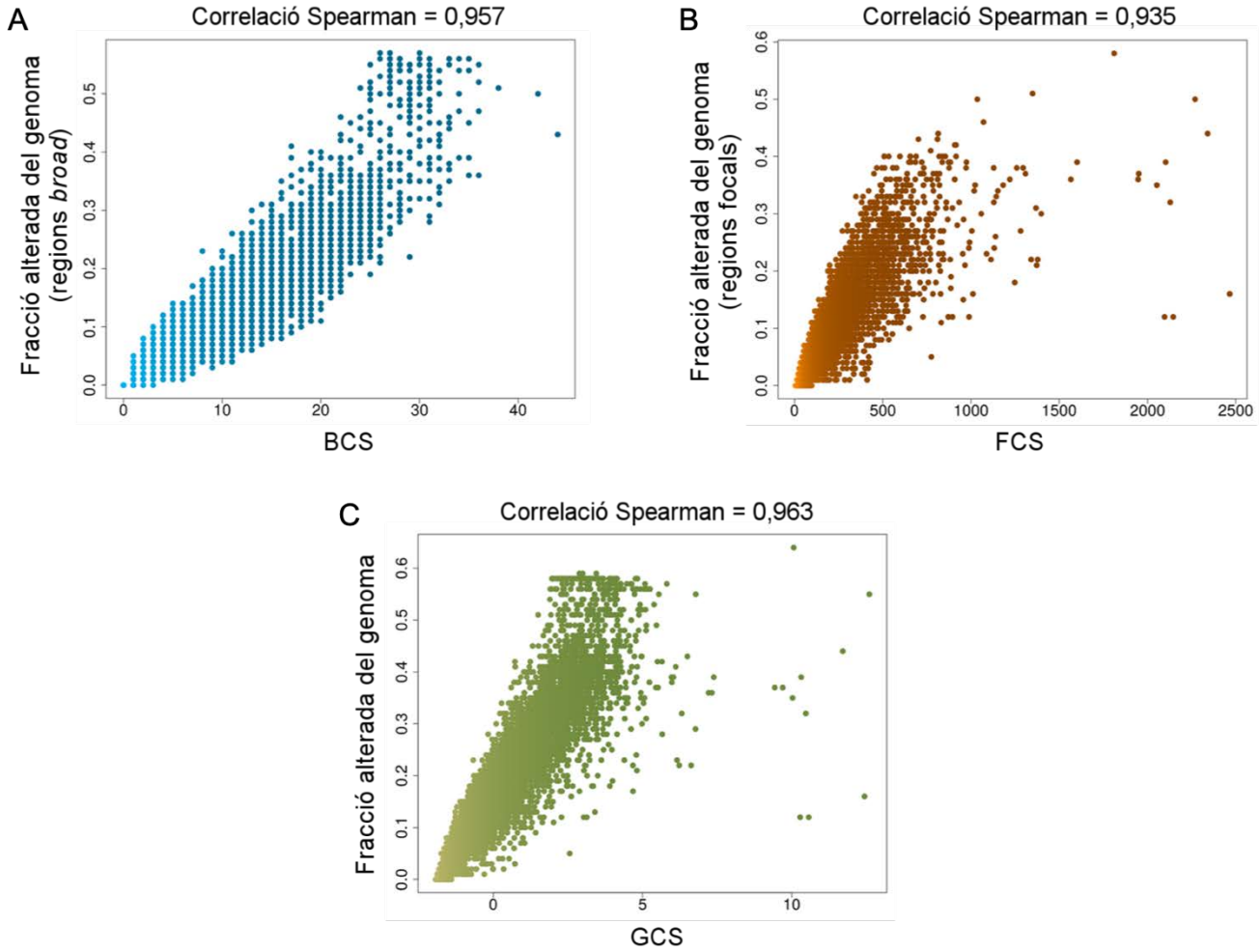


Figura 36. Correlació dels valors de CNA scores i la fracció alterada del genoma de les 10.635 mostres del Pan-cancer TCGA.
El test de correlació per rangs de Spearman va ser aplicat per tal d'evaluar la correlació entre els valors de *Broad CNA Score* (BCS) i la fracció alterada del genoma afectada per CNAs d'ampli rang (*broad – chromosomal* i *arm-level* -) (A); els valors de *Focal CNA score* (FCS) i la fracció alterada del genoma afectada per CNAs focals (B); i els valors del *Global CNA score* (GCS) i la fracció alterada global del genoma (considerant tots els tipus d'alteracions) per a cada mostra.

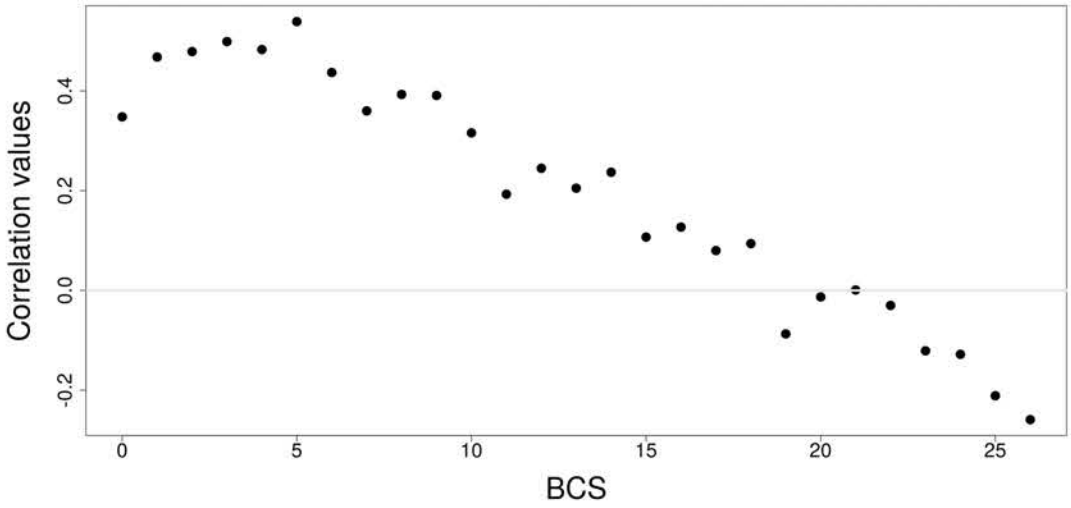


Figura 37. Valors de correlació entre BCS i FCS.
Per a cada valor x [0-26] es correlacionaren els valors de BCS de les mostres amb BCS = x ± 5 amb els seus respectius valors de FCS.

Caracterització genòmica dels subtipus de càncer

Els distints tipus de càncer presentaren distints nivells per als valors de BCS, FCS i GCS (**Figura 38**). Alguns d'aquests tipus de càncer mostraren nivells molts baixos en quant a càrregues globals de CNAs, com el cas de la leucèmia mieloide aguda (LAML, en anglès *acute myeloid leukemia*), el carcinoma de tiroides (THCA –*thyroid carcinoma*–) o el timoma (THYM –*thymoma*–) amb valors de mitjana per al GCS de -1,67 per LAML, -1,68 per THCA i -1,52 per THYM. Contràriament, el carcinosarcoma uterí (UCS –*uterine carcinosarcoma*–), el càncer d'ovari (OV –*ovarian serous cystadenocarcinoma*–) i el carcinoma cel·lular esquamós de pulmó (LUSC –*lung squamous cell carcinoma*–) presentaren els nivells més alts d'alteracions *broad* i focals, amb valors de mitjana per al GCS de 2,55 per UCS, 2,44 per OV i 0,97 per LUSC. Altres tipus de càncer mostraren preferència de càrrega de CNAs específics, ja fossin alteracions *broad* o focals. Per exemple, el càncer de ronyó cromòfob (KICH –*kidney chromophobe*–) mostrà la mitjana de BCS més alta (BCS = 27), mentre que el valor de mitjana per a FCS fou de 49. De forma contrària, les mostres de càncer de mama (BRCA –*breast carcinoma*–) presentaren valors molt alts de càrrega d'alteracions focals (mitjana de FCS de 150), però el seu valor de mitjana per al BCS fou de 7.

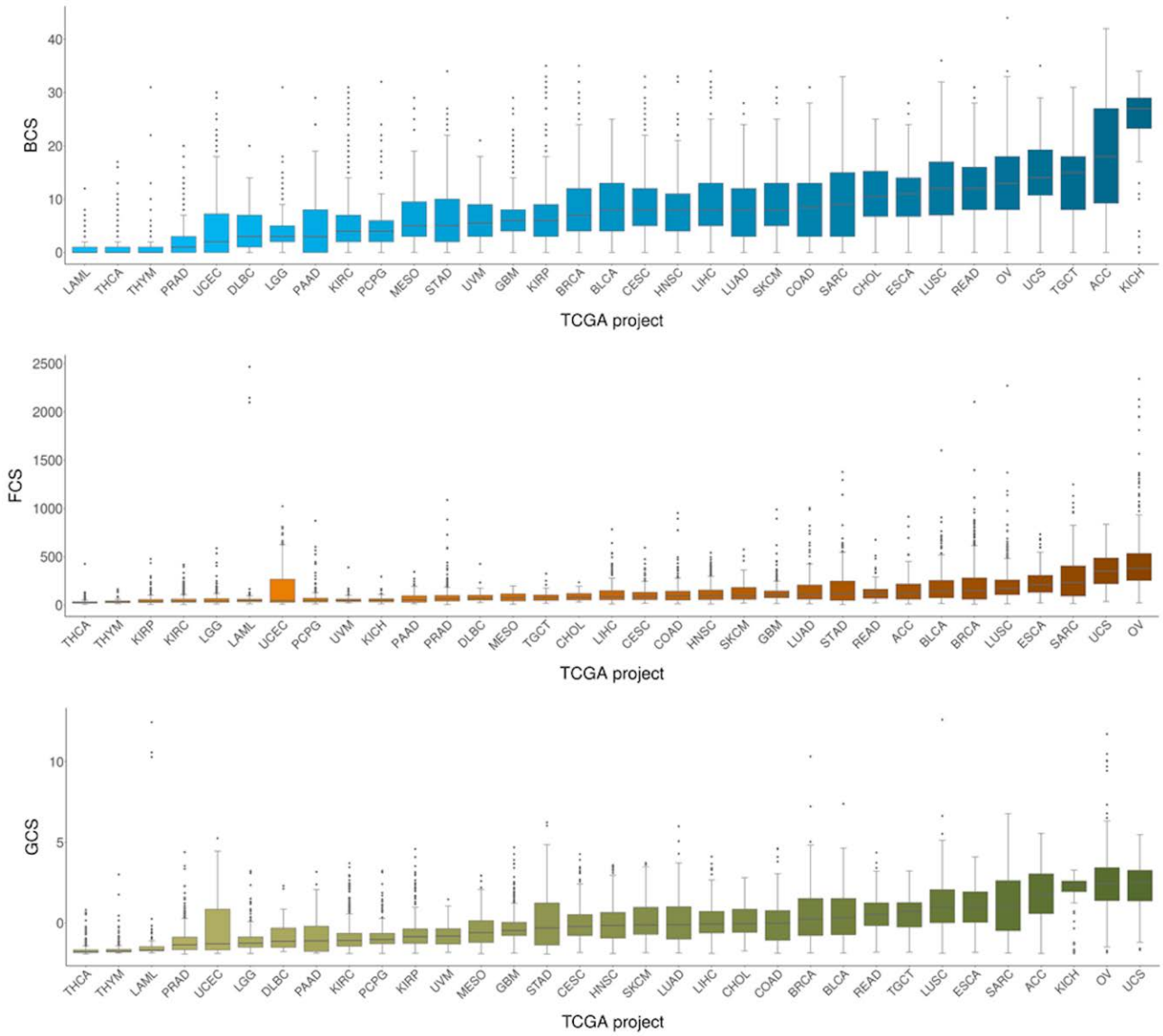


Figura 38. Distribució dels valors dels *CNA scores* (BCS, FCS i GCS) entre els 33 tipus de càncer del TCGA.
Es representen, mitjançant gràfiques de caixes, la distribució dels valors de BCS, FCS i GCS entre els 33 tipus de càncer per a les 10.635 mostres de la cohort del TCGA. Els distints tipus de càncer estan ordenats segons el valor de la mediana del valor de BCS. Les caixes representen la distribució entre el percentil 25 i 75; la línia mitja representa el valor de la mediana i els percentils 5 i 95 s'indiquen per les petites línies al final de la distribució del tipus de càncer.

En el següent pas, s'analitzaren els patrons d'espectre genòmic per cada un dels tipus de càncer utilitzant els braços cromosòmics com a finestres genòmiques en la secció *Region profile*, visualitzats mitjançant el *heatmap* de la **Figura 39**, i s'extragueren les freqüències de guanys i pèrdues corresponents en aquestes regions (**Figura 40**). S'observà que els braços de cromosoma alterats en més del 25% del total de mostres foren els braços 1q, 7p, 7q, 8q i 20q en quant a guanys, i 8p i 17p en quant a pèrdues de número de còpia. Contràriament, els braços cromosòmics afectats per CNAs en menys del 10% de mostres entre tots els tipus de càncer foren el 2q i el 19p.

Els *CNA scores* es representaren com a variables d'anotació per a visualitzar la càrrega de CNAs en els distints grups de mostres corresponents als tipus de càncer. Alguns patrons destaquen per la seva especificitat com, per exemple, el cas del GBM (*glioblastoma multiforme*), amb el guany del cromosoma 7 i la pèrdua del cromosoma 10; o el perfils de KICH i OV, altament aneuploïdies, en contraposició als tumors endocrí THCA i l'hematològic LAML, amb baixa densitat de CNAs.

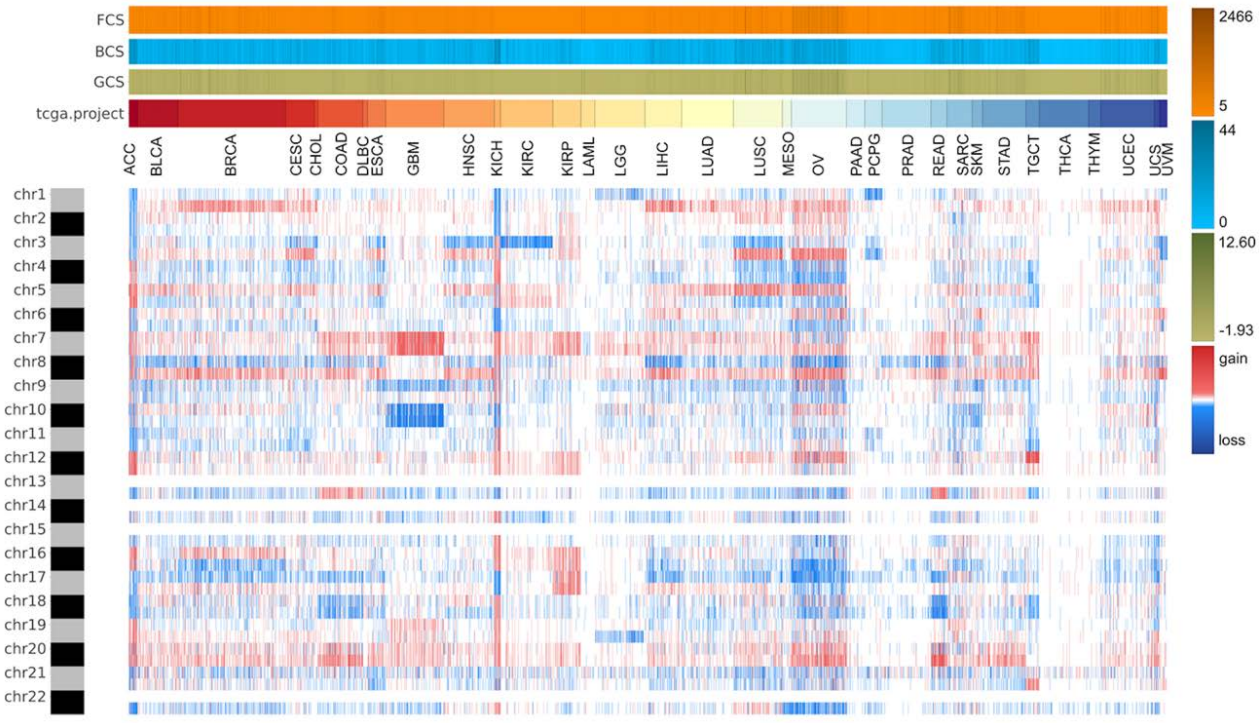


Figura 39. Heatmap dels perfils genòmics, per braços cromosòmics, que identifica els patrons característics de CNAs per tipus de càncer.
Els perfils genòmics generats en la secció *Region profile* del CNApp es poden observar mitjançant el heatmap principal que genera l'aplicació. S'identifiquen els patrons específics de CNAs a nivell genòmic i cromosòmic (eix y) per als distints tipus de càncer ordenant les mostres (eix X) segons aquests tipus de càncer. Els *CNA scores* (BCS, FCS i GCS) per mostra, així com la respectiva cohort del TCGA a la que corresponen, es troben il·lustrades en la part superior del heatmap.

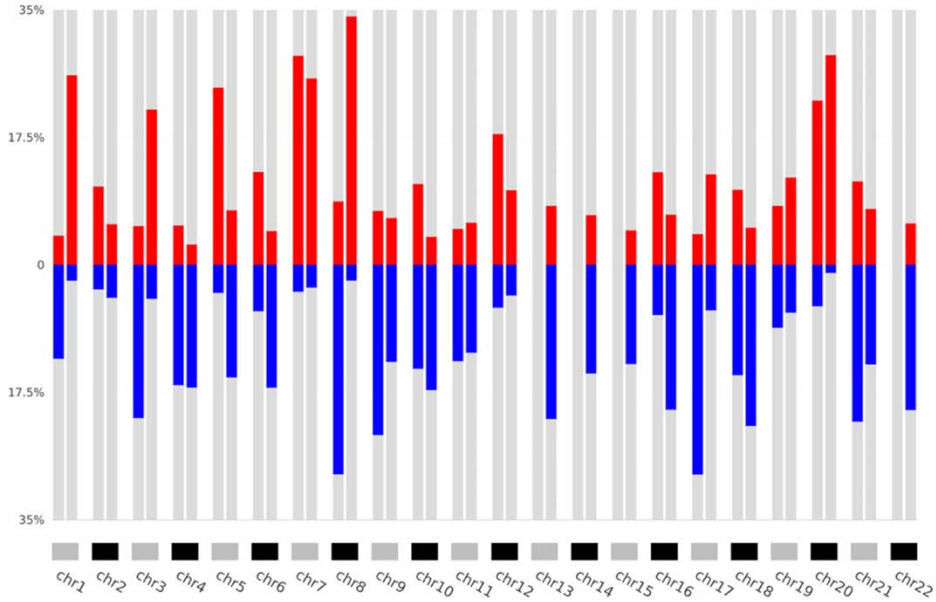


Figura 40. Freqüència d'alteració dels braços cromosòmics en la cohort de pan-cancer del TCGA.
Estudiant els perfils genòmics generats en la secció *Region profile* es calculen les freqüències (en forma de percentatge) en que cada regió genòmica es troba alterada (guany o pèrdua) en la cohort de mostres que s'estudia. El gràfic de barres acumulades expressa el percentatge de casos en que aquella regió apareix com a pèrdua (blau) o en forma de guany (vermell).

Clusterització per tipus tumoral

Utilitzant un sub-conjunt de mostres que incloïa 20 tipus de càncer dels quals es tenia informació del tipus tumoral (gastrointestinal, escamós o ginecològic), es calcularen les mitjanes dels valors de regions genòmiques dels braços cromosòmics per a cada un dels distints càncers que formaven els tipus tumorals coneguts, utilitzant, altra vegada, la secció *Region profile* del CNApp (**Figura 41**). Aquesta anàlisi va mostrar que els valors de correlació entre els distints perfils dels tipus tumorals clusteritzaven d'acord amb el tipus tumoral associat (**Figura 42**). Específicament, els tumors gastrointestinals (còlon, recte, estómac i pàncrees), ginecològics (ovari i úter) i escamosos (cèrvix, de cap i coll, i pulmó) formaren clústers ben definits per a cada un d'aquests grups, replicant estudis anteriors encaminats en aquesta direcció (Hoadley et al., 2018; Taylor et al., 2018).

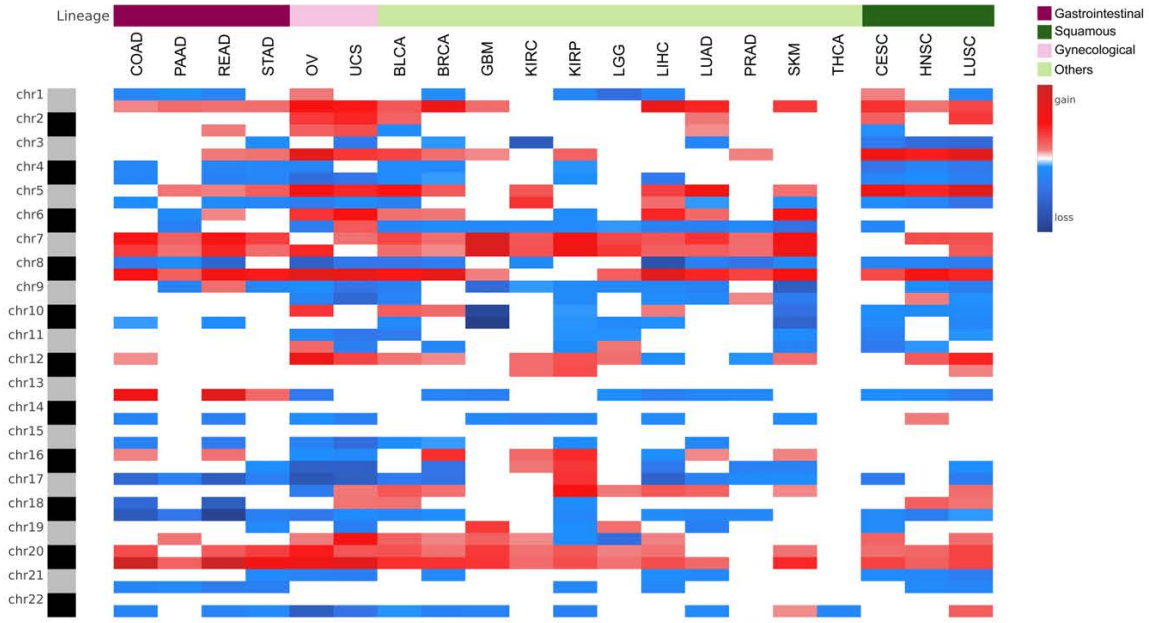


Figura 41. Perfils mitjans per als 20 tipus de càncers per a la posterior clusterització segons l'orígen tumoral.
Es calcularen els parefils mitjos per als diferents tipus de càncer mitjançant CNApp.

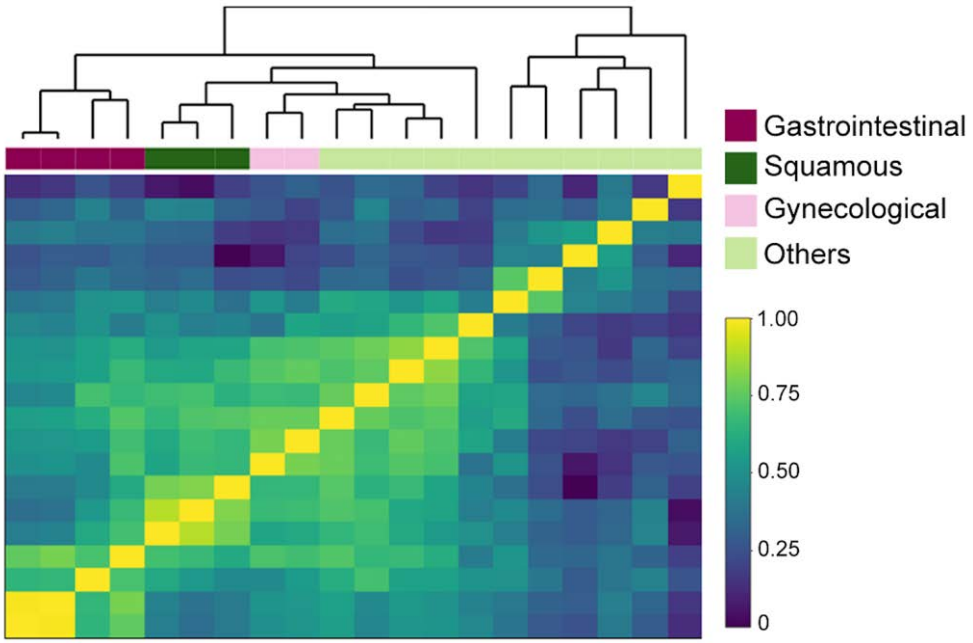


Figura 42. Heatmap de clusterització entre els orígens tumorals.
Els valors de correlació (test de correlació Pearson) entre els perfils mitjans de les distintes cohorts del TCGA per les que es va anotar l'orígen tumoral (anotació extreta del treball orginal de Taylor *et.al. Cell* 2018) clusteritzen entre les distintes anotacions (Gastrointestinal: còlon, recte, estómac i pàncrees; Escamós: cèrvix, cap i coll i pulmó; Ginecològic: ovari i úter; i altres)

Classificació del càncer de còlon per *CNA scores* i regions genòmiques

La classificació molecular del CCR més actualitzada i acceptada proposa una classificació taxonòmica d'aquest que inclou quatre subgrups moleculars consensuats (o CMS, per la seva nomenclatura en anglès: *consensus molecular subtypes*) (Guinney et al., 2015). Així, cada CMS es descriu per uns perfils moleculars definits per la presència de diferents característiques, entre elles: la inestabilitat de microsatèl·lits (en anglès *microsatellite instability*, MSI), nivells del fenotip metilador d'illes CpG (CIMP), CNAs somàtiques i mutacions no-sinònimes (aquelles que provoquen el canvi d'aminoàcid en la traducció del RNA a la seqüència proteica). Com ja s'ha explicat, el CMS1 inclou la majoria de tumors híper-mutats, mostrant MSI, alts nivells de CIMP i una càrrega de CNAs baixa; el CMS2 i CMS4 es componen dels tumors amb estabilitat de microsatèl·lits (en anglès *microsatellite stability*, o MSS) i alts nivells de CNAs; finalment, pel que fa al CMS3, aquest es compon d'una mescla de tumors MSI i MSS, amb nivells mitjans de CNAs i CIMP alta (**Taula 1**).

Mitjançant l'aplicació CNApp, s'analitzà un conjunt de 309 mostres de càncer de còlon extretes del projecte *TCGA Colon Adenocarcionma* (COAD), de les quals es coneixia la seva classificació entre els grups CMS (el nombre de mostres corresponent a cada grup s'indica a la **Taula 6**), així com també l'estat dels microsatèl·lits i la mutació en el gen *BRAF* per a cada mostra. S'aplicà el procés de re-segmentació estàndard, tot i que eliminant aquells segments genòmics més petits de 500 Kb per evitar incloure possibles errades tècniques.

Caracterització genòmica i alteracions recurrents

Es generaren els perfils genòmics de canvi del número de còpia utilitzant les finestres genòmiques per braços cromosòmics, i els patrons resultants per a cada grup de CMS es visualitzaren al *heatmap* de la **Figura 43**. També s'estudiaren els perfils genòmics de CNAs focals mitjançant la selecció de les finestres de sub-citobandes obtenint el *heatmap* corresponent (**Figura 44**). Per altra banda, es calcularen les freqüències d'alteració dels braços cromosòmics globals i per a cada subgrup CMS (**Figures 45 i 46**) dels patrons per mostra i els perfils de freqüència per a cada subgrup de CMS, i el mateix per a les regions de sub-citobandes (**Figures 47 i 48**).

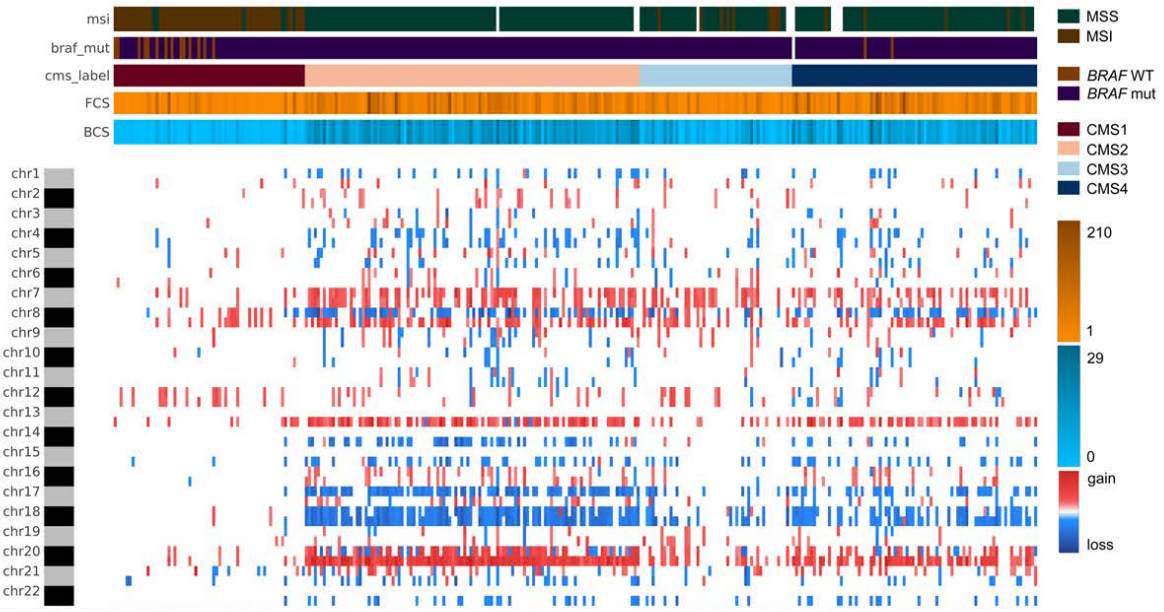


Figura 43. Heatmap dels perfils d'alteracions en braços cromosòmics per mostra i ordenades segons el subgrup de CMS.
Es visualitza perfil genòmic, definit pels braços cromosòmics, per a cada mostra. Es representen, en la zona superior, les variables d'anotació BCS, FCS, subtipus de CMS (CMS1/CMS2/CMS3/CMS4), mutació BRAF (mostra no-mutada [BRAF WT] / mostra mutada [BRAF mut]) i estat dels microsatèl·lits (MSS/MSI) per a cada mostra.

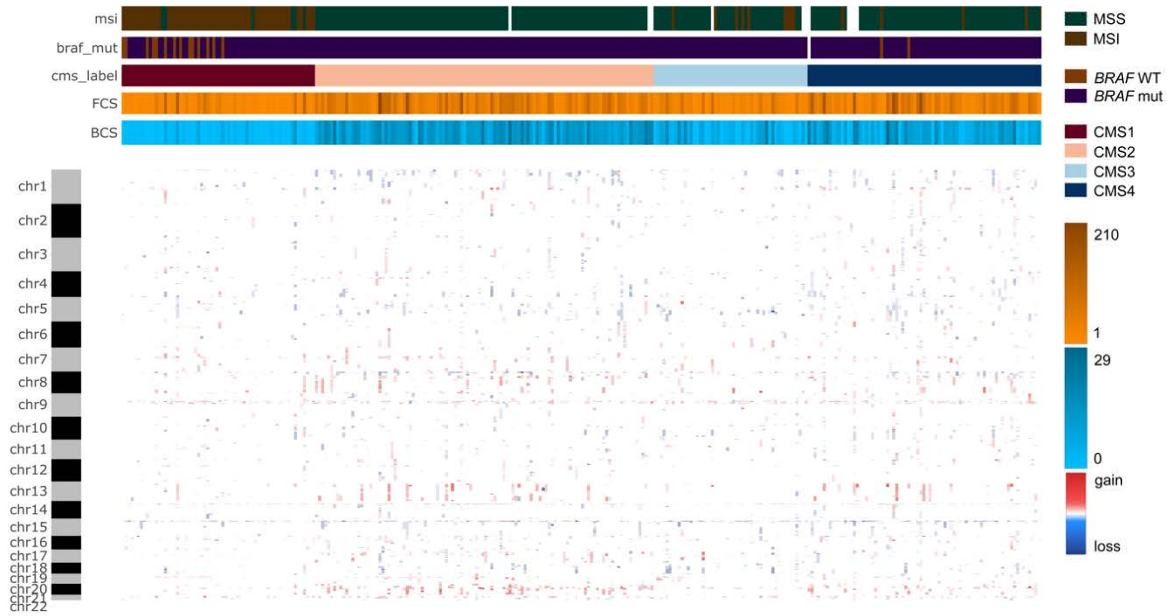


Figura 44. Heatmap dels perfils d'alteracions en sub-citobandes per mostra ordenada segons el subgrup de CMS.
Es visualitza perfil genòmic, definit les sub-citobandes genòmiques, per a cada mostra. Es representen, en la zona superior, les variables d'anotació BCS, FCS, subtipus de CMS (CMS1/CMS2/CMS3/CMS4), mutació BRAF (mostra no-mutada [BRAF WT] / mostra mutada [BRAF mut]) i estat dels microsatèl·lits (MSS/MSI) per a cada mostra.

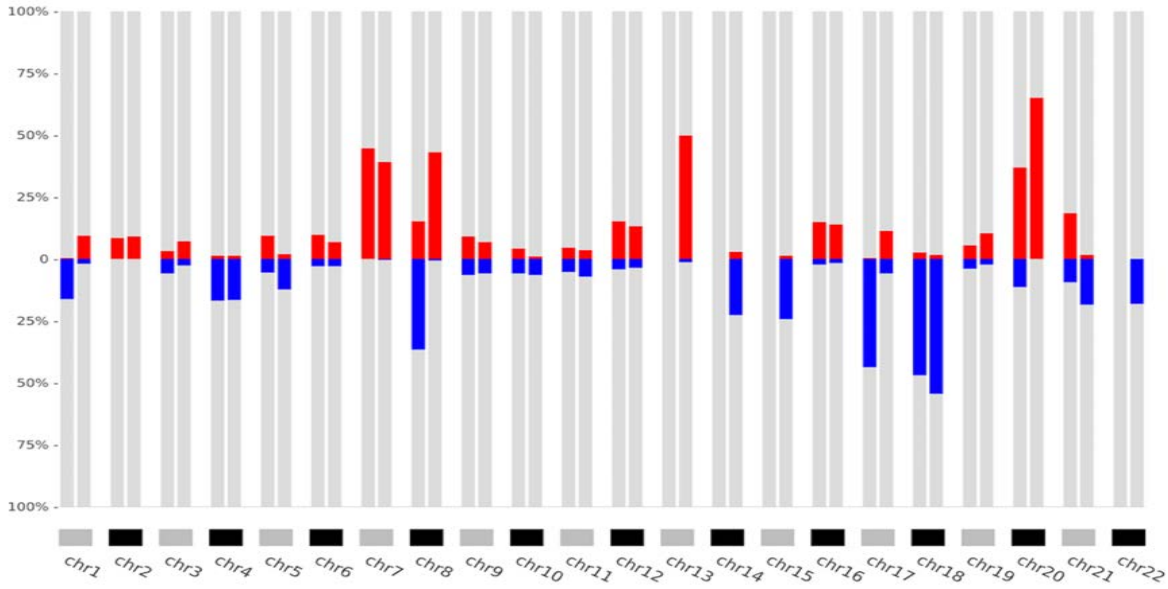
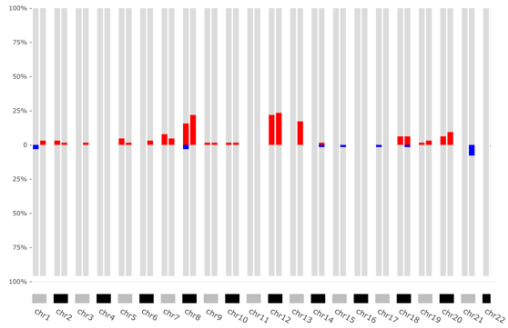
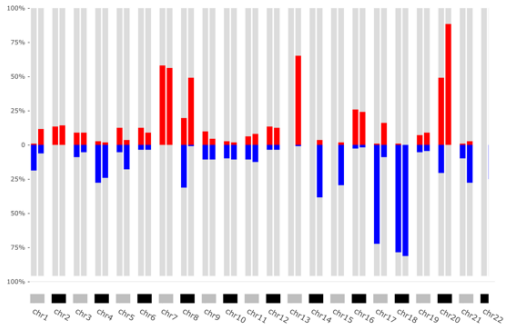


Figura 45. Percentatges d'alteracions de braços cromosòmics a la cohort de CMS.
Les regions genòmiques més alterades en la cohort de mostres del CMS, delimitades pels braços cromosòmics, foren 7p, 7q, 8q, 13q, 20p i 20q, pel que fa a guanys, i 8p, 17p, 18p i 18q, per a les pèrdues.

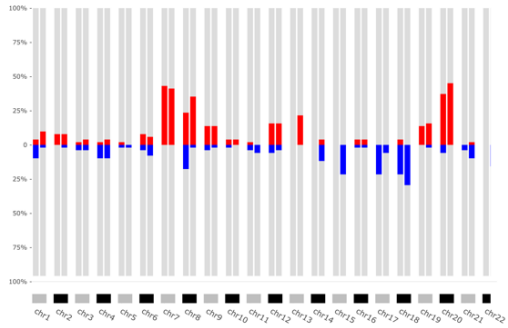
CMS1



CMS2



CMS3



CMS4

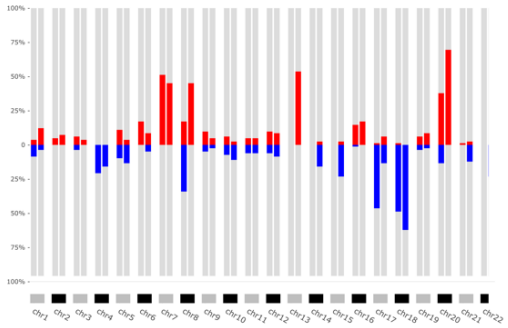


Figura 46. Perfils genòmics per braços cromosòmics per als grups CMS de les CNAs *broad*.

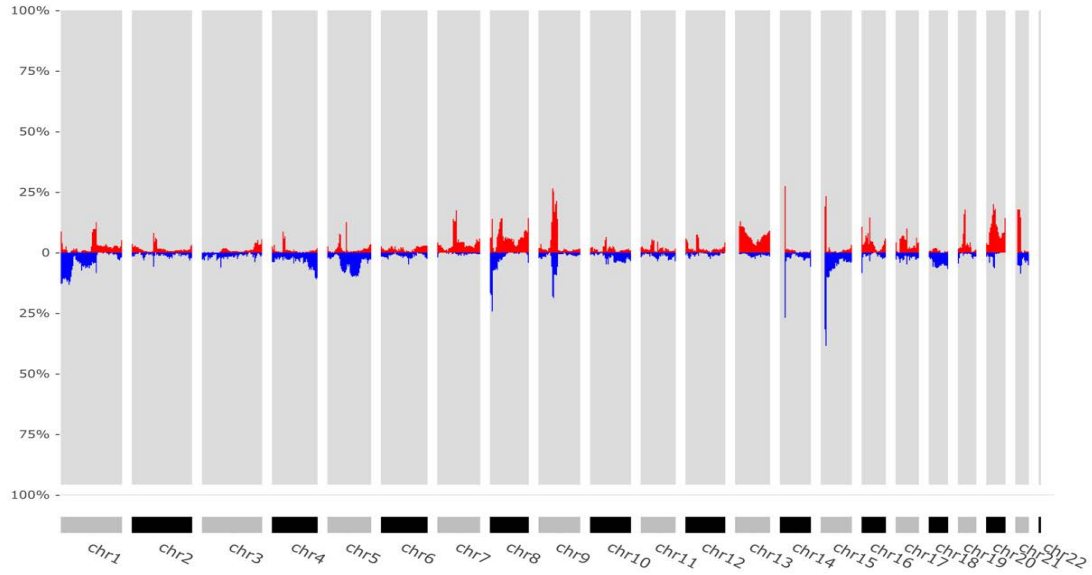
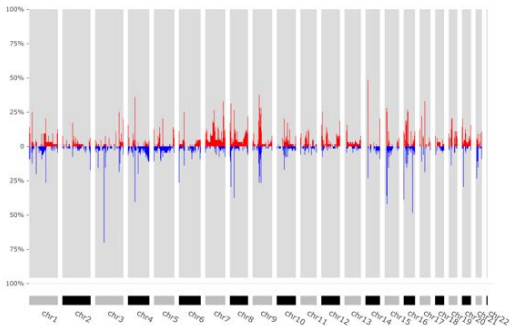
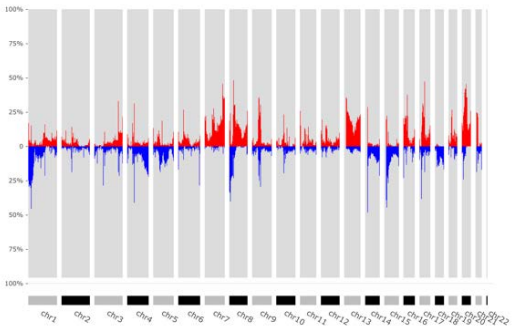


Figura 47. Perfil de freqüència global en les regions de sub-citobandes per a la cohort de CMS analitzant les CNAs focal.

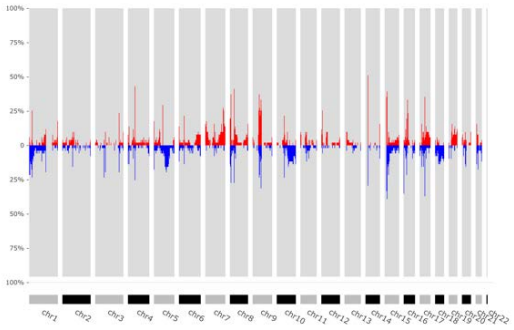
CMS1



CMS2



CMS3



CMS4

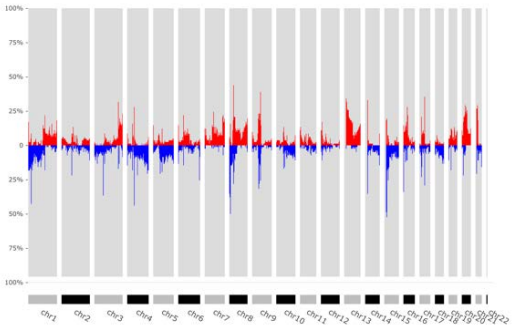


Figura 48. Perfils genòmics per sub-citobandesper als grups CMS considerant les CNAs focals.

Per que fa a l'anàlisi per braços cromosòmics, els guanys amb una freqüència d'alteració de més del 30 foren el 7p, 7q, 8q, 13q, 20p i 20q, mentre que les pèrdues amb més freqüència s'identificaren en els braços 8p, 17p, 18p i 18q (**Figura 43 i 45**), reproduint les regions genòmiques alterades més comuns associades al CCR esporàdic (Ried et al., 1996; Meijer et al., 1998; Douglas et al., 2004; Nakao et al., 2004). Per altra banda, d'entre les regions alterades obtingudes en l'anàlisi de les alteracions focals per sub-citobandes, cinc de les sis pèrdues genòmiques i cinc de les 18 alteracions en forma de guany coincidien amb les alteracions identificades mitjançant l'eina GISTIC2.0 en l'estudi del TCGA per al conjunt de mostres del projecte COAD (**Figura 44 i 47**) (The Cancer Genome Atlas, 2012).

Integració de les alteracions i variables d'anotació

CNApp ens va permetre realitzar un estudi d'integració dels perfils genòmics de CNAs amb les variables de classificació de CMS, l'estabilitat de microsatèl·lits i els *CNA scores*. D'aquesta manera, es va voler comprovar si aquests *CNA scores* tenien capacitat de classificar les mostres de càncer de còlon segons el seu subgrup CMS.

Els valors de BCS establiren diferències significatives entre els quatre grups de CMS (valoracions estadístiques mitjançant el T-test d'Student, $P \leq 0{,}0001$), mentre que els valors de FCS no establien diferències entre el CMS1 i CMS3, ni entre CMS2 i CMS4 (**Figura 49**). Per altra banda, la integració dels valors de BCS amb l'estabilitat de microsatèl·lits mostrà valors mitjans de BCS de 1,51±2,11 al grup de mostres amb MSI (N = 72) i 10,25±5,92 a les mostres amb MSS (N = 225) (**Figura 50**). A més, aplicant el valor de BCS = 4, corresponent al valor del percentil 90 en les mostres MSI, es va re-classificar les mostres de l'estudi: 186 de 225 (83%) de les mostres amb MSS mostraren valors de BCS > 4 i 39 mostres (17%) presentaren valors de BCS ≤ 4, corresponent a tres tumors de CMS1, sis de CMS2, 18 de CMS3 i 12 de CMS4 (**Taula 14**). De forma adjacent, i considerant el valor de FCS del percentil 90 en els tumors MSS (FCS = 37,2), es va determinar que vuit de les 39 mostres MSS amb BCS ≤ 4 presentaven valors de FCS > 37,2, reduint el percentatge de tumors MSS amb baixos nivells de CNAs al 13%.De forma interessant, set tumors MSI constaven amb valors de BCS > 4. Entre ells, cinc presentaven alteracions genòmiques típicament associades a la via clàssica del CCR, incloent amplificacions en la regió del gen *MYC*. Pel que fa als 51 tumors MSI classificats com CMS3 del nostre conjunt de mostres, dos d'aquests presentaven delecions focals en el cromosoma 2, afectant als gens *MSH2* i *MSH6*. A l'hora, el 43% dels tumors CMS3 amb MSS mostraven valors de BCS < 4.

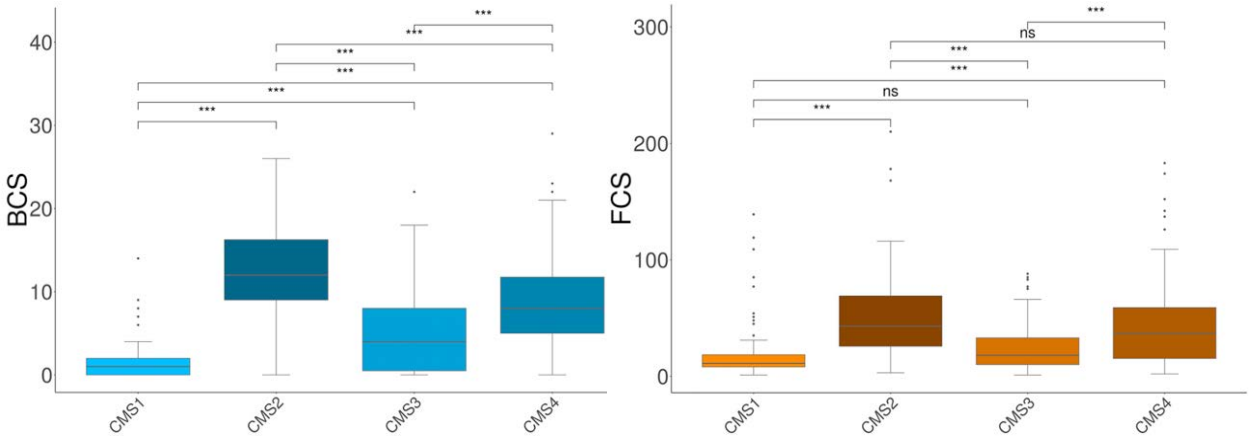


Figura 49. Distribució dels valors BCS i FCS en gràfica de caixes pels grups de CMS.
La distribució dels valors de BCS entre els distints grups de CMS fou significativament distinta entre totes les comparacions grup a grup, mentre que el FCS no fou capaç de diferenciar entre els grups CMS1 i CMS3, i CMS2 i CMS4. (Test d'Student: *P*-valor ≤ 0,001 (***); *P*-valor ≤ 0,01 (**); *P*-valor ≤ 0,05 (*); *P*-valor > 0,05 (ns)).

