GENS I LLENGÜES EN EVOLUCIÓ: UN RECORD PER A CAVALLI-SFORZA[1]

Jaume Bertranpetit
Institut de Biologia Evolutiva (UPF-CSIC)

Luca L. Cavalli Sforza ha estat un genetista amb una enorme capacitat de creuar disciplines, incloent la lingüística i molt especialment la lingüística comparativa i històrica. Nascut el 1922, fou professor de genètica a Parma i Pavia, treballà molts anys a Stanford (Califòrnia) i ens deixà l'any 2018. Des de la seva dedicació a la genètica de les poblacions humanes ha excel·lit en diferents camps com la genètica estadística o la iniciació de la disciplina de la transmissió i evolució cultural. El seu treball intentà, en diferents fases i regions, reconstruir la història de les poblacions humanes a partir de les dades genètiques i relacionar aquesta informació amb la proporcionada per la prehistòria, l'antropologia biològica i la lingüística històrica. Destaquen els treballs que lliguen la diversitat genètica a Europa amb l'expansió del Neolític i la possible expansió de les llengües indoeuropees; o els que mostren una coevolució de gens i llengües. Per altra banda, fou un gran lluitador contra el racisme.

EL MARC DEL TREBALL GENÈTIC

L'any 1859 Charles Darwin publica la primera edició del llibre *On the origin of species by means of natural selection, or the preservation of favoured races in the struggle for life*, que en el capítol 13 assenyala:

> If we possessed a perfect pedigree of mankind, a genealogical arrangement of the races of man would afford the best classification of the various languages now spoken throughout the world; and if all extinct languages, and all intermediate and slowly changing dialects, had to be included, such an arrangement would, I think, be the only possible one.... but the proper or even only possible arrangement would still be genealogical; and this would be

1. Agraeixo al professor Luca Cavalli-Sforza el mestratge fonamental de la meva carrera científica, i a la seva família per les llargues amistats. Gràcies a Mariona Costa per la correcció del text.

Estudis Romànics [Institut d'Estudis Catalans], vol. 42 (2020), p. 257-265
DOI: 10.2436/20.2500.01.295

> strictly natural, as it would connect together all languages, extinct and modern, by the closest affinities, and would give the filiation and origin of each tongue.

Darwin fa una reflexió comuna sobre els processos evolutius dels éssers vius i els de les llengües parlades i ens aporta dos conceptes diferents fonamentals: el primer és que la classificació de la diversitat donada per processos evolutius cal fer-la en un marc filogenètic, és a dir que, quan classifiquem entitats que han sorgit d'un procés de diferenciació (estrelles en l'univers, espècies o poblacions dins dels éssers vius, diversitat de les llengües parlades), ho hem de fer de tal manera que recapituli el procés evolutiu que ha donat lloc a la diversitat observada; no s'hi val a classificar en si mateix, ja que la diversitat té un sentit donat per uns processos de canvi temporals. El segon és que la classificació de poblacions humanes i de les llengües parlades és única i completa i, per tant, mirant les poblacions i les llengües que parlen, estem mirant dues cares d'una mateixa moneda, del mateix procés històric de la diferenciació entre els grups humans.

Grans desenvolupaments d'aquestes (i altres) idees han estat dutes a terme per Cavalli-Sforza, que ha estat un puntal per entendre les regles del procés evolutiu que, a partir de la diversitat d'un moment determinat, ens permet reconstruir el passat i ha fet confluir diferents disciplines envers la reconstrucció única del nostre passat.

EVOLUCIÓ, ARBRES I FILOGÈNIES

Ens cal plantejar primer com podem descriure la diversitat humana i ens centrem inicialment en la seva diversitat biològica. En l'època de Darwin es coneixia únicament la diversitat *morfològica* humana a través de l'aspecte visible dels individus, i les múltiples classificacions racials que s'havien fet es basaven en trets aparents: color de la pell i cabell, alçada, forma d'ulls, etc. No es contemplava ni tan sols la possibilitat que, observant i descrivint la diferència, s'estigués parlant d'evolució. Aquests trets morfològics tenen un gran inconvenient: no canvien de manera regular en el transcurs del temps i no s'hi pot aplicar un rellotge evolutiu: a vegades canvien molt de pressa i altres molt lentament, depenent de la pressió de la selecció natural. Com podem buscar una descripció biològica dels éssers vius que sí permeti entendre els mecanismes i les velocitats de canvi?

