

EURÓPAI ROMÁK GENETIKAI KAPCSOLATA KAUKÁZUSI ETNIKUMOKKAL, VÁNDORLÁSUK ÉS KEVEREDÉSÜK

XXII. OTDK Biológiai Szekció

készítette:

SZABÓ TITANILLA

Biológia MSc hallgató

témavezető:

Prof. Dr. MELEGH BÉLA

PTE ÁOK Orvosi Genetikai Intézet

konzulens:

Dr. HOFFMANN GYULA

PTE TTK Biológiai Intézet

Állatszervezettani és Fejlődésbiológiai Tanszék



Pécs, 2014

Tartalomjegyzék

1.	Bevezetés.....	3
2.	Célkitűzés	7
3.	Anyagok és módszerek	8
3.1.	Az SNP array-től a genotípusadatokig	8
3.2.	Programok	9
3.3.	Felhasznált adatkészletek	11
3.4.	Az analízisekhez összeállított adatkészletek.....	12
4.	Eredmények és értékelésük	13
4.1.	A romák világviszonylatban elfoglalt helye, kapcsolatuk az európaiakkal.....	13
4.2.	A romák kapcsolata a migrációs útvonalukba eső populációkkal	17
4.3.	Összegző analízisek.....	20
5.	Összefoglalás	24

1. Bevezetés

A roma kisebbség egy rendkívül diverz etnikai csoport. Nemzetállammal nem rendelkeznek, különböző nyelveket beszélnek, illetve számos, egymástól eltérő vallást gyakorolnak. Különböző kulturális csoportjaik egymástól eltérő, szigorú szabályok szerint élnek. Antropológiai és nyelvészeti tanulmányok bizonyítják a nagyfokú hasonlóságot a különböző indiai csoportok és a romák között. Számos örmény, görög, illetve perzsa kifejezést találunk a különböző roma dialektusokban, mely tükrözi vándorlásuk érintett területeit, illetve nyelvükben felfedezhető a közös szanszkrit és hindi eredet. A romák társadalmi szerveződése is rendkívül hasonló az indiaiak kasztrendszeréhez, ahol gyakran a szakmák határozzák meg az egyes emberek társadalmi ranglétrán elfoglalt helyét.

Genetikai kutatások és történelmi adatok is alátámasztják, hogy a romák tömeges vándorlása Európa felé körülbelül 1500 évvel ezelőtt India észak-nyugati részéről, Punjab tartományból és környékéről indult meg Nyugat-Perzsia felé ¹. A romák 8-10 millió fős csoportja Indiából elindulva és keveredve a migrációs útvonalukba eső népekkel megváltoztatta az érintett területek demográfiai jellemzőit. A vándorlás során populációjuk egyre jobban töredezett, egyedi genetikai profiljuk alakult ki ^{2,3}. Sokszínű, és egyedi populációjuk a migráció során eljutott a közel-keletre, a kaukázusi térségbe és Európába. Érintették többek között a mai Irán területét, Örményországot, Törökország területét és Görögországot.

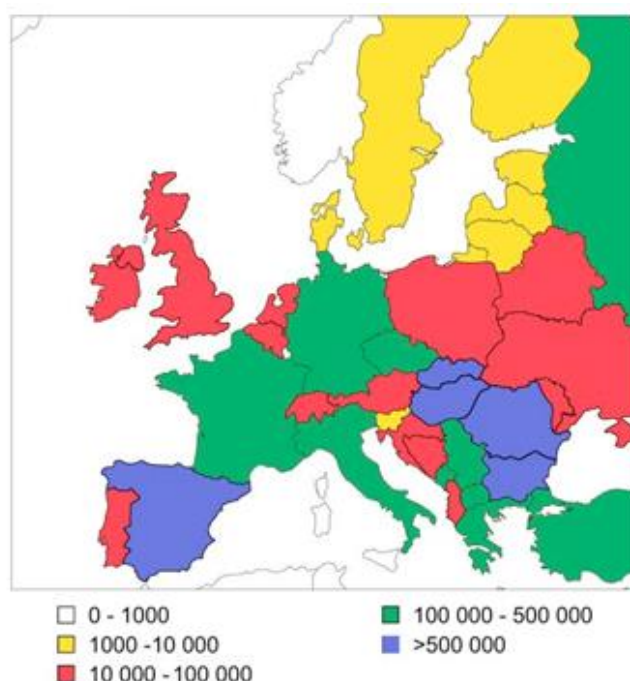
A romák becslések szerint 11 millióan élnek Európában, így ők alkotják a kontinens legnagyobb etnikai csoportját (**1. ábra**). A romaság genetikájának kutatása még rengeteg felfedeznivalót rejt magában ⁴.

1.1. A roma populáció történelme és vándorlása

A romaság kivándorlása Indiából 1000-1500 évvel ezelőtt vette kezdetét. Vándorlásuk három nagy migrációs hullámban történt meg. Az V-XI. század folyamán több roma csoport Észak-Irán területére vándorolt. A VII-XII. században azonban ezek a mai Kelet-Törökország területére migráltak. Egy másik csoportjuk Észak-Afrika illetve a Közel-Kelet felé vette útját. Az arab romaság egy része még ma is nomád életmódot folytat, velük főleg Egyiptomban, Izraelben és Irak területén találkozhatunk. A Bizánci Birodalomba a IX-X. században érkeztek meg. A perzsa területekről nyugat felé vándorlásuk a XIII. században főleg a mongol hódítás miatt kezdődött meg ^{5,6}. A Balkánra való betelepülésük a XIII-XIV. században kezdődött és végső betelepedésük Európába a XV. századra tehető. Legnagyobb arányban a

balkáni térség ad otthont az európai romaságnak. A balkáni roma telepések a Dunától délre telepedtek le az Oszmán Birodalomban. Az oláh romaság folytatta vándorlását a Dunától északra a havasalföldi fejedelemségbe, ahol a XIV. században a romákat rabszolgasorba taszították.

A romák migrációjában a rabszolgaság is nagy szerepet játszott. A XVI. század eleji török hódítás során Havasalföld engedélyezte az Oszmán Birodalomban élő romák ideiglenes bevándorlását területükre. A XIX. század végén illetve XX. század elején a rabszolgaság eltörlésével az oláh románok elvándoroltak Európa újabb területei felé. A legutóbbi migrációs hullám az 1990-es években a jugoszláviai háborúban történt, mely sok társadalmi és gazdasági változást vonzott maga után ^{2,7}.



1. ábra. Roma populáció megoszlása Európa országaiban

Mendizabal és munkatársai a romák indiai eredetének feltárását tűzték ki célul. Korábbi genetikai, antropológiai és nyelvészeti tanulmányok alapján megállapították, hogy az európai közösséget alkotó roma populáció az indiai szubkontinens területére vezethető vissza. Kutatásuk alátámasztotta, hogy a roma őshaza India. A vizsgálatok során 152 roma 800,000 SNP -jét (single nucleotide polymorphisms) vizsgálták és hasonlították össze európai, indiai, közép-ázsiai, illetve közel-keleti mintákkal. Az elemzés során kimutatták, hogy a romák alapító csoportja az indiai Punjab régióból indult meg és vándorolt Közép-Ázsia és a Közel-Kelet területeire és keveredett az ottani lakossággal ⁸.

2013-ban Moorjani és munkatársai meghatározták a romák indiai szubkontinensről való származását, illetve azonosították a roma népcsoportra jellemző genetikai struktúrák kialakításában szerepet játszó eurázsiai régiókat. Antropológiai vizsgálatok kapcsolatot feltételeztek a romák és az Indiában tartózkodó banjarai nomád roma csoport között. A roma történelem felderítéséhez szükséges genetikai következtetések fókuszában az Y-kromoszóma illetve a mitokondriális DNS tanulmányozása állt. Vizsgálataikból kiderült, hogy a romák ősei észak-nyugat indiai népcsoportok és kelet-európai populációk. Keveredésük megközelítőleg 870 évvel ezelőtt kezdődött ¹. Indiában ma is számos olyan népcsoport él, melyek a romákkal nagyfokú hasonlóságot mutatnak.

