

Note: This document contains the Catalan translation of the article “Inferring past demography and genetic adaptation in Spain using the GCAT cohort” by Garcia-Calleja et al. Scientific Reports (2025). <https://doi.org/10.1038/s41598-025-98272-w>. Please use this citation when referring to this work.

Nota: Aquest document conte la versió catalana de l'article “Inferring past demography and genetic adaptation in Spain using the GCAT cohort” per Garcia-Calleja et al. Scientific Reports (2025). <https://doi.org/10.1038/s41598-025-98272-w>. Si us plau, utilitzeu aquesta citació quan feu referència a aquest treball.

Inferència de la demografia passada i l'adaptació genètica a Espanya utilitzant la cohort del GCAT

Jorge Garcia-Calleja¹, Simone A Biagini^{1,2,3}, Rafael de Cid^{4,5}, Francesc Calafell ^{1*},
Elena Bosch^{1*}

1. Institute of Evolutionary Biology (UPF-CSIC), Department of Medicine and Life Sciences, Universitat Pompeu Fabra, Barcelona 08003, Spain
2. Department of Archaeology and Museology, Masaryk University, Brno, Czech Republic.
3. Center of Molecular Medicine, Central European Institute of Technology, Masaryk University, Brno, Czech Republic.
4. Genomes for Life-GCAT lab, CORE Program, Germans Trias i Pujol Research Institute (IGTP), Badalona 08916, Spain.
5. Grup de REcerca en Impacte de les Malalties Cròniques i les seves Trajectòries (GRIMTra), Germans Trias i Pujol Research Institute (IGTP), Badalona 08916, Spain

* Autors per a correspondència

francesc.calafell@upf.edu

elena.bosch@upf.edu

Resum

Situada a l'extrem sud-oest d'Europa, la Península Ibèrica està separada de la resta del continent pels Pirineus i d'Àfrica per l'estret de Gibraltar. Aquesta posició geogràfica podria haver condicionat pressions selectives diferenciades respecte a la resta d'Europa i influït en patrons específics de flux gènic. En aquest treball analitzem 704 seqüències de genomes complets del panell de referència del GCAT per a quantificar el flux gènic cap a Espanya des de diverses fonts històriques i identificar les principals empremtes de selecció positiva (adaptativa). Tot i que no vam trobar evidències clares d'un increment de la mescla genètica coincidint amb la diàspora francesa associada a les Guerres de Religió del segle XVI, vam detectar senyals de mescla nord-africana corresponents al període musulmà i a la posterior Reconquesta cristiana. Cal remarcar que, a més de trobar que hi ha gens candidats de selecció positiva ben coneguts i descrits prèviament en poblacions eurasiàtiques que també semblen tenir un paper adaptatiu a Espanya, vam identificar nous candidats possiblement relacionats amb la immunitat i la dieta (*UBL7*, *SMYD1*, *VAC14* i *FDFT1*). Finalment, l'anàlisi de la desviació local de l'ancestria va revelar que la regió genòmica MHCIII va experimentar selecció post-mescla després de la mescla postneolítica amb l'ancestria provinent de les estepes euroasiàtiques.

Paraules clau: selecció postmescla, població espanyola, escandall de selecció, selecció positiva, adaptació humana, demografia.

Introducció

Durant la migració "Out of Africa"¹, una petita part dels humans moderns es va dispersar des del seu hàbitat nadiu a Àfrica i va començar a ocupar una gran diversitat de nous entorns². Aquest important esdeveniment migratori, juntament amb les posteriors expansions locals arreu del món, explica la reduïda diversitat genètica de les poblacions no africanes en comparació amb les africanes. Malgrat això, la diversitat d'ecosistemes colonitzats pels humans moderns també va generar variació fenotípica i genètica entre poblacions, oferint un model adequat per a estudiar l'adaptació local³⁻⁸. En els darrers anys, s'han desenvolupat nombrosos tests estadístics per a identificar regions del genoma que han experimentat selecció positiva (adaptativa)⁹⁻¹⁵. Basant-se en el marc teòric de la genètica de poblacions^{16,17}, aquestes exploracions de selecció a escala genòmica poden utilitzar-se per identificar patrons de variació genètica que es desvien de les expectatives neutres tot essent compatibles amb la selecció a diferents escales temporals¹⁵.

L'anàlisi de les empremtes genòmiques de selecció entre poblacions humanes en diversos biomes i sota diferents pressions selectives ha permès aclarir l'origen i la base genètica d'alguns trets adaptatius, com la persistència de la lactasa¹⁸, la pigmentació clara de la pell en resposta a una radiació UV menor¹⁹ i l'adaptació a grans altituds²⁰, entre d'altres. A més, les respostes adaptatives geogràficament restringides als patògens locals i altres fonts ambientals d'estrès podrien explicar les diferències poblacionals en alguns fenotips relacionats amb la immunitat i altres trets vinculats a malalties com la hipertensió, l'obesitat i la diabetis, entre moltes altres^{8,21}.

La Península Ibèrica, situada a l'extrem sud-occidental d'Europa, està envoltada pel mar Mediterrani i l'oceà Atlàntic, i es troba a uns 13 km de la costa nord-africana en el punt més estret de l'estret de Gibraltar. Reconstruir la demografia passada i l'estructura genètica de la població espanyola a la Península Ibèrica ha estat tot un repte. Tot i això, els recents avanços en genètica de poblacions, combinats amb la disponibilitat de dades de

seqüenciació obtingudes en poblacions actuals²²⁻²⁵ i de conjunts de dades d'ADN antic²⁶⁻²⁷, han millorat significativament la nostra comprensió de les dinàmiques poblacionals i la diversitat genètica al llarg del temps. Com altres regions europees, la Península Ibèrica ha experimentat influència genètica de diverses poblacions humanes provinents del Llevant i el Caucas²⁸. Tanmateix, difereix en aspectes clau d'altres parts d'Europa. Durant el període glacial, la Península va servir de refugi per als caçadors-recol·lectors occidentals (WHG o *western hunter-gatherers* en anglès), que més tard també van contribuir a la diversitat genètica d'altres poblacions de caçadors-recol·lectors a Europa. L'ancestria WHG també és detectable en agricultors neolítics ibèrics, apuntant a esdeveniments de mescla entre agricultors anatòlics i caçadors-recol·lectors locals²⁷. La proximitat amb Àfrica també ha contribuït a les diferències genètiques entre la Península Ibèrica i la resta d'Europa continental²⁹. Estudis previs analitzant dades d'ADN antic han demostrat que van existir persones a la Península portadores d'una alta proporció d'ancestria nord-africana ja fa més de 4.000 anys (I4246 de Camino de las Yeseras)²⁶. Tanmateix, aquestes primeres contribucions nord-africanes van tenir un impacte limitat en el conjunt genètic ibèric. En canvi, el registre d'ADN antic mostra que, durant el domini musulmà de la Península Ibèrica (segles VIII-XV), la proporció d'ancestria nord-africana en individus ibèrics era superior a l'actualitat³⁰. Actualment, els nivells variables de mescla nord-africana probablement són el principal factor que explica la diferenciació genètica oest-est observada dins la Península Ibèrica³¹. Cal destacar que la disponibilitat d'un nou conjunt de dades genòmiques de diverses regions de França³² permetrà explorar l'impacte genètic d'una migració important, tot i que poc coneguda: la diàspora francesa provocada per les Guerres de Religió a finals del segle XVI, que hauria estat especialment notable a l'antiga Corona d'Aragó³³⁻³⁶, on va arribar a representar fins a una quarta part de la població local³⁶. Finalment, s'ha descrit que l'estructura geogràfica de la diversitat genètica a Espanya té implicacions per a trets de rellevància clínica³⁷.

Gràcies a la seva inclusió al conjunt de dades dels 1000 Genomes^{38,39}, la població espanyola ha estat àmpliament analitzada mitjançant exploracions genòmiques a gran escala per a selecció positiva. Tanmateix, fins ara no s'ha publicat cap senyal específic d'Espanya [farem servir en aquest article el terme *Espanya* degut al fet que no hem analitzat mostres portugueses, tot i que, en la seva configuració actual, és un terme històricament recent i, com a part de la Península Ibèrica, no representa una unitat geogràfica i ambiental]. Aquí, per explorar les empremtes de selecció positiva recent a la població espanyola, aprofitem un conjunt de dades excepcional que comprèn 785 genomes complets de residents a Catalunya, seqüenciats amb una cobertura de 30X com a part de la cohort GCAT|Genomes for Life⁴⁰⁻⁴². Donada la història recent de Catalunya, amb fluxos migratoris significatius, especialment des d'altres parts d'Espanya com Andalusia, podem assumir que la base de dades GCAT podria servir com a una referència adequada per a la població espanyola. Aquest enfocament difereix de treballs recents centrats en l'estructura microgeogràfica de mostres autòctones dels Pirineus catalans⁴³. A més, la grandària de la mostra i l'alta cobertura de seqüenciació del conjunt de dades GCAT haurien de proporcionar una millor potència estadística per a detectar la selecció positiva, incloent-hi senyals específics de la Península Ibèrica resultants d'escombrats selectius recents¹⁵.

En aquest context, primer vam investigar l'estructura genètica i la potencial influència del flux genètic de diverses fonts de població externes, tant antigues com actuals. A continuació, vam aplicar els mètodes SDS (*singleton density score*)¹⁵ i XP-EHH (*cross-population extended haplotype homozygosity*)¹⁰ per a detectar senyals de selecció adaptativa al llarg d'un extens període de temps (fins a 30.000 anys enrere). Aquest enfocament ens ha permès identificar senyals de selecció per a gens candidats nous, a més dels coneguts, que podrien ser deguts a una major potència estadística o a una especificitat genètica del sud-oest d'Europa.

Resultats

Estructura genètica i demografia

Quan vam explorar el conjunt de dades GCAT ($n=704$; vegeu detalls a Materials i Mètodes) en el context de les cinc grans regions geogràfiques cobertes en el Projecte 1000 Genomes (1000GP) (és a dir, AFR, AMR, EUR, SAS i EAS) utilitzant anàlisi de components principals (PCA), totes les mostres de GCAT es van agrupar dins del grup EUR (Figura S1). Quan el PCA es va restringir a les poblacions europees del 1000GP (IBS, TSI, CEU, GBR i FIN), els individus del GCAT es van agrupar estretament amb els ibèrics (IBS) i els toscans (TSI) (Figura S2). En un context mediterrani més ampli, i restringint l'anàlisi als individus del GCAT que tenien els quatre avis nascuts a la mateixa comunitat espanyola ($n=141$), es van superposar clarament amb altres poblacions europees, incloent-hi catalans, francesos i altres europeus (Figura 1A; Taula S1). Centrant-nos més específicament en les mostres d'Espanya i França, els individus GCAT amb els quatre avis procedents de la mateixa regió geogràfica es van agrupar amb mostres catalanes, balears i valencianes⁴⁴ (Figura S3). Incorporant dades antigues dels tres principals components mesolítics i neolítics que han configurat les poblacions europees actuals (caçadors-recol·lectors occidentals – WHG o *western hunter-gatherers*, agricultors europeus primerencs – EEF o *early European farmers* i nòmades primerencs de les estepes – ENS o *early nomad Steppe*)⁴⁵, es va mostrar que l'afinitat dels individus GCAT amb els EEF és més propera en comparació amb altres poblacions europees com els habitants de les Illes Òrcades, francesos o italians, tot i que no tan propera com els sards (Figura S4).

L'anàlisi d'ancestria amb ADMIXTURE en el marc de les poblacions veïnes contemporànies va confirmar una component d'ancestria europea majoritària dins del conjunt del GCAT, acompanyada de dues components menors presents principalment en poblacions del Pròxim Orient i del nord d'Àfrica en $K=7$ (clúster amb l'error de validació creuada més baix, seguit de $K=8$ i $K=6$) (Figura 1b, Figura S5). A $K=8$, una altra component,

caracteritzant principalment els mozabites però també present en altres poblacions nord-africanes i palestines, va aparèixer en totes les mostres espanyoles, així com en les de Provença (Bouches-du-Rhône, BdR) i Sardenya. Així doncs, mentre que els residents de Catalunya inclosos en el GCAT mostren un perfil genètic europeu continental típic, també presenten una petita proporció d'ancestria nord-africana (mitjana: 0.0442, desviació estàndard: 0.0023) i del Pròxim Orient (mitjana: 0.1010, desviació estàndard: 0.0019).

A continuació, vam explorar l'agrupació d'individus basada en haplotips dins el mateix conjunt regional del GCAT (n=141) i poblacions mediterrànies veïnes utilitzant fineSTRUCTURE (Figura S6A). La població espanyola es divideix en dues branques principals: Est i Oest (Figura S6B). La branca Est es divideix en dos clústers: un amb mostres de Catalunya i València (CAT, com es mostra a la Figura 1C) i un altre amb mostres d'Aragó (ARA). La branca Oest es subdivideix en Eivissa (no mostrada a la Figura 1C), un clúster Centre-Oest que inclou mostres d'Andalusia, Castella i Extremadura (WEST), i un tercer clúster amb algunes mostres de València (MED). La distribució geogràfica d'aquests clústers basats en haplotips mostra un gradient d'est a oest de diferenciació genètica, consistent amb anàlisis prèvies i el procés històric de la Reconquesta espanyola^{30,31}. A més, es van obtenir quatre branques principals per a França (Figura S6B): bascos francesos, mostres del nord procedents de París i Alsàcia, mostres del sud de Provença, Dordonya i els Pirineus, i mostres bretones de Rennes, aquestes últimes mostrant-se genèticament diferenciades d'altres poblacions franceses^{32,46}. Les mostres nord-africanes es van separar en un grup occidental i un d'oriental (Figura S6A), aquest últim amb una component més gran del Pròxim Orient, com era de preveure²⁹.

Posteriorment, es van inferir formalment esdeveniments de mescla per als clústers genètics de l'Est i l'Oest de la població espanyola utilitzant fastGLOBETROTTER⁴⁷. Ambdós clústers presenten un únic esdeveniment de mescla entre una font del sud-oest d'Europa (font 1, formada principalment per mostres de Provença, Dordonya, els Pirineus i

Bretanya) i una font menor de tipus africà (font 2). Les proporcions són del 98% i el 2% per al clúster de l'oest d'Espanya, i del 96% i el 4% per al clúster de l'est d'Espanya, respectivament (Figura S7A). Tot i que fastGLOBETROTTER va suggerir inicialment dos esdeveniments de mescla en el clúster de l'oest d'Espanya, aquest patró no es va mostrar després del *bootstrapping*. Curiosament, per al clúster de l'oest d'Espanya, el clúster del Nord-oest d'Àfrica es presenta com a població aproximada (o *surrogate* en anglès) tant a les fonts 1 com 2, possiblement reflectint un antic flux genètic entre la Península Ibèrica i el Magrib. A més, els bascos francesos són un *surrogate* a les fonts 1 i 2 per a ambdós clústers, apuntant a un esdeveniment de mescla basca, coherent amb publicacions anteriors³⁰. Les dates inferides per a aquests esdeveniments de mescla difereixen lleugerament entre clústers, sent aproximadament l'any 1153 (IC 95%: 1083-1242) per a l'est d'Espanya i aproximadament l'any 1211 (IC 95%: 1119-1287) per a l'oest d'Espanya, assumint 29 anys per generació (Figura S7B).

Empremtes de selecció

Els senyals de selecció es van explorar inicialment utilitzant SDS¹⁵ en el conjunt complet del GCAT (n=704) per investigar esdeveniments molt recents de selecció positiva. A més, vam emprar XP-EHH¹⁰ per a detectar haplotips d'alta freqüència que han estat recentment afavorits fins a freqüències relativament altes en la cohort GCAT, utilitzant la població YRI com a referència (Figures 2 i S8). Després d'anotar els gens corresponents i les variants SNP per a cada pic de selecció (per a més detalls, vegeu Materials i Mètodes), vam investigar amb detall les possibles variants candidates per a la selecció utilitzant iSAFE¹² i CLUES¹⁴ (Taules S2-S9). Com era d'esperar, la majoria de les variants candidates confirmades per CLUES i identificades inicialment amb SDS mostren freqüències al·lèliques més baixes i corresponen a escombrats selectius més recents en comparació amb els identificats per XP-EHH (Figures S9-S25). Anàlogament, mentre que 12 dels 40 pics de selecció SDS descrits al conjunt de dades GCAT inclouen regions candidates

conegudes per a selecció positiva prèviament identificades en europeus (Taula S3), la majoria de senyals detectats amb XP-EHH al GCAT corresponen a escombrats selectius ja descrits tant en europeus com en asiàtics en comparació amb la població YRI (Taula S7). Tal com discutirem a continuació, un nombre significatiu de regions candidates per a selecció positiva identificades al conjunt de dades GCAT estan relacionades amb la pigmentació de la pell més clara, la dieta i la resposta immune (Taula 1).

Selecció en gens de pigmentació

Tres dels pics principals de SDS identificats a les dades genòmiques del GCAT se superposen amb gens candidats ben coneguts que haurien facilitat el fenotip adaptatiu d'una pigmentació de pell més clara en europeus: *SLC45A2*⁴⁸, *OCA2-HERC2*¹¹ i la regió *GRM5-TYR*¹¹ (Taula 1, Figura 2). A *SLC45A2* es va trobar una forta evidència de selecció sobre la variant rs183671, associada a diversos trets de pigmentació de pell, cabell o ulls segons el Catàleg NHGRI-EBI de GWAS (<https://www.ebi.ac.uk/gwas/>; Taula S2).

Utilitzant CLUES per a estimar la trajectòria històrica de la freqüència de l'al·lel derivat seleccionat, es va detectar un escombrat selectiu en les últimes 500 generacions (Figura S9). A la regió genòmica *OCA2*, vam trobar una forta evidència de selecció sobre la variant intrònica rs7183877 de *HERC2*, també associada amb la pigmentació de la pell, el cabell i els ulls (Taula S2, Figura S10). Un altre pic de SDS potencialment relacionat amb una pigmentació de la pell més clara inclou una variant intrònica possiblement seleccionada al gen *GRM5* (rs7119749, Figura S11). Aquesta variant està situada a la regió 5' del gen *TYR*, que codifica l'enzim implicat en el primer pas de la síntesi de melanina.

En l'anàlisi XP-EHH que compara YRI i GCAT, es van detectar fins a tres regions candidates addicionals que podrien estar relacionades amb la pigmentació de la pell més clara i l'adaptació a una radiació UV més baixa, tal com s'ha descrit en estudis anteriors⁴⁹: *SLC12A1-DUT* (situada prop del gen *SLC24A5*), *MLPH-RAB17* i la regió *KITLG* (Taula 1,

Figures S12-S14). El gen *SLC24A5* codifica un intercanviador sodi-calci associat amb la pigmentació en peixos zebra i humans, possiblement facilitant el transport iònic als melanosomes^{50,51}, mentre que el gen *MLPH* (melanofilina) juga un paper clau en el transport de melanosomes i en la susceptibilitat al càncer de pròstata^{52,53}. *KITLG* codifica el lligand del receptor tirosina-cinasa, que s'ha demostrat que influeix en la pigmentació regulant la proliferació de melanòcits i la distribució de melanina⁵⁴.

Selecció en gens relacionats amb la dieta

Dos pics de selecció obtinguts amb SDS inclouen regions candidates ben conegudes per a facilitar adaptacions relacionades amb la dieta: *LCT-MCM6* i la regió que conté els gens *SLC22A4* i *SLC22A5* (Figura 2, Taula 1). Tal com s'esperava per a una població europea, vam trobar proves sòlides que recolzen la selecció positiva per a l'al·lel derivat de la variant intrònica rs4988235 de *MCM6*, associada amb la persistència de la lactasa en europeus. Utilitzant CLUES en les genealogies genòmiques inferides a partir del conjunt de dades GCAT, estimem que aquest esdeveniment de selecció va començar en les darreres 200 generacions (Figura S15). Al gen transportador d'ergotioneïna *SLC22A4* hi trobem una forta evidència de selecció positiva actuant sobre la variant de canvi de sentit rs1050152 (que codifica la substitució L503F) en les darreres 300 generacions (Figura S16). Es creu que aquest gen presenta un senyal de selecció degut a l'adaptació als baixos nivells dietètics d'ergotioneïna entre els primers agricultors del Neolític al Creixent Fèrtil^{55,56}. Curiosament, la variant seleccionada està associada amb una expressió reduïda del gen *SLC22A5* segons GTEx (Taula S2), la qual cosa resulta en nivells més baixos del transportador de carnitina *OCTN2*, que és important per al transport i l'oxidació d'àcids grassos als mitocondris⁵⁷.

