



XXXI. OTDK Biológia Genetika szekció

**Etnikai hovatartozás becslése
autoszómális SNP markerek alapján**

OTDK dolgozat

Készítette:

Bárány Gusztáv

Biológus MSc, II. évf.

ELTE-TTK, Biológiai Intézet, Genetikai Tanszék

Igazságügyi Szakértői és Kutató Intézetek, Budapesti Orvosszakértői Intézet; DNS
laboratórium

Témavezetők:

Dr. Pamzsav Horolma, igazságügyi genetikus szakértő, laborvezető

Dr. Völgyi Antónia, igazságügyi genetikus szakértő

TARTALOMJEGYZÉK

1. ÖSSZEFOGLALÁS	3
2. SUMMARY	4
3. BEVEZETÉS	5
3.1 A magyar populációk összetétele	6
3.2 Az SNP etnikai eredet becslésre való alkalmazhatósága.....	7
3.3 Kapcsolat a pigmentáció és az etnikai eredet között	8
4. CÉLKITŰZÉSEK	9
5. ANYAGOK és MÓDSZEREK.....	9
5.1 mintavétel, Felhasznált minták	9
5.2 Vizsgált SNP lokuszok	10
5.3 Genotipizálás	11
5.4 SNP szelektálás a makrókba.....	11
5.5 Predikciós makrók kifejlesztése	12
5.6 A makrók visszatesztelése	13
6. EREDMÉNYEK	13
6.1 kiválasztott SNP-k rangsora és SNP tesztek.....	13
6.2 Predikciós modellek	15
6.3 Makrók visszatesztelése - szimuláció.....	16
7. AZ EREDMÉNYEK ÉRTÉKELÉSE.....	18
8. TOVÁBBI TERVEK	19
8. SZERZŐI HOZZÁJÁRULÁS	20
9. KÖSZÖNETNYILVÁNÍTÁS	21
10. RÖVIDÍTÉSEK JEGYZÉKE	22
11. IRODALOMJEGYZÉK	23

1. ÖSSZEFOGLALÁS

A külső jellegeket meghatározó génekkel kapcsolt szekvencia polimorfizmusok meghatározása segíthet bizonyos fenotípus paraméterek becslésében. Az ezzel kapcsolatos korábbi kutatások sikeresen becsülték meg az írisz pigmentáltságát 6 ún. single-nucleotide polymorphism (SNP) lokusz segítségével. Egyes etnikai csoportok megfelelő – komplex fenotípus együttesét célzó – SNP lokuszok vizsgálatával lehetővé válik az ismeretlen eredetű minták rasszhoz, etnikumhoz tartozásának genetikai alapú becslése (EEO).

Céлом a legnépesebb magyar etnikai csoportok (roma, német, szlovák, horvát) etnikai besorolására alkalmas SNP markerekkel bővíteni az igazságügyi genetika eszköztárát. Jelen munkában a roma, mint legnépesebb magyar kisebbség genetikai alapú becslésére alkalmas SNP markerekkel foglalkozom.

Azonosítottam azokat a lokuszokat, melyeket allél-frekvenciájuk alapján szignifikáns eltérést mutatnak a két csoport között. Továbbá megmutattam az alkalmazott marker szettek szignifikáns megkülönböztető erejét. A dolgozatban 366 magyar és 147 roma mintát vizsgáltam 20 autoszomális SNP markerrel. A genotipizálást TaqMan próbával real-time PCR készüléken végeztem el. A 20 marker allélfrekvencia adataiból számolt G-teszt szignifikancia szintje alapján kiválasztottam 11 SNP lokuszt. Multinomiális logisztikus regresszió segítségével kiszámoltam az α és β együtthatókat, melyeket predikciós modellekbe (makrókba) építettem be. Továbbá igyekeztem optimalizálni, azaz csökkenteni a vizsgálatához szükséges SNP markerek számát. Ezért különböző próbák (G-teszt és Wald-teszt) eredményeit felhasználva több modellt is építettem.

A makrók specifikusságát modellépítésben nem szereplő 66 magyar és 53 roma mintán teszteltem vissza.

Az újonnan bevezetett SNP-k és a felállított makrók segítségével ismeretlen eredetű bűnügyi minták etnikai hovatartozása becsülhetővé vált magyar-roma relációban.

1. SUMMARY

The identification of external visible character-associated biallelic markers can be applied as a forensic phenotyping method. Previous studies predicted the human eye (iris) colour successfully using single-nucleotide polymorphisms (SNP) at six loci. With determination of SNP loci targeting ethnic-specific phenotype complexes it is possible to estimate the ethnic origin of unknown DNA samples found at crime scene.

The aim of this study was to investigate the possibility to estimate the ethnic origin of an unknown biological sample (suspect, victim, etc.) with determination of a set of autosomal SNPs. The long-term plan is to generate a toolkit for forensic genetics predicting the ethnic groups in Hungary.

In the present study I investigated the Romani – as the most abundant Hungarian ethnic group – and Hungarian populations in the above context. I tested models which were able to distinguish between Hungarian and Romani samples using a specific set of SNPs, and set up the one, which provides the best results. To conduct that 336 Hungarian and 147 Romani samples were tested with 20 autosomal SNP markers. For genotyping of biallelic markers real-time PCR with TaqMan probes was performed. I identified those loci that showed significant differences between Hungarian and Roma samples, based on allele-frequencies. 11 SNPs were chosen based on significance level using G-test. Different kinds of model buildings were constructed due to the given rank by various statistic probes such as Wald probe and G-test. A prediction model was set up with multinomial logistic regression to determine the α and β coefficients. Furthermore, I tried to reduce the number of SNPs applied to decrease the time and costs of the DNA tests. I verified the specificity of the models with 66 Hungarian and 53 Romani samples that are not used in model building based on the percentage of correct estimations of ethnic origin. Additionally, I demonstrated that several panel of markers provided significant statistical power for ethnic-affiliation estimation, but the rank based on G-test was the most useful for the model that estimated the ethnic origin the best.

In conclusion, these models using relevant SNPs, appear to be a promising approach to estimate the ethnical origin of unknown forensically samples in Hungarian-Romani relation.

2. BEVEZETÉS

Az igazságügyi genetika információt szolgáltat bűncselekmények felderítésében, ismeretlen személyazonosságú emberi maradványok azonosításában, valamint származás-megállapítási vizsgálatokban (pl. apasági vizsgálatok). Igazságügyi szempontból a tudományág tehát eszköz, melynek segítségével a kriminalisztikai feltételezések megerősíthetők, vagy elvethetők. Ezen alkalmazott tudományterület probléma felvetései, módszertani és technikai fejlesztései által nemcsak passzív felhasználóként, hanem aktív közreműködőként kapcsolódik a molekuláris genetikához, populációgenetikához, kriminalisztikai és jogalkalmazáshoz egyaránt.