1.2. Teljes genomra kiterjedő populációgenetikai vizsgálatok

A Genome Wide Association Study (GWAS) megjelenése a humángenetika területén óriási jelentőségű volt. A teljes genom vizsgálatán alapuló módszerek egyre nagyobb teret nyernek az orvostudományban és a természettudományokban, így a populációgenetikában is. Amellett, hogy lehetővé teszi populációkat érintő komplex betegségek genetikai hátterének vizsgálatát, populációk, népcsoportok eredetének, történelmének teljes genomi analízisén alapuló felderítése szintén elérhetővé válik ⁹. A populációk közötti genetikai kapcsolatok feltárásában nagy segítséget nyújtanak a különféle adatbázisok, mint a Human Genome Diversity Project (HGDP), a Haplotype Map (HapMap) Project adatbázisai, illetve az egyesült államokbeli National Human Genome Research Institute (NHGRI) Population Reference Samples (POPRES) nevű adatbázisa ¹⁰.

A Stanford Egyetem által elindított Human Genome Diversity Project a párizsi Center for the Study of Human Polymorphisms (CEPH) kutatóközponttal kollaborációban a humán populációk genetikai diverzitásának feltérképezését célozta meg. Ennek eredményeként létrejött a HGDP-CEPH Human Genome Diversity Cell Line Panel marker adatbázis. 52 populációból származó, összesen 1043 mintát gyűjtöttek össze, és genotipizálták SNP array chipen. Az adatbázis Afrika, Európa, Ázsia, Óceánia és Amerika legkülönbözőbb populációit foglalja magába ¹¹⁻¹⁴. Az adatbázist a Stanford Egyetem bárki számára szabadon elérhetővé tette.

A nemzetközi HapMap projekt célja a humán genom jellemző haplotípusainak feltérképezése, mely lehetővé teszi az egyes populációk gyakori genetikai variációinak leírását, jellemzését. A HapMap projektet 3 fázisban valósították meg, az adatbázis elkészítése óta a kóroki variánsok azonosításának fontos eszközévé vált. Elterjedését nagyban elősegítette, hogy a

HGDP adatbázisához hasonlóan az International HapMap Consortium a HapMap adatokat honlapján szabadon letölthetővé tette ¹⁵.

2. Célkitűzés

A romák történelmük során sokáig nomád életmódot folytattak, melynek eredményeként összetett genetikai struktúrára tettek szert. Számos tanulmány foglalkozott a romák származásával, illetve genetikai összetételük kialakulásában szerepet játszó főbb populációk azonosításával. Migrációjuk során nagyszámú populációval kerültek kontaktusba, melynek során a romák és a velük kapcsolatba került népcsoportok populáció-interakciók sokaságának lehettek kitéve. E populációs kölcsönhatások vizsgálatával a tudományos élet ezidáig részletesen nem foglalkozott. A roma populáció Európa népességének jelentős részét képezi, így eredetük, vándorlásuk, illetve más populációkkal való kapcsolataik megértése fontos része a tudománynak, genetikai hátterük megismerése érdekében.

Kutatásaink során a romák és a migrációs útvonalukra eső populációkkal való kapcsolatát vizsgáltuk. Arra kerestük a választ, hogy a romák jelenlegi genetikai struktúrájának kialakulásában az egyes dél-ázsiai, közel-keleti, kaukázusi és európai populációk, népcsoportok milyen mértékben és módon vettek részt.

3. Anyagok és módszerek

3.1. Az SNP array-től a genotípusadatokig

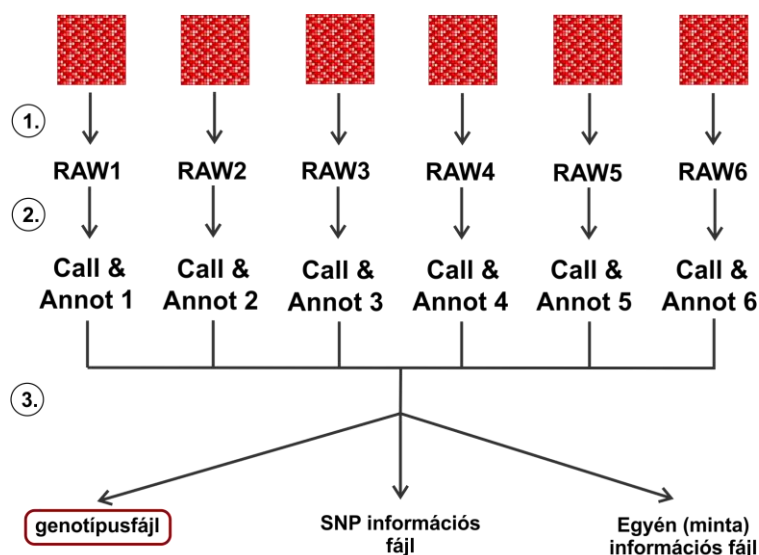
Az analízisekben használt populációk DNS mintái teljes genomra kiterjedő Affymetrix és Illumina SNP array platformokon kerültek meghatározásra.

A mintákkal hibridizált SNP chipeket a microarray scanner beolvassa, SNP-próbánkénti intenzitás- és konfidenciaértékeket egy szöveges vagy bináris fájlba írja. Ez az úgynevezett nyers (raw) fájl. A nyers fájl egy, a platform gyártója által fejlesztett genotípushívó algoritmus feldolgozza, amelynek eredményeként az intenzitásértékekből genotípushívásokat (genotype calls) tartalmazó fájlkat hoz létre. Affymetrix Genome-Wide Human SNP 6.0 chip esetében ilyen genotípushívó algoritmus a Birdseed, amelyet az Affymetrix Power Tools szoftvercsomag (APT) tartalmaz. Az Affymetrix és az Illumina platformjai egyjelöléses rendszert használnak, tehát a chipen lévő próbákkal referencia DNS-t nem hibridizálunk. Ezért a genotípushívó algoritmusnak szüksége van egy referenciákat tartalmazó könyvtárfájltra a hívások során. A genotípushívásokat tartalmazó fájl kezdetben csak olyan információt tartalmaz, hogy az egyes próbákhoz referencia vagy variáns allél kapcsolódott-e, illetve hogy az adott SNP-re nézve a minta homo- vagy heterozigóta-e. Ezért az APT szoftvercsomag segítségével egy annotációs könyvtárfájlból hozzáadhatjuk az egyes genotípusadatokhoz az egyes SNP-k adatait, úgy mint referencia-SNP azonosító (rs ID), a konkrét allélok, az SNP kromoszómális és fizikai pozíciója.

Ezután egy konvertáló program segítségével, a genotípushívások tartalmát 3 különböző funkciójú fájlba, adatkészletbe helyezzük át. Az első fájl tartalmazza a genotípusadatokat, a másik két fájl pedig a genotípusadatok azonosítására szolgál. A második fájl a genotípusfájlból szereplő SNP-k azonosítására és leírására szolgál. Az Affymetrix SNP 6.0 chip esetében egy SNP chip mindössze egyetlen egyén közel 1 millió SNP-jét tartalmazza. Ezzel szemben egy genotípusfájlból számos (gyakorlatilag bármennyi) minta genotípusadatait tartalmazhatja, ezért szükséges egy további, harmadik fájl is, amely azonosítja, hogy az egyes genotípusadatok mely mintához tartoznak, illetve leírja a minták releváns tulajdonságait (**2. ábra**).

Laboratóriumunk SNP array-t jelenleg nem végez, az általunk begyűjtött DNS-minták külső partnerek laboratóriumaiban kerültek SNP chipre és szkennelésre. Intézetünk az ezt követően kapott nyers adatokkal, a fent leírtak szerint dolgozott tovább. Az esetek másik részében,

például globális populációmintákat tartalmazó adatbázisok esetében a genotípusadatokat készen kaptuk vagy töltöttük le online forrásból szabványos formátumban. Ilyen formátum a széles körben elterjedt PLINK formátum, amelyet a későbbiekben ismertetésre kerülő PLINK szoftvercsomaggal hozhatunk létre.



2. ábra. Az SNP chip szkennelésétől az elkészült genotípusadatokig vezető munkafolyamat sematikus rajza. 1) A microarray scanner létrehozza az intenzitásértékeket tartalmazó nyers (RAW) fájlokat. 2) A genotípushívó algoritmus ezekből elkészíti a genotípushívásokat (Call), ehhez pedig hozzáadjuk az SNP annotációt (Annot). 3) Ezután egy megfelelő program segítségével konvertálhatjuk át adatainkat az analízisekhez szükséges formátumra, amely három, különböző funkciójú fájlból áll. A programok a genotípusokat tartalmazó fájlokkal dolgoznak, a másik két fájl információs fájlként szolgál.

