Česká zemědělská univerzita v Praze

Fakulta agrobiologie, potravinových a přírodních zdrojů

Katedra genetiky a šlechtění

Doplnění (imputace) chybějících genetických markerů SNP

Diplomová práce

Autor práce: Bc. Anita Kranjčevičová

Vedoucí práce: prof. Ing Josef Přibyl, DrSc.
Odborný konzultant: Ing. Alena Svitáková

Čestné prohlášení

Prohlašuji, že svou diplomovou práci "Doplnění (imputace) chybějících genetických markerů SNP" jsem vypracovala samostatně pod vedením vedoucího diplomové práce a s použitím odborné literatury a dalších zdrojů, které jsou citovány v práci a uvedeny v seznamu literatury na konci práce. Jako autorka uvedené diplomové práce dále prohlašuji, že jsem v souvislosti s jejím vytvořením neporušil autorská práva třetích osob.

V Praze dne 8. dubna 2016 _____

Poděkování

Ráda bych touto cestou poděkovala svému vedoucímu práce panu profesorovi Josefu Přibylovi za odborné vedení, cenné rady, velkou trpělivost a skvěle zvládnutou výuku na magisterském studiu. Dále bych chtěla poděkovat své rodině, která mě celou dobu studia vždy podporovala, finančně i psychicky. Za podkladové údaje bych chtěla poděkovat Českomoravské společnosti chovatelů a.s. a Svazu chovatelů holštýnského skotu. Nemalý dík patří také mým nejbližším přátelům, Daně, Barbaře, Zuzaně, Karolíně, Tomášovi, Míše, Alžbětě a Silvii, kteří mě cestou mého studia provázeli a spolu s mou rodinnou mi dodávali sílu jít dál. Děkuji také svému příteli Jakubovi, že to celou tu dobu statečně vydržel. Za cenné odborné rady, přátelský přístup a ochotu pomoci patří velký dík mé kolegyni a kamarádce, inženýrce Aleně Svitákové. Tato práce byla zpracována v návaznosti na výzkumný úkol NAZV Q1 1510139 „Celostátní informační systém genetického hodnocení hospodářských zvířat“, řešený ve VÚŽV.

Doplnění (imputace) chybějících genetických markerů SPN

Souhrn

Práce s genomovými informacemi se ve šlechtění skotu stala standardem. Diplomová práce se zabývá zohledněním chybějících genetických markerů SNP na genetických čipech. Jedná se o doplnění chybějících hodnot v souborech dat obsahujících informace o výskytu jednonukleotidových polymorfismů SNP v genomu skotu. Tyto polymorfismy se využívají při výpočtech genomických plemenných hodnot, při stanovení genomické příbuznosti, a tím i při vlastním hodnocení zvířat. Nejběžnější čipy pro genotypování jsou Illumina a Affymetrix, každá firma vyvíjí vlastní techniku získávání genotypů. Affymetrix má jednotné kódování jednotlivých SNP mezi čipy různých generací, a proto není obtížné použití i starších dat. Illumina využívá mnoho kódování mezi jednotlivými generacemi čipů, přímé porovnání SNP proto není možné. Ve své platformě má čipy o různé hustotě a různé finanční náročnosti. Čipy od Illuminy se staly světovým standardem a jsou používány všemi chovatelskými společnostmi. Nejčastěji využívané programy pro imputace chybějících SNP jsou Beagle, AlphaImpute, Impute 2, FindHap, DAGPHASE, FImputePedImpute a MaCH. Jednotlivé programy vyžadují příbuzenskou vazbu mezi genotypovanými jedinci. V běžném chovatelském provozu genotypování není ve sledu generací. Proto byl použit vlastní metodický postup. Cílem diplomové práce je zmapování stávajícího výzkumu ohledně doplňování chybějících genetických markerů na genetických čipech a ověření výpočetního postupu. Bylo vytvořeno celkem 8 modelů, lišících se počtem testovaných SNP. Otestováno bylo 10 až 100 sousedních lokusů. Testování probíhalo u zvolených lokusů na dvou souborech. Soubor A představoval 260 genotypů býků několika plemen z České republiky. Soubor B obsahoval 3 982 býků z devíti zemí, kteří splňovali podmínku 100% podíl holštýna. V prvním případě bylo dosaženo velmi dobrých výsledků. Předpověď chybějících hodnot se podařila téměř přesně se spolehlivostí modelu 100%, výjimkou byly téměř homozygotní lokusy, kde bylo dosaženo spolehlivosti modelu jen 55%. Při testování druhého souboru dat, který obsahoval mnohem více genotypů, se u nejrozsáhlejšího modelu podařilo dosáhnout spolehlivosti 80 – 90 % a to i v případě homozygotních lokusů. Chyba předpovědi byla vyšší než v prvním případě. Bylo dokázáno, že předpověď chybějících hodnot lze dopočítat pomocí sousedních SNP. Výsledky práce slouží jako základ k dalšímu studiu genomických dat.

Klíčová slova: imputace, genomická příbuznost, SNP markery

Imputation of missing genetic markers SNP

Summary

Working with genomic information in cattle breeding has become a standard procedure. This study is focused on completion of missing genetic markers - SNPs (single nucleotide polymorphisms) - on genetic chips. More specifically completion of missing values in datasets which contain pieces of information about SNP occurence in cattle genome. These polymorfisms are used for evaluation of genomic relationship, prediction of genomic breeding values and for the valuation of tested animals. The most common chips used for genotyping are Illumina and Affymetrix. Each company develops its own techniques of genotype obtaining. Affymetrix has unified coding type of SNPs among chips of different generations and thus even older data can be used. Illumina uses many coding types between different generations of chips. Thus, direct comparison of SNPs is not possible. Illumina has chips of different density and financial costingness. Illumina chips have become a standard all over the world and it is used by all breeding companies. The most used software programs for imputations are Beagle, AlphaImpute, Impute 2, FindHap, DAGPHASE, FImputePedImpute and MaCH. Each software requires a relationship between genotyped individuals. In common breeding business the genotyping is not in train of generations. That is why our own methodological process was used. The aim of this study is to map the current research about the completion of missing genetic markers on genetic chips and to verify the calculation process. In total, it was created 8 models with different amount of tested SNPs. From 10 to 100 neighbouring loci was tested. The testing was processed at chosen loci in two datasets. Dataset A contained 260 bull genotypes of different breeds from the Czech Republic. Dataset B contained 3982 genotypes of pure Holstein bulls from nine countries. In the first case a very good results were obtained. The prediction of missing values was almost accurate with model reliability 100%. The only exception was for almost entirely homozygous loci where the reliability reached only 55%. When the second dataset was tested, the most extensive model reached the reliability of 80 - 90% even in case of homozygous loci. The prediction error value was higher than in the first case. It was proven that missing values prediction is possible to calculate using the neighbouring SNPs. The outputs of this study are to be the base for further study of genomic data.

Keywords: Imputation , genetic relationship, SNP markers

Obsah

1 Úvod

Výsledkem šlechtitelské práce by měl být zisk a jeho trvalé zlepšování jak u chovatele, tak u celého odvětví. Základem rentability je správný a včasný výběr zvířat do plemenitby. Tato zvířata vynikají svými produkčními vlastnostmi založenými na výborném genotypu, který zaručuje předání vhodných genů do další generace, a tím určitou výhodu chovatele v tržním prostředí.

Pro hodnocení a výběr zvířat se používá plemenná hodnota, která představuje číselné vyjádření genetické odchylky jedince v dané vlastnosti od průměru populace. Čím vyšší je plemenná hodnota, tím vyšší je pravděpodobnost geneticky lepšího potomstva. Na každou vlastnost však současně působí vlivy prostředí, ve kterém se daný jedinec pohybuje. Není proto jisté, zda své genetické schopnosti dokáže plně využít a zda svoje genetické založení prokáže užitkovostí. Podíl jednotlivých složek působících na užitkovost jedince shrnuje koeficient dědivosti (poměr genetické proměnlivosti k celkové fenotypové).

V současné době se s rozvojem molekulárních metod dá využít další informace o jedinci, a to přímo o zastoupení jednotlivých alel v jeho genomu. Byly vytvořeny tzv. čipy, pomocí kterých je sledován výskyt jednotlivých bodových mutací v genomu zvířete. Tímto vývojem byl poznamenán i vývoj metod pro hodnocení zvířat a dnešní doba je dobou genomiky.

V praxi se stanovují genomické plemenné hodnoty, dále jen GPH, které vycházejí ze znalostí jednotlivých bodových mutací v genomu, a které mají vliv na užitkové vlastnosti. Znalost těchto mutací je klíčová především pro upřesnění genetických vazeb mezi jednotlivými zvířaty a lze ji použít k upřesnění příbuznosti mezi zvířaty. Je tedy možné vypočítat GPH u mladých zvířat s vyšší spolehlivostí než tomu bylo u běžné předpovědi plemenné hodnoty, a proto lze tyto mladé jedince následně využít v plemenitbě i přes to, že neproběhlo testovací připařování (Schaeffer, 2006; Pešek et al. 2014).

Pro určení bodových mutací (single nucleotide polymorphism), dále jen SNP, se využívají komerčně vyráběné čipy o různé hustotě. Nejběžnější čipy jsou od Affymetrix a Illumina. Čipy od Illuminy se v podstatě staly v chovatelství světovým standardem využívaným všemi chovatelskými společnostmi. Tím je umožněna i vzájemná výměna výsledků genotypování mezi zeměmi a chovatelskými společnostmi. Finanční náročnost čipů závisí na jejich hustotě. Nejčastěji jsou zvířata genotypována na BovineSNP50 BeadChipu obsahující zhruba 54 000 SNP (54K). Některá zvířata jsou však genotypována na čipech s odlišným počtem SNP. Odlišně ogenotypovaní jedinci jsou často vyřazováni z výpočtu

GPH, nebo jsou jejich chybějící údaje buď nahrazeny průměrnou hodnotou SNP v daném lokusu pro celou populaci, nebo se údaje doplňují na základě znalosti hodnot stejných lokusů u příbuzných jedinců a u lokusů sousedních, které jsou se sledovaným lokusem ve vazbě.

S ohledem na různou finanční náročnost jednotlivých čipů a vzhledem k možnostem dalšího využití informací z nich získaných bylo nutné vyvinout metodu dopočtu (imputace) chybějících údajů (Zhang a Druet, 2010). Z ekonomického hlediska se v České republice genotypují zpravidla jen plemenní býci. Vzhledem k ekonomické náročnosti 54K čipů se jako schůdné řešení jeví genotypovat některé jedince na čip o nižší hustotě, tedy i nižší ceně. Genotypování zvířat je finančně náročné, a proto je na dobrovolnosti chovatelů. Vzhledem k tomu, že se genotypují pouze býci, není možné komplexně propojit genotypy zvířat s jejich původem. Je tedy snahou využít programy nebo obecně dopočítat chybějící SNP bez informací o původu zvířete.

2 Vědecká hypotéza a cíl práce

2.1 Vědecká hypotéza

Doplnění chybějících SNP při genotypování plemeníků skotu umožní přesnější stanovení genomické příbuznosti.

2.2 Cíl práce

Práce se zabývá doplněním bodových mutací. Cílem diplomové práce je zmapování stávajícího výzkumu ohledně doplňování chybějících genetických markerů na genetických čipech a ověření výpočetního postupu. Výsledkem by měla být práce, která by podala dostatek informací, popřípadě způsobů řešení, aby mohla být následně využita v praxi.

3 Literární rozbor

Výpočet genomických plemenných hodnot založený na technice microarray pomocí SNP genetických čipů je v dnešní době u dojeného skotu standardem. Základní metodou předpovědi plemenných hodnot je BLUP animal model s určitými úpravami podle chovatelských podmínek a druhu hodnocené vlastnosti (Plemdat, 2009).

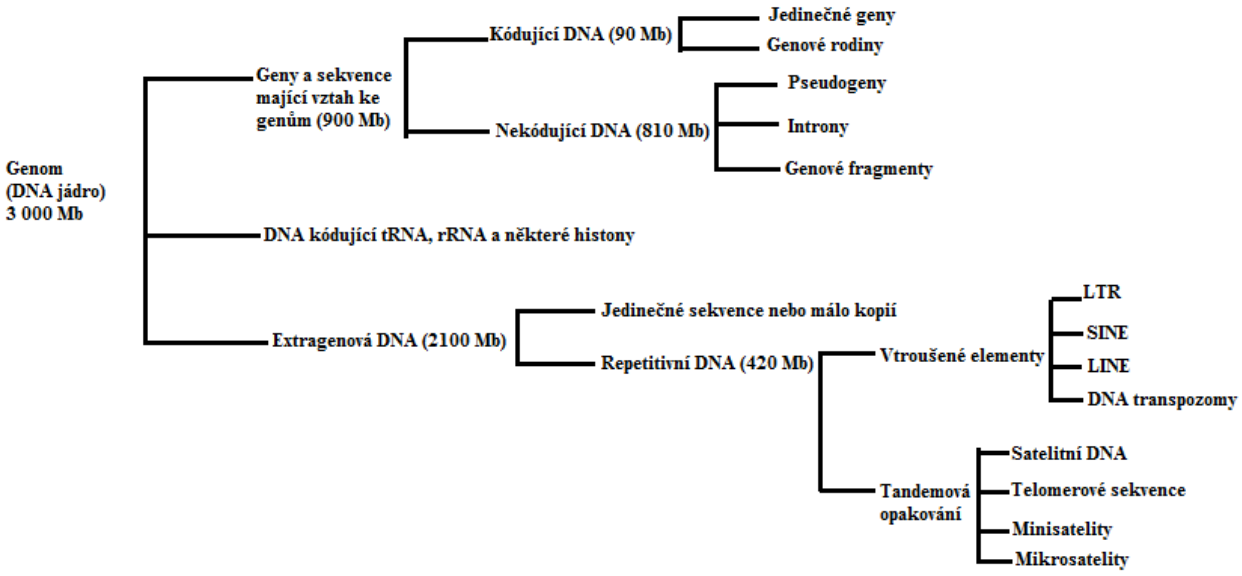
Informace o jednotlivých SNP mutacích je využívána pro výpočet genetických regresních koeficientů SNP markerů na jednotlivé vlastnosti nebo pro upřesnění příbuznosti jednotlivých zvířat (Pešek et al., 2015). Je snaha o výrobu čipů nižší hustoty pro snížení nákladů a efektivnější využití této metody. Jednou z klíčových strategií je doplnění - imputace SNP markerů z čipů o nižší hustotě na čipy s vyšší hustotou (Zhang a Druet, 2010).

V posledních letech se počet genotypovaných zvířat pomocí SNP čipů značně zvýšil. Kromě toho jsou nyní dostupné i čipy obsahující libovolný počet SNP, které si zákazník zvolí sám. Díky tomu dostala důležitou úlohu imputace chybějících markerů. Vliv imputace na výpočet genomické plemenné hodnoty je zpravidla uváděn jako korelace mezi GPH a imputovanými genotypy (Dassonneville et al., 2011, Segelke et al., 2012).

3.1 Vlastnosti genomu skotu

Genom představuje veškerou genetickou informaci buňky. V buněčném jádře je uložen hlavní genom, vedlejší můžeme nalézt v mitochondriích. Genom je složen z genů kódujících určité proteiny, které představují zhruba 5 % genomu, a nekódujících sekvencí DNA, které zaujímají až 95 % genomu (Hruban a Majzlík, 2005). Struktura genomu je znázorněna na obrázku č. 1.

Mitochondriální genomy mají obvykle kružnicové uskupení a jejich velikost se pohybuje od 6 kb do 2 500 kb. Se zvyšující se složitostí eukaryotických organismů se snižuje množství kódujících genů, které se podílejí na vzniku proteinů. Funkce většiny nekódující DNA je neznáma (Snudstad a Simmons, 2009).



Obr. 1 : Schéma genomu savců

Díky fúzovaným buňkám myši a člověka byly vytvořeny postupy pro vývoj prvních komplexních map lidského genomu. Ukázalo se, že použití somatických buněk je velmi účinné a rychlé pro mapování genomu i ostatních savců, včetně skotu (Womack, 2012).

Genom skotu byl zmapován v roce 2009 na krávě plemene hereford. Genom skotu obsahuje zhruba 22 000 genů. Karyotyp skotu obsahuje celkem 29 páru chromozomů a jeden pár pohlavních chromozomů (Womack, 2012).

V současné době jsou známy genetické mapy u člověka, myši, psa, potkana, prasete a skotu (Snudstad a Simmons, 2009)

3.2 Jednonukleotidový polymorfismus SNP

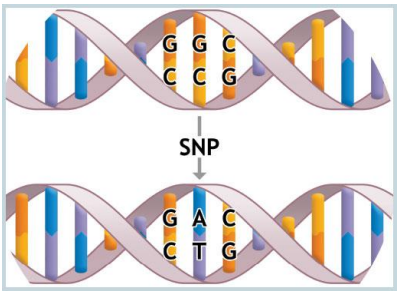
Nejčastějšími změnami v genomu jsou záměny jednotlivých nukleotidových párů, jako je tomu na obrázku č. 2. Například adenin - thymin za cytosin - guanin a opačně. Záměna těchto párů nukleotidů je příčinou velkého množství jednonukleotidových polymorfismů SNP. Udává se, že při porovnání sekvencí dvou jedinců lze nalézt jednonukleotidový polymorfismus na každých 1 200 párů nukleotidů (Snudstad a Simmons, 2009).

Pokud se polymorfismy vyskytují v kódujících oblastech, mohou ovlivnit činnost a vlastnosti výsledného proteinu. SNP tedy můžeme rozdělit podle míry vlivu, na synonymní a nesynonymní.

Synonymní SNP nemají vliv na výslednou sekvenci aminokyselin proteinu. Nesynonymní SNP způsobují změny v kódování aminokyselin a ovlivňují vlastnosti výsledného proteinu (Attia et al., 2009).

V nekódujících oblastech mohou být SNP součástí intronu a promotorových oblastí a mohou tedy v buňce ovlivňovat expresi genů nebo množství vytvořeného proteinu (Schork et al., 2000; Attia et al., 2009).

SNP se objevují v bialelickém zastoupení a jsou tvořeny alelami C a T nebo A a G (Ziegler et al., 2010).



Obr. 2. Schéma SNP - záměna nukleotidového páru (Ječmínková a Kyselová, 2015).

