Szegedi Tudományegyetem
Szent-Györgyi Albert Orvostudományi Kar
Multidiszciplináris Orvostudományok Doktori Iskola

Új számítástechnikai módszer rokonsági kapcsolatok meghatározására 4. fokig

Ph.D. Tézisfüzet

Nyerki Emil

Témavezető:

Dr. Maróti Zoltán, tudományos főmunkatárs

Szent-Györgyi Albert Klinikaközpont

Gyermekgyógyászati Klinika és Gyermekegészségügyi Központ

Genetikai Diagnosztikai Laboratórium

Szeged

2024

Publikációs lista

I. **Emil Nyerki**, Tibor Kalmár, Oszkár Schütz, M. Lima Rui, Endre Neparáczki, Tibor Török, Zoltán Maróti. *Correctkin: An Optimized Method to Infer Relatedness up to the 4th Degree from Low-Coverage Ancient Human Genomes*. Genome Biology 24, no. 1 (2023). https://doi.org/10.1186/s13059-023-02882-4. **IF:12.3 D1**

II. Gergely I.B. Varga, Lilla Alida Kristóf, Kitti Maár, Luca Kis, Oszkár Schütz, Orsolya Váradi, Bence Kovács, Alexandra Gînguță, Balázs Tihanyi, Péter L. Nagy, Zoltán Maróti, **Emil Nyerki,** Tibor Török, Endre Neparáczki. *The archaeogenomic validation of Saint Ladislaus' relic provides insights into the Árpád dynasty's genealogy* ,Journal of Genetics and Genomics, Volume 50, Issue 1, (2023) **IF:5,9 Q1**

Bevezetés

A rokonsági elemzés mind az archeogenetikában, mind az igazságügyi tudományokban alapvető fontosságú eszköz az egyének közötti kapcsolatok genetikai információkon alapuló meghatározásához.

Az archeogenetikában a rokonsági elemzés alapvető fontosságú az ősi népességek, a migrációs minták és a genetikai kapcsolatok megértéséhez. Az Y-kromoszóma markerek és mikrochip panelek használata értékesnek bizonyult az ősi maradványok rokonsági elemzése során, és fényt derített a rokonság jelenlétére vagy hiányára a múltból származó egyének között[1]. Emellett az új genetikai panelek és SNP markerek kifejlesztése lehetővé tette a kutatók számára a genetikai diverzitás, az ősi populációk haplotípusos szerkezetének feltárását, segítve a biogeográfiai származásra való következtetést és a rokonsági vizsgálatot. [2]

Az ősi populáció megértése mellett a rokonsági elemzés erősen ajánlott a populációgenetikai elemzések során, de a legtöbb populációgenetikai elemzésnek is előfeltétele,

hogy a közeli rokonokat kizárják az adathalmazokból (pl. ADMIXTURE, főkomponens-elemzés (Principal component analysis PCA)).

Célkitűzés

A PC-Relate algoritmus figyelembe veszi a minta egyedei közötti populációs struktúrát (származást) a rokonsági együttható becsléséhez használt főkomponensek (principal components, PC-k) segítségével. Következésképpen, bár a referenciaadatoknak jelen kell lenniük az elemzett adathalmazban, a legjobb referenciát nem kell kifejezetten megadni minden egyes potenciálisan rokon egyedhez. Az eredeti algoritmus nagyon jó pontosságú a teljesen genotipizált modern adatok felhasználásával, azonban nem alkalmas az ősi DNS-szekvenciákból származó, részben genotipizált haploid adatok elemzésére. Kutatásom során az volt a feladatom, hogy teszteljem azt a hipotézisünket, hogy a PC-Relate algoritmus nyers rokonsági együttható becslése pszeudohaploid részleges markeradatokból korrigálható a mintákban lévő genotipizált markerek aránya alapján. Az

algoritmus PCAngsd 0.99-es verziójú implementációját használtuk. Kutatásom során a következő feladatokban vettem részt.