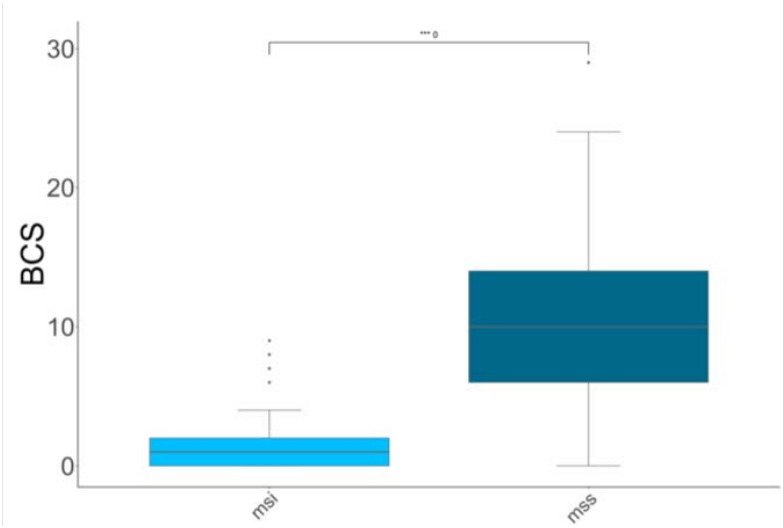


Figura 50. Distribució dels valors BCS entre les mostres MSI i MSS.
(Test d'Student: *P*-valor ≤ 0.001 (***); *P*-valor ≤ 0.01 (**); *P*-valor ≤ 0.05 (*); *P*-valor > 0.05 (ns)).

Taula 14. Resultats de la re-classificació de les mostres MSI/MSS segons el valor BCS=4.

	Total	BCS ≤ 4	BCS > 4
MSI	72	66 (*92%*)	6 (*8%*)
MSS	225	39 (*17%*)	186 (*83%*)

MSI: microsatellite instability; MSS: microsatellite stability.

Per altra banda, centrant-se amb la variable d'anotació que definia les mostres mutades pel gen *BRAF*, el grup de CMS1 estava enriquit per aquest tipus de mostres, mentre que dues mostres del CMS4 també es caracteritzaven per la mutació en aquest gen (**Figura 51**). Una d'aquests mostres presentava una càrrega alta d'alteracions *broad* (BCS = 11), contrastant amb l'altra mostra CMS4 i *BRAF*-mutada, que presentava MSI i un valor BCS igual a zero. De forma similar, quatre mostres no-mutades en *BRAF* i classificades com CMS4, presentaven MSI i BCS = 0.

Mitjançant la funció *Descriptive regions* de la secció *Region profile* del CNApp, s'estudiaren les regions genòmiques diferencialment alterades, en aquest cas els braços cromosòmics, entre els distints grups CMS. Valorant els canvis numèrics d'aquestes regions genòmiques (aplicant el T-test d'Student amb *P*-valor ajustat $P \leq 0{,}005$), s'observà que els grups CMS1 i CMS3 no presentaven regions descriptives entre ambdós, a banda del cromosoma 7q i la pèrdua del braç 18q (**Figura 52**). Les regions del 18q i el 20q destacaren com a aquelles amb més capacitat de diferenciació entre la majoria de grups: per una banda, els nivells de pèrdua del braç 18q diferenciaven entre tots els grups excepte CMS3 i CMS4, mentre que el cromosoma 20q diferenciava entre tots els CMS, però no era capaç de diferenciar entre el CMS1 i el CMS3 en aquest nivell de significança estadística, però sí es demostrava significativament diferent amb p-valor ajustat de $P \leq 0{,}05$ (**Figura 53**).

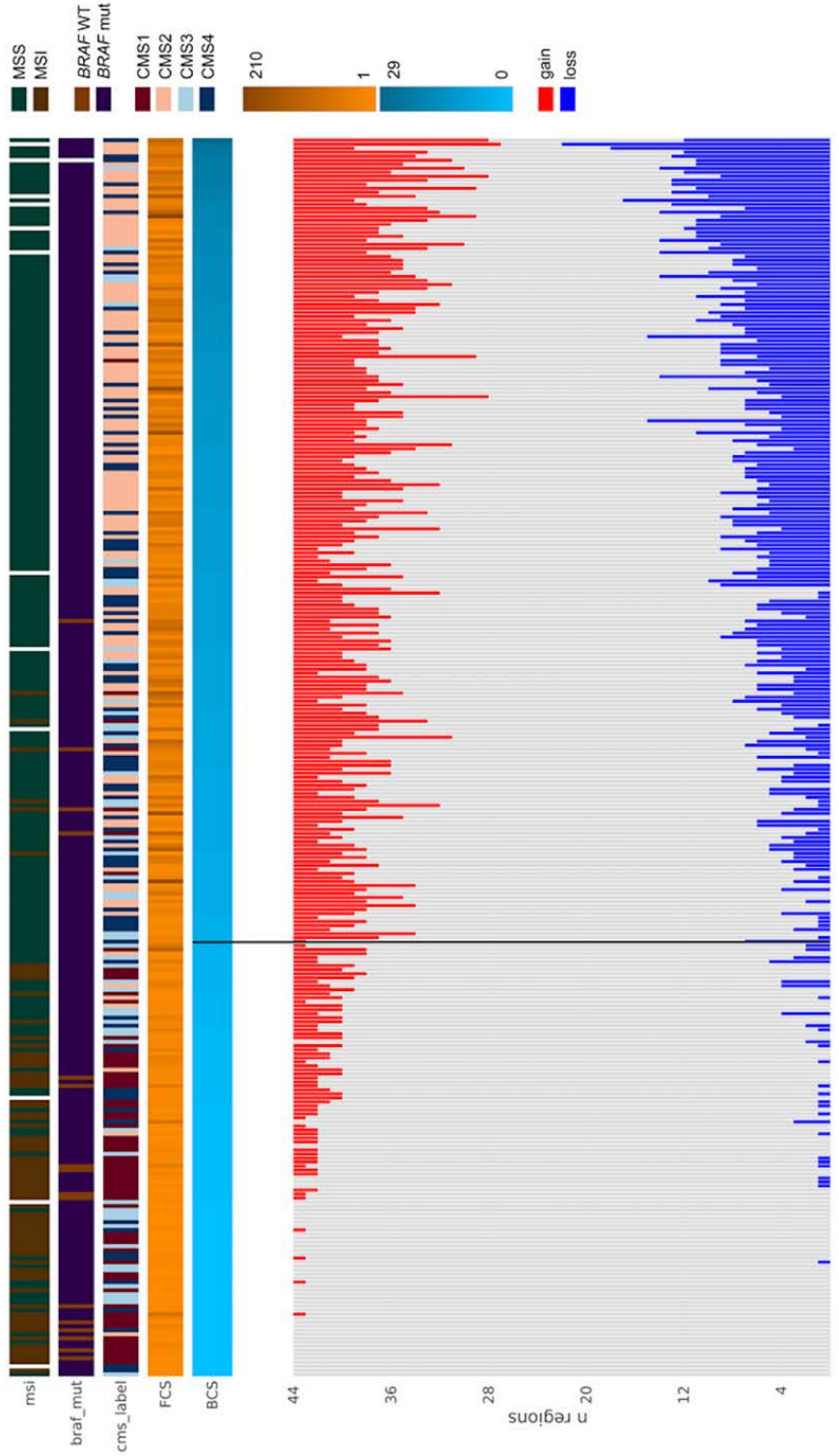


Figura 51. Visualització de guanys i pèrdues en les mostres de la cohort de CMS ordenades pel seu valor de BCS.
Es visualitza el nombre de regions amb guanyades (vermell/*gain*), perdudes (blau/*loss*) o sense canvi (gris) per a cada mostra. Es representen, en la zona superior, les variables d'anotació BCS, FCS, subtipus de CMS, mutació BRAF i estat dels microsatèl·lits (MSS/MSI) per a cada mostra. La línia fosca vertical representa el valor de BCS=4.

Aplicant la secció *Classifier model* de la nostra aplicació, es volgué estudiar la capacitat d'algunes de les variables de les que es disposava (*CNA scores*, regions de braços cromosòmics, regions genòmiques descriptives, etc.) a l'hora de classificar les mostres, tant a nivell d'inestabilitat de microsatèl·lits com per als grups de CMS (**Taula 15**). La variable BCS generada pel CNApp (encarregada de quantificar la càrrega d'alteracions *broad* de les mostres estudiades) fou testada per a classificar les mostres segons la variable MSI, demostrant una eficiència global del 82,2%. De forma consistent, el BCS també fou capaç de diferenciar les mostres dels grups CMS1 i CMS2, amb una eficiència del 88,40%, tot i que no demostrà una bona eficiència (47,60%) per a distingir entre els quatre grups de CMS alhora (**Taula 15**).

L'eficiència del braç 20q a l'hora de classificar les mostres entre els grups CMS tan sols assolí un 45,10%, tot i que la seva capacitat augmentava quan es tractava de classificar entre dos grups, essent el valor més baix el 56,90% a l'hora de diferenciar entre CMS3 i CMS4 (**Taula 15**). Posteriorment, es seleccionaren les regions genòmiques més descriptives entre els grups de CMS: braços 8p, 13q, 14q, 17p, 18p, 18q, i 20q (**Figura 52**). Aquestes regions assoliren una eficiència del 59.40% a l'hora de diferenciar entre els CMS, aproximadament la mateixa que s'aconseguia utilitzant tot el perfil genòmic de braços cromosòmics (**Taula 15**). De fet, quan s'utilitzaren les mateixes regions per a diferenciar entre els grups CMS per parelles, les eficiències foren altes. A destacar, les eficiències del 69,80% entre CMS2 i CMS4, 75,60% entre CMS1 i CMS3 i 95,90% entre CMS1 i CMS2.

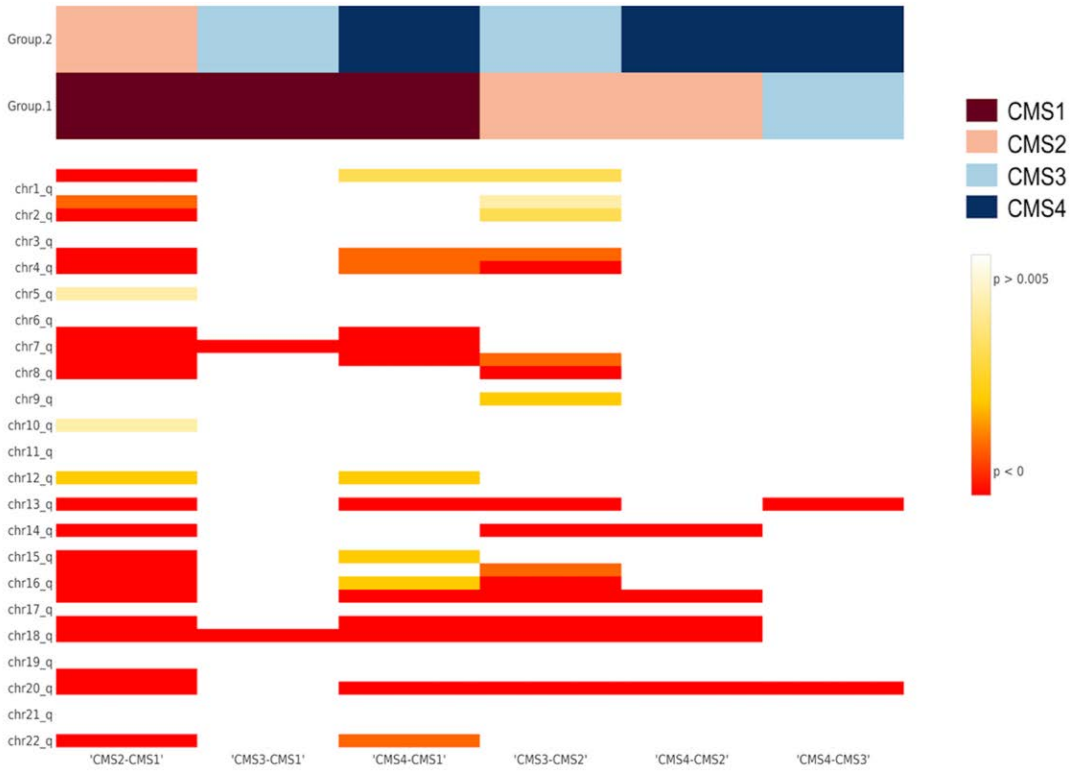


Figura 52. Heatmap de P-valors de les regions descriptives entre els grups de CMS.
Mitjançant l'aplicació del Test-t d'Student (P ≤ 0.005), s'avaluen la significança entre les distribucions de valors en les regions cromosòmiques en els distints grups de CMS. Aquelles regions genòmiques amb més diferències entre els grups s'identifiquen com a descriptives entre els grups de mostres.

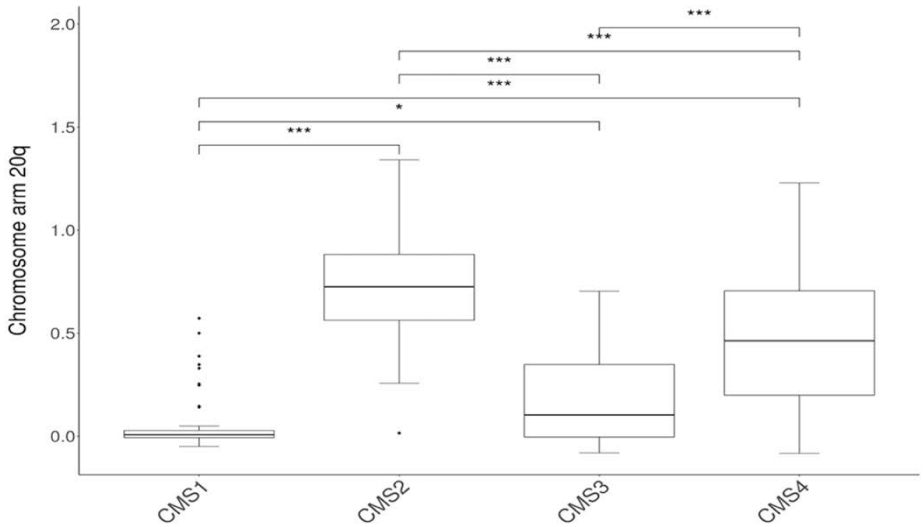


Figura 53. Distribució dels valors en la regió del cromosoma 20q als perfils genòmics per braços cromosòmics entre els grups de mostres dels CMS.
Els valors de la regió genòmica 20q és mostra significativament diferent entre els grups CMS (Test d'Student: *P*-valor ≤ 0,001 (***); *P*-valor ≤ 0,01 (**); *P*-valor ≤ 0,05 (*); *P*-valor > 0,05 (ns)).

Taula 15. Eficiències en les combinacions de variables i els models classificadors dels grups CMS.

	BCS	**FCS**	**BCS & FCS**	**All regions**	**Chr 20q**	***Descriptive regions*** 20q, 18q, 18p, 17p, 14q, 13q, 8p
CMS1-4	47,60	31,60	43,50	60,40	45,10	59,40
CMS1,2,3	71,10	48,20	66,30	83,70	71,90	81,90
CMS1,2,4	57,80	38,50	53,10	68,40	56,00	67,10
CMS1,3,4	51,60	40,10	51,30	69,40	53,10	67,80
CMS2,3,4	46,00	36,80	44,00	60,00	47,10	57,20
CMS1 vs CMS2	88,40	67,30	87,40	96,00	92,30	95,90
CMS1 vs CMS3	68,80	56,50	66,00	73,10	63,00	75,60
CMS1 vs CMS4	79,40	56,90	75,50	85,20	76,80	86,20
CMS2 vs CMS3	73,70	59,30	75,60	88,40	79,40	86,40
CMS2 vs CMS4	62,30	49,40	58,10	73,60	59,30	69,80
CMS3 vs CMS4	47,70	58,20	54,10	70,00	56,90	67,50

Anàlisi de les dades de seqüenciació de l'exoma amb CNApp

Per tal de comprovar l'aplicabilitat de l'eina CNApp utilitzant dades de segments genòmics resultat de plataformes NGS i la seva capacitat de replicar la identificació de la duplicació del cromosoma 1 del primer estudi , es varen analitzar les dades de seqüenciació d'aquest primer estudi (projecte FAMCOLON) on s'havien analitzat les dades de WES de DNA germinal en individus de famílies amb forta agregació per CCR familiar. En concret, es varen introduir els segments genòmics resultat de l'aplicació del paquet d'R ExomeDepth (**Figura 25**). L'opció de re-segmentació del CNApp es va ometre.

Com a resultat d'aquesta anàlisi es varen obtenir les quantificacions de les CNAs per mostra, il·lustrades amb els valors dels *CNA scores*. La distribució de valors pel BCS va ésser inexistent, indicant la falta d'alteracions de tipus *broad*, mentre que els valors de FCS presenten una mitjana de FCS = 80 (**Figura 54**). La falta d'identificació d'alteracions àmplies va decantar l'anàlisi en la direcció d'identificar els perfils associats a les alteracions focals. Per això es seleccionaren les finestres genòmiques mínimes (1Mb) per a generar els perfils genòmics.

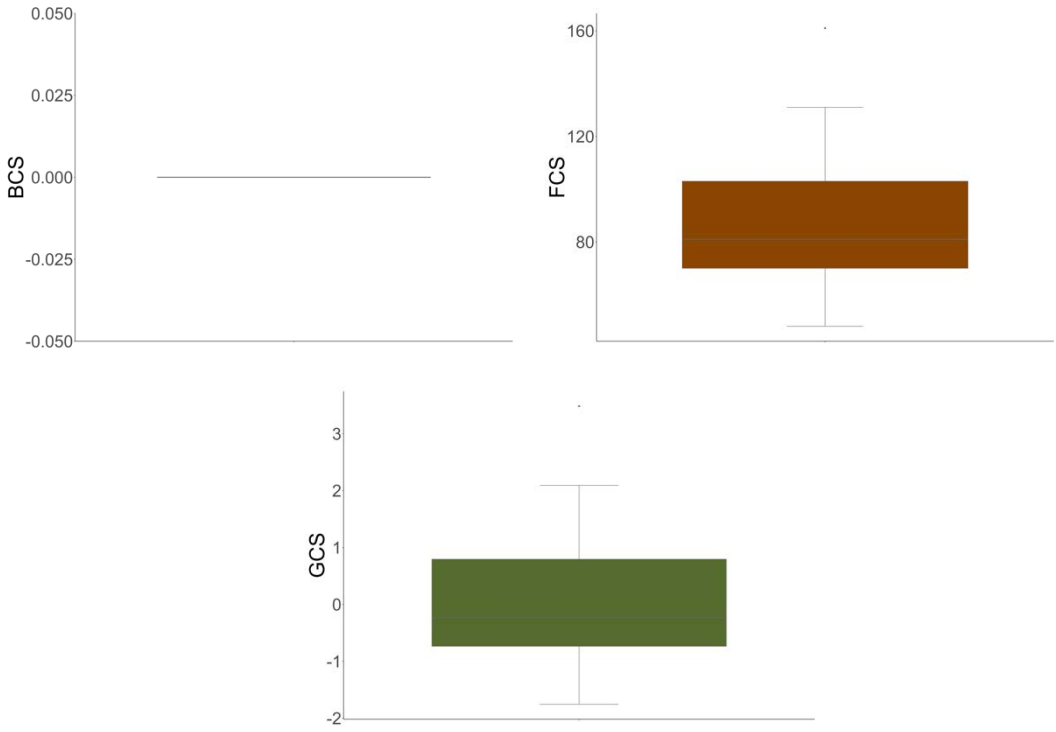


Figura 54. Valors dels *CNA scores* BCS, FCS i GCS en les mostres seqüenciades per exoma del projecte FAMCOLON.
Donada la falta d'alteracions àmplies (*broad*), els valors de BCS són inexistents, mentre que els valors de FCS presenten una mitjana de 80.

La característica *Descriptive regions*, de la secció *Region profile*, ens permetia analitzar aquelles regions diferencialment representades entre les famílies amb dos o més membres seqüenciats. Així, es varen poder identificar aquelles regions genòmiques que es presentaven alterades en les comparacions família a família. En aquests cas, es varen ometre de l'anàlisi els nuclis familiars compostos només d'un individu.

La regió *118|chr1:117000000-118000000* dels perfils genòmics generats a partir de les finestres d'1Mb, la qual conté els gens de la duplicació del cromosoma 1 identificada i estudiada al primer estudi (**Taula 16**), presentava significança estadística en quant a la seva alteració diferencial en totes les comparacions de les quals hi participava la família 7 (fam7). Pel que fa als valors d'alteració de la regió per a cada una de les famílies analitzades, tan sols la família 7 presentà valors d'amplitud allunyats de la neutralitat (**Figura 54**).

Per altra banda, aprofitant les dades del TCGA utilitzades en el segon estudi, es varen observar les freqüències en que la regió estudiada (*118|chr1:117000000-118000000*) es mostrava alterada en les mostres tumorals corresponents a les cohort de COAD i READ (de l'anglès, *rectal adenocarcionama*). Es calcularen els perfils genòmics per a les dues

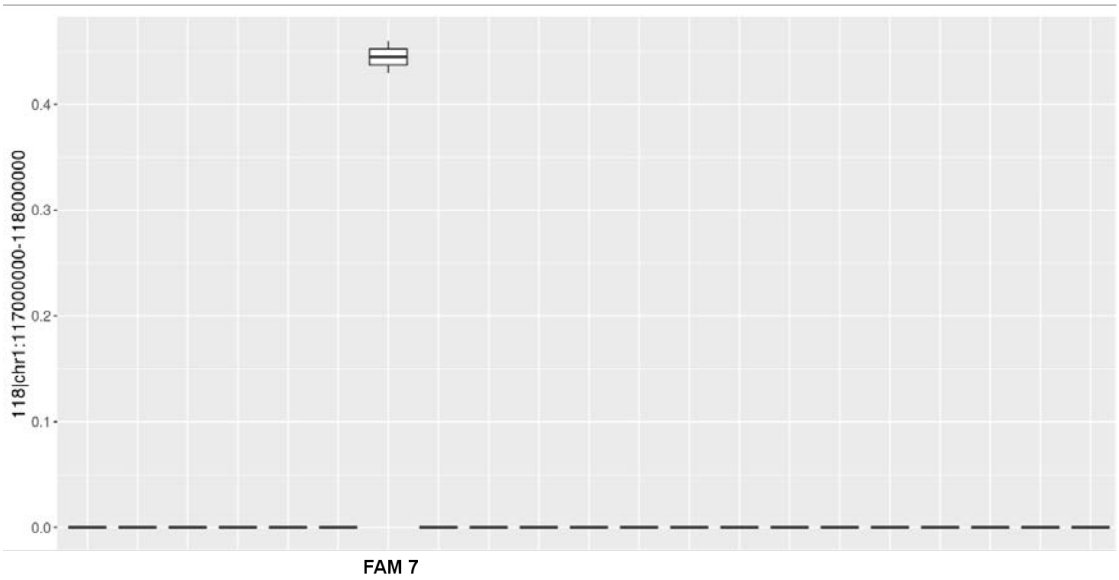


Figura 55. Valors d'alteració per a la regió 118 en les famílies amb més d'un individu seqüenciat.
Es visualitza la distribució dels valors mitjans d'alteració per a la regió 118 dels perfils genòmics d'una megabase per a cada família del projecte FAMCOLON amb més d'un individu seqüenciat. La regió 118 es veu inalterada en totes les famílies excepte en la portadora (familia 7 -FAM 7-) de la duplicació del cromosoma 1 identificada en l'estudi 1.

cohorts mitjançant finestres d'una megabase i només tenint en compte aquelles CNAs classificades com a focals. Els perfils de freqüència es mostren en la **Figura 56**.

Taula 16. Llistat dels gens implicats en la regió 118|chr1:117000000-118000000 dels perfils genòmics de finestres d'una Mb de CNApp.

Gen	Cromosoma	Inici	Final
CD58	chr1	117057155	117113715
IGSF3	chr1	117117019	117209578
MIR320B1	chr1	117214370	117214449
C1orf137	chr1	117236733	117249225
CD2	chr1	117297051	117311851
PTGFRN	chr1	117452544	117532980
CD101	chr1	117544371	117579173
LOC101929099	chr1	117568103	117602112
TTF2	chr1	117602948	117645491
MIR942	chr1	117637264	117637350
TRIM45	chr1	117653676	117664411
VTCN1	chr1	117686208	117746458
LINC01525	chr1	117838087	117863958

En les mostres de COAD (N=462), els perfils genòmics calculats a partir de finestres d'1Mb presentaren freqüències d'alteració de la regió específica de 1,73% en forma de guany, i 5,63% per a pèrdues (**Taula 17**). Pel que fa a la cohort de READ (N=164), les freqüències foren de 3,66% per al guany i 9,15% en la pèrdua.

Per altra banda, també s'interrogà la regió *536|chr1:45000000-46000000* , la qual conté el gen *TMEM158* (situat en la regió p21.31 del cromosoma 3), per tal d'esbrinar si la regió es mostrava alterada en mostres de CCR esporàdic. En aquest cas, les freqüències de guany i pèrdua per a la regió 536 foren de 0,22% i 2,66% en les mostres de COAD i 2,44% i 3,66 en la de READ (**Taula 17**).

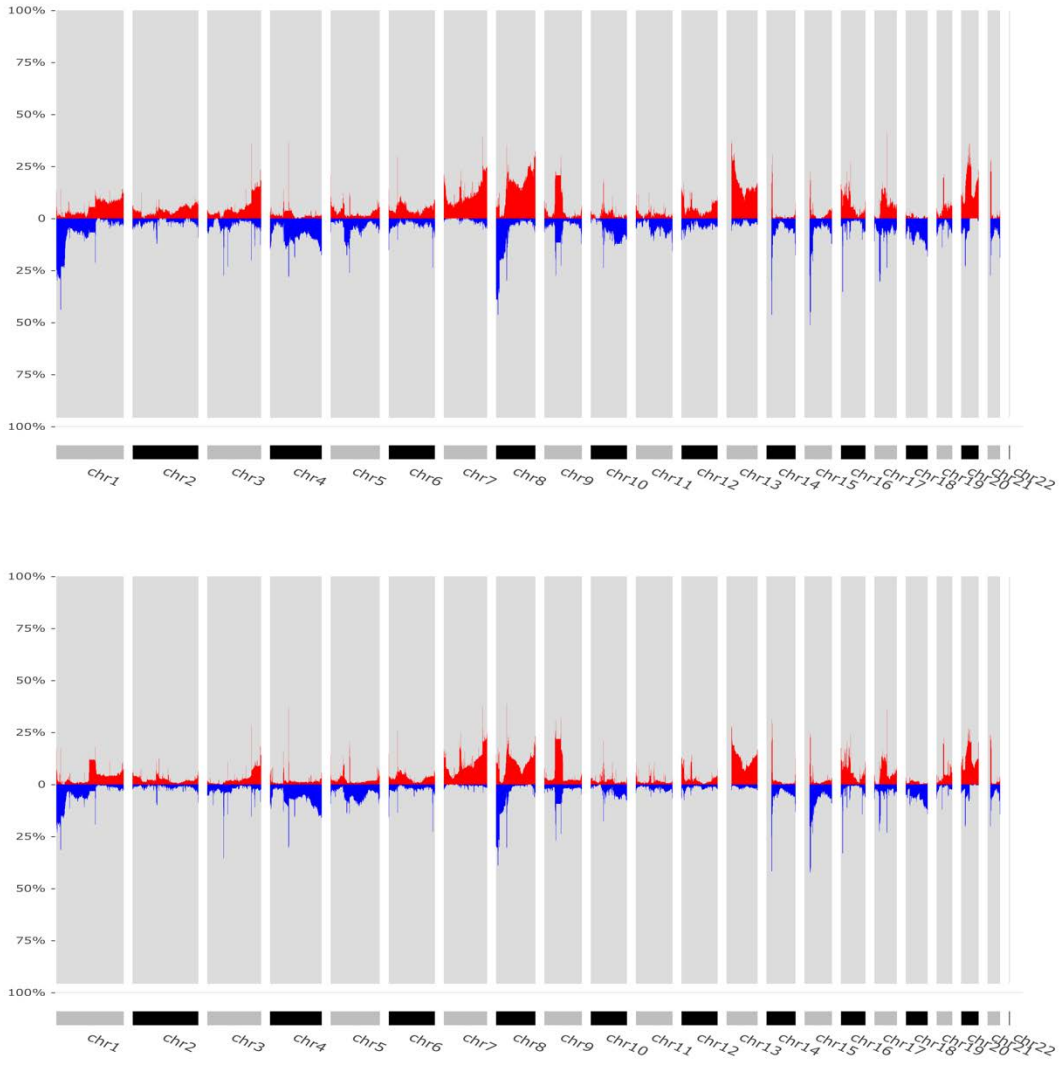


Figura 56. Perfils genòmics per a les cohorts READ i COAD en finestres d'una Mb i tenint en compte les CNAs focals.
Mitjançant l'aplicació CNApp, es varen generar els perfils genòmics de CNAs focals per a les mostres de les cohorts READ (superior) i COAD (inferior), respectivament, utilitzant les finestres genòmiques d'una megabase i es calcularen les freqüències de guany i pèrdua en cada una d'aquestes regions del genoma.

Taula 17. Freqüència d'alteració de la regió 118|chr1:117000000-118000000 i la 536|chr3:45000000-46000000 en les cohorts COAD i READ.

		% guany	% pèrdua
118|chr1:117000000-118000000	**COAD** (N=462)	1,73	5,63
	READ (N=164)	2,44	9,15
536|chr3:45000000-46000000	**COAD** (N=462)	0,22	2,66
	READ (N=164)	2,44	3,66

Discussió

Estudi 1

En l'estudi 1 s'ha realitzat el procediment d'identificació de CNVs potencialment implicades en la predisposició al CCR en famílies amb forta agregació per aquesta patologia. Això s'ha dut a terme mitjançant la inferència d'aquests tipus de variants en les dades de WES del DNA germinal d'alguns membres d'aquestes famílies. Mitjançant aquesta inferència, s'ha identificat, validat i caracteritzat una duplicació de 392 Kb del cromosoma 1 (chr1:117591257-117982865) en una de les famílies estudiades. En la regió genòmica de la duplicació s'hi localitzen quatre gens (*TTF2*, *TRIM45*, *VTCN1* i *MAN1A2*) i un miRNA (*MIR942*), localitzat en l'intró 18 del gen *TTF2*. Per tal de caracteritzar els possibles efectes moleculars de la variant a sobre d'aquests gens, s'estudiaren els nivells d'expressió gènica i proteica d'aquests.

Inferència de les variants del número de còpia en dades de seqüenciació de l'exoma

La utilització de la seqüenciació de l'exoma com a plataforma per a la inferència de CNVs es troba en discussió constant. El fet de només seqüenciar les regions codificants (exons), representa la falta de disponibilitat de la informació genòmica de forma continuada al llarg del genoma, intercalant les regions seqüenciades, de les que s'obtenen els *reads*, amb les regions intròniques i intergèniques no-seqüenciades. Això fa que s'hagi de recórrer a distintes metodologies de inferència estadística per a la identificació de les CNVs i una posterior re-segmentació dels segments genòmics de número de còpia, normalment mitjançant la *circular binary segmentation* (Olshen et al., 2004), implementada en el paquet d'R DNAcopy (Venkatraman & Olshen, 2007).

Però, tot i que la detecció de CNVs mitjançant dades de WES és sub-òptima –sobretot comparada amb la detecció de variants estructurals en la WGS, on la continuïtat en la cobertura de les regions genòmiques facilita la seva inferència (Hehir-Kwa et al., 2015)–, l'avantatge econòmica que suposa la seva generació en comparació a les dades de WGS, així com la increïble quantitat de bibliografia i algoritmes d'inferència a l'abast de la comunitat científica, fan que la seva utilitat sigui, almenys, un valor inicial afegit en aquest tipus de recerca biomèdica i, fins i tot, d'aplicació en la clínica. De fet, en els últims anys, han anat apareixent estudis que proposen la implementació de la seqüenciació de l'exoma com a eina encaminada al diagnòstic clínic (Gambin et al., 2017; Pfundt et al., 2017). Tot i això, la inferència de CNVs amb dades de WES necessita de la combinació *a posteriori* de tècniques citogenètiques o de matrius genòmiques per tal de confirmar i validar les

variants detectades. Alguns estudis pre-clínics en l'àmbit del diagnòstic prenatal han aconsellat la implementació d'aquesta combinació entre WES i tècniques genòmiques de validació per al diagnòstic prenatal de variants potencialment patogèniques, a l'espera de que la WGS s'abarateixi i el seu processament i anàlisi sigui més assequible i fiable per la comunitat clínica (Lord et al., 2019; Petrovski et al., 2019).

Sense anar més lluny, en aquest estudi s'han aplicat tècniques d'alta resolució per tal de validar la variant prioritzada. La primera validació s'ha realitzat mitjançant aCGH, tècnica que també ens ha permès encaminar els estudis de segregació de la variant en membres addicionals de la família portadora. Posteriorment, s'ha seqüenciat el genoma complet per a un dels portadors de la duplicació caracteritzada, descartant l'existència de qualsevol tipus de variant estructural que pogués representar el mecanisme candidat responsable de la predisposició al CCR en la família. A més, la re-validació de la duplicació mitjançant les dades de WGS ens ha permès identificar amb un major grau de resolució la posició genòmica de la variant, facilitant el disseny de sondes per a la realització d'estudis de seqüenciació dirigida, per tal de caracteritzar la regió de ruptura de la duplicació a nivell nucleotídic.

Els estudis d'identificació de CNVs implicades en la predisposició al CCR utilitzen tècniques genòmiques d'alta resolució com les matrius de sondes genòmiques d'alta densitat (**Taula 4**). Per altra banda, apliquen distints tipus de validació genòmica i/o funcional de les variants identificades. La duplicació de 40 Kb a la zona promotora del gen *GREM1*, va ser identificada en dos individus amb HMPS mitjançant les tècniques genòmiques aCGH i aSNP, validant la desregulació de l'expressió gènica del gen mitjançant la monitorització dels nivells de mRNA (Jaeger et al., 2012). A més, els investigadors identificaren que aquesta regulació venia condicionada per la major activació de la zona promotora del gen degut a la presència de la duplicació mitjançant estudis de la conformació de la cromatina en línies cel·lulars portadores de la duplicació i línies control. Pel que fa als demés estudis, la validació de les CNVs identificades es duu a terme mitjançant experiments de MLPA, estudiant les regions genòmiques identificades de forma més directa.

Per altra banda, la inferència de CNVs en dades WES per a l'estudi de la seva implicació en la predisposició al càncer no està molt implementada en la comunitat científica. Un dels pocs exemples n'és l'estudi de Fewings i col·laboradors, on seqüencien pacients diagnosticats de càncer gàstric difús hereditari en famílies amb forta agregació per la malaltia i sense mutacions al gen *CDH1*, el qual ja es

troba descrit com a gen de predisposició a aquesta neoplàsia (Fewings et al., 2018). En aquest treball, els investigadors acaben proposant els gens *PALB2*, *MSH2* (entre d'altres), degut a la identificació de mutacions puntuals deletèries en aquests gens. Pel que fa a la detecció de potencials CNVs de predisposició, després de la inferència d'aquestes variants mitjançant l'anàlisi de les dades WES amb l'algoritme XHMM, els resultats no reporten cap CNVs interessant en les famílies.

Algoritme de treball per a la detecció de variants del número de còpia rares

Per tal d'assolir la identificació de CNVs rares involucrades en la predisposició al CCR en famílies que presentaven forta agregació per la patologia, s'ha generat un flux de treball que consisteix en la inferència mitjançant les eines CoNIFER i ExomeDepth, la posterior anotació de la freqüència poblacional per a les variants inferides, consultant la base de dades *the Database of Genomic Variants* i la generació d'una nova bases de dades de CNVs a partir del perfils genòmics de 500 individus controls provinents del consorci internacional EPICOLON, i una priorització final de les variants candidates depenent de la seva freqüències poblacional, la seva longitud i els gens involucrats (**Figura 25**).

CoNIFER ha estat utilitzat en diversos estudis i comparat amb altres algoritmes d'inferència de CNVs (Krumm et al., 2012; Guo et al., 2013; Samarakoon et al., 2014; Tan et al., 2014; Yao et al., 2017). Aquesta eina és capaç d'identificar aquelles variants més rares entre la cohort d'estudi gràcies al tipus d'anàlisi que realitza: per a cada mostra que s'estudia, el perfil de cobertura de les regions genòmiques seqüenciades és comparat amb el perfil mitjà de cobertura de la resta de mostres. Això resulta en la identificació d'aquelles regions que es desvien dels valors de cobertura mitja com a CNVs de la mostra. Per altra banda, ExomeDepth s'identificà com un dels algoritmes més utilitzats en la inferència de CNVs mitjançant les dades de WES en la bibliografia i, a més, ha estat objecte d'estudi en distints treballs on es comparava amb altres eines d'inferència (Plagnol et al., 2012; Guo et al., 2013; Vandrovcova et al., 2013; Zhao, Wang, Wang, Jia, & Zhao, 2013; Samarakoon et al., 2014; Tan et al., 2014; Kadalayil et al., 2015; Talevich et al., 2016; Ellingford et al., 2017). ExomeDepth calcula un subconjunt de mostres control, per a cada mostra estudiada, que serveix de base comparativa per a la identificació de les CNVs. Això fa que aquest algoritme sigui menys restrictiu, en comparació amb CoNIFER, a l'hora d'identificar esdeveniments potencialment compartits entre les mostres de la cohort d'estudi. De fet, aquesta característica podria respondre al fet de que ExomeDepth acabi identificant un major nombre de variants en comparació a CoNIFER (3700

CNVs identificades per ExomeDepth i 288 variants amb CoNIFER), com ja suggeria cert estudi comparatiu on s'analitzaven ambdues eines (Guo et al., 2013). Per altra banda, l'eficiència de detecció en la longitud de la CNV caracteritzada fou major per part d'ExomeDepth (360.302 pb) i no de CoNIFER (150.598 pb) quan es compara amb els resultats finals de la caracterització de la regió duplicada del cromosoma 1 (391.608 pb), com també es demostrava en l'estudi de Guo i col·laboradors.

Tot i així, existeixen un bon grapat d'eines d'inferència de CNVs en dades de WES i no existeix cap indicatiu clar de quina n'és la millor a l'hora d'inferir aquestes variants. Diversos estudis comparatius han intentat valorar les distintes capacitats de la multitud d'eines que es poden trobar a la comunitat científica (Guo et al., 2013; Zhao et al., 2013; Samarakoon et al., 2014; Tan et al., 2014). Entre elles, una de les més utilitzades és la XHMM (Fromer et al., 2012). Aquesta elimina artefactes tècnics i biaixos de les mostres que s'analitzen de forma individual aplicant anàlisis de components principals en la matriu de dades de *coverage* de les regions seqüenciades i aplica models de normalització a les dades per tal d'inferir el número de còpia a nivell d'exons. Per altra banda, XHMM necessita d'un gran nombre de mostres per a poder realitzar la inferència de les CNVs correctament (aproximadament 50 mostres) (Fromer et al., 2012; Yao et al., 2017). Mentre que CoNIFER com ExomeDepth poden començar a treballar amb cohorts petites, fins i tot pròximes a la desena (Krumm et al., 2012; Plagnol et al., 2012), XHMM treballa millor amb cohorts a partir de les 50 mostres i fins a les milers de mostres (Lek et al., 2016; Huang et al., 2018).

Per altra banda, s'ha de tenir en compte que els resultats obtinguts en aquests estudi, en quant a les variants candidates identificades en les famílies, responen a CNVs compartides entre els individus seqüenciats per a cada família. Per una banda, s'hauria obtingut un major nombre de variants si no s'hagués aplicat aquest filtre que, per altra banda, també ens ajudava a centrar-nos en l'estudi de l'herència de la predisposició al CCR. De manera similar, el fet de prioritzar aquelles variants inferides en ambdues eines, també ens ha facilitat la reducció del nombre de variants candidates, a més d'atorgar certa robustesa en quant a la certesa de que aquelles CNVs s'havien inferit correctament.

Degut al nostre objectiu d'identificar CNVs candidates a representar el mecanisme de predisposició per el que aquelles famílies presentaven forta agregació al CCR, ens havíem d'assegurar de prioritzar aquelles CNVs rares que respondrien a la hipòtesis de variants d'alta penetrància present en les famílies de CCR. Per això, es

consultaren les freqüències poblacionals de les variants candidates tant en la base de dades DGV com en la generada a partir de les mostres de l'EPICOLON. En el nostre cas, les variants que es presentaven una presència més elevada de deu vegades en les bases de dades era desestimada de l'estudi. Però altres treballs han utilitzat filtres distints. Per exemple, l'estudi de Brea-Fernández i col·laboradors, on s'identificaren dues delecions rares que involucraven els gens *AK3* i *SLIT2* en dos individus distints d'entre 27 pacients diagnosticats per CCR amb edats menors als 50 anys CNVs, també consultaren la DGV com a base de dades control (Brea-Fernandez et al., 2017). En aquest cas, els investigadors consideraven només les variants del DGV que es trobaven recolzades en dos estudis diferents de la base de dades i identificades almenys dues vegades per a considera la seva validesa en la DGV. Aplicant aquest criteri, podríem considerar que la duplicació del cromosoma 1 estudiada en el primer estudi és totalment exclusiva de la família portadora entre la nostra cohort, ja que tan sols s'havia identificat una sola vegada en un estudi de la DGV.

Mecanismes de generació de variants del número de còpia de predisposició

La duplicació del cromosoma 1 detectada tan sols s'ha identificat en un dels grans estudis del catàleg del DGV. En aquest es generà un mapa de CNVs al genoma mitjançant l'aplicació de aCGH en aproximadament 29.000 menors amb desordres de retard mental (Coe et al., 2014). El fet de que la duplicació d'aquesta regió del cromosoma 1 s'hagi identificat en només un individu entre tots el que conformaren aquesta cohort, fa que es pugui etiquetar com una variant rara. Això, i la presència de les seqüències de micro-homologia identificades en la zona de ruptura de la duplicació (extrems 5' i 3' de la duplicació validada), a més de la inserció de 72 pb detectada que provenia d'una regió genòmica pròxima (l'últim intró del gen *TTF2* [chr1:117642853-117642924]) (**Figura 29**), fa pensar en la implicació de mecanismes de generació de variants no-recurrents, fet que quadraria amb el seu caràcter (gairebé) exclusiu per a la família. Com hem vist abans, els mecanismes moleculars implicats en la generació de les CNVs no-recurrents poden ser processos de replicació induïda per ruptura a conseqüència de micro-homologies (com BIR o MMBIR), o el mecanisme NHEJ (C. E. Smith et al., 2007; Lieber, 2008; Hastings et al., 2009; C. M. B. Carvalho & Lupski, 2016) .

Als estudis d'identificació de delecions de la regió 3' del gen *EPCAM* en pacients en síndrome de Lynch també es detectà la implicació d'aquests tipus de mecanismes moleculars de generació de CNVs. En aquests, la caracterització de casos de síndrome de Lynch que presentaven híper-metilació del gen *MSH2* va identificar distintes formes de deleció de la regió 3' del gen

EPCAM com el mecanisme de predisposició genètica en aquests individus (Ligtenberg et al., 2009). Diversos estudis han pogut identificar fins a 19 versions distintes de la deleció (Kovacs, Papp, Szentirmay, Otto, & Olah, 2009; Kuiper et al., 2011). Algunes d'aquestes presentaren seqüències repetitives en les regions de ruptura, característiques del mecanisme de recombinació NAHR, normalment associat a la generació de CNVs recurrents al genoma (Stankiewicz & Lupski, 2002), mentre que d'altres també mostraren regions de micro-homologia (Kuiper et al., 2011), més associats al BIR/MMBIR o NHEJ.

La duplicació de la regió del cromosoma 1 com a potencial mecanisme de predisposició al CCR en la família 7

Els estudis de segregació per a la duplicació en la família 7 detectaren la variant en els dos individus seqüenciats per WES (II1 i II2) i en un dels membres addicionals estudiats (III1, fill de II1) (**Figura 28**). Mitjançant l'estudi de la història clínica dels membres de la família es detectà que el familiar I1 (pare dels membres pels qui s'havia seqüenciat l'exoma) havia estat diagnosticat de rectorràgia. Per altra banda, el familiar I2 (mare dels individus seqüenciats i que no presentava la duplicació) havia estat diagnosticat de CCR a l'edat de 84, fet que dona a pensar que aquest diagnòstic es tracte d'un cas de CCR esporàdic. Amb tot, la hipòtesi en aquesta família apuntaria a que la predisposició al CCR prové de la branca familiar del pare, presentant-se amb un patró d'herència dominant, ja que sovint s'ha associat la presentació de rectorràgies (sagnat rectal) com a signe clínic de CCR (Majumdar, Fletcher, & Evans, 1999; Davies et al., 2005; Kuipers et al., 2015).

La duplicació caracteritzada, localitzada en la regió p13.1-p12 del cromosoma 1, amb una longitud de 392 Kb (chr1: 117591257-117982865), implicava els gens *TTF2*, *MIR942*, *TRIM45*, *VTCN1* i *MAN1A2*. El primer, el gen *TTF2*, forma part de la família SWI2/SNF2 i codifica per a un factor de terminació responsable de la repressió mitòtica en la transcripció del DNA, ja que actua sobre de la polimerasa II de RNA (Jiang & Price, 2004). *MIR942* es troba situat en l'intró 18 del gen *TTF2* i la seva funció s'ha associat amb la promoció de les cèl·lules mare del càncer (Ge et al., 2015). *TRIM45* pertany a la família proteica de les TRIM, que desenvolupen funcions associades a la regulació de la carcinogènesis (Hatakeyama, 2011). Pel que fa al gen *VTCN1*, en estudis anteriors s'han detectat nivells d'expressió alts d'aquest gen en distints tipus de càncer en humans, inclòs el CCR (Peng et al., 2015). I, finalment, el gen *MAN1A2*, que pertany a la família de les α -

manosidases, família proteica que realitza funcions de maduració d'oligosacàrids a l'aparell de Golgi (Tremblay & Herscovics, 2000).

Per tal de descartar altres potencials mecanismes de predisposició al CCR en la família, es consultà el treball anterior del grup on s'analitzaven les mutacions puntuals en la majoria de famílies sotmeses a estudi en aquesta tesi. En ells no s'havia detectat cap variant interesant en la família afectada per la duplicació del cromosoma 1 (**Taula 18**) (Esteban-Jurado et al., 2015), descartant la possibilitat de que el mecanisme de predisposició genètica al CCR en la família vingués donat com a conseqüència de mutacions puntuals en altres gens relacionats o implicats en el desenvolupament de la malaltia. Per altra banda, també es descartaren potencials variants estructurals mitjançant la seva inferència en les dades de WGS d'un dels membres portadors de la duplicació estudiada i altres variants rellevants en gens implicats en CCR hereditari. Per tant, els resultats del present treball apuntarien a la duplicació del cromosoma 1 caracteritzada com el potencial mecanisme responsable de la predisposició a la malaltia en la família portadora.

Per altra banda, tan sols els gens *TTF2* i *MIR942* s'havien vist sobre-expressats en els estudis d'expressió gènica quan es comparaven els perfils dels individus portadors de la duplicació amb els no-portadors de la duplicació. En termes de desregulació d'aquestes gens, la infraexpressió del factor de terminació producte del gen *TTF2* provocaria la retenció de la transcripció en els cromosomes mitòtics (Jiang & Price, 2004), mentre que no existeixen estudis en quant a la seva sobreexpressió i seria difícil fer alguna hipòtesis sobre les seves conseqüències, donada la forta regulació gènica a qual es troba sotmès el gen, associada al cicle cel·lular (M. Liu, Xie, & Price, 1998). Pel que fa al gen *MIR942*, la seva sobreexpressió ha estat correlacionada positivament amb característiques associades a cèl·lules mare del càncer, concretament en la progressió del carcinoma cel·lular d'esòfag esquamós mitjançant la interacció amb dianes que regulen negativament la via de senyalització WNT/ß-catenin (Ge et al., 2015).

Donada la sobre-expressió d'aquest miRNA es procedí a estudiar les dianes predites per al miR-942 i el seu creuament amb els gens infra-expressats al pacient portador de la duplicació, centraren l'atenció a sobre del gen *TMEM158*, també conegut com *RIS1* (*Ras-induced sensescence 1*). Aquest s'ha descrit com a potencial TSG relacionat en la carcinogènesi del CCR i com a biomarcador de mala prognosis en situacions en que es presentava alterat (Barradas et al., 2002; Iglesias et al., 2006). A més, *TMEM158* s'ha identificat com a marcador molecular durant la senescència induïda per mutacions en la proteïna Ras -procés associat a la supressió tumoral degut a l'estancament de la proliferació cel·lular que provoca- i, per altra banda, la seva localització genòmica (3p21.3), anomenada CER1 (de l'anglès *common eliminated region*

1), també s'ha identificat com una regió delecionada de forma recurrent durant el procés de carcinogènesis del CCR (Hesson et al., 2007). Aquests aspectes, juntament amb els nivells de proteïna baixos TMEM158 al teixit tumoral de l'individu portador de la duplicació, comparat amb mostres d'individus pacients no-portadors (**Figura 32**), reforcen el plantejament de que el mecanisme de predisposició al CCR de la família 7 podria donar-se per una desregulació del gen *TMEM158,* com a conseqüència de la sobre-expressió del miR-942, seguint el mecanisme molecular de funció dels miRNA, pel qual aquestes unitats regulen a la baixa els nivells d'expressió dels gens sobre els quals actuen (Baek et al., 2008; Bartel, 2009).