3.3 Metoda DNA microarray

Rozbor a vyhodnocení DNA čipu spočívá v navázání oligonukleotidu - sondy - kovalentní vazbou na destičku. Vzorek DNA, který vyhodnocujeme, musí být nejdříve přečištěn pomocí ELFO nebo PCR a poté je reversní transkripcí převeden na cDNA. Následně se provede amplifikace pomocí metody PCR, aby bylo zajištěno dostatečné množství vzorku.

V dalším kroku je ke každé molekule připojeno několik molekul fluorescentní látky, kterou je posléze možné detekovat. Výsledkem tohoto procesu je vzorek obsahující jednovláknovou DNA, který je označen detekovatelnou látkou. Po kontaktu s čipem je vzorek hybridizován s komplementárními sondami a následným proplachováním očištěn od molekul, které se na sondy nepřichytily dostatečným množstvím vodíkových můstků.

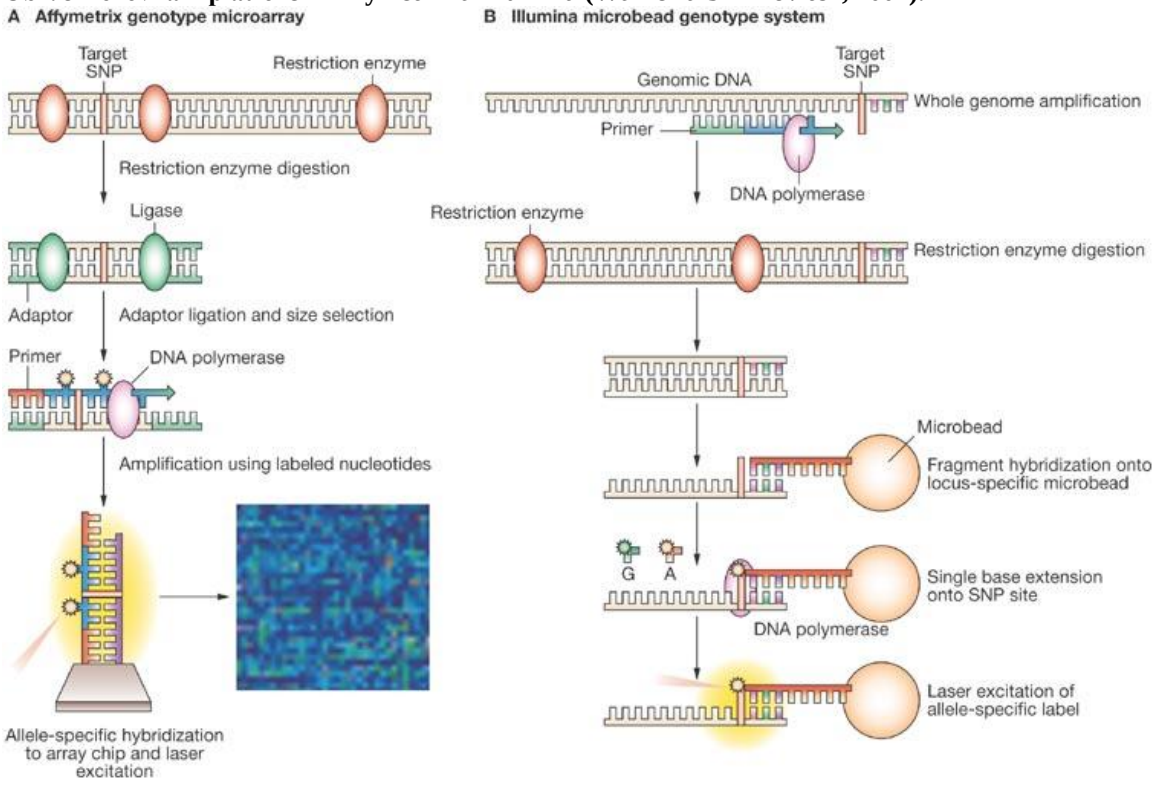
Následně je vzorek vyhodnocen laserem. Barvivy, přítomnými na vzorku, je vyzářeno světlo určité vlnové délky. Podle vyzářeného světla je možné určit množství komplementárních molekul ve vzorku (Elaswarapu a Starkley, 2010).

3.3.1 Porovnání platform Affymetrix a Illumina

U čipu společnosti Affymetrix jsou vzorky DNA štěpeny pomocí restrikčních enzymů na různě dlouhé úseky. Na konce úseků jsou navázány adaptory. Úseky o velikosti 250 – 1 000 párů bází jsou amplifikovány, fluorescenčně označeny a hybridizovány se sondami. Čipy jsou následně vyhodnoceny laserem a k určení genotypu je použit specializovaný počítačový program.

U čipu Illumina jsou vzorky DNA amplifikovány a následně rozštěpeny na menší části. Tyto části se hybridizují na specifické „korálky“ čipu, z nichž každý nese dvě sondy. Je tedy možné současně genotypovat obě alely v lokusu SNP. DNA je navázána v místě značené báze nacházející se před lokusem SNP. Prodloužené vzorky jsou obarveny a vyhodnoceny pomocí laseru a genotypy určeny ve specializovaném výpočetním programu (Walker a Siminovitch, 2007). Rozdíly v postupu genotypování znázorňuje obrázek 3.

Obr. 3 Porovnání platforem Affymetrix a Illumina (Walker a Siminovitch, 2007).



3.4 SNP čipy používané pro genotypování skotu

V současné době je pro skot dostupných 17 druhů čipů o různé hustotě, na dvou různých platformách. Jejich přehled nalezneme v tabulce č. 1.

Tabulka č. 1 Přehled čipů používaných pro genotypování skotu

Název čipu	Počet SNP
Illumina Bovine3k BeadChip	2 900
Illumina BovineLD BeadChip	6 909
Illumina BovineLD v1.1 BeadChip	6 912
Illumina BovineLD v.2 BeadChip	7 931
Illumina BovineSNP50v1 BeadChip	54 001
Illumina BovineSNP50v2 BeadChip	54 609
Illumina BovineSNP50v2 BeadChip	53 218
Illumina BovineHD BeadChip	777 962
GeneSeek Dairy Ultra LD v2	7 049
GeenSeek Genomic Profiler LD v1	8 610
GeneSeek Genomic Profiler LD v3	19 721
GeneSeek Genomic Profiler HD	26 151
GeneSeek Genomic Profiler HD v2	76 876
ICBF International Dairy and Beef v2	139 480
ICBF International Dairy and Beef v3	17 807
Affymetrix Axiom ® Bovine	648 875

Vývoj technologie genotypování pomocí čipů zvýšil poptávku po čipech o různých hustotách. Jen pro skot je v současné době komerčně vyráběno celkem 17 SNP čipů od velkých společností Illumina, Neogen-GeneSeek a Affymetrix. Existují dvě různé technologie genotypování, které poskytuje společnost Illumina a Affymetrix. Kromě toho je čím dál větší zájem o čipy dělané na zakázku, které nejsou komerčně dostupné, nebo je nutné požádat souhlas k jejich použití. Výhodou těchto čipů je především zahrnutí dodatečných SNP mutací v závislosti na dané populaci (mutace pro různé nemoci, atd.) (Nicholazzi et al., 2014).

Zvýšení počtu používaných SNP vedlo ke snaze tato data nějakým způsobem standardizovat. Kódování alel, jméno SNP a poloha SNP v genomu jsou údaje, které je těžké

získat a aktualizovat, a to zejména pokud se jedná o starší SNP čipy, které již nejsou komerčně dostupné.

Affymetrix své genotypy kóduje jako 0 pro homozygota AA, 1 pro heterozygota AB, 2 pro homozygota BB a -1 pro chybějící údaj. Uživatel tak může pomocí referenčního souboru jednoduše alely dekódovat. Na rozdíl od toho Illumina nabízí tři různé druhy kódování alel a to FORWARD/REVERSE, TOB/BOT a A/B.

Východiskem pro správnou standardizaci by mohl být SNP ID, číselný údaj, který je jedinečný pro každé SNP, nicméně tento údaj poskytuje pouze společnost Affymetrix. Tato skutečnost téměř znemožňuje propojení dat s veřejnými databázemi SNP. Standardizace dat je nezbytná, protože genotypy je mnohdy nutné kombinovat (Nicholazzi et al., 2015).

Prvním pokusem o řešení této problematiky byl on-line přístupný nástroj SNAT (Jiang et al., 2011), který již není dostupný. Byl zaměřen spíše na popis SNP než na jejich standardizaci. V roce 2014 Nicolazzi et al. publikovali projekt SNPchiMpv.1, který byl jako první zaměřen na standardizaci údajů o SNP skotu. Od svého prvního vydání tento nástroj téměř zdvojnásobil počet SNP a byl rozšířen do všech šesti hlavních druhů hospodářských zvířat. Nástroj skutečně řeší výše uvedené problémy. Zahrnuje všechny komerčně vyráběné SNP čipy a poskytuje první pokus o standardizaci, integraci a úplné zveřejnění údajů týkajících se čipů SNP (Nicholazzi et al., 2014).

3.4.1 BovineSNP50 BeadChip

Jedná se nejpoužívanější čip pro genotypování skotu. Na tomto čipu je navrženo více než 54 000 sond, které jsou schopny rovnoměrně pokrýt celý genom skotu. Společností Ilumina je garantována 99% přesnost a rovnoměrné rozložení SNP. Medián rozestupu mezi SNP je zhruba 37,4 kb. Jsou pokryta všechna SNP, která jsou významná pro dojený skot s četností méněčetné alely (MAF) alespoň 25% napříč všemi lokusy (Illumina, 2016).

V současné době je možné použít již třetí verzi čipu BovineSNP50 v. 3, mimo již zmiňované první dvě verze BovineSNP50v.1 a BovineSNP50v.2 (Nicholazzi, 2014).

V následující tabulce č. 2 je možné vidět rozdíly mezi všemi verzemi tohoto čipu, v počtu použitých sond, které ukazují přítomnost SNP.

Tabulka č. 2 Porovnání verzí čipu BovineSNP50 podle počtu SNP sond

Zdroj	BovineSNP50 v.1	BovineSNP50 v.2	BovineSNP50 v.3
Illumina GenomAnalyzer	23 840	24 181	22 299
Bovine HapMap Data set	12 298	12 342	11 607
Btau Assembly SNPs	9 381	9 404	9 086
Whole - Genom ShotGun Reads [a]	5 808	6 038	5 485
Holstein BAC Sequences Data	1 409	1 411	1 238
Původy [b]	116	120	200
Ostatní [c]	1 169	1113	3 384
Celkem	**54 001**	**54 609**	**53 218**

a. Odvozeno od šesti plemen skotu - norský červený skot, holštýn, brahman, angus, jerseyský skot a limousin v porovnání s Btau. 2.0.; b. Rodičovské markery ; c. Zahrnují SNP ověřené Institute for Food and Agricultural Sciences Alberta, INRA a French International Institute of Agriculture (Illumina, 2016). Zdroj uvadí projekt, v rámci něhož sondy vznikly.

3.5 Neúplné údaje

Příčinu vzniku neúplných dat je možné obecně rozdělit do tří skupin.

1. Jestliže mají chybějící hodnoty stejnou pravděpodobnost výskytu pro všechny záznamy a data s chybějícími hodnotami nejsou nijak odlišitelná od dat bez chybějících hodnot, potom se jedná o MCAR hodnoty – „Missing Completely at Random“.

2. Pokud se data s chybějícími hodnotami liší od dat bez chybějících hodnot, ale existuje možnost, že je proměnná závislá na jiných pozorovaných proměnných, jedná se MAR hodnoty – „Missing at Random“.

3. Pokud nedojde k naměření hodnoty, popřípadě není-li hodnota doplněna z jiného zdroje a její výskyt je závislý na hodnotě samotné, označujeme ji jako MNAR hodnotu – „Missing Not a Random“ (Rubin, 1976).

3.5.1 Jednorozměrné imputace dat

O jednorozměrných imputacích hovoříme, pokud doplňujeme hodnoty právě jedné proměnné. Řadíme sem metody, kdy chybějící hodnotu můžeme ve sloupci doplnit na základě její polohy, a to nejčastěji aritmetickým průměrem, mediánem nebo modusem.

Aritmetický průměr a medián je možné použít v případě, jedná – li se o proměnnou kvantitativní. Pokud doplňujeme proměnnou kvalitativní, použijeme modus.

Do jednorozměrných imputací patří také takzvaná deduktivní imputace, která je založená na odvození logických vztahů mezi proměnnými, kdy známe celkovou hodnotu proměnné a můžeme jednotlivé chybějící hodnoty dopočítat (Little a Rubin, 2002).

3.5.2 Vícerozměrné imputace dat

Vícerozměrnou imputací doplňujeme chybějících hodnoty ve více proměnných současně. Mezi metody vícerozměrné imputace patří hot-deck, cold-deck, metody založené na maximální věrohodnosti a regresní imputace.

Hot-deck imputace jsou založené na rozdělení hodnot do jednotlivých tříd s následným doplněním chybějících údajů na základě příslušnosti k dané třídě.

Cold-deck imputace je založená na podobném principu jako předchozí, s tím rozdílem, že na doplňování chybějících hodnotnou použity jiné datové soubory než je ten, který chybějící hodnoty obsahuje.

Metody založené na maximální věrohodnosti spočívají v počátečním nastavení parametrů, následným doplněním chybějících hodnot vytvořeným modelem a zpětném přetvoření parametrů, jehož cílem je nalézt parametry, které maximalizují věrohodnost modelu.

Regresní imputací lze doplňovat chybějící hodnoty na základě závislostí v datovém souboru. Pomocí nezávislých proměnných X lze vysvětlit vztah k závislé proměnné Y (Baraldi, 2009).

Maticový zápis:

$$Y = \mu + X\beta + \varepsilon$$

$$Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}, \quad X = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix}, \quad \beta = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_n \end{pmatrix}, \quad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

Kde:

Y = vektor závisle proměnné
X = matice nezávisle proměnné
β = vektor parametrů - regresních koeficientů
ε = chybový vektor¨
1 = průměr

Odhad parametru β lze vypočítat pomocí metody nejmenších čtverců pouze za předpokladu plné hodnosti matice X, pomocí následujícího vzorce.

$$\beta = (X^T X)^{-1} X^T Y$$

Kde:

X^T= transponovaná matice X

Pokud používáme vícerozměrné modely, je nutné provádět výpočet chybějících hodnot iterativně. Lineární regresi je možné použít jen v případě spojitých proměnných. Toto omezení můžeme obejít, pokud použijeme obecný lineární model (GLM), ve kterém nezáleží na tom, zdali jsou proměnné spojité či kategoriální (Yuang, 2016).

3.6 Imputační techniky použité při práci s genotypy ve šlechtění zvířat

Pokrok v biotechnologiích poskytl nové nástroje pro zvýšení užitkovosti a pohody hospodářských zvířat. Genomická selekce, při níž je počítána GPH z celého genomu pomocí SNP, se používá u mléčných plemen skotu v řadě zemí (Hayes et al., 2009). Zájem o tuto technologii byl ovlivněn dostupností komerčních SNP čipů s vysokou hustotou, zvýšením genetického zisku oproti testování užitkovosti zvířat v důsledku snížení generačního intervalu a zvýšení přesnosti výběru plemenných zvířat v mladším věku (Schaeffer, 2006; Pešek et al. 2014).

Imputační techniky využívané ve šlechtění hospodářských zvířat je možné obecně rozdělit do dvou skupin. Techniky založené na vazebné nerovnováze (LD), k nimž se používají programy Impute2 (Howie a Marchini, 2009), Beagle (Browning a Browning, 2007, 2013), Mach (Li et al., 2010) a techniky založené na původu a segregaci (LE) nebo kombinaci původu, segregace a populace, kde se využívá zpravidla AlphaImpute (Hickey et al., 2012), Findhap (VanRaden et al., 2011), DAGPHASE (Druet a Georges, 2010), FImpute (Sargolzaei et al., 2008, 2014), PedImpute (Nicolazzi et al., 2013).

Přesnost imputace je ovlivněna několika činiteli, například počtem a složením jedinců v referenční skupině, efektivní velikostí populace, frekvencí alel a rozdíly mezi hustotami referenčních a imputovaných genotypů (Sargolzaei et al., 2014).

3.6.1 Programy používané pro imputaci genotypů

3.6.1.1 Beagle

Beagle je počítačový program, který byl vytvořen Brianem Browningem z Washingtonské univerzity. Je určen pro práci s genotypy, fázování genotypů, doplňování chybějících markerů a určení úseků IBD – shodných podle původu. Patří mezi nejvyužívanější softwary pro dopočet chybějících markerů.

Používá se k fázování genotypů nepříbuzných jedinců, rodičovských párů a příbuzných zvířat. Doplňuje chybějící markery v genotypech a slouží k určení homozygotů, popřípadě genetických oblastí shodných podle původu u dvou jedinců.

Beagle v modelu pro výpočet pravděpodobnosti používá shlukový model lokalizovaných haplotypů (LHCM), které mohou být označovány jako speciální případ tichého Markovova modelu (HMM) (Browning, 2013).

3.6.1.2 AlphaImpute

Program AlphaImpute byl vytvořen Johnem Hickeym jako nástroj pro doplňování chybějících dat genotypů za použití informací o původu zvířete. Tento program kombinuje segregační analýzu, imputaci pomocí haplotypových knihoven a fázování genotypu.

SNP jsou pro výpočet v tomto programu upraveny na čtyři hodnoty 0, 1, 2, 3. Recesivní homozygot je značen jako 0, heterozygot jako 1, dominantnímu homozygotovi je přiřazena 2 a pro chybějící údaj se uvádí 3.

Program nejdříve oddělí zvířata, která jsou genotypována na čipech o vysoké hustotě, pomocí těchto zvířat je vytvořena haplotypová knihovna. Alely jsou dopočítávány, pokud je jejich pravděpodobnost výskytu vyšší než 0.99.

Tento proces lze považovat za fázování jednoho lokusu. Jakmile je jednoduché a dlouhé fázování alel dokončeno, pokračuje tento proces od nejstaršího po nejmladšího jedince v rodokmenu.

Druhá část programu zahrnuje ověření, jestli haplotypy nepříbuzných zvířat existují u zvířat v knihovně. Haplotypové knihovny jsou během procesu několikrát obnoveny.

Pomocí obnovených knihoven obsahujících velké množství SNP jsou dopočítány alely pro ostatní zvířata v rodokmenu. Tento program umožňuje dopočet SNP pro jedince z rodokmene, kteří jsou genotypováni na malých čipech nebo nemají stanovený žádný genotyp (Hickey et al., 2012).