Un altre senyal de selecció identificat amb SDS possiblement relacionat amb la detoxificació adaptativa es va observar al gen *ABCC1* al cromosoma 16, on es va detectar

selecció positiva actuant sobre la variant intrònica rs212086, encara que amb menys suport estadístic (Taula 1, Figura S17). És destacable que *ABCC1* codifiqui una proteïna de resistència multifàrmac ben coneguda (MRP1), que té un paper en la detoxificació biliar de diversos fàrmacs anticancerígens⁵⁸ i que prèviament s'ha detectat sota selecció positiva a la població CEU⁵⁹. A més, vam trobar proves moderades de selecció positiva actuant sobre la variant rs2404955 situada a la regió 3' del gen *CYP3A4* (Figura S18), que també està implicat en detoxificació, així com en la metabolització de nombrosos fàrmacs⁶⁰.

Selecció en gens de resposta immune

Una alta proporció de les regions candidates a selecció positiva identificades al conjunt de dades GCAT contenen gens presumptament relacionats amb la resposta immune (Figura 2, Taula 1). Per exemple, una gran regió candidata detectada amb SDS, que abasta ~1,474 Mb al cromosoma 6, inclou la regió MHC III, que conté diversos gens relacionats amb la immunitat, com *HCG20*, *HCG21*, *AIF1*, *GPANK1*, *ABHD16A*, *LY6G6F*, *C2* i *PBX2*. Utilitzant CLUES per analitzar diverses variants funcionals en aquesta regió, vam detectar evidències moderades de selecció actuant durant les darreres 150 generacions sobre rs204991, un SNP situat a la regió 5' de *PBX2* (Figura S19). Segons GTEx, aquest SNP influeix fortament en l'expressió dels components del complement 4A i 4B en diversos teixits. A més, utilitzant l'estadístic XP-EHH, vam replicar un senyal de selecció prèviament conegut al gen *TMEM232* en eurasiàtics⁶¹, que s'ha demostrat que promou la resposta inflamatòria en la dermatitis atòpica⁶². Entre diverses variants candidates en la regió *TMEM232*, es van trobar les proves més sòlides de selecció per al SNP exònic no codificant rs10038763 (Figura S20).

Nous candidats a adaptació

L'anàlisi dels principals senyals SDS al conjunt del GCAT (Figura 2) també va revelar diverses regions candidates no documentades anteriorment i noves variants funcionals

presumptament seleccionades en aquestes (Taula 1). Per exemple, vam trobar proves moderades de selecció durant els darrers 6.000 anys per al SNP intrònic rs35662596 de *SMYD1* (Figura S21), que arriba a la seva freqüència global més alta a la població IBS (14%). És destacable que mutacions en *SMYD1* poden provocar l'absència a la superfície dels glòbuls vermells de l'antigen AnWj⁶³, un receptor per a *Haemophilus influenzae*⁶⁴. La quantitat d' AnWj es correlaciona amb la capacitat de *H. influenzae* d'adherir-se a cèl·lules epitelials⁶⁵. També es van identificar evidències moderades de selecció actuant sobre el SNP intrònic rs1296025 de *FDFT1* (Figura S22), que té les freqüències al·lèliques més altes en les poblacions del 1000 GP CEU (17%), GBR (17%) i IBS (21%). És destacable que *FDFT1* codifica el primer enzim específic en la via de biosíntesi del colesterol⁶⁶ i és una diana molecular de la resposta fisiològica al dejuni⁶⁷, mentre que rs1296025 s'ha associat amb els nivells de colesterol no HDL⁶⁸. Un altre nou candidat per a selecció positiva es troba dins la regió *UBL7*, que conté la variant reguladora rs750607 (Figura S23). Segons Ensembl⁶⁹, aquest SNP està associat amb l'expressió diferencial de diversos gens de la regió, incloent Semaphorin 7A (*SEMA7A*) en neutròfils, que està implicat en la resposta immune, la inflamació⁷⁰ i la regulació de cèl·lules NK⁷¹, cèl·lules T⁷² i mastòcits⁷³.

A la regió *PRAG1* es van detectar evidències més dèbils de selecció per a la variant rs55852693, que és probable que alteri un lloc d'unió de factors de transcripció (Figura S24) i que s'ha associat amb la preferència per al menjar picant⁷⁴. Tot i que les freqüències al·lèliques de rs55852693 són similars a totes les poblacions europees dins del 1000 GP, cosa que suggereix que podria no correspondre a una adaptació específica de la població espanyola, *PRAG1* també s'ha identificat en una exploració de selecció adaptativa d'una població brasilera mestissa en una regió genòmica enriquida en ancestria nativa americana^{75,76}. Finalment, dins del pic SDS *MTSS2-VAC14*, es van trobar evidències febles de selecció positiva actuant sobre el SNP intrònic rs11075777 de *MTSS2* (Figura S25). L'al·lel seleccionat és un eQTL per al gen *VAC14*, que juga un paper en la susceptibilitat a la bacterièmia^{77,78}. No obstant això, atès que la seva freqüència és aproximadament del 50%

a totes les poblacions europees del 1000 GP, també podria representar una empremta de selecció més àmplia a Europa.

Desviacions locals d'ancestria

A continuació, vam explorar regions genòmiques que presenten desviacions locals d'ancestria (LAD o *local ancestral deviation*) associades a ancestries específiques antigues o modernes que se solapen amb els pics de selecció identificats amb SDS o XP-EHH, ja que aquestes regions podrien representar possibles casos de selecció post-mescla. En el conjunt regional del GCAT (n=141), les LAD modernes es van avaluar utilitzant mostres de referència del nord d'Àfrica per a l'ancestria nord-africana, Palestina per a l'ancestria del Pròxim Orient, i el sud de França com a població europea més propera per representar el rerefons autòcton del GCAT (per a més detalls, vegeu Materials i Mètodes). Es van identificar tres regions LAD coincidents ($SD > 4,42$) al cromosoma 6, indicant un excés d'ancestria nord-africana i del Pròxim Orient, que se solapen amb tres pics SDS de selecció positiva (Taula S10). Dos d'aquests pics SDS mostraven LAD per a ambdues ancestries, nord-africana i del Pròxim Orient, i se solapaven amb diversos gens a la regió MHC III, incloent-hi el SNP rs204991, que es troba a la regió 5' de *PBX2* i regula l'expressió de *C4A* i *C4B* (Figura 3).

Per investigar les regions LAD resultants d'extensos esdeveniments de barreja durant els períodes del Mesolític i Neolític a Europa, vam utilitzar el Mesoneo Dataset⁴⁵, emprant mostres EEF, WHG i ENS com a referents per estimar les corresponents proporcions d'ancestria antiga de cada individu del GCAT (n=704). Aquesta anàlisi va revelar un patró general similar a l'observat en individus ibèrics de l'Edat del Bronze (Figura S26). Amb un llindar estricte de $4,42 SD^{79}$, no es van trobar regions LAD per a aquestes ascendències antigues, però considerant un llindar més permissiu de 3 SD, es van detectar fins a cinc regions amb un excés d'ancestria WHG, dues amb excés d'ancestria ENS i una amb excés

d'ancestria EEF (Taula S10 i Figures S27-S28). Entre aquestes regions LAD, només tres es van solapar amb les 40 empremtes de selecció més fortes detectades amb SDS al conjunt de dades GCAT. Cal destacar que aquestes incloïen el pic *LCT-MCM6* (Figura S29), així com el tercer pic de selecció SDS al cromosoma 6 que comprèn la regió MHC III (Figura 3), ambdós mostrant desviacions d'ancestria ENS. La tercera LAD antiga que se solapava amb un pic SDS contenia un grup de gens *zinc finger* al cromosoma 19 i mostrava un excés d'ancestria WHG (Figura S29). Curiosament, tot i que no presentava empremtes de selecció positiva recents al conjunt de dades GCAT, l'única LAD associada amb ancestria EEF incloïa el gen *FADS2*. Aquest gen ha estat prèviament identificat com una diana de selecció forta en diverses ascendències, incloent-hi caçadors-recol·lectors orientals i occidentals, així com poblacions d'agricultors anatòlics⁶.

Discussió

Les nostres anàlisis demostren que la cohort del GCAT no només captura les característiques genètiques d'una població europea típica, sinó que també serveix com a *proxy* robust per a la població espanyola general dins de la Península Ibèrica. Com que les mostres del GCAT cobrien diverses regions d'Espanya, vam poder detectar la ben documentada contribució genètica del Nord d'Àfrica, que es presenta de manera diferent entre l'est i l'oest de la Península Ibèrica. No obstant això, no hem pogut identificar la possible contribució de la diàspora francesa vinculada a les Guerres de Religió del segle XVI al *pool* genètic actual d'Espanya. Atès que la distància genètica entre el nord-est de la Península Ibèrica i el sud de França és curta (la qual cosa reflecteix les similituds lingüístiques i de cognoms que van facilitar l'assimilació d'aquests nousvinguts) i la mida petita de la mostra d'individus que es podrien assignar a una regió determinada segons el lloc de naixement dels seus quatre avis, el nostre disseny pot no haver tingut prou potència per capturar aquesta contribució genètica.

Aprofitant la millor potència estadística del conjunt de dades GCAT degut a la seva grandària mostral total, vam poder identificar nous gens candidats sota selecció positiva i replicar diversos casos ben coneguts de selecció adaptativa relacionats amb la pigmentació, la dieta i la resposta immunitària, prèviament descrits mitjançant diferents mètodes estadístics. Com era d'esperar, les regions candidates que vam replicar mitjançant l'estadístic XP-EHH normalment corresponen a empremtes de selecció compartides entre diverses poblacions no africanes, probablement reflectint pressions ambientals comunes després de la migració *Out-Of-Africa*. En canvi, les empremtes identificades amb l'estadístic SDS són predominantment compartides amb altres poblacions europees o, en alguns casos, són específiques de la població espanyola, indicant esdeveniments selectius més recents i localitzats. Per tant, l'ús combinat de SDS i XP-EHH en el GCAT ens ha permès identificar un conjunt molt complet de petjades genòmiques de la selecció natural en la població espanyola contemporània, que foren modelades per les diverses pressions evolutives experimentades pels seus avantpassats a través de diferents períodes temporals.

Tot i que moltes de les empremtes de selecció identificades i variants candidates subjacents són ben conegudes i es troben compartides amb altres poblacions europees, es van observar diferències clares en el temps i la intensitat de la selecció en el conjunt de dades GCAT. Aquestes diferències poden provenir d'històries demogràfiques, influències externes i condicions ambientals lleugerament diferents entre les poblacions europees. Per exemple, el coeficient de selecció estimat aquí per a rs4988235 a *LCT* ($s=0.00862$) és lleugerament inferior però semblant a les estimacions prèvies ($s = 0.0194$, $s = [0.01019, 0.01056]$)^{6,80}. A més, l'al·lel de persistència de lactasa sembla haver emergit abans al nord d'Europa que a la Península Ibèrica²⁶, i la seva freqüència és més baixa a la població espanyola. En canvi, la freqüència actual de l'al·lel seleccionat a l'SNP rs1050152 a *SLC22A4* sembla ser més alta a la població espanyola (45%) en comparació amb altres poblacions europees en el 1000GP (36-41%). Aquesta diferència pot atribuir-se al

component més alt d'ancestria EEf inferit en el conjunt de dades GCAT. Curiosament, mentre que Irving-Pease et al. (2024)⁶ van inferir que la selecció sobre la variant rs1050152 va acabar fa uns ~1500 anys, la nostra anàlisi indica un augment continu de la freqüència al·lèlica fins a temps molt recents, consistent amb els resultats de Mathieson et al. (2015)⁸¹. Estudis similars de selecció positiva a Itàlia han revelat empremtes adaptatives diferencials entre les poblacions del nord i del sud d'Itàlia, tot i que el mecanisme més plausible implicat per aquestes diferències sigui probablement el nivell variable de deriva genètica⁸². Malauradament, la cohort del GCAT no inclou prou individus amb els quatre avis nascuts al mateix lloc geogràfic com per realitzar una exploració robusta latitudinal i ambiental de la selecció positiva a Espanya.

Com altres regions de la Europa continental, la Península Ibèrica ha estat influïda per les transicions culturals i els canvis demogràfics de l'Holocè. El canvi de societats caçadores-recol·lectores a sistemes agrícoles neolítics, seguit per l'arribada de grups de pastors nòmades, no només va transformar el paisatge genètic de la població ibèrica sinó que probablement va introduir noves pressions selectives, deixant empremtes genètiques diferencials. A més, com que aquestes poblacions entrants probablement estaven ben adaptades als seus respectius estils de vida i pràctiques culturals, també podrien haver contribuït amb variants adaptatives al *pool* genètic espanyol mitjançant la mescla. La nostra anàlisi del conjunt de dades GCAT mostra tres regions amb LAD significatives cap als components WHG i ENS, que se superposen amb senyals recents de selecció. Curiosament, la regió LAD del cromosoma 19 (Figura S29), enriquida pel component d'ancestria WHG, compren un conjunt de gens de *zinc fingers*. No obstant això, la funció exacta i els possibles fenotips adaptatius associats a aquests gens de *zinc fingers* segueixen sent desconeguts i requereixen més recerca per a millorar la nostra comprensió de les adaptacions genètiques pre-neolítiques en les poblacions WHG.

A més, l'anàlisi de LAD va mostrar que el pic de selecció SDS a la regió MHC III està significativament enriquit en l'ancestria ENS. Quan només es van considerar les poblacions contemporànies, aquesta regió va mostrar un enriquiment paral·lel de components d'ancestria del Nord d'Àfrica i del Pròxim Orient. És important destacar que les dues poblacions utilitzades com a *proxies* per a aquests components externs també van mostrar un LAD significatiu per al component d'ancestria ENS, suggerint una pressió selectiva compartida potencialment relacionada amb la domesticació i les transmissions zoonòtiques⁸³⁻⁸⁵. Cal assenyalar que la gran diversitat genètica de la regió HLA podria confondre potencialment els mètodes d'inferència d'ancestria local, com s'ha observat en la recerca sobre selecció post-mescla a les poblacions llatinoamericanes, que van mostrar un excés d'ancestria africana als senyals de HLA⁸⁶. No obstant això, no s'espera que aquest biaix afecti la comparació amb ENS al nostre estudi. Finalment, la regió *LCT-MCM6* també va mostrar una desviació significativa per a l'ancestria ENS en el conjunt de dades GCAT. Aquest patró coincideix amb treballs previs en els que se suggereix que les empremtes de selecció positiva a *LCT* es poden traçar fins a l'ancestria de les estepes.

Tot i que el nostre estudi proporciona informació valuosa, és important reconèixer algunes limitacions inherents i possibles biaixos. En primer lloc, la cohort del GCAT presenta un desequilibri geogràfic, amb una sobrerrepresentació d'individus del sud i l'est d'Espanya en comparació amb el nord i l'oest. Aquesta asimetria limita la nostra capacitat d'analitzar de manera exhaustiva els diferents períodes de mescla del Nord d'Àfrica a través de la Península Ibèrica, ja que el moment de la transició de domini musulmà a cristià va variar significativament entre regions. Per tant, és possible que no haguem capturat del tot les diferències regionals en els patrons de mescla derivades de diferents interaccions històriques amb les poblacions del Nord d'Àfrica. En segon lloc, la caracterització de les principals regions candidates per a la selecció (o pics de selecció més forts) depèn de dos paràmetres definits per l'usuari, els quals podrien introduir biaixos. En aquest cas, serien la longitud de la finestra utilitzada per detectar al·lels amb valor

extrems donat cada estadístic de selecció i el nombre d'al·lels amb valors estadístics significatius a nivell genòmic necessaris per a classificar una regió genòmica com a part d'un pic. Aquests límits arbitraris poden influir tant en la identificació com en la interpretació dels senyals de selecció detectats en els escaneigs genòmics realitzats mitjançant SDS i XP-EHH. Per resoldre-ho, vam intentar validar totes les regions candidates per a la selecció utilitzant iSAFE, permetent la identificació dels al·lels afavorits. A més, vam aplicar CLUES per a visualitzar les trajectòries temporals de la freqüència de cada al·lel presumiblement seleccionat i estimar el coeficient de selecció i la probabilitat de selecció positiva corresponents en cada cas. Finalment, la mida petita de la mostra de les poblacions utilitzades com a *proxies* del Nord d'Àfrica pot haver limitat el poder de les nostres anàlisis de LAD i les estimacions del temps de mescla. En un futur, es podria aconseguir una comprensió més completa del procés de mescla del Nord d'Àfrica incorporant dades de genoma complet de poblacions del Nord d'Àfrica, particularment dels grups amazics, que van jugar un paper important en la Conquesta Islàmica de la Península Ibèrica, tal com es documenta en els registres històrics com el Muqaddimah de Ibn Khaldūn⁸⁷.

En conclusió, els nostres resultats demostren que diversos gens candidats prèviament identificats com a adaptatius en altres parts d'Europa van estar subjectes a selecció positiva en les poblacions ancestrals dels actuals espanyols. A més, hem identificat nous gens candidats per a la selecció positiva, els quals podrien ser deguts a la mida més gran de la mostra utilitzada en el nostre estudi o a la seva especificitat per al sud-oest d'Europa. No obstant això, aquests gens han de ser considerats com a candidats provisionals fins que la funcionalitat de la seva variació genètica i la seva rellevància evolutiva siguin completament caracteritzades i enteses.

Materials i mètodes

La cohort del GCAT i el processament de les dades genòmiques

Els fitxers VCF per a seqüències de genoma complet (WGS) d'Illumina 30X de 785 individus actuals de la cohort del GCAT⁴² es van obtenir de l'European Genome-phenome Archive (EGA) sota el número d'accés EGAD00001007774. La cohort GCAT es va reclutar (del 2014 al 2018) entre residents a Catalunya d'entre 40 i 65 anys amb accés al sistema sanitari públic nacional. Està formada per 19.140 participants. Les característiques de la cohort^{40,41,88} i del conjunt de dades seqüenciades es descriuen en altres llocs⁴². Es disposava de seqüències genòmiques completes per a 785 voluntaris; d'aquests, segons la informació associada, 141 tenien els quatre avis nascuts a la mateixa regió espanyola, fos Catalunya o una altra (vegeu detalls a les Taules S1 i S11). El control de qualitat i el filtratge d'individus d'orígens mesclats ja s'havia realitzat prèviament⁴². Per a centrar-nos en els SNPs bial·lèlics, es va utilitzar BCFtools per excloure variants estructurals i indels. A més, es van eliminar 81 individus del GCAT que van indicar una ètnia no caucàsica (tot i que el significat de “caucàsic” en el context català no és gens clar; Taula S12). La mida final de la mostra va ser de 704 individus (Conjunt de dades A).

Els fitxers VCF es van traslladar a hg38 utilitzant Picard tools⁸⁹ i es van fusionar amb 1000 GP fase 3³⁹ utilitzant la comanda isec de BCFtools⁹⁰ (Conjunt de dades B). A continuació, es van eliminar variants rares, SNPs amb fortes desviacions de Hardy-Weinberg i es va realitzar una poda per desequilibri de lligament (LD) utilitzant PLINK2⁹¹ en finestres lliscants de 200 kb, amb un pas de 25 SNPs i un llindar de coeficient de correlació al quadrat (r^2) de 0.5 (--maf 0.05 --max-maf 0.95 --hwe 1e-50 --indep-pairwise 200 25 0.5). En aquest punt, el conjunt de dades constava de 679.677 variants i 3.208 individus. Les anàlisis de components principals (PCA) es van realitzar utilitzant l'eina smartPCA del paquet EIGENSOFT⁹² i la correcció EIGENSTRAT⁹³ sense eliminar valors atípics (Figures S1-2).

Anàlisis amb ADMIXTURE i fineSTRUCTURE

El conjunt de dades GCAT es va filtrar per retenir individus amb els quatre avis nascuts a la mateixa comunitat autònoma i es va fusionar amb conjunts de dades disponibles que contenien poblacions de referència adequades per a detectar contribucions externes a la Península Ibèrica. D'aquesta manera, les 141 mostres amb els quatre avis de la mateixa comunitat autònoma es van fusionar inicialment amb el panell HGDP⁹⁴ utilitzant el mateix procediment que per al conjunt de dades B. També es van incloure dades addicionals de genotipat d'arrays de SNPs de França³², Catalunya⁴⁴ i el nord d'Àfrica⁹⁵. No es van trobar mostres emparentades fins al tercer grau.