3.2. Programok

Analíziseinknél kétféle, gyökeresen eltérő módszert alkalmazó leszármazásbecslő szoftvert használtunk. Az egyik a Harvard Medical School, az MIT és Broad Institute közös együttműködéseként megszületett EIGENSOFT szoftvercsomag SMARTPCA¹⁶ elnevezésű, a leszármazás-becslést algoritmusos alapon megvalósító programja. A második az ADMIXTURE¹⁷, amely egy modell alapú leszármazás-becslő program. Adataink szerkesztéséhez a PLINK¹⁸ programcsomagot és az EIGENSOFT programcsomag CONVERTF és MERGEIT szoftvereit alkalmaztuk.

A SMARTPCA populációk egyedszintű struktúrájára, a populációk egymás közötti rokonsági, leszármazási kapcsolatainak felderítésére használható leszármazás-becslő szoftver. Ezt, szemben több modellen alapuló módszerrel, algoritmusos alapon végzi, amelynek során a

főkomponens-analízis lineáris algebrai módszerét alkalmazza. A genotípus-adatokból statisztikai módszerek segítségével négyzetes mátrixot képezve meghatározza annak sajátvektorait, azaz főkomponenseit, illetve a hozzájuk tartozó sajátértékeket, és kiszámolja a sajátértékek által meghatározott új adathalmaz adatainak koordinátáit. Az egyes sajátvektorok statisztikai szignifikanciáját is megadja Tracy-Widom statisztika alkalmazásával, kiszámolja az egyes sajátvektorokhoz tartalmazó sajátértékek valószínűség-eloszlását. A főkomponens-analízis lehetővé teszi, hogy egy sokdimenziós adathalmazban mintázatot, adatok közti összefüggéseket fedezzünk fel azáltal, hogy a módszer a sokdimenziós adathalmazt minimális információvesztéssel néhány dimenzióra transzformálja, ahol a dimenziókat a kiszámított főkomponensek képezik. A program segítségével tehát egy többszázezer SNP genotipizálásának eredményeként létrejött adathalmazban populációstruktúrát, populációk közötti leszármazási kapcsolatokra utaló mintázatokat találhatunk. A SMARTPCA-t be tudjuk állítani, hogy hány főkomponenst határozzon meg, amelyeket a sajátértékük nagysága szerint csökkenő sorrendbe rak. Minél kisebb egy főkomponens sajátértéke, az adatok annál kisebb strukturáltságot mutatnak. Analíziseinkben az adatokat mindig a két legnagyobb főkomponensen, kétdimenzióban ábrázoltuk. A SMARTPCA a vizsgált populációk páronkénti allélfrekvencia-differenciációk átlagértékeinek kiszámítására is használható, amelynek eredményeként egy F_{ST} mátrixot kapunk. Ennek mérete értelemszerűen az analízisbe bevont populációk számától, az adatkészlet méretétől függ.

Az ADMIXTURE a SMARTPCA-hoz hasonlóan egyes populációk struktúrájának és azok leszármazási kapcsolatainak vizsgálatára használható. A SMARTPCA-vel ellentétben az ADMIXTURE a legnagyobb valószínűség elvét (maximum likelihood estimation) felhasználva egy statisztikai modell alapú klaszterezési vizsgálati módszerrel kivitelezte ezt, amely az adatok két dimenzióra való kivetítése helyett közvetlenül ősi klasztereket próbál meghatározni az adatokban.

Az ADMIXTURE analízishez meg kell becsülnünk K értékét, azaz a vizsgált populációk hipotetikus őseinek számát. Az ADMIXTURE analízis az egyes vizsgált csoportokat, mint K számú hipotetikus ősi csoport elegyét jeleníti meg. A legmegfelelőbb K értéket egy keresztvalidációs algoritmus beiktatásának segítségével tudjuk meghatározni. Így a vélt K érték helyességét ellenőrizni tudjuk, illetve előzőleg nem vizsgált, ismeretlen adatokhoz tartozó K értéket is meg tudunk becsülni a programmal.

A továbbiakban leírásra kerülő, általunk összeállított adatkészleteket az EIGENSOFT szoftvercsomag CONVERTF és MERGEIT programjaival, illetve a PLINK programcsomaggal hoztuk létre. A programokkal különböző genotípusadat-formátumok

között konvertálhatjuk adatainkat, adatkészleteket vonhatunk össze egyetlen adatkészletté, illetve adatainkat szűrhetjük SNP-k, minták és különböző genetikai szempontok szerint. A genotípusadatok esetleges konvertálására azért van szükség, mert az EIGENSTRAT szofvercsomag, illetve a PLINK saját formátumot használ. A CONVERTF és a MERGEIT egyszerű konvertáló- illetve adatkészletek összevonására szolgáló programok, amelyek az adatok ellenőrzésén túl azok különböző szempontú, korlátozott szűrésére szintén használhatók. Ezzel szemben a PLINK egy teljes-genomra kiterjedő adatok menedzselésére és analízisére szolgáló komplett szofvercsomag. Kutatásunkban a PLINK szofvercsomagot adatkészletek létrehozására, illetve adatkészletek SNP szűrésére egyaránt használtuk.

3.3. Felhasznált adatkészletek

Munkánk során a romák és a kaukázusi népek kapcsolatának felderítéséhez több adatkészlet egyes populációit is felhasználtunk. A nemzetközi HapMap projekt eredményeként elkészült adatkészlet mellett a Human Genome Diversity Cell Line Panel adatkészletéből is szerepelnek különböző közel-keleti, kaukázusi és európai populációk az analíziseinkben^{11,12}. A POPRES adatkészletből szintén felhasználtunk néhány indiai és európai populációt¹⁰. A HapMap adatok (n=1184 minta 11 populáció Affymetrix 1M chipen genotipizálva, 613218 SNP) minden populációját felhasználtuk az analízishez. A HGDP adatokból (n=1043 minta 52 populáció Illumina 650Y chipen genotipizálva, 660918 SNP) egy kaukázusi csoportot (adige), 8 közel keleti csoportot (beludz, brahui, burusó, hazara, makrani, pastu, sindhi, ujugur) és 3 európai populációt (baszk, francia, olasz) használtunk fel. A POPRES adatokból (n=4077 minta 57 populáció Affymetrix 500k SNP chipen genotipizálva, 453617 SNP) 4 indiai csoportot (gujarati, hindi, punjabi, urdu) és egy európai populációt (német) használtunk fel.

A vizsgálatban résztvevő kaukázusi populációk, a tádzsikok és a törökök bárki számára elérhető adatbázisból származnak (n=204 minta 14 populáció Illumina 610K SNP chipen genotipizálva, 555767 SNP)¹⁹. Az indiai populációk másik részét a bostoni Harvard Medical School bocsátotta rendelkezésünkre (n=121 minta 23 etnikai csoport Affymetrix SNP 6.0 chipen genotipizálva, 524053 SNP)²⁰. Az európai populációk között magyar minták is szerepelnek, amelyet a Debreceni Egyetem Orvostudományi karának jóvoltából használhattunk fel (n=299 minta, Affymetrix SNP 6.0 chipen genotipizálva, 898723 SNP).

A munka középpontjában álló romák adatkészlete két forrásból származik. Az adatkészlet egy része saját, közép-európai roma mintákat tartalmaz, amelyek a Harvard Medical School-lal kollaborációban kerültek genotipizálásra (n=27 minta Affymetrix 1M SNP chipen

genotipizálva, 726 404 SNP). A roma adatkészlet másik részét a rotterdami Erasmus Egyetem Igazságügyi Molekuláris Biológiai Intézetének munkatársaitól kaptuk, nyugat- és közép-európai roma mintákat egyaránt tartalmaz (n=152 Affymetrix SNP 6.0 chipen genotipizálva, 868165 SNP)⁸.