3.6.1.3 Impute 2

Základní myšlenkou tohoto programu je předfázování genotypu k vytvoření haplotypu s nejlepší shodou a následné doplnění předpřipravených genotypů v samostatném programu. Předfázování vede sice k malé ztrátě přesnosti, protože je opominut nejistý odhad haplotypů, ale je umožněna velmi rychlá imputace dat.

V jednom kroku je možné imputovat genotypy v rámci celého chromozomu, je však lepší chromozom rozdělit na více menších částí. Důvodem je, že program pracuje s větší přesností na menších úsecích genomu, a také možnost imputovat na více procesorech paralelně, což umožňuje zkrácení reálného času výpočtu a omezení množství paměti počítače potřebné pro práci programu.

Výsledkem jsou tabulky poskytující statistické údaje o imputovaných datech a údaje o ověření správné činnosti programu (Howie et al., 2014).

3.6.1.4 FindHap

Program Findhap.f90, napsaný v jazyce fortran, byl vytvořen, aby sloučil a využil informace o populaci a haplotypování z rodokmenu. Genotypy jsou číselně kódovány jako 0, pokud se jedná o homozygota v první alele a 2 pro homozygota ve druhé alele. Pokud není údaj znám nebo se jedná o heterozygota, je značen jako 1. Haplotypy jsou číselně označeny jako 0 pro první alelu, 2 pro druhou alelu a 1 jako neznámou pro párování.

Algoritmus nejprve vytvoří seznam haplotypů z daných genotypů a následně se postup iterativně opakuje. Genotypy, které jsou v souboru, mohou být doplněny použitím haplotypů, které vznikly později.

První dvě opakování slouží pro tvorbu haplotypů pouze z populace. První opakování používá genotypy s nejvyšší hustotou, následně jsou použity všechny. Třetí a čtvrté opakování již používá k tvorbě haplotypů informace z populace i z rodokmene.

Haplotypy nebo genotypy rodičů či prarodičů jsou kontrolovány jako první. Pokud nějaký z haplotypů není při této kontrole nalezen, začíná kontrola znovu od začátku uspořádaného seznamu (Van Raden et al., 2011).

Tvorba haplotypů probíhá v několika krocích:

- Každý chromozom je rozdělen do několika částí podle počtu hodnocených markerů.
- První genotyp je zapsán do seznamu haplotypů, jako by se jednalo o haplotyp.
- Každý další genotyp, který sdílí daný haplotyp, je později použit pro přičlenění předchozích genotypů k haplotypům.
- Každý genotyp je porovnán se seznamem, shoda je potvrzena, pokud není žádný homozygotní lokus v rozporu s uloženými haplotypy.
- Všechny další neznámé alely v tomto haplotypu jsou dopočteny z homozygotních alel v genotypu.
- Druhý haplotyp jedince je získán odečtením prvního haplotypu od genotypu a je následně porovnán s ostaními haplotypy v seznamu.
- Tam, kde není nalezena shoda, je nový genotyp, popřípadě haplotyp přidán na konec seznamu. Neznámé alely genotypu jsou uloženy jako neznámé alely haplotypu.
- Seznam aktuálně známých haplotypů je setříděn od nejvíce k nejméně častým, více se opakující haplotypy jsou při doplňování upřednostněny (Van Raden et al., 2011).

3.6.1.5 DAGPHASE

Program DAGHPHASE byl použit pro výpočet chybějících hodnot v genotypech zvířat pomocí informací o mendelistické dědičnosti alel a informací o rodokmenu testovaných zvířat. Pracuje na podobném základu jako Beagle, pracuje tedy na základě LHCM modelu (Druet a Georges, 2010).

3.6.1.6 FImpute

Program FImpute byl vyvinut pro doplňování chybějících markerů genotypů hospodářských zvířat, která jsou genotypována často na rozdílných čipech.

Postup imputace začíná zachycením podobnosti mezi haplotypy příbuzných zvířat. Algoritmus programu předpokládá skutečnost, že haplotypy všech jedinců jsou k sobě navzájem připojeny v různých stupních. Příbuzní jedinci mají obvykle o něco delší haplotypy než nepříbuzní.

Pokud jsou informace o rodokmenu zvířete známé, program pracuje přesněji. Klíčem přesné imputace je vložení dostatečně kvalitního čipu s vyšší hustotou (Sargolzaei et al., 2014).

3.6.1.7 PedImpute

PedImpute je program napsaný v jazyce Fortran, který slouží k rychlému tvoření haplotypů a imputaci genotypů mléčného skotu, pomocí rodokmenů.

Program bere v úvahu všechna genotypovaná zvířata a úzce je propojuje negenotypovanými zvířaty (alespoň s jedním genotypovaným rodičem či potomkem). Výpočet probíhá iterativně a jsou při něm střídavě využívány údaje o původu zvířete a údaje o haplotypech v populaci.

Část programu využívající rodokmen zvířat používá úseky DNA pro každou dvojici rodič – potomek. Pro část, která využívá informace o haplotypech v populaci jsou úseky pevně dané. Počet iterací je předem stanoven nastavením různě dlouhých haplotypů (Nicholazzi et al., 2013).

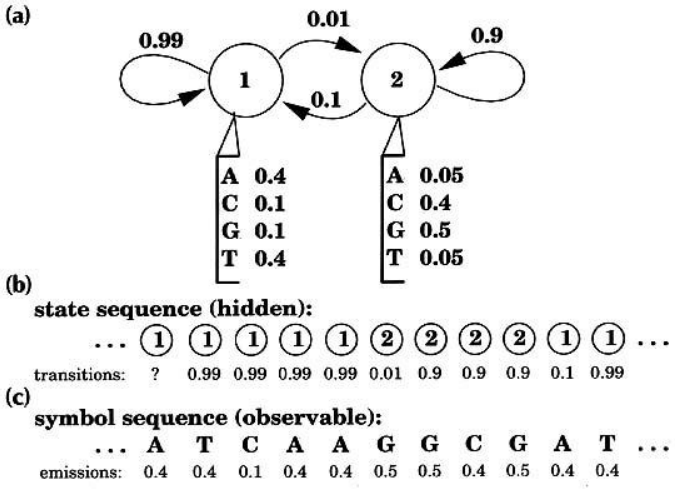
PedImpute zpracovává v jednom kroku jeden chromozom. Doplňování více chromozomů je možné, závisí však na možnostech procesoru a dostatečná velikost paměti v počítači (Jansen, 2012).

3.6.1.8 MaCH

MaCH je nástrojem pro haplotypování a imputace genotypů, pomocí skrytého Markovova modelu - HMM. Haplotyp je odhadnut pomocí genotypu pro každého jedince. Výpočet probíhá iterativně, kde při každé iteraci je nový pár haplotypů každého jedince pomocí HMM popsán, jako neúplná mozaika ostatních haplotypů.

Modelové parametry charakterizují pravděpodobnost změny mozaikové struktury mezi párem po sobě jdoucích markerů a pravděpodobnost pozorování neúplných údajů. Po mnoha iteracích (zpravidla 20 – 100) dochází ke sloučení vzorků haplotypů z každého opakování výpočtu (Li et al., 2010).

HMM je model popisující rozdělení pravděpodobností přes nekonečný počet možných sekvencí. Podstata je patrná z obrázku č. 4, popisující model pro sekvenci DNA (Eddy, 1996).



Obr. 4. Jednoduchý HMM model pro vznik sekvence DNA. A) Případ 1, který vytváří sekvence, kde jsou bohatě zasoupeny A- T a případ 2, při kterém jsou vytvářeny sekvence bohaté na G- C. Pravděpodobnosti změny případů jsou uvedeny šipkami a pravděpodobnost výskytu jednotlivých bází ve sloupci pod schématem. B) Model vytváří sekvenci jako Markovův řetězec, přestavující pravděpodobnost změny. C) Představuje sekvenci s pravděpodobnostmi výskytu jednotlivých bází v závislosti na výskytu v segmentu bohatém na A- T 1 nebo segmentu bohatém na G- C 2 (Eddy, 1996).

Pro naše podmínky a údaje o genotypech býků od Svazu chovatelů holštýnského skotu v návaznosti na výše uvedené metody vyplývá, že je nutné pro dopočet chybějících SNP lokusů zvolit vlastní metodický postup.

4 Materiál a metody

4.1 Vstupní data

Pro účely této diplomové práce byla použita databáze genotypovaných zvířat, kterou poskytl Svaz chovatelů holštýnského skotu. Údaje jsou uloženy na Plemdatu, který patří pod Českomoravskou společnost chovatelů a.s. (ČMSCH, 2016). Celá databáze čítá zhruba 6 500 zvířat. Jedná se ve velké části o býky, kteří se podíleli na tvorbě populace holštýnského skotu v České republice.

Zvířata byla genotypována na čipech Illumina Bovine SNP50K BeadChip. Většina genotypů byla získána ze zahraničí. Ne u všech byly získány údaje o kvalitě čipů, jako je například G - skóre. Při zpracování bylo proto nutno používat jednodušší postupy, za určitého snížení přesnosti výsledků.

4.2 Úprava dat

Testování vhodných modelů probíhalo na dvou souborech. První soubor - A obsahoval genotypy 260 býků všech kříženců s vysokým podílem H, původem z České republiky. Druhý soubor – B obsahoval genotypovaná zvířata z 9 zemí, které má Svaz chovatelů holštýnského skotu k dispozici. V tomto souboru byla provedena úprava dat přidáním plemene a výběrem býků, kteří mají 100% podíl holštýnského plemene. Toto plemeno bylo vybráno, protože zaujímalo největší zastoupení. Celkem bylo vybráno 3982 býků.

Pro výpočet bylo nutné k dostupným informacím přidat umístění jednotlivých SNP na určitých chromozomech. Pomocí databáze NCBI (2015) byl vytvořen seznam jednotlivých SNP s jejich umístěním a chromozomem, který byl následně připojen k souboru dat.

Bylo použito přečíslování alel, které se u nás běžně používá při práci s SNP a to 0 pro homozygota BB, 1 pro heterozygota AB a 2 pro homozygota AA (Pešek et al., 2015).

Data byla následně roztříděna do 29 samostatných souborů, představujících jednotlivé chromozomy. Lokusy SNP byly seřazeny podle umístění na chromozomu. Pro modelové testování byly z jednotlivých souborů vyřazeni jedinci, kteří měli více než 10 % chybějících údajů, a lokusy, které obsahovaly více než 5% chybějících údajů.

Výsledný soubor obsahoval tabulku (matici) hodnot jednotlivých lokusů, přičemž každý řádek zahrnoval lokusy jednoho býka.

4.3 Výpočet

Nebyly použity údaje o původech zvířat z důvodu nedostatku genotypů rodičů genotypovaných zvířat. Nedostatek informací o formátech čipů nedovolil efektivní využití některého z existujících imputačních softwarů.

Pro upravené údaje byl použit obecný lineární model (GLM) s regresními koeficienty. Regresní analýza souboru umožnila odhad regresních koeficientů sousedních lokusů k lokusu hodnocenému a zpětné odhadnutí hodnot závisle proměnné.

Postup vychází ze závislosti sousedních lokusů mezi sebou na základě genové vazby. Lokusy jsou poměrně hustě vedle sebe. Předpokládá se jen mizivý výskyt crossing-overů uvnitř skupiny sousedních lokusů. Z tohoto důvodu byla data rozčleněna do souborů, které představovaly jednotlivé chromozomy.

4.3.1 Postup výpočtu

Výpočet probíhal v programu SAS (2002) pomocí procedury GLM. Procedura GLM představuje lineární model, kde jsou pomocí metody nejmenších čtverců odhadnuty regresní koeficienty. Tyto koeficienty byly následně použity ke zpětnému odhadu závisle proměnné. Spolu s regresními koeficienty byl vypočítán koeficient determinace, který udává spolehlivost modelu. Dalším výpočtem byly udělány zpětné předpovědi hodnot závisle proměnné a byla zjištěna chyba a absolutní chyba předpovědi.

Maticový zápis:

$$Y = \mu + X\beta + \varepsilon$$

Kde:

$$Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}, \quad X = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix}, \quad \beta = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_n \end{pmatrix}, \quad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

Y = vektor závisle proměnné – sledovaný lokus

X = matice nezávisle proměnné – sousední lokusy

β = vektor parametrů – regresních koeficientů

1 = průměr

ε = chybový vektor

Za předpokladu, že

$$E(\varepsilon) = o \;\; a \; var(\varepsilon) = \sigma^2 I_n,$$

kde o představuje nulový sloupcový vektor a I jednotkovou matici řádu n.

V modelu lineární regrese má odhad parametrů metodou nejmenších čtverců tvar:

$$\beta = (X'X)^{-1}X'Y$$

Reziduální součet čtverců se následně vypočítal jako:

$$Se = (Y - Xb)'(Y - Xb) = Y'Y - b'X'Y$$

Koeficient determinace je vypočítán jako:

$$R^2 = 1 - \frac{S_e}{S_t}$$

Kde S_e představuje reziduální součet čtverců a S_t, jako celkový součet čtverců. Hodnota R^2 leží v intervalu <0, 1> a udává, jaký podíl rozptylu v pozorování závisle proměnné se podařilo modelem vysvětlit. Čím vyšší je hodnota, tím vyšší je spolehlivost modelu.

Absolutní chybu modelu vypočítáme jako rozdíl předpovězené hodnoty (M) a skutečné hodnoty (S) měřené proměnné.

$$\Delta = M - S$$

4.3.2 Modely

Testování datových souborů probíhalo celkem na 8 modelech. Modely se od sebe lišily na základě počtu lokusů sousedících s vybraným lokusem, které představovaly nezávisle proměnné. Snahou bylo vytvořit model, který by efektivně a s dostatečnou spolehlivostí odhadoval hodnoty závisle proměnné.

Závisle proměnná je v modelu značena jako **li**. Představuje lokus, který je v daném cyklu programu testován. Výpočet probíhá iterativně a poloha závisle proměnné se každým cyklem posouvá ve směru zleva doprava.

Nezávisle proměnné jsou rozděleny do dvou skupin podle polohy k testovanému lokusu. Proměnné **l1 – l50** představují lokusy, které se nacházejí na pravé straně od testovaného lokusu podle pravidla **l1 = li + 1, l2= li+2… l50= li+ 50**. Proměnné **ll1 – ll50** představují proměnné, které se nacházejí na levé straně od testované lokusu podle pravidla **ll1= li – 1, ll2= li- 2 … ll50= li – 50**.

Největší model představuje testování závisle proměnné při okolí 100 lokusů. Jedná se tedy o 50 lokusů z levé a 50 lokusů z pravé strany.

Obecně lze statistický model s **j** sousedními lokusy zapsat z výše uvedeného maticového zápisu jako:

$$li = \mu + ll_{(i-j)} + ll_{(i-j+1)} + \ldots + ll_{(i-1)} + l_{(i+1)} + \ldots + l_{(I+j-1)} + l_{(i+j)} + \varepsilon,$$

Kde:

li = sledovaný lokus

μ = celkový průměr

i = číslo sledovaného lokusu

j = počet sousedních lokusů na jedné straně

ll= sousední lokus zleva

l = sousední lokus zprava

ε = chyba

Postup byl ověřen na zvolených lokusech. Bylo vybráno celkem šest lokusů, které se nacházejí na prvním chromozomu. Tři lokusy ze souboru A a tři lokusy ze souboru B. Tyto lokusy byly náhodně vybrány na základě procentuálního zastoupení alely A, respektive průměrné hodnoty lokusu.

Soubor A:

Lokus 201 s 50% zastoupením alely A a průměrnou hodnotou lokusu 1.05 (heterozygotní lokus).

Lokus 716 s 75% zastoupením alely A a průměrnou hodnotou lokusu 1.5.

Lokus 133 s 95% zastoupením alely A a průměrnou hodnotou lokusu 1.9 (téměř homozygotní lokus).

Soubor B:

Lokus 760 s téměř 50% zastoupením alely A a průměrnou hodnotou lokusu 1.04.

Lokus 893 se 75% zastoupením alely A a průměrnou hodnotou lokusu 1.5.

Lokus 201 s 95% zastoupením alely A a průměrnou hodnotou lokusu 1.9.

V případě lokusu 201, který je zastoupen v souboru A i B je nutné podotknout, že se nejedná o stejný lokus. Při úpravě dat, které vstupují do modelů, dochází k úpravám a třídění. Čísla lokusů vznikají až po této úpravě. Jedná se o vnitřní označení lokusů uvnitř souboru. Pro přípravu vstupních údajů a výpočet byly připraveny programy v prostředí SAS (2002) – příloha 1 a 2.

5 Výsledky

5.1 Výsledky pro soubor A

Z testování modelů je patrné, že předpověď SNP byla nejúspěšnější u heterozygotních lokusů, které měly 50% zastoupení alely A. Stačilo pouze 50 sousedních lokusů pro téměř přesnou předpověď SNP při spolehlivosti 100 % (lokus č. 201).

Při klesajícím zastoupení alely A, bylo nutné okolí rozšířit. U lokusů se 75% (lokus č. 716) zastoupením A se podařilo dosáhnout stejného výsledku při dvojnásobném zvětšení okolí, tedy 100 sousedních lokusů, při spolehlivosti 99 %.

U téměř homozygotního lokusu (lokus č. 133) se zastoupením alely A 95%, bylo posledním modelem dosaženo spolehlivosti pouze 56 %. Odhady SNP kolísají kolem skutečných hodnot. Tabulka č. 3 představuje číselný souhrn modelů. Sledovali jsme spolehlivost, chybu předpovědi a předpověď jednotlivých SNP.

Tabulka č. 3 Výsledky sledovaných hodnot pro soubor A.

