Univerzita Karlova

Přírodovědecká fakulta

Studijní program: Bioinformatika

Studijní obor: Bioinformatika

Sára Simandlová

Vyvozování demografické historie populací z genomových dat

Inferring the demographic history of populations from genomic data

Bakalářská práce

Školitel: RNDr. Radka Reifová, Ph.D.

Praha, 2022

Poděkování

Poděkování patří mé školitelce RNDr. Radce Reifové, Ph.D. za pomoc a cenné rady při psaní této práce. Ráda bych také poděkovala mé mamince MUDr. Martině Simandlové za podporu při studiu a za pomoc při výběru studijního oboru.

Prohlášení

Čestně prohlašuji, že jsem svoji bakalářskou práci na téma „Vyvozování demografické historie populací z genomových dat“ vypracovala sama. K zpracování mé práce jsem používala doporučenou literaturu a vše jsem konzultovala s mojí školitelkou RNDr. Radkou Reifovou, Ph.D.. Tato práce ani její podstatná část nebyla předložena k získání jiného nebo stejného akademického titulu.

V . . . . . , dne . . . . . Podpis autora . . . . . .

Abstrakt

V současné době není obtížné získat genomová data i z nemodelových organismů. Tato data nám mohou přinést informace o demografické historii populací. Bylo vyvinuto mnoho statistických vyvozovacích postupů k odvození demografické historie populací z genomových dat, jejichž popisem se zabývám v této bakalářské práci. V úvodu čtenáře seznamuji s důležitými pojmy při analýze demografické historie populací. Dále popisuji různé typy genomových dat, které se dají použít k vyvozování demografické historie populací. Následně představuji průběh experimentu při vyvozování demografické historie populací, kde rozvádím jednotlivé kroky a uvádím přístupy a metody, které se v dnešní době využívají. Poskytuji základní přehled teorie a logiky každého přístupu. Dále čtenáře seznamuji s nejpoužívanějšími softwarovými balíčky pro vyvozování demografické historie populací a poskytuji jejich srovnání.

Klíčová slova: populační genetika, demografická inference, statistická inference, celogenomová data

Abstract

Currently, it is not difficult to obtain genomic data even from non-model organisms. These data can give us information about the demographic history of populations. Many statistical inference procedures have been developed to infer the demographic history of populations from genomic data, and I describe them in this thesis. In the introduction, I introduce the reader to important concepts in the analysis of the demographic history of populations. I then describe the different types of genomic data that can be used to infer the demographic history of populations. I then present the flow of an experiment in inferring the demographic history of populations, where I elaborate on the steps and present the approaches and methods that are used today. I provide a basic overview of the theory and logic behind each approach. I also introduce the reader the most commonly used software packages for inferring demographic histories of populations and provide a comparison between them.

Keywords: population genetics, demographic inference, statistical inference, whole genome data

Seznam použitých zkratek:

ABC – *approximate Bayesian computation;* aproximační Bayesovský výpočet

DNA – *deoxyribonucleic acid;* deoxyribonukleová kyselina

G-PhoCS – *A Generalized Phylogenetic Coalescent Sampler;* Generalizovaný Fylogenetický Koalescenční Vzorkovač

HMM – *hidden Markov model; skrytý Markovův model*

IBD – *Identity by descent;* Identita podle původu

IBS – *Identity by state;* Identita podle stavu

MAGIC - *Minimal-Assumption Genomic Inference of Coalescence*; Minimální Předpoklad Genomické Inference Koalescence

MCMC – *Markov Chain Monte Carlo;* Monte Carlo pomocí Markovova řetězce

MRCA – *Most Recent Common Ancestor;* Poslední Společný Předek

MSMC – *The Multiple Sequentially Markovian Coalescent;* Vícenásobná sekvenčně markovská koalescence

PCR – *polymerase chain reaction*; *polymerázová řetězová reakce*

PSMC – *The Pairwise Seqentially Markovian Coalescent;* Párová sekvenčně markovská koalescence

SFS – *site frequency spectrum;* frekvenční spektrum míst

SNP – s*ingle-nukleotide polymorphism;* jednonukleotidový polymorfismus

STR – *short tandem repeats;* krátké tandemové repetice

Obsah

1. Úvod

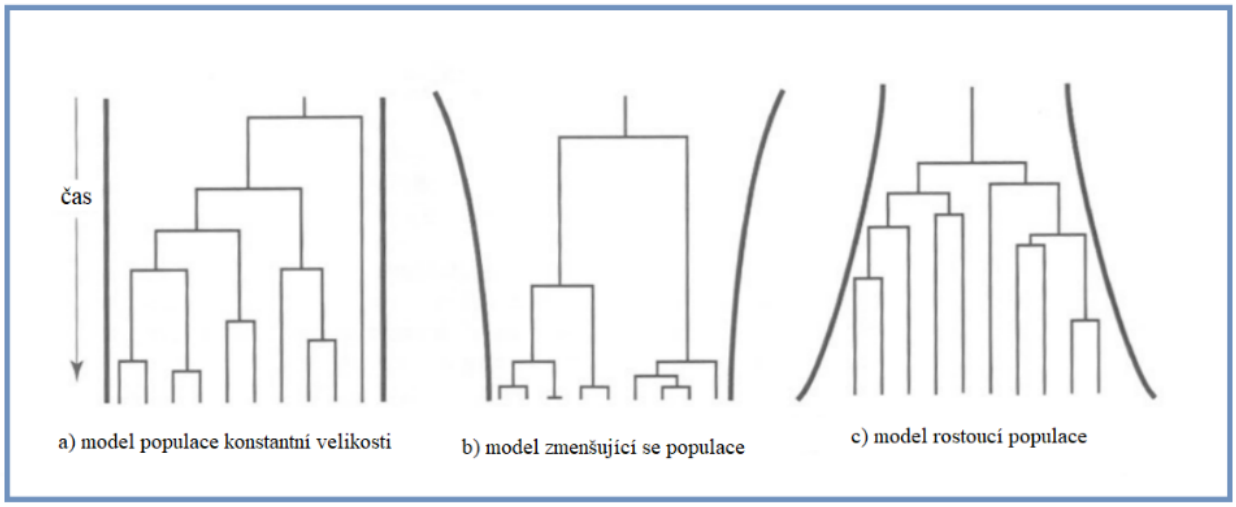
Biologové již dlouhá léta pozorují a zkoumají, jak se jednotlivé populace vyvíjejí a co je historicky mohlo ovlivnit. Výsledky těchto pozorování dokazují, že současné rozložení populací je následkem mnoha složitých historických a prehistorických demografických událostí, které formovaly, nejen, variabilitu genomů.

Populaci můžeme obecně popsat jako soubor jedinců téhož druhu, který se nachází na jednom určitém místě v daném čase. Demografie je obor zabývající se velikostí populace, její strukturou, vývojem a dalšími charakteristikami *(Achille Guilard, 1855).* Analýza demografického procesu umožňuje zobecnit zákony o populačním vývoji, najít vzory nebo formulovat předpoklady o budoucím vývoji populace.

Myšlenka, že by informace o demografické historii mohly nést sekvenční data vznikla na počátku dvacátého století *(Hirschfeld 1919),* avšak byla realizována až v sedmdesátých letech 19.století z důvodu, že až v této době byly vynalezeny vhodné statistické nástroje a začalo se pracovat se souhrnnými statistikami *(viz kap. Průběh vyvozování a souhrnné statistiky)*, což umožnilo populačním genetikům poprvé využívat sekvenční data k odhadu demografické historie populací.

Sekvenční data v sobě nesou velké množství informací, ve kterých se učíme číst a snažíme se získat přesnější a jednoznačnější výsledky. Kromě toho jsou tyto populačně genetické přístupy široce použitelné pro jakýkoli druh živočichů nebo rostlin.

Populační genetici sledují demografickou historii populací skrze sekvenční data pomocí genových genealogií. Genealogie popisují vztahy mezi kopiemi určitého genu v populaci napříč generacemi. Termín „genealogie“ je kombinací dvou řeckých slov: *genea* = původ/rodová linie a *logos* = vědění *(Farmer, 19.století).* Demografické faktory, jako je například změna ve velikosti populace v minulosti nebo čas divergence populací, ovlivňují tvar genealogií *(Obr. 1)*.



Obr. 1 : Tvar genealogie v závislosti na modelu populace

Na obrázku je zobrazeno, jakým způsobem ovlivňují změny ve velikosti populace tvar genealogie: a) tvar genealogie v případě konstantní velikosti populace, b) tvar genealogie v případě klesající velikosti populace, c) tvar genealogie v případě rostoucí velikosti populace (Garrigan et al., 2002, převzato a upraveno).

K vyvozování demografické historie používáme genomová data. Jak se čtenář dozví dále, dají se využít celé sekvence genomů, sekvence transkriptomů, nebo data získaná z RAD-sekvenování.

Při vyvozování především odhadujeme demografické parametry, mezi které například řadíme již zmíněnou změnu velikosti populace v minulosti *(Obr. 1)*, aktuální velikost populace, míru migrace mezi populacemi, nebo například čas divergence populací. George E. P. Box, jeden z největších statistiků 20.století, v roce 1976 prohlásil: „V podstatě všechny modely jsou špatné, ale některé jsou užitečné", *(Box, 1976).* Přesně takhle funguje demografické vyvozování historie populací. Nikdy se nám nepovede model dané populace/populací vyvodit přesně, ale postupným učením a zdokonalováním se dostáváme k přesnějším a jednoznačnějším výsledkům.

Při zpracování své bakalářské práce jsem si kladla rovnou několik cílů. V první řadě jsem chtěla čtenáře seznámit s vhodnými genomickými daty pro vyvozování demografické historie populací. Následujícím cílem bylo obecně popsat a přiblížit průběh experimentu demografického

vyvozování populací. Následně jsem chtěla čtenáře seznámit s vybranými přístupy při demografickém vyvozování a stručně popsat logiku a princip každého přístupu. Hlavním cílem této práce bylo srovnání popsaných softwarových balíčků.

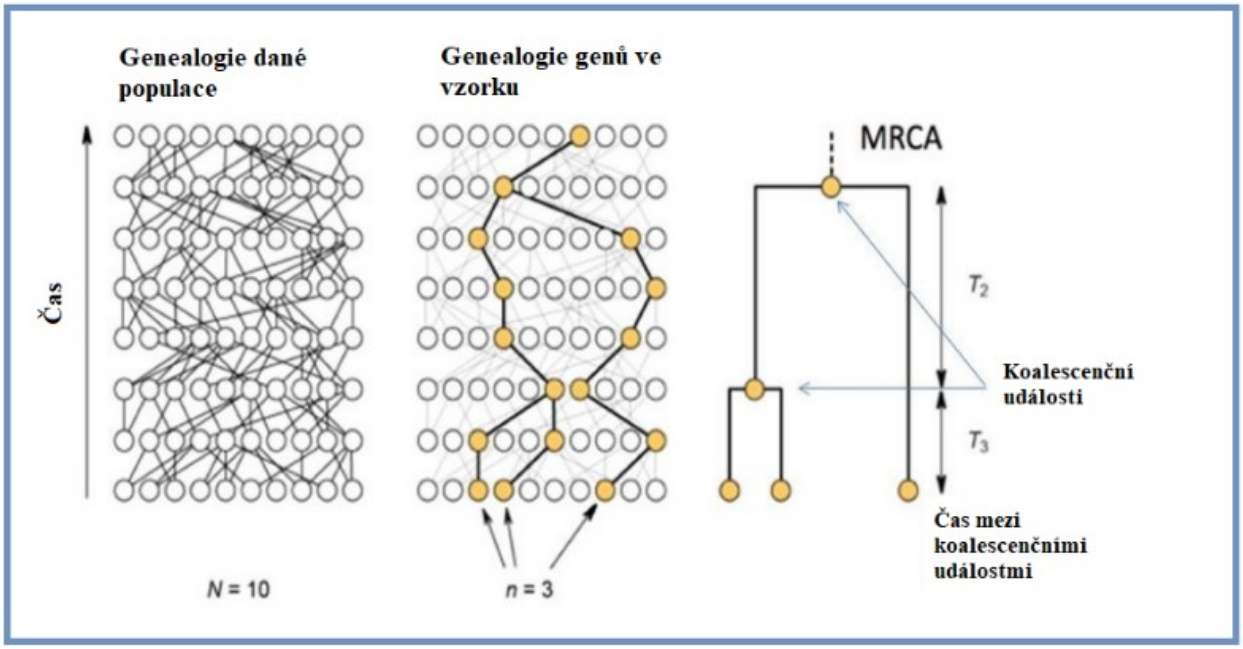
1.1. Genealogie a Teorie koalescence

Jak jsem zmiňovala výše, biologové používají k zobrazování vztahů mezi kopiemi určitého genu genové genealogie *(Obr. 2).* Gen může mít v populaci více variant a každá tato varianta se nazývá alela. Jednotlivé alely se pak mezi sebou liší nukleotidovou sekvencí. Sekvenováním alel určitého genu v populaci lze genealogický strom konstruovat.

