# Assemblatge de genomes a escala cromosòmica per redescobrir i conservar la biodiversitat catalana

Jèssica Gómez-Garrido,[1] Fernando Cruz,[1] Marc Palmada-Flores[2] i Tyler Alioto[1, 3]

[1] Centre Nacional d'Anàlisi Genòmica - Centre de Regulació Genòmica (CNAG-CRG), Barcelona Institute of Science and Technology (BIST)
[2] Department of Medicine and Life Sciences (MELIS), Institut de Biologia Evolutiva, Universitat Pompeu Fabra - CSIC
[3] Universitat Pompeu Fabra (UPF)

Correspondència: Tyler Alioto. Centre Nacional d'Anàlisi Genòmica. C. de Baldiri Reixac, 4. 08028 Barcelona. Tel.: +34 934 037 098. Adreça electrònica: *tyler.alioto@cnag.crg.eu*.

DOI: 10.2436/20.1501.02.214
ISSN (ed. impresa): 0212-3037
ISSN (ed. digital): 2013-9802
*http://revistes.iec.cat/index.php/TSCB*
Rebut: 31/01/2022
Acceptat: 24/03/2022

## Resum

Conèixer el genoma de les espècies que ens envolten és crucial per preservar la biodiversitat del territori. Assemblar un genoma consisteix a reconvertir les lectures fragmentades produïdes pels seqüenciadors en una seqüència contigua que representa el genoma complet de l'individu seqüenciat. Abans de l'arribada de les tecnologies de seqüenciació de lectura llarga, la majoria d'assemblatges produïts eren molt fragmentats, cosa que en limitava algunes de les utilitats. La incorporació de les noves tecnologies al camp de l'assemblatge de genomes ha permès una simplificació del procés i una millora de la qualitat dels assemblatges produïts. Els passos per obtenir un genoma de referència són: elaboració de blocs de seqüències consensuades, correcció de la seqüència, reconstrucció de cromosomes i perfeccionament de l'assemblatge. Un assemblatge de referència ens permet fer moltes anàlisis posteriors, com descobrir trets únics d'una espècie, que poden beneficiar les estratègies de conservació.

**Paraules clau:** assemblatge de genomes, refinament, reconstrucció de cromosomes, seqüenciació, assemblatges a escala cromosòmica.

## Chromosome-level genome assemblies to rediscover and conserve Catalonia's biodiversity

## Abstract

It is very important to have a knowledge of the genomes of the species around us in order to preserve a region's biodiversity. Assembling a genome involves combining the fragmented reads produced by sequencers into a contiguous sequence that represents the complete genome of the sequenced individual. Before the incorporation of long-read sequencing technologies, most of the genome assemblies that were produced were highly fragmented, limiting their utility for many downstream genomic analyses. The appearance of new technologies in the field of genome assembly has simplified the process and improved the quality of the resulting assemblies. The steps for producing a reference genome include contig assembly, sequence polishing, chromosome-level scaffolding and manual curation of the final assembly. A reference genome assembly allows multiple genomic analyses, which can greatly benefit the design of conservation plans.

**Keywords:** genome assembly, polishing, chromosome-level scaffolding, sequencing, chromosome-level assemblies.

## Introducció

Conèixer el genoma de les espècies que ens envolten pot aportar molt a l'hora de desenvolupar estratègies per conservar-les i protegir-les. A més, ens pot ajudar a saber com interactuen entre elles, potenciar-ne l'aplicació en ramaderia i agricultura o ajudar-nos a entendre com s'han originat determinats trets i com responen al canvi climàtic, entre d'altres. Projectes com la iniciativa catalana de l'Earth BioGenome Project (EBP), la CBP (de l'anglès Catalan Initiative for the Earth BioGenome Project), que té com a objectiu seqüenciar i assemblar el genoma de totes les espècies eucariotes dels territoris de parla catalana, són crucials per tal de garantir l'existència de dades genòmiques d'alta qualitat de les espècies que habiten la regió. La qualitat de l'assemblatge final és molt important perquè en depenen, en gran manera, la robustesa i fiabilitat de les anàlisis posteriors.

Com que la seqüenciació d'un genoma produeix lectures fragmentades (més o menys llargues segons la tecnologia de seqüenciació), és necessari un procés d'assemblatge capaç de transformar aquests fragments en una seqüència contigua que representi el genoma complet (com a mínim, el genoma nuclear i els orgànuls) de l'individu seqüenciat. L'assemblatge ideal tindria tantes seqüències contínues (també anomenades *còntigs*) com cromosomes contingui el genoma, i tots els nucleòtids d'aquests cromosomes serien coneguts. Desafortunadament, aconseguir el genoma perfecte és molt complicat i el que normalment obtenim són assemblatges fragmentats i incomplets.

El procés de seqüenciació i assemblatge de genomes ha canviat molt al llarg dels anys. Es va trigar més d'una dècada a produir els primers assemblatges del genoma humà (International Human Genome Sequencing Consortium *et al.*, 2001; Venter *et al.*, 2001) i d'altres espècies model, com el ratolí (Mouse Genome Sequencing Consortium *et al.*, 2002). Per obtenir-los, calia seqüenciar múltiples clons mitjançant el mètode de Sanger (500-1.000 nucleòtids per lectura) i van ser necessaris molts esforços manuals, així com el desenvolupament de programes informàtics. Més endavant, la seqüenciació del genoma complet amb lectures curtes d'Illumina (30-150 nucleòtids) va donar pas a un procés d'assemblatge més assequible, que va conduir a una revolució genòmica en la qual es va publicar l'assemblatge de moltes espècies, com per exemple el del panda gegant (Li *et al.*, 2010). Desafortunadament, l'ús de lectures tan curtes portava a l'obtenció d'assemblatges força fragmentats, amb problemes en certes zones del genoma, sobretot en les més repetitives, la qual cosa provocava l'absència o fragmentació d'alguns gens. Recentment,

l'aparició de les tecnologies de seqüenciació de lectura llarga ha fet que l'obtenció d'un assemblatge a escala cromosòmica d'una espècie qualsevol sigui possible en només unes quantes setmanes, la qual cosa ha canviat les normes del joc (Rhie *et al*., 2021). Per a més detall sobre les tecnologies de seqüenciació, convidem el lector a llegir l'article «Avenços en les tecnologies de seqüenciació del DNA», de Fusté *et al*., present en aquest mateix monogràfic.