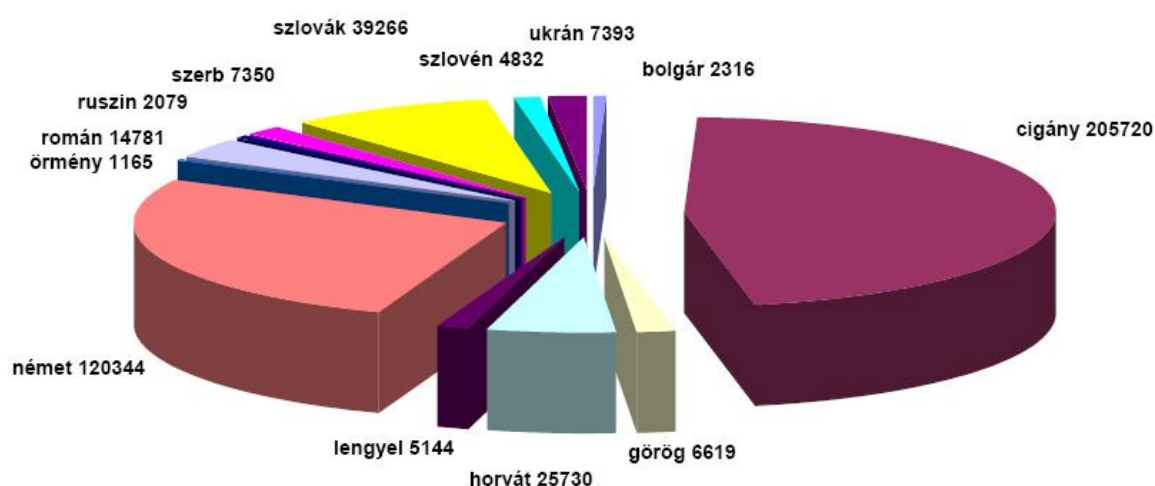
Az elképzelést, hogy léteznek olyan genetikai markerek (lokuszok) melyek egyes populációkban jelen vannak, míg másokban hiányoznak először Neel (1973) mutatta be, aki mutációs ráta becslésére használta ezeket a markereket. Ezen markerek jelenlétén, vagy hiányán túlmenően allél-gyakoriságuk közti különbségek is mutatkoznak a földrajzilag, illetve etnikailag eltérő populációk között. A lokuszok polimorfizmusának mértéke tehát populációról populációra változhat (Dean et al; 1994, Cavalli-Sforza et al; 1994.) A legjobb megkülönböztető erővel bíró leszármazási markerek (AIMs) azok, melyek két populáció között a legnagyobb allél-frekvencia eltérésekkel rendelkeznek (Chakraborty et al; 1992; Stephens et al; 1994; Shriver et al; 1997). A biológiai nyomot hátrahagyó személy etnikai eredetének megbecslésére (EEO, Estimating the Ethnic Origin) irányuló vizsgálatoknak világszerte (USA, Japán, Hollandia) elterjedt gyakorlata van (Shriver et al;1997) és hazánkban is folytak ilyen irányú vizsgálatok (Klitsch et al; 2003).

Törvényszéki eljárásban genetikai vizsgálatot először 27 évvel ezelőtt, 1985-ben alkalmaztak (Jeffreys et al; 1985a, 1985b). Az igazságügyi genetikai gyakorlatban, számos esetben a gyanúsított és/vagy a sértett személyazonossága nem ismert, így fenotípusa sem. A molekuláris biológiai technológia és genetika informatikai fejlődésével napjainkban lehetővé vált a helyszínen biztosított biológiai anyagmaradványok alapján bizonyos fenotípusos jellegek becslése, mivel egyes szekvencia polimorfizmusok kapcsolatosan öröklődnek fenotípust meghatározó génekkel. A szekvencia polimorfizmusok és fenotípusok statisztikai kiértékelése után lehetséges a tévedés lehetőségének figyelembe vételével a jelleg becslése. A bűnjelekről biztosított ismeretlen eredetű biológiai anyagmaradványok igazságügyi genetikai vizsgálata, etnikai eredetének becslése igen jelentős és hathatós nyomozati információval szolgálhat a hatóságok számára. Az is lehetséges, hogy valós tanú hiányában csak a genetikai információ „beszél”, vagyis a genetikai szemtanúra támaszkodhat egy nyomozati anyag.

2.1 A MAGYAR POPULÁCIÓK ÖSSZETÉTELE

A Központi Statisztikai Hivatal 2001-es népszámlálási adatai alapján nevesített főbb magyar kisebbségek a cigány, német, szlovák és horvát. Ezen kisebbségek arányát az 1. ábra szemlélteti. A hatályos adatvédelmi törvény (1992. évi LXIII. törvény) alapján az etnikai hovatartozást az érintett írásos beleegyezése nélkül nem lehet nyilvántartani, mivel az alapvető emberi jognak tekinti az identitás meghatározását. Tehát az etnikai hovatartozás definiálása személyes identitás kérdése.

1. ábra: A törvényben nevesített kisebbségek száma és aránya 2001-ben.



Tóth és Vékás, 2001 alapján

A cigány vagy roma népcsoportok, indiai eredetükből adódóan (Pamjav et al; 2011), többnyire armenid típus jegyeik alapján jól megkülönböztethetők a turáni típusra jellemző morfológiai karaktereket is viselő magyar népcsoportoktól (Coon et al, 1939; Kiszely, 1979, Henkei,). Ezek alapján a megfelelő – komplex fenotípusos jegyek együttesét célzó – SNP lokuszok vizsgálatával lehetséges a roma és magyar etnikai csoportok elkülönítése csupán biológiai anyagmaradványok alapján.

Mivel ma Magyarországon a legnépesebb kisebbség a roma, jelen dolgozatban az etnikai eredet becslésének magyar-roma relációjával foglalkozom. Trivialitás, hogy az etnikai eredet etnográfiai kérdés is. A magyarországi romák nyelvi és történeti szempontból, valamint önmaguk meghatározása szerint öt nagy csoportba sorolhatók. Az öt magyarországi csoport közül 1) a magyar cigányok, őket említik "romungróként", a 2) az oláh cigányok, ők jelentik a

tulajdonképpen romákat, és a 3) a beások képezik a legnépesebb csoportot. A 4) csoport a szintók és az 5) a román cigányok (Kemény, 2000). Ezek az etnikai csoportok nem különülnek el élesen egymástól. A keveredés demográfiai tekintetben legalább olyan mértékű, mint köznyelvi fogalomhasználat esetén. Vannak, akik etnikai önmeghatározásukkor „roma” kifejezést használnak és elvetik a „cigány” megnevezést, mivel azt durvának és derogálónak tartják. A másik oldalról viszont, egyesek kifejezetten cigánynak vallják és nevezik önmagukat, mint ahogy ez etnográfiai szempontból alátámasztott.

Látható, hogy a megfelelő pontos és elfogadható fogalomhasználat megválasztása igen nehéz és kényes kérdés. Mivel dolgozatomban egy fenotípusos jelleg-együttesen alapuló etnikai hovatartozás becslésével foglalkozom, a precíz etnográfiai alapú definíció használata kevésbé indokolt. Mindezen okoknál fogva jelen dolgozatban a magyarországi cigány kisebbségre, melynek része a roma csoport is, a továbbiakban egyöntetűen a politikailag korrektnek tekinthető „roma” kifejezést fogom alkalmazni.

2.2 AZ SNP ETNIKAI EREDET BECSLÉSRE VALÓ ALKALMAZHATÓSÁGA

A genom egy meghatározott pontján, azaz lokuszon egy nukleotidot érintő szekvencia variánsok az ún. single-nucleotide polymorphisms (SNPs), vagy SNP-k. Az SNP-k tehát pontmutációk eredményei. Ezen mutációk adják a humán genetikai variációk közel 90%-át. A humán genomban hozzávetőlegesen 1000 bp-onként található egy definiált SNP (The International HapMap Consortium, 2007). A kérdéses DNS szekvencia-variációt akkor tekinthetjük polimorfizmusnak, ha az a populáció legalább 1%-ában jelen van. Az SNP-k kapcsolt molekuláris markerként jól használhatók genetikai betegségek és egyéb genetikai alapú tulajdonságok követésére, térképezésére (Brookes et al; 1999), valamint alkalmazni lehet ismeretlen bűnügyi minták leszármazásának tisztázására is (Sturm & Frudakis, 2004).

Az SNP-k alacsonyabb mutációs rátával rendelkeznek, mint az igazságügyi genetikában széleskörűen elterjedt mikroszatellita markerek (STR), ezért „jobban rögzülnek” egy adott populációban. Az SNP lokuszok – mutációs rátájuk szerint – minden 108. generációban mutálódnak (Brookes, 1999). Relatív alacsony mutációs rátájuknak köszönhetően számos SNP populáció-specifikus (AIM, Ancestry Informative Marker). Az SNP-k többnyire két allélvariánssal rendelkeznek, az adott populációt a minor allél frekvenciájával jellemezhetjük, mely az alacsonyabb gyakoriságú allél. Az ilyen SNP-k a bűncselekmények felderítése során

igen hasznosak lehetnek a leszármazási vonalakon túlmenően egy fenotípusában is elkülönülő szűkebb etnikai csoport becslésére (Butler, 2005).

