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, 2021

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.

Praha 2021 Sára Simandlová

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, která se dají použít k vyvozování demografické historie populací. Následně diskutuji statistické metody, mezi které patří metody založené na datech z frekvenčního spektra míst, metody využívající aproximační Bayesovský výpočet, metody pro určování identity a sekvenční Markovovy koalescenční metody. Poskytuji základní přehled teorie a logiky každého přístupu. Poté uvádím postupy při výběru vyvozovacích metod.

Klíčová slova: populační genetika, demografická inference, statistická inference, celo-genomová 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 methods have been developed to infer the demographic history of populations from genomic data, which I describe in this bachelor thesis. At first, I introduce the reader to important concepts in analyzing 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. Next, I discuss statistical methods, which include methods based on site frequency spectrum data, methods using approximate Bayesian computation, methods for determining identity, and sequential Markov coalescent methods. I provide a basic overview of the theory and logic of each approach. I then present procedures for selecting inference methods.

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 1

1.2. Genealogie a teorie koalescence 3

2. Data 5

3. Metody pro vyvozování demografické historie populací 7

3.1. Metody založené na SFS 7

3.1.1. Softwarové balíčky a jejich použití 11

3.2. Metody využívající ABC 13

3.2.1. ABC softwarové balíčky 17

3.3. Metody pro určování identity 20

3.4. Sekvenční Markovovy koalescenční metody 22

3.5. Novější koalescenční metody 26

4. Postup při výběru vyvozovací metody 28

4.1. Selekce 29

5. Závěr 30

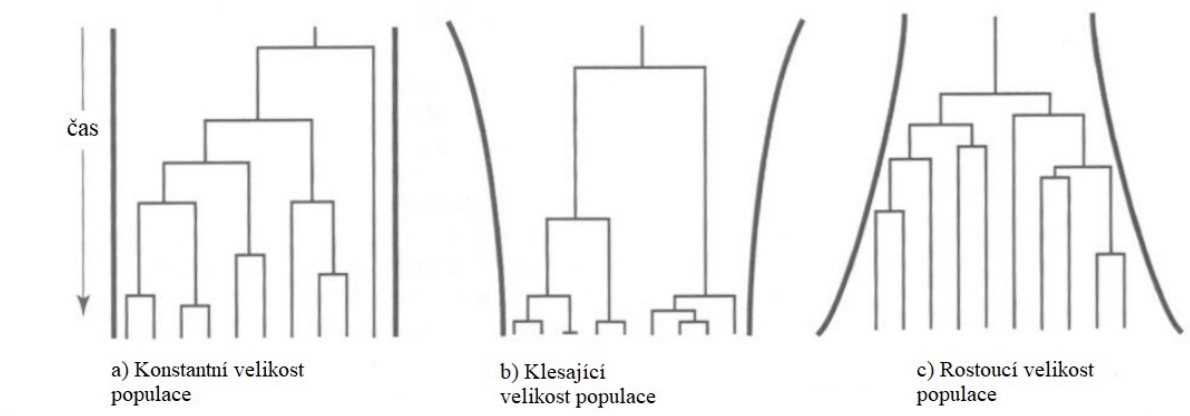
6. Seznam použité literatury 31

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. Analýza demografického procesu umožňuje zobecnit zákony o populačním vývoji, najít vzory nebo formulovat předpoklady o budoucím vývoji populace. Dnešní molekulární technologie nám dovolují ze sekvencí DNA vyčíst údaje o demografické historii populace, která bychom ani z historických pramenů často nezískali. 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. 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).

Jak jsem již zmínila, k vyvozování demografické historie používáme data ze sekvencí DNA neboli genomová data. Dají se využít celé sekvence genomů, sekvence transkriptomů, nebo data získaná z RAD-sekvenování.

Hlavní část této práce je věnovaná metodám, které k vyvozování demografické historie populací používáme. 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 čas divergence populací. V textu popisuji metody založené na frekvenčním spektru míst (*site frequency spectrum,* SFS), metody založené na aproximačním Bayesovském výpočtu (*approximate Bayesian computation,* ABC), metody pro určování identity a sekvenční Markovovy koalescenční metody.

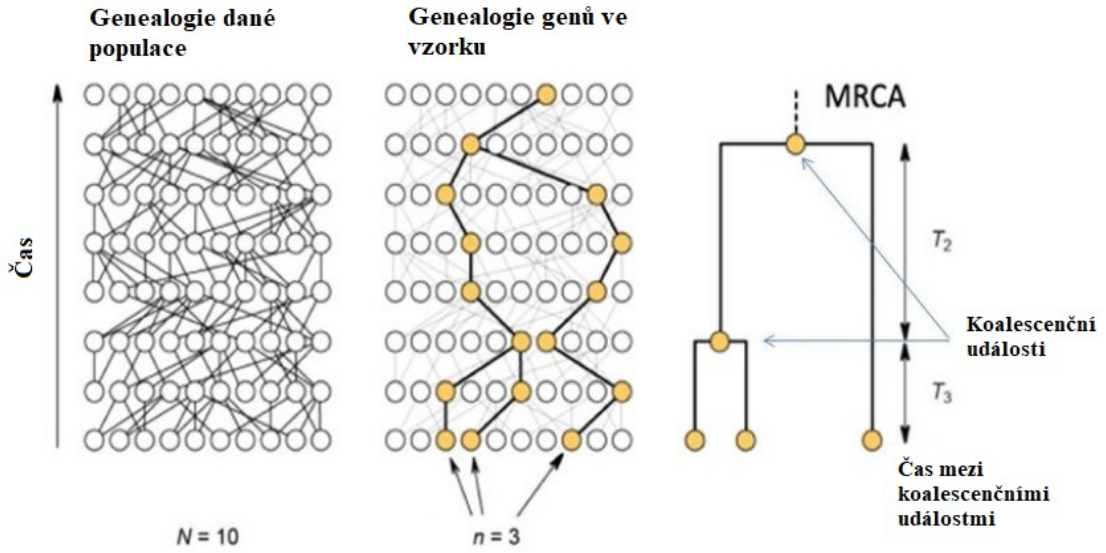
1.2. Genealogie a teorie koalescence

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 $P = 1/N$ pro haploidní organismy a $P = 1/(2N)$ pro diploidní organismy, kde N je počet jednotlivých kopií genů v populaci.

Důležitým 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. 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. Jedná se o chromozom 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 ležící daleko od sebe na stejném chromozomu (mimo chrY nebo mtDNA), nebo ležící na jiných chromozomech mají v podstatě nezávislé genealogie. Sekvenční data z jednoho lokusu poskytují jednu nezávislou realizaci evolučního procesu. Jelikož tato realizace je pouze jedna z nekonečného počtu možných genealogií pro základní demografii, závěry založené na jednom lokusu přináší značnou nejistotu a zvyšování počtu sekvenovaných jednotlivců tento problém nevyřeší. Další jedinci totiž budou součástí stále stejného rodokmenu. Jediný způsob, jak snížit nejistotu demografického modelu je studium genealogií více nezávislých genů. Vzorkování mnoha lokusů v celém genomu přináší řadu téměř nezávislých genealogií, které obsahují bohaté informace pro demografický závěr.



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, převzato a upraveno).

Je patrné, že koalescenční procesy vedou k vysoce variabilním genovým genealogiím, které lze využít k vyvozování demografické historie populace. V následujících částech bakalářské práce popisuji data a metody, které se za tím účelem využívají.

2. Data

Genom je kompletním souborem DNA organismu. Zahrnuje všechny geny a nekódující sekvence DNA. 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. 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.

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.

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). 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. 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. 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.