## Avaluació d'un assemblatge

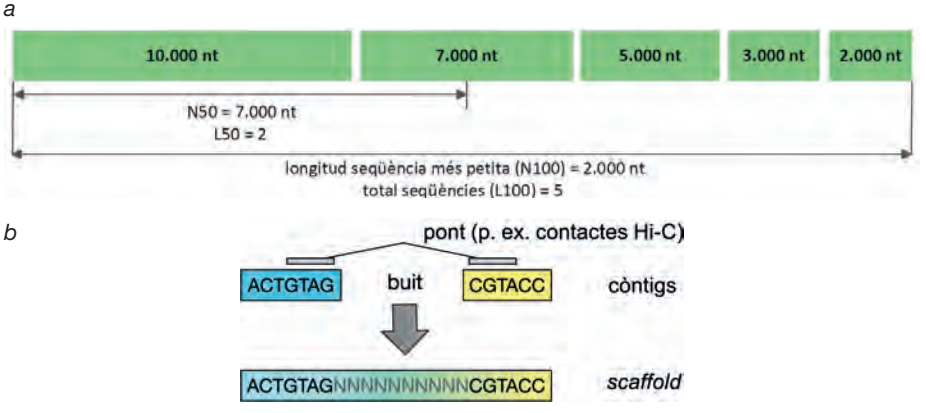
Abans d'entrar en la descripció dels passos que hem de seguir durant el procés d'assemblatge, definirem les diferents maneres que tenim per avaluar la qualitat d'un assemblatge *de novo* (utilitzant únicament dades de seqüenciació, sense fer servir cap altra referència prèviament coneguda). Però com podem avaluar la qualitat del nostre assemblatge si desconeixem com és realment el genoma? Els dos aspectes bàsics que cal tenir en compte són la contigüitat i la integritat (proporció del genoma que s'ha inclòs) de la seqüència assemblada.

Hem dit que l'assemblatge ideal tindria tantes seqüències com cromosomes tingui el genoma; així doncs, per avaluar-ne la contigüitat es té en compte el nombre total de seqüències assemblades i la longitud. Per a això es fan servir mètriques com el L50 i el N50 (Earl *et al*., 2011). El L50 és el menor nombre de fragments la suma de longituds dels quals conforma el 50 % del total de nucleòtids assemblats. El N50 es defineix com la longitud del fragment més petit d'aquest mateix conjunt. Per calcular-los, s'ordenen tots els fragments segons les longituds, de més gran a més petit, i es van comptabilitzant fins que la suma acumulada és més gran o igual que la meitat de l'assemblatge (figura 1*a*).

Un altre aspecte clau per determinar la contigüitat de l'assemblatge és el nombre de regions buides d'informació que conté. De vegades, es pot determinar que dues seqüències formen part del mateix fragment genòmic, però desconeixem la seqüència de nucleòtids que les uneix; en aquests casos, es construeix un pont entre elles afegint-hi unes quantes «N» entremig. Aquests ponts de seqüència desconeguda s'anomenen *buits* (*gaps*). Les seqüències assemblades que no contenen cap N s'anomenen *còntigs* i les formades per diversos còntigs connectats per buits són conegudes com a *scaffolds* (figura 1*b*).

Pel que fa a la integritat de l'assemblatge, es pot avaluar buscant quants gens d'entre un conjunt de gens coneguts i que, en principi, hauríem de trobar al genoma, són presents a l'assemblatge, o bé calculant la fracció de lectures del seqüenciador que han estat assemblades. CEGMA (*core eukaryotic genes mapping approach*) (Parra *et al*., 2007) i BUSCO (*benchmarking universal single-copy orthologs*) (Simão *et al*., 2015) són dos exemples de mètodes basats en la primera estratègia. CEGMA va ser el primer mètode d'aquest tipus que es va desenvolupar i utilitza com a referència 456 gens altament conservats en tots els eucariotes. D'altra banda, BUSCO fa servir bases de dades d'OrthoDB (Zdobnov *et al*., 2021) que contenen grups de gens conservats i presents una sola vegada en una branca filogenètica concreta (ortòlegs de còpia única). Aquests programes reporten quants dels gens presents a la base de dades es troben a l'assemblatge i quants estan duplicats o fragmentats. Totes aquestes dades són crucials per determinar la qualitat del genoma assemblat, ja que si falten molts gens que en teoria haurien d'estar presents o apareixen fragmentats, això pot voler dir que hi ha una part del genoma que no es troba en l'assemblatge, que no l'hem assemblat correctament o que la qualitat de la seqüència no és gaire bona, fet que provoca que els gens continguin errors de seqüència i no es puguin identificar. D'altra banda, si trobem molts gens duplicats que en teoria haurien de ser presents únicament una vegada, pot ser indicatiu de la presència de regions duplicades artificialment en el nostre assemblatge, tot i que també podria ser degut a altres raons biològiques, com, per exemple, una duplicació recent del genoma.