Teorie koalescence je teorie zabývající se průběhem genealogií. Pokud sledujeme genealogii alel určitého genu zpět v čase, dochází postupně ke splývání jednotlivých linií (koalescencím) až se dostaneme k takzvanému poslednímu společnému předku (*most recent common ancestor*, MRCA) *(Obr. 2)*. Matematickou teorii koalescence původně vyvinul na počátku 80. let minulého století John Kingman. Teorii koalescence můžeme vylíčit jako matematický model, který popisuje průběh genealogií, kdy postupujeme opačně v čase než u klasických modelů populační genetiky, jako je například Wright-Fisher model. Pravděpodobnost koalescenční události v předchozí generaci vyjadřujeme vzorcem $P = \frac{1}{N}$ pro haploidní organismy a $P = \frac{1}{2N}$ pro diploidní organismy, kde N je počet jednotlivých kopií genů v populaci *(Kingman, 1982).*

Důležitý faktor, který je třeba zohledňovat při vyvozování demografické historie populací, je genetická rekombinace, která způsobuje, že evoluční osudy genů, které jsou od sebe v genomu dostatečně daleko, mají nezávislé evoluční historie, nezávislé genealogie *(Stumpf et al., 2003)*. Proto lze vytvářet genealogie jen pro krátké oblasti na chromozomech, v rámci nichž je pravděpodobnost rekombinace malá. S touto limitací se však pojí fakt, že tyto krátké sekvence mohou postrádat dostatek polymorfismů nutných k vyvození genealogie. Každý lokus, čímž myslíme pozici genu v molekule DNA, má základní genealogii popisující jeho historii. Lokusy blízko sebe, které postrádají historickou rekombinaci, budou sdílet přesně stejné genealogie. Stejné genealogie také samozřejmě sdílí geny v oblastech genomu, které nerekombinují. V tomto případě mluvíme o chromozomu Y (chrY), který je děděný z otce na syna a mitochondriální DNA (mtDNA), která se v drtivé většině dědí pouze po matce. Z tohoto důvodu lze použít pro konstrukci genealogie celou sekvenci ChrY nebo mtDNA. Proto se sekvence mtDNA a chrY často používají při studiu genealogií v populační genetice a především je výhodné, že mtDNA nám umožňuje studovat historii mateřské linie, kdežto chrY otcovské linie.

Lokusy, které od sebe leží daleko (v rámci jednoho chromozomu, nyní nemluvím o chrY nebo mtDNA) nebo lokusy ležící na jiných chromozomech mají v podstatě nezávislé genealogie. Sekvenční data z jednoho lokusu nám přinesou jen jednu realizaci evolučního modelu. Kvůli této limitaci se vyplatí vzorkování více lokusů napříč celým genomem, jelikož toto vzorkování zvýší jistotu našeho výsledku *(Beichmann et al., 2018).*



Obr. 2: Genealogie a teorie koalescence

V levé části obrázku je genealogie určitého genu. Jednotlivá kolečka znázorňují jedince. V tomto případě se jedná o diploidní jedince, proto je každé kolečko spojené s předchozí generací dvěma čarami. Uprostřed je zobrazena genová genealogie tří vybraných alel, kde pozorujeme proces splynutí linií v posledního společného předka (MRCA). V pravé části je zobrazen rozbor koalescenčních událostí (Leblois 2010, "La théorie de la coalescence et ses applications", prezentace přednášky, ENS Lyon, převzato a upraveno).

Je patrné, že koalescenční procesy vedou k vysoce variabilním genovým genealogiím, které lze využít k vyvozování demografické historie populací. V následujících částech bakalářské práce popisuji celkový proces při vyvozování demografické historie populací, jehož logika a princip vychází z koalescenční teorie.

2. Data

Při vyvozování demografické historie nás zajímají určité oblasti genomu, které dokážou posloužit jako správná vodítka pro vyvozování. V první řadě se jedná o jednonukleotidové polymorfismy (*Single nukleotide polymorphism,* SNP). Jde o variaci v jediném nukleotidu v určité pozici genomu *(Brzezinski et al., 2006)*. Pomocí SNP jsme schopni odhadnout genealogie pro jednotlivé sekvence *(Obr. 3)*. Dalším z genetických markerů využívaných při vyvozování jsou krátké tandemové repetice (*Short tandem repeats,* STR). Jsou to specifické sekvence DNA, které se vyskytují ve velkém množství rozptýlené po celém genomu. Jedná se o repetice nukleotidů, jejichž počet opakování je pro každého jedince jedinečný, a proto jsou tato data velice užitečná pro vyvozovací metody *(Shi et al., 2010)*.

Nyní si pojďme povědět něco více o sekvenačních technikách, pomocí kterých data získáváme. Od počátku 21. století dominují na trhu sekvenační techniky hromadně nazývané jako sekvenování nové generace (*next generation sequencing,* NGS). První techniky sekvenování nové generace vytvořily komerční firmy – 454 Life Sciences (Roche), Illumina a Applied Biosystems. Tyto techniky jsou oproti starším metodám, jako bylo například Sangerovo sekvenování, mnohem rychlejší, levnější, ale především umožňují v jednom běhu získat mnohem více dat ve srovnání se staršími metodami. V dnešní době se nejčastěji setkáme s technikou od firmy Illumina. V první fázi této techniky se vytvoří sekvenační knihovna, dále se DNA naláme na krátké fragmenty (kolem 300pb). Pomocí můstkové PCR dochází k amplifikaci fragmentů. Samotná detekce nukleotidů probíhá pomocí fluorescenčních záblesků nově dosedajícího nukleotidu, neboli každý ze čtyř nukleotidů nese jiný fluorofor, a tedy vyzáří světlo jiné barvy. Jasnou výhodou této metody je nízká cena sekvenování a nízká chybovost *(Van Dijk et al., 2014)*.

Kromě metod sekvenování nové generace vznikají také takzvané metody třetí generace, které jsou také nazývány jako sekvenování jedné molekuly (*single-molecule sequencing,* SMC). V dnešní době se nejvíce využívají techniky vytvořené firmou Pacific Biosciences a Oxford Nanopore Technologies. Výhodou těchto metod jsou delší získaná čtení, ale značnou nevýhodu činí vysoká chybovost a vysoká cena sekvenování. Tím pádem se tyto metody pro získání dat pro vyvozování demografické historie populace příliš nehodí.

Sekvenční data využívaná při demografickém vyvozování můžeme rozdělit na dva typy: celogenomová data a redukovaná genomová data. Redukovaná genomová data můžeme získat například použitím restrikčních enzymů (RAD-seq), sekvenováním RNA, nebo pomocí takzvaných sequence capture.

Celogenomová data jsou data získaná sekvenováním veškeré DNA organismu. Přesněji řečeno, sekvenuje se veškerá chromozomální DNA organismu a DNA obsažené v mitochondriích (pro rostliny v chloroplastech, ekvivalentně pro RNA organismy). V praxi jsou genomové sekvence, které jsou téměř úplné, také nazývány celogenomovými sekvencemi.

RAD-seq (*Restriction-site associated DNA sequencing*) je přístup, jak z genomu sekvenujeme jen jeho určitou část. V tomto případě se sekvenují oblasti genomu, které byly vybrány na základě délky po restrikčním štěpení DNA. Pomocí RAD-seq jsou objevovány SNP v náhodných a především nekódujících oblastech genomu. Metoda zahrnuje stříhání genomu s alespoň jedním restrikčním enzymem, který specificky rozeznává úsek dlouhý 5-6 nukleotidů (pokud jsou použity dva různé restrikční enzymy najednou, hovoříme o double digest RAD sequencing (ddRAD-seq)). Poté jsou na základě velikostní selekce vybrány zájmové fragmenty určité délky a sekvenovány metodou sekvenování nové generace (přednostně na sekvenceru Illumina). RAD-seq poskytuje až miliony sekvencí délky 50-600bp *(Davey a Blaxter, 2010).*

Při RNA sekvenování není sekvenován celý genom, ale jen ty části, které jsou přepisované do RNA *(Zhang et al., 2017)*. RNA molekuly jsou izolovány, a reverzně přepsané do cDNA (DNA komplementární k mRNA) pomocí reverzní transkriptázy a následně sekvenovány.

Třetí možnost získání redukovaných dat jsou sequence capture. Tento přístup využívá dlouhé biotinylované oligonukleotidové sondy k hybridizaci se zájmovými oblastmi *(Grover et al., 2012)*. Jedná se například o sekvence konkrétních genů, nebo o místa s konkrétními oligonukleotidovými sekvencemi.

V neposlední řadě se stále používá již zmiňovaná starší Sangerova metoda sekvenování k získání sekvencí menšího množství lokusů, které lze také využít pro vyvozování demografické historie populací *(Gutenkunst et al., 2009)*.

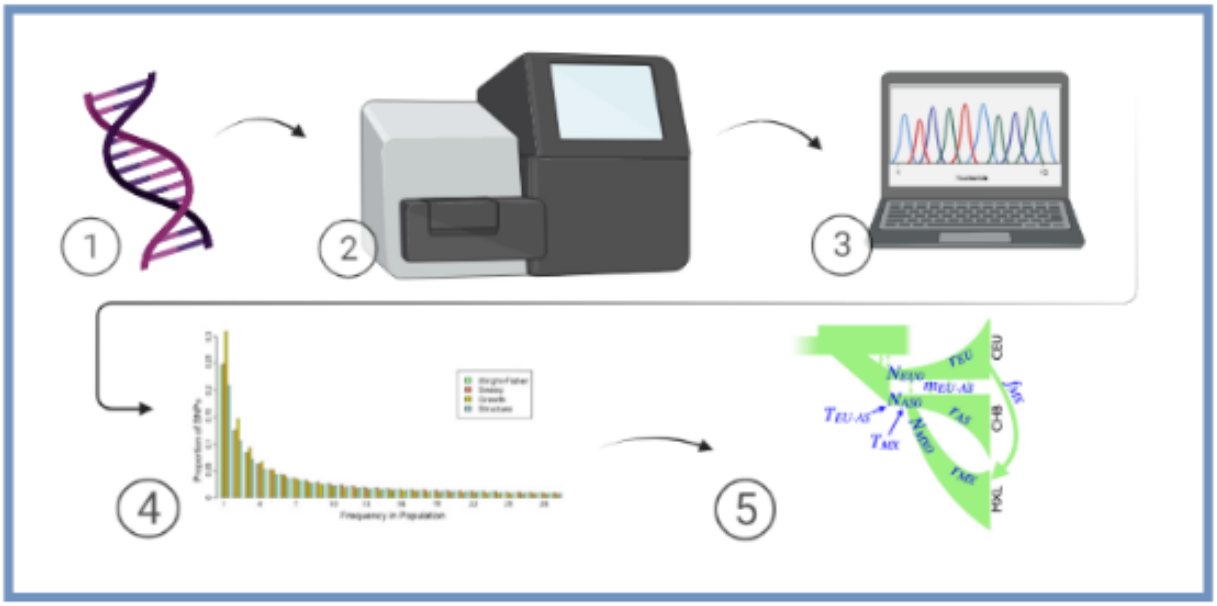
3. Průběh vyvozování a souhrnné statistiky

Na začátku experimentu odebereme vzorek, upravíme ho pro sekvenování a pomocí vhodného sekvenátoru osekvenujeme *(viz kap. 2. Data)*. Jakmile získáme sekvence, koukáme se na ty úseky nebo části genomu, které nesou důležité informace pro náš experiment. Dále z těchto získaných dat vypočítáváme souhrnné statistiky. Souhrnné statistiky jsou hodnoty vypočtené z dat tak, aby reprezentovaly maximální množství informací v co nejjednodušší podobě *(Beamount et al., 2002)*. Volba správné souhrnné statistiky je jedním z nejdůležitějších kroků při demografickém vyvozování. Na základně předpokládaného modelu volíme vhodné souhrnné statistiky. Jelikož poměrně velké množství metod, které se využívají při demografickém vyvozování a budu je popisovat ve svojí práci později, používají jako souhrnnou statistiku frekvenční spektrum míst *(SFS, site frequency spectrum),* ráda bych se v této úvodní části věnovala jiným, podobně intenzivně využívaným statistikám v těchto analýzách.

