DOI: 10.2436/20.1501.02.77

Cent cinquanta anys després de *L'origen de les espècies*, de Darwin
(Arcadi Navarro i Carmen Segarra, ed.)

Treballs de la SCB. Vol. 60 (2009) 57-80

ESTRUCTURA DEL GENOMA

Tomàs Marquès-Bonet [1] i Lluís Armengol [2]

[1] *Department of Genome Sciences, Universitat de Washington.*
[2] *Genes and Disease Program, Centre de Regulació Genòmica i Quantitative Genomic Medicine Laboratories (qGenomics).*

Adreces per a la correspondència. Tomàs Marquès-Bonet. Department of Genome Sciences, Universitat de Washington. Foege Building S-250, Box 355065. 1705 NE Pacific St. Seattle WA 98195-5065 (EUA). Adreça electrònica: *tmarques@u.washington.edu*.
Lluís Armengol i Dulcet. Quantitative Genomic Medicine Laboratories. Parc de Recerca Biomèdica de Barcelona, Planta 4. Despatx 423. Dr. Aiguader, 88. 08003 Barcelona. Adreça electrònica: *lluis.armengol@qgenomics.com*.

RESUM

El genoma és una estructura altament dinàmica amb una certa tendència a la inestabilitat, i està, per tant, subjecte a l'escrutini de la selecció natural. En el camí d'entendre el genoma hem pogut observar el paper clau que tenen les repeticions (de tota mena) per comprendre l'evolució estructural del genoma humà i com es relaciona l'estructura i la funció. Així, recentment, hem pogut apreciar que, a part de les variacions clàssiques i els SNP (polimorfismes d'un sol nucleòtid), els mamífers (com a mínim els ratolins i els humans) tenim una estructura dels nostres genomes altament variable. L'estudi de les regions variants en nombre de còpia (*structural variants* o *copy number polymorphism*) ens ha permès observar que els canvis en l'estructura tenen repercussió en l'expressió dels gens, que es tradueixen tant en variabilitat fenotípica entre individus com, en casos més extrems, en malalties. En aquest capítol donarem una visió sobre l'estructura i el dinamisme del genoma, centrant-nos en aspectes evolutius del polimorfisme humà i la malaltia.

Paraules clau: genoma, repeticions comunes, variants estructurals (SV), duplicacions segmentàries (DS), variants de nombre de còpia (CNV).

GENOME STRUCTURE

SUMMARY

Every mammalian genome is an unstable and highly dynamic structure and therefore subjected to the strict evaluation of natural selection. In the last decades, we have changed

our vocabulary and we have been using *genomics* instead of *genetics*, mainly because of the revolution on our research techniques. In the way of the understanding of the genomes, we realized how important repeats are and their role in the evolution of the structure of the human genome and how this structure and different functions are related. We have seen for instance that besides of the variation caused for SNPs (Single Nucleotide Polymorphism), the mammals (and specifically humans) are highly structural variant. The study of SVs (Structural Variant regions) and CNPs (Copy Number Polymorphisms) has allowed us to see that changes in structure have important consequences, from genomic diseases to simply affecting genes and its expression and hence, being responsible of the huge phenotypical variability observed among individuals. In this chapter, we will give our view on the structure and dynamism of the genome, from an evolutionary point of view, covering human polymorphism and diseases.

Key words: genome, common repeats, structural variants (SVs), segmental duplications (SDs), copy number polymorphism (CNVs).

INTRODUCCIÓ

Charles Darwin desconeixia les paraules *genoma*, *gen* o *cromosoma*. Nosaltres, per contra, estem familiaritzats des de fa gairebé dos segles amb el terme *genètica*, entesa com la ciència de l'herència o la branca de la biologia que es dedica a l'estudi de l'herència, tot fent èmfasi en l'estructura i funció dels gens i en com aquests es transmeten d'una generació a la següent. D'alguna o altra manera hem estat utilitzant la genètica des que els nostres avantpassats van abandonar la caça com a base del seu *modus vivendi* i van començar a socialitzar-se i a escollir les variants de plantes més productives i els animals més manyacs per fer-los companyia. Tot i així, no vam ser-ne plenament conscients com a científics fins que, a principis del segle XIX i després d'anys d'oblit, es van popularitzar els treballs sobre l'herència de la planta del pèsol, del monjo austríac Gregor Mendel. El seu coetani, Charles Darwin, amb la seva teoria sobre l'origen de les espècies, també fonamentava llavors una de les teories més acceptades en la biologia moderna, encara que sense el coneixement de la genètica (fusió que arribaria unes dècades més tard en el que es va anomenar la *teoria sintètica de l'evolució*). Sense ser-ne gaire conscients, ambdós van donar un nou sentit al món en què vivien en establir les bases de l'herència i el motor de l'evolució. Malgrat tot, aquests experiments només van representar les primeríssimes passes conscients envers un coneixement que s'ha desenvolupat de manera vertiginosa en les dues darreres dècades i que, de la mà de la revolució tecnològica, suposa una empenta sense precedents per a l'avenç no solament de la biologia, sinó també de camps tan aparentment allunyats com l'enginyeria, l'economia o la medicina.

PERSPECTIVA HISTÒRICA I DESENVOLUPAMENT TECNOLÒGIC

La rellevància que ha tingut i té la genètica per al coneixement de la humanitat i la seva vida pràctica s'ha vist traduïda en molts reconeixements en forma de premis Nobel a investigadors que han contribuït de manera decisiva a l'avenç d'aquest camp. El desenvolupament de la genètica ha estat estretament lligat a l'avenç d'altres branques del coneixement, essencialment la física i

la química en els seus inicis, de les quals els científics van prendre coneixements per resoldre els problemes biològics que se'ls plantejaven. Després d'una etapa d'evolució «pròpia», en què la biologia i la genètica molecular van evolucionar *per se*, fa algun temps que la gran quantitat de dades generades han obligat els genètics a valdre's d'altres branques del coneixement per avançar. Som en l'era de la bioinformàtica...

Van haver de passar més de vint-i-cinc anys des de la primera observació documentada al microscopi d'unes estructures que es tenyien amb colorants durant la mitosi, que en Walter Fleming va anomenar *cromosomes* (1882), fins que es va emprar el terme *gen* per referir-se a les unitats d'herència descrites per Mendel. Utilitzant mosques del vinagre com a organisme model, Thomas Hunt Morgan i els seus estudiants van demostrar que els gens, alineats en els cromosomes, són les unitats bàsiques de l'herència, van descobrir el lligament genètic i van descriure el procés de recombinació cromosòmica (Morgan, 1915) principis essencials per al posterior desenvolupament de la genètica mèdica. El 1944, Avery, McLeod i McCarty van demostrar que una molècula d'alt pes molecular, rica en àcids nucleics, que no era ni l'RNA, ni les proteïnes ni els lípids, era el principi transformant, capaç de convertir les soques R no virulentes de *Streptococcus pneumoniae* en soques S virulentes. Que els gens estaven formats majoritàriament per àcid desoxiribonucleic (DNA) tampoc no va fer-se evident fins algun temps després, quan Hershey i Chase, sabent que el DNA era ric en fòsfor i no tenia sofre, i que les proteïnes eren riques en sofre però no tenien fòsfor, van emprar la microscòpia electrònica i medis radioactius per demostrar que el bacteriòfag T4 només utilitza molècules riques en fòsfor per infectar el seu hoste (Hershey i Chase, 1952). Un any després, Watson i Crick publicarien el treball que descrivia l'estructura de doble hèlix de la molècula de DNA (Watson i Crick, 1953). En els anys posteriors els avenços en genètica van començar a ser vertiginosos: el descobriment d'enzims capaços de replicar el DNA, el descobriment del nombre de cromosomes humans i les primeres relacions entre malalties i alteracions en el material genètic. Els anys seixanta van estar marcats pel descobriment de l'àcid ribonucleic (RNA) com a «transmissor» de la informació genètica per mitjà del codi genètic. En els anys setanta es van identificar els primers enzims de restricció, es van produir les primeres molècules de DNA recombinant i es va descriure la tecnologia Sanger per obtenir la seqüència de nucleòtids del DNA. A principis dels vuitanta es van generar els primers ratolins i mosques transgèniques, es va mapar el primer gen causant d'una malaltia, la corea de Huntington (Gusella, *et al*., 1983), i es va descriure la tecnologia de la reacció en cadena de la polimerasa (PCR). L'any 1987 es va obtenir el primer mapa genètic, basat en polimorfismes de longitud de fragments de restricció (de l'anglès *restriction fragment length polymorphism* o RFLP) i el 1989 es van descriure els microsatèl·lits, utilitzats a bastament per construir mapes genètics que permetrien el clonatge posicional de nombrosos gens causants de malalties. A principis dels noranta, la popularització de les tecnologies de seqüenciació permeté desxifrar els primers genomes bacterians. La posada a punt de metodologies per clonar fragments de genoma humà en vectors de llevat i de bacteris (YAC i BAC, respectivament), juntament amb l'evolució de tecnologies de biologia molecular aplicades a la microscòpia (hibridació *in situ* fluorescent o FISH, de l'anglès *fluorescent in-situ hybridization*) permeten plantejar l'elaboració de mapes genètics detallats del genoma humà i, eventualment, l'obtenció de

la seqüència completa de diferents organismes. L'any 1996 es completa el primer mapa de gens del genoma humà i un consorci internacional comença la seqüenciació massiva d'aquest genoma (l'anomenat Projecte Genoma Humà) i el 1998, l'empresa privada Celera Genomics anuncia (i s'entén com una amenaça) que persegueix el mateix objectiu. A finals de 1999, es publica la seqüència completa del primer cromosoma humà (el vint-i-dos). El 2000 és l'any del cromosoma 21, i de la seqüenciació dels genomes de dos dels organismes model més utilitzats en genètica, *Drosophila* i *Arabidopsis*. La carrera entre el sector públic i el privat acaba el 2001 amb la publicació dels dos primers esborranys de la seqüència del genoma humà (Lander *et al*., 2001; Venter *et al*., 2001). Durant la primera meitat de la dècada del tombant de segle els avenços continuen a un ritme frenètic i a voltes difícil de digerir: centenars de genomes seqüenciats, identificació de la base genètica de desenes de malalties, utilització de l'enginyeria genètica en la vida quotidiana, estudi i descobriment de la gran variabilitat existent en el genoma humà en forma de SNP (de l'anglès *single nucleotide polymorphism*) i CNV (de l'anglès *copy number variant*), i els intents de trobar el paper d'aquesta variabilitat en la predisposició a malalties comunes i complexes. I la carrera del coneixement continua cada dia més ràpidament... Recentment, l'aparició de noves tecnologies de seqüenciació alternatives a la clàssica Sanger obren un nou horitzó que ens ha de permetre obtenir la major part (95 %) de la seqüència completa de genomes individuals per determinar les variacions normals i les alteracions causants de malalties. Ja tenim proves que això és factible i no pot deixar de sorprendre'ns que, el que desenes de centres de tot el món i centenars d'investigadors van trigar gairebé set anys, ara pot fer-se en setmanes en un únic laboratori!