La segona estratègia es basa a calcular quina part de la seqüència present a les lectures ha acabat continguda a l'assemblatge. Aquesta estratègia és més universal, ja que no se centra només en unes regions concretes del genoma, sinó que n'analitza totes les posicions. Durant aquest procés s'acostumen a dividir les lectures en subseqüències de longitud *k* (sovint al voltant de 21 nucleòtids), anomenades *k-mers*. Així doncs, extraiem totes les subseqüències de longitud *k* de les lectures i les busquem a l'assemblatge per tal de determinar el percentatge de *k*-mers que comparteixen i saber, així, com és de complet el nostre assemblatge. A més, comparant els *k*-mers presents en les seqüències assemblades i en les lectures, també podem obtenir un valor de qualitat (QV), que ens proporciona informació sobre la quantitat d'errors produïts durant la seqüenciació que es traslladen a l'assemblatge. La mètrica QV (Rhie *et al*., 2020) reflecteix de manera logarítmica la precisió de la seqüència assemblada, de manera que com més alt sigui el valor, més precisa serà la seqüència. Per exemple, un QV de 30 correspon a una precisió del 99,9 %, és a dir, un error cada 1.000 nucleòtids; un QV de 40, a una precisió del 99,99 %, etc.



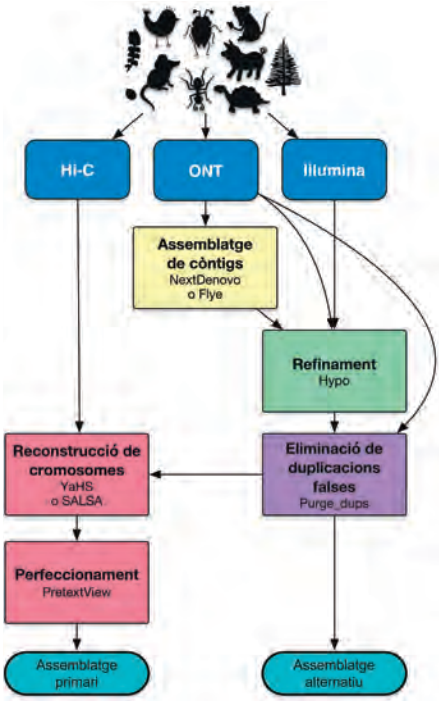

↑ Figura 1. *a*) Representació gràfica de les mètriques. *b*) Representació gràfica de còntigs, *scaffolds*, buits. Elaboració pròpia.

## El procés d'assemblatge

Ara que ja coneixem alguns dels conceptes més importants darrere dels assemblatges genòmics, ens endinsarem en el procés d'elaboració. Podem dividir el procés bàsic d'assemblatge en quatre passos principals:

1. obtenció de blocs de seqüències contigües (còntigs);
2. correcció de la seqüència de cada un dels còntigs (refinament);
3. reconstrucció de cromosomes (*chromosome scaffolding*);
4. perfeccionament de l'assemblatge.

La figura 2 mostra un exemple d'esquema d'un procés d'assemblatge real:




↑ Figura 2. Esquema del procés d'assemblatge que mostra els passos que es podrien seguir per assemblar el genoma de qualsevol organisme amb dades de seqüenciació d'ONT de lectura llarga, Illumina de lectura curta i dades de contacte de Hi-C. Elaboració pròpia.

## 1. Obtenció de blocs de seqüències contigües (còntigs)

Un cop s'han eliminat adaptadors (seqüències tècniques utilitzades en el procés de seqüenciació) i seleccionat les lectures de seqüenciació amb millor qualitat, aquestes es fan servir per generar blocs de seqüències contigües, que, com hem dit anteriorment, s'anomenen *còntigs*.

Fins fa pocs anys, el més freqüent era construir els còntigs a partir de lectures curtes mitjançant programes d'assemblatge que primer extreien tots els *k*-mers presents en les lectures i després construïen un graf de Bruijn connectant els diferents *k*-mers extrets (Pevzner *et al.*, 2001) en funció de les seves seqüències superposades. Entre els programes d'assemblatge que es van dissenyar per assemblar genomes amb lectures curtes i que es basen en grafs de Bruijn destacaríem Velvet (Zerbino i Birney, 2008), ABySS (Simpson *et al.*, 2009), SPAdes (Bankevich *et al.*, 2012) i SOAPdenovo (Li *et al.*, 2010). Com s'ha comentat, aquest procés produïa uns assemblatges molt fragmentats. Amb l'avenç en les tecnologies de seqüenciació de lectura llarga, el procés d'assemblatge no només ha millorat notablement, sinó que s'ha simplificat considerablement. Com que les lectures de tercera generació cobreixen regions molt més llargues del genoma en una sola lectura (hem passat d'aproximadament 150-200 nucleòtids a milers de nucleòtids), és molt més senzill resoldre el trencaclosques i saber on col·locar cada peça, fins i tot, les procedents de zones repetitives del genoma. En contraposició, l'alta taxa d'error de les tecnologies de lectura llarga provoca que hi hagi errors a les seqüències assemblades i fa necessari un pas posterior de correcció d'aquestes seqüències per millorar-ne la precisió.

Com es comenta a l'article sobre tecnologies de seqüenciació present en aquest monogràfic, les dues principals companyies amb seqüenciació de lectura llarga són Oxford Nanopore Technologies (ONT) i Pacific Biosciences (PacBio). El fet que construir grafs de Bruijn a partir de lectures llargues doni lloc a estructures molt enrevessades ha provocat que s'hagin desenvolupat programes d'assemblatge específics per a lectures llargues basats en altres mètodes. En alguns casos, s'ha mantingut la idea del graf de Bruijn, però s'ha adaptat a les característiques i peculiaritats de les lectures llargues; un exemple en seria el programa Flye (Freire *et al.*, 2021), que es basa en grafs *A*-Bruijn, una modificació dels grafs de Bruijn que té en compte els encavalcaments entre les lectures. En altres casos, s'apliquen mètodes jeràrquics (*hierarchical genome assembly process*) (Al-Okaily, 2016), en els quals inicialment es produeixen diversos assemblatges amb les lectures de més qualitat i longitud i després es corregeixen i s'allarguen aquests miniassemblatges amb la resta de lectures que pertanyen a cada regió. Els programes Canu (Koren *et al.*, 2017), Falcon-Unzip (Chin *et al.*, 2016) i Hifiasm (Cheng *et al.*, 2021) es basen en aquesta idea.