Jednou z používaných souhrnných statistik je statistika π. Jde o běžně používanou statistiku v populační genetice, kterou poprvé uvedli Nei a Li v roce 1979. Tato statistika je definována jako průměrný počet rozdílů nukleotidů na lokus mezi sekvencemi DNA v populaci . Slouží k empirickým odhadům genetické diverzity a můžeme ji také definovat jako průměrnou heterozygositu . Další souhrnnou statistikou, kterou při vyvozování demografické historie populace využíváme, je hodnota genetického polymorfismu, kterou značíme θ. Genetický polymorfismus definujeme jako existenci dvou nebo více alel v jednom lokusu, převyšující svým výskytem 1 % v populaci. Míra genetického polymorfismu je přímo úměrná mutační rychlosti a efektivní velikosti populace (N_e). Mutační rychlost můžeme určit jako frekvenci nových mutací na generaci a efektivní velikost populace, neboli N_e, jako velikost ideální panmiktické populace, ve které by všechny genetické procesy probíhaly stejnou rychlostí jako v dané reálné populaci. Neméně důležitou statistikou je *Tajima's D test*. Tato statistika je pojmenována po Fumio Tajimovi, který ji vytvořil. Statistika srovnává hodnoty dvou výše zmiňovaných odhadů genetické diverzity, θ – hodnotu genetického polymorfismu a π – průměrnou heterozygositu *(Tajima, 1983)*. Tato souhrnná statistika se v populační genetice například hojně využívá při studiu stupně hvězdicovitosti rodokmenů. *(Lohse et al., 2009)*. Poslední souhrnnou statistikou, kterou zmíním, je F_{st}. F_{st} vyjadřuje míru genetické diferenciace mezi populacemi a odráží rozdíly ve frekvencích alel mezi populacemi, proto se hodí při určování

míry divergence. Souhrnných statistik, které se využívají, je samozřejmě více, ale rozsáhlejší popis by byl už předmětem jiné bakalářské práce.

Po správném zvolení souhrnné statistiky vytváříme scénáře (podoby), které by dle znalostí nejvíce odpovídaly našim datům. Jedním z posledních kroků, je testování navrhovaných scénářů, ve kterém, jak se později dozvíme, hraje velice důležitou roli pravděpodobnost a statistika. Samotné testování je časově nejnáročnější část experimentu.



Obr. 3: Postup při vyvozování demografické historie populací

Na obrázku je zobrazeno přibližné schéma postupu experimentu při vyvozování demografické historie populací. Z jedince/jedinců odebereme vzorek DNA, dále genetickou informaci osekvenujeme na vhodném sekvenátoru, ze získaných dat určíme souhrnnou statistiku našeho zájmu, kterou následně porovnáváme se simulacemi a snažíme se co nejvíce přiblížit reálným datům. Výsledkem je model populace, který například znázorňuje změnu velikosti populace v minulosti. (část obrázku použita z review Gutenkunst et al., 2009, vytvořeno pomocí webové aplikace https://biorender.com/).

3.1. Frekvenční spektrum místa, neboli SFS

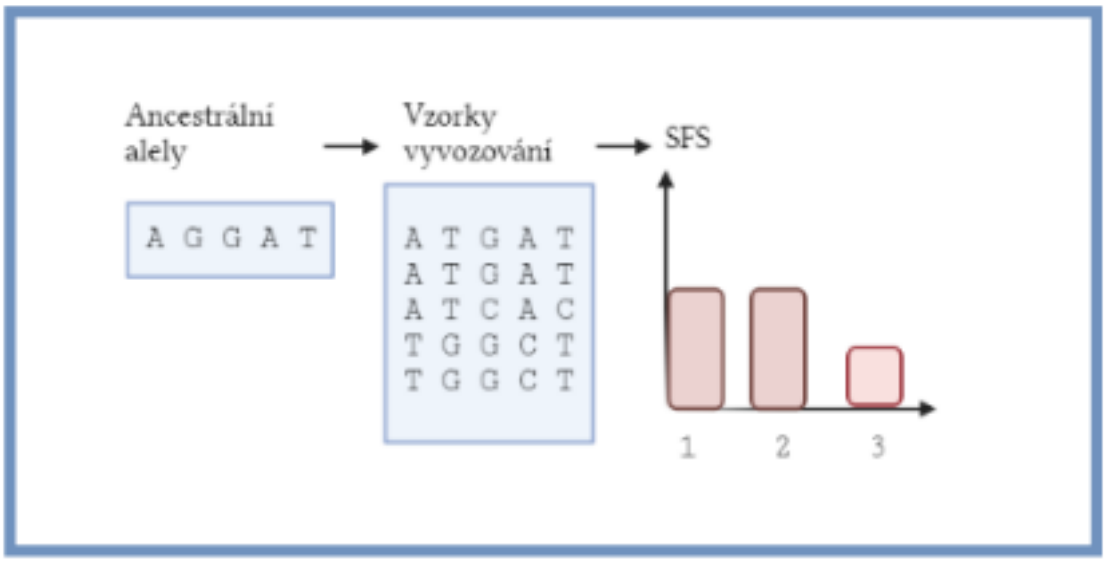
Jak jsem již v kapitole o průběhu experimentu zmínila, hojně využívanou souhrnnou statistikou při demografickém vyvozování historie populací je právě SFS. Pojďme si na této sumární statistice více přiblížit a ukázat, jak na základně dat vyvodit určitý model. Začneme ale od úplného začátku a nejdříve si vysvětlíme, co nám SFS přesně určuje.

Je parné, že i když mají genové kopie společného předka, liší se mutacemi. V tomto případě se zajímáme především o SNP. Očekávaný počet mutací oddělující jednotlivé kopie genu můžeme vyjádřit následujícím vztahem:

$$\theta = 4N\mu,$$

kde N je velikost populace a μ je mutační rychlost na lokus za generaci. Je zřejmé, že očekávaný počet mutací je ovlivněn velikostí populace. Čím menší populace je, tím méně jsou jednotlivé geny variabilní a zároveň je kratší čas koalescence, tedy kratší čas k poslednímu společnému předku (MRCA).

SFS tedy můžeme jednoduše definovat jako distribuci frekvencí pro jednotlivé mutované alely *(Obr. 3)*. Tento přístup není tolik náročný na data, jelikož vyžaduje nezávislá SNP, hodí se pro něj například i data získaná z RAD sekvenování *(viz kap. 2.2. Data)*.

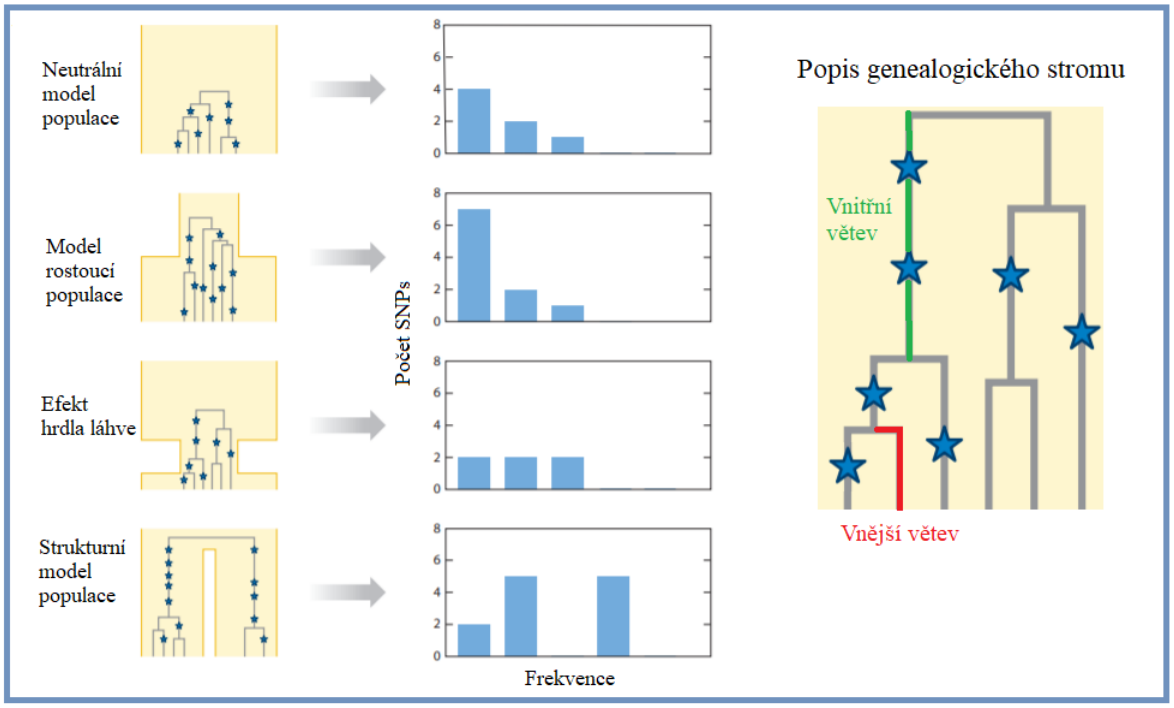


Obr. 4: Frekvenční spektrum míst

Na obrázku je zjednodušeně ukázáno, jakým způsobem vyhodnocujeme SFS. Na základě ancestrálních alel vyhodnocujeme počty SNP. Utvoříme alignment a díváme se, jakým způsobem se alely měnily. Na obrázku je zobrazena situace, kde vyobrazený lokus nese 5 SNP. V SFS osa y určuje počet mutovaných alel s danou frekvencí a na ose x je zobrazen absolutní počet mutovaných alel v populaci (vytvořeno pomocí webové aplikace, https://biorender.com/, inspirováno Beichman et al., 2018).

Při vyvozování demografické historie populací je důležitým faktorem i to, od kolika jedinců vyžadujeme data. V tomto případě platí jednoduchá přímá úměrnost, neboli čím více osekvenovaných jedinců k dispozici máme, tím směrodatnější a přesnější výsledky dostáváme. (*Beichman et al. 2018).*

Na základě koalescenční teorie, kterou jsme si představili na začátku práce *(viz kap. 1.1. Genealogie a Teorie koalescence)*, odvozujeme, jakým způsobem demografie ovlivňuje SFS. Různé demografické scénáře mění tvar a délku větví genealogií, které díky tomu mění i samotné SFS, což si přiblížíme na následujícím obrázku *(Obr. 5).*



Obr. 5: SFS při určitých modelech populace, popis genealogického stromu

Historie populace ovlivňuje tvar genealogií a SFS. Žluté oblasti vlevo označují historii každé populace. Tyto historické situace vedou k určitému tvaru genealogií v každém modelu. Modré hvězdy označují mutace v rodokmenech. Histogramy uprostřed obrázku znázorňují SFS pro dané modely. V pravé části obrázku je popis větví genealogického stromu (Beichman et al., 2018, převzato a upraveno).

Pojďme si tedy přiblížit, jakým způsobem ovlivňují změny ve velikosti populace SFS. Jak si můžeme všimnout na obrázku (*Obr. 5)* při neutrálním modelu populace je SFS rozložen rovnoměrně. Pokud se podíváme na SFS velké a malé populace, ke koalescencím bude docházet spíše u malých populací. Genealogie rostoucí populace mají dlouhé vnější větve a krátké větve vnitřní. U dlouhých větví dochází ke koalescencím méně často, než u větví krátkých. Celkově tedy při modelu rostoucí populace nese SFS hodně nízkofrekvenčních SNP *(Beichman et al. 2019)*.

