

Szegedi Tudományegyetem
Természettudományi és Informatikai Kar

Biológia Intézet
Genetika Tanszék
Archeogenetikai kutatócsoport

**Imputáció az Archeogenetikában: Módszertani kihívások
és lehetőségek a genom rekonstrukció javítására**



Bán Botond

Molekuláris bionika mérnök alapszak

TDK dolgozat

Témavezetők:
Maróti Zoltán PhD
Török Tibor PhD

Szeged, 2024

A KULTURÁLIS ÉS INNOVÁCIÓS MINISZTERIUM EGYETEMI KUTATÓI ÖSZTÖNDÍJ PROGRAMJÁNAK
A NEMZETI KUTATÁSI, FEJLESZTÉSI ÉS INNOVÁCIÓS ALAPBÓL FINANSZÍROZOTT SZAKMAI
TÁMOGATÁSÁVAL KÉSZÜLT.



Tartalom

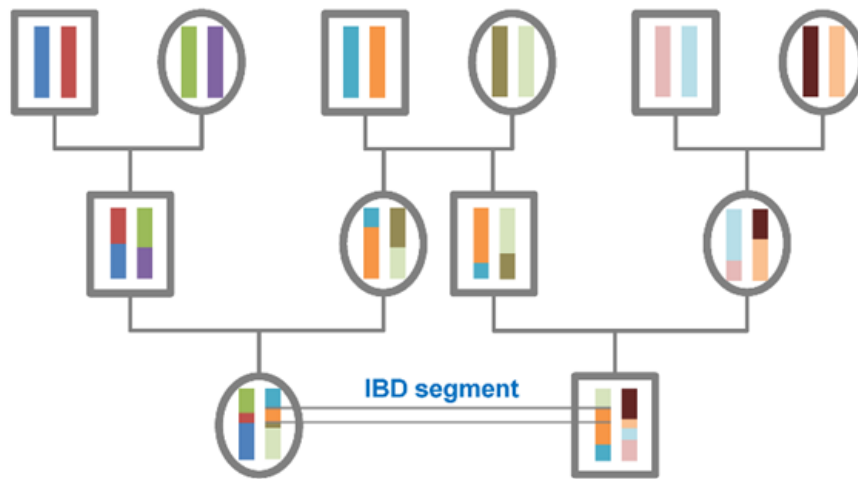
Tartalom.....	2
Bevezetés.....	3
Archeogenetika	3
Szekvencia formátumok a szekvenálástól az imputálásig	8
Imputáció	10
Az imputáció létjogosultsága.....	11
Célkitűzés	13
Anyagok és módszerek.....	14
Molekuláris biológia laboratórium	14
Algoritmusok, szoftverek	17
Önálló imputálás.....	29
Minták és eljárások.....	31
Eredmények értékelése.....	33
Köszönetnyilvánítás	36
Irodalomjegyzék.....	37

Bevezetés

Archeogenetika

Az archeogenetika egy olyan tudományág, amely ősi DNS elemzésével foglalkozik annak érdekében, hogy megértse a múltbeli populációk genetikai összetételét, keveredését, eredetét és migrációját. Interdiszciplináris tudomány, mert az adatok értelmezése során ötvözi a genetika, az archeológia és az antropológia ismereteit. A mi munkánkban ősi emberi populációkat vizsgálunk. A Humán Genom Projektben több ezer ma élő ember genom szekvenciáiból több 100 ezer változékony SNP-t (Single Nucleotid Polimorfizmus) azonosítottak, és az analízisek ezekre összpontosulnak. Ezek túlnyomó többsége (>99%-a) evolúciós szempontból indifferens (Auton és mtsai., 2015), vagyis nem hat rájuk a szelekció. Így öröklődésük teljes mértékben a mendeli törvényeknek megfelelő, allélgyakoriságukat elsősorban genetikai sodródás és a migráció befolyásolja és tökéletesen alkalmasak statisztikai elemzésre. Az archaikus genomokból megállapíthatók az egykor élt egyének rokonsági viszonyai, valamint az egyéb, nem rokon genomokhoz való hasonlóságuk szintje. Utóbbi lehetővé teszi a populáció-genomikai elemzést, mert a genomok összehasonlíthatók az adatbázisokban lévő ismert genomokkal, és hasonlóságuk alapján csoportosíthatók az egykori populációk. Megállapítható, hogy az egyes egyének genomja mely ősi populációkból származik és milyen arányban. Ezen eljárások matematikai módszereket, statisztikai analízist és valószínűségszámítást használnak. A leggyakrabban használt eljárások: PCA (Principal Component Analysis), F2-, F3-, F4-statistics, ADMIXTURE analysis (biogeographical ancestry analysis), qpAdm, IBD (Identity by Descent l.: 1. ábra), kinship Analysis. Fontos megjegyezni, hogy az archaikus genomok túlnyomó többségét alacsony lefedettséggel szekvenálják, így azok hiányosak. Ennek ellenére a fenti algoritmusok többsége jól működik ezeken a hiányos pszeudohaploid genomokon is. Azonban a legújabb eljárásokra, mint például az IBD sharing (Browning, 2008), kizárólag diploid, lehetőleg phaselt genomok alkalmasak, mivel ezek nem egyedi markereket vizsgálnak hanem egy hosszabb egybefüggő genom szakasz összes markerét tartalmazó haplotípusokat, amelyben lehetnek kisebb arányban hibás adatok is. Ezért az IBD analízis archaikus genomokra csak akkor alkalmazható, ha előtte a hiányos pszeudohaploid genomokat diploid, phaselt genomokká tudjuk kiegészíteni. Ezt a kiegészítést az úgynevezett genom imputációval lehet végrehajtani. Az imputáció azt a biológiai tényt használja ki, hogy a markerek többsége kapcsoltnak, nem egymástól függetlenül öröklődik. Az egyes genom lokuszokban minden populációra jellegzetes haplotípus kombinációk jellemzőek. Az imputáció során így egy megfelelően nagy referencia populáció haplotípus adataiból, az adott mintában a két valós allél

véletlenül mintázott, hiányos genotípus adatai alapján, a hiányos adatokat legjobban leíró egy vagy két haplotípust (allélt) igen nagy pontossággal meg lehet találni. Így az imputációval phase-elt, diploid genotípusokat kaphatunk, ami alap feltétele az IBD sharing analízisnek.



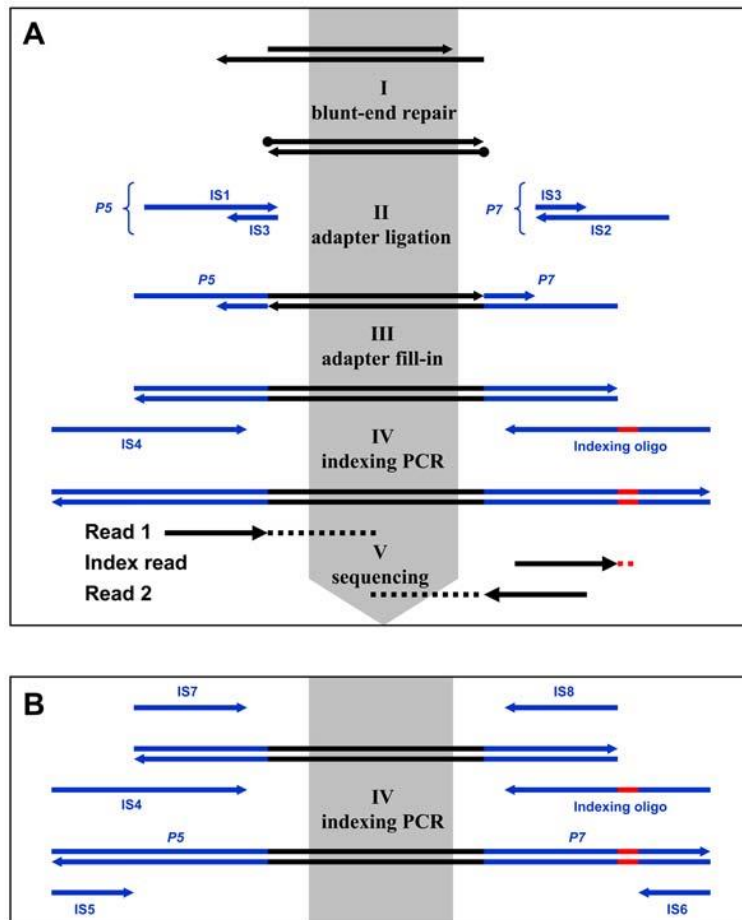
1. ábra: az IBD elemzés lényege a közös felmenőktől örökölt azonos genom szakaszok azonosítása. Az ábra egy 3 generációs családban öröklődő egyik IBD szakasz öröklésmenetét mutatja be. (https://isogg.org/wiki/Identical_by_descent Hozzáférés dátuma: 2024.11.08.)

A régészek által feltárt maradványokból nyerhető DNS sokban eltér az élő szövetből vett örökítőanyagtól. A földfelszín alatt töltött több száz vagy akár ezer év elsősorban öt különböző módon rontja a kinyerhető DNS minőségét (Garrido Marques és mtsai., 2024):

- Szennyeződés vagy kontamináció: Az elhalt szövetekben egy idő után túlnyomó többségben idegen, nem humán eredetű örökítőanyag (első sorban talajbaktériumok DNS-e) halmozódik fel.
- Láncszakadás, feldarabolódás: a DNS lánc a halált követően véletlenszerűen darabolódik. Ez a folyamat az idő előrehaladtával egyre rövidebb DNS szakaszokat hoz létre, melyek többnyire 5' vagy 3' túlnyúló végekkel rendelkeznek.
- Abázikus helyek: nem csak a cukor-foszfát gerincben található észter kötések, hanem a bázisok glikozidos kötései is felhasadhatnak, melyek abázikus helyeket, depurinációt, depirimidinizációt hoznak létre.
- Oxidáció, metiláció, dezamináció, tranzíció (ritkán transzverzió): oxidatív körülmények között, illetve különböző reagensek jelenlétében a bázisok szerkezete, funkciós csoportjai megváltozhatnak, leszakadhatnak illetve kicserélődhetnek, melynek következtében más bázisok, vagy bázis analógok jöhetnek létre, melyek a

szekvenáláskor hibás leolvasást eredményeznek. Az aDNS leggyakoribb módosulása a citozin dezaminációja, melynek következtében uracillá alakul, ami a szekvencia olvasatban timinként jelentkezik.

- Egyéb szerkezeti elváltozások: az oxidáció károsíthatja a cukor-foszfát gerincet is, mely nem feltétlen eredményez szakadást, de megnehezíti a szekvenálást. Ezen kívül létrejöhetnek timin dimerek is (első sorban UV hatására), de a sugárzások, reagensek és mikroorganizmusok számtalan más módokon is megváltoztathatják a DNS szerkezetét.

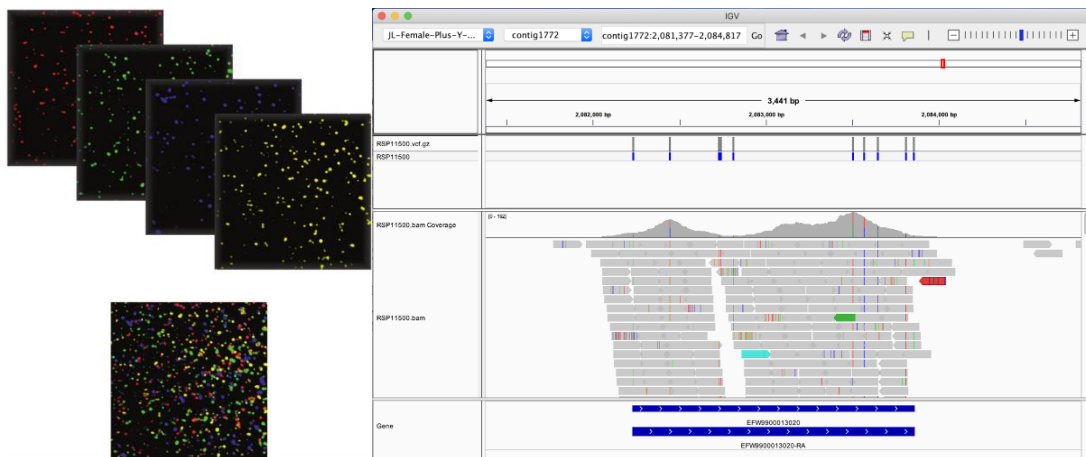


2. ábra: ősi DNS NGS szekvenálás előkészítő protokoll (Meyer & Kircher, 2010).

Az archeogenetikai vizsgálatok elvégzéséhez mindenekelőtt ezen hibákat kell kezelni, hogy minél megbízhatóbb legyen a szekvencia olvasat. Precíz mintavételezéssel, csökkenthető a kontamináció, UDG (Uracil-DNA Glycosylase) kezeléssel eltávolíthatók az uracilok, hibajavító enzimekkel (Bst polimeráz) kijavíthatók az abázikus helyek, de az összes hiba javítása így sem lehetséges, így ezeket csak informatikai módszerekkel lehet korrigálni. Ez után következhet az NGS előkészítése (l.: 2. ábra).