Atès que les lectures llargues poden contenir molts errors, per tal de simplificar els grafs que es construeixen durant el procés d'assemblatge, alguns programes fan un pas de correcció de les lectures llargues (p. ex., Nextdenovo, *https://github.com/Nextomics/NextDenovo*). Alternativament, hi ha la possibilitat de corregir les lectures prèviament i després donar-les als programes d'assemblatge. Quina és la millor opció varia en funció de les dades de les quals disposem i de l'organisme que estiguem assemblant.

## 2. Correcció de la seqüència de cada un dels còntigs (refinament)

Els còntigs obtinguts a partir de lectures llargues sovint contenen errors en forma de petites insercions, delecions o canvis de nucleòtids. Abans de poder fer servir aquests blocs de seqüència en anàlisis posteriors, com, per exemple, en l'anotació de gens, cal corregir aquests errors. De fet, si intentéssim anotar gens directament amb els assemblatges no corregits, no podríem definir correctament les seqüències codificants, ja que en molts casos trobaríem errors en la codificació dels aminoàcids i codons de terminació enmig de la seqüència.

El procés de correcció de la seqüència conegut com a *refinament* o *polishing* consisteix a determinar un consens per cada posició a partir de totes les lectures que cobreixen els còntigs. Les lectures emprades en aquest pas poden ser les mateixes amb les quals s'ha fet l'assemblatge base o diferents. En assemblatges obtinguts amb ONT o PacBio SMRT és freqüent fer diverses rondes de correcció amb les mateixes lectures llargues i, després, afegir-hi unes quantes rondes de correcció amb un altre tipus de lectures més precises, generalment, lectures curtes d'Illumina (que tenen una taxa d'error molt més baixa). També és possible fer servir programes com Hypo (Kundu *et al.*, 2019), que poden corregir amb diversos tipus de lectures al mateix temps.

En assemblatges obtinguts a partir de lectures de tipus HiFi de PacBio no és necessari corregir amb altres tecnologies, ja que elles mateixes ja són fruit d'un procés de consens. De fet, s'obtenen a partir de múltiples lectures sobre la mateixa molècula de DNA, fet que fa que la seva taxa d'error sigui força baixa (<1 %). Per produir assemblatges a partir d'aquestes lectures s'acostuma a fer servir el programa HiFiasm (Cheng *et al.*, 2021), que utilitza les mateixes lectures per corregir els possibles errors i produeix assemblatges contigus que, en principi, no cal corregir.

Un cop obtingudes les seqüències consensuades corregides, encara ens queda un pas de correcció addicional: eliminar duplicacions falses i separar els haplotips. Tot i que la majoria d'organismes contenen més d'una còpia del genoma, els assemblatges de referència idealment en contenen una sola còpia, és a dir, són una representació haploide del genoma. No obstant això, atès que ambdues còpies d'un mateix individu no són idèntiques, és possible que els programes d'assemblatge no detectin tots els casos i sovint més d'una còpia de certes regions acaba introduïda a l'assemblatge inicial. Per eliminar aquestes duplicacions, es tornen a alinear les lectures llargues i, basant-se en els alineaments, s'intenta detectar quines parts assemblades corresponen a la mateixa

regió del genoma. Com a resultat del procés de detecció de duplicats (per exemple, amb el programa Purge_dups (Guan *et al.*, 2020), s'obté un assemblatge primari que hauria de correspondre a una de les còpies del genoma. A més, si les lectures són prou acurades, és possible obtenir assemblatges secundaris amb tots els haplotips alternatius.

## 3. Reconstrucció de cromosomes (*chromosome scaffolding*)

Tot i que amb les lectures llargues podem reconstruir blocs molt llargs, és poc freqüent aconseguir assemblar cromosomes sencers sense ajuda d'informació suplementària. Per tant, un cop el nostre assemblatge és prou contigu i complet (per exemple, amb diverses megabases [Mb] de N50, un QV de més de 40 i amb més d'un 90 % de gens BUSCO i *k*-mers presents), podem procedir a emplaçar els blocs seqüència en estructures més llargues, idealment de la mida dels cromosomes. És el moment, doncs, d'ordenar i orientar les peces per crear superestructures (*super-scaffolds*) amb diversos blocs de seqüència connectats per un nombre arbitrari de *n* (sovint un valor fix, p. ex. 100). Hi ha diferents aproximacions que permeten aquest pas: mapes genètics, mapes òptics (p. ex. Bionano), mapes de contacte (p. ex. Hi-C), etc.

Durant dècades, els biòlegs han estudiat àmpliament nombroses espècies, de les quals, tot i no conèixer-ne amb exactitud la totalitat del genoma, han estat capaços d'extreure molta informació. Gràcies als esforços de molts investigadors, disposem de mapes genètics per a moltes d'aquestes espècies, que consisteixen en llistes de marcadors dels quals coneixem la seqüència i localització en el genoma. Aquests mapes es poden fer servir per ordenar i orientar les seqüències assemblades i, si aquestes són prou llargues, ens permeten reconstruir els cromosomes (p. ex. Guerrero-Cózar *et al.*, 2021).