A különböző filogenetikai háttérrel rendelkező emberek nem feltétlenül mutatják az egyébként várható fenotípus jegyeket, karaktereket. Ezért a leszármazás, vagy etnikai eredet becslését célzó genetikai tesztek eredményeit mindig körültekintően – a becslés bizonytalanságának figyelembevételével –, valamint más megbízható bizonyítékokkal összefüggésben kell értelmezni (Butler, 2005).

2.3 KAPCSOLAT A PIGMENTÁCIÓ ÉS AZ ETNIKAI EREDET KÖZÖTT

Korábbi tanulmányok alapján (Bouakaze et al; 2009, Liu et al; 2009, Walsh et al.; 2010) a laboratórium 2011-ben indította el az ún. „pigment projektet”. Ennek célja volt olyan alkalmazás kifejlesztése, amely csak egyes autoszómális SNP markerek vizsgálata alapján becsli a humán írisz pigmentáltságát. Az ezzel kapcsolatos vizsgálataink sikeresek voltak, lehetségessé vált a szemszínt csupán 6 SNP lokusz segítségével megbecsülni (Völgyi et al; 2011). A vizsgálatba bevont 20 SNP marker egy része alkalmas az ázsiai és európai eredet elkülönítésére is (Keyser et al; 2009). Egy korábbi feltételezés szerint a populáció specifikussága a pigmentációért felelős génekkel kapcsolt (Sturm et Frudakis, 2004, Pulker et al; 2007). Frudakis és munkatársai (2003) szerint a pigmentáltságot meghatározó génekhez asszociált SNP-k az ún. PIM (Phenotype Informative Marker) több, mint 80%-a alkalmas etnikai eredet és populációk szubstruktúrájának becslésére, melyet több későbbi forrás is alátámasztott (Branicki et al; 2005, Walsh et al; 2010). Ez valószínűleg azzal magyarázható, hogy az egyes etnikai csoportok pigmentáltságukban jelentősen eltérhetnek.

Mivel a legnépesebb magyar kisebbség fenotípusosan jól megkülönböztethető a magyartól, a fentiek fényében indokoltnak látszott a pigmentáltsághoz kapcsolt SNP-kre, mint populáció-specifikus allélvariánsokra (PSA, Population-Specific Alles) tekintve olyan alkalmazás kifejlesztése, amely képes különbséget tenni magyar és roma etnikai csoportok között. A roma etnikai eredet becslésére már korábban is folytak magyar vonatkozású kutatások, ennek során allélfrekvenciára alapozva STR markerekkel 88,8% érzékenységű eljárást fejlesztettek ki (Klitschar et al; 2003). Ugyanakkor fontos megjegyezni, hogy globalizált társadalmakban teljesen izolált, etnikai közösségek gyakorlatilag nem léteznek (Füredi, 2003). Következésképpen, etnikai eredetük becslése elméletileg korlátozott.

3. CÉLKITŰZÉSEK

Jelen munkám célja, hogy egy autoszómális SNP marker együttes felhasználásával olyan predikciós modellt készítek, mely szenzitív a magyar és roma populációk között mutatkozó allélgyakoriságok különbségeire, ezáltal egy ismeretlen minta genotípusa alapján becslést adhassak annak etnikai eredetére. Elképzelésem az, hogy a közeljövőben a módszer igazságügyi alkalmazására sor kerül, és az hasznos kiegészítő nyomozati információval szolgálhat.

A 366 magyar és 147 roma mintán 20 SNP genotípusának meghatározása után, a markerek allélfrekvenciáit számoltam ki. A 20 SNP marker közül kiválasztottam azokat az SNP-eket, amelyek a magyar és roma minták közt G-teszttel szignifikáns eltéréseket mutattak. A kiválasztott SNP-eket a G-teszt és Wald teszt p értékei és a multinomiális logisztikus regresszió β együtthatói alapján külön-külön rangsorokat határoztam. Így tehát 3 féle rangsort kaptam az SNP magarázó erejére nézve. Az összes szignifikanciát mutató 12 SNP-t tartalmazó makró után készítettem még a legjobbnak látszó 4 és 7 SNP-t tartalmazó makrót is. A sorrendek szerint multinomiális logisztikus regresszióval készítettem 7 predikciós modellt. Végül a modellépítésben fel nem használt mintákkal visszatesztelve ellenőrzöm a makrók érzékenységét, és több paraméter (pl. érzékenység, specifikusság) alapján kiválasztom a legjobb becslést adó makrót.

4. ANYAGOK és MÓDSZEREK

4.1 MINTAVÉTEL, FELHASZNÁLT MINTÁK

A laboratóriumban párhuzamosan futó két kutatásban („pigment projekt” és etnikai eredet becslése) használt minták egymással átfednek. Mindkét vizsgálatban egyazon 366 magyar mintát vizsgálatuk. A jelen dolgozatomban további 147 roma eredetű mintát is vizsgáltam.

A roma személy etnikai csoportához tartozását külső látható jelleg együttesek alapján határozzuk meg. Ennek megfelelően magyarként kezeljük a turáni típusjegyeket hordozó személyeket, míg romaként kezeljük az armenid típusjegyeit mutató személyeket. A roma minták tulajdonosai – írásos beleegyezésükkel – a rutin apasági vizsgálatok során hozzájárultak mintáik felhasználására kutatási célokra. A magyar eredetű mintákat (továbbiakban: magyar minta) önkéntesek biztosították az írásos beleegyezésükkel. Donorjaink túlnyomó többségben budapesti és pécsi egyetemisták voltak, akikről részletes

adatlap készült az írisz, bőr és hajszín, valamint etnikai eredet vonatkozásában. A mintákat azonosítókkal láttuk el, a továbbiakban ez alapján kezeltük azokat. A minták típus szerint vér minták és szájnyálkahártya kenetek.

4.2 VIZSGÁLT SNP LOKUSZOK

A „pigment projekt” során Liu és munkatársai (2009), Walsh és munkatársai (2010), Bouakaze és munkatársai (2009), valamint Eriksson és munkatársai (2010) nyomán kiválasztott 20 SNP-t vizsgáltunk. Ezek az SNP lokuszok egyértelműen asszociáltak a humán szemszínt meghatározó génekhez, melyek közül Walsh és munkatársai szerint 8 alkalmas európai és ázsiai eredet elkülönítésére is (Walsh et al., 2010). Az írisz pigmentáltságának genetikai komplexitásának köszönhetően a kiválasztott SNP-k a következő génekhez kapcsolódtak: *CDK10*, *DCT*, *FANCA*, *HERC2*, *IRF4*, *MC1R*, *SLC45A2*, *SLC24A4*, *SLC24A5*, *TYR*, *OCA2*. A 20 SNP jellemzőit a 1. táblázat foglalja össze. A könnyebb laboratóriumi azonosítás érdekében az SNP „rs” azonosítói helyett a labor saját kódolásával azonosítottuk (Pig#), melyet a továbbiakban is követek.