1. A kezdeti vizsgálat során arra kerestük a választ, hogy az általánosan alkalmazott véletlenszerű pszeudohaploidizáció jelentősen befolyásolja-e a főkomponens-elemzés (PCA) eredményeit. A későbbi vizsgálatok során szimulációkkal vizsgáltuk a véletlenszerű pszeudo-haploidizáció (RPsH) hatását a rokonsági fokok számításaira.

2. Az archaikus minták genomlefedettsége óriási változatosságot mutat az általában alacsony (változó) endogén DNS-tartalom miatt. A második tanulmányban ismert genomlefedettségi adatok visszaskalázásával szimulációkat használtunk fel annak vizsgálatára, hogy a részlegesen genotipizált markerfrakció hogyan befolyásolja a becsült rokonsági együtthatót.

3. A harmadik vizsgálatban a véletlen hibák és a PMD hiba hatásának szimulálása volt a cél a korrigált rokonsági együtthatóra.

4. Célunk volt annak tesztelése, hogy a pontos referenciapopuláció hiánya (ami gyakran előfordul ősi adatok esetében) nem érvényteleníti-e nagymértékben az eredményeinket. Továbbá azt is tesztelni kívántuk, hogy keveredő rokonság esetén a nem kevert forráspopulációkat is fel lehet-e használni referenciaként.

5. A továbbfejlesztett algoritmus validálása a READ-hez képest. Az algoritmus validálása az AADR-adatsor ismert rokonain.

6. A Szent László ereklye archeogenomikai validálása.

Anyag és Módszertan

Használt adatok

Minden elvégzett elemzésben az Allen Ancient DNA Resource (AADR) 1240K SNP-készletének genomkoordinátáit használtuk. [3] A markerek átfedésének szimulációjához két különböző, teljesen tipizált modern adathalmazt használtunk: az 1000 Genomes Project Phase

3 adatai [4], és egy jelentős méretű, Zöldfok szigeteki-magyar vegyes család ismert családfával, első (testvérek), második (féltestvérek) és ötödik fokú rokonokkal, amelyek egy anonimizált klinikai biobankból származnak.

Annak vizsgálatára, hogy a genom lefedettség milyen hatással van a valódi ősi adatokból származó becsült rokonsági együtthatókra, egy dokumentált középkori szülő-utód pár nem publikált 1240K genotípus adatait használtuk. A nyilvános AADR V42.4 [3] 1240K adatállomány szolgált a módszertan validálására az ókori egyének széles skáláján. Konkrétan olyan ősi mintákat vontunk be, amelyek több mint 100K genotipizált markerrel rendelkeztek (N=2810), míg az i.e. 8000 előtti mintákat (N=216) kizártuk a ritkaságuk és a referenciapopulációként való nem megfelelő reprezentációjuk miatt. ([5] Kiegészítő fájl 7: S4 Ábra). Továbbá, hogy megakadályozzák az elegendő vagy megfelelő referenciapopulációval nem rendelkező egyedek elemzését, a mintákat földrajzi koordinátáik alapján korlátozták (földrajzi hosszúság 12-120 és szélesség 28-65), ami 458 egyed kizárását eredményezte.

([5] Kiegészítő fájl 7, S5 ábra). A szűrést követően az adathalmaz 2136 ősi egyént tartalmazott. ([5] Kiegészítő fájl 6, S5 táblázat).

Új bioinformatikai eszközök

A genotípusadatok zökkenőmentes importálásának, manipulálásának, és elemzésének megkönnyítése érdekében a javasolt munkafolyamaton belül alapvető eszközöket fejlesztettünk ki, többek között a következőket:

- importHaploCall: Az ANGSD-ből származó pszeudo-haploid genotípushívások importálására tervezve.

- pseudoHaplo: az RPsH elvégzésére diploid adatállomány felhasználásával.

- markerOverlap: A páronkénti markerátfedés-frakció mátrixának kiszámítása.

- filterRelates: A rokonsági együttható korrigálása és a rokonok szűrése hibamodellek és/vagy szigorú rokonsági együttható küszöbértékek alapján.

A részlegesen genotipizált markerek hatásának kontrollált vizsgálatához és a teljesen genotipizált modern minták elemzésével való összehasonlításhoz a következőket használtuk:

- depleteMarkers: A kiválasztott minták között a kívánt marker átfedési hányad szimulálása.