Com que no hi ha mapes genètics per a totes les espècies, és important poder reconstruir els cromosomes amb altres mètodes. Amb aquesta finalitat, podem seqüenciar amb tecnologies com Bionano (*https://bionanogenomics.com/*) o Hi-C (Belton *et al.*, 2012).

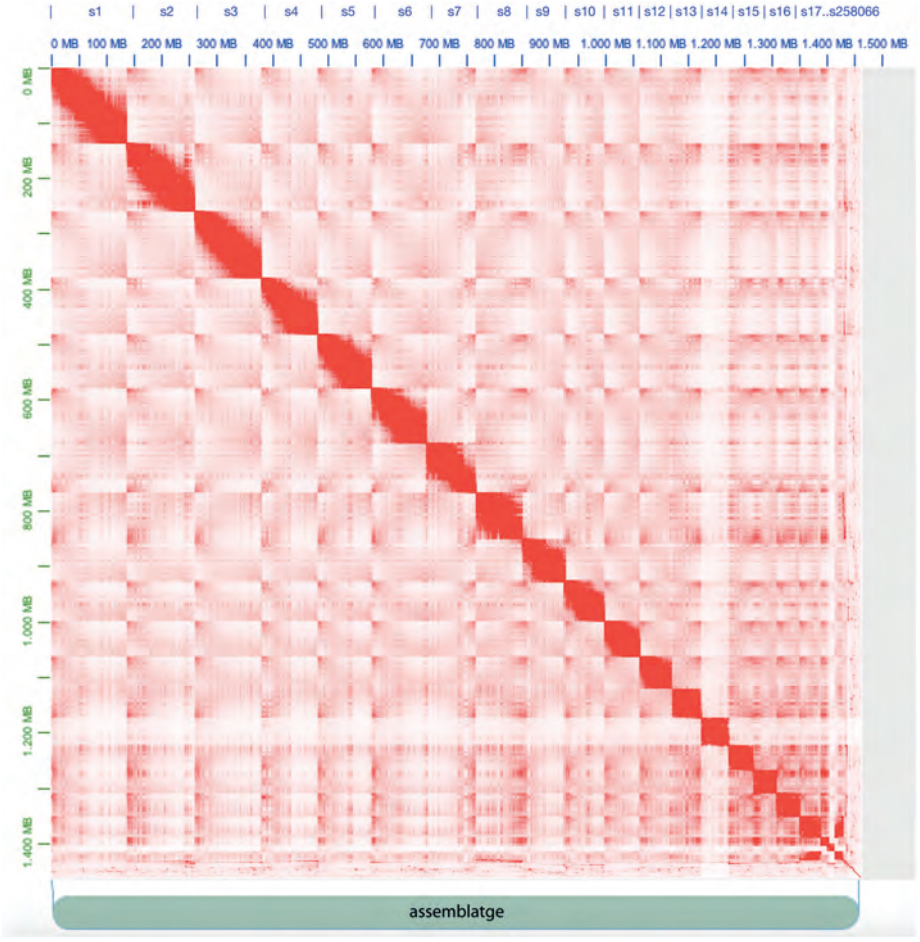
Bionano Genomics ha desenvolupat una tecnologia que permet fer «fotografies» d'una molècula de DNA. El procés consisteix a marcar el DNA en determinades seqüències, capturar la imatge amb un instrument especialitzat a detectar-ne el senyal i fer-ne mapes òptics. A continuació, podem comparar els mapes òptics i les seqüències que hem assemblat prèviament per tal de detectar llocs d'unió entre els nostres blocs de seqüència i així crear superestructures.

L'altra opció àmpliament emprada per a la reconstrucció de cromosomes són els mapes de contacte. Tècniques com el Hi-C permeten detectar les interaccions de la cromatina al nucli cel·lular i construir matrius de distància que reflecteixen la conformació en tres dimensions del genoma. Aquestes tècniques es basen en el fet que la probabilitat d'interacció disminueix ràpidament amb l'augment de la distància genòmica. L'anàlisi d'aquestes matrius de distància fa possible detectar tant zones properes com regions separades per diversos centenars de megabases en el mateix cromosoma. Un cop detectades les interaccions entre els fragments assemblats, podem unir els que interactuen i crear una macroestructura que corresponen al cromosoma (figura 3). Aquest procés es coneix com a *Hi-C scaffolding* i els millors programes per dur-lo a terme són 3Ddna (Dudchenko *et al.*, 2017), SALSA2 (Ghurye *et al.*, 2019) i YaHS (Chenxi Zhou *et al.*, 2022). Aquest darrer, que s'ha desenvolupat recentment, incorpora noves maneres de netejar el graf que donen lloc a superestructures de molta qualitat.

Les tècniques descrites anteriorment són complementàries, de manera que és freqüent combinar diversos mètodes per tal d'arribar a obtenir el millor assemblatge possible, és a dir, el més contigu, complet i amb menys errors.

## 4. Perfeccionament de l'assemblatge

Un cop finalitzat el pas de *scaffolding*, és possible que encara quedin algunes seqüències curtes sense col·locar o que algunes de les peces no




↑ Figura 3. Mapa de contactes Hi-C que mostra com la intensitat dels contactes es correlaciona amb la proximitat al llarg de cada seqüència. Aquest cas en concret mostra el genoma d'1,5 Gb d'un vertebrat heterogamètic. S'observa una reducció del nombre de contactes a la seqüència 14 (s14), que correspon al cromosoma sexual més llarg, degut a la presència d'una sola còpia d'aquest cromosoma al genoma. Per conveniència, s'han marcat només les 17 seqüències més llargues. El nombre total de seqüències assemblades és 258.066, la gran majoria de les quals són petites porcions del genoma que romanen sense col·locar pel seu alt contingut en repeticions. Elaboració pròpia.