Známým termínem v demografickém vyvozování je efekt hrdla láhve. Efekt nastává při velkém poklesu jedinců v populaci po kratší dobu. Při efektu hrdla láhve je SFS rovnoměrně rozložen právě z důvodu, že dochází ke snížení genetické variability v populaci, tudíž i ke kontrakci

populace. Množství nízkofrekvenčních mutací bude v tomto případě velice nízké a k těmto mutacím bude docházet srovnatelně často jako k vícefrekvenčním mutacím *(Beichman et al. 2019).*

A jak se SFS chová, pokud ho provádíme pro dvě populace? Pokud testujeme dvě populace, které se od sebe oddělily před dlouho dobou a tudíž je míra migrace mezi těmito dvěma populacemi téměř nulová, je zřejmé, že se jedinci budou pářit výhradně s jedinci ze stejné subpopulace. Proto je tvar SFS více než odpovídající, a jak je zřejmé z obrázku *(Obr. 5).* SFS nese především mutace se střední frekvencí *(Beichman et al. 2019).*

Jak je popsáno výše, demografická historie populace může mít dopad na SFS, tudíž je SFS užitečná souhrnná statistika k odvození demografických parametrů. V prvním kroku analýzy sestavují biologové empirické SFS ze sekvenčních dat. Poté je koalescenční teorie použita ke generování predikované SFS pro konkrétní demografický model. Jakmile je vygenerováno SFS předpovídaným demografickým modelem, posuzujeme vhodnost predikovaného SFS k empirickému SFS, obvykle v rámci pravděpodobnosti *(Rosen et al. 2018).*

Nyní se podíváme na odhad SFS pomocí koalescenčních simulací. Pravděpodobnost daného SFS vstupu i při modelu θ můžeme vyjádřit následujícím vztahem:

$$p_i = E(t_i \mid \theta)/E(T \mid \theta),$$

kde t_i je celková délka všech větví vedoucích k i koncovým uzlům a T je celková délka stromu. S lehkou matematickou úpravou se dá tento vzorec aplikovat přes všechny koalescenční simulace a pravděpodobnost vstupu i se tím pádem vypočítává pro všechny vytvořené simulace *(Spence et al. 2016)*.

SFS je v dnešní době velice využívaný přístup a jak se dozvíme později, pracuje s ním velké množství softwarových balíčků. Spektra frekvencí jednotlivých alel byla použita na skupinu Afričanů, Evropanů a Asiatů a bylo dokázáno, že u Asiatů a Evropanů došlo k efektu hrdla láhve, kdežto u Afričanů ne *(Marth et al. 2004)*.

3.2. Aproximační Bayesovský výpočet

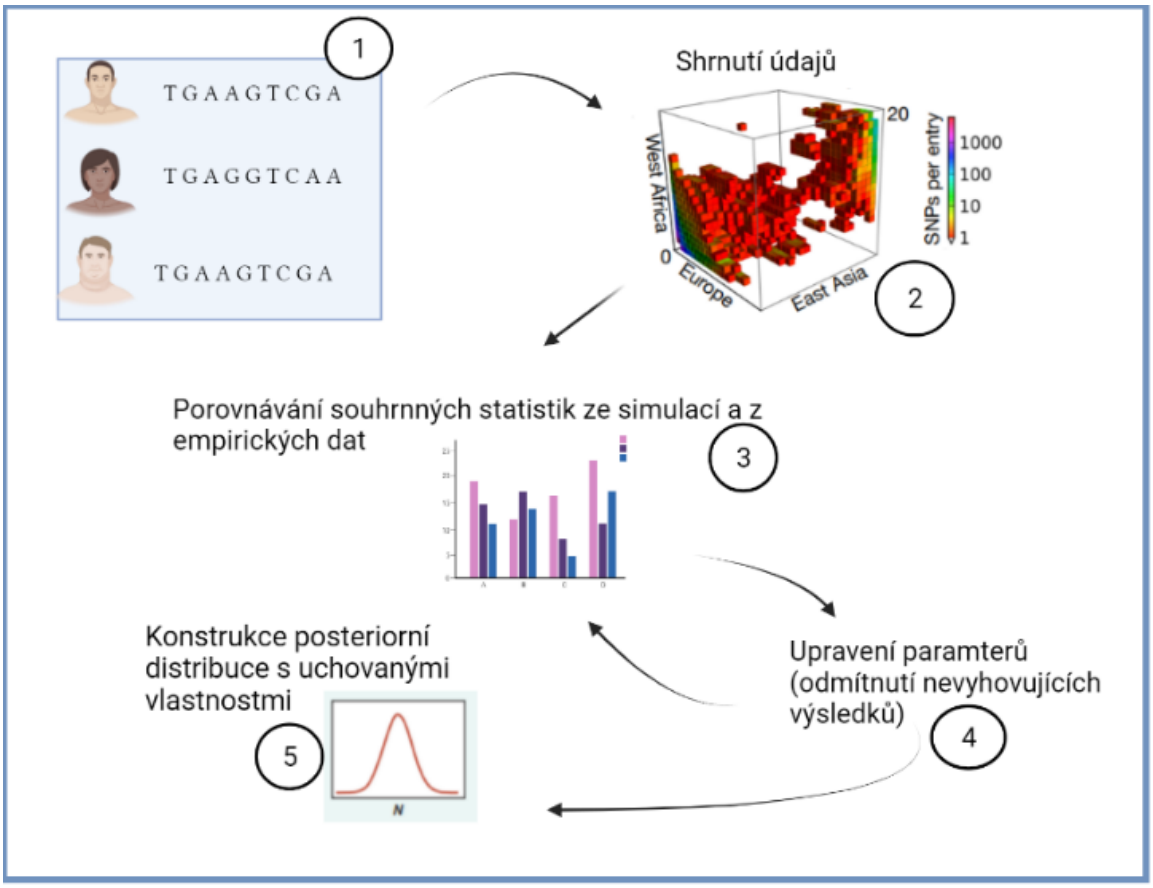
Jedním z přístupů pro vyvozování demografických historií je přibližný Bayesovský výpočet (*approximate Bayesion computation,* ABC). Přístup ABC obchází přesné výpočty pravděpodobností pomocí souhrnných statistik a simulací. Zásadní práce, která ovlivnila pohled na tuto problematiku byla publikována v roce 1997 Simonem Tavarém společně s jeho kolegy *(Tavaré et al., 1997).*

Autorem Bayesovy věty, která je jádrem Bayesovské inference, byl Thomas Bayes. Bayesovská inference nám umožňuje vypočítat pravděpodobnost struktury modelu při daných datech neboli $l(parametry \mid data)$. Označíme si parametrické hodnoty P a data jako D. Ponecháme $l()$ pro pravděpodobnostní funkce označující pravděpodobnosti nepozorovatelné náhodné veličiny P, neboli parametrů. Jako $f()$ označíme pravděpodobnostní rozdělení pro pozorovatelné náhodné veličiny neboli naše data. S tímto značením aplikujeme Bayesův vzorec:

$$l(P|D) = \frac{f(D|P)\, l(P)}{f(D)},$$

kde $l(P)$ je priorní funkce a nese informace o parametrických hodnotách na základě priorních informací. $f(D|P)$ je věrohodnostní funkce a jde o pravděpodobnost, při které pro dané $l(P)$ budou v našem modelu generována data D. $l(P|D)$ je posteriorní funkce a udává nám pravděpodobnost parametrických hodnot při započítání pozorovaných dat. $f(D)$ označujeme jako evidenci. Jde o celkovou hodnotu pravděpodobnosti dat D, kterou odvodíme sečtením všech hodnot parametru vážených jejich pravděpodobnostmi. Z posteriorní funkce získáváme posteriorní distribuci a z priorní funkce priorní distribuci *(převzato a upraveno, Sunnåker et al., 2013).*

Hodnoty demografických parametrů z priorních distribucí jsou poté přijímány, pokud vytvářejí souhrnnou statistiku, která je blízko hodnotám z empirických dat, čím se získá posteriorní parametr distribuce *(Beichman et al., 2018)*. Jelikož ABC přístup používá koalescenční simulace k vytváření souhrnné statistiky, dokáže si například solidně poradit i s hodně polymorfními sekvencemi DNA *(Cabrera et al., 2017)* . Další výhodou přístupu je jeho obecnost. Tento přístup se využívá nejen v demografické inferenci, ale můžeme se s ním setkat například v systémové biologii, ekologii a epidemiologii *(Beaumont et al., 2010)*.



Obr. 6: Schéma procesu při Aproximačním Bayesovském výpočtu

Na obrázku můžeme vidět zjednodušené schéma aplikace Aproximačního Bayesovského výpočtu při vyvozování demografické historie populací. V prvním kroku získáváme potřebná data, poté vytváříme souhrnné statistiky. Třetí krok je krokem porovnávacím, kdy porovnáváme souhrnné statistiky ze simulací a hodnoty souhrnných statistik z empirických dat. Pokud se výsledky blíží, statistiku přijímáme, nevyhovující výsledky odmítáme. V závěrečném kroku sestrojujeme posteriorní distribuce s uchovanými vlastnostmi. (Část obrázku z článku Gutenkunst et al., 2009, vytvořeno pomocí webové aplikace https://biorender.com/).

Pokud se rozhodneme vyvozovat demografickou historii populace pomocí ABC přístupu, v prvním kroku naší práce vybereme libovolně složitý demografický model, který by měl odpovídat našim datům. Z tohoto modelu odhadujeme parametry, které čerpáme ze získaných sekvenčních dat. Pro každý parametr, který sledujeme, je zvolena priorní distribuce. Dále vytváříme simulace, které jsou utvořené na základě hodnot paramerů z priorních distribucí.

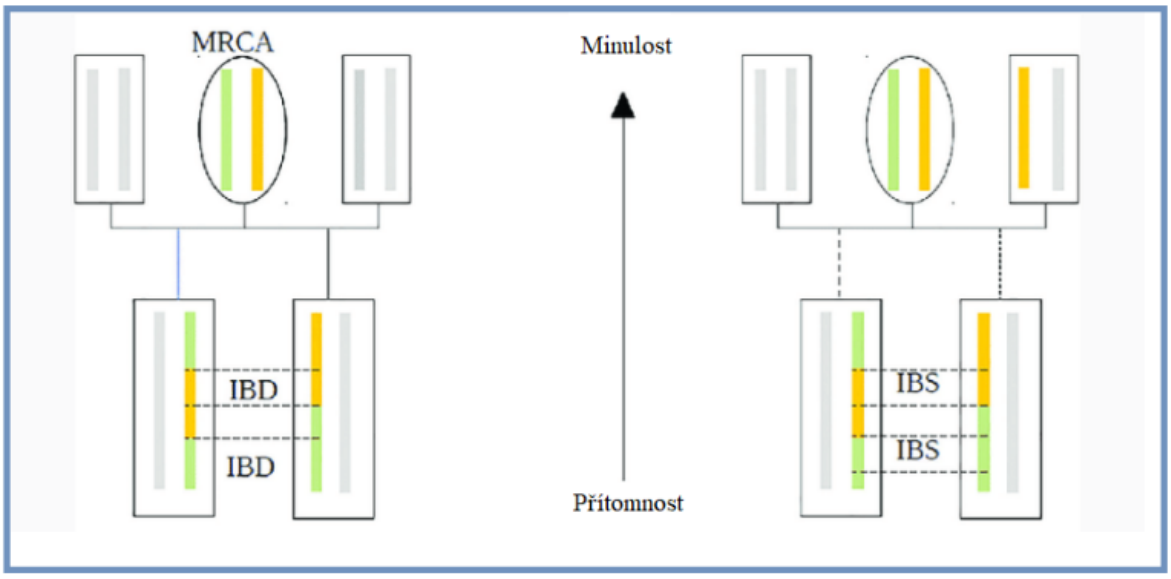
Simulace utváříme tak, aby odpovídaly specifikům empirických dat, která se používají v dané studii (dodržení počtu osekvenovaných jedinců, počet lokusů a tak dále). Výsledná souhrnná statistika ze simulované datové sady se porovnává se statistikami z empirických dat a pokud je tato statistika dostatečně blízko očekávané hodnotě, hodnoty parametrů z priorní distribuce jsou zachovány a přispívají k posteriorní distribuci. Pokud se tato souhrnná statistika neblíží hodnotě z empirických dat, parametry jsou odmítnuty. Tento postup opakujeme, dokud nemáme dostatečný počet replik, kde mluvíme o řádech tisíců. Distribuce přijatých hodnot parametrů bude představovat aproximační posteriorní distribuci *(Obr. 6) (Beichman et al. 2019).*

Některé ABC balíčky zahrnují Markovovy řetězce Monte Carlo *(Markov chain Monte Carlo*, MCMC) *(Beaumont et al., 2004)*. Jedná se o heuristický přístup, který umožňuje odhadnout posteriorní pravděpodobnosti bez toho, aniž by byly zkoumány všechny možné kombinace parametrů studovaných modelů, protože to by bylo výpočetně nemožné.

Jak si ukážeme v následujících částech bakalářské práce, existuje několik softwarových balíčků, které využívají ABC. V textu představím IMa2 *(„Isolation with migration“ metoda, Hey a Neilson, 2004*), DIY-ABC (*Do It Yourself Approximate Bayesian Computation, Cornuet 2015)*, ABCtoolbox *(Wegmann et al., 2009)* , G-PhoCS (*A Generalized Phylogenetic Coalescent Sampler, Gornau et al., 2011)*.

