

Manuál pro program

PxSNP v1.0

Program pro úpravu genomických dat pro účely
imputace genotypu.

2020

Autoři: Anita Klímová, Jakub Klíma

Výzkumný ústav živočišné výroby (Institute of Animal Science)

Přátelství 815 104 00 Praha – Uhřetěves

Česká republika

Vypracováno za podpory projektů MZe - Navýšení spolehlivosti celostátního genomického hodnocení dojeného skotu zařazením krav s domácí užitkovostí do genotypované referenční populace (QK1810253) a institucionální podpory (MZE-RO0718).

Úvod

Program PxSNP slouží pro úpravu surových genomických dat. Součástí programu je kontrola vstupních údajů pro použitelnost rozličných formátů k dalšímu zpracování a samotná příprava genomických údajů pro imputaci v programu FImpute (Sargolzaei et al. 2014). Součástí programu je také simulační balíček SIMPUT, který slouží k ověření přesnosti imputace v závislosti na velikosti referenční populace a množství chybějících SNP. Jedna z částí programu PxSNP je zaměřena na kontrolu rodokmenu a detekce cyklů, které se často v rodokmenech vyskytují. Program PxSNP je schopný připravit surová data k imputaci a následně je upravit pro další použití pomocí zpětného převodu. Zpětný převod upravuje data do původní podoby, případně do podoby vstupního souboru pro předpověď plemenných hodnot v programu BLUPf90 (Mizstal et al. 2018).

Systémové prostředí

Programy balíčku PxSNP jsou dostupné jako spustitelné (*.exe) soubory. Tyto soubory lze spustit pod operačními systémy Windows, Linux a MacOS ve variantách 32 a 64 bit.

Postup instalace

Po stažení je nutné rozbalit ZIP archiv PxSNP_v1.0.zip. Po rozbalení jsou ve složkách dostupné spustitelné soubory. Pro spuštění je nutné mít nainstalovaný .NET Core Runtime 3.1., ten lze stáhnout na webové stránce <https://dotnet.microsoft.com/download/dotnet-core/3.1>.

Popis a použití částí programu PxSNP

Balíček PxSNP je složen celkem ze 7 samostatně spustitelných částí. Každá část, kromě detekce cyklu v rodokmenu, obsahuje vedle spustitelného souboru také parametrový soubor XML, který je před spuštěním nutný vyplnit. (Pro spuštění programu nad jiným parametrovým souborem, zadejte cestu jako parametr příkazové řádky, např. „PxSNP.exe params2.xml“.)

Kontrola vstupu PxSNP

Část KontrolaVstupuPxSNP slouží pro prvotní seznámení s databází genotypů. Nahlíží do všech souborů v cílové složce a kontroluje přítomnost zvolených parametrů, které jsou nutné pro správný chod hlavní části PxSNP. Program se zaměřuje na hlavičku surových genomických dat a hledá označení čipu 1., řádek [Data] 2. (slouží k orientaci programu PxSNP a vyhledání řádku s názvy proměnných), proměnné s kódováním alel AB 3. (slouží pro přípravu genotypu pro imputaci) a proměnné s informacemi o pozici SNP Chr a Position 4 (důležité pro vytvoření mapového souboru SNP).

```
[Header]
GSGT Version,1.9.4
Content,,TestovaciCip3 1.
[Data] 2.
```

SNP Name	Sample ID	Allele1 - AB	Allele2 - AB	Chr	Position
SNP1	zvire2	B	B	1	6371334
SNP2	zvire2	B	B	1	7928189
SNP3	zvire2	A	B	1	31819743

Vstupní soubory

- složka se surovými genotypy

Parametrový soubor

```
<?xml version="1.0" encoding="utf-8" ?>
<root>

  <cestaKeSlozceVstupnichGenotypu>TestovaciData</cestaKeSlozceVstupnichGenotypu>
  <cestaKeSlozceVystupu>Vystupy</cestaKeSlozceVystupu>

</root>
```

TestovaciData – nahradit cestou ke složce se soubory s genotypy

Vystupy- cesta ke složce kam program uloží výstupy

Výstupní soubory

chybne_vstupy.txt – seznam s cestami k souborům, kterým chybí nějaký požadovaný parametr (včetně výpisu chybějících parametrů), chybějící parametry 1 a 2 lze doplnit ručně. Pokud chybí

parametr 3 – kodování alel AB, je soubor nepoužitelný. Pokud chybí parametr 4, lze využít další část programu – Připojení pozic SNP

ok_vstupy.txt – cesty k souborům, které splňují přítomnost požadovaných parametrů a lze je bez problému použít. Tento seznam souborů lze nadále použít jako vstup pro PxSNP.

output.txt – obsahuje informace o procházení jednotlivých souborů. V případě bezchybného chodu je zakončen řádkem – Hotovo.

