Duplicacions segmentàries a la regió cromosòmica humana 8p23.1: evolució i expansió d'una nova família gènica

Nina Bosch Pagès

Tesi doctoral

Barcelona, Octubre 2008

Duplicacions segmentàries a la regió cromosòmica humana 8p23.1: evolució i expansió d'una nova família gènica

Nina Bosch Pagès

Memòria presentada per optar al grau de Doctora
Per la Universitat Pompeu Fabra.

Aquesta tesi doctoral ha estat realitzada sota la direcció del Dr. Xavier Estivill Pallejà i la codirecció del Dr. Lluís Armengol Dulcet al programa Gens i Malaltia del Centre de Regulació Genòmica.

Tesi doctoral adscrita al Departament de Ciències Experimentals i de la Salut, Universitat Pompeu Fabra.

International PhD Program in Health and Life Sicences,
Bienni 2003-2005

Xavier Estivill Pallejà Lluís Armengol Dulcet Nina Bosch Pagès

Barcelona, Setembre 2008

Els treballs de recerca duts a terme en aquesta tesi doctoral han estat realitzats al laboratori del programa Gens i Malaltia del Centre de Regulació Genòmica (CRG), centre integrat al Parc de Recerca Biomèdica de Barcelona (PRBB).

La recerca duta a terme ha estat possible gràcies a la beca BEFI de l'Instituto de Salud Carlos III (FIS-ISCIII) i Genoma España.

Als meus pares

La ignorància afirma o nega rotundament;
la ciència dubta

Voltaire

Index

Agraïments

Gràcies al petit granet de sorra (en segons quins casos, rocs) que heu aportat cadascun de vosaltres, aquesta tesi ha estat possible. Per tant no seria just no mencionar-vos en un projecte del que d'alguna manera o altra heu estat partícips. Gràcies a tots de tot cor i espero no deixar-me a ningú...

En primer lloc, a en **Xavier Estivill** per donar-me l'oportunitat de realitzar aquesta tesi doctoral i d'aquesta manera. Per la seva capacitat de motivar als seus doctorands, per la confiança dipositada en mi, pel seu ampli ventall d'idees i punts de mira, per proposar i no imposar, i sobretot per la llibertat donada, ja que és la millor manera per aprendre dels propis errors i per madurar en l'exercici de la ciència. Per què en paraules seves, per dedicar-se al món de la recerca cal tenir un esperit quixotesc, i ell n'és un exemple.

A en **Lluís Armengol** per haver-me guiat al llarg d'aquest camí. Per què les seves múltiples aportacions a nivell científic i informàtic han servit de catalitzador pel desenvolupament del projecte. Per estar sempre disponible, amb ganes d'ajudar i per haver-me il·lustrat en l'apassionant món dels duplicons i CNVs. I per ser un gran científic i millor persona amb qui dóna gust treballar.

A en **Josep Vilardell**, per haver-me rebut i orientat en els meus inicis. Sense els seus consells qui sap si avui aquesta tesi existiria.

Al **"camarot dels germans Marx – una gran família..."** que per mi va néixer essent Zulo-P13 i ha acabat amb CeGen-521

A la **Mònica Gratacòs**, per ajudar-me en el meu aterratge al laboratori amb els seus centenars de protocols, gels, quantificacions i "truquillos" que ha anat acumulant i que fan que en sàpiga un niu de tot. Per què tot i tenir l'aparença de sèria molt sèria, desborda sentit de l'humor i per haver-me ajudat no només a nivell científic.

A l'**Ester Ballana,** la veu de la consciència, per què "nant bé" no hauria pogut tenir millor predecessora. Per què la seva organització i sentit de la responsabilitat ens feia estar segurs al seu costat, llàstima que no fes un "allarguis". Per totes les coses que m'ha ensenyat dins del laboratori sobre mitocondris, sordesa, 8p23.1, haplotips, etc, etc. Però sobretot per l'amistat que ens ha unit fruit de les estones que hem compartit i compartirem fora del laboratori viatjant, jalant, ballant i opinant. Gràcies per poder comptar amb tu en tot moment!

A la **Celia Cerrato,** per l'agudesa dels seus comentaris, per la seva discreció, pels problemes tècnics compartits amb els gels d'acrilamida, el "pulsed-field" i

companyia. Per què costa trobar tan bones persones com ella, per ser de les que no fan soroll però ho capten tot i per ser una bona amiga.

A la **Marta Morell,** per què cal molta mà esquerra per ajudar a fer rutllar el laboratori i la Marta la té. Per totes les hores destinades al FISH per analitzar la inversió de 8p23.1, i aguantar estoicament i solucionar els múltiples mal de caps associats. Per què tot i compartir sovint "*inversions*", ja siguin cromosòmiques o de gustos i opinions, ens hem acabat entenent sempre.

A en **Rafa de Cid,** "Oh Leoncio!" per reconstruir totes les destrosses que feia amb els gels d'acrilamida la "nineta dels seus ulls".

A l'**Anna Carreras,** per moltes coses. Per la seva eficiència alhora de treballar, pel seu tarannà optimista, pel seu oli que ho fa tot més bo, per què té una manera de fer que fa fàcil el treball en equip, la convivència i el bon enteniment. Per haver-me aguantat algun dels meus maldecaps i haver compartit molts moments divertidíssims des de fa anys. En definitiva, per ser una gran amiga.

A en **Txema Mercader** per ser un amic únic. Per què per molt que el coneguis mai saps ben bé que li està passant pel cap, per ser una caixa de sorpreses continua, amb un brillant sentit de l'humor. Per què aquest últim any he substituït el "suport crg" pel "suport txema mercader", gràcies per ajudar-me amb múltiples dubtes estadístics, científics i burocràtics. Ha estat un plaer tenir-te "manu-manu" fent la tesi. Sort que el fil no es perdrà i ens veurem de congrés en congrés, je, je...

A l'**Imma Ponsa** per encomanar tranquil·litat. Per ensenyar-me citogenètica, informàtica i tantes d'altres coses. Per què torna a ser una d'aquelles persones que fan molt però sense fer soroll. Gràcies per ser-hi sempre!

A la **Maya Muiños** per haver vingut al P13. Per estar sempre disposada a ajudar pel que faci falta al laboratori. Pel seu bon rotllo encomanadís, per què és un piló d'alegria en forma de persona. Reuneix tantes qualitats humanes que no tinc paraules. Gràcies per tots els moments que hem passat juntes, per cuidar-

me tant aquests últims mesos i per fer-me veure la llum en plena ofuscació. I si us plau, no hem facis dormir en el traster quan vingui de visita!!!

A l'**Ester Saus** pel seu seny. Per saber escoltar, sospesar sempre les coses i buscar solucions. Per calmar al galliner tot sovint excepte quan perd ella el control que llavors ja podem tremolar tots...

A l'**Anna Brunet** per la seva punyeteria i per no enfadar-se ni quan se li venen les maletes.

A totes les que han anat arribant més tard, que hem coincidit poc però en tinc molt bons records, la **Silvia Porta**, la **Birguit**, la juerguista de l'**Elisa**, la **Susana** i la **Laia**.

A en **Manel** i en **Sergi** per aguantar tanta noia garlant tot el dia.

A tota la gent del CeGen: a en **Carles** i el que ens ha fet riure amb els seus comentaris tan insòlits com els seus ulls, a la **Cecilia**, a l'**Anna Puig** per la seva complicitat en les discussions dels cafès de després de dinar, a la **Kristin**, la **Magda**, la **Silvia** i la **Josiane**. Al **Roger** del servei de seqüenciació de la UPF per la seva predisposició a analitzar-ho tot quan fes falta.

Als post-docs pels seus consells de gat vell: en **Mario** per haver-me introduït en el món de la genòmica comparada i per la seva escrupolositat i rigorositat científica en el treball, a la **Yolanda** per donar-me sempre el seu punt de vista, a la **Kelly** per les seves classes de "Real time", a la **Mònica Bayés** per donar la seva opinió sempre crítica, a la **Geòrgia** per la seva paciència en explicar-me l'ABC de l'estadística, a la **Magda**, la **Mónica Bañez** i el seu salero , al salero de l'**Eulàlia Martí** (fooooooooooondo) per saber encomanar la passió per la recerca, per què gaudeix amb el que fa i es nota, i a l'**Eva**.

A tota la gent de **la resta de laboratoris del programa**, especialment a en **Jon**, en **Dani**, la **Kriszti** i al carismàtic **Ignasi**, per què treballar amb aquest bon ambient no té preu.

A les ladies dels dimarts, **Anna**, **Eva**, **Ànnia, Noe**, **Imma** i **Ruth** pels molts vespres de converses i discusions sempre interessants.

Als **xampis** de la facultat: **Magalí**, **Roger**, **Gemma**, **Kjell**, **Laura**, **Marta**, **Laia**, **Fabio**, **Biel**, **Anna**, **Esther** i **Susanna**. Per les trobades setmanals per Gràcia, finalment ha arribat el moment, gràcies pel suport i interès mostrat!!!!! Ah! I gràcies al senyor **Guillem**-nivell k amb l'assistència quasi a temps real amb els dubtes que m'han sorgit sobre la nostra estimada llengua i per totes les converses que hem tingut tot fen una cerveseta.

Als de sempre de Girona, la **Raquel**, en **David**, en **Miquel**, la **Fara**, en **Marc** i la **Marta**. Ja no us donaré més la tabarra amb el "no crec, estaré escrivint la tesi".

A la meva germana **Ariadna** per la paciència d'aguantar-me els meus atacs de mal geni. Per esperar-me els caps de setmana amb els braços oberts i fer-me costat.

Al meu germà **Ramon** i a la **Bet**, per saber que hi són sempre i per la seva curiositat per la meva "feina" que pot conduir a preguntar-me coses a priori tan òbvies com: si se sap de què està format el genoma de, posem per exemple una balena, i se sap l'ordre...per què no es poden crear balenes??

A la **Flora** per donar tanta vida i per què no deixa de sorprendre'm com en un cos tan petit i pot quebre tant sentit comú.

Al meu **pare**, per què tot i dedicar-se a una altra disciplina en té molt de biòleg, per què suposo que de ben petita ja em va encomanar la curiositat pels bitxos i el voler saber com funciona la natura. A la meva **mare** per donar-me el punt de vista més pràctic de les coses. A tots dos pel donae-me el seu suport incondicional, per ensenyar-me a conèixer altres maneres de fer tot viatjant, per aconsellar-me, ajudar-me i fer-me costat en totes les decisions que he anat prenent. Gràcies per tot.

A l'**àvia Manolita** pel seu carinyo i també a l'**Anna** i l'**Emma**.

A l'àvia **Maria** i els seus gens que l'han permès arribar mes enllà dels cent anys.

Finalment a la connexió 320 per economitzar-me l'escriptura d'aquesta tesi i moltes altres hores de navegació!

Preàmbul

Una de les grans fites de l'última dècada en el camp de la biomedicina ha estat l'obtenció de la seqüència del genoma humà. A partir d'aquest recurs cabdal s'ha pogut ampliar el coneixement del nostre genoma i s'han adreçat un gran nombre d'estudis dirigits a comprendre el seu funcionament. Una de les claus per assolir aquesta comprensió radica en l'observació de les diferències resultants de la comparació dels genomes d'espècies diferents, així com de diferents individus d'una mateixa espècie. D'aquesta manera la natura ens proporciona una eina de valor incalculable ja que gran part del procés evolutiu està basat en el mètode d'assaig i error alhora de crear noves variants o diferències a nivell genòmic. Així s'ha pogut determinar que aproximadament un 5% del nostre DNA està format per fragments que s'han duplicat recentment, les anomenades duplicacions segmentàries que, degut a la seva similitud de seqüència, propicien l'aparició de reordenaments intra i intercromosòmics. Arrel d'aquests reordenaments es poden produir guanys o pèrdues de material genòmic, que poden ser variables d'individu a individu i donar lloc al que es coneix com a variants en número de còpia.

La regió cromosòmica humana 8p23.1 és una regió especialment idònia per aprofundir en la plasticitat genòmica de les regions flanquejades per duplicacions segmentàries. Aquest fragment cromosòmic no tan sols

presenta una arquitectura genòmica que el fa susceptible a patir diferents reordenaments, si no que els gens que estan continguts en les duplicacions segmentàries de la regió són variables en número de còpia. Es tracta de gens implicats en la resposta immune com les alfa i beta defensines, i la seva variabilitat en número de còpia ha estat implicada en diverses malalties com la malaltia de Crohn. Val a dir que en l'inici d'aquest projecte les defensines eren uns gens pràcticament desconeguts, mentre que en l'actualitat existeixen més de mil publicacions relacionades amb aquests gens. La velocitat a la qual avança el camp de la genètica és trepidant, i l'estudi de la regió cromosòmica 8p23.1 i en concret de la família gènica específica de primats *FAM90A,* ha estat una oportunitat magnífica per analitzar aspectes genòmics de rellevància tan actual per entendre que ens diferencia com a humans com són els duplicons, les variants estructurals i les seves conseqüències a nivell transcripcional.

Introducció

1. El genoma humà

El genoma és el conjunt del material hereditari d'un organisme, la seqüència de nucleòtids que especifiquen les instruccions genètiques per al seu desenvolupament i funcionament i que són transmeses de generació en generació, de pares a fills. A més dels gens pròpiament dits, s'hi inclouen regions espaiadores, regions reguladores, restes de gens antigament funcionals i moltes seqüències més de funció o paper encara desconegut (Figura 1.1).

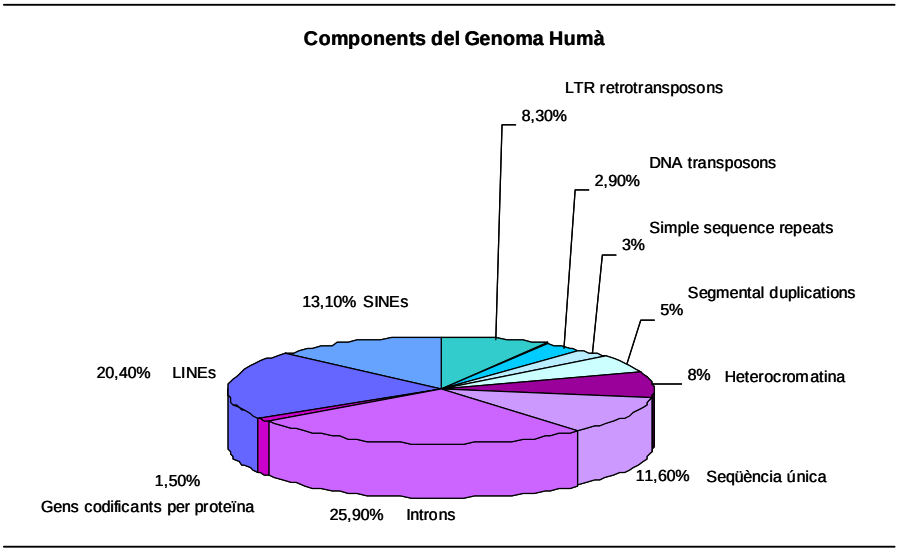

Figura 1.1 Components del genoma humà. El 45% consisteix en "paràsits genòmics": transposons de DNA, LTRs, LINEs i SINEs (extret de Gregory, T.R.[1]).

En el cas del genoma humà aquest material té un tamany estimat de 3.2 x10^9 parells de bases nucleotídiques[2]. Quan el Projecte del Genoma Humà va néixer, al 1990, ja es tenia coneixement del mecanisme de funcionament de molts gens, però l'esforç per seqüenciar el genoma sencer va permetre obtenir-ne una major comprensió. El fet de tenir avui en dia la seqüència del genoma humà complerta ha aportat a la comunitat científica un eina extremadament valuosa per explorar les nostres similituds i diferències com a humans.

1.1 Seqüenciació del genoma humà

El genoma humà està ple de seqüències de DNA duplicades i conté milions d'elements repetitius dispersos. Aquests elements són seqüències curtes i quasi idèntiques, de manera que dificulten l'assemblatge dels fragments on estan englobades. El principal problema sorgeix quan repeticions inconnexes s'assemblen com a contigües mitjançant eines computacionals[3-5]. Durant la passada dècada, amb el desenvolupament de protocols de seqüenciació a gran escala i mètodes d'anàlisi computacional avançats, ha estat possible generar assemblatges de seqüències que comprenen la majoria del genoma humà[2,6,7]. La seqüenciació del genoma humà es va dur a terme seguint dues estratègies diferents, la del Consorci públic per una banda i la de la iniciativa privada per l'altra (Figura 1.3).

1.1.1 Consorci Públic

Va ser a finals del 1990 quan es va idear el Projecte Genoma Humà amb la creació de centres de seqüenciació als Estats Units, Regne Unit, França i Japó i amb el suport de la Comunitat Europea (l'anomena't Consorci públic). L'estratègia adoptada pel consorci públic consistia en clonar el genoma en fragments de grandària adequada (de només uns centenars

de milers de bases en cromosomes artificials de llevat o BACs) que eren, al seu torn, seqüenciats mitjançant l'estratègia de la perdigonada ("*shotgun*") o seqüenciació a l'atzar. Per a fragments d'aquesta grandària, l'estratègia de la perdigonada necessita que cada nucleòtid sigui seqüenciat unes quantes vegades, quantificades amb el factor de cobertura o redundància, amb l'objectiu que no quedin regions de cap fragment sense seqüenciar almenys una vegada. A mesura que el fragment és major en grandària, hi ha una tendència a no augmentar la redundància tant com és necessari a fi de garantir que cap nucleòtid es quedi sense seqüenciar, per la qual cosa hi ha avui en dia alguns buits o "gaps" en la seva seqüència final.

En el projecte conduit pel consorci públic es van utilitzar limfòcits de dos donants de sexe masculí i dos donants de sexe femení, cadascú donant lloc a una llibreria de DNA. Una d'aquestes llibreries (RP11) es troba molt més representada que la resta en l'assemblatge final[2].

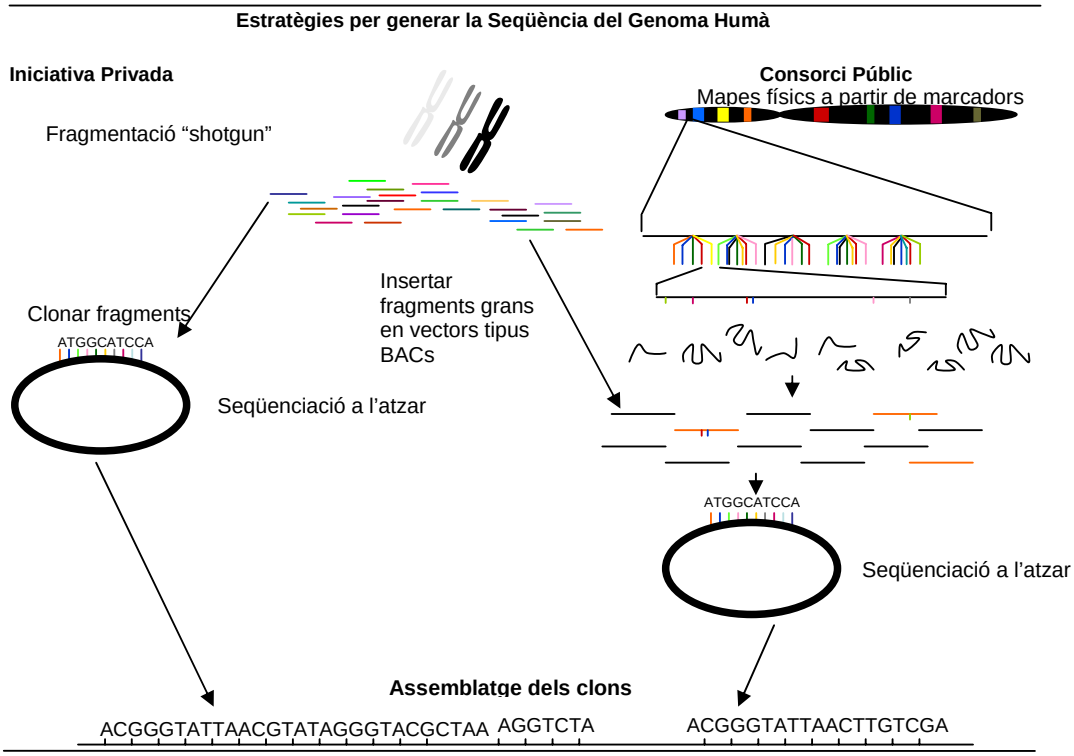

Figura 1.3 Esquema de les dues estratègies emprades per a la seqüenciació del genoma humà.

1.1.2 Iniciativa privada

Al 1998, Craig Venter, prèviament un dels líders del consorci públic i llavors ja en el sector privat (Celera Genomics), va llançar el desafiament que va conduir finalment a l'acceleració de tot el procés. La seva proposta consistia a abandonar les fases de mapatge físic i ordenació dels clons amb fragments grans per a passar directament a la seqüenciació completa del genoma mitjançant el mètode de perdigonada, deixant que nous algoritmes i ordinadors més potents s'encarreguessin de l'assemblatge de tota la seqüència. Rebuda amb escepticisme, la seva estratègia va demostrar la capacitat de seqüenciar un genoma complex en un temps rècord.

El seu grup va començar la seqüenciació del genoma humà (de fet, el genoma de cinc individus diferents) el 8 de setembre de 1999 i va concloure la fase d'obtenció de dades el 17 de juny del 2000. L'assemblatge sobre el qual es va basar la publicació, simultània al resultat del consorci públic el 15 de febrer del 2001, es va completar en tres mesos i mig[6].

Finalment mencionar que el 4 de setembre del passat any, es va publicar la seqüència complerta de DNA de Craig Venter. Aquesta publicació fou d'especial interès ja que es tractà del primer genoma diploide (ambdós cromosomes) que s'ha seqüenciat[8].

1.1.3 Components del genoma humà

La finalització del projecte de seqüenciació del genoma humà al 2001 també va permetre corroborar altres dades que ja feia temps que es coneixien, com ara la manca de correlació entre el tamany d'un genoma i la complexitat de l'organisme (altrament conegut com a "paradoxa del valor C"). D'aquesta manera els humans tenim 30.000 gens aproximadament , mentre que l'arròs per exemple en té 60.000. D'altra banda però, sabem que hi ha un nivell molt més elevat de "splicing" alternatiu en humans que no pas en la majoria d'espècies avui conegudes,

i és segurament aquesta font de generació d'isoformes gèniques, juntament amb complexos mecanismes de regulació gènics, els que ens confereixen complexitat a nivell de conducta, d'execució d'accions conscients, coordinació física, accions finament regulades en resposta a estímuls externs, capacitat d'aprenentatge, de memòria, etc[9].

En altres paraules, hi ha més proteïnes codificades per gen en humans, que a les altres espècies. Tot i així, una de les dades obtingudes amb la seqüenciació del genoma humà que més va sorprendre a la comunitat científica, va ser el menor nombre de gens existents respecte el que s'havia anticipat, indicatiu una vegada més, que la complexitat humana no només sorgeix del nombre de gens. De fet, del total del genoma, només un ~8% correspon a gens pròpiament. La resta, 2.95 Gb, correspon a zones lliures de gens o regions intergèniques i comprenen la major part de la seqüència del genoma humà[2].

Gran part del DNA intergènic pot ser un artefacte evolutiu sense una funció determinada en el genoma actual, pel que tradicionalment aquestes regions han estat denominades DNA "escombraria" (*Junk DNA*), denominació que inclou també les seqüències intròniques i pseudogens. No obstant això, aquesta denominació no és la més encertada donat el paper regulador conegut de moltes d'aquestes seqüències. A més el notable grau de conservació evolutiva d'algunes d'aquestes seqüències sembla indicar que posseeixen altres funcions essencials encara desconegudes o poc conegudes. Per tant, és preferible denominar-lo "DNA no codificant".

De fet, es postula que gran part de la seqüència de DNA considerada "única" deriva d'antics elements transposables que han divergit massa per poder ser reconeguts com a tals[2]. Els elements repetitius són de gran utilitat alhora d'estudiar el registre fòssil i

obtenir pistes al voltant d'esdeveniments i forces evolutives. Com a marcadors passius, serveixen com a experiments que ens proporciona la natura per estudiar processos de mutació i selecció. Com a elements actius, les repeticions han modelat el genoma causant reordenaments ectòpics (no al·lèlics), creant gens completament nous, modificant i reorganitzant gens ja existents, i variant el contingut global de GC. També han ajudat a conèixer millor l'estructura dels cromosomes i han servit d'eines per estudis de genètica mèdica i de genètica de poblacions.

D'aquesta manera, l'anàlisi de la seqüència crua del genoma és una primera aproximació per tenir una visió del nostre genoma. Ara sabem fins a quin punt ha estat colonitzada per elements de DNA parasítics. En el passat, aquests elements van proliferar de manera massiva, jugant un paper important en la modelació de l'evolució del que és ara el genoma humà actual. Bona part d'aquest DNA no codificant està compost per seqüències que es van repetint al llarg del genoma, els anomenats elements repetitius, classificables com a **repeticions disperses (apartat 1.2)** o **repeticions en tàndem (apartat 1.3)**[10].

1.2 Repeticions disperses o *"interspersed repeats"*

Aquesta mena de repeticions són elements genòmics derivats de transposons, és a dir seqüències de DNA que es poden desplaçar a diferents posicions del genoma. La majoria d'elements transposables en mamífers es poden classificar en 4 tipus, dels quals 3 es transposen a través de intermediaris de RNA i un es transposa directament com a DNA (Figura 1.2).

1.2.1 Elements transposables

LINEs (*"long interspersed nuclear elements"*)

Els LINEs són un dels elements més antics dels genomes eucariotes. En humans, aquests transposons s'estenen 6 kb aproximadament, tenen

un promotor de la polimerasa II i contenen dues pautes obertes de lectura (ORFs). Una de les pautes codifica per una proteïna que té la capacitat d'unió a ARN de cadena senzilla, l'altra té activitat de transcriptasa reversa i d'endonucleasa, capacitant als LINEs a copiar-se a ells mateixos i a d'altres elements. La transcripció reversa sovint falla alhora de processar l'extrem 5', fet que provoca insercions truncades i no funcionals. A més, la majoria de repeticions produïdes per LINEs són curtes, amb un tamany mitjà de 900 pb per totes les còpies dels LINE1, i un tamany mitjà de 1.070 bp per les còpies actualment actives dels LINE1 (L1Hs). Els llocs d'inserció estan flanquejats per un repeticions de 7-20 pb. Sembla ser que la maquinària dels LINE és la responsable per la majoria de la transcripció reversa del genoma, incloent la retrotransposició dels SINEs no-autònoms[11] i de la creació de pseudogens processats[12,13]. En el genoma trobem tres famílies de LINEs: LINE1, LINE2 i LINE3. D'aquests, només els LINE1 romanen actius.

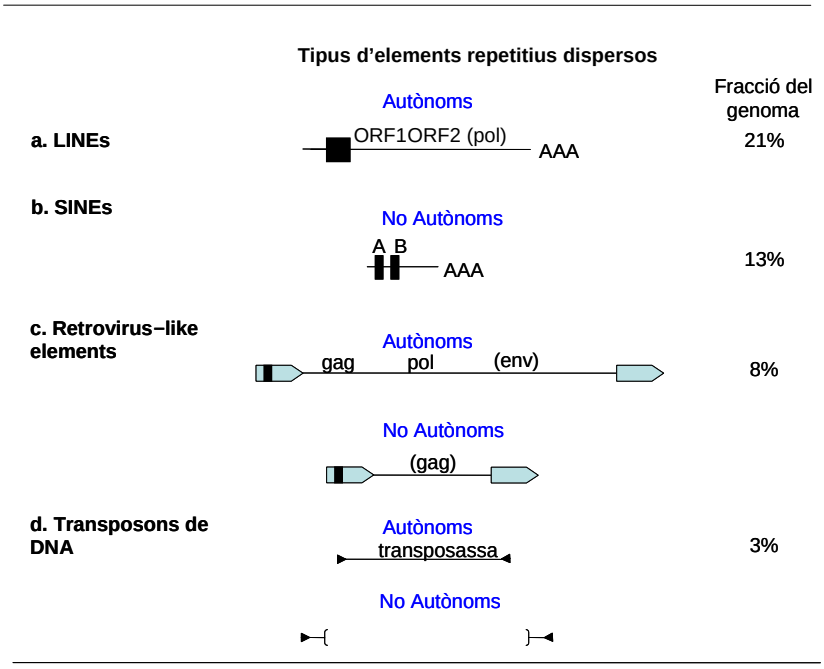

Figura 1.2 Estructura dels diferents tipus d'elements repetitius dispersos i percentatge que representen del total del genoma humà. Adaptada de "International Human Genome Sequencing Consortium".

SINEs ("*short interspersed nuclear elements*")

Els SINEs són elements curts (100-400 pb), contenen un promotor de polimerasa III a l'interior de la seva seqüència i no codifiquen per cap proteïna. Són transposons no autònoms ja que es pensa que usen la maquinària dels LINE per la transposició. La majoria dels SINE comparteixen l'extrem 3' amb un element LINE contigu[11]. La regió promotora de tots els elements SINE coneguts deriva de tRNAs, amb l'excepció d'una família monofilètica de SINEs que deriven del senyal de reconeixement 7SL. Aquesta família, que no comparteix l'extrem 3' amb cap LINE, engloba l'única família activa de SINEs del genoma humà: els elements Alu. El genoma humà conté tres famílies de SINEs: els Alus, els MIR inactius i els Ther2/MIR3.

LTRs ("*long terminal repeat*")

Els retrotransposons LTRs estan flanquejats per repeticions llargues disposades de manera directe i contenen tots els elements necessaris per a la regulació de la transcripció. Els elements autònoms (retrotransposons) contenen els gens gag i pol, que codifiquen per una proteasa, una transcriptasa reversa, una RNAse H i una integrasa. Els retrovirus exògens semblen provenir de retrovirus endògens per adquisició d'un gen d'embolcall cel·lular (env)[14]. La transposició ocorre mitjançant el mecanisme retrovíric on la transcripció reversa té lloc en una *virus-like particle* localitzada al citoplasma. Tot i que existeixen una gran varietat de LTRs, només els retrovirus endògens específics de vertebrats (ERVs) semblen haver estat actius en els genomes de mamífers. Els retrovirus de mamífers s'engloben en tres classes (I-III), cadascuna conté famílies d'orígens independents. Un 85% dels retrotransposons LTRs han esdevingut "fòssils" formats per un sol LTR, on la seqüència interna s'ha perdut per processos de recombinació homòloga entre els LTRs flanquejants.

Transposons de DNA

Els transposons de DNA s'assemblen als transposons bacterians, tenen repeticions terminals invertides i codifiquen per una transposasa que s'uneix prop de les repeticions invertides i permeten la mobilització a través d'un mecanisme de "tallar i enganxar". El genoma humà conté almenys set classes de transposons de DNA d'orígens independents[15]. Els retroposons de DNA solen tenir una vida mitja curta. Per tal de sobreviure, els transposons de DNA s'han de mobilitzar per transferència horitzontal a genomes verges, i existeixen evidències considerables d'aquesta transferència[16-18].

1.2.2 Pseudogens processats

Es tracta de còpies parcials inactives retrotransposades de gens cel·lulars i inclouen gens que codifiquen tant per proteïna, com per RNAs estructurals.

1.2.3 "*Simple sequence repeats*" (SSRs)

Són repeticions directes de petit tamany com (A)n, (CA)n o (CGG)n i segons el número de repeticions es classifiquen en:

a. Microsatèl·lits: n=1-13 bases

b. Minisatèl·lits: n=14-500 bases

Aquest tipus de repeticions sembla ser que apareixen degut a l'"slippage", és a dir, són errors que produeix la polimerasa durant el procés de replicació del DNA, que consisteix en un desplaçament sobre la cadena motlle[19,20] .