3. Metody pro vyvozování demografické historie populací

Jak jsem uvedla, velikost a struktura populace se v průběhu času mění, což má dopad na genetickou kompozici. Široká dostupnost molekulárních markerů a stále se zvyšující výkon počítačů podpořily vývoj sofistikovaných statistických metod, které se snaží co nejpřesněji vyvodit souhru genetické kompozice a demografických faktorů. Většina těchto technik je založena na konceptu pravděpodobnosti. V této kapitole se pokusím čtenáře seznámit s metodami, které se aktuálně nejvíce využívají.

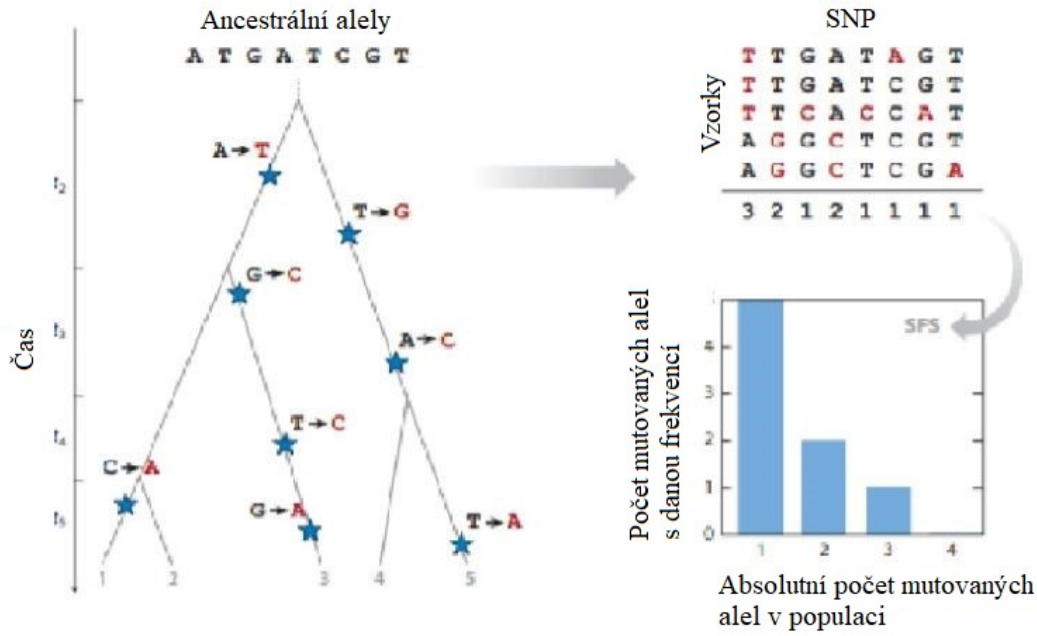
3.1. Metody založené na SFS

I když mají genové kopie společného předka, liší se mutacemi. V případě vyvozování demografické historie populací hovoříme 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 koalecence, tedy kratší čas k poslednímu společnému předku (MRCA).

Frekvenční spektrum míst (*site frequency specturum,* SFS), je běžný způsob, jak porozumět historickým událostem ovlivňující genetickou variabilitu. SFS můžeme jednoduše definovat jako distribuci frekvencí pro jednotlivé mutované alely *(Obr.3)*.



Obr.3: Frekvenční spektrum míst

Na obrázku je zobrazena situace, kde vyobrazený lokus nese osm SNP. V levé části vidíme, jak se nukleotidy v průběhu času měnily. V pravé části jsou zobrazené jednotlivé sekvence pod sebou a pod čarou je vypsán počet mutovaných alel v populaci. Graf znázorňuje závislost absolutního počtu mutovaných alel v populaci na počtu mutovaných alel s danou frekvencí. (Beichman et al. 2018, převzato a upraveno).

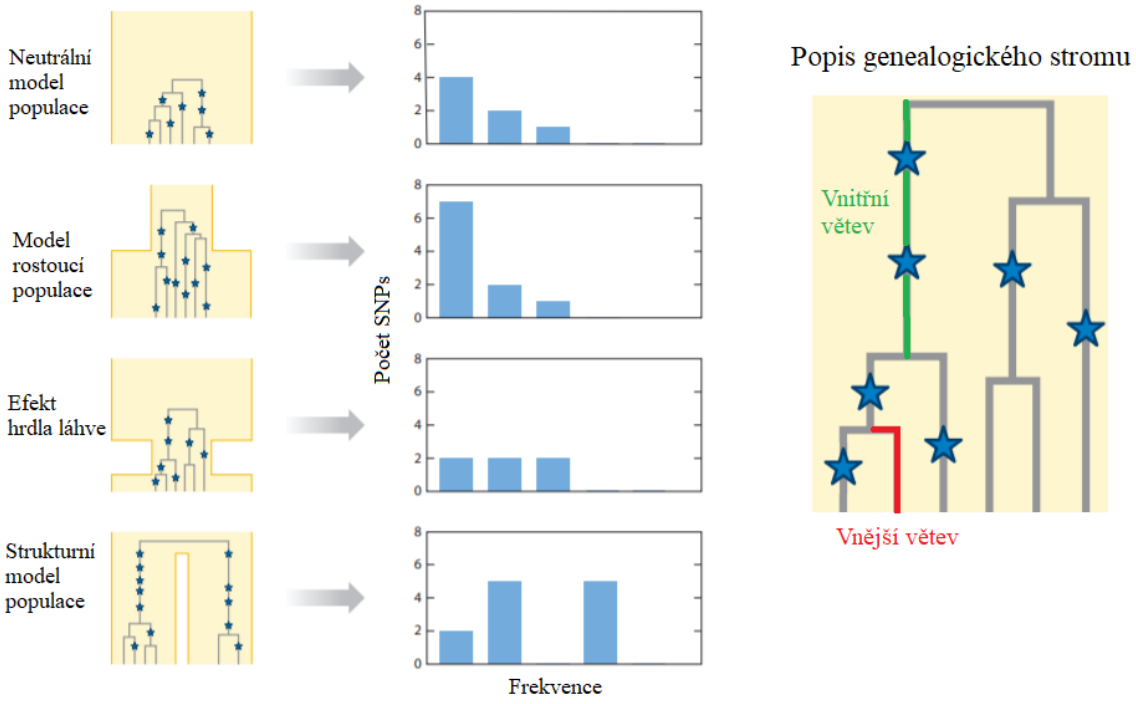
Jelikož SFS je funkcí velikosti vzorku, jsou požadována sekvenční data z více jedinců. I když u této metody není stanoveno, jaký je nejnižší počet osekvenovaných jedinců k získání směrodatných výsledků, vyšší počet jedinců zvýší schopnost odhalení demografických událostí v populaci. Jelikož SNP by měly být nezávislé (měly by pocházet z různých lokusů), SFS je možné počítat i z krátkých, náhodných fragmentů, jako jsou například sekvence vycházející z RAD sekvenování, *(viz kap. 2. Data)*.

Koalescenční teorie nám poskytuje určitá vodítka, jak demografie ovlivňuje SFS. Různé demografické scénáře mění tvar a délku větví základních genealogií, které mění i samotné SFS. Teorie očekávaného SFS v náhodném párování jedinců v populaci konstantní velikosti byla popsána v devadesátých letech minulého století *(Fu, 1995)* a následně byla rozšířena tím, že zahrnovala i aktuální velikost populace *(Griffiths & Tavaré, 1998)*.

Pojďme si tedy přiblížit, jakým způsobem ovlivňují změny ve velikosti populace SFS. Růst populace znamená, že současný počet jedinců v populaci je větší než počet jedinců v

minulosti *(Obr. 4)*. Pravděpodobnost koalescence v určité generaci je u velkých populací nižší než u populací menších. Podle scénáře populačního růstu, mají genealogie odpovídající rostoucím populacím dlouhé vnější větve. U dlouhých větví je pravděpodobnost koalescence v určité generaci menší než u kratších větví. V případě rostoucí populace mají genealogie kratší vnitřní větve. U těchto kratších vnitřních větví je pravděpodobnost koalescenčních událostí v určitě generaci větší. Nesmíme zapomenout na náhodné mutace v těchto genealogiích, které nazýváme singletony. Většina těchto mutací se vyskytuje na vnějších větvích (jelikož většina genealogie u rostoucí populace je tvořena dlouhými vnějšími větvemi). Růst populace tedy vede k tomu, že SFS je vychýlen směrem k většímu podílu nízkofrekvenčních SNP a singletonů *(Obr.4)*.