Třídění čipu

Slouží ke třídění surových souborů s genotypem podle použitého čipu.

Vstupní soubory

- složka se surovými genotypy

Parametrový soubor

```
<?xml version="1.0" encoding="utf-8" ?>
<root>

  <cestaDoSlozkyVstupu>C:\testovaci-slozka</cestaDoSlozkyVstupu>
  <cestaDoSlozkyVystupu>C:\testovaci-slozka_vystup</cestaDoSlozkyVystupu>

</root>
```

C:\testovaci-slozka – nahradit za cestu ke složce se soubory s genotypy

C:\testovaci-slozka_vystup – nahradit za cestu ke složce, kam program uloží výstupy

Výstupní soubory

Ve složce TrideniCipu jsou vytvořeny složky podle použitého čipu, kam program roztrídí soubory. Pokud soubor neobsahuje informace o použitém čipu, je uložen do složky BezNazvuCipu.

Vystup.txt – soubor obsahuje výčet čipů v databázi a počet souborů, které je obsahují

Připojení pozic SNP

Připojení pozic SNP slouží k připojení proměnných Chr a Position do souborů, které je neobsahují, případně přepíše tyto dvě proměnné podle externí databáze, kterou chcete použít.

Vstupní soubory

- složka se soubory s genotypy
- externí databáze pozic SNP ve tvaru:

SNP Name	Chr	Position
SNP1	1	6371334
SNP2	1	7928189
SNP3	1	31819743
SNP4	1	6133529
SNP5	1	27005721
SNP6	1	6616434
SNP7	1	17544926
SNP8	1	34639444
SNP9	1	31267746

Parametrový soubor

```
<?xml version="1.0" encoding="utf-8" ?>
<root>

  <cestaKeSlozceVstupnichGenotypu>TestovaciData/Genotypy</cestaKeSlozceVstupnichGenotypu>
  <cestaKSouboruSPozicemi>TestovaciData/externipozicetest.txt</cestaKSouboruSPozicemi>
  <cestaKeSlozceVystupu>Vystupy</cestaKeSlozceVystupu>

</root>
```

TestovaciData/Genotypy – nahradit za cestu ke složce se soubory s genotypem

TestovaciData/externipozicetest.txt – nahradit za cestu k souboru s externími pozicemi SNP

Vystupy – nahradit za cestu ke složce, kam program uloží výstupy

Výstupní soubory

Výstupem programu je složka PripojeniPozicSNP, která obsahuje upravené původní soubory s genotypem (přidané či přepsané proměnné Chr a Position).

PxSNP

PxSNP je hlavním programem pro přípravu genotypů pro imputaci v programu FImpute. Před jeho použitím doporučujeme spustit na požadované souboru opět program KontrolaVstupu PxSNP. Pokud soubory obsahují všechny požadované parametry, dojde k jejich úpravě ve dvou krocích. Nejprve jsou získány informace o SNP a je vytvořen mapový soubor SNP a poté jsou získány informace o genotypech zvířat pro vytvoření souboru s genotypem. Oba tyto soubory jsou zároveň vstupními soubory pro program FImpute. Genotyp zvířat je překódován dle následujících pravidel AA – 1, AB, BA – 2, BB – 3 a 5 pro chybějící SNP.