NGS

Az Újgenerációs Szekvenálás (Next Generation Sequencing, NGS) kiváló eszköz az archaikus DNS szekvenálására (l.: 3. ábra). A módszer több millió rövid, maximum néhány száz bázisból álló DNS szakasz párhuzamos szekvenálására alkalmas. Mivel az archaikus DNS természetéből adódóan rövid DNS darabokból áll, jól megfelel ezen paramétereknek. A hibridizációs genom dúsításokkal ellentétben itt nem csak az 1240k meghatározott régióiról nyerünk genetikai információt, hanem az elérhető genom teljes egésze szekvenálásra kerül. Mivel az ugyanazon DNS szakaszról nagyszámú párhuzamos olvasat (read) készül, az egyikben meglévő téves vagy bizonytalan olvasat kijavítható a párhuzamos információkból. A csoportunk által használt Novaseq6000 szekvenátor kapacitása olyan nagy, hogy egyszerre 40-60 egyén genomja is szekvenálható. Ezért minden egyes mintából készült könyvtárat egyedi 7 nukleotid méretű indexeket tartalmazó adapterekkel jelölünk (l.: 2. ábra).



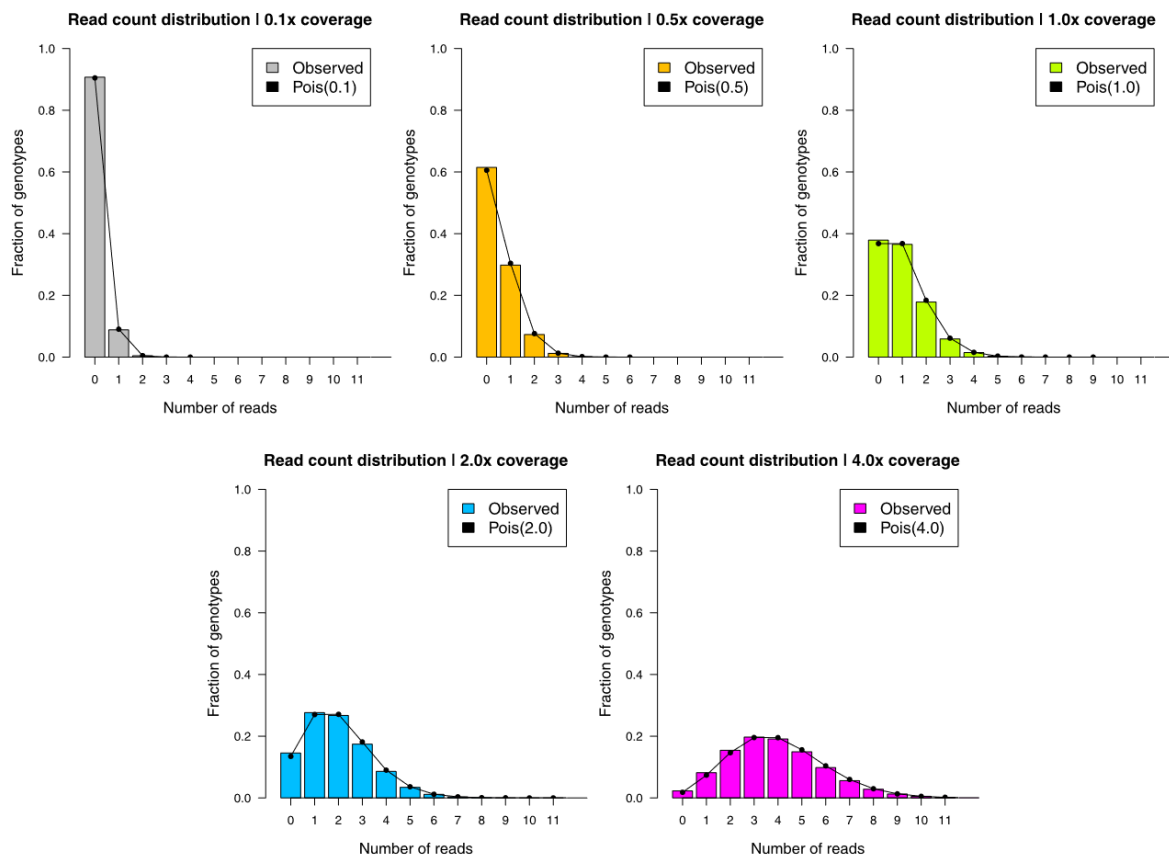
3. ábra: A: bázis lumineszcencia (Voelkerding és mtsai., 2009).

B: a szekvencia megjelenítése IGV formátumban.

(<https://help.medicinalgenomics.com/en/getting-started-igv> Hozzáférés dátuma: 08 11 2024)

Az alacsony humán DNS tartalmú minták szekvenálására kifejlesztették a genom dúsítás hibridizáció alapú (targeted oligo capture) módszerét (Lee és mtsai., 2009), mellyel kizárólag a változékony pontokat tartalmazó genom darabkák kerülnek dúsításra és szekvenálásra. A rutinszerűen alkalmazott hibridizációs dúsító kit 1240k markert tartalmazó target régiót dúsít, amely a leggyakoribb magas allél frekvenciájú (Alternatív vagy Minor Allele Frequency, MAF) variánsokat tartalmazza. Azonban ezzel az eljárással a nem dúsított genom szakaszok információja elvész. A jobb megtartású, magasabb endogén DNS tartalmú mintáknál jellemzően a shotgun WGS (Whole Genome Sequencing) eljárást alkalmazzák, amely a teljes

genomot szekvenálja. Az ősi DNS minták esetén pénzügyi okokból többnyire alacsony (0,1-2x) genom lefedettségű adatok keletkeznek. A kapott rövid szekvenciákat szoftveresen illesztik a referencia humán genomra, ennek során a nem humán DNS szennyeződés nagy része szoftveresen eltávolítható. Mivel az egyes DNS molekulák szekvenálása teljesen véletlenszerű, az alacsony lefedettség miatt előfordul, hogy egy DNS szakasról több olvasat is készül míg másokból egy sem (l.: 4. ábra). Az átlagosan egy bázisra eső 'read'-ek számát nevezzük lefedettségnek. Ha egy mintában alacsony az endogén DNS tartalom, akkor a szekvenálaskor többnyire kisebb lesz a genom lefedettsége, mivel sok lesz a feleslegesen megszekvenált exogén DNS. Ebből csak akkor nyerhetnénk információt, ha a többi minta rovására növeljük az alacsony endogén tartalmú minta koncentrációját, de ez rendkívül költséges, ezért inkább endogén tartalom alapján szelektáljuk a mintákat, és egy bizonyos százalék alatt nem végzünk WGS-t. A költséghatékonyság és a pontosság kompromisszumaként 1x és 2x átlag lefedettséggel szekvenáljuk a mintákat.



4. ábra: A readok eloszlása különböző átlag lefedettségek esetén (Rubinacci és mtsai., 2021).

Ennek következtében sok vizsgálandó pontról nem kapunk még csak haploid információt sem, diploid információ kinyerése pedig még ennél is ritkább. Ezért az ilyen hiányos humán genom olvasatot pseudo-haploidnak tekintjük az archeogenetikai elemzések során. A szekvenálás és illesztés során a szoftverek egyes nukleotid olvasatokhoz valószínűségi értékeket rendelnek, ami jellemzi az olvasat bizonytalanságát. Ezt a BAM fájlok a hiba valószínűségének logaritmusaként tárolják.

Alacsony lefedettség

A szekvenálás lefedettség olyan mérőszám, mely megmutatja, hogy a genom egy pontját átlagosan hányszor olvasta ki és azonosította a szekvenátor. Az archeogenetikában szokásos 1x és 2x átlagos lefedettség esetén a genomi pontok többségéről legalább 1 olvasatot kapunk, de szignifikáns számban fordulnak elő 3x, 4x, vagy akár 5-szörösen azonosított bázispárok is (l.: 4. ábra), azonban a továbbra is jelentős számú ismeretlen pont miatt még ez sem elegendő az IBD analízishez szükséges biztos diploid információ kinyerésére.

Szekvencia formátumok a szekvenálástól az imputálásig

1. Kép:

Szekvenálás: A szekvenálás során az elongációs ciklusban beépült új bázisokat fluoreszcens festékekkel jelölik. Digitális képeken rögzítik ezeket a fluoreszcens jeleket, amelyek a DNS szekvencia nyers adatait képviselik (l.: 3. ábra).

2. BCL (Base Call Files):

Szekvenálás utáni adatkimenet: A BCL fájlformátum a szekvenáló gépből származó nyers bázishívási (base calling) adatokat tartalmazza, beleértve az egyes bázisok hívásait és minőségi pontszámait. Ezek az adatok közvetlenül a szekvenálási folyamatból származnak.

3. FASTQ:

Szekvencia adatainak tárolása: A FASTQ fájlformátum a mintánként szétválogatott szekvencia adatokat és azok minőségi pontszámait tartalmazza. Minden egyes nukleotidhoz tartozik egy minőséget jelölő sor, amely az olvasatok megbízhatóságát jelzi.

4. SAM (Sequence Alignment/Map):

Illesztés (alignment): A SAM az egyes readok pozícióját, illesztési pontszámait és egyéb adatokat rögzíti. Szöveges formátum.

5. BAM (Binary Alignment/Map):

Illesztés és tárolás: A BAM fájl a SAM fájl bináris verziója, amely hatékonyabb tárolást biztosít a BGZF (Blocked GNU Zip Format) nevű tömörítési eljárásnak köszönhetően. Az illesztett szekvencia adatokat tartalmazza.

6. BAI (BAM Index):

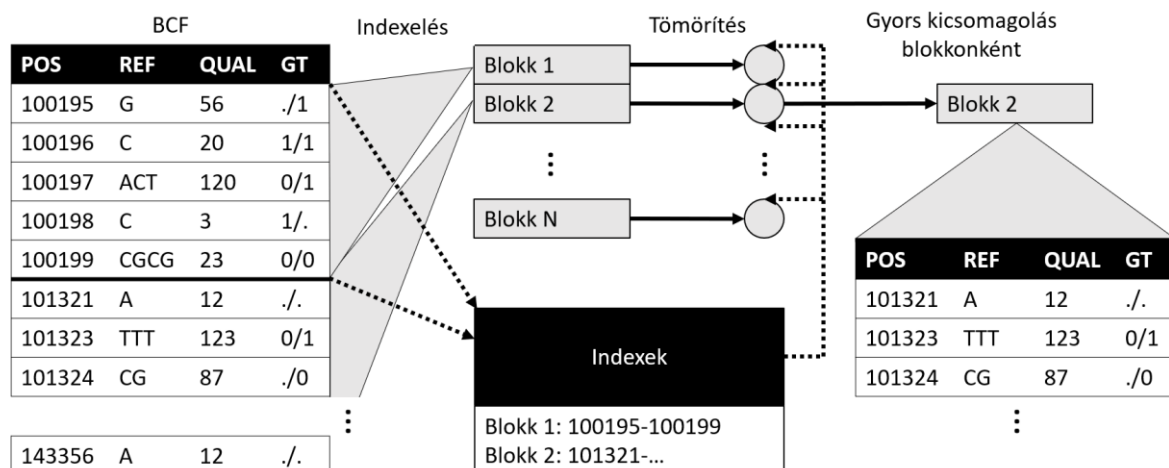
Gyors hozzáférés a BAM fájlokhoz: A BAI fájl a BAM fájl indexelt változata, amely lehetővé teszi a gyors hozzáférést a BAM fájl specifikus részeihez.

7. VCF (Variant Call Format)/BCF (Binary Call Format):

Variáns hívás: A VCF fájlformátum a szekvenciákban található genetikai variánsokat (SNP-k, inszerciók, deléciók) tartalmazza. Az egyes variánsok pozícióját és genotípusát (GT) rögzíti. VCF esetén Text formátum, amely tartalmazza a variánsok információit és a hozzájuk tartozó genotípusokat, A BCF a VCF bináris változata. Ugyanazokat az információkat tartalmazza, mint a VCF, de bináris formában tárolja, ami hatékonyabb tárolást és gyorsabb adatfeldolgozást tesz lehetővé.

8. BGZ (Block GZIP):

Adatkompresszió: A BGZ fájlformátum a fájlok (esetünkben VCF-ek) hatékony tárolására és tömörítésére szolgál. A ZIP és a „zipelés” egy informatikában elterjedt úgynevezett archiváló tömörítési eljárás, mely megőrzi a tömörített mappa struktúráját. A GZIP (GNU Zip) egy hatékonyabb, Unix által támogatott nem-archiváló tömörítési eljárás, mely egyszerre csak egy fájl tömörítését végzi, nem képes mappa struktúrákat kezelni. A GZIP és a BGZIP (Block GZIP) fájlok abban térnek el egymástól, hogy míg előbbi csak egy hatékonyan tömörített adatállományt hoz létre, utóbbi még a tömörítés előtt blokkokra osztja a fájlt, és azt blokkonként tömöríti. A BGZIP tömörítéshez tartozik egy index gyűjtemény, amely egy külön szöveges fájlban (.tbi/.csi) tárolja az eredeti fájl koordinátáinak megtalálhatóságát a tömörített blokkokban a tömörített állomány mellett (genom.bam + genom.bam.bai, genom.bcf.gz + genom.bcf.gz.tbi). Ez a fajta szakaszolt tárolás kis mértékben csökkenti a tömörítés hatékonyságát, ugyanakkor számottevően felgyorsítja a tömörített fájlok hozzáférését.



5. ábra: A BGZ tömörítés folyamata.

9. Egyéb segéd formátumok: TSV (Tab-Separated Values), TXT (Text Files):

Kiegészítő adatok tárolására: chunkolás, indexek, vagy akár két BAM fájl eltéréseinek tárolására szolgál.