Zmenšování velikosti populace směrem dopředu v čase vytváří opačný vzorec než růst populace. Při kontrakci je pravděpodobnost koalescence v určité generaci vyšší v současné populaci. Výsledkem populační kontrakce je nižší podíl nízkofrekvenčních variant ve srovnání s populací konstantní velikosti *(Obr.4)*.



Obr.4: 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).

Efekt hrdla láhve také ovlivňuje SFS *(Obr.4)*. Efekt nastává při velkém poklesu jedinců v populaci po kratší dobu. Efekt hrdla láhve má vliv na počet mutací v populaci, neboli dochází ke snížení genetické variability v dané populaci. Tento jev je nevratný, i při opětovném nárůstu velikosti populace. Pokud je efekt velký, tak naprostá většina linií splývá v době extrémního zmenšení populace, což má vliv na SFS. Pokud je efekt mírnější, některé linie se v úzkém hrdle nespojí. Proces splynutí před efektem hrdla ovšem trvá poměrně dlouho. Genealogie populace s efektem hrdla láhve budou mít proto delší vnitřní větve a SFS bude mít méně nízkofrekvenčních variant než SFS populace konstantní velikosti.

Struktura populace také ovlivňuje SFS *(Obr.4)* také. Pokud vzorkujeme stejný počet jedinců ze dvou oddělených populací (míra migrace je nízká a populace se od druhé oddělila před dlouhou dobou), pak je pravděpodobné, že linie v každé subpopulaci se budou navzájem spojovat přednostně než s liniemi z druhé subpopulace. Tyto genealogie proto mají dlouhé vnitřní větve. Mutace na těchto větvích budou výhradně neseny jedinci z dané subpopulace. V extrémním případě se dvě subpopulace budou lišit fixovanými mutacemi. Když jsou data při

SFS kombinována ze dvou populací, výsledkem pro tento typ populační struktury bude přebytek středních frekvencí SNP v SFS *(Beichman et al., 2018)*.

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, buď ve formě analytického výpočtu *(Marth et al., 2004)*, nebo jako simulace *(Nielsen, 2000)*, 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.

Pojďme si přiblížit 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 *(Kamm et al. 2015)*.

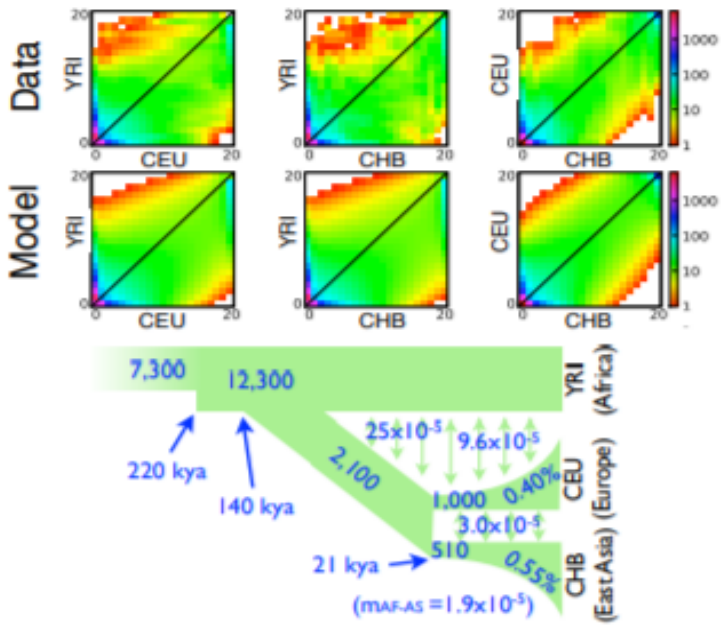
3.1.1. Softwarové balíčky a jejich použití

V dnešní době je k dispozici několik softwarových balíčků, které jsou schopné odvodit demografické parametry pomocí SFS.

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

FastNeutrino je dalším softwarovým balíčkem který používá SFS a koalescenční simulace pro demografické vyvozování. Analyticky vypočítává SFS pro model populace, která mění svoji velikost. FastNeutrino je rychlejší oproti Fastsimcoal2, hlavně pro velké vzorky, ale aktuálně zvládá výpočty pouze pro jednu populaci. Pokud jsou poskytovány délky lokusů, tak program zároveň zjišťuje míru mutace na lokus *(Bhaskar et al., 2015).*

Implementace ∂a∂i (*Diffusion Approximation for Demografic Infference*) odhaduje parametry složitých populačních modelů za použití SFS. Je schopný analyzovat jednotlivce z více populací. Implementace ∂a∂i 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. Důležité je, že ∂a∂iho rychlost umožňuje rozsáhlý bootstrapping pro statistickou charakterizaci modelu, včetně odhadu pro nejistotu parametrů. Tato metoda byla aplikována také na orangutany, rýži a dobytek (*Gutenkunst et al., 2011).*



Obr.5: **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., 2011, převzato a upraveno).

3.2. Metody využívající ABC

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. V předchozí části textu jsem předvedla jednu ze souhrnných statistik, kterou je SFS. Nyní si představíme další souhrnné statistiky, se kterými ABC i další vyvozovací metody pracují.

Souhrnné statistiky jsou hodnoty vypočtené z dat tak, aby reprezentovaly maximální množství informací v co nejjednodušší podobě. Využívají se různé genetické souhrnné statistiky. 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. Poslední souhrnnou statistikou, kterou zmíním, je F_{st}. F_{st} vyjadřuje míru genetické diferenciace mezi populacemi a odráží rozdíly ve frekvencích alel mezi populacemi. Souhrnných statistik, které se využívají, je samozřejmě více, ale rozsáhlejší popis by byl už předmětem jiné bakalářské práce.

Zpět k samotné inferenci. 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:

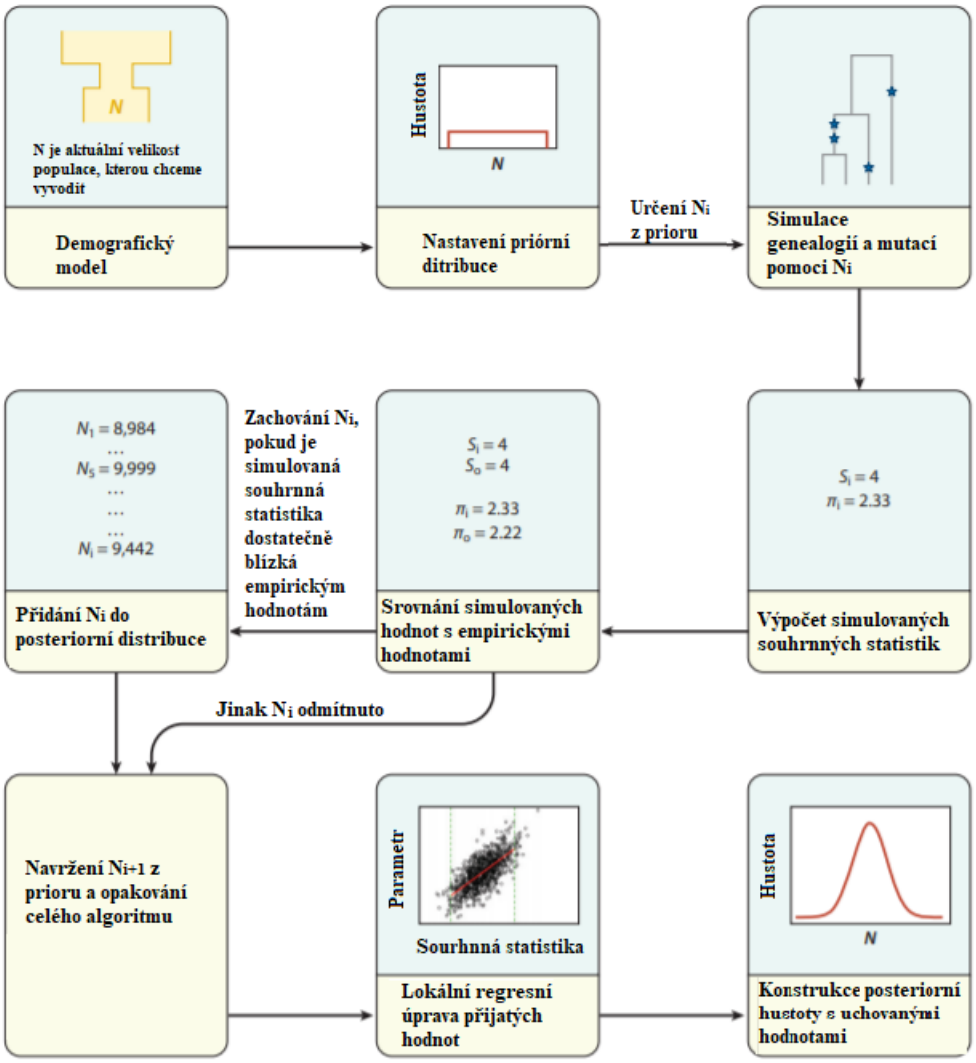
$$l(P|D) = \frac{f(D|P)\, l(P)}{f(D)},$$

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. Někdy evidenci postrádáme. Tato absence nám nedělá problémy při odhadování parametrů, ale při srovnávání modelů ano. Z posteriorní funkce získáváme posteriorní distribuci a z priorní funkce priorní distribuci *(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 *(Obr.6)*. Jelikož ABC přístup používá koalescenční simulace k vytváření souhrnné statistiky, dokáže si například solidně poradit i s hodně polymorfními sekvencemi DNA. Další výhodou přístupu je jeho obecnost. Tento přístup se využívá nejen v demografické inferenci, ale můžeme se s ním setkat například ve systémové biologii, ekologii a epidemiologii.

Při demografické inferenci přístup ABC začíná výběrem konkrétního, libovolně složitého demografického modelu, ze kterého se následně snažíme odhadnout jeho parametry na základě získaných sekvenčních dat. Pro každý sledovaný parametr je uvedena priorní distribuce. Často při tomto přístupu vybíráme uniformní priorní distribuce. Pro parametry, které mají tendenci se často měnit (míra migrace, selekční koeficient) volíme normální distribuce. Pokud máme pocit, že by distribuce mohly být chybné, naskýtá se možnost vícenásobných distribucí. Následné koalescenční simulace jsou prováděny na základě parametrů získaných z priorních distribucí a simulace by měly odpovídat specifikům empirických dat použitých ve studii (počet sekvenovaných jedinců, počet lokusů apod.). 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 (až několik tisíc). Distribuce přijatých hodnot parametrů bude představovat aproximační posteriorní distribuci *(Obr.6)*.



Obr.6: Postup při vyvozování pomocí ABC

Na obrázku je zobrazen pracovní postup ABC pro demografickou inferenci. V tomto konkrétním případě chceme odvodit N, aktuální velikost populace, v modelu úzkého hrdla láhve. Empirická data pro tento příklad se skládají z oblasti genomu, kde $S_o = 4$ a $\pi_o = 2{,}22$. (π je průměrná heterozygosita a S je počet segregujících míst). Souhrnná statistika počítaná ze simulační replikace i jsou označeny S_i a π_i (Beichman et al., 2018, převzato a upraveno).

Některé ABC balíčky zahrnují Markovovy řetězce Monte Carlo *(Markov chain Monte Carlo*, MCMC). 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.

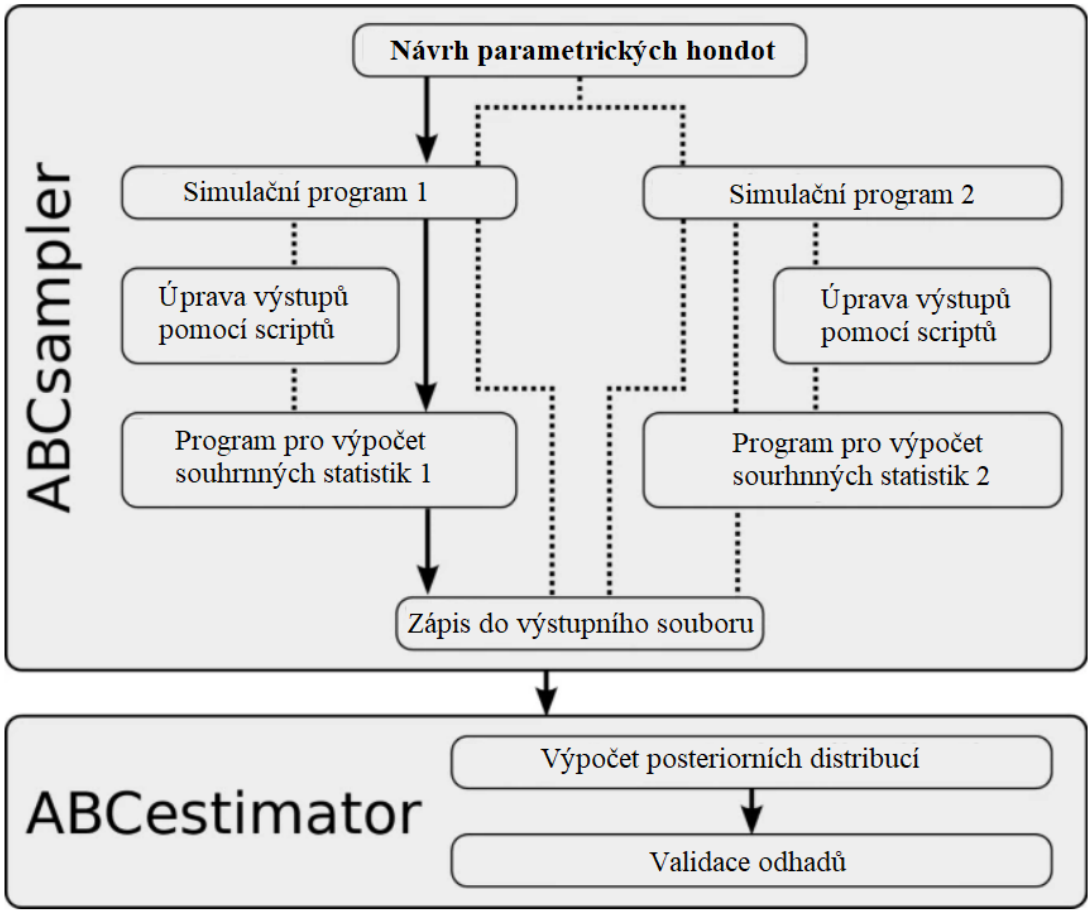
Metody ABC mají několik nevýhod. První nevýhodou je vysoká náročnost výpočtu. Metody počítají přes miliardy koalescenčních simulací, což může omezit parametrický prostor a díky tomu může tento typ výpočtu vést k nesprávnému závěru. Tento problém nám pomáhá řešit paralelizovatelnost nejjednodušších ABC přístupů a zvyšování výpočetní síly počítačů. Za druhé, přístupy ABC nemusí správně fungovat, pokud jsou priorní distribuce příliš široké nebo pokud zkoumaný typ modelu negeneruje sledovaná data. Největší nevýhodou tohoto přístupu je, že jeho správnost závisí na výběru odpovídající souhrnné statistiky. Pokud souhrnná statistika nenese dostatek informací z dat, pak se posteriorní distribuce bude jevit jako priorní distribuce, což nám naznačuje, že data nejsou dostatečně informativní. Správné rozhodování o tom, která souhrnná statistika je relevantní vyžaduje zkušenost a pochopení analyzovaných modelů.