El tipus de repeticions més freqüents són AC i AT, mentre que les repeticions de trinucleòtids són molt menys freqüents que les de dinucleòtids[21].

L'elevat grau de polimorfisme dels SSRs fa que siguin elements molt útils com a marcadors genètics i s'han utilitzat per a mapar diverses malalties humanes[22].

1.2.4 Duplicacions segmentàries (DSs)

Consisteixen en blocs de >1 kb que comparteixen més del 90% d'identitat de seqüència i que han estat copiats d'una regió del genoma a una altra[23]. Degut a la seva rellevància en el treball dut a terme, aquest tipus d'elements repetitius seran tractats a part en el capítol 2.

1.3 Blocs de seqüència repetits en tàndem

A la majoria de regions del genoma humà es tolera la presència d'elements repetitius, i en alguns lloc concrets, prop dels finals dels cromosomes, o prop de les constriccions dels cromosomes, anomenats telòmers i centròmers respectivament, s'agrupen formant fragments grans. Així doncs, els blocs de seqüència repetits en tàndem els trobem als centròmers, telòmers, als braços curts dels cromosomes acrocèntrics i als clústers de gens ribosomals. Aquestes repeticions també poden formar part de l'evolució de noves funcions i actuar com a loci de reordenament.

1.4 Una nova era per la seqüenciació a gran escala

Al 1977 Fred Sanger i Alan Coulson van publicar dos articles metodològics per a la seqüenciació de DNA[24] i van transformar la biologia generant una eina per desxifrar gens i més tard genomes sencers. El mètode va millorar considerablement les tècniques primerenques de seqüenciació de DNA desenvolupades per Maxam i

Gilbert publicades el mateix any[25]. Els avantatges consistien en reduir la manipulació de reactius tòxics. D'aquesta manera, el mètode inventat per Sanger es va convertir en la tècnica utilitzada per seqüenciar DNA durant 30 anys.

Amb l'objectiu de seqüenciar el genoma humà, la necessitat de crear mètodes de seqüenciació de DNA a gran escala va augmentar enormement, produint desenvolupaments com l'electroforesi capil·lar automatitzada. D'aquesta manera s'han creat centres sencers de seqüenciació on s'hi agrupen centenars de seqüenciadors. Nous instruments basats en noves tecnologies estan sent explorats arreu del món per desenvolupar centres de seqüenciació ràpids i econòmics. Les limitacions de la llargada dels fragments llegits, les errades de lectura i els algoritmes utilitzats, entre d'altres, són el principals problemes que calen optimitzar.

Al 2008, s'ha dut a terme la seqüenciació a nivell de tot el genoma utilitzant plataformes antigues i noves. En el Baylor College of Medicine es va seqüenciar el genoma de James Watson (premi Nobel 1965), amb la tecnologia **454 Life Sciences** (Roche), per un milió de dòlars[26]. Es tracta d'una tecnologia que combina la lligació dels fragments de DNA inclosos individualment en una gota d'emulsió. S'amplifica cada fragment individualment dins la gota i es seqüencïa mitjançant piroseqüenciació.

El genoma de dos altres individus ha estat també seqüenciat,: el de Craig Venter[8], en el institut fundat per ell, i el d'un individu xinés, al Beijing Genomics Institute. El Craig Venter Institute va utilitzar la tecnologia **Sanger** per seqüenciar el DNA de Venter, que té un cost aproximat de 70 milions de dòlars i s'ha tardat diversos anys.

A la Universitat de Yale, amb la col·laboració de la tecnologia 454 Life Sciences, han combinat la tecnologia de seqüenciació 454 amb el "paired-end mapping" per tal de detectar la variabilitat estructural al genoma

humà en un estudi publicat a l'octubre del 2007 a Science[27], i van detectar els punts de trencament d'inversions que s'haurien perdut utilitzant hibridació genòmica comparada (CGH) i fosmid-end sequencing.

Amb l'avantatge de tenir lectures curtes que comporta **Solexa** (Illumina), en un treball publicat a Cell al 2007, es va obtenir el perfil de modificacions a histones per immunoprecipitació de proteïnes[28].

Finalment l'última tecnologia desenvolupada va veure la llum a l'octubre del 2007, es tracta de **SOLiD** (Applied Biosystems), està basada en la lligació dels fragments i té una lectura final molt fidel.

Totes aquestes noves tecnologies no tan sols faciliten la seqüenciació a nivell genòmic, si no que a més són útils per mesurar el perfil de mRNAs, small RNAs, llocs d'unió a factors de transcripció, l'estructura de la cromatina i l'estat de metilació de DNA. D'aquesta manera ara és possible mesurar els diferents perfils d'un àcid nucleic pel mateix cost o inferior que el de la hibridació en microarrays. Així, podem dir que la nova era de seqüenciació que ve en camí és la era de la seqüenciació de la genòmica funcional.

Gràcies a tots aquests avenços tecnològics tenim un coneixement cada cop més acurat del nostre genoma, del que ens diferencia entre els diferents individus i del que ens diferencia de la resta d'espècies. En aquest terreny, el de l'evolució genòmica, tenen una importància cabdal els processos de duplicació que utilitza la natura per donar plasticitat als genomes.

2. Evolució i duplicacions

Les comparacions a nivell genòmic entre espècies distants són essencials per entendre el procés evolutiu. Però la comparació entre genomes d'espècies properes és encara més important. Els fragments de DNA que codifiquen per alguna funció tenen una tendència més marcada a retenir la mateixa seqüència durant l'evolució que no pas els fragments no funcionals. D'aquesta manera els segments que estan conservats entre espècies tenen més punts de tenir funcions importants. Per aquest motiu és òptim comparar espècies amb fisiologia i comportament similars però que els seus genomes hagin evolucionat suficientment per a què les seqüències no funcionals hagin tingut temps de divergir i puguem determinar l'estructura dels gens (exons introns). Tot i que les prediccions computacionals aporten estimacions en quant al nombre de gens i l'estructura del genoma, les respostes definitives requereixen de dades experimentals d'expressió gènica, tant en un aspecte temporal com posicional dels productes gènics.

Amb la informació obtinguda de tots els genomes que s'estan seqüenciant s'anirà completant el puzle evolutiu i es podran fer comparacions que aportaran pistes per entendre les diferències entre espècies. Per exemple, estudis de genòmica comparada entre humans i ximpanzés duts a terme fins ara ens han permès entendre que els trets

que ens diferencïen (parla, elaboració del lòbul frontal, dit gros oposable, postura bípeda, pensament abstracte, etc), provenen de canvis subtils en la regulació gènica, eficiència d'esplicing o interaccions proteïna-proteïna[29].

2.1 Complexitat proteica en vertebrats

Només 94 de les 1.278 famílies proteiques del nostre genoma són específiques de vertebrats. Les funcions cel·lulars més elementals com el metabolisme bàsic, la transcripció de DNA a RNA, la traducció de RNA a proteïna o la replicació del DNA, han evolucionat un sol cop i s'han quedat fixades des de l'evolució d'una sola cèl·lula com el llevat o el bacteri[29]. La diferència més gran entre humans i cucs o mosques és la complexitat de les proteïnes: més dominis per proteïna i combinacions noves de dominis. Així, més del 90% dels dominis identificats al proteoma humà també estan presents a la mosca del vinagre i al cuc, però han estat barrejats per crear prop del doble de combinacions en humans.

Trobem doncs que l'evolució dels vertebrats ha necessitat de l'invenció de pocs dominis nous. De les proteïnes humanes predites, el 60% tenen alguna similitud de seqüència amb proteïnes d'altres espècies seqüenciades. En vertebrats, observem l'elaboració i l'aparició *de novo* de dos tipus de gens: aquells específic per les habilitats específiques de vertebrats (com la complexitat neuronal, coagulació de la sang o l'adquisició de resposta immune), i aquells que confereixen un augment de les capacitats generals (com gens de senyalització intra i intercel·lular, desenvolupament, mort cel·lular programada o control de la transcripció gènica).

2.2 Duplicacions segmentàries al genoma humà

La comprensió de la biologia, patologia i evolució de les duplicacions segmentàries (DSs) requereix un estudi detallat d'aquestes regions. La presència i distribució d'aquests segments pot aportar informació evolutiva de processos com la barreja d'exons i l'increment general de diversitat proteica associada a l'augment de dominis proteic per pèptid. Així doncs és important tenir en compte no només la duplicació a gran escala, sinó també esdeveniments més puntuals de duplicació genòmica, com a forces en l'evolució dels genomes dels vertebrats.

Una característica destacable del genoma humà respecte d'altres genomes, és la porció de seqüència genòmica formada per DSs altrament anomenades duplicons, i que correspon aproximadament a un 5.3% del genoma eucromàtic[23,30,31]. La freqüència de DSs en el genoma humà és molt més elevada que no pas en d'altres espècies (veure apartat 2.2.3). Aquestes duplicacions corresponen a fragments de >1 kb de seqüència de DNA presents almenys dues vegades en el genoma i poden contenir gens, pseudogens i regions intergèniques. El seu origen és relativament recent (~40 milions d'anys), i tal i com s'ha esmentat anteriorment comparteixen entre elles un elevat grau d'homologia (>90%) i s'estima que gran part de les DSs són alhora variables en número de còpia. Cal tenir en compte però, que l'assemblatge erroni d'aquestes seqüències amb elevat nivell d'identitat provinents de clons no solapants pot subestimar la freqüència real d'aquests duplicons, sobretot entre aquells fragments que tenen una homologia més elevada[4]. La distribució de les DSs varia molt entre els diferents cromosomes, el cas més extrem és el del cromosoma Y, que conté DSs al llarg del 25% de la seva llargada i inclou blocs de fins a 1.45 Mb amb una identitat de seqüència del 99.97%. A més, moltes regions pericentromèriques i subtelomèriques són riques en DSs disperses i són resultat d'insercions degudes a translocacions. Tot i que avui en dia la majoria de regions que contenen DSs s'han

seqüenciat, aproximadament un 10% cauen en els *"gaps"* que encara queden per completar en l'assemblatge actual[32].

2.2.1 Classificació de les DSs

Atenent a la seva localització, les DSs es poden dividir en dos grups. D'una banda hi ha les DSs intercromosòmiques, que representen un ~2.37% del total del genoma[4,33]. Aquestes DSs són definides com a segments duplicats en cromosomes no homòlegs. La majoria de duplicons intercromosòmics tendeixen a agrupar-se en zones centromèriques i subtelomèriques dels cromosomes[34-37]. El fet que les DSs intercromosòmiques es trobin agrupades pot ser explicat si tenim en compte que el genoma té un mecanisme de control de reparació de dany on els productes de trencament dels cromosomes s'inserten preferentment als pericentròmers i, en menys proporció, a les regions subtelomèriques. Encara que senzillament pot ser que aquestes regions presentin major tolerància a la inserció de seqüències[38].

El segon grup comprèn un ~3.97% del genoma i són les DSs intracromosòmiques, que es troben dins un mateix cromosoma[33]. Aquest grup inclou diversos segments duplicats, que també es coneixen amb el nom de *"low copy repeats"* i estan implicats en reordenaments estructurals recurrents que poden anar associats a malalties genètiques[30,31]. Aquestes regions són doncs de gran interès mèdic, ja que la seva estructura inusual sovint les predisposa a delecionar-se o reordenar-se, fenòmens que comporten conseqüències fenotípiques. Exemples que ho certifiquen són la síndrome de Williams-Beuren a la regió cromosòmica 7q[39], Charcot-Marie-Tooth a la regió 17p[40] o la síndrome de DiGeorge al cromosoma 22q[41], entre d'altres (Taula 2.1).

Taula 2.1 Síndromes i trastorns genòmics, i reordenaments als quals estan associats.

Síndrome/Trastorn	Reordenament Cromosòmic	Síndrome/Trastorn	Reordenament Cromosòmic
Monosomia 1p	del(1)(p36)	Angelman	del(15)(q11.2q13)
Holoprosencefàlia 2	del(2)(p21p21)	Esclerosi tuberosa	del(16)(p13.3)
Nefronoftisi	del(2)(q13q13)	Ronyó poliquístic	del(16)(p13.3p13.3)
Holoprosencefàlia 6	del(2)(q37.1q37.3)	Rubinstein-Tabi	del(16)(p13.3p13.3)
Wolf-Hirshhorn	del(49(p16.3)	Charcot-Marie-Tooth	dup(17)(p12p12)
Cri-du-chat	del(5)(q15.2)	Neuropatia hereditària amb predisposició a paràlisi	del(17)(p12p12)
Sotos	del(5)(q35q35)	Smith-Magenis	del(17)(p11.2p11.2)
Sacthre-Chotzen	del(7)(p21p21)	dup(17)(p11.2p11.2)	dup(17)(p11.2p11.2)
Cefalopolisindactilia de Greig	del(7)(p13p13)	Neurofibromatosi I	del(17)(q11.2q11.2)
Williams-Beuren	del(7)(q11.23q11.23)	Holoprosencefàlia 4	del(18)(p11.3)
Silver-Russell	dup(7)(p12p13)	Allagile	del(20)(p12.2p12.2)
Holoprosencefàlia 3	del(7)(q36)	Holoprosencefàlia 1	del(21)(q22.3)
Kabuki	dup(8)(p22p23.1)	DiGeorge/Velocardiofacial	del(22)(q11.2q11.2)
Langer-Giedion	del(8)(q24.11q24.13)	Hipoplàsia adrenal congènita Reversió de sexe sensitiva a dosi	del(X)(p21p21) dup(X)(p21p21)
Holoprosencefàlia 7	del(9)(q22.3)	Pelizaeus-Merzbacher	del(X)(q22q22) i dup(X)(q22q22)
DiGeorge 2	del(10)(p13)	Microftàlmia amb defectes lineals de la pell	del(X)(p22.31q22.31)
Beckwith-Wiedemann	dup(11)(p15.5p15.5)	Distròfia muscular de Ducheme	del(X)(p21p21)
Tumor de Wilms	del(11)(p12p14)	Hiperglicerolèmia	del(X)(p21p21)
Aniridia genitourinària amb retard mental	del(11)(p12p14)	Ictiosi lligada al X	del(X)(p22.3p22.3)
Potocki-Shaffer	del(11)(p11.2p12)	Kallmann	del(X)(p22.3p22.3)
Noonan	del(12)(q24.1q24.31)	Estatura baixa	del(X)(p22.32)
Retinoblastoma	del(13)(q14q14)	Condroplàsia puntata recessiva lligada al X	del(X)(p22.3p22.3)
Holoprosencefàlia 5	del(13)(q32q32)	Albinisme ocular lligat al X	del(X)(p22.3p22.3)
Prader-Willi	del(15)(q11.2q13) paterna	Azoospèrmia	del(Y)(q11.2)

Per altra banda també s'ha vist que algunes regions que contenen DSs són punts calents evolutius on les seqüències codificants estan sota una forta selecció positiva[42]. L'anàlisi acurat de les DSs era fa un temps impossible ja que la seqüència crua del genoma contenia un elevat grau de duplicacions artefactuals però mica en mica l'assemblatge és cada cop més fidedigne.

2.2.3 Duplicacions segmentàries en els genomes de primats

Comparat amb d'altres mamífers, el genoma dels humans i d'altres primats es troba enriquit en DSs llargues amb elevat nivell d'identitat, degut a diferents processos duplicatius que han tingut lloc durant l'evolució dels primats. D'aquesta manera, han sorgit noves famílies gèniques específiques de primats, així com variabilitat gènica i fenotípica[43].

L'accés a la seqüència genòmica de diferents espècies ha ajudat a entendre la distribució, origen i els mecanismes d'evolució de les DSs de primats. La comparació entre espècies i dins d'una mateixa espècie, indica que les DSs han jugat un paper essencial en l'evolució dels primats, creant nous gens i modulant la variació genètica humana que sembla contribuir de manera important a la susceptibilitat a patir diferents malalties. Mitjançant anàlisis de genòmica comparada s'ha observat que les DSs en primats provenen de processos de duplicació ocorreguts fa aproximadament uns 35-40 milions d'anys, coincidint amb la divergència de les mones dels Nou i el Vell Món. El seu tamany pot arribar a ser d'una megabase[44], tot i que la majoria fan menys de 300 kb[45].

Comparat amb la mitjana en humans i ximpanzés, la seqüència d'altres genomes de mamífers (rata, ratolí i gos), mostren menys proporció de DSs[46-49]. En general les DSs dels mamífers seqüenciats tenen un tamany similar a les DSs humanes, però són més grans que en d'altres eucariotes com el cuc o la mosca[2,33].

Atenent a la seva distribució i organització, les DSs en humans i primats superiors (orangutans, goril·les, ximpanzés i bonobos) es poden classificar en tres categories: pericentromèriques, subtelomèriques i intersticials (Figura 2.1)[43].

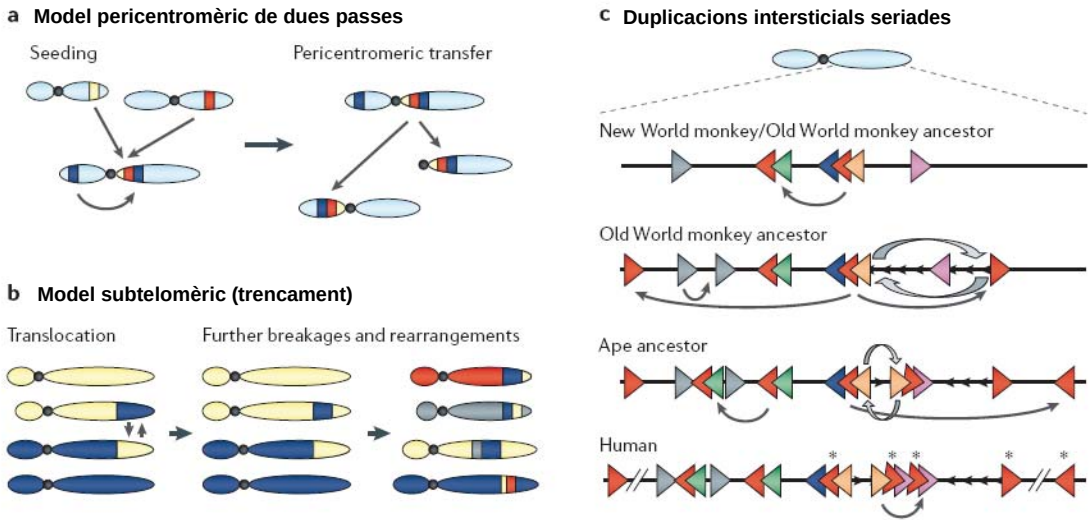

Figura 2.1 Models de formació de duplicacions segmentàries (DSs). a) Procés de dues passes[35] en regions pericentromèriques. DSs diverses (bandes blaves, grogues i vermelles) de duplicons contigus. Blocs de duplicons són transferits per duplicació no homòloga a regions pericentromèriques. b) DSs de regions subtelomèriques semblen provenir de mecanismes de reparació dels trencaments de DNA de doble cadena, donant lloc a translocacions a regions subtelomèriques. Diferents rondes de trencament i reparació donen lloc a un mosaic de duplicons amb DSs (els diferents colors corresponen a segments de diferents cromosomes). c) Les DSs intersticials provenen de múltiples rondes de duplicació seriada, on la seqüència flanquejant és duplicada en posteriors esdeveniments donant lloc a un patró complex de duplicons. L'exemple de la figura correspon als reordenaments de la regió on hi ha el gen *BRCA1*. Els triangles en color corresponen a duplicons, les fletxes negres primes processos duplicatius i les gruixudes, inversions. Els asteriscs indiquen la posició dels gens KIAA0563 que s'han anat expandint[43].

Un estudi recent compara les DSs en humans i ximpanzés[50]. D'aquest anàlisi es conclou que existeix un enriquiment de DSs recents en cromosomes que contenen reordenaments específics d'espècie. Així, set de les nou inversions flanquejades per DSs que hi ha entre humans i ximpanzés, han ocorregut en el llinatge dels ximpanzés (HSA4, HSA5, HSA9, HSA12, HSA15, HSA16 i HSA17)[51] (Figura 2.2).

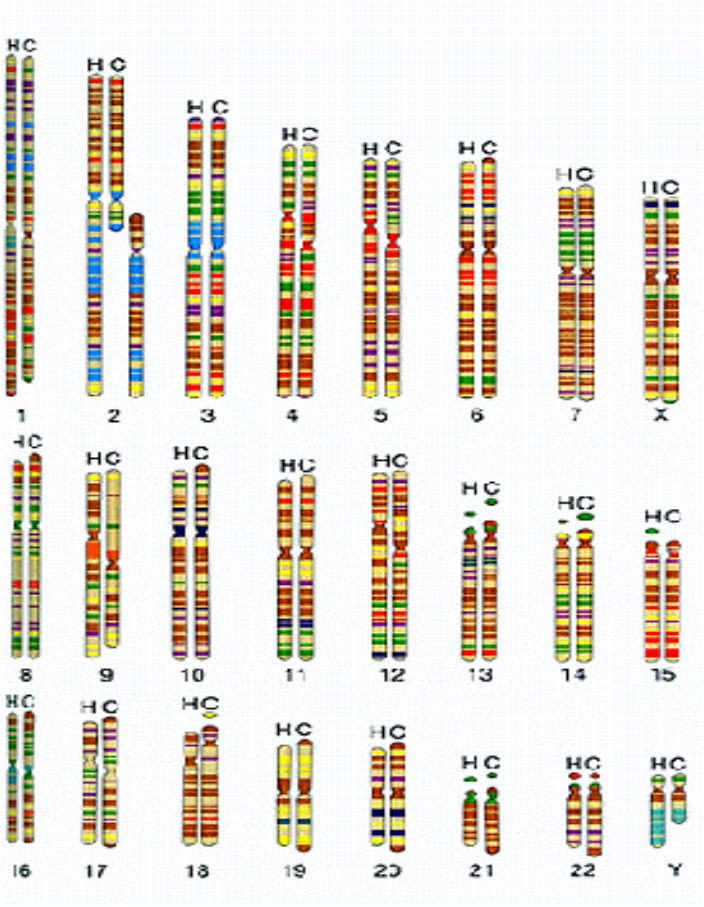

Figura 2.2 Comparació entre cromosomes humans (H) i cromosomes del ximpanzé (C).

2.3 Duplicacions com a font d'evolució genòmica

Les duplicacions segmentàries són una font important d'innovació evolutiva ja que l'aparició d'una duplicació permet que canviï la "nova" còpia del gen duplicat, i d'aquesta manera, es possibilita l'adopció d'una nova funció. També s'ha pogut observar que les DSs han tingut un paper important en la plasticitat del genoma durant l'evolució del primats. Les comparacions dutes a terme entre el genoma humà i el del ximpanzé han permès detectar la presència de DSs en els extrems flanquejants del 70-80% de les inversions i en el 40% de les delecions/duplicacions, tot i que

els mecanismes pels quals els produeixen aquests reordenaments són múltiples[52] .

Les DSs poden originar-se a partir de recombinació homòloga desigual (non-allelic homologous recombination o "NAHR") entre blocs de seqüència amb elevada homologia i crear així noves famílies gèniques en regions cromosòmiques particulars. Un altre mecanisme és el NHEJ *("non-homologous end joining")*que equival a la recombinació no homòloga per unió d'extrems que es produeix habitualment durant la reparació de ruptures de doble cadena del DNA[53]. Aquests mecanismes tant poden crear famílies o superfamílies, com és el cas dels clústers de receptors olfactoris, que en total contenen aproximadament 1000 gens i pseudogens escampats per tot el genoma humà. El destí d'aquests gens duplicats sol ser la desaparició però si es queden fixats es poden produir diverses situacions (Figura 2.3)[54]: a) en el cas que apareguin mutacions que anul·lin la funcionalitat del gen es parla de **no-funcionalització**. Amb el temps, la probabilitat de que s'introdueixin aquest tipus de mutacions augmenta.

b) Una altra possibilitat és l'aparició d'un nou al·lel avantatjós que comporta un guany de funció, em aquest cas parlem de **neofuncionalització**. Aquest procés sol anar acompanyat per una taxa accelerada de canvis d'aminoàcids després de la duplicació d'una de les còpies[55]. Una evidència de neofuncionalització és la que ha patit el gen d'activitat antibacteriana *EDN* en els micos del Vell Món i en hominoids. Primer va patir una sèrie de canvis aminoacídics i posteriorment van ocórrer diverses duplicacions en tàndem del gen progenitor *EDN*[56].

c) Finalment, més que un dels gens duplicats conservi la seva funció i l'altre es degradi o guanyi una nova funció, el que pot passar és que la funció del gen original quedi partida entre els dos duplicats, és quan parlem de **subfuncionalització**[54]. La subfuncionalització també pot donar lloc a la partició temporal, on l'expressió de cada còpia varia al llarg

del desenvolupament, aquest és el cas de les beta globines humanes explicat més endavant.

d) Tampoc és un fenomen estrany que tots dos gens (l'original i la còpia) siguin funcionals i cap dels dos retingui o parteixi la funció original. En aquesta situació parlem de **coevolució** i és especialment prevalent en les vies de senyalització.

Com ja s'ha dit, la recombinació homòloga entre seqüències paràlogues pot donar lloc a reordenaments, incloent duplicacions en tàndem i entre aquestes còpies paràlogues hi pot haver transferència de seqüència de manera no recíproca, el que es coneix com a **conversió gènica**. La conversió gènica homogenitza les seqüències paràlogues, retardant la seva divergència, i emmascarant la seva antiguitat Aquest fenomen permet observar "l'evolució concertada" on els duplicats d'una espècie concreta poden ser molt similars i en canvi divergir entre les diferents espècies[54].

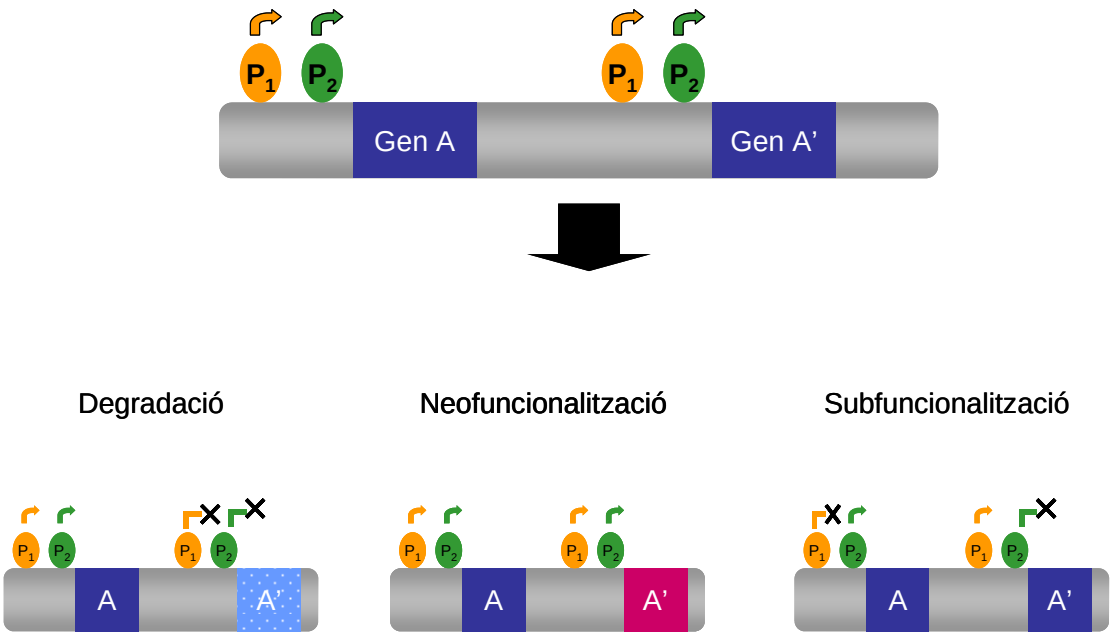

Figura 2.3 Una nova duplicació (gen A') amb dos promotors específics de teixit (P_1 i P_2). Un cop s'ha fixat la nova còpia (passa en una minoria de casos), la còpia A' es pot degradar o bé adoptar una nova funció (neofuncionalització), o bé que l'expressió del gen original A quedi partida en A i A' amb només un dels dos promotors actius en cada còpia (subfuncionalització).

El mecanisme més extrem de duplicació és la que ocorre a nivell de tot un genoma a partir d'un procés de **poliploidització,** pel qual un organisme diploide esdevé tetraploide. La poliploidització és un mecanisme comú en les plantes. En principi la poliploidització produeix una font d'innovació permetent la duplicació i divergència del material duplicat. Ohno va suggerir que el procés de duplicació de genomes sencers (WGD *"whole genome duplication"*) ha jugat un paper clau en l'evolució[57]. Hi ha evidències que mostren que una duplicació de tot el genoma va tenir lloc en un ancestre del llevat[58]. Les WGDs poden ser difícils de detectar ja que només una minoria dels loci duplicats es retenen, de manera que els gens en els segments duplicats no es poden alinear correctament. A més, els gens duplicats poden ser reordenats a posteriori.

Una de les hipòtesis més controvertida sobre l'evolució dels vertebrats està basada en dos processos de WGD al principi de l'aparició del llinatge dels vertebrats, al voltant de l'aparició dels peixos mandibulats fa uns 500 milions d'anys. Alguns autors recolzen aquesta teoria basant-se en el fet que diversos gens humans els trobem en 4 còpies homòlogues, l'exemple més notable són els 4 clústers de gens *Hox* als cromosomes 2, 7, 12 i 17[59,60]. Val a dir però, que amb l'anàlisi actual del genoma humà, de moment no es tenen suficients evidències per confirmar o descartar si aquestes dues rondes de WGD van tenir lloc o no.

2.4 Famílies gèniques i superfamílies

Gran part del DNA funcional del genoma està organitzat en famílies gèniques i superfamílies gèniques. El terme superfamília va ser concebut per descriure les relacions amb un ancestre comú que existeixen entre dos o més famílies gèniques. A mesura que es clonen més gens i es seqüencien, es van descobrint relacions més estretes i antigues entre

superfamílies. Totes aquestes superfamílies han evolucionat a partir de processos de recombinació desigual que han expandit el tamany dels clústers gènics i els mecanismes de transposició han actuat formant regions genòmiques distants amb nous gens o clústers gènics. La primera evidència que mostra que les duplicacions han jugat un paper vital en l'evolució de noves funcions gèniques és l'extensa existència de famílies gèniques.

Ja que gran part del treball realitzat en aquesta tesi doctoral tracta sobre l'aparició i expansió d'una nova família gènica, la família *FAM90A*, a continuació s'explica de manera breu com es poden originar i expandir les famílies gèniques a partir de tres exemples de famílies de gens àmpliament conegudes.

2.4.1 La superfamília de les globines

Un prototipus de superfamília gènica petita és la formada pels gens de les globines. Tots els membres funcionals d'aquesta superfamília juguen un paper en el transport d'oxigen. La superfamília conté tres famílies principals (o branques) representades per les beta-globines, les alfa-globines i la miosina (gen de còpia única). La duplicació i divergència d'aquestes tres branques principals va ocórrer durant l'evolució dels vertebrats. Els productes codificats per aquests gens en dues de les branques, la de les alfa-globines i les beta-globines, s'ajunten per formar el tetràmer de la proteïna funcional, l'hemoglobina, que actua transportant l'oxigen en sang. La mioglobina en canvi transporta l'oxigen en el teixit muscular.