Distints estudis han implicat les proteïnes de la família TMEM (de l'anglès, *transmembrane protein*) amb el càncer, ja sigui per la seva desregulació gènica (Zhou, Popescu, Klein, & Imreh, 2007; Cuajungco et al., 2012; Hrašovec, Hauptman, Glavač, Jelenc, & Ravnik-Glavač, 2013; Qiao et al., 2016), la seva funció de biomarcador (Wrzesiński et al., 2015) o el seu potencial paper en el desenvolupament del càncer i la resistència a distints fàrmacs, en forma de TSGs o d'oncògenes (Schmit & Michiels, 2018). Les proteïnes TMEM es localitzen en la bi-capa lipídica de la membrana cel·lular. De fet, moltes d'elles treballen com canals proteics que permeten el pas de substàncies específiques. Tot i així, les funcions de moltes de les proteïnes que composen aquesta família resten desconegudes (Vinothkumar & Henderson, 2010).

L'estudi de Venkatachalam i col·laboradors identificà diverses CNVs en una cohort de 41 pacients amb CCR abans dels 40 anys i que presentaven agregació familiar per la patologia mitjançant matrius genòmiques d'alta resolució (Venkatachalam et al., 2011). Entre els gens afectats per aquestes CNVs s'hi trobaven *CDH18*, *GREM1* i *BCR*, mentre que dues de les delecions identificades involucraven els miRNAs mir-491 mir-646. Els autors complementaren l'estudi amb cribratges mutacionals en una cohort independent de 96 pacients MSS diagnosticats per CCR abans dels 40 anys d'edat, on identificaren diversos polimorfismes en la seqüència del DNA que codifica pel miRNA mir-646, mentre que el mir-491 s'havia relacionat amb la inducció d'apoptosi en línies cel·lulars de CCR mitjançant la regulació d'un dels seus gens diana (Nakano, Miyazawa, Kinoshita, Yamada, & Yoshida, 2009). De fet, en els últims anys, els miRNAs s'han vingut relacionant en la susceptibilitat a distints tipus de neoplàsies (Calin et al., 2005; Jazdzewski et al., 2008; Iuliano et al., 2013).

Eines de predicció de dianes dels miRNAs

Des del seu descobriment, cap als anys 90, el nombre de miRNAs reportats ha incrementat de manera exponencial (R. C. Lee, Feinbaum, & Ambros, 1993). Actualment, la base de dades miRBase conté més de 38.000 seqüències de miRNAs (Kozomara, Birgaoanu, & Griffiths-Jones, 2019) i, tot i que any rere any es van coneixent més interaccions miRNA-mRNA a nivell biològic, la gran majoria d'aquestes encara no s'han validat a nivell experimental (Roberts & Borchert, 2017). Donada la forta regulació dels perfils d'expressió gènica per part dels miRNA, la necessitat de predicció de les seves potencials dianes és alta, mentre que la seva validació funcional és difícil i costosa (Alexiou, Maragkakis, Papadopoulos, Reczko, & Hatzigeorgiou, 2009). Les eines de predicció actuals, basades en la complementarietat de seqüència, produeixen una quantitat de falsos positius considerable, ja que aquesta complementarietat entre les petites seqüències nucleotídiques dels miRNA i les molècules de mRNA és altament probable, apuntant a una gran quantitat de seqüències compatibles (Barbato et al., 2009). A més, la diversitat genètica sorgida dels processos d'edició del mRNA i l'*splicing* alternatiu pot generar encara més diversitat de seqüències diana (Borchert et al., 2009; Yang, Zhang, & Li, 2012).

L'eina utilitzada en aquest estudi, TARGETSCAN (http://targetscan.org), treballa a nivell de complementarietat de seqüència per tal d'obtenir les dianes de predicció dels miRNAs i, per tant, és susceptible a presentar falsos positius entre les seves prediccions. Per altra banda, també incorpora informació sobre la conservació entre les distintes espècies dels organismes, per tal de reduir el nombre de falsos positius en les prediccions (Agarwal et al., 2015). Altres eines realitzen les seves prediccions a partir de l'estructura secundària dels miRNAs i el seu equilibri termodinàmic predit a partir d'aquesta estructura, tot i que això també suposa recolzar-se en l'estructura primària de les molècules; d'altres comencen a aplicar sistemes d'aprenentatge automàtic (en anglès, *machine-learning*) per predir les interaccions a partir de bases de dades amb interaccions miRNA-mRNA ja establertes (Roberts & Borchert, 2017). Estudis recents, aprofitant les millores en les plataformes genòmiques, s'han encaminat a la generació de noves dades de interacció per a la identificació de seqüències complementàries, aplicant les matrius d'immunoprecipitació per entrecreuament (en l'anglès *crosslinking immunoprecipitation*), validant seqüències d'interacció i aprofitant les dades generades per a millorar els sistemes de predicció (W. Liu & Wang, 2019).

Tot i aquests esforços en la millora de la seva predicció, la validació biològica d'aquestes interaccions miRNA-mRNA és obligatòria per a poder extreure'n conclusions aplicables a la recerca bàsica i/o a la clínica. En el nostre cas, la desregulació del gen *TMEM158* per part del miR-942 hauria d'ésser

validada a nivell molecular i, en aquest sentit, experiments com l'assaig de la luciferasa, amb posteriors estudis de co-expressió d'ambdós agents en les cèl·lules de l'epiteli colònic, es podrien dur a terme per tal de confirmar el mecanisme de desregulació (Kuhn et al., 2008).

Possible remodelació topològica de la cromatina

Un aspecte interessant en quant a la caracterització molecular dels gens implicats en la duplicació del cromosa 1 fou la desregulació gènica identificada en els individus portadors de la variant que tan sols afectava al gen *TTF2*, el més 5' de la regió duplicada, a més del miRNA miR-942 (**Figura 31**). Diversos estudis han descrit que esdeveniments de canvi del número de còpia del material genòmic en regions que afecten a diferents TADs poden provocar re-ordenaments estructurals i espacials de la cromatina al nucli, generant nous contactes genòmics a conseqüència d'aquesta remodelació, ja sigui mitjançant la formació de nous TADs en la regió o fusionant-ne d'altres que ja existien (Spielmann et al., 2018). En el nostre cas d'estudi, semblava que la duplicació caracteritzada podria estar afectant dos TADs distints: el primer, involucrant el gen *TTF2* i la seva regió promotora, a més d'altres en regions 5' d'aquest; i el segon, que contenia els gens *TRIM45*, *VTCN1*, *MAN1A2* i altres gens a 3' (**Figura 30**). Degut a aquesta implicació, la variant podria haver provocat l'aparició de nous contactes entre el gen *TTF2* i les regions promotores o elements reguladors d'altres gens com a conseqüència de la formació d'un nou TAD, alterant els nivells d'expressió gènica del mateix *TTF2*. A més, donat que el gen MIR942 es localitza en l'intró 18 de *TTF2*, aquest també podria veure's desregulat de la mateixa manera, provocant la seva sobre-expressió, identificada als estudis d'expressió gènica.

Un clar exemple de la desregulació gènica coma conseqüència de la remodelació genòmica n'és l'estudi de Franke i col·laboradors, els quals estudiaren la regió genòmica del gen *SOX9*, lligada a distints fenotips patològics a conseqüència de diferents formes de re-ordenaments cromosòmics (Franke et al., 2016). Comparant duplicacions en aquesta regió que afectaven només el TAD que implicava el gen *SOX9*, amb duplicacions inter-TAD que involucraven més d'un TAD de la regió, identificaren els efectes estructurals en la cromatina i les conseqüències a nivell dels contactes genòmics que es donaven en l'àrea genòmica. Així doncs, la duplicació genòmica dintre del TAD on es localitza el gen *SOX9* produïa el fenotip relacionat amb la reversió del sexe de l'individu, mentre que la duplicació que incloïa el gen *KCNJ2*, en

la regió 5' de *SOX9* i inserit en un TAD distint del de *SOX9*, provocava la formació d'un nou TAD (neo-TAD) que suposava el fenotip de la síndrome de Cooks (**Figura 57**). Per altra banda, si la duplicació inter-TAD no incloïa el gen *KCNJ2*, es generava un neo-TAD sense cap tipus de fenotip patològic relacionat.

De fet, per tal d'adreçar aquestes desregulacions genètiques com a conseqüència dels re-ordenaments estructurals de la cromatina, en els últims anys s'han desenvolupat algunes eines bioinformàtiques que prediuen de forma sistemàtica aquest tipus de desregulació de l'expressió gènica. Alguns exemples en són l'algoritme CESAM (Weischenfeldt et al., 2017), que es centra en la identificació de gens relacionats amb fenotips de càncer per la seva sobre-expressió com a conseqüència de l'alteració de l'arquitectura genòmica; el Genomiser (Smedley et al., 2016), que identifica regions genòmiques reguladores en malalties d'herència mendeliana; o l'eina PRISMR (Bianco et al., 2018), de modelat tri-dimensional de la cromatina mitjançant dades de HiC-seq i que pot utilitzar-se per la predicció d'interaccions genòmiques i així estudiar les variants estructurals potencialment patològiques.

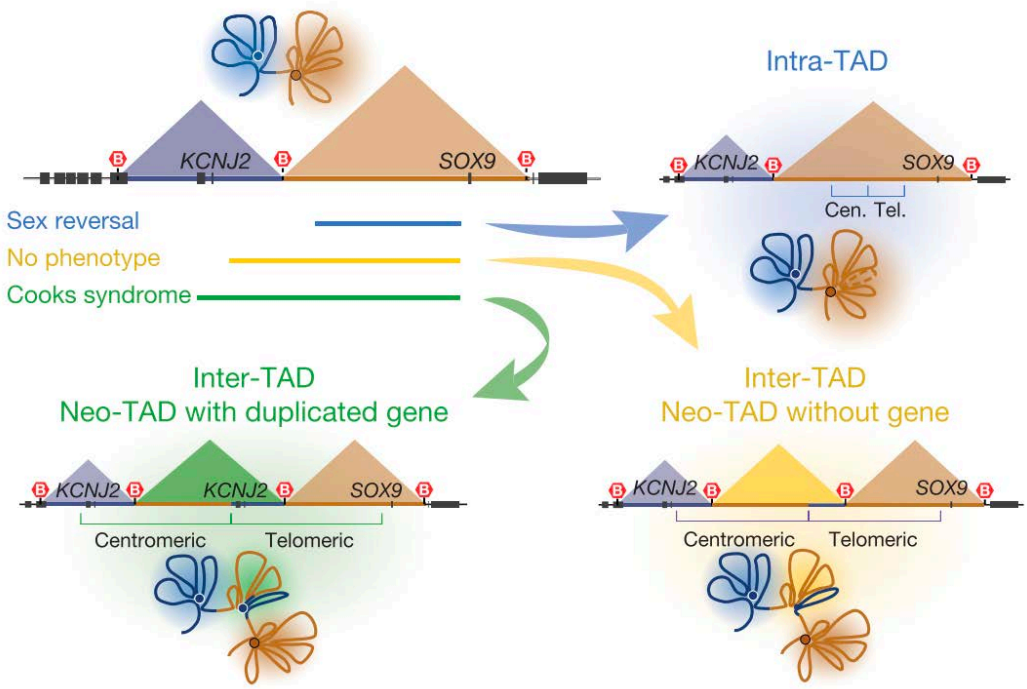


Figura 57. Efectes en l'organització de la cromatina segons la regió de duplicació i els fenotips que se'n deriven.
Les duplicacions que afecten un sol TAD (en blau, intra-TAD) no afecten a la conformació estructural del TAD, però tenen efectes en la dosis gènica dels gens afectats. Les duplicacions que afecten distints TADs (en verd i groc, inter-TAD) deriven en la formació de noves regions de contacte, o neo-TADs. La incorporació de gens i elements reguladors, externs al neo-TAD, condueixen a nous perfils de regulació que poden derivar en la desregulació gènica i distints fenotips (en verd, Síndrome de Cook). (Extreta de Franke M. et.al *Nature* 2016)

Taula 18. Variants puntuals detectades en els dos pacients de CCR seqüenciats per WES en la familia 7. SNVs detectades, prioritzades i validades en l'estudi Esteban-Jurado et al. 2015.

Gen	Chr	Pos	Ref	Alt	Canvi aminoàcid	ExAC	Freq. interna	Segregació	*In silico*	Funció biològica/ termes bibliogràfics/OMIM
ADC	1	33583621	T	G	V403G	1072/106708	0.1429	Incorrecta en altres famílies	5	*Metabolic processes; spermatogenesis*
HIST2H2BE	1	149858168	G	A	A8V	3/120728	0.0286	--	3	*Systemic lupus erythematosus; chromatin remodeling*
ITGA4	2	182396457	C	A	A913D	90/120308	0.0429	--	4	*Cell-cell adhesion*
MKRN2	3	12610401	A	C	E18D	3/121272	0.0286	--	4	*Protein ubiquitination*
CCKAR	4	26490952	C	T	M89I	1/121390	0.0286	--	4	*Neuronal processes*
PLK4	4	128811080	C	A	H507N	1/120056	0.0286	--	3	*Microcephaly and chorioretinopathy; autosomal recessive, 2*
FOXQ1	6	1314117	A	C	H393P	2/132	0.1857	Incorrecta en altres famílies	3	*Hair follicle morphogenesis*
HIST1H4B	6	26027246	G	A	R79C	16/121326	0.0286	--	4	*Systemic lupus erythematosus; chromatin remodeling*
GPR116	6	46851360	T	C	Q183R	37/120366	0.0286	--	3	*Fat cell differentiation; energy reserve metabolic process*
GSTA1	6	52664028	C	T	R13Q	188/121360	0.0286	--	3	*Xenobiotic metabolic process; epithelial cell differentiation*
TBX18	6	85446744	G	T	Q495K	0	0.0286	--	4	*Congenital anomalies of kidney and urinary tract 2*
HECA	6	139488148	G	C	R333S	8/119270	0.0286	--	3	*Respiratory tube development*
SYNJ2	6	158502214	G	T	V881L	1/121412	0.0286	--	3	*Metabolism of lipids and lipoproteins*
KANK1	9	732615	G	C	E1081D	1/120494	0.0286	--	3	*Cerebral palsy; spastic quadriplegic, 2*
CRTAC1	10	99644046	C	T	V517M	14/22688	0.0286	--	4	*Axonal fasciculation; olfactory bulb development; bone/cartilage metabolism*
PITX3	10	103990344	T	G	Y279S	0	0.0286	--	5	*Anterior segment mesenchymal dysgenesis; Cataract 11; multiple types*
PTPN5	11	18754841	C	A	V387L	0	0.0286	--	5	*MAPK signaling pathway; cognitive and morphological changes*
VEGFB	11	64005048	A	C	T156P	387/38646	0.1714	Incorrecta en altres famílies	3	*Angiogenesis*
OS9	12	58113927	G	A	R549H	17/120930	0.0286	--	3	*Protein ubiquitination*

MYO16	13	109535458	C	A	P471T	2/121382	0.0286	--	4	*Negative regulation of cell proliferation; cerebellum development*
CIDEB	14	24776647	C	T	R39H	5/121300	0.0286	--	5	*Induction of apoptosis*
PTGR2	14	74345884	T	C	F202S	67/121408	0.0571	--	5	*Prostaglandin metabolic process*
LDLRAD4	18	13612735	T	G	L11R	1/121346	0.0286	--	3	*Negative regulation of transforming growth factor beta receptor signaling pathway*
MAP1S	19	17838150	C	T	R653W	37/42994	0.0286	--	4	*Microtubule bundle formation; neuronal processes*
SLC5A5	19	17988795	G	A	G288S	42/121262	0.0286	--	4	*Thyroid dyshormonogenesis 1*
AKT1S1	19	50376485	C	T	R43Q	69/12598	0.0429	--	3	*Neuronal processes; epidermal growth factor receptor signaling pathway*
ATP6V1G3	1	198509776	G	T	T2K	19/119282	0.1286	Incorrecta en altres famílies	4	*Insulin receptor signaling pathway; Phagosomal maturation*

Chr: Cromosoma. Pos: coordenades genòmiques del genoma de referència hg19. Ref: Al·lel de referència. Alt: Al·lel alternatiu.
ExAC: Freqüència al·lèlica de la variant a la base Exome Aggregation Consortium (http://exac.broadinstitute.org).
Freqüència internafrequency: freqüència del genotip de la variant en la cohort interna de 71 pacients.
Segregació: if the variant segregates correctly with CRC in other families from our cohort.
In silico: Número de prediccions deletèries per a la variant en les eines de predicció consultades. Deletèries segons: CADD_phred>15; pp2>0.85; phylop>0.8; SIFT_score<0.05; LTR_score<0.1.
Funció biològica/ termes bibliogràfics/OMIM: Termes bibliogràfics anotats segons la funció biològica del gen (OMIM, Pubmed, GeneRIF) i les bases de dades (*Gene Onthology*, *KEGG pathway* i *REACTOME*).

Estudi 2

Durant els últims anys, el coneixement de les característiques genòmiques per als distints tipus de càncer s'ha pogut ampliar mitjançant l'estudi de les mutacions puntuals i les variants estructurals. El paper de les CNAs als processos de carcinogènesis ha estat sotmès a una àmplia caracterització en les mostres tumorals aplicant les últimes tecnologies de seqüenciació i genotipat, com les NGS o les matrius d'SNPs. El major exponent d'aquests estudis ha vingut de la mà del projecte TCGA, caracteritzant a nivell genòmic més de trenta tipus de càncer distints (The Cancer Genome Atlas, n.d.).

Superada aquesta fase d'identificació de les particularitats específiques genòmiques per molts tipus de càncer (tot i que encara resta moltíssima feina en el camp de la caracterització a mesura que es milloren les tècniques genòmiques i s'expandeix l'anàlisi mitjançant seqüenciació completa del genoma), estudis encaminats a extreure associacions funcionals per aquestes variants genòmiques específiques per als distints tumors i els diferents subgrups de mostres estan a l'ordre del dia (Wang et al. 2013; Beroukhim et al. 2010; Ciriello et al. 2013; Vogelstein et al. 2013). El gran *handicap* d'aquests tipus d'estudis ve donat per la necessitat d'anàlisi de les dades generades i la capacitat de relacionar-les amb les anotacions clíniques i/o moleculars que es tenen de les mostres que sotmeten a estudi. Per això, a banda de seguir millorant les eines de caracterització i identificació de les variants genòmiques (tant mutacions puntuals com els esdeveniments estructurals) a mesura que s'avança en les plataformes de seqüenciació i genotipat genòmic, també són necessàries eines que facilitin aquest procés d'anàlisi i associació, i que ho permetin fer d'una manera automatitzada, superant els impediments tècnics i de càrrega temporal que suposa aquest tipus d'estudis.

Al segon estudi d'aquesta tesi s'ha desenvolupat l'eina bioinformàtica CNApp, recolzant-se en llenguatge de programació R mitjançant la plataforma R-Studio i el paquet Shiny. L'objectiu prioritari de CNApp és el d'analitzar i integrar les CNAs detectades en qualsevol tipus de plataforma genòmica amb les anotacions molecular i/o clíniques disponibles sobre les mostres a estudiar. Per això, CNApp quantifica la càrrega que presenten les mostres en quant a alteracions *broad* i focals, presentant-la mitjançant els valors BCS (*broad CNA score*) i FCS (*focal CNA score*), respectivament. A més, també aporta valors de la càrrega d'alteracions global mitjançant els valors de GCS (*global CNA score*). L'eina ofereix un procés de re-segmentació opcional dels perfils de regions genòmiques introduïdes i, per altra banda, és capaç de generar perfils genòmics complets utilitzant finestres genòmiques pre-establertes (les distintes opcions entre les que es pot optar són: braços cromosòmics, mitjos braços, bandes

citogenètiques, sub-bandes i regions de 40 Mb fins 1 Mb). Aquest procés facilita la posterior comparació entre mostres o entre els distints grups de mostres definits per les variables d'anotació disponibles, i permet el reconeixement de regions específiques entre aquestes. Finalment, CNApp pot utilitzar les distintes variables generades o introduïdes per l'usuari per tal de generar models de classificació de les mostres basats en sistemes d'aprenentatge automàtic.

La facilitat de dur a terme un estudi genòmic personalitzat utilitzant CNApp proporciona major capacitat d'anàlisi i estudi de les CNAs al càncer. A més, el valor afegit de l'eina és el de permetre la integració de les CNAs amb anotacions clíniques i/o moleculars de les mostres que es puguin tenir a l'abast, amb l'objectiu d'identificar les relacions funcionals d'aquestes alteracions amb la possibilitat de, en un futur, extrapolar-les a la clínica.

Avantatges dels *CNA scores*

Un dels objectius clars de l'eina CNApp és la de prioritzar anàlisi de conjunt de dades per tal d'extreure característiques genòmiques referents a les CNAs que puguin associar-se a grups de mostres específics i a les seves característiques informatives. Tot i això, la primera secció de l'aplicació també realitza cert nivell d'estudi individual per a les mostres que es troben sota estudi. En un primer pas de re-segmentació opcional es pot utilitzar per a millorar la lectura dels perfils de CNAs de les mostres, en els casos que, per exemple, es cregui que puguin existir certes interferències tècniques degudes a la plataforma de genotipat utilitzada, o bé com a conseqüència d'una baixa qualitat de les mateixes mostres.

La quantificació de les alteracions que proporciona l'eina mitjançant el càlcul dels *CNA scores* (BCS per a les alteracions *broad*, FCS per a les focals i GCS com a mesura global) facilita una interpretació ràpida de la càrrega en quant a esdeveniments genòmics del número de còpia que presenta cada mostra. Així, es facilita una escala relativa dintre de la cohort d'estudi, per tal d'ordenar les mostres segons el número d'alteracions que presenta. Però, a més, i gràcies a la caracterització de les més de 10.000 mostres analitzades del TCGA, efectuada en aquest treball, també es té disponible una àmplia referència de valors dels diferents *CNA scores* i en distints tipus de càncer, per tal de que es pugui relativitzar la valoració quantitativa de nous estudis que apliquin CNApp.

Per altra banda, s'ha de tenir en compte que, a distintes plataformes de genotipat (per exemple WES o WGS), la regió coberta del genoma difereix i, per tant, també ho fa la quantitat de CNAs en les que s'està treballant. CNApp no realitza cap tipus de normalització o correcció tenint en compte la plataforma d'on provenen les dades que s'introdueixen. Això implica que, per a una mateixa mostra, s'identifiquin distints valors de *CNA scores* segons la plataforma genòmica que s'utilitza per a caracteritzar les CNAs: és a dir, la seqüenciació de l'exoma presentarà valors dels *CNA scores* més baixos en comparació a la seqüenciació completa del genoma de la mateixa mostra, per exemple, pel simple fet de que en la segona opció s'haurà cobert major quantitat de seqüència genòmica en l'estudi i, per tant, hi ha major probabilitat d'identificar CNAs.

Existeixen altres intents pel que fa a la quantificació de la càrrega de CNAs per mostra. Un dels més recents és l'estudi de Taylor i col·laboradors, on els investigadors generen un valor de quantificació d'aneuploïdia, el qual anomenen *aneuploidy score*, mitjançant el recompte de CNAs àmplies (*broad*) clusteritzades en els distints braços cromosòmics de les mostres de pan-cancer del TCGA (Taylor et al., 2018). A banda de correlacionar fortament amb valors de la fracció alterada del genoma per aneuploïdia en les mostres del seu estudi (com és per al nostre cas la correlació entre el CNA *score* BCS i la fracció alterada del genoma per segments *broad* de les mostres analitzades en aquest segon estudi), aquest *aneuploidy score* correlaciona negativament amb la càrrega mutacional de les mostres, com ja s'havia postulat anteriorment (Ciriello et al., 2013). Tot i això, els autors identifiquen que gran part d'aquesta correlació negativa ve aportada per mostres amb MSI o amb mutacions a *POLE* (sobretot en les cohorts de COAD i UCEC), per tant, mostres amb perfils hípermutats (Campbell et al., 2017). De fet, eliminades aquestes mostres, la correlació entre el *aneuploidy score* i la càrrega mutacional de les mostres que analitzen els autors apareix positiva en la majoria de tipus de càncer (Taylor et al., 2018). Per altra banda, també identifiquen correlació negativa entre els valors d'aneuploïdia (una altra vegada utilitzant el seu *aneuploidy score*) i mesures de fracció leucocitària, confirmant la relació directa entre els nivells d'aneuploïdia amb resposta immunitària reduïda (Davoli et al., 2017).

Per tant, seria d'esperar que el *CNA score* per a la quantificació de la càrrega d'esdeveniments *broad* (BCS), calculat mitjançant el CNApp, també pugues presentar el mateix tipus de relació amb la càrrega mutacional de les mostres i la resposta immunitària, ja que en aquest estudi s'ha validat la forta correlació entre el BCS i la fracció alterada del genoma en les mostres estudiades. De ser així, el càlcul del valor de BCS per mostra podria tenir certa capacitat predictiva en quant a la resposta immunitària, essent important a l'hora d'aplicar tractaments basats en immunoteràpia.

A més, la quantificació de la càrrega de CNAs focals mitjançant el càlcul del valor FCS aporta una capa més d'informació en quant al perfil genòmic estructural al càncer. L'estudi de Davoli i col·laboradors apuntà a una correlació positiva entre la càrrega de CNAs focals i firmes d'expressió de gens relacionats amb la proliferació cel·lular (Davoli et al., 2017). Seria interessant comprovar la relació entre els valors de FCS calculats i distints tipus de marcadors de proliferació cel·lular, amb la intenció d'identificar quins són aquells tumors amb major activació de la seva activitat proliferativa.

Region profile, regions recurrents i *descriptive regions*

Aquesta secció es dedica íntegrament a l'estudi, com ja es deia abans, de les mostres com a conjunts i subconjunts definits per les característiques i anotacions de variables informatives. La generació de perfils genòmics delimitats per finestres genòmiques prefixades permet la comparació directa entre mostres ja que, d'aquesta manera, totes elles tenen la mateixa quantitat de regions d'alteració a analitzar.

L'extracció dels valors de recurrència de les CNAs, és a dir, les freqüències en què es presenten les alteracions genòmiques entre el conjunt de mostres, segurament és una de les característiques més valorades a l'hora d'analitzar CNAs. Això és degut a que alteracions recurrents presents en mostres tumorals poden estar afectant gens implicats en vies de senyalització claus durant els processos de carcinogènesis (Camps et al., 2009; Stratton et al., 2009; The Cancer Genome Atlas, 2012; Zack et al., 2013).

Una de les eines més utilitzades a l'hora d'identificar les regions d'alteració genòmica recurrents és el GISTIC2.0 (Mermel et al., 2011). Aquesta eina aplica mètodes probabilístics a partir del càlcul de les regions recurrents per tal de valorar i identificar quins gens es troben afectats en elles i, per tant, implicats en la patogènesis del càncer sota estudi. El GISTIC2.0 s'ha utilitzat per a la caracterització dels perfils de CNAs de gran part de les cohorts genòmiques del TCGA i ha permès la identificació de gens implicats en els processos de carcinogènesis de les distintes neoplàsies mitjançant aquests tipus de metodologia probabilística enfocada a determinar la regió significativament alterada en les mostres. El CNApp, tot i no ésser una eina adreçada específicament en aquest tipus d'estudi de la recurrència o en la identificació de TSGs o oncogens implicats en la biologia dels tumors, ha demostrat certa capacitat en aquest sentit. Però, si les regions

genòmiques pre-definides del CNApp, per una banda, afavoreixen la comparació i correlació entre mostres, per l'altra, representen un inconvenient a l'hora de definir mínimes regions específiques de recurrència entre les mostres. Així doncs, l'única opció que resta és la de consultar quins són els gens involucrats en les regions pre-establertes més petites que proporciona el CNApp: les finestres genòmiques d'una megabase. Per altra banda, el GISTIC2.0 és capaç d'aportar les regions genòmiques mínimes comuns entre mostres (o *peaks*, en anglès), que es troben potencialment alterades en un alt percentatge de les mostres analitzades, aportant valors estadístics corregits i apuntant als gens implicats (Mermel et al., 2011).

La generació de perfils genòmics de les CNAs mitjançant les finestres genòmiques permet l'estudi d'aquelles regions dels perfils que es presenten alterades de forma distinta entre grups de mostres. Així doncs, mitjançant la característica *Descriptive regions*, de la secció *Region profile*, es poden estudiar quines són les regions genòmiques que diferencien de forma significativa els grups de mostres definits per variables d'anotació. Això adquireix major rellevància a l'hora de caracteritzar subgrups moleculars, per exemple entre un mateix tipus de càncer, com hem realitzat en aquest estudi amb els subtipus de CMS per a les mostres de còlon. En aquest sentit, l'aplicació CoNVaq també ens permet fer aquest tipus d'estudi de comparació entre dos grups de mostres diferents per tal d'associar les característiques que defineixen ambdós grups amb les CNAs característiques o compartides de cada un (Larsen et al., 2018). De la mateixa manera que fa CNApp, CoNVaq aplica el test de Fisher per identificar aquelles regions que són específiques per a un dels dos grups, diferenciant els perfils genòmics entre ambdós. A més, una segona característica d'aquesta eina permet realitzar anàlisi més detallats especificant els llindars de recurrència per a les CNVs (rares) que es volen detectar entre els grups (per exemple, consultant les CNVs més grans que es troben en >30% de les mostres del grup A i <5% a les mostres del grup B) (Larsen et al., 2018).

Caracterització de la cohort *pan-cancer* del TCGA

Per tal de testejar l'aplicació bioinformàtica CNApp, en aquest estudi s'han analitzat més de 10.000 mostres del projecte TCGA, corresponents a 33 tipus de càncers distints, mitjançant aquesta eina. Els distints apartats de CNApp han aportat resultats que vénen a confirmar observacions realitzades en estudis anteriors per la comunitat científica i d'altres que aporten certa novetat, deixant oberta la possibilitat per a futures observacions que es podrien dur a terme utilitzant la nostra eina.

En primer terme, s'han pogut quantificar els perfils de CNAs de les 10.635 mostres analitzades mitjançant els *CNA scores* de l'aplicació. La distribució dels valors de BCS, FCS i GCS per aquestes mostres entre els 33 tipus de càncer deixa entreveure la possibilitat de distintes relacions entre les alteracions *broad* i les focals per a les distintes neoplàsies (**Figura 38**).

Per altra banda, resta per explorar de forma més incisiva la correlació positiva entre els valors de BCS i FCS identificada en el present estudi (**Figura 37**). Aquesta es presentava en els primers valors de BCS (BCS = [0-5]), per a perdre's progressivament i, fins i tot, assolir valors de correlació negativa a partir de BCS = 19. Probablement, el fet d'assolir majors càrregues d'alteracions àmplies no permet l'adquisició d'esdeveniments focals, ja sigui per la restricció de l'espai físic al genoma, o per efecte de l'evolució clonal durant la carcinogènesis . De fet, tot i que els valors de BCS es distribueixen entre 0 i 44, els valors màxims de FCS són distints segons la neoplàsia que es consulta. Una hipòtesi seria l'existència d'un cert tipus de restricció en quant als nivells d'aneuploïdia estructural que les cèl·lules canceroses poden assolir (Gordon, Resio, & Pellman, 2012), mentre que, si aquests valors màxims d'esdeveniments amplis no s'han assolit, encara podrien presentar-se alteracions focals significatives (amb grans canvis d'amplitud del número de còpia, fet que genera valors tant alts de FCS).

Els 33 tipus de càncer estudiats mostren dinàmiques de relació diferents entre ambdós *CNA scores*, com hem pogut observar als resultats (**Figura 38**). Algunes es caracteritzen per presentar nivells baixos tant de BCS com de FCS (LAML, THCA o THYM) o alts nivells d'aneuploïdia per ambdós valors (UCS, OV o LUSC); mentre que d'altres es decanten per a presentar alts nivells de BCS però baixos de FCS, com el KICH, o a l'inrevés en el cas de BRCA. Per tant, és obvi pensar que les correlacions entre ambdós *CNA scores* podrien respondre a dinàmiques particulars segons els distints tipus de càncer. De fet, l'estudi de Beroukhim i col·laboradors ja deixava entreveure dinàmiques específiques entre la fracció de CNAs focals i la de CNAs *broad* entre els distints tipus de càncer (Beroukhim et al., 2010).

La caracterització dels perfils de CNAs elaborada pel CNApp utilitzant les finestres genòmiques de braços cromosòmics va permetre visualitzar de forma clara els patrons característics dels distints tipus de neoplàsia (**Figura 39**). Per altra banda, CNApp va ésser capaç de reproduir la clusterització entre els tipus de tumors que ja s'havia observat anteriorment (Hoadley et al., 2018; Taylor et al., 2018). Utilitzant les mateixes anotacions d'origen tumoral que l'estudi de

Taylor i col·laboradors, i treballant amb els 20 tipus de càncers en que s'havia identificat la clusterització, CNApp va ésser capaç de reproduir el resultat perfectament, reforçant la bona implementació del càlcul dels perfils de CNAs per finestres genòmiques. Aquesta dinàmica de clusterització a nivell d'aneuploïdia cromosòmica segons els patrons de braços ja es coneixia anteriorment gràcies a l'estudi de Beroukhim i col·laboradors, on els autors també van analitzar la clusterització mitjançant la identificació de les CNAs focals més rellevants en els distints tipus de càncer, finalment sense assolir resultat positiu en aquesta part (Beroukhim et al., 2010). Resta pendent l'anàlisi per esbrinar si els patrons d'alteracions focals en els distints tipus de càncer tenen capacitat per a clusteritzar segons l'origen tumoral, de la mateixa manera que ho fan els patrons de CNAs a nivell d'aneuploïdia de braços.

Com s'ha explicat a la introducció, l'origen tumoral tindria un fort impacte en la posterior classificació per al tipus de càncer desenvolupat. L'extensa caracterització genòmica dels distints tipus de càncer ha permès visualitzar els patrons d'alteració recurrent per a les neoplàsies amb origen cel·lular o tissular comú, provocant la conseqüent desregulació de certes vies de senyalització molecular que faciliten el desenvolupament dels tumors. Per exemple, els tumors gastrointestinals presenten els majors ratis d'alteració dels gens de la via TGF-ß , mentre que els tumors renal i cerebrals gairebé no presenten alteració d'aquesta via (Hoadley et al., 2018). Per tant, semblaria que els propis patrons de CNAs per als distints càncers contribueixen en aquesta desregulació condicionada per l'origen tumoral. D'altra banda, semblaria existir la possibilitat de que, en el cas dels patrons d'aneuploïdia dels braços cromosòmics, això suposi certa avantatge clonal prèvia al procés de carcinogènesis, on aquestes *broad* CNAs replicarien els patrons generals d'expressió dels teixits originals on es desenvolupen (Auslander et al., 2019).

Integració de les CNAs dels perfils de mostres CMS amb anotacions moleculars

Mitjançant la secció *Classifier model*, es va calcular l'eficiència de classificació de la variable BCS en quant als grup de mostres MSI o MSS, obtenint un valor d'eficiència del 82.2%. A més, això fou consistent amb la capacitat de discriminació de BCS entre el grups CMS1 (enriquit amb mostres MSI) i CMS2 (mostres amb perfils canònics de CNAs per al CCR), amb una eficiència del 88,40% (**Taula 15**). Aquesta capacitat de diferenciació entre les mostres amb majors nivells de CNAs (CMS2) i les que presenten perfils més plans en quant a aquests esdeveniments (CMS1) podria traslladar-se a la distinció entre mostres d'alta i baixa aneuploïdia en el CCR. De fet, l'aplicació del llindar BCS = 4 per a la re-classificació de les mostres entre els grups CMS ha demostrat l'existència de fins 39 tumors MSS (17%) amb valors de BCS

inferiors al valor de BCS=4, corroborant l'existència de tumors amb MSS amb baixa càrrega de CNAs (Jones et al., 2005; Camps et al., 2006). La major part d'aquestes mostres procedien del grup CMS3 (18 de 39), que es caracteritza per a presentar perfils amb baixos nivells de CNAs (Guinney et al., 2015; Dienstmann et al., 2017). Fins un 46% dels tumors CMS3 amb MSS de la cohort estudiada mostraren valors de BCS < 4.

Per altra banda, cinc de les set mostres MSI amb valors de BCS > 4 presentaven esdeveniments genòmics associats a les CNAs canòniques del CCR (guanys dels braços cromosòmics 8q, 13q i 20q, i pèrdues de 8p, 15q, 17p i 18q), incloent l'amplificació focal del gen *MYC*. Així, la facilitat en que les dades de segments genòmics i les anotacions per mostres s'integren en l'aplicació CNApp ha ajudat a corroborar, entre d'altres coses, l'existència de tumors que presenten MSI i alteracions genòmiques estructurals (Jones et al., 2005; Camps et al., 2006). Entre els tumors MSI classificats com a CMS3 es detectaren dos d'aquests que presentaven alteracions focals en forma de deleció en la regió del cromosoma 2, on es troben els gens *MSH2* i *MSH6*, els quals participen de la MMR (Fishel et al., 1993; Strand, Prolla, Liskay, & Petes, 1993; Miyaki et al., 1997; Bocker et al., 1999). Així doncs, aquesta pèrdua de les regions gèniques estarien suggerint una inactivació de la via MMR com a conseqüència de la deleció d'aquestes gens.

En quant a l'anotació corresponent a l'estat mutacional del gen *BRAF* en les mostres tumorals de còlon, com era d'esperar, el grup de CMS1 estava enriquit amb mostres amb el gen mutat, així com amb mostres MSI i amb fenotip CIMP (**Figures 43, 44 i Taula 7**). Curiosament, dues mostres del grup CMS4 també es caracteritzaven per la mutació en el gen *BRAF*. Una d'aquests mostres presentava una càrrega alta d'alteracions *broad* (BCS = 11), contrastant amb l'altra mostra CMS4 i *BRAF*-mutada, que presentava MSI i BCS=0, característiques pròpies d'una mostra CMS1. De forma similar, quatre mostres no-mutades en *BRAF* i classificades com CMS4, presentaven MSI i BCS=0, essent candidates a classificar-se com CMS1, basant-nos en la càrrega de CNAs que presentaven (**Figura 51**). Aquestes discrepàncies són d'importància rellevant, ja que estudis recents han descrit que alts nivells de CNAs correlacionarien amb una resposta reduïda a tractaments d'immunoteràpia (Davoli et al., 2017). De fet, s'ha suggerit que l'estat MSI de les mostres podria tenir capacitat predictiva a la resposta positiva de bloquejadors de controladors immunitaris en CCR avançat, segurament degut a aquesta presentació de nivells baixos de CNAs en els tumors MSI (Le et al., 2015). Per tant, seria de gran importància poder validar les re-classificacions assenyalades en aquest

treball aplicant el llindar BCS = 4 mitjançant l'estudi d'una cohort extra de validació amb tumors de CCR, per tal de comprovar que l'ús d'aquest llindar és capaç de discernir entre les mostres de baixa i alta aneuploïdia. De fet, altres treballs de caracterització genòmica han seleccionat satisfactòriament pacients per al tractament amb immunoteràpia sense recórrer a l'estudi de l'estat dels microsatèl·lits. En un d'aquests, es detectà un enriquiment per la signatura mutacional de deficiència en MMR en un pacient amb càncer de pròstata, mitjançant la implementació d'un panell genòmic on es seqüenciaven uns 400 gens relacionats amb càncer (Zehir et al., 2017).

Per altra banda, mentre que els valors de BCS presentaven diferències significatives entre els grups CMS ($P \leq 0.0001$, T-test d'Student), els valors de FCS no foren capaços d'il·lustrar diferències entre CMS1 de CMS3 i CMS2 de CMS4 (**Figura 49**). Per tant, semblaria ser que les alteracions *broad* tindrien més capacitat discriminatòria entre els quatre grups CMS. De fet, la distribució dels valors de BCS en la gràfica de caixes (**Figura 49**) reprodueix la distribució del comptatge d'alteracions somàtiques del número de còpia presentades en l'estudi original de Guinney i col·laboradors, extreta mitjançant l'eina GISTIC.2.0 (Guinney et al., 2015). Tenint en compte que la classificació dels grups de CMS es basa en l'expressió gènica de les mostres de còlon, una resposta a la major capacitat de discriminació entre els grups per part dels valors de BCS podria venir donada com a conseqüència de la desregulació gènica que suposa la presència d'alteracions cromosòmiques a nivell de braç i de cromosomes complets. Com ja s'ha explicat en la introducció, la desregulació gènica de les regions afectades per CNAs s'ha pogut confirmar en diversos estudis, on s'ha descrit una correlació positiva entre l'expressió gènica i les CNAs corresponents (Grade et al., 2006, 2007; Camps et al., 2008).

Regions descriptives entre els subtipus de CMS

Mitjançant l'aplicació dels testos estadístics T-test d'Student i de Fisher en l'apartat de *Descriptive regions* de la secció *Region profile* es varen poder determinar aquelles regions de braços cromosòmics diferencialment alterades entre els quatre grups de CMS en la cohort de COAD (**Figura 52**). S'identificaren com a regions més descriptives els braços cromosòmics 13q, 14q, 17p, 18p, 18q i 20q, presentant-se com a regions diferencialment alterades almenys en quatre de les sis comparacions grup a grup entre els subtipus de CMS (test d'Student amb *P*-valor ajustat $P \leq 0{,}005$). Durant els últims anys, aquestes sis regions han estat caracteritzades en diversos estudis entre les aneuploïdies cromosòmiques més recurrents al CCR (Ried et al., 1996; Meijer et al., 1998; Douglas et al., 2004; Nakao et al., 2004). A més, aquestes regions descriptives han demostrat una alta eficiència a l'hora de re-classificar les 309

mostres de COAD estudiades entre els grups de CMS (~59,40% d'eficiència), assolint la mateixa eficiència que s'aconseguí amb l'ús dels perfils de braços cromosòmics complets (**Taula 15**). L'estudi d'aquestes regions podria millorar la classificació dels subtipus de CMS, donada la forta correlació entre aquests esdeveniments i els efectes dels perfils d'expressió gènica dels gens involucrats en aquestes (Habermann et al., 2007), i potencialment utilitzar-se com a sistema de classificació.

Per altra banda, les regions que es presentaven diferencialment alterades en totes les comparacions dels grups CMS menys una eren els braços 18q i el 20q. El braç 18q fallava a l'hora de diferenciar entre els grups CMS3 i CMS4. La pèrdua d'aquesta regió, a banda de ser un dels esdeveniments freqüents en la carcinogènesi del CCR (~70% dels casos) (B Vogelstein et al., 1989; E R Fearon & Vogelstein, 1990), es considera un indicador de la via de carcinogènesis del CIN (Walther et al., 2009). Les mostres del subtipus CMS4 segueixen la via del CIN, mentre que la gran majoria de les del CMS3 també ho fan, per tant, és lògic que no existeixen moltes diferències de CNAs entre aquests dos grups (de fet, només les amplificacions del 13q i el 20q es presenten diferencialment alterades entre ells) (Dienstmann et al., 2017). Recordem que la classificació dels subtipus CMS es basa en els perfils d'expressió gènica, i no en els genòmics. Per altra banda, el braç 20q no era capaç de diferenciar entre CMS1 i CMS3 i, de fet, alguna cosa similar passa entre aquests dos grups (comparat amb el que passava entre CMS3 i CMS4). Les mostres del grup CMS1 presenten MSI i baixos nivells de CNAs, mentre que un subgrup considerable de les mostres de CMS3 també mostra les mateixes característiques , d'aquí que només les regions 7q i 18q es presenten diferencialment alterades (Dienstmann et al., 2017). De fet, aquestes regions alterades es presentaven de forma regular en el grup CMS3 en mostres que presentaven valors de BCS > 4. Per altra banda, el guany del braç cromosòmic 12q, present en el grup CMS1, diferenciava aquest amb el grup CMS2 i CMS4 ($P \leq 0.005$, T-test d'Student), coincidint en la descripció d'estudis anteriors que associaven aquest guany del cromosoma 12q amb els tumors MSI (Trautmann et al., 2006).

De forma interessant, ambdues regions es troben en distribucions de valors tant diferenciades entre els distints grups de CMS, que aconsegueixen replicar la distribució dels grups pel que fa al comptatge de CNAs reportat en l'estudi original dels CMS (Guinney et al., 2015), tot i que en el cas del 18q de forma invertida, ja que la regió sempre es perd. Especialment en el cas del braç cromosòmic 20q, el qual demostrà significança estadística entre totes les comparacions ($P \leq 0.05$, T-test

d'Student), la distribució diferencial dels valors es veia ben representada (**Figura 52**). El guany del cromosoma 20q, que conté diversos oncogens potencials, entre ells *BCL2L1*, *AURKA* i *SRC* (B Carvalho et al., 2009), és un dels esdeveniments característics implicats en el procés oncogènic del CCR i relacionat amb el procés de metàstasi (Diep et al., 2006). Tot i aquesta sobre-presentació de potencials oncogens al cromosoma 20q, sembla ésser que no tots ells es comporten de la mateixa manera en relació a altres marcadors moleculars. El cromosoma 20q representa un esdeveniment primerenc en la carcinogènesis del CCR, presentant-se recurrentment en tumors primaris i de metàstasi, com ho són les mutacions recurrents en els gens *KRAS* i *BRAF*. De fet, semblaria ésser que aquests dos tipus d'esdeveniments es relacionen de manera mútuament exclusiva entre ells en el conjunt de mostres de CCR. Per tant, els gens en el 20q haurien de presentar-se sobre-expressats en mostres sense mutacions als gens *KRAS* i *BRAF*. En aquest sentit, tot i que *BCL2L1* és el gen amb més freqüència de guany entre les mostres de CCR, el gen *SRC* demostra una major relació inversa entre la seva sobre-expressió i les mutacions en *KRAS* i *BRAF*, donant a entendre que la relació entre la variant estructural i la desregulació del gen no té perquè seguir sempre una relació directa (B Carvalho et al., 2009; Ptashkin et al., 2017).

Duplicació del cromosoma 1 estudiada amb el CNApp

Aprofitant un altre cop la característica *Descriptive regions* del CNApp, es varen analitzar les dades de WES germinal en individus de famílies amb forta agregació per CCR familiar del primer estudi. Mitjançant la generació dels perfils genòmics amb finestres d'una megabase es varen obtenir els valors per cada una de les regions del genoma. La regió *118|chr1:117000000-118000000*, la qual conté la regió duplicada validada en el primer estudi d'aquesta tesi, presentava significança estadística en quant a la seva alteració diferencial en totes les comparacions de les quals hi participava la família 7 (fam7), indicant la exclusivitat de l'alteració de la regió per aquesta família, com s'observa en la **Figura 55**.

En els perfils genòmics d'una megabase per a les cohorts del TCGA de COAD i READ, s'observaren les freqüències d'alteració de la regió 118 en ambdós sub-conjunts de dades (**Taula 17 i Figura 56**). Les freqüències enregistrades mostraren nivells molt baixos de recurrència en quant al guany de la regió 118 (1,73% per a la cohort de COAD i 2,44% per a la cohort de READ), confirmant de que la duplicació estudiada és un esdeveniment molt rar i, segurament, exclusiu de la família. Com a exemple de valors de recurrència, l'estudi de la caracterització genòmica del TCGA per al CCR situava amb un valor del 7% una de les CNAs focals més comunes en la cohort que afectava el

gen *IGF2* (The Cancer Genome Atlas, 2012). De fet, la pèrdua de la regió 118 semblaria ésser un esdeveniment més comú, donades les freqüències de pèrdua de les cohorts COAD (5,63%) i READ (9,15%).

Per altra banda, també es consultaren les freqüències d'alteració per a la regió *536|chr3:45000000-46000000*, que conté el gen *TMEM15*. Les freqüència d'alteració per aquesta regió en les cohorts de COAD i READ també resulten ésser baixes. Com ja hem vist, aquest gen es situa en la regió genòmica de la sub-citobanda 3p21.3, coneguda com CER1 (Hesson et al., 2007). Les sub-citobandes del cromosoma 3p.21 (3p21.31, 3p21.32 i 3p21.33) mostren valors de recurrència del ~3,5% en quan a la seva deleció en la cohort de CMS, entenent que no és una alteració comuna al CCR. De fet, aquesta regió on està inclòs el gen *TMEM158* pertany al 3p, un dels braços cromosòmics menys alterats de forma recurrent al casos de CCR.