3.3. Haplotypové metody

Většina souhrnných statistik, které při vyvozování demografické historie využíváme, předpokládá, že jsou lokusy nezávislé a statistiky nezohledňují jejich propojení. Metody, které se chystám popisovat v této kapitole, pracují se vzorky haplotypů. Haplotypem rozumíme kombinaci lokusů odkazujících na různá místa sekvence DNA, která jsou potomkům předávána preferenčně pohromadě *(viz kap. 1.2. Genealogie a teorie Koalescence) (Harris et al.,2013).* Podobnost mezi haplotypy může být způsobena identitou podle stavu *(Identity by state, IBS)* nebo identitou podle původu *(Identity by descent, IBD) (Beichmann et al., 2018) (Obr. 8)*.



Obr. 7: IBD & IBS

V levé části obrázku je zobrazena Identita podle původu (IBD) a v pravé části obrázku Identita podle stavu (IBS). Segmenty IBD jsou zobrazeny v případě nevlastního sourozence. IBS nemusí nutně vést k MRCA a mohou ji zdědit libovolní jedinci. Žlutá a zelená barva znázorňují úseky děděné od předků přerušené rekombinací v průběhu času (Maeva Leitwein et al. 2019, převzato a upraveno).

IBS je termín, který populační genetici používají k pojmenování segmentů, které jsou identické strukturou, ale ne z důvodu sdílení společného předka *(Obr. 7)*. Tyto metody nás velice spolehlivě informují o demografické historii, právě z důvodu, že identita mezi segmenty není způsobena sdíleným původem. Metody založené na IBS fungují na fázovaných datech. Fázovanými daty rozumíme taková genomová data, kdy jsou rozděleny mateřsky a otcovsky děděné kopie každého chromozomu do haplotypů, aby se získal úplný obraz genetické variability. Hlavní výhodou této metody je, že se její aplikace dá přizpůsobit i velice složitým modelům *(Browning et al., 2011)*. Aplikace těchto metod se osvědčila například v projektu 1000 genomů, kdy potvrdila rozsáhlý genový tok mezi Afrikou a Evropou (*Harris a Nielsen 2013).*

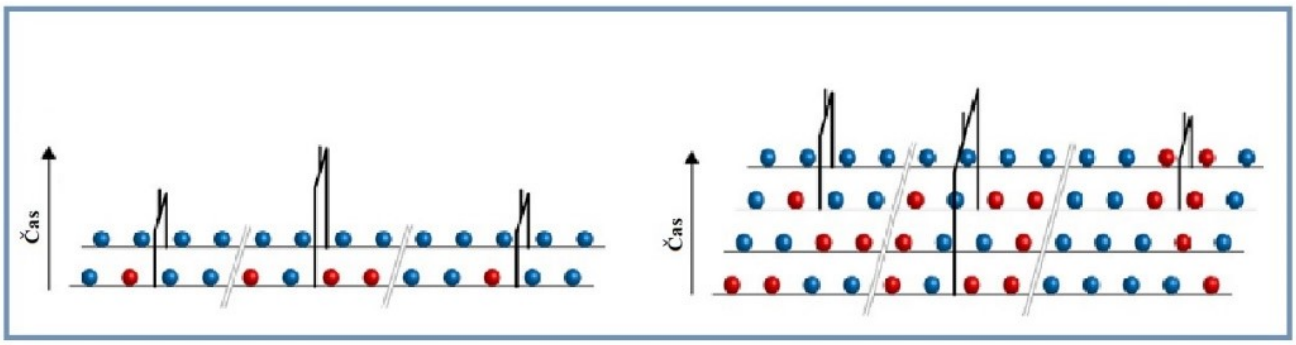
Jev nazveme IBD, pokud jsou dané nukleotidové sekvence děděny od společného předka. IBD však nemusí vždy znamenat IBS právě z důvodu, že v daném úseku DNA mohlo dojít ke tvorbě nových mutací, takže sekvence již nemusí být dále zcela identické (Obr. 7). IBD právě z důvodu vzniku mutací s sebou nese více kroků při vyvozování demografické historie. Pokud by ale předmětem našeho experiment bylo odhalování MRCA, IBD metody jsou jednou z nejlepších možností, protože hlavním vodítkem bude právě množství nově vniklých mutací. *(Beichman et al., 2018)*. Gusev a jeho spolupracovníci svojí studií potvrdili, že starší změny ve velikosti populace mají vliv na sdílení krátkých segmentů IBD mezi jednotlivci, zatímco novější změny ve velikosti populace ovlivňují sdílení dlouhých segmentů IBD *(Gusev et al. 2012)*.

Tyto metody jsou obecně spolehlivé právě v případě, pokud máme k dispozici kvalitní data a můžeme je spolehlivě rozfázovat a zarovnat. Tyto metody zvládají složité scénáře, osvědčili se při odhadu načasování příměsí a divergence (*Harris a Nielsen 2013).* Příkladem softwarového balíčku, který využívá tyto metody je DoRIS *(Palamara et al., 2012)* a budu se mu dále věnovat v kapitole o softwarových balíčcích *(<u>viz kap. 4.8. DoRIS</u>)*.

3.4. Sekvenční Markovovy koalescenční metody

Každý genom obsahuje velké množství lokusů, jejichž alely se při rekombinaci mohou přeorganizovat. Díky tomu mají různé lokusy odlišnou evoluční historii. Párová sekvenčně Markovská koalescence (*the pairwise sequentially Markovian coalescent,* PSMC, *Li et al., 2011*) a vícenásobná sekvenčně Markovská koalescence (t*he multiple sequentially Markovian coalescent,* MSMC, *Schiffels et al., 2014*) využívají tyto informace k rekonstrukci efektivní velikosti populace (Ne) při přijmutí určitých předpokladů o mutační rychlosti.

Tyto přístupy pracují s celogenomovými sekvencemi. Metodu PSMC lze použít k analýze nefázovaných sekvenčních dat z jednoho diploidního jedince, zatímco metoda MSMC používá sekvence z více jedinců *(Obr. 8) (Mather et al., 2019)*. Výsledné simulace těchto přístupů jsou generovány skrytým Markovovým modelem.



Obr. 8: PSMC versus MSMC

Na obrázku vidíme rozdíly v přístupech PSMC a MSMC. Pomocí modrých a červených kuliček rozeznáváme homozygotní a heterozygotní místa v genomu. Dvojité šedé čáry označují rekombinační breakpointy, které oddělují lokusy v genomu. Čas do MRCA dvou alel v každém lokusu je vyobrazen v lokálním stromě. V levé části obrázku u PSMC vidíme pouze dva haplotypy. Topologie lokálního stromu je tedy pevná, ale čas do MRCA mezi lokusy se liší. Jak je vidět v pravé části obrázku, u MSMC existuje více haplotypů. MSMC ignoruje topologie lokálních stromů a zaměřuje se na nejnovější koalescenční události v každém lokusu. Analýza více genomů je logicky výpočetně náročnější, ale MSMC zjednodušuje tento úkol pomocí vytváření podmnožiny lokálního stromu, který popisuje čas do MRCA dvou alel (obdobně jako u PSMC), které se na daném lokusu spojí jako první (Mather et al., 2019, převzato a upraveno).

Pojďme si nyní přiblížit matematické jádro těchto metod. Skrytý Markovův model (*Hidden Markov model,* HMM) je dvojice stochastických procesů X_t a Y_t, kde X_t je "skrytý proces", který nelze přímo pozorovat, a Y_t je pozorovatelný proces. V každém okamžiku t nabývá X_t jednoho z N možných stavů podle určitého rozdělení pravděpodobnosti. Protože X_t je Markovův proces, stav, který nabývá, závisí pouze na stavu v X_{t-1}. Poté, co X_t přejde do nového stavu, je hodnota Y_t generována pravděpodobnostním rozdělením, které závisí na hodnotě, kterou X_t v daném okamžiku nabývá. Hodnoty, kterých může Y_t nabývat, se obvykle označují jako "symboly pozorování" procesu.

Abychom mohli vytvořit skrytý Markovův model, musíme definovat klíčové složky procesu, které jsme popsali výše:
Možné stavy procesu X_t označíme q_i, kde $i \in \{1, \dots, N\}$,
možné symboly pozorování procesu Y_t označíme v_j kde $j \in \{1, \dots, M\}$.
Dále pravděpodobnostní rozdělení pro libovolné $k, l \in \{1, \dots, N\}$ "pravděpodobnosti přechodu", které popisuje, jak se pohybujeme mezi stavy X_t:

$$P(X_{t+1} = q_k | X_t = q_l),$$

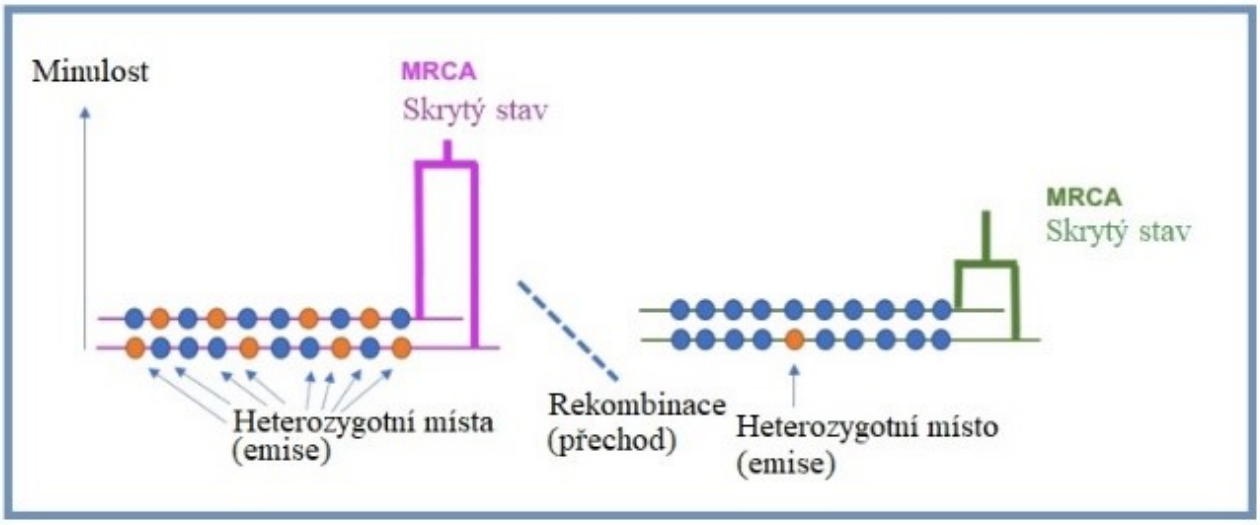
soubor pravděpodobnostních rozdělení pro libovolné $n \in \{1, \dots, N\}$ a $m \in \{1, \dots, M\}$ nazývaných "emisní pravděpodobnosti", který popisuje, jak stavy X_t generují hodnoty Y_t. Každé z nich bude mít tvar následující:

$$P(Y_t = v_m | X_t = q_n).$$

Pravděpodobnostní rozdělení popisující, jak by systém vypadal, když $t = 0$:

$$P(q_i | t = 0).$$

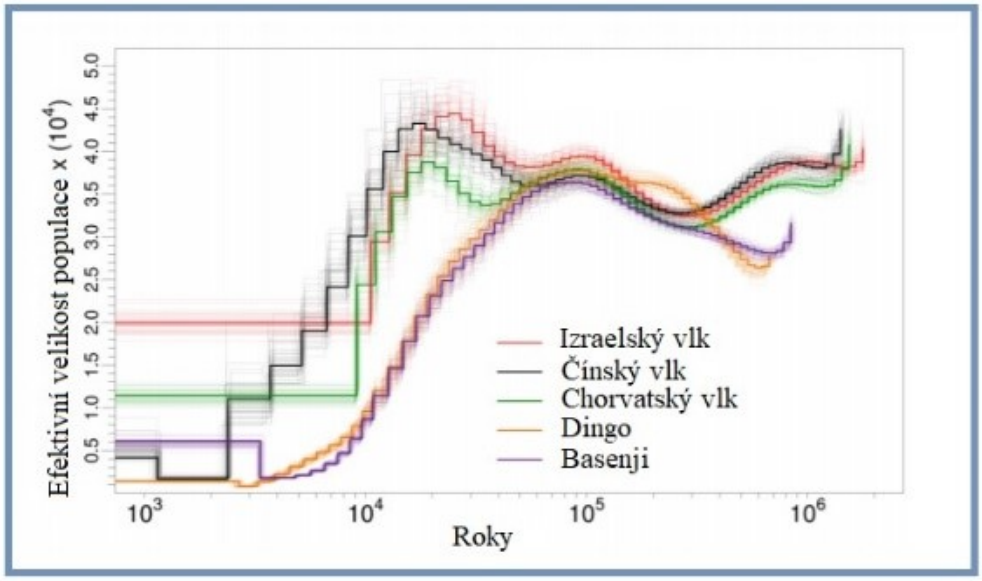
V případě sekvenčně Markovských koalescenčních modelů t indexuje jednotlivé lokusy. Skryté stavy jsou charakterizovány lokálními genealogiemi v lokusu *(Obr. 9)*. V případě PSMC jsou možnými stavy možné časy koalescence dvou alel. Pro MSMC je to čas koalescence dvou alel ve vzorku, které vytváří koalescence jako první. Symboly pozorování jsou vlastnosti genetických dat. Pro MSMC je zde několik dalších symbolů pozorování, aby se zohlednila složitost, kterou přináší více genomů. Pravděpodobnosti emise jsou určeny mírou mutace a pravděpodobnosti přechodu mírou rekombinace *(Obr. 9) (Mather et al., 2019).*