La superfamília de les beta-globines s'ha duplicat per múltiples processos de recombinació desigual i ha divergit en cinc gens funcionals i dos pseudogens, tots ells presents en un sol clúster del cromosoma 7 de ratolí. En el cas de les alfa-globines es van generar tres clústers, un és

funcional durant l'embriogènesi i dos funcionals en l'adult. A part d'aquests clústers d'alfa-globines també hi ha dos gens "alfa-like" no funcionals que s'han transposat en localitzacions disperses dels cromosomes 15 i 17[61]. Finalment, el gen de còpia única mioglobina es troba en el cromosoma 15 i no té cap altre membre relacionat al seu voltant ni en cap altre localització genòmica[62,63]. D'aquesta manera, la superfamília gènica de les globines dóna una visió dels diversos mecanismes utilitzats pel genoma per evolucionar la complexitat estructural i funcional.

2.4.2 La superfamília dels gens Hox

Per altra banda la superfamília de gens *Hox* exemplifica un altre prototipus d'expansió de número de gens. En aquest cas, els processos duplicatius primerencs (que daten abans de la divergència dels vertebrats i els insectes) van crear un clúster de gens relacionats que codifiquen per proteïnes d'unió a DNA que aporten la informació espaial necessària pel desenvolupament de l'embrió. Aquest clúster original s'ha duplicat en massa i s'ha dispersat donant un total de quatre localitzacions cromosòmiques, cadascuna conté de 9 a 11 gens[64]. Algunes insercions i delecions gèniques dins clústers concrets han succeït per recombinació desigual des de les duplicacions en massa, de manera que s'observen diferències en número i tipus de gens entre els diferents clústers.

2.4.3 La superfamília de les Immunoglobulines

Un últim exemple de superfamília gènica és la de les immunoglobulines (Ig). Es tracta d'una extensa família formada per receptors de membrana o solubles implicats en la resposta immune i processos d'interacció cèl·lula-cèl·lula. Aquest grup inclou les pròpies famílies d'Igs, els gens del complexe major d'histocompatibilitat i els gens dels receptors de les cèl·lules T, entre d'altres[65]. En aquest cas tenim gens dispersos i famílies

gèniques, clústers petits, clústers grans, i clústers dins de clústers, agrupats o dispersos. La dispersió ha ocorregut per transposició de clústers sencers en massa. A més, el domini original d'Ig pot estar present una vegada en alguns gens, però també ha estat duplicat de manera intragènica i hi ha membres que contenen dos, tres i fins a quatre dominis Ig en un sol polipèptid. La superfamília d'Igs, que conté centenars i potser milers de gens, il·lustra la manera com l'emergència inicial d'elements genètics versàtils pot ser explotada per les forces de l'evolució gènica donant com a conseqüència un creixement enorme de la complexitat genòmica de l'organisme.

Un nombre limitat de famílies gèniques de gens de còpia múltiple han evolucionat sota una forma especial de pressió selectiva que requereix que tots els membres de la família mantinguin essencialment la mateixa seqüència. En aquests casos, el propòsit de mantenir un número de còpies elevat no és crear variabilitat, si no més aviat poder proporcionar a la cèl·lula una quantitat suficient de producte idèntic en un període de temps curt. Exemples d'aquestes famílies gèniques inclouen els gens que codifiquen per RNA ribosomals i de transferència. També s'inclouen les histones que han de ser produïdes a nivells suficients de proteïna per embolcallar la nova còpia de tot el genoma que és replicat durant la fase S de cada cicle cel·lular.

En aquests casos, hi ha una forta pressió selectiva per mantenir intacte la seqüència de tots els membres de la família perquè tots són utilitzats per produir el mateix producte. En altres paraules, el funcionament òptim de la cèl·lula requereix que els productes de cadascun dels gens individuals sigui directament intercanviable en estructura i funció amb els productes de tota la resta de membres de la mateixa família. Aquesta situació és possible gràcies a **l'evolució concertada**[66]. Així, com ja s'ha dit, a partir de recombinació desigual i la

conversió gènica interal·lèlica juntament amb la selecció per homogeneïtat, tots els membres d'una família gènica es poden mantenir amb pràcticament la mateixa seqüència de DNA. Tot i així, l'evolució concertada també dóna lloc a un augment de la divergència entre famílies gèniques de diferents espècies.

Donada l'aparent importància de la duplicació gènica per l'evolució de noves funcions biològiques a diferents escales de temps evolutives, és important l'estudi de les diferències a nivell de segments duplicats que existeixen entre espècies i entre els nostres parents més propers, els primats superiors.

2.5 Famílies gèniques específiques de primats

El genoma dels primats superiors i dels humans és sorprenentment semblant. El treball dut a terme per Fortna et al. identifica un 3% dels gens humans com a gens que han patit canvis en número de còpia específics de llinatge en humans i primats superiors[67]. En aquest estudi detecten 140 gens amb canvis en número de còpia específics d'humans que s'espera que siguin investigats en els propers anys.

Una qüestió cabdal en la biologia evolutiva des de fa dècades és esbrinar què ens diferencia a nivell genètic de la resta de primats i en concret dels nostres parents més propers, els primats superiors. Les diferències genètiques poden ser a diferents nivells, des de grans alteracions en l'arquitectura citogenètica, reordenaments cromosòmics, duplicació de famílies gèniques, aparició i desaparició de gens i diferències en la transcripció gènica i l'splicing de mRNAs[68].

A nivell cariotípic s'ha pogut determinar que 18 dels 23 parells de cromosomes humans són molt semblants en els primats superiors a excepció dels cromosomes X, 4, 9 i 12 que han patit diversos reordenaments després de l'especiació dels primats[69]. D'altra banda també s'ha descrit en el mateix estudi l'origen del cromosoma 2 humà

provinent de la fusió de dos cromosomes, i la translocació recíproca en el goril·la entre els cromosomes homòlegs als cromosomes humans 15 i 17.

Una de les característiques genòmiques que ha tingut gran importància alhora de modelar els genomes de primats és la presència de les DSs, com ja s'ha esmentat. Per exemple, les DSs han jugat un paper important en l'adquisició de trets fenotípics com la visió tricromàtica, fet que va representar una millora considerable alhora de detectar els fruits grocs/vermells o fulles[70]. Aquest esdeveniment fou fonamental i va ocórrer fa uns 35 milions d'anys en un ancestre comú de les mones del Vell Món i els primats superiors[71]. Així els Cattarhini i els Hominoids (veure Figura 2.4 on es mostren les distàncies evolutives) contenen un gen autosòmic que codifica per la detecció del color blau i alhora tenen duplicats al cromosoma X que codifiquen per la capacitat de distingir els colors verd i vermell[72]. I són precisament aquests gens lligats al cromosoma X que es van originar a partir d'una duplicació posterior a la divergència de les mones del Nou Món. Aquests primats tenen només un sol gen al cromosoma X i un altre gen pel pigment del color (opsina) a l'autosoma[73].

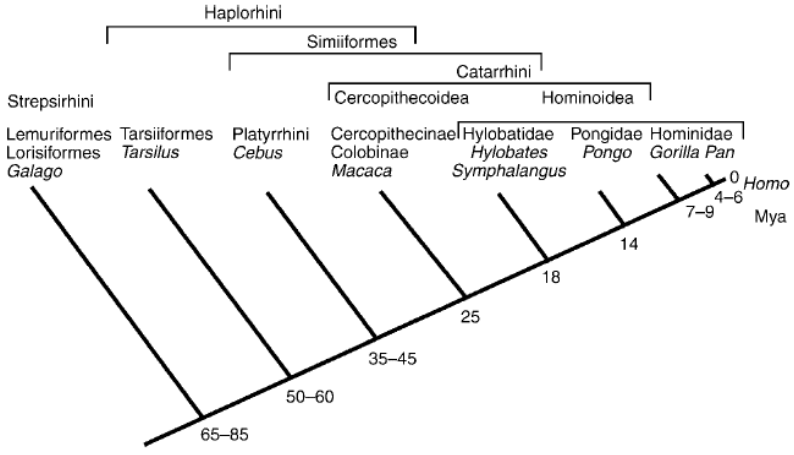

Figura 2.4 Relacions filogenètiques entre primats (Goodman 1999), Mya=milions d'anys.

La majoria de gens que codifiquen per proteïnes involucrades en la formació d'esperma, que tenen un paper en l'aparellament mascle-femella o les cèl·lules germinals humanes, són exemples de gens que han evolucionat ràpidament en els llinatges dels primats. En general, aquesta

evolució ràpida es troba sota selecció darwiniana positiva i probablement ha participat en l'evolució recent dels trets reproductius en mascles i femelles. La barreja d'exons ("*exon shuffling*") a traves de recombinació, la retrotransposició mediada per elements LINE-1 i l'expansió de DSs són mecanismes que han aportat en gran mesura gens quimèrics que codifiquen per mRNAs en primats. Un exemple de fusió gènica el trobem en el cas del gen *Kua-UEV* que codifica per una proteïna citoplasmàtica[74,75].

Tanmateix, també hi ha exemples de pèrdua de funció gènica, com en el cas dels receptors olfactoris que s'han anat pseudogenitzant en diferents llinatges antropoides[76]. Uns altres elements que juguen un paper important en l'aparició de nous gens són els LTRs de retrovirus endògens humans (HERVS), que per exemple, han permès l'expressió en primats del gen *AMY* que codifica per l'enzim amilasa[77,78].

Per altra banda tenim diversos exemples on els elements Alu inserits en regions promotores han modelat el control de l'expressió gènica dels gens del voltant[79,80].

Un exemple interessant en la creació de gens específics de primats el trobem en el cas de *PMCHL1/PMCHL2,* còpies truncades del gen *MCH ("melanin concentrating hormone") ,* que són gens que codifiquen per un neuropèptid que actua com a neurotransmissor i regulen un gran ventall de funcions[81]. Diversos estudis han aprofundit en l'estructura d'aquests gens i s'ha pogut postular que al llarg de l'evolució dels primats han patit un gran nombre de combinacions que han creat llocs d'splicing *de novo.* No tant sols això, si no que la regulació d'aquests gens es veu afectada per la retrotransposició de trànscripts antisentit i per processos de truncament que han creat una nova pauta de lectura.

Un treball dut a terme per Eichler et al. ha permès descriure l'important paper que han jugat les DSs en l'evolució dels genomes de primats amb la descripció de la DS anomenada "LCR 16a". Aquest estudi

revela un procés d'expansió d'aquests duplicons en primats superiors, així com una hipervariabilitat entre els paràlegs en humans[42].

Estudis de genòmica comparada han permès postular que l'aparició d'uns pocs gens nous en humans estan associats a trets específics d'humans com el bipedisme o funcions cognitives complexes[82], i que aquestes diferències s'haurien de trobar a nivell de regulació específica de teixit, a nivell del desenvolupament i a nivell d'expressió en les diverses espècies de primats[83,84], donat que el nivell d'identitat de seqüència entre humans i primats superiors a les zones codificants és del 99%[85].

Finalment cal mencionar que una font important de vairabilitat genètica en humans, ximpanzés, macacs, i ratolins són les variants estructurals, i variants en número de còpia (CNVs) tractades al següent capítol. De ben segur part d'aquestes variants estructurals contribueixen a l'ampli ventall fenotípic i genotípic entre individus.

3. Variants estructurals: CNVs i Inversions

Des de l'inici dels estudis citogenètics, es va observar que guanys o pèrdues de cromosomes sencers o alteracions visibles al microscopi que involucraven segments grans de cromosomes es trobaven de manera recurrent associats a diversos trastorns comuns com ara la síndrome de Down[86]. Cap al 1990, també es va fer patent que els guanys o pèrdues de fragments específics de certs cromosomes eren la causa recurrent de moltes altres malalties genètiques menys comuns. El que durant molts i molts anys ha passat desapercebut ha estat la variació en número de còpia del DNA i com aquesta ha contribuït al ventall de variació a nivell de seqüència i de fenotip entre individus aparentment sans.

Al 2004, dos estudis rellevants[87,88] van contribuir enormement a caracteritzar aquesta variació en número de còpia a gran escala en les poblacions humanes i a entendre quin paper juguen en la variabilitat fenotípica i en la susceptibilitat a patir diferents malalties. En paral·lel i utilitzant tecnologies semblants, s'han dedicat molts esforços a examinar el rol d'aquestes variants estructurals de gran tamany en malalties esporàdiques. Aquesta variabilitat estructural tant inclou insercions, duplicacions i delecions de DNA, conegudes com a variants en número de còpia o CNVs (copy number variants), com reordenaments cromosòmics balancejats com ara les inversions o translocacions, anomenant-se de manera genèrica variants estructurals.

Així l'any 2006 es va estimar que les CNVs representen entre 5 Mb i 24 Mb[89] de diferències genètiques entre individus, mentre que les variacions formades per un únic nucleòtid (single nucleotide polymorphisms o SNPs) només suposen 2.5 Mb. Les dades més recents semblen indicar que cada individu presenta ~1000 CNVs per genoma diploide[90]. Actualment hi ha més de 6000 CNVs descrites a la base de dades de variants genòmiques (http://projects.tcga.ca/variation), però moltes d'elles no han estat validades per diferents metodologies.

Així doncs, entendre diferències en número de còpia en les diferents poblacions humanes pot aportar grans avenços en el tractament de diverses malalties (Taula 3.1).

Taula 3.1 CNVs associats a malaltia i els gens implicats.(Adaptat d'Estivill i Armengol 2007)[91]

Malaltia amb CNV associat	Gen	Referència
VIH/ susceptibilitat a SIDA	*CCL3L1*	Gonzalez E *et al* , Science (2005)
Artritis reumatoide i Diabetis tipus I	*CCL3L1*	McKinney C *et al,* Ann. Rheum. Dis. (2007)
Lupus eritematós sistèmic (LES)	*C4A/C4B*	Yang Y *et al* Curr. Dir. Autoimmun. (2004)
LES/ Poliangiïtis microscòpica Granulomatosi de Wegener	*FCGR3B*	Atiman TJ *et al,* Nature (2006) Fanciulli M *et al,* Nat. Genet. (2007)
Malaltia de Crhon	*DEFB4*	Fellermann K *et al* , Am. J. Hum. Genet. (2006)
Trastorn bipolar	*GSK3B*	Lachman HM *et al,* Am. J. Med. Genet. B (2007)
Malaltia de Parkinson d'aparició temprana	*SNCA*	Ibanez P *et al,* Lancet (2004); Chartier-Harlin MC *et a* l, La Singleton AB *et al,* Science (2003)
Malaltia de Parkinson hereditària d'aparició temprana	*APP*	Rovelet-Lecrux A *et al,* Nat. Genet. (2006) Cabrejo L *et al,* Brain (2006)
Pancreatitis hereditària	*PRSS1*	Le Marechal C *et al,* Nat. Genet. (2006)
Trastorns d'aspectre autista	Múltiples	Sebat J *et al,* Science (2007); Weiss L *et al,* N. Engl. J. M Szatmari P *et al,* Nat. Genet. (2007)
Càncer de mama familiar	*MTUS1*	Frank B *et al,* BMC Cancer (2007)
Glaucoma	*FOXC1, FOXC2, FOXQ1*	Chanda B *et al,* Hum. Mol. Genet. (2008)
Trombocitopènia púrpura idiopàtica	*FCGR2C*	Breunis WB *et al,* Blood (2008)
Esquizofrènia	Múltiples	Walsh T *et al* , Science (2008)

Diverses innovacions tecnològiques (hibridació genòmica comparada "CGH" amb arrays de BACs, arrays d'oligonucleòtids, llibreries de fòsmids o la seqüenciació a gran escala per exemple) han obert la porta a un aspecte fonamental de la variació genòmica humana com a base genètica de malalties. Els mètodes per detectar CNVs a tot el genoma són útils per identificar els factors de risc de patir diverses malalties i superen les limitacions que hi havia amb el tradicional mapatge de gens[92]. La tecnologia dels arrays no és l'única que es pot usar per detectar CNVs i associar-los a malaltia. Altres mètodes com la PCR quantitativa, la MLPA ("*multiple ligation probe amplification*") o la DASH *("dynamic allele-specific hybridization")* també són útils per identificar CNVs tot i que totes tenen les seves limitacions en quant a capacitat i nivell de resolució.

Les aproximacions mitjançant tècniques de seqüenciació, com el mapatge comparant extrems de fòsmids clonats amb la seqüència de referència, són especialment idonis per detectar no només CNVs, si no reordenaments estructurals com inversions i translocacions[27,92,93]. Dissortadament, utilitzar els sistemes actuals de seqüenciació per capil·laritat resulten extremadament cars per seqüenciar els extrems d'una llibreria de fòsmids de tot un genoma. Tot i així, la nova generació de tecnologies de seqüenciació esmentades al primer apartat de la introducció, sembla que economitzaran el cost i possibilitaran aquesta mena d'estudis a gran escala en un futur no molt llunyà. Exemples que ho corroboren són la seqüenciació del genoma del Dr. James Watson (premi Nobel 1962) i el treball basat en la seqüenciació de vuit individus publicat aquest any per Kidd et *al.*[94].

3.1 CNVs i Duplicacions Segmentàries

Cal esmentar que existeix una estreta relació entre les seqüències duplicades a l'assemblatge de referència i la variació en número de còpia al genoma humà. Més de la meitat de nucleòtids anotats dins de les DSs (http://humanparalogy.gs.whashington.edu)[3] es solapen amb CNVs. La densitat mitjana de DSs en la majoria de regions amb CNVs del genoma és del ~25% , mentre que la densitat mitjana si es té en compte tot el genoma és del 4-5% i en zones pobres en CNVs és del 2-3%[3]. En altres paraules, les regions duplicades del genoma contenen de 4 a 10 vegades més CNVs[87,89,95,96]. De fet, moltes DSs representen els al·lels de CNVs de l'assemblatge de referència. Tot i així, només la meitat de CNVs aproximadament colocalitzen amb DSs[89,95].

Per altra banda, hi ha una gran relació entre les regions genòmiques que contenen CNVs i el contingut de gens. Les regions riques en gens tenen tendència a ser riques en CNVs i a l'inrevés[3,33,97]. També s'ha vist que el tipus de gens que es troben en CNVs associades a DSs estan preferentment involucrats en percepció sensorial com els receptors olfactoris i gens involucrats en la resposta immune[37,96,98,99]. Sovint s'ha suggerit, i en molts casos s'ha constatat, que aquests gens estan subjectes a canvis adaptatius ràpids al llarg de l'evolució dels mamífers[100]. També hi ha estudis que han trobat signes de selecció positiva a nivell de canvis d'aminoàcids en famílies gèniques com morpheus[42], RanBP2[101] i DUF1220[102] duplicades recentment. En canvi, les CNVs que no solapen amb DSs estan enriquides en gens de senyalització que regulen el desenvolupament i el creixement cel·lular[37,96,98,99]. Es tracta de gens que estan involucrats en respostes fisiològiques i en el desenvolupament, i per tant sovint desperten interès com a dianes terapèutiques de malalties com el càncer.

3.2 Inversions

Com s'ha explicat en l'apartat anterior el genoma conté gran quantitat de variants estructurals de gran tamany formades per duplicacions i delecions cada cop més ben caracteritzades. En el cas de les inversions polimòrfiques malauradament aquest coneixement és més pobre. El principal motiu és la manca de tècniques a gran escala per poder detectar inversions ja que es tracta de reordenaments equilibrats i els seus punts de trencament sovint cauen en regions complexes de DSs[103,104].

S'han proposat diversos mecanismes per explicar com la presència de *"low copy repeats"* en posició invertida poden causar les inversions al genoma[53,105,106] (Figura 3.1). També s'han pogut detectar parelles de repeticions invertides a prop dels punts de trencament de diferents inversions polimòrfiques[107,108].

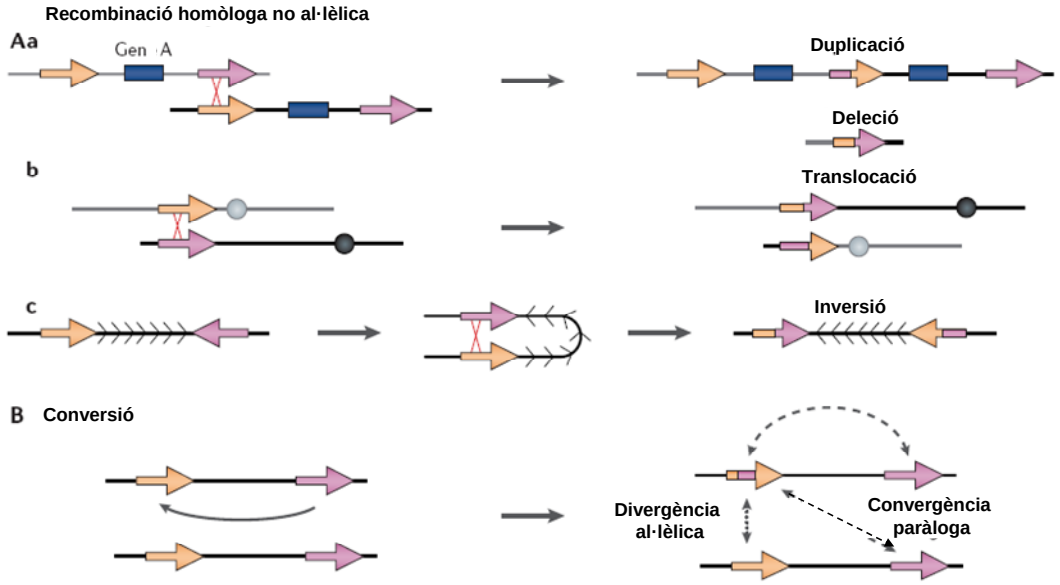

Figura 3.1 A) NAHR entre DSs causa reordenaments depenent de l'orientació de les DSs involucrades. (Aa) Duplicacions en tàndem i delecions com a resultat de NAHR entre seqüències contigües duplicades. (Ab) Les translocacions són el resultat de l'intercanvi entre DSs de cromosomes no homòlegs. (Ac) En Aa i Ab les DSs tenen la mateixa orientació. Quan es tracta de DSs d'orientació oposada, s'originen inversions. B)Conversió gènica entre DSs degut a l'intercanvi de seqüència entre còpies (evolució concertada). Adaptat de Bailey *et al.*[43]

Avui en dia existeixen pocs exemples d'inversions polimòrfiques de les quals s'hagi pogut establir la seva freqüència en una població

determinada. Una d'elles és la inversió de 900 kb del cromosoma 17q21.31[103]. En aquest cas trobem dos al·lels, un de "normal" i un altre en orientació invertida que és present en un 21% dels europeus però és poc comú en africans (6%) i asiàtics (1%). L'haplotip invertit data de fa ~3 milions d'anys però hi ha poques evidències de que hagi recombinat, d'aquesta manera ha pogut donar lloc a un haplotip diferent i a un gran desequilibri de lligament de la regió en la població d'origen europeu. A més, l'anàlisi fet a població islandesa mostra que les dones portadores de l'haplotip invertit tenen de mitjana més fills que no pas les no portadores, de manera que la conformació invertida estaria sota pressió selectiva.

L' estudi realitzat per Tuzun et al.[93], utilitzant la seqüenciació dels extrems de fòsmids d'una llibreria construïda a partir del DNA d'una dona d'origen Africà i comparant-la amb l'assemblatge de referència, va poder detectar 56 inversions putatives. Tot i que aquest mètode és eficaç alhora d'identificar inversions, no és viable utilitzar-lo per establir la freqüència d'aquests reordenaments en una població, ja que requereix múltiples re-seqüenciacions i comporta una elevada despesa econòmica. Una altre opció per detectar fragments invertits del genoma és utilitzar mètodes indirectes com la comparació per FISH i PCR entre el genoma humà i el del ximpanzé[107] (szamalek 2006). Un treball dut a terme utilitzant aquesta aproximació va tenir en compte 23 regions de les quals tres, a les regions cromosòmiques 7p22, 7q11 i 16q24, van resultar ser polimòrfiques en humans. Un altre estudi publicat al 2007 va utilitzar un mètode estadístic basat en la comparació dels blocs de desequilibri de lligament en individus HapMap respecte l'assemblatge de referència. Així van poder detectar 176 regions candidates a patir inversions tot i que caldran futures aproximacions experimentals per validar-les.

Diverses inversions han estat associades directament a fenotips deleteris degut a disrupcions d'elements reguladors crítics o de seqüències gèniques, o a predisposició a causar reordenaments en la

descendència, es tracta de trastorns genòmics com l'hemofília A[109], la síndrome de Prader Willi o d'Angelman, la síndrome de Williams-Beuren[39,110] o la síndrome de Hunter[111]. En canvi d'altres inversions semblen ser neutrals, seria el cas de les que trobem als cromosomes 9, 4p16[104] o 8p23[105]. Algunes d'elles es tracten amb més deteniment més endavant.

3.2.1 Classificació de les inversions

Les inversions es poden classificar en dos tipus atenent a les posicions relatives dels punts de trencament respecte del centròmer. Les inversions pericèntriques tenen un punt de trencament a cada braç del cromosoma i les inversions paracèntriques tenen els dos punts de trencament en el mateix braç cromosòmic. S'han trobat tant inversions pericèntriques com paracèntriques en els 22 autosomes humans[112]. La majoria d'elles es pensa que tenen punts de trencament únics a nivell de resolució citogenètica, però s'han identificat algunes inversions que presenten diferents punts de trencament de manera recurrent[113-115]. Aquestes últimes poden haver sorgit a través de múltiples reordenaments independents que van donar lloc a les diferents inversions o bé es van transmetre de manera idèntica a la descendència a partir d'un únic (o un número petit) d'ancestres comuns[116].

3.2.2 Recombinació en individus heterozigots per una inversió

Actualment se sap que les inversions cromosòmiques són més freqüents del que s'havia vist en els estudis citogenètics clàssics. Moltes de les inversions paracèntriques críptiques (no detectables citogenèticament) es troben flanquejades per DSs i, com ja s'ha explicat, poden causar malalties mendelianes per disrupció de gens als punts de trencament o bé ser presents en població general com a polimorfismes. En aquest últim cas, per la impossibilitat d'un aparellament correcte de la regió invertida

durant la meiosi, en els individus heterozigots existeix una predisposició a produir reordenaments desequilibrats com els reordenaments inv dup (inversió i duplicació del fragment invertit) o simplement delecions i duplicacions.

Així podem dir que les inversions no tenen conseqüències fenotípiques directes, excepte quan un o els dos punts de trencament disrupcionen un gen[117-121], però si poden estar associades a problemes reproductius o a un risc incrementat de produir gàmetes no balancejats a través d'entrecreuaments entre dos cromosomes homòlegs, l'un normal i l'altre invertit. Un número imparell de recombinacions meiòtiques dins una inversió pericèntrica dóna lloc a una deleció i una duplicació, i el mateix esdeveniment en una inversió paracèntrica dóna lloc a un cromosoma dicèntric i un fragment acrocèntric, que normalment no són viables i per tant la transmissió d'un recombinant no balancejat a la progènie és exclosa[116]. La importància d'aquest factor de susceptibilitat ha estat ben caracteritzat per alguns trastorns genòmics que afecten al cromosoma 8 i serveixen com a possible model per explicar les bases genètiques d'altres reordenaments cromosòmics recurrents. Per altra banda, estudis de segregació en famílies i d'anàlisi de cromosomes per FISH en esperma, mostren que hi ha casos on la recombinació meiòtica queda exclosa dins el fragment invertit[122] i en canvi augmenta en la resta del cromosoma[123]. Així, tot i que es desconeix el mecanisme que hi ha al darrera, les inversions en heterozigosi poden afectar la recombinació tant dins del fragment invertit com a la resta del cromosoma. Sembla ser que el factor que més influeix en l'existència o no de recombinació és el tamany de la inversió[124].

Les inversions humanes més comuns inclouen blocs centromèrics d'heterocromatina dels cromosomes 1, 9 i 16 i d'heterocromatina del braç llarg del cromosoma Y[112]. El més probables és que les inversions heterocromàtiques siguin conseqüència d'alteracions en la quantitat i la

distribució d'heterocromatina, i per tant, són considerades variants sense efectes fenotípics. També hi ha inversions polimòrfiques pericèntriques, que no engloben heterocromatina centromèrica, que afecten als cromosomes 2, 3, 5 i 10 que són comuns i no tenen conseqüències fenotípiques i per tant són considerades variants[112].

3.2.3 Inversions críptiques associades a DSs

En els últims 15 anys s'han descrit diverses inversions paracèntriques críptiques i la gran majoria estan mediades per LCRs o duplicons de >10 kb que actuen com a substrats de recombinació homòloga no al·lèlica (NAHR). Un cas interessant és el de la inversió que disrupciona el factor VIII de coagulació en els pacients amb hemofilia A severa[109]. La disrupció d'aquest gen és deguda a la NAHR entre l'intró 22 del gen A i una o ambdues còpies del mateix gen que es troben en direcció oposada a una distància de ~500 kb. Aquest mecanisme és el causant de la malaltia en un 50% dels pacients. El mateix passa en un 13% dels pacients amb la síndrome de Hunter, on la NAHR entre el gen *IDS* i el seu pseudogen localitzat a 900 kb dóna lloc a la disrupció d'aquest gen[111].

Altres inversions críptiques recurrents mediades per duplicons s'han vist en individus no afectes. Aquest és el cas de la inversió a Xq28, a la regió dels gens emerina/filamina[125]. Aquesta inversió es troba en heterozigosi en el 33% de les dones i en hemizigosi en el 19% dels homes i és mediada per duplicons que tenen un 99% d'identitat de seqüència que flanquegen els gens de l'emerina i la filamina. Tot i que aquest elevat nivell d'identitat podria suggerir un origen molt recent, un estudi recent realitzat en 116 espècies diferents mostra que aquests duplicons són presents en totes elles, indicant que el seu origen es remunta a més de 100 milions d'anys[126].Tot i que els portadors de la inversió són fenotípicament normals, s'ha hipotetitzat que algunes delecions de

l'emerina associades a la distròfia muscular d'Emery-Dreifuss són el resultat de reordenaments deguts a la inversió[127].

Una altra inversió críptica recurrent i benigna s'ha descrit a la regió que conté el locus *NPHP* al cromosoma 2q13. La deleció en homozigosis de 290 kb que conté aquest locus, és mediada per duplicons amb la mateixa orientació i dóna lloc a una malaltia renal coneguda com a nefronoptisi de tipus 1. Aquests mateixos duplicons quan es troben en orientació oposada són els causants d'una inversió de 500 kb que es troba en homozigosis en un 1.3% de la població[128].

Finalment en el cromosoma Y trobem una inversió de 3 Mb que propicia la translocació *PRKX/PRKY* que explicaria la majoria d'individus mascles XX i algunes femelles XY[129]. Aquest seria un clar exemple que demostra que aquesta inversió polimòrfica del cromosoma Y no és neutral. És més, sembla factible que d'altres inversions críptiques siguin les responsables del seguit de delecions i duplicacions complexes mediades per diferents duplicons associats al locus AZFc[130,131].

3.2.4 Altres reordenaments cromosòmics mediats per inversions paracèntriques críptiques

Osborne et al. van observar Inversions en heterozigosi a la regió de la síndrome de Williams-Beuren en quatre dels dotze pares transmissors del cromosoma relacionat amb la malaltia[110]. Aquests resultats han estat corroborats per Bayés et al.[39] en un estudi on es mostra que un ~25% dels progenitors transmissors eren heterozigots per una inversió mediada per les DSs disposades en sentit oposat de la regió 7q11.23. El mecanisme pel qual la inversió genera una deleció intersticial és degut probablement a que la sinapsi entre el cromosoma normal i l'invertit ocorre en tot el cromosoma, excepte en el loop de la zona invertida. D'aquesta manera es permet la recombinació homòloga no al·lèlica (NAHR) entre les DSs en direcció oposada que flanquegen la regió 7q11.23. Un estudi recent

mostra que dins aquests duplicons existeixen CNVs que actuen com a factor de predisposició a la deleció[132].