Vstupní soubory

- složka se soubory s genotypy, které obsahují požadované parametry (viz KontrolaVstupuSNP)
- případně textový soubor ok_vstupy.txt (výstup z KontrolaVstupuPxSNP) obsahující seznam souborů s genotypem obsahující požadované parametry

Parametrový soubor

```
<?xml version="1.0" encoding="utf-8" ?>
<root>

  <!-- Vstupy -->
  <pouzijOKVstupy>true</pouzijOKVstupy>
  <!-- Pri "true" se pouzije "cestaKSeznamuOKVstupu", pri "false" se pouzije "cestaKeSlozceVstupnichGenotypu". -->
  <cestaKSeznamuOKVstupu>TestovaciData/ok_vstupy.txt</cestaKSeznamuOKVstupu>
  <cestaKeSlozceVstupnichGenotypu>TestovaciData/genotypy</cestaKeSlozceVstupnichGenotypu>

  <!-- Vystupy -->
  <cestaKeSlozceVystupu>Vystupy</cestaKeSlozceVystupu>

</root>
```

Parametr pouzijOKVstupy – nastavení hodnotu true pro použití seznamu souborů s genotypem vygenerovaným programem KontrolaVstupuPxSNP, v případě, že je nastaven parametr true je

nutné za TestovacíData/ok_vstupy.txt zadat cestu k souboru ok_vstupy.txt. Pokud nechcete použít výše zmíněný soubor, napište do pouzijOKVstupy false a následně nahraďte TestovacíData/genotypy cestou ke složce se soubory s genotypem.

Vystupy – nahradit cestou ke složce kam program uloží výstupní soubory

Výstupní soubory

celkovySeznamSNP.txt - vytvořený mapový soubor obsahuje čtyři základní proměnné: SNP_ID – jedná se o název daného SNP, Chr – značí umístění SNP na chromozomu, Pos – genomická pozice SNP a poslední proměnnou Chip (1-10), kde Chip1 představuje referenční čip, který je nejhustší, a na který chceme imputovat. Chip2 až Chip10 (lze použít maximálně 10 různých čipů) odkazují na pozice SNP vůči referenčnímu čipu. Jedná se o vstupní soubor pro FImpute.

SNP_ID	Chr	Pos	Chip1	Chip2	Chip3	Chip4
SNP4	1	6133529	1	1	1	1
SNP1	1	6371334	2	2	2	2
SNP6	1	6616434	3	0	0	0
SNP2	1	7928189	4	3	3	3

genotyp.txt – vytvořený soubor s genotypem , kde ID – představuje označení zvířete, Chip – jedná se o číslo čipu, na kterém je dané zvíře genotypováno a slouží jako odkaz k mapovému souboru. Poslední proměnnou je Call... – v této proměnné nalezneme celý genotyp zvířete podle pořadí SNP. Jedná se o vstupní soubor pro FImpute.

ID	Chip	Call...
zvire1	1	122220211022022200002222212010222011121212
zvire10	3	122212202222201201122000020022010211102
zvire2	2	1222122022222012011220000200022010211102
zvire3	2	1222122022122012011210000200012010211002
zvire4	2	1222122012222012011220000100022010211102
zvire5	4	122212222010112000200022010211102

output_priprava_pro_fimpute.txt – v souboru je vypsán postup úpravy souborů po jednotlivých čipech. Následně je zde vypsán frekvence genotypů 0, 1, 2 a 5 a počet genotypovaných zvířat.

output_seznamy_snp.txt – v souboru je vypsán proces úpravy jednotlivých souborů s genotypem. Při správném průběhu je zakončen řádkem Hotovo.

prehled_o_cipech.txt – obsahuje názvy čipů použitých v databázi, počet SNP a počet genotypovaných zvířat,

prehled_o_cipech_vypis.txt – při správném průběhu je zakončen řádkem Hotovo.

prehledZvirat.txt – obsahuje výpis genotypovaných zvířat s informacemi o počtu SNP a čipu na kterém je zvíře genotypováno.

problemovaSNP.txt – obsahuje výpis SNP z mapového souboru, jejichž pozice na chromozomu a pozice se opakují u jiných SNP. Zároveň poukazují na pozici SNP na daných čipech.

seznam_snp.txt – obsahuje seznam SNP včetně proměnné Chr a Position

Zpětný převod

Zpětný převod slouží pro úpravu souboru s genotypem po imputaci v softwaru FImpute. Genotyp lze převést na původní podobu, kde každé SNP zvířete je na jednom řádku, nebo na vstupní soubor pro program BLUPf90.