Obr. 9: Princip HMM pro vyvozování

Na tomto obrázku je zobrazen princip skrytého Markovova modelu (HMM), který je jádrem metod MSMC a PSMC. Homozygotní místa v genomu jsou zaznamenána dvěma modrými kuličkami, heterozygotní místa jsou zaznamenána kombinací modré a oranžové kuličky. Lokální genealogie s MRCA jsou naznačeny fialovou a zelenou barvou. Zobrazené oblasti genomu se liší rekombinací, takže každý lokus má jiného MRCA. Lokus se starším MRCA (růžový) má více pozorovaných heterozygotů, jelikož je zaznamenána delší doba, ve které se mutace hromadily. (Beichmann et al., 2018, převzato a upraveno)

Obě metody jsou užitečné při studiu hlubších populačních časových harmonogramů, především pokud máme data z omezeného počtu jedinců. Mazet a kolektiv ve své studii z roku 2015 také ukázali, že výsledky poukazující na změnu velikosti populace mají analogický scénář jako strukturované populace se změnami v migraci. *(Mazet et al., 2015).* Pokud se podíváme na příklady reálných analýz, PSMC se například ve velké míře využívalo při studiích starověkého koně a starověkého vlka *(Skoglund et al., 2015).* Kromě rekonstrukce demografické historie jsou PSMC a MSMC často používány k odvození načasování divergence populace ze starých genomů. Určité studie však prokazují, že závěry těchto dvou metod mohou být citlivé na porušení základních předpokladů demografie *(Mazet et al., 2016)*.



Obr. 10: Historické změny efektivní velikosti populace u různých Canidae druhů

Na obrázku je zobrazen příklad publikovaných trajektorií PSMC, které ukazují změny efektivní velikosti populace (N_e) v čase pro různé Canidae druhy (Freedman et al., 2014). Osa x označuje minulé roky a osa y efektivní velikost populace. Tmavší čáry ukazují původní trajektorie, rozmazané čáry po stranách jsou výsledkem bootstrap analýzy. (Freedman et al., 2014, převzato a upraveno).

4. Softwarové balíčky a jejich aplikace

V této kapitole čtenáře seznámím s devíti balíčky, které se podle dostupné literatury aktuálně nejvíce používají, mají lehce odlišné přístupy a každá z metod má určité výhody a nevýhody a hodí na různé modely.

4.1. ‚Isolation with migration', neboli IMa metody

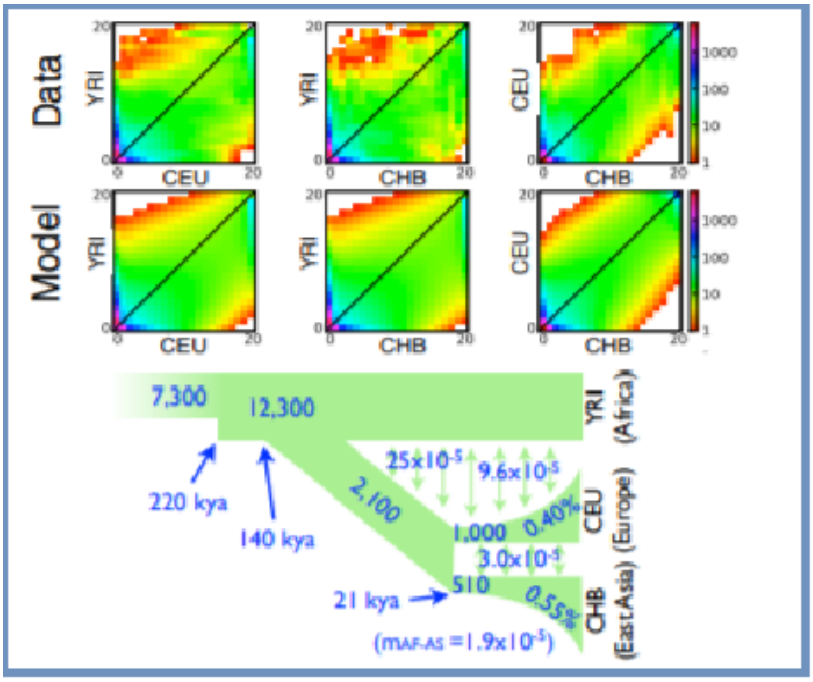
Jako první si představíme softwarové balíčky, který pracují s nerekombinujícími úseky DNA (mitochondriální DNA, chromozom Y, malé autozomální oblasti). Tyto přístupy obecně nazýváme ‚Isolation with migration' metody a v dnešní době se využívají různé varianty těchto balíčků. Aktuálně se z dostupné literatury nejvíce pracuje v programu IMa2. V průběhu analýzy jsou vytvářeny koalescentní simulace pro které je vypočítávána pravděpodobnost výskytu. V těchto balíčcích je použita Bayesovská inference založená na MCMC procházení, což, jak jsem již uváděla v předchozích kapitolách práce, může být časově velice náročné. Na druhou stranu tento přístup zvládá pracovat s libovolným počtem populací. Pomocí těchto balíčků je především odhadována míra migrace, velikost populace a čas divergence. Evans s kolegy použili program IMa2 k odhadnutí populačního modelu P. angustifolia *(Evans et al., 2015)*.

4.2. Fastsimcoal2

Fastsimcoal2 je software, který používá koalescenční simulace ke generování predikované SFS pro demografické parametry. Tento balíček dokáže pracovat s libovolným počtem populací. Metoda používá multidimenzionální SFS (v podstatě dvojdimensionální histogram) k odvození míry migrace mezi populacemi *(online zdroj č.1)*. Program je časově náročný pro velké počty populací, navíc k získání spolehlivých odhadů pravděpodobností musí být spuštěno mnoho replikačních simulací. Na druhou stranu Fastsimcoal2 zvládá komplexní evoluční scénáře zahrnující libovolné migrační matice mezi populacemi, různé historické události zapříčiňující změny ve velikosti populace. *(Excoffier et al., 2013).*

4.3. Implementace ∂a∂i

Implementace ∂a∂i (*Diffusion Approximation for Demografic Infference*) odhaduje parametry složitých populačních modelů za použití SFS. Implementace je schopná analyzovat jednotlivce z více populací a umí modelovat až tři interagující populace. Vývojáři použili ∂a∂i na lidská data z Afriky, Evropy a východní Asie a vybudovali nejsložitější, statisticky dobře charakterizovaný model migrace lidí z Afriky *(Gutenkunst et al., 2009)*. Důležité je, že ∂a∂iho rychlost umožňuje rozsáhlý bootstrapping pro statistickou charakterizaci modelu, včetně odhadu pro nejistotu parametrůa právě díky tomu se podařilo tento složitý model sestavit. Tato metoda byla dále aplikována také na modely populací Pongo abelii a Pongo pygmaeus, kde potvrdila, že strukturní evoluce probíhala mnohem pomaleji než u ostatních lidoopů (*Locke et al., 2011).*



Obr. 11: Implementace ∂a∂i

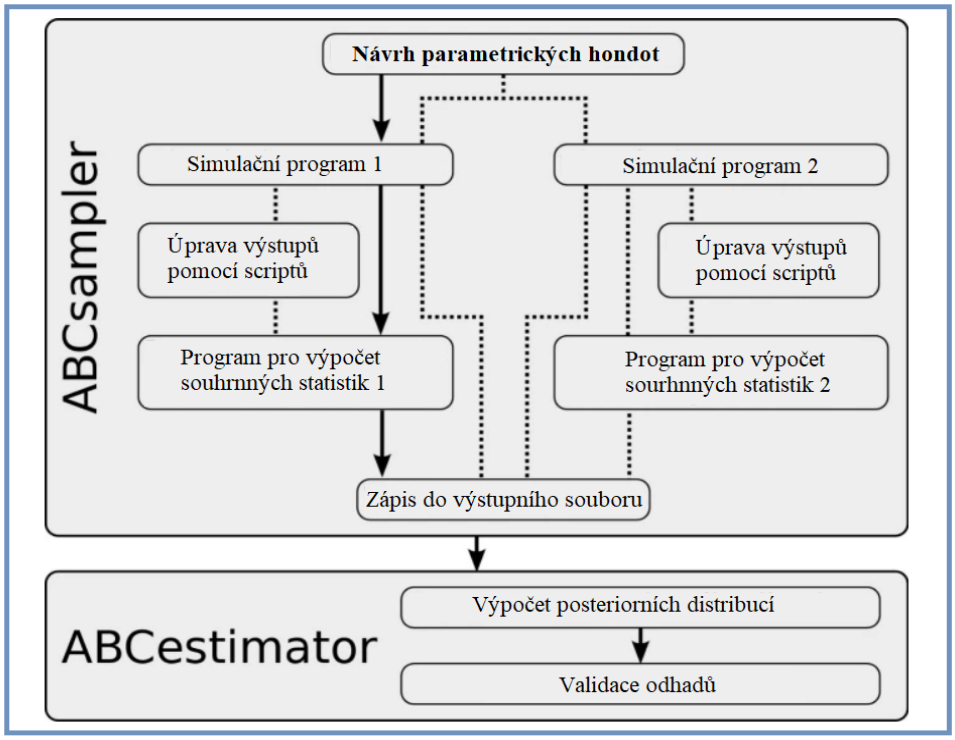
Na obrázku je zobrazen genetický model expanze lidské populace z Afriky. Pomocí ∂a∂i bylo 14 volných parametrů odhadnuto z 5Mb nekódující sekvence. Nejistota u parametrů je obvykle kolem 20 %. Mezi parametry autor uvádí například efektivní velikost populace (Ne), míru migrace sekvence pro danou generaci, nebo rychlost růstu populace (Gutenkunst et al., 2009, převzato a upraveno).

4.4. DIY-ABC

DIY-ABC (*Do It Yourself Approximate Bayesian Computation*) je softwarový balíček pro komplexní analýzu historie populací za použití aproximačního Bayesovského výpočtu na DNA datech. Program DIYABC má modulární podobu. Aktuální verze programu je sestavena ze čtyř modulů. První modul provádí koalescence v izolované populaci konstantní velikosti mezi dvěma danými časy. Druhý modul sdružuje genové linie ze dvou populací (divergence). Třetí modul dělá opak druhého modulu a rozděluje genové linie ze smíšené populace mezi dvě rodičovské populace. Čtvrtý modul byl přidán až v poslední verzi programu a provádí přidání vzorku genu do populace v dané generaci. Tento modul byl přidán proto, aby umožnil přesnější klasifikaci z více vzorků jedné populace, které byly odebrány v různých generacích populace. Kombinací výše uvedených čtyř modulů je program schopný simulovat genetická data zahrnující libovolný počet populací podle scénáře, který následně zohledňuje divergenci, genetické příměsi a změny velikosti populace. Navíc, díky poslednímu modulu, může být populace vzorkována více než jednou a v různých časech *(Cornuet et al., 2008)*. V roce 2014 představili Cornuet a jeho tým vylepšení programu (DIY-ABC 2.0) , které umožňuje zpracovávat soubory dat o více populacích s velkým počtem lokusů (např. několik tisíc až deset tisíc lokusů během několika hodin až několika dní) *(Cornuet et al., 2014*).

4.5. ABCtoolbox

ABCtoolbox byl navržen tak, aby prováděl ABC odhady za použití algoritmů zahrnující MCMC. ABC odhad je zde proveden ve dvou stejných paralelních krocích. Nejprve je proveden paralelně krok simulační, ve kterém vzniká velký počet simulací, které jsou následně použité k odhadu posteriorní distribuce. Balíček obsahuje dva hlavní programy *(Wegmann et al.,2010)*. První z nich je ABCsampler *(Obr. 12)*, který generuje simulace a počítá souhrnné statistiky pomocí vedlejších pomocných programů. Druhým programem je ABCestimator *(Obr. 12)*, který počítá marginální posteriorní rozdělení parametrů ze zaznamenaných simulací.