3.2.1. ABC softwarové balíčky

V dnešní době máme k dispozici několik softwarových balíčků, které využívají ABC přístupy. Mezi nejoblíbenější implementace patří DIY-ABC (*Cornuet et al., 2008*), a ABCtoolbox *(Wegmann et al., 2010)*.

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)*.

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. První z nich je ABCsampler, který generuje simulace a počítá souhrnné statistiky pomocí vedlejších pomocných programů. Druhým programem je ABCestimator, který počítá marginální posteriorní rozdělení parametrů ze zaznamenaných simulací, s regresní úpravou nebo bez ní.



Obr.7: ABCtoolbox

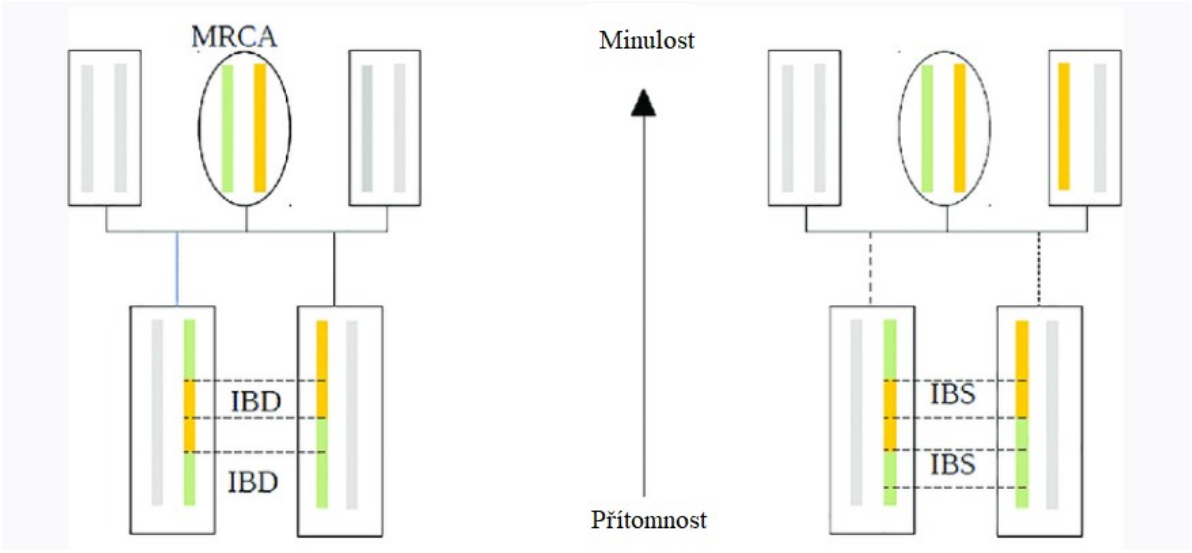
Diagram popisující jednotlivé kroky odhadu ABC pomocí ABCtoolboxu. Černé šipky označují standardní postup. Některé alternativní cesty jsou znázorněny tečkovanými čarami. Například je možné upravit výstup simulačního programu tak, aby bylo možné zohlednit specifické vlastnosti pozorovaných údajů (chybějící 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.7)*. Program ABCestimator přímo čte výstup programu ABCsampler a počítá posteriorní rozdělení na základě simulací, které jsou nejblíže pozorovaným datům. ABCestimator dále nabízí dva způsoby ověření

postupu odhadu. Zaprvé lze otestovat schopnost ABC odhadovat parametry analýzou velkého počtu simulovaných datových souborů se známými hodnotami parametrů vyvozenými z priorních rozdělení (generovaných pomocí ABCsampleru). Na základě takového testovacího souboru dat ABCestimator vypočítá míry přesnosti, jako je zkreslení, střední kvadratické chyby a vlastnosti pokrytí. Za druhé, ABCestimator nabízí nový způsob kontroly, zda jsou pozorovaná data v silném nesouladu s předpokládaným modelem. Jde o výpočet rozdělení mezních hustot pro všechny simulace ponechané pro posteriorní odhad. Mezní hustota pozorovaných dat se pak porovná s rozdělením mezních hustot simulací a vypočítá se hodnota, která ukazuje na schopnost modelu reprodukovat data *(Wegmann et al., 2010).*

3.3. Metody pro určování identity

Podobnost mezi alelami v dané oblasti genomu může být obecně způsobena identitou podle stavu (*Identity by state*, IBS) nebo identitou podle původu (*Identity by descent*, IBD) *(Obr.8)*.



Obr.8: Identita podle původu (IBD) a identita podle stavu (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 (Leitwein et al., 2019, převzato a upraveno).

IBS je termín používaný v genetice k popisu dvou identických alel (sekvencí DNA), které nejsou identické podle původu a nesdílejí společného předka *(Obr.8)*. Míra IBS se počítá přímo z pozorovaných dat, což usnadňuje její výpočet, ale je náchylnější k chybám v sekvenci. Harris & Nielsen (2013) odvodili komplexní demografické události, jako je čas divergence populací, nebo změny velikosti populace, z distribuce délek haplotypů (tj. haploidních sekvencí DNA) IBS mezi páry chromozomů. Výhodou použití IBS namísto IBD je, že IBS je přímo pozorovatelný, zatímco IBD je třeba odvodit a zároveň z IBS lze vyvodit jak dávné, tak nedávné genetické změny. 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 (tj. haploidních sekvencí), aby se získal úplný obraz genetické variability. Tyto metody byly úspěšně použity na soubory lidských dat z projektu 1000 genomů, genomů původních obyvatel Ameriky a na sekvenční data ledních medvědů *(Harris a Nielsen, 2013)*.

Pokud jsou nukleotidové sekvence děděny od společného předka, nazýváme tento jev IBD. IBD neznamená vždy IBS, protože v daném genetickém segmentu se mohly objevit nové mutace, takže není nutné mít stejné složení sekvencí navzdory sdílenému původu (podobně IBS neznamená vždy IBD) *(Obr.8)*. IBD je o něco složitější pro vyvozování, obsahuje více inferenčních kroků a využívá několik dalších nástrojů, které srovnávají přesnost určování IBD. Důležitým faktem je, že na základě znalostí o úrovní sdílení IBD mezi dvěma jedinci jsme schopni odhadnout MRCA, protože IBD nám dává přímé informace o původu. 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)*. Další demografické rysy, jako je ku příkladu zvýšená pravděpodobnost páření v rámci určitého společenství, zvýší IBD nad základní úroveň (*Beichman et al., 2018)*.

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 a Durbin, 2011*) a vícenásobná sekvenčně Markovská koalescence (t*he multiple sequentially Markovian coalescent,* MSMC, *Schiffels a Durbin, 2014*) využívají tyto informace k rekonstrukci efektivní velikosti populace (Ne) v čase 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.10) (Mather et al., 2019)*. Výsledné simulace těchto přístupů jsou generovány skrytým Markovovým modelem. Pojďme si nejprve popsat jeho matematický princip.