1. táblázat: Az SNP-k jellemzőinek összefoglalása

SNP ID	Labor ID	Kromoszóma	Pozíció*	Gén	Allél
rs4778138	Pig1	15.	26009415	<i>OCA2</i>	A/G
rs12913832	Pig2	15.	26039213	<i>HERC2</i>	A/G
rs7495174	Pig3	15.	26017833	<i>OCA2</i>	A/G
rs16891982	Pig4	5.	33987450	<i>SLC45A2</i>	C/G
rs1805008	Pig5	16.	89986144	<i>MC1R</i>	C/T
rs1805007	Pig6	16.	89986117	<i>MC1R</i>	C/T
rs4778241	Pig7	15.	26012308	<i>OCA2</i>	C/T
rs1800407	Pig8	15.	25903913	<i>OCA2</i>	C/T
rs1545397	Pig9	15.	28187772	<i>OCA2</i>	A/T
rs2031526	Pig10	13.	95100841	<i>DCT</i>	A/G
rs1426654	Pig11	15.	48426484	<i>SLC24A5</i>	A/G
rs12896399	Pig12	14.	91843416	<i>SLC24A4</i>	G/T
rs1393350	Pig13	11.	88650694	<i>TYR</i>	A/G
rs12203592	Pig14	6.	341321	<i>IRF4</i>	C/T
rs1597196	Pig15	15.	26009415	<i>OCA2</i>	G/T
rs2240204	Pig16	15.	28494032	<i>HERC2</i>	C/T
rs7183877	Pig17	15.	26039328	<i>HERC2</i>	A/C
rs8039195	Pig18	15.	28516084	<i>HERC2</i>	C/T
rs1667394	Pig19	15.	26203777	<i>HERC2</i>	C/T
rs258322	Pig20	16.	89755903	<i>CDK10</i>	C/T
rs12931267	Pig21	16.	89818732	<i>FANCA</i>	C/G

* Liu et al, 2009, és NCBI alapján

Mivel a két kutatási vonal („pigment projekt”, etnikai eredet becslése) ugyanabból a 20 SNP-ből indult ki, ezért a labormunkát a 366 magyar minta esetén Erős Ágnes hallgatótársammal közösen végeztük (lásd. Erős Ágnes TDK dolgozata, 2012).

4.3 GENOTIPIZÁLÁS

Az önkéntes mintaadóktól származó magyar mintákból (szájnyálkahártya törletek) PrepFiler™ kittel izoláltunk genomi DNS-t a gyártó utasításainak megfelelően (Applied Biosystems Inc. Foster City, CA). A roma minták DNS izolálása QIAamp DNA Blood Mini Kittel gyártói protokoll szerint (Qiagen, Hagen, Germany) történt. Az izolátumok DNS koncentrációmérését ABI 7500 real-time PCR készüléken Quantifiler Human DNA Kittel a gyártói protokoll szerint végeztük el (Applied Biosystems Inc. Foster City, CA).

A PCR reakcióhoz 3-5 ng templátot használtunk. Az amplifikációt a real-time PCR készüléken Förster-féle (fluoreszcens) rezonáns energiaátadáson (FRET) alapuló allél-specifikus oligonukleotid hibridizációs TaqMan próbával (Applied Biosystems, Foster City, CA) végeztük gyártói utasítás szerint. A minták genotipizálását – a PCR termék relatív floreszcencia értékeinek elemzését – az ABI 7500 készülék SDS szoftverével végeztük el. Az egyes SNP-eket csak egymástól független reakcióban tudtuk vizsgálni, így az 513 minta mindegyikét 20 SNP TaqMan próbájával vizsgálva 10260 adatot kaptunk eredményül. Az egyes minták genotípusát a 20 SNP allélvariáns kombinációi adják meg.

4.4 SNP SZELEKTÁLÁS A MAKRÓKBA

Első lépésként a 366 magyar és 147 roma minta genotípusa alapján kiszámoltam a 20 SNP marker mindegyiknek allélfrekvenciáit. Major-minor allél definiálását, a „pigment projekt” kompatibilitásából adódóan, a magyar minták szerint végeztem. Tehát minden marker major allélja a magyar mintákban gyakoribb allél, míg a minor a ritkább allélvariáns lett. A genotípusokat a predikciós modell számára kezelhető formába kódoltam át. A kódolás a következő volt: homozigóta major = 0, heterozigóta = 1, homozigóta minor = 2. A modellben a homozigóták lokuszonkénti β együtthatóját nullával szorozzuk, a homozigóta minor β értékét pedig kettővel. A β értékek összege arányos a „roma kimenetel” valószínűségével. Ennek köszönhetően lesz a modell becslése roma-specifikus.

A Pig5 és Pig6 markerek minden minta esetében ugyanazt a genotípust adták, ezért a Pig5 lokuszt önkényesen kivontam vizsgálatból. A mintacsoportok közötti allélfrekvencia

különbség szerint azonosítottam azokat, melyek az etnikai hovatartozás becslésekor jelentős statisztikai megkülönböztető erővel rendelkezhetnek. A 19 SNP marker közül G-tesztel kiválasztottam a magyar és roma minták allélfrekvenciái közti szignifikáns eltérést mutatókat ($p \leq 0,01$).

4.5 PREDIKCIÓS MAKRÓK KIFEJLESZTÉSE

A módszer hosszú távú gyakorlati alkalmazása előtt optimalizálni kell annak eredményességét és a ráfordított erőforrásokat, azaz minimalizált számú SNP alkalmazásával kell megalkotni a lehető legeredményesebb makrót. Ehhez szükséges a 12 darab G tesztel szignifikáns eltérést mutató SNP-ket a szignifikancia szerint rangsorolni (például 4 legjobb, 7 legjobb). Az így csoportosított SNP lokuszok használatával különböző makrók építhetők és tesztelhetők, ezek a Wald és G-tesztelkel rangsorolt 4 és 7 SNP lokuszt tartalmazó makrók. Természetesen az összes szignifikanciát mutató 12 SNP lokuszt tartalmazó makrót is kifejlesztettem, mivel várható, hogy a több SNP lokuszt tartalmazó makró predikciós eredményessége legjobb lehet. A statisztikai próbákat, a modellépítést és a szimulációkat Microsoft Excel és R-GUI 2.15.1 programokon végeztem (R, 2012).

A rangsor megállapítását három módon végeztem el:

- a G-teszt p-értékei szerint rendezve sorba (Shriver et al; 1997)
- a Wald-teszt p-értékei szerint rendezve sorba (Liu et al; 2009)
- β sorrendet alkalmazva, ebben az esetben a 12 SNP-t tartalmazó makrónál használt β együtthatók értékei szerint válogattam ki a többi csoportot képező SNP lokuszokat.

A megállapított rangsorok szerint, mindhárom sorrendből kiválasztottam azt az első 4 majd 7 lokuszt, amelyhez a legkisebb p érték, legnagyobb β együttható tartozik. Az értékek számottevő ugrásainál szubjektíven lett a 4 illetve 7 darab beválasztva a modellépítésbe. Ezek alapján három 4, három 7 és egy 12 SNP lokuszt tartalmazó makrót készítettem. A fenti 3 módszer más-más sorrendet állapított meg a 12 SNP-ből, ezért a különböző SNP csoportok más-más lokuszokat tartalmaznak. A predikciós modellek építése multinomiális logisztikus regresszió (1. és 2. egyenlet) alapszik. Regresszió esetén függvénykapcsolatot keresünk egy vagy több magyarázó változó (független változó) és egy függő változó között. A logisztikus regresszió megbecsüli egy bizonyos esemény bekövetkezésének valószínűségét, mely során maximum likelihood becslést alkalmazza a függő változó logit változóvá transzformálása után. Az adatsorok alapján kiszámoltam a regresszió α és β együtthatóit, az együtthatók ismeretében már kiszámolható a kimeneteli események valószínűsége, tehát egy ismeretlen

eredetű genotípus klasszifikációja elvégezhető, becsülhető. A modell azt tételezi fel, hogy a beírt változók az egyik kimenetel ($P(Y=1|x_i)$ magyar vagy roma) valószínűségét írják le.

1., 2. egyenlet: multinomiális logisztikus regresszió:

$$1. \quad \ln \frac{P(Y=1|x_i)}{1-P(Y=1|x_i)} = \ln \frac{P(Y=1|x_i)}{P(Y=0|x_i)} = a + \sum_{j=1}^m b_j * x_{i,j}$$

$$2. \quad P(Y=1|x_i) = \frac{\exp(a + \sum_{j=1}^m b_j * x_{i,j})}{1 + \exp(a + \sum_{j=1}^m b_j * x_{i,j})}$$

Ahol, $x_{i,j}$ = az i-edik személy j-edik jellemzőjének értéke,
 y_i = az i-edik személyek célváltozója,
 m = az osztályozási szempontok (jellemzők) száma.