3.4. Az analízisekhez összeállított adatkészletek

Analíziseinkhez több adatkészletet állítottunk össze az EIGENSOFT szoftvercsomag MERGEIT programjának segítségével. A vizsgálatok első részében a romák globális elhelyezkedésének felderítése érdekében egy roma és HapMap összevont adatkészletet készítettünk. A romák kaukázusi népekkel való kapcsolatának vizsgálatához roma, kaukázusi, indiai, európai és közel-keleti populációkat válogattunk össze az ismert adatokból. Ezeket felhasználva készítettük el a különböző egyesített adatkészleteket. Mivel a vizsgálatokban résztvevő populációk leszármazás szempontjából átmenetet képeznek, illetve közel helyezkednek el egymáshoz, az analízisekhez szükségünk volt egy olyan populációra, amely minden vizsgált populációval megközelítőleg egyformán távoli leszármazástani kapcsolatban áll. Ez az úgynevezett „outgroup” populáció, amely nélkül nehézkes szignifikáns és értelmezhető mintázatot találni a vizsgálni kívánt populációk között. Outgroup populációnak egy HGDP adatkészletben megtalálható távolkeleti populációt, a han kínaiakat választottuk, amely az előzetes klaszterező vizsgálatok alapján megfelelő kívülálló csoportnak bizonyult.

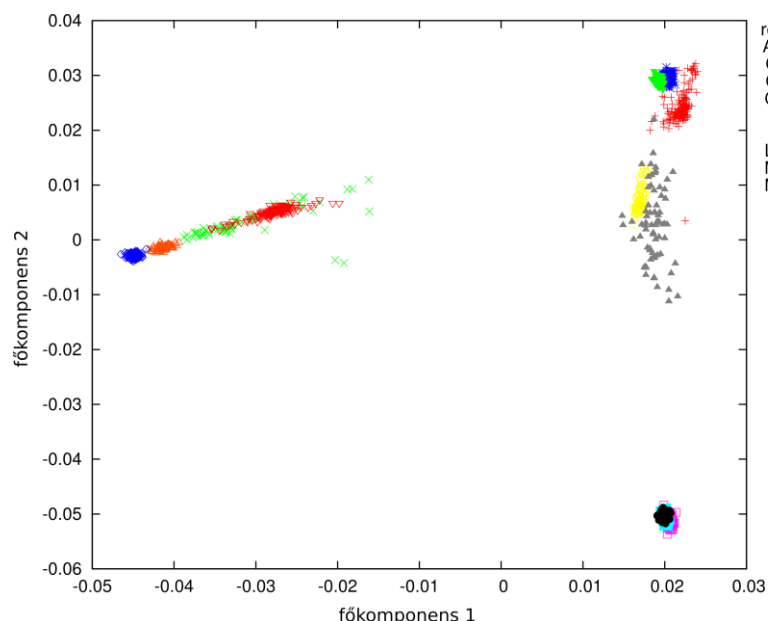
A genetikai sodródás eredményeként létrejövő háttér kapcsoltsági egyenlőtlenséget (background linkage disequilibrium, BLD) a különböző klaszterező algoritmusok nem veszik figyelembe, illetve könnyen összetéveszthető a genetikai keveredés okozta kapcsoltsági egyenlőtlenséggel, ezáltal az általunk használt leszármazásbecslő algoritmusok hibás, eltorzított eredményeket produkálhatnak. Ezért az adatkészletekből minden esetben eltávolítottuk a háttérkapcsoltságban lévő SNP-párok egyik tagját. Az LD-n alapuló SNP eltávolítást a PLINK programcsomaggal végeztük. A PLINK az eltávolítási műveletet tolóablakos megközelítéssel végzi. Az SNP eltávolítás során az SNP-k páronkénti analíziséhez és eltávolításához 50 SNP-re kiterjedő tolóablakméretet és 5 SNP/lépés tolóablakhaladást alkalmaztunk. A r^2 küszöbértéke szakirodalmi adatoknak megfelelően 0,1 volt²¹. Az SNP-eltávolítást követően a roma mintákat és HapMap populációkat tartalmazó adatkészlet 92779, a többi adatkészlet 79524 SNP-t tartalmazott.

4. Eredmények és értékelésük

4.1. A romák világviszonylatban elfoglalt helye, kapcsolatuk az európaiakkal

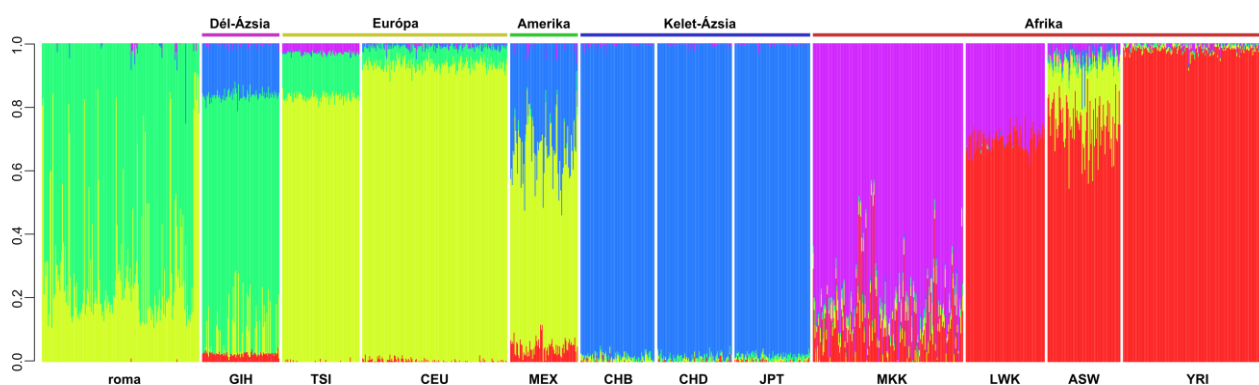
A romák világviszonylatban való elhelyezkedésének vizsgálata céljából elsőként a 11 HapMap populációval vetettük össze őket. A főkomponens-analízis (PCA) eredményét az első két főkomponensen ábrázolva jól látszik, hogy a romák a vártnak megfelelően az európai és a dél-ázsiai populációk közelében klasztereződnek. Minden más populáció a romáktól távolabb helyezkedik el. A romák legerősebben az európai populációkkal vannak leszármazástani kapcsolatban, az európaiakhoz klasztereződnek legközelebb. Ugyanakkor a klasztereződésük meglehetősen tág, amely erőteljes diverzitásra utal. Klasztereződésük az európai populációk, illetve az indiaiak felé irányul **(3. ábra)**. Látszólag a mexikói őslakosokból álló minták is közel helyezkednek el a romákhoz, azonban, ha a második két főkomponensen ábrázoljuk az adatokat, már jól elkülönülnek azoktól. A populációk a világ különböző kontinensein élő populációktól származnak, ilyen mértékű mintadiverzitásnál, már a második két főkomponens is szignifikáns mintázatot képes megjeleníteni az adatokban.

Az ADMIXTURE analízis a populációkat, mint 5 hipotetikus ősi csoport elegyét jeleníti meg ($K=5$). A halmozott oszlopdiagramon jól látszik, hogy a romák esetében az indiai klaszter mellett a nyugat- és dél-európai klaszter is igen erősen képviselteti magát. Az ADMIXTURE ábrán a PCA-hoz hasonlóan megfigyelhető, hogy a roma minták nem konzisztensek, jelentősen eltérő mértékben osztoznak a 2 hipotetikus klaszteren **(4. ábra)**.



3. ábra. A roma és HapMap populációk főkomponens-analízisének (PCA) ábrázolása az első két főkomponensen. Minden egyes pont (szimbólum) a diagramon egyetlen egyént jelöl. Az azonos szimbólummal jelölt egyének egy populációba vagy csoportba tartoznak.