Imputáció

Az alacsony lefedettségű szekvencia hiányosságainak kiegészítésére a közelmúltban kifejlesztették a genom imputáció módszerét. Az imputáció a statisztikában az a folyamat, amely során a hiányzó adatokat pótolják becsléssel vagy valószínűségi módszerekkel (Leonard E. Baum & Ted Petrie, 1966; Y. Li és mtsai., 2009, 2010). A genetikában az imputáció a hiányzó genotípusok becslésére szolgál, referencia panelek és a szekvenálásból származó meglévő adatok és valószínűségi értékek segítségével. A genetikai imputáció célja a hiányzó genetikai variánsok kiegészítése, melynek segítségével a hiányos pseudo-haploid genomokból diploid genomok készíthetők, ami sokkal mélyebb szintű genom elemzéseket tesz lehetővé. A meiózis során a kromoszómák átkereszteződnek, így az apai és anyai eredetű homológok kromoszóma szakaszai kicserélődhetnek (crossing-over). Minél közelebb van egymáshoz két lókus, annál valószínűbb, hogy együtt öröklődnek tovább, haplotípusonként nagy blokkokat képezve. Egy jó minőségű modern diploid referencia adatbázisában megtalálhatók az adott populációra jellemző marker sorozatok, haplotípus kombinációk. Az ősi DNS-ből csak pseud-haploid információt nyerhetünk, azaz bár két allélja van, de véletlenszerűen vagy egyik vagy másik allélról van marker információnk, és azt sem tudjuk, hogy melyik ezek homológokról származnak. Az imputációs algoritmusok képesek megtalálni a referenciában található

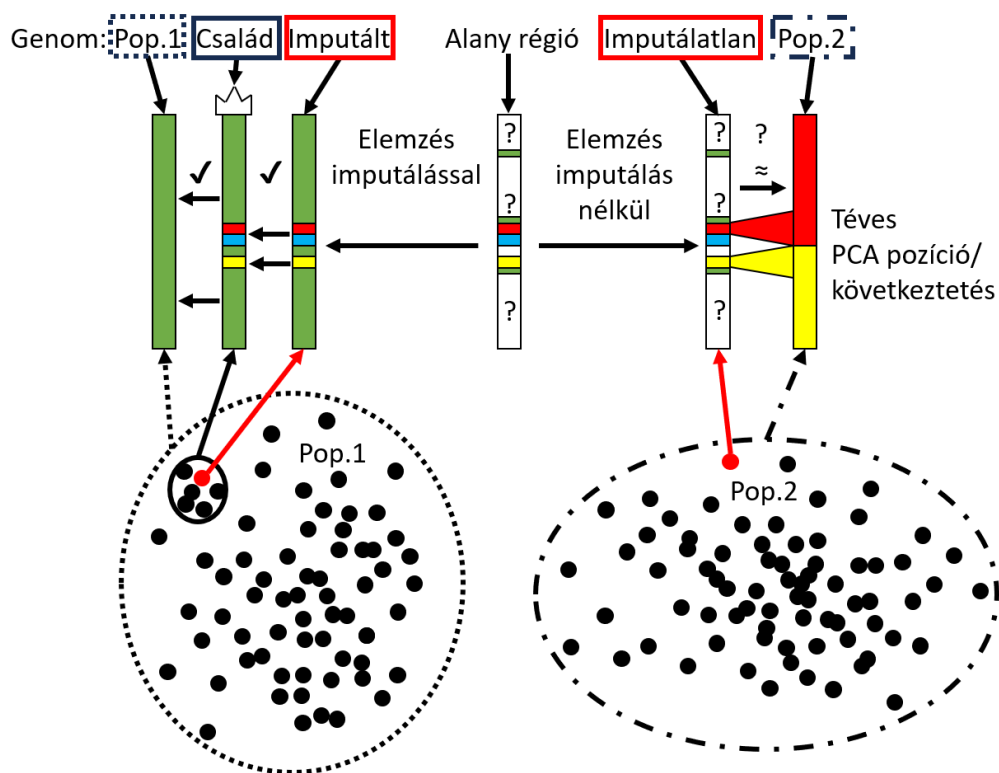
haplotípus kombinációkból azt az egy-két haplotípust, amivel a pseudo-haploid markerek legjobban leírhatók. Így végül az összes genom szakaszra megadja azokat a diploid markereket amivel az adott pseudo-haploid genotípusok legnagyobb valószínűséggel leírhatók. Az imputáció pontossága a lefedettség mértékétől, a mintának megfelelő referencia populáció meglététől, és az abból rendelkezésre álló genomok számától függ. Európai populációk esetén, melyekből nagyszámú referencia áll rendelkezésre, a $>1\times$ lefedettségű shotgun szekvenciák $>5\%$ MAF (alternatív allél gyakoriság, Minor Allele Frequency) értékű pontjai $>98\%$ -os pontossággal imputálhatók (Rubinacci és mtsai., 2021).

Az imputáció létjogosultsága

Felmerülhet a kérdés, mi értelme van referencia segítségével kipótolni egy hiányos adatszettet, majd az adatszettekben vett információk segítségével (beleértve a kipótolt adatpontokat is) vizsgálni azok hasonlóságait, milyenségét. Jogos a kérdés, hiszen bizonyos esetekben (alacsony lefedettség, alacsony MAF, ritka haplocsoport) ez az eljárás valóban a valóságtól eltérő genom szakaszokat rekonstruálhat. Például a nagyon kevés egyénre jellemző, nagyon alacsony MAF értékű SNP változatok többsége biztosan nincs reprezentálva az adtabázisban, így azok tévesen lesznek imputálva, a hozzájuk leghasonlóbb a haplotípusokból. Azonban ez nem okoz jelentős problémát, mert a populációgenetikában az egyénre jellemző pontmutációknak nem tulajdonítanak nagy jelentőséget, sokkal fontosabbak a magas MAF értékű SNP mintázatok (l.: 6. ábra), melyek a elterjedtek, és több egyedben is megtalálhatóak. Az ilyen magasabb MAF-val rendelkező variánsok, éppen elterjedésük miatt nagy valószínűséggel megtalálhatóak a referenciákban is. Emiatt lényeges a megfelelő referencia használata, kiváltképp, ha egy ritka genetikai összetételű populációt vizsgálunk.

Miért szükséges az imputáció? Miért nem mindig a hiányos, de biztos adatokat hasonlítjuk össze? Bizonyos esetekben, mint a PCA vagy az F statisztikák elegendő egy kis lefedettségű szekvenálás ahhoz, hogy kellő mennyiségű adatot kapjunk és azt össze tudjuk hasonlítani más hasonló minőségű adatszettekkel, viszont az IBD és a többi nagy egybefüggő szakaszokat váró algoritmus már nem fogja tudni összehasonlítani a két hiányos pseudo-haploid adatsort, mivel ezeknek szüksége van a diploid allélhívásra az öröklődési mintázat lekövetése céljából. Például előfordulhat, hogy egy minta magában hordozza részletekben a vizsgált királyi családra jellemző IBD mintázatot, de maga a teljes genom annyira hiányos, hogy ha ebben az állapotban próbálunk IBD szegmenseket találni, akkor csak kevés rövid szegmenst találunk, így a családtól tévesen távolra tesszük a mintát geneológiailag (l.: 6. ábra). Itt kap teret az imputáció, ami a lyukak környezetének ismeretében (amennyiben az IBD mintázat megtalálható a referenciában)

megfelelően ki tudja egészíteni a genomot saját populációjának szakaszaival, majd pedig az analízis immáron mint teljes, hiánytalan genomot, pontosan el tudja helyezni a királyi család közelében az apró imputációs hibák ellenére.



6. ábra: Egy fiktív alany vizsgálata. Magyarázatát lásd a szövegben.

Sokban hasonlít ez a képelemző algoritmusok működéséhez. Tegyük fel, hogy „P1” számláló program tökéletesen felismeri és körvonalazza egy forgalmas kereszteződésben készült képen az összes emberi alakot, bármilyen pozícióban, távolságban legyenek és bármilyen küllemi jellemzőkkel bírjanak (l.: 7. ábra, sárga színnel jelölt kiválasztás). Ezután a jól elkülönített alakokat megszámlolja. Ez felel meg az IBD analízisnek, mely nagy megbízhatósággal végzi a dolgát hiánytalan adatsoron. A probléma akkor adódik, mikor az elemzendő képen az egyik alakot szinte teljesen kitakarja például egy tábla, és csak egy csizma vagy egy kalap látszik a képen. Ekkor jön a képbe „P2” segéd program, mely úgy van megalkotva, hogy bár tökéletesen nem tudja körbe rajzolni az emberi alakokat, azt meg tudja mondani, hogy ahol egy csizma vagy kalap van, ahhoz nagy valószínűséggel tartozik egy ember is (l.: 7. ábra, piros színnel jelölt kiválasztás). P2 program P1 programnak továbbítja ezt az információt, ami így helyes következtetéseket képes levonni. Természetesen lehetséges, hogy csak önmagában volt egy csizma a képen, vagy olyan részletet ismert fel, amely nem tartozik személyhez, (l.: 7. ábra,

kérdőjellel jelezve) ami minimális bizonytalanságot visz a mérésbe, pontosan úgy, ahogyan az genomi imputáció esetén sem lehetünk biztosak az összes behelyettesített bázis genotípusában, de szélsőséges esetekben (=alacsony lefedettség), mikor túlnyomó részt lábbeliket és kilógó végtagokat látunk a képen egyértelmű, hogy egy P2 segéd program nélkül nem megvalósítható a számlálás, mivel a teljesen körülhatárolható alakok szám a elenyésző, viszont azt nagy bizonyossággal tudjuk, hogy a forgalmas kereszteződésben valójában nem hagytak gazdátlan lábbeliket és végtagokat. Tehát csak akkor alkalmazzuk az imputációt, ha tudjuk, hogy a vizsgálódásunk (IBD) nagyobb hibával dolgozna önmagában, mint amit az imputáció belevisz a mérésbe, ami alacsony lefedettség esetén megfelelő referenciával szinte mindig igaz. Ez természetesen egy komoly leegyszerűsítés, mivel az imputáció nem egyszerűen arról dönt, hogy van-e az adott ponton genotípus, hanem hogy milyen genotípus van ott (inkább sárga vagy inkább piros az átlátszó kabát).



7. ábra: P1 és P2 fiktív programok működésének szemléltetése.

(kép: <https://www.usnews.com/news/cities/articles/2020-10-09/the-us-cities-most-vulnerable-to-covid-19-and-poor-mental-health>. Hozzáférés: 2024.10.15.)

Célkitűzés

Kutatásunk egyik célja az volt, hogy megvizsgáljuk, hogy a Joint Model megfelelő alkalmazása javíthat-e a konkordancián. Ennek kivitelezéséhez öt minta lefedettségét mesterségesen lerontottuk, ezek imputáltuk párhuzamosan, valamint a referenciapanelhez hozzáadtunk további 352 korabeli, azonos területről származó (így a minták genotípusához vélhetően legjobban hasonló) archaikus mintát is. Minden imputálás csak a 22-es kromoszómán történt a korlátozott gépidő miatt. Az imputáció után a konkordanciát az eredeti (lerontás előtti) magas lefedettségű genomokkal összehasonlítva számoltuk.

Anyagok és módszerek

Molekuláris biológia laboratórium

Csontpor előállítása

A csontok és fogak tömörebb állományai jobb DNS-megőrzési képességgel bírnak, mint a lágy szövetek. A temporális koponyaacsont sziklacsontja, különösen a *labyrinthus osseus*, kiemelkedően tiszta endogén DNS-tartalmú. A sziklacsont rendkívül sűrű és ellenálló, ezért a halál után is hosszú időn keresztül konzerválja az endogén DNS-t, emiatt az ősi DNS-kutatások egyik legfontosabb mintavételi célpontja.

A sziklacsont mintákat tisztítás után Dremel magasfordulatszámú kézi maróval választottuk le. A folyamat során keletkező csontport elszívó berendezés távolítja el. A leletek állapotától függően különböző technikákat alkalmaztunk: teljes koponya esetén minimálisan invazív megközelítést, vagyis a csont megfúrását, különálló csontdarabnál közvetlen vágást. A darabolt mintákat golyósmalomban aprítottuk porrá, melyet steril mikrocentrifuga-csővekben tároltunk. Az eszközöket és az elszívófülkét minden használat után alaposan tisztítottuk 3%-os hipoklórossavval, 3%-os hidrogénperoxiddal és DNS mentes $\gamma\text{H}_2\text{O}$ -val (a továbbiakban: víz) valamint UV-fény alatt DNS-mentesítettük.

DNS extrakció

Az ősi DNS esetében nagyon alacsony endogén DNS-tartalomra lehet számítani. Negatív kontrollokat alkalmaztunk a szennyeződések kiszűrésére.

Az extrakció első lépése a csontpor előemésztése, mely során a por részecskéinek felszínén megtalálható, potenciálisan nagyobb mennyiségű szennyező (exogén) DNS eltávolítása történik. 200mg csontporhoz 1ml 0,5 M-os EDTA-t és 5 μl 20mg/ml-es Proteináz K szuszpenziót mértünk. A mintákat az előemésztés során 30 percig 48°C-on folyamatos rázatás mellett inkubáltuk. Ez inkubáció végeztével 1 perc 13000 rpm centrifugálás után leöntjük a felülúszót és így a szennyeződés nagyrésztét is.