Es van eliminar variants rares, SNPs amb fortes desviacions de Hardy-Weinberg i es va realitzar la poda de LD utilitzant PLINK2 en finestres lliscants de 200 kb, amb un pas de 25 SNPs i un llindar r^2 de 0.5 (--maf 0.05 --max-maf 0.95 --hwe 1e-50 --indep-pairwise 200 25 0.5), retenint 215.178 variants. Es va realitzar PCA utilitzant l'eina smartPCA del paquet EIGENSOFT amb la correcció EIGENSTRAT però sense eliminar valors atípics. Es van excloure mostres de Sud-amèrica, Oceania i l'Est d'Àsia per la seva manca de rellevància pel que fa als components d'ancestria a GCAT, com ja s'havia publicat en estudis anteriors^{24,30}. D'aquesta manera, el conjunt de dades final (Conjunt de dades C) constava de 1.181 individus (vegeu la Taula S1 per a més detalls). Les ancestries individuals en aquest conjunt de dades podades es van explorar amb ADMIXTURE 1.3²² en 10 execucions utilitzant el mode no supervisat i provant des de K=1 fins a K=12. Es va utilitzar PONG⁹⁶ per a representar els resultats d'ADMIXTURE.

El conjunt de dades no podades, que contenia 426.650 variants en 1.181 individus, es va fasejar utilitzant SHAPEIT 4.1.3⁹⁷ i el panell de referència d'haplotips 1000 GP³⁹.

Posteriorment, es va executar CHROMOPAINTER⁹⁸ en mode *tots contra tots* per als cromosomes 1, 4, 17 i 20 per estimar el paràmetre de taxa de canvi (N_e) i la taxa global de mutació (M) utilitzant 10 iteracions de l'algoritme d'expectativa-maximització (EM) de

CHROMOPAINTER. Utilitzant aquests valors, vam tornar a executar CHROMOPAINTER en mode *tots contra tots*, especificant que tots els individus copiessin d'altres individus per a tots els cromosomes. Es va aplicar el mètode de cadena de Markov Monte Carlo (MCMC) de fineSTRUCTURE⁹⁸ per assignar a cada individu un grup genètic utilitzant 1.000.000 d'iteracions de "burn-in" (paràmetre -x), i 2.000.000 d'iteracions de mostratge (paràmetre -y) de les quals només es van retenir cada deumil·lèsima iteració (paràmetre -z). A més, es va tornar a executar fineSTRUCTURE utilitzant el fitxer de forçat (-F) per a fixar els grups fora d'Espanya, França i Itàlia com a grups continentals.

Els dos principals grups genètics inferits a la Península Ibèrica (Oest i Est) es van analitzar més a fons utilitzant fastGLOBETROTTER⁴⁷ en mode *donant contra receptor*, exclouent els grups genètics espanyols (Oest, Est i Eivissa) com a donants i considerant tots els grups com a receptors (excepte els grups ibèrics no objectiu). A continuació, es va executar CHROMOPAINTER en mode *donant contra receptor*, utilitzant només el grup genètic espanyol objectiu com a receptor. Després, es va utilitzar fastGLOBETROTTER amb l'opció prop.ind: 1 per a inferir i datar els esdeveniments de mescla. Per a tenir en compte els patrons de desequilibri que podrien confondre els senyals de mescla, es va habilitar l'opció null.ind: 1. Es va realitzar una segona execució de fastGLOBETROTTER per realitzar una anàlisi de bootstrap i estimar un interval de confiança al voltant de la data de mescla inferida.

Components d'ancestria antiga

Per explorar l'estructura genètica del conjunt de dades GCAT en el context de mescles antigues, vam utilitzar el conjunt de dades públicament disponible d'ADN antic de Allentoft et al (2024)⁴⁵. Es van utilitzar els grups genètics inferits per Allentoft et al (2024)⁴⁵ com a referències per a les tres principals poblacions antigues que expliquen la diversitat genètica de l'Europa actual: caçadors-recol·lectors de l'oest (WHG), nòmades primerencs de les estepes (ENS) i agricultors europeus primerencs (EEF). Vam

preprocessar les dades seguint les directrius recomanades: descartar individus amb baixa cobertura i mantenir els llocs que passaven els filtres genòmics del 1000G, $MAF > 0.05$ i $INFO \geq 0.8$. Aquesta estratègia de filtratge va resultar en un conjunt de dades amb 2.997.159 SNPs. Restringint l'anàlisi només a llocs de transversió, es va obtenir un conjunt de dades amb 966.986 SNPs. Es va realitzar una PCA utilitzant el mateix procediment i *pipeline* que per al conjunt de dades C. Donada l'alta precisió de la imputació per a conjunts de dades antigues respecte a les dades de seqüenciació modernes⁴⁵, no es va aplicar la projecció PCA.

Anàlisi de selecció positiva

Per a investigar regions candidates a selecció positiva al conjunt de dades GCAT, vam emprar dos estadístics: (i) el *Singleton Density Score* (SDS) per a identificar esdeveniments de selecció molt recents; i (ii) el *Cross Population Extended Haplotype Homozygosity* (XP-EHH) per a detectar escombrats selectius on les variants afavorides han arribat recentment a altes freqüències (o fixació) a la cohort GCAT respecte a la població YRI.

Per a calcular l'SDS, es van polaritzar els SNPs en el conjunt de dades A basant-se en els seus al·lels ancestrals utilitzant *scripts* propis i el fitxer *fasta* d'Ensembl EPO (<http://May2024.archive.ensembl.org/info/genome/compara/mlss.html?mlss=2006>). Es van extreure els singletons en fitxers separats. Els SNPs de prova es van processar exclouent variants rares (`--maf 0.05 --max-maf 0.95`) utilitzant PLINK, mantenint llocs amb tres genotips, resultant en un conjunt de dades de 5.251.738 loci. Les regions centromèriques també es van excloure de l'anàlisi. Es va tractar l'observabilitat de cada variant com a igual. Es van inferir les formes gamma utilitzant un model de població europea basat en Tennesse et al (2012)⁹⁹ (implementat al repositori SDS de Github) amb una mida mostral de 1.408 cromosomes per a les freqüències al·lèliques que varien de 0,05 a 0,95 (en passos de 0,01). Els resultats bruts es van normalitzar per grups de freqüències d'al·lel derivades de 0,05 a 0,95 (passos de 0,01) i es van calcular els valors p.

Els SNPs es van classificar com a regions candidates per a selecció positiva si es trobaven en el percentil 99.99% dels valors SDS, acompanyats d'almenys 10 variants addicionals dins del percentil 99.995% en una finestra genòmica de 1 Mb.

En el mètode XP-EHH, primer es va fasejar el conjunt de dades B utilitzant SHAPEIT 4.1.3⁹⁷ i el panell de referència 1000 GP³⁹ i després es van calcular els valors XP-EHH normalitzats utilitzant selscan v1.3.0¹⁰⁰. Es va considerar que els SNPs eren en una regió candidata per a selecció positiva si queien dins del percentil 99.99% dels valors XP-EHH i estaven acompanyats per almenys 10 variants addicionals en el percentil 99.995% dins d'una finestra genòmica de 1 Mb.

Vam anotar funcionalment els SNPs en regions candidates per a selecció positiva utilitzant l'Ensembl Variant Effect Predictor (VEP¹⁰¹) i cada regió va ser explorada manualment per a possibles variants candidates (Taules S2 i S6). A més, vam executar iSAFE utilitzant finestres de 400 kb al voltant del SNP amb el senyal de selecció més alt de cada regió candidata, utilitzant el *flag* --IgnoreGaps i la població YRI del 1000 GP com a grup extern (Taules S4 i S8). Les possibles variants candidates per a la selecció van ser posteriorment validades amb CLUES. Per això, vam obtenir genealogies a nivell de genoma per a cada lloc amb Relate utilitzant els temps de coalescència inferits prèviament en un subconjunt de dades fusionades amb el 1000 GP que inclou totes les poblacions europees, xineses han i iorubes. La mida de la població es va estimar utilitzant el llindar -0.5 per a la població GCAT per obtenir coalescències específiques per a les dades de GCAT. Posteriorment, vam tornar a estimar les llargades de les branques de la genealogia utilitzant RelateCoalescentRate (-mode ReEstimateBranchLengths). Els gràfics d'anàlisi de recombinació ancestral (ARGs o *ancestral recombination graphs*) es van mostrear amb el *script* SampleBranchLengths, assumint temps de generació de 28 anys amb 100 mostres. A continuació, vam utilitzar CLUES per estimar els coeficients de selecció (Taules S5 i S9) i les respectives trajectòries de freqüència al·lèlica (Figures S9 a S25) utilitzant els temps de

coalescència inferits prèviament i totes les mostres europees del conjunt de dades d'ADN antic sense excloure els polimorfismes que no eren transversions. La inferència de selecció es va restringir al temps més antic mostrejat (és a dir, 528 generacions a San Teodoro 3 – ST3, de Sicília¹⁰²).

Desviacions locals d'ancestria (LAD)

Vam explorar les desviacions d'ancestria local (LAD o *local ancestral deviation*) al conjunt de dades GCAT utilitzant poblacions externes contemporànies i dades genètiques antigues. Per investigar LAD amb dades externes contemporànies, primer vam executar RFmix¹⁰³ sobre el Conjunt de Dades C utilitzant les següents opcions: -e 5 -n 5 --reanalyze-reference per tal d'aplicar l'algorisme d'iteració EM i corregir els individus mesclats en les poblacions de referència. Com a mostres de referència, vam utilitzar els grups genètics inferits per fineSTRUCTURE del Sud de França (SUD, que inclou principalment mostres de Provença i Dordonya) reduïdes a 30 mostres, Nord d'Àfrica Occidental (WNA, que inclou mostres d'Algèria, Tunísia i Marroc del conjunt de dades de Lazaridis⁹⁵), i Nord d'Àfrica Oriental (ENA, que inclou mostres egípcies i beduïnes) en una sola execució (Figura S30). Per comprovar si els senyals detectats podrien provenir d'una migració prèvia provinent de l'Orient Mitjà, vam repetir l'anàlisi, fusionant ENA i WNA en un únic grup d'Àfrica del Nord i utilitzant el grup palestí (PAL) com a referència del Pròxim Orient. El *flag* – reanalyze-reference es va utilitzar per tenir en compte una possible barreja en el panell de referència, que s'espera en les poblacions d'Àfrica del Nord. Per comprovar si els senyals detectats provenien d'una font ancestral comuna dels temps neolítics o post-neolítics, vam repetir la inferència d'ancestria local a tot el conjunt de dades GCAT, així com als grups d'Àfrica del Nord (ENA i WNA) i PAL, utilitzant els grups genètics⁴⁵ antics inferits EEF, ENS i WHG com a referències.

Declaració de disponibilitat de dades

Els WGS per la cohort del GCAT estan disponibles a EGA (<https://ega-archive.org/>) sota el codi d'accés EGAD00001007774.

Disponibilitat del codi

No es va escriure cap programari específic per a aquest projecte. El codi utilitzat en el projecte es pot descarregar des del següent repositori de github:

<https://github.com/JorgeGarciaC/SelDem-GCAT>

Agraïments

Volem agrair a Evan Irving-Pease i Laura Vilà-Valls els seus consells tècnics. Aquest estudi utilitza dades generades pel projecte GCAT Genomes for Life, un estudi de cohort dels Genomes de Catalunya, Fundació IGTP. IGTP forma part del Programa CERCA / Generalitat de Catalunya. Els autors d'aquest estudi volen reconèixer tots els investigadors del projecte GCAT que van contribuir a la generació de les dades GCAT. Una llista completa dels investigadors està disponible a www.genomesforlife.com. Agraïm al Banc de Sang i Teixits de Catalunya (BST) la seva col·laboració i a tots els voluntaris de GCAT la seva participació en l'estudi.

Finançament

Aquest treball va comptar amb el suport del projecte PID2023-147621NB-I00 finançat per MICIU/AEI/10.13039/501100011033 i per "ERDF A way of making Europe". JGC va comptar amb un contracte de doctorat FPI-MCIN/AEI (PRE2020-095762). GCAT va ser finançat per l'Acció de Dinamització del ISCIII-MINECO i pel Ministeri de Salut de la Generalitat de Catalunya [ADE 10/00026] i està finançat en part per l'Agència de Gestió d'Ajuts Universitaris i de Recerca (AGAUR) [SGR 01537]. SAB va ser finançat pel Czech

Ministry of Education, Youth and Sports [CZ.02.01.01/00/22_008/0004593, RES-HUM: Ready for the Future: Understanding the Long-Term Resilience of Human Culture grant].

Referències

1. Montinaro, F., Pankratov, V., Yelmen, B., Pagani, L. & Mondal, M. Revisiting the out of Africa event with a deep-learning approach. *Am J Hum Genet* **108**, 2037–2051 (2021).
2. Bergström, A., Stringer, C., Hajdinjak, M., Scerri, E. M. L. & Skoglund, P. Origins of modern human ancestry. *Nat.* **590**, 229–237 (2021).
3. Fumagalli, M. *et al.* Signatures of Environmental Genetic Adaptation Pinpoint Pathogens as the Main Selective Pressure through Human Evolution. *PLoS Genet* **7**, e1002355 (2011).
4. Fumagalli, M. *et al.* Greenlandic Inuit show genetic signatures of diet and climate adaptation. *Science* **349**, 1343–1347 (2015).
5. Caro-Consuegra, R. *et al.* Uncovering Signals of Positive Selection in Peruvian Populations from Three Ecological Regions. *Mol Biol Evol* **39**, msac158 (2022).
6. Irving-Pease, E. K. *et al.* The selection landscape and genetic legacy of ancient Eurasians. *Nature* **625**, 312–320 (2024).
7. Sinigaglia, B. *et al.* Exploring Adaptive Phenotypes for the Human Calcium-Sensing Receptor Polymorphism R990G. *Mol Biol Evol* **41**, msae015 (2024).
8. Rees, J. S., Castellano, S. & Andrés, A. M. The Genomics of Human Local Adaptation. *Trends in Genetics* **36**, 415–428 (2020).
9. Garud, N. R., Messer, P. W., Buzbas, E. O. & Petrov, D. A. Recent Selective Sweeps in North American *Drosophila melanogaster* Show Signatures of Soft Sweeps. *PLOS Genet.* **11**, e1005004 (2015).
10. Sabeti, P. C. *et al.* Genome-wide detection and characterization of positive selection in human populations. *Nature* **449**, 913–918 (2007).

11. Voight, B. F., Kudaravalli, S., Wen, X. & Pritchard, J. K. A Map of Recent Positive Selection in the Human Genome. *PLoS Biol* **4**, e72 (2006).
12. Akbari, A. *et al.* Identifying the favored mutation in a positive selective sweep. *Nat Methods* **15**, 279–282 (2018).
13. Speidel, L., Forest, M., Shi, S. & Myers, S. R. A method for genome-wide genealogy estimation for thousands of samples. *Nat Genet* **51**, 1321–1329 (2019).
14. Stern, A. J., Wilton, P. R. & Nielsen, R. An approximate full-likelihood method for inferring selection and allele frequency trajectories from DNA sequence data. *PLoS Genet* **15**, e1008384 (2019).
15. Field, Y. *et al.* Detection of human adaptation during the past 2000 years. *Science* **354**, 760–764 (2016).
16. Smith, J. M. & Haigh, J. The hitch-hiking effect of a favourable gene. *Genet Res* **23**, 23–35 (1974).
17. Kaplan, N. L., Hudson, R. R. & Langley, C. H. The ‘hitchhiking effect’ revisited. *Genetics* **123**, 887–899 (1989).
18. Enattah, N. S. *et al.* Identification of a variant associated with adult-type hypolactasia. *Nat Genet* **30**, 233–237 (2002).
19. Jablonski, N. G. & Chaplin, G. Epidermal pigmentation in the human lineage is an adaptation to ultraviolet radiation. *J Hum Evol* **65**, 671–675 (2013).
20. Huerta-Sánchez, E. *et al.* Altitude adaptation in Tibetans caused by introgression of Denisovan-like DNA. *Nature* **512**, 194–197 (2014).
21. Farré, X. *et al.* Skin Phototype and Disease: A Comprehensive Genetic Approach to Pigmentary Traits Pleiotropy Using PRS in the GCAT Cohort. *Genes* **14**, 149 (2023).
22. Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* **19**, 1655–1664 (2009).
23. Loh, P.-R. *et al.* Inferring Admixture Histories of Human Populations Using Linkage Disequilibrium. *Genetics* **193**, 1233–1254 (2013).

24. Hellenthal, G. *et al.* A Genetic Atlas of Human Admixture History. *Science* **343**, 747–751 (2014).
25. Salter-Townshend, M. & Myers, S. Fine-Scale Inference of Ancestry Segments Without Prior Knowledge of Admixing Groups. *Genetics* **212**, 869–889 (2019).
26. Olalde, I. *et al.* The genomic history of the Iberian Peninsula over the past 8000 years. *Science* **363**, 1230–1234 (2019).
27. Villalba-Mouco, V. *et al.* Survival of Late Pleistocene Hunter-Gatherer Ancestry in the Iberian Peninsula. *Curr Biol* **29**, 1169–1177.e7 (2019).
28. Lazaridis, I. *et al.* Ancient human genomes suggest three ancestral populations for present-day Europeans. *Nature* **513**, 409–413 (2014).
29. Moorjani, P. *et al.* The History of African Gene Flow into Southern Europeans, Levantines, and Jews. *PLoS Genet* **7**, e1001373 (2011).
30. Bycroft, C. *et al.* Patterns of genetic differentiation and the footprints of historical migrations in the Iberian Peninsula. *Nat Commun* **10**, 551 (2019).
31. Hernández, C. L. *et al.* Human Genomic Diversity Where the Mediterranean Joins the Atlantic. *Mol Biol Evol* **37**, 1041–1055 (2020).
32. Biagini, S. A., Ramos-Luis, E., Comas, D. & Calafell, F. The place of metropolitan France in the European genomic landscape. *Hum Genet* **139**, 1091–1105 (2020).
33. Salas Ausens, J. A. *En busca de El Dorado : inmigración francesa en la España de la Edad Moderna.* (Universidad del País Vasco, Bilbao, 2009).
34. Millàs i Castellví, C. *Els altres catalans dels segles XVI i XVII: la immigració francesa al Baix Llobregat en temps dels Àustria.* (2005).
35. Rumech, R. S. i. Quan la terra promesa era al sud. La immigració francesa al Maresme als segles XVI i XVII. Fundació Huro. *Paratge* 132–132 (2015).
36. Barquer i Cerdà, A., Congost i Colomer, R. & Mutos Xicola, C. El reto de reconstituir procesos migratorios. Diferentes modelos de migraciones francesas en la diócesis de

- Girona en la época moderna. *Revista de Demografía Histórica-Journal of Iberoamerican Population Studies* **40**, 61–88 (2022).
37. Dopazo, J. *et al.* 267 Spanish Exomes Reveal Population-Specific Differences in Disease-Related Genetic Variation. *Mol Biol Evol* **33**, 1205–1218 (2016).
 38. Auton, A. *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
 39. Byrska-Bishop, M. *et al.* High-coverage whole-genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios. *Cell* **185**, 3426–3440.e19 (2022).
 40. Obón-Santacana, M. *et al.* GCAT|Genomes for life: a prospective cohort study of the genomes of Catalonia. *BMJ Open* **8**, e018324 (2018).
 41. Galván-Femenía, I. *et al.* Multitrait genome association analysis identifies new susceptibility genes for human anthropometric variation in the GCAT cohort. *J Med Genet* **55**, 765–778 (2018).
 42. Valls-Margarit, J. *et al.* GCAT|Panel, a comprehensive structural variant haplotype map of the Iberian population from high-coverage whole-genome sequencing. *Nucleic Acids Res* **50**, 2464–2479 (2022).
 43. Fibla, J. *et al.* The power of geohistorical boundaries for modeling the genetic background of human populations: The case of the rural catalan Pyrenees. *Front. Genet.* **13**, (2023).
 44. Biagini, S. A. *et al.* People from Ibiza: an unexpected isolate in the Western Mediterranean. *Eur J Hum Genet* **27**, 941–951 (2019).
 45. Allentoft, M. E. *et al.* Population genomics of post-glacial western Eurasia. *Nature* **625**, 301–311 (2024).
 46. Flores-Bello, A. *et al.* Genetic origins, singularity, and heterogeneity of Basques. *Curr Biol* **31**, 2167–2177.e4 (2021).