Počet sousedních lokusů		**10**	**20**	**30**	**40**	**50**	**60**	**70**	**100**
Lokus 201 – 50% alely A									
Spolehlivost		0.985	0.989	0.997	0.998	1	1	1	1
Průměrná abs. chyba		0.040	0.032	0.018	0.014	4.1E-14	1.4E-13	1.6E-13	3.9E-13
Maximální chyba		0.786	0.589	0.196	0.129	3.5E-13	1.4E-12	1.4E-12	4.0E-12
Hodnota SNP	**min**	0	0	0	0	0	0	0	0
	max	2	2	2	2	2	2	2	2
Předpověď SNP	**min**	0.0148	-0.006	-0.008	-0.004	-1.2E-13	-3.5E-13	-3.3E-13	-3.0E-13
	max	2.074	2.151	2.071	2.102	2	2	2	2
Lokus 716 – 75% alely A									
Spolehlivost		0.623	0.847	0.917	0.954	0.983	0.990	0.996	0.999
Průměrná abs. chyba		0.282	0.180	0.117	0.101	0.051	0.038	0.0243	0.014
Maximální chyba		1.453	0.880	0.801	0.568	0.325	0.254	0.197	0.064
Hodnota SNP	**min**	0	0	0	0	0	0	0	0
	max	2	2	2	2	2	2	2	2
Předpověď SNP	**min**	0.226	-0.109	-0.089	-0.068	-0.019	-0.024	-0.022	-0.022
	max	2.138	2.639	2.238	2.234	2.197	2.208	2.196	2.062

Lokus 133 – 95% alely A									
Spolehlivost		0.087	0.243	0.327	0.369	0.423	0.499	0.528	0.558
Průměrná abs. chyba		0.124	0.121	0.122	0.119	0.114	0.105	0.099	0.098
Maximální chyba		0.992	0.942	0.855	0.819	0.797	0.740	0.754	0.729
Hodnota SNP	**min**	1	1	1	1	1	1	1	1
	max	2	2	2	2	2	2	2	2
Předpověď SNP	**min**	1.729	1.359	1.163	1.037	1.050	0.969	1	1
	max	2.010	2.108	2.219	2.173	2.231	2.222	2.259	2.219

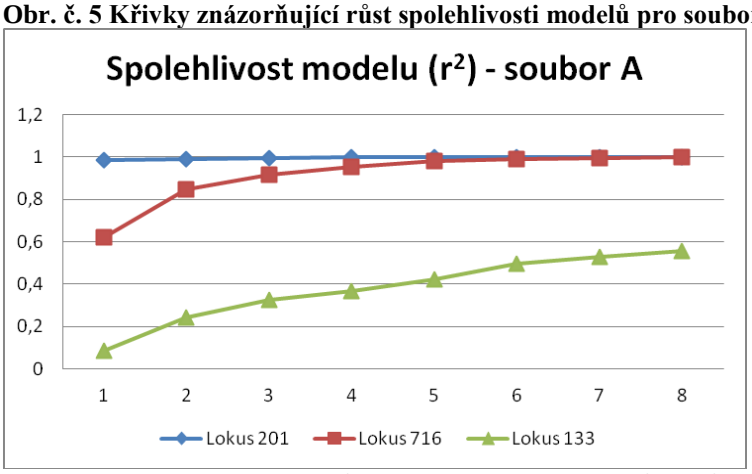
Lokus 201 nabýval skutečných hodnot 0, 1, 2. Spolehlivost modelu se při testování pohybovala mezi 98 % a 100 %. Stoprocentní spolehlivost modelu byla pozorována u testování 50 sousedních lokusů. Průměrná absolutní chyba u prvního modelu byla 0.04 a měla klesající tendenci, u 50 sousedních lokusů nabývala téměř nulové hodnoty. Maximální chyba v prvním modelu dosahovala 0.786 a stejně tak jako u absolutní chyby dosahovala téměř nuly u modelu s 50 sousedními lokusy. Minimální skutečná hodnota lokusu byla 0, minimální hotnota předpovědí nabývala hodnot 0.0148 až -3.0E-13. U maximální hodnoty, která byla 2, se předpověď pohybovala mezi 2 až 2.15. Jakým způsobem se předpovídané hodnoty shodovaly se skutečnými lze pozorovat z obrázků č. 6, 9, 12.

Lokus 716 nabýval skutečných hodnot 0, 1, 2. Spolehlivost modelů se zde pohybovala mezi 62 % až 99 %, z tabulky č. 3 je tedy patrné, že bylo potřeba dvojnásobné množství sousedních lokusů než v prvním případě, aby spolehlivost modelu dosáhla 100 %. Průměrná absolutní chyba se v průběhu testování pohybovala mezi hodnotami 0.282 až 0.014. Maximální chyba nabývala hodnot 1.453 až 0.064. Minimální hodnota předpovědi byla 0.226 až -0.022. Maximální hodnoty předpovědi se pohybovaly v rozpětí 2.639 až 2.062. Shoda předpovědi se skutečnými hodnotami pro lokus 716 je znázorněna na obrázcích č. 7, 10, 13.

U téměř homozygotního lokusu 133 spolehlivost modelu vzrostla z 8,7 % na konečných 56%, při průměrné absolutní chybě 0.124 až 0.098. Maximální chyba se u tohoto lokusu v průběhu testování pohybovala mezi 0.992 až 0.729. Skutečné hodnoty v lokusu 133 nabývaly hodnot pouze 1 a 2. Minimální hodnota předpovědi z hodnoty 1.729 v průběhu testování klesala až 0.969, u modelu se 70 a 100 sousedními lokusy nabývala hodnot 1. Maximální hodnota předpovědi měla kolísavou tendenci a pohybovala se v rozpětí 2.010 až 2.219. Shoda předpovědi se skutečnou hodnotou lokusu 133 znázorňují obrázky č. 8, 11, 14.

Vývoj spolehlivostí modelů pro všechny hodnocené lokusy znázorňuje obrázek č. 5.

Obr. č. 5 Křivky znázorňující růst spolehlivosti modelů pro soubor A



U všech hodnocených lokusů má spolehlivost modelů vzrůstající tendenci. U lokusů 201 a 716 se spolehlivost vyšplhala k hodnotě 1, tedy 100 %. U lokusu 133 se podařilo v modelu se 100 sousedními lokusy dosáhnout spolehlivosti 0,55, tedy 55 %.

V tabulkách č. 5, 6 a7 můžeme vidět vývoj regresních koeficientů při rozšiřování modelu. Jedná se o prvních 10 sousedních lokusů, které jsou v každém modelu zastoupeny.

Hodnota jednotlivých regresních koeficientů reaguje na celkový počet lokusů zahrnutých do analýzy. U uniformních lokusů (např. lokus 128 nebo lokus 130 v tabulce č. 7) je hodnota regresních koeficientů nulová od první předpovědi, tyto lokusy se tedy nezapojují do předpovědi hodnot zkoumaného lokusu.

Během testování modelů je také zřejmé, že při zahrnutí většího množství sousedních lokusů může dojít ke ztrátě důležitosti sousedního lokusu (lokus 196, 198 atd. tabulka č. 5) a hodnota regresního koeficientu je poté nulová. Odhadnuté hodnoty regresních koeficientů je třeba hodnotit jako celkovou hodnotu předpovědi pro daný lokus, a proto změny v jednotlivých koeficientech jsou očekávatelné.

Tabulka č. 5 Vývoj regresních koeficientů prvních 10 sousedních lokusů pro hodnocený lokus 201

Lokus 201	10 l	20 l	30 l	40 l	50 l	60 l	70 l	100 l
Spolehlivost	0.985	0.989	0.997	0.998	1	1	1	1
Lokus 196	0.045	-0.094	0.028	0.253	0	-1	-1	-1
Lokus 197	0.182	0.539	0.999	1.239	0	3	3	3
Lokus 198	-0.860	-0.496	-0.218	-0.400	-1	0	0	0
Lokus 199	0	0	0	0	0	0	0	0
Lokus 200	-0.908	-0.646	-0.703	-0.294	-1	0	0	0
Lokus 202	-0.991	-0.818	-0.894	-2.091	0	0	0	0
Lokus 203	-0.190	-0.257	-0.732	-0.784	0	0	0	0
Lokus 204	-0.146	-0.265	-0.395	-1.074	0	0	0	0
Lokus 205	0	0	0	0	0	0	0	0
Lokus 206	0.028	-0.143	-0.04	0	0	0	0	0

Tabulka č. 6 Vývoj regresních koeficientů prvních 10 sousedních lokusů pro hodnocený lokus 716

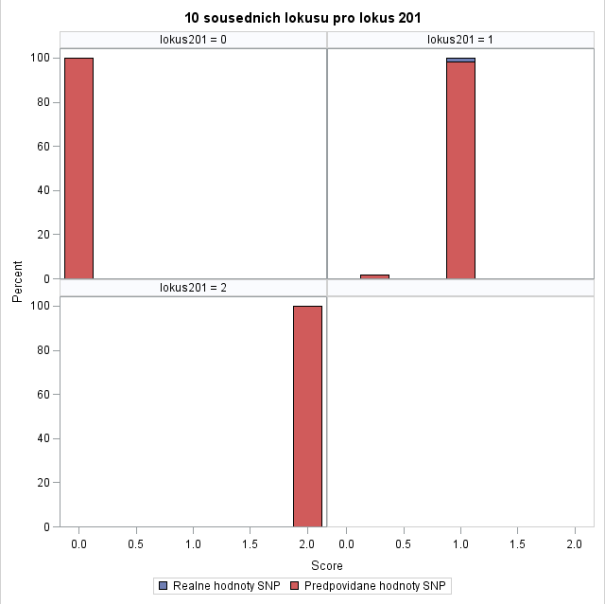
Lokus 716	10 l	20 l	30 l	40 l	50 l	60 l	70 l	100 l
Spolehlivost	0.623	0.847	0.917	0.954	0.983	0.990	0.996	0.999
Lokus 711	0	0	0	0	0	0	0	0
Lokus 712	0.038	0.004	- 0.016	-0.016	0.001	0.011	0.007	0.005
Lokus 713	-0.013	0.122	0.115	-0.186	0.042	0.118	0.417	0.292
Lokus 714	-0.189	-0.219	-0.229	-0.122	-0.404	-0.410	-0.147	1.102
Lokus 715	-0.071	-0.288	-0.367	-0.299	0.467	-0.031	-4.179	3.979
Lokus 717	-0.099	-0.293	0.168	0.199	0.004	0.120	0.173	1.699
Lokus 718	0.748	0.913	0.789	0.638	1.519	0.760	-3.357	4.382
Lokus 719	0	0	0	0	0	0	0	0
Lokus 720	0.028	-0.695	-0.657	-0.447	-0.352	-0.346	-0.202	-1.299
Lokus 721	0.243	-0.590	-0.336	-0.291	-0.553	-0.516	-0.514	-0.163

Tabulka č. 7 Vývoj regresních koeficientů prvních 10 sousedních lokusů pro hodnocený lokus 133

Lokus 133	10 l	20 l	30 l	40 l	50 l	60 l	70 l	100 l
Spolehlivost	0.087	0.243	0.327	0.369	0.423	0.499	0.528	0.558
Lokus 128	0	0	0	0	0	0	0	0
Lokus 129	-0.108	0.005	0.156	0.015	0.020	1.992	2.562	0
Lokus 130	0	0	0	0	0	0	0	0
Lokus 131	-0.024	0.582	0.511	0.029	1.298	3.168	4.725	1.109
Lokus 132	0	0	0	0	0	0	0	0
Lokus 134	0.065	-0.345	-0.481	0.061	-0.619	-1.858	-2.296	0
Lokus 135	0	0	0	0	0	0	0	0
Lokus 136	0.088	0.506	0.070	0.093	0.162	-1.058	-1.679	-0.403
Lokus 137	-0.022	-0.234	-0.104	0.118	0.109	-0.179	0.202	0.318
Lokus 138	0	0	0	0	0	0	0	0

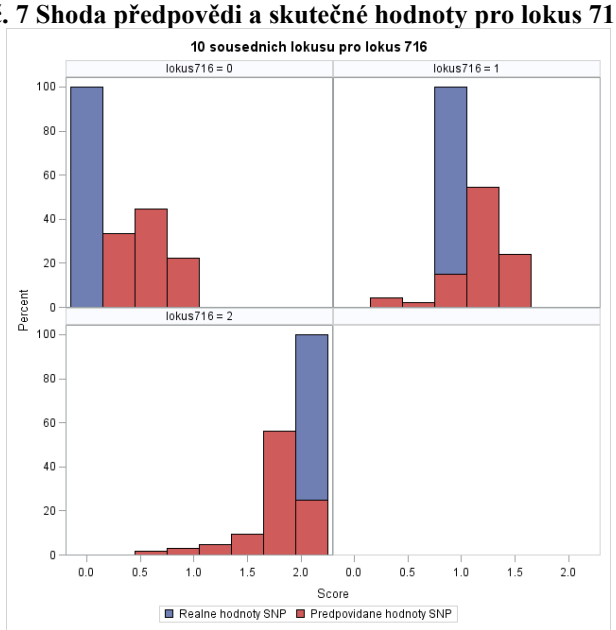
Grafy na obrázcích č. 6 až 14 představují procentuální shodu skutečných a předpovězených hodnot pro každou hodnotu sledovaného lokusu při použití modelu pro 10, 50, a 100 sousedních lokusů. Skutečnou hodnotu lokusu představuje modrý sloupec, Předpovídané hodnoty červený sloupec. Shoda předpovědi a skutečné hodnoty je znázorněna překrytím modrého sloupce červeným.

Obr. č. 6 Shoda předpovědi a skutečné hodnoty pro lokus 201 v 1. modelu



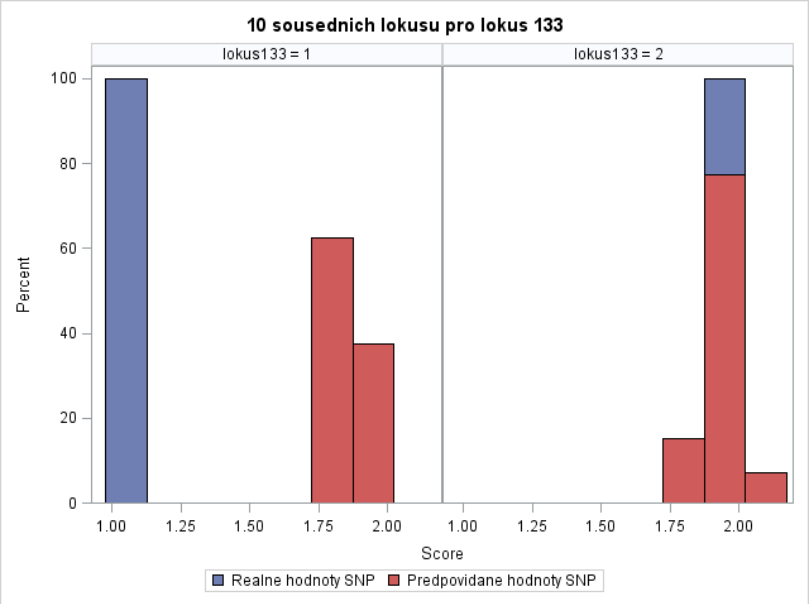
Na obrázku můžeme vidět graf pro každou hodnotu sledovaného lokusu a její shodu s předpovědi této hodnoty v %. U lokusu 201 se předpověď 100 % shoduje v hodnotách 0 a 2, u hodnoty 1 se předpověď shoduje téměř na 100 %.

Obr. č. 7 Shoda předpovědi a skutečné hodnoty pro lokus 716 v 1. modelu



Pro hodnotu 0 se předpověď blíží skutečné hodnotě, ale neshoduje se. U hodnoty 1 se shoduje zhruba 15 %, zbytek předpovězených hodnot je v blízkosti hodnoty skutečné v intervalu od 0.02 do 1.6. U hodnoty 2 se předpověď shoduje přibližně 20 % , ostatní předpovědi jsou v intervalu od 0.04 do 2.

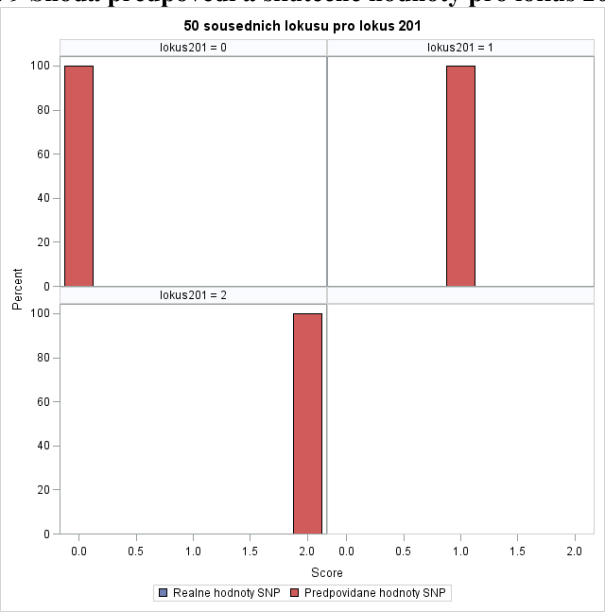
Obr. č. 8 Shoda předpovědi a skutečné hodnoty pro lokus 133 v 1. modelu



U téměř homozygotního lokusu 133 se pro hodnotu 1 předpověď neshoduje se skutečnou hodnotou. Předpovězené hodnoty se nacházejí v intervalu 1.70 až 2. U hodnoty 2 se shoda blíží 80 % zbylé předpovězené hodnoty se pohybují v intervalu 1.70 a 2.13.

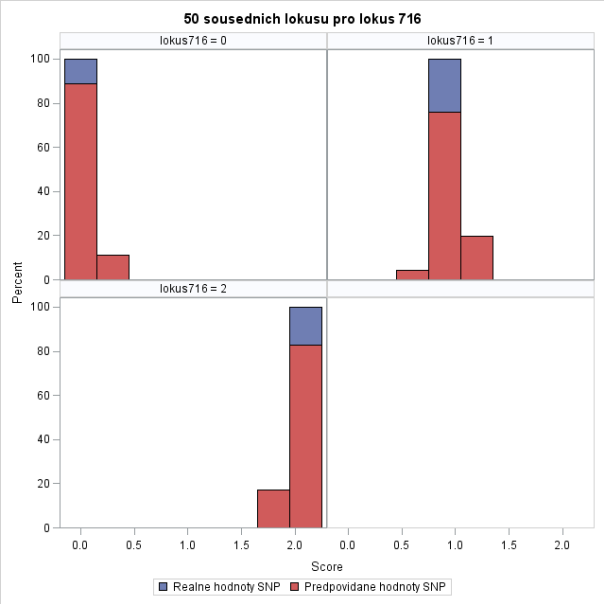
U prvního modelu je z grafů patrné, že nejlepší předpověď hodnot vychází u lokusu 201, kde již při nízkém počtu sousedních lokusů se předpověď hodnoty téměř rovná skutečné hodnoty lokusu.