Un altre exemple és la translocació recurrent t(4;8)(p16;p23) que és mediada per dos parells de clústers de receptors olfactoris localitzats a 4p16 i 8p23[104,133]. En cinc casos de novo d'aquests tipus de translocació, tots d'origen matern, es va observar de nou que les mares eren dobles heterozigotes per les inversions de les regions flanquejades per clústers de receptors olfactoris a 4p16 i 8p23. Semblaria doncs que la parella de cromosomes homòlegs no es poden aparellar bé degut a la inversió, i es potencia la recombinació entre els cromosomes no homòlegs 4 i 8 en les regions de gran homologia constituïdes per receptors olfactoris. Estudis realitzats en aquestes regions mostren que la freqüència de la inversió en heterozigosi a 4p16 en individus control és del 12.5% i del 26% a 8p23, mentre que la freqüència de dobles heterozigots és només del 2.5%.

La presència d'inversions críptiques paracèntriques com a base de reordenaments no balancejats també s'ha observat en mares de nens amb la deleció a 15q11-q13 amb la síndrome d'Angelman[134]. La inversió es va identificar en quatre de sis mares de pacients Angelman amb la deleció. En aquest cas la inversió es troba present en un 9% de la població general. A través d'aquesta informació es pot postular que les inversions en heterozigosi entre DSs són propenses a patir de NAHR i per tant aquest tipus d'inversions constitueixen un factor de susceptibilitat important per reordenaments cromosòmics no equilibrats. Tot i això, no tots els reordenaments recurrents en zones flanquejades per DSs poden ser explicats per aquest mecanisme. Per exemple, la deleció de la regió 22q11.2 en pacients amb la síndrome de Di George/ Velocardiofacial, no s'ha trobat en cap cas associada a cap inversió[135], com tampoc s'ha trobat cap inversió en els progenitors de nens amb la deleció causant de la síndrome de Prader Willi.

3.2.5 Origen de les inversions

Tot i la seva freqüència i importància, es coneix molt poc la proporció d'inversions amb punts de trencament únics o múltiples. Un estudi dut a terme pel Wessex Regional Genetics Laboratory durant 40 anys es va centrar en la detecció d'inversions i a partir de les troballes resultants, un estudi recent s'ha centrat en les més recurrents per esbrinar si s'han originat a partir d'un ancestre comú o bé de manera independent més d'una vegada[112]. De les 188 combinacions de punts de trencament que es van detectar, només 35 semblen ser recurrents, tot i que els punts de trencament van ser detectats citogenèticament i per tant, pot ser que en realitat tinguin diferents punts de trencament a un nivell més alt de resolució.

Per finalitzar aquest apartat, cal destacar una publicació recent on a partir de la creació de llibreries genòmiques de fòsmids subclonats pertanyents a DNA de 8 individus, s'ha obtingut un mapa perfecte d'alta resolució de la variació estructural d'aquests individus[94]. Així, les dades més recents mostren l'existència de 224 inversions, així com un gran enriquiment d'aquests tipus de reordenament en certes regions del cromosoma X.

4. La regió cromosòmica 8p23.1

La regió 8p23.1 engloba 6.5 Mb (de 6.2 Mb a 12.7 Mb) a la part distal del braç curt del cromosoma 8 i és d'especial interès ja que es troba flanquejada als extrems per diversos grups de duplicons. Aquest fet li confereix una enorme inestabilitat, donat que aquesta arquitectura predisposa a la regió a patir diferents tipus de reordenaments cromosòmic[105,136-143]. Entre les reorganitzacions descrites a la literatura que afecten a 8p23.1 trobem duplicacions en tàndem, duplicacions invertides, delecions, inversions pericèntriques (p23q22), inversions paracèntriques i també translocacions. Tot i que les reorganitzacions a 8p23.1 són freqüents, el fenotip al qual estan associades no és del tot clar (tant existeixen casos amb fenotip normal, com casos associats a malalties psiquiàtriques[144-147]), i la complexitat de les DSs d'aquesta regió dificulta poder delimitar acuradament els punts de trencament.

4.1 Arquitectura genòmica de 8p23.1

L'arquitectura genòmica de 8p23.1 es caracteritza per la seva complexitat i inestabilitat, essent una de les regions del genoma humà més difícils de desentrellar, com ho demostra el fet que en l'actual assemblatge encara hi ha "*gaps*" de seqüència. Aquest fet és degut a l'existència de DSs altament complexes a banda i banda de la regió que contenen múltiples repeticions en tàndem d'orientació tan directe com oposada, que alhora poden ser variables en número de còpia. Les duplicacions més distals es coneixen amb el nom de REPD i les proximals REPP (Figura 4.1). Originàriament es coneixia aquestes DSs com a una zona rica en receptors olfactoris, però a mesura que s'ha anat obtinguent un coneixement més detallat de la regió, els clústers d'altres gens com les defensines representen una extensió molt més significativa d'aquestes DSs.

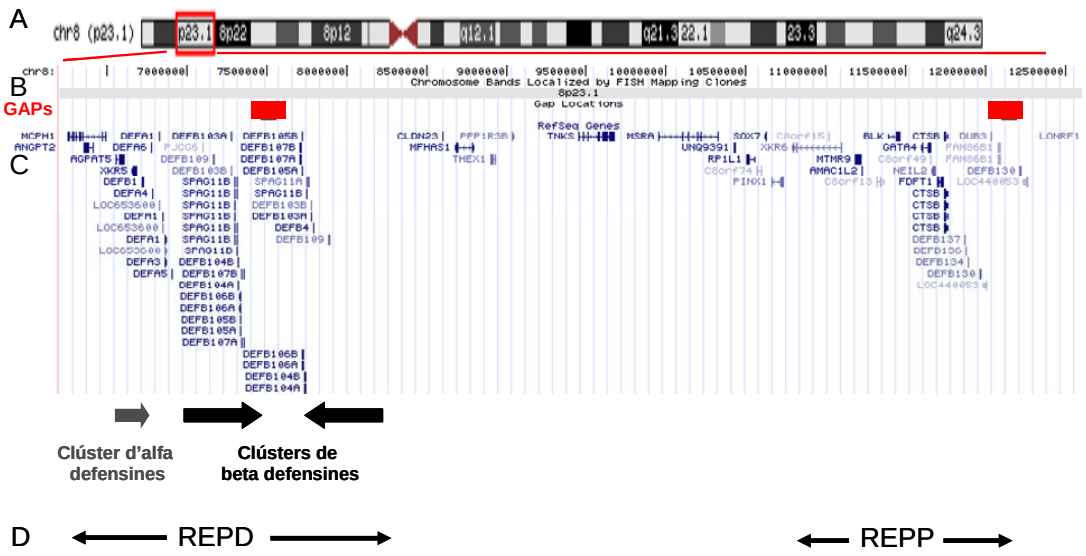

Figura 4.1. Regió cromosòmica 8p23.1. A) Imatge del cromosoma 8 amb la regió 8p23.1 marcada en vermell. B) Localització dels gaps (rectangles vermells) de 100 kb als dos extrems de la regió. C) Representació dels gens de 8p23.1. La fletxa grisa mostra l'orientació del clúster d'alfa defensines, i les negres dels de beta defensines a costa i costat del gap. D) Localització dels duplicons distals REPD i proximals, REPP. La regió compresa entre REPD i REPD és la que es troba invertida en població general.

La regió corresponent a REPD s'estén des de 6.8 Mb fins a 8.1 Mb (NCBI build 36.1; UCSC version hg18), mentre que REPP s'estén des de 11.5 Mb fins a 12.6 Mb. És important tenir en compte que aquesta regió s'engloba dins el conjunt de duplicacions segmentàries no resoltes de l'actual seqüència del genoma humà ja que encara a l'assemblatge del Març del 2006 existeix un gap de 100 kb a 7,5 Mb i el mateix passa a la regió REPP a 12 Mb.

Un motiu pel qual aquests gaps encara no s'han cobert és que tot i que hi ha clons de la zona perfectament aptes per seqüenciar, les seqüències resultants no poden ser mapades de manera no ambigua, ja que quan dues seqüències comparteixen >99% d'identitat és pràcticament impossible discernir si representen loci diferents o són diferents al·lels d'un mateix locus[4]. Els diversos intents de cobrir aquests dos gaps s'han vist dificultats per la presència de duplicons a les seqüències flanquejants.

Si ens centrem en REPD, trobem un grup de gens relacionats amb resposta immune anomenats defensines que s'han classificat com a polimòrfics pel que fa a número de còpies. Les defensines es divideixen en alfa i beta-defensines. Les alfa-defensines d'aquest clúster de 19 kb corresponen a *DEF1A3* i el pseudogen *DEFTP*[148]. El clúster beta-defensines de 240 kb està format per nou beta-defensines diferents i es troba duplicat en orientació reversa a costat i costat del gap[149] (Figura 4.1).

4.2 CNVs a la regió cromosòmica 8p23.1

Tal i com s'ha esmentat en apartats previs, gran part del genoma correspon a polimorfismes estructurals i molts d'ells són variants en número de còpia. Una de les CNVs més ben estudiades correspon al clúster de defensines[149,150] a la regió

humana de 8p23.1 degut a la seva potencial rellevància en la immunitat innata, la inflamació i el càncer[151-154]. La regió pot ser dividida en dos subclústers que s'estenen ~2 Mb i que contenen bàsicament gens corresponents a les alfa i a les beta-defensines. Per tal de determinar el número de còpies de les defensines s'han utilitzat diferent mètodes en grans cohorts[148,149,155-158]. La majoria d'aquestes tècniques comparen el locus de les defensines amb un gen de referència de número de còpia conegut i invariable.

La variabilitat en número de còpia és un aspecte essencial de la dinàmica del genoma humà, i alguns al·lels poden estar implicats en el desenvolupament de certes malalties, especialment si la dosi gènica influeix en l'expressió gènica. Les CNVs es troben escampades en els genomes dels mamífers[87,96,99] i per tant semblen ser tan importants alhora de modelar els genomes com ho són els SNPs. En el cas de les CNVs polimòrfiques s'han de tenir en compte ambdues coses, tant el número de copies en un individu com la variació de seqüència entre còpies. Tot i així, la caracterització del número de còpies dels gens compresos en CNVs comporta dificultats a nivell metodològic[159,160].

4.2.1 Variabilitat de les alfa-defensines

Dins el clúster de les alfa-defensines trobem, entre d'altres gens de la família, la *DEFA1* i *DEFA3,* que en realitat no són més que variants al·lèliques i per tant és més correcte parlar de *DEFA1A3.* Diversos estudis han constatat que *DEFA1A3* són les úniques variables en número de còpia (de 4 a 14 còpies per genoma diploide) mentre que la resta de defensines del clúster (*DEFB1, DEFA6, DEFA4, DEFT1 i DEFA5*) són invariables presentant-se sempre en dos còpies per genoma diploide[148,157,161]. Així s'ha pogut arribar a determinar que el que en realitat varia és una unitat de repetició de 19 kb que engloba els loci *DEFA1A3* i *DEFTP*. En aquest mateix estudi es mostra la presència en elevat número de còpies de

DEFA1 també en altres primats superiors[148]. I no només existeix aquesta variabilitat si no que una proporció significativa d'individus de totes les poblacions (del 10% al 37%) presenten una absència total de l'al·lel *DEFA3* [148,157,161].

4.2.2 Variabilitat de les beta-defensines

Per altra banda, tots els loci de les beta-defensines (*DEFB109p, DEFB108, DEFB4, DEFB103, DEFB104, DEFB106, DEFB105 i DEFB104*) han estat genotipats i s'ha trobat que varien de dos a set copies per genoma diploide, mentre que els portadors de duplicacions corresponents variants eucromàtiques visibles al microscopi però que no comporten cap fenotip clínic, presenten de 9 a 12 còpies d'aquest clúster de 240 kb[157,162].

En quant a la variabilitat de cadascun dels gens d'aquest clúster, un estudi descriu que aquesta és independent per cadascun dels gens[142]. Però a la literatura trobem més evidències que recolzen la situació contrària on tot el clúster varia com una entitat, és a dir, que existeix una concordança estricta en quant al número de copies dels tots els loci de beta-defensines en cada individu[157,163]. Aquesta concordança ajuda a simplificar la determinació del número de copies d'aquesta regió ja que permet tipar la variabilitat a partir d'un únic assaig d'un sol locus com a representant de tot el clúster. Aquest avantatge comporta però una limitació, i és que en estudis on es troba una associació positiva entre el número de còpies del clúster de beta-defensines i una malaltia concreta, és difícil d'establir quin és el gen de dins el clúster responsable d'aquesta associació.

Una solució a aquest problema és establir si hi ha correlació entre el número de còpies dels diversos gens i els seus nivells d'expressió. En aquest sentit l'estudi dut a terme per Hollox *et al.* va mostrar una correlació positiva entre el número de còpies de *DEFB4* i els seus nivells d'expressió en limfòcits[149]. Per altra banda Fellerman *et al.* van

correlacionar el número de còpies de *DEFB4* amb la malaltia de Crohn, de manera que a menor nombre de còpies més predisposició existeix a que els individus pateixin la malaltia inflamatòria. És més, van observar que l'expressió de *DEFB4* en la mucosa d'individus amb un número de còpies menor a 4 es troba reduïda[155].

4.3 Inestabilitat genòmica: Resum dels reordenaments més destacats a la regió cromosòmica 8p23.1

4.3.1 Duplicacions

a) duplicació a la regió 8p23.1

El fenotip clínic dels individus amb duplicacions a 8p23.1 és ambigu ja que s'han trobat casos associats a un fenotip completament normal, d'altres a retard mental, trets autístics, o patologies lleus (cap petit, dismorfisme facial), així com també casos amb patologies cardíaques severes[164]. Aquesta variabilitat tan pot ser deguda al fet que les duplicacions poden no compartir els mateixos punts de trencament o bé a que l'aneuploidia de 8p23.1 pot estar compensada de manera diferent per altres al·lels en cada individu.

b) inversió duplicació deleció de 8p23.1

Aquest tipus de duplicació és de mida variable (el punt de trencament proximal no és fixe)**,** té una freqüència de 1/10000-30000 naixements[105,141], les seves conseqüències fenotípiques inclouen retard mental, problemes cardíacs i hipotonia[139,141,142,165-173]. Es tracta d'un reordenament mediat pels receptors olfactoris presents a banda i banda de 8p23.1, i consta d'una deleció de quasi 6 Mb a la part distal de 8p. Entre les dues còpies hi ha una regió de còpia única de ~3.4 Mb. Les mares dels pacients inv dup (8p) presenten una inversió en heterozigosi a

la regió 8p23.1. Els casos d'inv dup (8p) s'originen a la meiosi materna probablement degut a un mal aparellament entre cromosomes homòlegs.

c) cromosomes marcadors

Un cromosoma marcador és un "cromosoma addicional" generalment de petit tamany, l'origen del qual no pot ser detectat per les tècniques citogenètiques habituals. Els cromosomes marcadors es troben en individus normals i també en individus amb diferent grau d'anomalies (des de característiques dismòrfiques fins a retard psicomotor). S'han descrit cromosomes marcadors amb neocentròmers relacionats amb la regió 8p23.1, tot i que, degut a la seva difícil detecció el nombre de casos i les conseqüències que comporten estan subestimats[105].

4.3.2 Delecions

Les delecions de la part més distal de 8p són un tipus de reordenament recurrent on el punt de trencament sembla conservat i la pèrdua sol ser de 8p23.1 i 8p23.2. Com a conseqüència de la deleció, els individus

presenten haploinsuficiència del gen *GATA4*, localitzat a 8p23.1, donant lloc a problemes cardíacs, retard mental i problemes de comportament[136-138,140,174-179].

4.3.3 Translocacions

Entre les diferents translocacions associades a 8p23.1 el reordenament t(4;8)(p16;p23) podria tractar-se de la segona translocació més freqüent en humans després de t(11q;22q). Les regions 8p23.1 i 4p16 estan flanquejades per receptors olfactoris. Aquest tipus de translocació pot ser balancejada o no. En els casos amb der (4) els individus pateixen la síndrome de Wolf-Hirschhorn, mentre que els individus amb un cromosoma derivatiu del 8 presenten un fenotip molt més lleu. De nou,

les mares analitzades amb fills amb der (4) van resultar ser heterozigotes per les inversions de 8p23.1 i de 4p16[104].

4.3.4 Inversió polimòrfica

La inversió de 8p23.1 és la inversió polimòrfica trobada a les mares dels pacients amb els reordenaments prèviament esmentats. Es tracta d'una inversió d'unes 3.4 Mb mediada per les DSs que flanquegen 8p23.1 amb una freqüència de l'1% en homozigosi i del 26% en heterozigosi en població general europea i japonesa[105,108]. La presència de la inversió no està associada a cap fenotip concret, tot i que sembla que hi ha prou evidències per afirmar que la inversió de 8p23.1 predisposa a un mal aparellament entre cromosomes homòlegs en meiosi desencadenant diferents tipus de reordenaments a la descendència. El perquè aquests reordenaments transmesos a la descendència no són més freqüents i el fet que la seva incidència sigui més elevada en les mares que en els pares no està del tot clar. Com tampoc se sap si ser portador de la inversió representa un tipus d'avantatge o desavantatge selectiu.

Tot i que la inversió polimòrfica a 8p23.1 per si sola no té efectes patològics, en dones portadores de la inversió en heterozigosi es poden produir recombinacions desiguals durant la meiosi que donen lloc majoritàriament a tres tipus de reordenaments ja esmentats : del (8p), der (8) (8p23.1pter), i del (8)(p23.1p23.2), relacionats amb fenotips severs[180].

Per tal d'esbrinar quins són els mecanismes que donen lloc a la inversió, quines seqüències hi participen, si existeix pèrdua o guany de material genòmic arrel de la inversió, i quins efectes pot tenir sobre l'expressió dels gens tant de les DSs que promouen el reordenament, com dels gens presents enmig dels dos grans blocs de duplicons, com de gens més allunyats, cal aprofundir en l'estudi de l'estructura genòmica del conjunt de DSs de 8p23.1.

Podem dir doncs que les inversions, i més específicament les inversions críptiques flanquejades per DSs orientades inversament poden ser molt més freqüents que el que s'havia estimat prèviament. A més, les inversions críptiques flanquejades per "*low copy repeats*" (LCRs), també anomenats duplicons, són un factor important de susceptibilitat per la ocurrència de reordenaments cromosòmics constitucionals no balancejats, però no l'únic. Després d'anys d'obscuritat de les causes moleculars d'anomalies cromosòmiques estructurals i dels possibles factors de susceptibilitat que hi ha al seu darrera, sembla que se'n comença a treure l'entrellat.

Les DSs són la causa d'inestabilitat genòmica i els portadors d'inversions críptiques tenen un risc incrementat de formar gàmetes amb alteracions cromosòmiques. Les investigacions sobre la presència de variants heterozigotes respecte a variacions a gran escala en número de còpia en els pares que transmeten anomalies cromosòmiques estructurals, probablement determinin si aquestes variants representen un altre factor de risc. Cal no oblidar però, que la complexitat genòmica deguda a la presència de variants estructurals i seqüències paràlogues, és un tret essencial de 8p23.1 que cal tenir en compte alhora d'interpretar els estudis de lligament i d'associació realitzats en la regió. La riquesa de variants genètiques i genòmiques que presenta 8p23.1 la converteix en una regió que confereix una elevada plasticitat al genoma interessant per entendre amb major detall el procés d'evolució dels genomes.

Objectius

Basant-nos en la hipòtesi que les duplicacions segmentàries (DSs) tenen un paper substancial en la creació de nous gens, en la dinàmica evolutiva dels cromosomes i, en darrera instància, en el procés d'especiació, els objectius d'aquesta tesi doctoral van encaminats a desxifrar la participació dels duplicons presents a 8p23.1, i en concret els clústers de la família gènica *FAM90A,* com a dianes potencials de la inestabilitat genòmica de la regió.

Els objectius específics que s'han plantejat es resumeixen en:

1- Caracterització de les DSs que flanquejen la regió 8p23.1
   - 1.1 Descripció a nivell genòmic de la família gènica *FAM90A* en humans.
   - 1.2 Descripció a nivell genòmic de la família gènica *FAM90A* en primats superiors no humans.
   - 1.3 Perfil d'expressió de la família gènica *FAM90A*.
   - 1.4 Estudi del passat evolutiu de les DSs de 8p23.1 a partir d'anàlisis de genòmica comparada.
2- Quantificació de la variant gènica en número de còpia *FAM90A* en diferents poblacions humanes i estudi de la correlació entre el número de còpia i els nivells d'expressió.
3- Estudi de la freqüència de la inversió de la regió 8p23.1 en diferents poblacions i com afecta aquest reordenament a l'expressió dels gens de la regió.

Resultats

Aquesta secció és el compendi dels resultats obtinguts durant la present tesi doctoral. La major part d'aquests resultats s'han publicat o estan sotmesos a publicació en revistes científiques indexades. Aquesta secció està organitzada en diferents apartats, cadascun correspon a un treball de recerca publicat, tots ells concebuts per respondre els diferents objectius d'aquesta tesi doctoral. Cada apartat va precedit d'una petita introducció que resumeix l'objectiu del treball, com es va dur a terme i les seves conclusions.

Els diferents apartats són:

1- "Characterization and evolution of the novel gene family *FAM90A* in primates originated by multiple duplication and rearrangement events"

2- "Analysis of the multi-copy gene family *FAM90A* as a copy number variant in different ethnic backgrounds"

3- " Frequency of the 8p23.1 polymorphic inversion in HapMap populations and its impact on gene expression levels"

Characterization and evolution of the novel gene family *FAM90A* in primates originated by multiple duplication and rearrangement events

Nina Bosch, Mario Cáceres, Maria Francesca Cardone, Anna Carreras, Ester Ballana, Mariano Rocchi, Lluís Armengol and Xavier Estivill.

Aquest estudi és el resultat de la primera aproximació per tal d'obtenir una caracterització detallada de les duplicacions segmentàries (DSs) que es troben flanquejant la regió cromosòmica 8p23.1. En el moment de l'inici d'aquest estudi, i encara a dia d'avui, aquestes duplicacions contenen "gaps" en el seu interior, ja que la seva complexa arquitectura genòmica la converteix en una regió especialment difícil d'assemblar. Amb l'objectiu d'esbrinar quins elements es repetien a l'interior de d'aquestes duplicacions segmentàries es va identificar un fragment d'aproximadament 7 kb que s'anava repetint reiteradament al llarg d'aquests duplicons. Una anàlisi més profunda va permetre confirmar que es tractava d'un gen, i d'aquesta manera es va poder descriure l'existència d'una nova família gènica, *FAM90A*, fins al moment desconeguda.

Un cop identificat el gen, es va detallar l'estructura dels paràlegs d'aquesta família i es va determinar l'existència de dues subfamílies de *FAM90A*, la I i la II, atenent al seu tamany. Aquestes subfamílies estan formades per un total de 25 membres diferents per genoma haploid en l'individu de referència utilitzat per l'assemblatge del genoma humà. Aquests gens estan distribuïts en "clústers" al llarg dels duplicons distals que flanquejen 8p23.1, REPD, mentre que als duplicons proximals, REPP, i al cromosoma 12p13 trobem 3 còpies senzilles. A més, experiments utilitzant electroforesi de camps pulsants en 20 individus diferents van revelar que els "clústers" de *FAM90A* són extremadament polimòrfics. Estudis de genètica comparada van permetre concloure que es tracta

d'un gen específic de primats, i els estudis de FISH i PCR quantitativa realitzats en diferents espècies de primats van mostrar una expansió en quant al número de còpies d'aquesta família gènica al llarg de l'escala evolutiva. La majoria de dominis de la proteïna predita són desconeguts i l'estudi a nivell de l'RNA missatger (mRNA) de 13 teixits diferents va permetre observar que es tracta d'un gen àmpliament expressat. Finalment, es va hipotetitzar un possible origen d'aquestes dues subfamílies gèniques que inclouria diversos reordenaments genòmics, des de delecions fins a fusions i duplicacions.

Bosch N, Cáceres M, Cardone MF, Carreras A, Ballana E, Rocchi M, Armengol L, Estivill X.
Characterization and evolution of the novel gene family FAM90A in primates originated by multiple duplication and rearrangement events.
Hum Mol Genet. 2007 Nov 1;16(21):2572-82. Epub 2007 Aug 7.

Analysis of the multi-copy gene family *FAM90A* as a copy number variant in different ethnic backgrounds

Nina Bosch, Geòrgia Escaramís, Josep Maria Mercader, Lluís Armengol and Xavier Estivill.

En el treball anterior, on es descrivia detalladament la familia gènica *FAM90A*, també es mostrava l'elevat grau de polimorfisme existent entre els diferents "clústers" de *FAM90A* en humans. És per això que es va dur a terme aquest altre estudi on, mitjançant PCR quantitativa, es va quantificar la variabilitat en número de còpia d'aquesta familia gènica. La quantificació es va realitzar en un total de 260 individus de les poblacions HapMap. Els resultats d'aquest treball demostren que la població d'origen europeu (CEU) i asiàtic (ASN) presenten diferències estadísticament significatives en quant a número de còpies de *FAM90A* respecte de la població d'origen africà (YRI). És en aquesta última població on els individus tenen un número més elevat de còpies de *FAM90A*. Per altra banda també es va fer un estudi per explorar si existia correlació entre el número de còpies de *FAM90A* i els seus nivells d'expressió. Amb aquesta finalitat es van utilitzar les dades d'espressió d'individus HapMap publicades per Stranger et. al. {Stranger, 2007 #529}Els resultats obtinguts van indicar que la correlació entre el número de còpies de *FAM90A* de cada individu i els seus nivells d'expressió és molt baixa (R^2=0.05, p=0.002). En conseqüència, el treball conclou que els nivells d'expressió dels membres de la família FAM90A no estan correlacionats amb el nombre de còpies i postula que han d'existir altres factors que en siguin responsables.

Bosch N, Escaramís G, Mercader JM, Armengol L, Estivill X.
Analysis of the multi-copy gene family FAM90A as a copy number variant in different ethnic backgrounds.
Gene. 2008 Sep 1;420(2):113-7. Epub 2008 May 13.

Analysis of 8p23.1 polymorphic inversion in HapMap populations and its impact on gene expression levels

Nina Bosch, Marta Morell, Imma Ponsa, Josep Maria Mercader, Magda Montfort, Lluís Armengol and Xavier Estivill.

L'objectiu d'aquest treball és aprofundir en el coneixement de la inversió polimòrfica que afecta a la regió cromosòmica 8p23.1. Per una banda es va analitzar mitjançant FISH la presència de la inversió en 24 individus control, d'origen espanyol. Els resultats van mostrar un elevat grau de mosaïcisme en les cèl·lules de tots els individus estudiats. També es va veure que la inversió tenia una freqüència del 79% en aquesta població. A partir de 6 individus homozigots per la inversió es va realitzar un "whole genome scan" per tal de trobar SNPs marcadors de la inversió. Es van definir dos blocs d'homozigositat que correlacionaven amb la presència de la inversió en tots sis individus. Posteriorment, es va predir l'estat de la inversió de 8p23.1 en els individus de HapMap a partir dels haplotips disponibles i utilitzant els blocs d'homozigositat descrits com a marcadors. El resultat de les prediccions va mostrar una elevada presència de la inversió en població tant d'origen europeu (CEU) com d'origen asiàtic ASN, mentre que l'absència dels blocs d'homozigositat en africans (YRI) va fer que aquests marcadors no poguessin ser utilitzats per fer les prediccions en aquesta població. Per confirmar la validesa d'aquests marcadors es van genotipar per FISH un subgrup d'individus. Els resultats van mostrar la presència de mosaïcisme per la inversió de 8p23.1 i que, en tots el casos, el genotip d'inversió observat per FISH corresponia amb el genotip predit pels marcadors. Finalment, es va fer un estudi d'associació entre el fenotip d'inversió i els nivells d'expressió de 26 gens continguts a la regió 8p23 i ambdós extrems flanquejants, i quatre gens van resultar ser diferentment expressats segons la conformació de la regió.

Analysis of 8p23.1 polymorphic inversion in HapMap populations and its impact on gene expression levels

Nina Bosch[1, 2], Marta Morell[1, 2], Imma Ponsa[3], Anna Carreras[1,2], Josep Maria Mercader[4], Magda Montfort[2, 5], Lluís Armengol[1,2], Xavier Estivill[1, 2, 6]

Abstract

The 8p23.1 region is a 6.5 Mb fragment on the short arm of chromosome 8 which can be found in different orientations among human individuals. The identification of markers linked to a particular orientation of the region is useful to characterize the homozygous or heterozygous status of the orientation of this 6.5 Mb fragment in a certain individual. Here we present a whole genome scan performed in homozygous individuals for the 8p23.1 inversion previously genotyped by FISH. As a result of this analysis we have been able to select 16 polymorphic loci which are in high linkage disequilibrium with the inversion. The combination of these 16 correlated markers are a reliable surrogate for the chromosome 8p inversion status of an individual as reported by FISH. The pattern of geographic variability of the 8p23.1 inversion allowed the use of these markers in CEU and ASN populations. Thus, we have predicted the status of the 8p23.1 inversion in 150 HapMap samples and we have verified the predictions by FISH analysis in a subset of individuals. Moreover, the presence of different genotypes, with respect to the 8p23.1 inversion, within all the 33 individuals studied by FISH points out the interesting phenomenon of mosaicism, which brings more complexity to the already complicate genomic structure of the region. In addition, we have also analyzed the effects of the allelic variation between the inverted and normal orientation of the genes contained in the 8p23.1 fragment. The association study between gene expression levels and the status of the inversion revealed that four genes (*NEIL2, MSRA, CTSB* and *BLK*) are differentially expressed ($p<0.0005$) according to the orientation of the 8p23.1 region.

Introduction

Among the different classes of structural variations, knowledge of the prevalence of the inversions in the human genome is poor. One of the reasons is that technologies developed for the discovery of genome-wide structural variants have limitations to detect balanced rearrangements such as the different orientations of the inverted regions. However, several studies based on fosmid subcloning and re-sequencing have succeeded in mapping different inversion breakpoints throughout the whole genome (Tuzun et al., 2005; Eichler et al., 2007; Korbel et al., 2007). Recently, a detailed study on eight human genomes using this method has described 224 different inversions which have mainly arisen by non-allelic homologous recombination (NAHR) between segmental duplications (SDs) that can be found flanking the inverted segments (Kidd et al., 2008).