Ezután következik az emésztés, melynek célja az előemésztés után megmaradt porszemcsék teljes feloldása, a bennük lévő magas endogén tartalmú DNS felszabadítása. Az emésztési reakció: a centrifugálásból megmaradt tisztított csontpor; 1,35 ml 0,5 M-os EDTA; 18,75 μl 20 mg/ml Proteináz K; 15 μl 10%-os Triton X-100® detergens; 1,5 ml víz. A reakciót 72 óráig 48°C-on folyamatos rázatás mellett hagytuk végbe menni.

A DNS-t szilika mátrix segítségével vontuk ki. A mátrixhoz való kötést kaotróp só, guanidinium-hidroklorid segíti. 4×45 ml kötőpuffer, melyből minden mintához 7,5 ml-t használtunk fel 24 mintára összesen: 21,5 g kristályos guanídium-hidroklorid; 1320 µl 3 M-os nátrium-oxálacetát; 18 ml 100%-os izopropanol; 22,5 µl 100%-os Tween-20[®] detergens; valamint a 45 ml teljes térfogat eléréséhez szükséges víz. A nátrium-oxálacetát savas pufferként szolgál, valamint egyértékű kationokat biztosít, melyek vízelvonó hatásuk révén facilitálják a DNS szilika mátrixhoz kötődését. A mintákat vákuummal szívtuk át a MinElute[®] szilika oszlopokon.

A DNS megkötése után mosási lépések következnek, amit 750 µl a gyártó által biztosított mosópufferrel hajtottunk végre, ami vízzel hígított 80%-os etil-alkoholt tartalmaz. A mosópuffert centrifugálással nyomtuk keresztül a mátrixon.

A mátrixon megkötődött DNS-t 2×25 µl TRIS elúciós pufferrel oldottuk le, ami 10 mM-os TRIS és 1mM-os EDTA elegye. Az elúciót centrifugában hajtottuk végre.

A tisztítás végén ellenőriztük az oldataink DNS koncentrációit illetve a kontroll esetleges szennyezettségét Qubit[®] fluorometriás műszer segítségével.

Hibajavítás

Az UDG-USER (Uracil-DNS-Glycosylase, Uracil-Specific Excision Reaction) a citozin bázisok deaminációval történő uracil tranzíciója korrigálható, melyet a DNS ősi jellegének bizonyításaképpen csak részlegesen hajtunk végre.

Az UDG reakció egy mintára, 30 µl DNS kivonathoz: 6 µl 10-szeres töménységű Buffer Tango, 0,24 µl 25 mM-os dNTP mix (4×0,1 M-os dNTP törzsoldatból); 0,6 µl 100 mM-os ATP; 1,8 µl USER reagens 1 U/ µl NEB törzsoldatból; 21,36 µl víz. A reakciót 37°C-on 30 percig inkubáltuk PCR gépben, így nem minden uracil kerül kijavításra, ami később validálja az adatainkat, bizonyítja a DNS archaikus jellegét. A reakció megfelelően időzített leállításához a mintákat párhuzamosan, időben eltolva indítjuk, így a 30 percek pontos leteltével éppen annyi időnk van a reakciót leállító UGI (Uracil-DNS-Glycosylase Inhibitor) rámerésére, mint amennyit kezdetben a reakció összemérésére felhasználtunk.

Az adapterek sikeres ligálása csak ép, foszforilált végű DNS-fragmentumokon végezhető el. Ezért a degradált DNS-re jellemző egyszálú túlnyúló végeket vagy ki kell egészíteni, vagy el kell távolítani. A Blunt-end-repair: 3 µl 10 U/µl T4 polinukleotidkináz; 1,2 µl 5 U/µl T4 DNS polimeráz. A T4 DNS polimeráz 3'-5' exonukleáz aktivitásának köszönhetően a túlnyúló 3'

végeket leemészt, míg az 5' túlnyúló végeket kettős szálúvá egészíti ki. Az 5' végek foszforilációjának helyreállítása érdekében T4 polinukleotid kinázt adunk a reakcióelegyhez, így a DNS alkalmas lesz a ligálásra. Az elegyet 25°C-on 15 percig PCR gépben inkubáltuk. A következő lépés előtt a korábbihoz hasonló módon MinElute oszlopokon tisztítjuk a DNS oldatot, annyi eltéréssel, hogy a kisebb mennyiségű kiindulási reagens megfelelő kihígításához ezzel elegendő 350 µl kötőpuffert alkalmazni, valamint a folyamat végén csak egyszer eluáltunk 20 µl TE pufferrel.

Adapter ligálás

A kapott tompavégű polinukleotidokhoz IS1-IS3 és IS2-IS3 adaptereket ligáltunk (l.: 2. ábra). Az ehhez szükséges reakció: 4 µl 10-szeres T4 DNS ligáz puffer; 4 µl polietilénlikol (50%-os PEG-4000); 1 µl adapter nukleotid mix (50 µM mindegyikből); 1 µl 5 U/µl-es T4 DNS ligáz; 10 µl víz; 30 perc inkubálás 22°C-on. Újabb tisztítás következett 200 µl kötőpufferrel 20 µl TE pufferbe.

Az adapter túlnyúló végeit Bst polimerázzal töltöttük fel a következőképpen: 1,5 µl 8 U/µl polimeráz; 0,4 µl dNTP mix; 4 µl 10-szeres ThermoPol reakció puffer; 14,1 µl víz. Ezt a reakciót is tisztítás követi.

Index szekvenciák ligálása

Az adapterekhez már csak a megfelelő indexkombinációkat kellett hozzáligálni. Mi 7 bázisos indezeket használtunk, melyeket úgy osztottunk el, hogy egy flowcellen semmiképp ne legyen félig vagy teljesen összetéveszthető minta. A ligáláshoz az imént megtisztított DNS oldatokhoz hozzáadtunk: 28,5 µl víz; 20 µl 5-szörös GoTaq puffer; 10 µl 25 mM-os MgCl₂ oldat; 10 µl 2,5 mM-os dNTP mix; 0,5 µl 5 U/µl-es G2GO Taq polimeráz; 5 µl 10 mg/ml-es BSA oldat; 6-6 µl a saját P5 és P7 10 µM-os indexekből. Minden reakciót a keveredés után három egyenlő térfogatban felvittünk egy-egy wellbe egy 96-os plate-en. A PCR program 12 ciklusból állt. Ezt követően ismét kitisztítottuk a DNS-t az oldatból MinElute oszlop segítségével, majd ellenőriztük az oldataink DNS koncentrációját szintén Qubit eljárással.

Ezt követően hajtottuk végre az NGS szekvenálást (l.: 3. ábra).

Algoritmusok, szoftverek

BCFtools (H. Li és mtsai., 2009)

Az imputáció során és általában a genomok kezelésében segített a SAMtools (Sequence Alignment/Map tools) BCFtools névre hallgató algoritmus csomagja, mely számos feladat végrehajtására képes a genomokon, mint szoftveres adatállományokon, vagyis BCF és VCF formátumú variáns leíró fájlakon. Az algoritmusok futását a HTSLib könyvtár segíti, mely jól optimalizált C programozási nyelven írt algoritmusokat tartalmaz. A csomag nyújtotta eszközök közül első sorban a következőket használtam:

mpileup:

A SAMtools mpileup-on alapuló parancs. A szekvenálási olvasatokból generál pileup formátumú adatokat, amelyek később variáns hívásra használhatók. Nyers adatokat generál a variáns híváshoz, de sokféle paraméter beállítható, például minimális lefedettség, minőségi pontszám. Több alany összevonható az mpileup során egy VCF/BCF fájlba és egy alany szekvenátuma is szétbontható külön VCF/BCF fájlba.

call:

Ez a parancs a variánsok hívására szolgál például a generált pileup fájlkból. Különböző algoritmusok és beállítások használhatóak a variáns hívásra. A hívott variánsokat VCF formátumban tárolja.

view:

Elsődleges funkciója, ahogyan a neve is sugallja, a genom megjelenítése egyszerű parancssori környezetben, de pipeline-nal továbbadhatjuk a szöveges információt, kimenthetjük más szöveges állományokba, TXT-be vagy konvertálhatjuk a BCF fájlokat VCF-be és fordítva. A megjelenítést vagy konvertálást különböző szűrőkkel is megvalósíthatjuk. A szűrés történhet kromoszómára, régióra, de más paraméterekre is szűrhetünk: a pozíció minőségi pontszáma, referencia/alternate allél, de az INFO/DP>X segítségével akár az X-szeres lefedettségénél nagyobb pozíciókat is listázhatjuk.

norm:

Indelek normalizálása. Ez az algoritmus végigmegy az adott genomon és kezeli az inszerciákat, deléciókat, multiallélikus helyeket. Magában a BCF/VCF fájlban balra rendezi a szöveget, illetve ha még nincs, pozícióként (site) sorokba tördeli azt. Indelek esetén ahol multiallélikus helyek vannak több sorba tördelheti az egy sorba tömörített allélokat, illetve fordítva, ha a

későbbi folyamatokban ilyen formában tudjuk kezelni a genomot, egy sorba tömöríti az azonos helyhez tartozó allélokat. Hasonló módon képes összevonni a duplikált pozíciókat, vagyis ahol nem eltérő allélok, hanem ugyanannak az allélnak duplikáltjai vannak jelen.

index:

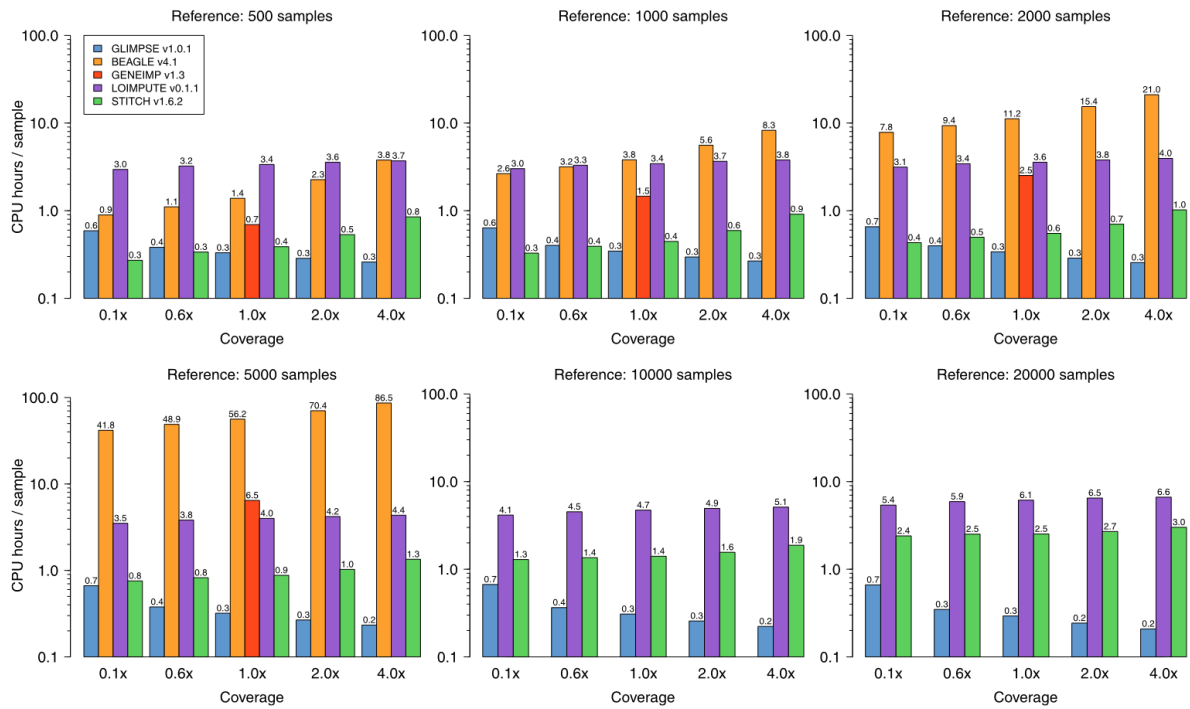
A tömörített fájlok indexelését végzi (lásd korábban: BGZ). Az algoritmus alapértelmezett módon a CSI (coordinate-sorted index, kromoszóma hossz: max. 2^{31} bp) vagyis a koordináta alapú indexelést támogatja és ebben a rendszerben indexeli a bemeneti állományokat. Ha ez nem lehetséges, a program TBI (tabix index, kromoszóma hossz: max. 2^{29} bp) fájlt hoz létre a HTSlib könyvtár tabix algoritmusának meghívásával, de akár szándékosan is létrehozhatunk ilyet.

query:

Lehetővé teszi a VCF/BCF fájllokból történő adatkivonást és formázást, specifikus információk kinyerését, adatok tetszőleges formátumú szövegfájlba történő írását. Hasonló a view funkcióhoz, de itt nem teljes rekordokat kapok vissza a megadott szűrők alapján, hanem a fájl egészéből nyerhetek ki információkat, például bizonyos szűrőknek (mint a genotípus vagy a lefedettség) megfelelő hívások számát, a multi BCF-ek esetén listázthatjuk a minták neveit, stb. valamint formázhatom a kimenet megjelenítését, például csak a kromoszómaszámot (#), pozíciót (#), referencia- és alternate allélt (A/G/C/T) és a genotípust (0/1) kívánom megjeleníteni tabulátor szeparált oszlopokban, hívásonként tört sorokban. Mindezek természetesen továbbíthatóak és kimenthetőek standard szövegállományokba.