Obr. č. 9 Shoda předpovědi a skutečné hodnoty pro lokus 201 v 5. modelu



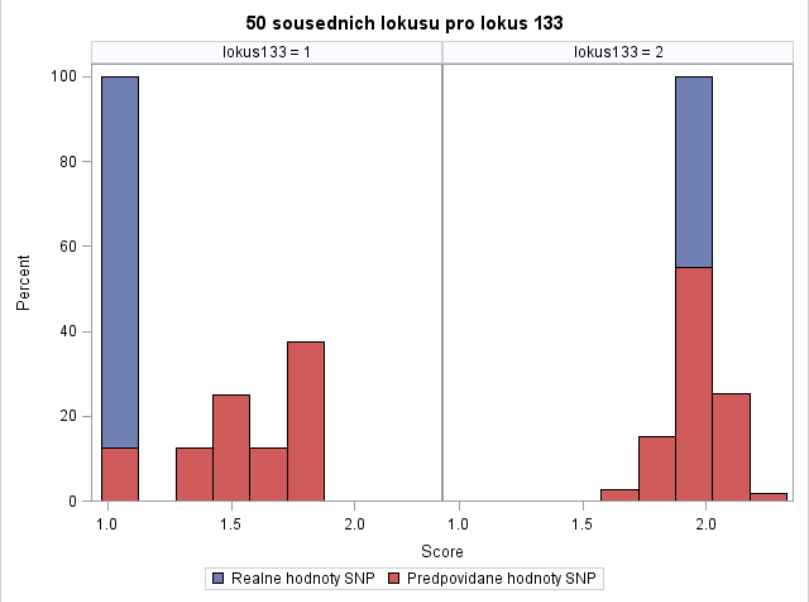
V modelu pro 50 sousedních lokusů se podařilo dosáhnout shody předpovědi 100 % pro všechny hodnoty sledovaného lokusu 201

Obr. č. 10 Shoda předpovědi a skutečné hodnoty pro lokus 716 v 5. modelu



U lokusu se 75% zastoupením alely A se u hodnoty 0 předpověď shoduje zhruba na 90 % zbylých 10 % se pohybuje v intervalu 0 až 0.5. Pro hodnotu 1 se shoda rovná 80 % zbytek hodnot se pohybuje v intervalu 0.5 až 1.5. Předpověď pro hodnotu 2 se se skutečnou hodnotou lokusu shoduje na 80% , ostatních 20 % leží v intervalu 1.6 až 2.

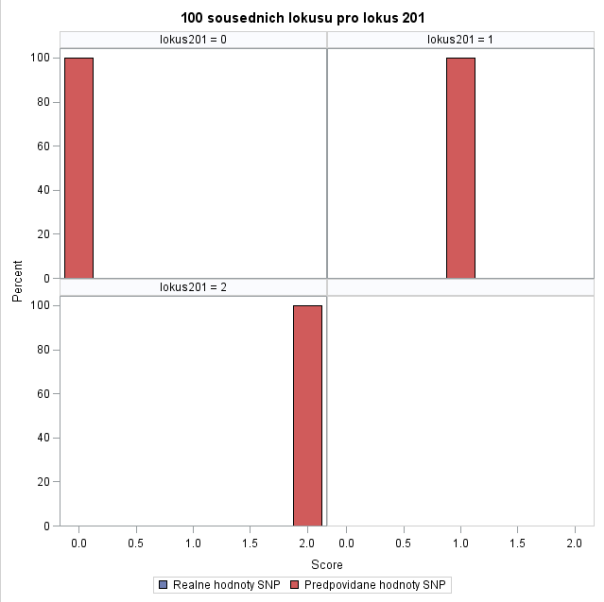
Obr. č. 11 Shoda předpovědi a skutečné hodnoty pro lokus 133 v 5. modelu



U lokusu 133 pro hodnotu 1 nastalo zlepšení, předpověď se shoduje zhruba na 15 %. Zbylé hodnoty se oproti prvnímu přiblížily hodnotě skutečné a nacházejí se v intervalu 1.25 až 1.75. Shoda u hodnoty 2 je přes 50% zbytek předpovězených hodnot je v intervalu 1.6 až 2.5.

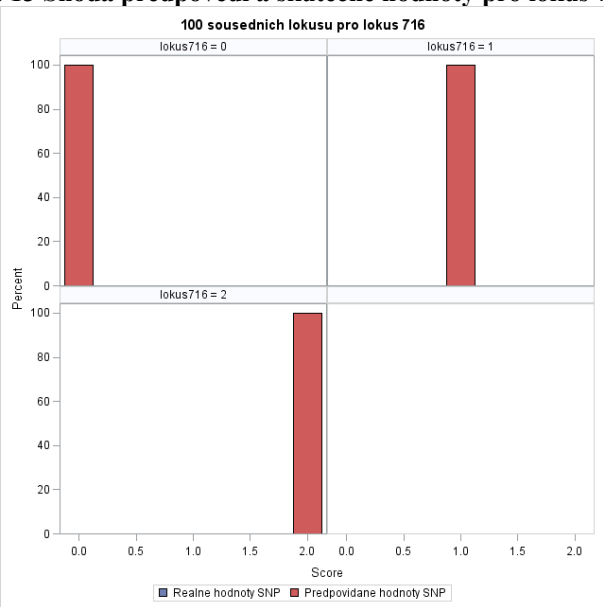
Při rozšíření modelu na 50 sousedních lokusů se v případě lokusu 201 předpověď zcela shoduje se skutečnou hodnotou lokusu. Stav se také výrazně zlepšil u lokusu 716. Předpovídané hodnoty se z větší části shodují se skutečnými hodnotami, nebo se jim velmi blíží. Mírné zlepšení lze pozorovat také u lokusu 133, kde se předpovědi blíží skutečné hodnotě lokusu.

Obr. č. 12 Shoda předpovědi a skutečné hodnoty pro lokus 201 v 8. modelu



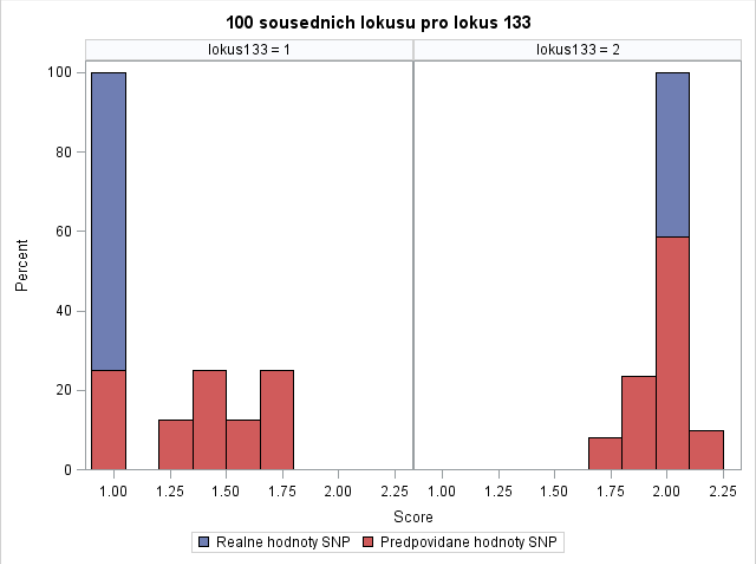
Shoda předpovědi při zahrnutí 100 sousedních lokusů se skutečnými hodnotami je ve všech případech 100%.

Obr. č. 13 Shoda předpovědi a skutečné hodnoty pro lokus 716 v 8. modelu



Shoda předpovědi se skutečnými hodnotami je ve všech případech 100%.

Obr. č. 14 Shoda předpovědi a skutečné hodnoty pro lokus 133 v 8. modelu



Pro hodnotu 1 je shoda předpovědi přes 20% zbylé hodnoty se nacházejí v intervalu od 1. 2 do 1.8. U hodnoty 2 shoda předpovědi se skutečnou hodnotou dosáhla zhruba 60%. Zbylé předpovězené hodnoty se nacházejí v intervalu od 1.6 do 2.25.

V posledním modelu, ve kterém je testováno 100 sousedních lokusů, lze pozorovat shodu předpovědi se skutečnou hodnotou u lokusu 716. U téměř homozygotního lokusu 133 lze pozorovat mírné zlepšení předpovědi.

5.2 Výsledky pro soubor B

Tabulka č. 8 Výsledky sledovaných hodnot pro soubor A.

Počet sousedních lokusů		**10**	**20**	**30**	**40**	**50**	**60**	**70**	**100**
Lokus 760 – 50% alely A									
Spolehlivost		0.466	0.666	0.764	0.796	0.822	0.845	0.852	0.878
Průměrná abs. chyba		0.402	0.280	0.213	0.199	0.190	0.176	0.171	0.156
Maximální chyba		1.742	1.834	2.067	1.904	1.659	1.765	1.543	1.407
Hodnota SNP	**min**	0	0	0	0	0	0	0	0
	max	2	2	2	2	2	2	2	2
Předpověď SNP	**min**	0.387	0.219	0.192	0.096	0.084	0.057	0.017	-0.068
	max	2.144	2.372	2.402	2.568	2.461	2.456	2.505	2.322
Lokus 893 – 75% alely A									
Spolehlivost		0.900	0.922	0.943	0.947	0.952	0.954	0.957	0.964
Průměrná abs. chyba		0.099	0.090	0.078	0.076	0.072	0.073	0.071	0.063
Maximální chyba		1.033	1.059	1.129	1.174	1.069	1.050	0.940	0.995
Hodnota SNP	**min**	0	0	0	0	0	0	0	0
	max	2	2	2	2	2	2	2	2
Předpověď SNP	**min**	-0.221	-0.372	-0.606	-0.586	-0.659	-0.765	-0.711	-0.407
	max	2.489	2.764	3.346	3.188	2.766	2.779	2.764	2.819
Lokus 201 – 95% alely A									
Spolehlivost		0.407	0.680	0.703	0.764	0.780	0.787	0.804	0.856
Průměrná abs. chyba		0.130	0.098	0.092	0.077	0.075	0.074	0.071	0.067
Maximální chyba		1.415	1.347	1.392	1.389	1.228	1.141	1.106	0.964
Hodnota SNP	**min**	0	0	0	0	0	0	0	0
	max	2	2	2	2	2	2	2	2
Předpověď SNP	**min**	0.387	0.219	0.192	0.096	0.084	0.057	0.017	-0.068
	max	2.144	2.372	2.402	2.568	2.461	2.456	2.505	2.322

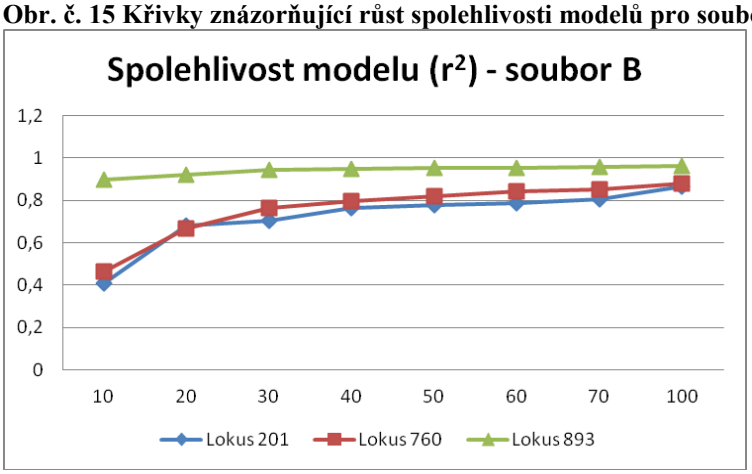
Výsledky výpočtů pro soubor B jsou shrnuty v tabulce č. 8. U lokusu 760 s 50% zastoupením alely A se spolehlivost modelů pohybovala mezi 47 % až 88 %. Hodnoty průměrné absolutní chyby klesaly z 0.402 až na 0.156, zatímco maximální chyba se v modelech od 20 do 30 sousedních lokusů zvýšila z hodnoty 1.742 až na hodnotu 2.067 a následně klesla až na hodnotu 1.407. Minimální hodnota předpovědi se pohybovala v intervalu od 0.387 do -0.068. Maximální hodnota předpovědi se do modelu se 40 sousedními lokusy zvyšovala z 2.144 na hodnotu 2.568 a následně klesla až na hodnotu 2.322. Shoda předpovědí se skutečnými hodnotami je patrná z obrázků č. 16, 19 a 22.

Lokus 893 vyšel, co se týče spolehlivosti modelů, nejlépe. Spolehlivost modelů se pohybovala mezi 90 % až 96 %, při průměrné absolutní chybě od 0.099 do 0.063 a maximální chybě od 1.033 do 0.995. Opět bylo zaznamenáno zvýšení hodnot maximální chyby v modelech od 20 do 40 sousedních lokusů z hodnoty 1.033 na hodnotu 1.174. Minimální hodnoty předpovědi kolísaly mezi hodnotami -0.221 a -0.765 s konečnou hodnotou -0.407. Stejně tak byl zaznamenán nárůst maximálních hodnot předpovědi u modelů s 20 až 30 sousedními lokusy z hodnoty 2.489 až na hodnotu 3.346 s následným poklesem na konečnou hodnotu 2.819. Shoda předpovědí se skutečnými hodnotami je patrná z obrázků č. 17, 20 a 23.

U téměř homozygotního lokusu 201 bylo dosaženo spolehlivosti mezi 41 – 88 % s průměrnou absolutní chybou, která se pohybovala v rozmezí 0.130 až 0.067. Maximální chyba opět kolísala mezi hodnotami 1.415 až 0.964. Minimální hodnota předpovědi v průběhu testování klesla z hodnoty 0.387 až na hodnotu -0.068. Maximální hodnota předpovědi se v modelech od 20 do 40 sousedních lokusů zvýšila z 2.144 na hodnotu 2.568, u 50 až 70 sousedních lokusech se objevilo další zvýšení hodnot z 2.461 na 2.505 a u 100 sousedních lokusů klesla na konečnou hodnotu 2.322. Shoda předpovědí se skutečnými hodnotami je patrná z obrázků č. 18, 21 a 24.

Porovnání růstu spolehlivostí napříč modely u všech sledovaných lokusů je znázorněno na obrázku č. 15.

Obr. č. 15 Křivky znázorňující růst spolehlivosti modelů pro soubor B



U souboru B je, co se týče spolehlivosti modelů, dosaženo vyrovnanějších hodnot. Téměř vyrovnané hodnoty se nacházejí u posledního modelu se 100 sousedními lokusy.

V tabulkách č. 9, 10 a 11 můžeme pozorovat změny regresních koeficientů pro prvních 10 sousedních lokusů, které se pro daný sledovaný lokus vyskytují ve všech modelech.

Tabulka č. 9 Vývoj regresních koeficientů prvních 10 sousedních lokusů pro hodnocený lokus 760

Lokus 760	10 l	20 l	30 l	40 l	50 l	60 l	70 l	100 l
Spolehlivost	0.466	0.666	0.764	0.796	0.822	0.845	0.852	0.878
Lokus 755	-0.153	0.066	0.024	0.008	-0.044	-0.035	-0.024	-0.015
Lokus 756	0.200	0.164	-0.300	-0.171	-0.178	-0.163	-0.162	-0.119
Lokus 757	-0.010	0.052	0.210	0.231	0.156	0.105	0.155	0.177
Lokus 758	-0.123	0.045	0.028	0.027	0.035	0.052	0.061	0.117
Lokus 759	-0.051	0.044	-0.036	-0.066	-0.026	0.066	0.152	0.171
Lokus 761	0.203	-0.055	0.547	0.545	0.275	0.341	0.330	0.393
Lokus 762	0.472	0.251	0.204	0.279	0.207	0.077	0.085	0.031
Lokus 763	0.101	0.440	0.257	0.154	0.112	0.119	0.126	0.084
Lokus 764	-0.355	-0.213	-0.223	-0.256	-0.237	-0.048	-0.088	-0.129
Lokus 765	0.425	-0.229	-0.111	-0.140	-0.138	-0.238	-0.203	-0.095

Tabulka č. 10 Vývoj regresních koeficientů prvních 10 sousedních lokusů pro hodnocený lokus 893

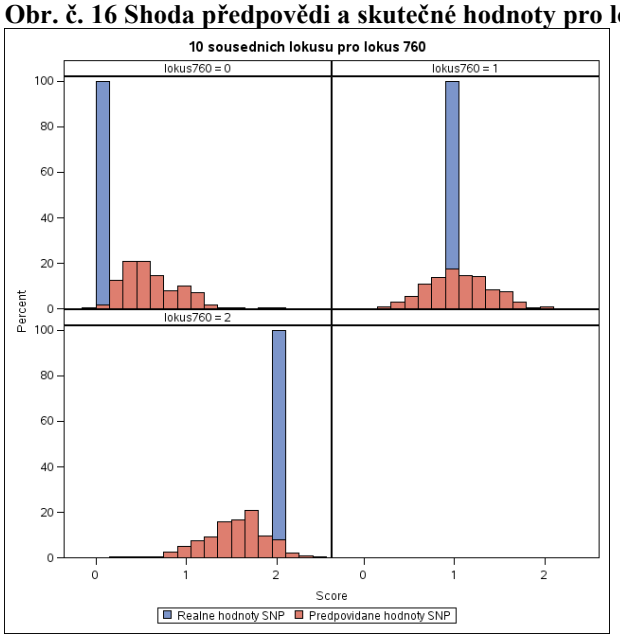
Lokus 893	10 l	20 l	30 l	40 l	50 l	60 l	70 l	100 l
Spolehlivost	0.900	0.922	0.943	0.947	0.952	0.954	0.957	0.964
Lokus 888	-0.030	-0.092	-0.101	-0.101	-0.114	-0.125	-0.113	-0.092
Lokus 889	-0.368	-0.347	-0.325	-0.320	-0.287	-0.279	-0.269	-0.237
Lokus 890	0.315	0.241	0.241	0.226	0.205	0.209	0.194	0.196
Lokus 891	-0.021	-0.104	-0.120	-0.151	-0.147	-0.127	-0.116	-0.115
Lokus 892	0.545	0.723	0.756	0.722	0.733	0.737	0.685	0.586
Lokus 894	0.432	0.415	0.546	0.563	0.513	0.511	0.497	0.433
Lokus 895	-0.541	-0.432	-0.334	-0.347	-0.356	-0.354	-0.386	-0.411
Lokus 896	-0.524	-0.423	-0.356	-0.349	-0.313	-0.320	-0.343	-0.397
Lokus 897	-0.029	-0.024	-0.128	-0.133	-0.158	-0.161	-0.206	-0.194
Lokus 898	-0.006	-0.007	0.006	0.005	0.007	-0.006	-0.001	-0.006