Another approach to delineate human inversions, which was first described on *Drosophila* studies (Navarro et al., 2000), is the characterization of the extended blocks of linkage disequilibrium (LD) that are created due to the lack of recombination in the heterozygous individuals (Pritchard and Przeworski, 2001). Following this criteria, haplotype subgroups can be defined in polymorphic inversions because different alleles are maintained in the different arrangements. Such is the case for the polymorphic inversion on chromosome 17q21.31, for which one of the haplotypes associated to one of the conformations has been found to be under positive selection in Europeans (Stefansson et al., 2005)(Zody et al., 2008)

The polymorphic inversion on chromosome 8p23.1 encompasses ~3.8 Mb and includes at least 50 coding genes. It was first described to have a frequency of 26% in European population (Giglio et al., 2001) and 27% in the Japanese population (Sugawara et al., 2003), assuming that the reference assembly corresponds to the non-inverted conformation. However, different studies have found an increased frequency of the inversion, around 60%, in populations of European ancestry, indicating that the human reference

assembly corresponds to the less common arrangement of the region (Chen et al., 2006; Deng et al., 2008). Moreover, 8p23 is an intricate DNA segment flanked by two large sets of SDs, REPP and REPD, where genes variable in copy number, such as defensins or *FAM90A*, are located (Taudien et al., 2004; Aldred et al., 2005; Barber et al., 2005; Linzmeier and Ganz, 2005; Fellermann et al., 2006; Ballana et al., 2007; Bosch et al., 2008; Hollox et al., 2008). Thus, the size of the flanking duplicons is also variable and it seems plausible that this phenomenon could play a role in the distinct rearrangements affecting the 8p23.1 fragment. Although considered a neutral inversion polymorphism, the 8p23.1 inverted status has been found in mothers of children with an associated phenotype suffering from different rearrangements involving 8p23.1 segment. The same phenomenon has been described in the parents of children carrying genomic disorders (Lupski, 1998) like Prader-Willi or Angelman syndrome (Gimelli et al., 2003), Williams-Beuren syndrome (Osborne et al., 2001; Bayes et al., 2003) and Hunter syndrome(Bondeson et al., 1995). To bring more complexity to the already complicated 8p23.1 region, it has been recently reported to be the genomic fragment showing the highest concentration of structural variants, increasing the difficulty to interpret the outputs when analyzing this chromosome region (Kidd et al., 2008).

Here we present a whole genome analysis performed in 6 Spanish control individuals which were first genotyped by FISH and found to be homozygous for the 8p23.1 inversion. In our aim to identify surrogate markers for the inversion, we have identified two tracks of homozygosity which perfectly correlate with the inversion status of the region in European and Asian populations. Thus, we predicted the genotype of the inversion for 150 HapMap individuals, and we confirmed our predictions by FISH analysis. Moreover, we have detected the presence of different genotypes with respect to the inversion within almost all the samples analyzed by FISH. In addition, we have explored the effect of the inversion rearrangement on gene

expression levels and we found four genes with statistically significant different expression levels ($p<0.0005$) depending on the inversion status.
The study described here can be used to investigate the frequency and the effects on gene expression of any other inversion in the human genome.

Methods

Genotype data and the delineation of homozygosity tracks

A whole genome analysis was performed using the HumanCNV370-Duo chip from Illumina in a sample of 6 Spanish general population individuals homozygous for the 8p23.1 inversion plus NA10861 HapMap individual as a control for genotype concordance. Genotyping was carried out following manufacturer's protocol (Illumina Inc.) at the CeGen genotyping center. This BeadChip uses SNPs to cover the majority of the common variation described for the CEU, CHB/JPT, and YRI populations based on HapMap Phase I and II data. Among these, we focused on 770 SNPs spanning the 8p23.1 region, where two tracks of homozygosity containing 16 SNPs could be defined between 8.5-8.7 Mb and 10.2-11 Mb in all 6 homozygous inverted individuals.
Genotype data of 210 unrelated HapMap individuals, including 60 parents of 30 trio samples from CEPH in Utah residents with ancestry from northern and western Europe (CEU), 60 parents of 30 trio samples from Yoruba in Ibadan, Nigeria (YRI), 45 unrelated Han Chinese from Beijing, China (CHB), and 45 Japanese from Tokyo, Japanese (JPT) individuals, were also used in the analysis. Due to the significant genetic similarity between Chinese and Japanese groups, we pooled CHB and JPT data and denoted these individuals as Asian (ASN) as seen in other studies (Voight et al., 2006).
HapMap phase II data was downloaded from the website (HapMap Data Rel23a/phaseII Mar08; www.hapmap.org). Phased haplotypes were used to predict the

genotype of the 150 individuals with respect to the 8p23 inversion based on the conserved homozygosity tracks previously described. The 150 individuals analyzed included the 60 parents from CEU samples and the 90 ASN individuals, the two populations were the 16 SNP markers are in linkage disequilibrium with the inversion.

FISH analysis

FISH using BAC-derived DNA was performed on metaphase chromosomes prepared from lymphoblastoid cell lines obtained from 24 control Spanish individuals. Two BAC clones that fall within the 8p23 inversion and that are free of segmental duplications (RP11-399J23; RP11-589N15) were used for the experiments. DNA from BAC clones was isolated by standard procedure and labeled by nick translation with biotin (detected with fluorescein) for RP11-399J23, and digoxigenin (detected with rhodamine) for RP11-589N15. Labeled probes were precipitated and resuspended following standard protocols. Next, metaphase chromosomes and probes were denatured and hybridized overnight at 37º C. Chromosomes were counterstained with 4', 6-diamino-2-phenylindole (DAPI). Images were analyzed using the Isis software from Metasystems. A minimum of 20 metaphases with clearly interpretable signals on both chromosomes were counted per individual.

HapMap cell lines were obtained from Coriell Repositories (Camdem, NJ). FISH analysis following the same procedure was performed in 9 HapMap samples to confirm the predictions for the 8p23.1 inversion based on the described surrogate SNP markers. Individuals NA11992, NA12057, NA11839 were used for the confirmation of the non-inverted conformation of 8p23.1. Next group was formed by heterozygous individuals NA11993, NA06993 and NA11994. Finally the homozygous status of the inversion was confirmed by this technique in individuals NA11831, NA12815 and NA12155.

Association study of 8p23.1 inversion and gene expression levels

Normalized gene expression values from the different populations (CEU, CHB/JPT) were downloaded from the Sanger Genevar webpage (http://www.sanger.ac.uk/humgen/genevar/). Data retrieved by 31 Illumina™ different probes targeting 26 genes located around 8p23.1 region (Supplementary Table 1) were used to explore the correlation between gene expression levels and the three different status of the inversion (non-inverted homozygous, heterozygous and homozygous inverted).

Individuals included in this analysis were the 150 samples corresponding to the CEU and ASN populations were the status of the inversion could be predicted using surrogate markers. The association analysis between gene expression values and the genotype for the inversion was tested for the different genetic models. P values were derived from likelihood ratio tests, and a significance level of 5% (two sided) was used for the analyses. All these analyses were performed using the SNPassoc R package (Gonzalez et al., 2007).

Results

Frequency of 8p23.1 inversion in Spanish population

To establish the frequency of 8p23.1 inversion in the Spanish population we genotyped a total of 24 Spanish control individuals by FISH analysis. For that we used BAC clones RP11-399J23 and RP11-589N15, localized at each end of the 3.8 Mb inversion on chromosome 8 and outside the segmental duplications flanking 8p23.1 (Sugawara et al., 2003). Based on the human reference assembly Human NCBI Build 36 we consider the telomere-to-

centromere orientation RP11-589N15 (Red) and RP11-399J23 (Green) as the inverted conformation ("Build36-inverted"). From the 24 individuals that were

genotyped, 50% resulted heterozygous for the inversion, 29% homozygous for the inverted conformation and 21% homozygous for the normal conformation (Fig. 1). Thus, the conformation present in the reference assembly exemplifies the less represented orientation in the Spanish population.

As shown in Table 1, this study also revealed that although a predominant genotype could be established for each control sample studied, metaphases corresponding to different genotypes are observed in almost all individuals. Therefore, from the 12 samples that were heterozygous for the inversion, the percentage of metaphases homozygous for either the normal or the inverted conformation ranged from 14% to 38%. Whereas among the seven individuals homozygous for the inversion, 12% to 22% of the metaphases resulted heterozygous for the inversion or homozygous for the normal conformation. Finally, among the 5 individuals that were found to be homozygous for the normal conformation, up to 18% of the metaphases corresponded to the heterozygous genotype.

These results point out not just the extremely high instability of the region, probably due to the two sets of large segmental duplications on both 8p23.1 extremes, but also highlights the difficulty when interpreting the genotype for the 8p23 inversion in terms of presence or absence. According to our results, a predominant genotype can be assigned to individuals but it is important to take into account that a percentage of mosaicism is extremely common, at least in samples derived from lymphoblastoid cell lines.

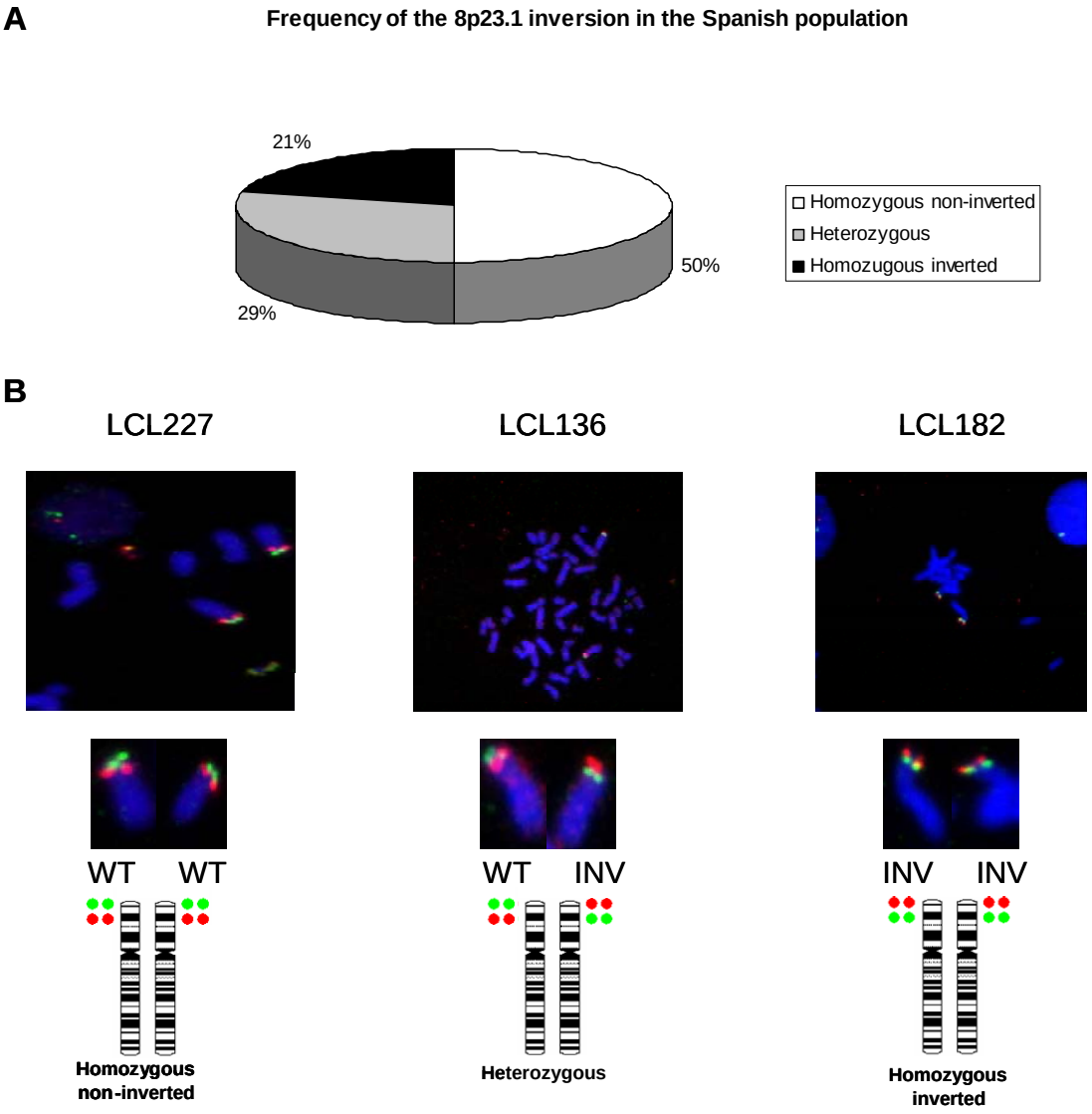

Figure 1. FISH analyses for the 8p23.1 inversion. DNA probes were made from BAC clones RP11-399J23 (Green) and RP11-589N15 (Red). **A** Frequencies of each of the three genotypes extracted from 24 Spanish control individuals. **B** Metaphase FISH of three Spanish control individuals, LCL227 as an example of "Build-36-non-inverted individual"; LCL136 is heterozygous for the 8p23.1 inversion and LCL182 corresponds to a homozygous "Build36-inverted individual".

Table 1. Percentages of each of the three possible genotypes observed within each Spanish individual in a subset of 10 samples. The predominant genotype is represented as a grey box. In individuals LCL159, LCL183 and LCL198 the predominant genotype is the non-inverted status. Individuals LCL146, LCL161, LCL184, LCL241 and LCL247 have most of the metaphases heterozygous for the 8p23.1 inversion. In samples LCL194 and LCL339 the most observed genotype is the homozygous inverted.

	METAPHASES		
Samples	Homozygous non-inverted	Heterozygous	Homozygous inverted
LCL159	87%	10%	3%
LCL183	82%	18%	-
LCL198	92%	8%	-
LCL146	14%	62%	24%
LCL161	19%	75%	6%
LCL184	13%	84%	3%
LCL241	5%	86%	9%
LCL247	20%	70%	10%
LCL194	4%	8%	88%
LCL339	-	22%	78%

Conservation of homozygosity blocks in 8p23.1 inverted alleles

If we consider that haplotype subgroups are created by suppression of recombination in heterozygous status at inverted regions, we should be able to describe surrogate markers for 8p23 inversion. To disclose the surrogate markers, six of the Spanish control individuals that were found to be homozygous for the 8p23.1 inversion by FISH analysis were genotyped using the Illumina's HumanCNV370-Duo chip. By this procedure we were able to delineate these markers as homozygosity tracks which include 16 SNPs, near

by the SDs, within the 8p23.1 inverted segment. The first homozygosity track expands ~172 kb and contains 8 SNPs (Fig. 2) which conform the "CGTCGAGG" haplotype in all 6 individuals (Table 2). This signature is located at 10.5 Mb, close to the REPD distal segmental duplications that are flanking 8p23.1 segment. A second block of 8 homozygous SNPs spans ~181 kb and in this case the conserved haplotype is "TCACGAGA" (Table 2) and lays at 10.8 Mb, close to REPP the proximal set of segmental duplications on 8p23.1 (Fig. 2).

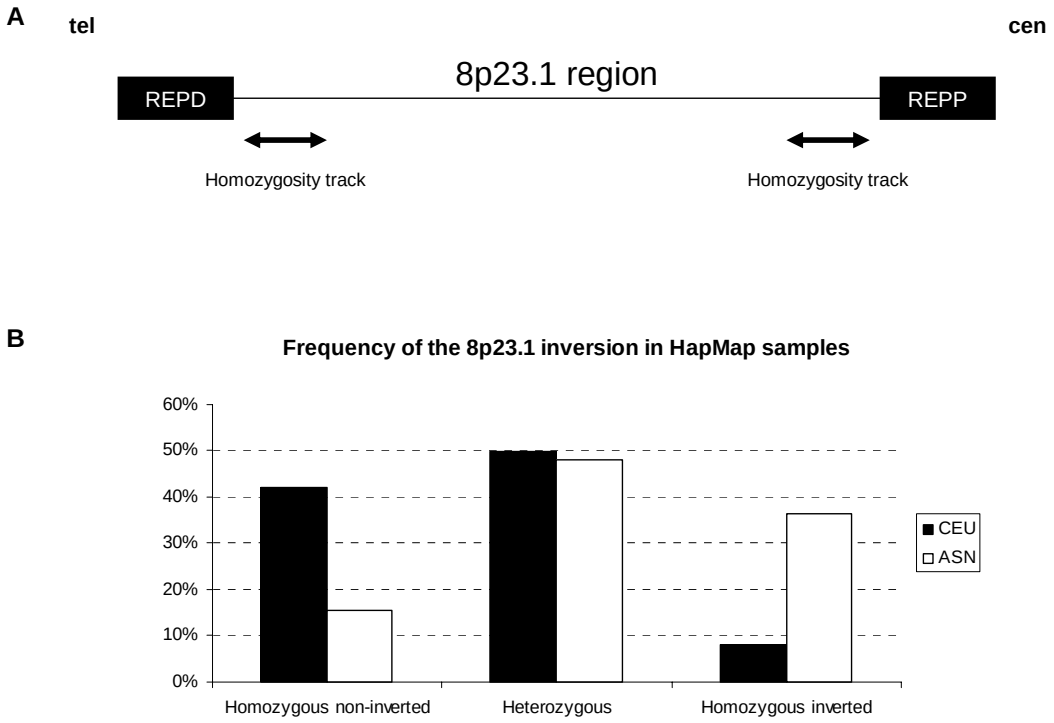

Figure 2. A Scheme of the localization of the homozygosity tracks used as surrogate markers to predict the status of the 8p23.1 inversion in HapMap samples. **B** Frequencies predicted for 8p23.1 inversion in CEU (white bars) and ASN (black bars) populations.

Table 2. Tracks of homozygosity in 8p23.1 region extracted from the whole genome scan data performed in homozygous inverted individuals. The 8 SNPs that serve as surrogate markers close to the REPD duplicons are shown on the left. The 8 SNPs that serve as surrogate markers close to the REPP duplicons are shown in the right.

REPD		REPP	
SNP	Homozygous inverted haplotype	SNP	Homozygous inverted haplotype
rs17627505	C	rs1178061	T
rs1769237	G	rs1178247	C
rs2428	T	rs3885690	A
rs11774860	C	rs2409691	C
rs3827811	G	rs13266785	G
rs17154769	A	rs10282848	A
rs1876836	G	rs10503417	G
rs1039916	G	rs2409719	A

Assuming that meiotic recombination is suppresed in the heterozygous individuals for the inversion, we postulate that these 16 SNPs can be used as markers for the 8p23.1 inversion. Following this strategy we have downloaded the phased haplotypes from the 210 HapMap samples (including only the parents from the CEU and YRI trios) to predict the status of the inversion in these individuals.

In the population of European ancestry we found that 50% of the individuals are heterozygous for the inversion, that 8% are homozygous and 42% don't present the inverted allele (Fig. 2). Among the population of Asiatic origin the presence of the inversion is extremely high, with a 48% of the samples being heterozygous and a 36.5% homozygous for the 8p23.1 inversion (Fig. 2). Finally, the Yoruban individuals lack the presence of the 16 conserved markers. This is probably because our predictions are based on Caucasian ancestry individuals, and YRI samples might have different SNP profiles

segregating with the inversion which can not assessed by our approach, altought it has recently been reported by other authors that ~76% of yoruban individuals have the inversion .

Confirmation of the predictions for 8p23.1 inversion by FISH analysis

In order to confirm the validity of the 16 SNP markers for the presence of 8p23.1 inversion, we have genotyped the inversion by FISH analysis in a subset of HapMap samples. For that we chose 9 individuals, three of them (NA11831; NA12815; NA12155) which were predicted to be homozygous for the 8p23.1 inversion; another three (NA11993; NA06993; NA1994) heterozygous and the last three (NA11992; NA12057; NA1839) homozygous for the non-inverted status (Fig. 3).

The predicted genotype was confirmed in all 9 samples, although again some degree of mosaicism could be detected (Table 3). Thus, we conclude that the defined tracks of SNP homozygosity are reliable markers to predict the inverted allele of the 8p23.1 region.

Table 3. Percentages of each of the three possible genotypes observed within HapMap individuals in a subset of 9 samples. The predominant genotype is represented as a grey box.

HapMap sample	Homozygous non-inverted	Heterozygous	Homozygous inverted
NA11992	95%	5%	-
NA12057	67%	33%	-
NA11839	97%	3%	-
NA11993	-	96%	4%
NA06933	-	91%	9%
NA11994	-	75%	25%
NA11831	-	38%	62%
NA12815	-	35%	65%
NA12155	-	46%	54%

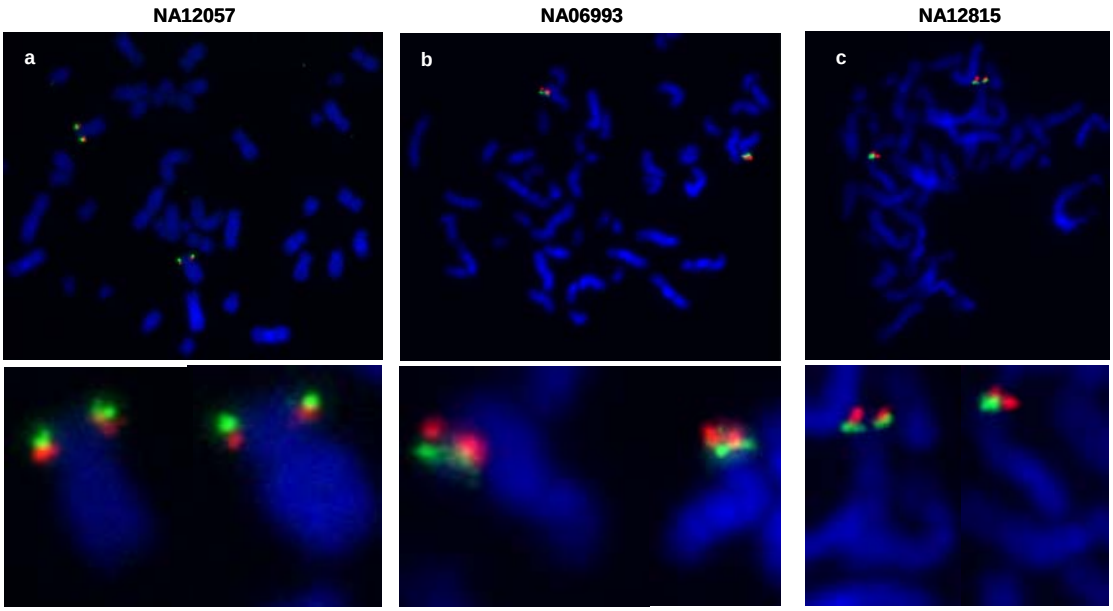

Figure 3. FISH analysis in HapMap samples. DNA probes were made from BAC clones RP11-399J23 (Green) and RP11-589N15 (Red). **A** Percentages of each of the three possible genotypes observed within HapMap individuals in a subset of 9 samples. The predominant genotype is represented as a grey box. In individuals NA11992, NA12057 and NA11839 the predominant genotype is the non-inverted status. Individuals NA11993, NA06933 and NA11994 have most of the metaphases heterozygous for the 8p23.1 inversion. In samples NA11831, NA12815 and NA12155 the most observed genotype is the homozygous inverted. **B** Metaphase FISH of three HapMap individuals, NA12057 as an example of non-inverted individual; NA06993 is heterozygous for the 8p23.1 inversion and NA12815 corresponds to a homozygous inverted individual.

Gene expression analysis of 8p23.1 genes in HapMap populations

Another aim of this study was to investigate if the inverted conformation has an effect on the expression of the genes contained in the 8p23.1 region. For this purpose we made use of the data produced by Stranger et al. regarding gene expression levels in HapMap individuals (Stranger et al., 2007) and we performed an association study. We used gene expression values from 26 genes (Supplementary Table 1) located around and within the 8p23.1 region

and we searched for any association between gene expression levels and the presence of the inversion. These analyses were carried out running the

SNPassoc R package on the CEU and ASN populations, where the 16 SNP signature to predict the inversion status can be observed.

Interestingly 4 genes NEIL2 (p=0.0003), MSRA (p=0.001), CTSB ($p=3.46x10^{-5}$) and BLK ($p=2.08x10^{-5}$) exhibit statistically significant expression level differences after Bonferroni correction for multiple testing ($p<0.001$), depending on the genotype of the inversion. The model of inheritance under which these genes were found to be differentially expressed is an additive model, with the exception of NEIL2 which follows a dominant inheritance pattern. Under the additive model the effect of the inversion on gene expression is gradually greater in individuals that have both chromosomes non-inverted, then in the heterozygous samples and finally the homozygous-inverted individuals. This data suggest that the inverted conformation has some effect on the expression of the genes embedded in the inverted fragment.

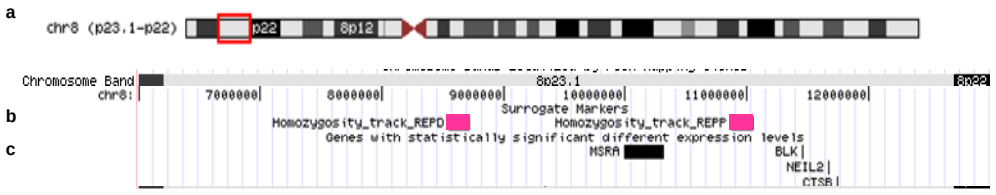

Figure 4. Scheme of the 8p23.1 region. **a** Red box indicates the localization of the 8p23.1 chromosome band. Coordinates of the region are also specified in base pairs. **b** Representation of the surrogate markers. The homozygosity tracks close to the distal (REPD) and proximal (REPP) segmental duplications are depicted as pink boxes. **c** Genes that are differentially expressed depending on the orientation of the 8p23.1 region are represented. **d** Image of the linkage disequilibrium blocks in the three HapMap populations. **e** Data reported by several studies corresponding to copy number polymorphisms and intermediate-sized structural variation found in the 8p23.1 region. On the bottom part the different sets of segmental duplications flanking 8p23.1 are shown.

Discussion

As it has already been reported in locus specific studies, inversions frequently occur in the human genome (Iafrate et al., 2004; Sebat et al., 2004; Feuk et al., 2005; Sharp et al., 2005; Tuzun et al., 2005; Conrad et al., 2006; McCarroll et al., 2006). A recent study has described that there are at least 35 recurrent inversions that are visible by light microscopy and six of them are known to disrupt a gene (Thomas et al., 2008). An important issue emerging from these studies is the mechanism underlying such common variants. As mentioned in the introduction, it is well known that the presence of high identity genomic sequences (low copy repeats or SDs) arranged in opposite orientations that flank certain regions of the genome, predispose to inversion events by non allelic homologous recombination (NAHR). What it is not so often taken into consideration is the presence of mitotic recombination leading to mosaicism in regions showing this type of genomic architecture. Mitotic recombination is an important mechanism under study that can complicate the interpretation of the rearrangements mediated by complex SDs, specially when the most commonly used material for study are transformed lymphoblastoid cell lines that exhibit a very high mitotic ratio. It has also been demonstrated that little requirement for long, identical homology blocks between paralogous DNA fragments is needed to produce exchanges by ectopic recombination (Lam and Jeffreys, 2006). This may be the reason why other studies focused on the same 8p23.1 region we have analyzed show some discrepancies with our findings (Deng et al., 2008)

Our FISH results in immortalized lymphocyte cells prepared from Spanish control individuals as well as HapMap cell lines, point out the presence of mosaicism regarding the 8p23.1 inversion that could appear during the lympholastoid cultures or either be due to a FISH artifact. Although the occurrence of somatic rearrangements seems an expectable event due to the extension and high level of homology between the SDs flanking the 8p23.1,

little is known about the incidence of this phenomenom in most regions of the genome. Several studies have suggested that this phenomenon, exemplified in the case of the alpha-globins where somatic deletions encompassing these genes arise by intrachromosomal homologous exchange, is common in blood and sperm (Lam and Jeffreys, 2006). Results form Flores also indicate that some cells within blood samples from normal individuals can undergo genomic rearrangements such as inversions and create genomic structural mosaics (Flores et al., 2007). These findings should expand our views about the plasticity of the genome.

Another aspect that arises from our FISH study of the 8p23.1 inversion is the elevated frequency of the inverted allele; by means of whole genome scan data, haplotype-based computational analyses and FISH experiments, we could infer and verify the orientation status of alleles in the 8p23.1 region. Chromosomes were initially studied by FISH, and surrogate markers for the inversion were identified by analyzing allelic association of 16 SNPs in the whole genome scan analysis, which delineated two tracks of homozygosity, in a group of individuals with known status for the 8p23.1 inversion, alleviating the need for further FISH. These 16 SNP markers tag the inversion and perfectly correlate with it in European and Asian ancestry HapMap samples. As long as these 16 markers were selected from a sample of Spanish individuals, the relatedness of CEU and ASN populations made them useful to predict the inversion in these individuals. Therefore, the method described here can be used in CEU and ASN population to tag the inversion, but in the YRI population, which exhibits a higher SNP diversity (Kidd et al., 2008), the absence of the 16 surrogate markers in linkage disequilibrium with inversion does not mean that the inversion is not present in these individuals. This fact argues in favor of different origins for this rearrangement. Thus, a 48% of ASN population and 50% of the CEU HapMap samples are found to be heterozygous for the inverted allele. This frequency is similar to what we found in the Spanish controls we have genotyped, where 50% are

heterozygous for the inverted conformation. Regarding the homozygous status of the inversion, this is present in a 36.5% of the ASN HapMap individuals, and in a 8% and 29% of the CEU HapMap and Spanish control samples, respectively.

These results altogether indicate that the recurrent 8p23.1 inversion is much more frequent than previously reported (Giglio et al., 2001), at least in the European population and that the current human genome reference assembly (Build 36) corresponds to the less common orientation of the 8p23.1 region.

Although inversions are generally found as neutral variants regarding their phenotypic effects, there are some exceptions where specific genes at the breakpoints are interrupted (Lakich et al., 1993; Iida et al., 2000; Saito-Ohara et al., 2002; Beiraghi et al., 2003; Sood et al., 2004; Tadin-Strapps et al., 2004). To investigate the possible effects of the allelic variation between the inverted and normal forms of the genes on 8p23.1 we performed and association study of gene expression levels and the genotype of the inversion in 150 HapMap individuals. We found four genes (*NEIL2, MSRA, CTSB, BLK*) that show statistically significant different expression levels ($p<0.0005$). Two of this genes, *NEIL2* and *MSRA*, are related to repair of oxidative damage (Moskovitz et al., 1996; Bandaru et al., 2002), and *CTSB* gene has been suggested to play a role in Alzheimer's disease (Esch et al., 1990). To which extend these gene expression differences we observe can influence the function of these genes and if they are directly related to the inversion of the region remains to be proved. Moreover, we can not discard that the degree of mosaicism present in the lymphoblastoid cell lines can modulate the expression levels of the genes analyzed. In addition we still have no evidences on how this inversion accompanied by any degree of mosaicism can affect the regulation of the transcription of the genes contained in the region.

In summary, using whole genome scan in homozygous inverted individuals previously genotyped by FISH, we have been able to describe surrogate markers to tag the 8p23.1 inversion in CEU and ASN populations. Moreover,

among the 26 genes we analyzed we could observe gene expression differences in four of them depending on the conformation of the region. We also highlight the presence of mosaicism regarding the inversion in most of the individuals genotyped by FISH. We postulate that this might be a common phenomenon which could occur in regions flanked by SDs, whose consequences have not been described, and need to be investigated.

Acknowledgements

Financial support was received from the Department of Universities, Research and Information Society (2005SGR00008) ("Generalitat de Catalunya"), from the European Union AnEUploidy project (grant number 037627). N. B. is a recipient of a BEFI fellowship from "Instituto de Salud Carlos III FIS-ISCIII"; J.M.M. was supported by the CRG under project SAF2002-00799 (Spanish Ministry of Science and Education) and by fellowship of the Danone Institute (Spain).

References

Aldred, P.M., Hollox, E.J. and Armour, J.A.: Copy number polymorphism and expression level variation of the human {alpha}-defensin genes DEFA1 and DEFA3. Hum Mol Genet (2005).

Ballana, E., Gonzalez, J.R., Bosch, N. and Estivill, X.: Inter-population variability of DEFA3 gene absence: correlation with haplotype structure and population variability. BMC Genomics 8 (2007) 14.