ESTRUCTURES REPETITIVES EN EL GENOMA HUMÀ

A mesura que el coneixement s'ha acumulat han sorgit noves preguntes, i així, una de les paradoxes que els biòlegs evolutius de mitjan segle xx van haver d'afrontar va ser l'observació que la complexitat dels organismes no estava correlacionada amb la mida del seu DNA. En aquella època es pensava que tot el material genètic estava constituït exclusivament per seqüència codificant (que, per tant, es convertiria en proteïna, la unitat funcional dels éssers vius), i així, era sorprenent veure que una ameba tenia un genoma fins a dues-centes vegades més gran que el dels éssers humans. Aquesta paradoxa va ser solucionada en els anys setanta, quan es va veure que no tot el genoma era codificant, sinó que tenia un elevat contingut repetitiu (Thomas, 1971; Gregory, 2005). De fet, ja des de les primeres descripcions del genoma humà es va veure que l'estrella principal de la genètica, les regions codificants, eren només aproximadament el 5 % del genoma (Lander *et al*., 2001) i almenys el 50 % del genoma estava compost per seqüències repetitives.

Sovint maltractades i conegudes com a *DNA escombraria* (*junk DNA*), les seqüències repetitives són de vital importància per a l'estructura d'un genoma i, a més, són extremadament útils per als científics que l'estudiem per entendre'n l'evolució. Les repeticions són usades com un registre paleontològic d'esdeveniments evolutius que van passar milers (i milions) d'anys enrere i que van deixar alguna petjada en el genoma que avui dia encara podem detectar. D'altra banda, i en ser considerades moltes vegades com a seqüències neutres, són elements molt adients i altament utilitzats, encara avui dia, per fer tota mena d'estudis, com tests de paternitat, en criminologia, o

reconstruccions filogenètiques. En general, les repeticions poden ser classificades en diferents famílies, que estan relacionades perquè en algun moment van compartir un avantpassat comú (de la mateixa manera que tots els éssers vius de la Terra estem classificats en grups o «calaixos» segons els nostres avantpassats). En aquest capítol distingirem entre dos grans tipus de repeticions: d'una banda, les que s'han anomenat *duplicacions segmentàries* (DS, de l'anglès *segmental duplications*, o LCR, de *low copy repeats*), que ocupen al voltant d'un 5 % del genoma humà i, d'altra banda, tots els altres tipus de repeticions, anomenades de manera genèrica *repeticions comunes* (de l'anglès *common repeats*), que conformen al voltant d'un 45 %.

Repeticions comunes

Les seqüències repetitives o repeticions comunes són fragments de DNA homòlegs que es troben en múltiples còpies (fins a milers) en el genoma. La manera com classifiquem generalment les repeticions està basada en els seus avantpassats, aquells elements actius que van començar un procés mutacional que va donar lloc a les grans famílies de repeticions que observem avui dia. De manera general, les podem agrupar en quatre grups diferents:

a) Elements transposables (ET), distribuïts esparsament en el genoma (amb una longitud de fins a 30 kb) i que ocupen aproximadament el 45 % del genoma humà.

b) Còpies retrotransposades inactives de gens cel·lulars, també anomenades *pseudogens processats*, que ocupen al voltant del 2 % del genoma humà.

c) Repeticions simples i en tàndem d'un nombre reduït de nucleòtids (entre un i sis), ocupen un 3 % del genoma humà.

d) Blocs de repeticions en tàndem, típiques d'estructures cromosòmiques com centròmers i telòmers. El seu percentatge en el genoma humà és difícil d'avaluar amb exactitud, perquè són extremadament polimòrfics en mida i generalment es troben infrarepresentats en qualsevol projecte de seqüenciació.

El grup més nombrós d'elements repetitius en els genomes dels mamífers són els *elements transposables* (ET). Els ET són fragments de DNA o RNA que són capaços de reproduir-se i inserir-se en el genoma hoste, i ocupen el 45 % del genoma humà. En mamífers es poden agrupar en quatre categories principals: *elements llargs esparsos* (LINE, *long interspersed nuclear elements*), que ocupen al voltant del 20 % del genoma humà, incloent-hi fins a 516.000 còpies del seu element més comú LINE-1; *elements curts esparsos* (SINE, *short interspersed nuclear elements*), que ocupen prop del 13 % del genoma humà, incloent-hi més d'un milió de còpies d'un element conegut com *Alu*; *retrotransposons amb repeticions terminals llargues* (LTR, *long terminal repeat retrotransposons*), que són elements que es mobilitzen mitjançant una molècula de RNA, generalment flanquejats per altres seqüències repetitives, que ocupen un 8 % del genoma humà i, finalment, *transposons de DNA* (o elements mòbils), que conformen aproximadament el 3 % del nostre genoma. Els tres primers es transposen mitjançant intermediaris de RNA, mentre que l'últim es transposa directament com a DNA.

El segon major component en quantitat de contingut repetitiu del genoma són les *repeticions simples*. Aquestes són repeticions perfectes o imperfectes en tàndem de petits motius de seqüència. Si la unitat de repetició és petita (entre una i tretze bases) aquestes repeticions s'anomenen *microsatèl·lits*, i si tenen entre catorze i cinc-centes bases són anomenades *minisatèl·lits*. Aquestes repeticions es creu que s'originen per errors en la

replicació del DNA per «relliscades» (*slippage*) de l'enzim DNA-polimerasa.

Malgrat que moltes d'aquestes repeticions són considerades neutres, hi ha una colla de casos en els quals podem observar que tenen conseqüències funcionals i, per tant, afecten el fenotip. Així, hi ha repeticions que es troben dins de regions codificants o en les seves regions reguladores. Les mateixes repeticions dels telòmers o de les regions pericentromèriques són importants perquè són elements constituents dels cromosomes. A més, i com a exemple del paper funcional que aquestes repeticions poden arribar a tenir, cal destacar que entre cinquanta i cent gens del genoma humà han evolucionat a partir d'ET i de retrotransposons. Segurament, un dels exemples més antics és el del gen *RAG1*. Aquest gen es creu que es va crear fa uns 500 milions d'anys (en l'origen dels vertebrats amb mandíbula) i té un paper clau en la recombinació V(D)J (recombinació entre membres de famílies gèniques amb un paper important en el sistema immunitari) (Kapitonov i Jurka, 2005). Així mateix, molts microRNA (petits RNA que regulen l'expressió gènica), sembla que han evolucionat a partir d'ET i del seu paper en la competència per la regulació d'expressió de seqüències semblants (Lu *et al*., 2005).

La distribució d'ET en el genoma no és homogènia, ja que tenen tendència a acumular-se preferentment en regions genòmiques que reuneixen característiques concretes, com per exemple els ET més antics, que es troben preferentment en regions amb alt contingut G + C (Lander *et al*., 2001). Aquesta distribució no aleatòria podria ser deguda a insercions no atzaroses en el genoma humà o a eliminacions selectives de les insercions atzaroses en certes regions. La inserció direccionada d'elements mòbils, malgrat la seva correlació amb altres elements, com poden ser les DS, no pot explicar tota la gran distribució que observem en el genoma, i és per això que es creu que no tots els llocs del genoma tenen la mateixa capacitat d'acceptar insercions. Atès que molts ET són elements no neutres que poden afectar l'expressió gènica de gens adjacents, la selecció natural eliminaria aquelles insercions que destrueixen (o empitjoren) la funcionalitat del genoma.

*Duplicacions segmentàries***.** Les duplicacions segmentàries (DS) (conegudes antigament com LCR o *low copy repeats*, repeticions de baix nombre de còpia) són fragments contigus de DNA que es localitzen en dos o més llocs del genoma. Poden contenir qualsevol dels elements intrínsecs del genoma, des de gens sencers a trossos d'exons o introns, seqüències reguladores, o regions no codificants (Bailey i Eichler, 2006). En la cerca de la relació entre estructura i funció del genoma, les DS són de gran importància, ja que tenen un paper rellevant en ambdós aspectes; així, des del punt de vista de l'estructura, molts dels reordenaments i de les estructures polimòrfiques que conté el genoma estan relaciones amb DS (Armengol *et al*., 2003). I, d'altra banda, les DS tenen un paper cabdal en processos de recombinació homòloga no al·lèlica (RHNA, vegeu la figura 1), pels quals regions altament idèntiques però no ortòlogues es recombinen tot creant duplicacions, delecions o inversions d'un dels trossos de DNA implicats en el procés. Les DS recents (i, per tant, altament, idèntiques) tindrien tendència a convertir el genoma en regions fràgils amb tendència a fracturar-se. Això trenca amb una idea que havia estat estesa no fa gaires anys, segons la qual el genoma es fragmentava de manera atzarosa. D'altra banda, i ja des dels anys seixanta, es va reconèixer la importància de la duplicació gènica com a motor evolutiu capaç de crear novetat (Ohno *et al*., 1968). Diferents estudis han demostrat que les DS tenen un paper important en la crea-

ció de nous gens, no solament mitjançant la duplicació de còpies íntegres de gens, sinó a causa de les dinàmiques de propagació de les DS, que sovint han resultat en la creació de noves variants gèniques per fusió gènica o per exaptació d'exons duplicats (Eichler, 2001). Finalment, i des del punt de vista mutacional, les duplicacions són també interessants ja que, al contrari que d'altres regions del genoma, són regions que permeten l'acumulació d'un elevat nombre de canvis nucleotídics. I no solament això: segons sembla, i a causa de la RHNA, regions properes a les DS també augmenten la probabilitat de veure's implicades en processos de reestructuració (Cheng *et al*., 2005).