ASW: Afrikai leszármazással rendelkező egyének az Egyesült Államok délnyugati részéről, **CEU:** Észak- és nyugat-európai leszármazással rendelkező egyének a CEPH kollekciónból (Utah, USA), **CHB:** Han kínaiak Pekingből (Kína), **CHD:** Kínaiak Denverből (Colorado, USA), **GIH:** Gujara indiaiak Houstonból (USA), **LWK:** Luhya-k Webuye-ből (Kenya), **JPT:** Japánok Tókióból (Japán), **MEX:** Mexikói őslakos leszármazással rendelkezők Los Angelesből (California, USA), **MKK:** Maasi-k Kinyawa-ből (Kenya), **TSI:** Toszkánok Olaszországból, **YRI:** Jorubák Ibadanból (Nigéria)



4. ábra. A roma és HapMap populációk ADMIXTURE analízisének eredménye. $K=5$, azaz a vizsgált populációk 5 hipotetikus őson osztoznak. A hipotetikus őseket a grafikon oszlopaiban különböző színek jelölik. Minden oszlop az adott populáció vagy csoport egy mintáját (egyén) reprezentálja, tehát a képzeletbeli X tengelyen az egyének szerepelnek, az Y-tengely pedig az egyes egyének genotípusának hipotetikus ősök közötti megoszlását jelenti.

A vizsgált csoportok páronkénti allélfrekvencia-differenciációinak átlagértékeiből táblázatot készítettünk. Ezt röviden F_{ST} mátrixnak nevezzük. A roma és HapMap populációk F_{ST} mátrixa igazolja, hogy a roma és az európai (CEU és TSI) csoportok között világviszonylatban nézve is lényegesen kisebb a populáció-differenciáció, tehát ezen csoportok genetikai struktúrája között nagyobb a hasonlóság, mint a romák és az indiaiak (GIH) esetében (**1. táblázat**).

	roma	ASW	CEU	CHB	CHD	GIH	JPT	LWK	MEX	MKK	TSI	YRI
roma	0.000											
ASW	0.098	0.000										
CEU	0.017	0.102	0.000									
CHB	0.095	0.138	0.110	0.000								
CHD	0.096	0.138	0.111	0.001	0.000							
GIH	0.024	0.093	0.035	0.075	0.076	0.000						
JPT	0.095	0.138	0.110	0.007	0.008	0.076	0.000					
LWK	0.136	0.011	0.143	0.168	0.168	0.128	0.168	0.000				
MEX	0.036	0.098	0.036	0.070	0.071	0.038	0.070	0.134	0.000			
MKK	0.095	0.016	0.102	0.138	0.139	0.092	0.138	0.017	0.098	0.000		
TSI	0.014	0.098	0.005	0.109	0.110	0.034	0.109	0.138	0.036	0.096	0.000	
YRI	0.147	0.011	0.154	0.177	0.177	0.139	0.177	0.009	0.145	0.028	0.150	0.000

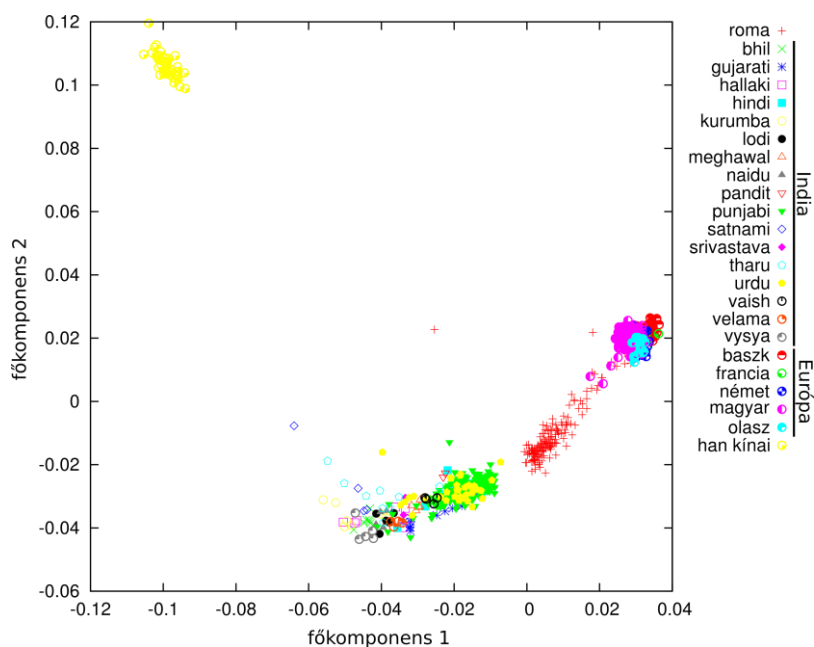
1. táblázat. A páronkénti allélfrekvencia-differenciációk átlagértékeiből készített mátrix (F_{ST} mátrix) a roma és a HapMap populációkat tartalmazó adatkészlet esetében.

A romák, európaiak és indiaiak kapcsolatát eurázsiai szinten is megvizsgáltuk. Ennek során a roma mintákkal különböző, Indiában élő etnikai csoportokat és európai populációmintákat hasonlítottunk össze. Az indiai mintákat úgy válogattuk össze, hogy az egész indiai szubkontinenst lefedjék, az európai minták kiválasztásánál pedig az volt a cél, hogy Európa minden olyan részét reprezentálják populációk, ahol jelentősebb mennyiségben élnek roma csoportok. Így esett a választásunk nyugat-, közép- és dél-európai populációkra. A világviszonylatban történt analízisből is kiderült, hogy az európai, roma és indiai populációk genetikai struktúrája relatíve közel áll egymáshoz. Ezért annak érdekében, hogy a PCA és ADMIXTURE analízis szignifikáns mintázatot legyen képes találni a bevitt adatokban, a vizsgálathoz ezektől a csoportoktól genetikailag távol álló kelet-ázsiai, han kínai populációmintákat is felhasználtunk.

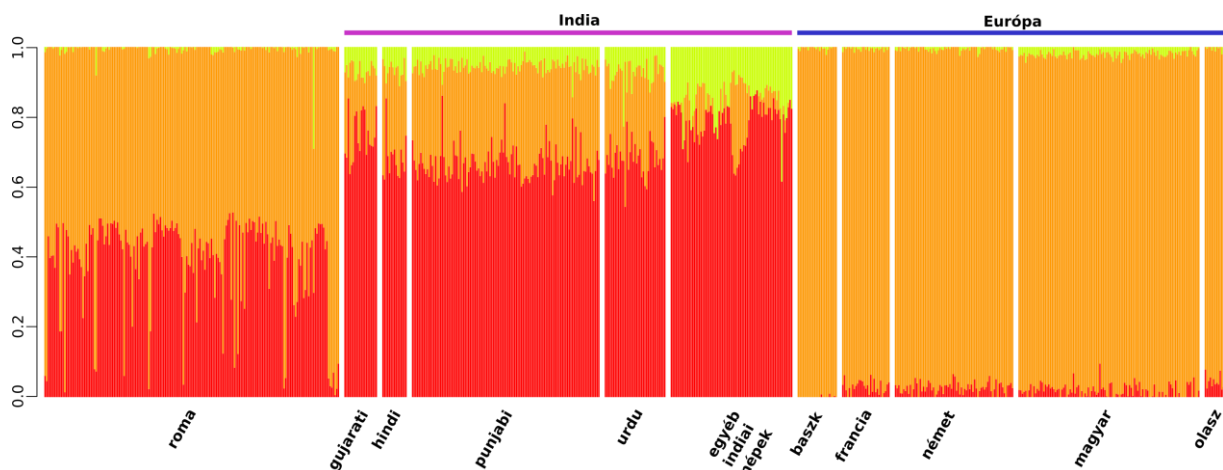
Az így készült PCA analízis megerősítette az eddigi eredményeinket, a romák genetikai jellemzői az európaiak és indiaiak között megközelítőleg egyenlően oszlik meg. Azonban

jelentős számú roma egyén mutat nagyobb hasonlóságot az európai populációkkal, továbbá megfigyelhetünk számos, nem roma európaival extrém módon keveredett roma egyént is, amelyek még erősebben az európai populációk közelébe klasztereződnek **(5. ábra)**. A grafikonon emellett az is megfigyelhető, hogy a romák az India északnyugati részén élő etnikai csoporthoz, a punjabikhoz állnak legközelebb. Ez egybeesik a szakirodalomban leírt eredményekkel, amely szerint a romák megközelítőleg 1500 évvel ezelőtti exodusa a Pakisztánra is áttérjedő Punjab régióból indult el.

Ugyanezen adatkészlet ADMIXTURE grafikonja még látványosabban szemlélteti a három csoport közötti hasonlóságokat, illetve a roma csoport rendkívüli diverzitását, inkonzisztenciáját **(6. ábra)**.