Conclusions

Estudi 1

1. La identificació de variants del número de còpia pot realitzar-se satisfactòriament mitjançant l'anàlisi de les dades de seqüenciació completa de l'exoma, tot i que són necessaris estudis posteriors per a la seva validació i la caracterització de la regió genòmica específica.

2. La duplicació del cromosoma 1, caracteritzada en una de les famílies estudiades, podria representar l'esdeveniment mutacional implicat en la predisposició al CCR en aquesta família, provocant la sobreexpressió del gens *TTF2* i *MIR942*.

3. El miR-942, regulat a l'alça com a conseqüència de la duplicació del cromosoma 1, estaria regulant a la baixa la presència de la proteïna TMEM158, potencial element supressor tumoral en la via mutagènica de la carcinogènesis del CCR.

4. La inclusió de dos dominis associats topològicament en la regió genòmica de la duplicació del cromosoma 1 podria explicar la desregulació específica de l'expressió gènica dels gens *TTF2* i *MIR942* en els individus portadors de la duplicació i que l'expressió dels demés gens en la regió es presenti inalterada.

Estudi 2

5. CNApp facilita la integració de perfils genòmics d'alteracions del número de còpia amb variables clíniques i moleculars per extreure'n noves implicacions funcionals.

6. CNApp ha caracteritzat a nivell genòmic les 10.635 mostres del projecte TCGA satisfactòriament, reproduint els patrons generals d'esdeveniments del número de còpia per als distints tipus de càncer i la seva clusterització entre els principals orígens tumorals, i aportant nova informació com la quantificació de la càrrega d'alteracions del número de còpia amb el càlcul dels distints CNA scores, diferenciant entre alteracions àmplies (broad) i focals.

7. La integració dels perfils genòmics de càncer de còlon i la variable d'anotació de l'estat dels microsatèl·lits per aquestes mostres ha permès la determinació del valor BCS = 4 com a potencial llindar de discriminació entre els tumors estables i inestables.

8. Els braços cromosòmics 8p, 13q, 14q, 17p, 18p, 18q i 20q s'han identificat com les regions genòmiques amb valors de canvi del número de còpia més diferenciats entre els grups CMS, presentant una eficiència de classificació de les mostres entre els 4 subtipus del 60%, similar a l'eficiència assolida quan s'utilitzen els perfils complets de braços cromosòmics. L'estudi de les alteracions del número de còpia en aquestes regions podria millorar la definició genòmica dels grups CMS.

9. CNApp ha estat capaç d'identificar la regió genòmica que involucrava la duplicació del cromosoma 1 caracteritzada anteriorment, mitjançant l'anàlisi de les regions descriptives entre les famílies que presentaven forta agregació per CCR. A més, la baixa freqüència d'alteració en forma de guany de la regió en les cohorts somàtiques de COAD i READ confirmaria el caràcter exclusiu de la variant caracteritza.

Bibliografia

Abyzov, A., Urban, A. E., Snyder, M., & Gerstein, M. (2011). CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. *Genome Research*, *21*(6), 974–984. https://doi.org/10.1101/gr.114876.110

Adam, R., Spier, I., Zhao, B., Kloth, M., Marquez, J., Hinrichsen, I., … Aretz, S. (2016). Exome Sequencing Identifies Biallelic MSH3 Germline Mutations as a Recessive Subtype of Colorectal Adenomatous Polyposis. *The American Journal of Human Genetics*, *99*(2), 337–351. https://doi.org/10.1016/j.ajhg.2016.06.015

AECC. (2019). Observatorio del Cáncer. Retrieved October 5, 2018, from http://observatorio.aecc.es/

Affymetrix - Thermofisher. (n.d.). Affymetrix Power Tools. Retrieved from https://www.thermofisher.com/es/es/home/life-science/microarray-analysis/affymetrix.html

Agarwal, V., Bell, G. W., Nam, J.-W., & Bartel, D. P. (2015). Predicting effective microRNA target sites in mammalian mRNAs. *ELife*, *4*. https://doi.org/10.7554/eLife.05005

Agrawal, S., Bhattacharya, A., Manhas, J., & Sen, S. (2019). Molecular Diagnostics in Colorectal Cancer. In *Molecular Diagnostics in Cancer Patients* (pp. 143–155). https://doi.org/10.1007/978-981-13-5877-7_9

Ahlquist, D. A. (2010). Molecular Detection of Colorectal Neoplasia. *Gastroenterology*, *138*(6), 2127–2139. https://doi.org/10.1053/j.gastro.2010.01.055

Al-Tassan, N., Chmiel, N. H., Maynard, J., Fleming, N., Livingston, A. L., Williams, G. T., … Cheadle, J. P. (2002). Inherited variants of MYH associated with somatic G:C→T:A mutations in colorectal tumors. *Nature Genetics*, *30*(2), 227–232. https://doi.org/10.1038/ng828

Alaei-Mahabadi, B., Bhadury, J., Karlsson, J. W., Nilsson, J. A., & Larsson, E. (2016). Global analysis of somatic structural genomic alterations and their impact on gene expression in diverse human cancers. *Proceedings of the National Academy of Sciences of the United States of America*, *113*(48), 13768–13773. https://doi.org/10.1073/pnas.1606220113

Alexandrov, L. B., Nik-Zainal, S., Wedge, D. C., Aparicio, S. a J. R., Behjati, S., Biankin, A. V, … Stratton, M. R. (2013). Signatures of mutational processes in human cancer. *Nature*, *500*(7463), 415–421. https://doi.org/10.1038/nature12477

Alexiou, P., Maragkakis, M., Papadopoulos, G. L., Reczko, M., & Hatzigeorgiou, A. G. (2009). Lost in translation: an assessment and perspective for computational microRNA target identification. *Bioinformatics*, *25*(23), 3049–3055. https://doi.org/10.1093/bioinformatics/btp565

Alkan, C., Coe, B. P., & Eichler, E. E. (2011). Genome structural variation discovery and genotyping. *Nature Reviews Genetics*, *12*(5), 363–376. https://doi.org/10.1038/nrg2958

Alkodsi, A., Louhimo, R., & Hautaniemi, S. (2015). Comparative analysis of methods for identifying somatic copy number alterations from deep sequencing data. *Briefings in Bioinformatics*, *16*(2), 242–254. https://doi.org/10.1093/bib/bbu004

Arnold, J. (1879). Beobachtungen über Kerntheilungen in den Zellen der Geschwülste. *Archiv Für Pathologische Anatomie Und Physiologie Und Für Klinische Medicin*, *78*(2), 279–301. https://doi.org/10.1007/BF01878412

Auslander, N., Heselmeyer-Haddad, K., Patkar, S., Hirsch, D., Camps, J., Brown, M., … ried, thomas. (2019). Cancer-type specific aneuploidies hard-wire chromosome-wide gene expression patterns of their tissue of origin. *BioRxiv*, 563858. https://doi.org/10.1101/563858

Baek, D., Villén, J., Shin, C., Camargo, F. D., Gygi, S. P., & Bartel, D. P. (2008). The impact of microRNAs on protein output. *Nature*, *455*(7209), 64–71. https://doi.org/10.1038/nature07242

Balaguer, F., Castellví–Bel, S., Castells, A., Andreu, M., Muñoz, J., Gisbert, J. P., … Gastrointestinal Oncology Group of the Spanish Gastroenterological Association. (2007). Identification of MYH Mutation Carriers in Colorectal Cancer: A Multicenter, Case-Control, Population-Based Study. *Clinical Gastroenterology and Hepatology*, *5*(3), 379–387. https://doi.org/10.1016/j.cgh.2006.12.025

Barbato, C., Arisi, I., Frizzo, M. E., Brandi, R., Da Sacco, L., & Masotti, A. (2009). Computational Challenges in miRNA Target Predictions: To Be or Not to Be a True Target? *Journal of Biomedicine and Biotechnology*, *2009*, 1–9. https://doi.org/10.1155/2009/803069

Barber, T. D., McManus, K., Yuen, K. W. Y., Reis, M., Parmigiani, G., Shen, D., … Hieter, P. (2008). Chromatid cohesion defects may underlie chromosome instability in human colorectal cancers. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(9), 3443–3448. https://doi.org/10.1073/pnas.0712384105

Barradas, M., Gonos, E. S., Zebedee, Z., Kolettas, E., Petropoulou, C., Delgado, M. D., … Serrano, M. (2002). Identification of a Candidate Tumor-Suppressor Gene Specifically Activated during Ras-Induced Senescence. *Experimental Cell Research*, *273*(2), 127–137. https://doi.org/10.1006/excr.2001.5434

Bartel, D. P. (2009). MicroRNAs: Target Recognition and Regulatory Functions. *Cell*, *136*(2), 215–233. https://doi.org/10.1016/j.cell.2009.01.002

Bass, A. J., Lawrence, M. S., Brace, L. E., Ramos, A. H., Drier, Y., Cibulskis, K., … Meyerson, M. (2011). Genomic sequencing of colorectal adenocarcinomas identifies a recurrent VTI1A-TCF7L2 fusion. *Nature Genetics*, *43*(10), 964–968. https://doi.org/10.1038/ng.936

Bass, A. J., Watanabe, H., Mermel, C. H., Yu, S., Perner, S., Verhaak, R. G., … Meyerson, M. (2009). SOX2 is an amplified lineage-survival oncogene in lung and esophageal squamous cell carcinomas. *Nature Genetics*, *41*(11), 1238–1242. https://doi.org/10.1038/ng.465

Bauman, J. G., Wiegant, J., Borst, P., & van Duijn, P. (1980). A new method for fluorescence microscopical localization of specific DNA sequences by in situ hybridization of fluorochromelabelled RNA. *Experimental Cell Research*, *128*(2), 485–490.

Beach, R., Chan, A. O.-O., Wu, T.-T., White, J. A., Morris, J. S., Lunagomez, S., … Rashid, A. (2005). BRAF mutations in aberrant crypt foci and hyperplastic polyposis. *The American Journal of Pathology*, *166*(4), 1069–1075. https://doi.org/10.1016/S0002-9440(10)62327-9

Bellido, F., Sowada, N., Mur, P., Lázaro, C., Pons, T., Valdés-Mas, R., … Valle, L. (2018). Association Between Germline Mutations in BRF1 , a Subunit of the RNA Polymerase III Transcription Complex, and Hereditary Colorectal Cancer. *Gastroenterology*, *154*(1), 181-194.e20. https://doi.org/10.1053/j.gastro.2017.09.005

Beroukhim, R., Getz, G., Nghiemphu, L., Barretina, J., Hsueh, T., Linhart, D., … Sellers, W. R. (2007). Assessing the significance of chromosomal aberrations in cancer: methodology and application to glioma. *Proceedings of the National Academy of Sciences of the United States of America*, *104*(50), 20007–20012. https://doi.org/10.1073/pnas.0710052104

Beroukhim, R., Mermel, C. H., Porter, D., Wei, G., Raychaudhuri, S., Donovan, J., … Meyerson, M. (2010). The landscape of somatic copy-number alteration across human cancers. *Nature*, *463*(7283), 899–905. https://doi.org/10.1038/nature08822

Bianco, S., Lupiáñez, D. G., Chiariello, A. M., Annunziatella, C., Kraft, K., Schöpflin, R., … Nicodemi, M. (2018). Polymer physics predicts the effects of structural variants on chromatin architecture. *Nature Genetics*, *50*(5), 662–667. https://doi.org/10.1038/s41588-018-0098-8

Binefa, G., Rodríguez-Moranta, F., Teule, A., & Medina-Hayas, M. (2014). Colorectal cancer: from prevention to personalized medicine. *World Journal of Gastroenterology*, *20*(22), 6786–6808. https://doi.org/10.3748/wjg.v20.i22.6786

Bird, A. P. (1986). CpG-rich islands and the function of DNA methylation. *Nature*, *321*(6067), 209–213. https://doi.org/10.1038/321209a0

Bocker, T., Rüschoff, J., & Fishel, R. (1999). Molecular diagnostics of cancer predisposition: hereditary non-polyposis colorectal carcinoma and mismatch repair defects. *Biochimica et Biophysica Acta*, *1423*(3), O1–O10.

Boeva, V., Zinovyev, A., Bleakley, K., Vert, J. P., Janoueix-Lerosey, I., Delattre, O., & Barillot, E. (2011). Control-free calling of copy number alterations in deep-sequencing data using GC-content normalization. *Bioinformatics*, *27*(2), 268–269. https://doi.org/10.1093/bioinformatics/btq635

Boland, C. R., & Goel, A. (2010). Microsatellite Instability in Colorectal Cancer. *Gastroenterology*, *138*(6), 2073-2087.e3. https://doi.org/10.1053/j.gastro.2009.12.064

Boparai, K. S., Dekker, E., van Eeden, S., Polak, M. M., Bartelsman, J. F. W. M., Mathus–Vliegen, E. M. H., … van Noesel, C. J. M. (2008). Hyperplastic Polyps and Sessile Serrated Adenomas as a Phenotypic Expression of MYH-Associated Polyposis. *Gastroenterology*, *135*(6), 2014–2018.

https://doi.org/10.1053/j.gastro.2008.09.020

Borchert, G. M., Gilmore, B. L., Spengler, R. M., Xing, Y., Lanier, W., Bhattacharya, D., & Davidson, B. L. (2009). Adenosine deamination in human transcripts generates novel microRNA binding sites. *Human Molecular Genetics*, *18*(24), 4801–4807. https://doi.org/10.1093/hmg/ddp443

Bosetti, C., Rosato, V., Gallus, S., Cuzick, J., & La Vecchia, C. (2012). Aspirin and cancer risk: a quantitative review to 2011. *Annals of Oncology*, *23*(6), 1403–1415. https://doi.org/10.1093/annonc/mds113

Botteri, E., Iodice, S., Bagnardi, V., Raimondi, S., Lowenfels, A. B., & Maisonneuve, P. (2008). Smoking and Colorectal Cancer. *JAMA*, *300*(23), 2765. https://doi.org/10.1001/jama.2008.839

Boyle, T., Keegel, T., Bull, F., Heyworth, J., & Fritschi, L. (2012). Physical Activity and Risks of Proximal and Distal Colon Cancers: A Systematic Review and Meta-analysis. *JNCI: Journal of the National Cancer Institute*, *104*(20), 1548–1561. https://doi.org/10.1093/jnci/djs354

Brady, C. A., Jiang, D., Mello, S. S., Johnson, T. M., Jarvis, L. A., Kozak, M. M., … Attardi, L. D. (2011). Distinct p53 transcriptional programs dictate acute DNA-damage responses and tumor suppression. *Cell*, *145*(4), 571–583. https://doi.org/10.1016/j.cell.2011.03.035

Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R. L., Torre, L. A., & Jemal, A. (2018). Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians*. https://doi.org/10.3322/caac.21492

Brea-Fernandez, A. J., Fernandez-Rozadilla, C., Alvarez-Barona, M., Azuara, D., Ginesta, M. M., Clofent, J., … Ruiz-Ponte, C. (2017). Candidate predisposing germline copy number variants in early onset colorectal cancer patients. *Clinical and Translational Oncology*, *19*(5), 625–632. https://doi.org/10.1007/s12094-016-1576-z

Breasted, J. H., & University of Chicago. Oriental Institute. (1930). *The Edwin Smith surgical papyrus, published in facsimile and hieroglyphic transliteration with translation and commentary in two volumes*. University of Chicago, Oriental Institute.

Brenner, H., Kloor, M., & Pox, C. P. (2014). Colorectal cancer. *The Lancet*, *383*(9927), 1490–1502. https://doi.org/10.1016/S0140-6736(13)61649-9

Briggs, S., & Tomlinson, I. (2013). Germline and somatic polymerase ε and δ mutations define a new class of hypermutated colorectal and endometrial cancers. *The Journal of Pathology*, *230*(2), 148. https://doi.org/10.1002/PATH.4185

Broad Institute of MIT and Harvard. (2019). Picard tools. Retrieved from http://broadinstitute.github.io/picard

Brookes, A. J. (1999). The essence of SNPs. *Gene*, *234*(2), 177–186.

Buchanan, D. D., Clendenning, M., Zhuoer, L., Stewart, J. R., Joseland, S., Woodall, S., … Rosty, C. (2017). Lack of evidence for germline RNF43 mutations in patients with serrated polyposis syndrome from a large multinational study. *Gut*, *66*(6), 1170–1172. https://doi.org/10.1136/gutjnl-2016-312773

Burrell, R. A., McGranahan, N., Bartek, J., & Swanton, C. (2013). The causes and consequences of genetic heterogeneity in cancer evolution. *Nature*, *501*(7467), 338–345. https://doi.org/10.1038/nature12625

Burt, R. W., Leppert, M. F., Slattery, M. L., Samowitz, W. S., Spirio, L. N., Kerber, R. A., … White, R. L. (2004). Genetic testing and phenotype in a large kindred with attenuated familial adenomatous polyposis. *Gastroenterology*, *127*(2), 444–451.

Cairns, J. (1975). Mutation selection and the natural history of cancer. *Nature*, *255*(5505), 197–200. https://doi.org/10.1038/255197a0

Calin, G. A., Ferracin, M., Cimmino, A., Di Leva, G., Shimizu, M., Wojcik, S. E., … Croce, C. M. (2005). A MicroRNA Signature Associated with Prognosis and Progression in Chronic Lymphocytic Leukemia. *New England Journal of Medicine*, *353*(17), 1793–1801. https://doi.org/10.1056/NEJMoa050995

Calle, E. E., & Kaaks, R. (2004). Overweight, obesity and cancer: epidemiological evidence and proposed mechanisms. *Nature Reviews Cancer*, *4*(8), 579–591. https://doi.org/10.1038/nrc1408

Campbell, B. B., Light, N., Fabrizio, D., Zatzman, M., Fuligni, F., de Borja, R., … Shlien, A. (2017). Comprehensive Analysis of Hypermutation in Human Cancer. *Cell*, *171*(5), 1042-1056.e10. https://doi.org/10.1016/j.cell.2017.09.048

Camps, J., Armengol, G., del Rey, J., Lozano, J. J., Vauhkonen, H., Prat, E., … Miró, R.

(2006). Genome-wide differences between microsatellite stable and unstable colorectal tumors. *Carcinogenesis*, *27*(3), 419–428. https://doi.org/10.1093/carcin/bgi244

Camps, J., Grade, M., Nguyen, Q. T., Hormann, P., Becker, S., Hummon, A. B., … Ried, T. (2008). Chromosomal Breakpoints in Primary Colon Cancer Cluster at Sites of Structural Variants in the Genome. *Cancer Research*, *68*(5), 1284–1295. https://doi.org/10.1158/0008-5472.CAN-07-2864

Camps, J., Pitt, J. J., Emons, G., Hummon, A. B., Case, C. M., Grade, M., … Ried, T. (2013). Genetic amplification of the NOTCH modulator LNX2 upregulates the WNT/ß-catenin pathway in colorectal cancer. *Cancer Research*, *73*(6), 2003–2013. https://doi.org/10.1158/0008-5472.CAN-12-3159

Camps, J., Tri Nguyen, Q., Padilla-Nash, H. M., Knutsen, T., McNeil, N. E., Wangsa, D., … Difilippantonio, M. J. (2009). Integrative genomics reveals mechanisms of copy number alterations responsible for transcriptional deregulation in colorectal cancer. *Genes, Chromosomes and Cancer*, *48*(11), 1002–1017. https://doi.org/10.1002/gcc.20699

Carethers, J. M., & Jung, B. H. (2015). Genetics and Genetic Biomarkers in Sporadic Colorectal Cancer. *Gastroenterology*, *149*(5), 1177–1190. https://doi.org/10.1053/j.gastro.2015.06.047

Carvalho, B, Ouwerkerk, E., Meijer, G. A., & Ylstra, B. (2004). High resolution microarray comparative genomic hybridisation analysis using spotted oligonucleotides. *Journal of Clinical Pathology*, *57*(6), 644–646. https://doi.org/10.1136/JCP.2003.013029

Carvalho, B, Postma, C., Mongera, S., Hopmans, E., Diskin, S., van de Wiel, M. A., … Meijer, G. A. (2009). Multiple putative oncogenes at the chromosome 20q amplicon contribute to colorectal adenoma to carcinoma progression. *Gut*, *58*(1), 79–89. https://doi.org/10.1136/gut.2007.143065

Carvalho, Beatriz, Diosdado, B., Terhaar Sive Droste, J. S., Bolijn, A. S., Komor, M. A., de Wit, M., … Meijer, G. A. (2018). Evaluation of Cancer-Associated DNA Copy Number Events in Colorectal (Advanced) Adenomas. *Cancer Prevention Research (Philadelphia, Pa.)*, *11*(7), 403–412. https://doi.org/10.1158/1940-6207.CAPR-17-0317

Carvalho, C. M. B., & Lupski, J. R. (2016). Mechanisms underlying structural variant formation in genomic disorders. *Nature Reviews Genetics*, *17*(4), 224–238. https://doi.org/10.1038/nrg.2015.25

Casper, J., Zweig, A. S., Villarreal, C., Tyner, C., Speir, M. L., Rosenbloom, K. R., … Kent, W. J. (2017). The UCSC Genome Browser database: 2018 update. *Nucleic Acids Research*, *46*(D1), D762–D769. https://doi.org/10.1093/nar/gkx1020

Caspersson, T., Zech, L., & Johansson, C. (1970). Analysis of human metaphase chromosome set by aid of DNA-binding fluorescent agents. *Experimental Cell Research*, *62*(2), 490–492.

Celsus, C. A. (50AD). *De Medicina*.

Chaisson, M. J. P., Huddleston, J., Dennis, M. Y., Sudmant, P. H., Malig, M., Hormozdiari, F., … Eichler, E. E. (2015). Resolving the complexity of the human genome using single-molecule sequencing. *Nature*, *517*(7536), 608–611. https://doi.org/10.1038/nature13907

Chaisson, M. J. P., Wilson, R. K., & Eichler, E. E. (2015). Genetic variation and the de novo assembly of human genomes. *Nature Reviews Genetics*, *16*(11), 627–640. https://doi.org/10.1038/nrg3933

Chan, A. T., & Giovannucci, E. L. (2010). Primary Prevention of Colorectal Cancer. *Gastroenterology*, *138*(6), 2029-2043.e10. https://doi.org/10.1053/j.gastro.2010.01.057

Chang, W., Cheng, J., Allaire, J., Xie, Y., & McPherson, J. (2018). *shiny: Web Application Framework for R* (p. R package version 1.1.0). p. R package version 1.1.0.

Chavez, J. A., & Summers, S. A. (2010). Lipid oversupply, selective insulin resistance, and lipotoxicity: Molecular mechanisms. *Biochimica et Biophysica Acta (BBA) - Molecular and Cell Biology of Lipids*, *1801*(3), 252–265. https://doi.org/10.1016/j.bbalip.2009.09.015

Check Hayden, E. (2014). Technology: The $1,000 genome. *Nature*, *507*(7492), 294–295. https://doi.org/10.1038/507294a

Chubb, D., Broderick, P., Dobbins, S. E., Frampton, M., Kinnersley, B., Penegar, S., … Houlston, R. S. (2016). Rare disruptive mutations and their contribution to the heritable risk of colorectal cancer. *Nature Communications*, *7*(1), 11883. https://doi.org/10.1038/ncomms11883

Cingolani, P., Patel, V. M., Coon, M., Nguyen, T., Land, S. J., Ruden, D. M., & Lu, X. (2012). Using Drosophila melanogaster as a model for genotoxic chemical mutational studies with a new program, SnpSift. *Frontiers in Genetics*, *3*(MAR), 35. https://doi.org/10.3389/fgene.2012.00035

Cingolani, P., Platts, A., Wang, L. L., Coon, M., Nguyen, T., Wang, L., … Ruden, D. M. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w 1118; iso-2; iso-3. *Fly*, *6*(2), 80–92. https://doi.org/10.4161/fly.19695

Ciriello, G., Miller, M. L., Aksoy, B. A., Senbabaoglu, Y., Schultz, N., & Sander, C. (2013). Emerging landscape of oncogenic signatures across human cancers. *Nature Genetics*, *45*(10), 1127–1133. https://doi.org/10.1038/ng.2762

Cirulli, E. T., & Goldstein, D. B. (2010). Uncovering the roles of rare variants in common disease through whole-genome sequencing. *Nature Reviews Genetics*, *11*(6), 415–425. https://doi.org/10.1038/nrg2779

Cleary, S. P., Cotterchio, M., Jenkins, M. A., Kim, H., Bristow, R., Green, R., … Gallinger, S. (2009). Germline MutY human homologue mutations and colorectal cancer: a multisite case-control study. *Gastroenterology*, *136*(4), 1251–1260. https://doi.org/10.1053/j.gastro.2008.12.050

Coe, B. P., Witherspoon, K., Rosenfeld, J. A., van Bon, B. W. M., Vulto-van Silfhout, A. T., Bosco, P., … Eichler, E. E. (2014). Refining analyses of copy number variation identifies specific genes associated with developmental delay. *Nature Genetics*, *46*(10), 1063–1071. https://doi.org/10.1038/ng.3092

Conrad, D. F., Pinto, D., Redon, R., Feuk, L., Gokcumen, O., Zhang, Y., … Hurles, M. E. (2010). Origins and functional impact of copy number variation in the human genome. *Nature*, *464*(7289), 704–712. https://doi.org/10.1038/nature08516

Consortium, I. H. G. S. (2001). Initial sequencing and analysis of the human genome. *Nature*, *409*(6822), 860–921. https://doi.org/10.1038/35057062

Consortium, T. E. P. (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature*, *489*(7414), 57–74. https://doi.org/10.1038/nature11247

Cuajungco, M. P., Podevin, W., Valluri, V. K., Bui, Q., Nguyen, V. H., & Taylor, K. (2012). Abnormal accumulation of human transmembrane (TMEM)-176A and 176B proteins is associated with cancer pathology. *Acta Histochemica*, *114*(7), 705–712. https://doi.org/10.1016/j.acthis.2011.12.006

Cunningham, D., Atkin, W., Lenz, H.-J., Lynch, H. T., Minsky, B., Nordlinger, B., & Starling, N. (2010). Colorectal cancer. *The Lancet*, *375*(9719), 1030–1047. https://doi.org/10.1016/S0140-6736(10)60353-4

Cunningham, D., Humblet, Y., Siena, S., Khayat, D., Bleiberg, H., Santoro, A., … Van Cutsem, E. (2004). Cetuximab Monotherapy and Cetuximab plus Irinotecan in Irinotecan-Refractory Metastatic Colorectal Cancer. *New England Journal of Medicine*, *351*(4), 337–345. https://doi.org/10.1056/NEJMoa033025

Davies, R. J., Miller, R., & Coleman, N. (2005). Colorectal cancer screening: prospects for molecular stool analysis. *Nature Reviews Cancer*, *5*(3), 199–209. https://doi.org/10.1038/nrc1569

Davoli, T., & de Lange, T. (2011). The Causes and Consequences of Polyploidy in Normal Development and Cancer. *Annual Review of Cell and Developmental Biology*, *27*(1), 585–610. https://doi.org/10.1146/annurev-cellbio-092910-154234

Davoli, T., Denchi, E. L., & de Lange, T. (2010). Persistent telomere damage induces bypass of mitosis and tetraploidy. *Cell*, *141*(1), 81–93. https://doi.org/10.1016/j.cell.2010.01.031

Davoli, T., Uno, H., Wooten, E. C., & Elledge, S. J. (2017). Tumor aneuploidy correlates with markers of immune evasion and with reduced response to immunotherapy. *Science*, *355*(6322), eaaf8399. https://doi.org/10.1126/science.aaf8399

de Voer, R. M., Geurts van Kessel, A., Weren, R. D. A., Ligtenberg, M. J. L., Smeets, D., Fu, L., … Kuiper, R. P. (2013). Germline Mutations in the Spindle Assembly Checkpoint Genes BUB1 and BUB3 Are Risk Factors for Colorectal Cancer. *Gastroenterology*, *145*(3), 544–547. https://doi.org/10.1053/j.gastro.2013.06.001

de Voer, R. M., Hahn, M.-M., Mensenkamp, A. R., Hoischen, A., Gilissen, C., Henkes, A., … Kuiper, R. P. (2015). Deleterious

Germline BLM Mutations and the Risk for Early-onset Colorectal Cancer. *Scientific Reports*, *5*(1), 14060. https://doi.org/10.1038/srep14060

de Voer, R. M., Hahn, M.-M., Weren, R. D. A., Mensenkamp, A. R., Gilissen, C., van Zelst-Stams, W. A., … Kuiper, R. P. (2016). Identification of Novel Candidate Genes for Early-Onset Colorectal Cancer Susceptibility. *PLOS Genetics*, *12*(2), e1005880. https://doi.org/10.1371/journal.pgen.1005880

DeRycke, M. S., Gunawardena, S. R., Middha, S., Asmann, Y. W., Schaid, D. J., McDonnell, S. K., … Goode, E. L. (2013). Identification of Novel Variants in Colorectal Cancer Families by High-Throughput Exome Sequencing. *Cancer Epidemiology Biomarkers & Prevention*, *22*(7), 1239–1251. https://doi.org/10.1158/1055-9965.EPI-12-1226

Dickinson, B. T., Kisiel, J., Ahlquist, D. A., & Grady, W. M. (2015). Molecular markers for colorectal cancer screening. *Gut*, *64*(9), 1485–1494. https://doi.org/10.1136/gutjnl-2014-308075

Dienstmann, R., Vermeulen, L., Guinney, J., Kopetz, S., Tejpar, S., & Tabernero, J. (2017). Consensus molecular subtypes and the evolution of precision medicine in colorectal cancer. *Nature Reviews Cancer*, *17*(2), 79–92. https://doi.org/10.1038/nrc.2016.126

Diep, C. B., Kleivi, K., Ribeiro, F. R., Teixeira, M. R., Lindgjaerde, O. C., & Lothe, R. A. (2006). The order of genetic events associated with colorectal cancer progression inferred from meta-analysis of copy number changes. *Genes, Chromosomes and Cancer*, *45*(1), 31–41. https://doi.org/10.1002/gcc.20261

Ding, L., Bailey, M. H., Porta-Pardo, E., Thorsson, V., Colaprico, A., Bertrand, D., … Mariamidze, A. (2018). Perspective on Oncogenic Processes at the End of the Beginning of Cancer Genomics. *Cell*, *173*(2), 305-320.e10. https://doi.org/10.1016/j.cell.2018.03.033

Dixon, J. R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., … Ren, B. (2012). Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, *485*(7398), 376–380. https://doi.org/10.1038/nature11082

Dobbins, S. E., Broderick, P., Chubb, D., Kinnersley, B., Sherborne, A. L., & Houlston, R. S. (2016). Undefined familial colorectal cancer and the role of pleiotropism in cancer susceptibility genes. *Familial Cancer*, *15*(4), 593–599. https://doi.org/10.1007/s10689-016-9914-4

Douglas, E. J., Fiegler, H., Rowan, A., Halford, S., Bicknell, D. C., Bodmer, W., … Carter, N. P. (2004). Array Comparative Genomic Hybridization Analysis of Colorectal Cancer Cell Lines and Primary Carcinomas. *Cancer Research*, *64*(14), 4817–4825. https://doi.org/10.1158/0008-5472.CAN-04-0328

Du, P., Kibbe, W. A., & Lin, S. M. (2008). lumi: A pipeline for processing Illumina microarray. *Bioinformatics*, *24*(13), 1547–1548. https://doi.org/10.1093/bioinformatics/btn224

Ellingford, J. M., Campbell, C., Barton, S., Bhaskar, S., Gupta, S., Taylor, R. L., … Black, G. C. M. (2017). Validation of copy number variation analysis for next-generation sequencing diagnostics. *European Journal of Human Genetics*, *25*(6), 719–724. https://doi.org/10.1038/ejhg.2017.42

Esteban-Jurado, C., Franch-Expósito, S., Muñoz, J., Ocaña, T., Carballal, S., López-Cerón, M., … Castellví-Bel, S. (2016). The Fanconi anemia DNA damage repair pathway in the spotlight for germline predisposition to colorectal cancer. *European Journal of Human Genetics*, *24*(10), 1501–1505. https://doi.org/10.1038/ejhg.2016.44

Esteban-Jurado, C., Giménez-Zaragoza, D., Muñoz, J., Franch-Expósito, S., Álvarez-Barona, M., Ocaña, T., … Castellví-Bel, S. (2017). POLE and POLD1 screening in 155 patients with multiple polyps and early-onset colorectal cancer. *Oncotarget*, *8*(16). https://doi.org/10.18632/oncotarget.15810

Esteban-Jurado, C., Vila-Casadesús, M., Garre, P., Lozano, J. J. J. J., Pristoupilova, A., Beltran, S., … Castellví-Bel, S. (2015). Whole-exome sequencing identifies rare pathogenic variants in new predisposition genes for familial colorectal cancer. *Genetics in Medicine*, *17*(2), 131–142. https://doi.org/10.1038/gim.2014.89

Etemadmoghadam, D., deFazio, A., Beroukhim, R., Mermel, C., George, J., Getz, G., … AOCS Study Group, D. (2009). Integrated genome-wide DNA copy number and expression analysis identifies distinct mechanisms of primary chemoresistance in ovarian carcinomas. *Clinical Cancer Research : An Official Journal of the American Association for Cancer*

Research, *15*(4), 1417–1427. https://doi.org/10.1158/1078-0432.CCR-08-1564

Fearon, E R, & Vogelstein, B. (1990). A genetic model for colorectal tumorigenesis. *Cell*, *61*(5), 759–767.

Fearon, Eric R. (1995). Molecular Genetics of Colorectal Cancer. *Annals of the New York Academy of Sciences*, *768*(1), 101–110. https://doi.org/10.1111/j.1749-6632.1995.tb12114.x

Fearon, Eric R. (2011). Molecular Genetics of Colorectal Cancer. *Annual Review of Pathology: Mechanisms of Disease*, *6*(1), 479–507. https://doi.org/10.1146/annurev-pathol-011110-130235

Fernandez-Rozadilla, C, Brea-Fernández, A., Bessa, X., Álvarez-Urturi, C., Abulí, A., Clofent, J., … Ruiz-Ponte, C. (2013). BMPR1A mutations in early-onset colorectal cancer with mismatch repair proficiency. *Clinical Genetics*, *84*(1), 94–96. https://doi.org/10.1111/cge.12023

Fernandez-Rozadilla, Ceres, Cazier, J.-B., Tomlinson, I. P., Carvajal-carmona, L. G., Palles, C., Lamas, M. J., … Ruiz-Ponte, C. (2013). A colorectal cancer genome-wide association study in a Spanish cohort identifies two variants associated with colorectal cancer risk at 1p33 and 8p12. *BMC Genomics*, *14*(1), 55. https://doi.org/10.1186/1471-2164-14-55

Feuk, L., Carson, A. R., & Scherer, S. W. (2006). Structural variation in the human genome. *Nature Reviews Genetics*, *7*(2), 85–97. https://doi.org/10.1038/nrg1767

Fewings, E., Larionov, A., Redman, J., Goldgraben, M. A., Scarth, J., Richardson, S., … Tischkowitz, M. (2018). Germline pathogenic variants in PALB2 and other cancer-predisposing genes in families with hereditary diffuse gastric cancer without CDH1 mutation: a whole-exome sequencing study. *The Lancet. Gastroenterology & Hepatology*, *3*(7), 489–498. https://doi.org/10.1016/S2468-1253(18)30079-7

Firestein, R., Bass, A. J., Kim, S. Y., Dunn, I. F., Silver, S. J., Guney, I., … Hahn, W. C. (2008). CDK8 is a colorectal cancer oncogene that regulates beta-catenin activity. *Nature*, *455*(7212), 547–551. https://doi.org/10.1038/nature07179

Fishel, R., Lescoe, M. K., Rao, M. R., Copeland, N. G., Jenkins, N. A., Garber, J., … Kolodner, R. (1993). The human mutator gene homolog MSH2 and its association with hereditary nonpolyposis colon cancer. *Cell*, *75*(5), 1027–1038.

Flora, M., Piana, S., Bassano, C., Bisagni, A., De Marco, L., Ciarrocchi, A., … Bisagni, G. (2012). Epidermal growth factor receptor (EGFR) gene copy number in colorectal adenoma-carcinoma progression. *Cancer Genetics*, *205*(12), 630–635. https://doi.org/10.1016/j.cancergen.2012.10.005

Fodde, R., Kuipers, J., Rosenberg, C., Smits, R., Kielman, M., Gaspar, C., … Clevers, H. (2001). Mutations in the APC tumour suppressor gene cause chromosomal instability. *Nature Cell Biology*, *3*(4), 433–438. https://doi.org/10.1038/35070129

Fodde, R., Smits, R., & Clevers, H. (2001). APC, Signal transduction and genetic instability in colorectal cancer. *Nature Reviews Cancer*, *1*(1), 55–67. https://doi.org/10.1038/35094067

Franke, M., Ibrahim, D. M., Andrey, G., Schwarzer, W., Heinrich, V., Schöpflin, R., … Mundlos, S. (2016). Formation of new chromatin domains determines pathogenicity of genomic duplications. *Nature*, *538*(7624), 265–269. https://doi.org/10.1038/nature19800

Fromer, M., Moran, J. L., Chambert, K., Banks, E., Bergen, S. E., Ruderfer, D. M., … Purcell, S. M. (2012). Discovery and statistical genotyping of copy-number variation from whole-exome sequencing depth. *American Journal of Human Genetics*, *91*(4), 597–607. https://doi.org/10.1016/j.ajhg.2012.08.005

Gambin, T., Yuan, B., Bi, W., Liu, P., Rosenfeld, J. A., Coban-Akdemir, Z., … Stankiewicz, P. (2017). Identification of novel candidate disease genes from de novo exonic copy number variants. *Genome Medicine*, *9*(1), 83. https://doi.org/10.1186/s13073-017-0472-7

Garraway, L. A., & Lander, E. S. (2013). Lessons from the Cancer Genome. *Cell*, *153*(1), 17–37. https://doi.org/10.1016/j.cell.2013.03.002

Ge, C., Wu, S., Wang, W., Liu, Z., Zhang, J., Wang, Z., … Song, X. (2015). miR-942 promotes cancer stem cell-like traits in esophageal squamous cell carcinoma through activation of Wnt/ß-catenin signalling pathway. *Oncotarget*, *6*(13), 10964–10977. https://doi.org/10.18632/oncotarget.3696

Gillen, C. D., Walmsley, R. S., Prior, P., Andrews, H. A., & Allan, R. N. (1994). Ulcerative colitis and Crohn's disease: a comparison of the colorectal cancer risk in extensive colitis. *Gut*, *35*(11), 1590–

1592.

Gingras, D., & Béliveau, R. (2011). Colorectal cancer prevention through dietary and lifestyle modifications. *Cancer Microenvironment : Official Journal of the International Cancer Microenvironment Society*, *4*(2), 133–139. https://doi.org/10.1007/s12307-010-0060-5

Giovannucci, E. (2002). Epidemiologic Studies of Folate and Colorectal Neoplasia: a Review. *The Journal of Nutrition*, *132*(8), 2350S-2355S. https://doi.org/10.1093/jn/132.8.2350S

González, J. R., Rodríguez-Santiago, B., Cáceres, A., Pique-Regi, R., Rothman, N., Chanock, S. J., … Pérez-Jurado, L. A. (2011). A fast and accurate method to detect allelic genomic imbalances underlying mosaic rearrangements using SNP array data. *BMC Bioinformatics*, *12*, 166. https://doi.org/10.1186/1471-2105-12-166

Goodier, J. L., & Kazazian, H. H. (2008). Retrotransposons revisited: the restraint and rehabilitation of parasites. *Cell*, *135*(1), 23–35. https://doi.org/10.1016/j.cell.2008.09.022

Goodwin, S., McPherson, J. D., & McCombie, W. R. (2016). Coming of age: ten years of next-generation sequencing technologies. *Nature Reviews Genetics*, *17*(6), 333–351. https://doi.org/10.1038/nrg.2016.49

Gordon, D. J., Resio, B., & Pellman, D. (2012). Causes and consequences of aneuploidy in cancer. *Nature Reviews Genetics*, *13*(3), 189–203. https://doi.org/10.1038/nrg3123

Grade, M., Ghadimi, B. M., Varma, S., Simon, R., Wangsa, D., Barenboim-Stapleton, L., … Difilippantonio, M. J. (2006). Aneuploidy-Dependent Massive Deregulation of the Cellular Transcriptome and Apparent Divergence of the Wnt/ß-catenin Signaling Pathway in Human Rectal Carcinomas. *Cancer Research*, *66*(1), 267–282. https://doi.org/10.1158/0008-5472.CAN-05-2533

Grade, M., Hörmann, P., Becker, S., Hummon, A. B., Wangsa, D., Varma, S., … Ried, T. (2007). Gene Expression Profiling Reveals a Massive, Aneuploidy-Dependent Transcriptional Deregulation and Distinct Differences between Lymph Node–Negative and Lymph Node–Positive Colon Carcinomas. *Cancer Research*, *67*(1), 41–56. https://doi.org/10.1158/0008-5472.CAN-06-1514

Grady, W. M. (2003). Genetic testing for high-risk colon cancer patients. *Gastroenterology*, *124*(6), 1574–1594.

Grasso, C. S., Wu, Y.-M., Robinson, D. R., Cao, X., Dhanasekaran, S. M., Khan, A. P., … Tomlins, S. A. (2012). The mutational landscape of lethal castration-resistant prostate cancer. *Nature*, *487*(7406), 239–243. https://doi.org/10.1038/nature11125

Grilley, M., Holmes, J., Yashar, B., & Modrich, P. (1990). Mechanisms of DNA-mismatch correction. *Mutation Research*, *236*(2–3), 253–267.

Groden, J., Thliveris, A., Samowitz, W., Carlson, M., Gelbert, L., Albertsen, H., … Robertson, M. (1991). Identification and characterization of the familial adenomatous polyposis coli gene. *Cell*, *66*(3), 589–600.

Grossman, R. L., Heath, A. P., Ferretti, V., Varmus, H. E., Lowy, D. R., Kibbe, W. A., & Staudt, L. M. (2016). Toward a Shared Vision for Cancer Genomic Data. *New England Journal of Medicine*, *375*(12), 1109–1112. https://doi.org/10.1056/NEJMp1607591

Guinney, J., Dienstmann, R., Wang, X., de Reyniès, A., Schlicker, A., Soneson, C., … Tejpar, S. (2015). The consensus molecular subtypes of colorectal cancer. *Nature Medicine*, *21*(11), 1350–1356. https://doi.org/10.1038/nm.3967

Guo, Y., Sheng, Q., Samuels, D. C., Lehmann, B., Bauer, J. A., Pietenpol, J., & Shyr, Y. (2013). Comparative study of exome copy number variation estimation tools using array comparative genomic hybridization as control. *BioMed Research International*, *2013*, 915636. https://doi.org/10.1155/2013/915636

Gylfe, A. E., Katainen, R., Kondelin, J., Tanskanen, T., Cajuso, T., Hänninen, U., … Aaltonen, L. A. (2013). Eleven Candidate Susceptibility Genes for Common Familial Colorectal Cancer. *PLoS Genetics*, *9*(10), e1003876. https://doi.org/10.1371/journal.pgen.1003876

Habermann, J. K., Paulsen, U., Roblick, U. J., Upender, M. B., McShane, L. M., Korn, E. L., … Ried, T. (2007). Stage-specific alterations of the genome, transcriptome, and proteome during colorectal carcinogenesis. *Genes, Chromosomes and Cancer*, *46*(1), 10–26. https://doi.org/10.1002/gcc.20382

Hahn, M. M., de Voer, R. M., Hoogerbrugge, N., Ligtenberg, M. J. L., Kuiper, R. P., & van Kessel, A. G. (2016). The genetic heterogeneity of colorectal cancer

predisposition - guidelines for gene discovery. *Cellular Oncology*, *39*(6), 491–510. https://doi.org/10.1007/s13402-016-0284-6

Hajdu, S. I. (2011). A note from history: Landmarks in history of cancer, part 1. *Cancer*, *117*(5), 1097–1102. https://doi.org/10.1002/cncr.25553

Hanahan, D, & Weinberg, R. A. (2000). The hallmarks of cancer. *Cell*, *100*(1), 57–70. https://doi.org/10.1016/S0092-8674(00)81683-9

Hanahan, Douglas, & Weinberg, R. A. (2011). Hallmarks of cancer: The next generation. *Cell*, *144*(5), 646–674. https://doi.org/10.1016/j.cell.2011.02.013

Hansemann, D. (1890). Ueber asymmetrische Zelltheilung in Epithelkrebsen und deren biologische Bedeutung. *Archiv Für Pathologische Anatomie Und Physiologie Und Für Klinische Medicin*, *119*(2), 299–326. https://doi.org/10.1007/BF01882039

Haseman, J. K., & Elston, R. C. (1972). The investigation of linkage between a quantitative trait and a marker locus. *Behavior Genetics*, *2*(1), 3–19.