En els darrers anys s'ha vist que diferents genomes tenen diferent contingut de DS. Per exemple, el genoma del ratolí té més o menys el mateix contingut relatiu de DS que l'humà (~ 5 %), però la seva distribució en els cromosomes (similar a la que es pot veure en el genoma de la rata) no és esparsa, sinó que es troben localitzades en un nombre limitat de regions i en estructura de tàndem (fins al 70-90 % del contingut duplicat és en tàndem, comparat amb el 40 % en el genoma humà) (She *et al*., 2008). En humans, contràriament, el que predominen són les duplicacions entre cromosomes (43 %, percentatge que es redueix fins al 13 % en el genoma del ratolí) (Bailey *et al*., 2002). Així, en els genomes dels primats (com a mínim a humans, ximpanzés, goril·les, orangutans i macacos) hi ha una clara tendència de les DS a acumular-se en les regions subtelomèriques i pericentromèriques dels cromosomes. Aquesta distribució esbiaixada i l'estudi de com s'ha generat ha suggerit diferents models de distribució en el quals la transferència de material entre aquestes complicades (i altament repetitives) regions del genoma seria un fet comú i ajudaria a explicar el format de «trencaclosques» que tenen molts dels blocs de DS (Bailey i Eichler, 2006). Estudis recents han demostrat que les DS es troben organitzades al voltant de seqüències de DNA concretes (*core elements*), que sovint es troben associades amb regions codificants que han patit o que pateixen episodis de selecció positiva (Jiang *et al*., 2007). A més, s'ha pogut veure que l'evolució dels *core elements* no ha estat uniforme al llarg de l'evolució dels grans simis, sinó que hi ha hagut períodes de gran activitat duplicativa, contrastats amb altres períodes de relativa calma (Marquès-Bonet *et al*., 2009). En aquest sentit, la comparació de dues espècies tan properes com són humans i ximpanzés va demostrar que gairebé un terç de les duplicacions del genoma humà no eren presents en el genoma del ximpanzé (~ 25 Mb) (Cheng *et al*., 2005), i això posa de manifest l'impacte i la importància estructural i evolutiva de les DS, que són clau per entendre l'evolució dels genomes i, en conseqüència, de les espècies.

VARIANTS ESTRUCTURALS EN EL GENOMA HUMÀ

No fa gaires anys, amb la publicació de les primeres seqüències del genoma humà (Lander *et al*., 2001; Venter *et al*., 2001), s'anuncià que dos humans qualssevol eren 99,9 % idèntics pel que fa a la seqüència de DNA. En aquells moments, era ben coneguda l'existència de seqüències repetitives polimòrfiques en la població, en forma de microsatèl·lits, minisatèl·lits i d'altres, i se suposava que els milions de SNP existents eren la font definitiva de variabilitat i un dels principals responsables de modular el fenotip i la predisposició a la malaltia en humans. El projecte HapMap (TIH Consortium64, 2005) es va endegar per tal d'avançar en l'estudi d'aquest tipus de polimorfismes en diferents poblacions humanes i d'establir mapes d'haplotips per utilitzar-los amb

finalitats de mapatge de gens causants de malalties humanes. Més recentment, amb l'avenç que va suposar la tecnologia de la hibridació genòmica comparada sobre suport sòlid (o *arrayCGH*) (Pinkel *et al*., 1998; Pollack *et al*., 1999), s'ha descobert l'existència d'una quantitat totalment inesperada de variants estructurals en el nostre genoma (Iafrate *et al*., 2004; Sebat *et al*., 2004). Amb el nom de *variants estructurals* (de l'anglès *structural variants* o SV) es coneixen aquelles regions del genoma de mida superior a 1 kb, per diferenciar-les dels polimorfismes d'inserció/deleció, que es troben en nombre, localització o orientació variables en diferents individus d'una mateixa espècie. Així doncs, sota aquest nom genèric trobem insercions, delecions, duplicacions, translocacions i inversions de material genètic. A dia d'avui, el grup més nombrós i conegut de variants estructurals el conformen les anomenades *variants de nombre de còpia* (de l'anglès *copy number variants* o CNV) que, com el seu nom indica, són fragments de genoma que es troben en un nombre de còpia variable entre individus i en comparació de la seqüència de referència del genoma humà. El març de 2008 s'havia informat d'un total de 15.466 CNV, que representen 5.083 locus independents i ocupen una fracció propera al 20 % de la seqüència eucromàtica del nostre genoma (vegeu la taula suplementària). Aquestes dades, però, són el resultat de l'anàlisi d'un nombre limitat d'individus amb unes tecnologies encara poc precises que resulten en una sobrerepresentació de CNV de mida gran, una infraestimació quant al nombre global i una sobrestimació quant a la mida que ocupen (Hurles *et al*., 2008; Kidd *et al*., 2008). El nombre de regions variables reconegudes en el genoma de cada individu ha anat variant a mesura que s'han emprat tecnologies amb una major resolució, i hem passat de desenes (Iafrate *et al*., 2004; Sebat *et al*., 2004) a centenars (Tuzun *et al*., 2005; Redon *et al*., 2006; Korbel *et al*., 2007) i, segurament, es comptaran per sobre del miler (Hurles *et al*., 2008). Les CNV estan distribuïdes de manera aleatòria per tot el genoma, però mostren un clar enriquiment en regions que contenen DS (Redon *et al*., 2006) i se suposa que aquest fet està relacionat amb un dels mecanismes pel qual s'originen (vegeu la figura 1) i perquè les CNV no deixen de ser DS que encara no s'han fixat en la població. Cal remarcar que, amb el coneixement que tenim ara mateix, la definició d'una seqüència concreta com a CNV o com a DS dependrà només de si les diferents còpies es troben o no en el genoma de referència.

El reconeixement de la importància funcional d'aquestes variants és una mostra del seu paper en certes malalties (vegeu la taula 2), però està molt lluny de ser complet. Atesa la seva mida, moltes es troben superposades amb gens i altres elements funcionals del genoma (promotors, *enhancers*, etc). Almenys en teoria, l'alteració del genoma que representen les variants estructurals pot tenir efectes importants en l'expressió de gens (vegeu la figura 3) i, per tant, pot ser determinant en el fenotip dels individus. Tot i així, un estudi recent fet en línies cellulars derivades de limfòcits de dos-cents setanta individus ha demostrat que tan sols un 18 % de la variabilitat d'expressió mesurable és atribuïble a les CNV, i la resta als SNP (Stranger *et al*., 2007). Això no obstant, hi ha un cert consens que aquestes dades representen una infravaloració del paper de les CNV, ja que es van obtenir emprant dades que avui sabem que estaven esbiaixades (Hurles *et al*., 2008).

La implicació de les CNV en malalties ha estat provada en una bona colla de casos (vegeu la taula 2). En alguns es tracta de variants de nombre de còpia rares que causen directament malalties mentre que, en altres, són variants freqüents en la pobla-

ció que confereixen un increment en la susceptibilitat a patir certes malalties. És també destacable, igual com succeeix amb els SNP i altres variants, que la seva distribució poblacional no és homogènia, i certes variants són més freqüents en unes poblacions humanes que en unes altres i representen un possible indicador de processos recents d'adaptació a l'entorn (Perry *et al*., 2007; Kidd *et al*., 2008). En aquest sentit, les variants estructurals representen un nou món de variabilitat a explorar que ha estat obviada en els grans estudis d'associació a escala de genoma sencer que s'han fet fins avui, i que van tenir el seu màxim exponent el darrer any amb la publicació de nombrosos treballs d'associació entre malalties complexes i SNP (Saxena *et al*., 2007; Tomlinson *et al*., 2007).

Malgrat la velocitat a què s'ha avançat en els darrers quatre anys, d'ençà que es va reconèixer que les SV eren molt més freqüents del que es creia, hi ha certs aspectes tècnics que encara cal solucionar per poder estudiar la implicació d'aquest tipus de variants en malalties complexes a escala del genoma complet. Les tecnologies per estudiar-ne un nombre reduït en milers d'individus estan al nostre abast (vegeu Estivill i Armengol, 2007). Però fer-ho a escala de genoma global ja són figues d'un altre paner. Aspectes com poder diferenciar la mida exacta de la variant estructural, el nombre de còpies d'un al·lel determinat o la seva orientació