4.6 A MAKRÓK VISSZATESZTELÉSE

A genotipizált a 366 magyar és 147 roma mintából 300 magyar és 94 roma mintát használtam fel a predikciós modell felállítására. A maradék 66 (18%) magyar és 53 (36%) roma mintát az elkészült modellek tesztelésére, a becslések szimulálására használtam fel. A szimulációban a tesztelt személy (akinek etnikai hovatartozása ismert) genotípusa felel meg az ismeretlen etnikai eredetű biológiai nyom genotípusának. A szimulációk összessége információt nyújt a tesztelt makró hatékonyságáról.

5. EREDMÉNYEK

5.1 KIVÁLASZTOTT SNP-K RANGSORA ÉS SNP SZETTEK

A G tesztek p-értékei alapján a két populáció között 12 SNP mutatott szignifikáns eltérést. (2. táblázat).

2. táblázat: A genotípus adatsor frekvencia adatai és a G-teszt p értéke

SNP markerek		Allél	Allélgyakoriság		G-test P érték (p<0.01)
NCBI azonosító	Laborazonosító		Magyar minták	Roma minták	
rs4778138	Pig1	A G	0,774456522 0,225543478	0,60544218 0,39455782	5,5409E-08
rs12913832	Pig2	A G	0,34421365 0,65578635	0,67346939 0,32653061	4,5212E-25
rs7495174	Pig3	A G	0,888732394 0,111267606	0,76190476 0,23809524	4,7686E-07
rs16891982	Pig4	G C	0,936141304 0,063858696	0,65646259 0,34353741	3,4946E-28
rs1805007	Pig6	C T	0,974114441 0,025885559	0,97959184 0,02040816	5,9168E-01
rs4778241	Pig7	A C	0,244565217 0,755434783	0,41836735 0,58163265	5,7172E-08
rs1800407	Pig8	C T	0,938858696 0,061141304	0,96258503 0,03741497	1,1196E-01
rs1545397	Pig9	A T	0,922343324 0,077656676	0,08163265 0,91836735	7,9149E-01
rs2031526	Pig10	A G	0,141304348 0,858695652	0,21768707 0,78231293	3,7912E-03
rs1426654	Pig11	A G	0,989130435 0,010869565	0,82993197 0,17006803	2,8784E-21
rs12896399	Pig12	G T	0,607336957 0,392663043	0,65986395 0,34013605	1,2891E-01
rs1393350	Pig13	A G	0,206521739 0,793478261	0,12244898 0,87755102	1,1735E-03
rs12203592	Pig14	C T	0,89673913 0,10326087	0,92857143 0,07142857	1,1386E-01
rs2240204	Pig16	A G	0,081521739 0,918478261	0,21768707 0,78231293	8,2125E-09
rs7183877	Pig17	A C	0,080163043 0,919836957	0,74468085 0,25531915	1,1024E-114
rs8039195	Pig18	C T	0,17798913 0,82201087	0,44897959 0,55102041	1,3195E-18
rs1667394	Pig19	A G	0,786684783 0,213315217	0,47959184 0,52040816	9,2076E-22
rs258322	Pig20	A G	0,054347826 0,945652174	0,04081633 0,95918367	3,5135E-01
rs12931267	Pig21	G C	0,027173913 0,972826087	0,0170068 0,9829932	3,1566E-01

A táblázatban zöld színnel emeltem ki a szignifikáns p értékkel rendelkező SNP lokuszokat.

A három módszerrel kapott SNP rangsort (G-teszt, Wald-teszt, β rendezés) a 3. táblázat szemlélteti.

3. táblázat: A modelleket alkotó SNP összeállítások						
SNP szett	Lab. ID	G-teszt p-érték	Lab. ID	Wald-teszt p-érték	Lab. ID	$\beta_{(i,j)}$ együttható
12SNP szett	Pig17	1,1024E-114	Pig11	7,9238E-08	Pig19	4,2916E+01
	Pig4	3,4946E-28	Pig2	4,3139E-03	Pig16	2,4651E+01
	Pig2	4,5212E-25	Pig13	1,5618E-02	Pig11	1,6364E+01
	Pig19	9,2076E-22	Pig19	5,0156E-02	Pig17	6,5501
	Pig11	2,8784E-21	Pig1	1,5589E-01	Pig4	2,5266
	Pig18	1,3195E-18	Pig7	2,8759E-01	Pig1	1,7445
	Pig16	8,2125E-09	Pig17	4,3992E-01	Pig18	1,5032
	Pig1	5,5409E-08	Pig18	9,5824E-01	Pig10	1,3326
	Pig7	5,7172E-08	Pig16	9,7427E-01	Pig13	3,7554E-02
	Pig3	4,7686E-07	Pig4	9,8419E-01	Pig7	-4,1831E-01
	Pig13	1,1735E-03	Pig3	9,8611E-01	Pig3	-1,4150E+01
	Pig10	3,7912E-03	Pig10	9,9382E-01	Pig2	-4,2183E+01

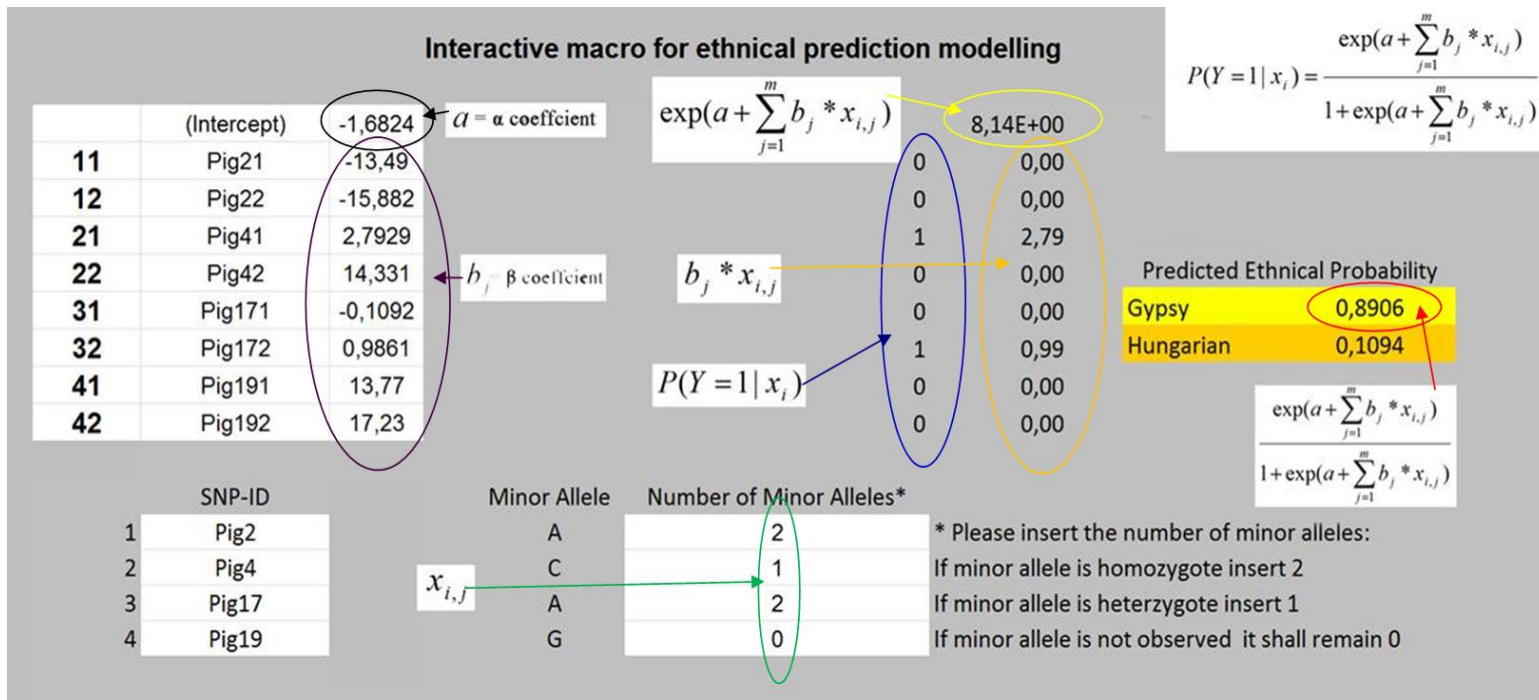
Az SNP-k rangsorolását 3 módszerrel végeztem el. Ezek a következők: G-teszt, Wald-teszt, és a regresszió β együtthatójának sorrendje alapján. A kiválogatott SNP szettekkel készítettem a modelleket.