es trobin correctament orientades en el cromosoma. Generalment, es tracta de blocs massa curts que no han pogut ser emplaçats correctament en les superestructures per manca d'informació en els mapes que els relacioni amb la resta de l'assemblatge. Per tal d'intentar col·locar aquestes peces, podem fer una revisió del mapa de contactes i reorganitzar els blocs de seqüència, assegurant-nos que els contactes siguin més freqüents en regions veïnes. A més, una altra manera d'intentar trobar-los el seu lloc és tornant a recórrer a les lectures llargues. Si tornem a buscar les lectures a l'assemblatge on tenim els cromosomes i els fragments restants, podem trobar lectures que continguin parts de la seqüència present a les superestructures i parts als còntigs o *scaffolds* deslocalitzats. Programes com Dentist (Ludwig *et al.*, 2021) o RagTag (Alonge *et al.*, 2019) permeten omplir els buits que s'han generat durant el procés de *scaffolding* amb els fragments curts que no s'havien pogut col·locar. Aquest procés es coneix com a *emplenament de buits* (*gap filling*, en anglès).

Un altre motiu pel qual alguns fragments no poden ser col·locats als cromosomes és perquè no pertanyen al genoma que estem assemblant. En alguns casos, és possible trobar contaminació, ja sigui de genomes d'orgànuls cel·lulars, d'organismes que viuen dins del mateix individu (endosimbionts) o bé seqüències contaminants d'altres organismes. Quan detectem la presència d'aquests fragments exògens, els podem eliminar de l'assemblatge final durant el procés de descontaminació. Una bona eina per a la detecció d'aquests contaminants és Blobtools (Challis *et al.*, 2020).

Després de tots aquests passos, és possible que encara quedin algunes peces sense col·locar. Aquestes seqüències, que sovint són curtes, altament repetitives i pobres en gens, es poden mantenir al final del fitxer amb l'assemblatge, ja que poden ser útils per a determinats tipus d'anàlisis.

## Selecció de l'assemblatge final

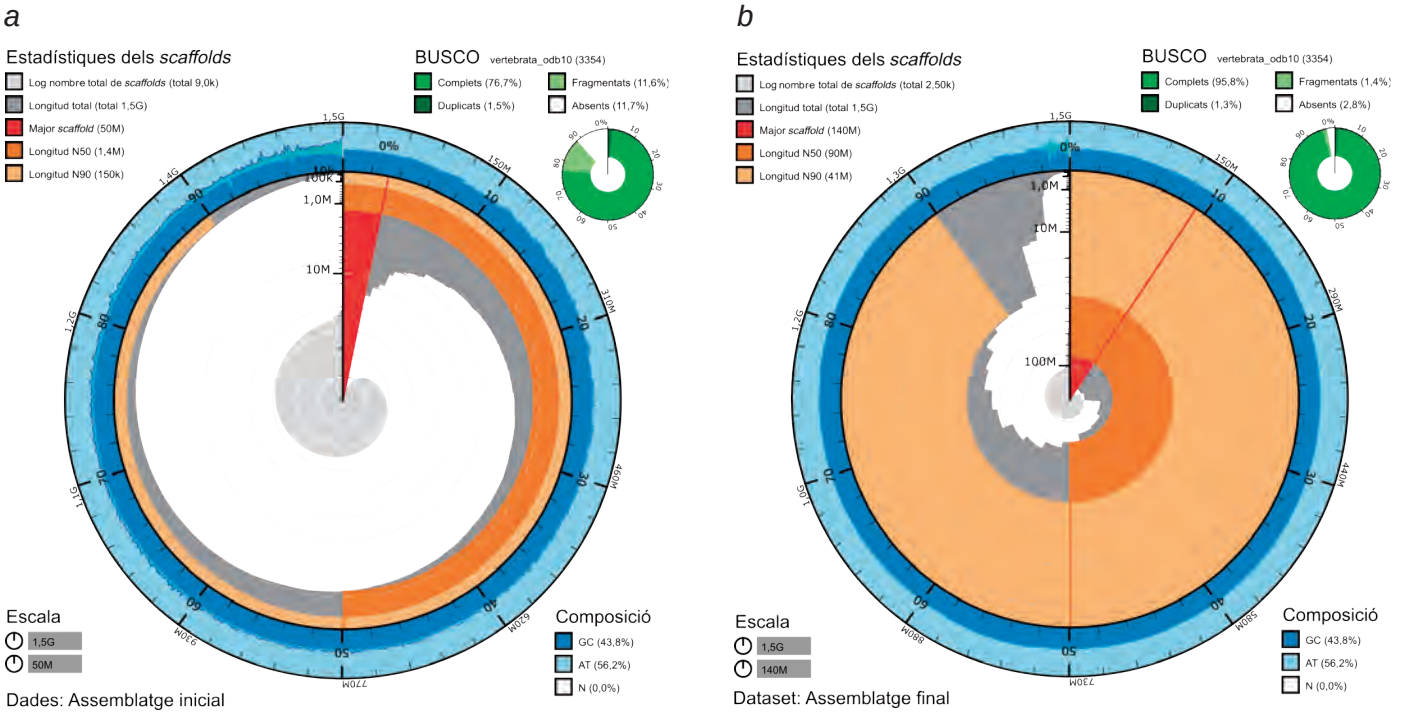
Durant tot el procés d'assemblatge d'un genoma es produeixen assemblatges intermedis que poden ser avaluats amb les tècniques descrites a l'apartat «Avaluació d'un assemblatge». A més, com que les característiques del genoma de diferents organismes poden ser molt variades (per exemple, aspectes com la mida, l'heterozigositat, la ploïdia, les regions repetitives, la manera de determinar el sexe, etc.), és difícil seleccionar una única estratègia universal que sigui la millor en tots els casos. Per tant, sovint es proven diferents tècniques d'assemblatge i es comparen entre elles per tal de determinar la combinació òptima en cada cas.