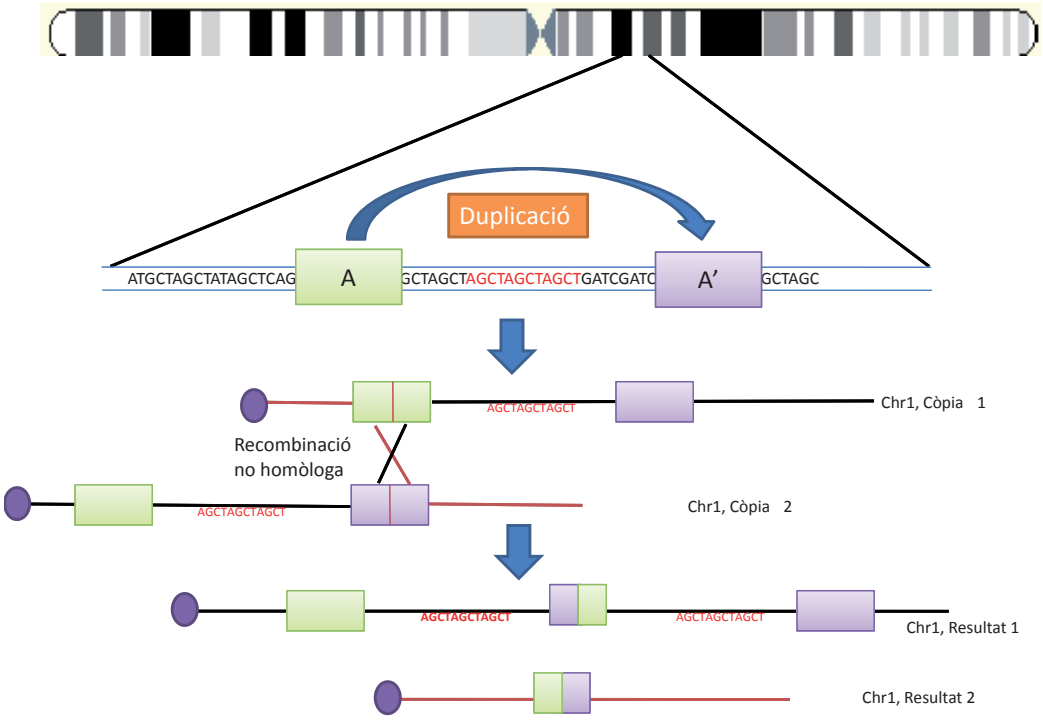

FIGURA 1. Recombinació homòloga no al·lèlica, reorganitzacions i duplicacions. L'encreuament entre còpies no homologues durant la meiosi pot potenciar duplicacions i delecions de segments de DNA (en vermell en l'esquema). Aquests tipus de reorganitzacions genòmiques poden conduir a polimorfismes estructurals capaços de conferir susceptibilitat per a certes malalties.

i localització cromosòmiques són crucials per poder emprar aquesta informació en estudis d'associació. Tot indica que una de les solucions de futur passa per la seqüenciació completa de genomes individuals. De la mà de les tecnologies de seqüenciació de nova generació (ABI-Solid, Illumina-Solexa, Roche-454) i ultraseqüenciació (Helicos) serà una realitat en els propers anys. Alguns treballs recents ja han emprat aplicacions d'aquestes tecnologies per a la identificació de variants estructurals en el genoma (Kidd *et al*., 2008).

VARIANTS ESTRUCTURALS EN ALTRES MAMÍFERS

Avui dia encara coneixem poca cosa de les CNV en mamífers. Pels pocs estudis efectuats, s'ha vist que les CNV en mamífers tampoc no estan distribuïdes a l'atzar en el genoma, sinó que es troben també regularment associades amb DS i, a més, es troben enriquides amb gens (Cooper *et al*., 2007). A l'hora de veure diferències entre llinatges, i amb l'estudi de llocs polimòrfics en ratolins de laboratori, s'ha vist que la taxa de creació de llocs polimòrfics és més alta que en primats (com a mínim un ordre de

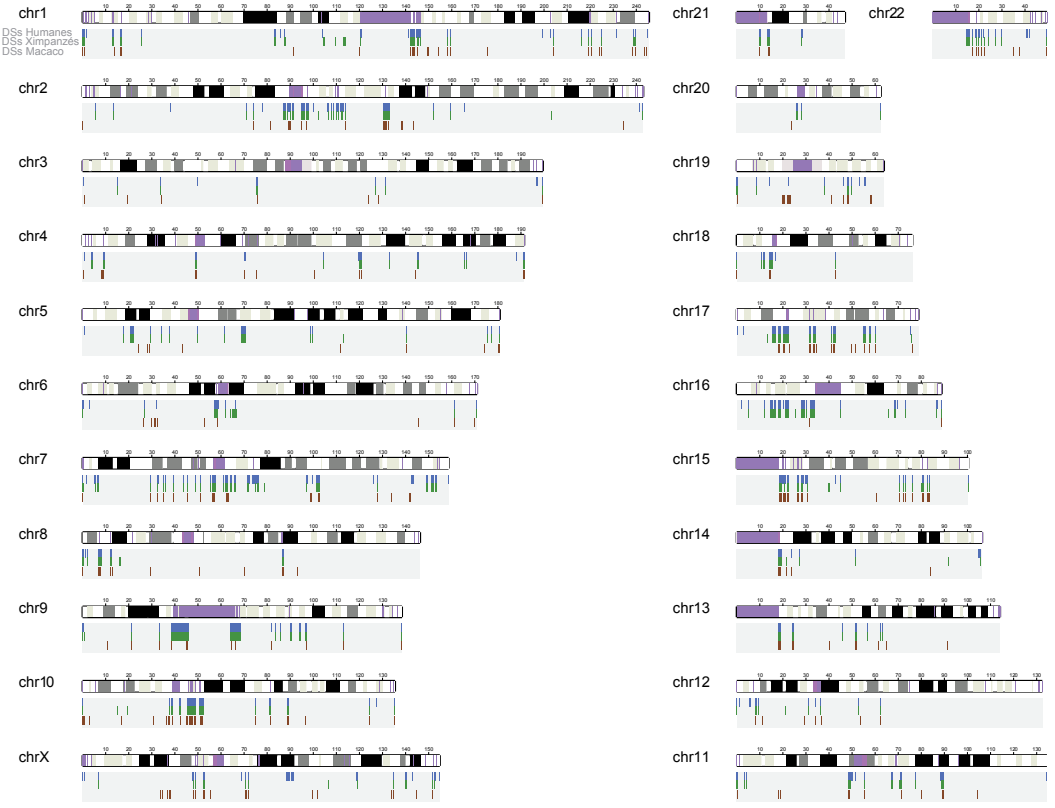

FIGURA 2. Distribució cromosòmica de les duplicacions segmentàries en humans i altres primats. Les duplicacions segmentàries tenen tendència a acumular-se en les regions subtelomèriques i pericentromèriques i, a més, les més noves acostumen a crear-se al voltant de duplicacions més antigues. Això es pot observar en aquests mapes, on hem dibuixat les duplicacions segmentaries (> 20 kb) en el genoma humà (blau), del ximpanzé (verd) i del macaco (marró).

magnitud) i que molts d'aquests llocs variants es van creant recurrentment en determinats locus del genoma (Egan *et al*., 2007). Aquest resultat concorda amb altres mesures de divergència com les taxes de substitució, per a les quals s'ha observat que en els rosegadors són marcadament més altes que en la resta de mamífers (Waterston *et al*., 2002), segurament a causa de la seva taxa metabòlica (més alta) o al temps generacional (més curt). Només alguns estudis han adreçat aquest tema en primats no humans (ximpanzés i macacos), per tal de poder fer la comparació més directa amb els humans (Perry *et al*., 2006, 2008; Lee *et al*., 2008). Segons sembla, aquests primats tenen un nombre de llocs variants més alt que els humans (sense arribar al nivell dels rosegadors), resultat també congruent tenint en compte que els ximpanzés i macacos són també més divergents entre si en seqüència (CSA Consortium, 2005; Gibbs *et al*., 2007; Hernandez *et al*., 2007). És interessant veure que molts dels llocs que són estructuralment variables en ximpanzés o macacos ho són també en humans i, a més, acostumen a estar superposats amb DS comunes a les dues espècies, i sembla que la RHNA és el mecanisme causal per a algunes de les CNV.

Aquestes observacions, òbviament, obren un escenari segons el qual l'estructura del genoma convertiria certes regions en llocs extremadament fràgils, amb més probabilitat de patir reorganitzacions recurrentment al llarg de l'evolució. Les conseqüències d'aquests moviments no són banals, ja que els llocs estructuralment polimòrfics s'ha demostrat que *a*) podrien tenir un impacte global en el genoma més gran que altres variants com ara els SNP i *b*) tot just es comença a descobrir la seva importància

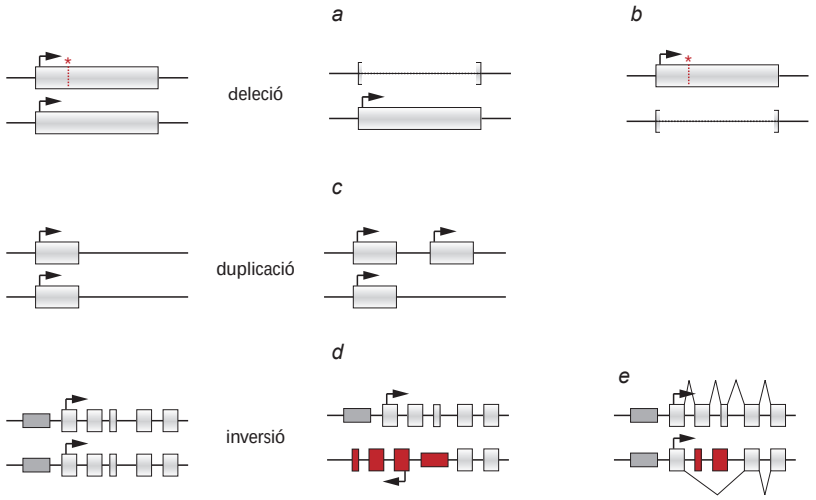

FIGURA 3. Diferents possibles efectes dels CNV sobre elements funcionals del genoma. *a*) Una variant de deleció elimina per complet l'expressió d'un dels dos al·lels. *b*) La variant de deleció desemmascara una mutació recessiva present en l'al·lel no delecionat. *c*) Una variant de duplicació crea una nova còpia d'un gen. *d*) Una inversió d'un element regulador i els tres primers exons d'un gen provoquen la pèrdua d'una pauta oberta de lectura en l'al·lel invertit i l'alteració de l'expressió normal d'aquest gen. *e*) La variant d'inversió, que afecta només un dels al·lels, reorganitza els exons del gen i provoca un canvi en el patró d'empalmament que té com a conseqüència l'expressió d'un nou transcrit.