Cavalli-Sforza va ser pioner en la proposta de fer servir la informació genètica per entendre les diferències entre els grups humans. Els organismes vius i, és clar, els humans, poden ser vistos com l'única part del món natural en què els seus membres contenen una descripció interna d'ells mateixos: tenim la nostra pròpia informació codificada en el genoma, format per ADN. I llegint i comparant l'ADN podem tenir una aproximació fantàstica de la diferència entre individus i entre poblacions humanes. Això és el que va portar Cavalli-Sforza primer a desenvolupar un cos doctrinal teòric sobre la mesura de les diferències genètiques entre poblacions i el seu modelat per processos evolutius estocàstics, fonamentalment la deriva genètica, que tendeix a diferenciar poblacions quan són petites. Podem considerar el sistema genètic com a paradigma d'un sistema que pot

canviar en el temps segons un model estocàstic depenent de pocs paràmetres (especialment, grandària de població).

En el moment que Cavalli-Sforza es proposa reconstruir la història de les poblacions humanes amb dades genètiques, la informació disponible era poca i va poder utilitzar, només, el que ara anomenem «marcadors genètics clàssics», sense poder observar directament la variació a l'ADN. Però va poder recollir moltes dades genètiques de poblacions de tot el món. Un cop hem mesurat diferències, ve una segona part: com condensar-les per tal que ens mostrin un procés de diversificació i, per tant, reconstrueixin una filogènia. Entre les solucions que l'estadística ofereix, Cavalli-Sforza va abraçar-ne una: construir arbres que alhora ens donen una topologia de les relacions entre les unitats (poblacions, llengües) que permet inferir el procés evolutiu i que ens en dona el marc temporal.

ELS MAPES SINTÈTICS

Cavalli-Sforza era molt conscient que agafar poblacions com a unitats d'estudi podia ser problemàtic, fonamentalment perquè els arbres, en la seva estructura fonamental, no admeten flux (genètic) entre les branques. L'arbre és un gran model de diversificació d'un conjunt d'entitats que evolucionen independentment. La solució l'aportà a través d'una anàlisi geogràfica, senzilla de fer quan es mira la variació en l'espai d'un únic caràcter (una variant genètica, una paraula), però que llavors no ens informa de la població, sinó del caràcter. Per això cal considerar molts caràcters alhora i per observar els canvis de tots ells necessitem tècniques de simplificar la complexitat, com és el cas de l'anàlisi de components principals (PCA) o tècniques similars, en què es divideix la diversitat en components independents, additius i ordenats per la importància relativa a l'hora d'explicar la variació total. Així, si posem conjuntament la variació en 1.000 caràcters, el primer component principal pot explicar, per exemple, un 10 % del conjunt de la variació, i els components següents, que seran independents, aniran disminuint la proporció de variació explicada.

Els mapes sintètics tenen un objectiu de fer servir molta informació alhora per no dependre dels marcadors genètics concrets i no ens importa com canvia cada caràcter, sinó el conjunt de tots ells. Mirant qualsevol part del genoma humà trobem molta variació, i el que hem de fer és veure patrons general de variació, els quals ens explicaran processos de canvi de les poblacions i no del genoma. El genoma humà és molt gran i té de l'ordre de tres mil milions d'unitats químiques d'informació (les bases de l'ADN: A, T, C o G); per altra banda, per raons purament evolutives, l'espècie humana és molt poc diversa, ja que som una espècie molt recent (no més enllà de 200.000 anys d'història) i trobem només una variant cada 1.000 unitats d'informació de mitjana. Amb tot això podem analitzar la variació genètica en milions de posicions del genoma i avui (no en el moment que Cavalli va fer els grans estudis) disposem de milions de variants que poden analitzar-se per un preu reduït i amb tècniques ràpides i eficients.