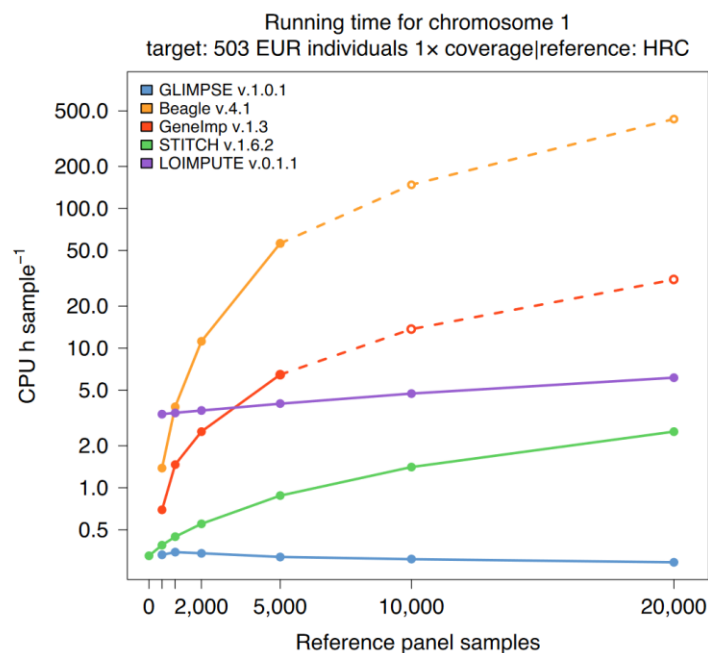
GLIMPSE v1.1.1 (Delaneau és mtsai., 2012, 2017; Rubinacci és mtsai., 2021)

A GLIMPSE (Genotype Likelihoods IMputation and PhaSing mEthod) egy imputációs szoftver, amit a svájci Simone Rubinacci és munkatársai publikáltak 2021-ben a nívós Nature Genetics c. folyóiratban (Rubinacci és mtsai., 2021). A GLIMPSE1-ben alkalmazott algoritmus a többi algoritmushoz képest sokkal kisebb gépidő mellett képes nagy pontosság elérésére (l.: 8., 9. ábra) (Escobar Rodríguez & Veeramah, 2024; Rubinacci és mtsai., 2023; Sousa da Mota és mtsai., 2023).



8. ábra: A GLIMPSE1 processzorigénye a többi imputációs szoftverhez hasonlítva (Rubinacci és mtsai., 2021).

A módszer bemutatása érdekében a paraméterek széles spektrumát tesztelték a GLIMPSE-en. 564 genom imputálását hajtották végre az 1000 Genome Project New York Genome Center által újrászekvenált 30-szoros lefedettségű genomjait használva referenciaként. A lefedettséget 0,1x-től 8-szorosig, a MAF-t 0,01%-os gyakoriságtól 50%-ig vizsgálták. A minták lefedettségét mesterségesen is csökkentették a 0,1-szeres lefedettség eléréséhez. A program futásidejét és annak skálázódását is behatóan tanulmányozták. Ennek érdekében változtatták a párhuzamosan imputált minták számát és a referencia adatbázis méretét.

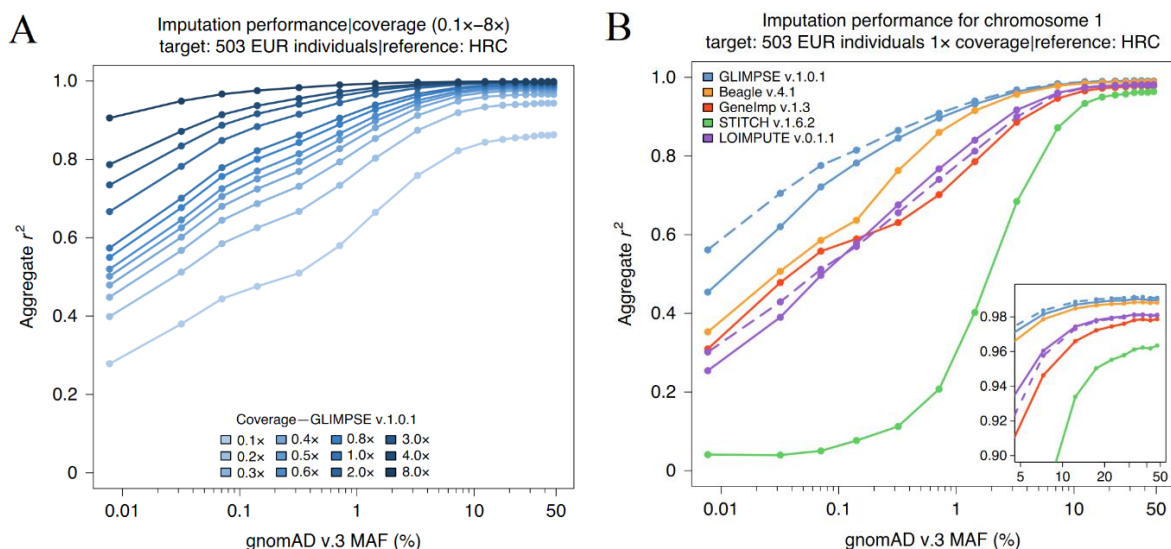


9. ábra: A vizsgált algoritmusok processzorigényének skálázódása a referenciapanel méretének függvényében (Rubinacci és mtsai., 2021).

A mérések a következőt mutatták a GLIMPSE v1.1.1 imputáló képességének vizsgálatakor:

Lefedettség	MAF	r^2	Értékelés
0,1x	>5%	>0,7	Nagy relatív pontosság extrém alacsony lefedettségre is.
0,3x	>5%	>0,9	Alkalmazható mértékű pontosság kiemelkedően alacsony lefedettség mellett.
1,0x	=0,1%	$\approx$ 0,8	Nagy relatív pontosság ritka variánsok esetén alacsony lefedettség mellett.
8,0x	=0,1%	>0,95	Kiemelkedően nagy pontosság ritka variánsok esetén magas lefedettség mellett.

1. táblázat: A GLIMPSE1 megbízhatósága (Rubinacci és mtsai., 2021).



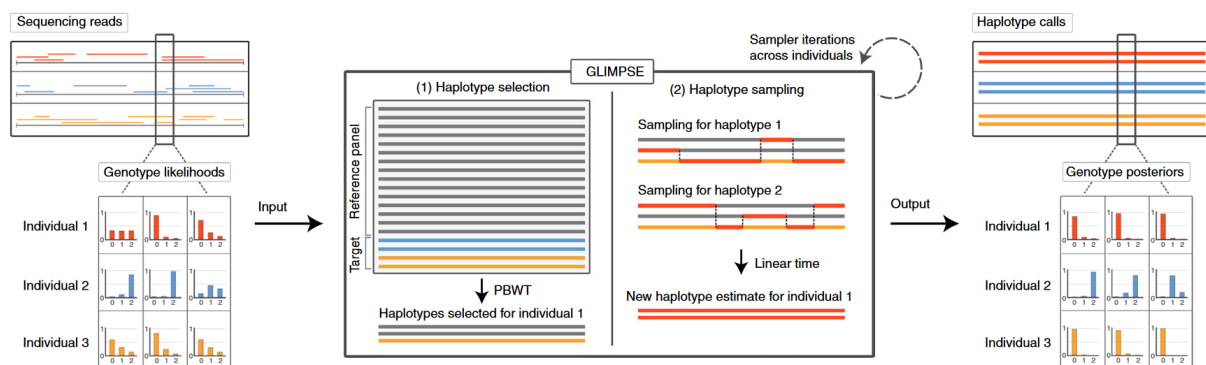
10. ábra: A GLIMPSE1 konkordancia vizsgálata a lefedettség függvényében (A) valamint a többi algoritmushoz viszonyítva (B) (Rubinacci és mtsai., 2021).

A GLIMPSE imputáló képessége felülmúlta az összes további state-of-the-art szoftver képességeit 1-szeres lefedettség mellett a ritka és a gyakori variánsok esetén is (l.: 10. ábra).

A program mintánkénti processzorigényének skálázódását tekintve messze felülmúlja kortársait jelentősen csökkentve az alacsony lefedettségű szekvenciák imputálásának időigényét és költségeit. Ezzel új lehetőségeket nyit meg a genome-wide studies területén.

GLIMPSE Joint Modell alapú imputálás

A GLIMPSE kiemelkedő hatékonysága és pontossága mellett abban is eltér társaitól, hogy képes egyszerre egy kisebb referencia adatszett mellett több mintát egyszerre imputálni úgy, hogy közben az imputálandó mintákat referenciaként is felhasználja a többi minta imputálásához az iteratív GL (Genotype Likelihood) javító lépés során. (l.: 11. ábra). Ez az eljárás abból a megfontolásból indul ki, hogy amennyiben az imputált minták populációs struktúrája valamelyest eltér a felhasznált referenciákétól, akkor ez elméletileg javíthatja az imputáció pontosságát.

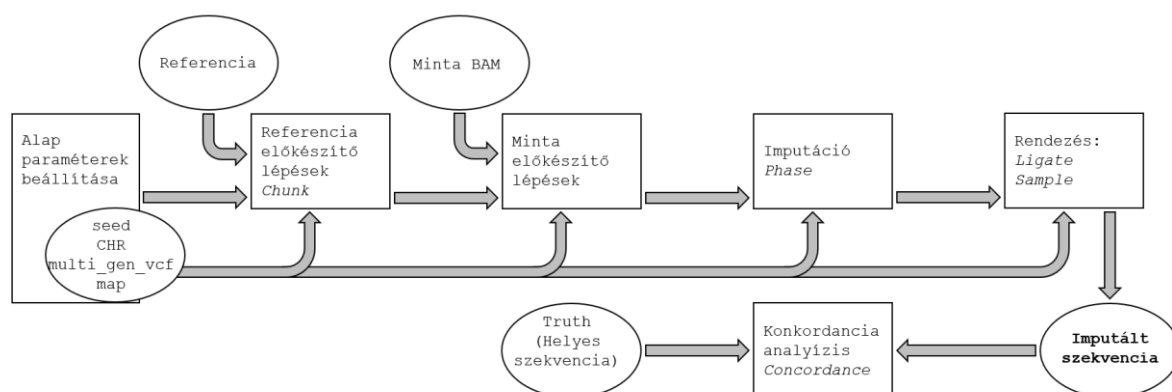


11. ábra: Az imputálás folyamata, a Joint modell szemléltetése (Rubinacci és mtsai., 2021).

Kutatásunk egyik fő célja volt, hogy kontrollált körülmények között, mesterségesen lecsökkentett lefedettségű mintákkal megvizsgáljuk, hogy Joint model alkalmazása több azonos populáció struktúrájú minta együttes imputálásával esetleg javítja-e a konkordanciát (Két genetikai adatforrás egyezőségi szintjének statisztikai mérőszáma.).

A GLIMPSE v1.1.1 működésének folyamata (Delaneau és mtsai., 2012, 2017; Rubinacci és mtsai., 2021)

Az imputáláshoz szükséges referencia adatok összeállításához, az imputálandó markerek szűréséhez, transzformációjához és a teszt minták adatainak kezeléséhez, konverziójához is szükség van a SAMtools és a BCFtools csomagokra. A GLIMPSE öt fő algoritmusra a Chunk, Phase, Ligate, Sample illetve a kiértékelést szolgáló Concordance. A GLIMPSE-el való imputálás általunk végzett folyamatát a következő ábra szemlélteti, egyúttal bemutatja az öt algoritmus szerepét a művelet során (dőlt betűvel kiemelve azokat).



12. ábra: Az általunk alkalmazott imputációs folyamat. Részletes magyarázatát lásd a szövegben.

A seed integer változó a randomszám generálás alapjául szolgáló szám, melyet az eljárás vizsgálata során rögzítettük a folyamatot befolyásoló tényezők felkutatása érdekében (Értéke kutatásunk során: 1).

A CHR integer változó a vizsgált kromoszóma számát tárolja (Értéke kutatásunk során: 22).

A multi_gen_vcf nevű string változó a referencia nevét és elérését.

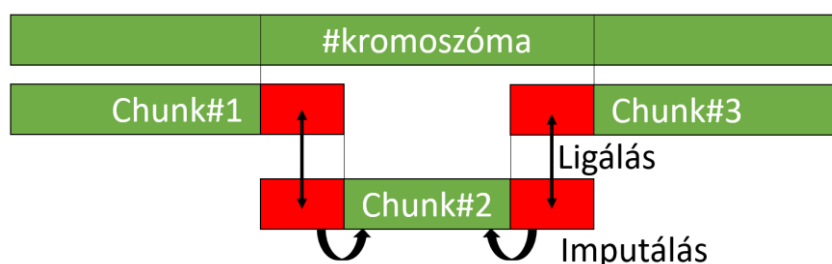
Szintén string változó a map, mely az éppen használt genetikai térképet és elérését tartalmazza (Ez kutatásunk során egy GRCh 37-es map volt).

Említetteken kívül még számos változó beállítása történt meg, mint a bam_prefix, a sample_name_prefix, a bam_dir vagy a ref_gen, melyek vagy a készített fájlok nevezéktanát (előtagjait, prefixeit), vagy elérési útvonalakat (dir=directory) határoztak meg.