Taula 1. *Contingut en duplicacions d'alguns dels genomes seqüenciats*

	Humà	*Ximpanzé*	*Macaco*	*Caenorhabditis elegans*	*Drosophila melanogaster*	*Ratolí*	*Rata*	*Pollastre*
DS > 1 Kb	5,20 %	~5 %	2,3 %	4,30 %	1,20 %	4,94 %	1,60 %	2,70 %
Mida del genoma Mb)	2,866	2,866	2,864	97	123	2,506	2,566	1,040

El contingut de duplicacions segmentàries, així com la seva distribució, canvia entre organismes. En aquesta taula hem resumit el contingut de duplicacions detectades en parells (WGAC (Bailey *et al*., 2001; She *et al*., 2006), amb més del 90 % d'identitat i majors d'1 kb). Els segments de DNA que no han estat assignats a cromosomes en el procés d'acoblament no van ser tinguts en compte (modificat de Bailey i Eichler, 2006).

en l'expressió i regulació dels gens adjacents, així com en malalties genètiques (Perry *et al*., 2006; Kidd *et al*., 2008; Lee *et al*., 2008).

ESTRUCTURA DEL GENOMA HUMÀ I MALALTIA

El genoma nuclear humà es troba repartit en un total de vint-i-dos parells d'autosomes i un parell de cromosomes sexuals. En totes i cadascuna dels milers de milions de cèl·lules del nostre cos, el genoma es troba en un equilibri dinàmic que combina estadis d'activitat per mantenir la funció cellular amb replicacions per donar lloc a les cèl·lules filles. Els cicles de replicació estan finament regulats per una gran quantitat de mecanismes que asseguren la replicació fidel del DNA i que, en cas de no ser possible, condueixen la cèl·lula a una mort programada. Així doncs, el control de l'arquitectura del genoma és crucial per a la supervivència de la cèl·lula i, en gran mesura, també de la viabilitat i capacitat de reproducció de l'organisme adult. L'alteració de l'estructura del genoma té conseqüències molt variables: algunes són aparentment innòcues (vegeu el cas de les CNV a dalt) però d'altres tenen greus efectes sobre els individus portadors o la descendència. Fins i tot s'ha postulat que aquesta arquitectura seria crucial per marcar esdeveniments d'especiació simpàtrica dels organismes (Navarro i Barton, 2003).

Les malalties causades per alteracions en l'estructura dels cromosomes són un clar exemple que la selecció natural deixa poc marge de maniobra a l'experimentació amb la informació genètica i que actua, en la majoria dels casos, per preservar una configuració determinada que, almenys, podem dir que funciona de manera adequada en l'entorn on es desenvolupa l'organisme. Aquestes malalties són el resultat de canvis en l'expressió dels gens, originats per l'alteració de l'estructura del genoma, que pot afectar el nombre de cromosomes (aneusomies o aneuploïdies) o fragments de cromosomes que contenen diversos gens (aneusomies segmentàries), i es pot trobar dins de gens (expansions de triplets), fins i tot com a conseqüència d'errors en els mecanismes que controlen la conformació i estructura normals del DNA (les modificacions epigenètiques de la cromatina). Potser el cas paradigmàtic d'alteracions en l'arquitectura del genoma com a font de malaltia el trobem amb les anomenades *malalties genòmiques* (Lupski, 1998), que resulten de la translocació, inversió, duplicació o deleció de regions cromosòmiques que contenen un o més gens sensibles a dosi. L'origen d'aquests reordenaments cal buscar-lo, en la majoria dels casos, en processos meiòtics de recombinació homòloga no al·lèlica (RHNA, vegeu la figura 1) entre còpies parà-

logues de duplicons (Lupski, 1998; Mazzarella i Schlessinger, 1998), tot i que també poden originar-se per processos de fusió d'extrems no homòlegs o, en mitosi, per processos de *fork stalling and template switching* (Lee *et al*., 2007). En els processos de RHNA, el grau d'identitat dels duplicons, la seva orientació i la localització cromosòmica determinaran el tipus de reordenament que té lloc entre les còpies paràlogues, que poden estar en cromosomes diferents, en cromosomes homòlegs, entre cromàtides germanes i, fins i tot, dins d'una mateixa cromàtida. Aquests tipus de reordenament dóna lloc a malalties com poden ser les α-talassèmies, la neuropatia hereditària amb hipersensibilitat a la pressió (HNPP) o les síndromes d'Angelman i Prader-Willi (PWAS), per esmentar només unes quantes de les més de cinquanta malalties monogèniques i sindròmiques que es coneixen. Malgrat que inicialment no estaven previstes, per l'àmplia definició del terme *malaltia genòmica*, *sensu stricto* també hauríem de comptar les malalties que recentment s'han descrit associades a la presència de diferents al·lels de nombre de còpia, com poden ser la pancreatitis hereditària, l'Alzheimer, el lupus eritematós disseminat, la malaltia de Chron o certes formes de glomerulonefritis, autisme o esquizofrènia (vegeu la taula 2).

ESTRUCTURA DEL GENOMA HUMÀ I EVOLUCIÓ

La inestabilitat i plasticitat del genoma és una característica fenotípica més i, com a tal, està subjecta a la selecció natural. Això no obstant, no totes les reestructuracions o regions variants del genoma tenen un efecte deleteri (nociu) sobre l'organisme que les porta (si no, no les veuríem de manera massiva en organismes viables avui dia). Per tant, l'estructura mateixa del genoma ens dóna informació sobre episodis del passat, de la relació entre espècies i de com hi ha actuat la selecció.

Selecció en regions duplicades i repetitives

Per exemple, els elements repetitius no són només agents mutacionals actius (tenen unes taxes de mutació molt més elevades que les regions de DNA de còpia única), sinó que són capaços de *a*) remodelar genomes sencers mitjançant recombinació no homòloga (i, per tant, generar reorganitzacions cromosòmiques), *b*) crear, de nous gens, la combinació de noves variants gèniques i *c*) modular el contingut G + C (amb la repercussió posterior que això pot tenir en altres taxes de mutació). Una característica debatuda, però de vital importància, és que moltes repeticions (com el cas de les DS) poden no solament canviar la taxa de mutació seva mateixa, sinó també de les seqüències adjacents, i aporten aleshores una nova visió, ja que no totes les regions del genoma tindrien la mateixa probabilitat de ser duplicades i, per tant, de ser subjectes a la selecció. Concretament, i parlant en escala temporal, l'activitat dels transposons sembla que no ha estat homogènia i que ha declinat marcadament en els darrers 35-50 milions d'anys (excepte potser la família LINE1). És interessant veure que els elements *Alu* sembla que han patit una increïble explosió uns quaranta milions d'anys enrere just en la radiació dels primats (Lander *et al*., 2001; Venter *et al*., 2001; CSA Consortium, 2005; Gibbs *et al*., 2007). L'estudi dels genomes d'altres organismes, com ara el del ratolí, demostra que aquesta disminució d'activitat de transposició podria ser específica de primats (Lander *et al*., 2001).

Els genomes són estructures conservadores que intenten preservar la seva integritat

TAULA 2. *Malalties associades a variants de nombre de còpia*

Malaltia o condició	*SD*	*Gen*	*Efecte*	*Allel de risc*	*Referència*
Malalties inflamatòries i autoinmunitàries					
Susceptibilitat infecció VIH-1/sida	Sí	CCL3L1	Dosi	Nombre de còpia baix	Gonzalez *et al.*, 2005
Artritis reumatoide	Sí	CCL3L1	Dosi	Nombre de còpia elevat	McKinney *et al.*, 2007
Diabetis de tipus 1	Sí	CCL3L1	Dosi	Nombre de còpia elevat	McKinney *et al.*, 2007
Púrpura trombocitopènica idiopàtica	Sí	FCGR2C	Dosi	Activació de FCGR2C-ORF	Breunis *et al.*, 2008
Lupus eritematós sistèmic	Sí	FCGR3B	Dosi	Nombre de còpia baix	Aitman *et al.*, 2006; Fanciulli *et al.*, 2007
Lupus eritematós sistèmic	Sí	C4A/C4B	Dosi	Nombre de còpia baix	Yang *et al.*, 2007
Poliangiitis microscòpica	Sí	FCGR3B	Dosi	Nombre de còpia baix	Fanciulli *et al.*, 2007
Granulomatosi de Wegener	Sí	FCGR3B	Dosi	Nombre de còpia baix	Fanciulli *et al.*, 2007
Malaltia de Crohn	Sí	DEFB4	Dosi	Nombre de còpia baix	Fellermann *et al.*, 2006
Psoriasi	Sí	DEFB	Dosi	Nombre de còpia elevat	Hollox *et al.*, 2007
Psoriasi	Sí	LCE3C/B	Dosi	Nombre de còpia baix	De Cid *et al.* [en prep.]
Pancreatitis hereditària	No	SPINK1	Dosi	Deleció	Masson *et al.*, 2006
Pancreatitis hereditària	No	PRSS1	Dosi	Duplicació / Triplicació	Le Maréchal *et al.*, 2006
Malalties neurològiques, psiquiàtriques i del neurodesenvolupament					
Trastorn bipolar	No	GSK3B	Posicional	Nombre de còpia elevat	Lachman *et al.*, 2007
Malaltia de Parkinson (forma juvenil)	No	SNCA	Dosi	Duplicació / Triplicació	Singleton *et al.*, 2003
Malaltia d'Alzheimer (forma juvenil)	No	APP	Dosi	Duplicació	Rovelet-Lecrux *et al.*, 2006
Autisme (i trastorns associats)	NA	Múltiple	Dosi	*De novo* CNV, múltiples	Sebat *et al.*, 2007; Szatmari, 2007
Esquizofrènia	Variable	Múltiple	Dosi	Múltiple	TISC, 2008; Walsh *et al.*, 2008
Malalties genòmiques [1]	Sí	Múltiple	Dosi	Delecions i duplicacions	Diversos autors