Amb tot, doncs, podem obtenir les representacions de la diversitat dels genomes humans dins d'un marc geogràfic, veient com les diferències genètiques en conjunt poden resseguir fronteres geogràfiques, polítiques o lingüístiques i ens ofereixen un marc interpretatiu de l'evolució demogràfica de les poblacions humanes, en la qual els processos de creixement, d'expansió cap a altres àrees, i de migracions estaran a la base de les explicacions de la variació genètica (Cavalli-Sforza *et al*., 1993).

És important assenyalar, en aquest context, que estem defugint les explicacions estrictament genòmiques de la variació genètica observada (no volem el mapa de variació d'un caràcter), i molt especialment de les regions del genoma que puguin haver estat sotmeses a selecció natural. La variació en aquestes regions no depèn de la història de la població, sinó de la història adaptativa de la regió genòmica concreta, per exemple que hi hagi una variant que s'hagi seleccionat positivament perquè confereix resistència a una malaltia infecciosa.

Tenim, doncs, la visió de la variació genètica dins d'un marc geogràfic concret i mostrant extrems de variació que han d'haver estat produïts per factors demogràfics i per tant produïts per la història de la població. Aquest és un dels punts essencials que impulsà Cavalli-Sforza per demostrar per què amb els gens podem reconstruir la història.

DIVERSITAT GENÈTICA I HISTÒRIA DE LES POBLACIONS

Hi ha un marc històric de les poblacions humanes en el qual esdeveniments del passat, molts d'ells amb implicacions demogràfiques, ens expliquen l'estructura geogràfica de la diversitat genètica i, de fet, molts aspectes d'aquest passat els hem pogut reconstruir gràcies als coneixements genètics. No només l'origen africà i modern de tota la humanitat actual, sinó molts aspectes més reduïts i puntuals de la història de les poblacions. Això és el que va mostrar Cavalli-Sforza en la seva obra màxima, *The History and Geography of Human Genes*. En aquest extens llibre (*il librone*, en deia ell bromejant) hi ha una detallada explicació d'allò que sabíem en el seu moment, el 1994, de la reconstrucció del passat de les poblacions humanes fent servir molts tipus d'informació, però posant per davant la informació genètica, que en molts casos era original (Cavalli-Sforza et al., 1994).

Va basar-se en una recopilació de variació genètica que va dur a terme amb els seus col·laboradors fent servir les variants genètiques conegudes a l'època, els marcadors clàssics, en un nombre que ara trobem molt baix d'uns 120 polimorfismes, i que tots aquests son variants funcionals (com els grups sanguinis), que no són els òptims per a aquests estudis. Sobre aquestes dades, s'hi aplicà tant una anàlisi d'arbres com de components principals i, donat el patró de variació, es pot atribuir una explicació causal (però hipotètica) en el passat de les poblacions que sigui compatible amb allò conegut des dels estudis de la seva història. De fet, normalment es referiran a la prehistòria, ja que els esdeveniments que més han impactat en el nostre passat genètic són els que tingueren lloc quan les poblacions eren molt reduïdes i es podien donar més processos de diferenciació especialment per deriva genètica. Així, per exemple, Cavalli-Sforza va explicar el paisa-

tge genètic d'Europa a partir dels components principals i va inferir els factors que els haurien produït (Figura 1). El primer factor i més important de tota la variació genètica (n'explica un 28 %) mostra un gradient entre Orient Mitjà i el nord-oest d'Europa i va postular que hauria estat produït per l'expansió del Neolític a Europa; de fet el mapa genètic és pràcticament el mateix que s'obté posant les dates d'expansió del Neolític, que es va iniciar fa uns 10.000 anys a l'Orient Mitjà i s'estengué paulatinament pel continent.