Tabulka č. 11 Vývoj regresních koeficientů prvních 10 sousedních lokusů pro hodnocený lokus 201

Lokus 201	10 l	20 l	30 l	40 l	50 l	60 l	70 l	100 l
Spolehlivost	0.407	0.680	0.703	0.764	0.780	0.787	0.804	0.856
Lokus 196	-0.074	-0.113	-0.173	-0.138	-0.113	-0.109	-0.125	-0.120
Lokus 197	-0.067	-0.056	-0.061	0.035	0.032	0.009	-0.038	-0.029
Lokus 198	-0.005	-0.027	0.011	-0.056	-0.116	-0.112	-0.092	-0.111
Lokus 199	-0.159	-0.225	-0.199	-0.233	-0.208	-0.212	-0.211	-0.163
Lokus 200	0.488	0.380	0.308	0.269	0.244	0.202	0.121	0.184
Lokus 202	-0.733	-0.453	-0.435	-0.319	-0.247	-0.238	-0.234	-0.220
Lokus 203	-0.684	-0.402	-0.399	-0.280	-0.265	-0.262	-0.243	-0.206
Lokus 204	0.132	0.406	0.483	0.573	0.605	0.609	-0.625	0.614
Lokus 205	-0.0002	-0.127	-0.167	-0.249	-0.206	-0.226	0.213	-0.256
Lokus 206	0.067	0.169	0.171	0.268	0.253	0.288	-0.248	0.294

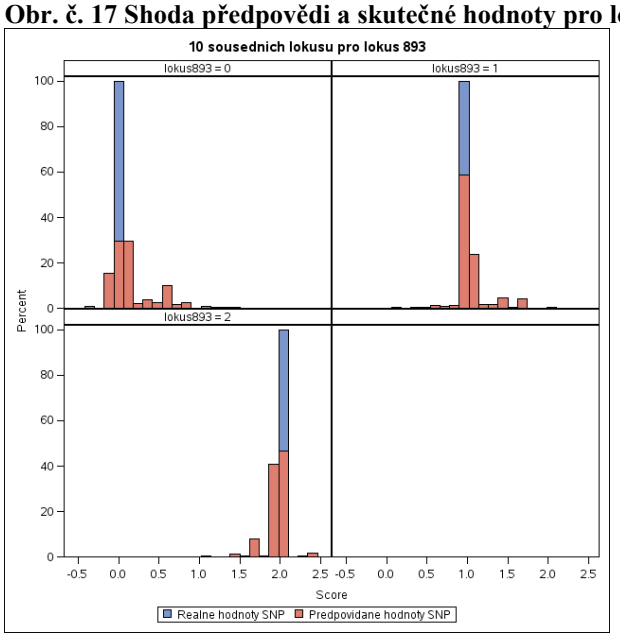
Shodu předpovídaných a skutečných hodnot představují obrázky č. 16 až 24, stejně tak, jako v případě souboru A.

Obr. č. 16 Shoda předpovědi a skutečné hodnoty pro lokus 760 v 1. modelu



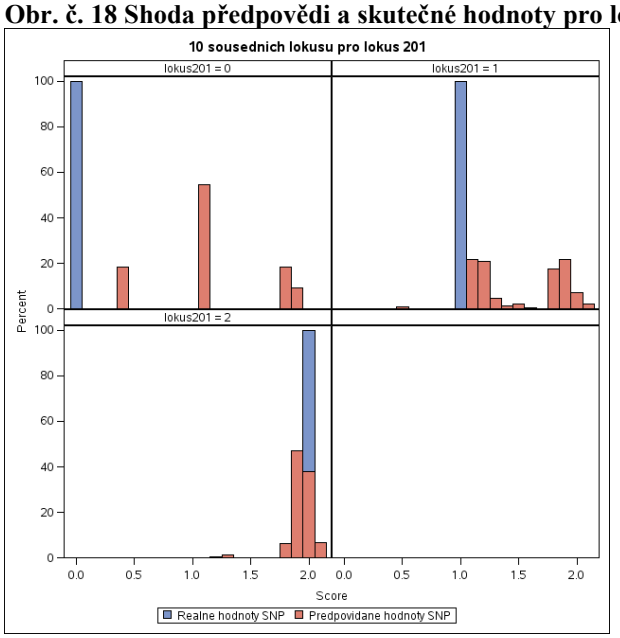
Shoda předpovědi u hodnoty nula je kolem 2 %, u hodnoty 1 se shoduje zhruba 16 % hodnot a při hodnotě 2 se shoda předpovědi se skutečnou hodnotou shoduje zhruba 10 %

Obr. č. 17 Shoda předpovědi a skutečné hodnoty pro lokus 893 v 1. modelu



Shoda předpovědi u hodnoty 0 se pohybuje kolem 30 % u hodnoty 1 téměř 60 % a u hodnoty 2 dosahuje téměř 45 % shody předpovídaných hodnot.

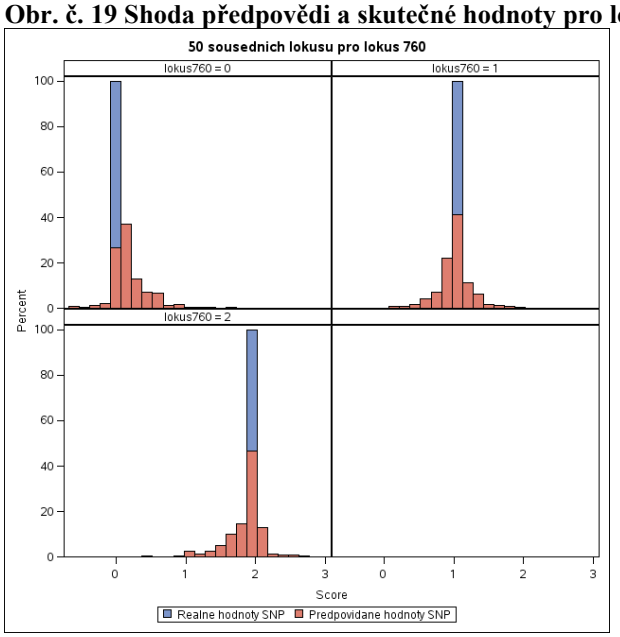
Obr. č. 18 Shoda předpovědi a skutečné hodnoty pro lokus 201 v 1. modelu



U hodnot 0 a 1 se nepodařilo dosáhnout shody předpovědi. U hodnoty 2 se shoda pohybuje kolem 38 %.

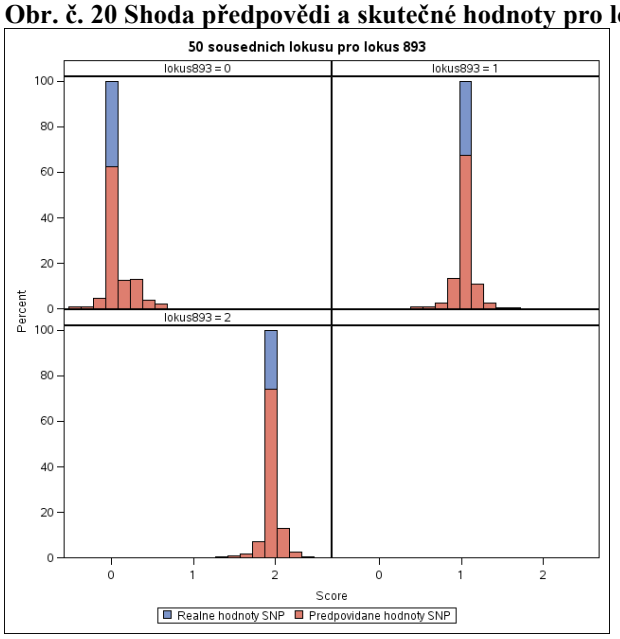
Stejně tak jako u testování souboru A, je z grafů patrné, že shoda předpovědi se skutečnou hodnotou SNP vychází stejně jako v prvním případě lépe u lokusů se zastoupením alely A 50% až 75 %.

Obr. č. 19 Shoda předpovědi a skutečné hodnoty pro lokus 760 v 5. modelu



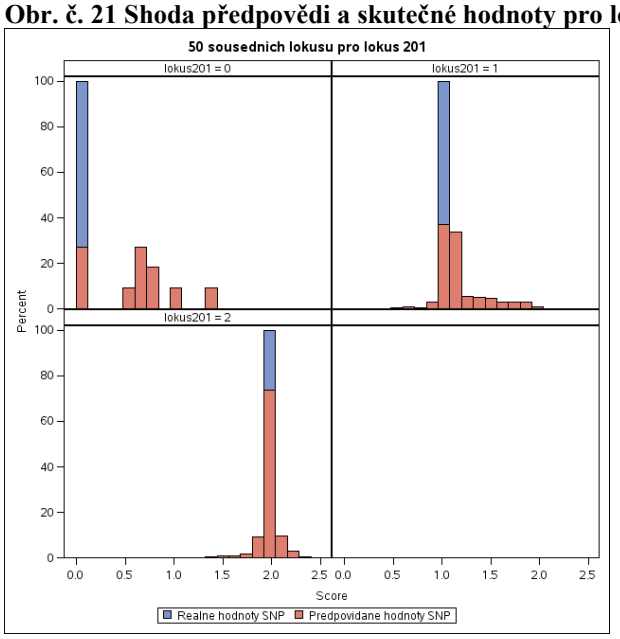
Pro hodnotu 0 se shoda předpovědi pohybuje kolem 25 %, pro hodnotu 1 40 % a pro hodnotu 2 dosahuje shoda předpovědi téměř 45 %.

Obr. č. 20 Shoda předpovědi a skutečné hodnoty pro lokus 893 v 5. modelu



Pro hodnoty 0 a 1 je shoda předpovědi se skutečnou hodnotou lokusu 60 a 65 %. U hodnoty 2 dosahuje až 75 %.

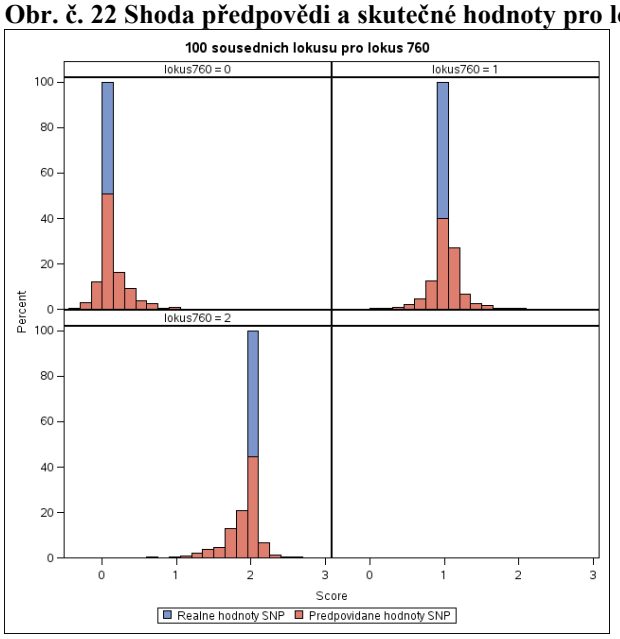
Obr. č. 21 Shoda předpovědi a skutečné hodnoty pro lokus 201 v 5. modelu



U hodnoty 0 se shoda předpovědi zvýšila oproti prvnímu modelu na 25 %, výrazný nárůst se projevil také u hodnoty 1, kde se shoduje zhruba 35 % hodnot. U hodnoty 2 se shoduje zhruba 70 % hodnot.

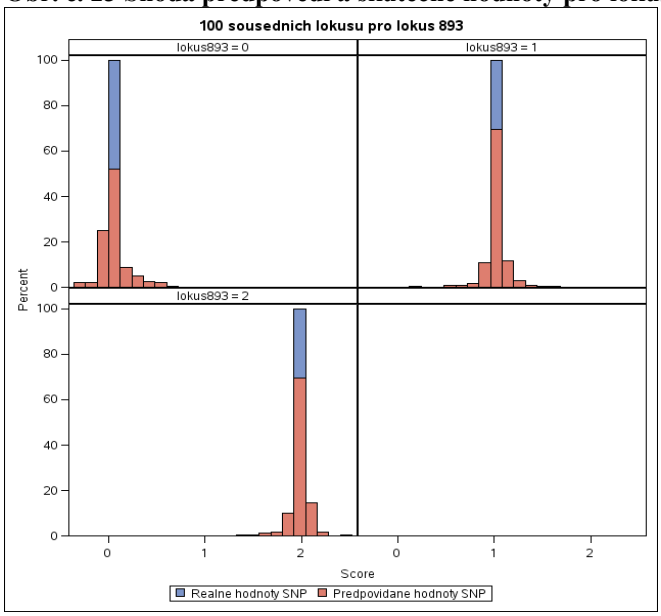
Při testování 50 sousedních lokusů lze pozorovat, že hodnoty předpovědi se u všech lokusů pohybují blízko skutečné hodnotě.

Obr. č. 22 Shoda předpovědi a skutečné hodnoty pro lokus 760 v 8. modelu



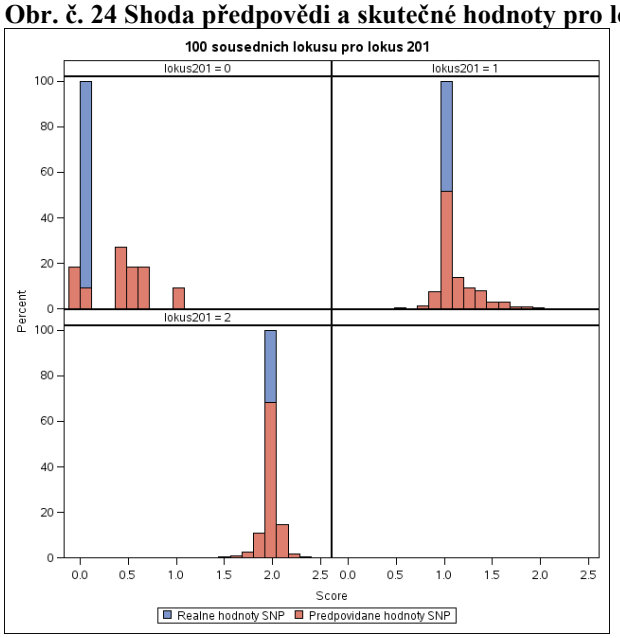
U hodnoty 0 se shoda předpovědi se skutečnou hodnotou v posledním modelu pro lokus 760 vyšplhala na 50 %, pro hodnotu 1 na 40% a pro hodnotu 2 na 45 %.

Obr. č. 23 Shoda předpovědi a skutečné hodnoty pro lokus 893 v 8. modelu



U lokusu 893 se shoda předpovědi pro hodnotu 0 dostala nad 50 %, pro hodnoty 1 a 2 na 70 %

Obr. č. 24 Shoda předpovědi a skutečné hodnoty pro lokus 201 v 8. modelu



U téměř homozygotního lokusu 201 se v posledním modelu podařilo u hodnoty nula dosáhnout shody předpovědi se skutečnou hodnotou lokusu zhruba 10 %, pro hodnotu 1 50 % a pro hodnotu 2 přes 65 %.

V porovnání se souborem A se nepodařilo zopakovat absolutní shodu předpovědi se skutečnou hodnotou. Důvodem jsou trochu jiné sledované lokusy, jiný soubor zvířat chybovost modelu, která dosahuje podstatně vyšších hodnot než v případě testování souboru A.

Výsledky na uvedených obrázcích je třeba hodnotit nikoliv pouze jako shodu skutečné a předpovídané hodnoty vyjádřené přesně jedním číslem, ale jako předpověď, která je v blízkém okolí skutečné hodnoty. Budeme-li jako správnou předpověď uvažovat i výskyt v bezprostředním okolí, vyjádřeném například podílem z celkového rozpětí <0, 2>, pak je prakticky využitelná shoda vyšší.

Důležitým výstupem jsou programy (příloha 1 a 2) vytvořené v programovém prostředí SAS (2002), které lze upravovat a použít k opakovaným výpočtům.

6 Diskuze

6.1 Stávající výzkum imputací a úpravy genomických dat

Imputace chybějících markerů v SNP čipech je v současné době ve světě velmi probírané téma. Důvodem je rozvoj genomiky a snaha o maximální využití již nashromážděných dat. Tato práce je první, která se v České republice danou věcí zabývá.

Existují obecně dva přístupy k předpovědi genomické plemenné hodnoty. Vícekroková metoda, kde jsou hodnoceni pouze genotypovaní jedinci, a jednokroková metoda, kde je možné hodnotit pomocí předpovědi plemenných hodnot s využitím genetických markerů celou populaci (Misztal et al., 2009; Bauer et al., 2014).

Výpočet genomických plemenných hodnot v České republice (Přibyl et al., 2014) probíhá jednokrokovou metodou, která pracuje s realizovanou genomickou příbuzností a nevyžaduje přímé doplnění chybějících hodnot. Proto můžeme pozorovat, že studie, které se tímto tématem zabývají, vznikají převážně v zemích, kde výpočet GPH probíhá pomocí dvoukrokové metody (Meuwissen et al., 2001). V takovém případě je totiž imputace dat pro správnost výpočtu nezbytná. Nicméně by bylo vhodné ověřit, zda imputace chybějících SNP zpřesní genomickou příbuznost.

V literární rešerši lze nalézt informace o programech, které byly k imputacím genotypů vytvořeny. Tyto programy nebyly v práci použity. Důvodem je, že k efektivnímu otestování programů nebyly dostupné potřebné údaje jako například ogenotypování rodičů a potomků a údaje o způsobu a kvalitě vyhodnocení čipů.

Zásadní obtíží bylo, že většina genotypů postrádala údaje o kódování alel, které jsou důležité pro sjednocení dat ve většině programů. Chybění některých údajů je zapříčiněna tím, že národní svazy plemen skotu si často vzájemně vyměňují genotypy se svazy v zahraničí. Dochází tedy k tomu, že některé údaje zahraniční svazy při výměně neposkytují a není je možné dohledat.