Bandaru, V., Sunkara, S., Wallace, S.S. and Bond, J.P.: A novel human DNA glycosylase that removes oxidative DNA damage and is homologous to Escherichia coli endonuclease VIII. DNA Repair (Amst) 1 (2002) 517-29.

Barber, J.C., Maloney, V., Hollox, E.J., Stuke-Sontheimer, A., du Bois, G., Daumiller, E., Klein-Vogler, U., Dufke, A., Armour, J.A. and Liehr, T.: Duplications and copy number variants of 8p23.1 are cytogenetically indistinguishable but distinct at the molecular level. Eur J Hum Genet (2005).

Bayes, M., Magano, L.F., Rivera, N., Flores, R. and Perez Jurado, L.A.: Mutational mechanisms of Williams-Beuren syndrome deletions. Am J Hum Genet 73 (2003) 131-51.

Beiraghi, S., Zhou, M., Talmadge, C.B., Went-Sumegi, N., Davis, J.R., Huang, D., Saal, H., Seemayer, T.A. and Sumegi, J.: Identification and characterization of a novel gene disrupted by a pericentric inversion inv(4)(p13.1q21.1) in a family with cleft lip. Gene 309 (2003) 11-21.

Bondeson, M.L., Dahl, N., Malmgren, H., Kleijer, W.J., Tonnesen, T., Carlberg, B.M. and Pettersson, U.: Inversion of the IDS gene resulting from recombination with IDS-related sequences is a common cause of the Hunter syndrome. Hum Mol Genet 4 (1995) 615-21.

Bosch, N., Escaramis, G., Mercader, J.M., Armengol, L. and Estivill, X.: Analysis of the multi-copy gene family FAM90A as a copy number variant in different ethnic backgrounds. Gene 420 (2008) 113-7.

Conrad, D.F., Andrews, T.D., Carter, N.P., Hurles, M.E. and Pritchard, J.K.: A high-resolution survey of deletion polymorphism in the human genome. Nat Genet 38 (2006) 75-81.

Chen, G.K., Slaten, E., Ophoff, R.A. and Lange, K.: Accommodating chromosome inversions in linkage analysis. Am J Hum Genet 79 (2006) 238-51.

Deng, L., Zhang, Y., Kang, J., Liu, T., Zhao, H., Gao, Y., Li, C., Pan, H., Tang, X., Wang, D., Niu, T., Yang, H. and Zeng, C.: An unusual haplotype structure on human chromosome 8p23 derived from the inversion polymorphism. Hum Mutat (2008).

Eichler, E.E., Nickerson, D.A., Altshuler, D., Bowcock, A.M., Brooks, L.D., Carter, N.P., Church, D.M., Felsenfeld, A., Guyer, M., Lee, C., Lupski, J.R., Mullikin, J.C., Pritchard, J.K., Sebat, J., Sherry, S.T., Smith, D., Valle, D. and Waterston, R.H.: Completing the map of human genetic variation. Nature 447 (2007) 161-5.

Esch, F.S., Keim, P.S., Beattie, E.C., Blacher, R.W., Culwell, A.R., Oltersdorf, T., McClure, D. and Ward, P.J.: Cleavage of amyloid beta peptide during constitutive processing of its precursor. Science 248 (1990) 1122-4.

Fellermann, K., Stange, D.E., Schaeffeler, E., Schmalzl, H., Wehkamp, J., Bevins, C.L., Reinisch, W., Teml, A., Schwab, M., Lichter, P., Radlwimmer, B. and Stange, E.F.: A chromosome 8 gene-cluster polymorphism with low human beta-

defensin 2 gene copy number predisposes to Crohn disease of the colon. Am J Hum Genet 79 (2006) 439-48.

Feuk, L., MacDonald, J.R., Tang, T., Carson, A.R., Li, M., Rao, G., Khaja, R. and Scherer, S.W.: Discovery of human inversion polymorphisms by comparative analysis of human and chimpanzee DNA sequence assemblies. PLoS Genet 1 (2005) e56.

Flores, M., Morales, L., Gonzaga-Jauregui, C., Dominguez-Vidana, R., Zepeda, C., Yanez, O., Gutierrez, M., Lemus, T., Valle, D., Avila, M.C., Blanco, D., Medina-Ruiz, S., Meza, K., Ayala, E., Garcia, D., Bustos, P., Gonzalez, V., Girard, L., Tusie-Luna, T., Davila, G. and Palacios, R.: Recurrent DNA inversion rearrangements in the human genome. Proc Natl Acad Sci U S A 104 (2007) 6099-106.

Giglio, S., Broman, K.W., Matsumoto, N., Calvari, V., Gimelli, G., Neumann, T., Ohashi, H., Voullaire, L., Larizza, D., Giorda, R., Weber, J.L., Ledbetter, D.H. and Zuffardi, O.: Olfactory receptor-gene clusters, genomic-inversion polymorphisms, and common chromosome rearrangements. Am J Hum Genet 68 (2001) 874-83.

Gimelli, G., Pujana, M.A., Patricelli, M.G., Russo, S., Giardino, D., Larizza, L., Cheung, J., Armengol, L., Schinzel, A., Estivill, X. and Zuffardi, O.: Genomic inversions of human chromosome 15q11-q13 in mothers of Angelman syndrome patients with class II (BP2/3) deletions. Hum Mol Genet 12 (2003) 849-58.

Gonzalez, J.R., Armengol, L., Sole, X., Guino, E., Mercader, J.M., Estivill, X. and Moreno, V.: SNPassoc: an R package to perform whole genome association studies. Bioinformatics 23 (2007) 644-5.

Hollox, E.J., Huffmeier, U., Zeeuwen, P.L., Palla, R., Lascorz, J., Rodijk-Olthuis, D., van de Kerkhof, P.C., Traupe, H., de Jongh, G., den Heijer, M., Reis, A., Armour, J.A. and Schalkwijk, J.: Psoriasis is associated with increased beta-defensin genomic copy number. Nat Genet 40 (2008) 23-5.

Iafrate, A.J., Feuk, L., Rivera, M.N., Listewnik, M.L., Donahoe, P.K., Qi, Y., Scherer, S.W. and Lee, C.: Detection of large-scale variation in the human genome. Nat Genet 36 (2004) 949-51.

Iida, A., Emi, M., Matsuoka, R., Hiratsuka, E., Okui, K., Ohashi, H., Inazawa, J., Fukushima, Y., Imai, T. and Nakamura, Y.: Identification of a gene disrupted by inv(11)(q13.5;q25) in a patient with left-right axis malformation. Hum Genet 106 (2000) 277-87.

Kidd, J.M., Cooper, G.M., Donahue, W.F., Hayden, H.S., Sampas, N., Graves, T., Hansen, N., Teague, B., Alkan, C., Antonacci, F., Haugen, E., Zerr, T., Yamada, N.A., Tsang, P., Newman, T.L., Tuzun, E., Cheng, Z., Ebling, H.M., Tusneem, N., David, R., Gillett, W., Phelps, K.A., Weaver, M., Saranga, D., Brand, A., Tao, W., Gustafson, E., McKernan, K., Chen, L., Malig, M., Smith, J.D., Korn, J.M., McCarroll, S.A., Altshuler, D.A., Peiffer, D.A., Dorschner, M., Stamatoyannopoulos, J., Schwartz, D., Nickerson, D.A., Mullikin, J.C., Wilson, R.K., Bruhn, L., Olson, M.V., Kaul, R., Smith, D.R. and Eichler, E.E.: Mapping and sequencing of structural variation from eight human genomes. Nature 453 (2008) 56-64.

Korbel, J.O., Urban, A.E., Affourtit, J.P., Godwin, B., Grubert, F., Simons, J.F., Kim, P.M., Palejev, D., Carriero, N.J., Du, L., Taillon, B.E., Chen, Z., Tanzer, A., Saunders, A.C., Chi, J., Yang, F., Carter, N.P., Hurles, M.E., Weissman, S.M., Harkins, T.T., Gerstein, M.B., Egholm, M. and Snyder, M.: Paired-end mapping reveals extensive structural variation in the human genome. Science 318 (2007) 420-6.

Lakich, D., Kazazian, H.H., Jr., Antonarakis, S.E. and Gitschier, J.: Inversions disrupting the factor VIII gene are a common cause of severe haemophilia A. Nat Genet 5 (1993) 236-41.

Lam, K.W. and Jeffreys, A.J.: Processes of copy-number change in human DNA: the dynamics of {alpha}-globin gene deletion. Proc Natl Acad Sci U S A 103 (2006) 8921-7.

Linzmeier, R.M. and Ganz, T.: Human defensin gene copy number polymorphisms: Comprehensive analysis of independent variation in alpha- and beta-defensin regions at 8p22-p23. Genomics (2005).

Lupski, J.R.: Genomic disorders: structural features of the genome can lead to DNA rearrangements and human disease traits. Trends Genet 14 (1998) 417-22.

McCarroll, S.A., Hadnott, T.N., Perry, G.H., Sabeti, P.C., Zody, M.C., Barrett, J.C., Dallaire, S., Gabriel, S.B., Lee, C., Daly, M.J. and Altshuler, D.M.: Common deletion polymorphisms in the human genome. Nat Genet 38 (2006) 86-92.

Moskovitz, J., Jenkins, N.A., Gilbert, D.J., Copeland, N.G., Jursky, F., Weissbach, H. and Brot, N.: Chromosomal localization of the mammalian peptide-methionine sulfoxide reductase gene and its differential expression in various tissues. Proc Natl Acad Sci U S A 93 (1996) 3205-8.

Navarro, A., Barbadilla, A. and Ruiz, A.: Effect of inversion polymorphism on the neutral nucleotide variability of linked chromosomal regions in Drosophila. Genetics 155 (2000) 685-98.

Osborne, L.R., Li, M., Pober, B., Chitayat, D., Bodurtha, J., Mandel, A., Costa, T., Grebe, T., Cox, S., Tsui, L.C. and Scherer, S.W.: A 1.5 million-base pair inversion polymorphism in families with Williams-Beuren syndrome. Nat Genet 29 (2001) 321-5.

Pritchard, J.K. and Przeworski, M.: Linkage disequilibrium in humans: models and data. Am J Hum Genet 69 (2001) 1-14.

Saito-Ohara, F., Fukuda, Y., Ito, M., Agarwala, K.L., Hayashi, M., Matsuo, M., Imoto, I., Yamakawa, K., Nakamura, Y. and Inazawa, J.: The Xq22 inversion breakpoint interrupted a novel Ras-like GTPase gene in a patient with Duchenne muscular dystrophy and profound mental retardation. Am J Hum Genet 71 (2002) 637-45.

Sebat, J., Lakshmi, B., Troge, J., Alexander, J., Young, J., Lundin, P., Maner, S., Massa, H., Walker, M., Chi, M., Navin, N., Lucito, R., Healy, J., Hicks, J., Ye, K., Reiner, A., Gilliam, T.C., Trask, B., Patterson, N., Zetterberg, A. and Wigler, M.: Large-scale copy number polymorphism in the human genome. Science 305 (2004) 525-8.

Sharp, A.J., Locke, D.P., McGrath, S.D., Cheng, Z., Bailey, J.A., Vallente, R.U., Pertz, L.M., Clark, R.A., Schwartz, S., Segraves, R., Oseroff, V.V., Albertson, D.G., Pinkel, D. and Eichler, E.E.: Segmental duplications and copy-number variation in the human genome. Am J Hum Genet 77 (2005) 78-88.

Sood, R., Bader, P.I., Speer, M.C., Edwards, Y.H., Eddings, E.M., Blair, R.T., Hu, P., Faruque, M.U., Robbins, C.M., Zhang, H., Leuders, J., Morrison, K., Thompson, D., Schwartzberg, P.L., Meltzer, P.S. and Trent, J.M.: Cloning and characterization of an inversion breakpoint at 6q23.3 suggests a role for Map7 in sacral dysgenesis. Cytogenet Genome Res 106 (2004) 61-7.

Stefansson, H., Helgason, A., Thorleifsson, G., Steinthorsdottir, V., Masson, G., Barnard, J., Baker, A., Jonasdottir, A., Ingason, A., Gudnadottir, V.G., Desnica, N., Hicks, A., Gylfason, A., Gudbjartsson, D.F., Jonsdottir, G.M., Sainz, J., Agnarsson, K., Birgisdottir, B., Ghosh, S., Olafsdottir, A., Cazier, J.B., Kristjansson, K., Frigge, M.L., Thorgeirsson, T.E., Gulcher, J.R., Kong, A. and

Stefansson, K.: A common inversion under selection in Europeans. Nat Genet 37 (2005) 129-37.

Stranger, B.E., Forrest, M.S., Dunning, M., Ingle, C.E., Beazley, C., Thorne, N., Redon, R., Bird, C.P., de Grassi, A., Lee, C., Tyler-Smith, C., Carter, N., Scherer, S.W., Tavare, S., Deloukas, P., Hurles, M.E. and Dermitzakis, E.T.: Relative impact of nucleotide and copy number variation on gene expression phenotypes. Science 315 (2007) 848-53.

Sugawara, H., Harada, N., Ida, T., Ishida, T., Ledbetter, D.H., Yoshiura, K., Ohta, T., Kishino, T., Niikawa, N. and Matsumoto, N.: Complex low-copy repeats associated with a common polymorphic inversion at human chromosome 8p23. Genomics 82 (2003) 238-44.

Tadin-Strapps, M., Warburton, D., Baumeister, F.A., Fischer, S.G., Yonan, J., Gilliam, T.C. and Christiano, A.M.: Cloning of the breakpoints of a de novo inversion of chromosome 8, inv (8)(p11.2q23.1) in a patient with Ambras syndrome. Cytogenet Genome Res 107 (2004) 68-76.

Taudien, S., Galgoczy, P., Huse, K., Reichwald, K., Schilhabel, M., Szafranski, K., Shimizu, A., Asakawa, S., Frankish, A., Loncarevic, I.F., Shimizu, N., Siddiqui, R. and Platzer, M.: Polymorphic segmental duplications at 8p23.1 challenge the determination of individual defensin gene repertoires and the assembly of a contiguous human reference sequence. BMC Genomics 5 (2004) 92.

Thomas, N.S., Bryant, V., Maloney, V., Cockwell, A.E. and Jacobs, P.A.: Investigation of the origins of human autosomal inversions. Hum Genet 123 (2008) 607-16.

Tuzun, E., Sharp, A.J., Bailey, J.A., Kaul, R., Morrison, V.A., Pertz, L.M., Haugen, E., Hayden, H., Albertson, D., Pinkel, D., Olson, M.V. and Eichler, E.E.: Fine-scale structural variation of the human genome. Nat Genet (2005).

Voight, B.F., Kudaravalli, S., Wen, X. and Pritchard, J.K.: A map of recent positive selection in the human genome. PLoS Biol 4 (2006) e72.

Zody, M.C., Jiang, Z., Fung, H.C., Antonacci, F., Hillier, L.W., Cardone, M.F., Graves, T.A., Kidd, J.M., Cheng, Z., Abouelleil, A., Chen, L., Wallis, J., Glasscock, J., Wilson, R.K., Reily, A.D., Duckworth, J., Ventura, M., Hardy, J., Warren, W.C. and Eichler, E.E.: Evolutionary toggling of the MAPT 17q21.31 inversion region. Nat Genet (2008).

Supplementary Table

Supplemetary Table 1. Genes analyzed in the association study between gene expression levels and the genotype for the 8p23.1 inversion, and their corresponding probes (http://www.sanger.ac.uk/humgen/genevar/)**.**

GENE	ILLUMINA PROBE
CSMD1	GI_41393594-S
MCPH1	GI_13375792-S
ANGPT2	GI_4557314-S
AGPAT5	GI_8922941-S
DEFB4	GI_13124885-S
THEX1	GI_31543183-S
TNKS	GI_4507612-S
MSRA	GI_13259538-S
RP1L1	GI_40255277-S
SOX7	GI_30581119-S
PINX1	GI_31982866-S
MTMR9	GI_33598962-S
AMAC1L2	GI_16876446-S
CTSB	GI_22538429-A
BLK	GI_33469981-S
GATA4	GI_33188460-S
NEIL2	GI_21450799-S
FDFT1	GI_31542631-S
DUB3	GI_41349931-S
FAM86B1	GI_42476336-I
FAM86B1	GI_37541806-S
FAM86B1	GI_41350215-S
LONRF1	GI_40217795-S
FAM90A	GI_42658889-S
SGCZ	GI_21040252-S
TUSC3	GI_30410789-A
TUSC3	GI_30410787-I
MSR1	GI_20357511-I
MSR1	GI_20357509-I
MSR1	GI_20357509-A
MYOM2	GI_4505314-S

Discussió

A continuació s'examinen detingudament els resultats presentats anteriorment, tenint en compte les principals contribucions d'aquesta tesi doctoral, així com els punts més controvertits. L'ordre seguit és el mateix que el dels resultats exposats per tal de facilitar la comprensió.

- Caracterització de les duplicacions segmentàries (DSs) que flanquegen la regió cromosòmica 8p23.1.

Tot i que avui en dia disposem de diferents estudis que han estat adreçats a obtenir un major coneixement de la complexa estructura genòmica que conforma la regió 8p23.1 i que han aportat informació de gran valor que ha ajudat a entendre la inestabilitat que pateixen les regions flanquejades per duplicons[105,108,148,149,180,181], cal posar en context des d'on va arrencar el projecte presentat.

En el moment que es van iniciar els estudis d'aquesta tesi la majoria d'aquests treballs encara no havien vist la llum. Així, la primera aproximació a l'estudi de la regió 8p23.1 en dur-se a terme va ser la realització d'un anàlisi minuciós de les DSs que la flanquegen per tal d'obtenir una millor caracterització d'aquests duplicons.

Un cop emmascarats els elements repetitius clàssics del genoma presents a la regió 8p23.1, a partir de l'alineament de la seqüència d'aquesta regió contra ella mateixa, es va poder detectar tota una sèrie de fragments molt idèntics localitzats a diferents porcions de la regió i que, per tant, conformen les DSs d'aquesta regió. Alguns d'ells pertanyien als gens de les alfa i beta defensines ja descrites, però fou una sorpresa

identificar un bloc de ~7 kb, repetit reiteradament un gran nombre de vegades, i que corresponia a un gen fins aleshores desconegut i anotat a l'assemblatge de referència com a gen de còpia única (*FAM90A1*) al cromosoma 12, i no pas al cromosoma 8. A partir d'aquest punt es van analitzar les DSs del voltant del gen *FAM90A1* al cromosoma 12 i es va poder comprovar que tenien una similitud important amb les seqüències de la regió 8p23.1 (DSs intercromosòmiques). Així doncs, amb la sospita que la formació de les DSs de 8p23.1 havien permès l'expansió d'aquest gen, es va iniciar la caracterització del que avui en dia és la família gènica específica de primats *FAM90A*. L'anàlisi es va començar en la versió de l'abril del 2003 (UCSC hg15, NCBI build 33) de l'assemblatge, però és perfectament vàlid en les versions de l'assemblatge posteriors. Cal tenir present que la descripció de *FAM90A* duta a terme està basada en l'individu de l'assemblatge de referència i que no es pot fer extensible a qualsevol individu, ja que els resultats aquí presentats mostren que aquesta família és altament polimòrfica entre individus.

Expansió en tàndem de la família gènica *FAM90A*

Els fenòmens de duplicació juguen un paper essencial en l'evolució dels genomes (veure capítol 2 de la introducció)[43,182]. La descripció duta a terme de la família *FAM90A,* composada per almenys 25 membres per genoma haploide en humans (UCSC hg18), representa un cas extrem de duplicació gènica que sembla haver-se expandit en tàndem. Els mecanismes que hi ha al darrera de l'expansió de variants en número de còpia no estan del tot clars, i es necessiten més estudis per tal d'aclarir quin paper juguen les duplicacions segmentàries en l'evolució d'aquestes variants.

Família gènica *FAM90A*: Subfamília I i Subfamília II

A partir de la descripció de les 24 còpies paràlogues a *FAM90A1* presents en les DSs de 8p23.1 es van poder identificar dues subfamílies, subfamília I (formada per la còpia *FAM90A1* del cromosoma 12 i les dues còpies úniques de REPP), i subfamília II (formada per les 22 còpies distribuïdes en clústers dels duplicons distals REPD). La principal diferència entre els membres d'aquestes dues subfamílies és el seu primer exó, la regió 5' no traduïda i la regió més upstream.

Aquestes diferències fan pensar en la possibilitat que aquestes subfamílies tinguin un perfil d'expressió específic de teixit donat que la regió "upstream" on hi acostuma a haver els elements reguladors de l'expressió són diferents en l'una i l'altra. Els experiments que es van realitzar però, no ho van poder ni constatar ni desmentir. Una de les grans dificultats alhora de dissenyar experiments en regions duplicades, és la impossibilitat de ser específic, ja que l'elevada identitat de seqüència entre paràlegs no permet distingir-ne uns dels altres. Aquest fet es veu encara més agreujat quan el disseny dels experiments està basat en l'individu/s de l'assemblatge de referència i després en canvi, l'individu/s analitzats són uns altres. Si a part de les variants de seqüència entre còpies paràlogues, hi afegim la dificultat que sovint les DSs contenen gens que són variables en número de còpia d'individu a individu, com mostren els resultats de l'electroforesi de camp polsant per *FAM90A,* mai es pot tenir la convicció que el disseny de l'experiment és 100% acurat i vàlid per a totes les mostres. Per tant, en l'estudi d'aquestes regions cal anar molt en compte i utilitzar diferents estratègies, de manera que els resultats convergeixin en una mateixa direcció per tal que siguin concloents. Així, tot i la dificultat en el disseny dels experiments destinats a amplificar fragments concrets situats en regions de DSs, es van arribar a obtenir oligonuclèotids per amplificar membres de les subfamílies I i II de manera específica. Amplificacions per RT-PCR en 13 teixits diferents van

mostrar l'expressió en tots els teixits de les dues subfamílies sense observar cap patró d'expressió diferent entre elles. És probable que la utilització de mètodes quantitatius sí que mostrin diferències en l'expressió de la Subfamília I i la Subfamília II en molts teixits, i tampoc no es pot descartar que aquests perfils d'expressió siguin variables en diferents estadis del desenvolupament pels membres de les dues subfamílies.

Pseudogenització de membres de FAM90A

Un cop generades per un procés de duplicació, algunes de les còpies que conformen la família de *FAM90A* van anar patint mutacions que en van alterar la funcionalitat i les van convertir en pseudogens. Això és el que es postula per cinc membres: Ψcòpia 3, Ψcòpia 7, Ψcòpia 9, Ψcòpia 23 i Ψcòpia 24. En el cas de les Ψcòpies 3, 23 i 24, les mutacions causen l'aparició d'un codó STOP, és a dir que es tracta de mutacions sense sentit. Per altra banda en les Ψcòpies 7 i 9 el que trobem són insercions i delecions d'un sol parell de bases nucleotídiques, donant lloc a l'alteració de la pauta de lectura. Si en canvi ens centrem en els acceptors i donadors d'splicing, trobem 10 còpies (HsaCopy1-5, 10-11 i 20-22) amb una mutació T→C en el donador d'splicing de l'exó 5. Aquest canvi però produeix la creació d'un nou donado d'splicing alternatiu (GC) que no tindria per què afectar la proteïna codificada per aquestes còpies[183]. Aquesta situació pot ser diferent en cada individu ja que aquest gen correspon a una variant en número de còpia, a més, també és més que probable que hagi ocorregut i ocorri conversió gènica entre les diverses còpies.

Potencial codificant de la família *FAM90A*

Les anàlisis de genòmica comparada mostren que *FAM90A* és una família gènica específica de primats i és absent en el genoma d'altes mamífers

com el ratolí, la rata, el gos, la vaca o la gallina. Els estudis basats en la comparació de les taxes de mutació que dónen lloc a canvi d'aminoàcid (Ka o substitucions no sinònimes) i les que no (Ks o substitucions sinònimes), van permetre determinar el grau de conservació al llarg de l'evolució de les seqüències proteiques codificades per aquests gens. Com a mínim un membre de *FAM90A* a cada espècie mostra una taxa Ka/Ks <1, és a dir que el més probable és que codifiquin per proteïnes funcionals. Aquest és el cas de *FAM90A1* (Ka/Ks=0.35), HsaCopy10-11 i HsaCopy21-22 en humans (mitjana de Ka/Ks=0.63), i els clústers de la Subfamília II que s'han pogut identificar en ximpanzés (mitjana Ka/Ks=0.48), i la còpia de macac MmuCopy1 (Ka/Ks=0.78). En general la mitjana de Ka/Ks de la família *FAM90A* és força alta (0.91), de manera que es pot assumir que segueix un model d'evolució neutre. Finalment, els resultats obtinguts d'analitzar les taxes de Ka/Ks de la regió codificant de *FAM90A* van mostrar que la regió C-terminal presenta una taxa inferior (Ka/Ks=0.77) que no pas la regió central (Ka/Ks=0.92) o N-terminal (Ka/Ks=0.99).

Expansió *de FAM90A* en primats

A partir de l'hibridació amb una sonda per detectar la família gènica *FAM90A* amb la tècnica de Southern Blot, es van comparar els patrons corresponents als diferents clústers en primats superiors i humans. Així es va observar que l'orangutan (~14 milions d'anys de divergència amb els humans), el goril·la (~10 milions d'anys de divergència) i el ximpanzé (~5 milions d'anys de divergència) tenen menys còpies de *FAM90A* que no pas els humans, tot i que els ximpanzés mostren un patró de bandes (clústers de *FAM90A*) més ampli i pròxim als humans.

Aquestes troballes suggereixen que ha existit una expansió de la família *FAM90A* en el llinatge dels primats a partir de processos de duplicació, que és una de les principals forces alhora de generar nous

gens que contribueixen en diversos processos biològics i faciliten l'evolució a partir d'organismes primitius a organismes complexes[182,184]. Tant els experiments per FISH, com de PCR quantitativa van permetre corroborar que aquesta família gènica s'ha anat expandint des del macac fins als humans.

LTRs que flanquegen els gens *FAM90A*

L'arquitectura genòmica de les seqüències que flanquegen cadascuna de les còpies de la Subfamília II consisteix en dos LTRs (veure capítol 1) que podrien haver jugat un paper en l'expansió dels duplicats de *FAM90A.*

Aquests LTRs en concret, pertanyen a una família de retrovirus endògens (HERV), que conformen un 8% del genoma humà i provenen d'elements semblant a retrovirus que es van quedar fixats en la línea germinal fa milions d'anys després d'una infecció per retrovirus exògens [185]. D'entre les diferents famílies d'HERVs, les que es troben en els clústers de *FAM90A* són del tipus HML-5, que corresponen a seqüències províriques detectades en les mones del Vell i Nou Món però no en els prosimis, essent els betaretrovirus més antics dels genoma de primats coneguts a data d'avui[186]. Amb aquestes dades podem postular que a partir d'aquest moment la família *FAM90A* hauria pogut iniciar el procés d'expansió.

Per altra banda, el primer i segon exó de la Subfamília II i I respectivament, contenen un element repetitiu LINE (veure capítol 1) de la família L1, l'única família LINE que roman activa en humans. Si tenim en compte que a la maquinària dels LINE se li atribueix la majoria de la transcripció reversa que existeix al genoma, incloent la que facilita el procés de transposició, és temptador especular sobre el paper dels LINE com a elements propiciadors de la duplicació i expansió de *FAM90A* en primats.

Hipòtesi sobre l'origen i evolució de *FAM90A* en primats

A partir de l'anàlisi de "contigs" corresponents a 8p23.1 i les regions sintèniques en macac i ximpanzé, juntament amb les dades de genòmica comparada obtingudes per FISH i Q-PCR en macac, orangutan, goril·la i ximpanzé, es va reconstruir el passat evolutiu de les DSs de 8p23.1, intentant seguir els diferents reordenaments que han patit els membres de la família *FAM90A* i s'ha hipotetitzat un model segons el qual aquesta família es podria haver originat i expandit.

Sembla clar doncs, a partir de dades genòmiques i experimentals, que els membres de la subfamília II haurien precedit els de la Subfamília I. En el model proposat, gràcies a la informació obtinguda de l'assemblatge de la seqüència del macac, podem postular que en algun ancestre dels Catarrhini trobaríem per una banda una còpia corresponent a la subfamília II en el cromosoma VIII, però no hi hauria la presència de cap element LTR en els extrems del gen. Per altra banda, no existiria encara cap còpia de la Subfamília I, però si la seqüència 5' "extra" (característica de la Subfamília I), que no comparteixen els membres d'ambdues famílies. Aquesta seqüència està formada per l'exó 1, un element AluSx i un element MiRb, i en el macac la trobem en un intró del gen *ALG1* localitzat en el cromosoma XVI. Segons el model proposat, el següent esdeveniment que hauria succeït seria la inserció d'un LTR a l'extrem 3' de la còpia de la subfamília II en el llinatge dels hominoides. A partir d'aquí s'hauria produït una duplicació de tot el segment i les dues subfamílies de *FAM90A* haurien divergit. Sota aquest model, el membres de la Subfamília I haurien nascut al voltant de la divergència de l'orangutan amb els primats superiors africans, fa uns 15 milions d'anys, per un procés de fusió de la "seqüència extra" de l'intró del gen *ALG1* amb una de les còpies de la Subfamília II. Llavors, diferents processos duplicatius, i esdeveniments de pèrdua gènica haurien afectat a la regió, donant lloc a

les diferents còpies de la Subfamília I que es poden observar avui en dia en els genomes dels humans i ximpanzés.

Per una altre costat, el precursor de la Subfamília II, a través de recombinació homòloga no al·lèlica (NAHR) o per processos semblants a l'"slippage" de la polimerasa durant la replicació, hauria creat duplicacions en tàndem d'aquests gens.

De fet, a la literatura tenim un altre exemple d'aquest procés de creació d'un nou gen específic de primats a partir d'un procés de fusió de dues seqüències no relacionades. Seria el cas del gen *MCH* de l'hormona concentradora de melanocortina[187]. També hi ha antecedents de famílies gèniques que s'han anat expandint en els primats, com els zinc fingers que es troben dispersos al cromosoma 19[188], o la família gènica *NBPF* ("neuroblastoma breakpoint gene family")[189], la família morfeus del cromosoma 16[42] o l'amplificació extrema que han patit els dominis proteics DEF1220 en humans[102].

Per què tant d'interès en una família gènica de funció desconeguda?

Si bé és veritat que l'únic domini proteic conegut de la família gènica *FAM90A* és un domini zinc finger del tipus C2H2, amb afinitat d'unió a DNA i RNA que ens podria encaminar a pensar que pot desenvolupar alguna funció semblant a la dels factors de transcripció, no existeixen evidències experimentals que ho corroborin. Així, aquesta família roman amb el nom descriptiu que fa al·lusió a la seva identitat de seqüència ("family with sequence similarity 90") i no pas a la seva funció.

Tot i no conèixer la seva funció, hi ha una sèrie de motius que fan especialment atractiu l'estudi de *FAM90A.* Per una banda, l'elevat número de còpies presents a l'assemblatge de referència (estimat a partir d'aquest treball en 25 per genoma haploide) fa que es tracti d'una família gènica poc corrent. Aquest fet tan podria ser degut a un origen molt recent d'aquestes duplicacions, i per tant les còpies no haurien tingut temps de

desaparèixer, o bé, a la fixació dels diferents membres de *FAM90A* degut a l'adquisició d'algun avantatge evolutiu. Altres factors que fan de *FAM90A* una família gènica interessant són: la seva especificitat en el genoma dels primats (podria tractar-se d'un gen amb una funció clau per l'especiació dels primats), la seva localització en una de les regions més dinàmiques del genoma humà (8p23.1), l'expansió que ha patit en els últims 15 milions d'anys i el seu elevat grau de polimorfisme.