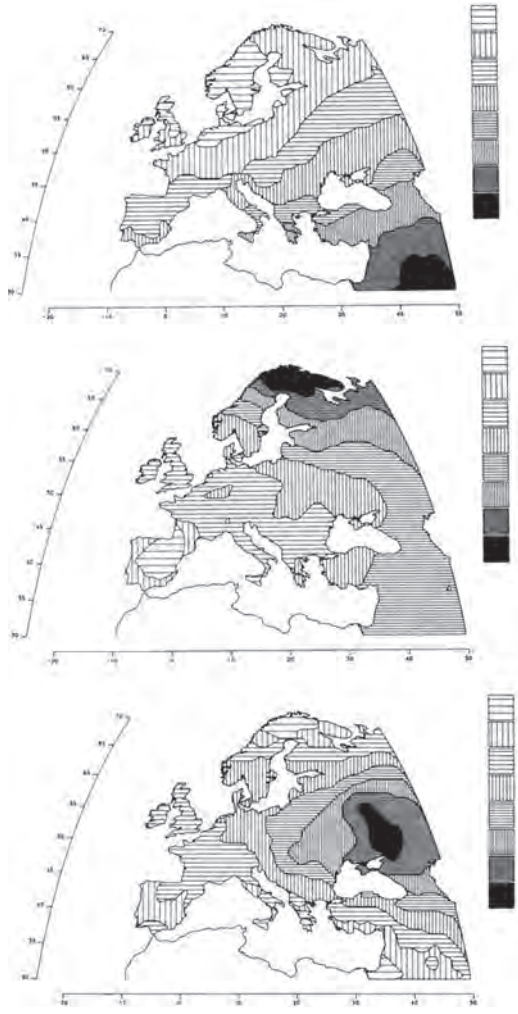

Figura 1. Patrons amagats de la variació genètica a Europa vistos pels primers components principals. El primer, imposable a l'expansió del Neolític a Europa. El segon mostra la diferència del nord, d'influència uràlica. El tercer mostra la regió dels Urals com a centre de radiació arqueològica amb els nòmades de les cultures Kurgan i possible bressol de les llengües indoeuropees.

Clarament hi ha una correlació entre els dos fenòmens, però les dades genètiques que Cavalli-Sforza va fer servir no eren capaces de posar dates concretes i sabem bé que una correlació no es un fenomen causal. Hi ha hagut molts, potser centenars d'estudis que han reanalitzat la variació genètica a Europa i les interpretacions han anat canviant a mesura que les anàlisis genètiques s'han fet més exhaustives, i els tractaments numèrics més sofisticats. Avui, gràcies a l'estudi de genomes sencers tant de les poblacions actuals com del passat (ADN antic), es considera provat l'impacte genètic de l'expansió del Neolític a Europa que Cavalli-Sforza va proposar. Però no en tots els seus detalls.

El que sí ha quedat és com l'empremta molt antiga roman en els nostres gens. En un dels seus treballs ens deia que, de la mateixa manera que quan mirem els estels estem veient la llum que varen projectar en un passat molt llunyà, quan mirem la diversitat genètica estem veient els efectes d'esdeveniments molt antics en la nostra història.

ELS PROCESSOS COMUNS EN GENS I LLENGÜES

Cavalli-Sforza va estar sempre fascinat per la possible estructuració de la diversitat lingüística. Tenia grans amics en la lingüística històrica i anava sovint a seminaris i llegia articles. Sens dubte van tenir-hi una gran influència tant el seu amic i company de Stanford Joseph Greenberg (1915-2001) com el continuador i divulgador dels seus estudis Meritt Ruhlen. I alhora va discutir molt el treball d'arqueòlegs que s'havien interessat en la lingüística històrica, especialment Marija Gimbutas i Colin Renfrew.

Influït pel treball d'aquest darrer (Renfrew 1987), en la interpretació que feu de l'impacte demogràfic i genètic del Neolític a Europa, va acceptar com a versemblant (sense proves clares) que aquesta expansió podria haver anat acompanyada de l'expansió de les llengües indoeuropees a Europa.

Tot i això, no hi ha hagut consens en aquesta hipòtesi. De fet Marija Gimbutas ja havia postulat la hipòtesi Kurgan (o de l'estepa o del túmuls), que és la proposta més acceptada actualment per identificar el lloc d'origen del protoindoeuropeu, a l'estepa pòntica al nord del Mar Negre, i a partir d'aquí es van estendre les llengües indoeuropees per tota Europa i algunes parts d'Àsia. Tant el lloc (estepa Pòntica i no Anatòlia) com el període (Edat del Bronze i no Neolític) no coincideixen i cal remarcar que les darreres dades genètiques, especialment del grup de David Reich amb moltes dades de genomes antics, enforteixen la proposta de Gimbutas, tot i que en la literatura es fa servir el terme de cultura Yamnaya i no el conjunt, potser heterogeni, de la cultura Kurgan. Tot i això, cal remarcar que el tercer dels components principals de Cavalli-Sforza (Figura 1) mostra aquest patró.