Dalším znevýhodněním je skutečnost, že máme většinou pouze genotypy býků, a tudíž nelze propojit genotypy jedinců s genotypy jejich rodiči. Tato skutečnost vylučuje využití programů založených na původu zvířete jako je AlphaImpute (Hickey et al., 2012), Findhap (VanRaden et al., 2011), DAGPHASE (Druet a Georges, 2010), FImpute (Sargolzaei et al., 2008, 2014), PedImpute (Nicolazzi et al., 2013). Z tohoto důvodu by bylo logické použít program, ve kterém údaje o původu zvířete a genotypy jeho rodičů můžeme vynechat.

Nejlepší volbou se jeví programy Beagle a Impute 2, které zároveň patří k nejpoužívanějším programům. Obtíží jsou však již zmiňované chybějící údaje o kódování alel jednotlivých čipů.

Zajímavé porovnání imputace pomocí programů Beagle, FindHap, FImpute, AlphaImpute a IMPUTE2 provedl ve své studii Ma et al. (2013). Jednalo se o imputaci genotypů z Illumina BovineSNP50 BeadChip na Illumina BovineHD BeadChip. Pracoval s 3 902 genotypovanými zvířaty na 54K čipu a 458 zvířaty, genotypovanými na 54k a HD čipu. Testovaná populace byla rozdělena do 4 skupin podle toho, zda byl v referenční populaci přítomen otec jedince nebo otec matky. Tabulka č. 12 uvádí úspěšnost imputace v % pro všechny 4 skupiny referenční populace.

Tabulka č. 12 Porovnání úspěšnosti imputace při použití různých programů.

	Beagle	Findhap	AlphaImpute	FImpute	IMPUTE2
GRPsmgs	99.4	99.2	98.8	99.6	99.6
GRPsire	99.1	98.1	98.1	99.2	99.3
GRPmgs	99.1	97.7	96.1	99.2	99.3
GRPnone	98.7	95.7	95.6	98.9	99.0
Dohromady	99.1	97.6	97.1	99.2	99.3

GRPsmgs = přítomni oba genotypovaní rodiče GRPSsire = pouze genotypovaný otec, GRPmgs = pouze genotypovaný otec matky, GRPnone = ani otec, ani otec matky nebyli genotapovaní .

Nejlépe v tomto testování dopadl program Impute 2. Nejmenší rozdíl v použití programů bylo ve skupině, kde byl znám otec jedince i otec matky jedince. Největší rozdíl byl naopak ve skupině, kde jedinec neměl v referenční skupině žádného předka (Ma et al., 2013).

6.2 Vlastní metoda imputace

Jak již bylo zmíněno, pro účely této práce nebyl použit žádný zmiňovaný program. Bylo nutné vyvinout vlastní možný postup předpovědi chybějících hodnot bez použití rodokmenu. Byla zvolena procedura GLM v systému SAS. Jednalo se tedy o výpočet regresních kofecientů mezi závisle proměnnou - testovaným lokusem - a nezávisle proměnnými – sousedními lokusy.

Údaje ve výše uvedené tabulce č. 12 (Ma et al., 2013) jsou souhrnné údaje za všechny lokusy. Obvykle se dělá před imputacemi mezi lokusy předvýběr a pracuje se s lokusy, které vykazují proměnlivost (Hickey et al., 2012). Rovněž se vyřadí jedinci s vysokou pravděpodobností chybného původu, nebo chybného genotypování (Přibyl et al., 2014). V naší práci jsme se zaměřili na vybrané lokusy, heterozygotní a téměř homozygotní. U téměř homozygotních lokusů jsou předpovědi méně spolehlivé. Při výpočtu přes všechny lokusy a

při práci jen s lokusy, které mají proměnlivost, lze očekávat, že bychom se přibližovali k hodnotám ve výše uvedené tabulce.

Výpočty probíhaly na vybraných lokusech z chromozomu č. 1 a dvou souborech genotypovaných býků. Výsledky pro jednotlivé druhy lokusů (heterozygotní x homozygotní) vycházely podobně i na jiných chromozomech.

Testovány byly dvě populace. V prvním případě se jednalo o 260 býků z České republiky. Tento soubor dat prvotně sloužil, jako modelová populace pro vyzkoušení výpočtu, zdali je vůbec možné touto metodou postupovat. Bylo dosaženo přesného odhadu hodnot pro lokusy se zastoupením alely A až 75%. Homozygotní lokusy na tom byly hůře, spolehlivost modelu vystoupala maximálně na 55%. U druhého testovaného souboru, který obsahoval 3982 genotypů býků holštýnského plemene, byly výsledky předpovědí vyrovnanější. Nicméně oproti prvnímu souboru se viditelně zvýšila chyba předpovědi a bylo zaznamenáno její kolísání v průběhu testování všech modelů

Možné vysvětlení by mohlo být to, že nebyla zohledněna skutečná vzdálenost mezi testovanými lokusy, takže hustota výskytu se může v modelech lišit. Z toho také vyplývá, že v určitých místech může docházet ke crossing-overu, takže ve skutečnosti by nedocházelo ke společnému předávání těchto lokusů.

Pokud bychom použili skutečnou vzdálenost mezi lokusy, tak bychom pravděpodobně tuto metodu mohli upřesnit a imputace chybějících dat by mohla být spolehlivější. I z tohoto testování je patrné, že podstata výpočtu není špatná.

Výpočet s vytvořenými a v příloze uvedenými počítačovými programy lze upravit pro výpočet jiných lokusů i na dalších chromozomech.

Vyslovenou hypotézu, že doplnění chybějících SNP při genotypování plemeníků skotu umožní přesnější stanovení genomické příbuznosti, nelze vyvrátit. Ale tato skutečnost není podložena předešlými výpočty. V současné době se za chybějící hodnoty SNP dosazuje průměr celého lokusu populace. Je tedy logické si myslet, že hodnota, která by byla vypočítána individuálně pro jednotlivé lokusy, by měla být přesnější, než jednotná hodnota průměru.

Na začátku diskuze bylo řečeno, že v současné době se pro potřeby výpočtu genomických plemenných hodnot v České republice, data neimputují. Tato práce může sloužit jako výchozí k dalšímu studiu využití genomických údajů ve šlechtění skotu.

Do výpočtu GPH je nyní používána zhruba polovina dostupných genotypů. Zbytek genotypů je z určitých důvodů vyřazen. Studiem imputačních technik bylo zjištěno, že

základem je správná úprava dat. Existuje mnoho formátů čipů, několik způsobů kódování alel, a to by měl být podnět k dalšímu prozkoumání.

Dalším studiem standardizace genotypů by se mohlo docílit začlenění nevyužitých genotypů do výpočtu GPH, a tím i upřesnění genomických plemenných hodnot mladých zvířat zařazovaných do plemenitby.

7 Závěr

- Chybějící markery SNP lze doplnit na základě výskytu sousedních SNP.
- Tato diplomová práce poskytuje výchozí bod k dalšímu studiu okolností úpravy genomických údajů před začleněním do genetického hodnocení (předpovědi PH a GPH) hospodářských zvířat.
- Byly vytvořeny počítačové programy v programovém prostředí SAS, které lze opakovaně použít pro daný druh práce.

8 Seznam použité literatury

Attia J., Ioannidis, J.P., Thakkinstian, A., McEvoy, M., Scott, R.J., Minelli, C., Thompson, J., Infante-Rivard, C., Guyatt, G. 2009. How to use an article about genetic association A: Background concepts. Journal of the American Medical Association. 301 (1). 74 -81.

Baraldi, A. N., Enders, C. K. 2009 An introduction to modern missing data analyses. Journal of School Psychology. 48 (2010). 5- 37.

Bauck, S., Rekaya, R., Wang, H., Woodward, B. 2011. Imputation of missing SNP genotypes using low density panels. Livestock Science 12 (146). 80 - 83.

Bauer, J., Vostrý, L., Přibyl, J. CERTIFIKOVANÁ METODIKA: Odhad spolehlivosti jednokrokových genomických plemenných hodnot pro dojený skot, [online]. 2014 [cit. 2016-03 - 13]. Dostupné z < https://www.cmsch.cz/store/2014-metodika-bauer.pdf>.

Boison, S. A., Neves, H. R. R., Pérez O'Brien, A. M., Utsunomiya, Y. T., Carvalheiro, R., da Silva, M.V.G.B., Sölkner, J., Garcia, J.F. 2014. Imputation of non-genotyped individuals using genotyped progeny in Nellore, a Bos Indicus cattle breed. Livestock science. 14 (166). 176-189.

Browning, B. L., and S. R. Browning. 2013. Improving the accuracy and efficiency of identity-by-descent detection in population data. Genetics. 194 (2). 459–471.

Browning, S. R., Browning, B. L. 2007. Rapid and accurate haplotype phasing andmissing-data inference for whole-genome association studies by use of localized haplotype clustering. The American Journal of Human Genetics. 81(5). 1084–1097.

ČMSCH. Organizační struktura Českomoravské společnosti chovatelů, a.s. [online]. 2016 [cit. 2016-03 - 13]. Dostupné z < http://www.cmsch.cz/cs/cinnosti-cmsch-a-s/>

Dassonneville, R., Brøndum, R. F., Druet, T., Fritz, S., Guillaume, F., Guldbrandtsen, B., Lund, M. S., Ducrocq, V., Su, G. 2011. Effect of imputing markers from a low-density chip on the reliability of genomic breeding values in Holstein populations. Journal Dairy Science. 94(7). 3679–3686.

Druet, T., Georges, M. 2010. A hidden markov model combining linkage and linkage disequilibrium information for haplotype reconstruction and quantitative trait locus fine mapping. Genetics. 184(3). 789–798.

Eddy, S. R., 1996. Hidden Markov models. Current opinion in structural biology. 6. 361 -365.

Elaswarapu, R., Starkey, M. 2010. Genomics: Essential Methods. John Wiley & Sons. p. 360. ISBN: 9780470711576.

Hayes B. J., Bowman P. J., Chamberlain A. J., Goddard M. E. 2009 Invited review: Genomic selection in dairy cattle: progress and challenges. Journal Dairy Science. 92(2). 433–443.

Hickey, J. M., Kinghorn, B. P., Tier, B., van der Werf, J. H. J., Cleveland, M. A. 2012. A phasing and imputation method for pedigreed populations that results in a single-stage genomic evaluation. Genetics Selection Evolutiom. 44. 9.

Howie, B., Donnelly, P., Marchini, J. 2009. A Flexible and Accurate Genotype Imputation Method for the Next Generation of Genome-Wide Association Studies. PLOS Genetics. 5(6).

Howie, B.,Marchini, J. Impute 2 [online]. 23. december 2014 [cit. 2016-02-03]. Dostupné z <https://mathgen.stats.ox.ac.uk/impute/impute_v2.html#whole_chroms>.

Hruban, V., Majzlík, I., 2002, Obecná genetika. Powerprint. ISBN: 978 -80 213- 0600- 4

Illumina., Bovine SNP50 genotyping BeadChip [online]. 9th February 2016. [cit. 2016-08-03]. Dostupné z <http://www.illumina.com/Documents/products/datasheets/ datasheet_bovine_snp5O.pdf >.

Jansen, G. PEDIMPUTE [online]. 2012 [cit. 2016-03-15]. Dostupné z <http://dekoppel.eu/pedimpute/> .

Jiang J., Jiang L., Zhou B., Fu W., Liu J.F., Zhang, Q. 2011. SNAT: a SNP annotation tool for bovine by integrating various sources of genomic information. BMC Genetics. 12. 85.

Ječmínková, K., Kyselová, J., Jak lze molekulární genetiku využít ve šlechtění skotu? [online]. 23. května 2015 [cit. 2016-02-02]. Dostupné z < http://naschov.cz/jak-lze-molekularni-genetiku-vyuzit-ve-slechteni-skotu/ >

Li, Y., Willer, C. J., Ding, J., Scheet, P., Abecasis, G. R. 2010. MaCH: Using sequence and genotype data to estimate haplotypes and unobserved genotypes. Genetics Epidemiology. 34. 816–834.

Little, R. J. A., Rubin, D. B., 2002, Statistical Analysis with Missing data, 2nd edition, John Wiley & Sons. p. 408. ISBN: 9780471183860.

Ma, P., Brøndum, R. F., Zhang, Q., Lund, M. S., Su, G. 2013. Comparison of different methods for imputing genome-wide marker genotypes in Swedish and Finnish Red Cattle. Journal of Dairy Science, 96(7), 4666-4677.

Misztal, I., Legarra, A., Aguilar, I. 2009 Computing procedures for genetic evaluation including phenotypic, full pedigree and genomic information. Journal of Dairy Science. 92(9). 4648 – 4655.

Meuwissen, T., Hayes, B. J., Goddard, M. E. 2001. Prediction of total genetic value using genome – wide dense marker maps. Genetics. 157(4). 1819 – 1829.

NCBI. Database of single nucleotide polymorphisms (SNPs) and multiple small-scale variations that include insertions/deletions, microsatellites, and non-polymorphic variants. [online]. 2015 [cit. 2015-07 - 17]. Dostupné z < http://www.ncbi.nlm.nih.gov/snp>

Nicolazzi, E. L., Biffani, S., Biscarini, F., Orozco ter Wengel, P., Caprera, A., Nazzicari, N., Stella, A. 2015 Software Solutions For The Livestock Genomics Snp Array Revolution, Animal Genetics, 46(4):343-53

Nicolazzi, E. L.,Biffani, S., Jansen, G. 2012. PEDIMPUTE: Imputing genotypes using a fast algorithm combining pedigree and population information. Interbull Bulletin. 46. 33-38.

Nicolazzi, E. L.,Biffani, S., Jansen, G. 2013. Short communication: Imputing genotypes using PedImpute fast algorithm combining pedigree and population information. Journal of Dairy Science. 96. 2649–2653.

Nicolazzi, E. L., Picciolini, M., Strozzi, F., Schnabel, R. D., Lawley, C., Pirani, A., Brew, F., Stella, A. 2014. SNPchiMp: a database to disentangle the SNPchip jungle in bovine livestock. BMC Genomics. 15. 123.

Pešek, P., Přibyl, J. 2015. Genetic variances of SNP loci for milk yield in dairy cattle. Journal of Applied Genetics. 56(3). 339 – 347.

Pešek, P., Přibyl, J. 2014. Genomika – možnost zvýšení zisku u dojeného skotu. Náš chov. 74(1). 54- 60.

Plemdat. Popis stanovení plemenné hodnoty pro exteriér. [online]. 22. července 2009 [cit. 2016-03-08]. Dostupné z < http://www.plemdat.cz/cz/pages/Popis_exterier.pdf>

Přibyl, J., Bauer, J., Pešek, P., Přibylová, J., Vostrý, L., Zavadilová, L. 2014. Domestic and Interbull information in the single step genomic evaluation of Holstein milk production. Czech journal of Animal Science. 59. 409 – 415.

Rubin, D. B. 1976. Inference and Missind Data. Biometrika. 63, 581-592.

Ruvinsky, A., Graves, J. A. M. 2004. Mammalian Genomics. CABI. p. 612. ISBN: 9780851999104.

Sargolzaei, M. Chesnais, J. P., Schenkel, F. S. 2014. A new approach for efficient genotype imputation using information from relatives. BMC Genomics. 15. 478.

Sargolzaei, M., Schenkel, F. S., Jansen, G. B., Schaeffer, L. R., 2008. Extent of linkage disequilibrium in Holstein cattle in North America. Journal of Dairy Science. 91. 2106–2117.

SAS. 2002. The MIXED Procedure, The GLM Procedure. SAS/STAT Software. SAS Institude Inc.

Segelke, D., Chen, J., Liu, Z., Reinhardt, F., Thaller, G., Reents, R. 2012. Reliability of genomic prediction for German Holsteins using imputed genotypes from low-density chips. Journal Dairy Science. 95(9). 5403 – 5411.

Schaeffer L. R. 2006. Strategy for applying genome-wide selection in dairy cattle. Journal of Animal Breeding Genetics. 123(4). 218–223.

Schork N., Fallin D., Lanchbury, J. 2000. Single nucleotide polymorphisms and the future of genetic epidemiology. Clinical Genetics. 58(4). 250 - 264.

Snudstad P., Simmons M. 2009. Genetika. Masarykova univerzita, Brno, 894 pp ISBN: 978-80-210-4852-2

VanRaden, P. M., O'Connell, J. R., Wiggans, G. R., Weigel, K. A. 2011. Genomic evaluations with many more genotypes. Journal of Dairy Science. 95. 5403–5411.

Walker E. J., Siminovitch K.A. 2007. Primer: genomic and proteomic tools for the molecular dissection of disease. Nature Clinical practise rheumatology. 3. 580 – 589.

Womack, J. 2012. Bovine Genomics. John Wiley & Sons. p. 285. ISBN: 9780813821221.

Yuang, Y. C. Multiple Imputation for Missing Data: Concepts and New Development (Version 9.0) [online]. 5th January 2016 [cit. 2016-11-03]. Dostupné z <http://www.math.montana.edu/jimrc/classes/stat506/notes/>

Zhang, Z., Druet, T. 2010. Marker imputation with low-density marker panels in Dutch Holstein cattle. Journal of Dairy Science. 93. 5487–5494.

Ziegler A., König I., Pahlke F. 2010. A Statistical Approach to Genetic Epidemiology: Concepts and Applications, with an e-learning platform. John Wiley & Sons, p. 144. ISBN: 978527323890.

9 Seznam použitých zkratek a symbolů

A – adenin

BLUP – Best linear unbiased prediction – nejlepší lineární nevychýlená předpověď

C – cytosin

cDNA – komplementární DNA vzniklá transkripcí (přepisem) RNA do DNA

DNA – deoxyribonukleová kyselina, genetická informace jedince

ELFO – elektroforéza

G – guanin

GPH – genomická plemenná hodnota

HMM – Hidden Markov model – Skrytý Markovův model

IBD – identity by descent – shodný podle původu

LD – Linkage disequilibrium – vazebná nerovnováha

LE – Linkage equilibrum – vazebná rovnováha

LHCM - Localized haplotype clustering model – Shlukový model lokalizovaných haplotypů

MAR - Missing at Random – hodnoty náhodně chybějící

MCAR – Missing Completely at Random – hodnoty kompletně náhodné chybějící

MNAR – Missing Not a Random – hodnoty nenáhodně chybějící

NCBI – National Centre For Biotechnology Information

PH – plemenná hodnota

PCR – Polymerase chain reaction – polymerázová řetězová reakce

SNP – Single nucleotide polymorphism – jednonukleotydový polymorfismus

T - thymin

10 Přílohy

V příloze jsou k nahlédnutí programy v SAS (2002), které byly vytvořeny pro potřeby výpočtu této diplomové práce. Pro zájemce na vyžádání k opakovanému použití.