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:

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

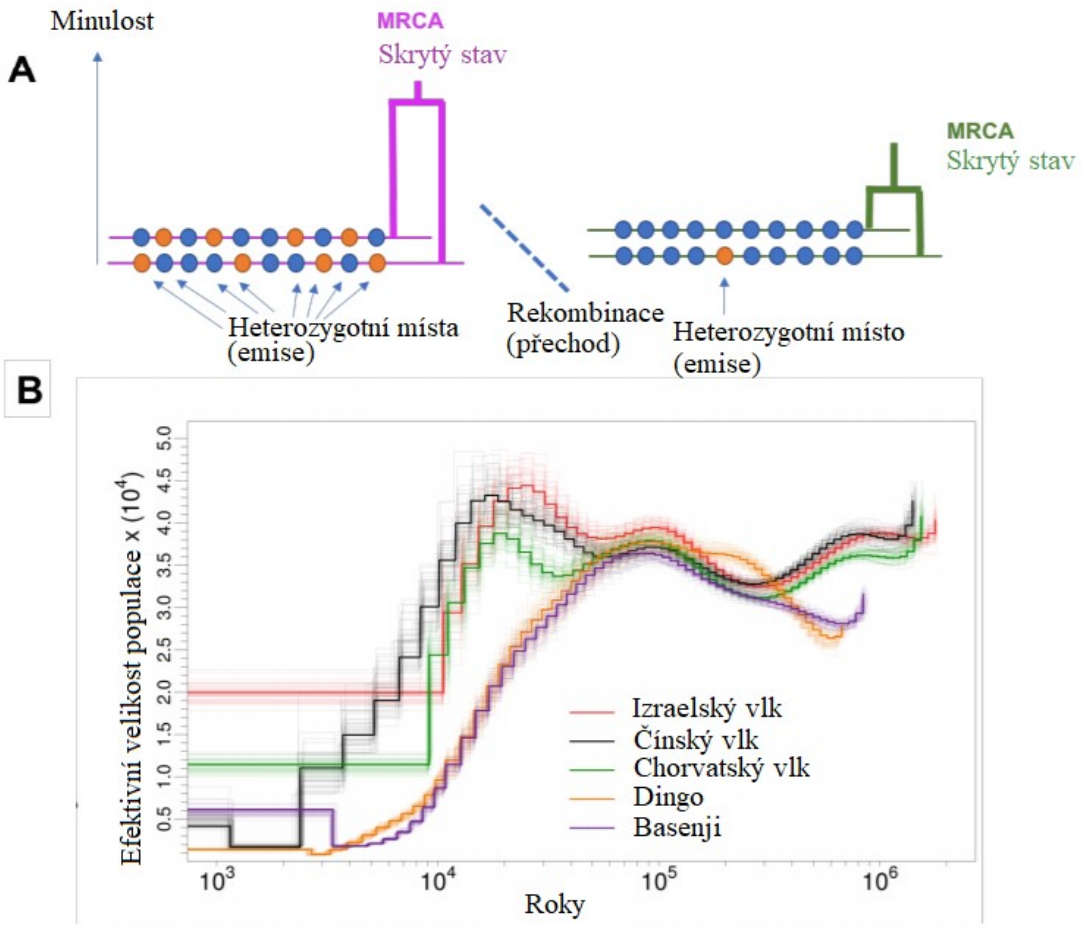
soubor pravděpodobnostních rozdělení pro libovolné $n \in \{1, \ldots, N\}$ a $m \in \{1, \ldots, M\}$ nazývaných "emisní pravděpodobnosti", který popisuje, jak stavy X_t generují hodnoty Y_t. Každé z nich bude mít tvar následující:

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

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).*

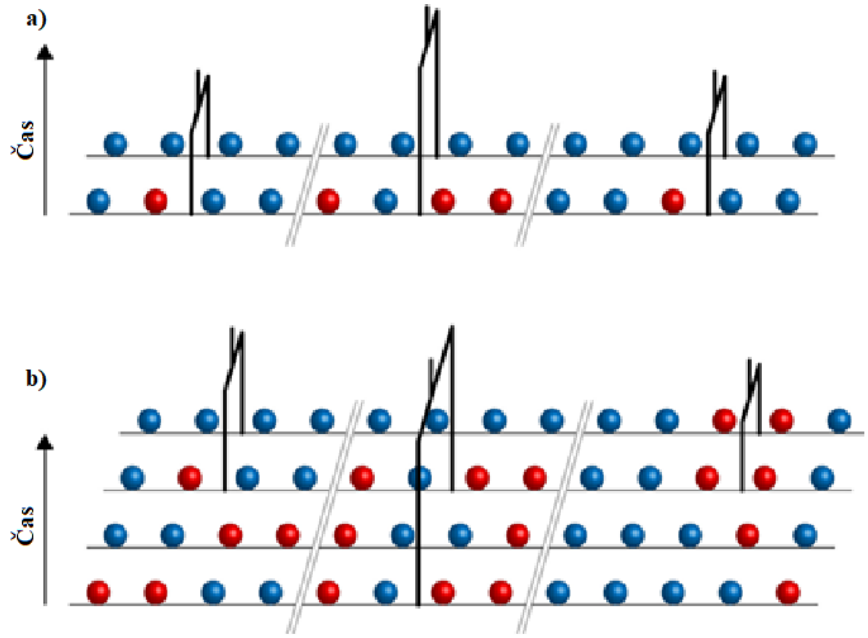


Obr.9: Princip HMM pro vyvozování, PSMC trajektorie pro různé Canidae druhy

*V části **A** 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.*

*V části **B** je 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 (Beichman et al., 2018, převzato a upraveno).*

Obě metody jsou užitečné při studiu hlubších populačních časových harmonogramů, především pokud máme data z omezeného počtu jedinců. Například PSMC se ve velké míře využívalo při studiích starověkého koně a starověkého vlka *(Skoglund, Ersmark, Palkopoulou a Dalén, 2015)*. Kromě rekonstrukce demografické historie byly PSMC a MSMC použity 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 a Rodríguez, 2016)*.



Obr.10: 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 části ***a)*** *u PSMC vidíme pouze dva haplotypy. Topologie lokálního stromu je tedy pevná, ale čas do MRCA mezi lokusy se liší. Jak je vidět u obrázku* ***b)****, u MSMC existuje více haplotypů. MSMC ignoruje topologie lokálních stromů a zaměřuje se na nejnovější koalescenční události v každém lokusu.*

U PSMC je lokální genealogie kompletně charakterizovaná časem do MRCA dvou alel, protože existuje pouze jedna možná topologie stromu. 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).

3.5. Novější koalescenční metody

Novější metody nadále využívají distribuci míry koalescence, ale především kombinují výhody výše zmiňovaných metod.

Jedním z novějších přístupů je **MAGIC** (*Minimal assumption genomic inference of coalescence, Weissman a Hallatschek, 2017*). Tento přístup používá celou řadu velikostí snímacích oken napříč genomem k odvození distribuce koalescenčních časů jednoho lokusu z libovolného počtu genomů, aniž by výslovně modeloval rekombinaci. MAGIC umožňuje uživatelům používat k testování výsledků simulací modely proti empirickým shrnutím koalescenčního procesu. Tato metoda je výkonnostně srovnatelná s PSMC a MSMC metodou, tudíž se používá, jak je uvedeno v následujících částech této práce *(Obr.11)*, při stejných podmínkách pro vyvozování.