Vstupní soubory

- genotyp.txt – po imputaci
- celkovySeznamSNP.txt

Parametrový soubor

```
<?xml version="1.0" encoding="utf-8" ?>
<root>

  <cestaKeGenotypu>C:\testovaci-slozka\genotyp.txt</cestaKeGenotypu>
  <cestaKMapovemuSouboru>C:\testovaci-slozka\celkovySeznamSNP.txt</cestaKMapovemuSouboru>
  <cestaKeSlozceVystupu>C:\testovaci-slozka\vystupy</cestaKeSlozceVystupu>

  <provedNaPuvodniFormat>true</provedNaPuvodniFormat>
  <provedNaFormatProBLUPf90>true</provedNaFormatProBLUPf90>

</root>
```

C:\testovaci-slozka\genotyp.txt – nahradit za cestu k imputovanému genotypu

C:\testovaci-slozka\celkovySeznamSNP.txt – nahradit za cestu k SNP mapovému souboru

C:\testovaci-slozka\vystupy – nahradit za cestu ke složce výstupů

provedNaPuvodniFormat – true – provede převod, false – nevytváří výstup

provedNaFormatProBLUPf90 – true – provede převod, false – nevytváří výstup

Výstupní soubory

Při vytvoření BLUPf90 formátu:

celkovySeznamSNP_formatProBLUPf90.txt – upravený SNP mapový soubor pro BLUPf90

Index	SNP_Name	SNP_Name_2	SNP_Name_3
1	SNP-4	SNP4	SNP-4
2	SNP-1	SNP1	SNP-1
3	SNP-6	SNP6	SNP-6
4	SNP-2	SNP2	SNP-2
5	SNP-7	SNP7	SNP-7

genotyp_formatProBLUPf90.txt – upravený soubor s genotypem pro BLUPf90

ID	Call...
zvire1	122220211022022200022222120102220111212;
zvire10	122212202222201201122000020022010211102;
zvire2	122212202222201201122000020002201021110;
zvire3	122212202212201201121000020001201021100;
zvire4	122212201222201201122000010002201021110;
zvire5	122212222010112000200022010211102021202;
zvire6	122212202222201201122000020002201021110;

Při vytvoření původního formátu:

genotyp_puvodniFormat.txt – upravený soubor s genotypem do původního formátu

Sample ID	SNP Name	Chr	Position	Allele1 - AB	Allele2 - AB
zvire1	SNP-4	1	6133529	A	B
zvire1	SNP-1	1	6371334	B	B
zvire1	SNP-6	1	6616434	B	B
zvire1	SNP-2	1	7928189	B	B
zvire1	SNP-7	1	17544926	B	B
zvire1	SNP-10	1	17911906	A	A

vystup.txt – obsahuje informace o průběhu programu. Při správném průběhu je zakončen řádkem Hotovo.

Kontrola rodokmenu

Samostatný program pro kontrolu rodokmenu, který je potřebný v případě použití rodokmenové nebo kombinované imputace v FImpute. Dokáže najít zacyklená zvířata a vypsát je. Uživatel pak podle svého uvážení může rodokmen upravit.

Vstupní soubory

- soubor s rodokmenem se třemi proměnnými v tomto pořadí ID (označení zvířete), Sire (označení otce) a Dam (označení matky)

Spuštění

- program se spouští bez parametrového souboru, ale je nutné předat cestu k rodokmenu jako parametr příkazové řádky, např. „DetekceCykluVRodkmenu.exe vstupy/rodokmen.txt“.

Výstupní soubory

vystup.txt – v případě, že rodokmen obsahuje cykly (jednoduché nebo složité) dojde k vypsání konkrétních zvířat. Při správném průběhu je zakončen řádkem Hotovo.

SIMPOT

Část SIMPUT je nezávislou částí balíčku PxSNP v1.0. Jedná se o program, který jako jediný lze spustit jen v operačním systému Linux, protože pro jeho správné fungování je nutné mít staženou a nainstalovanou funkční verzi imputačního programu FImpute. Stažení programu je možné na webové adrese <http://animalbiosciences.uoguelph.ca/~msargol/fimpute/>. SIMPUT provádí stochastické simulace pro ověření přesnosti imputace. Sledovanými faktory jsou zde velikost referenční populace, tedy populace na čipu, na který imputujeme a druhým sledovaným faktorem je množství chybějících markerů. Výběr zvířat do referenční populace, lze vybírat náhodně ve stanovených hladinách nebo cíleně podle seznamu zvířat, obě tyto možnosti je nutné vyplnit v parametrovém souboru. Množství chybějících SNP lze opět zvolit náhodně v hladinách nebo cíleně pomocí seznamu SNP, podle kterého dojde k maskování genotypu na chybějící hodnotu. Výběr SNP k maskování je opět nutné nastavit v parametrovém souboru. Vzhledem k náhodnému výběru SNP či zvířat do referenční populace je možné nastavit počet opakování pro výpočet průměrných parametrů přesnosti imputace. Příklad použití tohoto programu lze najít v článku:

Kranjčevićová, Anita, Kašná, Eva, Brzáková, Michaela, Příbyl, Josef a Vostrý, Luboš. Impact of reference population and marker density on accuracy of population imputation. Czech Journal of Animal Science, 2019, 64, 405-410. ISSN 1212-1819.