L'ARBRE DEL CONJUNT DE LA HUMANITAT

La relació entre gens i llengües fou tractada per Cavalli-Sforza també a un nivell global, que va intentar portar a terme la missió de Darwin: l'arbre de totes les poblacions i totes les llengües humanes, incloent-hi les extingides.

En aquesta aproximació Cavalli-Sforza va construir, per una banda, i com ja havia fet abans, l'arbre de 42 poblacions humanes amb dades genètiques i les va ajuntar, per una altra banda, segons les afinitats de les llengües parlades, fent servir majoritàriament les propostes de Ruhlen (1991), algunes molt contestades. Els resultats de l'anàlisi (Cavalli-Sforza *et al*., 1988 i 1992) es mostren a la figura 2, a l'esquerre les afinitats genè-

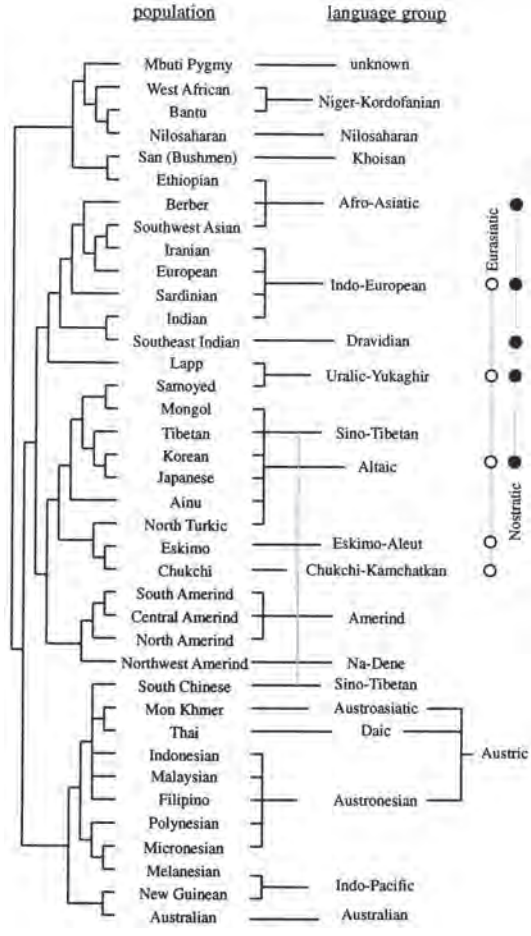

Figura 2. Arbre filogenètic de 42 poblacions humanes obtingut per dades genètiques (esquerre) que té coherència amb la classificació de la diversitat lingüística.

tiques i a la dreta les lingüístiques. Clarament hi ha una forta coincidència entre els dos arbres, tot i que és difícil de mesurar. Com era d'esperar, el treball fou molt criticat, tant des de la genètica (especialment pels pocs marcadors genètics emprats i per haver fet servir una estructura en arbre) com, especialment, des de la lingüística històrica (no hi ha acord amb les famílies emprades, algunes no es reconeixen i sobretot no hi ha acceptació de les entitats més grans, les "superfamílies" com el Nostràtic o Euroasiàtic).

Es tracta d'un treball pioner, amb clares mancances, però que no ha estat encara superat a nivell global. Per un costat l'arbre genètic ha estat superat per molts treballs que fonamentalment han fet un estudi aprofundit dels genomes fent servir centenars de milers de marcadors genètics, amb una quantitat realment enorme d'informació. El resultat presenta diferències amb l'arbre originalment proposat, però no gaires, i no gaire importants. El que resulta més discutit és la mateixa aproximació d'emprar un arbre per mostrar les semblances i diferències, quan sabem que els arbres no admeten migracions i sabem que són un factor essencial en la reconstrucció acurada de la història de les poblacions.