L'atenció dedicada a l'estudi d'aquesta família gènica ha estat de gran utilitat alhora d'aprofundir en el coneixement del proces d'aparició de nous gens, com la generació de la Subfamília I de *FAM90A* a partir d'un procés de fusió. O l'expansió de còpies gèniques a partir de processos duplicatius específics de cada espècie, que alhora generen grans extensions genòmiques d'elevada identitat que donen peu a la NAHR. I precisament són aquests mecanismes de NAHR que poden generar gens variables en número de còpia. En definitiva, el conjunt de processos que han permès la creació i expansió de *FAM90A* al llarg dels duplicons de 8p23.1, serveixen com a exemple del que pot haver ocorregut en d'altres regions del cromosòmiques, i són de gran utilitat alhora de comprendre una mica més l'evolució i grau de plasticitat del nostre genoma.

- Caracterització de la variabilitat en número de còpia de *FAM90A* en diferents poblacions humanes

Arrel dels resultats del primer treball, on a partir de l'anàlisi dels diferents clústers de *FAM90A* mitjançant electroforesi de camp polsant i Southern blot, es fa patent l'elevat grau de polimorfisme en el número de còpies de *FAM90A,* es va dur a terme un segon estudi per tal de determinar aquesta variabilitat en diferents poblacions humanes. Amb aquesta finalitat es van utilitzar les mostres del projecte HapMap, que són de gran utilitat alhora

de realitzar estudis sobre variants genètiques en poblacions provinents de diferents orígens geogràfics. L'estudi previ va permetre observar que la família gènica *FAM90A* s'expressa en almenys 13 teixits, incloent també línees cel·lulars de limfòcits immortalitzats.

Dificultats tècniques per quantificar la variabilitat de gens de còpia múltiple

Alhora de quantificar el número de còpies de *FAM90A* trobem diferents limitacions tècniques. Per una banda, els assajos per determinar la presència o absència de cadascun dels 25 membres de la família en un individu en concret no són viables ja que l'elevat grau d'identitat de seqüència entre ells (>93%) impedeix dissenyar sondes específiques. Per l'altra, la quantificació de manera absoluta mitjançant mètodes com l'amplificació simultània depenent de lligació (multiplex ligation-dependent probe amplification o MLPA) no són útils ja que la tècnica no és prou fina per a discriminar la dosi de segments amb números de còpia tan elevats, de 50 a 63 còpies posem per cas.

És per aquest motiu que es va triar la tècnica de PCR en temps real per a fer una quantificació relativa de la variabilitat de *FAM90A.* Tot i que normalment aquesta tècnica s'utilitza per quantificar nivells d'expressió (a nivell de mRNA), cada cop s'utilitza més per a quantificar número de còpia de DNA mitocondrial i CNVs al genoma nuclear. L'anàlisi es va realitzar en una mostra de 260 individus HapMap provinents de tres orígens poblacionals: els CEU d'origen europeu, el ASN (població asiàtica de la Xina i Japó que per similitud genètica es poden considerar un sol grup) i els YRI d'origen Africà.

PCR a temps real del gen *FAM90A* en diferents poblacions humanes

Per tal de augmentar el grau de certesa dels resultats de la variabilitat en número de còpia de *FAM90A* mitjançant PCR a temps real, es van utilitzar

dues sondes diferents que hibriden als exons 2 i 6 de les diferents còpies. La utilització de sondes provinents de la "Universal Probe Library" (Roche, Mannheim) van permetre estalviar temps alhora de posar a punt l'assaig.

Els resultats utilitzant una sonda i l'altre van mostrar una elevada correlació, de manera que es va confirmar la sensibilitat de les sondes i utilitat del mètode per dur a terme quantificacions de CNVs a nivell genòmic.

Variabilitat intra i inter-poblacional del número de còpia de *FAM90A*

La quantificació relativa del número de còpies de FAM90A amb la metodologia esmentada, mostrà que la població amb menor número de còpies és l'asiàtica, mentre que els individus africans tenen de mitjana una major quantitat de còpies gèniques de *FAM90A,* i fou justament aquesta última població la que presentà diferències estadísticament significatives en número de còpia de *FAM90A* respecte les poblacions CEU i ASN (F-test, $p<0.0001$). Aquests resultats corroboren altres estudis realitzats en mostres HapMap on els individus YRI resultaen ser els més heterogenis a nivell genètic[190].

Per tal de poder dur a terme la quantificació de gens variables en número de còpia, es necessita sempre un individu de referència a partir del qual comparar els valors de la resta d'individus. En el nostre cas es va utilitzar l'individu NA12892, d'origen europeu, ja que presentava un número de còpies proper a la mitjana de la població. A partir dels valors de quantificació en aquest individu es va poder establir que les diferents mostres HapMap contenen un número de còpies de *FAM90A* d'entre -6.3 vegades menys (l'individu que conté menys còpies), i fins a 2.1 vegades, l'individu que conté més membres de *FAM90A*. Tot i que no podem parlar de números de còpia absolut, si que podem tenir una idea de fins a quin punt aquesta família gènica és altament variable a nivell interindividual i interpoblacional.

Correlació entre el número de còpies de FAM90A i els seus nivells d'expressió

Una de les grans qüestions que sorgeixen a mesura que es van identificant variants en número de còpia, és si aquesta variabilitat es veu reflectida a nivell transcripcional. De no ser així, semblaria que les conseqüències d'aquestes variants tindrien poc impacte a nivell funcional. Sorprenentment, estudis a nivell de tot el genoma fets fins el moment han revelat que només entre un 8% i un 18% de la variabilitat heretable a nivell de trànscripts és deguda a CNVs[191], tot i que calen futures aproximacions amb punts de trencament de les CNVs més ben delimitats per determinar si aquests percentatges són prou acurats. El que si que sembla clar, és que la variabilitat a nivell d'expressió que es pot capturar a nivell de SNPs no es solapa amb la variabilitat que es captura a nivell de CNVs[191]. De manera que es necessiten estudis basats en les dues aproximacions per tal d'entendre la complexitat transcripcional del genoma humà. Fins i tot hi ha estudis on es mostra que nivells alts d'expressió correlacionen amb CNVs de baix número de còpia[191].

En el cas de *FAM90A,* sembla que la variabilitat en número de còpia explicaria un percentatge molt petit de la variabilitat a nivell d'expressió (r^2=0.05, p=0.002). De fet, per altres CNVs de la regió, concretament el locus *DEFA1A3* , tampoc sembla existir una correlació entre el número de còpies i els seus nivells d'expressió[148]. La baixa correlació entre els nivells d'expressió de *FAM90A* i el seu número de còpia podria ser degut a una regulació complexa a nivell de la transcripció dels mRNAs o a la presència de factors reguladors en trans[192]. Per altra banda cal tenir en compte que l'estudi s'ha dut a terme en limfòcits, i caldria explorar d'altres teixits i a ser possible, en diferents estadis cel·lulars, i del desenvolupament de l'individu (període fetal, infància i edat adulta), per poder ampliar o limitar la baixa correlació trobada en limfòcits.

- Inversió polimòrfica de la regió cromosòmica 8p23.1

A la literatura trobem diversos estudis que mostren que les inversions són un tipus de reordenament present en tots els humans[49,87,88,96,107,193,194]. D'aquestes, se n'han descrit fins a 35 que es produeixen de manera recurrent i que són detectables al microscopi, de les quals vuit tenen algun efecte fenotípic, ja que quan s'originen disrupcionen algun gen (veure apartat 3 de la introducció)[112].

Com s'originen les inversions?

En el capítol 3 de la introducció s'explica com la presència de DSs en orientació oposada pot donar lloc a la inversió de tot el fragment comprés entre els duplicons mitjançant NAHR. El que no se sol contemplar, és la possibilitat que existeixi recombinació a nivell mitòtic que pugui originar igualment la inversió de les regions flanquejades per duplicons orientats en mirall. Si aquests fenòmens tenen lloc durant la divisió cel·lular, les regions amb l'arquitectura genòmica esmentada, podrien patir reordenaments a l'atzar, donant lloc a mosaics cel·lulars. La recombinació mitòtica doncs, és un fenomen poc estudiat però que pot complicar la interpretació dels reordenaments mediats per DSs complexes. El que si que es coneix avui en dia és que petites extensions d'homologia de centenars de parells de bases són suficients per donar lloc a intercanvis mitjançant recombinació ectòpica[195].

Estudis en diferents poblacions de la inversió polimòrfica de 8p23.1

Existeixen diversos estudis centrats en l'anàlisi de la inversió polimòrfica que afecta la regió cromosòmica 8p23.1. Al 2001, Giglio et *al.* van establir la seva freqüència en heterozigosi en el 26% en una mostra de 72 individus d'origen europeu[105]. En aquest estudi es postula que la presència de receptors olfactoris en els duplicons REPP i REPD de 8p23.1,

serien els que induirien aquesta reorganització cromosòmica. Avui en dia, es disposa d'una caracterització molt més acurada d'aquests duplicons, i tant els gens de les alfa defensines, com de les beta defensines, com de *FAM90A* podrien actuar com a substrats d'aquest reordenament [148,149,157,180,196]. Dos anys més tard, un estudi dut a terme en població japonesa va reconstruir el mapa físic de 8p23.1 a partir de BACs. La freqüència de la inversió en aquesta població és una mica superior, del 34% en heterozigosi (17 individus d'una mostra de 50) i del 10% en homozigosi (5 individus de 50). Un estudi recent realitzat en les mostres HapMap proposa que el percentatge d'individus heterozigots en població d'origen europeu augmenta fins al ~50%[197]. Si afegim els individus homozigots per la inversió, els resultats indiquen que la conformació present en la seqüència de referència representa l'orientació minoritària.

Mosaïcisme per la inversió de 8p23.1

En el nostre estudi de la inversió que afecta la regió 8p23.1, es va poder observar, mitjançant FISH, que existeix un cert grau de mosaïcisme pel que fa a la presència de la inversió. Aquest podria ser un motiu pel qual d'altres estudis centrats en aquesta regió mostren alguna discrepància respecte els nostres resultats[197]. És a dir, que si existeix recombinació a nivell mitòtic, en anar-se dividint les cèl·lules en cultiu s'anirien produint diferents mosaics de cèl·lules, amb la inversió i sense la inversió, contínuament. La proporció d'aquestes dues poblacions cel·lulars variarà en cada cultiu segons les condicions d'incubació, nombre de passatges, etc. Tant l'estudi realitzat per Deng et *al.* com el nostre, s'han dut a terme en línees cel·lulars de limfoblasts dels individus HapMap i per tant, l'estudi d'una mateixa mostra en un laboratori i en un altre pot ser que condueixi a resultats diferents degut a aquesta inestabilitat genòmica de la regió.

Tot i que a priori, degut a la llarga extensió i l'elevat nivell d'homologia de les DSs de 8p23.1, sembla prou factible l'ocurrència de reordenaments somàtics a la regió, es coneix ben poc la freqüència d'aquests fenòmens a la majoria de regions del genoma. Tot i això, d'altres exploracions han suggerit la l'ocurrència d'aquest fenomen, com és el cas de les alfa-globines, on l' intercanvi intracromosòmic de regions homòlogues en sang i esperma dóna lloc a delecions somàtiques[195]. També Flores et al. han detectat la presència de mosaics genòmics a nivell estructural, més concretament d'inversions, en línees sanguínies d'individus sans[198].

Tots aquests fenòmens de mosaïcisme, que fins ara semblaven limitats a produir-se a nivell de recombinació meiòtica, amplien l'espectre de plasticitat que posseeix el genoma.

Marcadors de la inversió del fragment genòmic 8p23.1

Per tal d'esbrinar si es podia prescindir de la tècnica de FISH alhora de detectar la inversió a 8p23.1, ja que tot i que conclusiva, és laboriosa i de llarga durada, es va idear una nova aproximació. A partir de les dades generades de la genotipació per FISH de la inversió de 8p23.1 en una mostra de 24 individus sans de població espanyola, es va hibridar un xip amb 370.000 sondes en sis dels individus que van resultar homozigots per la inversió. L'objectiu d'aquesta estratègia era identificar algun bloc d'SNPs que estigués en desequilibri de lligament amb la inversió. La conservació d'aquests blocs seria més plausible en aquestes regions gràcies al fenomen de supressió de la recombinació en la meiosi dels individus heterozigots per la inversió[199,200] (un 50% dels individus segons les nostres dades).

L'anàlisi detallat de la genotipació d'aquests sis individus va permetre identificar 16 SNPs que correlacionen amb la inversió. Aquests SNPs es troben en homozigosi en la conformació invertida de la regió i es

localitzen en l'interior del fragment invertit a prop, però no enmig, dels duplicons flanquejants de 8p23.1.

Predicció de la inversió en mostres de HapMap

A partir de la detecció dels 16 SNPs com a marcadors de la inversió, es va predir la conformació de la regió en les mostres HapMap. Aquest és un aspecte importat, ja que hi ha molt poques inversions la freqüència de les quals hagi estat ben estudiat en diferents poblacions.

Aquesta caracterització només es va poder dur a terme en individus CEU i ASN (150 individus en total), ja que en les mostres africanes no existeix l'haplotip conservat que conformen aquests 16 SNPs en població europea i asiàtica. Aquest fet pot ser indicatiu de que la inversió pot haver-se originat múltiples vegades, i segons el seu context geogràfic, entre d'altres factors, uns SNPs o uns altres, s'hauran quedat fixats en la conformació invertida.

Una vegada més, les prediccions van mostrar l'elevada freqüència de l'al·lel invertit respecte estudis anteriors. Si bé l'homozigositat per la inversió és més elevada en individus d'origen asiàtic (36.5%) que no pas en individus d'origen europeu (8%), la freqüència dels individus heterozigots volta el 50% en ambdues poblacions. L'anàlisi per FISH en un subgrup dels individus HapMap que havien estat predits per la inversió, va permetre validar els 16 SNPs com a marcadors fiables de la inversió.

Efecte de la inversió de 8p23.1 sobre l'expressió dels gens de la regió

Sembla evident que cal qüestionar-se si el canvi en l'orientació de tota una regió genòmica, en aquest cas d'unes 4 Mb, pot tenir conseqüències a nivell de la transcripció dels gens que es troben al seu interior i en el seu entorn.

Per aquest motiu, es va procedir a realitzar un estudi d'associació entre els nivells d'expressió dels gens de la regió 8p23.1 i la presència de

la inversió en els mateixos 150 individus pels quals es va predir l'estatus de la inversió (absència, heterozigosi o homozigosi). Els resultats van mostrar que els gens *NEIL2, MSRA, CTSB i BLK*, es troben expressats de manera diferencial atenent al genotip per la inversió (test de raó de verosimilitud, $p<0.0005$). Dos d'aquests gens, *NEIL2* i *MSRA* estan relacionats amb la reparació del dany oxidatiu[201,202], i *CTSB* s'ha postulat que pot jugar algun paper en la malaltia d'Alzheimer[203]. Fins a quin punt les diferències que s'observen a nivell d'expressió poden influenciar la funció d'aquests gens, i si són conseqüència directe de la inversió o de qualsevol altre variant existent a la zona, són aspectes que caldrà comprovar.

Per altra banda cal tenir en compte que la presència de mosaïcisme per la inversió en la gran majoria de mostres analitzades, pot jugar un paper en la modulació dels nivells d'expressió, i el mateix pot passar in vivo. Fins al moment, es desconeix com la inversió de la regió 8p23.1 acompanyada de diferents graus de mosaïcisme pot afectar la regulació de la transcripció dels gens de la regió.

Conclusions

Les conclusions generals que es poden extreure del treball presentat en aquesta tesi doctoral són les següents :

1. S'ha identificat una nova família gènica específica de primats, *FAM90A,* que en humans està formada per 22 membres agrupats en 4 clústers (A, B, C i D) als duplicons distals REPD (HsaCopy1-HsaCopy22), 2 membres als duplicons proximals REPP (HsaCopy23 i HsCopy24) i un membre al cromosoma 12p13.31 (*FAM90A1*), segons l'assemblatge de referència vigent (UCSC hg18).

2. La família gènica *FAM90A* es pot classificar en dues subfamílies: la Subfamília I, formada per les dues còpies dels duplicons proximals REPP i la còpia localitzada al cromosoma 12p13.31, i la Subfamília II, formada per la resta de membres agrupats en clústers a REPD. Aquestes dues subfamílies divergeixen en els 1036 parells de bases inicials que inclouen un exó no traduït, un element repetitiu AluSx, i un element repetitiu MIRb en els membres de la Subfamília I.

3. Existeix una pauta de lectura oberta conservada en 20 dels 25 membres que formen la família gènica *FAM90A.* La seqüència codificant comprendria els exons 3, 4, 5 i 6 i la seva pauta de lectura té el potencial per codificar per una proteïna de 464 aminoàcids.

4. Diferents mutacions a nivell de seqüència han conduit al truncament de la pauta de lectura dels membres HsaCopy3, HsaCopy7, HsaCopy9, HsaCopy23 i HsaCopy24.

5. Els clústers de *FAM90A* són altament polimòrfics entre els individus humans de la població general.

6. El membre *FAM90A1* del cromosoma 12 s'expressa en tots els teixits testats: colon, cor, cervell, ronyó, fetge, pulmó, ovari, placenta, pròstata, melsa, testicle i timus. Els membres HsaCopy23 i HsaCopy24 de la Subfamília I s'expressen en els mateixos teixits excepte cor, testicle, placenta i pròstata. Els membres de la Subfamília II s'expressen en tots els teixits analitzats excepte ronyó i pulmó.

7. L'estudi de l'evolució de les DSs que flanquegen 8p23.1 en macac, orangutan, goril·la i ximpanzé mostra que aquestes duplicacions, així com la família gènica *FAM90A,* estan presents en els cromosomes homòlegs a HSA8 i HSA12 en les diferents espècies, i que tant en humans com en ximpanzés, la família *FAM90A* està formada per múltiples membres.

8. La família gènica *FAM90A* es distribueix de manera molt semblant en els genomes d'humà i ximpanzé. En el genoma de ximpanzé existeixen seqüències homòlogues al clústers de la Subfamília II d'humans que contenen 7 (PtrCopy1-7), i 6 còpies (PtrCopy8-13). Membres de *FAM90A* homòlogues a la Subfamília I es troben al cromosoma 12 (PtrCopy14), al cromosoma 8 (PtrCopy15) i al cromosoma 11(PtrCopy16) del ximpanzé.

9. No s'ha identificat cap membre de la Subfamília I de *FAM90A* en macac, papió, ni orangutan.

10. Els estudis Ka/Ks de substitucions sinònimes i no sinònimes en la seqüència codificant dels membres de *FAM90A* en les diferents espècies suggereixen que almenys un membre de *FAM90A* a cada espècie (macac, ximpanzé i humà) codifica per una proteïna funcional.

11. En el genoma de macac la seqüència de 1036 parells de bases de la Subfamília I es troba en l' intró del gen *ALG1* del cromosoma 16. L'anàlisi a nivell de seqüència en aquesta espècie, juntament amb els experiments de genòmica comparada realitzats en humà, ximpanzé, goril·la, orangutan i macac, indiquen que els membres de la Subfamília I es devien originar aproximadament en la divergència entre els orangutans i els grans simis (~15 milions d'anys). Mitjançant un procés de fusió que inclouria els 1036 bp de l' intró *ALG1* i un membre de la Subfamília II, s'hauria creat el primer membre de la Subfamília I.

12. La família gènica *FAM90A* ha patit un procés d'expansió al llarg de l'evolució dels primats, tal i com evidencien les anàlisis fetes per Southern blot, FISH i PCR quantitativa en les diferents espècies de primats. S'ha hipotetitzat un possible model segons el qual, a partir d'una fusió i diverses delecions i duplicacions, s'hauria format i expandit la família gènica *FAM90A* fins a constituir les 25 còpies presents avui en dia en el genoma humà.

13. La família gènica *FAM90A* representa una variant en número de còpia (CNV) de la regió 8p23.1, juntament amb d'altres CNVs de la

regió com les alfa i les beta defensines, en les diferents poblacions humanes.

14. El global de la mitjana en número de còpies de FAM90A en població d'origen asiàtic és menor que en població d'origen europeu, mentre que la mitjana en població africana és superior. La població d'origen africà presenta diferències estadísticament significatives en quant al número de còpies de *FAM90A* respecte les altres dues poblacions.

15. La variabilitat en número de còpies de la família gènica *FAM90A* només explica un 5% de la variabilitat a nivell d'expressió en limfoblasts dels membres d'aquesta família gènica.

16. La inversió polimòrfica de la regió 8p23.1 es troba en heterozigosi en un 50% dels nuclis analitzats procedents d'individus de les tres poblacions analitzades, la població espanyola, les mostres HapMap d'origen europeu i les d'origen asiàtic. L'assemblatge de referència representa la orientació minoritària en aquestes tres poblacions.

17. Existeixen dos blocs d'homozigositat prop dels duplicons REPD i REPP de 8p23.1, que permeten predir la conformació de la regió a partir de 16 SNPs que es troben en desequilibri de lligament amb la inversió de 8p23.1.

18. L'anàlisi per FISH de la inversió en limfòcits de mostres d'individus sans espanyols i d'individus HapMap mostra la presència de mosaïcisme en tots els individus. La recombinació mitòtica podria ser un fenomen comú en les regions flanquejades per "low copy repeats".

19. La presència de la inversió de 8p23.1 en un estudi realitzat en 150 individus, correlaciona amb els nivells d'expressió dels gens *NEIL2, MSRA, CTSB i BLK* localitzats a 8p23.1.

Bibliografia

1. Gregory TR. Synergy between sequence and size in large-scale genomics. *Nat Rev Genet* 2005;**6**(9)**:**699-708.
2. Lander ES, Linton LM, Birren B, et al. Initial sequencing and analysis of the human genome. *Nature* 2001;**409**(6822)**:**860-921.
3. Bailey JA, Gu Z, Clark RA, et al. Recent segmental duplications in the human genome. *Science* 2002;**297**(5583)**:**1003-7.
4. Cheung J, Estivill X, Khaja R, et al. Genome-wide detection of segmental duplications and potential assembly errors in the human genome sequence. *Genome Biol* 2003;**4**(4)**:**R25.
5. She X, Jiang Z, Clark RA, et al. Shotgun sequence assembly and recent segmental duplications within the human genome. *Nature* 2004;**431**(7011)**:**927-30.
6. Venter JC, Adams MD, Myers EW, et al. The sequence of the human genome. *Science* 2001;**291**(5507)**:**1304-51.
7. Finishing the euchromatic sequence of the human genome. *Nature* 2004;**431**(7011)**:**931-45.
8. Levy S, Sutton G, Ng PC, et al. The diploid genome sequence of an individual human. *PLoS Biol* 2007;**5**(10)**:**e254.
9. Glusman G, Yanai I, Rubin I, Lancet D. The complete human olfactory subgenome. *Genome Res* 2001;**11**(5)**:**685-702.
10. Smit AF. Interspersed repeats and other mementos of transposable elements in mammalian genomes. *Curr Opin Genet Dev* 1999;**9**(6)**:**657-63.

Bibliografia

11. Kawana S, Watanabe G, Asami T, Okada T, Yamamoto M. Pericentric inversion of chromosome No. 8. *Tohoku J Exp Med* 1976;**119**(1)**:**65-70.

12. Esnault C, Maestre J, Heidmann T. Human LINE retrotransposons generate processed pseudogenes. *Nat Genet* 2000;**24**(4)**:**363-7.

13. Wei W, Gilbert N, Ooi SL, et al. Human L1 retrotransposition: cis preference versus trans complementation. *Mol Cell Biol* 2001;**21**(4)**:**1429-39.

14. Malik HS, Henikoff S, Eickbush TH. Poised for contagion: evolutionary origins of the infectious abilities of invertebrate retroviruses. *Genome Res* 2000;**10**(9)**:**1307-18.

15. Smit AF, Riggs AD. Tiggers and DNA transposon fossils in the human genome. *Proc Natl Acad Sci U S A* 1996;**93**(4)**:**1443-8.

16. Clark JB, Kidwell MG. A phylogenetic perspective on P transposable element evolution in Drosophila. *Proc Natl Acad Sci U S A* 1997;**94**(21)**:**11428-33.

17. Koga A, Hori H. Detection of de novo insertion of the medaka fish transposable element Tol2. *Genetics* 2000;**156**(3)**:**1243-7.

18. Haring E, Hagemann S, Pinsker W. Ancient and recent horizontal invasions of drosophilids by P elements. *J Mol Evol* 2000;**51**(6)**:**577-86.

19. Kruglyak S, Durrett RT, Schug MD, Aquadro CF. Equilibrium distributions of microsatellite repeat length resulting from a balance between slippage events and point mutations. *Proc Natl Acad Sci U S A* 1998;**95**(18)**:**10774-8.

20. Toth G, Gaspari Z, Jurka J. Microsatellites in different eukaryotic genomes: survey and analysis. *Genome Res* 2000;**10**(7)**:**967-81.

21. Ellegren H. Heterogeneous mutation processes in human microsatellite DNA sequences. *Nat Genet* 2000;**24**(4)**:**400-2.

22. Dib C, Faure S, Fizames C, et al. A comprehensive genetic map of the human genome based on 5,264 microsatellites. *Nature* 1996;**380**(6570)**:**152-4.

23. Eichler EE. Masquerading repeats: paralogous pitfalls of the human genome. *Genome Res* 1998;**8**(8)**:**758-62.

24. Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A* 1977;**74**(12)**:**5463-7.

25. Maxam AM, Gilbert W. A new method for sequencing DNA. *Proc Natl Acad Sci U S A* 1977;**74**(2)**:**560-4.

26. Wheeler DA, Srinivasan M, Egholm M, et al. The complete genome of an individual by massively parallel DNA sequencing. *Nature* 2008;**452**(7189)**:**872-6.

27. Korbel JO, Urban AE, Affourtit JP, et al. Paired-end mapping reveals extensive structural variation in the human genome. *Science* 2007;**318**(5849)**:**420-6.

28. Barski A, Cuddapah S, Cui K, et al. High-resolution profiling of histone methylations in the human genome. *Cell* 2007;**129**(4)**:**823-37.

29. Baltimore D. Our genome unveiled. *Nature* 2001;**409**(6822)**:**814-6.

30. Mazzarella R, Schlessinger D. Pathological consequences of sequence duplications in the human genome. *Genome Res* 1998;**8**(10)**:**1007-21.

31. Ji Y, Eichler EE, Schwartz S, Nicholls RD. Structure of chromosomal duplicons and their role in mediating human genomic disorders. *Genome Res* 2000;**10**(5)**:**597-610.

32. Bovee D, Zhou Y, Haugen E, et al. Closing gaps in the human genome with fosmid resources generated from multiple individuals. *Nat Genet* 2008;**40**(1)**:**96-101.

33. She X, Liu G, Ventura M, et al. A preliminary comparative analysis of primate segmental duplications shows elevated substitution rates and a great-ape expansion of intrachromosomal duplications. *Genome Res* 2006;**16**(5)**:**576-83.

34. Eichler EE, Lu F, Shen Y, et al. Duplication of a gene-rich cluster between 16p11.1 and Xq28: a novel pericentromeric-directed mechanism for paralogous genome evolution. *Hum Mol Genet* 1996;**5**(7)**:**899-912.

35. Eichler EE, Budarf ML, Rocchi M, et al. Interchromosomal duplications of the adrenoleukodystrophy locus: a phenomenon of pericentromeric plasticity. *Hum Mol Genet* 1997;**6**(7)**:**991-1002.

36. Trask BJ, Friedman C, Martin-Gallardo A, et al. Members of the olfactory receptor gene family are contained in large blocks of DNA duplicated polymorphically near the ends of human chromosomes. *Hum Mol Genet* 1998;**7**(1)**:**13-26.

37. Trask BJ, Massa H, Brand-Arpon V, et al. Large multi-chromosomal duplications encompass many members of the olfactory receptor gene family in the human genome. *Hum Mol Genet* 1998;**7**(13)**:**2007-20.

38. Hattori M, Fujiyama A, Taylor TD, et al. The DNA sequence of human chromosome 21. *Nature* 2000;**405**(6784)**:**311-9.

39. Bayes M, Magano LF, Rivera N, Flores R, Perez Jurado LA. Mutational mechanisms of Williams-Beuren syndrome deletions. *Am J Hum Genet* 2003;**73**(1)**:**131-51.

40. Reiter LT, Murakami T, Koeuth T, Gibbs RA, Lupski JR. The human COX10 gene is disrupted during homologous recombination between the 24 kb proximal and distal CMT1A-REPs. *Hum Mol Genet* 1997;**6**(9)**:**1595-603.

41. Edelmann L, Pandita RK, Morrow BE. Low-copy repeats mediate the common 3-Mb deletion in patients with velo-cardio-facial syndrome. *Am J Hum Genet* 1999;**64**(4)**:**1076-86.

42. Johnson ME, Viggiano L, Bailey JA, et al. Positive selection of a gene family during the emergence of humans and African apes. *Nature* 2001;**413**(6855)**:**514-9.

43. Bailey JA, Eichler EE. Primate segmental duplications: crucibles of evolution, diversity and disease. *Nat Rev Genet* 2006;**7**(7)**:**552-64.

44. Bailey JA, Yavor AM, Viggiano L, et al. Human-specific duplication and mosaic transcripts: the recent paralogous structure of chromosome 22. *Am J Hum Genet* 2002;**70**(1)**:**83-100.

45. She X, Horvath JE, Jiang Z, et al. The structure and evolution of centromeric transition regions within the human genome. *Nature* 2004;**430**(7002)**:**857-64.

46. Armengol L, Pujana MA, Cheung J, Scherer SW, Estivill X. 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* 2003;**12**(17)**:**2201-8.

47. Cheung J, Wilson MD, Zhang J, et al. Recent segmental and gene duplications in the mouse genome. *Genome Biol* 2003;**4**(8)**:**R47.

48. Bailey JA, Church DM, Ventura M, Rocchi M, Eichler EE. Analysis of segmental duplications and genome assembly in the mouse. *Genome Res* 2004;**14**(5)**:**789-801.

49. Tuzun E, Bailey JA, Eichler EE. Recent segmental duplications in the working draft assembly of the brown Norway rat. *Genome Res* 2004;**14**(4)**:**493-506.

50. Marques-Bonet T, Cheng Z, She X, Eichler EE, Navarro A. The genomic distribution of intraspecific and interspecific sequence divergence of human segmental duplications relative to human/chimpanzee chromosomal rearrangements. *BMC Genomics* 2008;**9**(1)**:**384.

51. Yunis JJ, Prakash O. The origin of man: a chromosomal pictorial legacy. *Science* 1982;**215**(4539)**:**1525-30.

52. Kehrer-Sawatzki H, Cooper DN. Molecular mechanisms of chromosomal rearrangement during primate evolution. *Chromosome Res* 2008;**16**(1)**:**41-56.

53. Shaw CJ, Lupski JR. Implications of human genome architecture for rearrangement-based disorders: the genomic basis of disease. *Hum Mol Genet* 2004;**13 Spec No 1:**R57-64.

54. Hurles M. Gene duplication: the genomic trade in spare parts. *PLoS Biol* 2004;**2**(7)**:**E206.

55. Zhang P, Gu Z, Li WH. Different evolutionary patterns between young duplicate genes in the human genome. *Genome Biol* 2003;**4**(9)**:**R56.