Vstupní soubory

- vstupními soubory jsou vstupní soubory pro FImpute, vytvořené programem PxSNP
celkovySeznamSNP.txt
genotyp.txt
- v případě určení konkrétních zvířat do referenční populace použijte textový soubor se seznamem zvířat

- v případě určení konkrétních SNP pro maskování použijte textový soubor se seznamem SNP

Parametrový soubor

```
<?xml version="1.0" encoding="utf-8" ?>
<root>

  <variantaVyberuReferencniPopulace>ZeSouboru</variantaVyberuReferencniPopulace>
  <!--Jedna z hodnot: Procenta, ZeSouboru nebo PocetZvirat-->
  <souborReferencniPopulace>C:\testovaci-slozka\refPopulace.txt</souborReferencniPopulace>
  <procentaReferencniPopulace>25, 50, 75</procentaReferencniPopulace> <!--Cisla oddelena carkou-->
  <pocetyReferencnichZvirat>100</pocetyReferencnichZvirat> <!--Cisla oddelena carkou-->

  <variantaMaskovaniSnp>ZeSouboru</variantaMaskovaniSnp> <!--Jedna z hodnot: ProcentaKeSmazani nebo ZeSouboru-->
  <souborSnpKeSmazani>C:\testovaci-slozka\snpKeSmazani.txt</souborSnpKeSmazani>
  <procentaSnpKeSmazani>15, 30, 45, 60, 75, 90</procentaSnpKeSmazani> <!--Cisla oddelena carkou-->

  <pocetIteraci>10</pocetIteraci>

  <cestaKSouborum>/home/testovaci-slozka/soubory</cestaKSouborum>
  <nazevParametrovehoSouboruFImpute>control_file.ctr</nazevParametrovehoSouboruFImpute>

  <cestaKFImpute>/home/testovaci-slozka/SIMPOT</cestaKFImpute>

</root>
```

variantaVyberuReferencniPopulace – vybrat parametr, kterým se bude vytvářet referenční populace, a to Procenta (zadáni procent v příkladu místo 25, 50, 75), ZeSouboru, kdy se místo C:\testovaci-slozka\refPopulace.txt vloží cesta k textovému souboru se seznamem zvířat a poslední možností je PocetZvirat, kdy se zadá konkrétní počet zvířat, který bude náhodně vybrán (v příkladu místo 100 zadat požadovaný počet zvířat).

variantaMaskovaniSnp – vybrat parametr, podle kterého se budou vybírat SNP k maskování, a to ProcentaKeSmazani (zadáni procent SNP k maskování, v příkladu nahradit 15, 30, 45, 60... za požadovaná procenta), nebo ZeSouboru, kde se C:\testovaci-slozka\snpKeSmazani.txt nahradí cestou k souboru se seznamem SNP.

pocetIteraci – lze zadat počet opakování simulace

cestaKSouborum – nahradit cestou k souborům celkovySeznamSNP.txt a genotyp.txt, případně k souboru s rodokmenem. Zároveň slouží jako cesta do složky, kde se vytvoří složka s výstupy.

nazevParametrovehoSouboruFImpute – vložit název parametrového souboru ke spuštění vybrané metody imputace v programu FImpute

cestaKFImpute – cesta do složky, která obsahuje spustitelný program FImpute

Výstupní soubory

Hlavními výstupy SIMPUT jsou soubory s korelací a s percentilem, v závislosti na zvolené velikosti referenční populace a počtu maskovaných SNP. Oba soubory obsahují tolik řádků s hodnotami, kolik je nastaveno iterací v parametrovém souboru. Zároveň jsou vytvářeny mezivýstupy jako je podoba genotypvého souboru pro danou referenční populaci a procento chybějících SNP pro danou iteraci. S každou provedou iterací, se tyto soubory promazávají.