47. Wangkumhang, P., Greenfield, M. & Hellenthal, G. An efficient method to identify, date, and describe admixture events using haplotype information. *Genome Res.* **32**, 1553–1564 (2022).
48. Norton, H. L. *et al.* Genetic Evidence for the Convergent Evolution of Light Skin in Europeans and East Asians. *Mol Biol Evol* **24**, 710–722 (2007).
49. Pickrell, J. K. *et al.* Signals of recent positive selection in a worldwide sample of human populations. *Genome Res.* **19**, 826–837 (2009).
50. Lamason, R. L. *et al.* SLC24A5, a Putative Cation Exchanger, Affects Pigmentation in Zebrafish and Humans. *Science* **310**, 1782–1786 (2005).
51. Ginger, R. S. *et al.* SLC24A5 Encodes a *trans*-Golgi Network Protein with Potassium-dependent Sodium-Calcium Exchange Activity That Regulates Human Epidermal Melanogenesis*. *J Biol Chem* **283**, 5486–5495 (2008).
52. Myung, C. H., Lee, J. E., Jo, C. S., Park, J. il & Hwang, J. S. Regulation of Melanophilin (Mlph) gene expression by the glucocorticoid receptor (GR). *Sci Rep* **11**, 16813 (2021).
53. Ermini, L. *et al.* Evolutionary selection of alleles in the melanophilin gene that impacts on prostate organ function and cancer risk. *Evol Med Public Health* **9**, 311–321 (2021).
54. Cario-André, M., Pain, C., Gauthier, Y., Casoli, V. & Taieb, A. In vivo and in vitro evidence of dermal fibroblasts influence on human epidermal pigmentation. *Pigment Cell Res* **19**, 434–442 (2006).
55. Huff, C. D. *et al.* Crohn's Disease and Genetic Hitchhiking at IBD5. *Mol Biol Evol* **29**, 101–111 (2012).
56. Mathieson, S. & Mathieson, I. FADS1 and the Timing of Human Adaptation to Agriculture. *Molecular Biology and Evolution* **35**, 2957–2970 (2018).
57. Longo, N., Frigeni, M. & Pasquali, M. Carnitine transport and fatty acid oxidation. *Biochim Biophys Acta Mol Cell Res* **1863**, 2422–2435 (2016).

58. Lam, Y. W. F. Chapter 1 - Principles of Pharmacogenomics: Pharmacokinetic, Pharmacodynamic, and Clinical Implications. in *Pharmacogenomics (Second Edition)* (eds. Lam, Y. W. F. & Scott, S. A.) 1–53 (Academic Press, 2019). doi:10.1016/B978-0-12-812626-4.00001-2.
59. Wang, Z. *et al.* Signatures of recent positive selection at the ATP-binding cassette drug transporter superfamily gene loci. *Hum Mol Genet* **16**, 1367–1380 (2007).
60. Yang, X. *et al.* Systematic genetic and genomic analysis of cytochrome P450 enzyme activities in human liver. *Genome Res.* **20**, 1020–1036 (2010).
61. Liu, X. *et al.* Detecting and Characterizing Genomic Signatures of Positive Selection in Global Populations. *Am J Hum Genet* **92**, 866–881 (2013).
62. Han, J. *et al.* TMEM232 promotes the inflammatory response in atopic dermatitis via the nuclear factor- κ B and signal transducer and activator of transcription 3 signalling pathways. *British Journal of Dermatology* **189**, 195–209 (2023).
63. Yahalom, V. *et al.* SMYD1 is the underlying gene for the AnWj-negative blood group phenotype. *European Journal of Haematology* **101**, 496–501 (2018).
64. Poole, J. & Van Alphen, L. Haemophilus influenzae receptor and the AnWj antigen. *Transfusion* **28**, 289–289 (1988).
65. van Alphen, L., van Ham, M., Geelen-van den Broek, L. & Pieters, T. Relationship between secretion of the Anton blood group antigen in saliva and adherence of *Haemophilus influenzae* to oropharynx epithelial cells. *FEMS Microbiol Lett* **47**, 357–362 (1989).
66. Brusselmans, K. *et al.* Squalene Synthase, a Determinant of Raft-associated Cholesterol and Modulator of Cancer Cell Proliferation*. *J Biol Chem* **282**, 18777–18785 (2007).
67. Weng, M. *et al.* Fasting inhibits aerobic glycolysis and proliferation in colorectal cancer via the Fdft1-mediated AKT/mTOR/HIF1 α pathway suppression. *Nat Commun* **11**, 1869 (2020).

68. Graham, S. E. *et al.* The power of genetic diversity in genome-wide association studies of lipids. *Nature* **600**, 675–679 (2021).
69. Harrison, P. W. *et al.* Ensembl 2024. *Nucleic Acids Res* **52**, D891–D899 (2024).
70. Körner, A. *et al.* Sema7A is crucial for resolution of severe inflammation. *Proc Natl Acad Sci U S A* **118**, e2017527118 (2021).
71. Ghofrani, J., Lucar, O., Dugan, H., Reeves, R. K. & Jost, S. Semaphorin 7A modulates cytokine-induced memory-like responses by human natural killer cells. *Eur J Immunol* **49**, 1153–1166 (2019).
72. Gras, C. *et al.* Semaphorin 7A protein variants differentially regulate T-cell activity. *Transfusion* **53**, 270–283 (2013).
73. Holmes, S. *et al.* Sema7A is a Potent Monocyte Stimulator. *Scand J Immunol* **56**, 270–275 (2002).
74. May-Wilson, S. *et al.* Large-scale GWAS of food liking reveals genetic determinants and genetic correlations with distinct neurophysiological traits. *Nat Commun* **13**, 2743 (2022).
75. Secolin, R. *et al.* Distribution of local ancestry and evidence of adaptation in admixed populations. *Sci Rep* **9**, 13900 (2019).
76. Secolin, R. *et al.* Exploring a Region on Chromosome 8p23.1 Displaying Positive Selection Signals in Brazilian Admixed Populations: Additional Insights Into Predisposition to Obesity and Related Disorders. *Front. Genet.* **12**, (2021).
77. Alvarez, M. I. *et al.* Human genetic variation in VAC14 regulates Salmonella invasion and typhoid fever through modulation of cholesterol. *Proc Natl Acad Sci U S A* **114**, E7746–E7755 (2017).
78. Gilchrist, J. J. *et al.* Genetic variation in VAC14 is associated with bacteremia secondary to diverse pathogens in African children. *Proc Natl Acad Sci U S A* **115**, E3601–E3603 (2018).

79. Bhatia, G. *et al.* Genome-wide Scan of 29,141 African Americans Finds No Evidence of Directional Selection since Admixture. *Am J Hum Genet* **95**, 437–444 (2014).
80. Hejase, H. A., Mo, Z., Campagna, L. & Siepel, A. A Deep-Learning Approach for Inference of Selective Sweeps from the Ancestral Recombination Graph. *Mol Biol Evol* **39**, msab332 (2021).
81. Mathieson, I. *et al.* Genome-wide patterns of selection in 230 ancient Eurasians. *Nature* **528**, 499–503 (2015).
82. Sazzini, M. *et al.* Genomic history of the Italian population recapitulates key evolutionary dynamics of both Continental and Southern Europeans. *BMC Biol* **18**, 51 (2020).
83. Key, F. M. *et al.* Emergence of human-adapted *Salmonella enterica* is linked to the Neolithization process. *Nat Ecol Evol* **4**, 324–333 (2020).
84. L'Hôte, L. *et al.* An 8000 years old genome reveals the Neolithic origin of the zoonosis *Brucella melitensis*. *Nat Commun* **15**, 6132 (2024).
85. Barrie, W. *et al.* Elevated genetic risk for multiple sclerosis emerged in steppe pastoralist populations. *Nature* **625**, 321–328 (2024).
86. Mendoza-Revilla, J. *et al.* Disentangling Signatures of Selection Before and After European Colonization in Latin Americans. *Mol Biol Evol* **39**, msac076 (2022).
87. Arezki Ferrad, M. Repaso de la historia de los amazighes en al-Ándalus desde la conquista hasta los reinos taifas. in *Los bereberes en la Península Ibérica: contribución de los Amazighes a la historia de al-Ándalus*, 2021, ISBN 978-84-338-6790-2, págs. 81-104 81–104 (Editorial Universidad de Granada, 2021).
88. Blay, N. *et al.* Disease prevalence, health-related and socio-demographic factors in the GCAT cohort. A comparison with the general population of Catalonia.
2023.09.08.23295239 Preprint at <https://doi.org/10.1101/2023.09.08.23295239> (2023).

89. Picard toolkit. *Broad Institute, GitHub repository* (2019).
90. Danecek, P. *et al.* Twelve years of SAMtools and BCFtools. *GigaScience* **10**, giab008 (2021).
91. Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience* **4**, s13742-015-0047-8 (2015).
92. Patterson, N., Price, A. L. & Reich, D. Population Structure and Eigenanalysis. *PLOS Genet* **2**, e190 (2006).
93. Price, A. L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* **38**, 904–909 (2006).
94. Bergström, A. *et al.* Insights into human genetic variation and population history from 929 diverse genomes. *Science* **367**, eaay5012 (2020).
95. Lazaridis, I. *et al.* Genomic insights into the origin of farming in the ancient Near East. *Nature* **536**, 419–424 (2016).
96. Behr, A. A., Liu, K. Z., Liu-Fang, G., Nakka, P. & Ramachandran, S. pong: fast analysis and visualization of latent clusters in population genetic data. *Bioinformatics* **32**, 2817–2823 (2016).
97. Delaneau, O., Zagury, J.-F., Robinson, M. R., Marchini, J. L. & Dermitzakis, E. T. Accurate, scalable and integrative haplotype estimation. *Nat Commun* **10**, 5436 (2019).
98. Lawson, D. J., Hellenthal, G., Myers, S. & Falush, D. Inference of Population Structure using Dense Haplotype Data. *PLOS Genet* **8**, e1002453 (2012).
99. Tennessen, J. A. *et al.* Evolution and Functional Impact of Rare Coding Variation from Deep Sequencing of Human Exomes. *Science* **337**, 64–69 (2012).
100. Szpiech, Z. A. & Hernandez, R. D. selscan: An Efficient Multithreaded Program to Perform EHH-Based Scans for Positive Selection. *Mol Biol Evol* **31**, 2824–2827 (2014).

101. McLaren, W. *et al.* The Ensembl Variant Effect Predictor. *Genome Biol* **17**, 122 (2016).
102. Scorrano, G. *et al.* Genomic ancestry, diet and microbiomes of Upper Palaeolithic hunter-gatherers from San Teodoro cave. *Commun Biol* **5**, 1–13 (2022).
103. Maples, B. K., Gravel, S., Kenny, E. E. & Bustamante, C. D. RFMix: A Discriminative Modeling Approach for Rapid and Robust Local-Ancestry Inference. *The American Journal of Human Genetics* **93**, 278–288 (2013).

Contribucions dels autors

E.B. i F.C. van concebre l'estudi. J.G. va realitzar totes les anàlisis computacionals. S.A.B. i R.C. van assessorar en les anàlisis genètiques, la interpretació i la discussió. J.G., E.B i F.C. van escriure el manuscrit. Tots els autors van contribuir al manuscrit final.

Aprovació ètica

L'estudi ha estat aprovat pel comitè de revisió institucional CEIm - PSMAR (número de referència 2021/9767/I) i pel Comitè institucional de revisió ètica de projectes (CIREP) de la Universitat Pompeu Fabra (número de referència 296).

Conflictes d'interès

Cap

Llegendes de Figures

Figura 1. Estructura genètica en el conjunt de dades GCAT. A) Anàlisi de components principals (PCA) realitzada amb 141.849 SNPs i 141 mostres GCAT els quatre avis de les quals eren originaris de la mateixa comunitat autònoma d'Espanya. Cada punt geomètric representa un individu d'una regió geogràfica particular. Les poblacions de referència es van recopilar de diverses fonts, cobrint Catalunya, les Illes Balears i el País Valencià a Espanya⁴⁴; França³²; i el panell HGDP de diversitat genòmica humana⁹²; per a detalls poblacionals addicionals, vegeu la Taula S1. El polígon vermell engloba totes les mostres del GCAT, mentre que les mostres espanyoles de Biagini et al. (2019)⁴⁴ es mostren dins el polígon blau. Abreviacions: CLM, Castella- La Mancha; CYL, Castella i Lleó. B) Anàlisi ADMIXTURE en el conjunt de dades GCAT als errors de validació creuada més baixos (K=7, seguits de K=8 i K=6) emprant poblacions de referència de Catalunya, les Illes Balears, i del País Valencià⁴⁴, França³², Nord d'Àfrica⁹³ i el panel HGDP de diversitat genòmica humana⁹²; per a detalls poblacionals addicionals, vegeu la Taula S1. C) Mapa mostrant la fracció de mostres GCAT en cada un dels grups d'haplotips definits amb fineSTRUCTURE quan se silencien les poblacions externes. Els gràfics circulars estan centrats a les comunitats autònomes a les quals pertanyen aquestes mostres. Abreviacions: CAT: Catalunya. ARA: Aragó. MED: Mediterrani. WEST: Regió Oest-Central d'Espanya.

Figura 2. Gràfic de Manhattan de les empremtes de selecció positiva recent al conjunt del GCAT. L'eix y indica el $-\log_{10}(\text{p-valor})$ de l'estadístic *singleton density score* (SDS) calculat al conjunt de genomes del GCAT obtinguts de 704 individus espanyols. En color taronja s'indiquen tots els SNPs amb valors SDS superiors al 99,99% que van acompanyats d'almenys 10 SNPs per sobre dels valors SDS del 99,995% dins una regió d'1 Mb (40 pics en total; vegeu detalls sobre valors SDS, gens i anotacions dels SNPs a les Taules S2-S5). En negre, gens associats a funcions biològiques versemblantment

adaptatives. En negreta, gens candidats prèviament coneguts i associats a fenotips adaptatius

Figura 3. Selecció posterior a la mescla en la regió MHC III. **A.** Proporció d'ancestria nord africana (NA) al llarg del cromosoma 6 en el conjunt del GCAT. **B.** Proporció d'ancestria palestina (PAL) al llarg del cromosoma 6 en el conjunt del GCAT. Les línies negres mostren la proporció mitjana de cada ascendència. Les línies vermelles indiquen 4,42 desviacions estàndard per sobre de la mitjana genòmica. Les regions per sobre de la línia es consideren significativament diferents de la mitjana genòmica. **C.** Proporció d'ancestria dels nòmades esteparis primerencs (ENS) al llarg del cromosoma 6 al Nord d'Àfrica. **D** Proporció d'ancestria ENS a palestins. **E.** Proporció d'ancestria ENS al conjunt del GCAT. Les línies negres mostren la proporció mitjana de cada ascendència. Les línies vermelles indiquen tres desviacions estàndard per sobre de la mitjana genòmica. Les regions per sobre de la línia es consideren significativament diferents de la mitjana genòmica. **F.** Gràfic que mostra els valors de SDS transformats després de la correcció de FDR, i els gens corresponents subjacents a la regió LAD detectada al cromosoma 6. La línia vermella mostra el llindar de significació amb $\alpha=0,05$. Vegeu els detalls a la Taula S10.

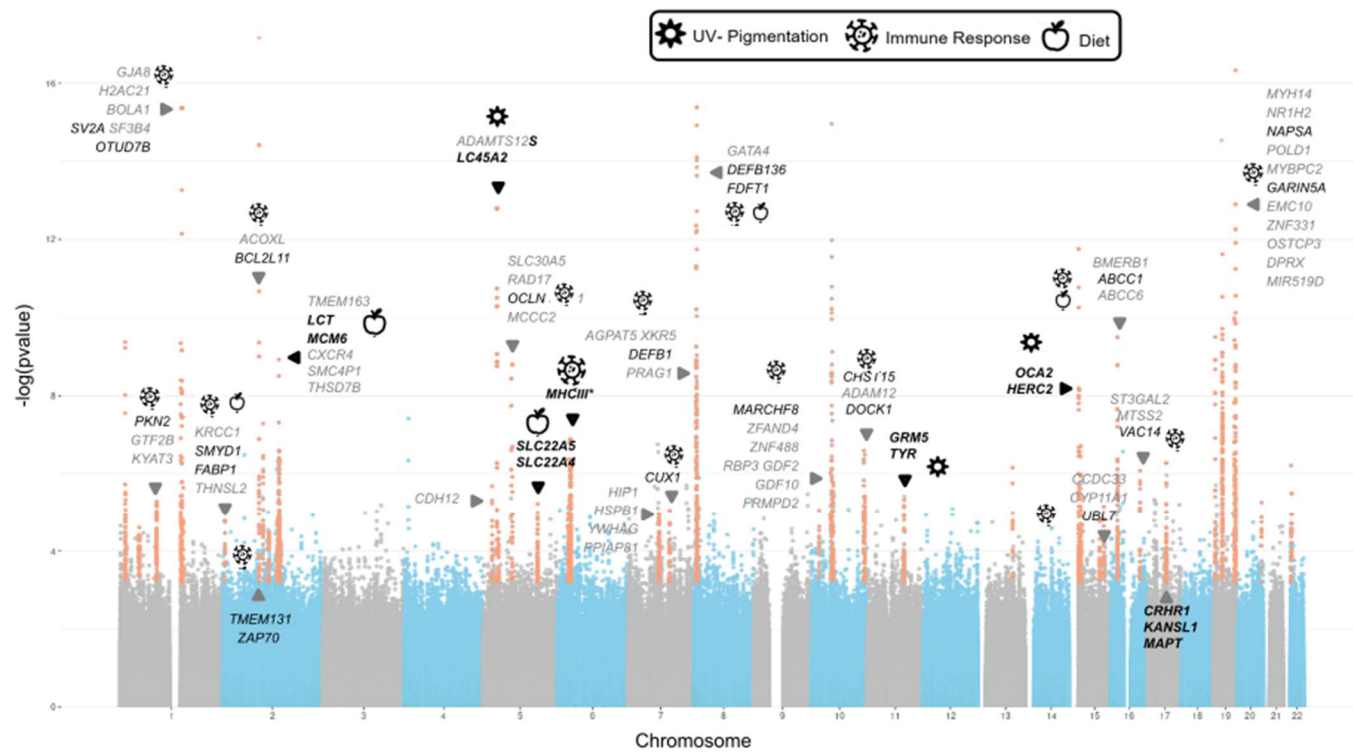


Figura 2

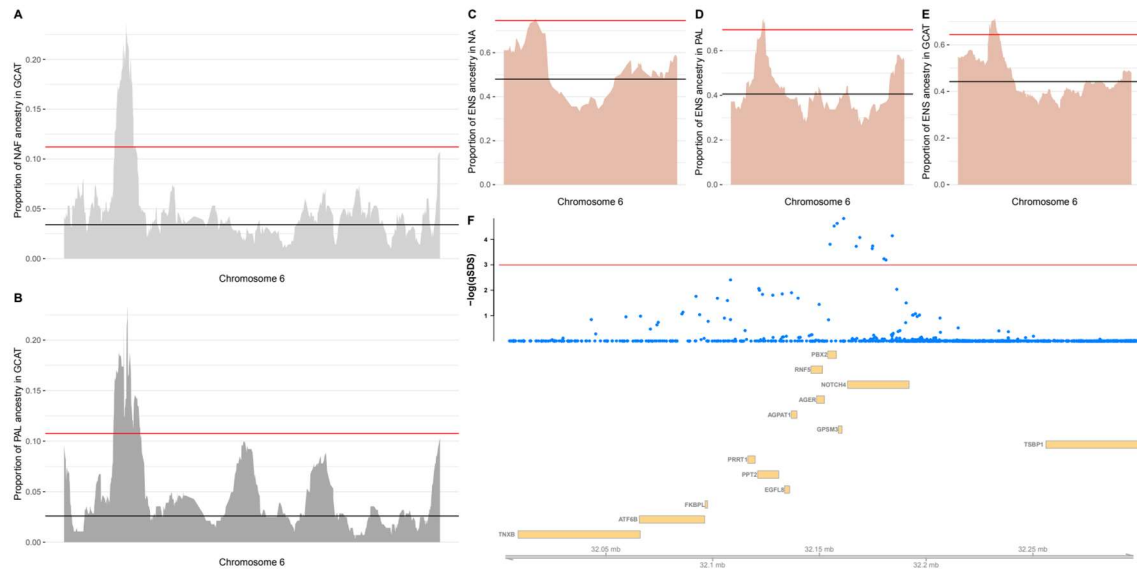


Figura 3

Taula 1. Principals gens candidats de selecció positiva en el conjunt del GCAT.