Malaltia o condició	*SD*	*Gen*	*Efecte*	*Al·lel de risc*	*Referència*
Càncer					
Leucèmia limfoblàstica aguda	Sí	Múltiple	Dosi	Múltiple	Mullighan *et al.*, 2007
Leucèmia limfoblàstica BCR–ABL1	No	Múltiple	Dosi	Múltiple	Mullighan *et al.*, 2008
Leucèmia mieloide aguda	Sí	Múltiple	UPD	Múltiple	Raghavan *et al.*, 2005
Leucèmia prolimfocítica de cèl·lules T	Sí	Múltiple	Dosi, posicional, UPD	Múltiple	Durig *et al.*, 2007
Leucèmia limfocítica crònica	Sí	Múltiple	Dosi, UPD	Múltiple	Pfeifer *et al.*, 2007
Limfoma de cèl·lules de mantell	Sí	Múltiple	Dosi, posicional, UPD	Múltiple	Bea *et al.* [en prep.]
Limfoma fol·licular	Sí	Múltiple	UPD	Múltiple	Fitzgibbon *et al.*, 2007
Mieloma múltiple	Sí	Múltiple	Dosi, UPDl	Múltiple	Walker *et al.*, 2006
Adenocarcinoma de pulmó	Sí	Múltiple	Dosi	Múltiple	Weir *et al.*, 2007
Neuroblastoma	Sí	Múltiple	Variable	Múltiple	Selzer *et al.*, 2005
Hepatoblastoma	Sí	Múltiple	Variable	Múltiple	Suzuki *et al.*, 2008
Càncer pancreàtic	Sí	Múltiple	Variable	Múltiple	Lucito *et al.*, 2007
Càncer de mama familiar	No	MTUS1	Posicional	Deleció d'exó comporta baix risc	Frank *et al.*, 2007
Càncer de còlon primari	SD	Múltiple	Dosi	Múltiple	Camps *et al.*, 2008

Més d'una trentena de malalties han estat associades a variants de nombre de còpia.

[1] Inclou més d'una cinquantena de malalties del neurodesenvolupament considerades com a malalties rares per la seva «baixa» freqüència. Entre aquestes podem destacar, per freqüència en la població, les síndromes de DiGeorge, de Prader-Willi/Angelman i de Williams-Beuren.

per continuar funcionant, i han evolucionat tota una sèrie de mecanismes per contrarestar els efectes d'aquest bombardeig d'insercions paràsites (i transgèniques). Per tant, els ET i els genomes hostes estan buscant sempre la manera de suprimir les insercions de codi forà (els genomes hostes) o de burlar els sistemes de suprimir activitats defensives dels genomes (per part dels ET) (Jurka *et al*., 2007).

Una pregunta que surt d'aquesta observació és: per què els genomes complexos i de tarannà conservatiu han permès la massificació d'elements forans (fins a un 45 % del seu genoma)? Sembla que alguns ET poden ser beneficiosos per als hostes i, per tant, poden haver evolucionat no de manera neutra o pertorbadora per a l'organisme que els accepta, sinó donant-los algun avantatge evolutiu que hauria d'augmentar-ne les possibilitats de supervivència. Pel que fa a les DS, i tal com abans s'ha esmentat, diferents evidències suggereixen que alguns dels gens que es troben a dins seu podrien haver estat sotmesos a processos de selecció positiva que haurien ajudat a la ràpida dispersió d'aquests elements en el genoma. Això és encara més clar en el cas de gens que es troben en els *core elements* de les duplicacions, que actuarien com a elements catalitzadors, duplicant-se i arrossegant altres seqüències contigües (Jiang *et al*., 2007).

Selecció en regions variants del genoma

Ja que les CNV estan distribuïdes per tot el genoma però mostren enriquiment en certes zones, s'espera que l'explicació d'aquest fenomen pugui provenir d'un biaix de la font mutacional que les crea, o bé del fet que han estat subjectes a un procés de selecció natural que no permet a totes les regions del genoma de ser tan «flexibles». Segurament (i com sempre) la resposta encertada pot ser una combinació de les dues. Malgrat que sabem que moltes CNV podrien estar subjectes a la selecció natural, algunes poden ser catalogades com a neutres i, per tant, poden tenir un efecte suau o nul en el fenotip dels individus. Entre dos individus catalogats com a fenotípicament «normals» podem trobar prop d'un miler de regions variants (Dermitzakis *et al*., 2008) i, a més, estudis en ratolins han demostrat que grans delecions (de fins a una megabase) no tenen un efecte fenotípic clar (Nobrega *et al*., 2004).

Com hem vist abans, les CNV tenen un paper decisiu en algunes malalties genètiques humanes. Així, els efectes de la selecció purificadora haurien de ser més visibles en les delecions, en què, per definició, l'efecte fenotípic hauria de ser més clar, i això és el que observem en la distribució de CNV, en què les delecions estan en franca minoria i tendeixen a evitar zones d'alt contingut gènic (Conrad *et al*., 2006). Sorprenentment, també hi ha casos de selecció darwiniana (adaptativa o positiva) en variants estructurals. Un estudi va trobar una inversió polimòrfica en el cromosoma 17 humà amb evidències de selecció positiva, potencialment lligada a un increment de la fertilitat (Stefansson *et al*., 2005; Zody *et al*., 2008). I no solament això: molts dels gens afectats per CNV es troben en categories típicament associades a episodis de selecció positiva en humans, com ara gens d'adaptació a l'entorn i immunitat. Un altre exemple de la possible no-neutralitat de moltes d'aquestes variants estructurals el trobem en l'evolució dels cromosomes. La reconstrucció dels cariotips ancestrals de mamífers ha demostrat que els punts de trencament de les reorganitzacions cromosòmiques no són a l'atzar, sinó que al llarg de l'evolució hi ha hagut certs punts del genoma (generalment associats a DS) que han estat recurrentment reorganitzats (Murphy *et al*., 2005). De tota

manera, la relació entre les DS i els punts de trencament no és necessàriament causal i, per tant, podrien estar relacionades per altres motius encara per determinar.

Evolució cromosòmica

Si ens fixem precisament en l'evolució de l'estructura dels cromosomes i, més concretament, en l'evolució en primats, podem veure que l'estructura no s'ha mantingut igual, i diverses configuracions han aparegut i desaparegut en l'evolució dels grans simis als humans. Un dels exemples més atractius i clarificadors per veure com l'evolució ha modelat l'estructura dels cromosomes és el cromosoma 2 humà. Tots els grans simis tenen vint-i-quatre cromosomes, i només els humans en tenim vint-i-tres. Aquesta reducció del nombre cromosòmic prové de la fusió de dos cromosomes ancestrals en un únic cromosoma. Les evidències per a aquesta observació provenen de *a*) els cromosomes anàlegs de ximpanzé tenen exactament el mateix patró de bandes que els dos braços del cromosoma 2, *b*) en el lloc de la hipotètica fusió dels dos cromosomes trobem (invertides) les típiques estructures repetitives dels telòmers —puntes dels cromosomes, on es creu que es va donar la fusió (Baldini *et al*., 1991)—, i *c*) com que cada cromosoma només té un centròmer, la fusió dels dos cromosomes hauria d'anar correlacionada amb la desactivació d'un d'aquests. I seguint la predicció, podem veure que el centròmer ancestral ara localitzat al braç *q* del cromosoma 2 està desactivat, malgrat que s'hi poden trobar restes d'elements repetitius típics d'aquesta estructura (Avarello *et al*., 1992). Precisament els centròmers no són elements fixos i estables en l'evolució d'un cromosoma, i fins a catorze noves activacions i consegüents desactivacions i relocalitzacions de centròmers han ocorregut en els darrers vint-i-cinc milions anys d'evolució (des de la separació de les mones del Vell Món i els grans simis) (Ventura *et al*., 2007). Tot i això, aquestes fissions i moviments centròmerics no són els únics moviments cromosòmics entre humans i altres primats. Ja des dels anys vuitanta es va veure que els genomes dels grans simis i els dels humans estaven separats per multitud de reorganitzacions (Yunis i Prakash, 1982). Concretament, entre humans i ximpanzés podem veure fins a nou grans inversions (detectades a escala citogenètica). En altres llinatges (com ara els hilobàtids, el gibó) trobem una taxa de reorganitzacions francament més alta i hi ha més de quaranta grans reorganitzacions cromosòmiques (Roberto *et al*., 2007). Les reorganitzacions cromosòmiques (inversions o translocacions) han estat des de fa temps en el punt de mira evolutiu pel seu possible paper en la creació d'espècies. Un dels mecanismes clàssics d'especiació cromosòmica (no l'únic) prediu que diferents reorganitzacions cariotípiques dins d'una espècie podrien contribuir a establir una barrera reproductiva mitjançant la noviabilitat (o viabilitat reduïda) de l'híbrid, o bé produint-ne l'esterilitat (Navarro i Barton, 2003). De tota manera, i malgrat que han estat demostrats en altres espècies (cavalls, gira-sols, mosquits, etc.) diferents papers de les reorganitzacions en processos d'especiació, avui creiem que en la nostra espècie no sembla haver estat un fet habitual (Marques-Bonet *et al*., 2007).

COROL·LARI

Però, quin és el motiu perquè el nostre genoma sigui tan flexible i variable? Per què l'evolució ha deixat que contingui aquesta mena de bombes de rellotgeria en forma de seqüències altament idèntiques i repetiti-

ves causants de reordenaments? Per què no han estat eliminades durant l'evolució? La presència d'aquestes seqüències representa una fulla de doble tall. D'una banda són regions dinàmiques que permeten fer experiments amb blocs del genoma que poden, a la llarga i en algunes circumstàncies, conferir avantatges als portadors mitjançant la creació de noves conformacions o de nous productes gènics, i són la manera de conservar el necessari dinamisme que permet al genoma respondre a canvis de l'entorn. Però d'altra banda, i en una manifestació del *yin-yang* de la natura, això no és gratuït i té conseqüències, indesitjables en molts casos i a curt termini, en forma de reordenaments que poden afectar la viabilitat dels organismes i causar malalties. Com en tantes altres coses de la vida, no hi ha ni bons ni dolents, ni blancs ni negres... Tot depèn del punt de vista de l'observador.