A GLIMPSE algoritmusainak ismertetése

1. Chunk

A genomot (azon belül a kromoszómákat) kisebb, kezelhető méretű szakaszokra osztja, amelyek önmagukban könnyebben feldolgozhatók (l.: 13. ábra). Ez a felosztás segít az adatkezelés és az analízis hatékonyságának növelésében, mivel a kisebb adatdarabok könnyebben kezelhetők és párhuzamosan imputálhatók. A chunking során biztosítani kell, hogy az egyes darabok átfedjenek egymással, hogy a határterületeken ne legyenek pontatlan becslések, az imputálandó szakasz két oldalán ne legyen aszimmetrikus az információk eloszlása, valamint a későbbi összefűzést („ligate”) érdekében.



13. ábra: A chunk függvény.

2. Phase

A cél az egyes chunkok imputálása (a hiányok kitöltése), valamint „phase”-elése, (annak meghatározása, hogy a diploidizált genom haplotípusai legnagyobb valószínűséggel melyik

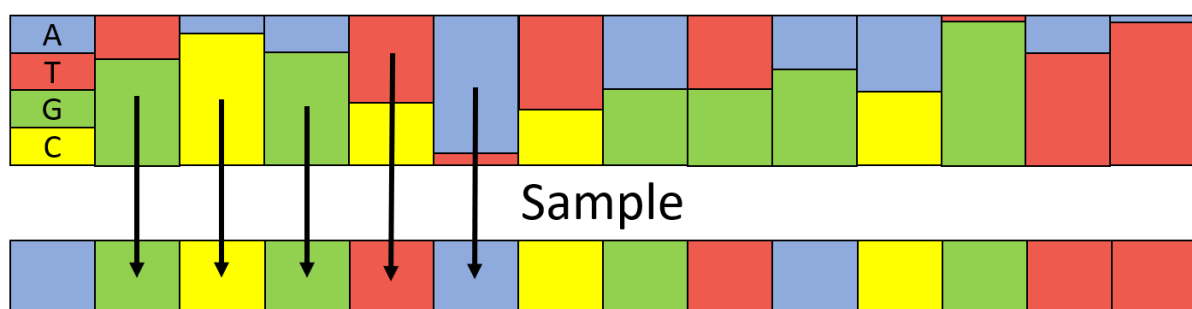
homológra esnek. Az egyes chunkokban található genotípus adatokból (egyrészt a lyukak közvetlen környezetéből, másrészt a teljes genomra kiterjedő kapcsolt valószínűségekből, vagyis a statisztikailag együtt öröklődő SNP-kből) meghatározza, hogy mi a keresett bázis, és az milyen „phase”-ben lehet. A fázist bayesi statisztikai modellek alapján állapítja meg, amelyek figyelembe veszik a genotípus valószínűségeket és a referencia panel adatait. Az algoritmus iteratív módon finomítja a becsléseket, hogy minél pontosabb eredményeket érjen el. A „Phase” lépés kimeneti értéke még nem genotípus, hanem a szakaszokhoz kapcsolt valószínűségi térkép minden pontra, a lyukakat (vagy akár más alacsony valószínűségű atipikus helyeket) belevéve.

3. Ligate

Az egyes „chunk”-okban meghatározott Genotype Likelihoodokat (GL) összeilleszti, hogy egy teljes genomi GL-ot hozzon létre. Ez a lépés biztosítja, hogy az egyes darabok közötti átfedések pontosan illeszkedjenek, és egy konzisztens teljes genomi szekvenciát eredményezzenek. A ligálás során figyelembe veszik a chunkok közötti átfedéseket és azok fázis információit, hogy a teljes genom fázisa helyes (a levalószínűbb) legyen.

4. Sample

Ennek a lépésnek a célja, hogy a valószínűségi modellek alapján mindenhova konkrét genotípust generáljon (l.: 14. ábra). A „Phase” során kapott valószínűségi eloszlásokból genotípus mintákat vesz, vagyis a valószínűségi térkép alapján konkrét genotípusokat választ. Ebben a lépésben van az egyik legnagyobb jelentősége a random szám generálásnak és annak „seed”-jének, melyet a későbbiekben részletesebben tárgyalok.



14. ábra: A Sample függvény.

5. Concordance

Az imputáció hatékonyságát úgy teszteljük, hogy magas lefedettségű, ismert diploid genomokat tetszőleges szintű alacsony lefedettségűre lerontunk (down sampling) úgy, hogy véletlenszerűen eltávolítjuk a szekvenciák egy részét, majd ezeken végezzük az imputációt. Végül a konkordancia vizsgálattal mérjük az imputált adatok pontosságát úgy, hogy összehasonlítjuk azokat az eredeti, valódi genotípusokkal (truth data). Ehhez kiválaszthatók azok a genomi pozíciók, ahol összehasonlítást szeretnénk végezni. Itt szűrhetünk nagyon magas lefedettségű pontokra a truth genomban így biztosítva a teszt hitelességét. A program összehasonlítja az imputált és a valódi genotípusokat minden egyes kiválasztott pozícióban, majd megszámolja, hány imputált genotípus egyezik meg a valódi genotípusokkal. A konkordanciát az egyező genotípusok százalékos arányával fejezzük ki. Emellett minden pontnak korábban meghatároztuk a MAF értékét a referencia alapján, így tudjuk ábrázolni a konkordanciát a MAF függvényében. A GLIMPSE_concordance mindezt önállóan elvégzi, majd ezen kívül alkalmazza a matplotlib Python csomagot és ábrázolja a kapott értékeket, majd elmenti egyszerű kép formátumban.

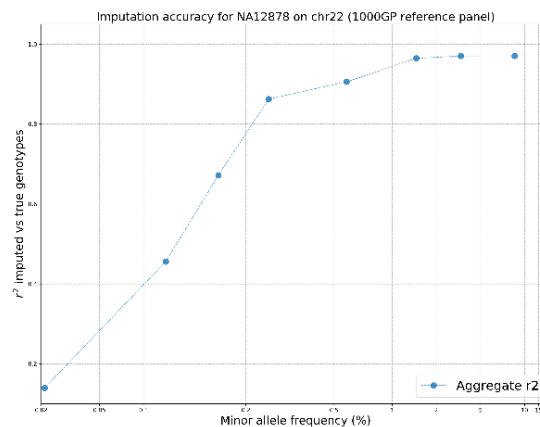
GLIMPSE 1 tutorial

A GLIMPSE 1.1.1-es verziójához készítették egy tutorialt, mely lényegében egy mappa benne shellscriptekkel, melyek tartalmazzák az egyes lépések végrehajtásához szükséges programsorokat. Ezek a lépések végig vezetnek egy teljes imputáció minden lépésén a referencia letöltésétől a konkordancia grafikus megjelenítéséig.

A tutorial alapértelmezetten a következő lépéseket foglalja magába:

- `step1_script_setup.sh`: Előzetes takarítást végez a későbbiekben felhasználásra kerülő mappákban, illetve jogosultságot biztosít az öt algoritmus bináris kódjához.
- `step2_script_reference_panel.sh`: Letölti, ellenőrzi, normalizálja és indexeli a referenciát, majd kiveszi belőle az imputálandó egyént.
- `step3_script_GL.sh`: Kiszámolja az imputálandó egyén (NA12878 egyed 1-szeres lefedettségűre rontott BAM fájlja) genotípus valószínűségeit (GL) minden pontra, valamint végrehajtja ugyanezt az alannal is. (Mivel ennek a lépésnek a fele csak a referenciát veszi igénybe, később praktikus okokból ketté bontottuk.)

- `step4_script_chunk.sh`: Végrehajtja a referencia feldarabolását. (GLIMPSE_chunk)
- `step5_script_impute.sh`: Egy ciklusban behívja az összes generált „chunk”-ot, majd végrehajtja az imputációt a korábban leírt iteratív módon. (GLIMPSE_phase)
- `step6_script_ligate.sh`: Összefűszli a keletkezett GL-okat. (GLIMPSE_ligate)
- `step7_script_sample.sh`: Genotípus generálás. (GLIMPSE_sample)
- `step8_script_concordance.sh`: Kiszámolja a konkordanciát minden megadott pontra, MAF szerint rendezi, majd ábrázolja a kapott eredményt és elmenti a grafikont. (GLIMPSE_concordance)



15. ábra: A tutorial Concordance vizsgálatának végeredménye.

- `step9_script_cleanup.sh`: Törli a tutorial futtatása során keletkezett és letöltött összes fájlt.

GLIMPSE2 (Rubinacci és mtsai., 2023)

Kutatásunkban összehasonlítottuk a GLIMPSE v1.1.1 egyszerű és Joint Modell alapú imputálását a GLIMPSE2.0-val, a GLIMPSE későbbi változatával is. A GLIMPSE2 fejlesztése során nagy hangsúlyt fektettek a szoftver optimalizálására, melynek köszönhetően csökkent az imputációhoz szükséges processzor igény, továbbá nőtt az alkalmazható referencia panel mérete. Emellett kivették az eredeti GLIMSE 1-ben található Joint Model funkciót.

QUILT (Davies és mtsai., 2021)

Az imputációkat a QUILT nevű programon is lefuttattuk. A QUILT egy újabb szoftver mint a GLIMPSE, és az ezt közlő publikációban gyorsabbnak, jobban skálázhatónak, esetenként pontosabbnak mutatták be.

Almalinux

Telepítettem a saját számítógémemre az Almalinux operációs rendszert, mely a kutatócsoport által használt szerveren is fut, hogy abban a környezetben tanuljam meg a GLIMPSE kezelését, amelyen majd az éles mérések is futnak. Ezután letöltöttem a GLIMPSE-t és a többi imputáláshoz szükséges további eszközöket.

Bash

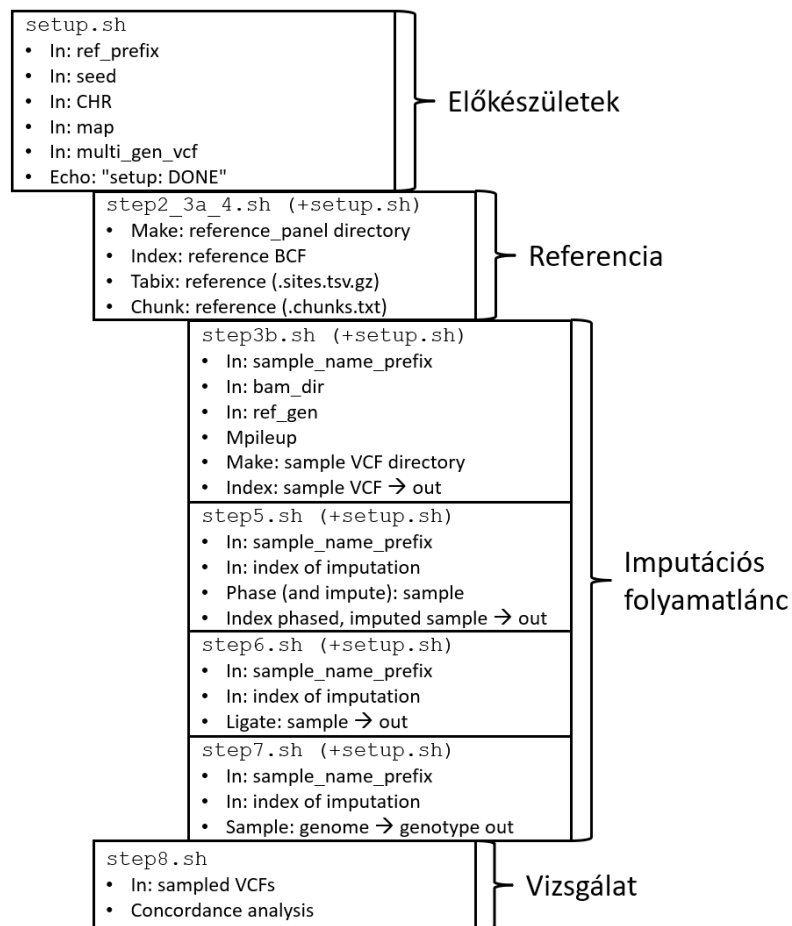
Az SZTE Molekuláris bionika mérnöki képzésben csak a C programozási nyelv kötelező (Programozás alapjai kurzus). Én ezen kívül elvégeztem Java és C# nyelveket érintő kurzusokat, de itt sem került sor a nyelv készség szintű elsajátítására, ezért a Linux kezelését és a Bash elsajátítását teljesen az alapoktól kezdtem. Meg kellett tanulni a mappák létrehozását, mozgatását, mappastruktúrák megőrzését, másolását, a programok és bináris kódok meghívását, futtatását, a futtatási jogosultságok kezelését. Ez után magával a GLIMPSE-szel kellett behatóbban foglalkoznom, átalakítani a mappa struktúrákat, hozzáféréseket, jogosultságokat, hogy akadálymentesen futhasson az imputálás.

Tutorial

A GLIMPSE-be épített tutorial futtatása sem magától értetődő. Kissé át kellett alakítani a benne foglalt letöltési linkeket, a fájlok nevezéktanát, a mappákat, hogy lefussanak az egyes lépések. Végül sikerült lépésről lépésre végigmenni a tutorialon, megtörtént a referencia letöltése, a példában szereplő minta imputálása és a konkordancia is a vártnak megfelelő értéket adott.

Gyakorlati alkalmazás

Végrehajtottam egy saját, mesterségesen lerontott (down sampled) minta (SZTLF, 1.: 3. táblázat) imputációját egy GRCh37-es referenciával, amit az archeogenetikai kutatásokban alkalmazunk. A folyamat során a tutorial lépéseire támaszkodtam, azokat alakítottam át (1.: 16. ábra).



16. ábra: A tutorial átszervezett lépései.