Funció	Gens	SNP	Tipus	Mètode	LogLR	s
Pigmentació	<i>SLC45A2</i>	rs183671	intrònic	SDS	8.06	0.00427
	<i>OCA2-HERC2</i>	rs7183877	intrònic	SDS	8.75	0.00104
	<i>KITLG</i>	rs556861	intrònic, transcript no-codificant	XP-EHH	6.53	0.00342
	<i>GRM5-TYR</i>	rs7119749	intrònic	SDS	9.81	0.00371
	<i>SLC12A1-DUT</i>	rs9920281	intrònic	XP-EHH	5.39	0.01865
	<i>MLPH-RAB17</i>	rs10176842	intrònic	XP-EHH	5.02	0.00255
Dieta	<i>MCM6-LCT</i>	rs4988235	intrònic	SDS	22.52	0.00862
	<i>SLC22A4</i>	rs1050152	canvi de sentit	SDS	11.68	0.00503
	<i>ABCC1</i>	rs212086	intrònic	SDS	3.71	0.00430
	<i>FDFT1*</i>	rs1296025	intrònic	SDS	4.72	0.00576
	<i>CYP3A4</i>	rs2404955	regió 3'	XP-EHH	7.90	0.00383
Resposta immune	<i>UBL7*</i>	rs750607	regió 3'	SDS	9.46	0.00361
	<i>SMYD1*</i>	rs35662596	intrònic	SDS	6.12	0.01294
	<i>PRAG1*</i>	rs55852693	regió 3'	SDS	5.25	0.00413
	<i>PBX2</i>	rs204991	regió 5'	SDS	6.68	0.00774
	<i>TMEM232</i>	rs10038763	intrònic, transcript no-codificant	XP-EHH	9.24	0.00374
	<i>MTSS2-VAC14*</i>	rs11075777	intrònic	SDS	3.58	0.00222

El coeficient de selecció (s) i la probabilitat de la inferència de selecció (LogRT) estimats per CLUES en el conjunt del GCAT s'indica a les dues darreres columnes. * Nous gens candidats identificats en aquest estudi. Per detalls addicionals, vegeu Taules S5 i S9.

Inferència de la demografia passada i l'adaptació genètica a Espanya utilitzant la cohort del GCAT

Jorge Garcia-Calleja¹, Simone A Biagini^{1,2,3}, Rafael de Cid^{4,5}, Francesc Calafell ^{1*}, Elena Bosch^{1*}

1. Institute of Evolutionary Biology (UPF-CSIC), Department of Medicine and Life Sciences, Universitat Pompeu Fabra, Barcelona 08003, Spain
2. Department of Archaeology and Museology, Masaryk University, Brno, Czech Republic.
3. Center of Molecular Medicine, Central European Institute of Technology, Masaryk University, Brno, Czech Republic.
4. Genomes for Life-GCAT lab, CORE Program, Germans Trias i Pujol Research Institute (IGTP), Badalona 08916, Spain.
5. Grup de REcerca en Impacte de les Malalties Cròniques i les seves Trajectòries (GRIMTra), Germans Trias i Pujol Research Institute (IGTP), Badalona 08916, Spain

* Autors per a correspondència

francesc.calafell@upf.edu

elena.bosch@upf.edu

Taules S1-S12

Taula S1. Poblacions, agrupacions i mides mostrals utilitzades a ADMIXTURE, FineSTRUCTURE i GLOBETROTTER.

Taula S2. Anotació funcional dels SNPs a les principals regions candidates per a selecció segons l'estadístic SDS al conjunt de dades GCAT.

Taula S3. Gens en les regions candidates principals per a selecció segons l'estadístic SDS al conjunt de dades GCAT.

Taula S4. Anàlisi iSAFE a les regions candidates principals per a selecció segons l'estadístic SDS al conjunt de dades GCAT.

Taula S5. Anàlisi CLUES als principals SNPs candidats per a selecció segons l'estadístic SDS al conjunt de dades GCAT.

Taula S6. Anotació funcional dels SNPs a les principals regions candidates per a selecció segons l'estadístic XP-EHH (GCAT vs YRI) al conjunt de dades GCAT.

Taula S7. Gens de les principals regions candidates per a selecció segons l'estadístic XP-EHH (GCAT vs YRI) al conjunt de dades GCAT.

Taula S8. Anàlisi iSAFE a les principals regions candidates per a selecció segons l'estadístic XP-EHH (GCAT vs YRI) al conjunt de dades GCAT.

Taula S9. Anàlisi CLUES als principals SNPs candidats per a selecció segons l'estadístic XP-EHH (GCAT vs YRI) al conjunt de dades GCAT.

Taula S10. Desviacions locals d'ancestria (LAD) detectades al conjunt de dades GCAT quan s'utilitzen dades d'ADN antic i modern de poblacions veïnes.

Taula S11. Procedències geogràfiques dels individus de la cohort del GCAT.

Taula S12. Mostres excloses de les anàlisis perquè provenien de donants que no es van identificar (ni els seus pares/avis) com a blancs/caucàsics.

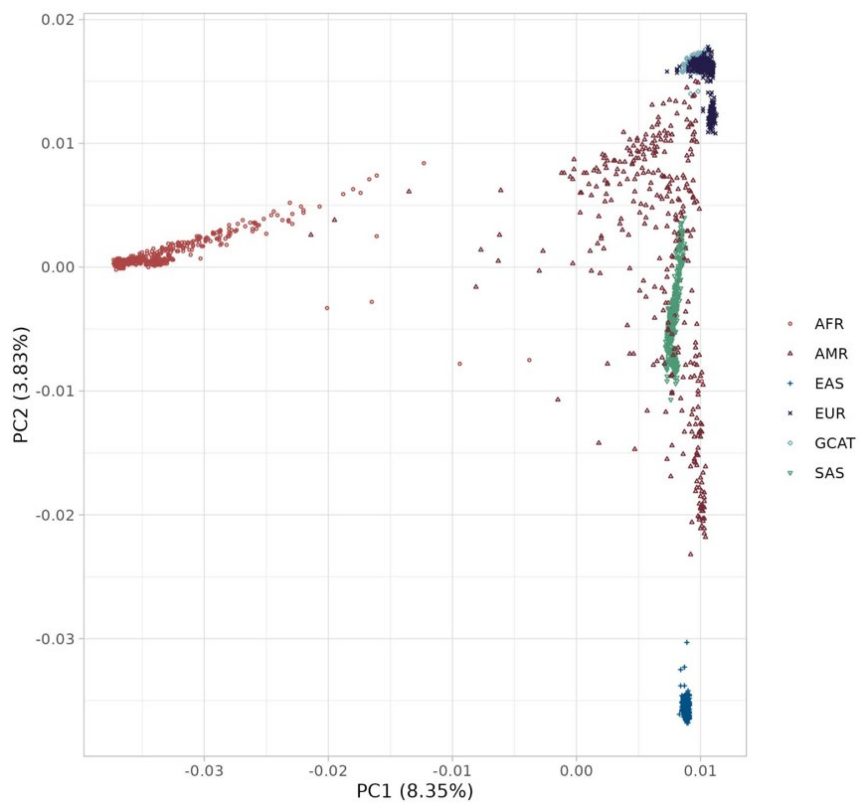


Figura S1. Conjunt de dades GCAT dins del context geogràfic del 1000 GP¹. Anàlisi de components principals (PCA) realitzada amb 679.677 SNPs. Cada punt representa un individu d'una regió continental. AFR, Àfrica; AMR, Amèrica; EAS, Àsia Oriental; EUR, Europa; SAS, Àsia del Sud.

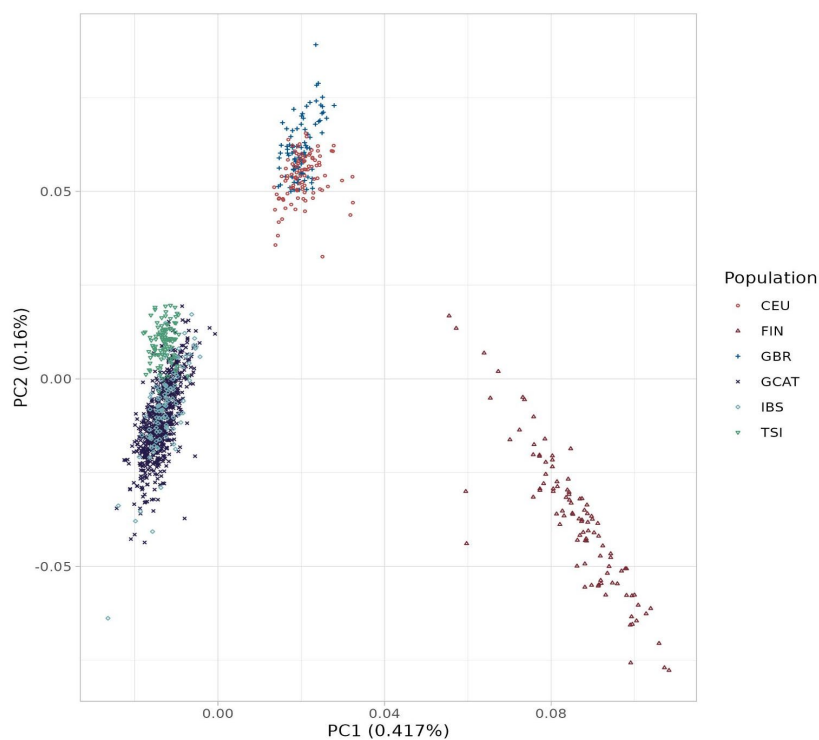


Figura S2. Conjunt de dades GCAT en el context europeu del 1000 GP¹. Anàlisi de components principals (PCA) realitzada amb 679.677 SNPs. Cada punt representa un individu d'una població europea. CEU, residents d'Utah amb ascendència d'Europa septentrional i occidental; FIN, finesos; GBR, britànics; IBE, ibèrics; TSI, toscans.



Figura S3. Conjunt de dades GCAT en el context franco-espanyol. Anàlisi de components principals (PCA) realitzada amb 141.690 SNPs i 141 mostres GCAT els quatre avis de les quals eren originaris de la mateixa comunitat autònoma d'Espanya. El polígon blau mostra les mostres del conjunt de dades GCAT², mentre que el polígon vermell mostra les mostres de Biagini et. al. (2019)³. Vegeu els detalls de les poblacions a la Taula S1.

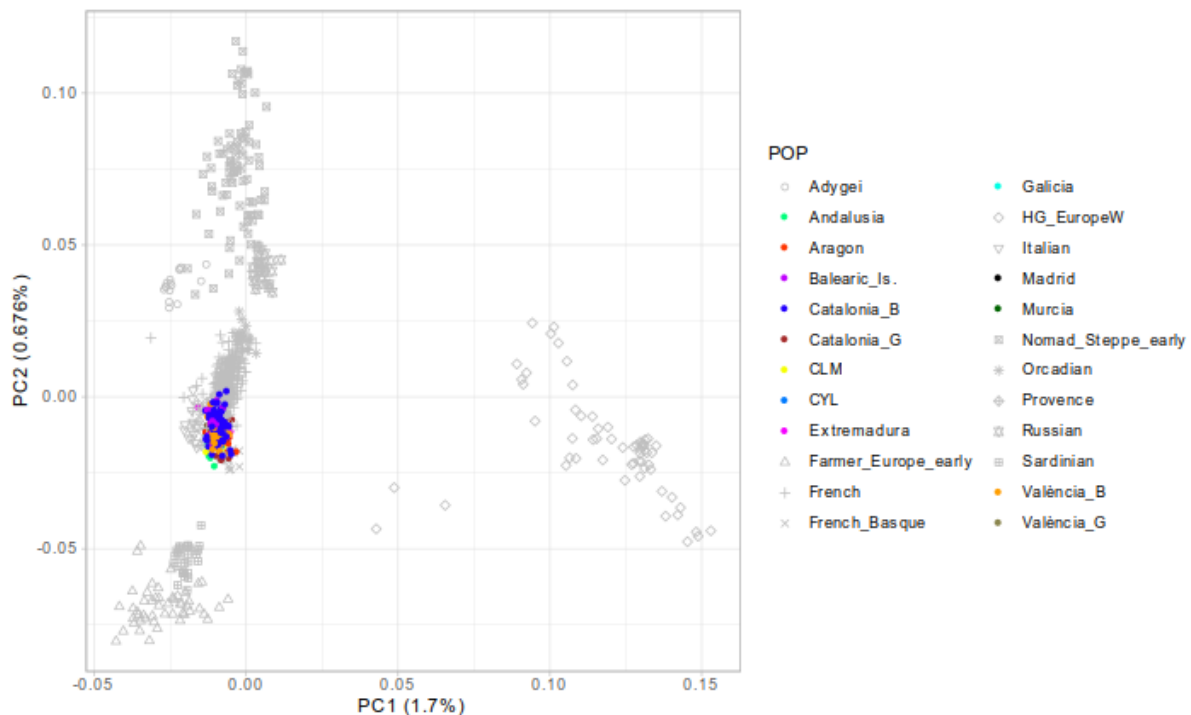


Figura S4. Conjunt de dades GCAT dins del context europeu antic. Anàlisi de components principals (PCA) realitzada amb 31.092 SNPs i 141 mostres de GCAT els quatre avis de les quals eren originaris de la mateixa comunitat autònoma d'Espanya. Es van extreure mostres de tres components europeus antics d'Allentoft et al. (2024)⁴: Farmer_Europe-early, agricultors europeus primerencs (EEF); Nomad_Steppe_early, nòmades primerencs de l'estepa (ENS); HG_EuropeW, caçadors-recol·lector occidentals (WHG). Vegeu els detalls i les abreviatures de les poblacions actuals a la Taula S1.

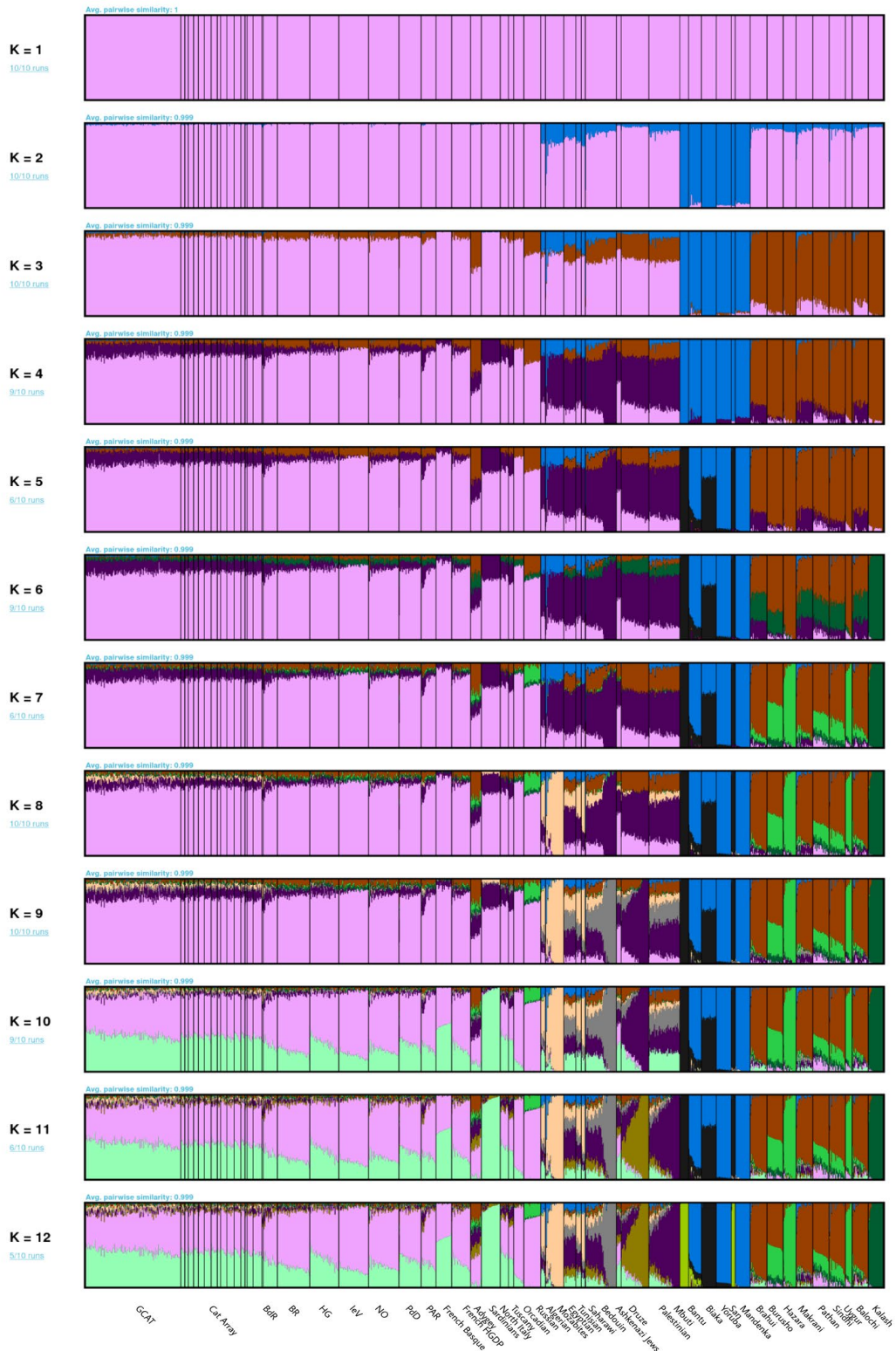
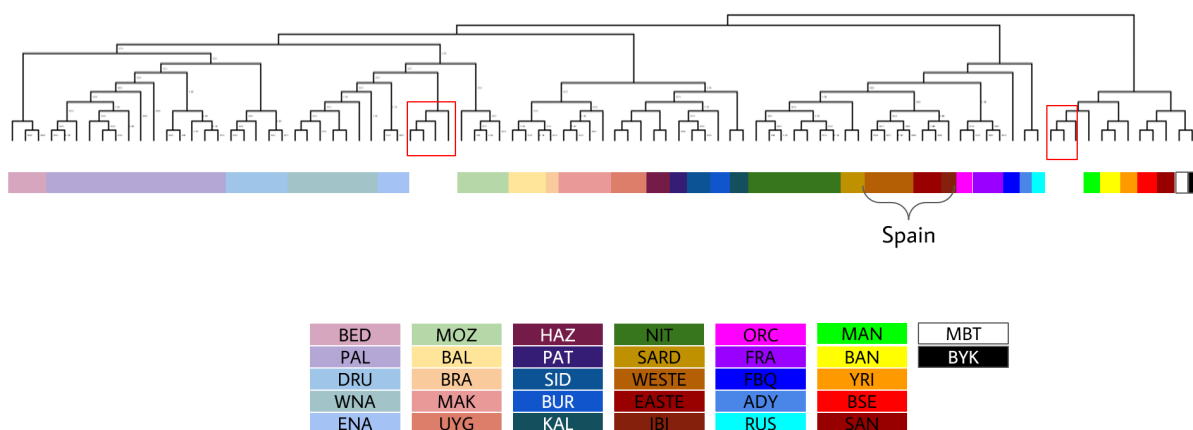


Figura S5. Conjunt de dades del GCAT en el context de les poblacions veïnes de França i el nord d'Àfrica, juntament amb les poblacions africanes, d'Àsia occidental i del sud d'Àsia de l'HGDP⁵. Errors de validació creuada més baixos a $K=7$, seguits de $K=8$ i $K=6$. Vegeu les referències i les abreviatures de les poblacions a la Taula S1.

A.



B.