A három módszer gyakorlatilag három teljesen eltérő rangsort alakított ki. A β értékek alapján meghatározott sorrend sokkal nagyobb mértékben tér el a G-teszt és Wald teszt eredményétől, mint azok egymástól. Bár a G és a Wald-tesztek más eredményt adtak, a Pig 19 és a Pig2 lokusz mindkét esetben előkelő pozícióban szerepel, Pig3 és Pig10 pedig egyaránt a táblázat alján szerepel. A sorrendek alapján határoztam meg a 4 és 7 SNP szetteket, amelyek a Pig2, Pig4, Pig11, Pig16, Pig17 és Pig19 SNP lokuszokat tartalmazzák (3. táblázat).

5.2 PREDIKCIÓS MODELLEK

R programcsomaggal végeztem a modellépítést, melynek eredménye az adott SNP együttesre jellemző alfa és béta együtthatók ismerete. Ezeket behelyettesítve a multinomiális logisztikus regresszió képletébe, megkapjuk a kész modellt, amelyen a szimulációk lefuttathatók. Hét ilyen makrót készítettem. Ezek a G-teszttel, Wald teszttel és β rendezéssel szelektált SNP lokuszok eredményeiből készült, valamint a 12 szignifikáns lokuszt tartalmazó makró. Az eredmények könnyebb alkalmazhatósága miatt a multinomiális logisztikus regressziót Excel felületre ültettem át (makró). A makró működését a 2. ábra mutatja be.

1. ábra: A Regressziós modell Excel felületen



Az ábrán a 4 SNP-t tartalmazó regressziós modell látható Excel felületen. A teljes képlet a jobb felső sarokban látható. X jelöli az adott makró kódolt genotípusát. A képlet egyes elemeinek helyzetét a nyilak jelölik. A példában az adott genotípusú mintát a modell 89,06% valószínűséggel romának becsli.

5.3 MAKRÓK VISSZATESZTELESE - SZIMULÁCIÓ

A modellek megbízhatóságát meghatározó paraméterek megválasztása az interpretáció sarkalatos pontja. Több olyan tanulmány is született a közelmúltban, amely az etnikai eredet meghatározásának megvalósíthatóságát az STR lokuszok allélgyakoriságai alapján tesztelte (Evett et al; 1997; Smouse és Chevillon, 1998). E megközelítések fő hiányossága, hogy az eredményekből a helyes és téves becslések várt aránya, vagy ehhez hasonló paraméter, így a procedura megbízhatósága nem kiolvasható.

Az etnikai eredet meghatározásának érzékenységet a helyes becslések százalékos arányával mértem oly módon, hogy a valós pozitív becslések számát elosztottam a valós pozitív és fals negatív becslések összegével. Ezt külön-külön számoltam a két mintacsoport esetén. Pozitívnak a $p \geq 0,5$ értéket tekintettem. Ebben az összefüggésben az etnikai származás becslésének megbízhatóságát a korrektnek tekintett EEO-k aránya fejezi ki. Az eredmények 4. táblázatban láthatók.

4. táblázat: A szimulációk eredménye, a makrók érzékenysége

Rendezés szempontja	Makrószett	minta eredete	helyes becslések aránya	makrók átlagos érzékenysége	rendezések átlagos makrók átlagos érzékenysége
nincs	12 SNP	magyar	92,75%	83,17%	nem értelmezhető
		roma	73,58%		
G-teszt	4 SNP	magyar	92,75%	84,11%	83,75%
		roma	75,47%		
	7 SNP	magyar	91,30%	83,39%	
		roma	75,47%		
Wald-teszt	4 SNP	magyar	97,10%	76,85%	77,40%
		roma	56,60%		
	7 SNP	magyar	89,86%	77,95%	
		roma	66,04%		
Log.reg. β együttható	4 SNP	magyar	95,65%	67,27%	73,80%
		roma	38,89%		
	7 SNP	magyar	92,75%	80,34%	
		roma	67,92%		

A roma és magyar minták becslési érzékenységében jelentős eltérés mutatkozik (5. táblázat). Az makrók optimalizálása szempontjából lényeges, hogy a kapott eredmény minél kevesebb vizsgálat alapján minél megbízhatóbb legyen. A makrók SNP darabszámának átlagos csökkenésének hatását a 6. táblázat szemlélteti.

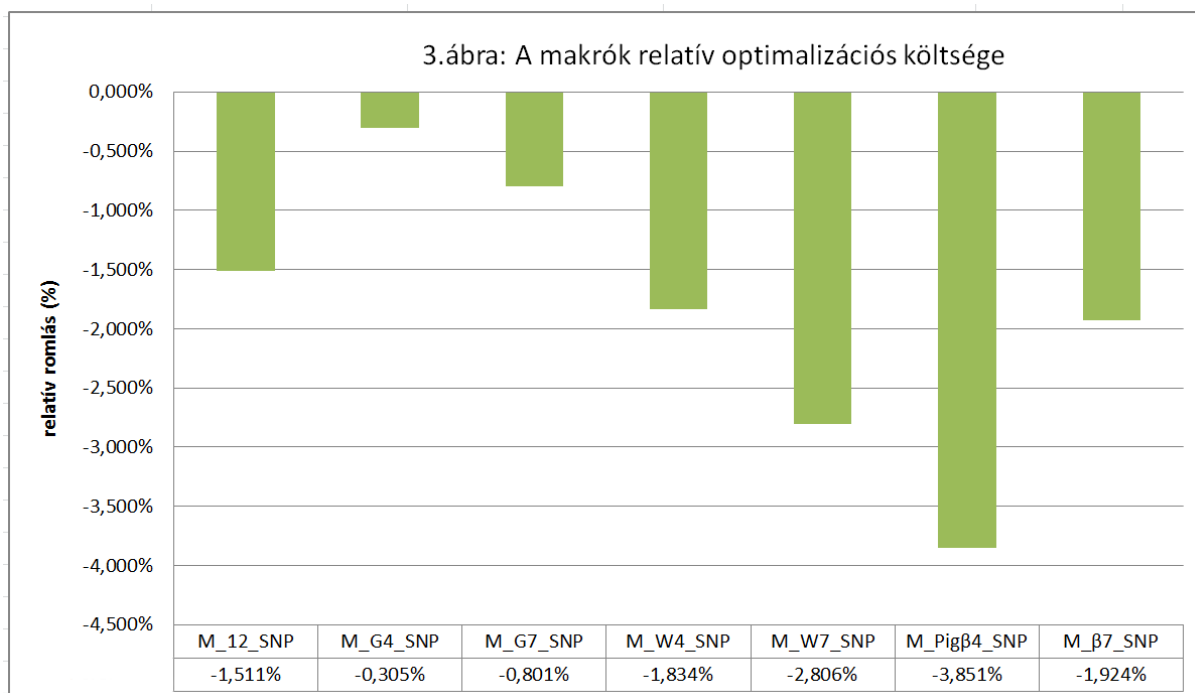
5. táblázat: A mintaszám hatása a makró szenzitivitására

	magyar minták	roma minták
Modellépítés mintaszáma (n_b)	300	94
Visszatesztelés mintaszáma (n_t)	66	53
Becslések átlagos érzékenysége	93,17%	67,92%

6. táblázat SNP szám csökkenés átlagos hatása a modellek érzékenységére

modellek átlagos érzékenysége	4 SNP	7 SNP	8 SNP	19 SNP
	76,08%	80,56%	83,17%	85,56%

Az SNP lokusz szám csökkentésével a becslés jóságában csökkenő tendencia látható. Ez alapján azonban nehéz felmérni, hogy vajon az SNP szám csökkenése közelebb visz-e az optimális megoldáshoz, vagy sem. Ha a becslés érzékenységének csökkenését súlyozom a SNP szám csökkenésének arányával a fenti kijelentés módosul. Az így kapott értéket optimalizációs „költségnek” nevezem. Ennek érzékeltetésére hivatott a 3. ábra.