56. Zhang J, Rosenberg HF, Nei M. Positive Darwinian selection after gene duplication in primate ribonuclease genes. *Proc Natl Acad Sci U S A* 1998;**95**(7)**:**3708-13.

57. Ohno S. Gene duplication and the uniqueness of vertebrate genomes circa 1970-1999. *Semin Cell Dev Biol* 1999;**10**(5)**:**517-22.

58. Wolfe KH, Shields DC. Molecular evidence for an ancient duplication of the entire yeast genome. *Nature* 1997;**387**(6634)**:**708-13.

59. Sidow A, Bowman BH. Molecular phylogeny. *Curr Opin Genet Dev* 1991;**1**(4)**:**451-6.

60. Sidow A. Gen(om)e duplications in the evolution of early vertebrates. *Curr Opin Genet Dev* 1996;**6**(6)**:**715-22.

61. Leder A, Swan D, Ruddle F, D'Eustachio P, Leder P. Dispersion of alpha-like globin genes of the mouse to three different chromosomes. *Nature* 1981;**293**(5829)**:**196-200.

62. Blanchetot A, Price M, Jeffreys AJ. The mouse myoglobin gene. Characterisation and sequence comparison with other mammalian myoglobin genes. *Eur J Biochem* 1986;**159**(3)**:**469-74.

63. Drouet B, Simon-Chazottes D. The microsatellite found in the DNA sequence with the code name MMMYOGG1 (GenBank) does not correspond to the myogenin gene (Myog) but to myoglobin (Mb) and maps to mouse chromosome 15. *Mamm Genome* 1993;**4**(6)**:**348.

64. McGinnis W, Krumlauf R. Homeobox genes and axial patterning. *Cell* 1992;**68**(2)**:**283-302.

65. Hood L, Kronenberg M, Hunkapiller T. T cell antigen receptors and the immunoglobulin supergene family. *Cell* 1985;**40**(2)**:**225-9.

66. Dover G. Molecular drive: a cohesive mode of species evolution. *Nature* 1982;**299**(5879)**:**111-7.

67. Fortna A, Kim Y, MacLaren E, et al. Lineage-specific gene duplication and loss in human and great ape evolution. *PLoS Biol* 2004;**2**(7)**:**E207.

68. Gagneux P, Varki A. Genetic differences between humans and great apes. *Mol Phylogenet Evol* 2001;**18**(1)**:**2-13.

69. Hacia JG. Genome of the apes. *Trends Genet* 2001;**17**(11)**:**637-45.

70. Moffat AS. Primate origins meeting. New fossils and a glimpse of evolution. *Science* 2002;**295**(5555)**:**613-5.

71. Jacobs GH, Neitz M, Deegan JF, Neitz J. Trichromatic colour vision in New World monkeys. *Nature* 1996;**382**(6587)**:**156-8.

72. Nathans J, Thomas D, Hogness DS. Molecular genetics of human color vision: the genes encoding blue, green, and red pigments. *Science* 1986;**232**(4747)**:**193-202.

73. Boissinot S, Tan Y, Shyue SK, et al. Origins and antiquity of X-linked triallelic color vision systems in New World monkeys. *Proc Natl Acad Sci U S A* 1998;**95**(23)**:**13749-54.

74. Long M. A new function evolved from gene fusion. *Genome Res* 2000;**10**(11)**:**1655-7.

75. Thomson TM, Lozano JJ, Loukili N, et al. Fusion of the human gene for the polyubiquitination coeffector UEV1 with Kua, a newly identified gene. *Genome Res* 2000;**10**(11)**:**1743-56.

76. Rouquier S, Blancher A, Giorgi D. The olfactory receptor gene repertoire in primates and mouse: evidence for reduction of the functional fraction in primates. *Proc Natl Acad Sci U S A* 2000;**97**(6)**:**2870-4.

77. Samuelson LC, Wiebauer K, Gumucio DL, Meisler MH. Expression of the human amylase genes: recent origin of a salivary amylase promoter from an actin pseudogene. *Nucleic Acids Res* 1988;**16**(17)**:**8261-76.

78. Samuelson LC, Phillips RS, Swanberg LJ. Amylase gene structures in primates: retroposon insertions and promoter evolution. *Mol Biol Evol* 1996;**13**(6)**:**767-79.

79. Mighell AJ, Markham AF, Robinson PA. Alu sequences. *FEBS Lett* 1997;**417**(1)**:**1-5.

80. Batzer MA, Deininger PL. Alu repeats and human genomic diversity. *Nat Rev Genet* 2002;**3**(5)**:**370-9.

81. Nahon JL. The melanin-concentrating hormone: from the peptide to the gene. *Crit Rev Neurobiol* 1994;**8**(4)**:**221-62.

82. Gibbons A. Which of our genes makes us human? *Science* 1998;**281**(5382)**:**1432-4.

83. Normile D. Comparative genomics. Gene expression differs in human and chimp brains. *Science* 2001;**292**(5514)**:**44-5.

84. Enard W, Khaitovich P, Klose J, et al. Intra- and interspecific variation in primate gene expression patterns. *Science* 2002;**296**(5566)**:**340-3.

85. Watanabe H, Hattori M, Fujiyama A, Sakaki Y. [Construction and analysis of a human-chimpanzee comparative clone map]. *Tanpakushitsu Kakusan Koso* 2002;**47**(7)**:**808-13.

86. Jacobs PA, Baikie AG, Court Brown WM, Strong JA. The somatic chromosomes in mongolism. *Lancet* 1959;**1**(7075)**:**710.

87. Sebat J, Lakshmi B, Troge J, et al. Large-scale copy number polymorphism in the human genome. *Science* 2004;**305**(5683)**:**525-8.

88. Iafrate AJ, Feuk L, Rivera MN, et al. Detection of large-scale variation in the human genome. *Nat Genet* 2004;**36**(9)**:**949-51.

89. Redon R, Ishikawa S, Fitch KR, et al. Global variation in copy number in the human genome. *Nature* 2006;**444**(7118)**:**444-54.

90. Hurles ME, Dermitzakis ET, Tyler-Smith C. The functional impact of structural variation in humans. *Trends Genet* 2008;**24**(5)**:**238-45.

91. Estivill X, Armengol L. Copy number variants and common disorders: filling the gaps and exploring complexity in genome-wide association studies. *PLoS Genet* 2007;**3**(10)**:**1787-99.

92. Carter NP. Methods and strategies for analyzing copy number variation using DNA microarrays. *Nat Genet* 2007;**39**(7 Suppl)**:**S16-21.

93. Tuzun E, Sharp AJ, Bailey JA, et al. Fine-scale structural variation of the human genome. *Nat Genet* 2005.

94. Kidd JM, Cooper GM, Donahue WF, et al. Mapping and sequencing of structural variation from eight human genomes. *Nature* 2008;**453**(7191)**:**56-64.

95. Wong KK, deLeeuw RJ, Dosanjh NS, et al. A comprehensive analysis of common copy-number variations in the human genome. *Am J Hum Genet* 2007;**80**(1)**:**91-104.

96. Sharp AJ, Locke DP, McGrath SD, et al. Segmental duplications and copy-number variation in the human genome. *Am J Hum Genet* 2005;**77**(1)**:**78-88.

97. Zhang L, Lu HH, Chung WY, Yang J, Li WH. Patterns of segmental duplication in the human genome. *Mol Biol Evol* 2005;**22**(1)**:**135-41.

98. Nguyen DQ, Webber C, Ponting CP. Bias of selection on human copy-number variants. *PLoS Genet* 2006;**2**(2)**:**e20.

99. Feuk L, Carson AR, Scherer SW. Structural variation in the human genome. *Nat Rev Genet* 2006;**7**(2)**:**85-97.

100. Waterston RH, Lindblad-Toh K, Birney E, et al. Initial sequencing and comparative analysis of the mouse genome. *Nature* 2002;**420**(6915)**:**520-62.

101. Ciccarelli FD, von Mering C, Suyama M, Harrington ED, Izaurralde E, Bork P. Complex genomic rearrangements lead to novel primate gene function. *Genome Res* 2005;**15**(3)**:**343-51.

102. Popesco MC, Maclaren EJ, Hopkins J, et al. Human lineage-specific amplification, selection, and neuronal expression of DUF1220 domains. *Science* 2006;**313**(5791)**:**1304-7.

103. Stefansson H, Helgason A, Thorleifsson G, et al. A common inversion under selection in Europeans. *Nat Genet* 2005;**37**(2)**:**129-37.

104. Giglio S, Calvari V, Gregato G, et al. Heterozygous submicroscopic inversions involving olfactory receptor-gene clusters mediate the

recurrent t(4;8)(p16;p23) translocation. *Am J Hum Genet* 2002;**71**(2)**:**276-85.

105. Giglio S, Broman KW, Matsumoto N, et al. Olfactory receptor-gene clusters, genomic-inversion polymorphisms, and common chromosome rearrangements. *Am J Hum Genet* 2001;**68**(4)**:**874-83.
106. Lupski JR. Genomic disorders: structural features of the genome can lead to DNA rearrangements and human disease traits. *Trends Genet* 1998;**14**(10)**:**417-22.
107. Feuk L, MacDonald JR, Tang T, et al. Discovery of human inversion polymorphisms by comparative analysis of human and chimpanzee DNA sequence assemblies. *PLoS Genet* 2005;**1**(4)**:**e56.
108. Sugawara H, Harada N, Ida T, et al. Complex low-copy repeats associated with a common polymorphic inversion at human chromosome 8p23. *Genomics* 2003;**82**(2)**:**238-44.
109. Lakich D, Kazazian HH, Jr., Antonarakis SE, Gitschier J. Inversions disrupting the factor VIII gene are a common cause of severe haemophilia A. *Nat Genet* 1993;**5**(3)**:**236-41.
110. Osborne LR, Li M, Pober B, et al. A 1.5 million-base pair inversion polymorphism in families with Williams-Beuren syndrome. *Nat Genet* 2001;**29**(3)**:**321-5.
111. Bondeson ML, Dahl N, Malmgren H, et al. Inversion of the IDS gene resulting from recombination with IDS-related sequences is a common cause of the Hunter syndrome. *Hum Mol Genet* 1995;**4**(4)**:**615-21.
112. Thomas NS, Bryant V, Maloney V, Cockwell AE, Jacobs PA. Investigation of the origins of human autosomal inversions. *Hum Genet* 2008;**123**(6)**:**607-16.
113. Sherman SL, Sutherland GR. Segregation analysis of rare autosomal fragile sites. *Hum Genet* 1986;**72**(2)**:**123-8.

114. Daniel A, Hook EB, Wulf G. Risks of unbalanced progeny at amniocentesis to carriers of chromosome rearrangements: data from United States and Canadian laboratories. *Am J Med Genet* 1989;**33**(1)**:**14-53.

115. Pettenati MJ, Rao PN, Phelan MC, et al. Paracentric inversions in humans: a review of 446 paracentric inversions with presentation of 120 new cases. *Am J Med Genet* 1995;**55**(2)**:**171-87.

116. Youings S, Ellis K, Ennis S, Barber J, Jacobs P. A study of reciprocal translocations and inversions detected by light microscopy with special reference to origin, segregation, and recurrent abnormalities. *Am J Med Genet A* 2004;**126A**(1)**:**46-60.

117. Iida A, Emi M, Matsuoka R, et al. Identification of a gene disrupted by inv(11)(q13.5;q25) in a patient with left-right axis malformation. *Hum Genet* 2000;**106**(3)**:**277-87.

118. Saito-Ohara F, Fukuda Y, Ito M, et al. The Xq22 inversion breakpoint interrupted a novel Ras-like GTPase gene in a patient with Duchenne muscular dystrophy and profound mental retardation. *Am J Hum Genet* 2002;**71**(3)**:**637-45.

119. Beiraghi S, Zhou M, Talmadge CB, et al. Identification and characterization of a novel gene disrupted by a pericentric inversion inv(4)(p13.1q21.1) in a family with cleft lip. *Gene* 2003;**309**(1)**:**11-21.

120. Sood R, Bader PI, Speer MC, et al. Cloning and characterization of an inversion breakpoint at 6q23.3 suggests a role for Map7 in sacral dysgenesis. *Cytogenet Genome Res* 2004;**106**(1)**:**61-7.

121. Tadin-Strapps M, Warburton D, Baumeister FA, et al. Cloning of the breakpoints of a de novo inversion of chromosome 8, inv (8)(p11.2q23.1) in a patient with Ambras syndrome. *Cytogenet Genome Res* 2004;**107**(1-2)**:**68-76.

122. Jaarola M, Martin RH, Ashley T. Direct evidence for suppression of recombination within two pericentric inversions in humans: a new sperm-FISH technique. *Am J Hum Genet* 1998;**63**(1)**:**218-24.

123. Zetka MC, Rose AM. The meiotic behavior of an inversion in Caenorhabditis elegans. *Genetics* 1992;**131**(2)**:**321-32.

124. Ciccone R, Mattina T, Giorda R, et al. Inversion polymorphisms and non-contiguous terminal deletions: the cause and the (unpredicted) effect of our genome architecture. *J Med Genet* 2006;**43**(5)**:**e19.

125. Small K, Iber J, Warren ST. Emerin deletion reveals a common X-chromosome inversion mediated by inverted repeats. *Nat Genet* 1997;**16**(1)**:**96-9.

126. Caceres M, Sullivan RT, Thomas JW. A recurrent inversion on the eutherian X chromosome. *Proc Natl Acad Sci U S A* 2007;**104**(47)**:**18571-6.

127. Small K, Warren ST. Emerin deletions occurring on both Xq28 inversion backgrounds. *Hum Mol Genet* 1998;**7**(1)**:**135-9.

128. Saunier S, Calado J, Benessy F, et al. Characterization of the NPHP1 locus: mutational mechanism involved in deletions in familial juvenile nephronophthisis. *Am J Hum Genet* 2000;**66**(3)**:**778-89.

129. Jobling MA, Williams GA, Schiebel GA, et al. A selective difference between human Y-chromosomal DNA haplotypes. *Curr Biol* 1998;**8**(25)**:**1391-4.

130. Repping S, Skaletsky H, Brown L, et al. Polymorphism for a 1.6-Mb deletion of the human Y chromosome persists through balance between recurrent mutation and haploid selection. *Nat Genet* 2003;**35**(3)**:**247-51.

131. Tyler-Smith C, McVean G. The comings and goings of a Y polymorphism. *Nat Genet* 2003;**35**(3)**:**201-2.

132. Cusco I, Corominas R, Bayes M, et al. Copy number variation at the 7q11.23 segmental duplications is a susceptibility factor for the Williams-Beuren syndrome deletion. *Genome Res* 2008;**18**(5)**:**683-94.

133. Wieczorek D, Krause M, Majewski F, et al. Unexpected high frequency of de novo unbalanced translocations in patients with Wolf-Hirschhorn syndrome (WHS). *J Med Genet* 2000;**37**(10)**:**798-804.

134. Gimelli G, Pujana MA, Patricelli MG, et al. Genomic inversions of human chromosome 15q11-q13 in mothers of Angelman syndrome patients with class II (BP2/3) deletions. *Hum Mol Genet* 2003;**12**(8)**:**849-58.

135. Saitta SC, Harris SE, Gaeth AP, et al. Aberrant interchromosomal exchanges are the predominant cause of the 22q11.2 deletion. *Hum Mol Genet* 2004;**13**(4)**:**417-28.

136. Giglio S, Graw SL, Gimelli G, et al. Deletion of a 5-cM region at chromosome 8p23 is associated with a spectrum of congenital heart defects. *Circulation* 2000;**102**(4)**:**432-7.

137. Pehlivan T, Pober BR, Brueckner M, et al. GATA4 haploinsufficiency in patients with interstitial deletion of chromosome region 8p23.1 and congenital heart disease. *Am J Med Genet* 1999;**83**(3)**:**201-6.

138. Devriendt K, Matthijs G, Van Dael R, et al. Delineation of the critical deletion region for congenital heart defects, on chromosome 8p23.1. *Am J Hum Genet* 1999;**64**(4)**:**1119-26.

139. Barber JC, Joyce CA, Collinson MN, et al. Duplication of 8p23.1: a cytogenetic anomaly with no established clinical significance. *J Med Genet* 1998;**35**(6)**:**491-6.

140. Wu BL, Schneider GH, Sabatino DE, Bozovic LZ, Cao B, Korf BR. Distal 8p deletion (8)(p23.1): an easily missed chromosomal abnormality that may be associated with congenital heart defect and mental retardation. *Am J Med Genet* 1996;**62**(1)**:**77-83.

141. Floridia G, Piantanida M, Minelli A, et al. The same molecular mechanism at the maternal meiosis I produces mono- and dicentric 8p duplications. *Am J Hum Genet* 1996;**58**(4)**:**785-96.

142. Guo WJ, Callif-Daley F, Zapata MC, Miller ME. Clinical and cytogenetic findings in seven cases of inverted duplication of 8p with evidence of a telomeric deletion using fluorescence in situ hybridization. *Am J Med Genet* 1995;**58**(3)**:**230-6.

143. Ohashi H, Wakui K, Ogawa K, Okano T, Niikawa N, Fukushima Y. A stable acentric marker chromosome: possible existence of an intercalary ancient centromere at distal 8p. *Am J Hum Genet* 1994;**55**(6)**:**1202-8.

144. Blouin JL, Dombroski BA, Nath SK, et al. Schizophrenia susceptibility loci on chromosomes 13q32 and 8p21. *Nat Genet* 1998;**20**(1)**:**70-3.

145. Kendler KS, MacLean CJ, O'Neill FA, et al. Evidence for a schizophrenia vulnerability locus on chromosome 8p in the Irish Study of High-Density Schizophrenia Families. *Am J Psychiatry* 1996;**153**(12)**:**1534-40.

146. Ophoff RA, Escamilla MA, Service SK, et al. Genomewide linkage disequilibrium mapping of severe bipolar disorder in a population isolate. *Am J Hum Genet* 2002;**71**(3)**:**565-74.

147. Pulver AE, Lasseter VK, Kasch L, et al. Schizophrenia: a genome scan targets chromosomes 3p and 8p as potential sites of susceptibility genes. *Am J Med Genet* 1995;**60**(3)**:**252-60.

148. Aldred PM, Hollox EJ, Armour JA. Copy number polymorphism and expression level variation of the human {alpha}-defensin genes DEFA1 and DEFA3. *Hum Mol Genet* 2005.

149. Hollox EJ, Armour JA, Barber JC. Extensive normal copy number variation of a beta-defensin antimicrobial-gene cluster. *Am J Hum Genet* 2003;**73**(3)**:**591-600.

150. Mars WM, Patmasiriwat P, Maity T, Huff V, Weil MM, Saunders GF. Inheritance of unequal numbers of the genes encoding the human neutrophil defensins HP-1 and HP-3. *J Biol Chem* 1995;**270**(51)**:**30371-6.

151. Zhang L, Yu W, He T, et al. Contribution of human alpha-defensin 1, 2, and 3 to the anti-HIV-1 activity of CD8 antiviral factor. *Science* 2002;**298**(5595)**:**995-1000.

152. Ganz T. Defensins: antimicrobial peptides of innate immunity. *Nat Rev Immunol* 2003;**3**(9)**:**710-20.

153. Chang TL, Vargas J, Jr., DelPortillo A, Klotman ME. Dual role of alpha-defensin-1 in anti-HIV-1 innate immunity. *J Clin Invest* 2005;**115**(3)**:**765-73.

154. Ericksen B, Wu Z, Lu W, Lehrer RI. Antibacterial activity and specificity of the six human {alpha}-defensins. *Antimicrob Agents Chemother* 2005;**49**(1)**:**269-75.

155. Fellermann K, Stange DE, Schaeffeler E, et al. A chromosome 8 gene-cluster polymorphism with low human beta-defensin 2 gene copy number predisposes to Crohn disease of the colon. *Am J Hum Genet* 2006;**79**(3)**:**439-48.

156. Chen GK, Slaten E, Ophoff RA, Lange K. Accommodating chromosome inversions in linkage analysis. *Am J Hum Genet* 2006;**79**(2)**:**238-51.

157. Linzmeier RM, Ganz T. Human defensin gene copy number polymorphisms: Comprehensive analysis of independent variation in alpha- and beta-defensin regions at 8p22-p23. *Genomics* 2005.

158. Hollox EJ, Davies J, Griesenbach U, Burgess J, Alton EW, Armour JA. Beta-defensin genomic copy number is not a modifier locus for cystic fibrosis. *J Negat Results Biomed* 2005;**4:**9.

159. Fredman D, White SJ, Potter S, Eichler EE, Den Dunnen JT, Brookes AJ. Complex SNP-related sequence variation in segmental genome duplications. *Nat Genet* 2004;**36**(8)**:**861-6.

160. Groth M, Huse K, Reichwald K, et al. Method for preparing single-stranded DNA templates for Pyrosequencing using vector ligation and universal biotinylated primers. *Anal Biochem* 2006;**356**(2)**:**194-201.

161. Ballana E, Gonzalez JR, Bosch N, Estivill X. Inter-population variability of DEFA3 gene absence: correlation with haplotype structure and population variability. *BMC Genomics* 2007;**8**(1)**:**14.

162. Barber JC. Directly transmitted unbalanced chromosome abnormalities and euchromatic variants. *J Med Genet* 2005;**42**(8)**:**609-29.

163. Huse K, Taudien S, Groth M, et al. Genetic variants of the copy number polymorphic beta-defensin locus are associated with sporadic prostate cancer. *Tumour Biol* 2008;**29**(2)**:**83-92.

164. Tsai CH, Graw SL, McGavran L. 8p23 duplication reconsidered: is it a true euchromatic variant with no clinical manifestation? *J Med Genet* 2002;**39**(10)**:**769-74.

165. Mitchell JJ, Vekemans M, Luscombe S, et al. U-type exchange in a paracentric inversion as a possible mechanism of origin of an inverted tandem duplication of chromosome 8. *Am J Med Genet* 1994;**49**(4)**:**384-7.

166. Dill FJ, Schertzer M, Sandercock J, Tischler B, Wood S. Inverted tandem duplication generates a duplication deficiency of chromosome 8p. *Clin Genet* 1987;**32**(2)**:**109-13.

167. Nevin NC, Morrison PJ, Jones J, Reid MM. Inverted tandem duplication of 8p12----p23.1 in a child with increased activity of glutathione reductase. *J Med Genet* 1990;**27**(2)**:**135-6.

168. Henderson KG, Dill FJ, Wood S. Characterization of an inversion duplication of the short arm of chromosome 8 by fluorescent in situ hybridization. *Am J Med Genet* 1992;**44**(5)**:**615-8.

169. Feldman GL, Weiss L, Phelan MC, Schroer RJ, Van Dyke DL. Inverted duplication of 8p: ten new patients and review of the literature. *Am J Med Genet* 1993;**47**(4)**:**482-6.

170. Minelli A, Floridia G, Rossi E, et al. D8S7 is consistently deleted in inverted duplications of the short arm of chromosome 8 (inv dup 8p). *Hum Genet* 1993;**92**(4)**:**391-6.

171. Barber JC, James RS, Patch C, Temple IK. Protelomeric sequences are deleted in cases of short arm inverted duplication of chromosome 8. *Am J Med Genet* 1994;**50**(3)**:**296-9.

172. Engelen JJ, de Die-Smulders CE, Fryns JP, et al. Partial trisomy and monosomy 8p due to inversion duplication. *Clin Genet* 1994;**45**(4)**:**203-7.

173. Plomp AS, Engelen JJ, Albrechts JC, de Die-Smulders CE, Hamers AJ. Two cases of partial trisomy 8p and partial monosomy 21q in a family with a reciprocal translocation (8;21)(p21.1;q22.3). *J Med Genet* 1998;**35**(7)**:**604-8.

174. Pecile V, Petroni MG, Fertz MC, Filippi G. Deficiency of distal 8p--report of two cases and review of the literature. *Clin Genet* 1990;**37**(4)**:**271-8.

175. Blennow E, Brondum-Nielsen K. Partial monosomy 8p with minimal dysmorphic signs. *J Med Genet* 1990;**27**(5)**:**327-9.

176. Fryns JP, Kleczkowska A, Vogels A, Van den Berghe H. Normal phenotype and slight mental retardation in de novo distal 8p deletion (8pter----8p23.1:). *Ann Genet* 1989;**32**(3)**:**171-3.

177. Fagan K, Wilkinson I, Allen M, Brownlea S. The coagulation factor VII regulator is located on 8p23.1. *Hum Genet* 1988;**79**(4)**:**365-7.

178. Hutchinson R, Wilson M, Voullaire L. Distal 8p deletion (8p23.1----8pter): a common deletion? *J Med Genet* 1992;**29**(6)**:**407-11.

179. Pettenati MJ, Rao N, Johnson C, et al. Molecular cytogenetic analysis of a familial 8p23.1 deletion associated with minimal dysmorphic features, seizures, and mild mental retardation. *Hum Genet* 1992;**89**(6)**:**602-6.

180. Taudien S, Galgoczy P, Huse K, et al. Polymorphic segmental duplications at 8p23.1 challenge the determination of individual defensin gene repertoires and the assembly of a contiguous human reference sequence. *BMC Genomics* 2004;**5**(1)**:**92.

181. Barber JC, Maloney V, Hollox EJ, et al. Duplications and copy number variants of 8p23.1 are cytogenetically indistinguishable but distinct at the molecular level. *Eur J Hum Genet* 2005.

182. Ohno S. The spontaneous mutation rate revisited and the possible principle of polymorphism generating more polymorphism. *Can J Genet Cytol* 1969;**11**(2)**:**457-67.

183. Wu Q, Krainer AR. AT-AC pre-mRNA splicing mechanisms and conservation of minor introns in voltage-gated ion channel genes. *Mol Cell Biol* 1999;**19**(5)**:**3225-36.

184. Samonte RV, Eichler EE. Segmental duplications and the evolution of the primate genome. *Nat Rev Genet* 2002;**3**(1)**:**65-72.

185. Hughes JF, Coffin JM. Human endogenous retrovirus K solo-LTR formation and insertional polymorphisms: implications for human and viral evolution. *Proc Natl Acad Sci U S A* 2004;**101**(6)**:**1668-72.

186. Lavie L, Medstrand P, Schempp W, Meese E, Mayer J. Human endogenous retrovirus family HERV-K(HML-5): status, evolution, and reconstruction of an ancient betaretrovirus in the human genome. *J Virol* 2004;**78**(16)**:**8788-98.

187. Courseaux A, Nahon JL. Birth of two chimeric genes in the Hominidae lineage. *Science* 2001;**291**(5507)**:**1293-7.

188. Eichler EE, Hoffman SM, Adamson AA, et al. Complex beta-satellite repeat structures and the expansion of the zinc finger gene cluster in 19p12. *Genome Res* 1998;**8**(8)**:**791-808.

189. Vandepoele K, Van Roy N, Staes K, Speleman F, van Roy F. A novel gene family NBPF: intricate structure generated by gene duplications during primate evolution. *Mol Biol Evol* 2005;**22**(11)**:**2265-74.

190. Foster MW, Sharp RR. Beyond race: towards a whole-genome perspective on human populations and genetic variation. *Nat Rev Genet* 2004;**5**(10)**:**790-6.

191. Stranger BE, Forrest MS, Dunning M, et al. Relative impact of nucleotide and copy number variation on gene expression phenotypes. *Science* 2007;**315**(5813)**:**848-53.

192. Morley M, Molony CM, Weber TM, et al. Genetic analysis of genome-wide variation in human gene expression. *Nature* 2004;**430**(7001)**:**743-7.

193. Conrad DF, Andrews TD, Carter NP, Hurles ME, Pritchard JK. A high-resolution survey of deletion polymorphism in the human genome. *Nat Genet* 2006;**38**(1)**:**75-81.

194. McCarroll SA, Hadnott TN, Perry GH, et al. Common deletion polymorphisms in the human genome. *Nat Genet* 2006;**38**(1)**:**86-92.

195. Lam KW, Jeffreys AJ. Processes of copy-number change in human DNA: the dynamics of {alpha}-globin gene deletion. *Proc Natl Acad Sci U S A* 2006;**103**(24)**:**8921-7.

196. Bosch N, Caceres M, Cardone MF, et al. Characterization and evolution of the novel gene family FAM90A in primates originated by multiple duplication and rearrangement events. *Hum Mol Genet* 2007.

197. Deng L, Zhang Y, Kang J, et al. An unusual haplotype structure on human chromosome 8p23 derived from the inversion polymorphism. *Hum Mutat* 2008.

198. Flores M, Morales L, Gonzaga-Jauregui C, et al. Recurrent DNA inversion rearrangements in the human genome. *Proc Natl Acad Sci U S A* 2007;**104**(15)**:**6099-106.

199. Navarro A, Barbadilla A, Ruiz A. Effect of inversion polymorphism on the neutral nucleotide variability of linked chromosomal regions in Drosophila. *Genetics* 2000;**155**(2)**:**685-98.

200. Pritchard JK, Przeworski M. Linkage disequilibrium in humans: models and data. *Am J Hum Genet* 2001;**69**(1)**:**1-14.

201. Moskovitz J, Jenkins NA, Gilbert DJ, et al. Chromosomal localization of the mammalian peptide-methionine sulfoxide reductase gene and its differential expression in various tissues. *Proc Natl Acad Sci U S A* 1996;**93**(8)**:**3205-8.

202. Bandaru V, Sunkara S, Wallace SS, Bond JP. A novel human DNA glycosylase that removes oxidative DNA damage and is homologous to Escherichia coli endonuclease VIII. *DNA Repair (Amst)* 2002;**1**(7)**:**517-29.

203. Esch FS, Keim PS, Beattie EC, et al. Cleavage of amyloid beta peptide during constitutive processing of its precursor. *Science* 1990;**248**(4959)**:**1122-4.

Abreviatures

ASN Individus del panell HapMap d'origen asiàtic.

BAC Cromosoma artificial de llevat (*bacterial artificial chromosome*).

CEU Individus del panell HapMap d'origen europeu.

CNV Variant en número de còpia (*copy number variant*).

DSs Duplicacions segmentàries.

FAM90A Família gènica (*family with sequence similarity 90*)

FISH Hibridació *in-situ* fluorescent (Fluorescent *in-situ* hybridazation).

Gb Gigabases. Milers de milions de parells de bases (10^9).

HERV Retrovirus endògens humans (human endogenous retroviruses).

HSA Seguit d'un número fa referència a l'autosoma humà especificat pel número (*Homo Sapiens Autosome*).

kb Kilobases. Milers de parells de bases (10^3).

LCR Repetició de baix número de còpies (*low copy repeat*).

LINE Element llarg repetitiu (*long interspersed repeat*).

LTR Repetició terminal llarga (long terminal repeat).

Mb Megabases. Milions de parells de bases (10^6).

NAHR Recombinació homòloga no al·lèlica.

NCBI National center for biotechnology information.

REPD Duplicons localitzats a l'extrem distal de la regió humana 8p23.1.

REPP Duplicons localitzats a l'extrem proximal de la regió humana 8p23.1.

SINE Element curt repetitiu dispers (*short interpersed nuclear element*).

SNP Polimorfisme d'un sol nucleòtid (*single nucleotide polymorphism*).

UCSC Universitat de Califòrnia, Santa Cruz.

WGD Duplicació de tot un genoma (*whole genome duplication*).

YRI Individus del panell HapMap d'origen africà.

Annex

Ballana E, González JR, Bosch N, Estivill X.
Inter-population variability of DEFA3 gene absence: correlation with haplotype structure and population variability.
BMC Genomics. 2007 Jan 10;8:14.