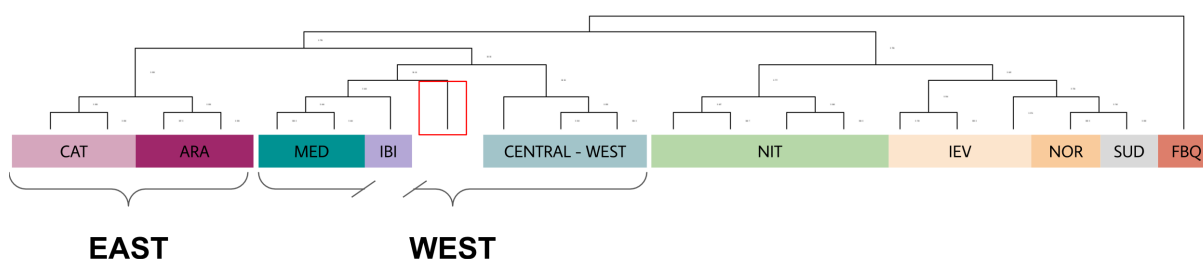


Figura S6. FineSTRUCTURE. A. Clústers detectats per l'algorisme fineSTRUCTURE. En la majoria dels casos, l'agrupació corresponia a l'origen geogràfic o a l'afinitat lingüística. Els quadres vermells corresponen a dos grups que es descarten a causa de la presència d'individus molt barrejats. En el primer, començant per l'esquerra, es van descartar 10 mostres del sud d'Àsia central: Sindhi (2), Balochi (3), Brahui (2), Makrani (3). En el segon, vam descartar 5 mostres pertanyents a mozabites (2), Bantu N.E. (2) i beduïns (1). Abreviatures: ADY, adygei; BED, beduïns; BAL, balutxi; BAN, bantus; BRA, brahui; EEB, parlants bantu del sud-est; BUR, burusho; BYK, biaka; DRU, drusos; EASTE, Est d'Espanya; ENA, Nord d'Àfrica oriental; FBQ, bascos francesos; FRA, França; HAZ, hazara; IBI, Eivissa; KAL, kalash; MAN, mandenka; MAK, makrani; MOZ, mozabites; MBT, mbuti; NIT, nord d'Itàlia; ORC, orcadians; PAL, palestins; PAT, patans; RUS, russos; SAN, san; SARD, sards; SID, sindhi; UYG, uigurs; WESTE, Oest d'Espanya; WNA, nord d'Àfrica occidental; YRI, ioruba. B. Clústers inferits de l'algoritme de sortida fineSTRUCTURE silenciant totes les poblacions externes amb l'opció *Force continents*. Els clústers s'assemblen molt a resultats anteriors^{6,7}. Abreviatures: ARA, Aragó; CAT, Catalunya; MED, Mediterrani; IBI, Eivissa; NIT, nord d'Itàlia; IEV, Bretanya; NOR, nord de França; SUD, Sud de França; FBQ, bascos francesos. El clúster d'Itàlia del Nord (NIT) conté TSI i Itàlia del Nord separats en dues branques. El quadre vermell correspon a una única mostra descartada d'Andalusia que no s'agrupa amb altres mostres. Per a l'anàlisi posterior, els clústers CAT i ARA es fusionen a l'est d'Espanya i els clústers MED i WEST es fusionen a l'oest d'Espanya.

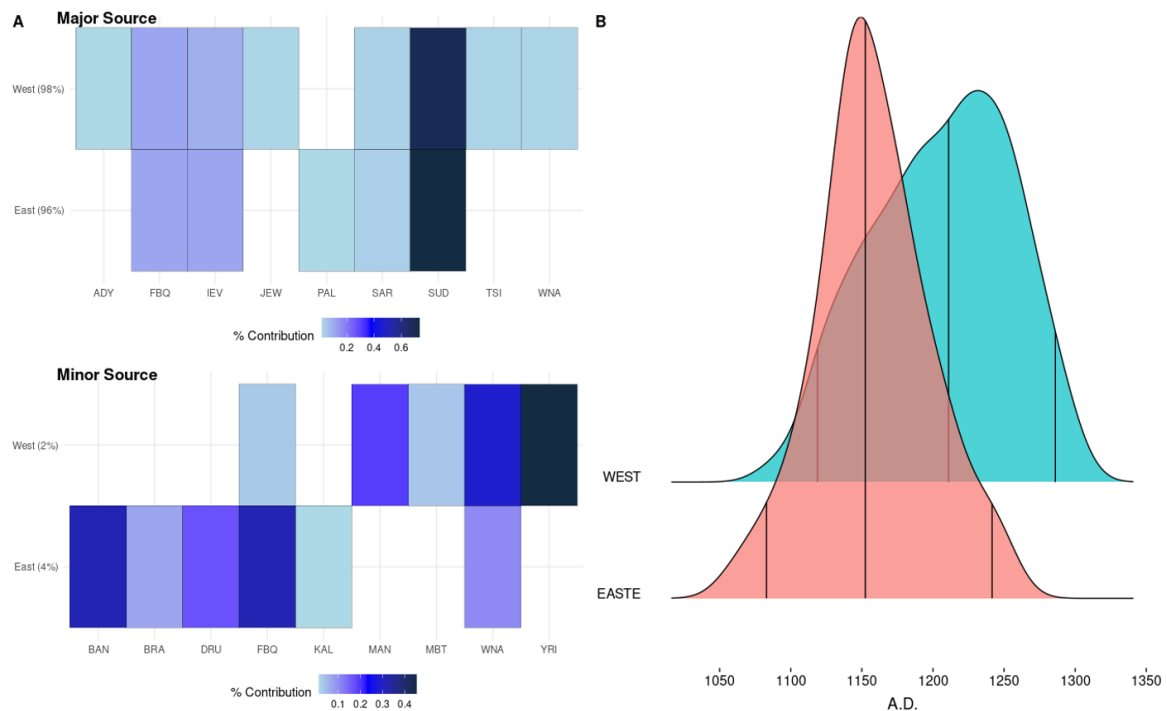


Figura S7. Estimació dels esdeveniments de barreja. A. Les proporcions de les millors fonts de donants de FastGLOBETROTTER es dedueixen per als clústers de l'oest i l'est d'Espanya des de la contribució baixa a la població presumptament barrejada en blau clar fins a una contribució més alta en blau fosc. A l'oest d'Espanya, les fonts majors representen un 98% de l'aportació mentre que les fonts menors representen el 2% de l'aportació total. A l'est d'Espanya, les fonts majors representen un 96% de l'aportació mentre que les fonts menors representen el 4% de l'aportació total. Abreviatures de població: ADY, adygei; FBQ, bascos francesos; IEV, Bretanya; JEUU, jueus ashkenazi; PAL, palestins; SAR, sards; SUD, Sud de França; TSI, toscans; WNA, nord d'Àfrica occidental; BAN, bantu; BRA, brahui; DRU, drusos; KAL, kalash; MAN, mandenka; MBT, mbuti; YRI, ioruba. B. Valors bootstrap de les dates de mescla inferides a partir de 100 rèpliques amb fastGLOBETROTTER per als clústers de l'oest i l'est d'Espanya. Les línies negres indiquen la mitjana, i els percentils del 25 i del 75%.

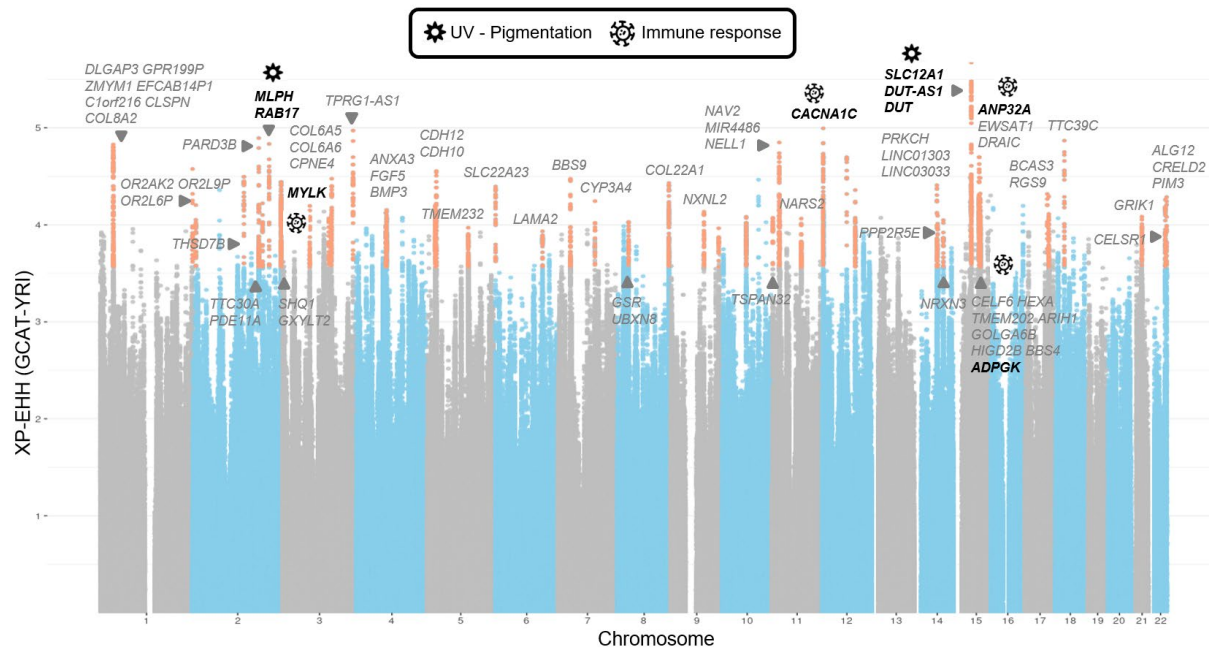


Figura S8. Gràfic de Manhattan de les empremtes de selecció positiva al conjunt de dades del GCAT. L'eix y indica els valors de l'estadístic XP-EHH al GCAT en comparació amb la població YRI. Els pics destacats indiquen tots els SNP per sobre dels valors XP-EHH superiors al 99,99% que van acompanyats d'almenys 10 SNPs per sobre dels valors XP-EHH del 99,995% dins una regió d'1 Mb (40 pics en total; vegeu detalls sobre valors XP-EHH, gens i anotacions dels SNPs a les Taules S6-S9). En negreta negra, gens associats a funcions biològiques adaptatives plausibles.

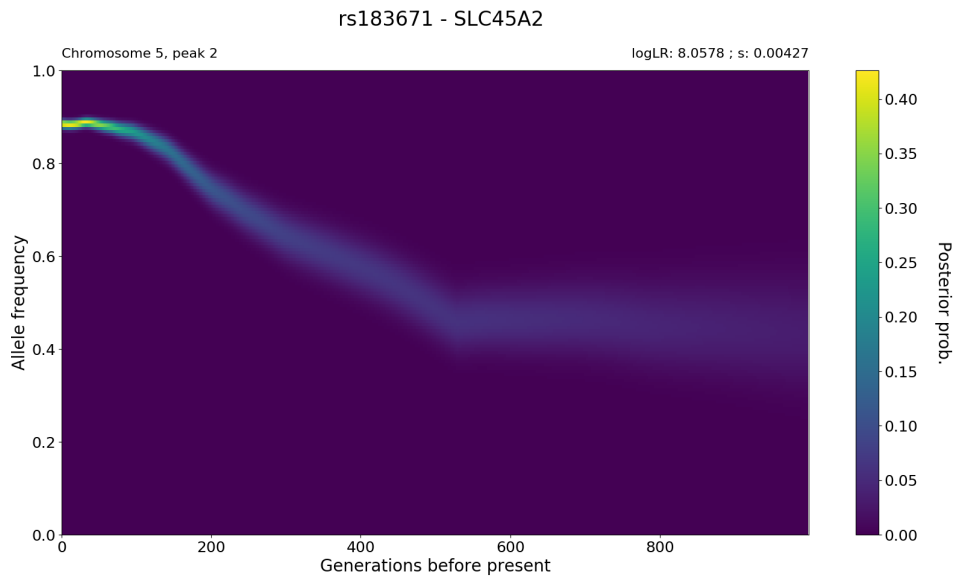


Figura S9. Trajectòria de la freqüència al·lèlica, coeficient (s) de selecció i probabilitat de selecció positiva (logLR) obtinguts a partir de CLUES per a la variant rs183671 dins del locus candidat *SLC45A2*.

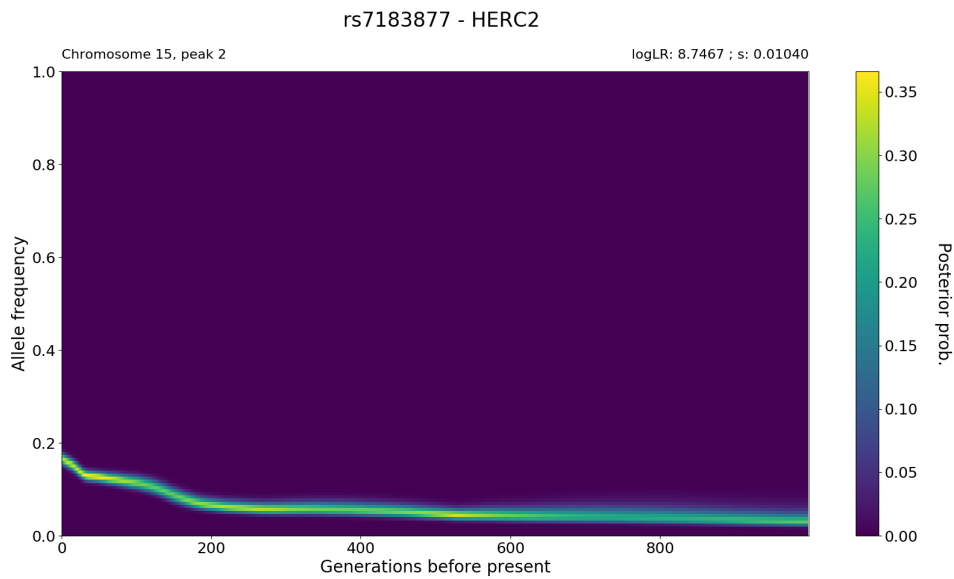


Figura S10. Trajectòria de la freqüència al·lèlica, coeficient (s) de selecció i probabilitat de selecció positiva (logLR) obtinguts a partir de CLUES per a la variant rs7183877 dins del locus candidat *OCA-HERC2*.

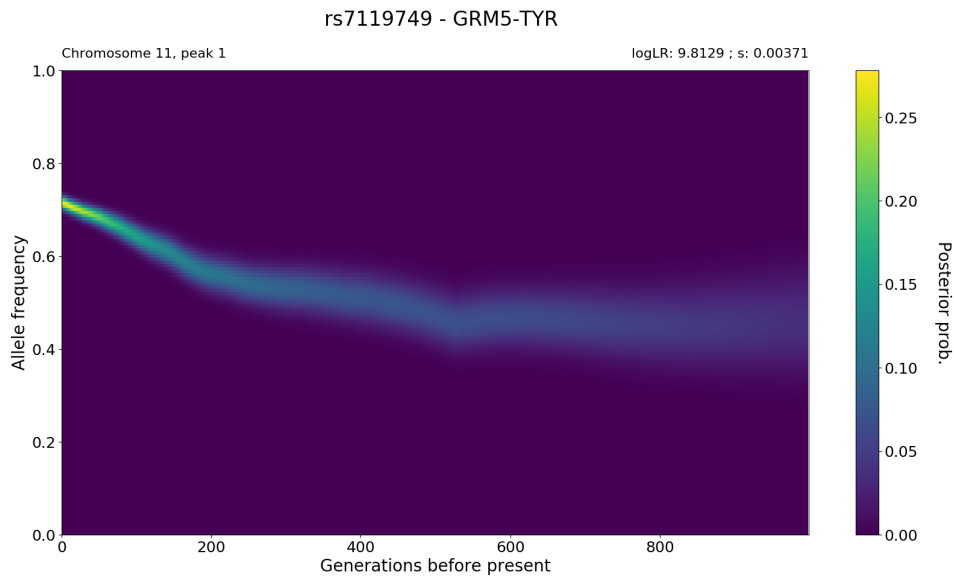


Figura S11. Trajectòria de la freqüència al·lèlica, coeficient (s) de selecció i probabilitat de selecció positiva (logLR) obtinguts a partir de CLUES per a la variant rs7119747 dins del locus candidat *GRM5-TYR*.

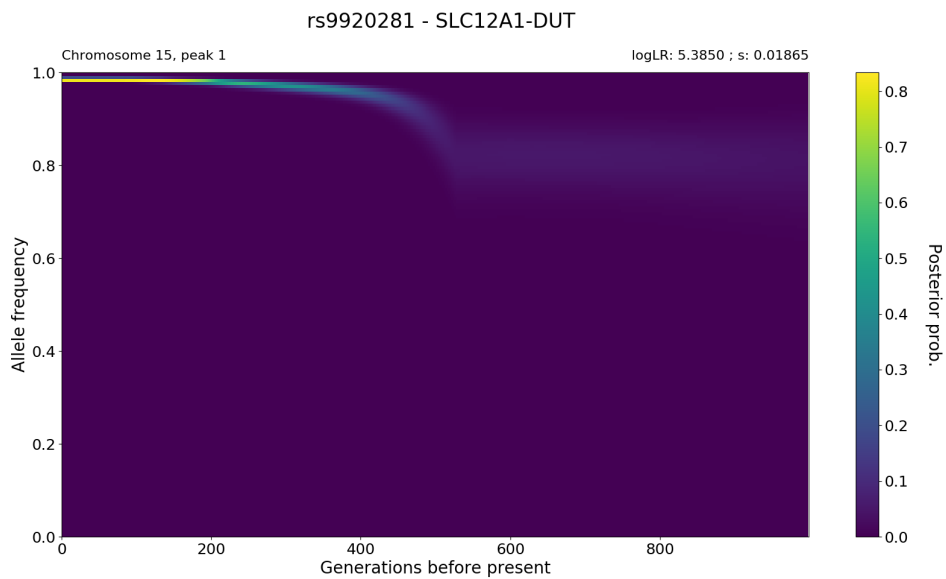


Figura S12. Trajectòria de la freqüència al·lèlica, coeficient (s) de selecció i probabilitat de selecció positiva (logLR) obtinguts a partir de CLUES per a la variant rs9920281 dins de la regió candidata *SLC12A1-DUT*.

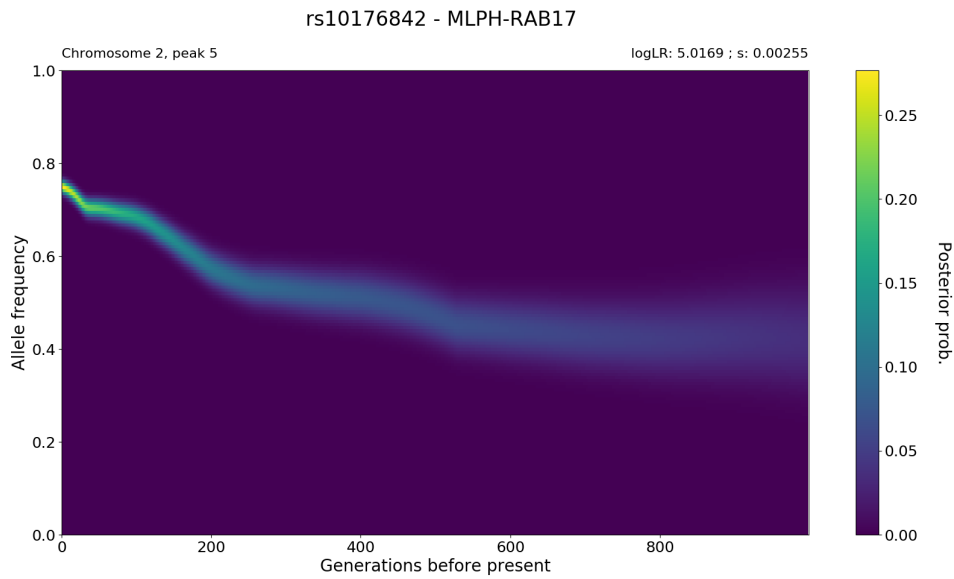


Figura S13. Trajectòria de la freqüència al·lèlica, coeficient (s) de selecció i probabilitat de selecció positiva (logLR) obtinguts a partir de CLUES per a la variant rs10176842 dins de la regió candidata *MLPH-RAB17*.