- depleteIndivs: Részlegesen genotipizált minták véletlenszerű kohorszának szimulálása.

Az alacsony lefedettség hatásának szimulálása teljesen tipizált modern adatkészletek alapján

A lefedettség és az ebből következő alacsonyabb genotipizálási százalékos arányok rokonsági együttható számításokra gyakorolt hatásának szisztematikus vizsgálatához a "depleteMarkers" módszerrel véletlenszerűen eltávolítottunk markereket egy teljesen tipizált adatállományból. Ez az eljárás lehetővé tette számunkra, hogy elérjük a kívánt százalékos markerátfedést két minta között. Ezzel az eszközzel az átfedő markerfrakciókat szimuláltuk a kiválasztott

mintákon belül 5 és 100% közötti tartományban, 5 százalékos lépésekkel.

A munkafolyamat során az alacsony vagy változó lefedettségű adatokkal kapcsolatos technikai hibák értékeléséhez az 1000 Genom Project 3. fázisú adatállományából 1020, különböző populációkból származó, teljesen tipizált diploid eurázsiai mintát választottunk. A "depleteIndivs" alkalmazásával egy véletlenszerű, részben tipizált minta kohortot hoztak létre, amelynek markerszámai 100 000 és a teljes, 1 150 639 markert tartalmazó készlet között mozogtak. A populációk a következők voltak: ibériai (IBS), nagy-britanniai (GBR), finn (FIN), Toszkán (TSI), észak- és nyugat-európai felmenőkkel rendelkező utahi lakosok (CEU), dai kínaiak (CDX), bejingi han kínaiak (CHB), déli han kínaiak (CHS), japánok (JPT) és khin vietnamiak (KHV).

Ezt követően a részben tipizált diploid adathalmazból a "pseudoHaplo" eszközzel pszeudohaploidizált adathalmazt hoztunk létre. A rokonsági elemzést a PCAngsd segítségével végeztük el, és a becsült rokonsági együtthatókat a részlegesen genotipizált adatkészleteken

belül a mintapárok marker-átfedési aránya alapján korrigáltuk. Ezeket az eredményeket összehasonlítottuk az eredeti, teljes mértékben tipizált diploid adathalmazon kapott eredményekkel az elemzéshez.

Az aDNS-sel kapcsolatos genotipizálási hibák szimulációja

A PLINK és az EIGENSTRAT adatformátumokat biallelikus markerekhez tervezték. Csak négy lehetséges allélállapot létezik (homozigóta major allél, homozigóta minor allél, heterozigóta major / minor allél és hiányzó), így a minor- vagy major alléltől eltérő bármely más nukleotid nem reprezentálható, és az érvénytelen alléllal rendelkező minták allélállapota az ilyen markerpozíciókban a "hiányzó" állapotra van beállítva.

Az adatformátum-korlátozás alapján a három tipikus aDNS-hez kapcsolódó genotípushiba a következő módon szimulálható a pszeudohaploid PLINK-adatkészletre:

- post mortem károsodás,

- exogén (nem emberi DNS) szennyeződés,

- endogén (emberi DNS) szennyeződés.

Konklúzió

Munkánk során megvizsgáltuk a PC-Relate algoritmust, valamint annak fejlesztési lehetőségeit. Megvizsgáltuk az algoritmus korlátait és hatékonyságát különböző paraméterek mellett.

Az egyik ilyen paraméter a véletlenszerű pszeudohaploidizáció hatása volt a főkomponens-elemzésen alapuló módszerre és a rokonsági együtthatók meghatározására, amelyről kimutattuk, hogy egyikre sincs jelentős hatása. Ezt követően megvizsgáltuk az átfedő markerek számának hatását a rokonsági együttható értékére véletlenszerű lefelé mintavételezéssel, ami lineáris kapcsolatot eredményezett. Ezt felhasználva kidolgoztunk egy korrekciós módszert, amelyben az átfedő markerek számát elosztjuk a markerek teljes számával, majd ezt az értéket elosztjuk a kapott rokonsági együtthatóval. Megmutattuk, hogy ezzel a korrekcióval pontos rokonsági becsléseket lehet kapni még az átfedő markerek alacsony száma esetén is.