Obr. 12: ABCtoolbox

Diagram popisující jednotlivé kroky odhadu ABC pomocí ABCtoolboxu. Černé šipky označují standardní postup. Některé alternativní cesty jsou znázorněny tečkovanými čarami. Například je možné upravit výstup simulačního programu tak, aby bylo možné zohlednit specifické vlastnosti pozorovaných údajů (chybějící údaje apod.). Kromě toho může ABCtoolbox v jedné iteraci volat několik simulačních programů, z nichž každý může být spuštěn se stejnými hodnotami parametrů. V jedné analýze tak lze pohodlně kombinovat různé typy dat (Wegmann, 2010, převzato a upraveno).

Interakce programu ABCsampler s externími programy probíhá prostřednictvím příkazového řádku, což umožňuje používat většinu z mnoha dostupných programů pro simulaci genetických dat. Program ABCsampler také nabízí možnost volat libovolný skript nebo program pro úpravu výstupu simulačního programu *(Obr. 12)*. Program ABCestimator přímo čte výstup programu ABCsampler a počítá posteriorní rozdělení na základě simulací, které jsou nejblíže pozorovaným datům. *(Wegmann et al., 2010).*

4.6. SMC ++

SMC ++ je poměrně nový přístup, který je však aktuálně hojně využíván, protože jak se dozvíme za chvíli, jeho koncept přináší mnoho výhod. SMC++ totiž kombinuje výpočetní efektivitu SFS a využívá informací o vazební nerovnováze v sekvenčních Markovových koalescenčních metodách. Tento přístup je navržen tak, aby využíval moderní datové soubory sestávající ze stovek nefázovaných celých genomů. Nefázovanými genomy rozumíme genotypy bez ohledu na to, který z páru chromozomů nese danou alelu. Přístup dokáže také analyzovat dvojice divergentních populací, což mu umožňuje sdružovat informace z obou populací a také přímo odhadovat dobu divergence. Nejspíše se jedná o první metodu demografické inference, která je schopna analyzovat nefázovaná data celých genomů velkého počtu jedinců výpočetně efektivním a stabilním způsobem, přičemž bere v úvahu informace o genových vazbách *(Terhorst et al., 2017).*

4.7. Lamarc

Dalším softwarovým balíčkem, pomocí kterého můžeme vyvozovat demografickou historii populací, je Lamarc. Dokáže odvodit efektivní velikost populace, míru růstu populace, nebo míru migrace populací. Aproximuje součet všech možných genealogií, které by mohly vysvětlit pozorovaný vzorek. Dle autora dokáže program pracovat s různými typy dat (celé sekvence, SNP data, mikrosatelitní data). Tento balíček taktéž používá MCMC. První verze tohoto programu vznikla již v roce 2001 *(online zdroj č.2)*. Aktuální verze opravují několik nedostatků verzí předchozích, zejména chyby v maximalizaci a chyby při zpracovávání vícelokusových dat *(Kuhner et al., 2006).* Poslední aktualizace programu proběhla v roce 2018, kde však sám autor uvádí, že došlo k opravám, které usnadňují kompilaci programu. LAMARC (a jeho sesterský program Migrate) je nástupce starších programů Coalesce, Fluctuate a Recombine, které už aktuálně nejsou podporovány. *(online zdroj č.2).*

4.8. DoRIS

DoRIS je softwarový balíček, který dokáže rekonstruovat demografické události, ke kterým došlo ve velmi nedávné minulosti (100 generací) u jedné nebo více populací *(online zdroj č.3)*. V kapitole a haplotypových metodách *(viz kap. 3.3. Haplotypové metody)* jsme si představili pojem IBD, pomocí kterého tento program vyvozuje. DoRIS tedy odvozuje parametry několika možných demografických modelů pomocí hustoty IBD segmentů různých délek. Tento typ analýzy může odhalit vztahy mezi skupinami v rámci malých geografických regionů, například v rámci jedné země *(Francioli et al., 2014, cit. dle online zdroje č.3), nebo* jemné výkyvy velikosti populace v nedávné minulosti, například efekty hrdla láhve v posledních 100 generacích *(Palamara et al. 2012, cit. dle online zdroje č.3).*

4.9. G-PhoCS

Posledním programem, který v této práci zmíním je generalizovaný fylogenetický koalescenční vzorkovač (*A Generalized Phylogenetic Coalescent Sampler,* G-PhoCS). **G-PhoCS** používá pro demografickou inferenci vzorky odebrané z několika jedinců populace *(Gronau et al., 2011).* Výhodou tohoto programu je, že dokáže pracovat s redukovanými genomovými daty i s celogenomovými daty *(viz kap. 2. Data)*. K odvozování pomocí G-PhoCS je zapotřebí sekvencí od několika jedinců. Tento balíček dokáže odvodit velikost ancestrální populace, čas divergence, nebo míru migrace *(online zdroj č. 4).* Program se spouští na několika nezávislých fragmentech a poté se jednotlivé genealogie generují podle konkrétních demografických modelů a následně se vypočítá pravděpodobnost pro daná sekvenční data pro každý fragment. G-PhoCS nevzorkuje genealogie náhodně, ale místo toho používá vzorkování Metropolis – Hastings *(Gronau et al., 2011).* Tento algoritmus je metodou typu MCMC, kterou jsem popisovala v kapitolách výše, tedy algoritmus preferenčně vzorkuje genealogie, které budou pravděpodobně kompatibilní se sledovanými daty, což zvyšuje účinnost inference. G-PhoCS využívá ABC a poskytuje posteriorní distribuci požadovaných demografických parametrů. Hlavní výhodou tohoto programu je, že zvládá složité vícepopulační modely, ale na druhou stranu je díky tomu výpočetně náročný a často vyžaduje až týdny času pro samotný výpočet. Další velkou nevýhodou je, že je méně schopný odvodit nedávné změny velikosti populace ve srovnání s jinými přístupy. (*Beichman et al., 2018).*

5. Srovnání softwarových balíčků

V této kapitole bych ráda shrnula a porovnala jednotlivé balíčky, pomocí kterých můžeme vyvozovat demografickou historii populací *(Tab. 1)*. V předchozí kapitole jsem podrobněji představila devět softwarových balíčků, které se aktuálně hojně využívají pro vyvozování demografické historie populací. Tyto balíčky zároveň pracují s různými druhy dat, využívají různé inferenční metody, každý software se hodí pro jiné demografické modely, umí pracovat s různým množstvím populací a každý přístup nese určité výhody a nevýhody. Všechny tyto parametry softwarových balíčků jsem shrnula do tabulky, aby si čtenář mohl přehledněji uvědomit a rozmyslet, který ze softwarů se hodí pro danou reálnou analýzu.

Jméno balíčku	**Data**	**Využívané metody**	**Vhodné pro modely**	**Pro kolik populací**	**Výhody**	**Nevýhody**	**Zdroje**
IMa2	SNP, STR	ABC, MCMC	míra migrace, čas divergence, velikost populací	více populací	nejpřesnější pro nerekombinující oblasti	nedá se aplikovat na všechny modely	*(Evans et al., 2015; Hey, 2011)*
Fastsimcoal2	SNP, STR	MSMC, SFS	změna velikosti populací, štěpení populací, změna migrační matice	více populací	zvládá složité evoluční scénáře, různé migrační matice mezi populacemi	časově náročné pro velké počty populací	*(Excoffier et al., 2013; online zdroj č.1)*
Implemen-tace ∂a∂i	SNP	SFS	míra migrace pro danou generaci, efektivní velikost populace, rychlost růstu populace	≤ 3 populace	výpočetní náklady nezávislé na počtu SNP, ale exponenciální v počtu populací, obecně rychlé	vyžaduje znalosti kódování v jazyce Python, maximálně pro tři populace	*(Gutenkunst et al. 2010, Locke et al., 2011)*
DIY-ABC	SNP	ABC	divergence, genetické příměsi, změny ve velikosti populace	více populací	populace může být vzorkována více než jednou a v různých časech	Jelikož je program hodně flexibilní k modelům a datům, někdy dlouho běží	*(Cornuet et al., 2008; Cornuet et al., 2014)*
ABCtoolbox	SNP, STR, různá ploidie (doprovodné programy)	ABC,MCMC	velice flexibilní, různé aplikace	více populací	dobře interaguje s doprovodnými programy	vyhodnocování trvá dlouho	*(Wegmann et al., 2010; Wegmann et al., 2009)*
SMC ++	nefázované celé genomy	SFS, MSMC	velikost populace v historii, divergence	více populací	malé nároky na zpracování dat	vyhodnocování trvá dlouho, někdy méně přesné	*(Terhorst et al., 2017)*
Lamarc	krátké fázované haplotypy	MCMC	efektivní velikost populace, míra migrace, míra růstu populace	více populací	velmi flexibilní	vyžaduje vzorkování metodou MCMC	*(Kuhner at al., 2006; online zdroj č.2)*

DoRIS	délky úseků IBD mezi dvojicemi jedinců	IBD	velikost populace	více populací	citlivé na nedávné události	IBD musí být odvozena pomocí dalších programů, jen nedávné události	*(Palamara et al. 2012; online zdroj č.3)*
G-PhoCS	krátké, (ne)fázované haplotypy, data z RAD seq	ABC, MCMC	velikost populace, divergence, míra migrace (komplexní demografické scénáře)	více populací	odvozuje komplexní demografické scénáře	dlouho počítá, nevhodné pro nedávné události	*(Gronau et al., 2011; Beichmann et al. 2018, online zdroj č.4)*
PSMC	diploidní genotypy z jednoho jedince	PSMC	hlubší časové populační harmonogramy	jedna populace	lze použít k analýze nefázovaných sekvenčních dat z jednoho diploidního jedince	někdy velice nepřesné	*(Mather et al., 2019; Beichmann et al., 2018; Li et al., 2011)*
MSMC	celý genom, fázované haplotypy	MSMC	hlubší časové populační harmonogramy	více populací	stačí malý počet osekvenovaných jedinců dané populace	někdy velice nepřesné	*(Mazet et al., 2016; Beichmann et al., 2018; Schiffels et al., 2014*)

Tab. 1: Srovnání softwarových balíčků používaných pro demografické vyvozování historie populací.

6. Závěr

Metody, které se při vyvozování demografické historie populací používají, podléhají rychlému vývoji. Vývoj vyvozování úzce souvisí s vývojem počítačových softwarů a neustále rostoucí výpočetní kapacitou. Dalším důležitým faktorem je skutečnost, že v posledních letech dochází k velkému rozvoji sekvenačních technik, díky čemuž můžeme získat data z velkého množství lokusů i jedinců, což má kladný vliv na správnost výsledků při vyvozování demografií.

Během své bakalářské práce jsem několikrát zmínila, že daný software/metoda může opravdu hodně chybovat a nemusí přinášet směrodatné výsledky. Pokud se tedy chystáme vyvozovat demografickou historii populací, měli bychom zvážit, jestli je námi navržený design experimentu vůbec vhodný pro vyvození daného demografického modelu. Dále bychom měli věnovat pozornost tomu, do jaké míry porušují genomická data předpokladaný demografický model a závěr analýzy zvážit se všemi úskalími, které vyvozování přináší.

Právě z důvodu ověření výsledků se hodí vyvodit daný demografický model pomocí více metod, nebo zkoumat shodu demografického modelu s více souhrnnými statistikami (*Beichman et al., 2018*).

7. Seznam použité literatury

BEAUMONT, M. A., (2010). Approximate Bayesian Computation in Evolution and Ecology. *Annual Review of Ecology, Evolution, and Systematics* . **41**(1), 379-406. doi: 10.1146/annurev-ecolsys-102209-144621.

BEAUMONT, M. A., ZHANG, W., & BALDING, D. J. (2002). Approximate Bayesian computation in population genetics. *Genetics*, **162**(4), 2025-2035. doi: 10.1093/genetics/162.4.2025.

BEICHMAN, A. C., SANCHEZ, E. H. a LOHMUELLER, K. E. (2018). Using Genomic Data to Infer Historic Population Dynamics of Nonmodel Organisms. *Annual Reviews*, **49**, 433-456. doi: 10.1146/annurev-ecolsys-110617-062431.

BOX, G. E. P. (1976). Parsimony. *Journal of the American Statistical Association,* **71**(356) 791-799. doi: 10.2307/2286841.