A la figura 4 es mostren, a tall d'exemple, els resultats obtinguts en assemblar el genoma d'1,5 × $10^9$ nucleòtids (Gb) d'un vertebrat. La primera figura (*a*) recull les mètriques després del primer pas del procés d'assemblatge, és a dir, després de la construcció dels còntigs (figura 2). A la segona figura (*b*) s'observen els estadístics obtinguts durant l'assemblatge del mateix vertebrat, però després de tot el procés, en concret, després de la reconstrucció de cromosomes amb dades de Hi-C. En comparar ambdues imatges veiem que el primer assemblatge conté moltes més seqüències i que són molt més curtes. A més, el percentatge de gens BUSCO complets és molt més baix en el primer cas perquè el pas de refinament o correcció de la seqüència encara no ha tingut lloc.

L'Earth BioGenome Project (Lawniczak *et al.*, 2022), que té com a objectiu seqüenciar i assemblar el genoma de totes les espècies del planeta, ha definit uns estàndards de qualitat mínims per als assemblatges obtinguts (Lewin *et al.*, 2022). En resum, es considera un genoma d'alta qualitat si compleix els requisits següents:

— més d'un 90 % de l'assemblatge emplaçat en cromosomes;

— N50 dels còntigs més gran d'una megabase;




↑ Figura 4. Gràfics de cargol (*snail plots*) (Challis *et al.*, 2020) que mostren les mètriques obtingudes durant l'assemblatge d'un vertebrat. *a*) Mostra de l'assemblatge inicial generat a partir de les lectures llargues i *b*) mostra de l'assemblatge final generat després de tot el procés exposat a la figura 2. Elaboració pròpia.

— QV superior a 40;

— més d'un 90 % de gens BUSCO complets;

— més d'un 90 % de *k*-mers presents.

Un cop finalitzat el procés i seleccionat el nostre assemblatge final, ja podem anotar-lo i fer-lo servir per a múltiples tipus d'estudis i anàlisis. Mitjançant aquest procés podem descobrir aspectes únics de la nostra espècie d'interès, comparar expressions gèniques entre diferents condicions, analitzar poblacions o, fins i tot, comparar el genoma de la nostra espècie amb el d'altres espècies. Gràcies a la presència de genomes de referència de bona qualitat és possible estudiar l'espectre complet de la variació genètica a la natura, fet que millora molt la nostra capacitat a l'hora de desenvolupar estratègies de conservació i també d'estudiar interaccions ecològiques dins d'un mateix ecosistema (Formenti *et al*., 2022). Aquesta informació serà clau per mantenir la diversitat a escala global.

## Genomes de referència assemblats a Catalunya

Molts grups catalans han assemblat i publicat els genomes de diverses espècies al llarg dels anys. El Centre Nacional d'Anàlisi Genòmica - Centre de Regulació Genòmica (CNAG-CRG) de Barcelona, un dels centres europeus de referència en seqüenciació i anàlisi de dades genòmiques, té un equip que es dedica a produir assemblatges i anotacions de bona qualitat de tot tipus d'organismes (*https://denovo.cnag.cat/*). Entre els genomes publicats en destaquen els d'algunes plantes com l'olivera (Cruz *et al*., 2016; Julca *et al*., 2020) i l'ametller (Alioto *et al*., 2020); vertebrats com el turbot (Figueras *et al*., 2016) i el linx ibèric (Abascal *et al*., 2016); insectes com la mosca *Drosophila guanche* (Puerma *et al*., 2018) i l'efemeròpter *Cloeon dipterum* (Almudi *et al*., 2020), i bivalves com el musclo (Gerdol *et al*., 2020).

En els darrers anys han aparegut múltiples iniciatives per generar genomes de referència de les espècies del planeta. Alguns d'aquests projectes se centren en branques taxonòmiques concretes, com el Vertebrate Genomes Project (*https://vertebrategenomesproject.org/*) o el Bird 10K (*https://b10k.genomics.cn/*), que assemblen genomes de vertebrats i d'ocells, respectivament. Per una altra banda, trobem els projectes enfocats a determinades zones geogràfiques, com el Darwin Tree of Life (*https://www.darwintreeoflife.org/*), l'African BioGenome Project (*https://africanbiogenome.org/*) o la iniciativa catalana per a l'Earth BioGenome Project (*https://www.biogenoma.cat/*). Tots aquests projectes es troben emmarcats dins de l'Earth BioGenome Project, que té l'objectiu ambiciós de generar almenys un genoma de referència per cada espècie eucariota de la Terra. Científics que desenvolupen la seva tasca professional al llarg de tots els territoris de parla catalana estan contribuint que aquesta fita sigui possible.

## Bibliografia


Abascal, F. [*et al.*] (2016). «Extreme genomic erosion after recurrent demographic bottlenecks in the highly endangered Iberian lynx». *Genome Biol.*, 17: 251.

Al-Okaily, A. A. (2016). «HGA: *De novo* genome assembly method for bacterial genomes using high coverage short sequencing reads». *BMC Genomics*, 17: 193.

Alioto, T. [*et al.*] (2020). «Transposons played a major role in the diversification between the closely related almond and peach genomes: Results from the almond genome sequence». *Plant J.*, 101: 455-472.

Almudi, I. [*et al.*] (2020). «Genomic adaptations to aquatic and aerial life in mayflies and the origin of insect wings». *Nat. Commun.*, 11: 2631.

Alonge, M. [*et al.*] (2019). «RaGOO: fast and accurate reference-guided scaffolding of draft genomes». *Genome Biol.*, 20: 224.

Bankevich, A. [*et al.*] (2012). «SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing». *J. Comput. Biol.*, 19: 455-477.