Az egyes téglalapok aszerint helyezkednek el balról jobbra, hogy a bennük található algoritmusok mennyire specifikusak a referenciára vagy a mintára, vagyis minden téglalapot meg lehet hívni több, a tőle jobbra elhelyezkedő téglalap alkalmazásakor. A setup.sh-t minden tőle jobbra található lépés meg is hívja, mert működésükhöz szükségesek az ebben tárolt adatok.

Az első blokk a setup, vagyis az állandó paramétereket beállító scriptet szemlélteti. Ezt a lépést azért hoztam létre, hogy letisztultabb és felhasználóbarátabb legyen az imputáció folyamata, egyúttal nekem is sokkal egyszerűbb a különböző imputálások alkalmával csak a szükséges paramétereket megváltoztatni. Lényegében ez egyfajta alapértelmezett paraméter hívás, ami az objektum orientált programozásban gyakorta használt eszköz.

A következő blokk már több lépést kölcsönöz a tutorialból. Ahogyan azt korábban említettem (Tutorial/step3_script_GL.sh), a harmadik lépés első, referenciára vonatkozó fele is itt szerepel, de a másik, alannyal kapcsolatos szakaszát át tettem egy későbbi lépésbe. Ennek célja az volt, hogy a step2_3a_4.sh futtatásakor elvégzésre kerüljön minden referencia előkészítő

folyamat, ugyanakkor ne kelljen már ebben a lépésben az alanyt is kezelni. Ezáltal egyrészt letisztultabbá és követhetőbbé válik a folyamat, másrészt ha több imputálást is ezzel a referenciával végzünk, akkor csak az ezt követő lépéseket kell mintánként megismételni. Maga a lépés a referenciapanel átalakítását, GL számítását, indexelését és darabolását végzi.

A következő lépés a `step3b.sh`, ami az előző lépésből kimaradt GL számítást végzi el ezúttal az aktuális mintára. Ennek kivitelezéséhez lekéri a minta nevét, elérhetőségét, az továbbiakban alkalmazandó azonosító előtagját ('prefix') és a referenciagenomot, amellyel a továbbiakban számolni fog, valamint létrehozza az eredmények tárolására alkalmazott mappákat.

Az ezt követő három lépés (Imputációs folyamatlánc) szinte pontosan a korábban leírt tutorial lépéseket hajtja végre annyi különbséggel, hogy ezúttal mindezt az adott mérés során meghatározott mintákkal, referenciákkal és paraméterekkel. Ez a négy blokk azért került az ábrán azonos hierarchiai szintre, mert mindegyik közvetlenül az öt megelőző blokk kimenetét ('out') fogadja egy összefüggő folyamatláncot (pipeline) alkotva.

Végül a folyamat utolsó lépéseként a konkordanciaszámítást is személyre szabtam a könnyebb futtathatóság érdekében. Mivel ez már ugyanúgy végrehajtható bármely másik genomra is, valamint futtatása előtt meg kell szabni a megjelenítés beállításait, így az ábrán egyet visszalépünk a hierarchiában.

Önálló imputálás

A fentiekben bemutatott, célfeladatra módosított imputációs algoritmust egy saját mintán teszteltem (SZTLF). A folyamat sikeresen lezajlott, az eredmények megfelelőek voltak, a GLIMPSE a várt pontossággal imputálta a mintát.

A randomizálás befolyásának vizsgálata

A GLIMPSE algoritmusok mindegyike tartalmaz egy seed paramétert, mely az adott algoritmusban felhasznált randomizációs eljárások alapját képezi. Az egyik legjobb példa erre a `sample`, mely a GL függvényében véletlenszerűen választ a genotípusok közül. Ez standard eljárás a genotípus előállító szoftvereknél a programfüggő torzítás (bias) kiküszöbölése céljából. Az eljárás nem teljesen random, mert figyelembe veszi a GL értékét, de ugyanakkor szükséges, hogy ne vigyünk olyan hibát az imputációba, mely az adott algoritmus működéséből adódóan torzíthatja az eredményeket. Például ha nem random módon döntenénk minden 50%-os valószínűségű A/G (Adenin/Guanin) genotípusú pont esetében, hanem mindig az adenint választanánk az ABC sorrend vagy egyszerűen az algoritmus működéséből kifolyólag,

akkor egyértelműen az adenin irányába torzítanánk a valóságot, ami egy programfüggő torzítást eredményezne. Ugyanez a bizonytalanság a 99%/1%: A/G valószínűség esetén is jelen van. Amennyiben a referencia adatbázis alapján a program 99%-os valószínűséggel adenint, 1%-os valószínűséggel guanint jósol, akkor az az instrukció, hogy a genom létrehozásakor ugyezen valószínűségekkel történjen a kiválasztás. Tehát 99% annak az esélye, hogy a betöltendő lyukban eredetileg is adenin volt és a genotípus kiválasztásakor is ezt a valószínűséget alkalmazza. Hasonlóképpen a guanin esetében, amikor 1% a valószínűsége, hogy az adott ponton guanin volt jelen és ezért 1%-os valószínűséggel választunk ide guanint. Ha biztosak lennénk abban, hogy a mintánk a gyakrabban előforduló major alléllal rendelkezett (100%/0%: A/G), vagyis esetünkben adeninnel, akkor látható, hogy ez a kiválasztás 1%-os valószínűséggel hibát eredményezne ($1 \times 0,99 = 99\%$ hogy helyes és $1 \times 0,01 = 1\%$ hogy helytelen). Viszont a valóságban nem ismerjük az eredeti genotípust, annak helyességére csak a referenciában előforduló gyakoriságok alapján tudunk következtetni, ami a kiválasztásunknak is alapjául szolgált (l.: 2. táblázat). Tehát 99% a valószínűsége, hogy az adott ponton adenin volt a helyes genotípus és 99%, hogy ezt választjuk, vagyis a kettő együttes valószínűsége $0,99 \times 0,99 = 99,01\%$. Ha helytelen lenne a genotípusról tett feltételezésünk, vagyis az adott helyen guanin volt az egyénben, aminek a valószínűsége 1%, akkor is még marad 1% valószínűsége annak, hogy mégis ezt választjuk, így $0,01 \times 0,01 = 0,01\%$ további valószínűség hozzáadódik ahhoz, hogy nem tévedtünk. A maradék két lehetőség az első és másodfajú hibák, vagyis hogy a jó feltételezésünk ellenére (99% adenin) mégis rossz allélt (1%) választottunk ($0,99 \times 0,01 = 0,99\%$), valamint hogy igyekeztünk helyesen dönteni (99%), de már a kezdeti feltevésünk sem állta meg a helyét (1%) így rosszul választottunk ($0,01 \times 0,99 = 0,99\%$).

Meghatározás pontossága	A (P=99%)	G (P=1%)
Helyes a feltett genotípus	98,01%	0,99%
Helytelen a feltett genotípus	0,99%	0,01%

2. táblázat: A randomizálás feltételes valószínűsége.

A randomizálást tehát egyfajta torzítás kiküszöbölésére használják, amikor feltételes valószínűséget használva a valószínűségek arányát figyelembe véve, de véletlenszerűen döntünk.

A csapatmunkában történő eredmények összehasonlíthatóságának tesztelésére megvizsgáltuk, hogy azonos seed esetében, a különböző munkaállomásokra telepített szoftver környezetben ugyanazt az eredményeket kapjuk-e. A vizsgálat során ugyanazt a mintát (SZTLF) imputáltuk

3 különböző munkaállomáson, eltérő időben, és helyen. Ez azért volt fontos, mert egyes randomszám generálási folyamatok beleforgathatják pl az időt, vagy a munkaállomásra jellemző egyéb egyedi paramétereket a véletlen számok generálásába. Az idő és az architektúra egy tipikusan ilyen példa, ugyanis egyes algoritmusok a készülék órája által adott számok bizonyos számjegyeinek manipulálásával (szorzás, maradékos osztás) állítják elő a szükséges random számokat. Ki kellett zárunk tehát, hogy ezek hatással lennének az imputálásra. Eredményeink igazolták, hogy azonos random seed megadása esetében az eltérő munka környezetekben végrehajtott imputációk determinisztikusan azonosak voltak.

Minták és eljárások

Számos archaikus minta közül kiszűrtünk 357-et, melyek legalább 2-szeres átlagos lefedettséggel bírtak 0 exogén kontamináció mellett, és azonos populációt képviseltek. Ezek közül az öt legmagasabb lefedettségű mintát választottuk imputálásra: (SZTLF, VZ-12673, CSPF-114, SZAK-4, K2-61). A folyamat első lépéseként a szekvenciákat mesterségesen lerontottuk, a readok megfelelő hányadának véletlenszerű eltávolításával úgy, hogy a minták 1x és 0.5x átlag lefedettségűek legyenek. Az összehasonlíthatóság és reprodukálhatóság érdekében a lefedettség csökkentése minden minta esetében azonos random seed-del történt. A fennmaradó 352 genomot a Joint modellben együtt analizáltuk a mesterségesen lerontott lefedettségű teszt mintákkal, míg a többi imputációnál nem vettük bele az analízisbe.

Minta kódja	Leírás	Genom típusa	Lefedettség
SZTLF	Szent László király	Európai	16,40x
VZ-12673	Elit hun	Kelet-ázsiai	9,98x
CSPF-114	Első generációs avar bevándorló	Kelet-ázsiai	11,47x
SZAK-4	Első generációs honfoglaló	50% európai, 50% szibériai	7,95x
K2-61	Második generációs honfoglalás kori egyén	50% Kelet-ázsiai, 50% szibériai	13,37x

3. táblázat: A felhasznált minták és a szekvenálás átlagos lefedettségai

Ezt követően a genomokat imputáltuk. Az öt genomot párhuzamosan a GLIMPSE Joint Model szerint, majd egyesével is GLIMPSE1, GLIMPSE2 és QUILT segítségével. A GLIMPSE2-nél pozitív kontrollként olyan imputációt is végrehajtottunk, mikor a GLIMPSE1 joint modellben imputált és phaselt 352 genom adatot is bevettük a GLIMPSE2 referencia adatszettjébe (mintha

egy előre elkészített modern + ancient joint modell lenne a referencia), és ezzel a joint referenciával is elvégeztük az imputációt.

A konkordancia vizsgálat során az eredeti genomokból (truth) is csak a legbiztosabb pontokat (≥ 15 -szörös lefedettség) vettük figyelembe, így a konkordanciát csak azon pozíciókban számoltuk, ahol az eredeti minta 15, vagy több readdel volt reprezentálva, azaz a diploid genotípus (GT) nagy biztonsággal lehívható volt.

Eredmények értékelése

Az részletek láthatósága érdekében a kapott konkordanciákat nagy formátumban illesztettem be, így azok mérete és számossága miatt csak az eredmények értékelése után kerülnek bemutatásra.

- A magas Allélfrekvenciájú lókuszok jól, míg az alacsony MAF-jú lókuszok rosszul imputálhatóak mind az öt mintára, mindhárom program alkalmazásával. A 0.5x lefedettség is viszonylag magas 0.9-0.95-ös konkordanciát eredményezett az 5-10%-os magas allélfrekvenciájú markereknél.
- A magas allélfrekvenciájú markerek konkordanciája közel azonos volt a különböző módszerekkel.
- A GLIMPSE1-es módszer tűnt legpontosabbnak vizsgált metodikák közül.
- A Joint modell csak kis mértékben javította csak az eredményeket, olykor nem is volt megfigyelhető javulás.
- Alacsony lefedettség esetén egyértelműen a GLIMPSE1 a legpontosabb a Joint modellel és anélkül is.
- A programok közül a GLIMPSE2 bizonyult legyorsabbnak, legkisebb CPU igénnyel, viszont kicsit alacsonyabb konkordanciát eredményezett.