Hastings, P. J., Ira, G., & Lupski, J. R. (2009). A microhomology-mediated break-induced replication model for the origin of human copy number variation. *PLoS Genetics*, *5*(1), e1000327. https://doi.org/10.1371/journal.pgen.1000327

Hatakeyama, S. (2011). TRIM proteins and cancer. *Nature Reviews Cancer*, *11*(11), 792–804. https://doi.org/10.1038/nrc3139

Hawk, E. T., & Levin, B. (2005). Colorectal cancer prevention. *Journal of Clinical Oncology : Official Journal of the American Society of Clinical Oncology*, *23*(2), 378–391. https://doi.org/10.1200/JCO.2005.08.097

Hehir-Kwa, J. Y., Pfundt, R., & Veltman, J. A. (2015). Exome sequencing and whole genome sequencing for the detection of copy number variation. *Expert Review of Molecular Diagnostics*, *7159*(April), 1–10. https://doi.org/10.1586/14737159.2015.1053467

Hermsen, M., Postma, C., Baak, J., Weiss, M., Rapallo, A., Sciutto, A., … Meijer, G. (2002). Colorectal adenoma to carcinoma progression follows multiple pathways of chromosomal instability. *Gastroenterology*, *123*(4), 1109–1119. https://doi.org/10.1053/gast.2002.36051

Hesson, L. B., Cooper, W. N., & Latif, F. (2007). Evaluation of the 3p21.3 tumour-suppressor gene cluster. *Oncogene*, *26*(52), 7283–7301. https://doi.org/10.1038/sj.onc.1210547

Hoadley, K. A., Yau, C., Hinoue, T., Wolf, D. M., Lazar, A. J., Drill, E., … Mariamidze, A. (2018). Cell-of-Origin Patterns Dominate the Molecular Classification of 10,000 Tumors from 33 Types of Cancer. *Cell*, *173*(2), 291-304.e6. https://doi.org/10.1016/j.cell.2018.03.022

Hodis, E., Watson, I. R., Kryukov, G. V, Arold, S. T., Imielinski, M., Theurillat, J.-P., … Chin, L. (2012). A landscape of driver mutations in melanoma. *Cell*, *150*(2), 251–263. https://doi.org/10.1016/j.cell.2012.06.024

Holland, A. J., & Cleveland, D. W. (2009). Boveri revisited: chromosomal instability, aneuploidy and tumorigenesis. *Nature Reviews Molecular Cell Biology*, *10*(7), 478–487. https://doi.org/10.1038/nrm2718

Holland, A. J., & Cleveland, D. W. (2012). Losing balance: The origin and impact of aneuploidy in cancer. *EMBO Reports*, *13*(6), 501–514. https://doi.org/10.1038/embor.2012.55

Hrašovec, S., Hauptman, N., Glavač, D., Jelenc, F., & Ravnik-Glavač, M. (2013). TMEM25 is a candidate biomarker methylated and down-regulated in colorectal cancer. *Disease Markers*, *34*(2), 93–104. https://doi.org/10.3233/DMA-120948

Huang, K.-L., Mashl, R. J., Wu, Y., Ritter, D. I., Wang, J., Oh, C., … Ding, L. (2018). Pathogenic Germline Variants in 10,389 Adult Cancers. *Cell*, *173*(2), 355-370.e14. https://doi.org/10.1016/j.cell.2018.03.039

Human Genome Sequencing Consortium, I. (2004). Finishing the euchromatic sequence of the human genome. *Nature*, *431*(7011), 931–945. https://doi.org/10.1038/nature03001

Huxley, R. R., Ansary-Moghaddam, A., Clifton, P., Czernichow, S., Parr, C. L., & Woodward, M. (2009). The impact of dietary and lifestyle risk factors on risk of colorectal cancer: A quantitative overview of the epidemiological evidence. *International Journal of Cancer*, *125*(1), 171–180. https://doi.org/10.1002/ijc.24343

Huyghe, J. R., Bien, S. A., Harrison, T. A., Kang, H. M., Chen, S., Schmit, S. L., … Peters, U. (2018). Discovery of common and rare genetic risk variants for colorectal cancer. *Nature Genetics*, 1. https://doi.org/10.1038/s41588-018-0286-6

Iafrate, A. J., Feuk, L., Rivera, M. N., Listewnik, M. L., Donahoe, P. K., Qi, Y.,

… Lee, C. (2004). Detection of large-scale variation in the human genome. *Nature Genetics*, *36*(9), 949–951. https://doi.org/10.1038/ng1416

Iglesias, D., Fernández-Peralta, A. M., Nejda, N., Daimiel, L., Azcoita, M. M., Oliart, S., & González-Aguilera, J. J. (2006). RIS1, a gene with trinucleotide repeats, is a target in the mutator pathway of colorectal carcinogenesis. *Cancer Genetics and Cytogenetics*, *167*(2), 138–144. https://doi.org/10.1016/j.cancergencyto.2005.12.002

Illumina CASAVA-1.8 FASTQ Filter. (2011). Retrieved from http://cancan.cshl.edu/labmembers/gordon/fastq_illumina_filter/

Ionov, Y., Peinado, M. A., Malkhosyan, S., Shibata, D., & Perucho, M. (1993). Ubiquitous somatic mutations in simple repeated sequences reveal a new mechanism for colonic carcinogenesis. *Nature*, *363*(6429), 558–561. https://doi.org/10.1038/363558a0

Ira, G., & Haber, J. E. (2002). Characterization of RAD51-independent break-induced replication that acts preferentially with short homologous sequences. *Molecular and Cellular Biology*, *22*(18), 6384–6392. https://doi.org/10.1128/MCB.22.18.6384-6392.2002

Iuliano, R., Vismara, M. F. M., Dattilo, V., Trapasso, F., Baudi, F., & Perrotti, N. (2013). The role of microRNAs in cancer susceptibility. *BioMed Research International*, *2013*, 591931. https://doi.org/10.1155/2013/591931

Jaeger, E., Leedham, S., Lewis, A., Segditsas, S., Becker, M., Cuadrado, P. R., … Tomlinson, I. (2012). Hereditary mixed polyposis syndrome is caused by a 40-kb upstream duplication that leads to increased and ectopic expression of the BMP antagonist GREM1. *Nature Genetics*, *44*(6), 699–703. https://doi.org/10.1038/ng.2263

Jasperson, K. W., Tuohy, T. M., Neklason, D. W., & Burt, R. W. (2010). Hereditary and Familial Colon Cancer. *Gastroenterology*, *138*(6), 2044–2058. https://doi.org/10.1053/j.gastro.2010.01.054

Jass, J. R. (2008). Colorectal polyposes: From phenotype to diagnosis. *Pathology - Research and Practice*, *204*(7), 431–447. https://doi.org/10.1016/j.prp.2008.03.008

Jazdzewski, K., Murray, E. L., Franssila, K., Jarzab, B., Schoenberg, D. R., & de la Chapelle, A. (2008). Common SNP in pre-miR-146a decreases mature miR expression and predisposes to papillary thyroid carcinoma. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(20), 7269–7274. https://doi.org/10.1073/pnas.0802682105

Jelsig, A., Qvist, N., Brusgaard, K., Nielsen, C., Hansen, T., & Ousager, L. (2014). Hamartomatous polyposis syndromes: A review. *Orphanet Journal of Rare Diseases*, *9*(1), 101. https://doi.org/10.1186/1750-1172-9-101

Jiang, Y., & Price, D. H. (2004). Rescue of the TTF2 knockdown phenotype with an siRNA-resistant replacement vector. *Cell Cycle*, *3*(9), 1151–1153. https://doi.org/10.4161/cc.3.9.1151

Jones, A. M., Douglas, E. J., Halford, S. E., Fiegler, H., Gorman, P. A., Roylance, R. R., … Tomlinson, I. P. M. (2005). Array-CGH analysis of microsatellite-stable, near-diploid bowel cancers and comparison with other types of colorectal carcinoma. *Oncogene*, *24*(1), 118–129. https://doi.org/10.1038/sj.onc.1208194

Kadalayil, L., Rafiq, S., Rose-Zerilli, M. J. J., Pengelly, R. J., Parker, H., Oscier, D., … Collins, A. (2015). Exome sequence read depth methods for identifying copy number changes. *Briefings in Bioinformatics*, *16*(3), 380–392. https://doi.org/10.1093/bib/bbu027

Kallioniemi, A., Kallioniemi, O. P., Sudar, D., Rutovitz, D., Gray, J. W., Waldman, F., & Pinkel, D. (1992). Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors. *Science (New York, N.Y.)*, *258*(5083), 818–821.

Kambara, T., Simms, L. A., Whitehall, V. L. J., Spring, K. J., Wynter, C. V. A., Walsh, M. D., … Leggett, B. A. (2004). BRAF mutation is associated with DNA methylation in serrated polyps and cancers of the colorectum. *Gut*, *53*(8), 1137–1144. https://doi.org/10.1136/gut.2003.037671

Katona, B. W., Yurgelun, M. B., Garber, J. E., Offit, K., Domchek, S. M., Robson, M. E., & Stadler, Z. K. (2018). A counseling framework for moderate-penetrance colorectal cancer susceptibility genes. *Genetics in Medicine*, *20*(11), 1324–1327. https://doi.org/10.1038/gim.2018.12

Kazazian, H. H., & Moran, J. V. (1998). The impact of L1 retrotransposons on the human genome. *Nature Genetics*, *19*(1), 19–24. https://doi.org/10.1038/ng0598-19

Keating, G. M. (2010). Panitumumab. *Drugs*,

70(8), 1059–1078. https://doi.org/10.2165/11205090-000000000-00000

Keenan, T. E., Burke, K. P., & Van Allen, E. M. (2019). Genomic correlates of response to immune checkpoint blockade. *Nature Medicine*, *25*(3), 389–402. https://doi.org/10.1038/s41591-019-0382-x

Kent, W. J., Sugnet, C. W., Furey, T. S., Roskin, K. M., Pringle, T. H., Zahler, A. M., & Haussler, D. (2002). The human genome browser at UCSC. *Genome Research*, *12*(6), 996–1006. https://doi.org/10.1101/gr.229102

Kinnersley, B., Chubb, D., Dobbins, S. E., Frampton, M., Buch, S., Timofeeva, M. N., … Houlston, R. S. (2016). Correspondence: SEMA4A variation and risk of colorectal cancer. *Nature Communications*, *7*(1), 10611. https://doi.org/10.1038/ncomms10611

Kinzler, K. W., Nilbert, M. C., Su, L. K., Vogelstein, B., Bryan, T. M., Levy, D. B., … McKechnie, D. (1991). Identification of FAP locus genes from chromosome 5q21. *Science (New York, N.Y.)*, *253*(5020), 661–665.

Kinzler, K. W., & Vogelstein, B. (1996). Lessons from hereditary colorectal cancer. *Cell*, *87*(2), 159–170.

Knudsen, A. L., Bisgaard, M. L., & Bülow, S. (2003). Attenuated familial adenomatous polyposis (AFAP). A review of the literature. *Familial Cancer*, *2*(1), 43–55.

Knudson, A. G. (1971). Mutation and cancer: statistical study of retinoblastoma. *Proceedings of the National Academy of Sciences of the United States of America*, *68*(4), 820–823.

Knudson, A. G. (2001). Two genetic hits (more or less) to cancer. *Nature Reviews Cancer*, *1*(2), 157–162. https://doi.org/10.1038/35101031

Koboldt, D. C., Zhang, Q., Larson, D. E., Shen, D., McLellan, M. D., Lin, L., … Wilson, R. K. (2012). VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Research*, *22*(3), 568–576. https://doi.org/10.1101/gr.129684.111

Kops, G. J. P. L., Foltz, D. R., & Cleveland, D. W. (2004). Lethality to human cancer cells through massive chromosome loss by inhibition of the mitotic checkpoint. *Proceedings of the National Academy of Sciences of the United States of America*, *101*(23), 8699–8704. https://doi.org/10.1073/pnas.0401142101

Kovacs, M. E., Papp, J., Szentirmay, Z., Otto, S., & Olah, E. (2009). Deletions removing the last exon of *TACSTD1* constitute a distinct class of mutations predisposing to Lynch syndrome. *Human Mutation*, *30*(2), 197–203. https://doi.org/10.1002/humu.20942

Kozomara, A., Birgaoanu, M., & Griffiths-Jones, S. (2019). miRBase: from microRNA sequences to function. *Nucleic Acids Research*, *47*(D1), D155–D162. https://doi.org/10.1093/nar/gky1141

Krumm, N., Sudmant, P. H., Ko, A., O'Roak, B. J., Malig, M., Coe, B. P., … Eichler, E. E. (2012). Copy number variation detection and genotyping from exome sequence data. *Genome Research*, *22*(8), 1525–1532. https://doi.org/10.1101/gr.138115.112

Kuhn, D. E., Martin, M. M., Feldman, D. S., Terry, A. V., Nuovo, G. J., & Elton, T. S. (2008). Experimental validation of miRNA targets. *Methods*, *44*(1), 47–54. https://doi.org/10.1016/j.ymeth.2007.09.005

Kuiper, R. P., Vissers, L. E. L. M., Venkatachalam, R., Bodmer, D., Hoenselaar, E., Goossens, M., … Ligtenberg, M. J. L. (2011). Recurrence and variability of germline EPCAM deletions in Lynch syndrome. *Human Mutation*, *32*(4), 407–414. https://doi.org/10.1002/humu.21446

Kuipers, E. J., Grady, W. M., Lieberman, D., Seufferlein, T., Sung, J. J., Boelens, P. G., … Watanabe, T. (2015). Colorectal cancer. *Nature Reviews Disease Primers*, *1*, 15065. https://doi.org/10.1038/nrdp.2015.65

Kurashina, K., Yamashita, Y., Ueno, T., Koinuma, K., Ohashi, J., Horie, H., … Mano, H. (2008). Chromosome copy number analysis in screening for prognosis-related genomic regions in colorectal carcinoma. *Cancer Science*, *99*(9), 1835–1840. https://doi.org/10.1111/j.1349-7006.2008.00881.x

Lamlum, H., Al Tassan, N., Jaeger, E., Frayling, I., Sieber, O., Reza, F. B., … Tomlinson, I. (2000). Germline APC variants in patients with multiple colorectal adenomas, with evidence for the particular importance of E1317Q. *Human Molecular Genetics*, *9*(15), 2215–2221. https://doi.org/10.1093/oxfordjournals.hmg.a018912

Langner, C. (2015). Serrated and non-serrated precursor lesions of colorectal cancer.

Digestive Diseases (Basel, Switzerland), *33*(1), 28–37. https://doi.org/10.1159/000366032

Larsen, S. J., do Canto, L. M., Rogatto, S. R., & Baumbach, J. (2018). CoNVaQ: a web tool for copy number variation-based association studies. *BMC Genomics*, *19*(1), 369. https://doi.org/10.1186/s12864-018-4732-8

Le, D. T., Uram, J. N., Wang, H., Bartlett, B. R., Kemberling, H., Eyring, A. D., … Diaz, L. A. (2015). PD-1 Blockade in Tumors with Mismatch-Repair Deficiency. *New England Journal of Medicine*, *372*(26), 2509–2520. https://doi.org/10.1056/NEJMoa1500596

Lee, J. A., Carvalho, C. M. B., & Lupski, J. R. (2007). A DNA Replication Mechanism for Generating Nonrecurrent Rearrangements Associated with Genomic Disorders. *Cell*, *131*(7), 1235–1247. https://doi.org/10.1016/j.cell.2007.11.037

Lee, R. C., Feinbaum, R. L., & Ambros, V. (1993). The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell*, *75*(5), 843–854.

Lek, M., Karczewski, K. J., Minikel, E. V, Samocha, K. E., Banks, E., Fennell, T., … Exome Aggregation Consortium, D. G. (2016). Analysis of protein-coding genetic variation in 60,706 humans. *Nature*, *536*(7616), 285–291. https://doi.org/10.1038/nature19057

Leppert, M., Dobbs, M., Scambler, P., O'Connell, P., Nakamura, Y., Stauffer, D., … Gardner, E. (1987). The gene for familial polyposis coli maps to the long arm of chromosome 5. *Science (New York, N.Y.)*, *238*(4832), 1411–1413.

Lewis, B. P., Shih, I., Jones-Rhoades, M. W., Bartel, D. P., & Burge, C. B. (2003). Prediction of mammalian microRNA targets. *Cell*, *115*(7), 787–798. https://doi.org/10.1016/S0092-8674(03)01018-3

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, *25*(14), 1754–1760. https://doi.org/10.1093/bioinformatics/btp324

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., … 1000 Genome Project Data Processing Subgroup. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, *25*(16), 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Li, J., Lupat, R., Amarasinghe, K. C., Thompson, E. R., Doyle, M. A., Ryland, G. L., … Gorringe, K. L. (2012). CONTRA: Copy number analysis for targeted resequencing. *Bioinformatics*, *28*(10), 1307–1313. https://doi.org/10.1093/bioinformatics/bts146

Liaw, A., & Wiener, M. (2002). Classification and Regression by randomForest. *R News*, *2*(3), 18–22.

Lichtenstein, P., Holm, N. V., Verkasalo, P. K., Iliadou, A., Kaprio, J., Koskenvuo, M., … Hemminki, K. (2000). Environmental and Heritable Factors in the Causation of Cancer — Analyses of Cohorts of Twins from Sweden, Denmark, and Finland. *New England Journal of Medicine*, *343*(2), 78–85. https://doi.org/10.1056/NEJM200007133430201

Lieber, M. R. (2008). The Mechanism of Human Nonhomologous DNA End Joining. *Journal of Biological Chemistry*, *283*(1), 1–5. https://doi.org/10.1074/jbc.R700039200

Ligtenberg, M. J. L., Kuiper, R. P., Chan, T. L., Goossens, M., Hebeda, K. M., Voorendt, M., … Hoogerbrugge, N. (2009). Heritable somatic methylation and inactivation of MSH2 in families with Lynch syndrome due to deletion of the 3' exons of TACSTD1. *Nature Genetics*, *41*(1), 112–117. https://doi.org/10.1038/ng.283

Lin, K. J., Cheung, W. Y., Lai, J. Y.-C., & Giovannucci, E. L. (2012). The effect of estrogen vs. combined estrogen-progestogen therapy on the risk of colorectal cancer. *International Journal of Cancer*, *130*(2), 419–430. https://doi.org/10.1002/ijc.26026

Lindor, N. M., Rabe, K., Petersen, G. M., Haile, R., Casey, G., Baron, J., … Seminara, D. (2005). Lower cancer incidence in Amsterdam-I criteria families without mismatch repair deficiency: familial colorectal cancer type X. *JAMA*, *293*(16), 1979–1985. https://doi.org/10.1001/jama.293.16.1979

Liu, B., Morrison, C. D., Johnson, C. S., Trump, D. L., Qin, M., Conroy, J. C., … Liu, S. (2013). Computational methods for detecting copy number variations in cancer genome using next generation sequencing: principles and challenges. *Oncotarget*, *4*(11), 1868–1881. https://doi.org/10.18632/oncotarget.1537

Liu, M., Xie, Z., & Price, D. H. (1998). A human RNA polymerase II transcription termination factor is a SWI2/SNF2 family member. *The Journal of Biological Chemistry*, *273*(40), 25541–25544.

Liu, W., & Wang, X. (2019). Prediction of functional microRNA targets by integrative modeling of microRNA binding and target expression data. *Genome Biology*, *20*(1), 18. https://doi.org/10.1186/s13059-019-1629-z

Loeb, L. A., & Harris, C. C. (2008). Advances in Chemical Carcinogenesis: A Historical Review and Prospective. *Cancer Research*, *68*(17), 6863–6872. https://doi.org/10.1158/0008-5472.CAN-08-2852

Loo, L. W. M., Tiirikainen, M., Cheng, I., Lum-Jones, A., Seifried, A., Church, J. M., … Le Marchand, L. (2013). Integrated analysis of genome-wide copy number alterations and gene expression in microsatellite stable, CpG island methylator phenotype-negative colon cancer. *Genes, Chromosomes & Cancer*, *52*(5), 450–466. https://doi.org/10.1002/gcc.22043

Lord, J., McMullan, D. J., Eberhardt, R. Y., Rinck, G., Hamilton, S. J., Quinlan-Jones, E., … Wilson, E. (2019). Prenatal exome sequencing analysis in fetal structural anomalies detected by ultrasonography (PAGE): a cohort study. *Lancet (London, England)*, *393*(10173), 747–757. https://doi.org/10.1016/S0140-6736(18)31940-8

Lynch, H., Lynch, P., Lanspa, S., Snyder, C., Lynch, J., & Boland, C. (2009). Review of the Lynch syndrome: history, molecular genetics, screening, differential diagnosis, and medicolegal ramifications. *Clinical Genetics*, *76*(1), 1–18. https://doi.org/10.1111/j.1399-0004.2009.01230.x

Lynch, H T, & Krush, A. J. (1971). Cancer family "G" revisited: 1895-1970. *Cancer*, *27*(6), 1505–1511.

Lynch, H T, Smyrk, T., & Lynch, J. (1997). An update of HNPCC (Lynch syndrome). *Cancer Genetics and Cytogenetics*, *93*(1), 84–99.

Lynch, Henry T., & de la Chapelle, A. (2003). Hereditary Colorectal Cancer. *New England Journal of Medicine*, *348*(10), 919–932. https://doi.org/10.1056/NEJMra012242

MacDonald, J. R., Ziman, R., Yuen, R. K. C., Feuk, L., & Scherer, S. W. (2014). The Database of Genomic Variants: a curated collection of structural variation in the human genome. *Nucleic Acids Research*, *42*(D1), D986–D992. https://doi.org/10.1093/nar/gkt958

Macintyre, G., Goranova, T. E., De Silva, D., Ennis, D., Piskorz, A. M., Eldridge, M., … Brenton, J. D. (2018). Copy number signatures and mutational processes in ovarian carcinoma. *Nature Genetics*, *50*(9), 1262–1270. https://doi.org/10.1038/s41588-018-0179-8

Magi, A., Pippucci, T., & Sidore, C. (2017). XCAVATOR: accurate detection and genotyping of copy number variants from second and third generation whole-genome sequencing experiments. *BMC Genomics*, *18*(1), 747. https://doi.org/10.1186/s12864-017-4137-0

Majumdar, S. R., Fletcher, R. H., & Evans, A. T. (1999). How does colorectal cancer present? symptoms, duration, and clues to location. *The American Journal of Gastroenterology*, *94*(10), 3039–3045. https://doi.org/10.1111/j.1572-0241.1999.01454.x

Malkova, A., Ivanov, E. L., & Haber, J. E. (1996). Double-strand break repair in the absence of RAD51 in yeast: a possible role for break-induced DNA replication. *Proceedings of the National Academy of Sciences of the United States of America*, *93*(14), 7131–7136.

Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., … Visscher, P. M. (2009). Finding the missing heritability of complex diseases. *Nature*, *461*(7265), 747–753. https://doi.org/10.1038/nature08494

Marco-Sola, S., Sammeth, M., Guigó, R., & Ribeca, P. (2012). The GEM mapper: fast, accurate and versatile alignment by filtration. *Nature Methods*, *9*(12), 1185–1188. https://doi.org/10.1038/nmeth.2221

Markowitz, S. D., & Bertagnolli, M. M. (2009). Molecular Basis of Colorectal Cancer. *New England Journal of Medicine*, *361*(25), 2449–2460. https://doi.org/10.1056/NEJMra0804588

Markowitz, S., Wang, J., Myeroff, L., Parsons, R., Sun, L., Lutterbaugh, J., … Vogelstein, B. (1995). Inactivation of the type II TGF-beta receptor in colon cancer cells with microsatellite instability. *Science (New York, N.Y.)*, *268*(5215), 1336–1338.

Martín-Morales, L., Feldman, M., Vershinin, Z., Garre, P., Caldés, T., & Levy, D. (2017). SETD6 dominant negative mutation in familial colorectal cancer

type X. *Human Molecular Genetics*, *26*(22), 4481–4493. https://doi.org/10.1093/hmg/ddx336

Martincorena, I., & Campbell, P. J. (2015). Somatic mutation in cancer and normal cells. *Science*, *349*(6255), 1483–1489. https://doi.org/10.1126/science.aab4082

Massagué, J., Blain, S. W., & Lo, R. S. (2000). TGFbeta signaling in growth control, cancer, and heritable disorders. *Cell*, *103*(2), 295–309.

Matano, M., Date, S., Shimokawa, M., Takano, A., Fujii, M., Ohta, Y., … Sato, T. (2015). Modeling colorectal cancer using CRISPR-Cas9–mediated engineering of human intestinal organoids. *Nature Medicine*, *21*(3), 256–262. https://doi.org/10.1038/nm.3802

McCarthy, M. I., Abecasis, G. R., Cardon, L. R., Goldstein, D. B., Little, J., Ioannidis, J. P. A., & Hirschhorn, J. N. (2008). Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nature Reviews Genetics*, *9*(5), 356–369. https://doi.org/10.1038/nrg2344

McGranahan, N., & Swanton, C. (2017). Clonal Heterogeneity and Tumor Evolution: Past, Present, and the Future. *Cell*, *168*(4), 613–628. https://doi.org/10.1016/j.cell.2017.01.018

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., … DePristo, M. A. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, *20*(9), 1297–1303. https://doi.org/10.1101/gr.107524.110

Meijer, G. A., Hermsen, M. A., Baak, J. P., van Diest, P. J., Meuwissen, S. G., Belien, J. A., … Walboomers, J. M. (1998). Progression from colorectal adenoma to carcinoma is associated with non- random chromosomal gains as detected by comparative genomic hybridisation. *Journal of Clinical Pathology*, *51*(12), 901–909. https://doi.org/10.1136/jcp.51.12.901

Mekenkamp, L. J., Tol, J., Dijkstra, J. R., de Krijger, I., Vink-Börger, M. E., van Vliet, S., … Nagtegaal, I. D. (2012). Beyond KRAS mutation status: influence of KRAScopy number status and microRNAs on clinical outcome to cetuximab in metastatic colorectal cancer patients. *BMC Cancer*, *12*(1), 292. https://doi.org/10.1186/1471-2407-12-292

Mermel, C. H., Schumacher, S. E., Hill, B., Meyerson, M. L., Beroukhim, R., & Getz, G. (2011). GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biology*, *12*(4), R41. https://doi.org/10.1186/gb-2011-12-4-r41

Minami, Y., Nishino, Y., Tsubono, Y., Tsuji, I., & Hisamichi, S. (2006). Increase of colon and rectal cancer incidence rates in Japan: trends in incidence rates in Miyagi Prefecture, 1959-1997. *Journal of Epidemiology*, *16*(6), 240–248.

Minoo, P., Moyer, M., & Jass, J. (2007). Role of BRAF-V600E in the serrated pathway of colorectal tumourigenesis. *The Journal of Pathology*, *212*(2), 124–133. https://doi.org/10.1002/path.2160

Miyaki, M., Konishi, M., Tanaka, K., Kikuchi-Yanoshita, R., Muraoka, M., Yasuno, M., … Mori, T. (1997). Germline mutation of MSH6 as the cause of hereditary nonpolyposis colorectal cancer. *Nature Genetics*, *17*(3), 271–272. https://doi.org/10.1038/ng1197-271

Morganella, S., Alexandrov, L. B., Glodzik, D., Zou, X., Davies, H., Staaf, J., … Nik-Zainal, S. (2016). The topography of mutational processes in breast cancer genomes. *Nature Communications*, *7*, 11383. https://doi.org/10.1038/ncomms11383

Nakano, H., Miyazawa, T., Kinoshita, K., Yamada, Y., & Yoshida, T. (2009). Functional screening identifies a microRNA, miR-491 that induces apoptosis by targeting Bcl-XL in colorectal cancer cells. *International Journal of Cancer*, *127*(5), 1072–1080. https://doi.org/10.1002/ijc.25143

Nakao, K., Mehta, K. R., Fridlyand, J., Moore, D. H., Jain, A. N., Lafuente, A., … Waldman, F. M. (2004). High-resolution analysis of DNA copy number alterations in colorectal cancer by array-based comparative genomic hybridization. *Carcinogenesis*, *25*(8), 1345–1357. https://doi.org/10.1093/carcin/bgh134

Nam, J.-Y., Kim, N. K. D., Kim, S. C., Joung, J.-G., Xi, R., Lee, S., … Park, W.-Y. (2016). Evaluation of somatic copy number estimation tools for whole-exome sequencing data. *Briefings in Bioinformatics*, *17*(2), 185–192. https://doi.org/10.1093/bib/bbv055

Nayak, L., Iwamoto, F. M., LaCasce, A., Mukundan, S., Roemer, M. G. M., Chapuy, B., … Shipp, M. A. (2017). PD-1 blockade with nivolumab in relapsed/refractory primary central nervous system and testicular lymphoma.

Blood, *129*(23), 3071–3073. https://doi.org/10.1182/blood-2017-01-764209

Nesic, K., Wakefield, M., Kondrashova, O., Scott, C. L., & McNeish, I. A. (2018). Targeting DNA repair: the genome as a potential biomarker. *The Journal of Pathology*, *244*(5), 586–597. https://doi.org/10.1002/path.5025

Network, T. C. G. A. (TCGA) R. (2008). Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature*, *455*(7216), 1061. https://doi.org/10.1038/NATURE07385

Nieminen, T. T., O'Donohue, M.-F., Wu, Y., Lohi, H., Scherer, S. W., Paterson, A. D., … Peltomäki, P. (2014). Germline mutation of RPS20, encoding a ribosomal protein, causes predisposition to hereditary nonpolyposis colorectal carcinoma without DNA mismatch repair deficiency. *Gastroenterology*, *147*(3), 595-598.e5. https://doi.org/10.1053/j.gastro.2014.06.009

Nigg, E. A. (2006). Origins and consequences of centrosome aberrations in human cancers. *International Journal of Cancer*, *119*(12), 2717–2723. https://doi.org/10.1002/ijc.22245

Nik-Zainal, S., Davies, H., Staaf, J., Ramakrishna, M., Glodzik, D., Zou, X., … Stratton, M. R. (2016). Landscape of somatic mutations in 560 breast cancer whole-genome sequences. *Nature*, *534*(7605), 47–54. https://doi.org/10.1038/nature17676

Ning, Y., Wang, L., & Giovannucci, E. L. (2010). A quantitative analysis of body mass index and colorectal cancer: findings from 56 observational studies. *Obesity Reviews*, *11*(1), 19–30. https://doi.org/10.1111/j.1467-789X.2009.00613.x

Nishisho, I., Nakamura, Y., Miyoshi, Y., Miki, Y., Ando, H., Horii, A., … Hedge, P. (1991). Mutations of chromosome 5q21 genes in FAP and colorectal cancer patients. *Science (New York, N.Y.)*, *253*(5020), 665–669.

Noone, A., Howlader, N., Krapcho, M., Miller, D., Brest, A., Yu, M., … Cronin, K. (2018). SEER Cancer Statistics Review, 1975-2015, National Cancer Institute. Bethesda, MD. Retrieved October 5, 2018, from https://seer.cancer.gov/csr/1975_2015/

Northcott, P. A., Nakahara, Y., Wu, X., Feuk, L., Ellison, D. W., Croul, S., … Taylor, M. D. (2009). Multiple recurrent genetic events converge on control of histone lysine methylation in medulloblastoma. *Nature Genetics*, *41*(4), 465–472. https://doi.org/10.1038/ng.336

Nowell, P. C. (1976). The clonal evolution of tumor cell populations. *Science (New York, N.Y.)*, *194*(4260), 23–28.

Olaharski, A. J., Sotelo, R., Solorza-Luna, G., Gonsebatt, M. E., Guzman, P., Mohar, A., & Eastmond, D. A. (2006). Tetraploidy and chromosomal instability are early events during cervical carcinogenesis. *Carcinogenesis*, *27*(2), 337–343. https://doi.org/10.1093/carcin/bgi218

Olschwang, S., Serova-Sinilnikova, O. M., Lenoir, G. M., & Thomas, G. (1998). PTEN germ-line mutations in juvenile polyposis coli. *Nature Genetics*, *18*(1), 12–14. https://doi.org/10.1038/ng0198-12

Olshen, A. B., Venkatraman, E. S., Lucito, R., & Wigler, M. (2004). Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics*, *5*(4), 557–572. https://doi.org/10.1093/biostatistics/kxh008

Ostertag, E. M., & Kazazian Jr, H. H. (2001). Biology of Mammalian L1 Retrotransposons. *Annual Review of Genetics*, *35*(1), 501–538. https://doi.org/10.1146/annurev.genet.35.102401.091032

Palles, C., Cazier, J.-B., Howarth, K. M., Domingo, E., Jones, A. M., Broderick, P., … Tomlinson, I. (2013). Germline mutations in the proof-reading domains of POLE and POLD1 predispose to colorectal adenomas and carcinomas. *Nature Genetics*, *45*(2), 136. https://doi.org/10.1038/NG.2503

Pan-Cancer Atlas. (n.d.). Pan-Cancer Atlas. Retrieved February 27, 2019, from https://www.cell.com/pb-assets/consortium/pancanceratlas/pancani3/index.html?utm_campaign=STMJ_1522957422_SC&utm_campaign=STMJ_26797_SC_197&utm_channel=email&utm_source=Other&utm_source=AC_7&dgcid=STMJ_1522957422_SC&dgcid=STMJ_26797_SC_197&utm_medium=emai

Peiffer, D. A., Le, J. M., Steemers, F. J., Chang, W., Jenniges, T., Garcia, F., … Gunderson, K. L. (2006). High-resolution genomic profiling of chromosomal aberrations using Infinium whole-genome genotyping. *Genome Research*, *16*(9), 1136. https://doi.org/10.1101/GR.5402306

Peng, H.-X., Wu, W.-Q., Yang, D.-M., Jing, R.,

Li, J., Zhou, F.-L., … Chu, Y.-M. (2015). Role of B7-H4 siRNA in Proliferation, Migration, and Invasion of LOVO Colorectal Carcinoma Cell Line. *BioMed Research International*, *2015*, 1–10. https://doi.org/10.1155/2015/326981

Peters, U., Bien, S., & Zubair, N. (2015). Genetic architecture of colorectal cancer. *Gut*, *64*(10), 1623–1636. https://doi.org/10.1136/gutjnl-2013-306705

Petrovski, S., Aggarwal, V., Giordano, J. L., Stosic, M., Wou, K., Bier, L., … Wapner, R. J. (2019). Whole-exome sequencing in the evaluation of fetal structural anomalies: a prospective cohort study. *Lancet (London, England)*, *393*(10173), 758–767. https://doi.org/10.1016/S0140-6736(18)32042-7

Pfundt, R., del Rosario, M., Vissers, L. E. L. M., Kwint, M. P., Janssen, I. M., de Leeuw, N., … Hehir-Kwa, J. Y. (2017). Detection of clinically relevant copy-number variants by exome sequencing in a large cohort of genetic disorders. *Genetics in Medicine*, *19*(6), 667–675. https://doi.org/10.1038/gim.2016.163

Pihan, G. A., Purohit, A., Wallace, J., Knecht, H., Woda, B., Quesenberry, P., & Doxsey, S. J. (1998). Centrosome defects and genetic instability in malignant tumors. *Cancer Research*, *58*(17), 3974–3985.

Pilarski, R. (2009). Cowden Syndrome: A Critical Review of the Clinical Literature. *Journal of Genetic Counseling*, *18*(1), 13–27. https://doi.org/10.1007/s10897-008-9187-7

Pinkel, D., Landegent, J., Collins, C., Fuscoe, J., Segraves, R., Lucas, J., & Gray, J. (1988). Fluorescence in situ hybridization with human chromosome-specific libraries: detection of trisomy 21 and translocations of chromosome 4. *Proceedings of the National Academy of Sciences of the United States of America*, *85*(23), 9138–9142.

Pinkel, D., Segraves, R., Sudar, D., Clark, S., Poole, I., Kowbel, D., … Albertson, D. G. (1998). High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. *Nature Genetics*, *20*(2), 207–211. https://doi.org/10.1038/2524

Pino, M. S., & Chung, D. C. (2010). The Chromosomal Instability Pathway in Colon Cancer. *Gastroenterology*, *138*(6), 2059–2072. https://doi.org/10.1053/j.gastro.2009.12.065

Plagnol, V., Curtis, J., Epstein, M., Mok, K. Y., Stebbings, E., Grigoriadou, S., … Nejentsev, S. (2012). A robust model for read count data in exome sequencing experiments and implications for copy number variant calling. *Bioinformatics*, *28*(21), 2747–2754. https://doi.org/10.1093/bioinformatics/bts526

Polakis, P. (1997). The adenomatous polyposis coli (APC) tumor suppressor. *Biochimica et Biophysica Acta*, *1332*(3), F127-47.

Postma, C., Terwischa, S., Hermsen, M. A. J. A., van der Sijp, J. R. M., & Meijer, G. A. (2007). Gain of chromosome 20q is an indicator of poor prognosis in colorectal cancer. *Cellular Oncology : The Official Journal of the International Society for Cellular Oncology*, *29*(1), 73–75.

Poynter, J. N., Gruber, S. B., Higgins, P. D. R., Almog, R., Bonner, J. D., Rennert, H. S., … Rennert, G. (2005). Statins and the risk of colorectal cancer. *The New England Journal of Medicine*, *352*(21), 2184–2192. https://doi.org/10.1056/NEJMoa043792

Pritchard, J. K., Stephens, M., & Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics*, *155*(2), 945–959.

Ptashkin, R. N., Pagan, C., Yaeger, R., Middha, S., Shia, J., O'Rourke, K. P., … Hechtman, J. F. (2017). Chromosome 20q Amplification Defines a Subtype of Microsatellite Stable, Left-Sided Colon Cancers with Wild-type RAS/RAF and Better Overall Survival. *Molecular Cancer Research*, *15*(6), 708–713. https://doi.org/10.1158/1541-7786.MCR-16-0352

Qiao, W., Han, Y., Jin, W., Tian, M., Chen, P., Min, J., … Lin, Q. (2016). Overexpression and biological function of TMEM48 in non-small cell lung carcinoma. *Tumor Biology*, *37*(2), 2575–2586. https://doi.org/10.1007/s13277-015-4014-x

R Core Team. (2018). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical computing.

Rao, S. S. P., Huntley, M. H., Durand, N. C., Stamenova, E. K., Bochkov, I. D., Robinson, J. T., … Aiden, E. L. (2014). A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell*, *159*(7), 1665–1680. https://doi.org/10.1016/j.cell.2014.11.021

Rausch, T., Zichner, T., Schlattl, A., Stutz, A. M., Benes, V., & Korbel, J. O. (2012).

DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics*, *28*(18), i333–i339. https://doi.org/10.1093/bioinformatics/bts378

Rayner, E., van Gool, I. C., Palles, C., Kearsey, S. E., Bosse, T., Tomlinson, I., & Church, D. N. (2016). A panoply of errors: polymerase proofreading domain mutations in cancer. *Nature Reviews Cancer*, *16*(2), 71–81. https://doi.org/10.1038/nrc.2015.12

Redon, R., Ishikawa, S., Fitch, K. R., Feuk, L., Perry, G. H., Andrews, T. D., … Hurles, M. E. (2006). Global variation in copy number in the human genome. *Nature*, *444*(7118), 444–454. https://doi.org/10.1038/nature05329

Ried, T., Hu, Y., Difilippantonio, M. J., Ghadimi, B. M., Grade, M., & Camps, J. (2012). The consequences of chromosomal aneuploidy on the transcriptome of cancer cells. *Biochimica et Biophysica Acta*, *1819*(7), 784–793. https://doi.org/10.1016/j.bbagrm.2012.02.020

Ried, T., Knutzen, R., Steinbeck, R., Blegen, H., Schröck, E., Heselmeyer, K., … Auer, G. (1996). Comparative genomic hybridization reveals a specific pattern of chromosomal gains and losses during the genesis of colorectal tumors. *Genes, Chromosomes and Cancer*, *15*(4), 234–245. https://doi.org/10.1002/(SICI)1098-2264(199604)15:4<234::AID-GCC5>3.0.CO;2-2

Rieder, C. L., Cole, R. W., Khodjakov, A., & Sluder, G. (1995). The checkpoint delaying anaphase in response to chromosome monoorientation is mediated by an inhibitory signal produced by unattached kinetochores. *The Journal of Cell Biology*, *130*(4), 941–948.

Risch, N. J. (2000). Searching for genetic determinants in the new millennium. *Nature*, *405*(6788), 847–856. https://doi.org/10.1038/35015718

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., & Smyth, G. K. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, *43*(7), e47. https://doi.org/10.1093/nar/gkv007

Roberts, J. T., & Borchert, G. M. (2017). *Computational Prediction of MicroRNA Target Genes, Target Prediction Databases, and Web Resources*. https://doi.org/10.1007/978-1-4939-7046-9_8

Rohlin, A., Eiengård, F., Lundstam, U., Zagoras, T., Nilsson, S., Edsjö, A., … Nordling, M. (2016). GREM1 and POLE variants in hereditary colorectal cancer syndromes. *Genes, Chromosomes & Cancer*, *55*(1), 95. https://doi.org/10.1002/GCC.22314

Rosty, C., Parry, S., & Young, J. P. (2011). Serrated polyposis: an enigmatic model of colorectal cancer predisposition. *Pathology Research International*, *2011*, 157073. https://doi.org/10.4061/2011/157073

Rothwell, P. M., Fowkes, F. G. R., Belch, J. F., Ogawa, H., Warlow, C. P., & Meade, T. W. (2011). Effect of daily aspirin on long-term risk of death due to cancer: analysis of individual patient data from randomised trials. *The Lancet*, *377*(9759), 31–41. https://doi.org/10.1016/S0140-6736(10)62110-1

Rubinfeld, B., Souza, B., Albert, I., Müller, O., Chamberlain, S. H., Masiarz, F. R., … Polakis, P. (1993). Association of the APC gene product with beta-catenin. *Science (New York, N.Y.)*, *262*(5140), 1731–1734.

Rustgi, A. K. (2007). The genetics of hereditary colon cancer. *Genes & Development*, *21*(20), 2525–2538. https://doi.org/10.1101/gad.1593107

Saito, T., Niida, A., Uchi, R., Hirata, H., Komatsu, H., Sakimura, S., … Mimori, K. (2018). A temporal shift of the evolutionary principle shaping intratumor heterogeneity in colorectal cancer. *Nature Communications*, *9*(1), 2884. https://doi.org/10.1038/s41467-018-05226-0

Samarakoon, P. S., Sorte, H. S., Kristiansen, B. E., Skodje, T., Sheng, Y., Tjønnfjord, G. E., … Lyle, R. (2014). Identification of copy number variants from exome sequence data. *BMC Genomics*, *15*(1), 661. https://doi.org/10.1186/1471-2164-15-661

Sampson, J. R., Jones, S., Dolwani, S., & Cheadle, J. P. (2005). *MutYH* (*MYH*) and colorectal cancer. *Biochemical Society Transactions*, *33*(4), 679–683. https://doi.org/10.1042/BST0330679

Sanchez-Vega, F., Mina, M., Armenia, J., Chatila, W. K., Luna, A., La, K. C., … Schultz, N. (2018). Oncogenic Signaling Pathways in The Cancer Genome Atlas. *Cell*, *173*(2), 321-337.e10. https://doi.org/10.1016/j.cell.2018.03.035

Sandhu, M. S., White, I. R., & McPherson, K.

(2001). Systematic review of the prospective cohort studies on meat consumption and colorectal cancer risk: a meta-analytical approach. *Cancer Epidemiology, Biomarkers & Prevention : A Publication of the American Association for Cancer Research, Cosponsored by the American Society of Preventive Oncology*, *10*(5), 439–446.

Sansregret, L., Vanhaesebroeck, B., & Swanton, C. (2018). Determinants and clinical implications of chromosomal instability in cancer. *Nature Reviews Clinical Oncology*, *15*(3), 139–150. https://doi.org/10.1038/nrclinonc.2017.198

Sathirapongsasuti, J. F., Lee, H., Horst, B. A. J., Brunner, G., Cochran, A. J., Binder, S., … Nelson, S. F. (2011). Exome sequencing-based copy-number variation and loss of heterozygosity detection: ExomeCNV. *Bioinformatics*, *27*(19), 2648–2654. https://doi.org/10.1093/bioinformatics/btr462

Schmit, K., & Michiels, C. (2018). TMEM Proteins in Cancer: A Review. *Frontiers in Pharmacology*, *9*, 1345. https://doi.org/10.3389/fphar.2018.01345

Schouten, J. P. (2002). *Multiplex ligatable probe amplification.*

Schouten, J. P., McElgunn, C. J., Waaijer, R., Zwijnenburg, D., Diepvens, F., & Pals, G. (2002). Relative quantification of 40 nucleic acid sequences by multiplex ligation-dependent probe amplification. *Nucleic Acids Research*, *30*(12), e57.

Schulz, E., Klampfl, P., Holzapfel, S., Janecke, A. R., Ulz, P., Renner, W., … Sill, H. (2014). Germline variants in the SEMA4A gene predispose to familial colorectal cancer type X. *Nature Communications*, *5*(1), 5191. https://doi.org/10.1038/ncomms6191

Seguí, N., Mina, L. B., Lázaro, C., Sanz-Pamplona, R., Pons, T., Navarro, M., … Valle, L. (2015). Germline Mutations in FAN1 Cause Hereditary Colorectal Cancer by Impairing DNA Repair. *Gastroenterology*, *149*(3), 563–566. https://doi.org/10.1053/j.gastro.2015.05.056

Shackney, S. E., Smith, C. A., Miller, B. W., Burholt, D. R., Murtha, K., Giles, H. R., … Pollice, A. A. (1989). Model for the genetic evolution of human solid tumors. *Cancer Research*, *49*(12), 3344–3354.

Shendure, J., & Ji, H. (2008). Next-generation DNA sequencing. *Nature Biotechnology*, *26*(10), 1135–1145. https://doi.org/10.1038/nbt1486

Shepherd, N. A., Bussey, H. J., & Jass, J. R. (1987). Epithelial misplacement in Peutz-Jeghers polyps. A diagnostic pitfall. *The American Journal of Surgical Pathology*, *11*(10), 743–749.

Singhal, S. S., Singhal, J., Yadav, S., Dwivedi, S., Boor, P. J., Awasthi, Y. C., & Awasthi, S. (2007). Regression of Lung and Colon Cancer Xenografts by Depleting or Inhibiting RLIP76 (Ral-Binding Protein 1). *Cancer Research*, *67*(9), 4382–4389. https://doi.org/10.1158/0008-5472.CAN-06-4124

Slatkin, M. (2008). Linkage disequilibrium — understanding the evolutionary past and mapping the medical future. *Nature Reviews Genetics*, *9*(6), 477–485. https://doi.org/10.1038/nrg2361

Smedley, D., Schubach, M., Jacobsen, J. O. B., Köhler, S., Zemojtel, T., Spielmann, M., … Robinson, P. N. (2016). A Whole-Genome Analysis Framework for Effective Identification of Pathogenic Regulatory Variants in Mendelian Disease. *American Journal of Human Genetics*, *99*(3), 595–606. https://doi.org/10.1016/j.ajhg.2016.07.005

Smith, C. E., Llorente, B., & Symington, L. S. (2007). Template switching during break-induced replication. *Nature*, *447*(7140), 102–105. https://doi.org/10.1038/nature05723

Smith, C. G., Naven, M., Harris, R., Colley, J., West, H., Li, N., … Cheadle, J. P. (2013). Exome Resequencing Identifies Potential Tumor-Suppressor Genes that Predispose to Colorectal Cancer. *Human Mutation*, *34*(7), 1026–1034. https://doi.org/10.1002/humu.22333

Snover, D. C., Jass, J. R., Fenoglio-Preiser, C., & Batts, K. P. (2005). Serrated Polyps of the Large Intestine: A Morphologic and Molecular Review of an Evolving Concept. *American Journal of Clinical Pathology*, *124*(3), 380–391. https://doi.org/10.1309/V2EP-TPLJ-RB3F-GHJL

Speicher, M. R., & Carter, N. P. (2005). The new cytogenetics: blurring the boundaries with molecular biology. *Nature Reviews Genetics*, *6*(10), 782–792. https://doi.org/10.1038/nrg1692

Spielmann, M., Lupiáñez, D. G., & Mundlos, S. (2018). Structural variation in the 3D genome. *Nature Reviews Genetics*, *19*(7), 453–467. https://doi.org/10.1038/s41576-018-0007-0

Spier, I., Kerick, M., Drichel, D., Horpaopan, S., Altmüller, J., Laner, A., … Aretz, S. (2016). Exome sequencing identifies potential novel candidate genes in patients with unexplained colorectal adenomatous polyposis. *Familial Cancer*, *15*(2), 281–288. https://doi.org/10.1007/s10689-016-9870-z

Stankiewicz, P., & Lupski, J. R. (2002). Genome architecture, rearrangements and genomic disorders. *Trends in Genetics : TIG*, *18*(2), 74–82. https://doi.org/10.1016/S0168-9525(02)02592-1

Stoffel, E., Mukherjee, B., Raymond, V. M., Tayob, N., Kastrinos, F., Sparr, J., … Gruber, S. B. (2009). Calculation of Risk of Colorectal and Endometrial Cancer Among Patients With Lynch Syndrome. *Gastroenterology*, *137*(5), 1621–1627. https://doi.org/10.1053/j.gastro.2009.07.039

Strand, M., Prolla, T. A., Liskay, R. M., & Petes, T. D. (1993). Destabilization of tracts of simple repetitive DNA in yeast by mutations affecting DNA mismatch repair. *Nature*, *365*(6443), 274–276. https://doi.org/10.1038/365274a0

Stratton, M. R., Campbell, P. J., & Futreal, P. A. (2009). The cancer genome. *Nature*, *458*(7239), 719–724. https://doi.org/10.1038/nature07943

Su, L. K., Vogelstein, B., & Kinzler, K. W. (1993). Association of the APC tumor suppressor protein with catenins. *Science (New York, N.Y.)*, *262*(5140), 1734–1737.

Sugimura, T., Wakabayashi, K., Nakagama, H., & Nagao, M. (2004). Heterocyclic amines: Mutagens/carcinogens produced during cooking of meat and fish. *Cancer Science*, *95*(4), 290–299.

Talevich, E., Shain, A. H., Botton, T., & Bastian, B. C. (2016). CNVkit: Genome-Wide Copy Number Detection and Visualization from Targeted DNA Sequencing. *PLOS Computational Biology*, *12*(4), e1004873. https://doi.org/10.1371/journal.pcbi.1004873

Tan, R., Wang, Y., Kleinstein, S. E., Liu, Y., Zhu, X., Guo, H., … Zhu, M. (2014). An Evaluation of Copy Number Variation Detection Tools from Whole-Exome Sequencing Data. *Human Mutation*, *35*(7), 899–907. https://doi.org/10.1002/humu.22537

Tanaka, K., & Hirota, T. (2016). Chromosomal instability: A common feature and a therapeutic target of cancer. *Biochimica et Biophysica Acta (BBA) - Reviews on Cancer*, *1866*(1), 64–75. https://doi.org/10.1016/j.bbcan.2016.06.002

Tang, Y.-C., & Amon, A. (2013). Gene Copy-Number Alterations: A Cost-Benefit Analysis. *Cell*, *152*(3), 394–405. https://doi.org/10.1016/j.cell.2012.11.043

Tanskanen, T., Gylfe, A. E., Katainen, R., Taipale, M., Renkonen-Sinisalo, L., Järvinen, H., … Aaltonen, L. A. (2015). Systematic search for rare variants in Finnish early-onset colorectal cancer patients. *Cancer Genetics*, *208*(1–2), 35–40. https://doi.org/10.1016/j.cancergen.2014.12.004

Taupin, D., Lam, W., Rangiah, D., McCallum, L., Whittle, B., Zhang, Y., … Cook, M. C. (2015). A deleterious RNF43 germline mutation in a severely affected serrated polyposis kindred. *Human Genome Variation*, *2*(1), 15013. https://doi.org/10.1038/hgv.2015.13

Taylor, A. M., Shih, J., Ha, G., Gao, G. F., Zhang, X., Berger, A. C., … Meyerson, M. (2018). Genomic and Functional Approaches to Understanding Cancer Aneuploidy. *Cancer Cell*, *33*(4), 676-689.e3. https://doi.org/10.1016/J.CCELL.2018.03.007

Tenesa, A., & Dunlop, M. G. (2009). New insights into the aetiology of colorectal cancer from genome-wide association studies. *Nature Reviews. Genetics*, *10*(6), 353–358. https://doi.org/10.1038/nrg2574

The Cancer Genome Atlas. (n.d.). The Cancer Genome Atlas. Retrieved February 27, 2019, from https://cancergenome.nih.gov/

The Cancer Genome Atlas. (2012). Comprehensive molecular characterization of human colon and rectal cancer. *Nature*, *487*(7407), 330–337. https://doi.org/10.1038/nature11252

The SAM/BAM Format Specification Working Group. (2018). *Sequence Alignment/Map Format Specification*.