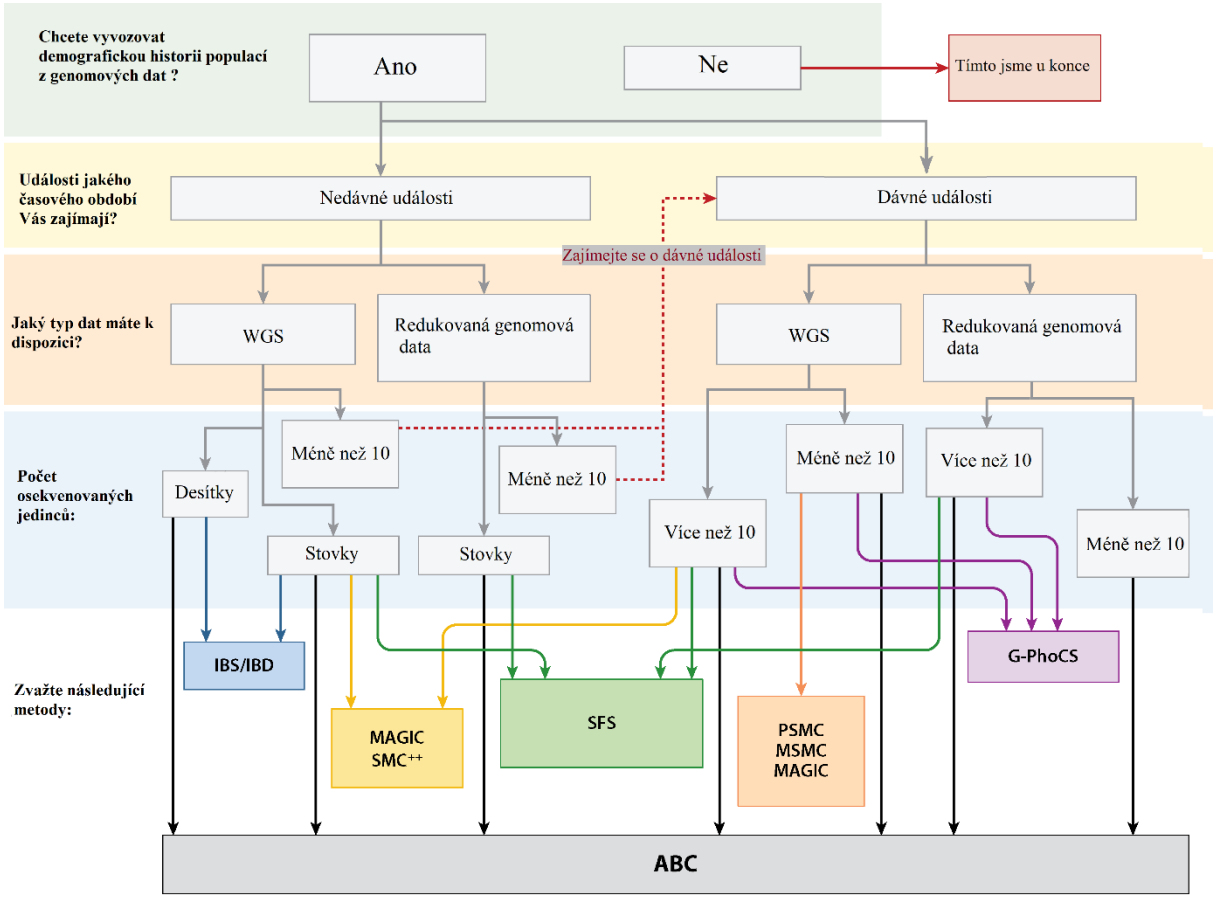
SMC ++ je další novou vyvozovací metodou. Kombinuje výpočetní efektivitu SFS a využívá informací o vazební nerovnováze v sekvenčních Markovových koalescenčních metodách. Tato metoda je navržena tak, aby využívala moderní datové soubory sestávající ze stovek nefázovaných celých genomů. Nefázovanými genomy rozumíme genotypy bez ohledu na to, který z páru chromozomů nese danou alelu. Dokáže také analyzovat dvojice divergentních populací, což jí 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).*

Posledním programem, který v této práci zmíním je generalizovaný fylogenetický koalescenční vzorkovač (*A Generalized Phylogenetic Coalescent Sampler,* G-PhoCS). **G-PhoCS** je typicky provozován na mnoha nezávislých krátkých genomových fragmentech rozptýlených po celém genomu. Data pro tento přístup můžeme získat z celogenomových sekvencí, sequence capture nebo z RAD-Seq *(viz kap. 2. Data)*. Předpokládá se, že v každém fragmentu neexistuje rekombinace. K odvozování pomocí G-PhoCS je zapotřebí sekvencí od několika jedinců. Vzhledem k tomu, že přístup využívá maximální množství informací z dostupných sekvencí, je ideální z hlediska statistické inference. G-PhoCS provádí dedukci podle multipopulačního demografického modelu, jehož parametry sestávají z časů rozdělení populace, relativních velikostí populace a migračních pásem (tj. míry migrace mezi

populacemi v konkrétních časových bodech). Genealogie se generují podle konkrétních demografických modelů a poté se vypočítá pravděpodobnost pro daná sekvenční data pro každý fragment. Jelikož většina genealogií má velmi nízkou pravděpodobnost toho, že budou generovat pozorovaná sekvenční data, G-PhoCS nevzorkuje genealogie náhodně, ale místo toho používá vzorkování Metropolis – Hastings *(Gronau I. 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. Kromě toho je tento přístup navržen pro odvozování v rámci komplexních demografických modelů, jako jsou multipopulační modely s migrací. Hlavní nevýhodou G-PhoCS je, že je výpočetně náročný a často vyžaduje až týdny času pro samotný výpočet. Navíc je méně schopný odvodit nedávné změny velikosti populace ve srovnání s jinými přístupy, což je důležitým parametrem při správném výběru vyvozovací metody (*Beichman et al., 2018).*

4. Postup při výběru vyvozovací metody

V této kapitole si ukážeme, jaký vyvozovací přístup je nejvhodnější použít na základě dostupných dat a časového úseku, který nás zajímá *(Obr.11)*. Postup při výběru vyvozovací metody neřeší otázku výběru vhodného demografického scénáře, který má být vložen do postupu vyvozování. Výběr vhodného demografického scénáře (např. úzké hrdlo, růst populace, divergence populace apod.) nemusí být triviální, protože závisí na demografické historii daného druhu organismu, výzkumné otázce a typu dat, která budou generována.



Obr.11: Rozhodovací strom pro výběr vyvozovací metody

Na obrázku je vyobrazen rozhodovací strom k určení, které metody demografického vyvozování jsou vhodné pro danou biologickou otázku a data. Volba metody závisí na časovém období, které biologa zajímá (např. nedávné události, k nimž došlo během posledních ~100 - 1 000 let), typu sekvenčních dat a velikosti vzorku. Je zajímavé, že velikost vzorku je hrubým doporučením, nikoliv pevnou hranicí (Beichman et al., 2018, převzato a upraveno).

4.1. Selekce

Výše diskutované vyvozovací metody předpokládají, že uvažované lokusy se vyvíjejí neutrálně. Některé lokusy v genomu jsou však také ovlivněny selekcí ať už pozitivní, vedoucí k fixaci výhodných mutací, nebo negativní, která naopak odstraňuje nevýhodné mutace. Každý typ může ovlivnit jak samotné funkční mutace, tak související neutrální mutace.

Selekce nemusí ovlivnit jen kódující oblasti genomu, ale díky genetickému svezení se také blízké okolní nekódující sekvence. Existují značné důkazy o tom, že místa v genomech Drosophily a člověka, která se nacházejí v blízkosti genů a v oblastech s nízkou rekombinací, byla ovlivněna selekcí *(Sella et al., 2009).* Ačkoli účinky selekce u mnoha nemodelových druhů je třeba ještě prozkoumat, je pravděpodobné, že měly alespoň nějaký vliv na vzorce genetické variability.

Přítomnost selekce může ovlivnit demografické závěry. Aby se zmírnily selekční účinky, je třeba se zaměřit na neutrálně se vyvíjející lokusy. Účinky výběru na demografické závěry budou pravděpodobně nejproblematičtější v kompaktních genomech, kde je obtížné vybírat místa vzdálená od genů. Správným prvním krokem v boji se selekcí při vyvozování demografické historie populací je vyhýbání se kódujícím sekvencím a místům, která se nacházejí v jejich blízkosti. Druhým doporučeným krokem je používání těch míst v genomu, které mají vysokou míru rekombinace *(Beichman et al., 2018).* Schrider et al. (2016) rovněž doporučují nepoužívat tzv. rozptyly v souhrnných statistikách v přístupech ABC, protože rozptyl může být též ovlivněný přítomností selekce.