5. ábra. A roma, indiai és európai csoportok főkomponens-analízisének eredménye az első két főkomponensen ábrázolva. Urdu, Hindi: India Uttar Pradesh, Uttarakhand, Haryana, Rajasthan tartományában élő urdu, illetve hindi nyelvet beszélő egyéneinek csoportja.

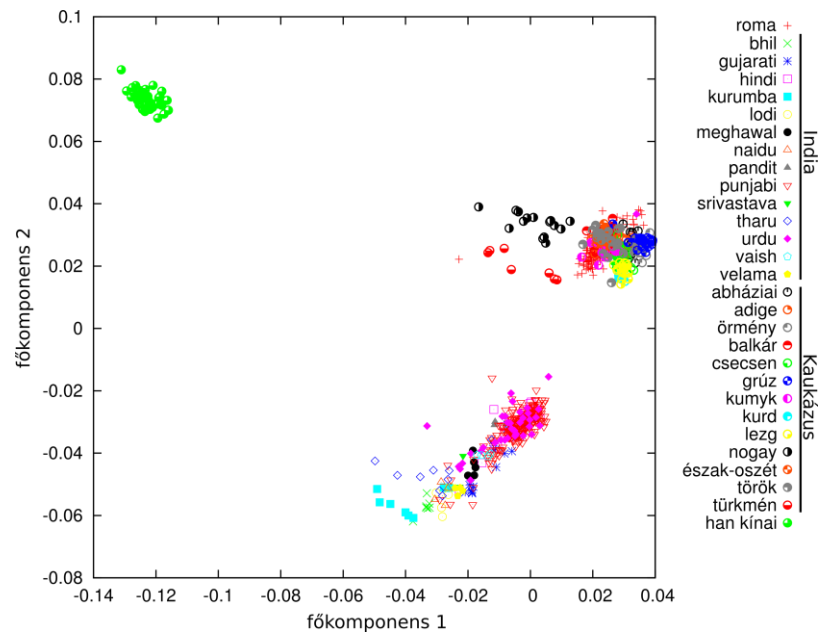


6. ábra. A roma, indiai és európai mintákat tartalmazó adatkészlet ADMIXTURE grafikonja ($K=3$). A vizsgált populációk 3 hipotetikus ősz között osztoznak, amelyek közül piros színnel a dél-ázsiai, narancs színnel az európai, illetve nyugat-ázsiai, sárga színnel pedig a kelet-ázsiai őst jelöltük. A kelet-ázsiai hipotetikus őst itt pontosan a vizsgálatba outgroup-ként bevont han kínaiakat jelenti, amelyet az ADMIXTURE grafikonon nem jelenítettünk meg.

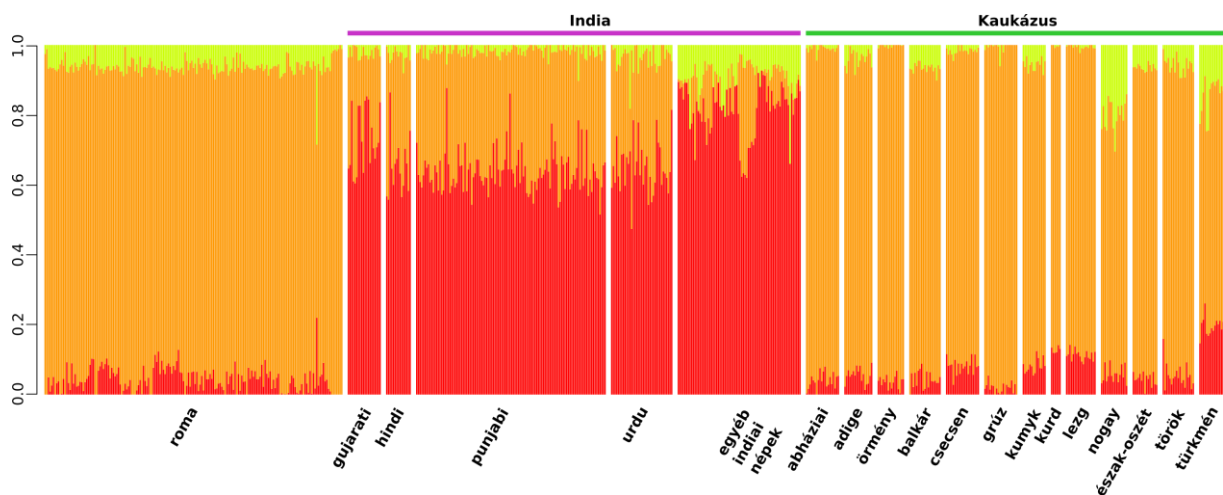
4.2. A romák kapcsolata a migrációs útvonalukba eső populációkkal

A romák kaukázusi populációkkal való viszonyának felderítésére az előbbiekhöz hasonló módszert alkalmaztunk. A roma, indiai és kaukázusi csoportokon végeztünk PCA és ADMIXTURE analíziseket. Az így kapott PCA ábrán az egyes populációk elhelyezkedése az európaiakkal való kapcsolat vizsgálatakor kapott PCA eredményre emlékeztet. Jelentős eltérés viszont, hogy a romák ebben az esetben egyértelműen a kaukázusi populációkkal együtt csoportosulnak az első és a második főkomponensen egyaránt, továbbá diverzitásuk kevésbé fejeződik ki. Ez arra enged következtetni, hogy a romák között a Nyugat-Ázsiából származó genetikai sajátosságok egyenletesebben oszlanak el az európainál, ami azt jelentheti, hogy a Nyugat- és Közép-Európában élő romák, amelyektől mintáink is származnak, szorosabb közös múlttal rendelkeznek. Migrációjuk során feltehetően egy csoportot alkottak, és később, az Európába való bejutásuk után szóródtak szét. Az ábrán továbbá azt is megfigyelhetjük, hogy a kaukázusi régióban élő nogay és türkmén csoportok jelentősebb kelet-ázsiai leszármazással rendelkeznek, mert elkülönülnek a különben egységesen, egy helyre szűken klasztereződő többi kaukázusi csoporttól (**7. ábra**). Az ADMIXTURE analízis eredménye megerősíti a PCA-t, a grafikonon nagyon jól látható, hogy kaukázusi csoportok alkalmazása esetén a romák jelentősen nagyobb genetikai egységességet mutatnak, illetve az is látható, hogy megközelítőleg egyformán osztoznak a dél-ázsiai, európai és nyugat-ázsiai és

kelet-ázsiai klasztereken, az említett erősebb kelet-ázsiai hatást kivéve nogay és türkmén csoportok esetében (8. ábra).



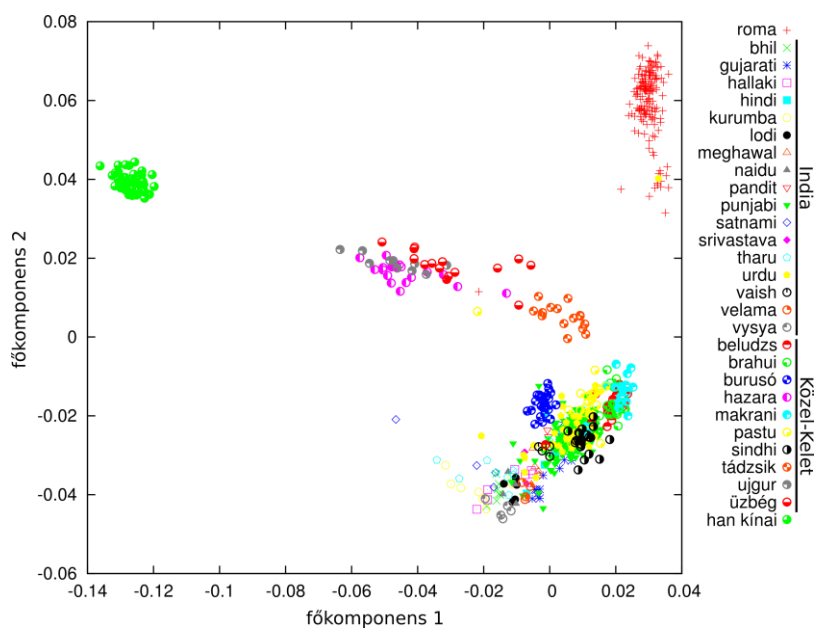
7. ábra. Roma, indiai és kaukázusi csoportok PCA eredményének diagramja az első két főkomponensen ábrázolva.