Thutkawkorapin, J., Picelli, S., Kontham, V., Liu, T., Nilsson, D., & Lindblom, A. (2016). Exome sequencing in one family with gastric- and rectal cancer. *BMC Genetics*, *17*(1), 41. https://doi.org/10.1186/s12863-016-0351-z

Tjio, J. H., & Levan, A. (2010). The chromosome number of man. *Hereditas*, *42*(1–2), 1–6. https://doi.org/10.1111/j.1601-

5223.1956.tb03010.x

Tomasetti, C., Li, L., & Vogelstein, B. (2017). Stem cell divisions, somatic mutations, cancer etiology, and cancer prevention. *Science*, *355*(6331), 1330–1334. https://doi.org/10.1126/science.aaf9011

Tomlinson, I. P. M., Carvajal-Carmona, L. G., Dobbins, S. E., Tenesa, A., Jones, A. M., Howarth, K., … Dunlop, M. G. (2011). Multiple common susceptibility variants near BMP pathway loci GREM1, BMP4, and BMP2 explain part of the missing heritability of colorectal cancer. *PLoS Genetics*, *7*(6), e1002105. https://doi.org/10.1371/journal.pgen.1002105

Toyota, M., Ahuja, N., Ohe-Toyota, M., Herman, J. G., Baylin, S. B., & Issa, J. P. (1999). CpG island methylator phenotype in colorectal cancer. *Proceedings of the National Academy of Sciences of the United States of America*, *96*(15), 8681–8686.

Trautmann, K., Terdiman, J. P., French, A. J., Roydasgupta, R., Sein, N., Kakar, S., … Waldman, F. M. (2006). Chromosomal Instability in Microsatellite-Unstable and Stable Colon Cancer. *Clinical Cancer Research*, *12*(21), 6379–6385. https://doi.org/10.1158/1078-0432.CCR-06-1248

Tremblay, L. O., & Herscovics, A. (2000). Characterization of a cDNA Encoding a Novel Human Golgi α1,2-Mannosidase (IC) Involved in *N* -Glycan Biosynthesis. *Journal of Biological Chemistry*, *275*(41), 31655–31660. https://doi.org/10.1074/jbc.M004935200

Trost, B., Walker, S., Wang, Z., Thiruvahindrapuram, B., MacDonald, J. R., Sung, W. W. L., … Scherer, S. W. (2018). A Comprehensive Workflow for Read Depth-Based Identification of Copy-Number Variation from Whole-Genome Sequence Data. *American Journal of Human Genetics*, *102*(1), 142–155. https://doi.org/10.1016/j.ajhg.2017.12.007

Umar, A., Boland, C. R., Terdiman, J. P., Syngal, S., de la Chapelle, A., Rüschoff, J., … Srivastava, S. (2004). Revised Bethesda Guidelines for hereditary nonpolyposis colorectal cancer (Lynch syndrome) and microsatellite instability. *Journal of the National Cancer Institute*, *96*(4), 261–268.

Umar, A., Boyer, J. C., Thomas, D. C., Nguyen, D. C., Risinger, J. I., Boyd, J., … Kunkel, T. A. (1994). Defective mismatch repair in extracts of colorectal and endometrial cancer cell lines exhibiting microsatellite instability. *The Journal of Biological Chemistry*, *269*(20), 14367–14370.

Valle, L. (2017). Recent Discoveries in the Genetics of Familial Colorectal Cancer and Polyposis. *Clinical Gastroenterology and Hepatology*, *15*(6), 809–819. https://doi.org/10.1016/j.cgh.2016.09.148

Valle, L., de Voer, R. M., Goldberg, Y., Sjursen, W., Försti, A., Ruiz-Ponte, C., … Hemminki, K. (2019). Update on genetic predisposition to colorectal cancer and polyposis. *Molecular Aspects of Medicine*. https://doi.org/10.1016/j.mam.2019.03.001

Valle, L., Hernández-Illán, E., Bellido, F., Aiza, G., Castillejo, A., Castillejo, M.-I., … Blanco, I. (2014). New insights into POLE and POLD1 germline mutations in familial colorectal cancer and polyposis. *Human Molecular Genetics*, *23*(13), 3506–3512. https://doi.org/10.1093/hmg/ddu058

Valle, L., Vilar, E., Tavtigian, S. V., & Stoffel, E. M. (2018). Genetic predisposition to colorectal cancer: syndromes, genes, classification of genetic variants and implications for precision medicine. *The Journal of Pathology*. https://doi.org/10.1002/path.5229

van Duijnhoven, F. J., Bueno-De-Mesquita, H. B., Ferrari, P., Jenab, M., Boshuizen, H. C., Ros, M. M., … Riboli, E. (2009). Fruit, vegetables, and colorectal cancer risk: the European Prospective Investigation into Cancer and Nutrition. *The American Journal of Clinical Nutrition*, *89*(5), 1441–1452. https://doi.org/10.3945/ajcn.2008.27120

Vandrovcova, J., Thomas, E. R. a, Atanur, S. S., Norsworthy, P. J., Neuwirth, C., Tan, Y., … Aitman, T. J. (2013). The use of next-generation sequencing in clinical diagnosis of familial hypercholesterolemia. *Genetics in Medicine : Official Journal of the American College of Medical Genetics*, *15*(May), 1–10. https://doi.org/10.1038/gim.2013.55

Vasen, H., Watson, P., J, M., & Lynch, H. (1999). New clinical criteria for hereditary nonpolyposis colorectal cancer (HNPCC, Lynch syndrome) proposed by the International Collaborative Group on HNPCC☆. *Gastroenterology*, *116*(6), 1453–1456. https://doi.org/10.1016/S0016-5085(99)70510-X

Venkatachalam, R., Ligtenberg, M. J. L., Hoogerbrugge, N., Schackert, H. K., Görgens, H., Hahn, M.-M., … Kuiper, R. P. (2010). Germline Epigenetic Silencing of the Tumor Suppressor Gene PTPRJ in

Early-Onset Familial Colorectal Cancer. *Gastroenterology*, *139*(6), 2221–2224. https://doi.org/10.1053/j.gastro.2010.08.063

Venkatachalam, R., Verwiel, E. T. P., Kamping, E. J., Hoenselaar, E., Görgens, H., Schackert, H. K., … Kuiper, R. P. (2011). Identification of candidate predisposing copy number variants in familial and early-onset colorectal cancer patients. *International Journal of Cancer*, *129*(7), 1635–1642. https://doi.org/10.1002/ijc.25821

Venkatraman, E. S., & Olshen, A. B. (2007). A faster circular binary segmentation algorithm for the analysis of array CGH data. *Bioinformatics*, *23*(6), 657–663. https://doi.org/10.1093/bioinformatics/btl646

Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., … Zhu, X. (2001). The sequence of the human genome. *Science (New York, N.Y.)*, *291*(5507), 1304–1351. https://doi.org/10.1126/science.1058040

Villacis, R. A. R., Miranda, P. M., Gomy, I., Santos, E. M. M., Carraro, D. M., Achatz, M. I., … Rogatto, S. R. (2016). Contribution of rare germline copy number variations and common susceptibility loci in Lynch syndrome patients negative for mutations in the mismatch repair genes. *International Journal of Cancer*, *138*(8), 1928–1935. https://doi.org/10.1002/ijc.29948

Vinothkumar, K. R., & Henderson, R. (2010). Structures of membrane proteins. *Quarterly Reviews of Biophysics*, *43*(1), 65–158. https://doi.org/10.1017/S0033583510000041

Vogelstein, B, Fearon, E., Kern, S., Hamilton, S. R., Preisinger, A., Nakamura, Y., & White, R. (1989). Allelotype of colorectal carcinomas. *Science*, *244*(4901), 207–211. https://doi.org/10.1126/science.2565047

Vogelstein, Bert, Papadopoulos, N., Velculescu, V. E., Zhou, S., Diaz, L. A., & Kinzler, K. W. (2013). Cancer Genome Landscapes. *Science*, *339*(6127), 1546–1558. https://doi.org/10.1126/science.1235122

Walther, A., Johnstone, E., Swanton, C., Midgley, R., Tomlinson, I., & Kerr, D. (2009). Genetic prognostic and predictive markers in colorectal cancer. *Nature Reviews Cancer*, *9*(7), 489–499. https://doi.org/10.1038/nrc2645

Wang, H., Liang, L., Fang, J.-Y., & Xu, J. (2016). Somatic gene copy number alterations in colorectal cancer: new quest for cancer drivers and biomarkers. *Oncogene*, *35*(16), 2011–2019. https://doi.org/10.1038/onc.2015.304

Wang, K., Lim, H. Y., Shi, S., Lee, J., Deng, S., Xie, T., … Xu, J. (2013). Genomic landscape of copy number aberrations enables the identification of oncogenic drivers in hepatocellular carcinoma. *Hepatology*, *58*(2), 706–717. https://doi.org/10.1002/hep.26402

Watson, J. D., & Crick, F. H. (1953). The structure of DNA. *Cold Spring Harbor Symposia on Quantitative Biology*, *18*, 123–131.

Weaver, B. A., & Cleveland, D. W. (2006). Does aneuploidy cause cancer? *Current Opinion in Cell Biology*, *18*(6), 658–667. https://doi.org/10.1016/j.ceb.2006.10.002

Weir, B. A., Woo, M. S., Getz, G., Perner, S., Ding, L., Beroukhim, R., … Meyerson, M. (2007). Characterizing the cancer genome in lung adenocarcinoma. *Nature*, *450*(7171), 893–898. https://doi.org/10.1038/nature06358

Weischenfeldt, J., Dubash, T., Drainas, A. P., Mardin, B. R., Chen, Y., Stütz, A. M., … Korbel, J. O. (2017). Pan-cancer analysis of somatic copy-number alterations implicates IRS4 and IGF2 in enhancer hijacking. *Nature Genetics*, *49*(1), 65–74. https://doi.org/10.1038/ng.3722

Weischenfeldt, J., Symmons, O., Spitz, F., & Korbel, J. O. (2013). Phenotypic impact of genomic structural variation: insights from and for human disease. *Nature Reviews Genetics*, *14*(2), 125–138. https://doi.org/10.1038/nrg3373

Weren, R. D. A., Ligtenberg, M. J. L., Kets, C. M., de Voer, R. M., Verwiel, E. T. P., Spruijt, L., … Hoogerbrugge, N. (2015). A germline homozygous mutation in the base-excision repair gene NTHL1 causes adenomatous polyposis and colorectal cancer. *Nature Genetics*, *47*(6), 668–671. https://doi.org/10.1038/ng.3287

Weren, R. D. A., Venkatachalam, R., Cazier, J.-B., Farin, H. F., Kets, C. M., de Voer, R. M., … Kuiper, R. P. (2015). Germline deletions in the tumour suppressor gene FOCAD are associated with polyposis and colorectal cancer development. *The Journal of Pathology*, *236*(2), 155–164. https://doi.org/10.1002/path.4520

What Is Cancer? - National Cancer Institute. (n.d.). Retrieved October 4, 2018, from https://www.cancer.gov/about-cancer/understanding/what-is-cancer

Wheeler, D. A., & Wang, L. (2013). From human genome to cancer genome: The first decade. *Genome Research*, *23*(7), 1054–1062. https://doi.org/10.1101/gr.157602.113

Whitelaw, S. C., Murday, V. A., Tomlinson, I. P., Thomas, H. J., Cottrell, S., Ginsberg, A., … Solomon, E. (1997). Clinical and molecular features of the hereditary mixed polyposis syndrome. *Gastroenterology*, *112*(2), 327–334.

Win, A. K., Jenkins, M. A., Dowty, J. G., Antoniou, A. C., Lee, A., Giles, G. G., … MacInnis, R. J. (2017). Prevalence and Penetrance of Major Genes and Polygenes for Colorectal Cancer. *Cancer Epidemiology, Biomarkers & Prevention : A Publication of the American Association for Cancer Research, Cosponsored by the American Society of Preventive Oncology*, *26*(3), 404–412. https://doi.org/10.1158/1055-9965.EPI-16-0693

Wolin, K. Y., Yan, Y., Colditz, G. A., & Lee, I.-M. (2009). Physical activity and colon cancer prevention: a meta-analysis. *British Journal of Cancer*, *100*(4), 611–616. https://doi.org/10.1038/sj.bjc.6604917

Wolman, S. R., Gundacker, H., Appelbaum, F. R., Slovak, M. L., & Southwest Oncology Group. (2002). Impact of trisomy 8 (+8) on clinical presentation, treatment response, and survival in acute myeloid leukemia: a Southwest Oncology Group study. *Blood*, *100*(1), 29–35.

Wood, L. D., Parsons, D. W., Jones, S., Lin, J., Sjoblom, T., Leary, R. J., … Vogelstein, B. (2007). The Genomic Landscapes of Human Breast and Colorectal Cancers. *Science*, *318*(5853), 1108–1113. https://doi.org/10.1126/science.1145720

Woods, M. O., Younghusband, H. B., Parfrey, P. S., Gallinger, S., McLaughlin, J., Dicks, E., … Green, R. C. (2010). The genetic basis of colorectal cancer in a population-based incident cohort with a high rate of familial disease. *Gut*, *59*(10), 1369–1377. https://doi.org/10.1136/gut.2010.208462

Wrzesiński, T., Szelag, M., Cieślikowski, W. A., Ida, A., Giles, R., Zodro, E., … Wesoly, J. (2015). Expression of pre-selected TMEMs with predicted ER localization as potential classifiers of ccRCC tumors. *BMC Cancer*, *15*(1), 518. https://doi.org/10.1186/s12885-015-1530-4

Xie, C., & Tammi, M. T. (2009). CNV-seq, a new method to detect copy number variation using high-throughput sequencing. *BMC Bioinformatics*, *10*(1), 80. https://doi.org/10.1186/1471-2105-10-80

Yan, H. H. N., Lai, J. C. W., Ho, S. L., Leung, W. K., Law, W. L., Lee, J. F. Y., … Leung, S. Y. (2017). RNF43 germline and somatic mutation in serrated neoplasia pathway and its association with BRAF mutation. *Gut*, *66*(9), 1645–1656. https://doi.org/10.1136/gutjnl-2016-311849

Yang, X., Zhang, H., & Li, L. (2012). Alternative mRNA processing increases the complexity of microRNA-based gene regulation in Arabidopsis. *The Plant Journal*, *70*(3), 421–431. https://doi.org/10.1111/j.1365-313X.2011.04882.x

Yao, R., Zhang, C., Yu, T., Li, N., Hu, X., Wang, X., … Shen, Y. (2017). Evaluation of three read-depth based CNV detection tools using whole-exome sequencing data. *Molecular Cytogenetics*, *10*(1), 30. https://doi.org/10.1186/s13039-017-0333-5

Ye, K., Schulz, M. H., Long, Q., Apweiler, R., & Ning, Z. (2009). Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics*, *25*(21), 2865–2871. https://doi.org/10.1093/bioinformatics/btp394

Yoon, S., Xuan, Z., Makarov, V., Ye, K., & Sebat, J. (2009). Sensitive and accurate detection of copy number variants using read depth of coverage. *Genome Research*, *19*(9), 1586–1592. https://doi.org/10.1101/gr.092981.109

Young, J., & Jass, J. R. (2006). The case for a genetic predisposition to serrated neoplasia in the colorectum: hypothesis and review of the literature. *Cancer Epidemiology, Biomarkers & Prevention : A Publication of the American Association for Cancer Research, Cosponsored by the American Society of Preventive Oncology*, *15*(10), 1778–1784. https://doi.org/10.1158/1055-9965.EPI-06-0164

Yu, L., Yin, B., Qu, K., Li, J., Jin, Q., Liu, L., … Cao, K. (2018). Screening for susceptibility genes in hereditary non-polyposis colorectal cancer. *Oncology Letters*, *15*(6), 9413–9419. https://doi.org/10.3892/ol.2018.8504

Yurgelun, M. B., Allen, B., Kaldate, R. R., Bowles, K. R., Judkins, T., Kaushik, P., … Syngal, S. (2015). Identification of a Variety of Mutations in Cancer Predisposition Genes in Patients With Suspected Lynch Syndrome.

Gastroenterology, *149*(3), 604-613.e20. https://doi.org/10.1053/j.gastro.2015.05.006

Yurgelun, M. B., Kulke, M. H., Fuchs, C. S., Allen, B. A., Uno, H., Hornick, J. L., … Syngal, S. (2017). Cancer Susceptibility Gene Mutations in Individuals With Colorectal Cancer. *Journal of Clinical Oncology*, *35*(10), 1086–1095. https://doi.org/10.1200/JCO.2016.71.0012

Zack, T. I., Schumacher, S. E., Carter, S. L., Cherniack, A. D., Saksena, G., Tabak, B., … Beroukhim, R. (2013). Pan-cancer patterns of somatic copy number alteration. *Nature Genetics*, *45*(10), 1134–1140. https://doi.org/10.1038/ng.2760

Zare, F., Dow, M., Monteleone, N., Hosny, A., & Nabavi, S. (2017). An evaluation of copy number variation detection tools for cancer using whole exome sequencing data. *BMC Bioinformatics*, *18*(1), 286. https://doi.org/10.1186/s12859-017-1705-x

Zarrei, M., MacDonald, J. R., Merico, D., & Scherer, S. W. (2015). A copy number variation map of the human genome. *Nature Reviews Genetics*, *16*(3), 172–183. https://doi.org/10.1038/nrg3871

Zehir, A., Benayed, R., Shah, R. H., Syed, A., Middha, S., Kim, H. R., … Berger, M. F. (2017). Mutational landscape of metastatic cancer revealed from prospective clinical sequencing of 10,000 patients. *Nature Medicine*, *23*(6), 703–713. https://doi.org/10.1038/nm.4333

Zhang, F., Gu, W., Hurles, M. E., & Lupski, J. R. (2009). Copy Number Variation in Human Health, Disease, and Evolution. *Annual Review of Genomics and Human Genetics*, *10*(1), 451–481. https://doi.org/10.1146/annurev.genom.9.081307.164217

Zhang, F., Khajavi, M., Connolly, A. M., Towne, C. F., Batish, S. D., & Lupski, J. R. (2009). The DNA replication FoSTeS/MMBIR mechanism can generate genomic, genic and exonic complex rearrangements in humans. *Nature Genetics*, *41*(7), 849–853. https://doi.org/10.1038/ng.399

Zhao, M., Wang, Q. Q., Wang, Q. Q., Jia, P., & Zhao, Z. (2013). Computational tools for copy number variation (CNV) detection using next-generation sequencing data: features and perspectives. *BMC Bioinformatics*, *14 Suppl 1*(11), S1. https://doi.org/10.1186/1471-2105-14-S11-S1

Zhou, X., Popescu, N. C., Klein, G., & Imreh, S. (2007). The interferon-α responsive gene TMEM7 suppresses cell proliferation and is downregulated in human hepatocellular carcinoma. *Cancer Genetics and Cytogenetics*, *177*(1), 6–15. https://doi.org/10.1016/j.cancergencyto.2007.04.007

Índex de figures

Índex de taules

Annexes

Rare germline copy number variants in colorectal cancer predisposition characterized by exome sequening analysis

Journal of Genetics and Genomics

20 de gener del 2018
(doi: 10.1016/j.jgg.2017.12.001)

Journal of Genetics and Genomics 45 (2018) 41–45

Contents lists available at ScienceDirect

Journal of Genetics and Genomics

Journal homepage: www.journals.elsevier.com/journal-of-genetics-and-genomics/

Letter to the editor

Rare germline copy number variants in colorectal cancer predisposition characterized by exome sequencing analysis

Colorectal cancer (CRC) is one of the most common neoplasms and an important cause of mortality worldwide (http://globocan.iarc.fr/). Approximately 35% of the variation in CRC susceptibility is likely due to heritable factors (Lichtenstein et al., 2000). Genetic variations in the human genome include single nucleotide variants (SNVs), short insertions and deletions, and larger structural variants resulting in gain or loss of genomic DNA larger than 1 kb, such as copy number variants (CNVs). Leaving aside the importance of CNVs in sporadic tumor development, these variants can also be present in the germline DNA of healthy individuals from the general population and be considered polymorphic. Common germline CNVs can confer a small increase in the risk of predisposition to disease, whereas rare CNVs have been linked to hereditary cancer predisposition including CRC. Recent examples include alterations involving *EPCAM*, *PTPRJ*, *CDH18*, *GREM1* and *FOCAD* (Ligtenberg et al., 2009; Venkatachalam et al., 2011; Jaeger et al., 2012; Weren et al., 2015).

A considerable number of cases with strong familial CRC aggregation and early disease onset remain unresolved at the genetic level. As a result of the research to discover new genes involved in hereditary CRC, next-generation sequencing (NGS) efforts have identified additional causative germline mutations in *POLE*, *POLD1*, *NTHL1* and *MSH3* (Peters et al., 2015). Previous studies analyzed SNVs and short insertions and deletions, but did not include CNV characterization. Traditional molecular methods for genotyping CNVs include hybridization-based techniques (e.g., fluorescent *in situ* hybridization, FISH), PCR-based techniques (e.g., multiplex ligation-dependent probe amplification, MLPA) and array-based approaches (e.g., array comparative genome hybridization, aCGH). However, the current availability of large NGS datasets has prompted its use in the determination of structural variation and recently several bioinformatic approaches have been developed to predict CNVs from whole-exome sequencing (WES) data (Tan et al., 2014).

Accordingly, the aim of the present study was to screen our current germline WES dataset of 71 patients from 38 families, in order to identify rare CNVs that could correspond to the mutational event for CRC predisposition. Families presented strong CRC aggregation compatible with an autosomal dominant pattern of inheritance and no alteration in the known hereditary CRC genes was detected (Esteban-Jurado et al., 2015, 2016). Clearly, our final objective is to facilitate genetic counseling in order to correctly address preventive strategies in these families.

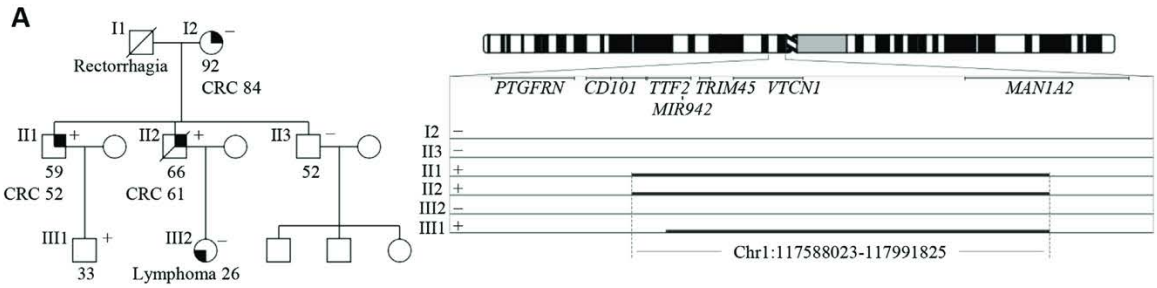
Many studies have evaluated multiple tools for CNV calling in order to compare their performance and reproducibility (Tan et al., 2014; Wang et al., 2015). In this study, CoNIFER (http://conifer.sourceforge.net/) and ExomeDepth (https://cran.r-project.org/web/packages/ExomeDepth) were used to infer CNVs from WES data. Further details are available in Supplementary data (see Supplementary Material and Methods). Variants shared among sequenced individuals from the same family were prioritized, and only variants detected by both bioinformatic tools were considered as candidate CNVs. Through this process, we detected 21 candidate CNVs corresponding to 16 duplications and 5 deletions (Table S1).

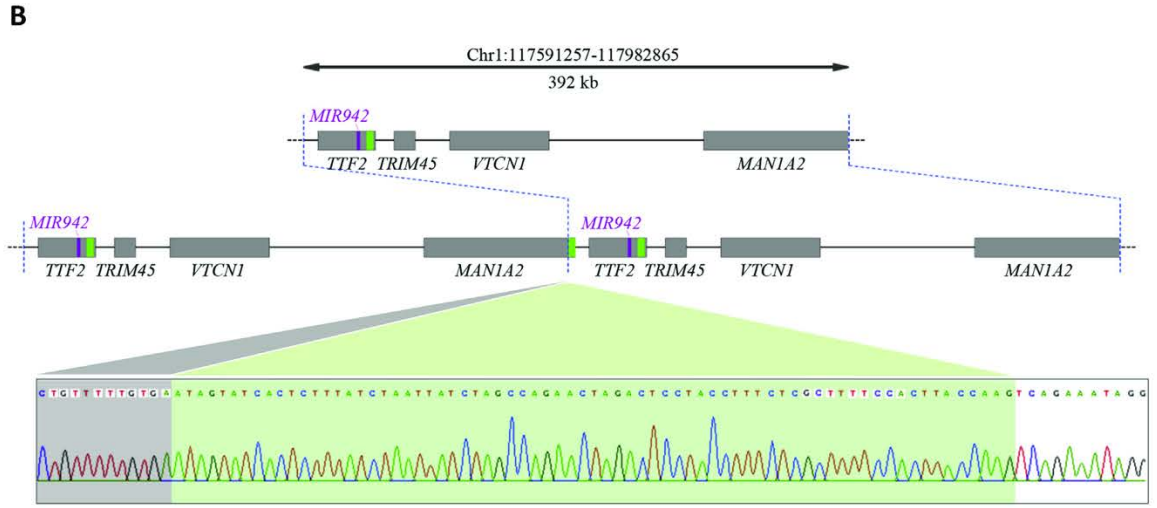
Frequency of inferred CNVs in the general population was annotated by using the DGV catalogue (The Database of Genomics Variants, http://dgv.tcag.ca). We also used an in-house CNV database (Fernandez-Rozadilla et al., 2013) consisting of data from 500 Spanish controls. When checking these control CNV datasets, a 20% increase/decrease in length was considered for each inferred start/end to avoid false negatives. Only rare CNVs (frequency $\leq$ 10 in either of the control databases) were further prioritized. Following this step, 14 rare variants were considered as plausible mutational events involved in germline CRC predisposition. Most CNVs involved genes whose function has not been associated with CRC (Table S1).

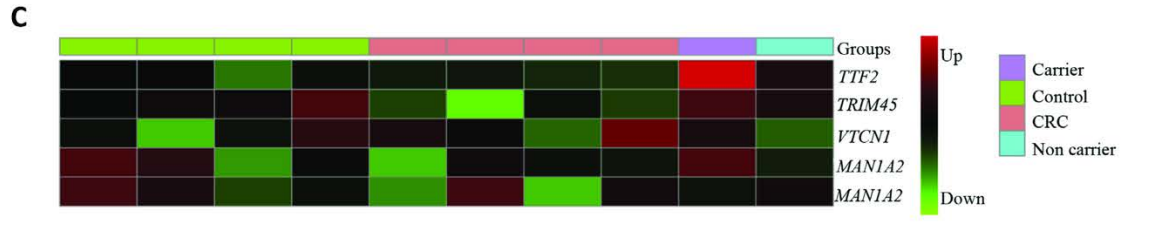
A duplication event on chromosome 1 identified in family 7 was the most relevant finding. It spans a region of 400 kb encompassing *TTF2* (transcription termination factor 2), *MIR942* (microRNA 942), *TRIM45* (tripartite motif containing 45), *VTCN1* (V-set domain containing T-cell activation inhibitor 1) and *MAN1A2* (mannosidase α class 1A member 2) partially. *TTF2* is a member of the SWI2/SNF2 family and is responsible for mitotic repression of transcription elongation (Jiang and Price, 2004). *MIR942* has been associated with cancer stem cell-like promotion (Ge et al., 2015), and *TRIM45* belongs to a protein family that functions as an important regulator for carcinogenesis (Hatakeyama, 2011). Additionally, *VTCN1* is expressed in various types of human cancer tissues including CRC and may be clinically relevant (Peng et al., 2015). This duplication was present only once in DGV and absent in our in-house Spanish database. Interestingly, in our previous studies, we did not detect any interesting SNVs in this family (Esteban-Jurado et al., 2015, 2016), as shown in Table S2.

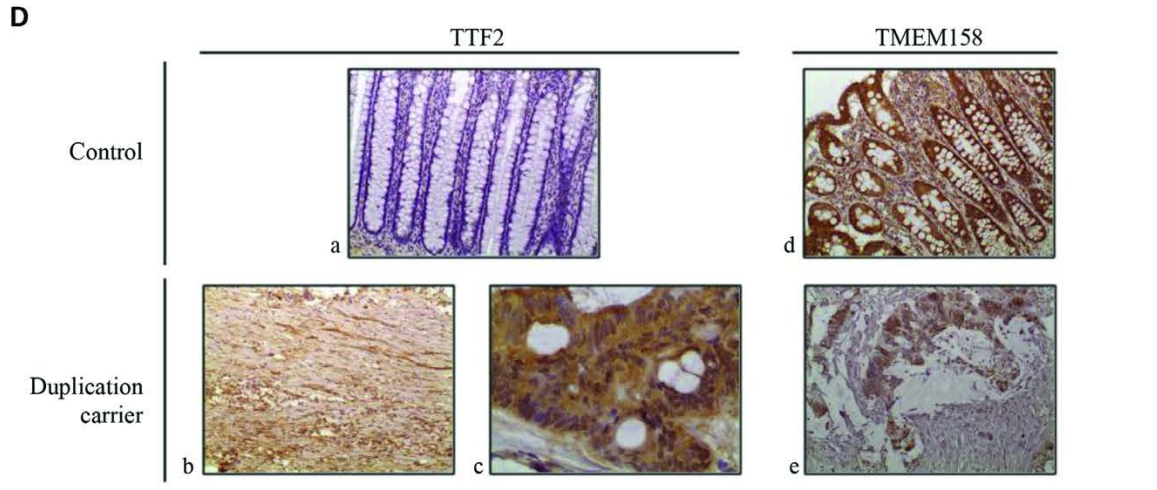
While it has been shown that NGS can be used for CNV detection, it is generally recommended that aCGH is used to validate detected CNVs (Guo et al., 2013; Wang et al., 2015). Therefore,

https://doi.org/10.1016/j.jgg.2017.12.001
1673-8527/Copyright © 2017, Institute of Genetics and Developmental Biology, Chinese Academy of Sciences, and Genetics Society of China. Published by Elsevier Limited and Science Press. All rights reserved.


A
I1
I2
Rectorrhagia
92
CRC 84
II1
II2
II3
59
CRC 52
66
CRC 61
52
III1
III2
33
Lymphoma 26
PTGFRN
CD101
TTF2
TRIM45
VTCN1
MIR942
MAN1A2
I2
II3
II1
II2
III2
III1
Chr1:117588023-117991825


B
Chr1:117591257-117982865
392 kb
MIR942
TTF2
TRIM45
VTCN1
MAN1A2


C
Groups
TTF2
TRIM45
VTCN1
MAN1A2
MAN1A2
Up
Down
Carrier
Control
CRC
Non carrier


D
TTF2
TMEM158
Control
Duplication carrier
a
b
c
d
e

we proceeded to validate the identified duplication on chromosome 1 using a 180K array aCGH on both individuals with WES data available and other available relatives. The duplication was confirmed to involve *TTF2*, *TRIM45*, *VTCN1* and the 5′-end of *MAN1A2*, and to be located at chr1:117588023-117991825 (404 kb) (Fig. 1A). The duplication was initially discovered in CRC cases II1 and II2 and was also found to be present in III1 (son of II1, unaffected). Relatives II3 (brother), I2 (mother, CRC at 84) and III2 (daughter of II2, lymphoma at 26) did not carry the duplication. The phenotype of I2 most likely corresponds to a sporadic CRC case in this family due to the late onset, whereas III2 is most likely not related to the CRC germline predisposition in this family. It is important to highlight that I1 (father) died of rectorrhagia at 78, a condition that could be related to CRC. Therefore, our hypothesis is that CRC predisposition may come from the father's family branch (Fig. 1A).

Whole-genome sequencing (WGS) data from a carrier (case II2) was generated and no other remarkable SNV or CNV variants besides the mentioned duplication could be found. Delly2 software (https://github.com/dellytools/delly) was applied on WGS data to extract structural variants, and the duplication on chromosome 1 was found again. The new breakpoint coordinates were more accurate and hence were used to design PCR primers aiming to characterize the breakpoint of the duplication at base-pair level. Final coordinates for this duplication were chr1:117591257-117982865 covering 392 kb (Fig. 1B). A 72-bp insertion coming from the last *TTF2* intron (chr1:117642853-117642924) was noticed inside the breakpoint by Sanger sequencing and short homologous sequences were found between the ends of the inserted intron fragment and the 5′- and 3′-ends of the duplicated sequence (Fig. 1B). This finding is in accordance with the observation that microhomology-mediated break-induced replication is the major contributor to nonrecurrent structural variants in genomic disorders (Carvalho and Lupski, 2016), though additional experiments would be necessary to further confirm this.

In order to identify the molecular effects of the detected duplication, gene expression was initially monitored using whole-genome arrays in blood RNA from a duplication carrier (II2) and a non-carrier (I2). Overexpression of the *TTF2* gene was evident in the carrier, whereas the other genes included in the duplication (*TRIM45*, *VTCN1* and *MAN1A2*) were not significantly altered when compared to non-carrier samples (I2, sporadic CRC and unaffected individuals) (Fig. 1C). *TTF2* overexpression in blood was further confirmed by real-time quantitative PCR. Additionally, gene expression of *MIR942* in tumor tissue also showed similar results (Fig. S1A).

TTF2 is an RNA polymerase II termination factor that is responsible for mitotic repression of transcription elongation. Underexpression of this factor is responsible for RNA polymerase II retention on mitotic chromosomes (Jiang and Price, 2004), while *TTF2* overexpression has not been studied, and it would be difficult to hypothesize its role due to its tight regulation during the cell cycle. On the other hand, overexpression of *MIR942* has been correlated with cancer stem cell-like traits in esophageal squamous cell carcinoma (ESCC) progression through targeting of negative regulators of the Wnt/β-catenin signaling pathway (Ge et al., 2015).

The TARGETSCAN (http://www.targetscan.org) database was used to find predicted target genes of *MIR942* among the differentially downregulated genes in the duplication carrier when compared to the non-carrier, as identified in whole-genome expression array data. Downregulated predicted targets included eight genes (Table S3). *TMEM158*, also known as *RIS1*, stood out as the most interesting candidate, having previously been claimed to be a tumor suppressor gene and target in the mutator pathway of colorectal carcinogenesis (Iglesias et al., 2006). CRC tumors with altered *TMEM158* were associated with poor prognosis, and its likely implication in *Ras*-induced senescence and its location at 3p21.3, a short DNA region called *CER1* (common eliminated region 1), support a potential role of *TMEM158* as a tumor suppressor gene (Hesson et al., 2007). The other seven predicted targets did not show a clear gene function related to cancer predisposition.

In order to confirm the upregulation of TTF2 and downregulation of TMEM158, immunohistochemistry experiments were performed on control colonic tissue and tumor tissue from a duplication carrier. The protein expression level of TTF2 was high in colon smooth muscle wall and tumor from a duplication carrier (Fig. 1D, b and c) when compared to the normal colonic mucosa of a control individual (Fig. 1D, a). Regarding TMEM158 expression, a reduction was evident when the tumor sample of the duplication carrier (Fig. 1D, e) was compared to the normal mucosa in a control sample (Fig. 1D, d).

Looking further into the topological architecture of the duplicated region, we observed that this region was divided into two topologically associated domains (TADs). One TAD contained the *TTF2* promoter, close to the 5′-end of the duplication, while the other TAD included *TRIM45*, *VTCN1*, and part of *MAN1A2* (Fig. S1B). It has been demonstrated that duplications affecting different TADs can lead to pathogenicity by reorganizing chromatin architecture and forming a neo-TAD (Franke et al., 2016). Due to the duplication, the *TTF2* promoter may be associated with additional gene regulatory elements that are involved in their upregulation. *MIR942*, located within intron 19 of *TTF2*, would be collaterally overexpressed. On the other hand, *TRIM45*, *VTCN1* and *MAN1A2* are probably not affected by this new TAD organization and therefore not overexpressed. It should be noted that *VTCN1* was previously reported to be overexpressed in CRC tissues (Peng et al., 2015).

Clearly, further efforts must be made in order to improve CNV prediction using WES data. However, it remains an interesting approach that increases the diagnostic capacity of exome-based sequencing tests, due to its potential to detect both SNVs and CNVs with the same data. Here, we prove that WES data can be used to search for CNVs with potential involvement in germline predisposition to CRC. The validated duplication in chromosome 1 may correspond to the mutational event involved in predisposition to CRC in the carrier family through overexpression of *TTF2* and *MIR942* and ultimate downregulation of *TMEM158*.

Fig. 1. Characterization of the duplication in chromosome 1 by aCGH, Sanger sequencing, transcriptomics and immunohistochemistry. **A**: Family 7 pedigree (left panel) and aCGH validation results in carriers (+) and non-carriers (−) (right panel). In the pedigree, top-right filled symbols denote individuals affected with CRC and a bottom-left filled symbol indicates an individual affected with lymphoma. Age at the time of study is shown for all relevant individuals (I2, II1, II2, II3, III1 and III2), whereas disease onset is shown in affected relatives next to the disease name (eg., II1 was affected with CRC at 52 and was 59 at the time of study). II1, II2 and III1 were confirmed as duplication carriers as shown. **B**: Schematic illustration of the chromosome 1 duplication. *TTF2*, *TRIM45*, *VTCN1* and *MAN1A2* are represented by grey boxes, *MIR942* by a purple box and a small 72-bp insertion by the green box. Sanger sequencing results for the duplication breakpoint are shown. Homologous base pairs are highlighted in white boxes. **C**: Gene expression array results from blood samples. In the heatmap, rows represent *TTF2*, *TRIM45*, *VTCN1* and *MAN1A2* probes and columns correspond to duplication carrier and non-carriers (I2, sporadic CRC and unaffected individuals). **D**: Immunohistochemistry results for TTF2 and TMEM158 in control colonic tissue and tumor tissue from a duplication carrier. TTF2 staining is shown in a normal colonic mucosa of a control individual (a), and in colon smooth muscle wall (b) and tumor tissue (c) of a duplication carrier. TMEM158 protein levels are shown in normal colonic mucosa of a control individual (d) and in tumor tissue of a duplication carrier (e).

Acknowledgments

This work was supported by CIBEREHD (to SFE, CEJ and JM), CIBERER, Fondo de Investigación Sanitaria/FEDER (14/00173, 14/00230 and 17/00878), Ministerio de Economía y Competitividad (SAF2014-54453-R), Fundación Científica de la Asociación Española contra el Cáncer (GCB13131592CAST), PERIS (SLT002/16/00398, Generalitat de Catalunya), COST Action BM1206, Beca Grupo de Trabajo "Oncología" AEG (Asociación Española de Gastroenterología), CERCA Programme (Generalitat de Catalunya) and Agència de Gestió d'Ajuts Universitaris i de Recerca (Generalitat de Catalunya, FI 2017 B00619 to MDG, 2014SGR255, 2014SGR135). We are sincerely grateful to the patients, Itziar Salaverria, CNAG, Biobank of Hospital Clínic–IDIBAPS, Biobanco Vasco and the Esther Koplowitz Centre.

Supplementary data

Supplementary data related to this article can be found at https://doi.org/10.1016/j.jgg.2017.12.001.

References

Carvalho, C.M.B., Lupski, J.R., 2016. Mechanisms underlying structural variant formation in genomic disorders. Nat. Rev. Genet. 17, 224–238.

Esteban-Jurado, C., Vila-Casadesús, M., Garre, P., Lozano, J.J., Pristoupilova, A., Beltran, S., Muñoz, J., Ocaña, T., Balaguer, F., López-Cerón, M., Cuatrecasas, M., Franch-Expósito, S., Piqué, J.M., Castells, A., Carracedo, A., Ruiz-Ponte, C., Abulí, A., Bessa, X., Andreu, M., Bujanda, L., Caldés, T., Castellví-Bel, S., 2015. Whole-exome sequencing identifies rare pathogenic variants in new predisposition genes for familial colorectal cancer. Genet. Med. 17, 131–142.

Esteban-Jurado, C., Franch-Expósito, S., Muñoz, J., Ocaña, T., Carballal, S., López-Cerón, M., Cuatrecasas, M., Vila-Casadesús, M., Lozano, J.J., Serra, E., Beltran, S., Brea-Fernández, A., Ruiz-Ponte, C., Castells, A., Bujanda, L., Garre, P., Caldés, T., Cubiella, J., Balaguer, F., Castellví-Bel, S., 2016. The Fanconi anemia DNA damage repair pathway in the spotlight for germline predisposition to colorectal cancer. Eur. J. Hum. Genet. 24, 1–5.

Fernandez-Rozadilla, C., Cazier, J.-B., Tomlinson, I.P., Carvajal-carmona, L.G., Palles, C., Lamas, M.J., López-fernández, L., a, Brea-fernández, A., Abulí, A., Baiget, M., López-fernández, L. a, Brea-fernández, A., Abulí, A., Bujanda, L., Clofent, J., Gonzalez, D., Xicola, R., Andreu, M., Bessa, X., Jover, R., Llor, X., Moreno, V., Castells, A., Carracedo, Á., Castellví-Bel, S., Ruiz-Ponte, C., 2013. A colorectal cancer genome-wide association study in a Spanish cohort identifies two variants associated with colorectal cancer risk at 1p33 and 8p12. BMC Genomics 14, 55.

Franke, M., Ibrahim, D.M., Andrey, G., Schwarzer, W., Heinrich, V., Schöpflin, R., Kraft, K., Kempfer, R., Jerković, I., Chan, W.-L., Spielmann, M., Timmermann, B., Wittler, L., Kurth, I., Cambiaso, P., Zuffardi, O., Houge, G., Lambie, L., Brancati, F., Pombo, A., Vingron, M., Spitz, F., Mundlos, S., 2016. Formation of new chromatin domains determines pathogenicity of genomic duplications. Nature 538, 265–269.

Ge, C., Wu, S., Wang, W., Liu, Z., Zhang, J., Wang, Z., Li, R., Zhang, Z., Li, Z., Dong, S., Wang, Y., Xue, Y., Yang, J., Tan, Q., Wang, Z., Song, X., 2015. miR-942 promotes cancer stem cell-like traits in esophageal squamous cell carcinoma through activation of Wnt/β-catenin signalling pathway. Oncotarget 6, 10964–10977.

Guo, Y., Sheng, Q., Samuels, D.C., Lehmann, B., Bauer, J.A., Pietenpol, J., Shyr, Y., 2013. Comparative study of exome copy number variation estimation tools using array comparative genomic hybridization as control. Biomed Res. Int. 1–7.

Hatakeyama, S., 2011. TRIM proteins and cancer. Nat. Rev. Canc. 11, 792–804.

Hesson, L.B., Cooper, W.N., Latif, F., 2007. Evaluation of the 3p21.3 tumour-suppressor gene cluster. Oncogene 26, 7283–7301.

Iglesias, D., Fernández-Peralta, A.M., Nejda, N., Daimiel, L., Azcoita, M.M., Oliart, S., González-Aguilera, J.J., 2006. *RIS1*, a gene with trinucleotide repeats, is a target in the mutator pathway of colorectal carcinogenesis. Cancer Genet. Cytogenet. 167, 138–144.

Jaeger, E., Leedham, S., Lewis, A., Segditsas, S., Becker, M., Cuadrado, P.R., Davis, H., Kaur, K., Heinimann, K., Howarth, K., East, J., Taylor, J., Thomas, H., Tomlinson, I., 2012. Hereditary mixed polyposis syndrome is caused by a 40-kb upstream duplication that leads to increased and ectopic expression of the BMP antagonist GREM1. Nat. Genet. 44, 699–703.

Jiang, Y., Price, D.H., 2004. Rescue of the *TTF2* knockdown phenotype with an siRNA-resistant replacement vector. Cell Cycle 3, 1151–1153.

Lichtenstein, P., Holm, N.V., Verkasalo, P.K., Iliadou, A., Kaprio, J., Koskenvuo, M., Pukkala, E., Skytthe, A., Hemminki, K., 2000. Environmental and heritable factors in the causation of cancer– analyses of cohorts of twins from Sweden, Denmark, and Finland. N. Engl. J. Med. 343, 78–85.

Ligtenberg, M.J.L., Kuiper, R.P., Chan, T.L., Goossens, M., Hebeda, K.M., Voorendt, M., Lee, T.Y.H., Bodmer, D., Hoenselaar, E., Hendriks-Cornelissen, S.J.B., Tsui, W.Y., Kong, C.K., Brunner, H.G., van Kessel, A.G., Yuen, S.T., van Krieken, J.H.J.M., Leung, S.Y., Hoogerbrugge, N., 2009. Heritable somatic methylation and inactivation of *MSH2* in families with Lynch syndrome due to deletion of the 3' exons of *TACSTD1*. Nat. Genet. 41, 112–117.

Peng, H.X., Wu, W.Q., Yang, D.M., Jing, R., Li, J., Zhou, F.L., Jin, Y.F., Wang, S.Y., Chu, Y.M., 2015. Role of B7-H4 siRNA in proliferation, migration, and invasion of LOVO colorectal carcinoma cell line. Biomed. Res. Int 2015, 326981.

Peters, U., Bien, S., Zubair, N., 2015. Genetic architecture of colorectal cancer. Gut 64, 1623–1636.

Tan, R., Wang, Y., Kleinstein, S.E., Liu, Y., Zhu, X., Guo, H., Jiang, Q., Allen, A.S., Zhu, M., 2014. An evaluation of copy number variation detection tools from whole-exome sequencing data. Hum. Mutat. 35, 899–907.

Venkatachalam, R., Verwiel, E.T.P., Kamping, E.J., Hoenselaar, E., Görgens, H., Schackert, H.K., van Krieken, J.H.J.M., Ligtenberg, M.J.L., Hoogerbrugge, N., van Kessel, A.G., Kuiper, R.P., 2011. Identification of candidate predisposing copy number variants in familial and early-onset colorectal cancer patients. Int. J. Canc. 129, 1635–1642.

Wang, X., Li, X., Cheng, Y., Sun, X., Sun, X., Self, S., Kooperberg, C., Dai, J.Y., 2015. Copy number alterations detected by whole-exome and whole-genome sequencing of esophageal adenocarcinoma. Hum. Genom. 9, 22.

Weren, R.D.A., Venkatachalam, R., Cazier, J.B., Farin, H.F., Kets, C.M., De Voer, R.M., Vreede, L., Verwiel, E.T.P., Van Asseldonk, M., Kamping, E.J., Kiemeney, L.A., Neveling, K., Aben, K.K.H., Carvajal-Carmona, L., Nagtegaal, I.D., Schackert, H.K., Clevers, H., Van De Wetering, M., Tomlinson, I.P., Ligtenberg, M.J.L., Hoogerbrugge, N., Geurts Van Kessel, A., Kuiper, R.P., 2015. Germline deletions in the tumour suppressor gene *FOCAD* are associated with polyposis and colorectal cancer development. J. Pathol. 236, 155–164.