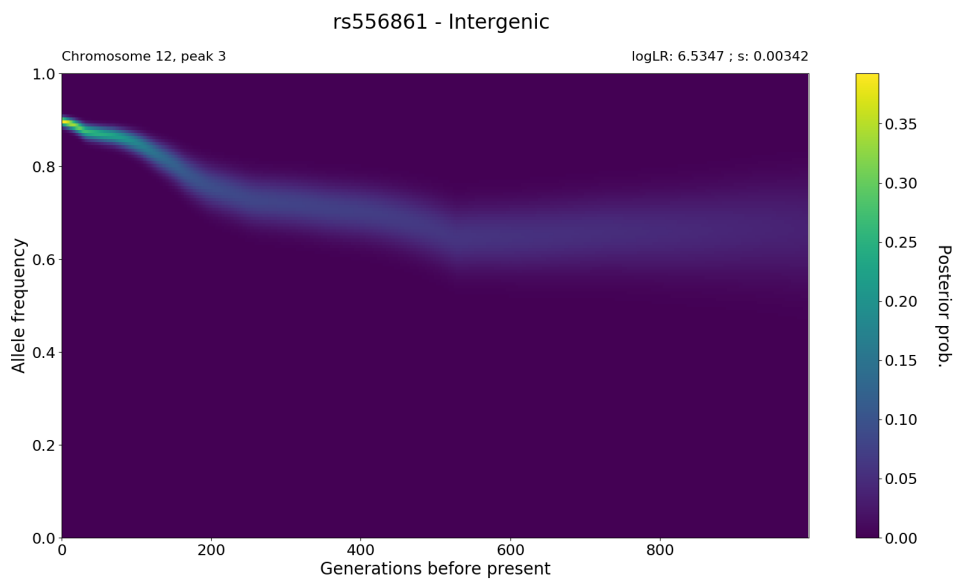


Figura S14. Trajectòria de la freqüència al·lèlica, coeficient (s) de selecció i probabilitat de selecció positiva (logLR) obtinguts a partir de CLUES per a la variant rs556861 dins de la regió candidata *KITLG*.

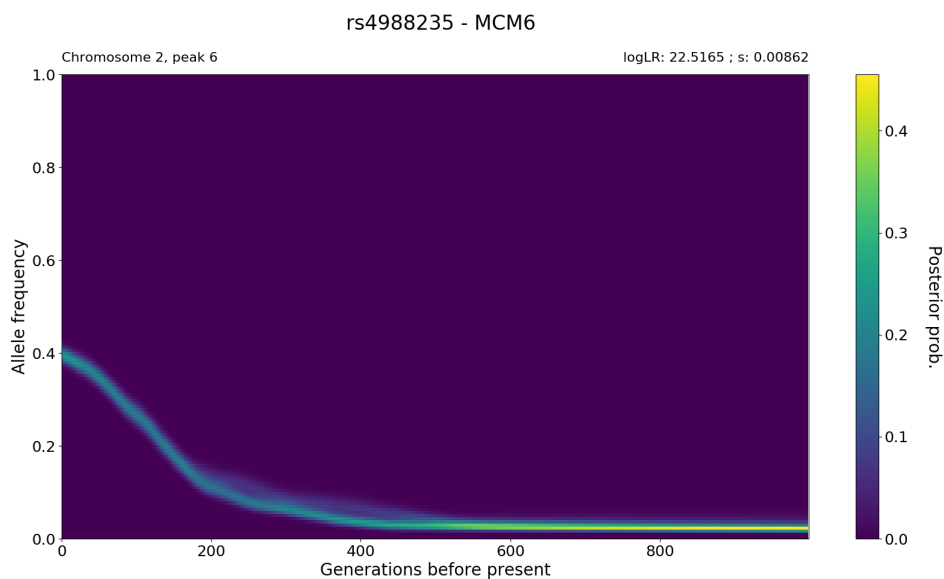


Figura S15. Trajectòria de la freqüència al·lèlica, coeficient (s) de selecció i probabilitat de selecció positiva (logLR) obtinguts a partir de CLUES per a la variant rs4988235 dins de la regió candidata *MCM6* - *LCT*.

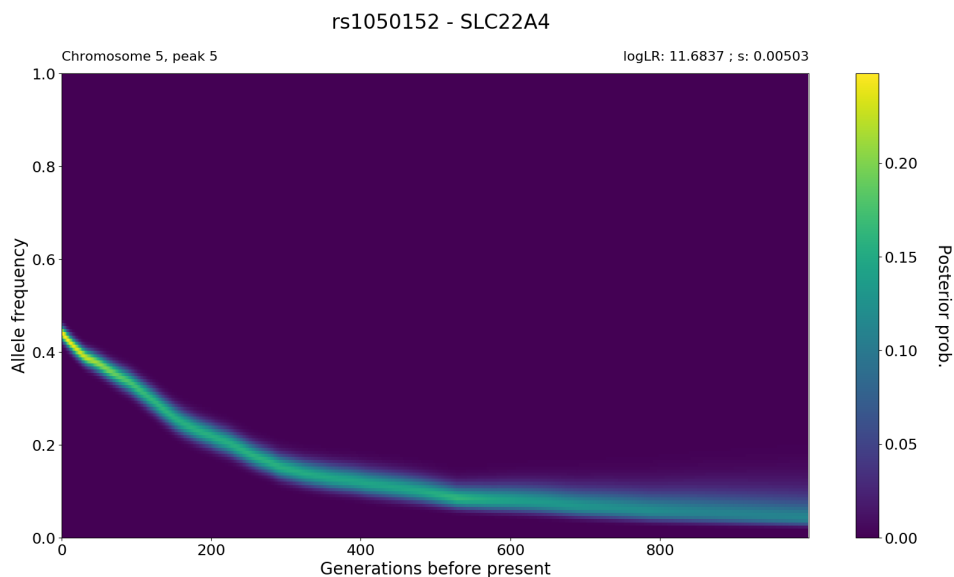


Figura S16. Trajectòria de la freqüència al·lèlica, coeficient (s) de selecció i probabilitat de selecció positiva (logLR) obtinguts a partir de CLUES per a la variant rs1050152 dins de la regió candidata *SLC22A4*.

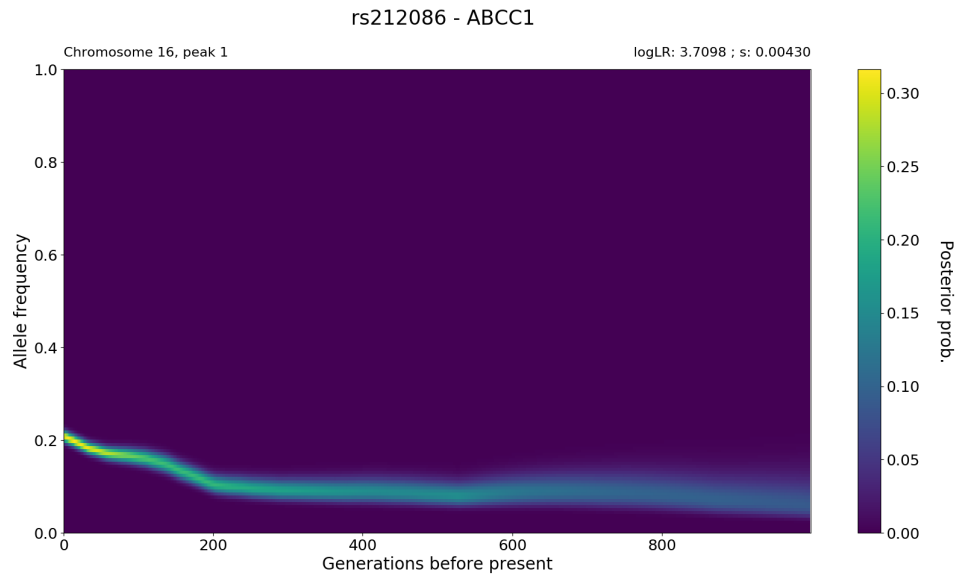


Figura S17. Trajectòria de la freqüència al·lèlica, coeficient (s) de selecció i probabilitat de selecció positiva (logLR) obtinguts a partir de CLUES per a la variant rs212086 dins de la regió candidata *ABCC1*.

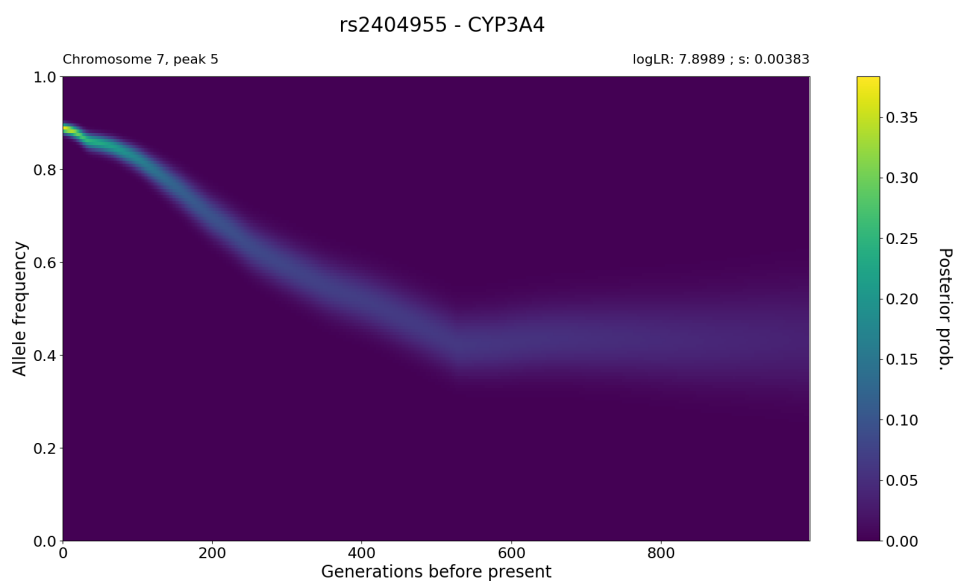


Figura S18. Trajectòria de la freqüència al·lèlica, coeficient (s) de selecció i probabilitat de selecció positiva (logLR) obtinguts a partir de CLUES per a la variant rs2404955 dins de la regió candidata *CYP3A4*.

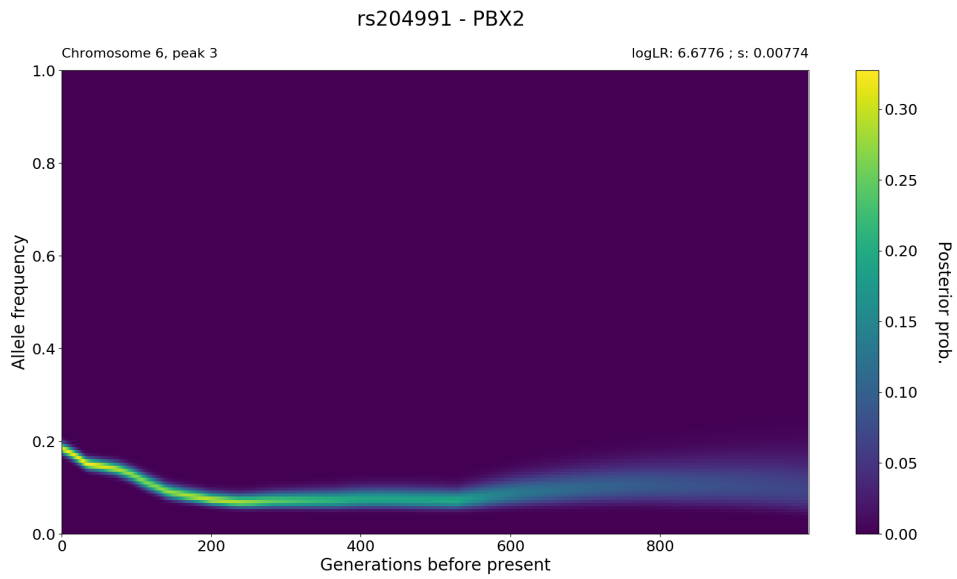


Figura S19. Trajectòria de la freqüència al·lèlica, coeficient (s) de selecció i probabilitat de selecció positiva (logLR) obtinguts a partir de CLUES per a la variant rs204991 dins de la regió candidata *PBX2*.

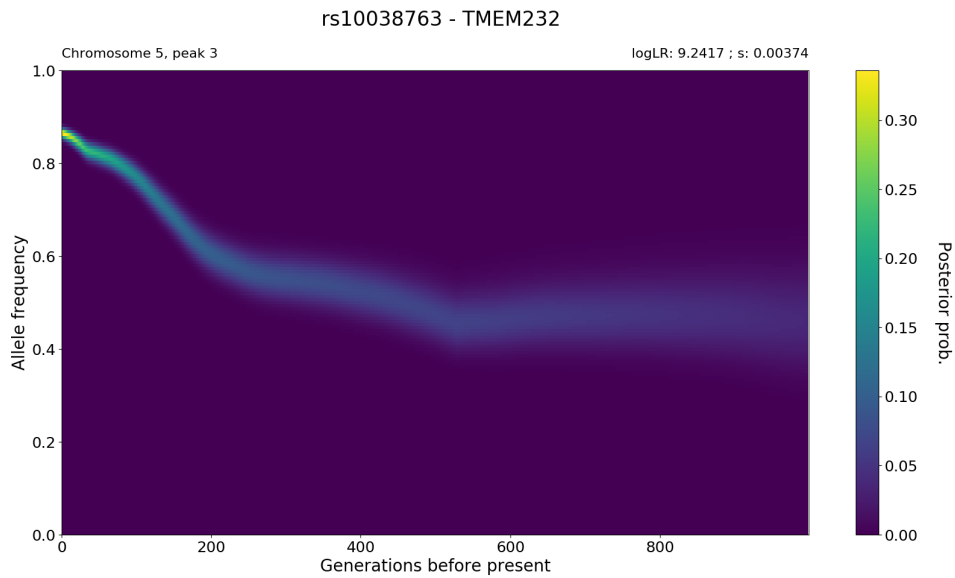


Figura S20. Trajectòria de la freqüència al·lèlica, coeficient (s) de selecció i probabilitat de selecció positiva (logLR) obtinguts a partir de CLUES per a la variant rs10038763 dins de la regió candidata *TMEM232*.

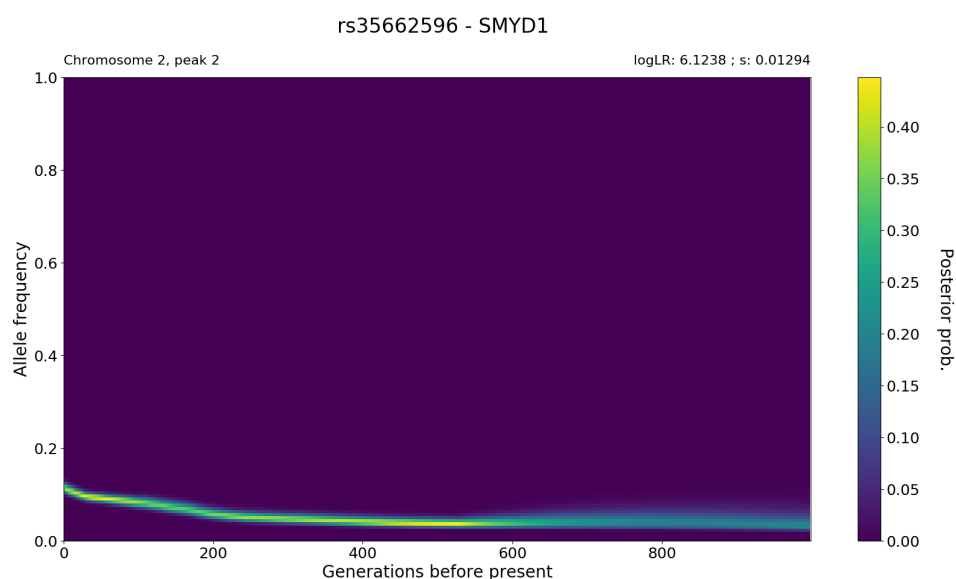


Figura S21. Trajectòria de la freqüència al·lèlica, coeficient (s) de selecció i probabilitat de selecció positiva (logLR) obtinguts a partir de CLUES per a la variant rs35662596 dins de la regió candidata *SMYD1*.

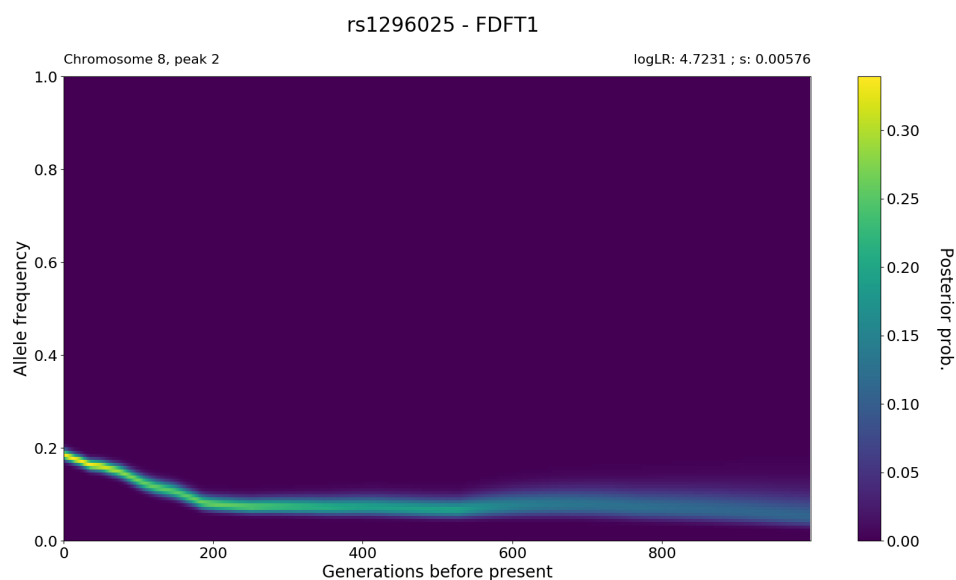


Figura S22. Trajectòria de la freqüència al·lèlica, coeficient (s) de selecció i probabilitat de selecció positiva (logLR) obtinguts a partir de CLUES per a la variant rs1296025 dins de la regió candidata *FDFT1*.

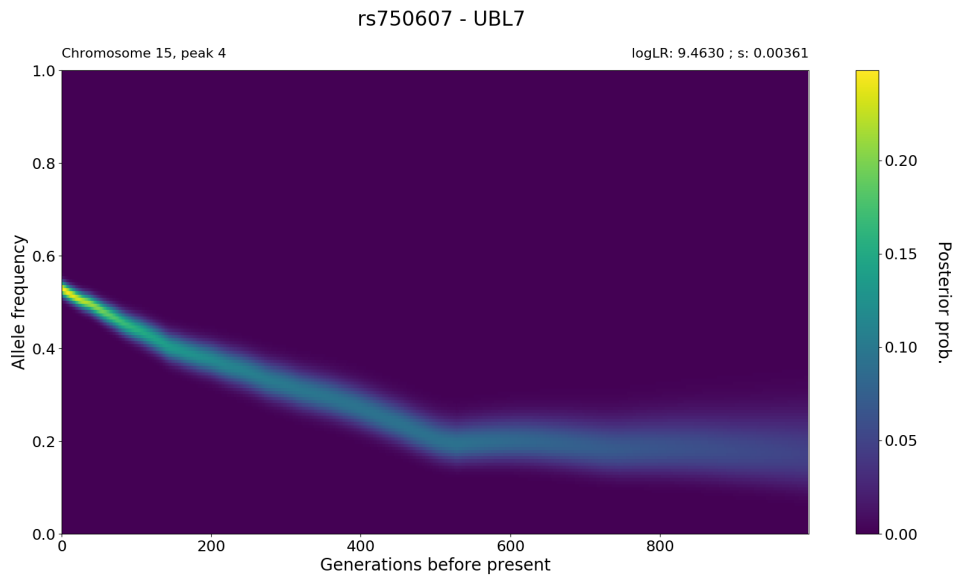


Figura S23. Trajectòria de la freqüència al·lèlica, coeficient (s) de selecció i probabilitat de selecció positiva (logLR) obtinguts a partir de CLUES per a la variant rs750607 dins de la regió candidata *UBL7*.

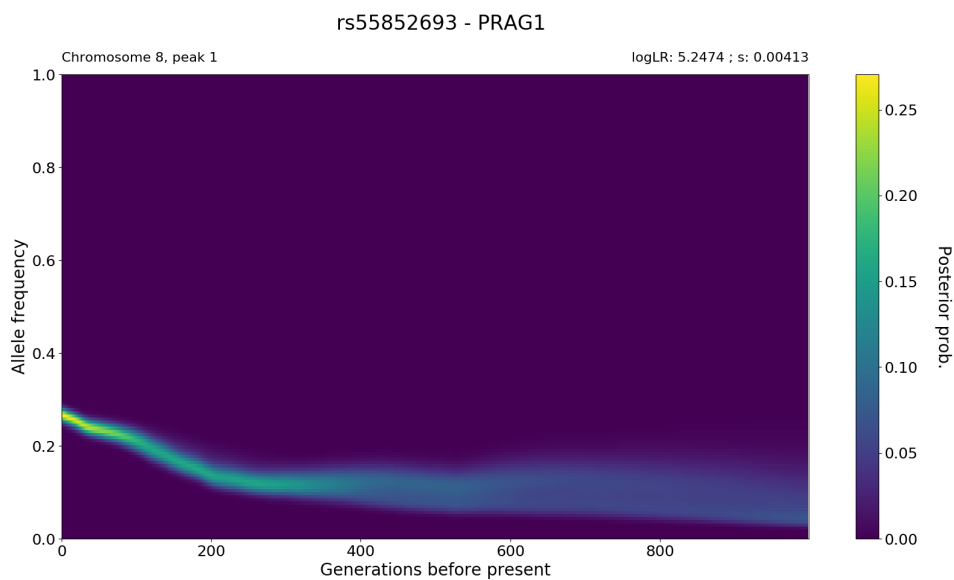


Figura S24. Trajectòria de la freqüència al·lèlica, coeficient (s) de selecció i probabilitat de selecció positiva (logLR) obtinguts a partir de CLUES per a la variant rs55852693 dins de la regió candidata *PRAG1*.