A relatív optimalizációs „költség” az adott makró érzékenységeinek romlása, a 19 SNP-t tartalmazó makróhoz képest (0%), melyet súlyoztam (szoroztam) a makró SNP szám csökkenésének arányával. Ez a mutató érzékelteti az SNP szám csökkentésének hatásfokát, melynek a gyakorlati alkalmazás során lesz jelentősége. M_G: G-teszt alapú makró, M_W: Wald-teszt alapú makró, M_β: β egygyűthetű alapú makró

6. AZ EREDMÉNYEK ÉRTÉKELÉSE

A kiválogatott markerek mindegyike pigmentáltságért felelős génhez kapcsolt. A pigmentáltság és az etnikai hovatartozás becslésére használt markerek egymástól egyértelműen nem elkülöníthetők (Frudakis et al; 2003), az általunk vizsgált SNP markerek a romákra jellemző fenotípusos jelleg együtttest valószínűsítene.

A 3. táblázat alapján lehetséges, hogy a modellépítéshez használt 94 roma minta még nem elegendő nagyszámú, mivel a becslési érzékenység átlagosan 25,25%-al alatta marad a helyes magyar becsléseknek, ahol a mintaszám 300 volt. Ennek oka persze abban is rejtőzhet, hogy a romák már régóta ebben a földrajzi régióban élnek, így keveredésük a turáni magyar lakossággal nem zárható ki. Amennyiben további minták vizsgálatára lehetőség nyílik a becslés eredményessége új makrók létrehozásával növelhető lesz.

A 4. táblázat adataiból jól látható, hogy a rangsorolások módszerek közül a legjobb eredményt adó sorrendet a G-test p értékei alapján kaptuk, vagyis a frekvenciaadatok különbözősége jobban visszatükrözte az egyén etnikai hovatartozását, mint a Wald-teszt p értéke (amely az egyes lokuszok szignifikanciájáról ad információt a predikcióban), vagy a β paraméterek

súlyozása. Tehát a matematikai megközelítést egy empirikusnak kell követnie ahhoz, hogy megfelelő lokusz készlethez jussunk egy megfelelő modellben.

A makrók SNP szám csökkenésével – ahogy az várható – csökken a makró megbízhatósága is (6. táblázat). Az optimalizációs „költség” értéket tekintve a legjobbnak bizonyuló, G–tesztrel rendezett 4 SNP lokuszt tartalmazó makró szett esetén csökken legkevésbé a relatív becslési megbízhatóság (3. ábra).

A fentiek figyelembevételével a becslés megbízhatósága és az optimalizációs „költség” alapján kijelenthetjük, hogy a G-tesztrel válogatott és rangsorolt 4 SNP marker együttes alkalmazásával (Pig2, Pig4, Pig17, Pig19) magyar-roma relációjú etnikai eredetet 84,11% érzékenységgel becselő predikciós modellt fejlesztettünk ki.

7. TOVÁBBI TERVEK

Rövidtávú terveim szerint növelni szeretném a makró becslési hatékonyságát, érzékenységet. Ezt a szelektáló és sorrendbe állító eljárások finomításával szeretném elérni.

A becslést minősítő paramétereket a becslés bizonytalanságával fogom bővíteni, amely a szenzitivitás és a specificitás alapján készített görbe (ROC) integrálja (AUC) (Liu et al, 2009).

Középtávú tervként szeretném további roma minták genotipizálásával növelni a modellépítés mintaszámát és ezzel, valamint a visszatesztelési logika továbbfejlesztésével javítani a makró becslését. Korábbi kutatásoknak köszönhetően rendelkezésünkre állnak mongol és malajziai-indiai minták. Szeretném e mintákat is genotipizálni, majd a kapott genotípusokkal tesztelni a makró specificitását.

Hosszútávon a többi jelentős arányú kisebbséget (német, szlovák, horvát) is vizsgálni szeretnénk. Azonban a romák közti szigorúbb házasodási szokások következtében kialakult izoláció, nem, vagy jóval kevésbé jellemző a többi népcsoportra. Ezért feltételezésünk szerint ezen populációk esetében egyáltalán nem, vagy kevés olyan SNP lokusz várható, mely alkalmas lehet az elkülönítésükre.

Mivel a vizsgálathoz nagy mennyiségű DNS minta szükséges és bűncselekmény helyszíni minták esetében ez sokszor nem áll rendelkezésünkre, a gyakorlati alkalmazás továbbfejlesztése lehet a kiválasztott SNP lokuszokra multiplex SNaPShot vizsgálat tervezése, így kisebb mennyiségű DNS minta is elegendő lesz majd a vizsgálatokhoz.

8. SZERZŐI HOZZÁJÁRULÁS

Munkám fő céljai a következők voltak:

1. Elsőként a mintavételek során - melyekben részt vettem - biztosított 366 magyar, és 147 roma eredetű mintát genotipizáltam 20 SNP tekintetében, mely vizsgálatokat a laborasszisztencia támogatásával, Erős Ágnes hallgató társammal közösen végeztük.
2. Ezt követően a magyar és roma allélfrekvenciák alapján általam végzett statisztikai próbák (G-teszt, Wald-teszt) segítségével kiválasztottam az egyes makrókat képező SNP csoportokat.
3. Készítettem több, multinomiális logisztikus regresszió alapú, algoritmust (makrót), melyek becslést adnak egy genotípus etnikai csoporthoz tartozásának valószínűségéről.
4. Az elkészített algoritmusokat 66 magyar és 53 roma mintával visszateszteltem és a legjobb becslést adó makró kiválasztásával meghatároztam az etnikai hovatartozást legjobban becslő négy SNP markert.

9. KÖSZÖNETNYILVÁNÍTÁS

Köszönetet szeretnék mondani témavezetőmnek, **Dr. Pamzsav Horolmának**, a Budapesti Orvosszakértői Intézet DNS laboratórium vezetőjének hogy szakmai fejlődésemet felkarolta és biztosította a kutatás megfelelő feltételeit. Így támogatva kutatómunkám mind megfelelőbb elvégzését.

Külön köszönettel tartozom **Dr. Völgyi Antónia** témavezetőmnek, akinek állhatatos és türelmes útmutatásai, tanácsai nélkül ez a dolgozat nem készülhetett volna el.

Köszönet illeti **Dr. Mádi-Nagy Gergelyt**, az ELTE TTK Operációkutatási Tanszék adjunktusát, amiért utat mutatott a valószínűségi becslések útvesztőjében és segített saját matematikai korlátaim kitágításában.

Ezúton szeretném megköszönni **Dr. Egyed Baláznak** értékes tanácsait és a gondolatébresztő konzultációkat, melyekre mindig szorított időm.

Köszönöm **Erős Ágnes** hallgatótársamnak a laborban végzett eredményes, megbízható és üdítő együttműködést.

Szeretném továbbá megköszönni mindazon személyeknek a hozzájárulását, akik biológiai mintájukkal megalapozták az adatsor elkészítését.

Végül, de nem utolsósorban szeretném megköszönni **Dr. Susa Évának az ISZKI** főigazgató asszonyának a nélkülözhetetlen antropológiai tanácsokat, és a kutatások finansziális alapjainak messzemenő biztosítását, melyek lehetővé tették a vizsgálatok kivitelezését.