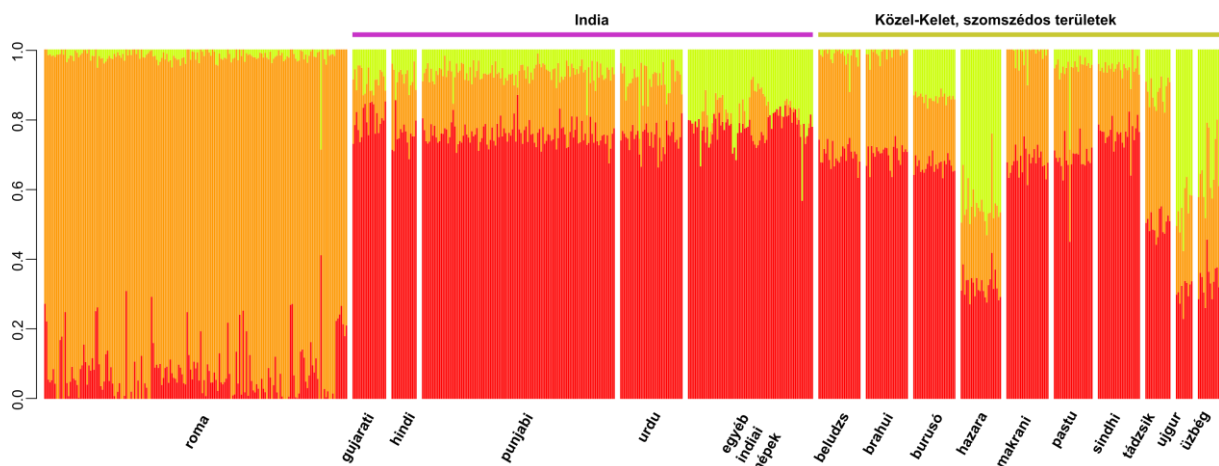
8. ábra. A roma, indiai és kaukázusi népeket tartalmazó adatok ADMIXTURE grafikonja (K=3). Az előzőekhez hasonlóan itt is a dél-ázsiai, európai, illetve nyugat-ázsiai és kelet-ázsiai hipotetikus csoportokat lehet elkülöníteni.

A romák kapcsolatát közel-keleti és Indiával szomszédos dél-ázsiai népekkel is megvizsgáltuk. A PCA eredménye jól mutatja, hogy a jelenlegi romák ma már genetikailag határozottan elkülönülnek a Dél-Ázsia népeitől. A közel-keleti és a nem indiai dél-ázsiai minták nagy része szorosan együtt klasztereződik India északi és nyugati részén élő csoportokkal, csak néhányuk tolódik el jelentősebben az első főkomponens mentén a han kínaiak irányába, amely kelet-ázsiai leszármazásra utal. A közel-keleti minták használatakor a romák erős diverzitása ismét visszatér, ugyanakkor annak eloszlása egyenletesebb, ami egyértelműen azt jelenti, hogy ezt a jelentősnek mondható diverzitást az európaiakkal való, eltérő mértékű keveredés okozta **(9. ábra)**.

Az ADMIXTURE grafikon megerősíti a PCA eredményeit, és jól látható az indiai és a környezetében élő populációk hasonló genetikai struktúrája, és a 4 populáció, amely jelentősebb kelet-ázsiai leszármazással bír **(10. ábra)**.



9. ábra. Roma, indiai és közel-keleti, illetve indiával szomszédos dél-ázsiai népek PCA eredménye az első és második főkomponenseken ábrázolva.



10. ábra. Roma, indiai és indiával szomszédos dél-ázsiai illetve közel-keleti népek ADMIXTURE grafikonja (K=3). A 3 hipotetikus ős megegyezik az eddigi analízisekben ismertettekkel.

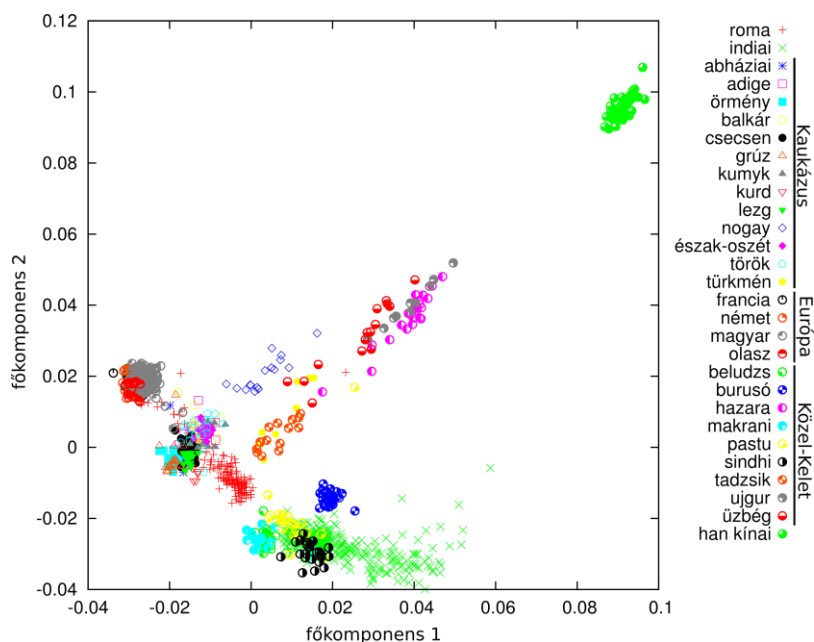
4.3. Összegző analízisek

Végül összevont analíziseket is végeztünk az eddig szerepelt összes populációt és csoportot egyetlen adatkészletté kombinálva. A PCA ábrán az egyes populációk megközelítőleg a tényleges topográfiai helyzetük szerint helyezkednek el egymáshoz képest, amelyen a néhány populációnál megfigyelt, erősebb kelet-ázsiai leszármazás kissé torzít. A romák a diagramon a kaukázusi és az indiai és velük szomszédos csoportok között helyezkednek el, a kaukázusi populációkhoz klasztereződve közelebb. A korábbi eredményekhez hasonlóan itt is jelentős számú roma helyezkedik el az európai populációk közelében vagy helyezkedik el azokkal egy csoportban. Az európai populációk használatával együtt visszatért a romák egyenlőtlen eloszlású diverzitása is (**11. ábra**).

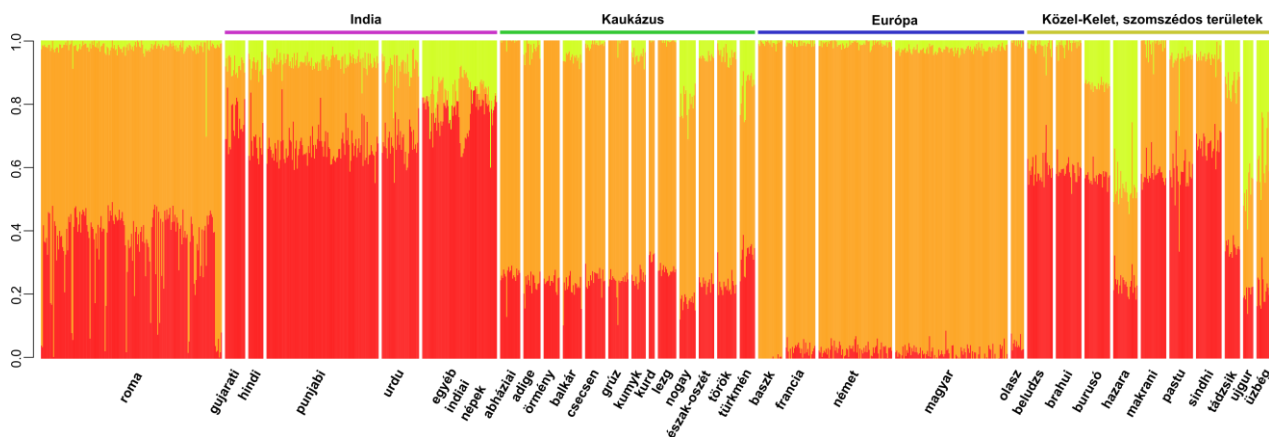
Míg a PCA ábra esetében topográfiai párhuzamot tudtunk vonni az egyes populációk PCA-n elfoglalt helyével, addig az ADMIXTURE ábrán a dél-ázsiai hipotetikus ős arányának folyamatos csökkenése jól megfigyelhető, ahogy az egyes populációk földrajzilag Dél-Ázsiától egyre távolabb helyezkednek el. Az ADMIXTURE ábrán is jól látszik, hogy a romák a kaukázusi populációk és a dél-ázsiai, közel-keleti területek népei között helyezkednek el (**12. ábra**).