5. 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.

V předchozích kapitolách byly představeny různé metody pro vyvozování demografické historie populací. Správný výběr vyvozovací metody je nelehký úkol. V rozhodovacím stromu *(Obr.11)* je téměř u každého scénáře uvedeno více metod, které lze vyzkoušet, přičemž ABC představuje záchytný bod, který lze použít v každé situaci. Vzhledem k tomu, že metody mohou být různě ovlivněny složitostmi demografie, je doporučeno kombinovat výsledky z více metod *(Freedman et al., 2014).*

Ačkoli může být náročné provést více než jeden druh demografického vyvozování, mnoho studií používá více přístupů k posílení výsledků. Pokud však nelze použít více metod demografického vyvozování, doporučuje se zkoumat shodu demografického modelu s více souhrnnými statistikami (*Beichman et al., 2018*).

V rámci své diplomové práce bych se ráda věnovala vyvozování demografické historie populace v praxi.

6. Seznam použité literatury

ADAMS, A. M. a HUDSON, R. R. (2004). Maximum-Likelihood Estimation of Demographic Parameters Using the Frequency Spectrum of Unlinked Single-Nucleotide Polymorphisms. *Genetics*, **168**(3), 1699–1712. doi: 10.1534/genetics.104.030171.

ANDERMANN, T., TORRES JIMENEZ, M. F., MATOS-MARAVÍ, P. et al. (2019). A Guide to Carrying Out a Phylogenomic Target Sequence Capture Project. *Frontiers in Genetics*. doi: 10.3389/fgene.2019.01407.

BAHLO, M. et al. (2000). Inferences from gene trees in a subdivided population. *Theoretical Population Biology*, **57**, 79-95. doi: 10.1006/tpbi.1999.1447.

BEICHMAN, A. C. et al. (2017). Comparison of single genome and allele frequency data reveals discordant demographic histories. *G3: Genes, Genomes, Genetics,* **7**(11), 3605–3620. doi: 10.1534/g3.117.300259.

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.

BHASKAR, A., WANG, Y. X. R. a SONG, Y. S. (2015). Efficient inference of population size histories and locus-specific mutation rates from large-sample genomic variation data. G*enome Research*, **25**(2), 268–79.

COURNET, 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.

CSILLÉRY, K., BLUM, M. G. B., GAGGIOTTI, E. O. a FRANÇOIS O. (2010). Approximate Bayesian Computation (ABC) in practice. *Trends in Ecology & Evolution*, **25**(7), 410-418. doi: 10.1016/j.tree.2010.04.001.

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.

DRUMMOND, A. J., NICHOLLS, G. K., RODRIGO, A. G. a SOLOMON, W. (2002). Estimating Mutation Parameters, Population History and Genealogy Simultaneously From

Temporally Spaced Sequence Data. *Genetics*, 161(3),1307–1320.
doi: 10.1093/genetics/161.3.1307.

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.

FU, X. Y. (1995). Statistical properties of segregating sites. *Theoretical population biology*, **48**(2), 172-97. doi: 10.1006/tpbi.1995.1025.

FUNGTAMMASAN, A., ANANDA, G., HILE, S. E. et al. (2015). Accurate typing of short tandem repeats from genome-wide sequencing data and its applications. *Genome research*, **25**(5), 736-749. doi: 10.1101/gr.185892.114.

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.

GRIFFITHS, R. C. a TAVARÉ, S. (2007). The age of a mutation in a general coalescent tree. *Stochastic models*, **14**(1-2), 273-295. doi: 10.1080/15326349808807471.

GRONAU, I., HUBISZ, M., GULKO, B. et al. (2011). Bayesian inference of ancient human demography from individual genome sequences. *Natural Genetics*, **43,** 1031–1034. doi: doi.org/10.1038/ng.937.

GUSEV, A., LOWE, J.K., STOFFEL, M. et al. (2009). Whole population, genome-wide mapping of hidden relatedness. *Genome Research*, **19**(2), 318–26.

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.

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

HARVEY, M. G., SMITH, B. T., GLENN, T. C., FAIRCLOTH, B. C. a BRUMFIELD R. T. (2016). Sequence Capture versus Restriction Site Associated DNA Sequencing for Shallow Systematics. *Systematic Biology*, **65**(5), 910–924. doi: 10.1093/sysbio/syw036.

CHEN, J., KÄLLMAN, T., MA, X. a ZAINA, G. (2017). Identifying Genetic Signatures of Natural Selection Using Pooled Population Sequencing in Picea abies. *G3: Genes, Genomes, Genetics*, **6**(7), 1979-1989. doi: 10.1534/g3.116.028753.

CHEN, J., KÄLLMAN, T., MA, X., ZAINA, G. et al. (2013). Inferring population size changes with sequence and SNP data: lessons from human bottlenecks. *G3: Genes, Genomes, Genetics*, **110**, 1979-1989. doi: 10.1534/g3.116.028753.

JOHNSTON, H. R., HU, Y. a CUTLER, J. D. (2015). Population Genetics Identifies Challenges in Analyzing Rare Variants. *Genetic Epidemiology*, **39**(3), 145-148. doi: 10.1002/gepi.21881.

JOYCE, P. a MARJORAM, P. (2008). Approximately Sufficient Statistics and Bayesian Computation. *Statistical Applications in Genetics and Molecular Biology*, **7**(1). doi: 10.2202/1544-6115.1389.

LAURENT S., PFEIFER, S. P., SETTLES, M. L., HUNTER, S. S., HARDWICK, K. M. et al. (2016). The population genomics of rapid adaptation: disentangling signatures of selection and demography in White Sands lizards. *Molecular ecology*, **25**(1), 306–23. doi: 10.1111/mec.13385.

LEITWEIN, M., DURANTON, M., ROUGEMONT, Q. et al. (2019). Using Haplotype Information for Conservation Genomics. *Trends in Ecology & Evolution*, **35**(3), 245-258. doi: 10.1016/j.tree.2019.10.012.

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.

LOUDOVÁ, M. (2015). Implemenation of the RAD sequencing methods to the population genetic studies of hedgehogs from the genus Erinaceus. Diplomová práce. *Univerzita Karlova v Praze*, Praha.

MARTH, G. T., CZABARKA, E., MURVAI, J. et al. (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., 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.

NIELSON, R. (2000). Estimation of population parameters and recombination rates from single nucleotide polymorphisms. *Genetics*, **154**(2),931–42. doi: 10.1093/genetics/154.2.931.

NOTOHARA, M. (1990). The coalescent and the genealogical process in geographically structured population. *Jurnal of Matemathical Biology,* **29**, 59-75.

RALPH, P. L. (2019). An empirical approach to demographic inference with genomic data. *Theoretical Population Biology*, 91-101. doi: 10.1016/j.tpb.2019.03.005.

RANNALA, B. (1997). Gene genealogy in a population of variable size. *Heredity*, **78,** 417–423. doi: 10.1038/hdy.1997.65.

SELLA, G., PETROV, D.A., PRZEWORSKI, M. a ANDOLFATTO, P. (2009). Pervasive natural selection in the Drosophila genome?. *PLOS Genetics,* **5**(6). doi: 10.1371/journal.pgen.1000495.

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.

SCHRIDER, D.R., SHANKU, A.G. a KERN, A.D. (2016). Effects of linked selective sweeps on demographic inference and model selection. *Genetics*, **204**(3), 1207–23. doi: 10.1534/genetics.116.190223.

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.

SUNNAKER, M., BUSETTO, A. G., NUMMINEN, E. et al. (2013). Approximate Bayesian Computation. *Plos computational biology*, **9**(1). doi: 10.1371/journal.pcbi.1002803.

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.

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.

WEISSMAN, D. B. a HALLATSCHEK, O. (2017). Minimal-assumption inference from population-genomic data. *ELife*, (6). doi.org/10.7554/eLife.24836.