BIBLIOGRAFIA

Aitman, T. J.; Dong, R. [*et al*.] (2006). «Copy number polymorphism in Fcgr3 predisposes to glomerulonephritis in rats and humans». *Nature*, 439: 851-855.

Armengol, L.; Pujana, M. A. [*et al*.] (2003). «Enrichment of segmental duplications in regions of breaks of synteny between the human and mouse genomes suggest their involvement in evolutionary rearrangements». *Hum. Mol. Genet*., 12: 2201-2208.

Avarello, R.; Pedicini, A. [*et al*.] (1992). «Evidence for an ancestral alphoid domain on the long arm of human chromosome 2». *Hum. Genet*., 89: 247-249.

Bailey, J. A.; Eichler, E. E. (2006). «Primate segmental duplications: crucibles of evolution, diversity and disease». *Nat. Rev. Genet*., 7: 552-564.

Bailey, J. A.; Gu, Z. [*et al*.] (2002). «Recent segmental duplications in the human genome». *Science*, 297: 1003-1007.

Bailey, J. A.; Yavor, A. M. [*et al*.] (2001). «Segmental duplications: organization and impact within the current human genome project assembly». *Genome Res*., 11: 1005-1017.

CSA Consortium (2005). «Initial sequence of the chimpanzee genome and comparison with the human genome». *Nature*, 437: 69-87.

Conrad, D. F.; Andrews, T. D. [*et al*.] (2006). «A high-resolution survey of deletion polymorphism in the human genome». *Nat. Genet*., 38: 75-81.

Cooper, G. M.; Nickerson, D. A. [*et al*.] (2007). «Mutational and selective effects on copy-number variants in the human genome». *Nat. Genet*., 39: S22-S29.

Cheng, Z.; Ventura, M. [*et al*.] (2005). «A genome-wide comparison of recent chimpanzee and human segmental duplications». *Nature*, 437: 88-93.

Egan, C. M.; Sridhar, S. [*et al*.] (2007). «Recurrent DNA copy number variation in the laboratory mouse». *Nat. Genet*., 39: 1384-1389.

Eichler, E. E. (2001). «Recent duplication, domain accretion and the dynamic mutation of the human genome». *Trends. Genet*., 17: 661-619.

Estivill, X.; Armengol, L. (2007). «Copy number variants and common disorders: filling the gaps and exploring complexity in genome-wide association studies». *PLoS. Genet*., 3: 1787-1799.

Gibbs, R. A.; Rogers, J. [*et al*.] (2007). «Evolutionary and biomedical insights from the rhesus macaque genome». *Science*, 316: 222-234.

Gonzalez, E.; Kulkarni, H. [*et al*.] (2005). «The influence of CCL3L1 gene-containing segmental duplications on HIV-1/AIDS susceptibility». *Science*, 307: 1434-1440.

Gregory, T. R. (2005). «Genome size evolution in animals». A: Gregory, T. R. [ed.]. *The evolution of the genome*. San Diego: Elsevier, 3-87.

Gusella, J. F.; Wexler, N. S. [*et al*.] (1983). «A polymorphic DNA marker genetically linked to Huntington's disease». *Nature*, 306: 234-238.

Hernandez, R. D.; Hubisz, M. J. [*et al*.] (2007). «Demographic histories and patterns of linkage disequilibrium in Chinese and Indian rhesus macaques». *Science*, 316: 240-243.

Hershey, A. D.; Chase, M. (1952). «Independent functions of viral protein and nucleic acid in growth of bacteriophage». *J. Gen. Physiol*., 36: 39-56.

Hollox, E. J.; Huffmeier, U. [*et al*.] (2008). «Psoriasis is associated with increased beta-defensin genomic copy number». *Nat. Genet*., 40: 23-25.

Hurles, M. E.; Dermitzakis, E. T. [*et al*.] (2008). «The functional impact of structural variation in humans». *Trends. Genet*., 24: 238-245.

Iafrate, A. J.; Feuk, L. [*et al*.] (2004). «Detection of large-scale variation in the human genome». *Nat. Genet*., 36: 949-951.

Jiang, Z.; Tang, H. [*et al*.] (2007). «Ancestral reconstruction of segmental duplications reveals punctuated cores of human genome evolution». *Nat. Genet*., 39: 1361-1368.

Jurka, J.; Kapitonov, V. V. [*et al.*] (2007). «Repetitive sequences in complex genomes: structure and evolution». *Annu. Rev. Gen. Hum. Genet*., 8: 241-259.

Kapitonov, V. V.; Jurka, J. (2005). «RAG1 core and V(D)J recombination signal sequences were derived from Transib transposons». *PLoS. Biol*., 3: e181.

Kidd, J. M.; Cooper, G. M. [*et al.*] (2008). «Mapping and sequencing of structural variation from eight human genomes». *Nature*, 453: 56-64.

Korbel, J. O.; Urban, A. E. [*et al.*] (2007). «Systematic prediction and validation of breakpoints associated with copy-number variants in the human genome». *Proc. Natl. Acad. Sci. USA*, 104: 10110-10115.

Lander, E. S.; Linton, L. M. [*et al.*] (2001). «Initial sequencing and analysis of the human genome». *Nature*, 409: 860-921.

Lee, A. S.; Gutierrez-Arcelus, M. [*et al.*] (2008). «Analysis of copy number variation in the rhesus macaque genome identifies candidate loci for evolutionary and human disease studies». *Hum. Mol. Genet*., 17: 1127-1136.

Lee, J. A.; Carvalho, C. M. [*et al.*] (2007). «A DNA replication mechanism for generating nonrecurrent rearrangements associated with genomic disorders». *Cell*, 131: 1235-1247.

Lu, C.; Tej, S. S. [*et al.*] (2005). «Elucidation of the small RNA component of the transcriptome». *Science*, 309: 1567-1569.

Lupski, J. R. (1998). «Genomic disorders: structural features of the genome can lead to DNA rearrangements and human disease traits». *Trends. Genet*., 14: 417-422.

Marques-Bonet, T.; Kidd, J. M. [*et al.*] (2009). «A burst of segmental duplications in the African great ape genome ancestor». *Nature*. [En premsa]

Marques-Bonet, T.; Sanchez-Ruiz, J. [*et al.*] (2007). «On the association between chromosomal rearrangements and genic evolution in humans and chimpanzees». *Genome Biol*., 8: R230.

Mazzarella, R.; Schlessinger, D. (1998). «Pathological consequences of sequence duplications in the human genome». *Genome Res*., 8: 1007-1021.

Morgan, T. H. (1915). «Localization of the hereditary material in the germ cells». *Proc. Natl. Acad. Sci. USA*, 1: 420-429.

Murphy, W. J.; Larkin, D. M. [*et al.*] (2005). «Dynamics of mammalian chromosome evolution inferred from multispecies comparative maps». *Science*, 309: 613-617.

Navarro, A.; Barton, N. H. (2003). «Chromosomal speciation and molecular divergence–accelerated evolution in rearranged chromosomes». *Science*, 300: 321-324.

Nobrega, M. A.; Zhu, Y. [*et al.*] (2004). «Megabase deletions of gene deserts result in viable mice». *Nature*, 431: 988-993.

Ohno, S.; Wolf, U. [*et al.*] (1968). «Evolution from fish to mammals by gene duplication». *Hereditas*, 59: 169-187.

Perry, G. H.; Dominy, N. J. [*et al.*] (2007). «Diet and the evolution of human amylase gene copy number variation». *Nat. Genet*., 39: 1256-1260.

Perry, G. H.; Tchinda, J. [*et al.*] (2006). «Hotspots for copy number variation in chimpanzees and humans». *Proc. Natl. Acad. Sci. USA*, 103: 8006-8011.

Perry, G. H.; Yang, F. [*et al.*] (2008). «Copy number variation and evolution in humans and chimpanzees». *Genome Research*, 18: 1698-1710.

Pinkel, D.; Segraves, R. [*et al.*] (1998). «High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays». *Nat. Genet*., 20: 207-211.

Pollack, J. R.; Perou, C. M. [*et al.*] (1999). «Genome-wide analysis of DNA copy-number changes using cDNA microarrays». *Nat. Genet*., 23: 41-46.

Redon, R.; Ishikawa, S. [*et al.*] (2006). «Global variation in copy number in the human genome». *Nature*, 444: 444-454.

Roberto, R.; Capozzi, O. [*et al.*] (2007). «Molecular refinement of gibbon genome rearrangements». *Genome Res*., 17: 249-257.

Saxena, R.; Voight, B. F. [*et al.*] (2007). «Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels». *Science*, 316: 1331-1336.

Sebat, J.; Lakshmi, B. [*et al.*] (2004). «Large-scale copy number polymorphism in the human genome». *Science*, 305: 525-528.

She, X.; Cheng, Z. [*et al.*] (2008). «Mouse segmental duplication and copy number variation». *Nat. Genet*. [En premsa]

She, X.; Liu, G. [*et al.*] (2006). «A preliminary comparative analysis of primate segmental duplications shows elevated substitution rates and a great-ape expansion of intrachromosomal duplications». *Genome Res*., 16: 576-583.

Stefansson, H.; Helgason, A. [*et al.*] (2005). «A common inversion under selection in Europeans». *Nat. Genet*., 37: 129-137.

Stranger, B. E.; Forrest, M. S. [*et al.*] (2007). «Relative impact of nucleotide and copy number variation on gene expression phenotypes». *Science*, 315: 848-853.

Thomas, C. A. Jr. (1971). «The genetic organization of chromosomes». *Annu. Rev. Genet*., 5: 237-256.

TIH Consortium64 (2005). «A haplotype map of the human genome». *Nature*, 437: 1299-320.

Tomlinson, I.; Webb, E. [*et al.*] (2007). «A genome-wide

association scan of tag SNPs identifies a susceptibility variant for colorectal cancer at 8q24.21». *Nat. Genet*., 39: 984-988.