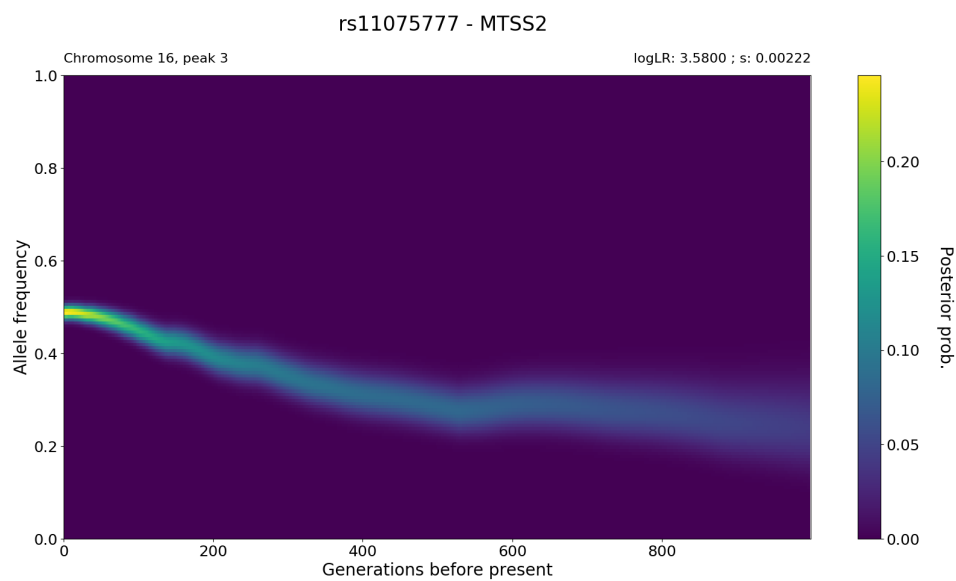


Figura S25. Trajectòria de la freqüència al·lèlica, coeficient (s) de selecció i probabilitat de selecció positiva (logLR) obtinguts a partir de CLUES per a la variant rs11075777 dins de la regió candidata *MTSS2*.

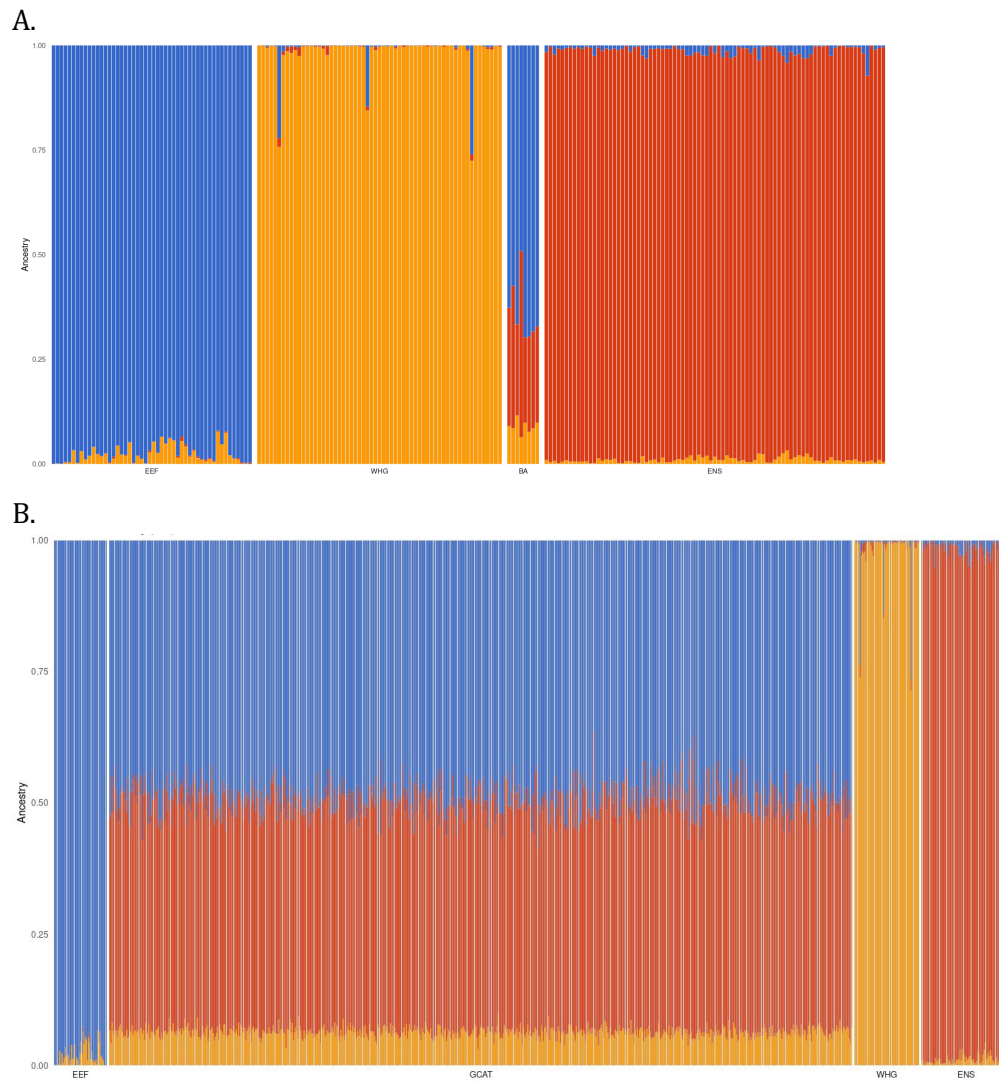


Figura S26. Dades GCAT en el context de dades antigues. A. Proporcions d'ancestría antiga global obtingudes a partir de RFmix mitjançant mostres imputades i fasejades usant genomes post-glacials de l'Euràsia occidental disponibles públicament (EEF, agricultors eurasiàtics primerencs; WHG, caçadors-recol·lectors occidentals; BA, Edat del Bronze; i ENS, nòmades esteparis primerencs)⁴. B. Proporcions globals d'ancestría antiga obtingudes amb RFmix al conjunt de dades GCAT utilitzant els tres components antics principals; BA no es va incloure com a referència ja que, com es veu al panell A, ells mateixos són una població mixta.

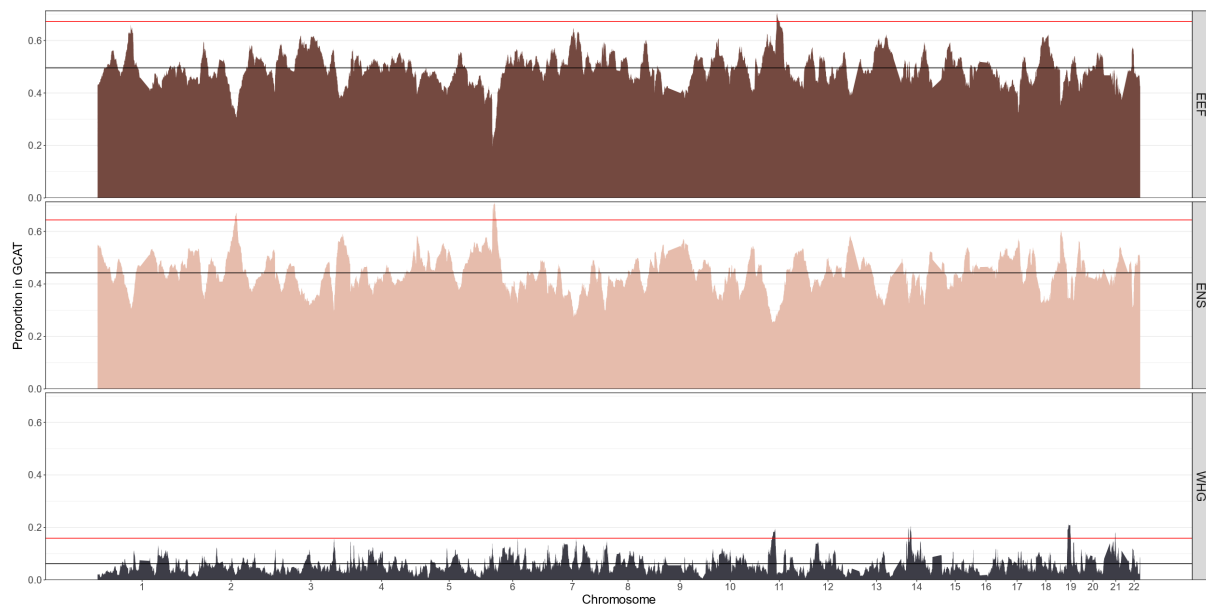


Figura S27. Inferència de components d'ascendència externa i cerca de desviacions

locals d'ancestria (LAD). Proporcions d'ancestria antiga al llarg del genoma obtingudes a

partir de RFmix utilitzant mostres imputades i fasejades usant genomes post-glacials de

l'Euràsia occidental disponibles públicament (EEF, agricultors eurasiàtics primerencs; WHG,

caçadors-recol·lectors occidentals; BA, Edat del Bronze; i ENS, nòmades esteparis primerencs)⁴.

Les línies negres mostren la proporció mitjana de cada ascendència. Les línies vermelles

indiquen tres desviacions estàndard per sobre de la mitjana genòmica. Les regions per sobre de

la línia es van considerar desviacions locals. Vegeu els detalls a la Taula S10.

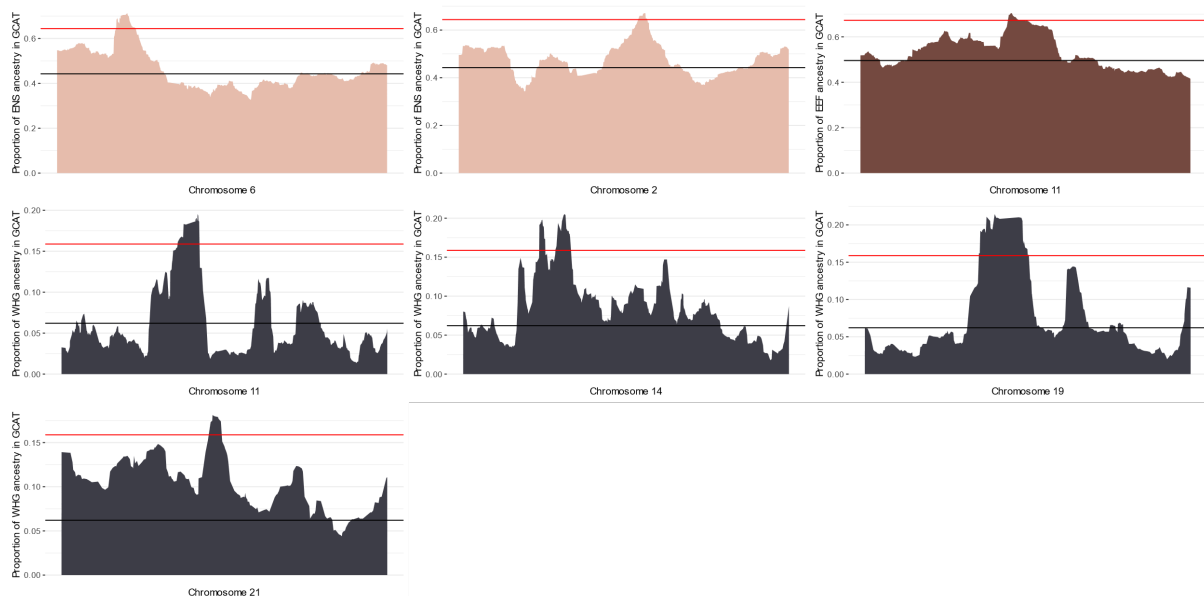


Figura S28. Regions enriquides per a components específics d'ancestria: detalls de totes les regions desviades obtingudes de RFmix utilitzant mostres imputades i fasejades usant genomes post-glacials de l'Euràsia occidental disponibles públicament (EEF, agricultors eurasiàtics primerencs; WHG, caçadors-recol·lectors occidentals; BA, Edat del Bronze; i ENS, nòmades esteparis primerencs)⁴. Les línies negres mostren la proporció mitjana de cada ascendència. Les línies vermelles indiquen tres desviacions estàndard per sobre de la mitjana genòmica.

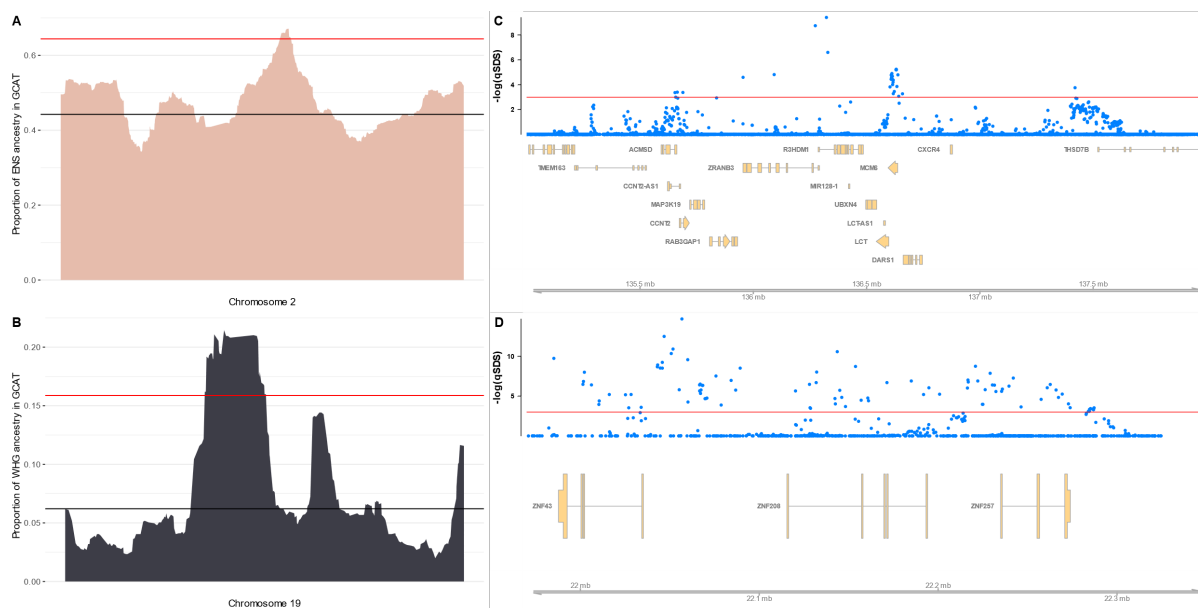


Figura S29. Selecció posterior a la mescla en gens *LCT* i *zinc finger*. Proporcions ancestrals a les dades del GCAT per a A. l'ascendència ENS a la regió LAD que se superposa a la regió *LCT-MCM6* al cromosoma 2 i B. l'ascendència WHG a la regió LAD que superposa un grup de gens de proteïnes *zinc finger* al cromosoma 19. La línia negra mostra la proporció mitjana de cada ascendència. La línia vermella indica tres desviacions estàndard per sobre de la mitjana genòmica. C. Gràfic que mostra els valors de SDS transformats després de la correcció FDR i els gens corresponents subjacents a la regió LAD que es mostra a A. D. Gràfic que mostra els valors SDS transformats després de la correcció FDR i els gens candidats subjacents a la regió LAD que es mostra a B. La línia vermella mostra el llindar de significació amb $\alpha=0,05$. Vegeu els detalls a la Taula S10.

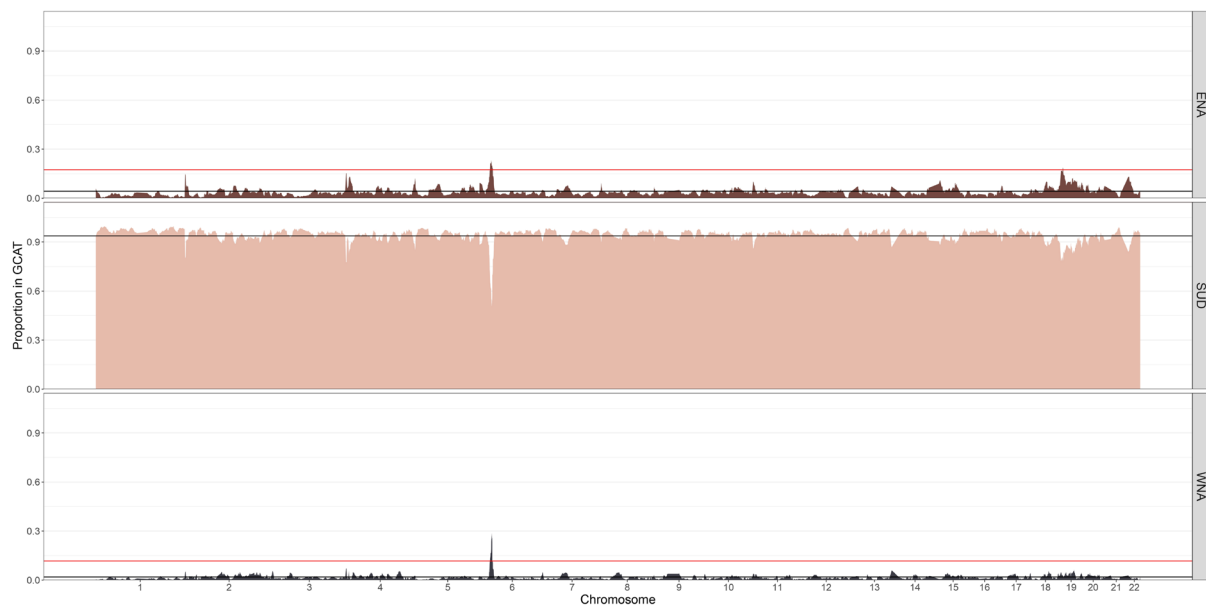


Figura S30. Inferència de components d'ancestria moderns externs i cerca de desviacions locals d'ancestria (LAD). Proporcions d'ancestria modernes a través dels cromosomes obtingudes a partir de RFmix mitjançant mostres fasejades de poblacions veïnes^{8,9} (vegeu els detalls a la Taula S1). Les línies negres mostren la proporció mitjana de cada ancestria. Les línies vermelles indiquen 4,42 desviacions estàndard per sobre de la mitjana genòmica. Les regions per sobre de la línia es consideren significativament diferents de la mitjana genòmica.

Referències

1. Byrska-Bishop, M. *et al.* High-coverage whole-genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios. *Cell* **185**, 3426-3440.e19 (2022).
2. Valls-Margarit, J. *et al.* GCAT|Panel, a comprehensive structural variant haplotype map of the Iberian population from high-coverage whole-genome sequencing. *Nucleic Acids Res* **50**, 2464–2479 (2022).
3. Biagini, S. A. *et al.* People from Ibiza: an unexpected isolate in the Western Mediterranean. *Eur J Hum Genet* **27**, 941–951 (2019).
4. Allentoft, M. E. *et al.* Population genomics of post-glacial western Eurasia. *Nature* **625**, 301–311 (2024).
5. Bergström, A. *et al.* Insights into human genetic variation and population history from 929 diverse genomes. *Science* **367**, eaay5012 (2020).
6. Bycroft, C. *et al.* Patterns of genetic differentiation and the footprints of historical migrations in the Iberian Peninsula. *Nat Commun* **10**, 551 (2019).
7. Hernández, C. L. *et al.* Human Genomic Diversity Where the Mediterranean Joins the Atlantic. *Mol Biol Evol* **37**, 1041–1055 (2020).
8. Biagini, S. A., Ramos-Luis, E., Comas, D. & Calafell, F. The place of metropolitan France in the European genomic landscape. *Hum Genet* **139**, 1091–1105 (2020).
9. Lazaridis, I. *et al.* Genomic insights into the origin of farming in the ancient Near East. *Nature* **536**, 419–424 (2016).