Belton, J. M. [*et al.*] (2012). «Hi-C: A comprehensive technique to capture the conformation of genomes». *Methods*, 58: 268-276.

Challis, R. [*et al.*] (2020). «BlobToolKit - Interactive Quality Assessment of Genome Assemblies». *G3*, 10: 1361-1374.

Cheng, H. [*et al.*] (2021). «Haplotype-resolved *de novo* assembly using phased assembly graphs with Hifiasm». *Nat. Methods*, 18: 170-175.

Chin, C. S. [*et al.*] (2016). «Phased diploid genome assembly with single-molecule real-time sequencing». *Nat. Methods*, 13: 1050-1054.

Cruz, F. [*et al.*] (2016). «Genome sequence of the olive tree, *Olea europaea*». *Gigascience*, 5: 29.

Dudchenko, O. [*et al.*] (2017). «*De novo* assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds». *Science*, 356: 92-95.

Earl, D. [*et al.*] (2011). «Assemblathon 1: A competitive assessment of de novo short read assembly methods». *Genome Res.*, 21: 2224-2241.

Figueras, A. [*et al.*] (2016). «Whole genome sequencing of turbot (*Scophthalmus maximus*; *Pleuronectiformes*): A fish adapted to demersal life». *DNA Res.*, 23: 181-192.

Formenti, G. [*et al.*] (2022). «The era of reference genomes in conservation genomics». *Trends Ecol. Evol.* [en línia], 37 (3): 197-202. <https://doi.org/10.1016/j.tree.2021.11.008>.

Freire, B. [*et al.*] (2021). «Memory-efficient assembly using flye». *IEEE/ACM Trans. Comput. Biol. Bioinform.*

Gerdol, M. [*et al.*] (2020). «Massive gene presence-absence variation shapes an open pan-genome in the Mediterranean mussel». *Genome Biol.*, 21: 275.

Ghurye, J. [*et al.*] (2019). «Integrating Hi-C links with assembly graphs for chromosome-scale assembly». *PLoS Comput. Biol.*, 15: e1007273.

Guan, D. [*et al.*] (2020). «Identifying and removing haplotypic duplication in primary genome assemblies». *Bioinformatics*, 36: 2896-2898.

Guerrero-Cózar, I. [*et al.*] (2021). «Chromosome anchoring in Senegalese sole (*Solea senegalensis*) reveals sex-associated markers and genome rearrangements in flatfish». *Sci. Rep.*, 11: 13460.

International Human Genome Sequencing Consortium [*et al.*] (2001). «Initial sequencing and analysis of the human genome». *Nature*, 409: 860-921.

Julca, I. [*et al.*] (2020). «Genomic evidence for recurrent genetic admixture during the domestication of Mediterranean olive trees (*Olea europaea* L.)». *BMC Biol.*, 18: 148.

Koren, S. [*et al.*] (2017). «Canu: Scalable and accurate long-read assembly via adaptive *k*-mer weighting and repeat separation». *Genome Res.*, 27: 722-736.

Kundu, R. [*et al.*] (2019). «Hypo: super fast & accurate polisher for long read genome assemblies». *BioRxiv*. <https://doi.org/10.1101/2019.12.19.882506>.

Lawniczak, M. K. N. [*et al.*] (2022). «Standards recommendations for the Earth BioGenome Project». *Proc. Natl. Acad. Sci. USA*, 119.

Lewin, H. A. [*et al.*] (2022). «The Earth BioGenome Project 2020: Starting the clock». *Proc. Natl. Acad. Sci. USA*, 119 (4): e2115635118.

Li, R. [*et al.*] (2010). «*De novo* assembly of human genomes with massively parallel short read sequencing». *Genome Res.*, 20: 265-272.

Ludwig, A. [*et al.*] (2021). «DENTIST – using long reads to close assembly gaps at high accuracy». *BioRxiv*. <https://doi.org/10.1101/2021.02.26.432990>.

Mouse Genome Sequencing Consortium [*et al.*] (2002). «Initial sequencing and comparative analysis of the mouse genome». *Nature*, 420: 520-562.

Parra, G. [*et al.*] (2007). «CEGMA: A pipeline to accurately annotate core genes in eukaryotic genomes». *Bioinformatics*, 23: 1061-1067.

Pevzner, P. A. [*et al.*] (2001). «An Eulerian path approach to DNA fragment assembly». *Proc. Natl. Acad. Sci. USA*, 98: 9748-9753.

Puerma, E. [*et al.*] (2018). «The high-quality genome sequence of the oceanic island endemic species *Drosophila guanche* reveals signals of adaptive evolution in genes related to flight and genome stability». *Genome Biol. Evol.*, 10: 1956-1969.

Rhie, A. [*et al.*] (2020). «Merqury: Reference-free quality, completeness, and phasing assessment for genome assemblies». *Genome Biol.*, 21: 245.

— (2021). «Towards complete and error-free genome assemblies of all vertebrate species». *Nature*, 592: 737-746.

Simão, F. A. [*et al.*] (2015). «BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs». *Bioinformatics*, 31: 3210-3212.

Simpson, J. T. [*et al.*] (2009). «ABySS: A parallel assembler for short read sequence data». *Genome Res.*, 19: 1117-1123.

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

Zdobnov, E. M. [*et al.*] (2021). «OrthoDB in 2020: Evolutionary and functional annotations of orthologs». *Nucleic Acids Res.*, 49: D389-D393.

Zerbino, D. R.; Birney, E. (2008). «Velvet: Algorithms for *de novo* short read assembly using de Bruijn graphs». *Genome Res.*, 18: 821-829.

Zhou, C. [*et al.*] (2022). «YaHS: yet another Hi-C scaffolding tool». *BioRxiv* [en línia], 495093. <https://doi.org/10.1101/2022.06.09.495093>.