10.1 Přípravný program pro úpravu dat

```
/*DIPLOMOVA PRACE - Bc. ANITA KRANJCEVICOVA*/
/*PRIPRAVNY PROGRAM PRO VYPOCET IMPUTACI*/
/*POSLEDNI ZMENA 2.2.2016*/

/*Nacteni souboru*/
%let cesta= /home/kranjcevicova/soubory;
%let csnp= /home/ kranjcevicova/Genotypy11-2014;
%let cchr= /home/ kranjcevicova/chromozom;

Filename zeme   "&cesta/zeme.txt";     /*vzornik zeme*/
Filename chrom  "&cesta/chrom";        /*umisteni na chromozomu*/
Filename puv      "&cesta/puvo";       /*PUVOD - PLEMENO*/
filename zdvjm   "&cesta/zdvjm.txt";  /*zdvojen jmen*/
Filename hrubci  "&cesta/hrubci";     /* HRUBY CISELNIK*/

/*soubory s snp*/
Filename snp1    "&csnp/kgnp42u.txt";
Filename snp2    "&csnp/kgnp41u.txt";
Filename snp3    "&csnp/kgnp43u.txt";
Filename snp4    "&csnp/kgnp44u.txt";
Filename snp5    "&csnp/kgnp45u.txt";
Filename snp6    "&csnp/kgnp46u.txt";
Filename snp7    "&csnp/kgnp47u.txt";
Filename snp8    "&csnp/kgnp48u.txt";
Filename snp9    "&csnp/kgnp49u.txt";
/*chromozom*/ /*VYSTUPNI SOUBORY PRO VYPOCET*/
filename ch1  "&cchr/ch/ch1.txt";
filename ch2 "&cchr/ch/ch2.txt";
filename ch3 "&cchr/ch/ch3.txt";
filename ch4 "&cchr/ch/ch4.txt";
filename ch5 "&cchr/ch/ch5.txt";
filename ch6 "&cchr/ch/ch6.txt";
filename ch7 "&cchr/ch/ch7.txt";
filename ch8 "&cchr/ch/ch8.txt";
filename ch9 "&cchr/ch/ch9.txt";
filename ch10 "&cchr/ch/ch10.txt";
filename ch11 "&cchr/ch/ch11.txt";
filename ch12 "&cchr/ch/ch12.txt";
filename ch13 "&cchr/ch/ch13.txt";
filename ch14 "&cchr/ch/ch14.txt";
filename ch15 "&cchr/ch/ch15.txt";
filename ch16 "&cchr/ch/ch16.txt";
filename ch17  "&cchr/ch/ch17.txt";
filename ch18 "&cchr/ch/ch18.txt";
filename ch19 "&cchr/ch/ch19.txt";
filename ch20  "&cchr/ch/ch20.txt";
filename ch21  "&cchr/ch/ch21.txt";
filename ch22  "&cchr/ch/ch22.txt";
filename ch23  "&cchr/ch/ch23.txt";
filename ch24  "&cchr/ch/ch24.txt";
filename ch25  "&cchr/ch/ch25.txt";
```

```
filename ch26  "&cchr/ch/ch26.txt";
filename ch27  "&cchr/ch/ch27.txt";
filename ch28  "&cchr/ch/ch28.txt";
filename ch29  "&cchr/ch/ch29.txt";
/*filename chx   "&cchr/ch/chx.txt";
filename chy "&cchr/ch/chy.txt";*/
dm  output "clear";dm log "clear";

proc printto print= " /home/kranjcevicova/stat.txt";
proc printto log = " /home/kranjcevicova/log.lst";

data zeme;   /*zeme puvodu vzor*/
infile zeme firstobs = 2 missover;
input czm $39-41 ze2 $42-43 ze3 $44-46;
if czm = "840" then do; ze2 ="US"; ze3 = "USA"; end;
if czm= "." then delete; xx= czm; if xx*1= 0 then delete; if czm = "" then
delete;
proc means;
proc sort; by czm;

data soub ; /* PUVODY - plemena*/
            title " rodokmen 32";
infile puv missover;
input nc 1-16 nco 18-33 ncm 35-50
podCj 52-54 podHj 56-58 podRj 60-62 podOj 64-66 podcel 68-70
roknaj 72-75 rokkj 77-80 zj $82-84 zo $86-88 zm $90-92 ;
proc means;


data snd; /*nacteni databaze snp*/
Infile chrom dlm="," firstobs=2 missover ;
Informat  ums $25.;
/*format ums $20.; */
attrib snj format = $40. length= $40.;
attrib chr format =$2. length= $2.;
Input  snj $  chr $   ums;

snj= compress(snj,"-");
snj= compress (snj, "_");
proc sort data= snd; by snj;


%macro nactisnp;/*nacitani souboru s SNP*/
%do i=1 %to 9;
data sn1;
infile snp&i missover;
input czm $1-3 cj $4-16 snj $18-57 sncis 59-63 al1 $65 al2 $67 gcsco 69-74
gtsco 76-81;
attrib ciz format= $16.  length=$16;
attrib snj format = $40. length= $40.;
ciz= czm||cj;
als= compress (al1||al2);
proc means; var sncis gcsco gtsco;
proc sort data= sn1; by czm; run;

/*.........................zdvojena jmena.....................*/
data ze ; title "zdvojena jmena";
attrib cdo ciz format=$16. length=$16;
infile zdvjm ;
input plin $1-3 cdo 5-20 ciz 22-37 ;
proc means ;
```

```
proc sort ; by ciz ;
/*.........................ciselnik.............................*/
data cis ;                                                  title " ciselnik
47";
infile hrubci dlm=";" ;
input ciz $1-16 nc 18-26 ;
proc sort data= cis; by ciz;
/*.........................oprava zeme.........................*/
data snpa;
merge sn1(in=zesn) zeme ; by czm ; if zesn ;
data snpb; set snpa;
if ze3 ne "" then ciz = ze3||cj;
proc sort ; by ciz ;
/*.......................oprava zdvojeneho jmena.............*/
data snpc;
merge snpb(in=zesn) ze ; by ciz ; if zesn ;

data snpp;
set snpc;
keep ciz snj sncis al1 al2 als;
proc sort data= snpp; by ciz;

data msn;
Merge  snpp (in=fromsni) cis (in=puv);if fromsni; if puv; by ciz;run;

data pm1;
set msn;
keep nc snj sncis al1 al2  als;
run;

proc sort data= pm1; by nc;
proc sort data= soub; by nc;
/*.........................pridani plemene.....................*/
data pod;
Merge pm1(in=sni) soub (in=cisl); if sni; if cisl; by nc; run;

data pod1; /*vyber 100% holstynu*/
set pod;
if podhj ne "100" then delete;
proc means;

data vst1;
set pod1;

keep nc snj sncis al1 al2 als;
/*....................... trideni snp...........................*/
data sn;   /*trideni snp*/
set vst1;
if al1 ne "A" & al1 ne "B" then al1= "_";
if al2 ne "A" & al2 ne "B" then al2= "_";
if al1= "_" & al2= "_" then delete;

proc sort data = sn; by snj; run;
data ds;
merge snd sn; by snj; if nc= "" then delete;
proc sort data= ds; by snj;
run;
data alely; title " Precislovani alel";
set ds;
format a1 a2 alely 1.;
/*A=1, B=2, 11=0, 12=1 ,21=1, 22=2, 00=., 0=., 0=.*/
```

```
      if al1= "A" then a1= 1;
else  if al1= "B" then a1= 2;
      if al2= "A" then a2= 1;
else  if al2= "B" then a2= 2;


if a1 =1 and a2 =1 then alely= 0;
else if a1= 1 and a2= 2 then alely= 1;
if a1= 2 and a1= 1 then alely= 1;
else if a1= 2 and a2= 2 then alely= 2;
if a1= 0 and a2= 0 then alely= .;
 if a1= 0            then alely= .;
 if            a2= 0 then alely= .;

proc freq;  tables chr;
run;
/*.............vymazani chr X a Y.................................*/
data snp&i;
set alely;
if chr ="X " then delete;
if chr = "Y " then delete; ch=chr*1 ;
if ch=0 then delete;
  run;
%end;%mend;%nactisnp; run;
                        data snp ;
                              set snp1 snp2 snp3 snp4 snp5 snp6 snp7 snp8
snp9 ; run;
%macro chromozom;
%do k=1 %to 29;
data chr;
set snp;
if ch  ne &k then delete;proc sort;by ums;

data lok;  /*ciselnik lokusu*/
set chr; by ums; if first.ums;data lok;set lok;
lokus= _n_;
keep  lokus ums;
proc means; Title "Ciselnik Chromozom &k";
data chl&k;
Merge  chr lok ; by ums;


data zapis;
set chl&k ;
File ch&k dlm= ";" ;
put  snj $  lokus chr $  ums a1 a2 alely sncis nc ;

%end;%mend;%chromozom;run;
```

10.2 Program pro výpočet imputací

```
/*************************DIPLOMOVA PRACE*****************************/
/*Anita Kranjcevicova
Imputace chybejicich dna markerů v snp cipech
2. cast programu- navazuje na pripravny program 2.2.2016
*********************************************************************/
%let cesta= /home/kranjcevicova;

filename ch "&cesta/ch1.txt";
```

```
filename souct "&cesta/souct.txt";
filename matice "&cesta/matice";
proc printto print= " /home/kranjcevicova/stat.txt";
proc printto log = " /home/kranjcevicova/log.lst";

/*Nacitani souboru ch - chromozom*/
/*vytvoreni pomocne promenne p - pro vymazani duplicitnich snp */
data nacti;
Infile ch dlm=";" missover;
attrib snj format = $40. length= $40.;
attrib chr format =$2. length= $2.;
attrib nc format=$16. length=$16.;
attrib p   format=$56. length=$56.;
Input snj $  lokus chr $  ums $ a1 a2 alely sncis nc  ;
p= snj||nc;
proc sort; by p;
data t;
set nacti;by p;  if first. p;
keep snj lokus alely nc; proc sort; by nc;run;

/*Precislovani byku*/
data t1;
set t; by nc; if first.nc;
proc sort; by nc;
data t1; set t1;
ncb= _n_;keep nc ncb;
proc means; run; Title "ciselnik byci";run;

/* zjisteni poctu byku v souboru a trideni*/
data t2;  Merge T T1; by nc;
keep snj lokus alely ncb; run;

data T; set t2; proc sort; by ncb;run;

data t3; set t2;
proc means noprint; var alely; by ncb;
output out=pru mean=;

data T31; set pru; pocet= _freq_;
keep ncb pocet;
proc means;run;

data t4; Merge T3 T31; by ncb;

data t5; set T4;
pom= 3429*0.9;
if pocet < pom then delete;
keep snj lokus alely ncb pocet;proc means; run;proc sort; by ncb lokus;

data pocb;
set t5; by ncb; if first.ncb; keep ncb;
proc means noprint; var ncb;
output out=pru mean=;

data sbc; set pru; celpocet=_freq_;
keep celpocet a;
a=1;
proc means;

/*zjisteni poctu lokusu v chromozomu a trideni*/
proc sort data=t5; by lokus;
```

```
data _null_;
keep lokus souc poct;
file souct;
set t5; by lokus;
if first.lokus then do;
souc=0; poct=0;
end;
if alely ne . then do;
pocet +1; souc + alely; end;
if last.lokus then
put lokus 1-6 poct 8-13 souc 15-20;

data b;  Title "cetnosti lokusu";
Infile souct;
Input lokus 1-6 souc 8-13 poct 15-20;
a=1;
proc means; run;

data sn;
merge b sbc; by a; drop a;
if poct = . or poct < 0.95*celpocet then delete;
prum=souc/poct;
roz= prum*(2-prum)/2;
proc means;run; proc sort; by lokus;proc sort data=t5; by lokus;
data b1;    Title "pretrizene alely";
merge  t5 sn(in=abc); by lokus; if abc;
keep snj lokus alely ncb; proc sort; by ncb;
proc means;


/* 2. precislovani byku */
data nc1;
set b1; by ncb; if first.ncb;
data nc2; set nc1;
nncb=_n_;
keep ncb nncb;
proc means; run; Title "cis byci";
data nck;
Merge b1 nc2; by ncb;
keep snj lokus alely nncb;
proc sort; by lokus; run;

/*2. precislovani lokusu*/
data l1;
set nck;by lokus; if first.lokus;

data l2; set l1;
nlokus=_n_;
keep lokus nlokus;
proc means; run; Title "cis lokus";

data ncl;
Merge l2 nck(in=acc); by lokus; if acc;
keep snj nlokus alely nncb;run;

data KV;
set ncl;
proc sort; by nncb nlokus;

/*********MACRO1***********************/
%macro vytvm;%do i= 1 %to 3982;
```

```
data v;set KV;
if nncb ne  &i then delete;
alely&i= alely;
drop alely;
filename byk "&cesta/byk//byk&i";
File byk dlm= ";";
put snj alely&i nlokus nncb;
%end; %mend; %vytvm;run;
/************************************/
/********MACRO2***********************/
/*spojeni dat byk1-n za sebou pro pripravu matice (sloupec = byk , radek=
lokus*/
%macro all;
data c1;
filename byk "&cesta/byk/byk1";
infile byk dlm= ";";
input snj $ alely1 nlokus nncb;proc sort; by nlokus;
%do i= 2 %to 3982;
filename byk "&cesta/byk//byk&i";
data byk;
infile byk dlm= ";";    /*spojeni dat byk&i za sebou  */
input snj $ alely&i nlokus nncb;proc sort; by nlokus;
data c1 ;
merge c1  byk ; by nlokus ;
%end;
data c1;
set c1; drop snj nlokus nncb;
%mend;%all;run;

proc iml;
use c1;
read all into x;
close pa;
a=x`;
create ab from a;
append from a;
run;

data aa1;
set ab;
array  col col1- col1898;
                        /*sloupec= lokus radek= byk*/
file matice dlm= ";";
        put col1 - col1898 ;
run;
%macro zj; data aa2; set aa1;
%do i=1 %to 1898;
RENAME col&i=lokus&i;
%end; %mend; %zj;run;
data az; set aa2;

%macro vyp;
%do i= 51 %to 1854;
data az1;
    set az;
array lll lokus1 - lokus1904;
li=lll[&i]; /*zavisle promenna*/
l1=lll[&i+1];l2=lll[&i+2];l3=lll[&i+3];
l4=lll[&i+4];l5=lll[&i+5];l6=lll[&i+6];l7=lll[&i+7];l8=lll[&i+8];l9=lll[&i+
9];l10=lll[&i+10];
l11=lll[&i+11];l12=lll[&i+12];l13=lll[&i+13];l14=lll[&i+14];l15=lll[&i+15];
l16=lll[&i+16];l17=lll[&i+17];l18=lll[&i+18];l19=lll[&i+19];l20=lll[&i+20];
```

```
l21=lll[&i+21];l22=lll[&i+22];l23=lll[&i+23];l24=lll[&i+24];l25=lll[&i+25];
l26=lll[&i+26];l27=lll[&i+27];l28=lll[&i+28];l29=lll[&i+29];l30=lll[&i+30];
l31=lll[&i+31];l32=lll[&i+32];l33=lll[&i+33];l34=lll[&i+34];l35=lll[&i+35];
l36=lll[&i+36];l37=lll[&i+37];l38=lll[&i+38];l39=lll[&i+39];l40=lll[&i+40];
l41=lll[&i+41];l42=lll[&i+42];l43=lll[&i+43];l44=lll[&i+44];l45=lll[&i+45];
l46=lll[&i+46];l47=lll[&i+47];l48=lll[&i+48];l49=lll[&i+49];l50=lll[&i+50];
ll1=lll[&i-1];ll2=lll[&i-2];ll3=lll[&i-3];ll4=lll[&i-4];ll5=lll[&i-
5];ll6=lll[&i-6];ll7=lll[&i-7];ll8=lll[&i-8];ll9=lll[&i-9];ll10=lll[&i-10];
ll11=lll[&i-11];ll12=lll[&i-12];ll13=lll[&i-13];ll14=lll[&i-
14];ll15=lll[&i-15];ll16=lll[&i-16];ll17=lll[&i-17];ll18=lll[&i-
18];ll19=lll[&i-19];ll20=lll[&i-20];
ll21=lll[&i-21];ll22=lll[&i-22];ll23=lll[&i-23];ll24=lll[&i-
24];ll25=lll[&i-25];ll26=lll[&i-26];ll27=lll[&i-27];ll28=lll[&i-
28];ll29=lll[&i-29];ll30=lll[&i-30];
ll31=lll[&i-31];ll32=lll[&i-32];ll33=lll[&i-33];ll34=lll[&i-
34];ll35=lll[&i-35];ll36=lll[&i-36];ll37=lll[&i-37];ll38=lll[&i-
38];ll39=lll[&i-39];ll40=lll[&i-40];
ll41=lll[&i-41];ll42=lll[&i-42];ll43=lll[&i-43];ll44=lll[&i-
44];ll45=lll[&i-45];ll46=lll[&i-46];ll47=lll[&i-47];ll48=lll[&i-
48];ll49=lll[&i-49];ll50=lll[&i-50];
proc glm data= az1; Title "Zavisle promenna= lokus&i";
model li = l1 l2 l3 l4 l5 l6 l7 l8 l9 l10
l11 l12 l13 l14 l15 l16 l17 l18 l19 l20
l21 l22 l23 l24 l25 l26 l27 l28 l29 l30
l31 l32 l33 l34 l35 l36 l37 l38 l39 l40
l41 l42 l43 l44 l45 l46 l47 l48 l49 l50
l11 l12 l13 l14 l15 l16 l17 l18 l19 l20
l11 l12 l13 l14 l15 l16 l17 l18 l19 l110
l111 l112 l113 l114 l115 l116 l117 l118 l119 l120
l121 l122 l123 l124 l125 l126 l127 l128 l129 l130
l131 l132 l133 l134 l135 l136 l137 l138 l139 l140
l141 l142 l143 l144 l145 l146 l147 l148 l149 l150;

output out=diag p=predpoved&i r=chyba&i;
data q&i;
set diag;
abschyba&i=abs(chyba&i); keep lokus&i predpoved&i chyba&i abschyba&i;

proc means;run;

%end;%mend;%vyp;
run;*/
```