Tuzun, E.; Sharp, A. J. [*et al*.] (2005). «Fine-scale structural variation of the human genome». *Nat. Genet*., 37: 727-732.

Venter, J. C.; Adams, M. D. [*et al*.] (2001). «The sequence of the human genome». *Science*, 291: 1304-1351.

Ventura, M.; Antonacci, F. [*et al*.] (2007). «Evolutionary formation of new centromeres in macaque». *Science*, 316: 243-246.

Waterston, R. H.; Lindblad-Toh, K. [*et al*.] (2002). «Initial sequencing and comparative analysis of the mouse genome». *Nature*, 420: 520-562.

Watson, J. D.; Crick, F. H. (1953). «Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid». *Nature*, 171: 737-738.

Yunis, J. J.; Prakash, O. (1982). «The origin of man: a chromosomal pictorial legacy». *Science*, 215: 1525-1530.

Zody, M. C.; Jiang, Z. [*et al*.] (2008). «Evolutionary toggling of the MAPT 17q21.31 inversion region». *Nat. Genet. Nature Genetics*, 40: 1076-1083

Taula suplementària. *Treballs que han identificat regions variants de nombre de còpia*

Autor de la publicació	*Núm. CNV*	*Any*	*Metodologia*[1]	*PubMedID*	*Títol del treball*
Sasso *et al.*	1	1995	RFLP + PCR	7657830	Ethnic differences of polymorphism of an immunoglobulin VH3 gene
McLellan *et al.*	1	1997	RFLP + PCR	9415705	Frequent occurrence of CYP2D6 gene duplication in Saudi Arabians
Small *et al.*	2	1997	RFLP + PCR	9140403	Emerin deletion reveals a common X-chromosome inversion mediated by inverted repeats.
Blanchong *et al.*	1	2000	RFLP + PCR	10859342	Deficiencies of human complement component C4A and C4B and heterozygosity in length variants of RP-C4-CYP21-TNX (RCCX) modules in caucasians. The load of RCCX genetic diversity on major histocompatibility complex-associated disease.
Franchina *et al.*	1	2000	Diversa[2]	10799969	Allele-specific variation in the gene copy number of human cytosine 5-methyltransferase.
Gilles *et al.*	1	2000	Diversa	11161787	dCloning and characterization of a Golgin-related gene from the large-scale polymorphism linked to the PML gene
Osborne *et al.*	2	2001	Diversa	11685205	A 1.5 million-base pair inversion polymorphism in families with Williams-Beuren syndrome
Robledo *et al.*	1	2002	Diversa	12504850	A 9.1-kb gap in the genome reference map is shown to be a stable deletion/insertion polymorphism of ancestral origin.
Townson *et al.*	1	2002	Diversa	12355456	Gene copy number regulates the production of the human chemokine CCL3-L1

Autor de la publicació	*Núm. CNV*	*Any*	*Metodologia*[1]	*PubMedID*	*Títol del treball*
Yu *et al.*	1	2002	Diversa	12058345	Presence of large deletions in kindreds with autism.
Giglio *et al.*	4	2002	Diversa	12058347	Heterozygous submicroscopic inversions involving olfactory receptor-gene clusters mediate the recurrent t(4;8)(p16;p23) translocation
Hikami *et al.*	1	2003	Diversa	12618865	Variations of human killer cell lectin-like receptors: common occurrence of NKG2-C deletion in the general population.
Hollox *et al.*	1	2003	Diversa	12916016	Extensive normal copy number variation of a beta-defensin antimicrobial-gene cluster.
Milunski *et al.*	1	2003	CGH[3] + FISH	14986831	Unmasking Kabuki syndrome: chromosome 8p22-8p23.1 duplication revealed by comparative genomic hybridization and BAC-FISH
Vissers *et al.*	1	2003	aCGH	14628292	Array-based comparative genomic hybridization for the genome-wide detection of submicroscopic chromosomal abnormalities.
Gimelli *et al.*	2	2003	Diversa	12668608	Genomic inversions of human chromosome 15q11-q13 in mothers of Angelman syndrome patients with class II (BP2/3) deletions
Conrad *et al.*	1	2004	PCR	15160257	Novel procedures for high-throughput analysis of a frequent insertion-deletion polymorphism in the human T-cell receptor beta locus.
Fredman *et al.*	2	2004	Genotipatge de SNP	15247918	Complex SNP-related sequence variation in segmental genome duplications
Martin *et al.*	2	2004	Diversa	15616553	The sequence and analysis of duplication-rich human chromosome 16.
Shaw-Smith *et al.*	5	2004	aCGH	15060094	Microarray based comparative genomic hybridisation (array-CGH) detects submicroscopic chromosomal deletions and duplications in patients with learning disability/mental retardation and dysmorphic features.
Sebat *et al.*	79	2004	aCGH	15273396	Large-scale copy number polymorphism in the human genome.

Autor de la publicació	*Núm. CNV*	*Any*	*Metodologia* [1]	*PubMedID*	*Títol del treball*
Iafrate *et al.*	190	2004	aCGH	15286789	Detection of large-scale variation in the human genome.
Aldred *et al.*	1	2005	Diversa	15944200	Copy number polymorphism and expression level variation of the human alpha-defensin genes DEFA1 and DEFA3
Jobanputra *et al.*	1	2005	ROMA [4]	15714078	Application of ROMA (representational oligonucleotide microarray analysis) to patients with cytogenetic rearrangements.
Le Caignec *et al.*	1	2005	aCGH	15689449	Detection of genomic imbalances by array based comparative genomic hybridisation in fetuses with multiple malformations.
Stefansson *et al.*	2	2005	Genotipatge de SNP	15654335	A common inversion under selection in Europeans.
Feuk *et al.*	4	2005	Computacional [5]	16254605	Discovery of human inversion polymorphisms by comparative analysis of human and chimpanzee DNA sequence assemblies.
Bejjani *et al.*	6	2005	aCGH	15723295	Use of targeted array-based CGH for the clinical diagnosis of chromosomal imbalance: is less more?
Hinds *et al.*	25	2005	SNP + computacional	16327809	Common deletions and SNPs are in linkage disequilibrium in the human genome.
Sharp *et al.*	160	2005	aCGH	15918152	Segmental duplications and copy-number variation in the human genome.
Tuzun *et al.*	297	2005	Seqüenciació + computacional	15895083	Fine-scale structural variation of the human genome.
McCarroll *et al.*	495	2005	SNPs + computacional	16468122	Common deletion polymorphisms in the human genome.
Conrad *et al.*	544	2005	SNP + computacional	16327808	A high-resolution survey of deletion polymorphism in the human genome.
Gilling *et al.*	2	2006	Diversa	16642442	Breakpoint cloning and haplotype analysis indicate a single origin of the common Inv(10)(p11.2q21.2) mutation among northern Europeans.
Urban *et al.*	10	2006	aCGH	16537408	High-resolution mapping of DNA copy alterations in human chromosome 22 using high-density tiling oligonucleotide arrays.
Locke *et al.*	253	2006	aCGH	16826518	Linkage disequilibrium and heritability of copy-number polymorphisms within duplicated regions of the human genome.

Autor de la publicació	*Núm. CNV*	*Any*	*Metodologia*[1]	*PubMedID*	*Títol del treball*
Mills *et al.*	1304	2006	Computacional	16902084	An initial map of insertion and deletion (INDEL) variation in the human genome.
Redon *et al.*	1892	2006	aCGH + aCGI	17122850	Global variation in copy number in the human genome.
Zogopoulos *et al.*	255	2007	aCGI	17638019	Germ-line DNA copy number variation frequencies in a large North American population.
Simon-Sanchez *et al.*	256	2007	aCGI	17116639	Genome-wide SNP assay reveals structural genomic variation, extended homozygosity and cell-line induced alterations in normal individuals
Levy *et al.*	728	2007	Seqüenciació Sanger + computacional	17803354	The diploid genome sequence of an individual human.
Korbel *et al.*	881	2007	Ultraseqüenciació + computacional	17901297	Paired-end mapping reveals extensive structural variation in the human genome.
Wong *et al.*	1005	2007	aCGH	17160897	A comprehensive analysis of common copy-number variations in the human genome
de Smith *et al.*	1034	2007	aCGH	17666407	Array CGH analysis of copy number variation identifies 1284 new genes variant in healthy white males: implications for association studies of complex diseases.
Pinto *et al.*	1044	2007	aCGI	17911159	Copy-number variation in control population cohorts.
Wang *et al.*	1296	2007	aCGI	17921354	PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data.
Jakobsson *et al.*	785	2008	aCGI	18288195	Genotype, haplotype and copy-number variation in worldwide human populations.
Kidd *et al.*	2725	2008	Ultraseqüenciació + computacional	18451855	Mapping and sequencing of structural variation from eight human genomes.
Perry *et al.*	2832	2008	aCGH	18304495	The fine-scale and complex architecture of human copy-number variation.
Shen *et al.*	94	2008	aCGI	18373861	Improved detection of global copy number variation using high density, non-polymorphic oligonucleotide probes.
Total no superposats	17735	Agost 2008			

Informació extreta parcialment de la base de dades Database of Genomic Variants (*http://projects.tcag.ca/variation/*). El sumatori no coincideix amb la suma de les variants estructurals individuals identificades en cada estudi perquè la suma reflecteix només les variants no superposades. Per a la identificació de CNV s'han emprat tecnologies molt diverses al llarg del temps.

[1] RFLP: de l'anglès *restriction fragment length polymorphism* (fragments de restricció de longitud variable).

[2] Diversa: fa referència a la utilització de diferents metodologies de biologia molecular, com poden ser les transferències *Southern*, les electroforesis de camps polsants (o PFGE) i els RFLP, entre d'altres.

[3] CGH: hibridació genòmica comparada.

[4] ROMA: *representational oligonucleotide microarray analysis* és un tipus de CGH sobre microxips.

[5] Computacional: fa referència a diferents aproximacions basades en algorismes bioinformàtics.