10. RÖVIDÍTÉSEK JEGYZÉKE

AIM: ancestry informative marker

AUC: area under curve

CDK10: cyclin-dependent kinase 10 [*Homo sapiens*]

DCT: dopachrome tautomerase (dopachrome delta-isomerase, tyrosine-related protein 2) [*Homo sapiens*]

EEO: estimating the ethnic origin

EVC: extrenal visible characters

FANCA: Fanconi anemia, complementation group A [*Homo sapiens*]

FRET: Fröster (fluorescence) resonance energy transfer

HERC2: HECT and RLD domain containing E3 ubiquitin protein ligase 2 [*Homo sapiens*]

IRF4: interferon regulatory factor 4 [*Homo sapiens*]

makró: multinomiális logisztikus regresszió modell.

MC1R: melanocortin 1 receptor (alpha melanocyte stimulating hormone receptor) [*Homo sapiens*]

nDNS: nukleáris DNS

PCR.: polimerase chain reaction

PIM: phenotype informative marker

PSA:Population-specific alleles

RT-PCR: RealTime PCR

ROC: receiver operating characteristic

SLC24A5: solute carrier family 24, member 5 [*Homo sapiens*]

SLC45A2: (MATP): solute carrier family 45, member 2 [*Homo sapiens*]

SLC24A5: solute carrier family 45, member 2 [*Homo sapiens*]

SNP: single nukleotid polymorphim

STR.: short tandem repeat

TYR: tyrosinase (oculocutaneous albinism IA) [*Homo sapiens*]

OCA2: oculocutaneous albinism II [*Homo sapiens*]

11. IRODALOMJEGYZÉK

Coon CS. (1939) The Races of Europe *Macmillan Com.* call number: 31695

Branicki W., Brudnik U., Woja-Pelc A., **(2005)** Genetic prediction of pigmentary traits in forensic studies. *Problems of Forensic Sciences*, LXIV, 343–357

Brookes, AJ. (1999) The essence of SNPs. Review, *Gene*, 234, 177–186.

Bouakaze C., Keyser Ch., Crubézy E., Montagnon D., Ludes B, **(2009)** Pigment phenotype and biogeographical ancestry from ancient skeletal remains: inferences from multiplexed autosomal SNP analysis *Int J Legal Med* DOI 10.1007/s00414-009-0348-5

Butler, JM. (2005) Forensic DNA Typing. Biology, technology Genetics of STR markers. 2nd edit. *Academic Press* ISBN 0-12-147952-8

Cavalli-Sforza LL, Menozzi P, Piazza A **(1994)** Geography of Human Genes. *Princeton University Press*

Chakraborty R, Kamboh MI, Nwankwo M, Ferrell RE **(1992)** Caucasian genes in American blacks: new data. *Am J Hum Genet*, 50:145-155

Dean M, Stephens JC, Winkler C, Lomb DA, Ramsburg M, Boaze R, Stewart C, et al **(1994)** Polymorphic admixture typing in human ethnic populations. *Am J Hum Genet*, 55: 788-808

Eriksson, NJ. Macpherson M., Tung JY., Hon1 LS., Naughton B., Saxonov S., Avey L., Wojcicki A., Pe'er I., Mountain J. **(2010)** Web-Based, Participant-Driven Studies Yield Novel Genetic Associations for Common Traits. *PLoS Genetics* V. 6. e1000993

Evett IW. Gill PD. Lambert JA. Frazier NR, Watson S., Panchal S, Connolly A. Kimpton C. **(1997)** Statistical analysis of data for three British ethnic groups from a new STR multiplex. *Int J Legal Med.* 110: 5-9

Füredi S., (2003) Humán polimorf mikroszatellita (short tandem repeat) lokuszok igazságügyi genetikai vizsgálata magyar populációkban. *Doktori értekezés. ELTE, Biológia Doktori Iskola*

Frudakis T., Venkateswarlu K., Matthew J. Thomas JM., Gaskin Z., Ginjupalli S., Gunturi S Ponnuswamy V., Natarajan S., and Nachimuthu PK. **(2003)** A Classifier for the SNP-Based Inference of Ancestry. *Journal Forensic Science*, Vol. 48, No. 4 Paper ID JFS2002079_484

Jeffreys AJ., Wilson V. & Thein SL. **(1985a)** Hypervariable 'minisatellite' regions in human DNA. *Nature* Vol. 314: 67-73

Jeffreys AJ., Wilson V. & Thein SL. **(1985b)** Individual-specific 'fingerprints' of human DNA. *Nature* Vol. 316: 76-79

Kemény I. [szerk.] (2000) A magyarországi romák. *Útmutató Kiadó*, 128 p., Változó világ sorozat, 31. ISBN:963-9001-47-3

Keyser Ch, Bouakaze C., Crubézy E., Nikolaev GV., Montagnon D., Reis T., Ludes B. (2009). Ancient DNA provides new insights into the history of south Siberian Kurgan people. *Hum Genet* 126: 395–410. DOI 10.1007/s00439-009-0683-0

Kiszely I.(1979) A föld népei. *Gondolat kiadó* ISBN: 963-280-689-1

Klintschar AM., Füredi S, Egyed B., Reichenpfader B., Kleiber M., (2003) Estimating the ethnic origin (EEO) of individuals using short tandem repeat loci of forensic relevance *International Congress Series* 1239 53– 56 PII: S0531-5131(02)00513-7

Liu F., Duijn K, Johannes RV., Hofman A., Uitterlinden AG., Janssens JW and Manfred Kayser M., (2009) Eye color and the prediction of complex phenotypes from genotypes. *Current Biology* Vol 19 No 5

Neel JV (1973) "Private" genetic variants and the frequency of mutation among South American Indians. *Proc Natl Acad Sci USA*, 70: 3311-3315

Pamjav H., Zalán A. Béres J., Nagy M. and Chang YM., (2011) Genetic Structure of the Paternal Lineage of the Roma People *Amercian J. of Physical Anthropology* 000:000–000

Pulker H., Lareu MA., hillips C., Carracedo A (2007) Finding genes that underlie physical traits of forensic interest using genetic tools. *Forensic Science International: Genetics* 1 100–104

R version 2.15.1 -- "Roasted Marshmallows" Copyright (C) 2012 The R Foundation for Statistical Computing ISBN 3-900051-07-0

Smouse PE. and Chevillon C. (1998) Analytical Aspects of Population-Specific DNA Fingerprinting for Individuals *Journal of Heredity*;89:143–150

Shriver DS., Smith MW., Jin L., Marcini A., Akey JM., Deka R., and Ferrell RE. (1997) Ethnic-Affiliation Estimation by Use of Population-Specific DNA Markers *Am. J. Hum. Genet.* 60: 957-964

Stephens JC, Briscoe D, O'Brien S (1994) Mapping by admixture linkage disequilibrium in human populations: limits and guidelines. *Am J Hum Genet* 55:809-824

Sturm AR. and Frudakis TN. (2004) Eye colour: portals into pigmentation genes and ancestry. *Trends in Genetics* Vol. 20

The International HapMap Consortium (2007) *Nature* Vol 449 doi:10.1038/nature06258

Tóth Á., Vékás J (2001). Nemzeti és etnikai kisebbségek Magyarországon (a 2001. évi népszámlási adatok rövid összefoglalása), Etnikai-nemzeti Kisebbségkutató Intézet Kiadványa

Völgyi A., Zalán A., Sándor G., Kis Z., Heinrich A., Toldy E., Mádi-Nagy G., Susa É. and Pamzsav H., (2011) ISFG conference poster Wien, Austria

Walsh S., Liu F., Ballantyne KN., Oven M., Lao O., Kayser M. **(2010)** IrisPlex: A sensitive DNA tool for accurate prediction of blue and brown eye colour in the absence of ancestry information *Forensic Science International: Genetics*; 572 doi:10.1016/j.fsigen.2010.02.004