Az adatkészletből F_{ST} mátrixot is készítettünk, amelyet két részletben tartalmaz a **2. és 3. táblázat**. Az F_{ST} értékekből leolvasható, hogy a romákhoz valóban a kaukázusi népek állnak legközelebb, illetve, hogy a romák genetikai struktúrája mára jobban hasonlít az európai populációkéra, mint az Indiában élő népekére. A romák a közel keleti és Indiával szomszédos dél-ázsiai populációktól állnak legtávolabb. A romák genetikai struktúrája a kaukázusi

populációk közül a balkárok, kumykok, nogay-okéhoz, illetve a Kaukázussal szomszédos törökökhöz áll legközelebb, az allélfrekvencia-differenciációk átlagértékei ezekkel a populációkkal szemben a legkisebb. A közel-keleti, illetve Indiával szomszédos népek közül a tádzsikokkal mutatnak még a kaukázusi népekhez hasonló alacsony differenciációs értéket.



11. ábra. A romák, és az eddigi analízisekben szereplő összes csoport PCA diagramja. A jobb átláthatóság végett minden indiai etnikumot egyetlen indiai csoportba vontunk össze.



12. ábra. A romák és az eddigi analízisekben szerepelt összes csoport ADMIXTURE analízisének grafikona (K=3). Az eddigiekhez hasonlóan piros színnel jelöltük a dél-ázsiai, narancs színnel az európai, illetve nyugat-ázsiai és sárga színnel a kelet-ázsiai hipotetikus ősi klasztert.

	roma	indiai	abháziai	adige	örmény	balkár	csecsen	örmény	kumyk	kurd	lezg	nogay	észak-oszét	török	türkmén	baszk	francia	német	magyar	olasz
roma	0.000																			
indiai	0.017	0.000																		
abháziai	0.015	0.020	0.000																	
adige	0.014	0.018	0.004	0.000																
örmény	0.014	0.020	0.004	0.005	0.000															
balkár	0.013	0.016	0.003	0.003	0.005	0.000														
csecsen	0.015	0.019	0.007	0.006	0.007	0.005	0.000													
örmény	0.016	0.021	0.001	0.004	0.003	0.003	0.007	0.000												
kumyk	0.012	0.014	0.004	0.003	0.004	0.002	0.003	0.004	0.000											
kurd	0.015	0.018	0.008	0.008	0.005	0.008	0.010	0.008	0.006	0.000										
lezg	0.015	0.017	0.006	0.006	0.006	0.005	0.005	0.006	0.003	0.008	0.000									
nogay	0.013	0.011	0.008	0.005	0.009	0.004	0.008	0.009	0.004	0.010	0.007	0.000								
észak-oszét	0.015	0.018	0.004	0.005	0.006	0.003	0.006	0.004	0.003	0.009	0.006	0.005	0.000							
török	0.011	0.015	0.004	0.004	0.002	0.003	0.006	0.004	0.002	0.004	0.005	0.005	0.004	0.000						
türkmén	0.014	0.012	0.012	0.010	0.012	0.008	0.012	0.013	0.007	0.011	0.010	0.004	0.010	0.007	0.000					
baszk	0.023	0.034	0.020	0.018	0.017	0.018	0.020	0.020	0.017	0.021	0.019	0.020	0.020	0.015	0.025	0.000				
francia	0.015	0.025	0.012	0.010	0.010	0.009	0.011	0.011	0.009	0.013	0.010	0.011	0.011	0.007	0.015	0.007	0.000			
német	0.015	0.025	0.012	0.010	0.011	0.009	0.011	0.012	0.009	0.013	0.010	0.011	0.012	0.008	0.015	0.009	0.001	0.000		
magyar	0.014	0.023	0.012	0.010	0.010	0.009	0.011	0.012	0.008	0.013	0.010	0.010	0.011	0.007	0.015	0.010	0.003	0.002	0.000	
olasz	0.015	0.026	0.010	0.009	0.007	0.008	0.010	0.010	0.008	0.011	0.010	0.012	0.011	0.005	0.016	0.008	0.002	0.003	0.004	0.000

2. táblázat. A roma, indiai, kaukázusi és európai populációk páronkénti allélfrekvencia-differenciációk átlagértékeiből készített mátrixa.

	roma	indiai	beludzs	brahui	burusó	hazara	makrani	pastu	sindhi	tádzsik	ujgur	üzbég
roma	0.000											
indiai	0.017	0.000										
beludzs	0.017	0.008	0.000									
brahui	0.019	0.010	0.004	0.000								
burusó	0.020	0.006	0.011	0.014	0.000							
hazara	0.030	0.017	0.025	0.027	0.017	0.000						
makrani	0.016	0.009	0.003	0.004	0.012	0.026	0.000					
pastu	0.014	0.003	0.005	0.007	0.006	0.018	0.006	0.000				
sindhi	0.017	0.003	0.005	0.007	0.008	0.021	0.006	0.003	0.000			
tádzsik	0.013	0.007	0.008	0.009	0.008	0.014	0.007	0.004	0.007	0.000		
ujgur	0.029	0.017	0.024	0.027	0.016	0.004	0.026	0.017	0.020	0.013	0.000	
üzbég	0.021	0.012	0.017	0.019	0.012	0.005	0.018	0.012	0.015	0.008	0.003	0.000

3. táblázat. A roma, indiai és közel-keleti, illetve egyéb dél-ázsiai populációk páronkénti allélfrekvencia-differenciációk átlagértékeiből készített mátrixa.

5. Összefoglalás

Munkánk során SNP array genotípusadatok felhasználásával behatároltuk a romák genetikai struktúra alapján világviszonylatban elfoglalt helyét. Megállapítottuk, hogy mára közelebb állnak az európai populációkhoz, mint az indiaiakhoz. Az európai populációkkal való keveredésük a kontinens különböző részein eltérő, ami rendkívül diverz csoporttá teszi őket. Ugyanakkor ez az európaiakkal való erős, olykor extrém keveredés felelős azért, hogy az indiai eredet rovására differenciációjuk az európai populációktól egyre csökken. A romák legközelebb a kaukázusi régióban élő etnikai csoportokhoz, illetve a velük szomszédos törökökkel mutatják a legnagyobb hasonlóságot, viszont a jelenség nem az európaiakra jellemző genetikai struktúra felé sodródás számlájára írható. A kaukázusiak és romák közötti relatíve nagymértékű hasonlóság az Európában élő romákban egyenletesen oszlik el, amely a romák migrációja során történt jelentősebb keveredésre, egy kaukázusi populációkkal közös nyugat-eurázsiai leszármazásra utalhat.

Vizsgálataink szerint a romák a migrációs útvonalukba eső népek közül a legkevesebb hasonlóságot az Indiával szomszédos, illetve közel-keleti csoportokkal mutatják, ami arra utalhat, hogy ezekkel a népekkel jelentős keveredés vándorlásuk során nem történt.

A romák a kaukázusi népek közül a balkárokhöz, kumykokhoz és nogay-okhoz állnak a legközelebb. Hasonlóan kis differenciációt az Indiával szomszédos területeken élő népek közül a tádzsikkal találtunk.

A kapott eredmények további vizsgálata hivatalos, genetikai keveredést feltáró statisztikai módszerekkel, a genetikai keveredés okozta kapcsoltsági egyenlőtlenség amplitúdójának becslése az egyes populációkban, illetve a leszármazás útján azonos (Identity by Descent, IBD) DNS szegmentek azonosítása és statisztikai úton való feldolgozása további betekintést nyújthat a romák és a migrációs útvonalukba eső népek között lezajlott populációgenetikai kölcsönhatásokba, segíthetik azok megértését, pontos leírását.