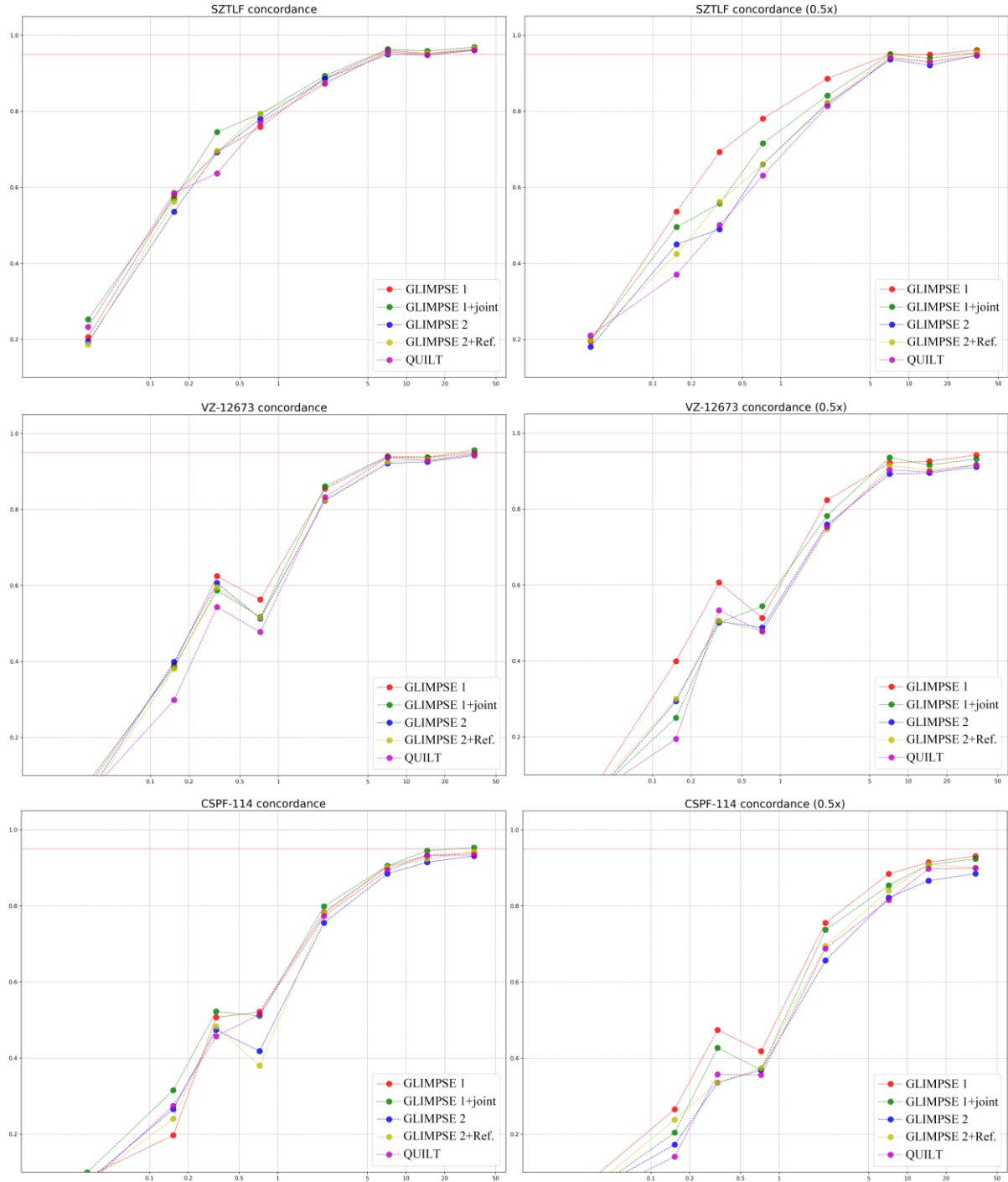
Összegezve, a GLIMPSE1 minimálisan pontosabb a másik két programnál, de a többi módszerrel kapott adatok is alkalmasak az IBD sharing analízisre. Magas MAF esetében bármelyik módszer megbízható eredményt ad. Bár az módszert publikáló tanulmányban azt állították, hogy a QUILT alacsony MAF-nél pontosabb mint a GLIMPSE (Davies és mtsai., 2021), ez vélhetőleg csak igen nagy referencia adatbázis esetében (például a nem publikus UK biobank) igaz, az 1KG referenciával mi nem kaptunk magasabb konkordanciát vele.

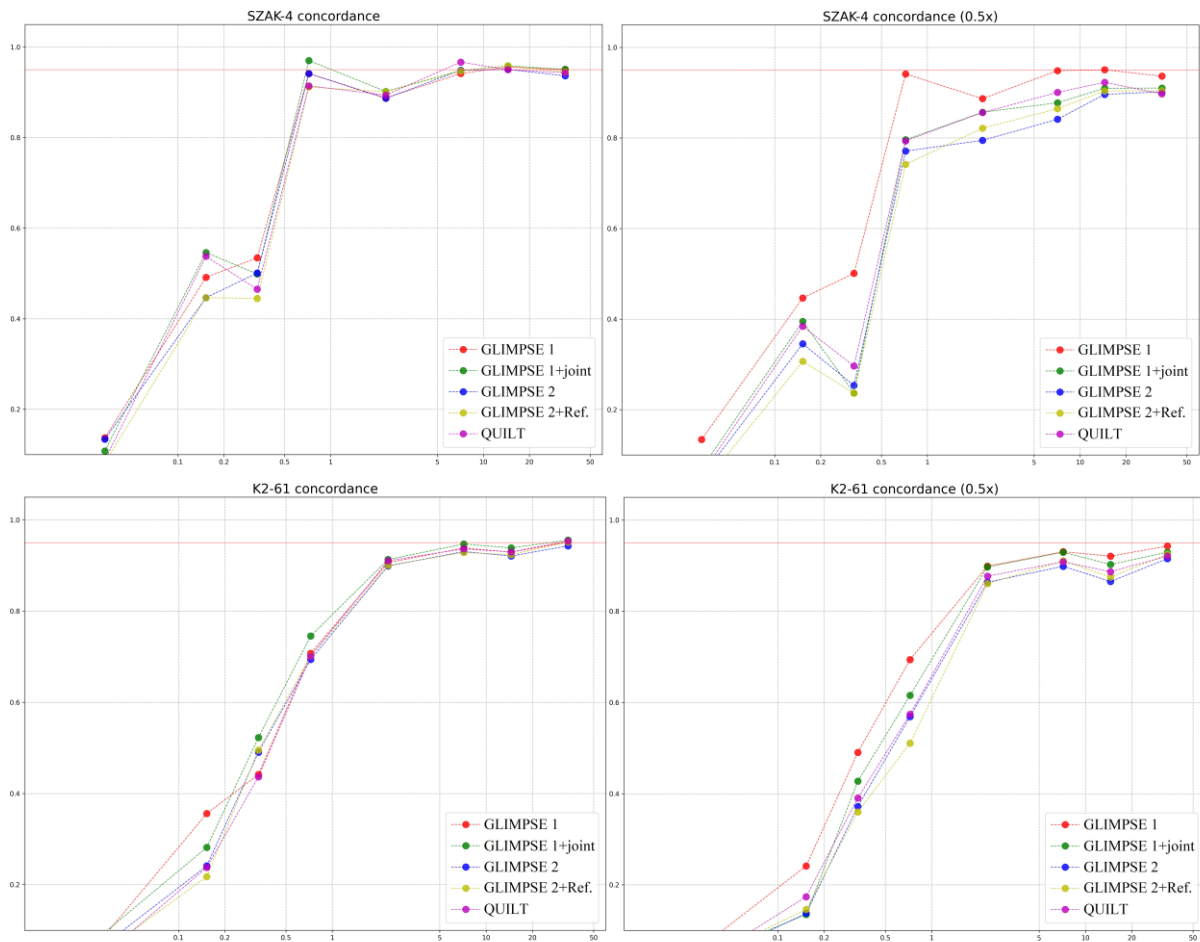
A kapott konkordanciák:

1x (bal) és 0.5x (jobb) lefedettségre csökkentett minták imputálása

Az Y tengelyen a konkordancia ($r^2 = 0-1$) látható az X logaritmusos tengely Alternatív

Allélgyakoriság függvényében (Minor Allele Frequency = 0.01-50%).





Köszönetnyilvánítás

Szeretnék köszönetet mondani Dr. Maróti Zoltánnak, aki végigvezetett engem ezen az úton, valahányszor bizonytalanság támadt bennem, türelmesen és részletesen elmagyarázta nekem és segített a nehezebb feladatok végrehajtásában.

Köszönöm Dr. Török Tibor tanár úrnak, hogy megteremtette számomra ezt a lehetőséget és mindig bíztatott, ha elakadtunk a munkában.

Hálával tartozom Schütz Oszkárnak, akit napszaktól függetlenül zaklathattam a kérdéseimmel és Dr. Varga Gergellyel együtt gondosan felügyelték a laborban minden mozdulatomat.

Irodalomjegyzék

- Auton, A., Abecasis, G. R., Altshuler, D. M., Durbin, R. M., Abecasis, G. R., Bentley, D. R., Chakravarti, A., Clark, A. G., Donnelly, P., Eichler, E. E., Flicek, P., Gabriel, S. B., Gibbs, R. A., Green, E. D., Hurles, M. E., Knoppers, B. M., Korbel, J. O., Lander, E. S., Lee, C., ... National Eye Institute, N. (2015). A global reference for human genetic variation. *Nature*, 526(7571), 68–74. <https://doi.org/10.1038/nature15393>
- Browning, S. R. (2008). Estimation of pairwise identity by descent from dense genetic marker data in a population sample of haplotypes. *Genetics*, 178(4), 2123–2132. <https://doi.org/10.1534/genetics.107.084624>
- Davies, R. W., Kucka, M., Su, D., Shi, S., Flanagan, M., Cunliffe, C. M., Chan, Y. F., & Myers, S. (2021). Rapid genotype imputation from sequence with reference panels. *Nature Genetics*, 53(7), 1104–1111. <https://doi.org/10.1038/s41588-021-00877-0>
- Delaneau, O., Marchini, J., & Zagury, J.-F. (2012). A linear complexity phasing method for thousands of genomes. *Nature Methods*, 9(2), 179–181. <https://doi.org/10.1038/nmeth.1785>
- Delaneau, O., Ongen, H., Brown, A. A., Fort, A., Panousis, N. I., & Dermizakis, E. T. (2017). A complete tool set for molecular QTL discovery and analysis. *Nature Communications*, 8(1), 15452. <https://doi.org/10.1038/ncomms15452>
- Escobar Rodríguez, M., & Veeramah, K. (2024). Evaluation of ancient DNA imputation: A simulation study. *Human Population Genetics and Genomics*, 1–23. <https://doi.org/10.47248/hpgg2404010002>
- Garrido Marques, A., Rubinacci, S., Malaspinas, A.-S., Delaneau, O., & Sousa da Mota, B. (2024). Assessing the impact of post-mortem damage and contamination on imputation performance in ancient DNA. *Scientific Reports*, 14(1), 6227. <https://doi.org/10.1038/s41598-024-56584-3>
- Lee, H., O'Connor, B. D., Merriman, B., Funari, V. A., Homer, N., Chen, Z., Cohn, D. H., & Nelson, S. F. (2009). Improving the efficiency of genomic loci capture using oligonucleotide arrays for high throughput resequencing. *BMC Genomics*, 10, 646. <https://doi.org/10.1186/1471-2164-10-646>
- Leonard E. Baum & Ted Petrie. (1966). Statistical Inference for Probabilistic Functions of Finite State Markov Chains. *The Annals of Mathematical Statistics*, 37(6), 1554–1563. <https://doi.org/10.1214/aoms/1177699147>
- Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., & Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics (Oxford, England)*, 25(16), 2078–2079. <https://doi.org/10.1093/bioinformatics/btp352>
- Li, Y., Willer, C. J., Ding, J., Scheet, P., & Abecasis, G. R. (2010). MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genetic Epidemiology*, 34(8), 816–834. <https://doi.org/10.1002/gepi.20533>
- Li, Y., Willer, C., Sanna, S., & Abecasis, G. (2009). Genotype imputation. *Annual Review of Genomics and Human Genetics*, 10, 387–406. <https://doi.org/10.1146/annurev.genom.9.081307.164242>
- Meyer, M., & Kircher, M. (2010). Illumina Sequencing Library Preparation for Highly Multiplexed Target Capture and Sequencing. *Cold Spring Harbor protocols*, 2010, pdb.prot5448. <https://doi.org/10.1101/pdb.prot5448>
- Rubinacci, S., Hofmeister, R. J., Sousa da Mota, B., & Delaneau, O. (2023). Imputation of low-coverage sequencing data from 150,119 UK Biobank genomes. *Nature Genetics*, 55(7), 1088–1090. <https://doi.org/10.1038/s41588-023-01438-3>
- Rubinacci, S., Ribeiro, D. M., Hofmeister, R. J., & Delaneau, O. (2021). Efficient phasing and imputation of low-coverage sequencing data using large reference panels. *Nature Genetics*, 53(1), 120–126. <https://doi.org/10.1038/s41588-020-00756-0>
- Sousa da Mota, B., Rubinacci, S., Cruz Dávalos, D. I., G. Amorim, C. E., Sikora, M., Johannsen, N. N., Szmyt, M. H., Włodarczyk, P., Szczepanek, A., Przybyła, M. M., Schroeder, H., Allentoft, M. E., Willerslev, E., Malaspinas, A.-S., & Delaneau, O. (2023). Imputation of ancient human genomes. *Nature Communications*, 14(1), 3660. <https://doi.org/10.1038/s41467-023-39202-0>
- Voelkerding, K. V., Dames, S. A., & Durtschi, J. D. (2009). Next-generation sequencing: From basic research to diagnostics. *Clinical Chemistry*, 55(4), 641–658. <https://doi.org/10.1373/clinchem.2008.112789>

**A 37. OTDK
BIOLÓGIA SEKCIÓ FELHÍVÁSÁNAK
2. SZÁMÚ MELLÉKLETE**

TÉTELES SZERZŐI HOZZÁJÁRULÁS

Alulírott **Bán Botond** nyilatkozom, hogy a 37. OTDK-ra benyújtott **Imputáció az Archeogenetikában: Módszertani kihívások és lehetőségek a genom rekonstrukció javítására** című pályamunkámban leírtak esetében az alábbi munkafolyamatok elvégzésében vettem részt¹:

Részfeladat²	A részvétel mértéke (ötfokozatú skálán[#] jellemezve)
Mintavételezés koponyából illetve sziklacsontból	5
DNS kivonás	5
DNS tisztítás	5
UDG kezelés és blunt end repair	4
Adapter és index szekvencia ligálás	4
Szekvenciák szoftveres kezelése	3
Linux és hash alapok	5
Imputáció (GLIMPSE1) ^{m1}	2
Következtetések	4
Irodalmi háttér	5

Kelt: Szeged, 2024.11.27.

¹ A kutatási projekt azon részfadatait, amelyekben egyáltalán nem vett részt, ne tüntesse fel a listában.

² Az alábbi lista mintául szolgál, szabadon módosítható, bővíthető.