BROWNING, S. R., & BROWNING, B. L. (2011). Haplotype phasing: existing methods and new developments. *Nature Reviews Genetics*, **12**(10), 703-714. doi: 10.1038/nrg3054.

BRZEZINSKI, M., MICHELOTTI, A. G., SCHWINN, A. D. (2006), Chapter 5 - Genomics and proteomics, *Foundations of Anesthesia 2ND edition*, 71-78. doi: 10.1016/B978-0-323-03707-5.50011-5.

CABRERA, A. & PALSBOLL, P. (2017). Inferring past demographic changes from contemporary genetic data: a simulation-based evaluation of the ABC methods implemented in DIYABC. *Molecular ecology resources*. **17**(6). doi: 10.1111/1755-0998.12696.

CORNUET, J. M. et al. (2008). Inferring population history with DIY ABC: a user-friendly approach to approximate Bayesian computation. *Bioinformatics*, **24**(23), 2713–2719. doi: 10.1093/bioinformatics/btn514.

CORNUET, J. M., PUDLO, P., VEYSSIER, J., DEHNE-GARCIA, A., GAUTIER, M., LEBLOIS, R., ... & ESTOUP, A. (2014). DIYABC v2. 0: a software to make approximate Bayesian computation inferences about population history using single nucleotide polymorphism, DNA sequence and microsatellite data. *Bioinformatics*, **30**(8), 1187-1189. doi: 10.1093/bioinformatics/btt763.

DAVEY, J.L. a BLAXTER, M.W. (2010). RADSeq: next-generation population genetics. *Briefings in Functional Genomics* **9**, 416-423. doi: 10.1093/bfgp/elq031.

ESTOUP, A., CORNUET, J. M., DEHNE-GARCIA, A., & LOIRE, E. (2015). New version of the computer program DIYABC (DIYABC v2. 1.0): a user-friendly approach to approximate Bayesian computation for inference on population history using molecular markers.

EVANS, L., ALLAN, G., DIFAZIO, S., SLAVOV, G., WILDER, J., FLOATE, K.D., ROOD, S. & WHITNAM, T. (2015). Evans et al. 2015 Heredity.

EXCOFFIER, L., MARCHI, N., MARGUER, D. A., MATTHEY-DORET, R., GOUY, A., & SOUSA, V. C. (2021). fastsimcoal2: demographic inference under complex evolutionary scenarios. *Bioinformatics*, **37**(24), 4882-4885. doi: 10.1093/bioinformatics/btab468.

FRANCIOLI, L. C., MENELAOU, A., PULIT, S. L., VAN DIJK, F., PALAMARA, P. F., ELBERS, C. C., ... & Lifelines Cohort Study. (2014). Whole-genome sequence variation, population structure and demographic history of the Dutch population. *Nature genetics*, **46**(8), 818-825. doi:10.1038/ng.3021.

FREEDMAN, A.H., GRONAU, I., SCHWEIZER, R.M. et al. (2014). Genome Sequencing Highlights the Dynamic Early History of Dogs. *PLOS Genetics* **10**(1). doi: 10.1371/journal.pgen.1004016.

GARRIGAN, D., HEDRICK, P. W. a LEE, R. N. (2002). Major histocompatibility complex variation in red wolves: evidence for common ancestry with coyotes and balancing selection. *Molecular ecology*, **11**(10), 1905-1913. doi: 10.1046/j.1365-294X.2002.01579.x.

GRONAU, I., HUBISZ, M.J., GULKO, B., DANKO, C.G., SIEPEL, A. (2011). Bayesian inference of ancient human demography from individual genome sequences. *Nature Genetics.* **43**, 1031–1034.

GROVER, C. E., SALMON, A., & WENDEL, J. F. (2012). Targeted sequence capture as a powerful tool for evolutionary analysis1. *American journal of botany*, **99**(2), 312-319. doi: 10.3732/ajb.1100323.

GUILLARD, A. (1855). Elements De Statistique Humaine: Ou Demographie Comparee.

GUTENKUNST, R. N., HERNANDEZ, R. D., WILLIAMSON, S. H. a BUSTAMANTE, C. D. (2010). Diffusion Approximations for Demographic Inference: DaDi. *Theoretical Population Biology*. doi: 10.1038/npre.2010.4594.1.

GUTENKUNST, R. N., HERNANDEZ, R. D., WILLIAMSON, S. H. a BUSTAMANTE, C. D. (2009). Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoS genetics*, **5**(10), e1000695. doi: 10.1371/journal.pgen.1000695.

HARRIS, K., & NIELSEN, R. (2013). Inferring demographic history from a spectrum of shared haplotype lengths. *PLoS genetics*, **9**(6), e1003521. doi: 10.1371/journal.pgen.1003521.

HEY, J., & NIELSEN, R. (2004). Multilocus methods for estimating population sizes, migration rates and divergence time, with applications to the divergence of Drosophila pseudoobscura and D. persimilis. *Genetics*, **167**(2), 747-760. doi: 10.1534/genetics.103.024182.
HIRSCHFELD, L. (1919). A New Germ of Paratyphoid. *The Lancet*, *193*(4982), 296-297.

KINGMAN, J. F. C. (1982). The coalescent. *Stochastic processes and their applications*, **13**(3), 235-248. doi: 10.1016/0304-4149(82)90011-4.

KUHNER, M. K. (2006). LAMARC 2.0: maximum likelihood and Bayesian estimation of population parameters. *Bioinformatics*, **22**(6), 768-770. doi: 10.1093/bioinformatics/btk051.

LEITWEIN, M., CAYUELA, H., FERCHAUD, A. L., NORMANDEAU, É., GAGNAIRE, P. A., & BERNATCHEZ, L. (2019). The role of recombination on genome-wide patterns of local ancestry exemplified by supplemented brook charr populations. *Molecular Ecology*, **28**(21), 755-4769. doi: 10.1111/mec.15256.

LI, H. a DURBIN, R. (2011). Inference of human population history from individual whole-genome sequences. *Nature*, **475**(7357), 493–96. doi: 10.1038/nature10231.

LOCKE, D. P., HILLIER, L. W., WARREN, W. C., WORLEY, K. C., NAZARETH, L. V., MUZNY, D. M., ... & WILSON, R. K. (2011). Comparative and demographic analysis of orang-utan genomes. *Nature*, **469**(7331), 529-533. doi: 10.1038/nature09687.

LOHSE, K.A., BROOKS, P.D., MCINTOSH, J.C., MEIXNER, T., & HUXMAN T.E. (2009): Interactions between biogeochemistry and hydrologic systems. *Annual Review of Environment and Resources. 34, 65-96*. doi: 10.1146/annurev.environ.33.031207.111141.

MARTH, G.T., CZABRAKA, E., MURVAI, J., SHERRY, S.T. (2004) The allele frequency spectrum in genome-wide human variation data reveals signals of differential demographic history in three large world populations. *Genetics,* **166**(1), 351-72. doi: 10.1534/genetics.166.1.351.

MATHER, N., TRAVES, S. M. a HO, S. Y. M. (2019). A practical introduction to sequentially Markovian coalescent methods for estimating demographic history from genomic data. *Ecology and Evolution*, 579-589. doi: 10.1002/ece3.5888.

MAZET, O., RODRÍGUEZ, W., & ChIKHI, L. (2015). Demographic inference using genetic data from a single individual: Separating population size variation from population structure. *Theoretical Population Biology*, **104**, 46-58. doi: 10.1016/j.tpb.2015.06.003.

MAZET, O., RODRIGUEZ, W., GRUSEA, S. et al. (2016). On the importance of being structured: instantaneous coalescence rates and human evolution—lessons for ancestral population size inference. *Heredity*, **116**(4), 362–71. doi: 10.1038/hdy.2015.104.

NEI, M., & LI, W. H. (1979). Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proceedings of the National Academy of Sciences*, **76**(10), 5269-5273. doi: 10.1073/pnas.76.10.5269.

NIELSEN, R., KORNELIUSSEN, T., ALBRECHTSEN, A., LI, Y. & WANG, J. (2012). SNP Calling, Genotype Calling, and Sample Allele Frequency Estimation from New-Generation Sequencing Data. *PloS one*. 7. e37558. doi: 10.1371/journal.pone.0037558.

PALAMARA, P. F., LENCZ, T., DARVASI, A., & PE'ER, I. (2012). Length distributions of identity by descent reveal fine-scale demographic history. *The American Journal of Human Genetics*, **91**(5), 809-822. doi: 10.1016/j.ajhg.2012.08.030.

ROSEN, Z., BHASKAR, A., ROCH, S., & SONG, Y. S. (2018). Geometry of the Sample Frequency Spectrum and the Perils of Demographic Inference. *Genetics*, **210**(2), 665–682. doi: 10.1534/genetics.118.300733.

SHI, W., AYUB, Q., VERMEULEN, M., SHAO, R. G., ZUNIGA, S., VAN DER GAAG, K., ... & TYLER-SMITH, C. (2010). A worldwide survey of human male demographic history based on Y-SNP and Y-STR data from the HGDP–CEPH populations. *Molecular biology and evolution*, **27**(2), 385-393. doi: 10.1093/molbev/msp243.

SCHIFFELS, S., DURBIN, R. (2014). Inferring human population size and separation history from multiple genome sequences. *Natural Genetics*, **46**(8), 919–25. doi: 10.1038/ng.3015.

SKOGLUND, P., ERSMARK, E., PALKOPOULOU. E. et al. (2015). Ancient wolf genome reveals an early divergence of domestic dog ancestors and admixture into high-latitude breeds. *Current Biology*, **25**(11), 1515-9. doi: 10.1016/j.cub.2015.04.019.

SPENCE, J. P., KAMM, J. A., & SONG, Y. S. (2016). The Site Frequency Spectrum for General Coalescents. *Genetics*, **202**(4), 1549–1561. doi: 10.1534/genetics.115.184101.

STUMPF, M., MCVEAN, G. (2003). Estimating recombination rates from population-genetic data. *Nat Rev Genet* **4,** 959–968 doi: 10.1038/nrg1227.

TAJIMA, F. (1983) Evolutionary relationship of DNA sequences in finite populations. *Genetics*, **105**(2), Pages 437–460, doi: 10.1093/genetics/105.2.437.

TAVARÉ, S., BALDING, D. J., GRIFFITHS, R. C., & DONNELLY, P. (1997). Inferring coalescence times from DNA sequence data. *Genetics*, **145**(2), 505-518.

TERHORST, J., KAMM, J. A. a SONG, Y. S. (2017). Robust and scalable inference of population history from hundreds of unphased whole-genomes. *Nature Genetics*, 303–309. doi: 10.1038/ng.3748.

VAN DIJK, E. L., AUGER, H., JASZCZYSZYN, Y., & THERMES, C. (2014). Ten years of next-generation sequencing technology. *Trends in genetics*, **30**(9), 418-426. doi: 10.1016/j.tig.2014.07.001.

WEGMANN, D., LEUENBERGER, C., & EXCOFFIER, L. (2009). Efficient approximate Bayesian computation coupled with Markov chain Monte Carlo without likelihood. *Genetics*, **182**(4), 1207-1218. doi: 10.1534/genetics.109.102509.

WEGMANN, D., LEUENBERGER, C., & EXCOFFIER, L. (2009). Using abctoolbox. *BMC Bioinformatics*, **11**, 116.
WEGMANN, D., LEUENBERGER, C., NEUENSCHWANDER, S. et al. (2010). ABCtoolbox: a versatile toolkit for approximate Bayesian computations. *BMC Bioinform.,* **11**(116). doi: 10.1186/1471-2105-11-116.

ZHANG, L., SU, W., TAO, R., ZHANG, W., CHEN, J., WU, P., ... & KUANG, H. (2017). RNA sequencing provides insights into the evolution of lettuce and the regulation of flavonoid biosynthesis. *Nature communications*, *8*(1), 1-12. doi: 10.1038/s41467-017-02445-9.

8. Online zdroje

1. Gronau Ilan. (2011) G-PhoCS - A Generalized Phylogenetic Coalescent Sampler [online] cit. [7.8.2021]. Dostupné z: http://compgen.cshl.edu/GPhoCS/
2. Palamara Lab. (2021) DoRIS [online] cit. [18.12.2021] Dostupné z: https://palamaralab.github.io/software/doris/
3. Swiss institute of bioinformatics. (2021) Fastsimcoal2 [online] cit. [4.1.2022]. Dostupné z: http://cmpg.unibe.ch/software/fastsimcoal27/
4. University of Washington Terms of Use. (2018) LAMARC [online] cit.[20.12.2021]. Dostupné z: https://evolution.genetics.washington.edu/lamarc/index.html