Per altra banda, des de la lingüística històrica, el punt de debat més fort és si es podrà, d'alguna manera, tirar enrere la barrera del temps que assenyala que les relacions entre les llengües no es poden reconèixer més enllà de 6.000-10.000 anys (Gray 2005), un temps que és curt per considerar el conjunt de la humanitat. Hi ha diversos estudis que intenten fer servir mètodes estadístics més potents o utilitzar informació que pugui tenir una taxa lenta d'evolució, més lenta que el lèxic. Hi ha debat sobre si fer servir dades tipològiques o estructurals de la llengua ajudarà a precisar relacions entre les llengües i tindran poder més enllà del sostre donat pel lèxic (Greenhill et al., 2017). Molt probablement la possible solució no sigui si triar el lèxic o la gramàtica, sinó quines estructures concretes dins de cada un d'aquests són més estables en el temps i per tant tenen una duració més llarga i poden ser emprats per establir parentius lingüístics més allunyats.

ALLÒ QUE CAVALLI-SFORZA JA NO VA VEURE

Hi ha molts conceptes que actualment, en biologia evolutiva, trobem normals i que formen part de la quotidianitat però que no existien fa poques dècades i sobre els quals Cavalli-Sforza tingué un fort impacte en el seu desenvolupament. Alguns serien:

– Coneixem la dinàmica dels processos subjacents a les diferències genètiques.

– La genètica pot recuperar la història de les poblacions humanes.

– Esdeveniments històrics antics poden ser a la base de la història de les poblacions (expansions, migracions, barreges).

– Cal reconstruir l'evolució humana mitjançant un enfocament multidisciplinari que inclou genètica, demografia, antropologia, arqueologia i lingüística.

Altres conceptes els va desenvolupar directament ell i n'hem repassat alguns exemples. Però en l'actualitat hi ha hagut un desenvolupament extraordinari en la biologia i computació que ha canviat radicalment les nostres aproximacions i que Cavalli-Sforza no va poder emprar. Podem fer seqüències completes de genomes humans amb

poc temps i amb un cost reduït, podem analitzar centenars de milers de variants genètiques també ràpidament; les tècniques de l'ADN antic es van optimitzant molt ràpidament i actualment molts laboratoris ja les fan servir amb garanties i tenim ordinadors i desenvolupament de software que permet anàlisis molt sofisticades. La lingüística històrica també millora en les bases de dades i en els mètodes, cada vegada més computacionals. Les confluències són possibles i desitjables.

Cavalli-Sforza ens va visitar a Barcelona l'any 1988 i va sembrar la llavor dels estudis genètics en profunditat. Actualment al país hi ha, en diferents universitats, bones escoles que segueixen el seu mestratge i que, gràcies a la seva inspiració i esperit obert a altres disciplines, han aconseguit assolir resultats que el mateix Cavalli no podia imaginar-se.

REFERÈNCIES

Cavalli-Sforza L. L. / Menozzi, P. / Piazza, A. (1993): «Demic expansions and human evolution», *Science*, 259, p. 639-646,

Cavalli-Sforza L. L. / Menozzi, P. / Piazza, A. (1994): *The History and Geography of Human Genes*. Princeton Univ. Press, Princeton, NJ.

Cavalli-Sforza L. L. / Menozzi, P. / Piazza, A. / Mountain, J. L. (1988): *Reconstruction of human evolution: bringing together genetic, archaeological and linguistic* data. Proc Natl Acad Sci USA 85: 6002-6006.

Cavalli-Sforza L. L. / Minch, E / Mountain, J. L. (1992) *Coevolution of genes and languages revisited*. Proc Natl Acad Sci USA 89: 5620-5624.

Gray, R. (2005) «Pushing the time barrier in the quest for language roots», *Science* 309, 2007-8.

Greenhill, S. J. / Wu, C. H. / Hua, X. / Dunn, M. / Levinson, S. C. / Gray, R. D. (2017): *Evolutionary dynamics of language systems*. Proc Natl Acad Sci USA. 114: E8822-E8829.

Renfrew, C. (1987) *Archeology and Language*. Cambridge: Univ. Press / U.K. Google Scholar.

Ruhlen, M. (1991) *A Guide to the Languages of the World* (Stanford Univ. Press, Stanford, CA).