Saját algoritmusunk segítségével megvizsgáltuk a korrigált rokonsági együtthatót különböző genotipizálási hibák mellett. Összesen öt kísérleti elrendezést vizsgáltunk: genotipizálási hiba nélküli vizsgálat; endogén szennyeződés; exogén szennyeződés; postmortem károsodás; és az összes genotipizálási hiba együttesen. A legnagyobb különbség az exogén szennyeződés esetén volt megfigyelhető, amely közel 10%-kal csökkentette az együttható értékét.

A szimulációs vizsgálatok utolsó csoportjában a referenciapopulációnak a vizsgálati eredményekre gyakorolt hatása érdekelt bennünket. Ebben az esetben az 1KG adatbázisból három ismert rokonságot és egy erősen kevert magyar-grönlandi családot vizsgáltunk. Az előbbi esetben az volt az eredmény, hogy ha nem ismerjük pontosan egy adott mintapár populációját, akkor csak a megfelelő szuperpopuláció felhasználásával érhetünk el kielégítő eredményeket. Az erősen kevert család esetében mindkét szuperpopuláció használata szükséges volt.

A módszer validálását két adathalmazon végeztük el. Az egyiket egy másik módszer, a READ segítségével

publikálták egy 5 férfitagot tartalmazó zsinórdíszes kultúra családon. Minden rokonsági kapcsolatot fel tudtunk fedezni, beleértve a másik algoritmus által azonosítottokat is, valamint a másodfokú és számos harmad- és negyedfokú rokonsági kapcsolatot. Egy másik esetben egy ismert apa-fiú kapcsolat középkori családjának több, különböző lefedettségű szekvenciáját hasonlítottuk össze, és sikeresen azonosítottuk a rokonsági kapcsolatot a lefedettségi különbségek ellenére.

A rokonsági kapcsolatokat az Allen Ancient DNA Repository 44.2-es verziójában vizsgáltuk meg, miután kiszűrtük a minta kor és a földrajzi elhelyezkedés alapján, ahogyan azt az „Anyagok és módszerek” részben említettük. Sikeresen kimutattunk számos új rokonsági kapcsolatot és mintaszennyezést is.

Végül a módszer segítségével sikeresen igazoltuk a Győrben őrzött Szent László koponya ereklye hitelességét, aki az első olyan katolikus szent, amit genetikailag igazoltunk.

Összefoglalva, az általunk javasolt módszer képes a rokonsági viszonyok 4. fokig történő megbízható

azonosítására alacsony lefedettségű genomadatokból, újradefiniálva a rokonsági analízis határait alacsony lefedettségű ősi vagy erősen degradált törvényszéki WGS adatokból.

Irodalomjegyzék

1. Shyla A, Borovko SR, Tillmar AO, et al. Belarusian experience of the use of FamLinkX for solving complex kinship cases involving X-STR markers. *Forensic Sci Int Genet Suppl Ser*. 2015;5:e539-e541. doi:10.1016/j.fsigss.2015.09.213

2. He G, Adnan A, Al-Qahtani WS, et al. Genetic admixture history and forensic characteristics of Tibeto-Burman-speaking Qiang people explored via the newly developed Y-STR panel and genome-wide SNP data. *Front Ecol Evol*. 2022;10(October):1-19. doi:10.3389/fevo.2022.939659

3. Allen Ancient DNA Resource. No Title. https://reich.hms.harvard.edu/allen-ancient-dna-resource-aadr-downloadable-genotypes-present-

day-and-ancient-dna-data

4. Auton A, Abecasis GR, Altshuler DM, et al. A global reference for human genetic variation. *Nature*. 2015;526(7571):68-74. doi:10.1038/nature15393

5. Nyerki E, Kalmár T, Schütz O, et al. correctKin: an optimized method to infer relatedness up to the 4th degree from low-coverage ancient human genomes. *Genome Biol*. 2023;24(1):1-21. doi:10.1186/s13059-023-02882-4