Sebastià Franch-Expósito[1], Clara Esteban-Jurado[1]
Gastroenterology Department, Hospital Clínic de Barcelona, August Pi i Sunyer Biomedical Research Institute, CIBER of Hepatic and Digestive Diseases, University of Barcelona, Barcelona 08036, Spain

Pilar Garre
Molecular Oncology Laboratory, Hospital Clínico San Carlos, Health Research Institute of the Hospital Clínico San Carlos, Madrid 28040, Spain

Isabel Quintanilla, Saray Duran-Sanchon, Marcos Díaz-Gay, Laia Bonjoch
Gastroenterology Department, Hospital Clínic de Barcelona, August Pi i Sunyer Biomedical Research Institute, CIBER of Hepatic and Digestive Diseases, University of Barcelona, Barcelona 08036, Spain

Miriam Cuatrecasas
Department of Pathology, Hospital Clinic de Barcelona, Barcelona 08036, Spain

Esther Samper, Jenifer Muñoz, Teresa Ocaña, Sabela Carballal, María López-Cerón, Antoni Castells
Gastroenterology Department, Hospital Clínic de Barcelona, August Pi i Sunyer Biomedical Research Institute, CIBER of Hepatic and Digestive Diseases, University of Barcelona, Barcelona 08036, Spain

EPICOLON consortium

Maria Vila-Casadesús
Bioinformatics Platform, CIBER of Hepatic and Digestive Diseases, Barcelona 08036, Spain

Sophia Derdak, Steven Laurie, Sergi Beltran
National Center of Genomic Analysis, Science Park of Barcelona, Barcelona 08028, Spain

Jaime Carvajal
Andalusian Developmental Biology Institute, CSIC-Pablo de Olavide University-Andalusian Regional Government, Sevilla 41013, Spain

Luis Bujanda
Gastroenterology Department, Hospital Donostia–Biodonostia Institute, CIBER of Hepatic and Digestive Diseases, University of the Basque Country (UPV/EHU), San Sebastián 20080, Spain

Clara Ruiz-Ponte
Galician Public Foundation of Genomic Medicine (FPGMX), CIBER of Rare Diseases, Genomics Medicine Group, Hospital Clínico Universitario, University of Santiago de Compostela, Santiago de Compostela 15706, Spain

Jordi Camps, Meritxell Gironella
Gastroenterology Department, Hospital Clínic de Barcelona, August Pi i Sunyer Biomedical Research Institute, CIBER of Hepatic and Digestive Diseases, University of Barcelona, Barcelona 08036, Spain

Juan José Lozano
Bioinformatics Platform, CIBER of Hepatic and Digestive Diseases, Barcelona 08036, Spain

Francesc Balaguer
Gastroenterology Department, Hospital Clínic de Barcelona, August Pi i Sunyer Biomedical Research Institute, CIBER of Hepatic and Digestive Diseases, University of Barcelona, Barcelona 08036, Spain

Joaquín Cubiella
Gastroenterology Department, Complexo Hospitalario Universitario de Ourense, Ourense Biomedical Research Institute, Ourense 32005, Spain

Trinidad Caldés
Molecular Oncology Laboratory, Hospital Clínico San Carlos, Health Research Institute of the Hospital Clínico San Carlos, Madrid 28040, Spain

Sergi Castellví-Bel*
Gastroenterology Department, Hospital Clínic de Barcelona, August Pi i Sunyer Biomedical Research Institute, CIBER of Hepatic and Digestive Diseases, University of Barcelona, Barcelona 08036, Spain

* Corresponding author.
E-mail address: sbel@clinic.cat (S. Castellví-Bel).

15 June 2017
Available online 20 December 2017

[1] These two authors contributed equally to this work.

CNApp: a web-based tool for integrative analysis of genomic copy number alteracions in cancer

(preprint) ***bioRxiv***

2 de desembre del 2018
(https://doi.org/10.1101/479667)

bioRxiv preprint first posted online Nov. 27, 2018; doi: http://dx.doi.org/10.1101/479667. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

CNApp: a web-based tool for integrative analysis of genomic copy number alterations in cancer

Sebastià Franch-Expósito[1*], Laia Bassaganyas[2*], Maria Vila-Casadesús[3], Eva Hernández-Illán[1], Roger Esteban-Fabró[2], Marcos Díaz-Gay[1], Juan José Lozano[3], Antoni Castells[1], Josep M. Llovet[2,4,5], Sergi Castellví-Bel[1#], Jordi Camps[1,6#]

1 Gastrointestinal and Pancreatic Oncology Team, Institut D'Investigacions Biomèdiques August Pi i Sunyer (IDIBAPS), Hospital Clínic de Barcelona, Centro de Investigación Biomédica en Red de Enfermedades Hepáticas y Digestivas (CIBEREHD), Universitat de Barcelona, Barcelona, Catalonia, Spain

2 Liver Cancer Translational Research Group, Liver Unit, Institut D'Investigacions Biomèdiques August Pi i Sunyer (IDIBAPS), Hospital Clínic, Centro de Investigación Biomédica en Red de Enfermedades Hepáticas y Digestivas (CIBEREHD), Universitat de Barcelona, Barcelona, Catalonia, Spain

3 Bioinformatics Unit, CIBEREHD, Barcelona, Catalonia, Spain

4 Mount Sinai Liver Cancer Program, Division of Liver Diseases, Tisch Cancer Institute, Icahn School of Medicine at Mount Sinai, New York, USA

5 Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Catalonia, Spain

6 Unitat de Biologia Cel·lular i Genètica Mèdica, Departament de Biologia Cel·lular, Fisiologia i Immunologia, Facultat de Medicina, Universitat Autònoma de Barcelona, Bellaterra, Catalonia, Spain

* equal contribution; # corresponding author

Running title: CNApp: copy number integrative analysis

Keywords: Copy number alterations, hepatocellular carcinoma, colorectal cancer, pan-cancer, machine learning

bioRxiv preprint first posted online Nov. 27, 2018; doi: http://dx.doi.org/10.1101/479667. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

ABSTRACT

Copy number alterations (CNAs) are a hallmark of cancer. Large-scale cancer genomic studies have already established the CNA landscape of most human tumor types and some CNAs are recognized as cancer-driver events. However, their precise role in tumorigenesis as well as their clinical and therapeutic relevance remain undefined, thus computational and statistical approaches are required for the biological interpretation of these data. Here, we describe CNApp, a user-friendly web tool that offers sample- and cohort-level association analyses, allowing a comprehensive and integrative exploration of CNAs with clinical and molecular variables. CNApp generates genome-wide profiles, calculates CNA levels by computing broad, focal and global CNA scores, and uses machine learning-based predictions to classify samples by using segmented data from either microarrays or next-generation sequencing. In the present study, using copy number data of well-annotated 10,635 genomes from The Cancer Genome Atlas spanning 33 cancer subtypes, we showed that patterns of CNAs classified tumor subtypes according to their tissue-of-origin and that broad and focal CNA scores correlated positively in those samples with low levels of chromosome and arm-level events. Moreover, CNApp allowed the description of recurrent CNAs in hepatocellular carcinoma further confirming previous results identified using other methods. Finally, we established machine learning-based models to predict colon cancer molecular subtypes and microsatellite instability based on broad and focal CNA scores and specific genomic imbalances. In summary, CNApp facilitates data-driven research and provides a unique framework to comprehensively assess CNAs and perform integrative analyses that enable the identification of relevant functional implications.

bioRxiv preprint first posted online Nov. 27, 2018; doi: http://dx.doi.org/10.1101/479667. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

INTRODUCTION

The presence of somatic copy number alterations (CNAs) is a ubiquitous feature in cancer. Indeed, the distribution of such CNAs is sufficiently tissue-specific to distinguish and enable the classification of tumor entities (Ried et al. 2012; Taylor et al. 2018a), and may allow identifying groups of tumors responsive to particular therapies(Cairncross et al. 2013; Davoli et al. 2017). Moreover, high levels of CNAs, which result from aneuploidy and chromosome instability, are generally associated with high-grade tumors and poor prognosis (Sansregret et al. 2018). Two main subtypes of CNAs can be discerned: broad CNAs, which are defined as whole-chromosome and chromosomal arm-level alterations, and focal CNAs, which are alterations of limited size ranging from part of a chromosome-arm to few kilobases (Krijgsman et al. 2014; Zack et al. 2013). Recently, it has been uncovered that while focal events mainly correlate with cell cycle and proliferation markers, broad aberrations are mainly associated with immune evasion markers (Taylor et al. 2018b; Davoli et al. 2017; Buccitelli et al. 2017). Nevertheless, the precise role of CNAs in tumor initiation and progression, as well as their clinical relevance and therapeutic implications remain still poorly understood.

Characterization and interpretation of CNAs is time-consuming and very often requires complex integrative analyses with clinical and molecular information. Moreover, visualization of complex data is usually essential to discriminate key results. Well-established CNA algorithms, such as the gold-standard circular binary segmentation, determine the genomic boundaries of copy number gains and losses based on signal intensities or read depth obtained from array comparative genomic hybridization and SNP-array or next-generation sequencing data, respectively (Olshen et al. 2004). Variability within gain or loss levels can be addressed with the algorithm CGHcall, which enables the identification of single copy number changes (van de Wie et al. 2007). In order to overcome the complex nature of tumor samples (Stratton et al. 2009), more recent

bioRxiv preprint first posted online Nov. 27, 2018; doi: http://dx.doi.org/10.1101/479667. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

segmentation methods improved the accuracy to identify copy number segments either by considering the B allele frequency (BAF), such as ExomeCNV (Sathirapongsasuti et al. 2011), Control-FREEC (Boeva et al. 2012) and SAAS-CNV (Zhang and Hao 2015), or through adjusting by sample purity and ploidy estimates, such as GAP (Popova et al. 2009), ASCAT (Van Loo et al. 2010) and ABSOLUTE (Carter et al. 2012). However, the state-of-the-art computational approach for CNA analysis is GISTIC2.0 (Mermel et al. 2011), which is a gene-centered probabilistic method that enables to define the boundaries of recurrent putative driver CNAs in large cohorts (Beroukhim et al. 2010). Nevertheless, despite ongoing progress on identifying CNAs, to our knowledge there is no bioinformatic tool readily available for integrative analyses to unveil the biological interpretation of these CNAs.

To address this issue, we developed CNApp, the first open-source application to comprehensively analyze and integrate CNA profiles with molecular and clinical variables. CNApp was built in Shiny R package (Chang et al. 2018) and provides the user with high-quality interactive plots and statistical correlations between CNAs and annotated variables in a fast and easy-to-explore interface. In particular, CNApp uses genomic segmented data to quantify CNA levels based on broad and focal genomic alterations, assess differentially altered genomic regions, and perform machine learning-based predictions to classify tumor samples. A dataset including 160 colon cancer samples with clinical annotation is loaded for demonstration purposes. To exemplify the applicability and performance of CNApp, we used publicly available segmented data from The Cancer Genome Atlas (TCGA) to (i) measure the burden of global, broad, focal CNAs as well as generate CNA profiles in a pan-cancer dataset spanning 33 cancer subtypes, (ii) identify cohort-based recurrent CNAs in hepatocellular carcinoma and compare it with previously reported data using different methods, and (iii) assess predicting models for colon cancer molecular subtype and microsatellite instability status

bioRxiv preprint first posted online Nov. 27, 2018; doi: http://dx.doi.org/10.1101/479667. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

classification based on CNA scores and specific genomic imbalances. CNApp is hosted at http://bioinfo.ciberehd.org/CNApp and the source code is freely available at GitHub (https://github.com/ait5/CNApp).

RESULTS

Implementation and basic usage

Functions of CNApp comprise three main sections: 1- *Re-Seg & Score: re-segmentation, CNA scores computation and variable association*, 2- *Region profile: genome-wide CNA profiling*, and 3- *Classifier model: machine learning classification model predictions*, (Figure 1). Each of these sections and their key functions are described below. The input file consists of a data frame with copy number segments provided by any segmentation algorithm. Mandatory fields and column headers are sample name (*ID*), chromosome (*chr*), start (*loc.start*) and end (*loc.end*) genomic positions, and the log2 ratio of the copy number amplitude (*seg.mean*) for each segment. If available, it is recommended to include sample purity (*purity*) and BAF values (*BAF*), which can improve the accuracy of CNA calls and will provide information of copy number neutral loss-of-heterozygosity (CN-LOH) events. Annotation of variables can be included in the input file (tagged in every segment from each sample) or by loading an additional file specifying new variables to every sample.

Section 1. Re-Seg & Score: re-segmentation, CNA scores computation and variable association

First, CNApp applies a re-segmentation approach aiming at correcting potential background noise and amplitude divergence due to technical variability. Default re-segmentation settings include *minimum segment length* (100 Kbp), *maximum distance between segments* (1 Mbp), *maximum amplitude (seg.mean) deviation between segments*

bioRxiv preprint first posted online Nov. 27, 2018; doi: http://dx.doi.org/10.1101/479667. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

(0.16), *minimum amplitude (seg.mean) deviation from segment to zero* (0.16), and *maximum BAF deviation between segments* (0.1). These parameters can be customized by the user to better adjust the re-segmentation and CNA calling for each particular dataset. Re-segmented data are then used to calculate the focal (FCS), broad (BCS) and global (GCS) CNA scores, which provide three different quantification of CNA levels for each sample. To compute these scores, CNApp classifies and weights CNAs based on amplitude and length. A weight is given to each segment according to its *seg.mean* value and by applying low-, medium- and high-level copy number amplitude thresholds. By considering the relative length of each segment to the whole-chromosome or chromosome arm, segments are tagged as *chromosomal* -by default, 90% or more of the chromosome affected-, as *arm-level* -50% or more of the chromosome arm affected-, or as *focal* -less than 50% of the chromosome arm affected. Percentages for relative lengths are also customizable. For each sample, BCS is computed by considering chromosome and arm-level segment weights according to the amplitude value. Likewise, calculation of FCS takes into account weighted focal CNAs and the amplitude and length of the segment. Finally, GCS is computed by considering the sum of normalized FCS and BCS values, providing an overall assessment of the CNA burden for each sample. To assess the reliability of CNA scores, we compared each score with the corresponding fraction of altered genome using a TCGA pan-cancer set of 10,635 samples. Both FCS (values ranging from 5 to 2,466) and BCS (ranging from 0 to 44) highly correlated with the fraction of altered genome by focal and broad copy number changes, respectively (Spearman's rank correlation for BCS = 0.957 and for FCS = 0.938) (Supplemental_Fig_S1 A and B). As expected, GCS (values ranged from -1.93 to 12.60) highly correlated with the fraction of altered genome affected by both focal and broad CNAs (Spearman's rank correlation for GCS = 0.963) (Supplemental_Fig_S1C).

bioRxiv preprint first posted online Nov. 27, 2018; doi: http://dx.doi.org/10.1101/479667. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

Additionally, parametric and non-parametric statistical tests are used to establish associations between CNA scores and annotated variables from the input file.

Section 2. Region profile: genome-wide CNA profiling

This section utilizes re-segmented data obtained from section 1 or uploaded segmented data without re-segmentation to generate genomic region profiling and sample-to-sample correlations. To conduct this, re-segmented data are transformed into genome region profiles according to a user-selected genomic window (i.e., chromosome arms, half-arms, cytobands, sub-cytobands or 40-1 Mbp windows). All segments, or either only broad or only focal can be selected for this analysis. Length-relative means are computed for each window by considering amplitude values from those segments included in each specific window. Default thresholds for low-level copy number gains and losses (i.e., |0.2|) are used as cutoffs to classify genome regions and to calculate their frequencies in this section. Genome-region profiles are presented in genome-wide heatmaps to visualize general copy number patterns. Up to six annotation tracks can be added and plotted simultaneously allowing visual comparison and correlation between CNA profiles and different variables, including the CNA scores obtained in section 1. Generation of hierarchical clusters by samples and regions is optional. CNA frequency summaries by genomic region and by sample are represented as stacked bar plots.

Importantly, assessing differentially altered regions between sample groups might contribute to discover genomic regions associated with annotated variables and thus unveil the biological significance of specific CNAs. To do so, CNApp interrogates descriptive regions associated with any sample-specific annotation variable provided in the input file. Student's t-test or Fisher's test are applied when considering CNAs as continuous alterations (*seg.mean* values) or as categorical events (presence of gains and losses), respectively. Default statistical significance is set to *P*-value lower than 0.1.

bioRxiv preprint first posted online Nov. 27, 2018; doi: http://dx.doi.org/10.1101/479667. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

However, p-value thresholds can be defined by the user and adjusted *P*-value is optional. A heatmap plot allows the visualization and interpretation of which genome regions are able to discriminate between sample groups. By selecting a region of interest, box plots and stacked bar plots are generated comparing *seg.mean* values and alteration counts in Student's t-test and Fisher's test tabs, respectively. Additionally, genes comprised in the selected region are indicated.

3. Classifier model: Machine learning classification model predictions

This section allows the user to generate machine learning-based classifier models by choosing a variable to define sample groups and one or multiple classifier variables. To do so, CNApp incorporates the *randomForest* R package (Liaw and Wiener 2002). The model construction is performed 50-times and bootstrap set is changed in each iteration. By default, only annotation variables from the input file are loaded to work either by group defining or by classifier variables. If *Re-Seg & Score* and/or *Region profile* sections have been previously completed, the user can upload data from these sections (i.e., CNA scores and genomic regions). Predictions for the model performance are generated and the global accuracy is computed along with sensitivity and specificity values by group. Classifier models can be useful to point out candidate clinical or molecular variables to classify sample subgroups. A summary of the data distribution and plots for real and model-predicted groups are visualized. A table with prediction rates throughout the 50-times iteration model and real tags by sample is displayed and can be downloaded.

Genomic characterization of cancer subtypes

First, we evaluated the capacity of CNApp to analyze and classify cancer subtypes according to distinct patterns of CNA scores, and assess whether CNApp was able to reproduce the distribution of cancer subtypes based on specific CNA profiles. To do so,

bioRxiv preprint first posted online Nov. 27, 2018; doi: http://dx.doi.org/10.1101/479667. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

level 3 publicly available Affymetrix SNP 6.0 array data from 10,635 tumor samples spanning 33 cancer types from TCGA pan-cancer database were used. We applied *Re-Seg & Score* and *Region profile* using default parameters to obtain re-segmented data, CNA scores, and cancer-specific CNA profiles. Correlations between CNA scores were assessed by computing Spearman's rank test, obtaining values of 0.59 between BCS and FCS, 0.90 between BCS and GCS, and 0.85 between FCS and GCS. In addition, we further assessed the correlation between BCS and FCS for each individual BCS value. While tumors with low BCS displayed a positive correlation between broad and focal alterations, tumors did not maintain such correlation in higher BCS values (Supplemental_Fig_S2A). BCS, FCS and GCS distributions across cancer subtypes supported the existence of distinct CNA levels between tumors from different origin (Figure 2A). While cancer subtypes such as acute myeloid leukemia (LAML), thyroid carcinoma (THCA) or thymoma (THYM) showed low levels of broad and focal events (GCS median values of -1.67 for LAML, -1.68 for THCA, and -1.52 for THYM), uterine carcinosarcoma (UCS), ovarian cancer (OV) and lung squamous cell carcinoma (LUSC) displayed high levels of both types of genomic imbalances (GCS median values of 2.55, 2.44, and 0.97 for UCS, OV, and LUSC, respectively). Some cancer subtypes displayed a preference for either broad or focal copy number alterations. For example, kidney chromophobe (KICH) tumors showed the highest levels of broad events (median BCS value of 27); however, they were amongst those subtypes with less focal CNAs (median FCS value of 49). In contrast, breast cancer (BRCA) samples displayed high values for FCS (median FCS value of 150), while BCS values were intermediate (median BCS value of 7).

Subsequent analysis aimed at generating genome-wide patterns for each cancer subtype based on chromosome-arm genomic windows and the overall corresponding frequencies (Figure 2B). We found that chromosome arms altered in more than 25% across all

bioRxiv preprint first posted online Nov. 27, 2018; doi: http://dx.doi.org/10.1101/479667. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

samples were 1q, 7p, 7q, 8q and 20q for copy number gains, and 8p and 17p for copy number losses. Conversely, chromosome arms affected by CNAs in less than 10% of all cancer subtypes included 2q and 19p (Figure 2C). By using a subset of 20 out of the 33 cancer types for which tumor type information was available, we asked CNApp to compute the average arm-region for each cancer type to assess if they clustered according to their CNA profile (Supplemental_Fig_S2B). Our analysis showed that correlation profiles resulting from Pearson's test were hierarchically clustered according to their tumor type (Figure 2D). Gastrointestinal (colon, rectum, stomach and pancreatic), gynecological (ovarian and uterine) and squamous (cervical, head and neck, and lung) cancers clustered together based on specific CNA profiles for each group (Figure S2B). These results strongly correlated with previously reported findings (Taylor et al. 2018b; Hoadley et al. 2018).

Identification of recurrent CNAs in liver hepatocellular carcinoma

Next, we attempted to test the ability of CNApp to identify recurrent broad and focal CNAs in a large cohort of samples. For that reason, we chose to perform CNA analysis of 370 samples from TCGA corresponding to the Liver Hepatocellular Carcinoma (LIHC) cohort, robustly reproducing previous findings reported by GISTIC2.0 (Ally et al. 2017). The overall pattern of recurrent broad and focal CNAs described in the TCGA study was similar to earlier reports, confirming the specific copy number profile for hepatocellular carcinoma (HCC) (Chiang et al. 2008; Guichard et al. 2012; Wang et al. 2013; Totoki et al. 2014; Schulze et al. 2015). By using GISTIC2.0, the most frequent broad alterations in LIHC were gains at 1q (61%) and 8q (52%), and losses at 8p (70%) and 17p (56%) (Supplemental_Table_S1). Recurrent focal amplifications involved the well-characterized driver oncogenes *CCND1* and *FGF19* (11q13.3), *MYC* (8q24.21), *MET* (7q31.2), *VEGFA* (6p21.1) and *MCL1* (1q21.3), and the most recurrent deletions

bioRxiv preprint first posted online Nov. 27, 2018; doi: http://dx.doi.org/10.1101/479667. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

included tumor suppressor genes such as *RB1* (13q14.2) and the *CDKN2A* (9p21.3) genes (Supplemental_Table_S2).

By applying the default parameters of CNApp to the LIHC dataset and selecting chromosome arms as genomic regions to assess broad events, we consistently found copy number gains at 1q (56%) and 8q (46%), and copy number losses at 8p (62%) and 17p (47%) as the most frequent alterations (Figure 3A). The slightly lower rate tendency of broad CNAs from CNApp as compared to GISTIC2.0 also appeared in the subsequent recurrent broad alterations (Supplemental_Table_S1). For instance, GISTIC2.0 significantly detected gains with rates between 25-40% on eight additional chromosome-arms, including 5p, 5q, 6p, 20p, 20q, 7p, 7q, and 17q, which were identified by CNApp in 20-30% of the samples. Similarly, GISTIC2.0 significantly detected broad deletions at frequencies between 20-40% on 18 additional chromosome-arms, of which 4q, 6q, 9p, 13q, 16p, and 16q losses were observed at ≥20% by CNApp, and the rest of them displayed rates between 10-20%. In this case, discrepancies in CNA frequencies were expected considering the lower copy number amplitude thresholds used by GISTIC2.0 in comparison with the CNApp default cutoffs (|0.1| vs |0.2|, corresponding to ~2.14/1.8 copies vs 2.3/1.7 copies, respectively). Indeed, previous reports analyzing CNAs in other HCC cohorts and using greater copy number thresholds, showed frequencies of alterations similar to those estimated by CNApp (Chiang et al. 2008; Guichard et al. 2012; Wang et al. 2013; Schulze et al. 2015). To assess the impact of modifying CNApp amplitude thresholds, we next re-run the software dropping the minimum copy number values to |0.1|. As expected, the overall number of broad alterations increased, reaching frequency values similar or even higher than those reported by GISTIC2.0 (Figure 3B and Supplemental_Table_S1). Of note, such drop from 0.2 to 0.1 might facilitate the identification of subclonal genomic imbalances, which are very frequent among tumor samples (McGranahan and Swanton 2017), though it can also increase the number of false

bioRxiv preprint first posted online Nov. 27, 2018; doi: http://dx.doi.org/10.1101/479667. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

positive calls. Furthermore, we assessed whether the identification of broad events was affected by two additional parameters: (i) the relative length to classify a segment as *arm-level* alteration, and (ii) the re-segmentation provided by CNApp. As expected, increasing the percentage of chromosome arm required to classify a CNA segment as *arm-level* (from ≥ 50% to ≥ 70%) or skipping the re-segmentation step led to an underestimation of some broad events, whereas decreasing the percentage of chromosome arm (from ≥50% to ≥40%) resulted in the opposite (Supplemental_Fig_S3A-C and Supplemental_Table_S1).

As far as focal CNAs are concerned, CNApp and GISTIC2.0 use different strategies to quantify their recurrence. Therefore, the comparison between the two methods was evaluated in a more indirect manner. GISTIC2.0 constructs minimal common regions (also known as 'peaks') that are likely to be altered at high frequencies in the cohort, which are scored using a Q-value and may present a wide variety of genomic lengths (Mermel et al. 2011). Instead, CNApp allows dividing the genome in windows of different sizes, calculating an average of the copy number amplitudes of segments included within the selected windows. We reasoned that considering the length of GISTIC2.0 reported 'peaks', CNApp might also be capable to identify focal recurrently altered regions by dividing the genome in windows of a relatively small size. To test our hypothesis, we asked CNApp to calculate the frequency of focal gains and losses by dividing the genome by sub-cytobands. As a result, CNApp consistently localized the most frequently altered sub-cytobands (found in 10-25% of samples), including gains at 1q21.3 (25%), 8q24.21 (17%, *MYC*), 5p15.33 (13%, *TERT*), 11q13.3 (12%, *CCND1*/*FGF19*) and 6p21.1 (11%, *VEGFA*), and losses at 13q14.2 (20%, *RB1*), 1p36.11 (18%, *ARID1A*), 4q35.1 (17%, *IRF2*) and 9p21.3 (14%, *CDKN2A*), which are in agreement with previous studies in HCC (Figure 3C and Supplemental_Table_S2) (Chiang et al. 2008; Guichard et al. 2012; Wang et al. 2013; Schulze et al. 2015).

bioRxiv preprint first posted online Nov. 27, 2018; doi: http://dx.doi.org/10.1101/479667. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

Compared to GISTIC2.0, CNApp reported 14 of the 27 significant amplifications and 14 of the 34 significant deletions at rates >10%, and the remaining alterations displaying rates between 4-10% (Supplemental_Table_S3) (Wang et al. 2013). Most importantly, regions with the highest frequency detected by CNApp showed a good match with lowest GISTIC2.0 Q-residual values, indicating that the most significant 'peaks' identified by GISTIC2.0 were actually included in the most recurrently altered sub-cytobands reported by CNApp.

As previously suggested, recurrent focal alterations often occur at lower frequencies than broad events (Beroukhim et al. 2010). However, previous studies describing the genomic landscape of HCC mostly focused on high-level focal CNAs (from >3 copies for gains and from <1.3 copies for losses), thus reporting lower frequencies than those estimated by CNApp (Chiang et al. 2008; Guichard et al. 2012; Schulze et al. 2015). Interestingly, excluding the low-level alterations and evaluating only the moderate and high-amplitude events (≥3 and ≤1 copies), frequencies dropped to values closer to those previously reported (Figure 3D and Supplemental_Table_S2). Amplifications reached maximum rates of 11%, whereas losses ended up at rates of ~2%, in consistence with the observation that high-level CNAs are relatively rare (Zack et al. 2013). Top recurrent gains involved sub-cytobands 1q21.3 (11%) and 8q24.21 (11%, *MYC*), 11q13.3 (7%, *CCND1*/*FGF19*), and 5p15.33 (5%, *TERT*). Recurrent losses estimated at ~2% of the samples included 13q14.2 (*RB1*), 9p21.3 (*CDKN2A*), 4q35.1 (*IRF2*), and 8p23.1. Slight discrepancies between frequencies might be explained by minimal variability in the copy number threshold.

Classification of colon cancer according to CNA scores and genomic regions

A proposed taxonomy of colorectal cancer (CRC) includes four consensus molecular subtypes (CMS), mainly based on differences in gene expression signatures. Accordingly,

bioRxiv preprint first posted online Nov. 27, 2018; doi: http://dx.doi.org/10.1101/479667. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

each CMS shows specific molecular features such as microsatellite instability (MSI) status, CpG island methylator phenotype (CIMP) levels, somatic CNAs and non-synonymous mutations. Briefly, CMS1 includes the majority of hypermutated tumors showing MSI, high CIMP, and low levels of CNAs; CMS2 and 4 typically comprise microsatellite stable (MSS) tumors with high levels of CNAs; and finally, mixed MSI status and low levels of CNAs and CIMP are associated with CMS3 tumors (Guinney et al. 2015). Using a representative cohort of 309 colon cancers from the TCGA Colon Adenocarcinoma (COAD) cohort (Cancer and Atlas 2012) with known CMS classification (CMS1, N = 64; CMS2 N = 112; CMS3 N = 51; CMS4 N = 82) and MSI status, we asked CNApp to generate a genome-wide frequency plot after re-segmentation using the default copy number thresholds and excluding segments smaller than 500 Kbp to avoid technical background noise. CNA profiles were generated using genomic regions defined by chromosome arms. As expected, the frequency plot displayed the most commonly altered genomic regions in sporadic CRC (Camps et al. 2008; Cancer and Atlas 2012; Ried et al. 1996; Meijer et al. 1998; Nakao et al. 2004). By assessing the broad CNA events in the entire cohort, we observed that the most frequently altered chromosome arms were gains of 7p, 7q, 8q, 13q, 20p, and 20q, and losses of 8p, 17p, 18p, and 18q, occurring in more than 30% of the samples (Figure 4A). Focal CNAs were obtained by generating genomic regions by sub-cytobands. Of note, five out of six genomic losses and five out of 18 genomic gains contained deletions and amplifications, respectively, identified by GISTIC2.0 in the COAD TCGA cohort.

Subsequently, we performed integrative analysis of genomic imbalances, CMS groups, and CNA scores. By using CNApp, we assessed whether CNA scores were able to classify colon cancer samples according to their CMS. While BCS established significant differences between CMS paired comparisons ($P \leq 0.0001$, Student's t-test), FCS poorly discern CMS1 from 3 and CMS2 from 4 (Figure 4B and Supplemental_Fig_S4A). Thus,

bioRxiv preprint first posted online Nov. 27, 2018; doi: http://dx.doi.org/10.1101/479667. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

we reasoned that broad CNAs rather than focal were able to better discriminate between different CMS groups. In fact, the distribution of CMS groups based on BCS resembled the distribution of somatic CNA counts defined by GISTIC2.0 (Guinney et al. 2015), which agrees with the observation that BCS highly correlates with the fraction of altered genome (Supplemental_Fig_S1A). Subsequently, we integrated the BCS and the CMS groups with the microsatellite status. Our results showed an average BCS of 1.51 2.11 and 10.25 5.92 for MSI (N = 72) and MSS (N = 225) tumors, respectively. In addition, a BCS of 4, corresponding to the 90th percentile in the MSI sample set, was able to differentiate MSI and MSS tumors. Applying this cutoff, 186 out of 225 (83%) of MSS tumors showed a BCS greater than 4 (Figure 4C). In contrast, 39 (17%) MSS tumors showed a BCS value of 4 or lower, corresponding to three CMS1, six CMS2, 18 CMS3 and 12 CMS4 tumors, further demonstrating the existence of MSS tumors with a very low CNA burden. When we assessed the level of focal alterations in this subset of MSS samples by considering the 90th percentile of FCS in the MSI group (37.2), we could determine that eight of these MSS tumors showed high FCS, thus reducing the percentage of MSS tumors with overall low copy number changes to 13%. On the other hand, seven MSI tumors showed BCS higher than 4. Among these, five samples displayed genomic imbalances typically associated with the CRC canonical pathway, including a focal amplification of *MYC*, unveiling tumors with co-occurrence of MSI and extensive genomic alterations (Trautmann et al. 2006). Our dataset comprised nine out of 51 CMS3 tumors with MSI. Intriguingly, two of them showed focal deletions on chromosome 2 involving *MSH2* and *MSH6*, suggesting the inactivation of these mismatch repair genes through a focal genomic imbalance. In fact, 46% of CMS3 MSS tumors showed BCS below 4, in agreement with the finding that CMS3 tumors display low levels of somatic CNAs.

bioRxiv preprint first posted online Nov. 27, 2018; doi: http://dx.doi.org/10.1101/479667. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

CNApp enable the identification of possible sample misclassifications by integrating CMS annotation and *BRAF*-mutated sample status.. As expected, CMS1 cases were enriched for *BRAF* mutation. Nevertheless, two CMS4 samples also showed mutations in *BRAF*. One of these samples showed a BCS of 11, displaying canonical CNAs. In contrast, the other CMS4 BRAF-mutated sample showed MSI and a BCS of 0, similar features as in CMS1. Likewise, four *BRAF* WT samples, classified within the CMS4 group, displayed MSI and a BCS of 0, thus being candidates to be labeled as CMS1 based on the levels of CNAs (Figure 4D). These disparities are of utmost importance since recent studies reported that high copy number alterations correlate with reduced response to immunotherapy (Davoli et al. 2017). Importantly, it has been suggested that MSI status might be predictive of positive immune checkpoint blockade response in advanced CRC, probably due to the low levels of CNA usually presented by MSI tumors (Le et al. 2015). We next asked CNApp to compare genomic regions differentially represented in the four CMS groups based on a Student's t-test or Fisher's test with adjusted p-value. By applying a Student's t-test, we could observe that CMS1 resembled CMS3, except for the gain of chromosome 7 and the loss of 18q, which were the alterations that commonly appeared in CMS3 samples with BCS above 4 ($P \leq 0.001$, Student's t-test) (Supplemental_Fig_S4B). Even though only subtle CNA differences between CMS2 and CMS4 were identified, the loss of 14q was significantly more detected in CMS2 (42%) than in CMS4 (17.1%) ($P \leq 0.005$, Student's t-test) (Supplemental_Fig_S4B). Visually exploring the heatmap plot and further analyzing specific regions, we observed that the gain of 12q was more frequently associated with CMS1 than CMS2 ($P \leq 0.005$, Student's t-test), in agreement with previous studies reporting that the gain of chromosome 12 is associated with microsatellite unstable tumors (Supplemental_Fig_S4B) (Trautmann et al. 2006). Intriguingly, the gain of the chromosome arm 20q alone mimicked the distribution of somatic CNAs defined by GISTIC2.0 across consensus subtype samples

bioRxiv preprint first posted online Nov. 27, 2018; doi: http://dx.doi.org/10.1101/479667. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

(Figure 4E) (Guinney et al. 2015). In fact, chromosome arm 20q was gained in 99.1%, 70.7%, 39.2%, and 10.9% of CMS2, CMS4, CMS3 and CMS1 tumors, respectively.

Finally, we applied machine learning-based prediction models to classify samples by their MSI status or CMS. BCS predicted MSI status with a global accuracy of 82.2%. This was consistent with the fact that BCS was able to distinguish CMS1 from CMS2 with 89.2% of accuracy. However, when we tested the performance of BCS to predict any CMS group, the accuracy was only 47.5%, indicating that BCS alone is a poor predictive variable to assess CMS. We then used the most discriminative descriptive regions among CMS groups (i.e., 13q, 17p, 18, and 20q), and reached an accuracy to correctly predict CMS of 55%. In fact, the occurrence of these genomic alterations was able to differentiate CMS2 from CMS4 with an accuracy of 70%, and CMS1 from CMS3 with a 72.3% accuracy. As expected, this set of genomic alterations distinguished CMS1 from CMS2 samples with an accuracy of 95%. Altogether, these data suggest that CNApp might provide insight into further classifying CRC samples in CMS groups.

DISCUSSION

Here we present CNApp, a web-based computational approach to analyze and integrate CNAs associated with molecular and clinical variables. CNApp calculates CNA scores to quantify focal, broad and global levels of alterations for each individual sample after an optional process of re-segmentation. Moreover, CNApp utilizes genomic imbalances selected by the user to assess classifier variables by computing machine learning-based models. Although CNApp has been developed using segmented genomic copy number data obtained from SNP-arrays, the software is also able to accommodate segmented data from next-generation sequencing.

Overall, CNApp was benchmarked by analyzing a pan-cancer TCGA dataset with more than 10,000 samples, being able to cluster major tumor types according to CNA patterns.

bioRxiv preprint first posted online Nov. 27, 2018; doi: http://dx.doi.org/10.1101/479667. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

Moreover, our results demonstrate the reliability of CNApp in identifying regions encompassing the most recurrent CNAs. The software successfully reproduced the well-characterized genomic profile of HCC and CRC, considering both broad and focal events. Although CNApp has not been developed to define the precise boundaries of focal events, the software is capable to detect which regions are likely to contain the most recurrent alterations. However, we acknowledge that the characterization of focal alterations potentially containing driver events performed by GISTIC2.0 is more accurate than the genomic windows provided by CNApp. Thus, despite the in-depth comparison described here, we consider CNApp as a complementary tool rather than a replacement for GISTIC2.0.

Finally, applying CNApp to a colon cancer dataset for which clinical features were known allowed the determination of a BCS value of 4 to potentially discriminate MSI from MSS tumors. Most importantly, due to the inverse correlation between MSI and aneuploidy in CRC, our results suggest that this BCS value could be established as a cutoff to define the edge between low and high aneuploid tumors. Nevertheless, these results ought to be further validated in an independent cohort. Since high levels of aneuploidy correlate with immune evasion markers, quantification of CNAs and their association with molecular and clinical features might be of extreme relevance. In fact, specific genomic regions defined by CNApp contributed to classify the consensus molecular subtypes. This is of clinical interest as it is known that CMS1 microsatellite unstable tumors might show a positive response to immuno-related treatments. Therefore, we believe that CNApp enables not only the fundamental analysis of CNA profiles, but also the functional understanding of CNAs in the context of clinical samples and their potential use as biomarkers.

METHODS

bioRxiv preprint first posted online Nov. 27, 2018; doi: http://dx.doi.org/10.1101/479667. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

Data set availability

CNA data from TCGA: pan-cancer cohort

Affymetrix SNP6.0 array copy number segmented data (Level 3) from 10,635 samples spanning 33 cancer types from TCGA pan-cancer dataset were downloaded from Genomic Data Commons (National Cancer Institute, NIH) (Grossman et al. 2016). This dataset included the 370 Liver Cancer-Hepatocellular Carcinoma (LIHC) samples used for the analysis of recurrent CNAs and the subset of 309 samples from Colon Adenocarcinoma (COAD) for which the colorectal cancer consensus molecular subtype (CMS) was known (Guinney et al. 2015).

GISTIC data from TCGA: LIHC cohort

GISTIC 2.0.22 (Ally et al. 2017) copy number results (Level 4) of the 370 LIHC samples, were downloaded from the Broad Institute GDAC Firehose. Parameters used for the analysis are detailed in the same GDAC repository. Specifically, parameters conditioning the definition of the CNAs and of interest for our comparison were publicly reported with the following values: *amplification* and *deletion thresholds*: 0.1; *broad length cutoff*: 0.7; *joint segment size*: 4.

Software and tool availability

CNApp can be accessed at http://bioinfo.ciberehd.org/CNApp. It was developed using Shiny R package (version 1.1.0), from R-Studio (Chang et al. 2018). The tool was applied and benchmarked while using R version 3.4.2 (2017-09-28) -- "Short Summer". List of packages, libraries and base coded are freely available at GitHub, and instructions for local installation are also specified.

CNA scores computation

bioRxiv preprint first posted online Nov. 27, 2018; doi: http://dx.doi.org/10.1101/479667. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

Segments resulted from re-segmentation (or original segments from input file when re-segmentation is skipped) are classified in *chromosomal*, *arm-level* and *focal* events by considering the relative length of each segment to the whole-chromosome or chromosome arm. Using default parameters, segments are tagged as *chromosomal* when 90% or more of the chromosome is affected; as *arm-level* when 50% or more of the chromosome arm affected; and as *focal* when affecting less than 50% of the chromosome arm. Percentages for relative lengths are customizable. Broad (chromosomal and arm-level) and focal alterations are then weighted according to their amplitude values (*seg.mean*) and taking into account copy number amplitude ranges defined by CNA calling thresholds and specified in Supplemental_Methods.

Broad CNA Score (BCS): for a total *N* of broad events in a sample (*x*), it equals to the summation of segments weights (*A*) in that corresponding sample and being *i* the corresponding segment:

$$BCS(x) = \sum_{i=1}^{N} A_i$$

Focal CNA Score (FCS): same as in BCS, with an additional pondering value *L* included to the summation, which captures the relative size of the chromosome-arm coverage of each focal CNA (according to weights specified in Supplemental_Methods):

$$FCS(x) = \sum_{i=1}^{N} A_i \cdot L_i$$

Global CNA Score (GCS): for a sample *x*, it is calculated as the summation of normalized BCS and FCS values, where *meanBCS* and *meanFCS* stand for mean values of BCS and FCS from total samples, respectively, and *sdBCS* and *sdFCS* stand for standard deviation values of BCS and FCS from total samples, respectively:

$$normBCS(x) = \frac{BCS(x) - meanBCS}{sdBCS} \qquad normFCS(x) = \frac{FCS(x) - meanFCS}{sdFCS}$$

bioRxiv preprint first posted online Nov. 27, 2018; doi: http://dx.doi.org/10.1101/479667. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

$$GCS(x) = \sum_{i=1}^{N} normBCS_i + normFCS_i$$

Genomic windows computation

Region profiling section allows genome segmentation analysis by user-selected windows (i.e. arms, half-arms, cytobands, sub-cytobands, and 40Mb till 1Mb). In order to do that, windows files were generated for each option and genome build (*hg19* and *hg38*). Cytobands file *cytoBand.txt* from UCSC page and for both genome builds was used as mold to compute regions (Casper et al. 2017).

Segmented samples are transformed into genome region profiles using genomic windows selected by user. Segments from each sample are consulted to assess whether or not overlap with the window region. Thus, window-means (W) are computed for each genomic window by collecting segments (t) overlapping with window-region (i). Segments with *loc.start* or *loc.end* position falling within the region are collected, as well as those segments embedding the entire region. At this point, the summation of each segment-mean (S) corrected by the relative window-length (L) affected by the segment length (l) is performed:

$$W(i) = \sum_{t=1}^{n} S_t \cdot \frac{l_t}{L(i)}$$

Descriptive regions assessment

Potential descriptive regions between groups defined by the annotated variables provided in the input file can be studied and *P*-values are presented to evaluate significance in differentially altered regions between those groups. The alterations can be considered as (1) numerical continuous (*seg.mean* values) and (2) categorical variables (gains, losses and non-altered). In the first case, to assess statistical significance between groups

bioRxiv preprint first posted online Nov. 27, 2018; doi: http://dx.doi.org/10.1101/479667. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

Student's T-test is applied, whereas in the second situation the significance is assessed by applying the Fisher's exact test. False discovery rate (FDR) adjustment is performed using the Benjamini-Hochberg (BH) procedure in both cases and corrected *P*-values (*Adj.p-value*) or non-corrected *P*-values (*p-values*) are displayed by user selection.

Machine learning-based classifier models

We used the *randomForest* R package (Liaw and Wiener 2002) to compute machine learning classifier models. Variables to define sample groups must be selected, as well as at least one classifier variable. Model construction is performed 50-times and training set is changed by iteration. In order to compute model and select training set, multiple steps and conditions have to be accomplished:

i. total *N* samples divided by *G* groups depicted by group-defining variable must be higher than n samples from the smaller group:

$$P = \frac{N}{G} \;;\; P > n$$

ii. If condition above is not accomplished, then *P* is set to 75% of n:

$$\text{if } P \leq n \text{ then } P = n \cdot 0.75$$

iii. *P* term must be higher than one, and *N* must be equal or higher than 20:

$$P > 1 \text{ or } N \geq 20$$

iv. Classifier variables, when categorical, shall not have higher number of tags (*Z*) than groups defined (*G*) by group-defining variable:

$$Z < G$$

bioRxiv preprint first posted online Nov. 27, 2018; doi: http://dx.doi.org/10.1101/479667. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

v. Training set (T) is computed and merged for each group (g) from groups (G) defined by group variable, extracting P samples from g as follows:

$$t\,(g) = P \text{ samples from } g \qquad\qquad T = \sum_{i=1}^{g} t_i$$

After model computation, contingency matrix with prediction and reference values by group is created to compute accuracy, specificity and sensitivity by group.

Ethics approval and consent to participate

Ethics approval was not required for this study.

Consent for publication

Not applicable.

Conflicts of interests

Dr. Llovet is receiving research support from Bayer HealthCare Pharmaceuticals, Eisai Inc, Bristol-Myers Squibb and Ipsen, and consulting fees from Eli Lilly, Bayer HealthCare Pharmaceuticals, Bristol-Myers Squibb, EISAI Inc, Celsion Corporation, Exelixis, Merck, Ipsen, Glycotest, Navigant, Leerink Swann LLC, Midatech Ltd, and Nucleix.

Funding

This work has been supported by grants from the European Commission (PCIG11-GA-2012-321937), the Instituto de Salud Carlos III and co-funded by the European Regional Development Fund (ERDF) (CP13/00160, PI14/00783, PI17/01304, PI17/00878), the CIBEREHD program, the CERCA Program (Generalitat de Catalunya), the Agència de

bioRxiv preprint first posted online Nov. 27, 2018; doi: http://dx.doi.org/10.1101/479667. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

Gestió d'Ajuts Universitaris i de Recerca, Generalitat de Catalunya (2017 SGR 1035, 2017 SGR 21, and 2017 SGR 653), PERIS (SLT002/16/00398, Generalitat de Catalunya), Fundación Científica de la Asociación Española Contra el Cáncer (GCB13131592CAST). SF-E was supported by a contract from CIBEREHD. LB and MD-G were supported by the Agència de Gestió d'Ajuts Universitaris i de Recerca - AGAUR- from Generalitat de Catalunya (2016BP00161 and 2018FI B1_00213, respectively). RE-F is supported by the Spanish National Health Institute (FPI grant BES-2017-081286). CIBEREHD is funded by the Instituto de Salud Carlos III. This article is based upon work from COST Action CA17118, supported by COST (European Cooperation in Science and Technology). JML is supported by the European Commission (EC)/Horizon 2020 Program (HEPCAR, Ref. 667273-2), U.S. Department of Defense (CA150272P3), National Cancer Institute (P30-CA196521), Samuel Waxman Cancer Research Foundation, Spanish National Health Institute (SAF2016-76390) and Generalitat de Catalunya/AGAUR (SGR-1162 and SGR-1358).

Acknowledgements

The authors would like to thank Dr. Rodrigo Dienstmann from Vall d'Hebron Institute of Oncology, Barcelona, Spain for providing the CMS and clinical information for the subset of samples included in the COAD cohort from TCGA. The work was carried out at the Esther Koplowitz Centre, Barcelona.

Authors contributions

SF-E, LB and JC designed the study and analyzed the data. SF-E, LB, MV-C and MD-G designed, generated and implemented the package, and analyzed the data. EH-I and RE-F tested the software. JJL supervised the software implementation. AC, and JMLl

bioRxiv preprint first posted online Nov. 27, 2018; doi: http://dx.doi.org/10.1101/479667. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

critically reviewed the software implementation and the data output. SF-E, LB, SC-B and JC wrote the manuscript. All authors read and approved the final manuscript.

REFERENCES

Ally A, Balasundaram M, Carlsen R, Chuah E, Clarke A, Dhalla N, Holt RA, Jones SJM, Lee D, Ma Y, et al. 2017. Comprehensive and Integrative Genomic Characterization of Hepatocellular Carcinoma. *Cell* **169**: 1327–1341.e23.

Beroukhim R, Mermel CH, Porter D, Wei G, Raychaudhuri S, Donovan J, Barretina J, Boehm JS, Dobson J, Urashima M, et al. 2010. The landscape of somatic copy-number alteration across human cancers. *Nature* **463**: 899–905.

Boeva V, Popova T, Bleakley K, Chiche P, Cappo J, Schleiermacher G, Janoueix-Lerosey I, Delattre O, Barillot E. 2012. Control-FREEC: a tool for assessing copy number and allelic content using next-generation sequencing data. *Bioinformatics* **28**: 423–425.

Buccitelli C, Salgueiro L, Rowald K, Sotillo R, Mardin BR, Korbel JO. 2017. Pan-cancer analysis distinguishes transcriptional changes of aneuploidy from proliferation. *Genome Res* **27**: 501–511.

Cairncross G, Wang M, Shaw E, Jenkins R, Brachman D, Buckner J, Fink K, Souhami L, Laperriere N, Curran W, et al. 2013. Phase III trial of chemoradiotherapy for anaplastic oligodendroglioma: Long-term results of RTOG 9402. *J Clin Oncol* **31**: 337–343.

Camps J, Grade M, Nguyen QT, Hormann P, Becker S, Hummon AB, Rodriguez V, Chandrasekharappa S, Chen Y, Difilippantonio MJ, et al. 2008. Chromosomal Breakpoints in Primary Colon Cancer Cluster at Sites of Structural Variants in the Genome. *Cancer Res* **68**: 1284–1295.

Cancer T, Atlas G. 2012. Comprehensive molecular characterization of human colon and

bioRxiv preprint first posted online Nov. 27, 2018; doi: http://dx.doi.org/10.1101/479667. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

rectal cancer. *Nature* **487**: 330–7.

Carter SL, Cibulskis K, Helman E, McKenna A, Shen H, Zack T, Laird PW, Onofrio RC, Winckler W, Weir BA, et al. 2012. Absolute quantification of somatic DNA alterations in human cancer. *Nat Biotechnol* **30**: 413–421.

Casper J, Zweig AS, Villarreal C, Tyner C, Speir ML, Rosenbloom KR, Raney BJ, Lee CM, Lee BT, Karolchik D, et al. 2017. The UCSC Genome Browser database: 2018 update. *Nucleic Acids Res* **46**: D762–D769.

Chang W, Cheng J, Allaire J, Xie Y, McPherson J. 2018. shiny: Web Application Framework for R. R package version 1.1.0.

Chiang DY, Villanueva A, Hoshida Y, Peix J, Newell P, Minguez B, LeBlanc AC, Donovan DJ, Thung SN, Sole M, et al. 2008. Focal Gains of VEGFA and Molecular Classification of Hepatocellular Carcinoma. *Cancer Res* **68**: 6779–6788.

Davoli T, Uno H, Wooten EC, Elledge SJ. 2017. Tumor aneuploidy correlates with markers of immune evasion and with reduced response to immunotherapy. *Science (80-)* **355**: eaaf8399.

Grossman RL, Heath AP, Ferretti V, Varmus HE, Lowy DR, Kibbe WA, Staudt LM. 2016. Toward a Shared Vision for Cancer Genomic Data. *N Engl J Med* **375**: 1109–1112.

Guichard C, Amaddeo G, Imbeaud S, Ladeiro Y, Pelletier L, Maad I Ben, Calderaro J, Bioulac-Sage P, Letexier M, Degos F, et al. 2012. Integrated analysis of somatic mutations and focal copy-number changes identifies key genes and pathways in hepatocellular carcinoma. *Nat Genet* **44**: 694–698.

Guinney J, Dienstmann R, Wang X, de Reyniès A, Schlicker A, Soneson C, Marisa L, Roepman P, Nyamundanda G, Angelino P, et al. 2015. The consensus molecular subtypes of colorectal cancer. *Nat Med* **21**: 1350–1356.

Hoadley KA, Yau C, Hinoue T, Wolf DM, Lazar AJ, Drill E, Shen R, Taylor AM,

bioRxiv preprint first posted online Nov. 27, 2018; doi: http://dx.doi.org/10.1101/479667. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

Cherniack AD, Thorsson V, et al. 2018. Cell-of-Origin Patterns Dominate the Molecular Classification of 10,000 Tumors from 33 Types of Cancer. *Cell* **173**: 291–304.e6.

Krijgsman O, Carvalho B, Meijer GA, Steenbergen RDM, Ylstra B. 2014. Focal chromosomal copy number aberrations in cancer-Needles in a genome haystack. *Biochim Biophys Acta - Mol Cell Res* **1843**: 2698–2704.

Le DT, Uram JN, Wang H, Bartlett BR, Kemberling H, Eyring AD, Skora AD, Luber BS, Azad NS, Laheru D, et al. 2015. PD-1 Blockade in Tumors with Mismatch-Repair Deficiency. *N Engl J Med* **372**: 2509–2520.

Liaw A, Wiener M. 2002. Classification and Regression by randomForest. *R News* **2**: 18–22.

McGranahan N, Swanton C. 2017. Clonal Heterogeneity and Tumor Evolution: Past, Present, and the Future. *Cell* **168**: 613–628.

Meijer GA, Hermsen MA, Baak JP, van Diest PJ, Meuwissen SG, Belien JA, Hoovers JM, Joenje H, Snijders PJ, Walboomers JM, et al. 1998. Progression from colorectal adenoma to carcinoma is associated with non- random chromosomal gains as detected by comparative genomic hybridisation. *J Clin Pathol* **51**: 901–909.

Mermel CH, Schumacher SE, Hill B, Meyerson ML, Beroukhim R, Getz G. 2011. GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biol* **12**: R41.

Nakao K, Mehta KR, Fridlyand J, Moore DH, Jain AN, Lafuente A, Wiencke JW, Terdiman JP, Waldman FM. 2004. High-resolution analysis of DNA copy number alterations in colorectal cancer by array-based comparative genomic hybridization. *Carcinogenesis*.

Olshen AB, Venkatraman ES, Lucito R, Wigler M. 2004. Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics* **5**: 557–572.

bioRxiv preprint first posted online Nov. 27, 2018; doi: http://dx.doi.org/10.1101/479667. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

Popova T, Manié E, Stoppa-Lyonnet D, Rigaill G, Barillot E, Stern MH. 2009. Genome Alteration Print (GAP): A tool to visualize and mine complex cancer genomic profiles obtained by SNP arrays. *Genome Biol* **10**: R128.

Ried T, Hu Y, Difilippantonio MJ, Ghadimi BM, Grade M, Camps J. 2012. The consequences of chromosomal aneuploidy on the transcriptome of cancer cells. *Biochim Biophys Acta* **1819**: 784–93.

Ried T, Knutzen R, Steinbeck R, Blegen H, Schröck E, Heselmeyer K, du Manoir S, Auer G. 1996. Comparative genomic hybridization reveals a specific pattern of chromosomal gains and losses during the genesis of colorectal tumors. *Genes, Chromosom Cancer* **15**: 234–245.

Sansregret L, Vanhaesebroeck B, Swanton C. 2018. Determinants and clinical implications of chromosomal instability in cancer. *Nat Rev Clin Oncol* **15**: 139–150.

Sathirapongsasuti JF, Lee H, Horst BAJ, Brunner G, Cochran AJ, Binder S, Quackenbush J, Nelson SF. 2011. Exome sequencing-based copy-number variation and loss of heterozygosity detection: ExomeCNV. *Bioinformatics* **27**: 2648–2654.

Schulze K, Imbeaud S, Letouzé E, Alexandrov LB, Calderaro J, Rebouissou S, Couchy G, Meiller C, Shinde J, Soysouvanh F, et al. 2015. Exome sequencing of hepatocellular carcinomas identifies new mutational signatures and potential therapeutic targets. *Nat Genet* **47**: 505–511.

Stratton MR, Campbell PJ, Futreal PA, Andrew F P. 2009. The cancer genome. *Nature* **458**: 719–724.

Taylor AM, Shih J, Ha G, Gao GF, Zhang X, Berger AC, Schumacher SE, Wang C, Hu H, Liu J, et al. 2018a. Genomic and Functional Approaches to Understanding Cancer Aneuploidy. *Cancer Cell* **33**: 676–689.e3.

Taylor AM, Shih J, Ha G, Gao GF, Zhang X, Berger AC, Schumacher SE, Wang C, Hu H, Liu J, et al. 2018b. Genomic and Functional Approaches to Understanding Cancer

bioRxiv preprint first posted online Nov. 27, 2018; doi: http://dx.doi.org/10.1101/479667. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

Aneuploidy. *Cancer Cell* **33**: 676–689.e3.

Totoki Y, Tatsuno K, Covington KR, Ueda H, Creighton CJ, Kato M, Tsuji S, Donehower LA, Slagle BL, Nakamura H, et al. 2014. Trans-ancestry mutational landscape of hepatocellular carcinoma genomes. *Nat Genet* **46**: 1267–1273.

Trautmann K, Terdiman JP, French AJ, Roydasgupta R, Sein N, Kakar S, Fridlyand J, Snijders AM, Albertson DG, Thibodeau SN, et al. 2006. Chromosomal instability in microsatellite-unstable and stable colon cancer. *Clin Cancer Res* **12**: 6379–6385.

van de Wie MA, Kim KI, Vosse SJ, van Wieringen WN, Wilting SM, Ylstra B. 2007. CGHcall: Calling aberrations for array CGH tumor profiles. *Bioinformatics* **23**: 892–894.

Van Loo P, Nordgard SH, Lingjaerde OC, Russnes HG, Rye IH, Sun W, Weigman VJ, Marynen P, Zetterberg A, Naume B, et al. 2010. Allele-specific copy number analysis of tumors. *Proc Natl Acad Sci* **107**: 16910–16915.

Wang K, Lim HY, Shi S, Lee J, Deng S, Xie T, Zhu Z, Wang Y, Pocalyko D, Yang WJ, et al. 2013. Genomic landscape of copy number aberrations enables the identification of oncogenic drivers in hepatocellular carcinoma. *Hepatology* **58**: 706–717.

Zack TTI, Schumacher SES, Carter SLS, Cherniack AD, Saksena G, Tabak B, Lawrence MS, Zhang C-Z, Wala J, Mermel CH, et al. 2013. Pan-cancer patterns of somatic copy number alteration. *Nat Genet* **45**: 1134–1140.

Zhang Z, Hao K. 2015. SAAS-CNV: A Joint Segmentation Approach on Aggregated and Allele Specific Signals for the Identification of Somatic Copy Number Alterations with Next-Generation Sequencing Data ed. E. Wang. *PLoS Comput Biol* **11**: e1004618.

bioRxiv preprint first posted online Nov. 27, 2018; doi: http://dx.doi.org/10.1101/479667. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

FIGURE LEGENDS

Figure 1: CNApp workflow. The diagram depicts the overall processes performed by CNApp and indicates the output for each section.

Figure 2: Analysis of the TCGA pan-cancer dataset and clustering by tumor type. CNApp outputs to characterize pan-cancer 10,635 samples including 33 TCGA cancer types. **A)** Broad, Focal and Global CNA scores (BCS, FCS and GCS, respectively) distribution across the 33 cancer types. **B)** Genome-wide chromosome arm CNA profile heatmap for 10,635 samples considering broad and focal events. Annotation tracks for FCS, BCS and GCS are presented. **C)** Arm regions frequencies as percentages relative to the TCGA pan-cancer dataset (red for gains and blue for losses). **D)** Heatmap plot showing 20 out of the 33 TCGA cancer type profile correlations, by Pearson's method, hierarchically clustered by tumor type. Gastrointestinal, gynecological and squamous types are clustering consistently in their respective groups.

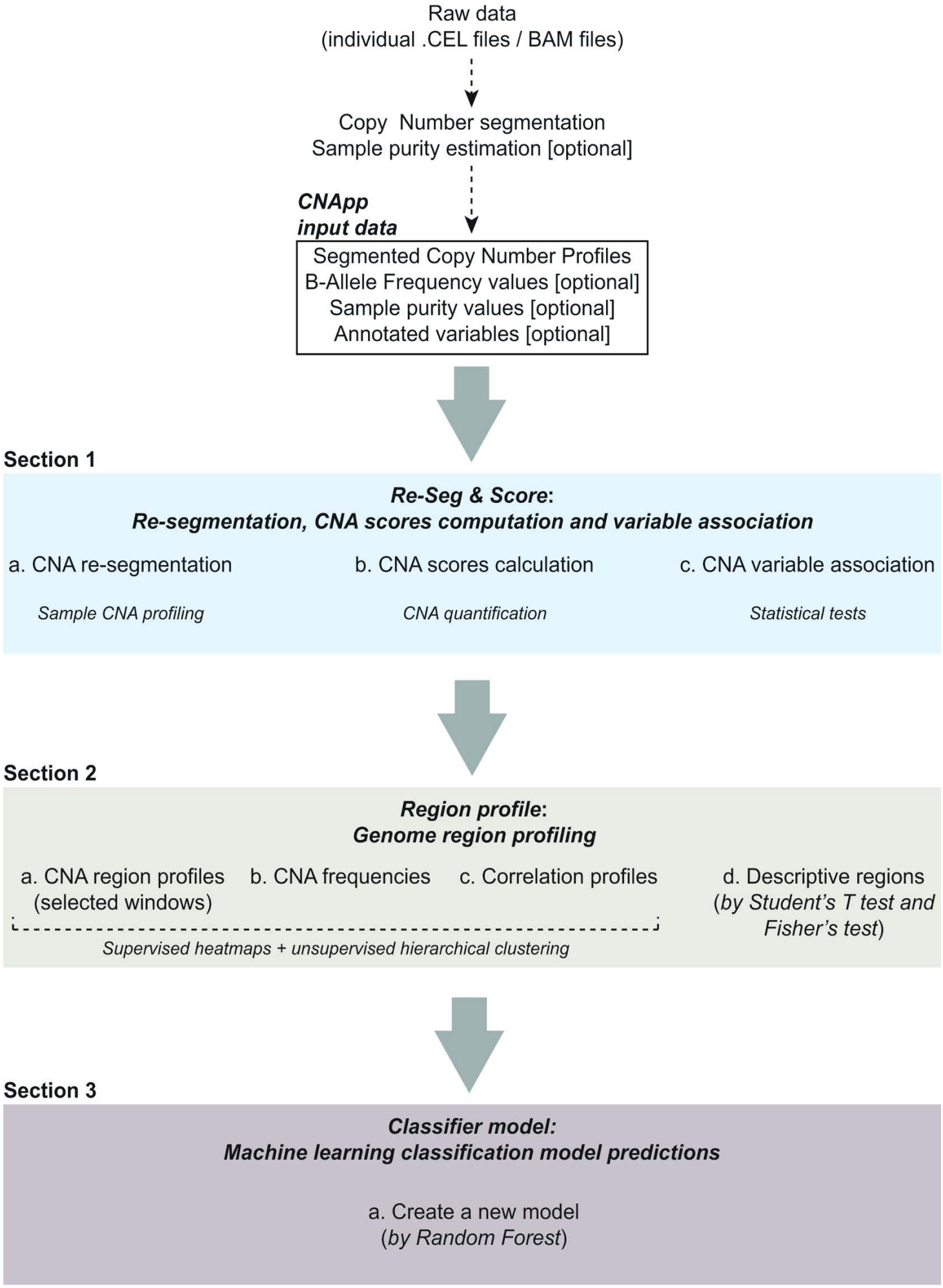
Figure 3: Identification of recurrent broad and focal CNAs. Calculation of broad and focal CNA frequencies using several parameters in CNApp in order to describe the genomic landscape of LIHC. **A)** CNApp frequencies for chromosome arm regions using default cutoffs, corresponding to 2.3/1.7 copies for gains and losses, respectively. **B)** CNApp frequencies for chromosome arm regions relaxing cutoffs to make them equivalent to those of GISTIC2.0. **C)** CNApp frequencies of focal events using default thresholds and sub-cytobands genomic regions. **D)** Frequencies of focal events from moderate- to high-amplitude levels using sub-cytobands genomic regions.

Figure 4: Genomic characterization of colon cancer according to the CMS

bioRxiv preprint first posted online Nov. 27, 2018; doi: http://dx.doi.org/10.1101/479667. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

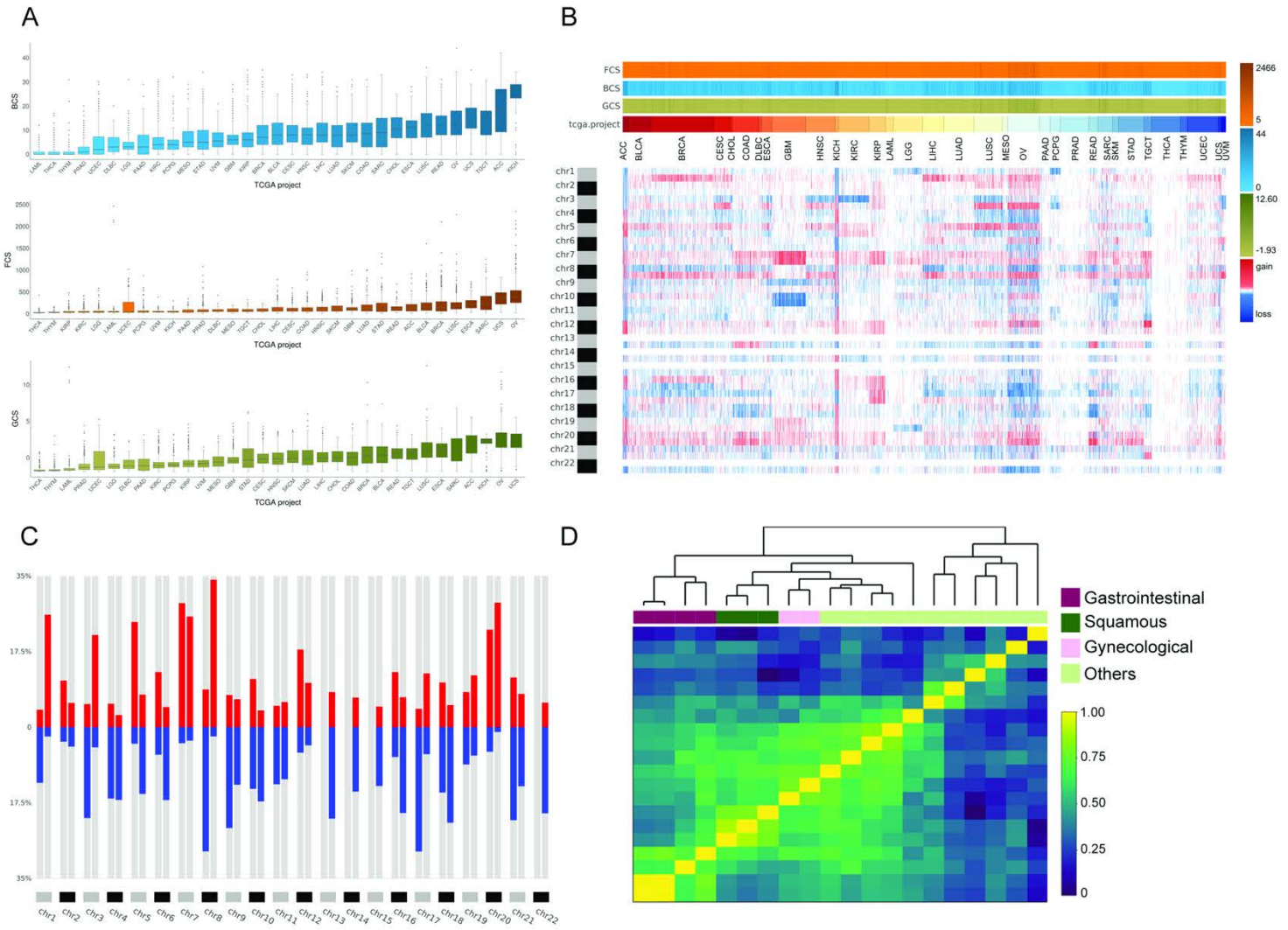
classification. A) Arm-region frequencies of 309 colon cancer samples using CNApp default thresholds for CNAs. **B)** BCS distribution by CMS sample groups. Significance is shown as p-value ≤ 0.001 (***); p-value ≤ 0.01 (**); p-value ≤ 0.05 (*); p-value > 0.05 (ns). **C)** Number of gained and lost chromosome arms for each sample distributed according to the BCS values. Note that a cutoff at 4 is indicated with a black line. Annotation tracks for microsatellite instability (msi), *BRAF* mutated samples (braf_mut), CMS groups (cms_label), FCS and BCS are displayed. **D)** Genome-wide profiling by chromosome arms distributed according to the CMS group. Annotation tracks for microsatellite instability (msi), *BRAF* mutated samples (braf_mut), CMS groups (cms_label), FCS and BCS are displayed. Sample-to-sample correlation heatmap plot by Pearson's method is shown below. **E)** Distribution of CNA values affecting 20q according to the CMS groups. Significance is shown as p-value ≤ 0.001 (***); p-value ≤ 0.01 (**); p-value ≤ 0.05 (*); p-value > 0.05 (ns).

Figure 1


Raw data
(individual .CEL files / BAM files)
Copy Number segmentation
Sample purity estimation [optional]
CNApp
input data
Segmented Copy Number Profiles
B-Allele Frequency values [optional]
Sample purity values [optional]
Annotated variables [optional]
Section 1
Re-Seg & Score:
Re-segmentation, CNA scores computation and variable association
a. CNA re-segmentation
b. CNA scores calculation
c. CNA variable association
Sample CNA profiling
CNA quantification
Statistical tests
Section 2
Region profile:
Genome region profiling
a. CNA region profiles (selected windows)
b. CNA frequencies
c. Correlation profiles
d. Descriptive regions (by Student's T test and Fisher's test)
Supervised heatmaps + unsupervised hierarchical clustering
Section 3
Classifier model:
Machine learning classification model predictions
a. Create a new model
(by Random Forest)

bioRxiv preprint first posted online Nov. 27, 2018; doi: http://dx.doi.org/10.1101/479667. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

Figure 2

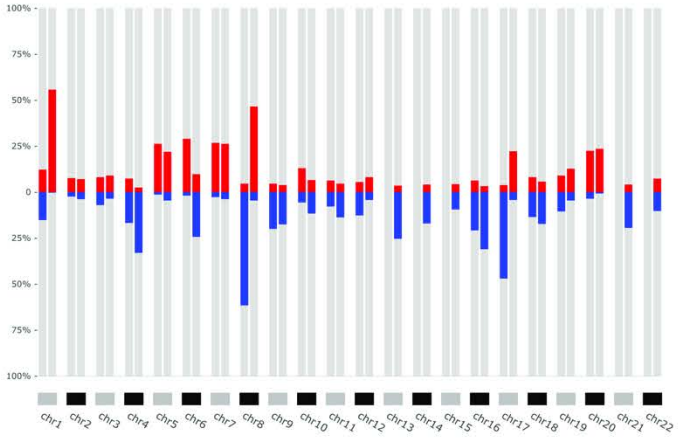


bioRxiv preprint first posted online Nov. 27, 2018; doi: http://dx.doi.org/10.1101/479667. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

Figure 3

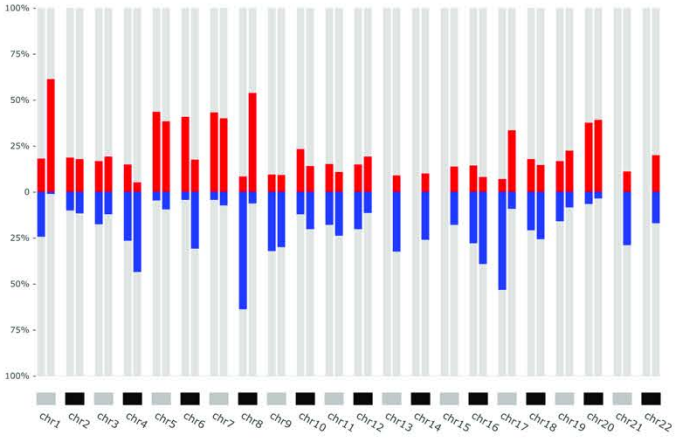
A

CNApp CNA default thresholds: ≥ 2.3 copies / ≤ 1.7 copies



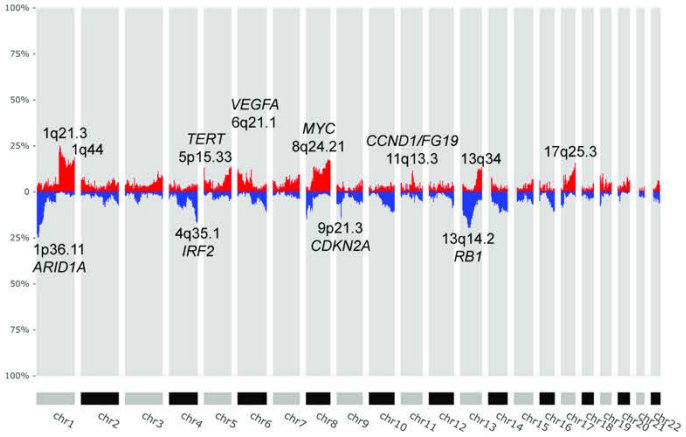
B

CNApp CNA relaxed thresholds: ≥ 2.1 copies / ≤ 1.8 copies



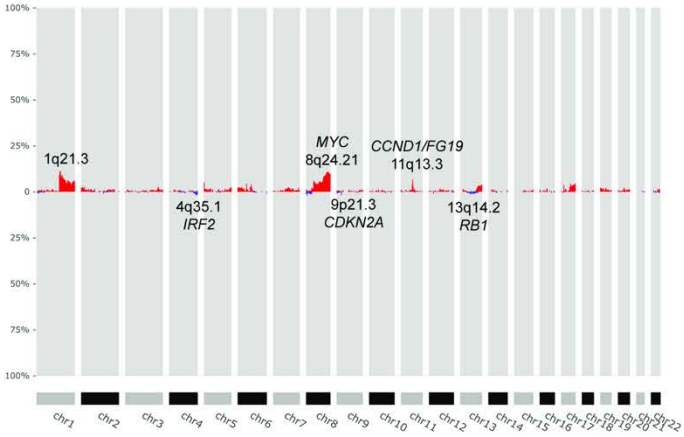
C

CNApp CNA default thresholds: ≥ 2.3 copies / ≤ 1.7 copies



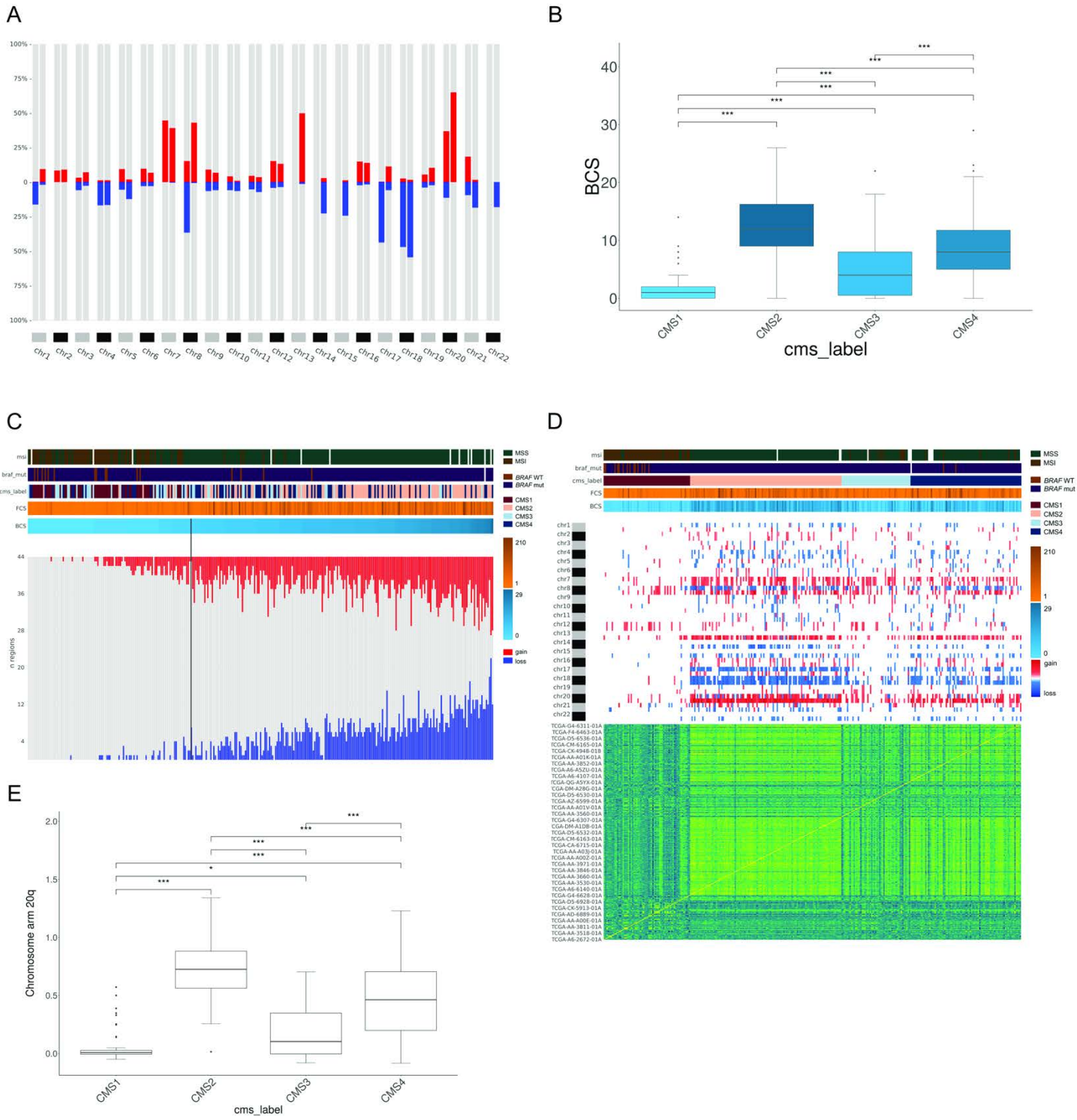
D

CNApp CNA moderate-amplitude thresholds: ≥ 3 copies / ≤ 1 copy



bioRxiv preprint first posted online Nov. 27, 2018; doi: http://dx.doi.org/10.1101/479667. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

Figure 4



Altres articles de participació

Whole-exome sequencing identifies rare pathogenic variants in new predisposition genes for familial colorectal cancer

Genetics in medicine.
17 - 2, pp. 131 - 142. 2015.

© American College of Medical Genetics and Genomics

ORIGINAL RESEARCH ARTICLE | Genetics inMedicine

Open

Whole-exome sequencing identifies rare pathogenic variants in new predisposition genes for familial colorectal cancer

Clara Esteban-Jurado, MSc[1], Maria Vila-Casadesús, MSc[2], Pilar Garre, MD, PhD[3], Juan José Lozano, PhD[2], Anna Pristoupilova, MSc[4,5], Sergi Beltran, PhD[4], Jenifer Muñoz, MSc[1], Teresa Ocaña, MSc[1], Francesc Balaguer, MD, PhD[1], Maria López-Cerón, MD, PhD[1], Miriam Cuatrecasas, MD, PhD[6], Sebastià Franch-Expósito, MSc[1], Josep M. Piqué, MD, PhD[1], Antoni Castells, MD, PhD[1], Angel Carracedo, MD, PhD[7], Clara Ruiz-Ponte, PhD[7], Anna Abulí, PhD[8], Xavier Bessa, MD, PhD[8], Montserrat Andreu, MD, PhD[8], the EPICOLON Consortium, Luis Bujanda, MD, PhD[9], Trinidad Caldés, PhD[3] and Sergi Castellví-Bel, PhD[1]

Purpose: Colorectal cancer is an important cause of mortality in the developed world. Hereditary forms are due to germ-line mutations in *APC*, *MUTYH*, and the mismatch repair genes, but many cases present familial aggregation but an unknown inherited cause. The hypothesis of rare high-penetrance mutations in new genes is a likely explanation for the underlying predisposition in some of these familial cases.

Methods: Exome sequencing was performed in 43 patients with colorectal cancer from 29 families with strong disease aggregation without mutations in known hereditary colorectal cancer genes. Data analysis selected only very rare variants (0–0.1%), producing a putative loss of function and located in genes with a role compatible with cancer. Variants in genes previously involved in hereditary colorectal cancer or nearby previous colorectal cancer genome-wide association study hits were also chosen.

Results: Twenty-eight final candidate variants were selected and validated by Sanger sequencing. Correct family segregation and somatic studies were used to categorize the most interesting variants in *CDKN1B*, *XRCC4*, *EPHX1*, *NFKBIZ*, *SMARCA4*, and *BARD1*.

Conclusion: We identified new potential colorectal cancer predisposition variants in genes that have a role in cancer predisposition and are involved in DNA repair and the cell cycle, which supports their putative involvement in germ-line predisposition to this neoplasm.

Genet Med advance online publication 24 July 2014

Key Words: colorectal neoplasm; genetic variant; genetic predisposition to disease; hereditary disease; next-generation sequencing

The Fanconi anemia DNA damage repair pathway in the spotlight for germline predisposition to colorectal cancer

European journal of human genetics.
24 - 10, pp. 1501 - 1505. 2016.

EJHG Open

European Journal of Human Genetics (2016) 24, 1501–1505
© 2016 Macmillan Publishers Limited, part of Springer Nature. All rights reserved 1018-4813/16
www.nature.com/ejhg

SHORT REPORT

The Fanconi anemia DNA damage repair pathway in the spotlight for germline predisposition to colorectal cancer

Clara Esteban-Jurado[1], Sebastià Franch-Expósito[1], Jenifer Muñoz[1], Teresa Ocaña[1], Sabela Carballal[1], Maria López-Cerón[1], Miriam Cuatrecasas[2], Maria Vila-Casadesús[3], Juan José Lozano[3], Enric Serra[4], Sergi Beltran[4], The EPICOLON Consortium, Alejandro Brea-Fernández[5], Clara Ruiz-Ponte[5], Antoni Castells[1], Luis Bujanda[6], Pilar Garre[7], Trinidad Caldés[7], Joaquín Cubiella[8], Francesc Balaguer[1] and Sergi Castellví-Bel*[,1]

Colorectal cancer (CRC) is one of the most common neoplasms in the world. Fanconi anemia (FA) is a very rare genetic disease causing bone marrow failure, congenital growth abnormalities and cancer predisposition. The comprehensive FA DNA damage repair pathway requires the collaboration of 53 proteins and it is necessary to restore genome integrity by efficiently repairing damaged DNA. A link between FA genes in breast and ovarian cancer germline predisposition has been previously suggested. We selected 74 CRC patients from 40 unrelated Spanish families with strong CRC aggregation compatible with an autosomal dominant pattern of inheritance and without mutations in known hereditary CRC genes and performed germline DNA whole-exome sequencing with the aim of finding new candidate germline predisposition variants. After sequencing and data analysis, variant prioritization selected only those very rare alterations, producing a putative loss of function and located in genes with a role compatible with cancer. We detected an enrichment for variants in FA DNA damage repair pathway genes in our familial CRC cohort as 6 families carried heterozygous, rare, potentially pathogenic variants located in *BRCA2*/*FANCD1*, *BRIP1*/*FANCJ*, *FANCC*, *FANCE* and *REV3L*/*POLZ*. In conclusion, the FA DNA damage repair pathway may play an important role in the inherited predisposition to CRC.
European Journal of Human Genetics (2016) **24,** 1501–1505; doi:10.1038/ejhg.2016.44; published online 11 May 2016

POLE and *POLD1* screening in 155 patients with multiple polyps and early-onset colorectal cancer

Oncotarget.
8 - 16, pp. 26732 - 26743. 2017.

www.impactjournals.com/oncotarget/ Oncotarget, 2017, Vol. 8, (No. 16), pp: 26732-26743

Research Paper

POLE and *POLD1* screening in 155 patients with multiple polyps and early-onset colorectal cancer

Clara Esteban-Jurado[1], David Giménez-Zaragoza[2], Jenifer Muñoz[1], Sebastià Franch-Expósito[1], Miriam Álvarez-Barona[3], Teresa Ocaña[1], Miriam Cuatrecasas[4], Sabela Carballal[1], María López-Cerón[1], Maria Marti-Solano[5], Marcos Díaz-Gay[1], Tom van Wezel[6], Antoni Castells[1], Luis Bujanda[7], Judith Balmaña[8], Victoria Gonzalo[9], Gemma Llort[10], Clara Ruiz-Ponte[3], Joaquín Cubiella[11], Francesc Balaguer[1], Rosa Aligué[2], Sergi Castellví-Bel[1]

ABSTRACT

Germline mutations in *POLE* and *POLD1* have been shown to cause predisposition to colorectal multiple polyposis and a wide range of neoplasms, early-onset colorectal cancer being the most prevalent. In order to find additional mutations affecting the proofreading activity of these polymerases, we sequenced its exonuclease domain in 155 patients with multiple polyps or an early-onset colorectal cancer phenotype without alterations in the known hereditary colorectal cancer genes. Interestingly, none of the previously reported mutations in *POLE* and *POLD1* were found. On the other hand, among the genetic variants detected, only two of them stood out as putative pathogenic in the *POLE* gene, c.1359 + 46del71 and c.1420G > A (p.Val474Ile). The first variant, detected in two families, was not proven to alter correct RNA splicing. Contrarily, c.1420G > A (p.Val474Ile) was detected in one early-onset colorectal cancer patient and located right next to the exonuclease domain. The pathogenicity of this change was suggested by its rarity and bioinformatics predictions, and it was further indicated by functional assays in *Schizosaccharomyces pombe*. This is the first study to functionally analyze a *POLE* genetic variant outside the exonuclease domain and widens the spectrum of genetic changes in this DNA polymerase that could lead to colorectal cancer predisposition.

GeMSTONE: orchestrated prioritization of human germline mutations in the cloud

Nucleic Acids Research.
Volume 45 - Issue W1, pp. W207 - W214. Oxford Academic, 2017.

Nucleic Acids Research, 2017 **1**
doi: 10.1093/nar/gkx398

GeMSTONE: orchestrated prioritization of human germline mutations in the cloud

Siwei Chen[1,2,3,†], Juan F. Beltrán[1,2,†], Clara Esteban-Jurado[4], Sebastià Franch-Expósito[4], Sergi Castellví-Bel[4], Steven Lipkin[5], Xiaomu Wei[2,5,*] and Haiyuan Yu[1,2,*]

[1]Department of Biological Statistics and Computational Biology, Cornell University, Ithaca, NY 14853, USA, [2]Weill Institute for Cell and Molecular Biology, Cornell University, Ithaca, NY 14853, USA, [3]Department of Molecular Biology and Genetics, Cornell University, Ithaca, NY 14853, USA, [4]Gastroenterology Department, Hospital Clínic, Institut d'Investigacions Biomèdiques August Pi i Sunyer (IDIBAPS), Centro de Investigación Biomédica en Red de Enfermedades Hepáticas y Digestivas (CIBEREHD), University of Barcelona, 08036 Barcelona, Catalonia, Spain and [5]Department of Medicine, Weill Cornell College of Medicine, NY 10021, USA

Received February 18, 2017; Revised April 18, 2017; Editorial Decision April 26, 2017; Accepted April 28, 2017

ABSTRACT

Integrative analysis of whole-genome/exome-sequencing data has been challenging, especially for the non-programming research community, as it requires simultaneously managing a large number of computational tools. Even computational biologists find it unexpectedly difficult to reproduce results from others or optimize their strategies in an end-to-end workflow. We introduce Germline Mutation Scoring Tool fOr Next-generation sEquencing data (GeMSTONE), a cloud-based variant prioritization tool with high-level customization and a comprehensive collection of bioinformatics tools and data libraries (http://gemstone.yulab.org/). GeMSTONE generates and readily accepts a shareable 'recipe' file for each run to either replicate previous results or analyze new data with identical parameters and provides a centralized workflow for prioritizing germline mutations in human disease within a streamlined workflow rather than a pool of program executions.

to organize, maintain and standardize the variant analysis workflows, increasing the time and monetary investment for less computationally oriented biologists and labs. Some integrative frameworks (4–7) have been developed to enhance the reproducibility and accessibility of NGS studies. This same initiative inspired the framework for GeMSTONE: recording all analysis metadata for reproducible computational experiments, specifically focusing on germline mutation prioritization in human disease.

Although other platforms bring together different bioinformatics tools and allow users to schedule their analyzes online, none of them are built with an emphasis on streamlined single-run scheduling and automatic fetching of the necessary supplementary public data. Platforms like Galaxy (4), for instance, allow the user to combine many different tools from an impressive catalog, but require the user to reformat their data depending on the particular input format of the database or tool that they want to add to their analysis. A major design goal in the development of GeMSTONE is the ability to maximize customization for studies in a streamlined workflow rather than a pool of program executions. Within the GeMSTONE interface, databases required by the user-selected tools are pre-loaded and the user-input data will be automatically reformatted

Zeb1 in stromal myofibroblasts promotes *Kras*-driven development of pancreatic cancer

Cancer Research.
78(10); 2624–37. AACR, 2018.

Tumor Biology and Immunology

Cancer Research

Zeb1 in Stromal Myofibroblasts Promotes *Kras*-Driven Development of Pancreatic Cancer

Irene Sangrador[1], Xavier Molero[2], Fiona Campbell[3], Sebastià Franch-Expósito[4], Maria Rovira-Rigau[5], Esther Samper[1], Manuel Domínguez-Fraile[1], Cristina Fillat[5], Antoni Castells[1,6], and Eva C. Vaquero[1,6]

Abstract

The transcription factor Zeb1 has been identified as a crucial player in Kras-dependent oncogenesis. In pancreatic ductal adenocarcinoma (PDAC), Zeb1 is highly expressed in myofibroblasts and correlates with poor prognosis. As Kras mutations are key drivers in PDAC, we aimed here to assess the necessity of Zeb1 for Kras-driven PDAC and to define the role of Zeb1-expressing myofibroblasts in PDAC development. Genetically engineered mice with conditional pancreatic *Kras*^{G12D} and *Trp53* mutations (KPC) were crossed with *Zeb1* haploinsufficient mice ($Z^{+/-}$). Extensive PDAC was prominent in all 20-week-old KPC;$Z^{+/+}$ mice, whereas only low-grade precursor lesions were detected in age-matched KPC;$Z^{+/-}$ littermates, with PDAC developing eventually in KPC;$Z^{+/-}$ aged animals. Zeb1 expression in myofibroblasts occurred early in tumorigenesis and *Zeb1* haploinsufficiency retarded native expansion of stromal myofibroblasts during precursor-to-cancer progression. *Zeb1* downregulation in mPSC repressed their activated gene profile, impaired their migratory and proliferative activity, and attenuated their tumor-supporting features. Conditioned media from $Z^{+/+}$ mouse-activated (myofibroblast-like) pancreatic stellate cells (mPSC) boosted Ras activity in pancreatic cancer cells carrying mutant *Kras*; this effect was not observed when using conditioned media from $Z^{+/-}$ mPSC, revealing a paracrinal cooperative axis between Zeb1-expressing PSC and oncogenic Kras-bearing tumor cells. We conclude that Zeb1-expressing stromal myofibroblasts enable a heterotypic collaboration with the Kras-fated epithelial compartment, thus supporting pancreatic malignancy.

Significance: Zeb1 expression in stromal myofibroblasts supports PDAC development via collaboration with the epithelial compartment bearing oncogenic Kras mutations. *Cancer Res; 78(10); 2624–37. ©2018 AACR.*

Mutational Signatures in Cancer (MuSiCa): a web application to implement mutational signatures analysis in cancer samples

BMC Bioinformatics
2018; 19:224.

https://doi.org/10.1186/s12859-018-2234-y

BMC Bioinformatics

SOFTWARE Open Access

Mutational Signatures in Cancer (MuSiCa): a web application to implement mutational signatures analysis in cancer samples

Marcos Díaz-Gay[1†], Maria Vila-Casadesús[2,3†], Sebastià Franch-Expósito[1†], Eva Hernández-Illán[1], Juan José Lozano[2] and Sergi Castellví-Bel[1*]

Abstract

Background: Mutational signatures have been proved as a valuable pattern in somatic genomics, mainly regarding cancer, with a potential application as a biomarker in clinical practice. Up to now, several bioinformatic packages to address this topic have been developed in different languages/platforms. MutationalPatterns has arisen as the most efficient tool for the comparison with the signatures currently reported in the Catalogue of Somatic Mutations in Cancer (COSMIC) database. However, the analysis of mutational signatures is nowadays restricted to a small community of bioinformatic experts.

Results: In this work we present Mutational Signatures in Cancer (MuSiCa), a new web tool based on MutationalPatterns and built using the Shiny framework in R language. By means of a simple interface suited to non-specialized researchers, it provides a comprehensive analysis of the somatic mutational status of the supplied cancer samples. It permits characterizing the profile and burden of mutations, as well as quantifying COSMIC-reported mutational signatures. It also allows classifying samples according to the above signature contributions.

Conclusions: MuSiCa is a helpful web application to characterize mutational signatures in cancer samples. It is accessible online at http://bioinfo.ciberehd.org/GPtoCRC/en/tools.html and source code is freely available at https://github.com/marcos-diazg/musica.

Keywords: Mutational signatures, COSMIC database, Single nucleotide variants, Cancer genomics, Web tool, Shiny, R language

Approaches to functionally validate candidate genetic variants involved in colorectal cancer predisposition

Molecular Aspects of Medicine.
Elsevier, 2019.

Contents lists available at ScienceDirect

Molecular Aspects of Medicine

journal homepage: www.elsevier.com/locate/mam

Approaches to functionally validate candidate genetic variants involved in colorectal cancer predisposition

Laia Bonjoch[a], Pilar Mur[b,c], Coral Arnau-Collell[a], Gardenia Vargas-Parra[b,c], Bahar Shamloo[d], Sebastià Franch-Expósito[a], Marta Pineda[b,c], Gabriel Capellà[b,c], Batu Erman[e], Sergi Castellví-Bel[a,*]

[a] *Gastroenterology Department, Hospital Clínic, Institut d'Investigacions Biomèdiques August Pi i Sunyer (IDIBAPS), Centro de Investigación Biomédica en Red de Enfermedades Hepáticas y Digestivas (CIBEREHD), University of Barcelona, Barcelona, Spain*
[b] *Hereditary Cancer Program, Catalan Institute of Oncology, Institut d'Investigació Biomèdica de Bellvitge (IDIBELL), ONCOBELL Program, L'Hospitalet de Llobregat, Barcelona, Spain*
[c] *Centro de Investigación Biomédica en Red de Cáncer (CIBERONC), Spain*
[d] *Molecular Biology, Genetics, and Bioengineering Department, Legacy Research Institute, Portland, OR, USA*
[e] *Molecular Biology, Genetics and Bioengineering Program, Faculty of Engineering and Natural Sciences, Sabanci University, Istanbul, Turkey*

ARTICLE INFO

Keywords:
Colorectal cancer
Genetic variant
Functional genomics
Disease model
CRISPR
Organoid

ABSTRACT

Most next generation sequencing (NGS) studies identified candidate genetic variants predisposing to colorectal cancer (CRC) but do not tackle its functional interpretation to unequivocally recognize a new hereditary CRC gene. Besides, germline variants in already established hereditary CRC-predisposing genes or somatic variants share the same need when trying to categorize those with relevant significance. Functional genomics approaches have an important role in identifying the causal links between genetic architecture and phenotypes, in order to decipher cellular function in health and disease. Therefore, functional interpretation of identified genetic variants by NGS platforms is now essential. Available approaches nowadays include bioinformatics, cell and molecular biology and animal models. Recent advances, such as the CRISPR-Cas9, ZFN and TALEN systems, have been already used as a powerful tool with this objective. However, the use of cell lines is of limited value due to the CRC heterogeneity and its close interaction with microenvironment. Access to tridimensional cultures or organoids and xenograft models that mimic the in vivo tissue architecture could revolutionize functional analysis. This review will focus on the application of state-of-the-art functional studies to better tackle new genes involved in germline predisposition to this neoplasm.

Integrated Analysis of Germline and Tumor DNA Identifies New Candidate Genes Involved in Familial Colorectal Cancer

Cancers.
2019; 11(3), 362.

Article

Integrated Analysis of Germline and Tumor DNA Identifies New Candidate Genes Involved in Familial Colorectal Cancer

Marcos Díaz-Gay [1], Sebastià Franch-Expósito [1], Coral Arnau-Collell [1], Solip Park [2], Fran Supek [3], Jenifer Muñoz [1], Laia Bonjoch [1], Anna Gratacós-Mulleras [1], Paula A. Sánchez-Rojas [1], Clara Esteban-Jurado [1], Teresa Ocaña [1], Miriam Cuatrecasas [4], Maria Vila-Casadesús [5], Juan José Lozano [5], Genis Parra [6], Steve Laurie [6], Sergi Beltran [6], EPICOLON Consortium [1,7,8], Antoni Castells [1], Luis Bujanda [7], Joaquín Cubiella [8], Francesc Balaguer [1] and Sergi Castellví-Bel [1,*]

Received: 25 January 2019; Accepted: 9 March 2019; Published: 13 March 2019

Abstract: Colorectal cancer (CRC) shows aggregation in some families but no alterations in the known hereditary CRC genes. We aimed to identify new candidate genes which are potentially involved in germline predisposition to familial CRC. An integrated analysis of germline and tumor whole-exome sequencing data was performed in 18 unrelated CRC families. Deleterious single nucleotide variants (SNV), short insertions and deletions (indels), copy number variants (CNVs) and loss of heterozygosity (LOH) were assessed as candidates for first germline or second somatic hits. Candidate tumor suppressor genes were selected when alterations were detected in both germline and somatic DNA, fulfilling Knudson's two-hit hypothesis. Somatic mutational profiling and signature analysis were also performed. A series of germline-somatic variant pairs were detected. In all cases, the first hit was presented as a rare SNV/indel, whereas the second hit was either a different SNV (3 genes) or LOH affecting the same gene (141 genes). *BRCA2*, *BLM*, *ERCC2*, *RECQL*, *REV3L* and *RIF1* were among the most promising candidate genes for germline CRC predisposition. The identification of new candidate genes involved in familial CRC could be achieved by our integrated analysis. Further functional studies and replication in additional cohorts are required to confirm the selected candidates.

UNIVERSITAT DE
BARCELONA